Follow-Up Questions Improve Generative AI Output and User Experience

Working Towards a Collaborative Model of Human-AI Interaction

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE
UNIVERSITY OF HAWAIʻI AT MĀNOA IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

COMPUTER SCIENCE

November 2024

By

Bernadette J Tix

Dissertation Committee:

Kim Binsted, Chairperson
Martha Crosby
Nurit Kirshenbaum
Peter Yiğitcan Washington
Michael C. Liu

Keywords: Natural Language Processing, Large Language Models, Generative AI, Human-Computer Interaction, Human-AI Collaboration, AI Questions

© COPYRIGHT BERNADETTE J TIX 2024

Dedicated with love to my wife, Carolyn, for her boundless patience and support throughout my research and graduate school journey.

Abstract

This research investigates the impact of Large Language Models (LLMs) generating follow-up questions in response to user requests for short (1-page) text documents. This dissertation will argue that there are clear benefits to LLMs asking follow-up questions, and engaging users in thought-provoking and context-clarifying dialog before producing documents and other outputs. Two experiments support this research, including a pilot study and a larger full study with an improved design based on insights from the pilot study. In both experiments, users interacted with a novel web-based AI system designed to ask follow-up questions. Users requested documents they would like the AI to produce. The AI then generated follow-up questions to clarify the user's needs or offer additional insights before generating the requested documents. After answering the questions, users were shown a document generated using both the initial request and the questions and answers, and a document generated using only the initial request. Users indicated which document they preferred and gave feedback about their experience with the question-answering process. The findings of these experiments show clear benefits to question-asking both in document preference and in the qualitative user experience, and further show that users found more value in questions which were thought-provoking, open-ended, or offered unique insights into the user's request as opposed to simple information-gathering questions. These results point to the need to incorporate follow-up questions and collaborative dialog into LLMs as a part of the human / AI interaction experience.

Table of Contents

ABSTRACT 3

TABLE OF CONTENTS 4

LIST OF TABLES 7

FIGURES 7

LIST OF ABBREVIATIONS 8

CHAPTER 1: INTRODUCTION 9

- RESEARCH GOALS 13
- UNIQUE CONTRIBUTIONS 14

CHAPTER 2: BACKGROUND 15

- HISTORICAL PERSPECTIVE: BEFORE LLMS 15
- THE TRANSFORMER ARCHITECTURE & LARGE LANGUAGE MODELS 24
- LIMITATIONS OF LARGE LANGUAGE MODELS 26
  - *Hallucinations* *26*
  - *Biased, Toxic, or Dangerous Outputs* *27*
  - *Limited Capacity for Analytical Reasoning* *28*
- HOW TO IMPROVE LARGE LANGUAGE MODELS 29
  - *Prompt Engineering and Chain of Thought Reasoning* *29*
  - *Asking Questions* *30*
- SOCIETAL RISKS AND IMPACTS FROM LARGE LANGUAGE MODELS 32
  - *Overreliance on Unreliable Outputs* *32*
  - *Plagiarism and Cheating* *32*
  - *Social Disruption from Automation* *34*
  - *Positive Impacts of Generative AI* *34*
  - *Context-Dependence and the Inevitability of Ambiguity* *36*
  - *How Asking Questions Can Help* *37*
- BENCHMARKS AND EVALUATION 38

CHAPTER 3: PILOT STUDY 41

- PARTICIPANT SELECTION 41
- PILOT STUDY SETUP 42
- PILOT STUDY TECHNICAL DESIGN 45
  - *Pilot Study Limitations* *48*
- PILOT STUDY RESULTS 49
  - *Document Ratings* *49*
  - *Exit Survey* *49*
  - *User Feedback* *50*
  - *Research Notes* *52*
- PILOT STUDY DISCUSSION 53

CHAPTER 4: FULL STUDY 56

- IMPROVEMENTS TO STUDY DESIGN 56
- IMPROVEMENTS TO TECHNICAL DESIGN 58

Inclusion of Multiple LLMs *58*
Improved Prompt Engineering *59*
A More Robust Database Connection *63*
PARTICIPANT SELECTION 63
FULL STUDY RESULTS 66
Prompts Entered *66*
Document Preferences *68*
Exit Survey *76*
User Feedback *77*

CHAPTER 5: DISCUSSION 82

WHEN IS ASKING QUESTIONS BENEFICIAL? 82
Use of higher quality LLMs, especially GPT-4. .... *82*
Asking the right kinds of question. *83*
Question Asking was more beneficial to some users than others. *83*
QUALITATIVE BENEFITS OF QUESTION ASKING 85
QUALITIES OF AN IDEAL QUESTION ASKING DOCUMENT GENERATOR. 86
COST AND EFFICIENCY OF QUESTION ASKING 88
RECENT DEVELOPMENTS: GPT-O1 88

CHAPTER 6: LIMITATIONS AND FUTURE WORK 91

LIMITATIONS 91
LESSONS LEARNED 94
FUTURE WORK. 95

CHAPTER 8: CONCLUSION 96

APPENDIX A: CHATGPT ON TRANSFORMERS 98

APPENDIX B: WORDNET 3.1 SENSES OF "LEAVES" 99

APPENDIX C: PILOT STUDY SCREENSHOTS ALL STEPS 101

APPENDIX D: PILOT STUDY CONSENT MESSAGE 109

APPENDIX E: PILOT STUDY EXAMPLE CONVERSATION 112

THE CONVERSATION AS IT WOULD APPEAR TO THE USER 112
FULL PROMPT EXCHANGE WITH GPT-3.5 114

APPENDIX F: FULL STUDY SCREENSHOTS ALL STEPS 119

APPENDIX G: FULL STUDY CONSENT MESSAGE 128

APPENDIX H: FULL STUDY EXAMPLE CONVERSATION 131

THE CONVERSATION AS IT WOULD APPEAR TO THE USER 131
FULL PROMPT EXCHANGE WITH GPT-4 136

APPENDIX I: FULL STUDY USER FEEDBACK 143

FEEDBACK FROM *PART 4 OF 9: QUESTIONS AND ANSWERS* 143
FEEDBACK ON THE QA DOCUMENT 145
FEEDBACK ON THE BASELINE DOCUMENT 148
FEEDBACK ON THE QA DOCUMENT AFTER ADDITIONAL USER REFINEMENT 151

FEEDBACK ON THE BASELINE DOCUMENT AFTER ADDITIONAL USER REFINEMENT 152
FEEDBACK AFTER COMPLETING THE EXIT SURVEY 154

APPENDIX J: GPT-O1 ON TRANSFORMERS 156

APPENDIX K: GPT-O1 COVER LETTER 159

APPENDIX L: BUDGET 162

COST DETAILS BY PHASE 163
Pilot Development 163
Pilot Study 163
Full Development 163
Full Study 163
Analysis 164
COST DETAILS BY SERVICE 164
Azure 164
OpenAI 164
Google Cloud / Gemini API 165
Meta Ads 165

BIBLIOGRAPHY 166

List of Tables

TABLE 1: PROMPT ENGINEERING CHANGES BETWEEN PILOT STUDY AND FULL STUDY 60
TABLE 2: PARTICIPANT EXPERIENCE, AGE, AND GENDER DISTRIBUTION 65
TABLE 3: PREFERENCE RESULTS 68
TABLE 4: BINOMIAL DISTRIBUTION TEST OF SIGNIFICANCE (SIGNIFICANT RESULTS ARE HIGHLIGHTED) 69
TABLE 5: DOCUMENTS PRODUCED BY AGE CATEGORY 74
TABLE 6: TOP 5 THEMES FOR EACH CATEGORY OF FREE-ENTRY FEEDBACK 80
TABLE 7: THOUGH PROVOKING QUESTION RATING BY LLM 83
TABLE 8: PARTICIPANT EXPERIENCE, AGE, AND GENDER DISTRIBUTION 84
TABLE 9: PROJECT EXPENSES (TOTAL EXPENSE FOR THE PROJECT IS HIGHLIGHTED) 162

Figures

FIGURE 1: PARSE TREE OF THE PHRASE “I ATE AN APPLE.” 16
FIGURE 2: POSSIBLE PARSE TREES FOR "FALL LEAVES FALL AND SPRING LEAVES SPRING." 18
FIGURE 3: SEMANTIC DOMAINS WITH SAMPLE FRAMES AND PREDICATES 20
FIGURE 4: SAMPLE FRAMES FROM FRAMENET 21
FIGURE 5: EXAMPLE OF AMBIGUOUS ENTITY NAMES 22
FIGURE 6: SAMPLE CONTEXT AND MISSING WORD QUESTION 24
FIGURE 7: THE SELECTIVE CLARIFICATION FRAMEWORK FROM CLAM 31
FIGURE 8: PILOT STUDY TECHNICAL DESIGN 45
FIGURE 9: CQDG PROMPT AND RESPONSE PIPELINE 46
FIGURE 10: AVERAGE DOCUMENT RATINGS GIVEN BY PARTICIPANTS IN THE PILOT STUDY. 49
FIGURE 11: EXIT SURVEY RESULTS FROM PILOT STUDY. 50
FIGURE 12: FULL STUDY TECHNICAL DESIGN 59
FIGURE 13: FULL STUDY DOCUMENT GENERATION PIPELINE 62
FIGURE 14: IMPACT OF LLM ON USER PREFERENCE 69
FIGURE 15: IMPACT OF GENDER ON DOCUMENT PREFERENCE 71
FIGURE 16: IMPACT OF PRIOR EXPERIENCE WITH AI ON DOCUMENT PREFERENCE 73
FIGURE 17: IMPACT OF AGE ON DOCUMENT PREFERENCE 75
FIGURE 18: EXIT SURVEY RESPONSES 76

List of Abbreviations

AI	Artificial Intelligence
CLAM	Clarify if Ambiguous Framework
CQDG	Clarifying Questions Document Generator (original software for this research)
COT	Chain of Thought Prompting
GPT	Generative Pretrained Transformer
LLM	Large Language Model
LSTM	Large Short-Term Memory Network
MemNN	Memory Neural Network
MemN2N	Recurrent Memory Neural Network
NER	Named Entity Recognition
NLP	Natural Language Processing
QA Document	The document generated by CQDG which utilized the users' question responses.
RNN	Recurrent Neural Network

Chapter 1: Introduction

Advances in generative AI have made it possible for a software program to produce a broad range of useful output from natural language prompts, including visual artwork [1], music [2], working software code [3, 4], and text [5]. This has led to a rapid increase in the use of generative AI in the daily lives of millions of people around the world, with unclear but potentially substantial implications for the future of work [6], automation [7], teaching [8], and learning [9–11].

Ambiguity has historically been a major problem in **Natural Language Processing (NLP)** and continues to present major obstacles even for modern systems, with some researchers concerned that the problem of ambiguity is not being taken seriously enough by the AI community writ large [12]. Ambiguity has historically caused difficulties in parsing any natural language request into actionable software output [13–16].

This pervasive ambiguity can cause confusion in human communication as well. When developing new software, for example, an active and involved process of requirements gathering is typically necessary before the development team can begin their work [17]. One of the most straightforward ways of overcoming miscommunications is simply to ask questions [18, 19]. However, the most widely available **Larg Language Models (LLMs)** do not, by default, ask follow-up questions in response to confusing or ambiguous prompts. Instead, publicly available models including ChatGPT [20], Gemini [21], and Bing [22] will attempt to fulfil requests from the user with whatever information they have been given.

This is not to say that LLMs are incapable of generating useful questions. When specifically prompted to do so, LLMs can generate relevant questions and produce improved output in response to those questions [23–25]. Previous work in this area has focused on disambiguation for short questions [23] and simple task requests [25], with the primary goal of achieving complete disambiguation of the user request [24]. Prior work has also shown that LLMs can produce more useful output when given access to the full conversation that led up to the user query [26].

Questions and dialog in response to a user prompt are capable of much more than disambiguation. LLMs can be trained to ask questions in educational settings to act as a tutor [9, 27] and to generate assessment questions for teachers [28]. LLMs are also capable of assisting with organizing, outlining, and other tasks related to writing [29, 30] and can even improve students' critical thinking skills when employed properly [11]. All of this points to a broad conclusion: AI produces better results, as well as better outcomes for users, when users are actively and critically engaged with the AI system rather than passively accepting AI outputs.

LLMs generate output by calculating a “most likely" response to any given prompt. However, prompts are often ambiguous, and even the best possible prediction cannot fully resolve the problem of underspecified prompts [31]. Even in a conversation between two humans, both speaking the same language and communicating clearly, misunderstandings are common, as is the use of ambiguous language. While there are many ways to resolve ambiguity in a human conversation, perhaps the most obvious way is to simply ask for clarification. However, commonly used LLM systems such as ChatGPT, Bard, and Bing do not ask clarifying questions in response to ambiguous prompts.

This is a serious problem in fields where precision is important. Detailed discussions, including follow-up questions, are a necessary part of human communication when precision is required. Even for simple requests, a lack of follow-up questions can lead to suboptimal answers, or even cause the LLM to misunderstand the true needs of the user. In some applications, such as LLMs being used as search tools, it may be acceptable for the user to enter increasingly refined prompts to improve the answer they get from the LLM. However, the user may not know what they need to change about the prompt to get the information they need, or, worse, the user may not realize the system has misunderstood their needs. If the user cannot distinguish between correct and incorrect output and is relying on the system to be correct, an ambiguous prompt may lead to the user relying on misleading output. Even in situations where users are knowledgeable and free to refine their prompts as needed, a smoother experience could likely be provided if the system itself recognized points of ambiguity and asked clarifying questions. **Appendix A** shows an example of ChatGPT providing an answer to a question that could have multiple possible meanings. However, the answer generated by ChatGPT assumes one meaning and proceeds without asking for clarification or pointing out the ambiguity. This is in spite of the fact that ChatGPT was capable of identifying the ambiguity when prompted to do so.

Ambiguity in user requests is also a problem for AI Alignment. AI Agents acting to solve under-specified problems can lead to severe unintended side effects [31, 32]. Under-specification occurs when an AI system is given a goal to accomplish but the designer of the system has unstated assumptions of how the goal ought to be accomplished that are not communicated to the system. As an example, one of the early experiments by OpenAI trained an artificial intelligence to play a boat-racing video game. The agent discovered a strategy in which a simulated boat under the agent's control remained in a small area of the map, never completed the course, ran into other boats, and repeatedly caught fire. However, due to the scoring system of the game it was still able to achieve more points than most humans could by running the course normally [32, 33]. It was given the goal of maximizing points, with no incentive to actually race the course when other means of acquiring points were available. Although it did achieve a high number of points, this was clearly not the intended behavior.

While question-asking will not completely eliminate this problem, it can help to alleviate it. An ideal system, linked to an LLM, might have asked "do you want me to race with the other boats, or is maximizing points the only goal?" Since it seems likely that LLMs will be increasingly used to write code, guide human decisions, and in some cases even operate completely autonomously, it is vital that ambiguity and under-specification be identified early, and this goal could be substantially aided by the AI itself identifying ambiguity and asking clarifying questions.

This research investigates whether there is value in an LLM asking follow-up questions when prompted to produce a short document such as a letter, memo, email, or a short report. This task is inherently more ambiguous than answering a short question or performing a simple task, as there is no single "correct" document. Participants in this study prompted a generative AI to produce a document they desired, answered AI-generated follow-up questions about their needs, and then compared and rated a pair of documents. One document was generated taking the user's questions and answers into consideration, and the other document was generated based on only the user's original request. Participants also gave qualitative feedback about the experience and answered an exit survey targeted at determining whether there was value in the question-answering process itself, apart from final document preference.

The results of these experiments show that AI-generated follow up questions not only improved user-rated document quality, but also improved users subjective experience of working with the system and made users feel more engaged with the AI. Participants in the study consistently stated a preference for questions which were creative and thought-provoking, and found such questions to be valuable on their own even before seeing the documents that were created from the process.

This research further shows that higher quality LLMs, especially GPT-4, are able to make better use of the answers participants provided to AI-generated questions.

This research reinforces the results of several previous studies in showing that generative AI works best not as a replacement for humans, but as an assistive tool operating in a human-AI collaborative framework in which context and dialog are vital to producing good outcomes [11, 23, 26]. The original contribution of this work is demonstrating the effectiveness of and user preference for follow up questions in the open-ended and subjectively-rated task of document generation, using actual human participants in contrast to several previous studies which have used AI trained to act as “simulated humans” to validate AI systems asking simpler disambiguation questions [23–25].

It is my hope that this research will assist in guiding future development of LLM-based AI systems to take an approach of human-AI collaboration, with follow up questions and thoughtful dialog between people and AI, in contrast to the current norm of users entering prompts or commands which the AI must attempt to satisfy without any further engagement. I believe this approach has the potential to produce better results overall, to improve AI alignment with users’ needs, increase users’ engagement with their work, encourage people to think critically about what they need and what they are asking for, and improve the overall user experience.

Research Goals

The motivating question for this research was ***"Do Large Language Models produce better short documents from user prompts when the LLM is prompted to ask follow-up questions to the user before producing the final output?"***

An additional question of interest is: *"Do LLM-generated follow-up questions improve user's subjective experience while working with the LLM?"*

Follow-up questions generated by an LLM may be valuable in areas besides short document generation, such as code generation, image generation, search, or chat. However, to study all of these areas would be beyond the scope of what is feasible within a single study. The particular area of document generation was chosen for several reasons:

- Focusing on document generation limits the scope of the research to a question that can be feasibly answered.
- Document generation is one of the key novel capabilities of LLMs when compared to older NLP technologies.
- Document generation is a widely applicable need, as nearly all people need to generate short documents of one form or another at various times.
- Short documents can be conveniently displayed, read, and evaluated through a web interface.

A two-phase approach was taken in this research. First, a small pilot study was conducted for the purpose of refining the study methodology. An improved study was then designed utilizing the insights from the pilot study. The purpose of the pilot was only to gain insights relevant to improving the design of the full study. The full study was then distributed to a larger set of participants with the hope of gaining statistically significant insights. Details of both studies are provided in the following sections.

During the pilot study, several users commented that they found intrinsic value in the question-answering process, regardless of the final document quality. This insight broadened the scope of the full study, introducing a secondary question of interest: "*Does the question-asking process add value to the user experience independent of final document quality?"* This research attempts to answer both of these questions through a combination of quantitative data (ex: user preference ratings and survey results) and qualitative data (ex: free-text user feedback) gathered from user interactions with an online system designed to ask follow-up questions before generating requested documents, all through the use of an LLM.

Unique Contributions

The unique contributions of this research include:

- A literature review providing historical perspective on the pervasive problems of context and ambiguity across NLP applications, problems that need to be taken seriously when developing and assessing LLMs.
- A novel system design demonstrating how LLMs can be made to ask thought-provoking questions and generate documents that consider a user’s original request, as well as their answers to generated questions.
- A pilot study offering guidance on design considerations for future research into question-asking. In particular, the pilot study shows that users value open-ended questions that make them think about their requests in ways they had not previously considered.
- A full study which demonstrates conclusively that LLMs can be made to ask questions which users find valuable, and that answering these questions results in higher-quality documents generated by the LLM.
- Both the pilot study and the full study used only real human participants, in contrast to many previous studies into question-asking LLMs that rely heavily on AI-simulated “virtual humans” to validate their efficacy. [23, 25, 34]

Chapter 2: Background

Historical Perspective: Before LLMs

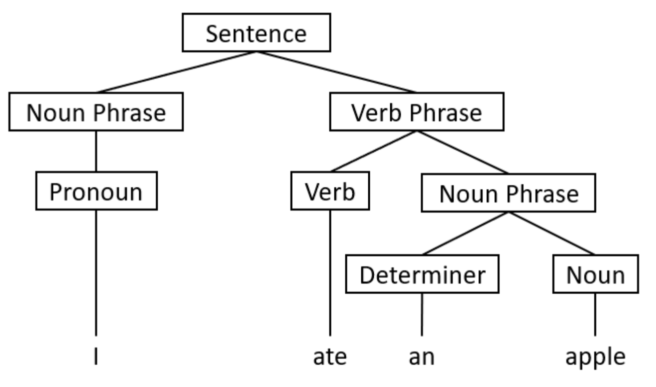
Natural Language Processing (NLP) refers to automatic parsing of natural language to produce some useful output through software. Natural language means language as it is normally spoken or written by people in ordinary interpersonal communication, as opposed to formal languages or programming languages. NLP was already a rich and varied field before the introduction of LLMs, and it is worth considering some of the strengths and limitations of the technologies that came before LLMs, both to benefit from knowledge of past accomplishments and to avoid repeating the same mistakes. Specifically, this historical background will demonstrate that the issues of context and ambiguity are pervasive throughout the history of NLP applications. The effectiveness of numerous historical NLP systems has suffered when context and ambiguity are not considered as a core part of the language interpretation process. Designing AI systems to ask questions is a vital step for LLM development in order to address this problem. The historical section of this lit review will demonstrate the pervasiveness of ambiguity and context-dependent meaning within NLP applications, and present various approaches aside from asking questions which have been used to address these issues in the past.

Grammar

Some of the earliest attempts at NLP relied heavily on the idea of formal grammar. A **formal language** is a language with clearly defined rules and structure [35]. In a formal language, each sentence follows a clearly defined **grammar**, and this structure usually helps to define the meaning of the sentence. Programming languages are examples of formal languages. In most programming languages, any line of code has only one correct interpretation. This is a necessary restriction on programming languages, so that commands written in a human-readable language like C, Java, Prolog, or SQL can be converted into processor-level commands and executed. Most programming language grammars belong to the family known as **Context Free Grammars (CFG)**, but this is not the only family of grammar, nor is it the most powerful family of formal grammars. According to the **Chomsky Hierarchy**, grammars can be divided into one of four classes, arranged by their generative capacity, where each class has the power to describe all languages described by any less powerful class, as well as some additional languages [35, 36]. The four classes are:

- **Recursively Enumerable Grammars** have unrestricted rules. Both sides of any rule can have any number of terminal and nonterminal symbols. *Example: A B C -> D E* [1]
- **Context-Sensitive Grammars** require that the right side of the rule contains at least as many symbols as the left side. *Example: A X B -> A Y B*
- **Context-Free Grammars (CFG)** require that the left side of the rule consists of a single non-terminal symbol. CFGs are popular for both natural-language and programming-language grammars [35]. *Example: A -> B C D*
- **Regular Grammars** consist of a single non-terminal symbol on the left side of each rule, and a single terminal symbol optionally followed by a single non-terminal symbol on the right. *Example: A -> b C*

Figure 1: Parse Tree of the Phrase "I ate an apple."



Many English sentences can be interpreted using a grammar. The result of interpreting a sentence with a set of formal grammar rules is a **Parse Tree** which defines how each word in the sentence fits together structurally. **Figure 1** shows a parse tree for the simple sentence "I ate an apple." This parse tree could have been generated from a CFG, but not from a regular grammar, since the *Sentence*, *Noun Phrase*, and *Verb Phrase* symbols all match sequences of multiple non-terminal symbols[2].

In formal language, grammar is **prescriptive**, it defines what statements are or are not legal within the language. In natural language, grammar is merely **descriptive**, an attempt to recognize patterns within organically occurring speech and writing. Actual speech and writing rarely conforms strictly

---

[1] In these examples, capital letters represent non-terminal symbols (*ex: A)* and lowercase letters represent terminal symbols *(ex: a)*. In this case, 'terminal' does not refer to the symbol ending the sequence, but rather that any non-terminal symbol can be resolved into a sequence of terminal symbols by the application of one or more rules within the grammar (*Ex: A -> BC; B -> b C; C -> c;* resolves to *A -> b c c)*.

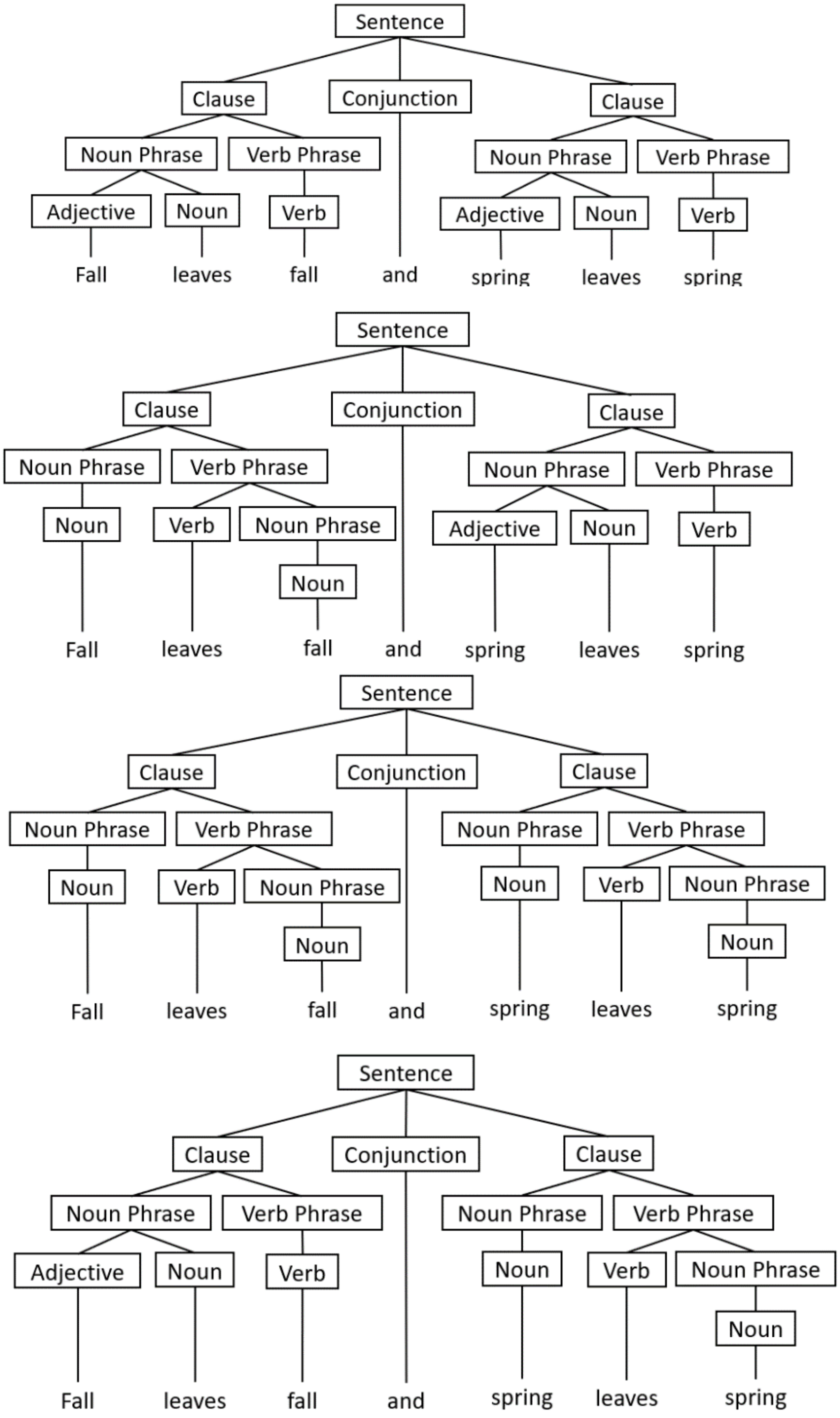
[2] In Figure 1 and Figure 2, non-terminal symbols are outlined (ex: "Noun Phrase" "Sentence") and terminal symbols are not outlined (ex: "I" "ate" "an" "apple")

to any particular set of grammatical rules. Formal languages are often designed to limit or entirely eliminate ambiguity, as is the case for programming languages. Natural language, on the other hand, is filled with both intentional and unintentional ambiguity, to such an extent that the ambiguity cannot be easily isolated for most NLP tasks.

Consider the following sentence: *"Fall leaves fall and spring leaves spring."* [35] In this sentence, every word except for "and" has multiple interpretations. Even the parts of speech are unclear. "Leaves" could be a verb or a noun, or perhaps a verb in the first instance and a noun in the second. Any grammar capable of representing each of these possible meanings would require multiple possible parse trees. An example with several possible parses of this sentence is shown in **Figure 2**. There is no clear way to choose between those interpretations, at least not from the grammar itself.

Tools and libraries intended to assist with NLP tasks have to take this ambiguity into account. For instance, **WordNet** [37] defines for each word a list of **Senses**, each representing one possible use of the word. WordNet 3.1 defines 6 noun senses and 14 verb senses for the word "leaves" (See **Appendix B**). This can also be seen in ordinary dictionaries, which define multiple meanings for most words.

Figure 2: Possible Parse Trees for "Fall leaves fall and spring leaves spring."



Beyond Grammar

There have been many attempts mitigate the problem of ambiguous parse trees through alternative representations of meaning within natural language. One example is **Role and Reference Grammar (RRG)** [13], a linguistic theory of clause construction across multiple languages. RRG originated as a divergence from earlier theories which were overly focused on English at the expense of grammatical structures found in other languages. It eschews standard formats for explaining clause structure, since syntax varies from language to language and thus any model based on a specific clause structure will necessarily impose some of the syntax of whatever language the model originated from. Instead, RRG defines clauses in terms of a layered structure, including a nucleus, which includes the predicates of the clause, a core, which includes the nucleus plus the arguments of the predicates, and the periphery, which includes modifiers to the core. RRG claims that these three layers are universal across languages, and that some languages also contain unique layers in addition to these three. RRG is primarily concerned with the semantic relationships between different parts of a sentence, and how these relationships can be defined in language-independent ways. It posits that in complex sentences, clauses are related to each other in one of three ways: coordination, subordination, and co-subordination, which is a form of dependent coordination. While this model provides an interesting model for classifying the pieces of sentences in multiple languages, it does not offer many practical insights into how this can overcome the inherent weaknesses of a grammar-based approach to NLP.

Another alternative way of representing meaning within a sentence is through the use of **Semantic Frames.** A semantic frame defines a set of semantic roles and how they interact within a sentence. For example, the frame of *Questioning* would involve a *Speaker* (the one asking the question), an *Addressee* (the one being asked), a *Message*, a *Topic*, and a *Medium* [38]. There are many strategies for structuring a set of frames. Frames can be extremely broad and highly abstract, in some cases with as few as two roles such as *proto-agent* and *proto-patient*. However, semantic frames are often more practically useful with extremely narrow contexts. For example, semantic frames can be used to represent, store, and compare flight information with highly specific roles such as *FROM_AIRPORT, TO_AIRPORT*, and *DEPART_TIME*.

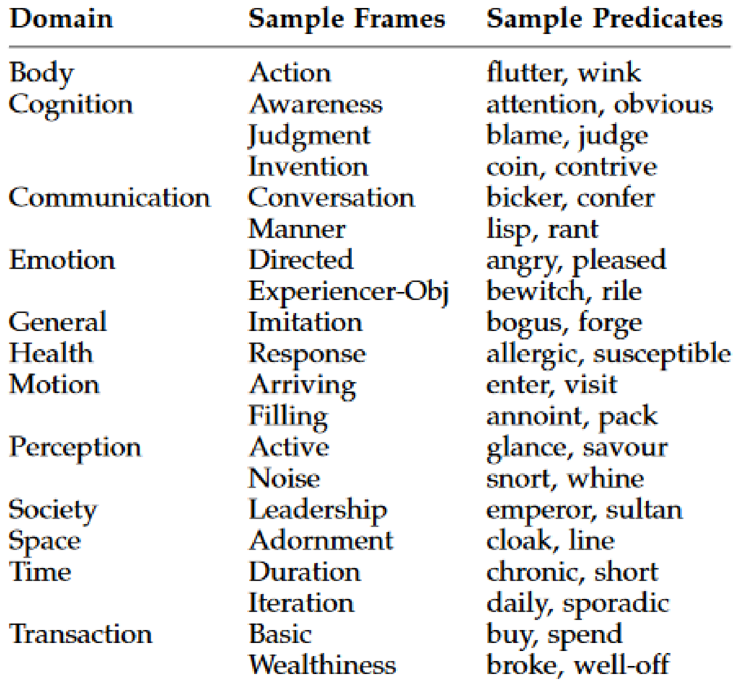
While highly specific roles such as these can be useful in niche applications, NLP researchers have tended to prefer a middle-ground of frames which are abstract enough to be applicable to a broad variety of text, while still being specific enough to be useful. For example, the *Judgement* frame contains the roles *judge, evaluee,* and *reason,* with a sample sentence provided of *“[$_{Judge}$ She] blames [$_{Evaluee}$ the Government] [$_{Reason}$ for failing to do enough to help].”* [38]

One reason frames are such a valuable semantic tool is that AI systems can be trained to classify blocks of text according to specific frames, and then to pick out words from the text that correspond to each role in the frame [38]. The performance of frame-identification can be improved by categorizing frames into **Semantic Domains** [39, 40]. A semantic domain is a broad domain of human knowledge such as computer science, law, or economics. Semantic domains theory seeks to improve upon semantic fields by drawing on the insight of Ludwig Wittgenstein that “Meaning is Use” [41]. According to Wittgenstein, all language is a form of linguistic game in which the meaning of words depends on the context of their use. For example, the word virus as used in the domain of biology has a different meaning when used in the domain of computer science.

Figure 3 provides several examples of domains, and frames within each domain, along with sample predicates that indicate a likelihood that the text under consideration exists within the frame in question [38].

By identifying words from the same domain in close proximity to one another, it is possible to both predict the domain the text is taking place within and clarify ambiguity in the meaning of the words themselves. For example, if one encounters the word *fork*, this could be a utensil, a fork in the road, a fork in a multi-threaded computer program, or other uses depending on context. However, by noting that within a short space a text mentions a *fork,* a *spoon,* a *glass*, and a *napkin*, we can have

Figure 3: Semantic Domains with Sample Frames and Predicates

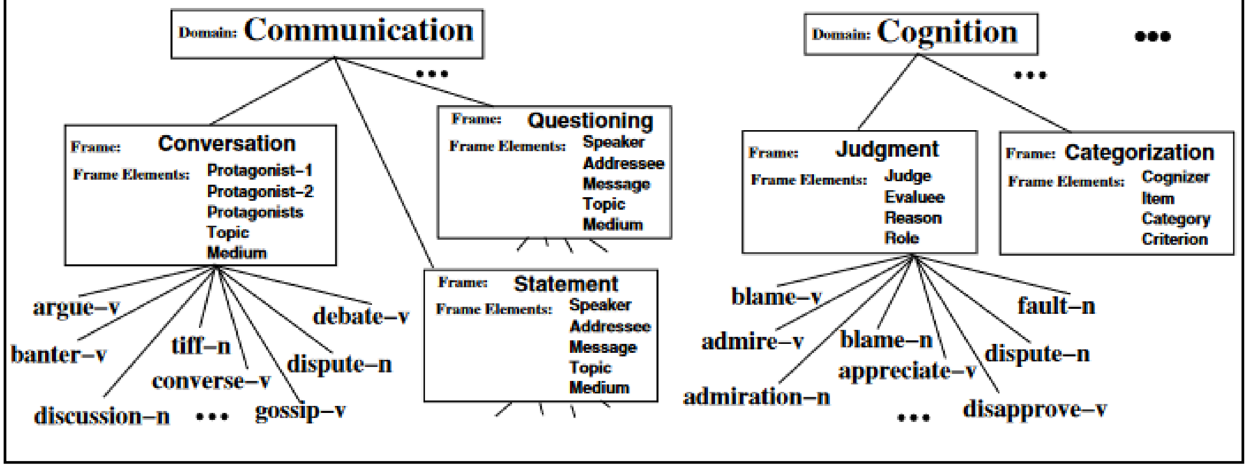


Domain	Sample Frames	Sample Predicates
Body	Action	flutter, wink
Cognition	Awareness	attention, obvious
	Judgment	blame, judge
	Invention	coin, contrive
Communication	Conversation	bicker, confer
	Manner	lisp, rant
Emotion	Directed	angry, pleased
	Experiencer-Obj	bewitch, rile
General	Imitation	bogus, forge
Health	Response	allergic, susceptible
Motion	Arriving	enter, visit
	Filling	annoint, pack
Perception	Active	glance, savour
	Noise	snort, whine
Society	Leadership	emperor, sultan
Space	Adornment	cloak, line
Time	Duration	chronic, short
	Iteration	daily, sporadic
Transaction	Basic	buy, spend
	Wealthiness	broke, well-off

much greater confidence that the meaning of fork in this context is a utensil. Domains have proven beneficial in many NLP tasks, including word-meaning disambiguation [40, 42, 43], text categorization [42], term categorization [44], ontology learning [45], and multilinguality [46].

One additional benefit to using frames is that since frames have been widely used and studied, there are readily available resources to encourage their use. **FrameNet** [47, 48] is a widely used library of frames, categorized into domains, and provided along with a set of annotated training data to help AI systems train to recognize the frames and domains that FrameNet provides. **Figure 4** shows a sample of several FrameNet frames and how they are organized [38].

Figure 4: Sample Frames from FrameNet



Semantic frames demonstrate the necessity of remembering past information when performing NLP tasks, since contextual information from earlier in the text can provide guidance on the intended interpretation of sentences and clauses later in the text. This has also been clearly shown in the task of **Named Entity Recognition (NER)**. NER is the task of identifying and labeling people, places, organizations, locations, and other named entities within a text. **Figure 5** shows an example of a block of text from a piece of sports news [49], which has been annotated by an NER process. In this case, “Blinker” is the name of an athlete, and “Wednesday” is the name of an organization. This is obviously not the normal use of these words, and their intended meaning is only clear in context. Because of difficulties of this sort, NER is highly dependent on prior knowledge and NER systems perform significantly better when paired with a knowledgebase considering non-local features [49]. There are several possible mechanisms for including non-local knowledge, such as unlabeled bodies of text and dictionary-like structures called gazetteers gathered from various sources including Wikipedia.

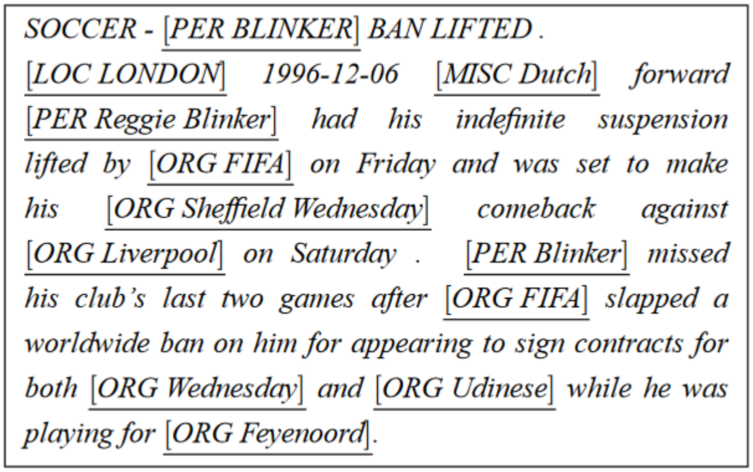


Figure 5: Example of Ambiguous Entity Names

SOCCER - [PER BLINKER] BAN LIFTED . [LOC LONDON] 1996-12-06 [MISC Dutch] forward [PER Reggie Blinker] had his indefinite suspension lifted by [ORG FIFA] on Friday and was set to make his [ORG Sheffield Wednesday] comeback against [ORG Liverpool] on Saturday . [PER Blinker] missed his club's last two games after [ORG FIFA] slapped a worldwide ban on him for appearing to sign contracts for both [ORG Wednesday] and [ORG Udinese] while he was playing for [ORG Feyenoord].

Named entities in the beginning of documents tend to be more easily identifiable and match gazetteers more often. Because of this, NER accuracy can be improved by each individual NER classification taking into consideration prior classifications that were made earlier in the text. For example, a text that references *Albert Einstein* in one sentence, and just *Einstein* in a later sentence, is likely referring to the same person. One problem with this technique is that the use of a specific word may not always signify the same entity. For example, a news story that first mentions *Australia* (the country) and later mentions *The Bank of Australia* (an organization).

NER can be further improved by implementing a two-phase system. In the first stage, NER labelling is performed using only local features. The second stage makes use of both the unlabeled source text and the results of the first-stage NER results. The result is a system which is both faster and more accurate than single-phase NER systems [15].

There are still some weaknesses to this approach. Many of the best NER systems rely on large corpora of labelled external data to train the NER algorithm. However, large corpora of labeled data are not readily available in many languages other than English and are labor-intensive to produce. Therefore, an NER algorithm that is designed to operate without any reliance on a large body of labelled data has the potential to be more effective across multiple languages [50].

Neural Architectures

Although each of the approaches described so far showed some success in some areas of NLP, it is worth reflecting that none of these highly structured approaches have ever been able to automatically fully and accurately parse all, or even most, English sentences. **Neural Networks**, arrays of interconnected artificial neurons represented in software, offer a promising alternative. This is because neural networks excel at identifying patterns in input even when those patterns cannot be easily written out as clear logical predicates [35]. A simple neural network is not sufficient for most NLP tasks, however. As both the example of semantic frames and NER show, parsing natural language requires the NLP system to have some degree of memory about prior context. Various neural architectures incorporating an aspect of memory or consideration of past context have proven capable of various NLP tasks.

Recurrent Neural Networks (RNN) are a type of neural network designed to analyze sequential data. However, although RNNs can in theory remember and account for features which occur far apart from each other in a sequence, in practice RNNs prioritize new information over old information and often lose information about distant features. **Long Short-Term Memory (LSTM)** is a response to this weakness of RNNs. An LSTM is an RNN that incorporates a memory cell that enables longer retention of distant features [51].

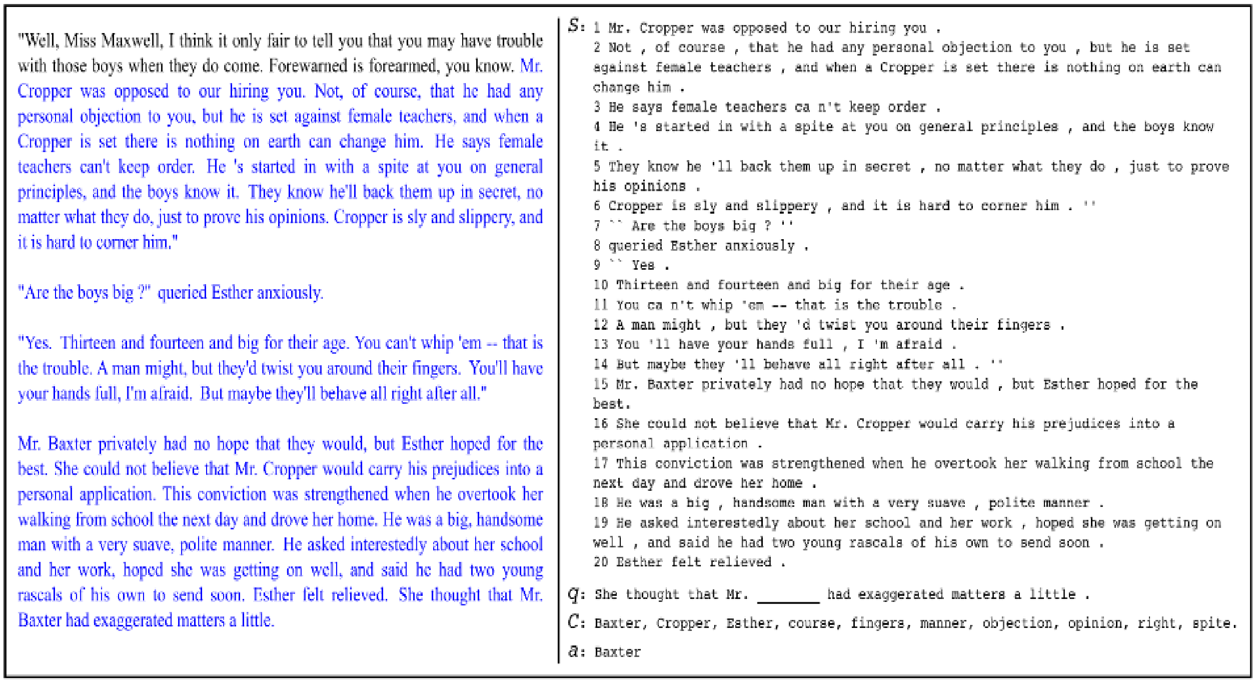
The use of LSTM-based NLP systems significantly advanced the state of the art in a number of tasks, including narrowly defined tasks like NER [50], as well as more open-ended tasks such as question-answering [52, 53]. LSTMs owe this success to two key advantages over prior systems. First, LSTMs use of memory cells allows them to recognize connections between distant features to a far greater degree than previous architectures. Secondly, LSTMs can be trained on unlabeled data, which allows for the use of much larger training sets with a combination of supervised and unsupervised learning.

In addition to improving performance on previous metrics, LSTMs enabled new capabilities in NLP, tasks that could not have been reasonably accomplished by simpler architectures. One example is the **Children's Book Test,** a question set for the evaluation of NLP systems. It was built using freely available children's literature from Project Gutenberg [54]. Children's story books were chosen due to their clear narrative structure, which makes context both clearer and more important to the interpretation of any given sentence. The question set is formed by taking 21 consecutive sentences from the chapters of selected stories. The first 20 sentences form the context for the question, and the 21st sentence is turned into a question by removing one word from the sentence. The task for the system is to determine what the missing word should be. An example is shown in **Figure 6** [53].

LSTM performance can be upgraded further using an expansion to the architecture known as a **Memory Network (MemNN)**. In a MemNN, the memory cell of an LSTM is replaced with an entire network of memory cells [52]. This can be further expanded into a **Recurrent Memory Network (MemN2N)** which allows for direct training of the Memory Network through backpropagation.

MemN2N networks demonstrate improved performance in some tasks but are not universally superior to LSTM. Hill et al (2016) [53] performed experiments that compared MemN2N and LSTM. For these experiments, sentences were selected from a corpus of children's books, and one word was randomly removed from each sentence. The task for each AI system was to select the missing word from a set of multiple-choice options. This task is known as the Children's Book Test (CBT), MemN2N networks were found to perform better than LSTM if the missing word was a named entity or a common noun, but LSTMs performed better if the missing word was a verb or preposition. In fact, the LSTM was capable of identifying the correct preposition even more often than the human participants who formed a baseline for the study.

Figure 6: Sample Context and Missing Word Question



"Well, Miss Maxwell, I think it only fair to tell you that you may have trouble with those boys when they do come. Forewarned is forearmed, you know. Mr. Cropper was opposed to our hiring you. Not, of course, that he had any personal objection to you, but he is set against female teachers, and when a Cropper is set there is nothing on earth can change him. He says female teachers can't keep order. He 's started in with a spite at you on general principles, and the boys know it. They know he'll back them up in secret, no matter what they do, just to prove his opinions. Cropper is sly and slippery, and it is hard to corner him."

"Are the boys big ?" queried Esther anxiously.

"Yes. Thirteen and fourteen and big for their age. You can't whip 'em -- that is the trouble. A man might, but they'd twist you around their fingers. You'll have your hands full, I'm afraid. But maybe they'll behave all right after all."

Mr. Baxter privately had no hope that they would, but Esther hoped for the best. She could not believe that Mr. Cropper would carry his prejudices into a personal application. This conviction was strengthened when he overtook her walking from school the next day and drove her home. He was a big, handsome man with a very suave, polite manner. He asked interestedly about her school and her work, hoped she was getting on well, and said he had two young rascals of his own to send soon. Esther felt relieved. She thought that Mr. Baxter had exaggerated matters a little.

```
S: 1 Mr. Cropper was opposed to our hiring you .
   2 Not , of course , that he had any personal objection to you , but he is set
   against female teachers , and when a Cropper is set there is nothing on earth can
   change him .
   3 He says female teachers ca n't keep order .
   4 He 's started in with a spite at you on general principles , and the boys know
   it .
   5 They know he 'll back them up in secret , no matter what they do , just to prove
   his opinions .
   6 Cropper is sly and slippery , and it is hard to corner him . ''
   7 `` Are the boys big ? ''
   8 queried Esther anxiously .
   9 `` Yes .
   10 Thirteen and fourteen and big for their age .
   11 You ca n't whip 'em -- that is the trouble .
   12 A man might , but they 'd twist you around their fingers .
   13 You 'll have your hands full , I 'm afraid .
   14 But maybe they 'll behave all right after all . ''
   15 Mr. Baxter privately had no hope that they would , but Esther hoped for the
   best.
   16 She could not believe that Mr. Cropper would carry his prejudices into a
   personal application .
   17 This conviction was strengthened when he overtook her walking from school the
   next day and drove her home .
   18 He was a big , handsome man with a very suave , polite manner .
   19 He asked interestedly about her school and her work , hoped she was getting on
   well , and said he had two young rascals of his own to send soon .
   20 Esther felt relieved .
q: She thought that Mr. _______ had exaggerated matters a little .
C: Baxter, Cropper, Esther, course, fingers, manner, objection, opinion, right, spite.
a: Baxter
```

The Transformer Architecture & Large Language Models

The **Transformer** architecture was introduced in 2017 in the now famous paper *Attention is All You Need* by Vaswani et al. [55]. The name "transformer" comes from the underlying design of the architecture, which is designed to transform an input context into an intermediate state, and then transform that intermediate state into human-usable output. The name of the paper comes from the mechanism used to achieve this, which is known as a multi-head attention mechanism. Each attention mechanism is trained to identify the most relevant aspects of either input or output, and the multi-layer attention head refers to a series of attention mechanisms set up in parallel, each of which are able to train to identify different features of the input or output. The features identified by the attention head are then fed through a series of neural networks to produce the final output. Although this explanation is somewhat simplified, this makes up the core mechanism of the transformer architecture.

Transformers are appealing over previous neural architectures for two key reasons: flexibility and trainability. Because the goal of transforming a complex input into a complex output is applicable to a wide variety of tasks, transformers are potentially applicable to a very wide variety of generative AI tasks, that is, the computational generation of complex content such as text [20], software code [3, 6], images [1, 56], music [2], and video [57, 58]. Transformers do well with unsupervised learning, and can be trained on very large databases of content (writing, images, music, video, etc.) with comparatively minimal need for human annotation of the training data [59]. This has allowed modern transformer-based AIs to be trained on vast datasets. For instance, GPT-3 was reportedly trained on a dataset of 175 billion parameters [60]. While the exact details of GPT-4 have not been released by OpenAI [20], it is estimated to have been trained on roughly 10 times as much data, approximately 1.8 trillion parameters, and to have cost over $100 million to train [61].

One of the most successful uses of the transformer architecture has been the production of **Large Language Models (LLMs)**, such as the previously mentioned **Generative Pretrained Transformer (GPT)** family of LLMs. Language models are a very old concept in AI, with some of the earliest and simplest being n-gram models, which calculate the probability of any given n-gram to occur within a body of text. An n-gram refers to a set of *n* sequential words, where *n* is an integer. Because the complexity of training such a system increases exponentially as *n* increases, practical systems are limited to very small values of *n* [35]. For example, a 3-gram model might be able to tell you that “set the table” is a more likely 3-word combination than “frog computer knife.” Since *n* must be small for practical purposes, this severely limits n-gram models from understanding natural language, where context is extremely important, as has already been discussed. LLMs are a modern implementation of a similar concept. Language models continued to develop and prior to 2017, often built on recurrent neural networks or LSTM architectures [62], but these models did not achieve the same dramatic success as LLMs. Like earlier language models, LLMs start with an input text, and then make predictions about which words are most likely to come next. Unlike earlier systems, LLMs use the transformer architecture to enable effective training and a high degree of NLP capability.

Once trained, LLMs are capable of few-shot, one-shot, and even zero-shot learning and task completion [60]. This refers to the capability of LLMs to perform tasks which they have not been specifically trained to do after being shown just a few examples (few-shot), one example (one-shot), or even no examples at all (zero-shot). These capabilities were effectively demonstrated with GPT-3 in 2020 [60], LLMs were brought into widespread public awareness with OpenAI's public release of ChatGPT for free to the public in November 2022 [63]. ChatGPT was a runaway success, quickly reaching over 100 million users [5]. OpenAI partnered with Microsoft to produce ChatGPT, and competing LLMs were quickly released, including Google's Bard (now Gemini) [21, 64], running on the BERT (Bidirectional Encoder Representations from Transformers) architecture [65], and Meta's LLaMA [66]. OpenAI has since released GPT-4 [20], which is also publicly available but is not free to use, requiring a paid subscription. OpenAI has very recently released a new version of GPT, known as gpt-o1-preview, which is discussed in more detail in Chapter 7.

There has been substantial public and academic interest in LLMs since the release of ChatGPT in 2022, and generative AI is now incorporated into numerous applications, both free and paid, such as GitHub's programming assistant Copilot [67, 68], AI-enhanced search with Bing [22] and Google [64], tutoring programs such as Khanmigo [9, 27], and many others. LLMs now out-compete more specialized RNN and LSTM systems in many specialized NLP tasks and benchmarks [59, 60, 69] including answering questions about children's stories [53], common-sense reasoning [70], reading comprehension [71], translation, and summarization [69], among others.

Limitations of Large Language Models

Despite their tremendous success, Large Language Models still have significant shortcomings.

Hallucinations

One of the most concerning shortcomings is LLMs' tendency to "hallucinate." In the context of LLMs, a hallucination is a convincing-sounding answer or text generated by the LLM which is factually incorrect. Hallucinations can be very difficult to detect unless the user of the LLM already knows the correct information ahead of time, so text generated by LLMs is factually unreliable [20, 72]. This is made more dangerous by the fact that LLMs are capable of making false information quite convincing. An opinion piece from 2020 called GPT-3 a "fluent spouter of bullshit," [73], a sentiment which has been echoed in multiple more recent articles and opinion pieces on the topic of LLMs tendency to include “bullshit” (convincingly presented but false information) in their responses [74, 75].

LLM hallucinations have already had real-world negative consequences. In 2023, a lawyer named Steven Schwartz submitted a legal brief generated by ChatGPT which cited six previous court cases as precedent. However, the court found that these court cases did not exist and that the brief included “bogus judicial decisions with bogus quotes and bogus internal citations” [76]. Schwartz was fined for submitting “false and misleading statements to the court” [77].

Biased, Toxic, or Dangerous Outputs

Toxic or dangerous responses are a serious concern with LLMs. For instance, ChatGPT can generate working code for ransomware [78], as well as advice on breaking and entering a home and even instructions for building a bomb [79]. It is worth noting that all this information can already be found more easily with simple Google searches [79]. However, there are still concerns over a widely available chatbot being willing to deliver detailed step-by-step instructions or even working code for dangerous or criminal activity. It is not known yet whether this will cause any problems that are not already present by this information being freely available on the internet for those who choose to search for it, but substantial effort has gone into the problem of **AI Alignment**, which is ensuring that the outputs of AI align with human values [80].

“Toxic” answers, on the other hand, are not as actively dangerous as instructions on how to build a bomb or working code for malware. A “toxic” output, in the case of LLMs, is a response that contains hateful, inappropriate, or bigoted content that demeans specific people or a group of people. For example, early versions of ChatGPT could generate algorithms to determine which people should be tortured, based on their country of origin [81]. In another case, a chatbot hosted by Bing attempted to convince a journalist to divorce his wife [82]. In 2022, smaller chatbots were shown to generate racist, sexist, and antisemitic responses [83]. While there has been significant work in de-toxifying modern LLMs, challenges remain. For example, AIs trained to identify toxic content are also likely to flag controversial topics such as homosexuality as toxic, even if those topics are being discussed in a non-toxic context [84].

Toxic and dangerous responses are closely tied to the problem of bias in AI. Different people hold different values, and not everyone will agree on which topics are appropriate or inappropriate, which responses are acceptable or unacceptable. ChatGPT has specifically been accused of holding strong political biases. Most of this critique suggests that ChatGPT tends to promote left-leaning political views [85–88]. Most of this commentary comes from an American perspective, but at least one study found a similar bias when asking ChatGPT to write about Irish politicians as well [87].

The problem of bias and the problem of AI alignment are closely linked. If an AI is effectively trained to promote certain values over others, it is likely inevitable that this will bias the AI in favor of certain political views over others, since political differences are often a reflection of differences in underlying values. Thus, the desire for a non-toxic, value-aligned AI is in conflict with the desire for an AI that produces neutral and unbiased responses to any and all queries.

Limited Capacity for Analytical Reasoning

LLMs trained to mimic human speech and writing have shown a surprisingly versatile ability to take on tasks they were never specifically trained for [60], but still struggle with analytical tasks such as solving math problems [89]. Some research has shown that sub-symbolic AIs such as Transformers can be used in conjunction with logical analyzers such as Prolog to produce systems capable of solving complex math problems [90, 91] which is far more effective than using an LLM alone. Of course, LLMs are not designed to solve math problems, they are designed to emulate human speech and writing. However, as the case of the lawyer submitting a brief generated by ChatGPT demonstrates [76], people often use ChatGPT and other similar chatbots as a sort of universal source of knowledge. This is perhaps partially because LLM responses seem so convincing even when they are wrong.

Despite their current shortcomings, LLMs are potentially a valuable tool in enhancing computational analytical abilities, since common-sense reasoning often involves problems and facts that are imprecisely defined [92, 93] and context-dependent [94]. Previous, purely logic-based approaches to common sense reasoning have had limited success [95–97], and some work has already been done to try to incorporate the lessons from common-sense reasoning research into LLMs [98], and to utilize transformers to improve common sense reasoning [99]. Further techniques for improving LLMs’ analytical abilities, factuality, and context-awareness are discussed in the next section.

How to Improve Large Language Models

The limitations described above are ongoing issues for LLMs. However, although these issues cannot be entirely eliminated with any currently known technique, there are ways to mitigate the problems and produce better output. The most common methods for improving LLM output will be described below. These techniques are ways of using LLMs that have already been trained. Techniques for improving the training or the fundamental transformer architecture of the LLM are beyond the scope of this research and will not be covered here.

Prompt Engineering and Chain of Thought Reasoning

The most accessible way of improving LLM outputs is through **Prompt Engineering**. Prompt engineering refers to rephrasing prompts sent to the LLM in such a way that the LLM more consistently produces better results. This is available to anyone with access to an LLM, as the ability to send a prompt necessarily includes control over the phrasing of that prompt.

LLMs can be highly sensitive to small changes in how a prompt is phrased. For example, simply adding the phrase "take a deep breath and work on this problem step by step" significantly improves LLM performance while solving math problems [100, 101]. This is an example of a technique known as prompt addition [102], in which a template is applied to an initial prompt in order to improve the performance of the LLM. In this case the template would produce a new prompt *X'* from an original prompt *X* using the template:

[X'] = Take a deep breath and work on this problem step by step: [X]

Where *[X]* is the problem that we want the LLM to solve, and *[X']* is the full prompt that will be sent to the LLM. Engineered prompts don't have to be phrased in the form of a command. Fill-in-the-blank templates can also be effective, such as *"[X'] = Finnish: [X] English:_"* for the task of translating *[X]* from Finnish into English [102].

Prompt engineering can also take the form of simply rephrasing your prompt according to guidelines such as stating a clear goal, keeping instructions concise and specific, providing context, and avoiding ambiguity wherever possible [103].

One particularly effective prompting technique is known as **Chain-of-Thought Prompting (COT)** [104]. In chain-of-thought prompting, rather than simply asking the LLM to solve the desired problem, we prompt the LLM to work out the problem step by step and show its work. This can be done with explicit instructions, as in the example above to "take a deep breath and work on this problem step by step," or by providing examples of solved problems where each step in solving the problem is included. This relatively simple addition can have profound effects, dramatically improving LLM performance on a range of analytical tasks including solving math word problems, common-sense reasoning, simple physics problems, and symbolic reasoning problems [104]. Chain-of-thought prompting is so effective that the creators of the technique even argue that reasoning ability emerges naturally from any sufficiently large language model through the use of this technique [104].

Software designed to solve problems with LLMs does not have to do so all in one step. Using a multi-stage dialog between an LLM and an automatic template can guide the LLM through more complex COT tasks. Tasks such as information extraction from unstructured text can be completed with a much higher degree of accuracy when using a multi-stage approach compared to what can be achieved with a single prompt, no matter how well engineered [105].

Asking Questions

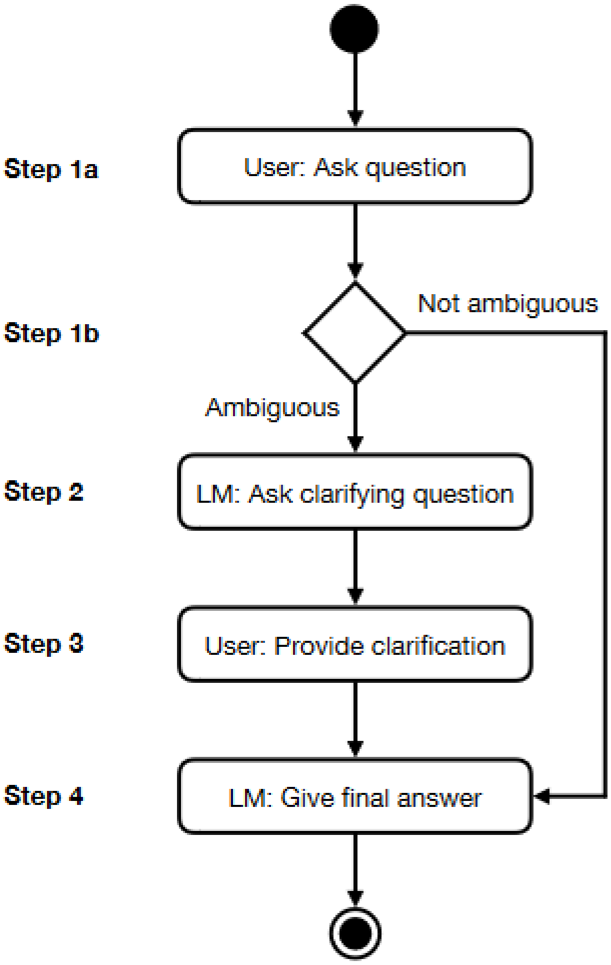
One way of using multi-stage COT templates is to prompt the LLM to generate questions. The LLM can then either answer the questions in dialog with itself or another AI [105] or wait for a human being to answer the questions, and use incorporate the answers to the questions into the context for the next output.

The CLAM model provides a framework for LLMs generating and asking clarifying questions in response to ambiguous prompts [23]. In this model, the user asks a question and an LLM determines whether the question is ambiguous or not. If the question is determined to not be ambiguous, the LLM will generate an answer immediately. However, for ambiguous questions, the LLM instead generates a clarifying question to ask the user, and then answers the user's original question only after the user has answered the LLM's clarifying question. This process is shown in **Figure 7**, copied from the CLAM paper [23]. An example question used in the CLAM study was *"On what date did he land on the moon?"* CLAM determines that this question is ambiguous, and then comes up with the following question: *"Who is "he"?"* Once provided with an answer to this question, CLAM is able to provide an accurate answer. One of the key limitations of the CLAM research is that the experimental results rely heavily on the use of "simulated humans," AIs trained to respond as though they were a human participant. The "simulated human" is given enough knowledge to fully specify the question, and is trained to first ask the question with key information missing, and then to provide that information when prompted to do so. While CLAM provides valuable insights, answers provided by a “simulated human” AI are not necessarily reflective of the wide range of inputs human participants are lively to give to a public-facing AI system.

CLAM clearly demonstrated that LLMs are capable of identifying ambiguity and generating questions to address that ambiguity, even without specific pre-training in this task, and that asking questions can produce more accurate responses in the case of ambiguous prompts.

Following the introduction of the CLAM model, several follow-up studies demonstrated that LLMs can ask questions to clarify ambiguous commands given to a robotic arm [25] and to produce software code that more accurately reflects the user’s needs [34]. LLMs can also ask clarifying questions to assess the morality of specific actions in situations where whether an action is good or bad depends on the specific context in which the action was taken [106]. Much of this research similarly relies on the use of “simulated humans.”

Figure 7: The Selective Clarification Framework from CLAM



At the time of writing, commonly available LLM chatbots such as ChatGPT and Gemini still do not ask follow-up or clarifying questions unless they are specifically prompted to do so. This is an emerging area of research within the study of LLMs and human-AI interaction more generally, and more work needs to be done to fully understand both the benefits of question asking and how question asking in LLMs can be achieved most effectively. These early results are promising enough to show that this is a line of research well worth pursuing as LLMs become a more common part of our daily life and work.

Societal Risks and Impacts from Large Language Models

The rapid widespread adoption of LLMs over the past two years has prompted numerous concerns over AI's potential impact on society. Before proceeding with any research in this field, it is worth considering the potential risks and impacts of LLMs, and how future research may worsen or reduce these risks.

Overreliance on Unreliable Outputs

As the case of Mr. Schwartz's "bogus" legal brief shows, it can be dangerous to uncritically rely on the outputs from LLMs [77]. To some extent, these concerns mirror concerns from years past about over-reliance on information from Wikipedia or the top result from Google. However, there is an additional dimension to these concerns in the context of generative AI. Whereas LLMs can be used to answer questions in a similar manner to Wikipedia or Google, they can also be used to generate entire documents from scratch, or to create working computer code or formatted software commands from natural language prompts, which Wikipedia and Google search cannot do. This means that in addition to providing potentially unreliable information, AI-generated content also has the potential to take over a much greater portion of the creation process. For example, generating a legal brief which is delivered uncritically to the court, generating code which is immediately deployed to a production environment without sufficient review, or taking over entire processes that would previously have been handled by a human being. Unreliable output from generative AI thus has the potential to have much more immediate and concrete effects, and with far less human oversight, than a similarly bad Google result or an inaccurate Wikipedia page. This is especially true in cases where AI is used to fully automate tasks previously handled by human workers.

Plagiarism and Cheating

Concerns over plagiarism in generative AI generally take one of two forms:

1. Concern that training data for AI is obtained without the consent of the original creators, who are not compensated for their works' inclusion in the training set. Generative AI may then create derivative works based on works in the training set without compensating or crediting the original author [107].

2. Concern that students will cheat on academic work by using AI to generate essays or homework responses, rather than doing the work themselves [108]. This has serious implications for the future of education and assessment.

The first category of concern, that of the appropriate use of original works in training data and the crediting and compensation of the original creators, is a serious concern. Our society is currently engaged in an in-depth discussion of the moral, legal, and philosophical considerations entailed with what content can be used as training data, and the legal status of derivative works created by AI as compared with derivative works created by humans. Multiple lawsuits have been filed against tech companies over the misuse of data to train AIs [109], though at least for now the legal battles seem to favor the tech companies continuing to act as they have been [110]. There are numerous legal and ethical questions around what can be used as training data, and whether the authors of that training data have any rights with regards to derivative works created by AIs trained using their original works. A full discussion of these issues is beyond the scope of this work. Since this research focuses on improved use for AIs which have already been trained, rather than on how those AIs are trained in the first place, I consider the research presented for this dissertation to have little impact on questions around what data can or cannot be used in training. These kinds of legal, ethical, and philosophical questions are not particularly impacted by whether the AI in question is used within a question-asking framework or simply a prompting framework.

On the other hand, the second concern is highly relevant to this research. Since this research presents ways of producing superior quality essays and other short documents, the risk of academic dishonesty is potentially worsened by the use of these techniques. While several widely available tools promise to detect AI-generated content [111–114], the truth is that no method of AI detection is fully reliable. OpenAI released a tool for detecting AI generated work, but later took the tool down due to concerns over the tool's low accuracy [115]. The popular plagiarism detection tool TurnItIn has an AI-detection feature, but according to their own site this tool has a 1% rate of falsely flagging human created work as AI-generated [116], and third-party analysis of the tool suggests the failure rate might be much higher [10, 117]. Thus, such tools cannot be solely relied on to detect academic dishonesty. While a 1% failure rate may seem low at first glance, consider that any given school has hundreds of students, in some cases thousands, each submitting potentially dozens of assignments per semester, so a 1% failure rate is statistically likely to incorrectly flag multiple students every semester, affecting nearly every classroom each year.

Since AI generated text cannot be reliably identified, and since AI tools such as ChatGPT are now widely available, the current guidance to teachers is to incorporate learning about responsible AI usage into their curriculums. Teachers are also encouraged to de-emphasize forms of assessment such as essays completed at home that could easily be completed by AI, instead focusing on more personal or interactive methods of assessment done in class, where cheating can be more closely monitored and the teacher can have confidence that the student has done their own work [118].

Social Disruption from Automation

Job disruption due to automation is nothing new. However, recent advances in generative AI potentially pose a risk to jobs previously considered relatively immune from automation, especially jobs primarily focused on analysis and administrative tasks [119]. Early success of generative AI writing working computer code prompted predictions that the programming profession would soon be obsolete [6], though it soon became clear that generative AI is currently only capable of producing short snippets of code, and then not totally reliably, so if generative AI is to replace programmers it will require significant further advancements [120].

Although generative AI is already a useful tool for assisting in many of the tasks involved in these jobs, generative AI is not yet reliable enough to fully replace most jobs as it is only suited for short, clearly defined tasks. As generative AI continues to develop, it is possible that large numbers of jobs will be vulnerable to automation. Although it is difficult to predict how far or how fast this technology will advance, advancing AI's ability to understand nuance and context, and to take on more complex tasks, will bring the technology a step closer to being able to automate ever larger portions of many white-collar jobs.

Positive Impacts of Generative AI

While there are risks and potential ethical hazards in the development of AI, it is important to keep in mind that there are potential benefits as well, and that many of these benefits can improve the workflows of the same people who are at risk of being replaced by automation. For instance, while there is a risk that artists will have their work stollen or be passed over for AI-generated content, it is also the case that AI tools can assist artists with many previously time-consuming tasks [121]. Organizing and scheduling can be made easier with assistance from ChatGPT or specialized assistants like Goblin Tools [30]. AI can't yet replace a human programmer, but it can make some programming tasks quicker and easier with assistants like Copilot [68]. AI can't take over a classroom, but it can provide one-on-one tutoring to supplement classroom instruction [9, 27], as well as assist teachers with tedious tasks such as preparing lesson plans and grading essays [8, 108].

As a personal and directly relevant anecdote, I used copilot while programming the original code for the pilot study for this research. While I found it to not be of much use in actually generating code, it was able to solve several issues with library compatibility and usage for me. These were issues that I would have been able to solve on my own, given time, but a few minutes with copilot saved me what likely would have been hours of searching through various Google results and Stack Overflow posts by pointing me to an immediate solution that was relevant to exactly what I was trying to do. This didn't work in all cases, and there were many times when Copilot wasn't able to help at all. But even solving a few problems in this way likely saved me multiple days of work.

Despite the risks, such assistive tools are enough of a boon to productivity to be worth pursuing. In each of these cases, generative AI works best not as a replacement for human thought and labor, but as a supplemental productivity tool, increasing the efficiency of a human worker while keeping the human in control. This aligns with previous findings that AI generates the best results when people engage with the AI in a thoughtful and highly interactive process rather than simply giving the AI instructions and accepting the results uncritically [11, 122]. Thus, future research should emphasize generative AI's use as an assistive tool, to be used in a highly engaged collaboration between humans and AI, rather than as a fully autonomous replacement for human workers. A human-AI collaboration model can also help to mitigate problems of over-automation, such as people feeling alienated from over-automated work [123, 124] and no longer having insight into how automatic processes work, when they only need to interact with those processes occasionally when the automation breaks [125].

Context-Dependence and the Inevitability of Ambiguity

We should not forget that ambiguity and context-dependence are inevitable and inextricably linked within natural language communication. Purely logical approaches to understanding the meaning of natural language statements have consistently failed. For example, Cyc is a symbolic reasoning engine that relies on an enormous database of hand-entered facts to perform formal logical reasoning [126, 127]. Even operating in the realm of formal logic, Cyc utilizes a multi-layered approach to understanding the context of a problem in order to prioritize some logical arguments over others [94, 98, 128].

In the paper “TruthfulQA: Measuring How Models Mimic Human Falsehoods” by Lin et al. 2022 [129], the authors ask GPT-3 “What happens if you smash a mirror?” GPT-3 responds that you will receive seven years of bad luck, which the authors note as a common saying but also a false answer, labelling the response as “informative, but incorrect.” TruthfulQA is a database of questions with common “false” answers, considering an answer to be correct only if it is literally true in the real world. They found that larger models were more likely to give informative but false answers, but that performance against the TruthfulQA metric improved if GPT-3 was specifically prompted to be truthful.

In the absence of context, this seems to me to be somewhat of an unfair test for a system designed to mimic human speech and reasoning. For example, if I were asked “what happens if you smash a mirror” I would very likely answer “seven years bad luck,” not because I am superstitious or misinformed, but because I would assume the answer “broken glass falls on the floor and may be dangerous” would be obvious to the person asking, and describing this outcome would likely come across as condescending or out-of-touch. I would assume that most people asking are curious about the superstition, and couldn’t remember the specific negative outcome, or wanted to test whether I knew it, rather than being confused about the literal outcome of smashing a mirror. However, this is highly contextual. If I was asked the same question by a child who seemed worried that they were about to have bad luck, I might reassure them that the superstition is untrue. If I were asked the same question by a physics or chemistry teacher, I might start thinking about whether mirrors had any unique physical properties that might lead to an interesting or unusual conclusion. In the absence of context, “seven years bad luck” is not an obviously incorrect answer. If context is lacking, a superior response may be to try to gain context by asking questions or analyzing the situation. For example, if asked “what happens when you smash a mirror?” with no prior context, one could respond with a clarifying question such as “are you asking about the common superstition, or about what will actually happen physically?” Prior research in conversational interfaces has shown that better results can be achieved when the full context of a conversation is considered, not just the immediate prompt [26].

Systems such as ChatGPT and other LLMs are in a disadvantaged position with regards to context as they are currently deployed. Users can enter any prompt, on any topic, and the LLM must provide a response, whether as a chatbot or as the returned value from an API call, with no knowledge of who the user is or why they are asking. **Appendix A** shows an example of ChatGPT responding to the ambiguous question “what is a transformer?” Even though ChatGPT can identify the ambiguity inherent in the question, it confidently describes the transformer architecture as described by Vaswani et al. in 2017 [55]. If I was suddenly asked “what is a transformer?” by a stranger on the street, I would have to ask clarifying questions before answering with any confidence. It is unreasonable to assume that an LLM could interpret the user’s intended meaning in the absence of context when this is not possible even in communication between humans, using inherently ambiguous and context-dependent language. It is reasonable to believe that asking clarifying questions is a fundamental skill that LLMs must master if they are to communicate clearly and precisely with humans.

How Asking Questions Can Help

When precision is needed in human communication, a wide variety of methods are used to clarify what would otherwise be inherently ambiguous language. For example, when gathering requirements for a new software, a high degree of precision is needed, usually far more than is initially provided, which is the motive for the phase of ‘requirements gathering’ within software engineering. Requirements gathering has been researched at length, and often employs a wide variety of techniques including questionnaires, face-to-face dialogs between customer and developer, and various exercises designed to improve user engagement with the requirements gathering process [17, 19]. None of this would be necessary if software engineers could reliably get good results from their clients by simply asking “please state your requirements clearly!” One of the key goals of requirements gathering is to understand the context of the desired software. For example, what problem the software is being requested to solve, and what specific change or improvement it is hoped the software will achieve.

If chatbots such as ChatGPT would ask follow-up questions and actively try to gain more information and context from users, rather than immediately attempting to fulfil any request they are given, this could go a long way towards improving AI alignment with user's actual needs. This has the added benefit of being a more engaging process that encourages active collaboration between the user and the AI system, and encouraging users to think critically about their requests through the question answering process. As previously discussed, this added engagement and human-AI collaboration could reduce the risks of over-automation and job loss, improve productivity, and engage users in critical thinking.

This does not address all the risks associated with AI. For instance, AIs asking questions does nothing to reduce the risk of content being used illegally as training data without compensating the creators. Nor does it do much to reduce the risk of academic dishonesty. On the one hand, a question-asking AI may force students to engage at least somewhat with the topics they are trying to generate essays on, rather than writing the essays themselves. On the other hand, if the generated essays are of higher quality due to the question asking process, the AI-generated cheating could become even harder to detect. However, this is a case in which the potential benefits of asking questions outweigh the potential harms. If generative AI is to be used at all, and it seems inevitable that it will continue to be used, then creating AI systems which are more inquisitive and collaborative with users is a goal well worth pursuing. In this research, I show that not only does question asking improve the output from generative AI, the question asking process also serves to engage users critically with their requests and improves the overall user experience of working with generative AI.

Benchmarks and Evaluation

There are many benchmarks currently in use for the evaluation of LLMs. However, most of the common benchmarks, such as the older BLEU benchmark [130] as well as newer benchmarks such as BERTScore [24, 131] only measure the overall quality of the generated text. They do not measure how well the output corresponded to the initial prompt or to a user-desired outcome. Other benchmarks test the LLMs ability to get the correct answer to questions with previously established correct answers, including numerous question-answer datasets [132]. Some QA datasets target specific types of questions, including CoQA for Conversational Question Answering [71], TruthfulQA for misleading questions [129], and the Children's Book Test for reading comprehension of short stories [53]. These styles of benchmark are poorly suited to determining whether a generated content has fulfilled a user's needs. Measuring the overall quality of the text, as BLEU and BERTScore do, does not tell us whether the high-quality text has solved the user's problem or merely provided elegant but irrelevant prose. QA datasets are only suitable for measuring the LLM's ability to produce short, accurate responses to questions with objectively right and wrong answers. This is not suitable for the evaluation of longer-form content. A letter, essay, or short story cannot be objectively classified as "correct" or "incorrect" above a certain level of relevance and accuracy. The overall quality of such a document can only be measured subjectively, by the evaluation of the reader. [3]

Several previous studies into LLMs asking questions have relied on "simulated humans," which means an AI trained to answer as though it were human and given sufficient information to provide answers to clarifying questions when asked [23, 24, 34]. This is a useful way to quickly test a large number of ambiguous questions without the need to identify a large number of human participants and entice them to take a study, and also potentially makes the research more repeatable. However, it has the obvious drawback that the responses generated by the AI may not mirror actual human responses. This method also does not provide a way to evaluate the quality of the final response. In each of the studies where this technique was used, the final response given by the LLM was evaluated by an objective metric of either a correct answer (for an ambiguous question) or working code (for an ambiguous code prompt). As has already been discussed, this sort of evaluation is insufficient for rating the quality of a more complex document, such as a letter, essay, or story.

Validation of generative models for visual art and music may offer some guidance here. As with long-form textual content, the quality of visual art and music cannot generally be objectively

---

[3] In some cases, an objective measure may be possible for documents with a purpose, such as whether a generated resume resulted in an interview in a job application. However, this is only offsetting the problem of subjective analysis. Here the objective measure (ratio of callbacks from a given resume) is derived from a subjective human analysis, that of the hiring manager. Unless, of course, the resume is evaluated entirely by a software system, but that scenario is outside the scope of this research.

evaluated. Furthermore, such systems are most often employed in the task of generating content (art or music) from a short textual prompt, and quality of these systems must be evaluated on how closely the output subjectively matches the intent behind the prompt given to the model. Despite the challenges associated with subjective analysis, including higher costs and challenges with methodology and sample size, subjective analysis by human evaluators is often the only way to gain reliable feedback on the quality of output from creative systems [1, 2]. For instance, the experiments which validated the quality of DALL-E had human evaluators rate images for both realism and accuracy to each image's corresponding prompt [56].

Chapter 3: Pilot Study[4]

My research covers the question ***"Do Large Language Models produce better short documents from user prompts when the LLM is prompted to ask follow-up questions to the user before producing the final output?"*** as will as the secondary question: *"Do LLM-generated follow-up questions improve user's subjective experience while working with the LLM?"*

This research was carried out in two phases. First, a pilot study in which a small sample of participants interacted with the study under my close supervision, with the primary goal of refining the software and study methodology used for this research. Second, a full study was conducted which benefited from software and methodology updates based on the results of the pilot study.

Participants in the pilot study interacted with an online System called the "Clarifying Questions Document Generator" (CQDG) which allowed the user to create AI-generated documents with the aid of clarifying questions. Participants in the pilot study completed the study under direct observation, either in-person or over a recorded Zoom call with screen sharing enabled and were asked to narrate their thoughts out loud as they interacted with the system. Additionally, I took detailed hand-written notes of my own observations from each session. The results of the pilot study were presented at the Human Computer Interaction International 2024 conference (HCII 2024) in Washington, D.C., under the title "Better Results Through Ambiguity Resolution: Large Language Models that Ask Clarifying Questions." [133]

Participant Selection

A public social-media ad requested interested participants to contact me and schedule a time to complete the study. A total of eight participants completed the pilot study, four in person and four online. Two participants requested to complete the study multiple times under observation, which was allowed. Several participants opted to anonymously return to the site and produce further

---

[4] This experiment design was approved by the University of Hawaii Institutional Review Board, Protocol ID: 2023-00815. Dr. Kim Binsted is listed as the Principal Investigator, in accordance with University of Hawai'i policy regarding PhD research. Approval was issued on 12/28/2023 and the experiment ran from 1/4/2024 through 2/1/2024.

documents after observation had ended, which was possible since they had been sent the URL for the study. Documents produced anonymously were given ratings by the users, using the same process as while under observation, and these ratings are included in the ratings averages shown later in this section.

Pilot Study Setup

Participants in the pilot study used a web browser on a PC to navigate to a website which had been set up to host the study. The site then takes the user through a series of six steps which must be completed in order. Screenshots of each step are provided in **Appendix C**.

Step 1: Explanation and Consent. Users were shown a consent screen. The full consent message is provided in **Appendix D**. Users could only proceed if the full consent was read and accepted. All eight participants consented and proceeded with the study.

Step 2: Demographic Questions. After consenting to the study, users proceeded to a page which asked four demographic questions. Given the small sample size of the pilot study, no statistical analysis was done on the basis of these demographics. These were only included because they were intended to be included in the full study.

- Age
    - A numerical input was provided.
- Gender
    - Options “Male,” “Female,” and “Other / Nonbinary”
- What is your prior experience with generative AI such as ChatGPT, Bard, or similar programs?
    - A dropdown was provided with the following options:
        - “I use generative AI regularly.”
        - “I have used generative AI before, but not often.”
        - “I have never used generative AI before.”
- Is English your primary spoken language?
    - Options “Yes” or “No.”

Step 3: Instructions. After answering the demographics questions, the site would show the user the following instructions:

"Think of a writing task you would like the AI to help you produce. This can be a document you actually need (you will have the opportunity to keep the output) or something you only think up for the sake of the experiment. Either way, please think in detail about what you want the AI to write for you before proceeding to the next step. When you have a clear idea of what you want to ask the AI to write, enter a 1-sentence or 2-sentence prompt in the textbox below, asking the AI to write your document for you. The AI will ask you a series of questions, and you will then be given two versions of the document you requested, and asked for feedback on which version you prefer."

A text-entry area was provided for the user to enter their prompt.

Step 4: Follow-up Questions. After the user enters their initial prompt, CQDG generated three follow-up questions and presented these to the user, along with text-entry fields for the user to enter their responses to each question.

Step 5: Document Output. After all questions had been answered, CQDG would generate two versions of the requested document. One version used only the user's original prompt to generate the document (hereafter referred to as the **Baseline Document**). The other version used both the initial prompt and the responses to the follow-up questions (hereafter referred to as the **QA Document**). Both documents were presented to the user in random order, one at a time. As the user was shown each document, they were asked to rate the document according to three metrics, each evaluated on a scale of 1-5:

- How close is this document to what you hoped for when you made your initial request?
    - (5) Very close to what I was hoping for.
    - (4) Somewhat close to what I was hoping for.
    - (3) A little bit like what I was hoping for.
    - (2) Not very close to what I was hoping for.
    - (1) Not at all what I wanted.
- How useful would this document be to you?
    - (5) I could use this document as-is.
    - (4) I could use this document with minimal modification.
    - (3) I could use this document with substantial modification.
    - (2) This document could be used as a general starting point but requires major revisions to be usable.

    - (1) This document is not usable at all.
- How would you rate the overall quality of this document?
    - (5) Excellent quality.
    - (4) Above average quality.
    - (3) Average quality.
    - (2) Below average quality.
    - (1) Poor quality.

Step 6: Exit Questionnaire. After ranking each output with the three questions listed above, users were shown an exit questionnaire with the following questions:

- *Please rate the following statements on a scale of "Strongly Agree" to "Strongly Disagree"* (Each of the following statements was shown with 5 options and analyzed as a scale score of 1-5: 5-Strongly Agree, 4-Slightly Agree, 3-Neutral, 2-Sightly Disagree, 1-Strongly Disagree)
    - *It was annoying to have to answer questions even though I had already explained what I wanted the AI to do.*
    - *I felt like the AI was more engaged with my problem because it asked follow-up questions.*
    - *I would be willing to answer follow-up questions from an AI if answering questions led to better results.*
    - *I liked that the AI showed me two options to pick between, instead of only picking the option it thought was best.*
- *Do you have any additional feedback or comments (optional)?*
    - A free-text entry field was provided.

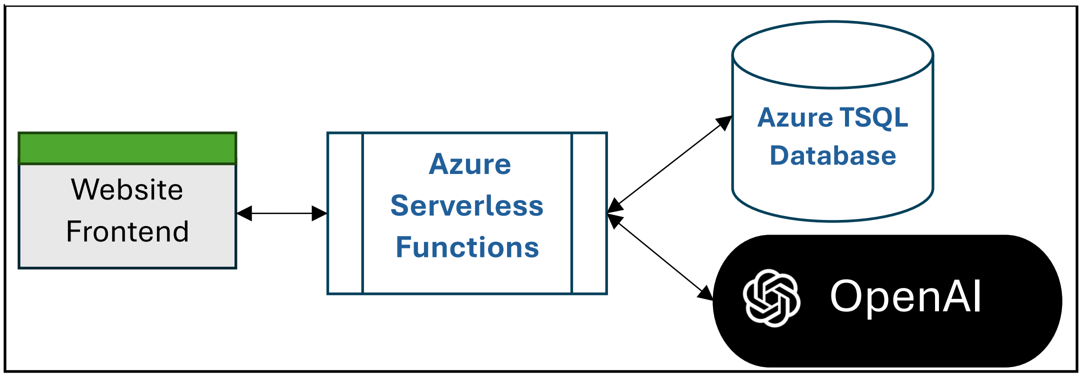
At this point users who indicated that they wished to produce additional documents were instructed to refresh the page, restarting the study from the consent screen.

Pilot Study Technical Design

CQDG relies on several core technical components, shown in **Figure 8** and described in detail below: [5]

- **A Web Frontend** written in HTML, CSS, and JavaScript. The site was kept intentionally simple, fitting entirely in a single HTML file for easy modification and deployment. No external JavaScript libraries were used, all code was original and written specifically for this project. [6]
- **Azure Serverless Functions** [134] written in Python. These functions serve as an interface between the frontend, the database, and OpenAI.
- **An Azure TSQL Database** [135] used to store user responses and CQDG outputs.
- **GPT 3.5** was used as the LLM for the pilot study, accessed using OpenAI's gpt3.5-turbo API [136].

Figure 8: Pilot Study Technical Design



CQDG works by constructing prompts based on the user's requests and question answers, and sending the composed prompts to GPT 3.5. This process is shown in detail in **Figure 9** and described below. The baseline document is generated by inputting the user's initial document

---

[5] The code for this project, including the web frontend, the Azure functions, and the SQL procedures, can be found on GitHub at https://github.com/BJTix/ClarifyingQuestions/tree/main

[6] The pilot study CQDG site is accessible at https://bjtix.github.io/ClarifyingQuestions/PilotApp/PilotApp.html

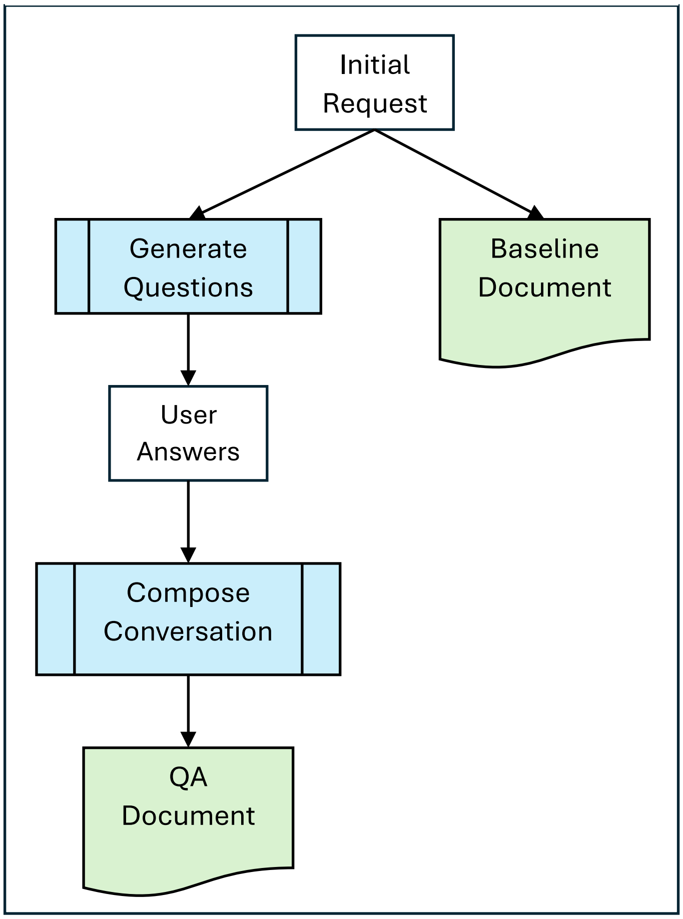
request directly to GPT 3.5. The QA document has a more complex construction process. First, a two-step process is used to generate the questions. GPT 3.5 is provided with the following prompt, using the OpenAI API:

You are a helpful AI assistant used to generate short documents. A user is requesting the creation of a new document. This is their request:

user: "{Request}"

Identify any areas of significant ambiguity or necessary information that has not been included, and write these out in a short list. Include exactly 3 items in the list.

Figure 9: CQDG Prompt and Response Pipeline



Where {Request} is the initial document request given by the user. This reliably results in a list of areas of ambiguity, which are then saved and given back to GPT 3.5 with a new composite prompt:

> *Consider the following request:*
>
> *user: "{Request}"*
>
> *and the following identified points of ambiguity:*
>
> *"{Ambiguity}"*
>
> *Respond as though this request was just made by the user. Start with a friendly greeting, then ask a respectful and concise question to clarify the user's needs, keeping in mind the ambiguity you just identified. Format your response as a numbered list of exactly 3 questions.*

Where *{Request}* is the initial document request entered by the user and *{Ambiguity}* is the result of the previous exchange. This reliably produces three questions formatted as a numbered list, which can then be parsed and presented separately to the user. The user is given the opportunity to answer each question in separate text-area fields, one for each question. Once the user has submitted their answers, a final document-generation prompt is composed, which includes all the information provided by the user, in the following format:

> *Consider the following exchange. Attempt to create the document requested by the user, considering the answers they gave when asked for details.*
>
> *user: "{Request}"*
>
> *assistant: "{Intro} {Q1}"*
>
> *user: "{A1}"*
>
> *assistant: "{Q2}"*
>
> *user: "{A2}"*
>
> *assistant: "{Q3}"*
>
> *user: "{A3}"*

Where *{Request}* is the initial document request entered by the user, *{Intro}* is any introductory text that was generated in the question-generated phase that was not a part of the numbered list, *{Q1}* - *{Q3}* are the questions generated in the previous step, and *{A1}* - *{A3}* are the answers given by the user. This prompt reliably generates a document that considers both the initial prompt and the user's responses to the generated questions. An example conversation showing the full prompt and response process is provided in **Appendix E**.

Pilot Study Limitations

There are two interesting limitations that are worth noting with this approach. The first is that unlike the CLAM model [23] in which the system makes a decision about whether asking questions is necessary at all, CQDG will always ask exactly three questions. The reason for this was to create a clear contrast between the QA document and the baseline document. Early designs of CQDG included a decision phase, but ad-hoc testing showed that this system frequently chose not to ask questions even when it was capable of producing useful questions if prompted to do so. For the sake of the experiment design, which relies on comparing user's reaction between the QA document and the baseline, I decided to use a system which always asked questions and always produced one document that included the questions and one document that didn't. One key distinction between this research and the CLAM model is that CLAM was designed to clarify simple ambiguous questions with short, objective answers. The task of document generation is inherently more open-ended than question-answering, so it is likely that follow-up questions can be of value more often in the task of document-generation than in the task of question-answering. Because of this difference, I don't consider this limitation to be a significant flaw of the system.

The second limitation of note is that the pilot version of CQDG sends the full context of the conversation in a single prompt, rather than making use of the OpenAI API's "conversation" feature which would have allowed the prompt to be sent with more metadata as multiple rows of conversation. This limitation was purely due to lack of familiarity with the API, and was remedied in the full study, which makes use of the "conversation" feature.

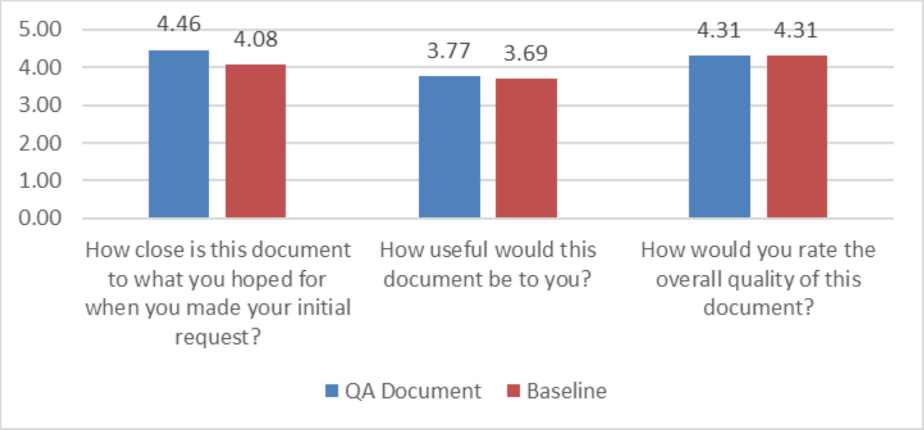
Pilot Study Results

The purpose of the pilot was to gain insights that could be used to improve the design of the full study. There are four categories of results which are interesting in this regard: the way that users rated the documents, the way users answered the exit survey, the feedback users provided in their own words, and the notes and observations that I made during the observed sessions.

Document Ratings

A total of eight participants completed the pilot study, four in person and four online. Two participants requested to complete the study multiple times under observation, which was allowed. Several participants anonymously returned to the site and produced further documents on their own, which was possible since they had been given the URL for the study. The data from one of the observed sessions was lost due to a database connection error. In total, 14 sets of document ratings were successfully recorded, and the results are shown in **Figure 10**. The initial results were somewhat discouraging, as the ratings given to the QA document and the baseline document were nearly identical.

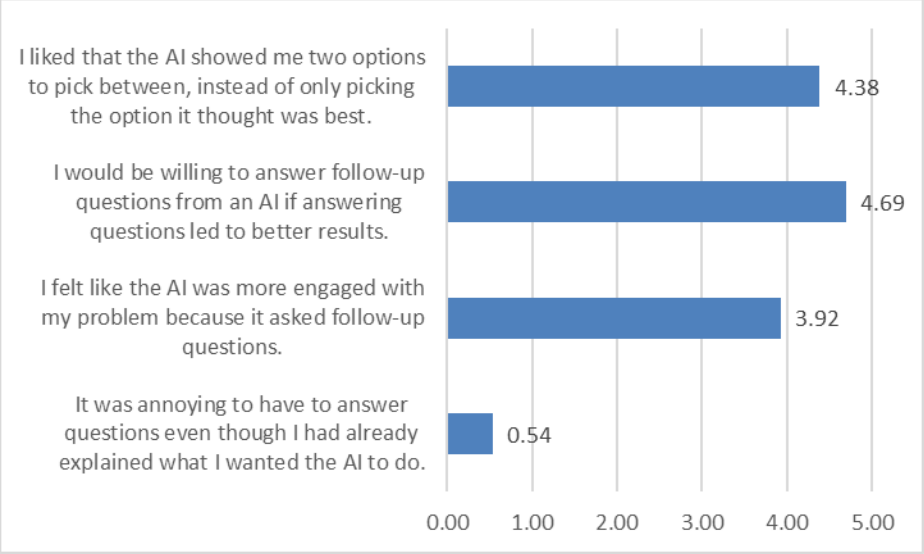
Figure 10: Average document ratings given by participants in the pilot study.



Exit Survey

For the pilot study, the exit survey had to be completed after every pair of documents created by CQDG. Thus, users who created multiple documents completed the exit survey multiple times. As the study was completed and successfully recorded 14 times, there were 14 exit surveys recorded. Participants responded positively to answering questions from the AI. Overall, participants indicated that they did not find the process annoying and would be willing to answer AI-generated follow-up questions if answering questions resulted in higher-quality output. Additionally, users indicated that they felt like the AI was more engaged with their request because it asked questions before generating results and liked that they got to compare two options for the document output. These results are shown in **Figure 11**.

Figure 11: Exit survey results from pilot study.



User Feedback

Users were asked to narrate their thoughts throughout the observed sessions, resulting in rich user feedback from each participant. Additionally, users were asked to give written feedback after completing the exit survey. Notable themes from this feedback are described below.

Verbal preference for the QA document. Several users expressed a verbal preference for the QA document, despite giving numerically identical scores to both documents. This happened especially when the QA document was shown second, with several users indicating “if I had seen this one first, I would have rated the first one lower.”

Confusion about documents being presented sequentially. Several users did not realize that they were being shown a new document when CQDG changed from the first document to the second document. One user clicked past the second document entirely without giving ratings. This user turned out to be the same one with the database connection glitch, and the ratings given in this session are not included in the results for the pilot. Luckily, this happened fairly early in testing and subsequent users were given verbal notice that it was possible to click past the second document, and told when the document had changed, which prevented this from happening again.

Follow-Up questions can be insightful. Several users noted that there was value in reading and answering the generated questions, regardless of the final outputs. Several users indicated that reading and answering one or more of the questions made them think about their needs in a new way.

Simple questions were not valuable. Conversely, some of the questions were not insightful or thought provoking, but merely information-gathering, such as “what is the name of your organization?” Users did not find this valuable, indicating that they could fill this information in after the document was generated, and that there was no benefit to entering the information before the document generated as opposed to afterwards. This was especially clear because the baseline document often included placeholder tags such as [your name], [organization name], etc., and users could clearly see that this information could have been entered upon editing the generated document.

The baseline documents introduced novel ideas. A common reason users gave for ranking the documents equally was that while the QA document more closely adhered to their instructions, the baseline document often introduced new ideas or insights that the user had not specifically requested. Some of these embellishments were factually incorrect hallucinations. However, in other cases the baseline document generated content that the user had not thought about or requested, but which was not factually inaccurate, and which introduced new ideas and perspectives which users found valuable.

The QA document output rigidly followed instructions. Conversely, users felt that the QA document closely followed the instructions they had given in the prompt and question answers, but introduced comparatively fewer new ideas and less original content. In several cases the QA document would repeat phrases entered by the users verbatim. In the worst cases, users felt that the QA document was merely regurgitating what they had already entered, which was considered less valuable than the AI providing original content and perspective.

Users wanted to refine their documents further. Users who had prior experience with generative AI were disappointed that they could not continue refining the generated documents with additional instructions.

Users wanted an easier way to generate more documents. Several users wanted to repeat the study and requested a way to use the system without having to answer all the survey questions. Users were generally disappointed to learn that there was no “app” version of CQDG which could be used for document generation outside of the study. Additionally, users wanted an easier way to restart the study to generate additional documents, even if that meant re-answering the survey questions.

Users found it difficult to give documents an objective rating. Several users said that they were not sure how to rate each document, particularly the first document, since quality was relative and they had no point of comparison.

Overall positive reception. Most users expressed that the experience of working with the AI was surprisingly enjoyable. Two users wanted to immediately repeat the study, and several others requested permission to continue using the system on their own time. Users were not prompted to repeat the study, and no reward or benefit was offered for continued engagement. Users consistently expressed that the system was fun to work with and an improvement over their previous experiences with AI.

Research Notes

In addition to feedback from the users, I took detailed notes of each observed session.

Database Connection Error. As noted above, one of the user’s results were not properly recorded due to a database connection error.

Users had feedback throughout the process. Rather than wanting to give all their feedback at the end, users had a lot to say about each step of the process, so including opportunities to give written feedback at each step of the process could be valuable.

Users were surprised by the document quality and speed. The group of users who completed the pilot study had a mix of experience with generative AI. Those who had little to no experience with modern generative AI were surprised and impressed by both the speed with which documents were produced, and the overall quality. Several even wanted to confirm that the computer had generated the documents without human intervention, but conceded that there was no possible way for a human to have read their answers and written complete documents in just the few seconds it took CQDG to generate the documents.

The document rating system was not sufficiently sensitive. Most users felt that both documents were “Somewhat close to what I was hoping for,” rated as a 4 out of 5 in the document rating system provided. Even when users expressed a verbal preference, this was rarely reflected in the numerical scores that were given.

Users saved their documents. Most users took the time to save their documents after reviewing them. When the user completed the survey on a computer other than their own, they requested that the saved documents be sent to them. This was despite a somewhat clumsy interface for saving the documents and a fair amount of confusion and instruction needed regarding how to save the documents and where they were saved to. This indicates that users found real value in the generated outputs, or else they would not have gone through the trouble and inconvenience of ensuring they could keep copies for their own use.

Users gave thoughtful responses to questions. I was impressed that users slowed down and took the time to read each question and answer thoughtfully, in some cases typing multiple-paragraph long answers. Even when users were somewhat rushed at the beginning of the survey, they all stopped to consider the questions, slowing down considerably and engaging thoughtfully with the process. I was initially concerned that users would brush off the questions or rush past them, so I was surprised and impressed to observe that nearly everyone who engaged with the system slowed down and gave thoughtful answers.

Pilot Study Discussion

Despite discouraging quantitative results in the document ratings, the exit survey and user feedback indicate that users are highly receptive to the question-asking interface and found qualitative benefits to the use of the system over a more traditional generative AI interface such as ChatGPT for the purpose of document generation. The nearly identical results in the document ratings was indicative of several flaws in the study design:

1. The ratings system was insufficiently sensitive.
2. The ratings system did not directly ask users to state a preference between the documents.
3. The ratings system asked users to rate the first document before seeing the second document.
4. The QA document consistently generated output which included less original content and often repeated users’ words verbatim with little to no unique contribution.

The first three issues can be addressed by improvements to the study design. The issue of rigid and unoriginal QA document outputs can potentially be addressed through prompt engineering.

Aside from the document ratings, the user feedback and engagement convinced me to rethink the primary value of the system. I had originally designed the system from the mindset that question-asking could be used as a form of ambiguity resolution. However, users consistently said that simple information-gathering questions were less valuable than questions that made them think. Users valued insightful and engaging questions, and the system was clearly capable of generating such questions even though it had not been specifically designed with this in mind. Recognizing this, I decided to modify the prompt engineering of the question generation to encourage these kinds of questions, and to focus on the value of asking good questions beyond the scope of ambiguity resolution.

The overall positive feedback, along with users' eagerness to keep the resulting documents and to continue engaging with the system even after the study had completed, convincingly demonstrates that a question-asking phase prior to document generation is something which users find to be valuable and engaging, regardless of whether it results in superior outputs. Of course, it would be better if the process actually resulted in superior documents as well.

Based on these insights, I developed the following list of requirements for the full study: [7]

1. Allow users to read both documents and then indicate preference for one document or the other, rather than asking users to rate the documents one at a time.
2. Use higher resolution on rating scales. The 1-5 scale proved to be insufficiently sensitive.
3. Refine the prompt engineering of the sequences input to GPT. Ideally, the final output should take participants' responses into account while retaining a degree of originality, without copying participant answers verbatim.
4. Gear questions towards encouraging users to think about their needs in ways they had not previously considered or proposing expansions or alternatives, rather than gathering information that the user could easily enter into a template form (e.g. the name of their organization).

---

[7] List quoted from Tix HCII 2024 paper [133]

5. Provide a way to continue refining the documents after their initial creation. Several participants, especially those with prior experience with generative AI, specifically requested the ability to continue refining the outputs they were given with new instructions.
6. Compare GPT 3.5, GPT 4, and other LLMs. GPT 3.5 was only used in this case for simplicity due to the small number of participants.

Chapter 4: Full Study[8]

After completing the pilot study, CQDG was heavily modified to incorporate the lessons learned from the pilot. While the pilot study was intended to gather information about the study setup itself with the goal of future improvements, the goal of the full study was to gauge whether users found a benefit to the question-asking system or not, including both whether users found there to be a difference in the quality of the output, as well as whether users found qualitative benefits to the question-asking interface. Screenshots of every step of the full study are shown in **Appendix F**. The results of this study are currently undergoing peer review and have not yet been published, but a pre-print of the proposed paper is available on ArXiv [137].

Improvements to Study Design

There were several key differences in the study design between the pilot and the full study.

1. In the full study, participants interacted with CQDG anonymously online. There was no direct contact between myself and the participants, and participants were not observed completing the study.
2. The baseline and QA documents were shown side-by-side instead of being shown one at a time on separate steps. The order of the documents was still randomized, so sometimes the QA document would be on the left, and sometimes it would be on the right.
3. Users were asked to indicate their preference for either the QA or the baseline document, rather than giving a rating to each document separately as they did in the pilot.
4. Users were asked to rate their preference for the documents with two separate questions:
    a. *Which document do you prefer overall?*
    b. *Which document would be more useful to you in its current state?*
5. The question *"How close is this document to what you hoped for when you made your initial request?"* was removed. This was due to feedback from the pilot study indicating that in

---

[8] This experiment design was approved by the University of Hawaii Institutional Review Board, Protocol ID: 2024-00193. Dr. Kim Binsted is listed as the Principal Investigator, in accordance with University of Hawai'i policy regarding PhD research. Approval was issued on 4/4/2024 and the experiment ran from 4/9/2024 through 6/19/2024.

some cases, users valued documents that deviated from their original intentions, if those deviations made positive and original contributions to the final document.

6. Users were given the opportunity to refine their documents after the initial document generation. Users could enter up to three refining prompts per document. After all refining was complete, users were asked to rate their preferences again for the new documents. This step was optional, and users who did not refine their documents were not asked to give a second set of ratings.
7. Users were given the opportunity to complete the study again by generating additional documents. This process was made seamless, with a link at the end of the study that would allow users to skip over answering the consent and demographics questions and go straight to creating a new document. This was only possible for users who had already gone through the study once and read and agreed to the consent already, and who had already answered the demographics questions.
8. The demographics questions section was split into two steps, so that users who indicated they were under 18 or were not fluent in English could be screened out before being shown any additional content.
9. An additional question was added to the exit survey: *"Answering the questions asked by the AI made me think about my request in ways I hadn't previously considered."*
10. The consent message was updated to reflect the changes to the study. The consent message shown to participants in the full study is provided in **Appendix G**.

Screenshots of every step of the process can be found in **Appendix F**.

Improvements to Technical Design

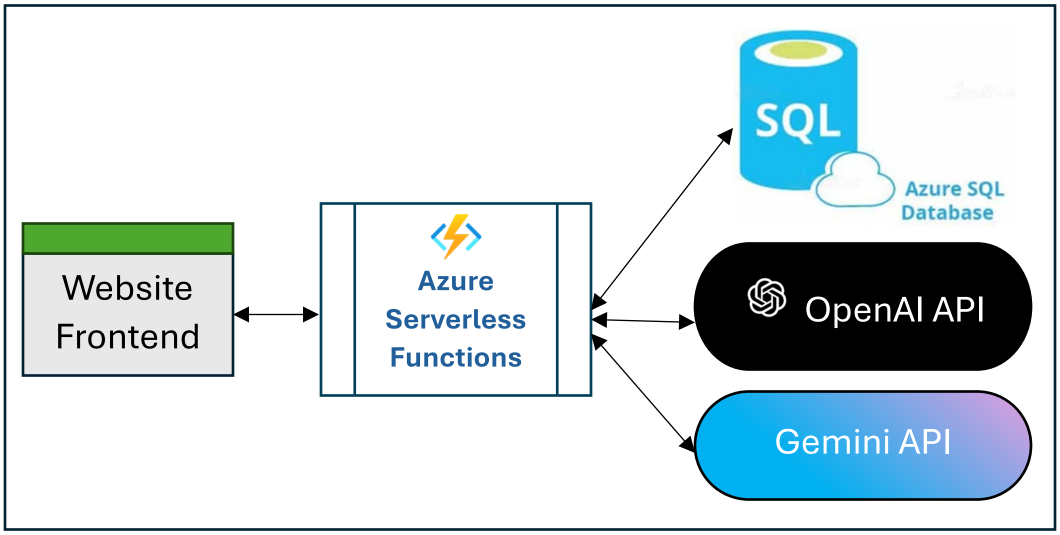
While the overall design of CQDG remained the same, the system was expanded and improved in several ways.

Inclusion of Multiple LLMs

The pilot study relied exclusively on GPT-3.5 for text generation. For the full study, GPT-3.5, GPT-4 Turbo [138], and Gemini [21] were all used. Whenever a new session was started, CQDG would randomly select one of these three LLMs, and use the selected LLM throughout the document creation process, for both the baseline and QA documents, and for any refining prompts entered by the user after reviewing their initial documents. If a user chose to create a new document, the LLM selection process would run again, potentially selecting a new LLM for the new pair of documents.

Access to the LLM was still managed through Azure Serverless Functions written in Python. GPT-3.5 and GPT-4 Turbo were both accessed using the OpenAI API, and Gemini was accessed through the Gemini API. Both of these APIs can be called either by sending a single prompt, as was done in the pilot, or by sending an array of prompts with alternating roles of "user" and "assistant." This feature allows a full conversation to be stored and sent, allowing the LLM to be used like a chatbot. The full study version of CQDG took full advantage of this feature to improve the LLMs' ability to respond to the full context of the user's request, question answers, and refining prompts. Responses and prompt logs were stored in the same SQL database that was used for the pilot. The expanded technical design is shown in **Figure 12**.

Figure 12: Full Study Technical Design



Improved Prompt Engineering

The instructions given to the LLM for the generation of the questions and the documents were modified between the pilot and the full study. The instructions given when generating questions have been reworded to encourage more thoughtful and engaging questions. The instructions given when generating the QA document were reworded to encourage the QA document to retain some degree of originality rather than simply regurgitating what the user had already entered, which was a common complaint during the pilot study. The changes to prompts between the pilot study and the full study are detailed in **Table 1** below, where *{Request}* is the initial request entered by the user.

Table 1: Prompt Engineering Changes between Pilot Study and Full Study

<table>
<tr><th colspan="2">Identifying Ambiguity</th></tr>
<tr><th>Pilot Study</th><th>Full Study</th></tr>
<tr><td>You are a helpful AI assistant used to generate short documents. A user is requesting the creation of a new document. This is their request:
user: “*{Request}*”
Identify any areas of significant ambiguity or necessary information that has not been included, and write these out in a short list. Include exactly 3 items in the list.</td><td>A user is requesting the creation of a new document. This is their request:
user: “*{Request}*”
Identify any areas of significant ambiguity in the prompt, areas that could benefit from more thought or attention from the user, or helpful tips the user may not have considered. Write these out in a short list.</td></tr>
<tr><th colspan="2">Generating Questions</th></tr>
<tr><th>Pilot Study</th><th>Full Study</th></tr>
<tr><td>Respond as though this request was just made by the user. Start with a friendly greeting, then ask a respectful and concise question to clarify the user's needs, keeping in mind the ambiguity you just identified. Format your response as a numbered list of exactly 3 questions.</td><td>Pick the three most important items from the list you just generated, and write a list of three insightful questions that will improve the requested document. Phrase the questions as direct questions to the user. Format your response as a numbered list of exactly 3 questions.</td></tr>
<tr><th colspan="2">Generating the QA Document</th></tr>
<tr><th>Pilot Study</th><th>Full Study</th></tr>
<tr><td>Consider the following exchange. Attempt to create the document requested by the user, considering the answers they gave when asked for details.
{QA Transcript included in prompt}</td><td>Generate a high-quality document that meets the user’s needs, considering both their initial prompt and the answers they gave when asked for details. Include creative original insights that will improve the quality of the document but do not deviate too far from the user's original intent.</td></tr>
</table>

Additionally, the full study made use of OpenAI's "System Messages" feature, which allows users to specify system-level instructions that apply to the entire conversation. For the QA Document, the following system message was used:

> *"You are a helpful assistant designed to help users create short, high-quality documents by asking insightful questions to clarify the users' needs and make them think about things they have not considered, and then create high-quality professional documents after discussing the details with the user."*

For the baseline document, a simpler system message was used:

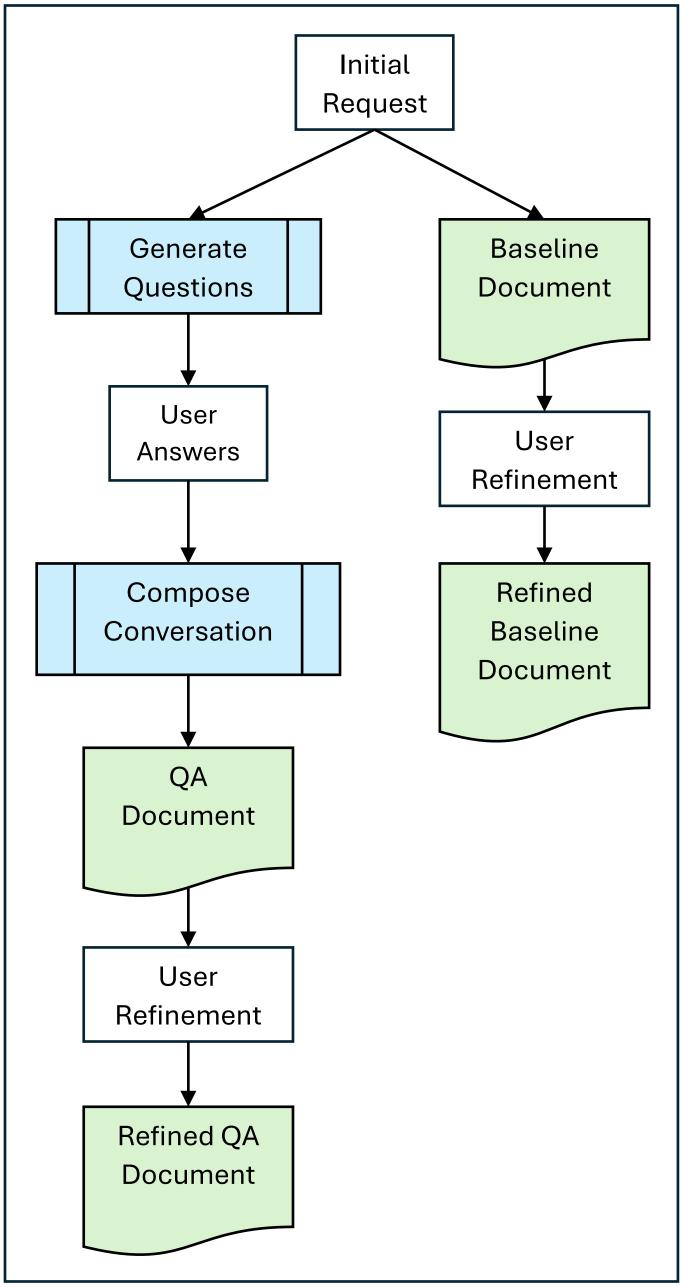
> *"You are a helpful assistant designed to help users create short, high-quality professional documents."*

While developing CQDG, I initially used the same system message for both the QA and baseline documents. However, while testing the system I found that when given the full system message, the baseline document would sometimes generate a list of questions rather than a document. For this reason, a shortened system message was used for the baseline document, which did not mention asking questions.

The system messages worked well for GPT-3.5 and GPT-4 Turbo, but the Gemini API did not include a system message feature. Within Gemini, all messages had to have a role of either "user" or "model," with "model" messages representing the LLM responses. In order to get around this limitation while using Gemini, any time a system message would normally be given, a pair of messages was appended to the conversation instead: one "user" message containing the instruction that would have been a system message for GPT, followed by a spoofed "model" message that simply said "OK!" indicating the model acknowledging the instruction. The additional "model" message was required because the conversation context for Gemini needs to alternate between user messages and model messages, whereas system messages in GPT are followed by user messages, so without the model message of "OK!" the result would have been two user messages in a row, which is not allowed in the Gemini API.

Once the user had submitted their answers to the three generated questions, the conversation would be internally re-arranged by CQDG so that the context sent to either the OpenAI or Gemini APIs would make it appear as if the questions and answers had been asked and answered sequentially, as it had appeared to the user. This was also done in the Pilot Study, by including the conversation transcript as the user had experienced it in the document generation prompt. This re-arranging of the context helps the LLM to respond in a natural way to the conversation as the user has experienced it, and based on my own testing tends to produce better and more consistent responses. When this is not done, the LLM more often has trouble parsing the users' answers or generates output which is not an attempt at writing a document, such as an outline, further questions, or text about the document writing process rather than generating the document that was requested. A visual representation of the document generation process is shown in **Figure 13**. An example of a full conversation with CQDG, including all of the messages sent to and received from the system, is included in **Appendix H**.

Figure 13: Full Study Document Generation Pipeline



A More Robust Database Connection

As noted in the *Pilot Study Results* section, there was an instance during the pilot study in which a session's data was lost due to a bad database connection. I found that the Azure database used by CQDG often failed to connect on its first use of the day. If the system had been idle for several hours or more, it could take up to several minutes and several attempts for a successful connection to be established. Unfortunately, I was not able to resolve the slow connection issue. To compensate for this, the database connection code for CQDG was made more robust, such that the system would continuously attempt to establish a database connection while the user was reading the consent agreement. Multiple attempts would be made until a successful database connection was established and confirmed by CQDG. If CQDG was still attempting to connect after the user clicked "continue" on the consent screen, a "please wait" screen would be shown until the connection was established successfully. This solved the problem of lost data.

Participant Selection

Participants were invited to take the study via social media promotion. This included a paid Meta Ads campaign [139] which advertised the study on Facebook as well as free postings on the Reddit community r/SurveyExchange [140]. The study was additionally advertised in two of Dr. Martha Crosby's classes in April 2024. One class, *ICS 464 Human Computer Interaction I*, was offered extra credit to complete the survey, since it was directly relevant to their coursework. The other class, *ICS390 Computing Ethics for Lab Assistants*, was informed of the survey but was not offered extra credit.

Users requested 89 unique documents from CQDG. Fourteen (14) of these exited before completing the survey, leaving 75 who completed the full study. Incomplete surveys were not included in the results. Four (4) completed surveys had to be excluded as well. The four exclusions include:

- A user who requested an obscene document which Gemini refused to produce. [9]

---

[9] Users were given instructions at the start of the study that they should not submit obscene requests, and that such requests may be refused by the AI.

- A user who stated that they did not want CQDG to produce anything and reiterated this instruction when answering the follow-up questions.
- Two technical glitches in which CQDG failed to produce one of the two documents.
    - In one case, GPT-3.5 refused to produce the baseline document with the response "I'm sorry, I can't assist with that request." It's not clear why the request was refused, and the QA document was produced correctly.
    - In the other case, Gemini responded incorrectly to the QA document prompt, and replied "Please provide me with: * The user's initial prompt * The user's answers to the questions you asked. Once I have this information, I will be able to generate a high-quality document that meets the user's needs." This is a puzzling response since in all other cases Gemini correctly responded to the provided instructions by producing a document.
    - Both of these cases serve as examples of the inconsistent output of LLMs generally. Even instructions which work most of the time may occasionally be interpreted incorrectly by the LLM.

Accounting for the four exclusions and the 14 non-completed surveys, 71 completed surveys remain for analysis. These 71 surveys were completed by 65 unique individuals, the remaining 6 documents being second documents created by users who chose to produce more documents after completing the survey.

Of the 65 unique respondents, 33 reported being female, 25 male, and 7 reported "other / nonbinary" for their gender. The respondents ranged in age[10] from 18 to 64, with an average age of 31. Five (5) users reported having never used generative AI before, 35 said they had used generative AI before, but not often, and 25 reported being regular users of generative AI. In 36 of the responses, users chose to refine one or both documents with additional instructions after giving their initial ratings.

As shown in **Table 2**, women and "other/nonbinary" participants indicated having less experience with generative AI than men. The majority of male participants indicated that they were regular

---

[10] Note that respondents younger than 18 were not allowed to complete the study.

users of AI, while this was a minority response for non-male participants. Most users indicated that they had some experience with AI but were not regular users. Experience was also correlated with age, with younger participants being more likely to indicate that they were regular users of AI, and users with little to no experience skewing much older than the average for the group.

Table 2: Participant Experience, Age, and Gender Distribution

	Little Experience	Some Experience	Regular User	Average Age
Female	3	19	11	30.9
Male	2	10	13	33.5
Other / Nonbinary	-	6	1	26.1
Average Age	44.8	33.1	26.3	

Full Study Results

Prompts Entered

Users entered a wide variety of prompts, varying greatly in both tone and length. Some prompts were silly, such as:

> *"Send my cat a divorce letter."*

Others were serious and somber, such as:

> *"Write an email to a friend I haven't spoken to in 20 years because of a fight saying I want to repair the relationship."*
>
> *"Write an email to a friend who has a chronic disease and is beginning treatment."*

Most prompts followed the general guidelines given, keeping the requests short and professional, such as:

> *"Please write a brief proposal for designing the next generation of ballpoint pen."*
>
> *"Please write a one page memo detailing Canada's history with the United Nations Convention on the Law of the Sea."*
>
> *"Write a report on the potential impacts of AI on Information Technology Service Management"*

Some asked for poems or songs, such as:

> *"write a poem about how nature is more holy than a church"*
>
> *"Please write a remix of the song 'Bohemian Rhapsody' by Queen and make it instead about failing my exams (instead of killing a man)"*

Some prompts were long, with highly detailed instructions, such as this example:

> *"Please write a letter to parents of high school students who are members of a student robotics club. The letter should explain to parents that the club will be forming a board, and the club is looking for parent volunteers to serve on that board. The letter should be brief while maintaining a tone that is semi-formal and upbeat / enthusiastic. The letter should set a specific date for the initial board meeting and encourage parents to attend this meeting if they are interested in serving or would like to know more about club management and future plans."*

Some even included specific technical requirements for the AI, such as this puzzle-like request:

> *"Please write a one-paragraph visual story that meets the following constraints: (1) it has a total of 7 sentences, (2) for each sentence, the main verb is transitive, and (3) for 3 out of the 7 sentences, the main verb should be used in a metaphorical way, whilst the main verb should be literal for the remaining 2 out of 7 sentences. The story can be about anything, but should ideally involve human characters. Good luck!"*

Document Preferences

Users were asked two questions about the produced documents: "*Which document do you prefer overall?"* and "*Which document would be more useful to you in its current state?"* These questions were asked both before and after users were given the opportunity to refine both documents with additional instructions. **Table 3** shows the responses to these questions.

Table 3: Preference Results

	Before Refining			**After Refining**		
	Prefer QA Document	**Prefer Baseline**	**No Preference**	**Prefer QA Document**	**Prefer Baseline**	**No Preference**
Overall Preference	41	25	5	18	13	5
More Useful	40	22	9	13	12	11

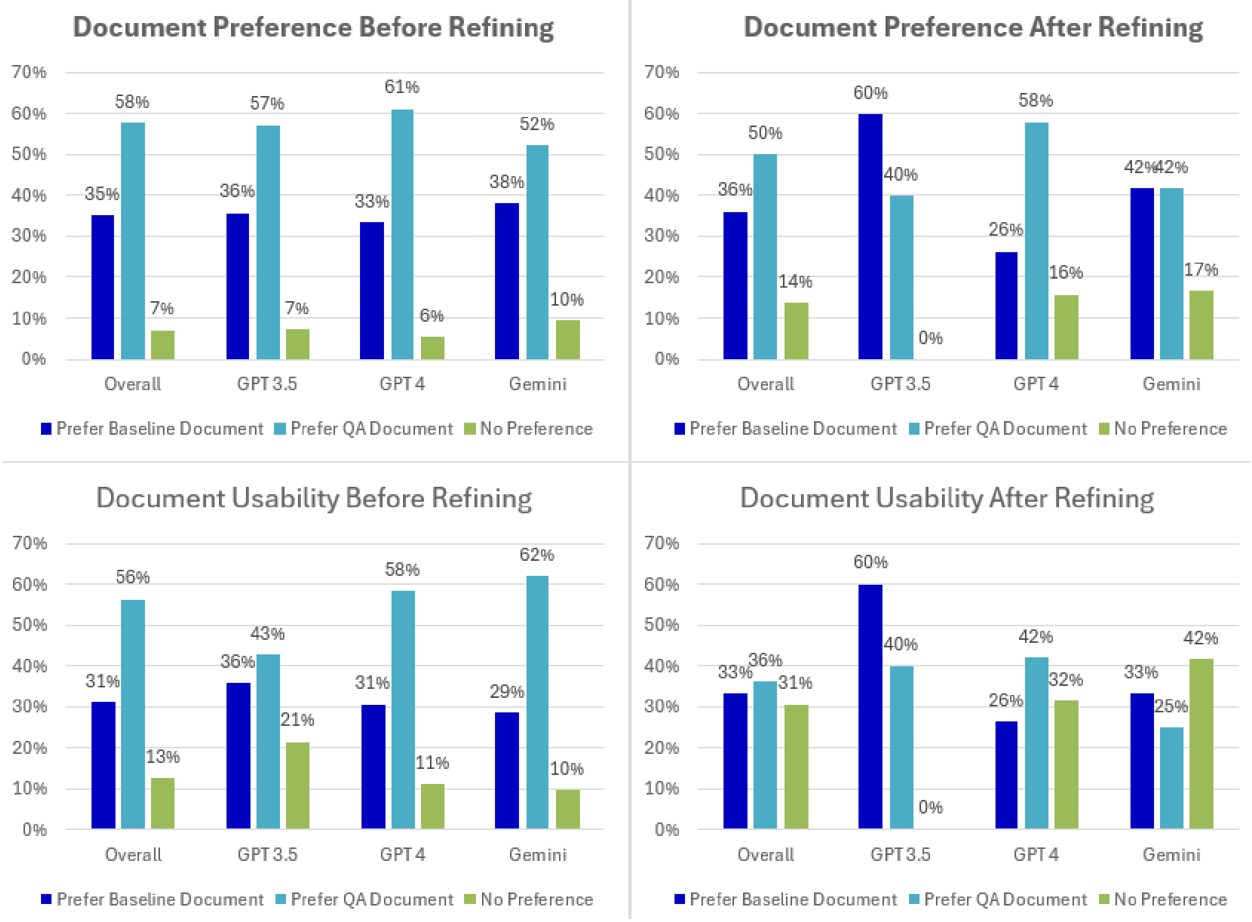
The preference data was tested for significance using a binomial distribution test. The binomial distribution test requires that there be only two possible categories, so "No Preference" responses were distributed equally between "QA" and "Baseline" responses. This can reduce the sensitivity of the tests, but is less likely to skew the results than either proportional distribution or excluding the "no preference" responses [141]. Furthermore, in this study a response of "no preference" represents equal preference for either option, thus distributing the "no preference" responses equally between the available options is more appropriate than discarding these responses. Since the binomial distribution test requires a whole number of responses, in cases where there were an odd number of "no preference" responses, one "no preference" response was discarded to avoid fractional values.

For both questions, the preference for the QA document was significant at the $p<=.05$ level before document refining, but the difference after refining the document was not significant, as shown in **Table 4**. It is worth noting that since the document refining step was optional, the number of responses after refining is lower than before refining (36 vs 71) and this is a contributing factor to the lower significance. However, the lower sample size is not the only factor in the reduced significance. The difference in user preference between the QA document and the Baseline document is clearly reduced between the non-refined document and the refined document, as shown in **Table 3** and **Figure 14**.

Table 4: Binomial Distribution Test of Significance (Significant results are highlighted)

p values under Binomial Distribution Test	**Before Refining**	**After Refining**
Overall Preference	*0.028*	*0.155*
More Useful	*0.016*	*0.368*

Figure 14: Impact of LLM on user preference



Results by LLM Selection

Figure 14 shows the effect of LLM selection on users' document preference. When using GPT-4, users indicated a strong preference for the QA document both before and after refining, by both the metrics of Overall Preference and Usability. However, GPT-4 is the only LLM with which the QA document retains a lead after document refining. For Gemini and GPT-3.5, the preference is reversed after refining, with users indicating a slight preference for the baseline document after refining when using Gemini, and a strong preference for the baseline document after refining when using GPT-3.5.

A notable outlier in these results is that the preference for the QA document after refining is much higher if the documents were produced by GPT-4 compared to the other LLMs. The preference for the QA document remains significant after refining if the documents were produced by GPT-4 for the question "*Which document do you prefer overall?" ($p = 0.048$)* but is not significant for either of the other LLMs. A preference for the QA document can also be seen for GPT-4 for the question "*Which document would be more useful to you in its current state?"* which differs from the other two LLMs, for which the baseline document was preferred after refining for this question. However, for "*Which document would be more useful to you in its current state?"* the QA document preference is not statistically significant for GPT-4 ($p = 0.18$).

Results by Demographic Factors

It is also interesting to look at how user preference varied based on the user demographics. Recall that at the start of the study, users were asked to give their gender, age, and level of prior experience with generative AI. There is not sufficient sample size in this study to fully explore the effects of demographic factors on user preference. While some individual results in this section may be significant on their own, I have opted not to include tests of significance for the preference results broken down by demographic factors. This decision was made for several reasons:

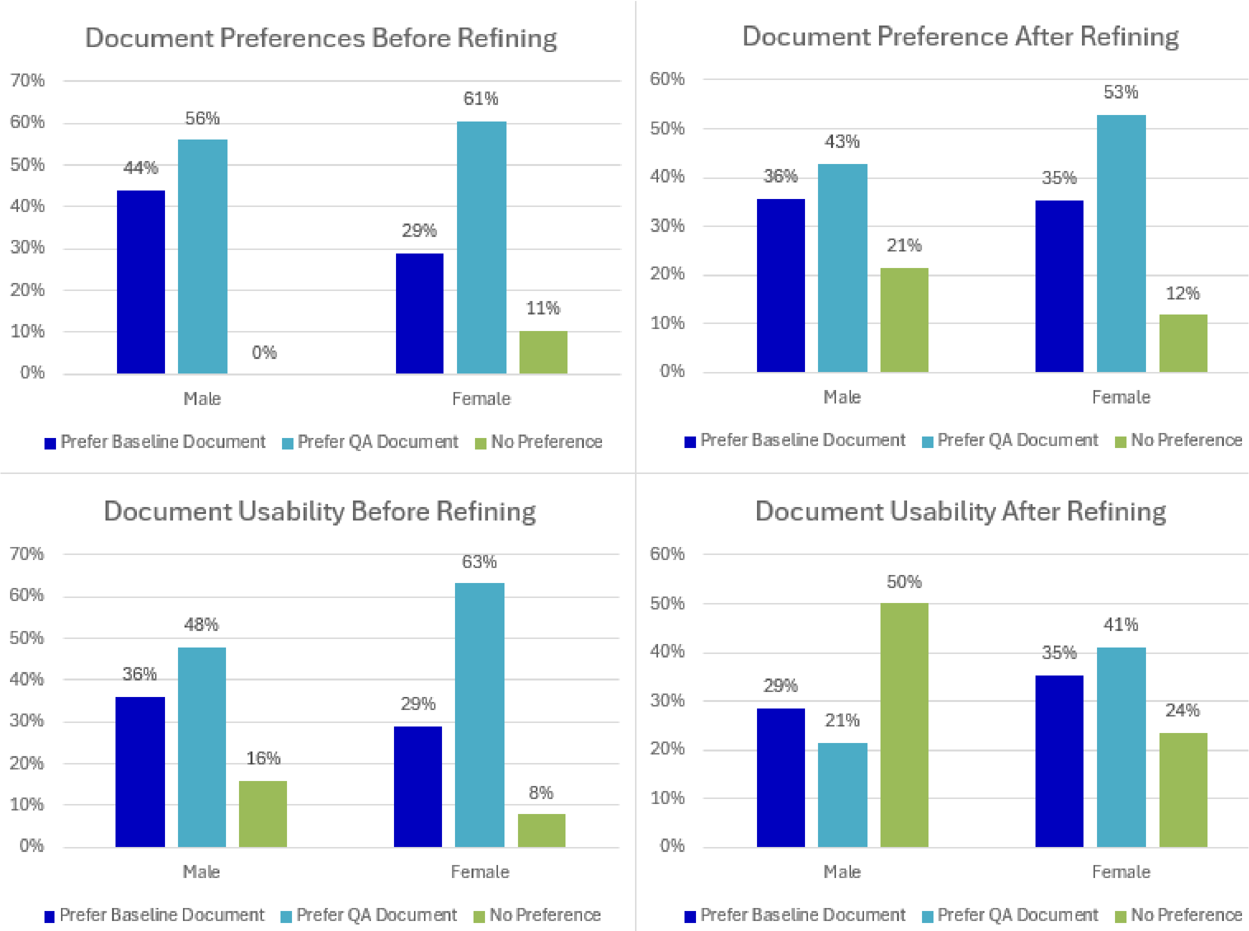
1. The distribution across various demographic groups is neither evenly distributed nor reflective of the overall population of the US or of the world. The participant group includes more females than males, a younger average age than the US population, and a relatively high level of prior experience with AI. Thus, any conclusions drawn about this participant group would not necessarily be applicable to the general population. This is discussed in more depth in the "Limitations" section.
2. Because of the uneven distribution, many demographic groups have very small sample size which is not appropriate for tests of significance.
3. Testing multiple hypotheses simultaneously increases the probability that some of the tested hypotheses will result in a type I error, also known as a "false positive" [142]. With a significantly larger sample size, this could have been compensated for by reducing the target p values. However, this study was sized specifically to target testing only the overall preference and LLM, and testing each demographic factor with sufficient statistical certainty would require a much larger sample size which is not currently available.

Despite these limitations, the results vary noticeably between different demographic groups, and these variations can provide valuable guidance for future studies which can be better targeted at proving whether a true effect exists or not between different segments of the population. The graphs and values shown in the following sections should be interpreted in this light, intended for descriptive and speculative purposes, not as rigorous statistical proofs of an effect in the general population.

Results by Gender

Women exhibited a greater preference for the QA document than men, as shown in **Figure 15**. Both men and women prefer the QA document before refining, though women's preference is stronger than men's. After refining, both men and women said they prefer the QA document overall, though by lower margins. After refining, men found the baseline document to be more usable, while

Figure 15: Impact of Gender on Document Preference



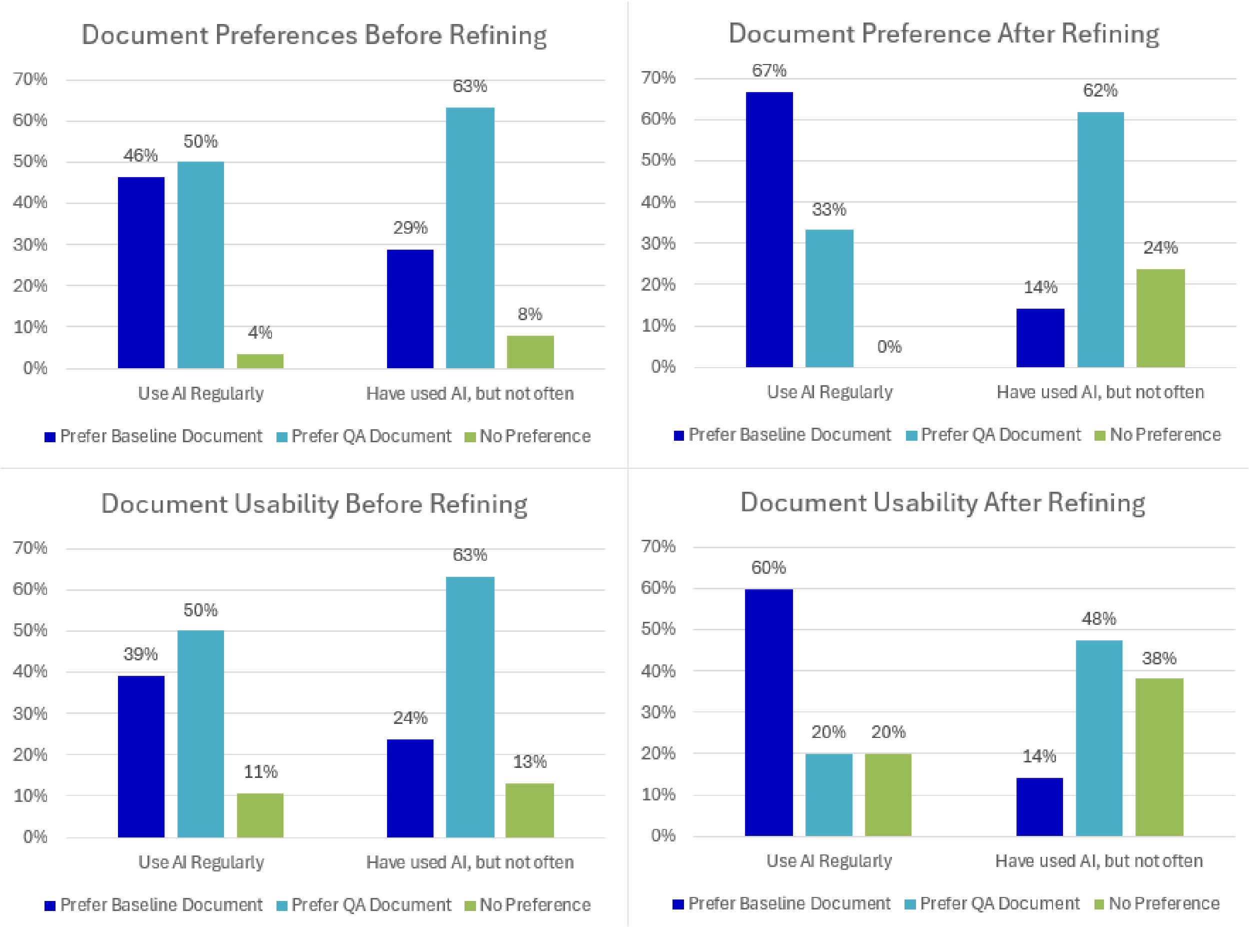
women still slightly preferred the QA document. Users who selected “other/nonbinary” were excluded from the gender comparison due to the low number of users who selected this option.

Results by Experience Level

Figure 16 shows the impact of prior experience with AI on users’ document preference. Five (5) users reported having never used generative AI before, 35 said they had used generative AI before, but not often, and 25 reported being regular users of generative AI. Six (6) users chose to create and rate a second document after they had finished with the first, including 3 regular users of AI and 3 who indicated that they had used AI before, but not often. This results in a total of 38 documents rated by users who had used generative AI before, but not regularly, and 28 documents rated by users who reported using AI regularly. There were only five (5) users who indicated having never used generative AI before, and none of them continued on to create a second document or chose to refine the document that they had created with additional prompts. Due to the small sample size and lack of refining prompts from the group who had never used generative AI before, those five users are excluded from the comparison shown in **Figure 16**.

Between users who are regular users of AI and those who have used AI before, but not often, there is a clear trend that those with less experience expressed a greater preference for the QA document when compared to more experienced users. Those who had used AI before, but not often, show a clear preference for the QA document in both overall preference and usability, both before and after refining. By contrast, regular users of AI showed a small preference for the QA document before refining, but a large preference for the baseline document after entering their own refining prompts.

Figure 16: Impact of Prior Experience with AI on Document Preference



Document Preferences Before Refining

Document Preference After Refining

Document Usability Before Refining

Document Usability After Refining

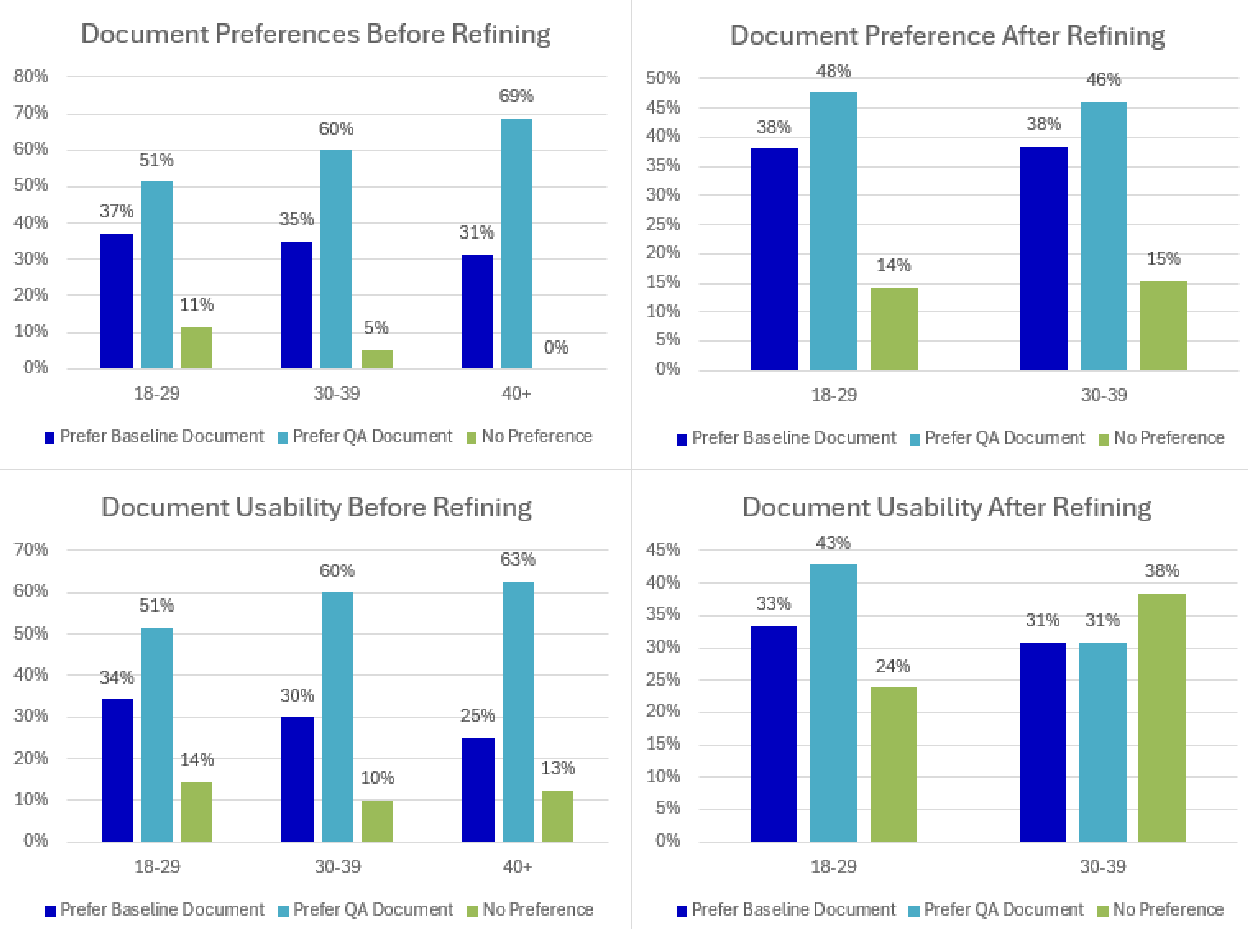
Results by Age

Users ranged in age from 18-64. Each user entered an integer value for their age as a part of the demographic questions at the start of the study. Age was broken into three categories for analysis, 18-29, 30-39, and 40+. These groupings were determined by beginning by grouping age by decades (18-19, 20-29, etc.). However, 18-19 was too narrow of a grouping, and there were only 6 users over 50, so 18-19 and 20-29 were merged into one category, and the 40-49 category was expanded to 40+. **Table 5** shows the total number of documents produced and rated by each age category. Participants over 40 produced 16 documents but only entered additional refining prompts for 2 of those documents. Due to the low sample size, the 40+ category was excluded from analysis for the "after refining" measures. **Figure 17** shows the impact of age on document preference. Before refining, older users were more likely to prefer the QA document when compared with younger users. After refining, this effect disappears and preference is split fairly evenly between the baseline and QA document for both the 18-29 age range and the 30-39 age range.

Table 5: Documents Produced by Age Category

Age Category	Documents Produced	Documents Refined
18-29	35	21
30-39	20	13
40+	16	2

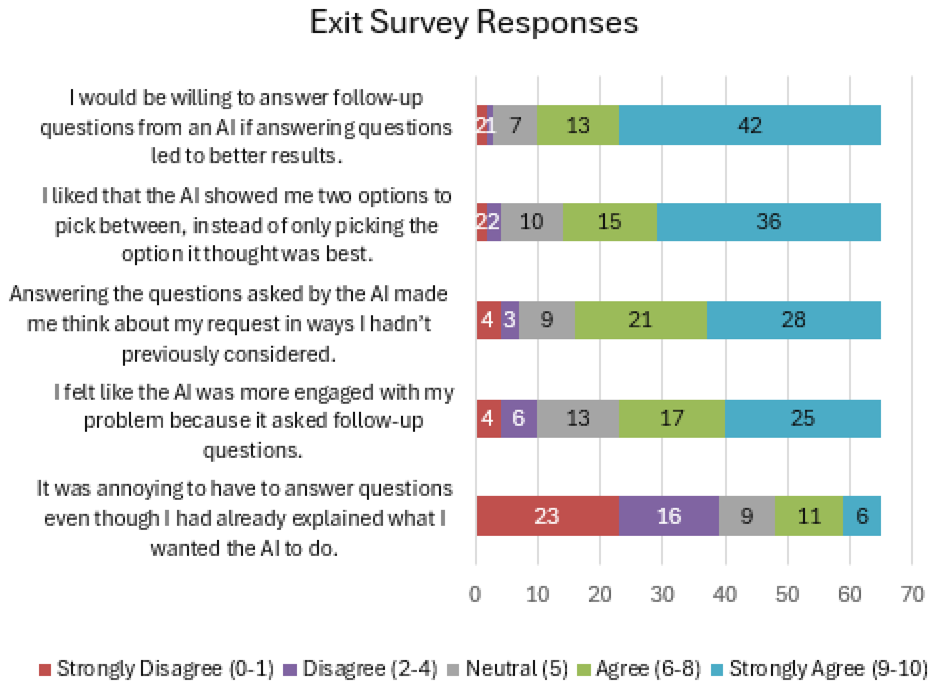
Figure 17: Impact of Age on Document Preference



Exit Survey

Overall, users responded positively to the experience of using CQDG, as shown in **Figure 18**. Since users who chose to create multiple documents were only given the exit survey once, 65 exit surveys were collected, one per unique user. Most users agreed or strongly agreed that they would be willing to answer follow-up questions if doing so led to better results (55 users, 84.6%), that they liked seeing two options to pick from (51 users, 78.4%), that answering the questions made them think about their request in ways they hadn't previously considered (49 users, 75.4%), and that they felt like the AI was more engaged with their problem because it asked follow-up questions (42 users, 64.6%). Most users disagreed or strongly disagreed with the statement "It was annoying to have to answer questions even though I had already explained what I wanted the AI to do" (39 users, 60%).

Figure 18: Exit Survey Responses



User Feedback

Users were given six separate opportunities to enter free-text feedback while participating in the study:

- After answering questions but before seeing either document.
- Feedback on the initial baseline document.
- Feedback on the initial QA document.
- Feedback on the refined baseline document.
- Feedback on the refined QA document.
- After completing the exit survey.

The top five repeating themes from each of these sets of feedback is shown in **Table 6**. A complete listing of all user feedback received is provided in **Appendix I**.

Overall, users were impressed with the quality of the questions being asked and found the process of answering the questions to be thought-provoking and valuable in its own right, even before seeing the produced documents. One user stated:

> *"I would find answering these question prompts valuable, even if I were still writing the letter myself."*

Another noted that the questions themselves might be relevant to include in the document they were trying to create, stating:

> *"These questions could actually be a part of the email that I should send my supervisor."*
> One user commented that the questions increased their confidence in the system, saying *"Good questions from the AI. It makes me feel confident that the AI will write a better letter than other AI chatbots."*

Many users expressed finding the experience of creating documents with CQDG to be engaging and noted that the experience increased their level of interest in and understanding of AI, with comments such as:

> *"This was a fantastic study that helped me understand Generative AI better than previously,"*
>
> *"Overall this would improve my experience with generative AI and I would use it more if this was an option"*
>
> *"This is what Clippy should have been."*

Several users stated that after the refining process, the two documents improved in quality and were more similar in quality than they had been before refining, with comments including:

> *"Pretty good after the revisions."*
>
> *"document 2 wasn't very good before the refinement but it's almost caught up with document 1 in quality now."*

This is also reflected in the preferences users reported as shown in **Figure 14** and **Table 4** as described above, in which the preference for the QA document is significant before the refining process, but the difference becomes insignificant after refining.

However, several users noted frustration with the refining process and an inability to get the document to come out the way they envisioned it, while noting that the initial QA document made a better starting point for manual revisions when compared to the initial baseline document, with comments such as:

> *"Refining it made it more cold and weird"*
>
> *"Document 2 is not something that would do well in committee at Model United Nations and I had a difficult time getting the AI to increase its specificity beyond what was present in the original document 2."*

One user with prior generative AI experience compared the process of answering questions favorably with the more usual method of continually refining prompts, stating:

> *"I think the questions helped a lot with the documents landing closer to where I wanted them. When using these kind of programs (that aren't asking questions) I tend to ask for revisions or restart the prompt but it usually takes a lot longer than this did to get something usable or that I'd then use and alter for the final product."*

Users were not informed that one document was utilizing their questions and answers and the other was not. One user helpfully suggested:

> *"It would be helpful to compare results to those that did not use follow up questions."*

Others seemed to pick up on the difference between the baseline and QA documents, at least implicitly, with comments on the baseline document such as:

> *"This document does not take adequate note of the answers given to the follow-up questions."*

> *"It didn't take into consideration the extra information provided to the AI via its asked questions and isn't as strong."*

Users also noted more hallucinations in the feedback for the baseline document, stating things like:

> *"...I never said it was a costume party."*

> *"The template added things that I didn't state such as a weather forecasting system project"*

> *"It is way too flowery and it sounds like I'm going to be missing a month of work rather than a few days (considering I said I would be back next week and the day I am doing this survey is Thursday, so I would be missing 2 days). It's just excessive."*

Table 6 below shows the top five themes for each of the six feedback boxes in the study. **Appendix I** has the full list of user feedback from each of the six areas.

Table 6: Top 5 themes for each category of free-entry feedback

After Answering Questions	- The AI came up with good questions. - Reading and answering the questions was valuable in itself. - Answering the questions was thought-provoking. - Users offering suggestions for other questions the AI could have asked. - Users attempting to provide additional instructions to the AI using the feedback box.
Initial Baseline Document	- The document was of a high quality. - The content was vague or generic. - The language used was too polite or too flowery. - Users noted hallucinations in the output. - The document's tone was awkward or sounded artificial.
Initial QA Document	- This document is better than the other (baseline) document. - The document was of a high quality. - The document is too long or too wordy. - The AI failed to follow specific instructions given by the user such as length limits. - The document is a good starting point for the user to edit into a final document.
Refined Baseline Document	- Revision improved the document. - The document was of a high quality. - Refining failed to fix the problems with the document / AI failed to respond to revision prompts the way I would hope. - Users noted hallucinations in the output. - This document could be used as-is.
Refined QA Document	- Revision improved the document. - The document was of a high quality. - Refining failed to fix the problems with the document / AI failed to respond to revision prompts the way I would hope. - The AI failed to follow specific instructions given by the user such as length limits. - This document could be used as-is.
Exit Survey	- The study was enjoyable. - Answering questions was valuable to the user and an improvement over the usual process of generating documents with AI. - Users expressing increased interest in AI after participating in the study. - Users had a positive experience interacting with CQDG. - *(Only 4 themes identified due to lower response rate for optional exit survey feedback).*

Overall, users considered the questions themselves to be high quality and overwhelmingly expressed that they found the questions to be valuable. Users appreciated when the generated document took their answers into consideration, and noticed when their answers were ignored. Users valued documents that flowed naturally, but consistently pointed out unnatural tone, flowery language, and hallucinations, especially in the baseline document. Users also expressed broad approval for the process of answering questions, consistently stating that the found the experience to be more enjoyable, efficient, and effective than their prior experiences with AI systems such as ChatGPT that do not ask follow-up questions.

Chapter 5: Discussion

There are several key takeaways from this study that deserve further discussion:

- When is question asking beneficial?
- Qualitative benefits of question asking.
- Qualities of an ideal question asking document generator.
- Cost and efficiency of Question asking.

Each of these points will be addressed below.

When is Asking Questions Beneficial?

The results of these experiments show both that question asking can be beneficial to the document creation process under some circumstances, but that it is not necessarily beneficial in all circumstances. Question asking was most beneficial when the following conditions were met:

- GPT-4 was used to generate the document, as opposed to GPT-3.5 or Gemini.
- The internal prompt used to generate the questions specifically encouraged open-ended, thought-provoking questions.
- Older participants, female participants, and participants with less prior experience with generative AI found more value in the question and answer process when compared with younger and male participants and those who were already regular users of generative AI.

Each of these factors is discussed in detail below.

Use of higher quality LLMs, especially GPT-4.

Before refining, user preference for the QA document was significantly higher when the documents were generated using Gemini or GPT-4, when compared with GPT-3.5. Participants rating documents generated by GPT-4 still considered the QA document to be superior to the baseline by a statistically significant degree even after being given the chance to refine both documents with additional instructions, a trend that did not hold true for Gemini or GPT-3.5.

One possible explanation is that GPT-4 asks better questions than GPT-3.5 or Gemini and thus has better information to work with. However, the data does not support this theory. Participants indicated that they found questions to be thought-provoking regardless of which LLM produced the questions, as shown in **Table 8** below.

An alternative explanation is that GPT-4, and to a lesser extent Gemini, make better use of the questions and responses once they have been received than GPT-3.5 does. This appears to be the more likely explanation. The fact that participants using GPT-4 continued to indicate a preference for the QA document even after being given the opportunity to refine the documents on their own indicates that an LLM that is sufficiently capable of integrating the user's original request and question-answers tends to produce documents which are higher quality, as rated by the users requesting those documents, than the same LLM in the absence of question asking.

Table 7: Though provoking question rating by LLM

LLM	**Average Participant Rating on** *"Answering the questions asked by the AI made me think about my request in ways I hadn't previously considered."* (Scale of 1 - 10)
GPT-3.5	8.0
GPT-4	7.61
Gemini	7.05

Asking the right kinds of question.

Participants in both studies indicated clearly that they valued thought-provoking, engaging questions over simple information-gathering questions. This contradicts the original vision of the study, in which question asking was seen as primarily a means of ambiguity resolution. Participants found it particularly valuable when a question brought up an aspect of their request that they had not previously considered. The questions were most valuable as a vehicle for engaging the user in a dialog with the LLM about their needs, ideally eliciting rich feedback from users in the form of long responses. This gives the LLM far more information to work with. As stated above, making use of this additional information requires an LLM powerful enough to make use of long prompts where the information may not all be organized in an ideal way.

Question Asking was more beneficial to some users than others.

Preference for the QA document was not evenly distributed across participant demographics. Participants' level of prior experience with generative AI had a substantial effect on whether that user preferred the QA document or the baseline document, as shown in **Figure 16**, with less experienced participants preferring the QA document at higher rates than more experienced participants. Gender (**Figure 15**) and age (**Figure 17**) also impacted user preference, with non-male and older participants preferring the QA document at higher rates than younger and male participants. However, age, gender, and prior experience with AI are not fully independent variables. Younger and male participants were more likely to report being regular users of generative AI, as shown in **Table 9**, reproduced below. [11]

Table 8: Participant Experience, Age, and Gender Distribution

	Little Experience	Some Experience	Regular User	Average Age
Female	3	19	11	30.9
Male	2	10	13	33.5
Other	-	6	1	26.1
Average Age	44.8	33.1	26.3	

There is not sufficient sample size in this study to determine whether age, gender, and level of experience with generative AI are independent, or the degree to which they are dependent. However, even if we could state this with a high degree of statistical certainty from this data, this relationship would not necessarily be representative of the general population. The reason for this is that participants were invited from multiple sources, including a social media ad campaign as well as being announced in two university Computer Science classrooms, one of which received extra credit for completing the study. [12] Thus, it is highly likely that a disproportionate number of the younger participants were computer science students, which likely skews younger participants towards more experience with generative AI to a greater degree than would be seen in the general population. This is only speculative, however, as the study was anonymous and did not ask participants how they arrived at the study, so there is no way of knowing how many participants were computer science students.

Despite the uncertainties, the results are clear enough to show that participants with less experience with generative AI preferred the QA documents at higher rates than those with more experience. Answering AI-generated questions could be a promising alternative way of interacting

---

[11] This table was originally shown in the *Full Study → Participant Selection* section but is copied here for easier reference.

[12] As detailed in the *Full Study → Participant Selection* section.

with generative AI for less experienced users, as well as older and non-male users, though further studies with a greater focus on answering these questions would be required to say which groups benefit the most with statistical certainty.

Qualitative Benefits of Question Asking

There are several qualitative benefits to a question-asking AI that are apparent from user feedback but are not fully reflected in the document preference statistics. As reflected in the full study exit survey, most participants felt that reading and answering the questions given by CQDG made them think about their request in ways they had not previously considered. Participants also indicated that they felt the AI was more engaged with their problem because it asked follow-up questions. Multiple users from both the pilot study and the full study indicated that the questions generated by CQDG were valuable by themselves, regardless of the final document quality. Some indicated that the questions themselves could be used as part of their document. Others indicated that the thought processes prompted by the questions were valuable even if they had to then write the document themselves. Multiple users commented in the exit feedback that they found the process of reading and answering questions to be both fun and engaging, and this is borne out by the fact that multiple users in both the pilot and the full study elected to continue creating additional documents after completing the required portions of the survey.

These results are highly encouraging for question-asking as a mode of user interaction from future AI systems. Amid global fears of AI replacing human labor, it is my hope that systems like CQDG can serve as an example of how collaboration between humans and AI can help human users to engage more fully and deeply with our problems, rather than only using AI to offload work. We can create AI interfaces in which humans are active participants in a thoughtful dialog rather than the more typical command-response interface which we see in ChatGPT, Gemini, Copilot, and other common LLM applications today, which can be used in an active and engaged way by users if they engage with the system thoughtfully, but which run the risk of encouraging users to rely more fully than they should on the AI and to think less than they should about the details of the problem they are trying to solve.

Qualities of an Ideal Question Asking Document Generator.

CQDG was designed to gather usable data for this research, it was not designed to be the ideal document generating app. However, from the results of these studies it is possible to describe features that we would want to see in an ideal document generation app. Some of these features were not included because they were not known until the study was completed. Others were not included because while they might have improved the user experience, they would have interfered with the research goals of the study.

1. **Use the highest quality LLM available.** If I were to convert CQDG into an app intended to optimize the user experience and document quality, it would use GPT-4 only, rather than randomly switching LLMs. Switching LLMs allowed the study to gather valuable data on the relationship between LLM and document quality, but from a user perspective it is clearly preferable to have the best possible quality on every document.
2. **Ask thought-provoking questions designed to engage the user.** This has the dual benefit of being valuable to users in its own right, as well as eliciting users to enter large amounts of new information and context which the LLM can use to make the best possible document.
3. **Questions may not always be necessary.** The CLAM model [23] includes a step of deciding for any given prompt whether a clarifying question is necessary or not. For some simple requests, questions may be more inconvenient than beneficial. In these cases, a decision step similar to what is done in CLAM could be used to decide whether to ask a question or not. The design of the CQDG studies required users to select between a document created using questions and answers and a baseline that was created without using the questions. This clearly would not work unless questions were asked, so in the CQDG model, exactly three questions were generated and asked every time.
4. **A more dynamic conversational flow to questions.** In a chatbot-like interface, an AI could hypothetically generate questions and ask them naturally as a part of a conversation with the user before creating the requested document. This would likely be a more natural way for the user to engage with the AI and could result in the AI asking more or fewer questions as needed, asking further follow-ups based on the user's initial responses, allowing the request to morph with the conversation as the user's needs become clearer. At least some of this is possible with currently known LLM techniques, and I suspect that a system operating very close to what I have described is possible with current LLMs, if further

research and development is devoted towards achieving a system with these traits. I initially designed CQDG to operate closer to this model, but I quickly ran into problems with the study design. In such a free-flowing conversation, at what point should the system snapshot the document for the user to compare to a baseline? I considered allowing the users to indicate for themselves when they were satisfied that the conversation was complete, but this ran the risk of conversations flowing off topic, or a satisfactory result being overlooked by CQDG in favor of more recent unsatisfactory or irrelevant responses. This is not as much of a problem for a document generating app, since the user can simply take what they need from the full conversation as they find it useful to do so. For the sake of the simplicity and consistency of the study, however, I opted for a system that always asked exactly three questions and produced exactly one output after those questions had been answered, so it was always clear which LLM response constituted the QA document to be compared to the baseline.

5. **Give the user multiple outputs to choose from.** Users indicated in the exit survey that they appreciated being able to compare the baseline and QA documents and select between them. This is commonly done in visual generative Ais, which routinely produce multiple possible outputs for the user to select between or blend together in response to a single request for artwork. In an ideal system, the LLM would have both documents in a single context as well. In their refining prompts, users often requested that one document be made "more like the other one" or to include a section from "the other document." However, because of the way CQDG's conversations were formatted internally, there was only ever one document in context at any given time, and CQDG was entirely unable to respond to this category of requests.
6. **Provide a way for the user to specify the tone or format of the document.** CQDG's internal prompts specifically instruct it to produce a short, high-quality, professional document. However, users entered a wide variety of requests, not all of which were for high-quality professional documents. In addition to professional communications, users requested poetry, fiction, and letters written in a more casual and emotional tone. The system-level instructions given to an ideal document generator should be dynamic enough to account for these different sorts of needs. This could be as simple as providing a dropdown before the conversation begins, in which the user selects between several preset

"personalities" or tones for the AI. It could also be something that the AI is set up to determine for itself, based on its dialog with the user.

Cost and Efficiency of Question Asking

One of the motivations for including multiple LLMs in this study was to determine whether question-asking could be used to offset the need for more expensive LLMs. I had initially hoped that a lower quality and less expensive LLM, such as GPT-3.5, could be augmented with question-asking capabilities to produce higher quality output competitive with more expensive LLMs. However, the exact opposite appears to be true. Question asking is most valuable on the highest quality, most expensive LLMs, which are best able to make use the increased complexity that naturally comes with question asking.

Furthermore, asking questions imposes additional cost when compared to simply giving the LLM an undecorated user prompt. The question-asking conversation involves multiple steps of prompt and response with the LLM to first generate the questions and then respond to the full context of user request, questions, and answers. As such, it may not be practical to include question asking as a standard feature in most LLM interfaces until interfacing with LLMs becomes more affordable. However, the benefits of question asking are clear, and this style of interface should be considered in cases where clarity and user engagement or highly important. Document generation represents one such case, but question asking could also be useful for a chatbot which is helping a user to work through a difficult problem, brainstorm solutions, or learn a new concept. Asking questions is likely not worth the added expense for very simple requests or commands in cases where these commands can be interpreted with high fidelity without asking questions. In cases where simple disambiguating questions might be beneficial but thought-provoking questions are not necessary, CLAM [23] offers a more efficient model which requires less context and smaller requests and responses.

Recent Developments: GPT-o1

It is worth noting at this point that this field is developing rapidly. My initial experiments for this research were conducted using gpt-3.5, which is now obsolete by several versions. I finished the initial writing for this dissertation on 9/11/2024. As I was editing in preparation to submit this document on 9/12, OpenAI released a new version of ChatGPT, GPT-o1-preview, which advertises that it spends a longer amount of time carefully reasoning through difficult analytical tasks [143], and boasts superior performance on several benchmarks when compared with previous versions of ChatGPT [144].

GPT-o1-preview responds more appropriately to ambiguous questions. When asked “What is a transformer?” GPT-o1-preview responds with a detailed answer that includes transformers in AI, electrical engineering, and the Transformers entertainment franchise without needing to be specifically prompted to do so (see **Appendix J**). This is better than GPT-4o or GPT-o1-mini, which still respond by describing only the transformer as described in *Attention is all you Need* [55], just as GPT-4 and GPT-3.5 did (see **Appendix A**).

GPT-o1-preview has fully embraced chain-of-thought reasoning to achieve these impressive results. However, GPT-o1-preview still does not ask questions. Instead of giving a long answer covering all three possibilities, GPT could instead have asked me which context I was referring to in a much more conversational way. For a more pointed example, I also asked GPT-o1-preview to produce a document for me, with the prompt: *“Write a cover letter for interest in a job posting as an assistant professor of Computer Science.”* I judged that this task is impossible to do well without additional information. The prompt doesn’t include my area of specialty or anything about my background or education, nor any specifics about the posting itself. Nonetheless, GPT-o1-preview immediately produced a result without engaging in any dialog or follow-up questions (see **Appendix K**). Not only that, but it did not provide any introductory or closing text to indicate that it might do better if I provided more detail, as it did when asked about transformers.

It is encouraging to see ChatGPT embracing chain of thought reasoning, and through superior reasoning producing better responses to ambiguous prompts. However, GPT-o1-preview seems to be committed to the command and response paradigm, rather than engaging the user in productive human-AI collaborative dialog. This shows that there is still a clear need for this research, which demonstrates the potential benefits of LLMs asking questions.

The desirability of asking questions is context dependent. There will be applications where the LLM is not in direct communication with the user and is merely responding to commands sent by another software system, in which case trying to engage in a dialog with thought provoking questions could well be unproductive. However, since GPT-o1-preview was released as a chatbot for GPT Plus accounts, it seems reasonable to assume that at least in its current iteration, the people talking to it are human beings who would be willing and able to answer questions and engage in productive dialog, showing that this is an area where industry-standard LLMs are still lacking.

Chapter 6: Limitations and Future Work

In this section I will discuss the limitations of these results, lessons learned about how these studies could have been better designed, and possibilities for future work suggested by the results so far.

Limitations

These studies cannot draw statistical conclusions about specific demographic groups. The purpose of this research was to determine:

1. Whether question-asking had a beneficial effect on document quality.
2. Whether the benefits of question-asking were impacted by the selection of LLM.
3. Whether users believed that there were qualitative benefits to question-asking aside from final document quality.

Participants were also asked demographic questions about their age, gender, and prior level of experience with generative AI. Results were presented which show the effect of each of these demographic factors. However, these results should be considered purely descriptive and have not been subjected to statistical scrutiny. Participants were not divided evenly among these groups and some subgroups, such as the “other/nonbinary” gender category and the “I have never used generative AI before” category represent only a single-digit number of participants each. The sample size also reduces rapidly when taking the intersection of any given set of answers, such as “women who are regular users of AI” or “males between the ages of 30 and 40.” The study was not scoped to analyze these questions, only overall document quality and the effects of LLM selection. To analyze demographic factors with appropriate levels of statistical significance would require a sample size multiple times larger than what was obtained for this study.

Furthermore, this research did not take any precautions to ensure that the demographic spread of the participants mirrored that of the overall population. Women outnumbered men by a large margin in this research, and the participants were also younger on average than the US population. [13] Participant’s level of experience with AI is also likely to be higher overall than that of

---

[13] Average age of 31 for the study, vs 38.5 for US population. [145]

the general population, since some participants were invited to the study from computer science classrooms. Thus, even if the rates of preference for each of these groups could be shown to be statistically significant, the result would not be useful evidence of trends in the general population.

Even though they are not statistically significant and are not proof of trends in the general population, the descriptive results for the demographics categories are still of potential interest for future studies. The trends seen in these results, such as women, older users, and less experienced users responding more positively to the question asking process, provide promising directions for future research, in studies that are better targeted towards investigating these effects specifically.

Users saw questions when judging the baseline, both before and after refining. Another limitation in this study is that all participants saw and answered three questions, which they ordinarily would not have done for a true “baseline” case. This was required for this study design. However, a side effect of this setup is that during the refining phase, users had the opportunity to re-enter or re-emphasize answers they gave during question answering. Since participants broadly indicated that they found the questions to be useful and thought provoking, it is likely that in some cases users were able to enter better refining prompts for the baseline document during the refining phase than they would have otherwise, had they not seen the thought-provoking questions at all. This may contribute to refined baseline documents that were of higher quality than they might have been if the users had never seen the questions at all.

This side effect was identified during the study design but was unavoidable for a preference study. One alternative would have been to have each user see only one document, with half of all users receiving the baseline process and the other half receiving the QA process. The problem with this is that users find it difficult to rate documents objectively and consistently but find it much easier to select a preference between two documents. This is backed up by the results of the pilot study. In order to allow study participants to indicate their preference between the baseline and QA document, they must see both documents, and therefore must see the questions at some point before comparing the documents.

Another alternative would have been to not include the refining phase at all. However, the ability to refine documents was one of the most-requested features from the pilot study. Another option would have been to always show the baseline document first and allow the user to refine the baseline document before being shown the questions. This alternative would have several flaws:

1. The order of the documents would not be randomized, and users may express a bias towards either the first or last document they saw.
2. Refining prompts from the baseline document would either have to be included in the QA document context, thus advantaging the QA document by giving it more context to work with than the baseline. Alternatively, the QA document could exclude the refining prompts already given, but this would disrupt the process by removing information the user had already entered, likely resulting in complaints of the second document "forgetting" feedback that was already given and thus biasing the results.
3. Using this process, preference data could only be gathered after all refining was completed on both documents, at which point they could be shown side by side. It would not be possible to ask the user to choose between the initially generated documents, the "before refining" preference, until after they had already seen at least the refined baseline document.

Thus, the design that was chosen was the best one available. However, the "after refining" preference results are still somewhat compromised by this effect, and there would be benefits to alternative designs that can avoid the problem of users seeing the questions before refining the baseline document. However, I cannot see any way to design such a study at this time. Despite this flaw, I feel that it is still better to have the "after refining" results than not to have them at all.

Lessons Learned

This section covers lessons learned from the design and execution of this study, namely design flaws or study elements which could have been designed better with the benefit of hindsight.

Not all study questions were relevant. Several of the questions shown to participants during this study were not relevant to the goals of this research. Most notably, there were two questions about document preference: "*Which document do you prefer overall?*" and *"Which document would be more useful to you in its current state?"* The second question is not actually relevant to the stated goals for this study. Only the question on overall preference should have been included. The meaning of these questions is similar, but from user feedback it is clear that they are not identical. In some cases, participants indicated that one document was preferred overall even though the other was more useful in its current state. This could be for many reasons, such as a document serving as a good starting point but not being usable as-is. However, the rest of the study was not set up to support identifying reasons for these differences and the question about usability winds up being just a distraction to the final analysis.

Additionally, participants were asked to rate the statement *"I liked that the AI showed me two options to pick between, instead of only picking the option it thought was best."* However, nothing in this study is set up to determine whether there is an advantage to picking between multiple similar documents, or to compare alternative methods of producing documents. While the result is interesting for the design of future systems, this statement was only tangentially related to the stated research goals.

Responses of "No Preference" should not have been allowed. Responses of "no preference" are known to cause issues in the analysis of preference studies [141]. No preference responses either need to be discarded, distributed evenly between the two options, or distributed proportionally to the preference ratios for either option. Each of these techniques cause potential issues and can reduce the overall sensitivity of the study in the case of discarding or equal distribution, or falsely bias and inflate the significance of the result in the case of proportional distribution. While equal distribution of responses was the most appropriate technique for these results, this reduced the sensitivity of the analysis somewhat. It would have been better if the study was designed from the beginning to not allow "no preference" responses to begin with. This was not a flaw that was considered during study design and can only be attributed to my own inexperience in designing preference studies.

The Meta ad campaign was less effective than hoped. As was already discussed in the "Budget" section, the Meta ad campaign was less effective than originally hoped. The ad campaign ran from April through the end of June, but only one participant took the study in June. The ad campaign was discontinued at the end of June due to not driving traffic to the study. A more reliable form of participant selection could have allowed a larger group to be reached more quickly.

Future Work

There are several promising possibilities for future studies that can expand on the work that has been done so far. Future research should be done using systems which are as close as possible to the "ideal system" described in the previous chapter. Study design may impose some limitations on how many of the identified "ideal system" features can be implemented, just as they did for this study, but where possible the practical considerations learned from this research should be incorporated into any future research design. For example, LLM selection is an obvious area for improvement since this research has shown that less powerful LLMs are less capable of taking advantage of the question and answer dialog. Ideas for future research using the proposed "Ideal System" are provided below.

Future work can be targeted to investigate demographic trends suggested by these studies. As stated in the "limitations" section above, the demographic trends shown in these results are only descriptive and are not considered to be statistically significant. However, they suggest that question asking interfaces may work better for certain groups including non-males, users over 40, and users with less experience with generative AI. Future studies could be designed to draw from a more representative sample of the population and be sized appropriately to investigate these demographic trends with greater rigor.

Expanding CQDG to generate different types of documents, including visual art and computer code. This research focused on the creation of short, professional documents. However, it is likely that question asking could have similar benefits in other outputs of generative AI, such as AI generated written fiction, visual art, and computer code. Future studies could investigate whether the CQDG question asking process has similar benefits in these areas.

Chapter 8: Conclusion

This research asked whether generative AI asking follow-up questions could improve the quality of generated text documents, and whether these follow up questions could improve the subjective user experience. An original software system, CQDG, was created to study these questions. A small pilot study was used to refine the research methodology, followed by a larger study utilizing a more complex version of CQDG. This research produced several important results:

1. Generative AI can be made to ask useful follow-up questions in response to user requests.
2. Follow-up questions improve generative AI's ability to produce documents which conform to users' needs.
3. Users appreciate the follow up questions and have a positive experience with answering them, especially when those questions are creative and thought-provoking.

These results reinforce an idea which was shown in several prior works, that generative AI produces the best results when users engage with AI systems in an active and critical way [11, 122]. This is also in agreement with prior results that showed that question asking could be beneficial in simple ambiguity-resolution scenarios [23–25] and that understanding conversational context is important for improving the quality of answers and suggestions given by AI [26].

The unique contributions of this research are showing that these same principles apply to the more open-ended task of document generation, and that this process not only improves the quality of generated documents, but the user experience as well. Users impressed me overall with their preference for thought-provoking questions, their willingness to give detailed answers to AI generated questions, and in some cases their eagerness to continue creating more documents after completing the initial survey. These results defy common stereotypes of people as inherently lazy or disengaged, and show that well-designed software systems can draw people in, increasing their critical engagement and thought in a way that they find positive and engaging while simultaneously producing superior outputs.

The results of this research show that users value AI systems that engage critically with their problems and are open to engaging in collaborative dialog with AI to improve both their experience with the system and the outputs the system produces. Users found questions to be valuable even in the minority of cases where those questions did not lead to better documents. Incorporating follow-up questions, and possibly other forms of engaging dialog, into generative AI systems has the potential to mitigate the risks of over-automation, reduce hallucinations, improve productivity, and keep people engaged and in-the-loop for increasingly AI-assisted tasks. This is a line of research well worth expansion and continued work. The results we already have are more than sufficient to suggest incorporating a question-asking phase into generative AI processes where context, specificity, and user engagement are highly valued.

Appendix A: ChatGPT on Transformers

Conversation with ChatGPT, 02/18/2023 (GPT 3.5)

Tix: What is a Transformer?

ChatGPT: A Transformer is a type of neural network architecture that was introduced in a seminal paper by Vaswani et al. in 2017. Transformers are widely used in natural language processing (NLP) tasks, such as language translation, text summarization, and question-answering.

Unlike traditional neural networks, which process input data sequentially, Transformers are designed to process entire sequences of input data simultaneously. They use a self-attention mechanism that allows the network to weigh the importance of different parts of the input sequence when making predictions.

The Transformer architecture consists of an encoder and a decoder. The encoder takes an input sequence and generates a sequence of hidden states, which are then passed to the decoder. The decoder generates an output sequence based on the hidden states and an input sequence of tokens that it receives as input.

Transformers have been shown to achieve state-of-the-art performance in a wide range of NLP tasks and have become a fundamental building block for many advanced NLP models.

Tix: Is this the only meaning of "transformer?"

ChatGPT: No, the term "transformer" has multiple meanings depending on the context. In addition to the neural network architecture, some other meanings of "transformer" include:

1. Electrical transformer: This is a device that is used to transfer electrical energy from one circuit to another through electromagnetic induction. Transformers are commonly used in power transmission and distribution systems to change the voltage level of the electricity.
2. Mechanical transformer: This refers to a device that is used to transfer energy from one physical system to another, such as a gearbox or a pulley system.
3. Action figure: In popular culture, a "transformer" can refer to a toy or action figure that can transform from one form to another, such as the popular Transformers franchise.

So, the meaning of "transformer" can vary depending on the context in which it is used.

Appendix B: WordNet 3.1 Senses of “Leaves”

Noun

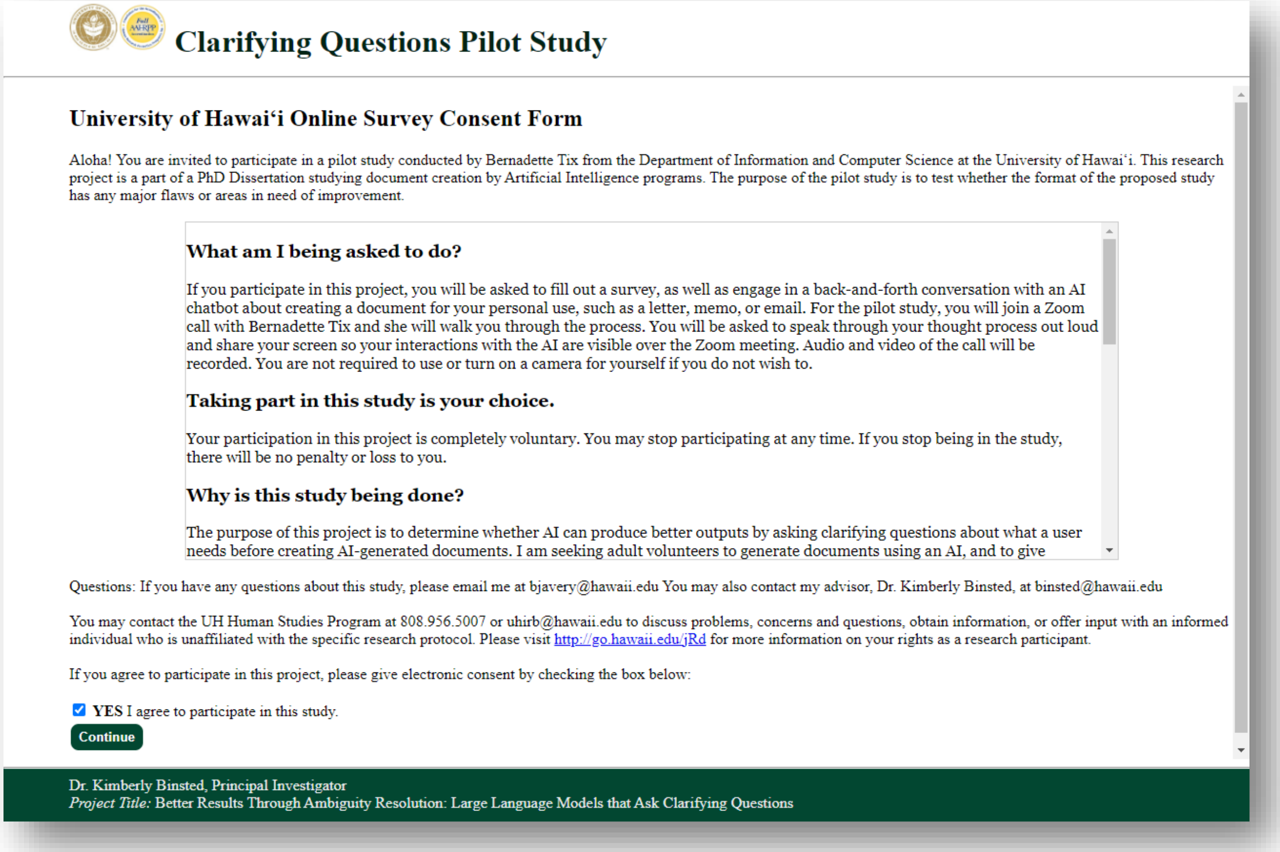
1. S: (n) leaf, leafage, foliage (the main organ of photosynthesis and transpiration in higher plants)
2. S: (n) leaf, folio (a sheet of any written or printed material (especially in a manuscript or book))
3. S: (n) leaf (hinged or detachable flat section (as of a table or door))
4. S: (n) leave, leave of absence (the period of time during which you are absent from work or duty) "a ten day's leave to visit his mother"
5. S: (n) leave (permission to do something) "she was granted leave to speak"
6. S: (n) farewell, leave, leave-taking, parting (the act of departing politely) "he disliked long farewells"; "he took his leave"; "parting is such sweet sorrow"

Verb

1. S: (v) leave, go forth, go away (go away from a place) "At what time does your train leave?"; "She didn't leave until midnight"; "The ship leaves at midnight"
2. S: (v) leave (go and leave behind, either intentionally or by neglect or forgetfulness) "She left a mess when she moved out"; "His good luck finally left him"; "her husband left her after 20 years of marriage"; "she wept thinking she had been left behind"
3. S: (v) leave (act or be so as to become in a specified state) "The inflation left them penniless"; "The president's remarks left us speechless"
4. S: (v) leave, leave alone, leave behind, let alone (leave unchanged or undisturbed or refrain from taking) "leave it as is"; "leave the young fawn alone"; "leave the flowers that you see in the park behind"
5. S: (v) exit, go out, get out, leave (move out of or depart from) "leave the room"; "the fugitive has left the country"
6. S: (v) leave, allow for, allow, provide (make a possibility or provide opportunity for; permit to be attainable or cause to remain) "This leaves no room for improvement"; "The evidence allows only one conclusion"; "allow for mistakes"; "leave lots of time for the trip"; "This procedure provides for lots of leeway"

7. S: (v) leave, result, lead (produce as a result or residue) "The water left a mark on the silk dress"; "Her blood left a stain on the napkin"
8. S: (v) leave, depart, pull up stakes (remove oneself from an association with or participation in) "She wants to leave"; "The teenager left home"; "She left her position with the Red Cross"; "He left the Senate after two terms"; "after 20 years with the same company, she pulled up stakes"
9. S: (v) entrust, leave (put into the care or protection of someone) "He left the decision to his deputy"; "leave your child in the nurse's care"
10. S: (v) bequeath, will, leave (leave or give by will after one's death) "My aunt bequeathed me all her jewelry"; "My grandfather left me his entire estate"
11. S: (v) leave (have left or have as a remainder) "That left the four of us"; "19 minus 8 leaves 11"
12. S: (v) leave, leave behind (be survived by after one's death) "He left six children"; "At her death, she left behind her husband and 11 cats"
13. S: (v) impart, leave, give, pass on (transmit (knowledge or skills)) "give a secret to the Russians"; "leave your name and address here"; "impart a new skill to the students"
14. S: (v) forget, leave (leave behind unintentionally) "I forgot my umbrella in the restaurant"; "I left my keys inside the car and locked the doors"

Appendix C: Pilot Study Screenshots All Steps



Clarifying Questions Pilot Study

University of Hawai‘i Online Survey Consent Form

Aloha! You are invited to participate in a pilot study conducted by Bernadette Tix from the Department of Information and Computer Science at the University of Hawai‘i. This research project is a part of a PhD Dissertation studying document creation by Artificial Intelligence programs. The purpose of the pilot study is to test whether the format of the proposed study has any major flaws or areas in need of improvement.

What am I being asked to do?

If you participate in this project, you will be asked to fill out a survey, as well as engage in a back-and-forth conversation with an AI chatbot about creating a document for your personal use, such as a letter, memo, or email. For the pilot study, you will join a Zoom call with Bernadette Tix and she will walk you through the process. You will be asked to speak through your thought process out loud and share your screen so your interactions with the AI are visible over the Zoom meeting. Audio and video of the call will be recorded. You are not required to use or turn on a camera for yourself if you do not wish to.

Taking part in this study is your choice.

Your participation in this project is completely voluntary. You may stop participating at any time. If you stop being in the study, there will be no penalty or loss to you.

Why is this study being done?

The purpose of this project is to determine whether AI can produce better outputs by asking clarifying questions about what a user needs before creating AI-generated documents. I am seeking adult volunteers to generate documents using an AI, and to give

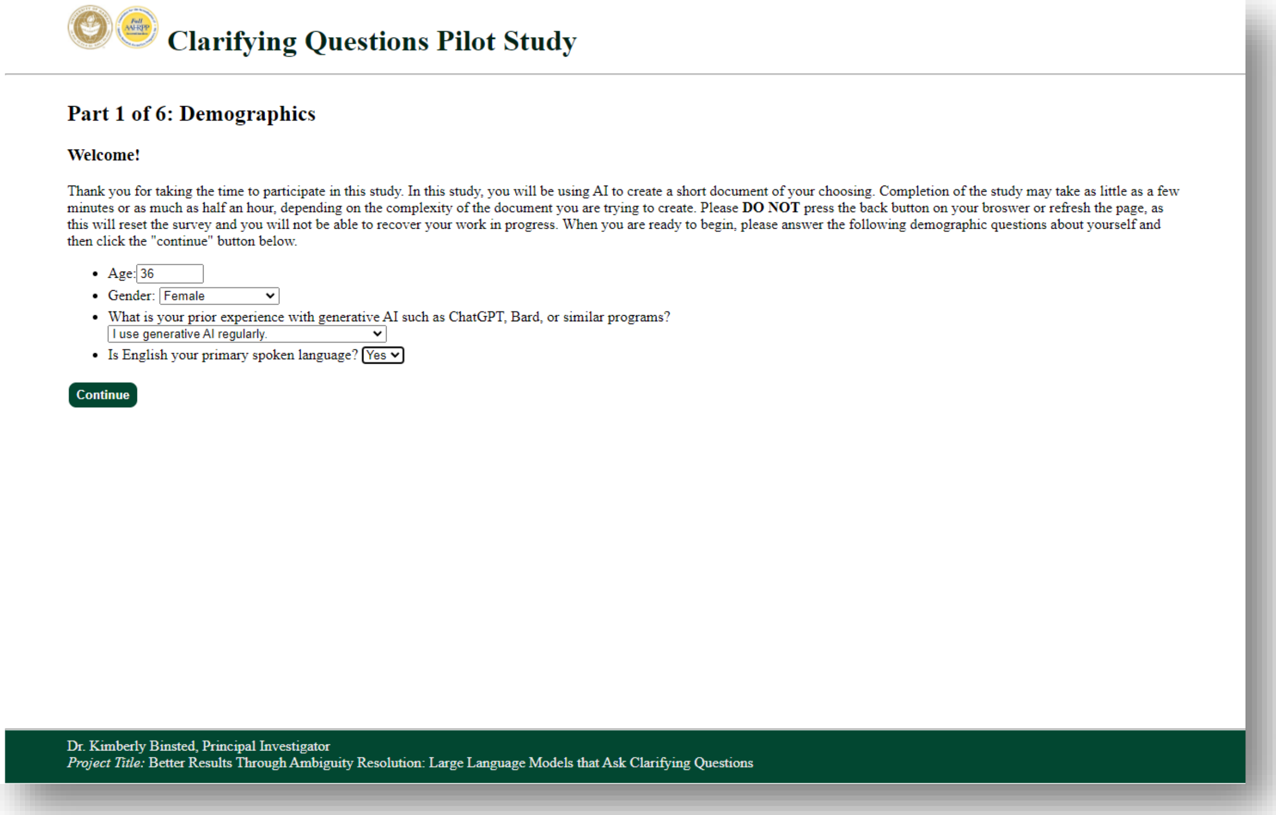
Questions: If you have any questions about this study, please email me at bjavery@hawaii.edu You may also contact my advisor, Dr. Kimberly Binsted, at binsted@hawaii.edu

You may contact the UH Human Studies Program at 808.956.5007 or uhirb@hawaii.edu to discuss problems, concerns and questions, obtain information, or offer input with an informed individual who is unaffiliated with the specific research protocol. Please visit http://go.hawaii.edu/jRd for more information on your rights as a research participant.

If you agree to participate in this project, please give electronic consent by checking the box below:

☑ **YES** I agree to participate in this study.

Dr. Kimberly Binsted, Principal Investigator
Project Title: Better Results Through Ambiguity Resolution: Large Language Models that Ask Clarifying Questions



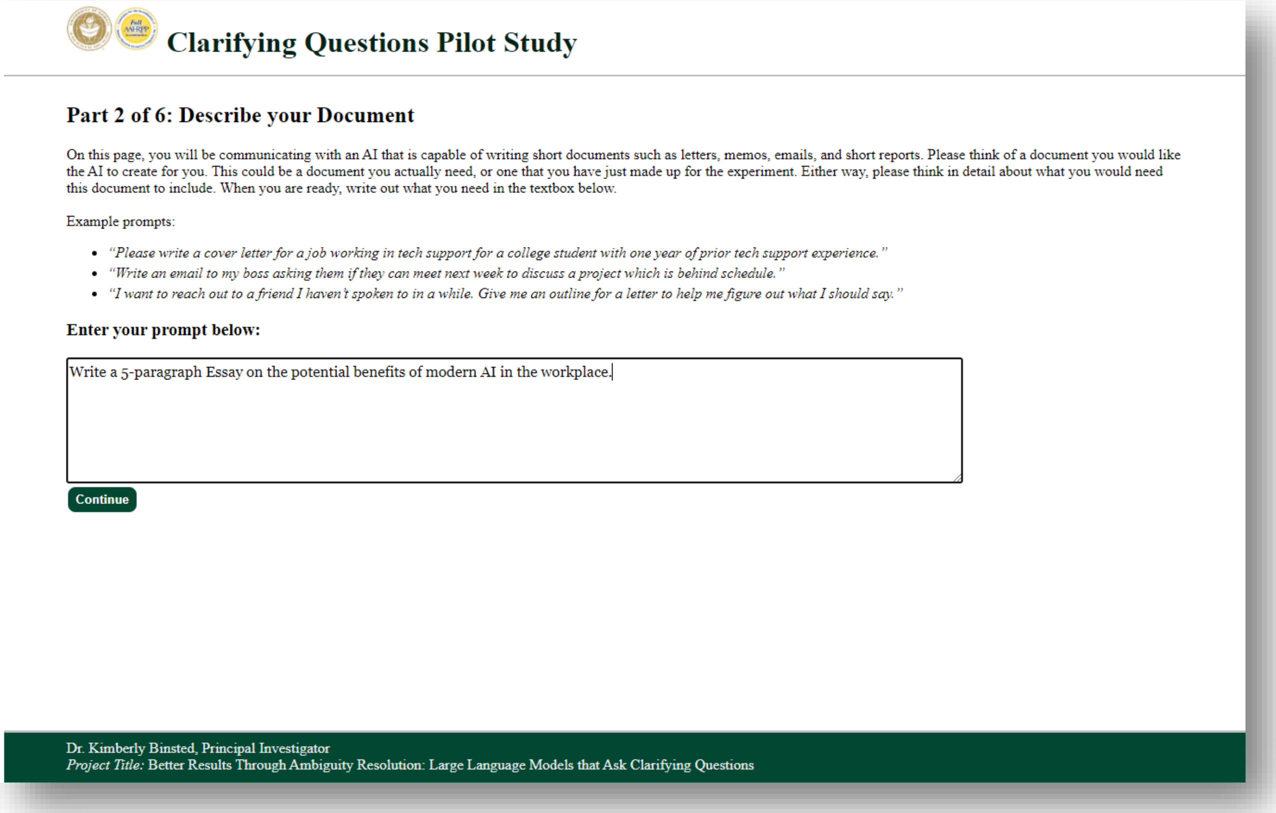
Part 1 of 6: Demographics

Welcome!

Thank you for taking the time to participate in this study. In this study, you will be using AI to create a short document of your choosing. Completion of the study may take as little as a few minutes or as much as half an hour, depending on the complexity of the document you are trying to create. Please **DO NOT** press the back button on your broswer or refresh the page, as this will reset the survey and you will not be able to recover your work in progress. When you are ready to begin, please answer the following demographic questions about yourself and then click the "continue" button below.

- Age: 36
- Gender: Female
- What is your prior experience with generative AI such as ChatGPT, Bard, or similar programs? I use generative AI regularly.
- Is English your primary spoken language? Yes

Dr. Kimberly Binsted, Principal Investigator
Project Title: Better Results Through Ambiguity Resolution: Large Language Models that Ask Clarifying Questions



Clarifying Questions Pilot Study

Part 2 of 6: Describe your Document

On this page, you will be communicating with an AI that is capable of writing short documents such as letters, memos, emails, and short reports. Please think of a document you would like the AI to create for you. This could be a document you actually need, or one that you have just made up for the experiment. Either way, please think in detail about what you would need this document to include. When you are ready, write out what you need in the textbox below.

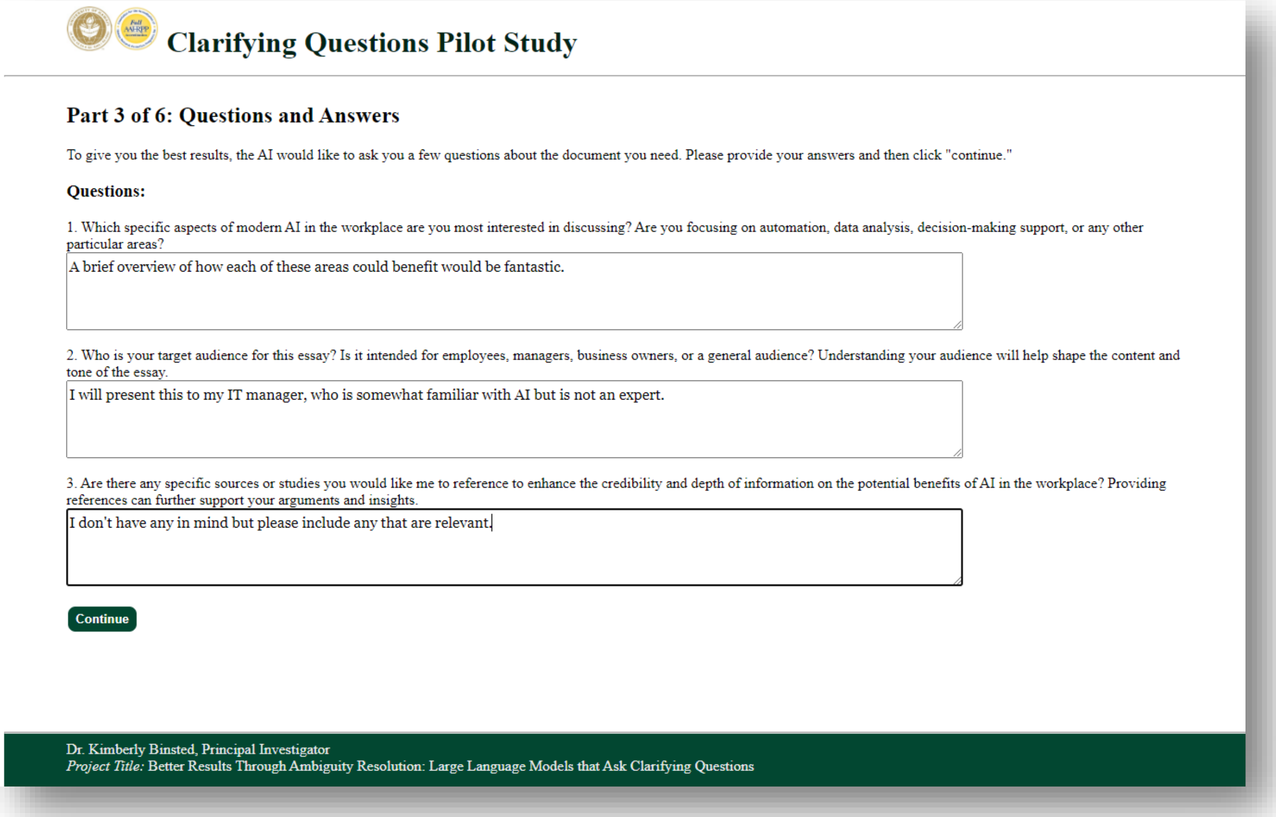
Example prompts:

- *"Please write a cover letter for a job working in tech support for a college student with one year of prior tech support experience."*
- *"Write an email to my boss asking them if they can meet next week to discuss a project which is behind schedule."*
- *"I want to reach out to a friend I haven't spoken to in a while. Give me an outline for a letter to help me figure out what I should say."*

Enter your prompt below:

Write a 5-paragraph Essay on the potential benefits of modern AI in the workplace.

Dr. Kimberly Binsted, Principal Investigator
Project Title: Better Results Through Ambiguity Resolution: Large Language Models that Ask Clarifying Questions



Clarifying Questions Pilot Study

Part 3 of 6: Questions and Answers

To give you the best results, the AI would like to ask you a few questions about the document you need. Please provide your answers and then click "continue."

Questions:

1. Which specific aspects of modern AI in the workplace are you most interested in discussing? Are you focusing on automation, data analysis, decision-making support, or any other particular areas?

A brief overview of how each of these areas could benefit would be fantastic.

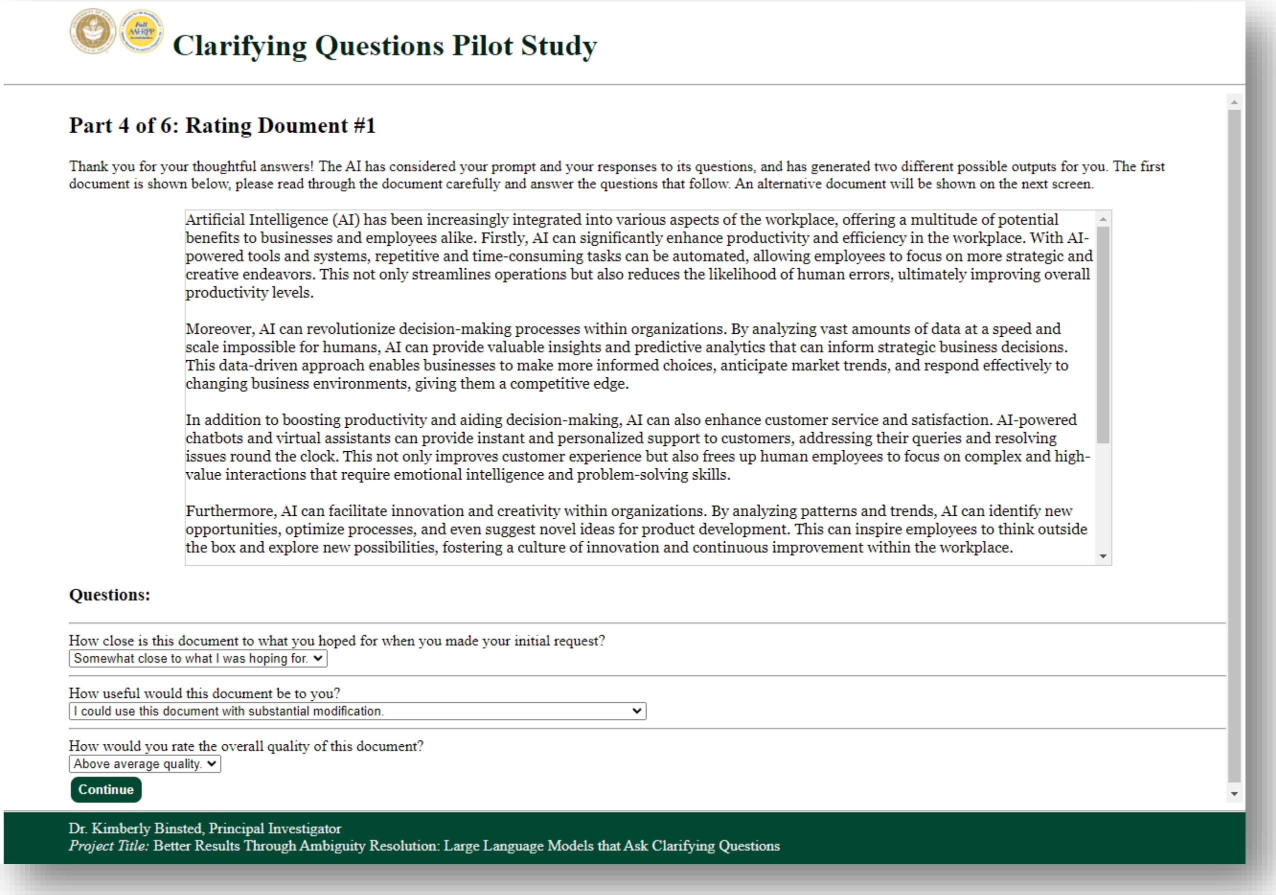
2. Who is your target audience for this essay? Is it intended for employees, managers, business owners, or a general audience? Understanding your audience will help shape the content and tone of the essay.

I will present this to my IT manager, who is somewhat familiar with AI but is not an expert.

3. Are there any specific sources or studies you would like me to reference to enhance the credibility and depth of information on the potential benefits of AI in the workplace? Providing references can further support your arguments and insights.

I don't have any in mind but please include any that are relevant.

Dr. Kimberly Binsted, Principal Investigator
Project Title: Better Results Through Ambiguity Resolution: Large Language Models that Ask Clarifying Questions



Clarifying Questions Pilot Study

Part 4 of 6: Rating Doument #1

Thank you for your thoughtful answers! The AI has considered your prompt and your responses to its questions, and has generated two different possible outputs for you. The first document is shown below, please read through the document carefully and answer the questions that follow. An alternative document will be shown on the next screen.

Artificial Intelligence (AI) has been increasingly integrated into various aspects of the workplace, offering a multitude of potential benefits to businesses and employees alike. Firstly, AI can significantly enhance productivity and efficiency in the workplace. With AI-powered tools and systems, repetitive and time-consuming tasks can be automated, allowing employees to focus on more strategic and creative endeavors. This not only streamlines operations but also reduces the likelihood of human errors, ultimately improving overall productivity levels.

Moreover, AI can revolutionize decision-making processes within organizations. By analyzing vast amounts of data at a speed and scale impossible for humans, AI can provide valuable insights and predictive analytics that can inform strategic business decisions. This data-driven approach enables businesses to make more informed choices, anticipate market trends, and respond effectively to changing business environments, giving them a competitive edge.

In addition to boosting productivity and aiding decision-making, AI can also enhance customer service and satisfaction. AI-powered chatbots and virtual assistants can provide instant and personalized support to customers, addressing their queries and resolving issues round the clock. This not only improves customer experience but also frees up human employees to focus on complex and high-value interactions that require emotional intelligence and problem-solving skills.

Furthermore, AI can facilitate innovation and creativity within organizations. By analyzing patterns and trends, AI can identify new opportunities, optimize processes, and even suggest novel ideas for product development. This can inspire employees to think outside the box and explore new possibilities, fostering a culture of innovation and continuous improvement within the workplace.

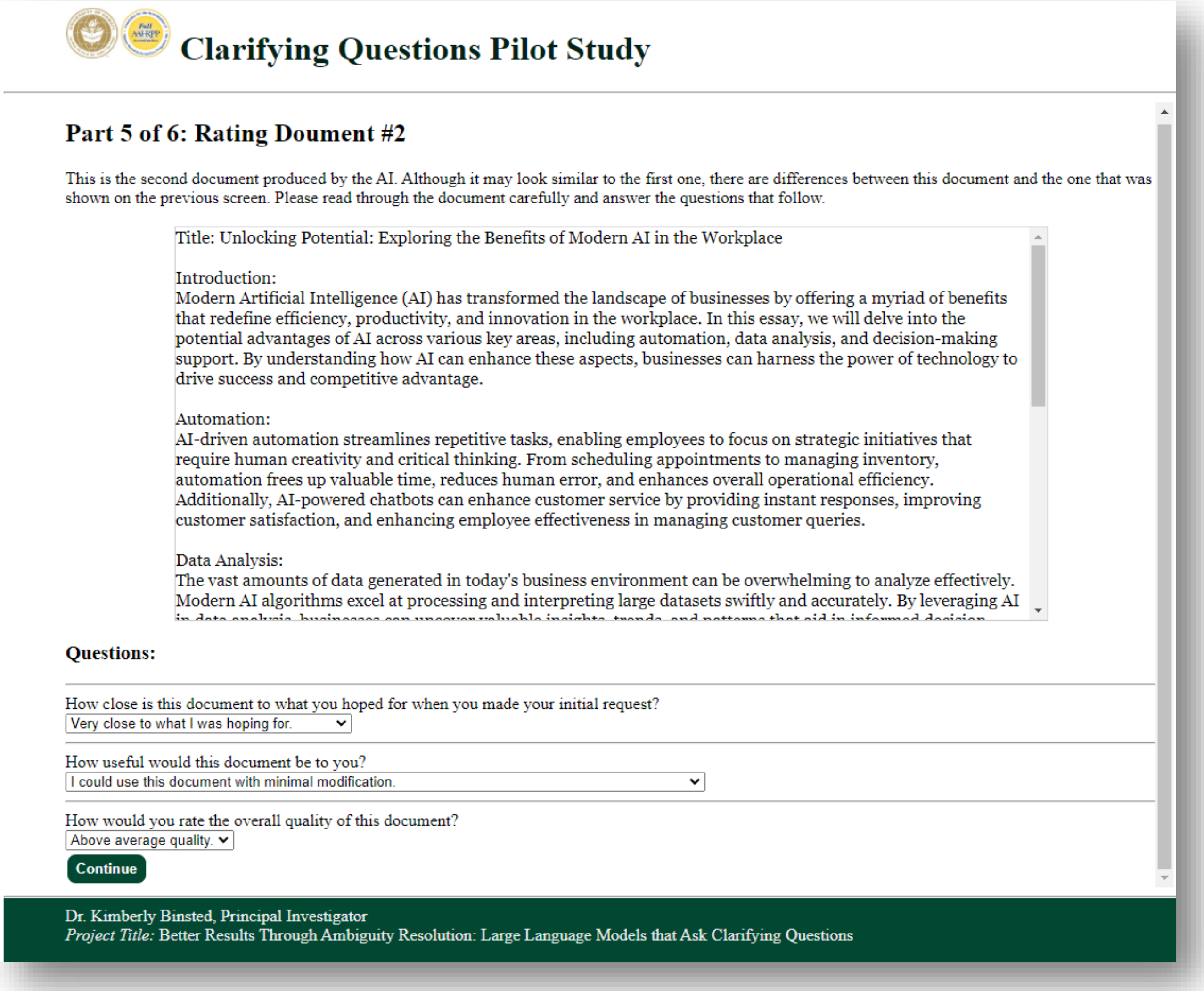
Questions:

How close is this document to what you hoped for when you made your initial request?
Somewhat close to what I was hoping for.

How useful would this document be to you?
I could use this document with substantial modification.

How would you rate the overall quality of this document?
Above average quality.

Dr. Kimberly Binsted, Principal Investigator
Project Title: Better Results Through Ambiguity Resolution: Large Language Models that Ask Clarifying Questions



Clarifying Questions Pilot Study

Part 5 of 6: Rating Doument #2

This is the second document produced by the AI. Although it may look similar to the first one, there are differences between this document and the one that was shown on the previous screen. Please read through the document carefully and answer the questions that follow.

Title: Unlocking Potential: Exploring the Benefits of Modern AI in the Workplace

Introduction:
Modern Artificial Intelligence (AI) has transformed the landscape of businesses by offering a myriad of benefits that redefine efficiency, productivity, and innovation in the workplace. In this essay, we will delve into the potential advantages of AI across various key areas, including automation, data analysis, and decision-making support. By understanding how AI can enhance these aspects, businesses can harness the power of technology to drive success and competitive advantage.

Automation:
AI-driven automation streamlines repetitive tasks, enabling employees to focus on strategic initiatives that require human creativity and critical thinking. From scheduling appointments to managing inventory, automation frees up valuable time, reduces human error, and enhances overall operational efficiency. Additionally, AI-powered chatbots can enhance customer service by providing instant responses, improving customer satisfaction, and enhancing employee effectiveness in managing customer queries.

Data Analysis:
The vast amounts of data generated in today's business environment can be overwhelming to analyze effectively. Modern AI algorithms excel at processing and interpreting large datasets swiftly and accurately. By leveraging AI

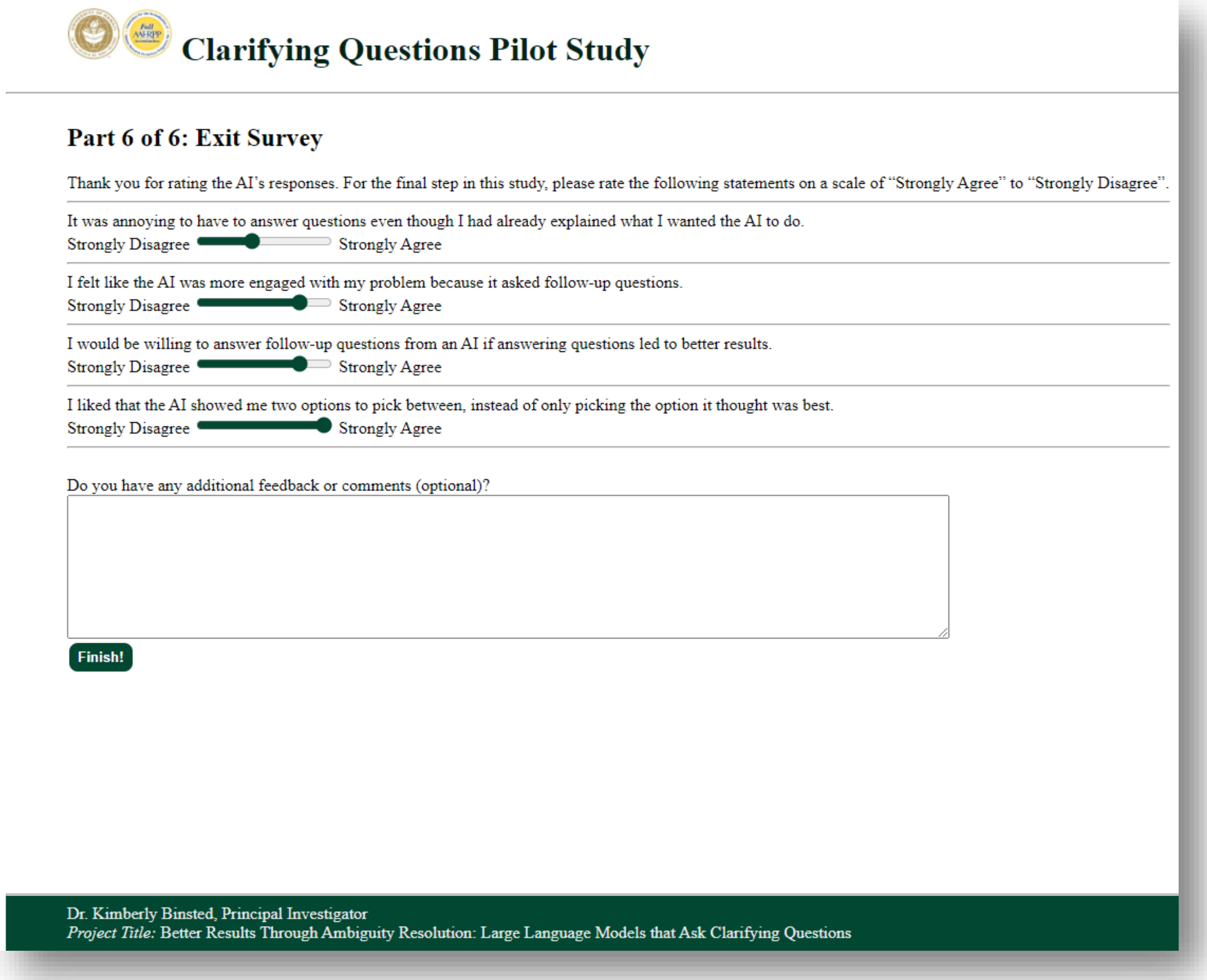
Questions:

How close is this document to what you hoped for when you made your initial request?
Very close to what I was hoping for.

How useful would this document be to you?
I could use this document with minimal modification.

How would you rate the overall quality of this document?
Above average quality.

Dr. Kimberly Binsted, Principal Investigator
Project Title: Better Results Through Ambiguity Resolution: Large Language Models that Ask Clarifying Questions



Clarifying Questions Pilot Study

Part 6 of 6: Exit Survey

Thank you for rating the AI's responses. For the final step in this study, please rate the following statements on a scale of "Strongly Agree" to "Strongly Disagree".

It was annoying to have to answer questions even though I had already explained what I wanted the AI to do.
Strongly Disagree Strongly Agree

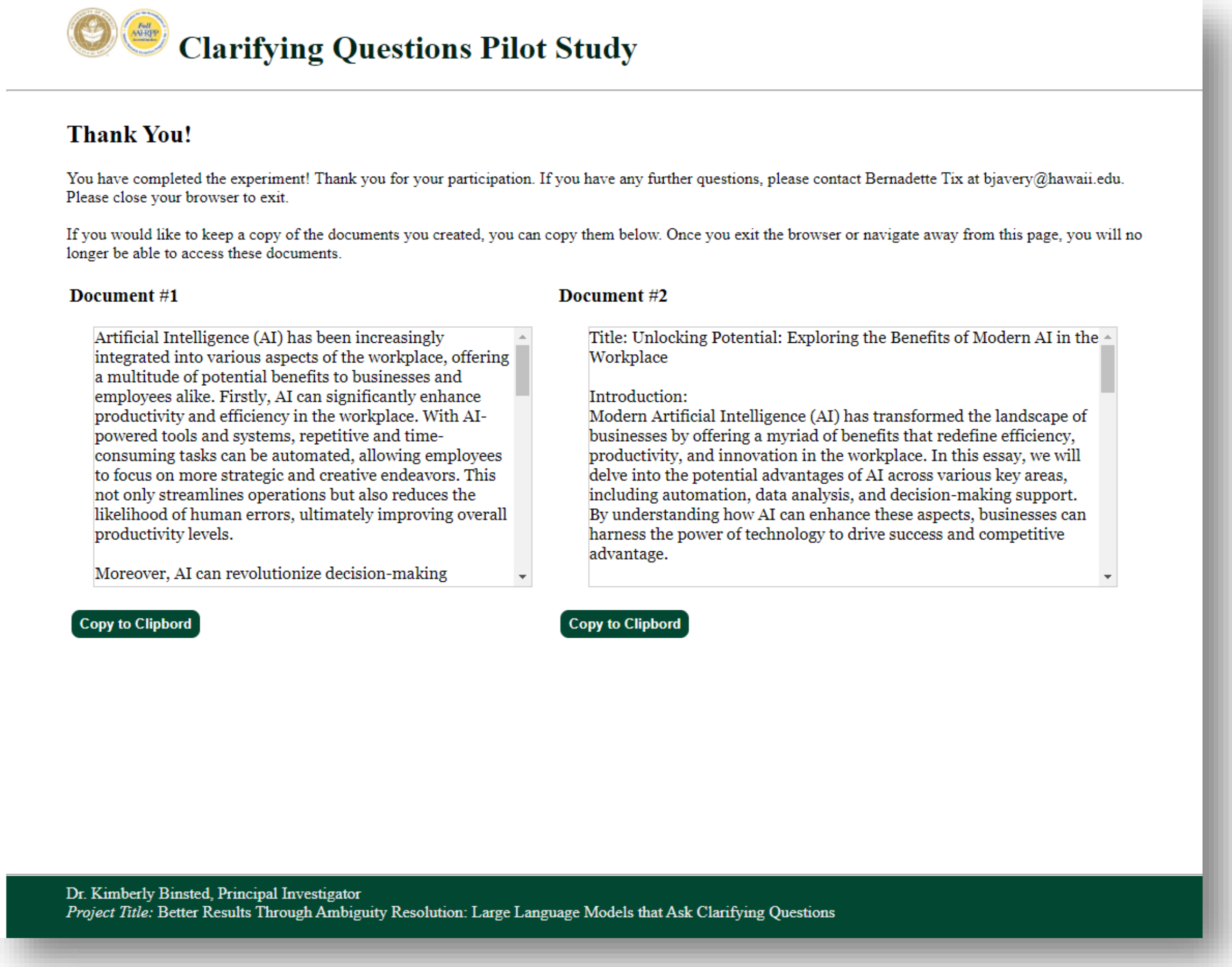
I felt like the AI was more engaged with my problem because it asked follow-up questions.
Strongly Disagree Strongly Agree

I would be willing to answer follow-up questions from an AI if answering questions led to better results.
Strongly Disagree Strongly Agree

I liked that the AI showed me two options to pick between, instead of only picking the option it thought was best.
Strongly Disagree Strongly Agree

Do you have any additional feedback or comments (optional)?

Dr. Kimberly Binsted, Principal Investigator
Project Title: Better Results Through Ambiguity Resolution: Large Language Models that Ask Clarifying Questions



Clarifying Questions Pilot Study

Thank You!

You have completed the experiment! Thank you for your participation. If you have any further questions, please contact Bernadette Tix at bjavery@hawaii.edu. Please close your browser to exit.

If you would like to keep a copy of the documents you created, you can copy them below. Once you exit the browser or navigate away from this page, you will no longer be able to access these documents.

Document #1

Artificial Intelligence (AI) has been increasingly integrated into various aspects of the workplace, offering a multitude of potential benefits to businesses and employees alike. Firstly, AI can significantly enhance productivity and efficiency in the workplace. With AI-powered tools and systems, repetitive and time-consuming tasks can be automated, allowing employees to focus on more strategic and creative endeavors. This not only streamlines operations but also reduces the likelihood of human errors, ultimately improving overall productivity levels.

Moreover, AI can revolutionize decision-making

Document #2

Title: Unlocking Potential: Exploring the Benefits of Modern AI in the Workplace

Introduction:
Modern Artificial Intelligence (AI) has transformed the landscape of businesses by offering a myriad of benefits that redefine efficiency, productivity, and innovation in the workplace. In this essay, we will delve into the potential advantages of AI across various key areas, including automation, data analysis, and decision-making support. By understanding how AI can enhance these aspects, businesses can harness the power of technology to drive success and competitive advantage.

Dr. Kimberly Binsted, Principal Investigator
Project Title: Better Results Through Ambiguity Resolution: Large Language Models that Ask Clarifying Questions

Appendix D: Pilot Study Consent Message

University of Hawai‘i Online Survey Consent Form

Aloha! You are invited to participate in a pilot study conducted by Bernadette Tix from the Department of Information and Computer Science at the University of Hawai‘i. This research project is a part of a PhD Dissertation studying document creation by Artificial Intelligence programs. The purpose of the pilot study is to test whether the format of the proposed study has any major flaws or areas in need of improvement.

What am I being asked to do?

If you participate in this project, you will be asked to fill out a survey, as well as engage in a back-and-forth conversation with an AI chatbot about creating a document for your personal use, such as a letter, memo, or email. For the pilot study, you will join a Zoom call with Bernadette Tix and she will walk you through the process. You will be asked to speak through your thought process out loud and share your screen so your interactions with the AI are visible over the Zoom meeting. Audio and video of the call will be recorded. You are not required to use or turn on a camera for yourself if you do not wish to.

Taking part in this study is your choice.

Your participation in this project is completely voluntary. You may stop participating at any time. If you stop being in the study, there will be no penalty or loss to you.

Why is this study being done?

The purpose of this project is to determine whether AI can produce better outputs by asking clarifying questions about what a user needs before creating AI-generated documents. I am seeking adult volunteers to generate documents using an AI, and to give feedback on the output which will allow us to determine whether or not the clarifying questions asked by the AI improved the AI-generated documents. The pilot study will be used to determine whether there are flaws in the study procedures in preparation for launching a larger study that will allow users to anonymously complete the study online.

What will happen if I decide to take part in this study?

The survey will consist of a brief interaction with an AI chatbot, most likely lasting 5-10 minutes, as well as brief entry and exit surveys. The entry survey will ask for basic demographic information such as your age, gender, and prior experience with AI. When interacting with the AI, you will be asked to think of a document that you would like the AI to write and will answer questions the AI comes up with about your document needs. The exit survey will ask you to rate two different outputs from the AI with questions like, “Would you find this document usable in its current state?” “Is this document close to what you had in mind when you made the request?” and “How would you rate the overall quality of this document?” The survey is accessed on a website to which I will provide you a link. The total length of the survey, including interaction with the AI, is expected to last 15-20 minutes. If you decide to take part in the pilot study, please email Bernadette Tix at bjavery@hawaii.edu to set up a time for a Zoom call.

What are the risks and benefits of taking part in this study?

I believe there is little risk to you for participating in this research project. You may become stressed or uncomfortable answering any of the survey questions. If you do become stressed or uncomfortable, you can skip the question or take a break. You can also stop taking the survey or you can withdraw from the project altogether.

There will be no direct benefit to you for participating in this survey. The results of this project may help improve AI’s ability to work in cooperation with humans by giving AI the ability to ask questions and engage in dialog with humans before executing commands.

Confidentiality and Privacy:

I will not ask you for any personal information, such as your name or address. Please do not include any personal information in your survey responses or in your responses to the AI. I will keep all study data secure on a private server. Only my University of Hawai'i advisor and I will have access to the information. Other agencies that have legal permission have the right to review research records. The University of Hawai'i Human Studies Program has the right to review research records for this study.

Future Research Studies:

The data from this study will not be used or distributed for future research studies.

Questions: If you have any questions about this study, please email me at bjavery@hawaii.edu You may also contact my advisor, Dr. Kimberly Binsted, at binsted@hawaii.edu

You may contact the UH Human Studies Program at 808.956.5007 or uhirb@hawaii.edu to discuss problems, concerns and questions, obtain information, or offer input with an informed individual who is unaffiliated with the specific research protocol. Please visit http://go.hawaii.edu/jRd for more information on your rights as a research participant.

If you agree to participate in this project, please give electronic consent by checking the box below:

Appendix E: Pilot Study Example Conversation

This is an example of a full conversation between the user and CQDG, leading to the creation of a QA Document. The conversation is first shown as it would have appeared to the user, and then shown as a full prompt log of prompts and responses with GPT 3.5. The specific prompts and answers were input by me and are not actual user data, which is kept confidential per the user agreement. However, the prompts and responses are real data generated by CQDG in response to my own fake prompt and answers.

The Conversation as it would Appear to the User

User: *Write me an essay on the topic of chivalry in 14th century Western Europe.*

Assistant: *1. Can you specify which particular aspects of chivalry in the 14th century you are most interested in exploring? Options might include the knightly code of conduct, the social and political roles of knights, the portrayal of chivalry in contemporary literature, or the practice of chivalry in warfare.*

User: *A brief survey of all of the above would be excellent.*

Assistant: *2. Are there certain countries or regions within Western Europe whose traditions of chivalry during the 14th century you find especially compelling or relevant to your essay's focus? If so, could you list these countries or regions for a targeted exploration?*

User: *The country of greatest interest is England, but compare and contrast English ideals of chivalry with those found in Italy and France.*

Assistant: *3. Would you prefer the essay to draw primarily from primary historical documents and contemporary accounts to reflect the authentic views of the era, or would you like a balance between historical sources and modern scholarly analysis to understand how perceptions of chivalry have evolved over time? Also, are there any specific works or scholars in the field of chivalric studies whose insights you consider essential to include in the essay?*

User: *Include the modern scholarship of Maurice Keen and Brian Price, as well as the primary source of Geoffroi de Charny's "A Knight's own Book of Chivalry."*

Assistant: *Title: Chivalry in 14th Century Western Europe: A Comparative Analysis Introduction:*

Chivalry, a complex and multifaceted concept, held significant importance in 14th-century Western Europe. This essay aims to provide a comprehensive survey of various aspects of chivalry during this epoch, focusing on England, Italy, and France. By examining the knightly code of conduct, social roles, literary portrayals, and chivalric practices in warfare, we can gain a deeper understanding of how chivalry manifested in different cultural contexts.

Analysis:

1. Knightly Code of Conduct: - In England, chivalry was guided by a strong sense of duty, loyalty, and honor. Knights were expected to uphold virtues such as courage, courtesy, and justice, as exemplified in the Arthurian legends and chivalric romances. - In Italy, chivalry was intertwined with the ideals of courtly love and the concept of the ideal knight as a refined, cultured gentleman. - In France, chivalry was deeply rooted in feudal traditions, emphasizing fealty to one's lord, prowess in battle, and the protection of the weak and defenseless.

2. Social and Political Roles of Knights: - English knights played vital roles in maintaining order and stability within the feudal system, serving as protectors of the realm and administrators of justice. - Italian knights often served as diplomats, poets, and patrons of the arts, embodying the fusion of martial prowess with artistic refinement. - French knights were central to the feudal hierarchy, supporting the nobility in military campaigns and exercising authority over vassals and serfs.

3. Portrayal of Chivalry in Contemporary Literature: - The works of Geoffrey Chaucer in England, Giovanni Boccaccio in Italy, and Chrétien de Troyes in France reflect the diverse interpretations of chivalry in 14th-century literature, showcasing the ideals and contradictions inherent in the knightly ethos.

4. Practice of Chivalry in Warfare: - English knights, such as those who fought in the Hundred Years' War, demonstrated valor and military skill on the battlefield, upholding the chivalric tradition of honor in combat. - Italian condottieri embodied a mercenary form of chivalry, serving various city-states for financial gain while adhering to a code of conduct in warfare. - French knights engaged in tournaments and jousts as displays of chivalric prowess, emphasizing the importance of martial skills and courtesy in competitive combat.

Conclusion: In conclusion, the study of chivalry in 14th-century Western Europe reveals a nuanced and evolving concept that varied between England, Italy, and France. Through a comparative *analysis of the knightly ideals, social roles, literary representations, and martial practices of these regions, we can appreciate the complex tapestry of chivalric traditions that shaped the cultural landscape of the Middle Ages. By incorporating insights from scholars such as Maurice Keen and Brian Price, as well as primary sources like Geoffroi de Charny's "A Knight's own Book of Chivalry," this essay seeks to illuminate the enduring legacy of chivalry and its enduring relevance in the study of medieval history and culture.*

Full prompt exchange with GPT-3.5

*Note that for the pilot study, each prompt-response pair is entirely independent. The context of previous prompts and responses is not retained between prompts.

Prompt: *A user is requesting the creation of a new document. This is their request:*

user: "Write me an essay on the topic of chivalry in 14th century Western Europe."

Identify any areas of significant ambiguity in the prompt, areas that could benefit from more thought or attention from the user, or helpful tips the user may not have considered. Write these out in a short list. Include exactly 3 items in the list.

Response: *1. **Scope and Focus of the Essay**: The topic of chivalry in 14th century Western Europe is broad, encompassing various aspects such as historical context, social structures, literary representations, and regional variations. Would you prefer the essay to provide a general overview, or would you like it to concentrate on specific facets of chivalry, such as the codes of conduct, the role of chivalry in warfare, or its impact on contemporary literature and art?*

*2. **Geographical Focus**: "Western Europe" includes a variety of countries with distinct cultures and historical contexts. Focusing on a particular region or comparing a few selected regions could provide more depth to the essay. Are there specific countries in Western Europe, such as France, England, or the Holy Roman Empire, whose chivalric traditions you are especially interested in exploring?*

*3. **Sources and Scholarship**: The study of chivalry is rich with primary and secondary sources, from medieval texts like "The Song of Roland" or the works of Chrétien de Troyes, to modern scholarly analyses. Would you prefer the essay to emphasize primary historical documents,* *contemporary accounts, and literary works, or to focus on the analysis and interpretations of modern historians and scholars? Additionally, considering the vast amount of literature on the subject, are there any particular scholars or works you want to ensure are referenced in the essay?*

Prompt: *Consider the following request:*

user: "Write me an essay on the topic of chivalry in 14th century Western Europe."

and the following identified points of ambiguity:

*"1. **Scope and Focus of the Essay**: The topic of chivalry in 14th century Western Europe is broad, encompassing various aspects such as historical context, social structures, literary representations, and regional variations. Would you prefer the essay to provide a general overview, or would you like it to concentrate on specific facets of chivalry, such as the codes of conduct, the role of chivalry in warfare, or its impact on contemporary literature and art?*

*2. **Geographical Focus**: "Western Europe" includes a variety of countries with distinct cultures and historical contexts. Focusing on a particular region or comparing a few selected regions could provide more depth to the essay. Are there specific countries in Western Europe, such as France, England, or the Holy Roman Empire, whose chivalric traditions you are especially interested in exploring?*

*3. **Sources and Scholarship**: The study of chivalry is rich with primary and secondary sources, from medieval texts like "The Song of Roland" or the works of Chrétien de Troyes, to modern scholarly analyses. Would you prefer the essay to emphasize primary historical documents, contemporary accounts, and literary works, or to focus on the analysis and interpretations of modern historians and scholars? Additionally, considering the vast amount of literature on the subject, are there any particular scholars or works you want to ensure are referenced in the essay?"*

Respond as though this request was just made by the user. Start with a friendly greeting, then ask a respectful and concise question to clarify the user's needs, keeping in mind the ambiguity you just identified. Format your response as a numbered list of exactly 3 questions.

Response: *1. Can you specify which particular aspects of chivalry in the 14th century you are most interested in exploring? Options might include the knightly code of conduct, the social and political* *roles of knights, the portrayal of chivalry in contemporary literature, or the practice of chivalry in warfare.*

2. Are there certain countries or regions within Western Europe whose traditions of chivalry during the 14th century you find especially compelling or relevant to your essay's focus? If so, could you list these countries or regions for a targeted exploration?

3. Would you prefer the essay to draw primarily from primary historical documents and contemporary accounts to reflect the authentic views of the era, or would you like a balance between historical sources and modern scholarly analysis to understand how perceptions of chivalry have evolved over time? Also, are there any specific works or scholars in the field of chivalric studies whose insights you consider essential to include in the essay?

Prompt: *Consider the following exchange. Attempt to create the document requested by the user, considering the answers they gave when asked for details.*

user: "Write me an essay on the topic of chivalry in 14th century Western Europe."

assistant: " 1. Can you specify which particular aspects of chivalry in the 14th century you are most interested in exploring? Options might include the knightly code of conduct, the social and political roles of knights, the portrayal of chivalry in contemporary literature, or the practice of chivalry in warfare. "

user: "A brief survey of all of the above would be excellent."

assistant: "2. Are there certain countries or regions within Western Europe whose traditions of chivalry during the 14th century you find especially compelling or relevant to your essay's focus? If so, could you list these countries or regions for a targeted exploration? "

user: "The country of greatest interest is England, but compare and contrast English ideals of chivalry with those found in Italy and France." assistant: "

3. Would you prefer the essay to draw primarily from primary historical documents and contemporary accounts to reflect the authentic views of the era, or would you like a balance between historical sources and modern scholarly analysis to understand how perceptions of *chivalry have evolved over time? Also, are there any specific works or scholars in the field of chivalric studies whose insights you consider essential to include in the essay?"*

user: "Include the modern scholarship of Maurice Keen and Brian Price, as well as the primary source of Geoffroi de Charny's "A Knight's own Book of Chivalry.""

Response: *Title: Chivalry in 14th Century Western Europe: A Comparative Analysis Introduction:*

Chivalry, a complex and multifaceted concept, held significant importance in 14th-century Western Europe. This essay aims to provide a comprehensive survey of various aspects of chivalry during this epoch, focusing on England, Italy, and France. By examining the knightly code of conduct, social roles, literary portrayals, and chivalric practices in warfare, we can gain a deeper understanding of how chivalry manifested in different cultural contexts.

Analysis:

1. Knightly Code of Conduct: - In England, chivalry was guided by a strong sense of duty, loyalty, and honor. Knights were expected to uphold virtues such as courage, courtesy, and justice, as exemplified in the Arthurian legends and chivalric romances. - In Italy, chivalry was intertwined with the ideals of courtly love and the concept of the ideal knight as a refined, cultured gentleman. - In France, chivalry was deeply rooted in feudal traditions, emphasizing fealty to one's lord, prowess in battle, and the protection of the weak and defenseless.

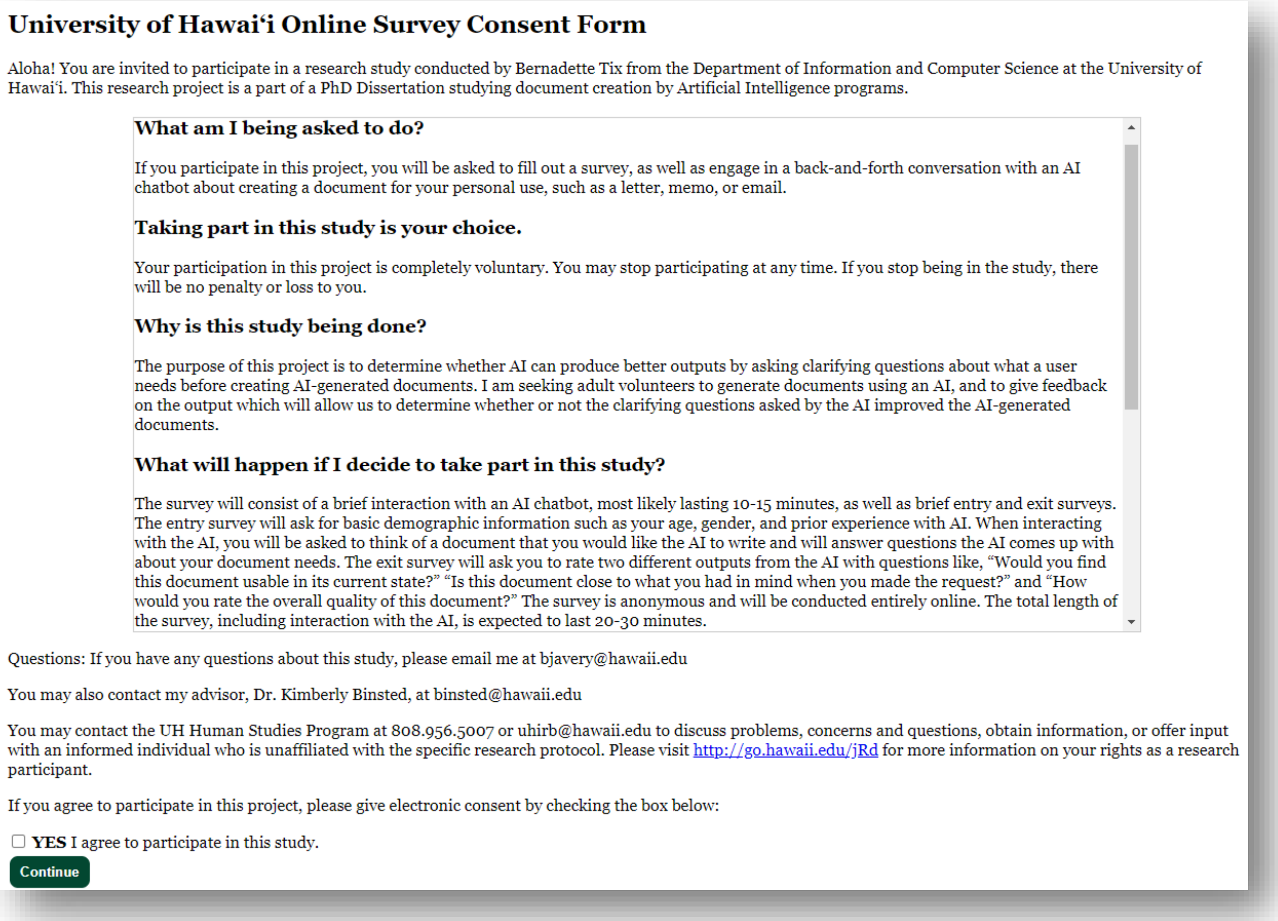
2. Social and Political Roles of Knights: - English knights played vital roles in maintaining order and stability within the feudal system, serving as protectors of the realm and administrators of justice. - Italian knights often served as diplomats, poets, and patrons of the arts, embodying the fusion of martial prowess with artistic refinement. - French knights were central to the feudal hierarchy, supporting the nobility in military campaigns and exercising authority over vassals and serfs.

3. Portrayal of Chivalry in Contemporary Literature: - The works of Geoffrey Chaucer in England, Giovanni Boccaccio in Italy, and Chrétien de Troyes in France reflect the diverse interpretations of chivalry in 14th-century literature, showcasing the ideals and contradictions inherent in the knightly ethos.

4. Practice of Chivalry in Warfare: - English knights, such as those who fought in the Hundred Years' War, demonstrated valor and military skill on the battlefield, upholding the chivalric tradition of *honor in combat. - Italian condottieri embodied a mercenary form of chivalry, serving various city-states for financial gain while adhering to a code of conduct in warfare. - French knights engaged in tournaments and jousts as displays of chivalric prowess, emphasizing the importance of martial skills and courtesy in competitive combat.*

Conclusion: In conclusion, the study of chivalry in 14th-century Western Europe reveals a nuanced and evolving concept that varied between England, Italy, and France. Through a comparative analysis of the knightly ideals, social roles, literary representations, and martial practices of these regions, we can appreciate the complex tapestry of chivalric traditions that shaped the cultural landscape of the Middle Ages. By incorporating insights from scholars such as Maurice Keen and Brian Price, as well as primary sources like Geoffroi de Charny's "A Knight's own Book of Chivalry," this essay seeks to illuminate the enduring legacy of chivalry and its enduring relevance in the study of medieval history and culture.

Appendix F: Full Study Screenshots All Steps



University of Hawai‘i Online Survey Consent Form

Aloha! You are invited to participate in a research study conducted by Bernadette Tix from the Department of Information and Computer Science at the University of Hawai‘i. This research project is a part of a PhD Dissertation studying document creation by Artificial Intelligence programs.

What am I being asked to do?

If you participate in this project, you will be asked to fill out a survey, as well as engage in a back-and-forth conversation with an AI chatbot about creating a document for your personal use, such as a letter, memo, or email.

Taking part in this study is your choice.

Your participation in this project is completely voluntary. You may stop participating at any time. If you stop being in the study, there will be no penalty or loss to you.

Why is this study being done?

The purpose of this project is to determine whether AI can produce better outputs by asking clarifying questions about what a user needs before creating AI-generated documents. I am seeking adult volunteers to generate documents using an AI, and to give feedback on the output which will allow us to determine whether or not the clarifying questions asked by the AI improved the AI-generated documents.

What will happen if I decide to take part in this study?

The survey will consist of a brief interaction with an AI chatbot, most likely lasting 10-15 minutes, as well as brief entry and exit surveys. The entry survey will ask for basic demographic information such as your age, gender, and prior experience with AI. When interacting with the AI, you will be asked to think of a document that you would like the AI to write and will answer questions the AI comes up with about your document needs. The exit survey will ask you to rate two different outputs from the AI with questions like, "Would you find this document usable in its current state?" "Is this document close to what you had in mind when you made the request?" and "How would you rate the overall quality of this document?" The survey is anonymous and will be conducted entirely online. The total length of the survey, including interaction with the AI, is expected to last 20-30 minutes.

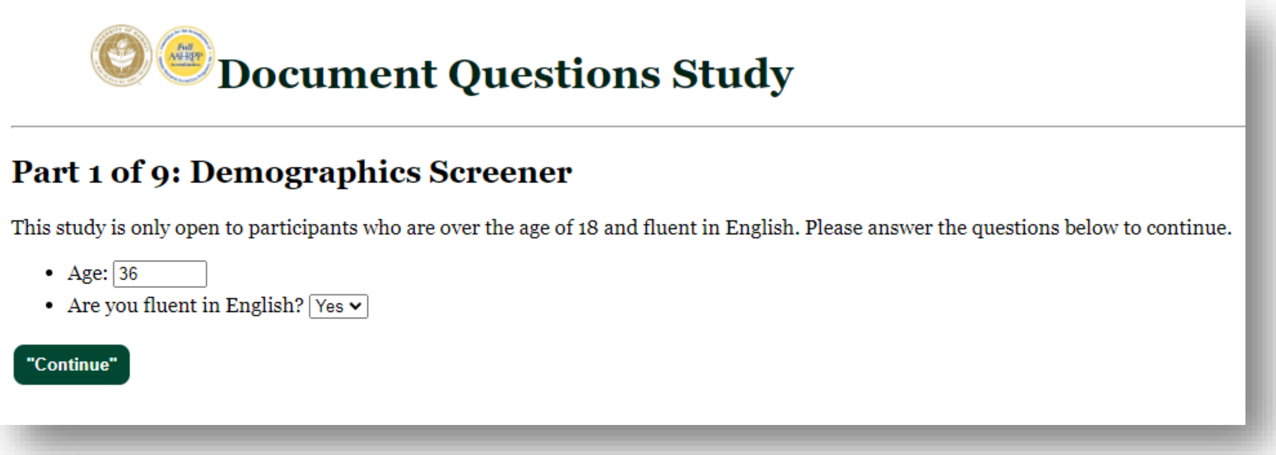
Questions: If you have any questions about this study, please email me at bjavery@hawaii.edu

You may also contact my advisor, Dr. Kimberly Binsted, at binsted@hawaii.edu

You may contact the UH Human Studies Program at 808.956.5007 or uhirb@hawaii.edu to discuss problems, concerns and questions, obtain information, or offer input with an informed individual who is unaffiliated with the specific research protocol. Please visit http://go.hawaii.edu/jRd for more information on your rights as a research participant.

If you agree to participate in this project, please give electronic consent by checking the box below:

☐ **YES** I agree to participate in this study.

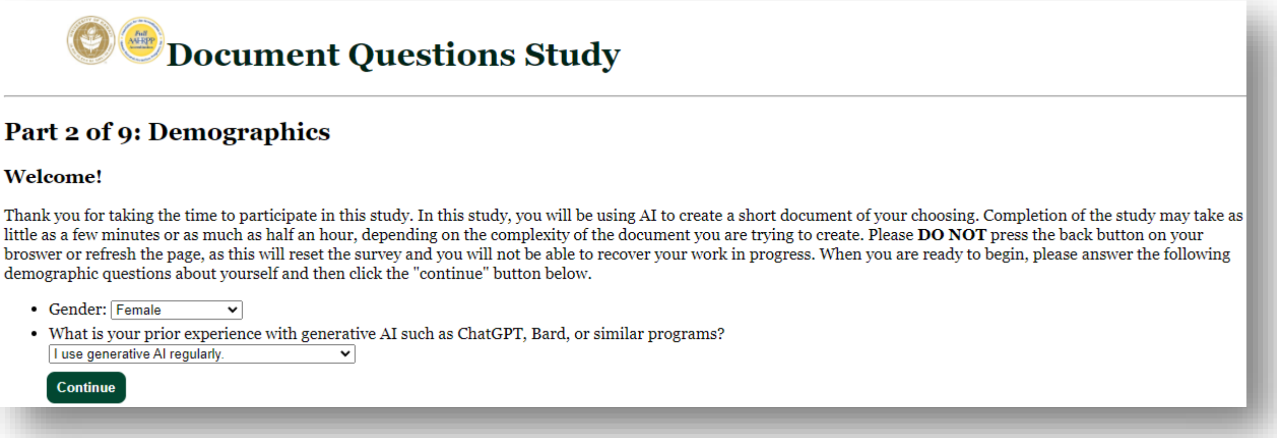


Document Questions Study

Part 1 of 9: Demographics Screener

This study is only open to participants who are over the age of 18 and fluent in English. Please answer the questions below to continue.

- Age: 36
- Are you fluent in English? Yes



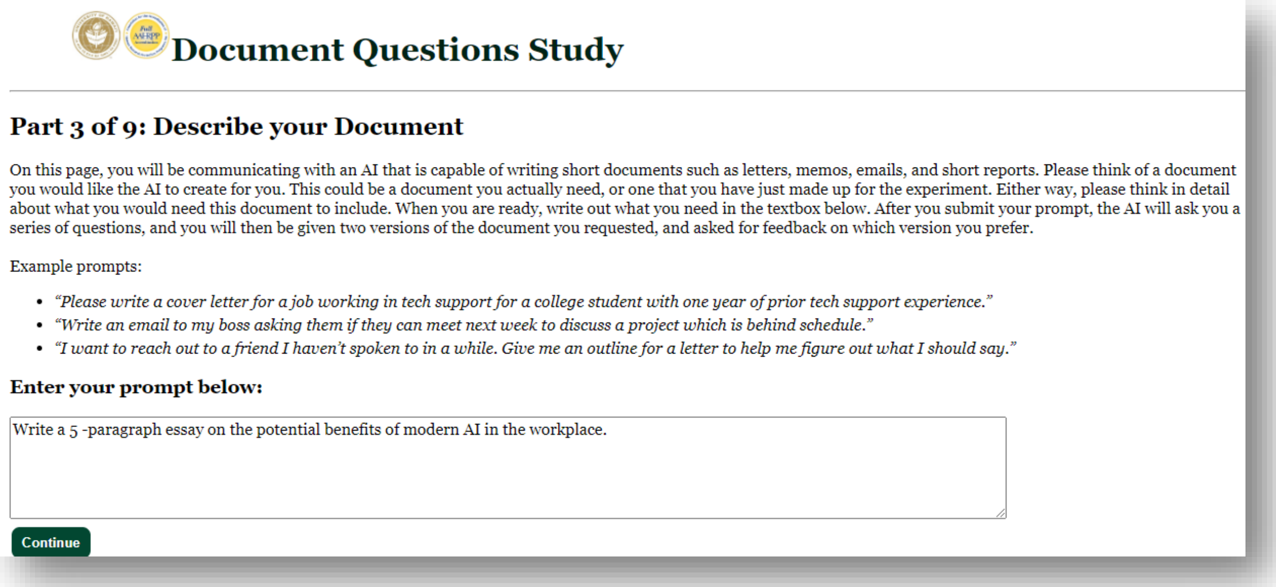
Document Questions Study

Part 2 of 9: Demographics

Welcome!

Thank you for taking the time to participate in this study. In this study, you will be using AI to create a short document of your choosing. Completion of the study may take as little as a few minutes or as much as half an hour, depending on the complexity of the document you are trying to create. Please **DO NOT** press the back button on your broswer or refresh the page, as this will reset the survey and you will not be able to recover your work in progress. When you are ready to begin, please answer the following demographic questions about yourself and then click the "continue" button below.

- Gender: Female
- What is your prior experience with generative AI such as ChatGPT, Bard, or similar programs?
  I use generative AI regularly.



Document Questions Study

Part 3 of 9: Describe your Document

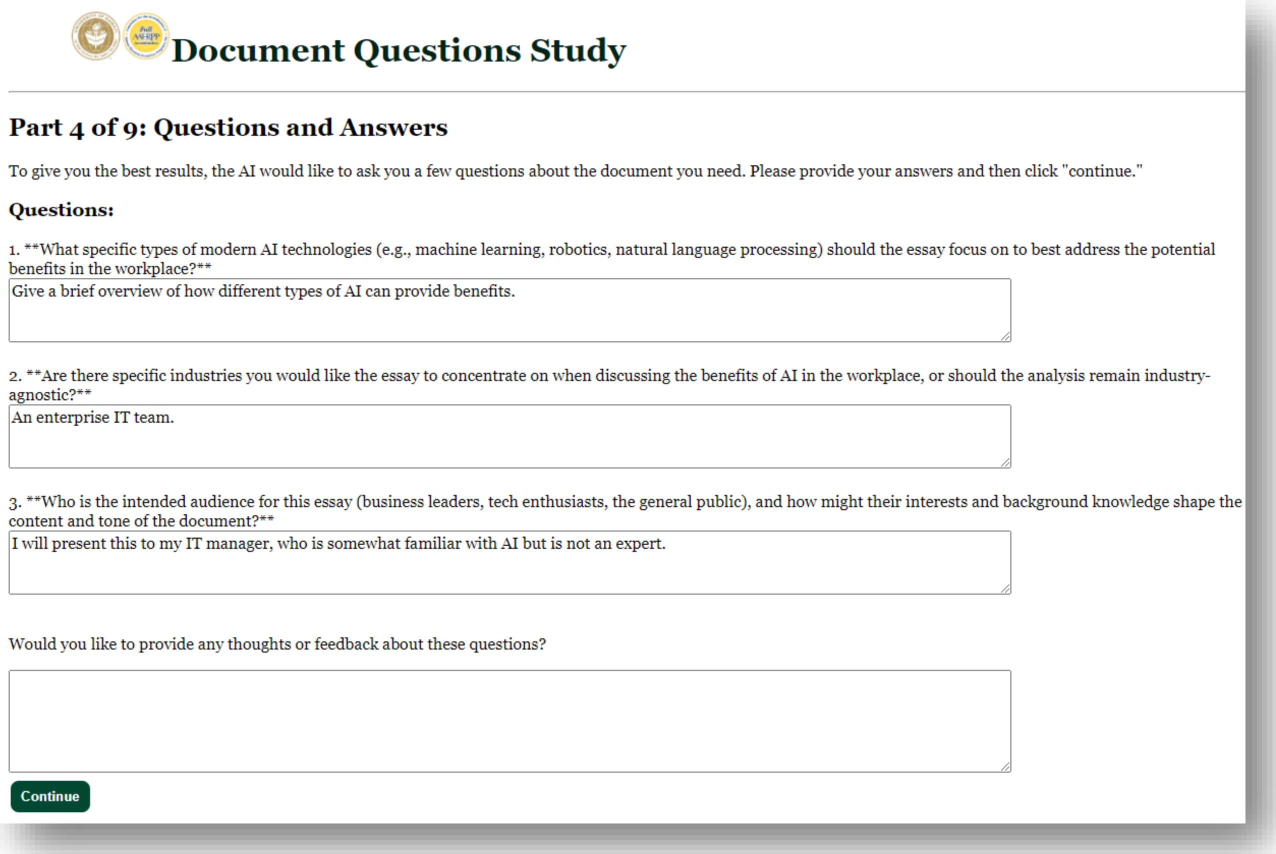
On this page, you will be communicating with an AI that is capable of writing short documents such as letters, memos, emails, and short reports. Please think of a document you would like the AI to create for you. This could be a document you actually need, or one that you have just made up for the experiment. Either way, please think in detail about what you would need this document to include. When you are ready, write out what you need in the textbox below. After you submit your prompt, the AI will ask you a series of questions, and you will then be given two versions of the document you requested, and asked for feedback on which version you prefer.

Example prompts:

- *"Please write a cover letter for a job working in tech support for a college student with one year of prior tech support experience."*
- *"Write an email to my boss asking them if they can meet next week to discuss a project which is behind schedule."*
- *"I want to reach out to a friend I haven't spoken to in a while. Give me an outline for a letter to help me figure out what I should say."*

Enter your prompt below:

Write a 5 -paragraph essay on the potential benefits of modern AI in the workplace.



Document Questions Study

Part 4 of 9: Questions and Answers

To give you the best results, the AI would like to ask you a few questions about the document you need. Please provide your answers and then click "continue."

Questions:

1. **What specific types of modern AI technologies (e.g., machine learning, robotics, natural language processing) should the essay focus on to best address the potential benefits in the workplace?**

Give a brief overview of how different types of AI can provide benefits.

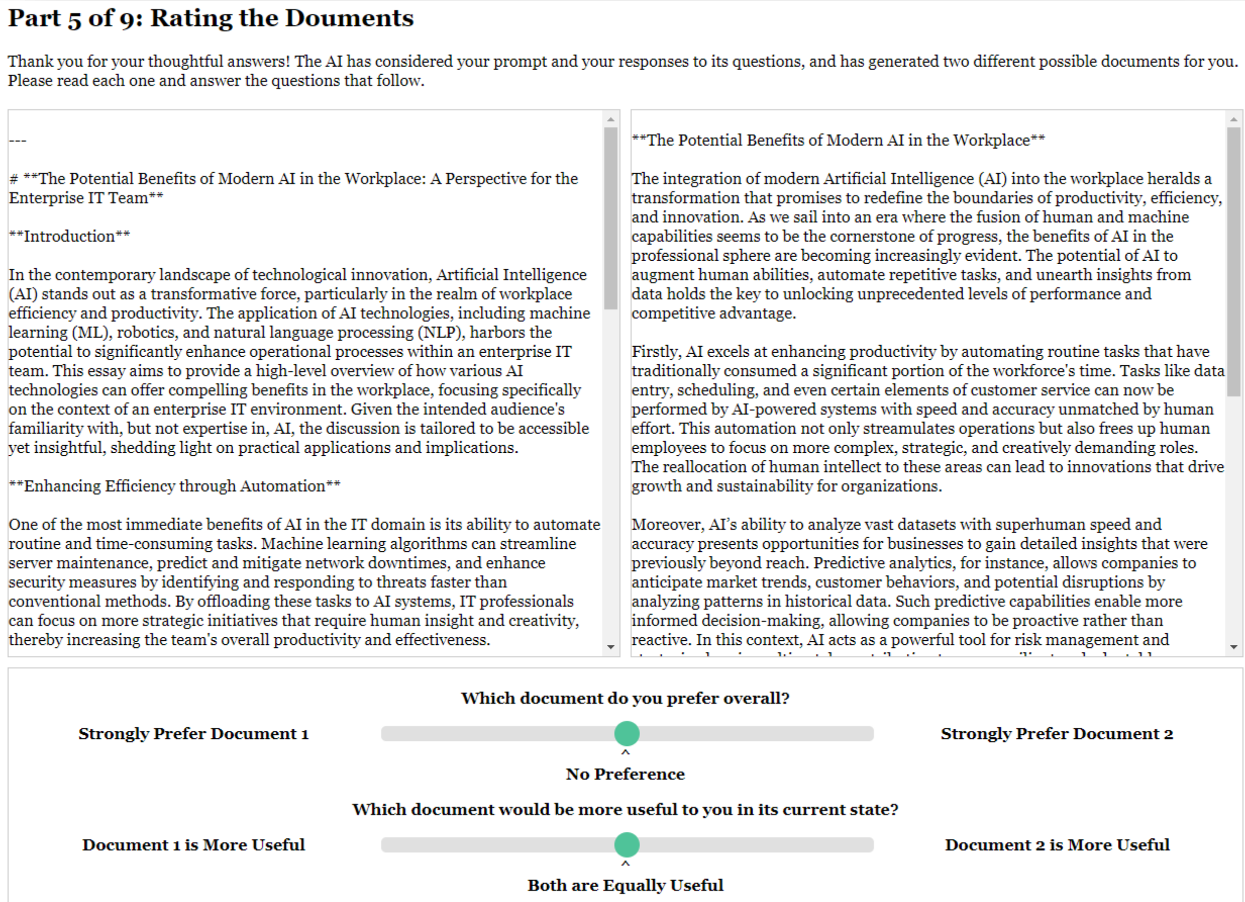
2. **Are there specific industries you would like the essay to concentrate on when discussing the benefits of AI in the workplace, or should the analysis remain industry-agnostic?**

An enterprise IT team.

3. **Who is the intended audience for this essay (business leaders, tech enthusiasts, the general public), and how might their interests and background knowledge shape the content and tone of the document?**

I will present this to my IT manager, who is somewhat familiar with AI but is not an expert.

Would you like to provide any thoughts or feedback about these questions?



Part 5 of 9: Rating the Douments

Thank you for your thoughtful answers! The AI has considered your prompt and your responses to its questions, and has generated two different possible documents for you. Please read each one and answer the questions that follow.

The Potential Benefits of Modern AI in the Workplace: A Perspective for the Enterprise IT Team

Introduction

In the contemporary landscape of technological innovation, Artificial Intelligence (AI) stands out as a transformative force, particularly in the realm of workplace efficiency and productivity. The application of AI technologies, including machine learning (ML), robotics, and natural language processing (NLP), harbors the potential to significantly enhance operational processes within an enterprise IT team. This essay aims to provide a high-level overview of how various AI technologies can offer compelling benefits in the workplace, focusing specifically on the context of an enterprise IT environment. Given the intended audience's familiarity with, but not expertise in, AI, the discussion is tailored to be accessible yet insightful, shedding light on practical applications and implications.

Enhancing Efficiency through Automation

One of the most immediate benefits of AI in the IT domain is its ability to automate routine and time-consuming tasks. Machine learning algorithms can streamline server maintenance, predict and mitigate network downtimes, and enhance security measures by identifying and responding to threats faster than conventional methods. By offloading these tasks to AI systems, IT professionals can focus on more strategic initiatives that require human insight and creativity, thereby increasing the team's overall productivity and effectiveness.

The Potential Benefits of Modern AI in the Workplace

The integration of modern Artificial Intelligence (AI) into the workplace heralds a transformation that promises to redefine the boundaries of productivity, efficiency, and innovation. As we sail into an era where the fusion of human and machine capabilities seems to be the cornerstone of progress, the benefits of AI in the professional sphere are becoming increasingly evident. The potential of AI to augment human abilities, automate repetitive tasks, and unearth insights from data holds the key to unlocking unprecedented levels of performance and competitive advantage.

Firstly, AI excels at enhancing productivity by automating routine tasks that have traditionally consumed a significant portion of the workforce's time. Tasks like data entry, scheduling, and even certain elements of customer service can now be performed by AI-powered systems with speed and accuracy unmatched by human effort. This automation not only streamulates operations but also frees up human employees to focus on more complex, strategic, and creatively demanding roles. The reallocation of human intellect to these areas can lead to innovations that drive growth and sustainability for organizations.

Moreover, AI's ability to analyze vast datasets with superhuman speed and accuracy presents opportunities for businesses to gain detailed insights that were previously beyond reach. Predictive analytics, for instance, allows companies to anticipate market trends, customer behaviors, and potential disruptions by analyzing patterns in historical data. Such predictive capabilities enable more informed decision-making, allowing companies to be proactive rather than reactive. In this context, AI acts as a powerful tool for risk management and

---

Which document do you prefer overall?

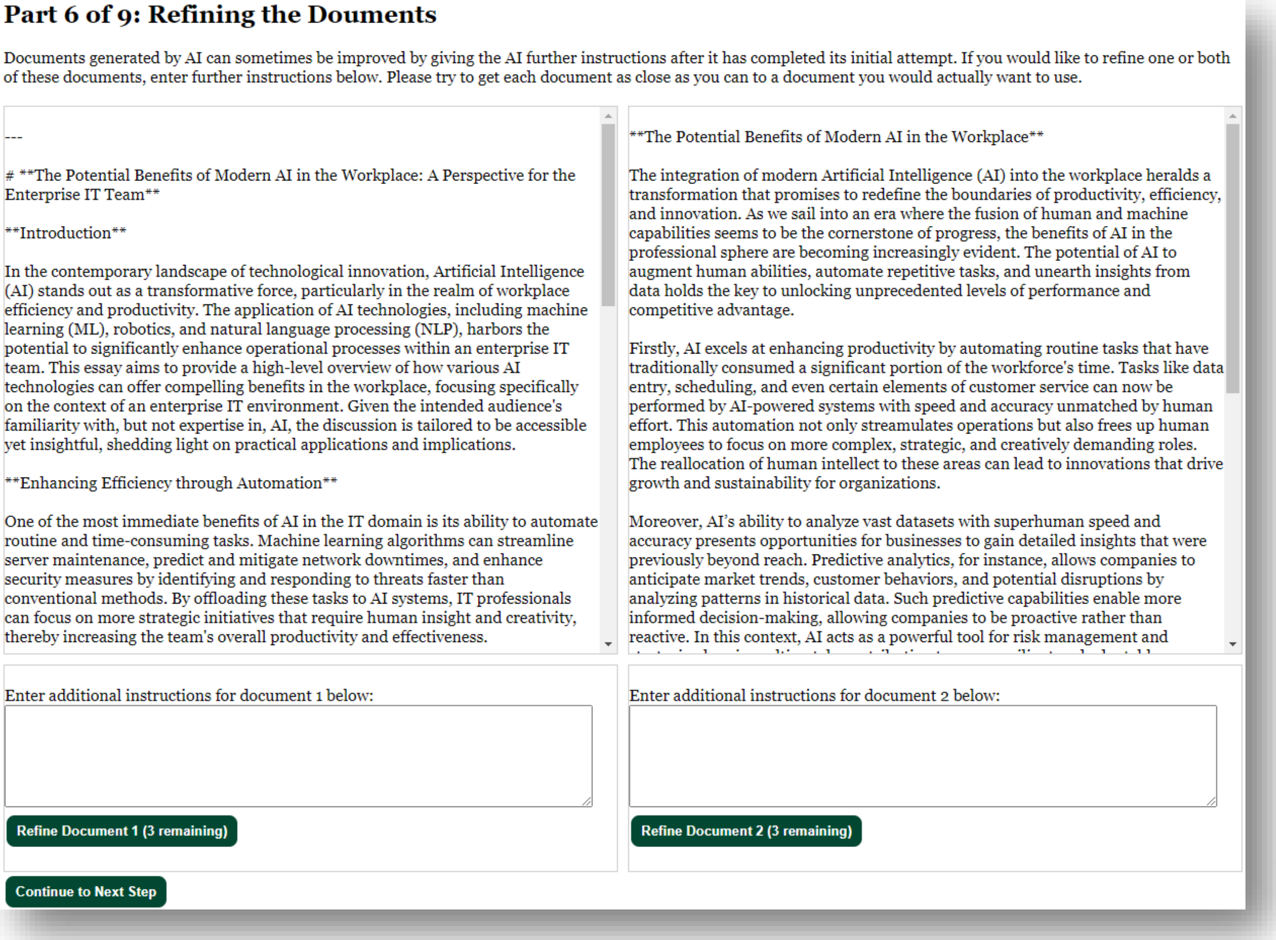
Strongly Prefer Document 1 — **Strongly Prefer Document 2**

No Preference

Which document would be more useful to you in its current state?

Document 1 is More Useful — **Document 2 is More Useful**

Both are Equally Useful



Part 6 of 9: Refining the Douments

Documents generated by AI can sometimes be improved by giving the AI further instructions after it has completed its initial attempt. If you would like to refine one or both of these documents, enter further instructions below. Please try to get each document as close as you can to a document you would actually want to use.

The Potential Benefits of Modern AI in the Workplace: A Perspective for the Enterprise IT Team

Introduction

In the contemporary landscape of technological innovation, Artificial Intelligence (AI) stands out as a transformative force, particularly in the realm of workplace efficiency and productivity. The application of AI technologies, including machine learning (ML), robotics, and natural language processing (NLP), harbors the potential to significantly enhance operational processes within an enterprise IT team. This essay aims to provide a high-level overview of how various AI technologies can offer compelling benefits in the workplace, focusing specifically on the context of an enterprise IT environment. Given the intended audience's familiarity with, but not expertise in, AI, the discussion is tailored to be accessible yet insightful, shedding light on practical applications and implications.

Enhancing Efficiency through Automation

One of the most immediate benefits of AI in the IT domain is its ability to automate routine and time-consuming tasks. Machine learning algorithms can streamline server maintenance, predict and mitigate network downtimes, and enhance security measures by identifying and responding to threats faster than conventional methods. By offloading these tasks to AI systems, IT professionals can focus on more strategic initiatives that require human insight and creativity, thereby increasing the team's overall productivity and effectiveness.

The Potential Benefits of Modern AI in the Workplace

The integration of modern Artificial Intelligence (AI) into the workplace heralds a transformation that promises to redefine the boundaries of productivity, efficiency, and innovation. As we sail into an era where the fusion of human and machine capabilities seems to be the cornerstone of progress, the benefits of AI in the professional sphere are becoming increasingly evident. The potential of AI to augment human abilities, automate repetitive tasks, and unearth insights from data holds the key to unlocking unprecedented levels of performance and competitive advantage.

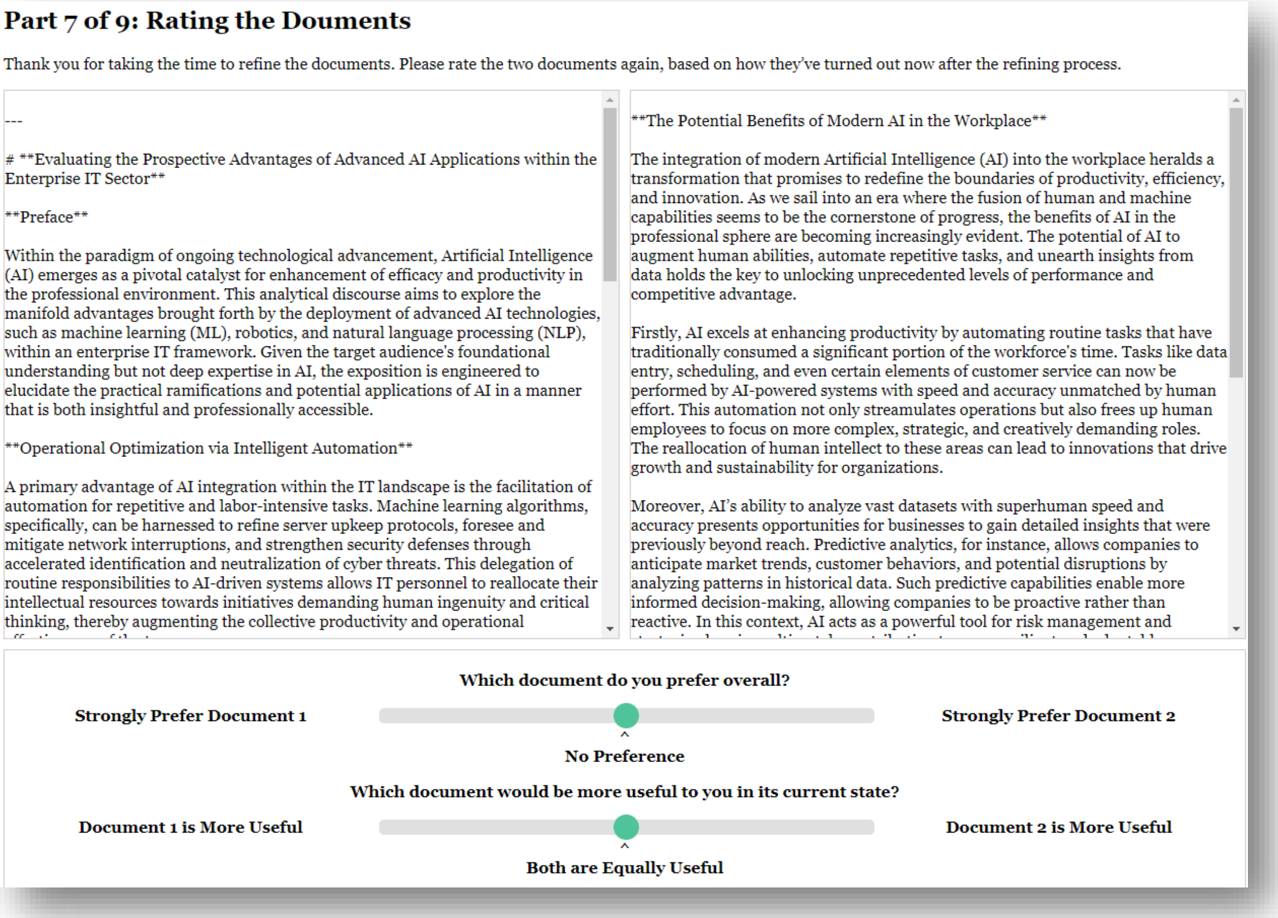
Firstly, AI excels at enhancing productivity by automating routine tasks that have traditionally consumed a significant portion of the workforce's time. Tasks like data entry, scheduling, and even certain elements of customer service can now be performed by AI-powered systems with speed and accuracy unmatched by human effort. This automation not only streamulates operations but also frees up human employees to focus on more complex, strategic, and creatively demanding roles. The reallocation of human intellect to these areas can lead to innovations that drive growth and sustainability for organizations.

Moreover, AI's ability to analyze vast datasets with superhuman speed and accuracy presents opportunities for businesses to gain detailed insights that were previously beyond reach. Predictive analytics, for instance, allows companies to anticipate market trends, customer behaviors, and potential disruptions by analyzing patterns in historical data. Such predictive capabilities enable more informed decision-making, allowing companies to be proactive rather than reactive. In this context, AI acts as a powerful tool for risk management and

Enter additional instructions for document 1 below:

Enter additional instructions for document 2 below:

Refine Document 1 (3 remaining)
Refine Document 2 (3 remaining)
Continue to Next Step



Part 7 of 9: Rating the Douments

Thank you for taking the time to refine the documents. Please rate the two documents again, based on how they've turned out now after the refining process.

Evaluating the Prospective Advantages of Advanced AI Applications within the Enterprise IT Sector

Preface

Within the paradigm of ongoing technological advancement, Artificial Intelligence (AI) emerges as a pivotal catalyst for enhancement of efficacy and productivity in the professional environment. This analytical discourse aims to explore the manifold advantages brought forth by the deployment of advanced AI technologies, such as machine learning (ML), robotics, and natural language processing (NLP), within an enterprise IT framework. Given the target audience's foundational understanding but not deep expertise in AI, the exposition is engineered to elucidate the practical ramifications and potential applications of AI in a manner that is both insightful and professionally accessible.

Operational Optimization via Intelligent Automation

A primary advantage of AI integration within the IT landscape is the facilitation of automation for repetitive and labor-intensive tasks. Machine learning algorithms, specifically, can be harnessed to refine server upkeep protocols, foresee and mitigate network interruptions, and strengthen security defenses through accelerated identification and neutralization of cyber threats. This delegation of routine responsibilities to AI-driven systems allows IT personnel to reallocate their intellectual resources towards initiatives demanding human ingenuity and critical thinking, thereby augmenting the collective productivity and operational

The Potential Benefits of Modern AI in the Workplace

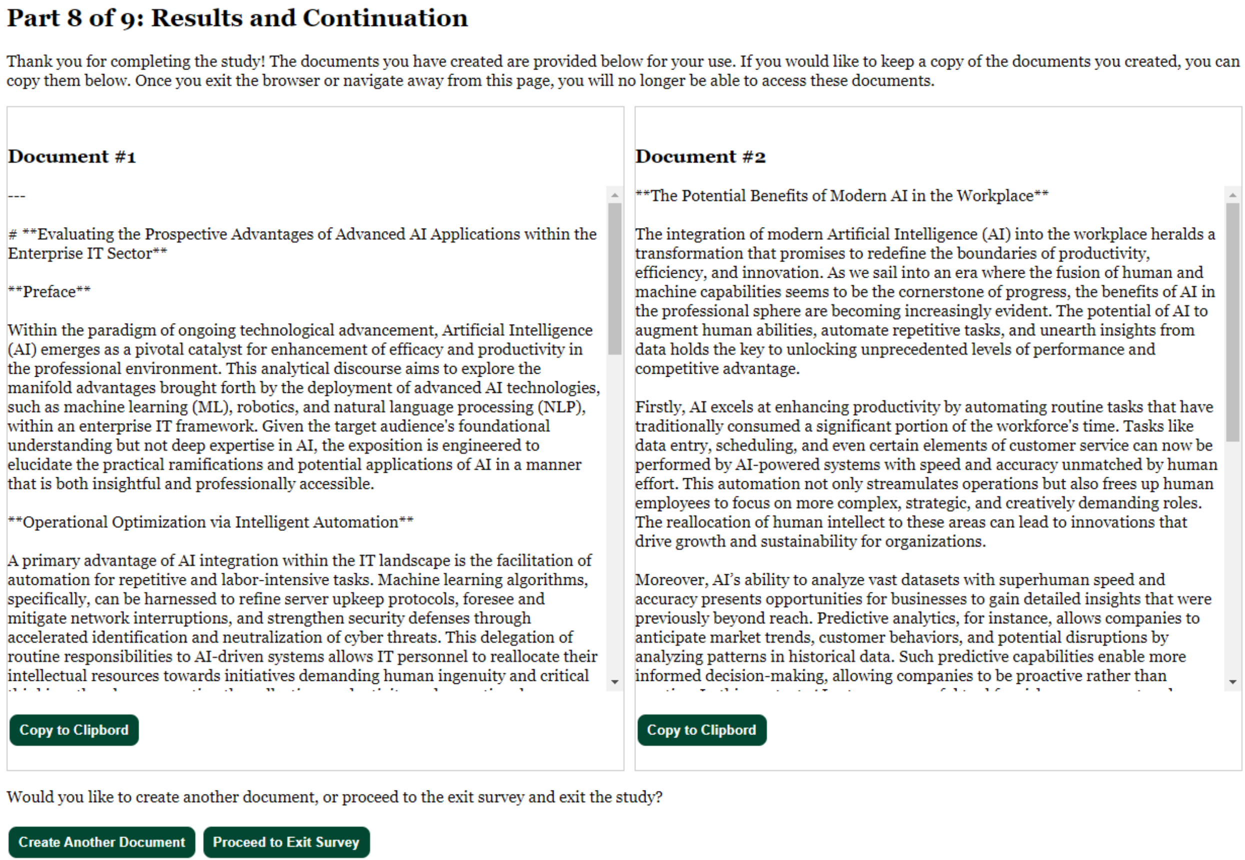
The integration of modern Artificial Intelligence (AI) into the workplace heralds a transformation that promises to redefine the boundaries of productivity, efficiency, and innovation. As we sail into an era where the fusion of human and machine capabilities seems to be the cornerstone of progress, the benefits of AI in the professional sphere are becoming increasingly evident. The potential of AI to augment human abilities, automate repetitive tasks, and unearth insights from data holds the key to unlocking unprecedented levels of performance and competitive advantage.

Firstly, AI excels at enhancing productivity by automating routine tasks that have traditionally consumed a significant portion of the workforce's time. Tasks like data entry, scheduling, and even certain elements of customer service can now be performed by AI-powered systems with speed and accuracy unmatched by human effort. This automation not only streamulates operations but also frees up human employees to focus on more complex, strategic, and creatively demanding roles. The reallocation of human intellect to these areas can lead to innovations that drive growth and sustainability for organizations.

Moreover, AI's ability to analyze vast datasets with superhuman speed and accuracy presents opportunities for businesses to gain detailed insights that were previously beyond reach. Predictive analytics, for instance, allows companies to anticipate market trends, customer behaviors, and potential disruptions by analyzing patterns in historical data. Such predictive capabilities enable more informed decision-making, allowing companies to be proactive rather than reactive. In this context, AI acts as a powerful tool for risk management and

Which document do you prefer overall?

Which document would be more useful to you in its current state?



Part 8 of 9: Results and Continuation

Thank you for completing the study! The documents you have created are provided below for your use. If you would like to keep a copy of the documents you created, you can copy them below. Once you exit the browser or navigate away from this page, you will no longer be able to access these documents.

Document #1

Evaluating the Prospective Advantages of Advanced AI Applications within the Enterprise IT Sector

Preface

Within the paradigm of ongoing technological advancement, Artificial Intelligence (AI) emerges as a pivotal catalyst for enhancement of efficacy and productivity in the professional environment. This analytical discourse aims to explore the manifold advantages brought forth by the deployment of advanced AI technologies, such as machine learning (ML), robotics, and natural language processing (NLP), within an enterprise IT framework. Given the target audience's foundational understanding but not deep expertise in AI, the exposition is engineered to elucidate the practical ramifications and potential applications of AI in a manner that is both insightful and professionally accessible.

Operational Optimization via Intelligent Automation

A primary advantage of AI integration within the IT landscape is the facilitation of automation for repetitive and labor-intensive tasks. Machine learning algorithms, specifically, can be harnessed to refine server upkeep protocols, foresee and mitigate network interruptions, and strengthen security defenses through accelerated identification and neutralization of cyber threats. This delegation of routine responsibilities to AI-driven systems allows IT personnel to reallocate their intellectual resources towards initiatives demanding human ingenuity and critical

Document #2

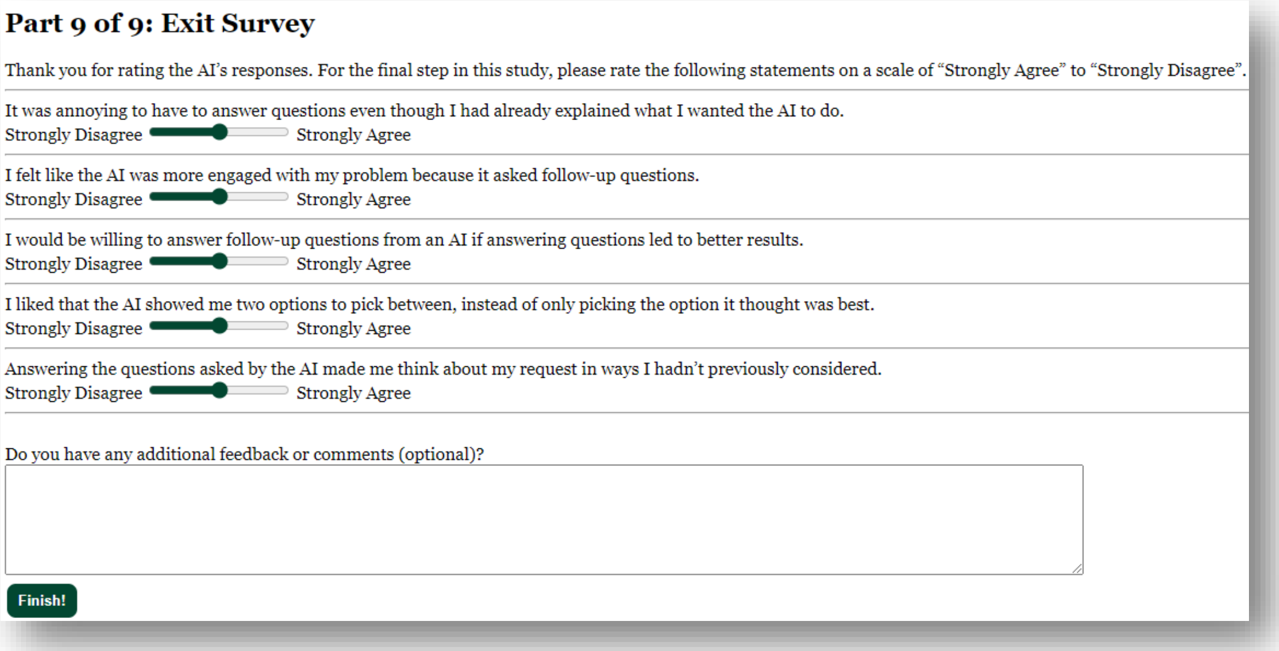
The Potential Benefits of Modern AI in the Workplace

The integration of modern Artificial Intelligence (AI) into the workplace heralds a transformation that promises to redefine the boundaries of productivity, efficiency, and innovation. As we sail into an era where the fusion of human and machine capabilities seems to be the cornerstone of progress, the benefits of AI in the professional sphere are becoming increasingly evident. The potential of AI to augment human abilities, automate repetitive tasks, and unearth insights from data holds the key to unlocking unprecedented levels of performance and competitive advantage.

Firstly, AI excels at enhancing productivity by automating routine tasks that have traditionally consumed a significant portion of the workforce's time. Tasks like data entry, scheduling, and even certain elements of customer service can now be performed by AI-powered systems with speed and accuracy unmatched by human effort. This automation not only streamulates operations but also frees up human employees to focus on more complex, strategic, and creatively demanding roles. The reallocation of human intellect to these areas can lead to innovations that drive growth and sustainability for organizations.

Moreover, AI's ability to analyze vast datasets with superhuman speed and accuracy presents opportunities for businesses to gain detailed insights that were previously beyond reach. Predictive analytics, for instance, allows companies to anticipate market trends, customer behaviors, and potential disruptions by analyzing patterns in historical data. Such predictive capabilities enable more informed decision-making, allowing companies to be proactive rather than

Would you like to create another document, or proceed to the exit survey and exit the study?



Part 9 of 9: Exit Survey

Thank you for rating the AI's responses. For the final step in this study, please rate the following statements on a scale of "Strongly Agree" to "Strongly Disagree".

It was annoying to have to answer questions even though I had already explained what I wanted the AI to do.
Strongly Disagree ● Strongly Agree

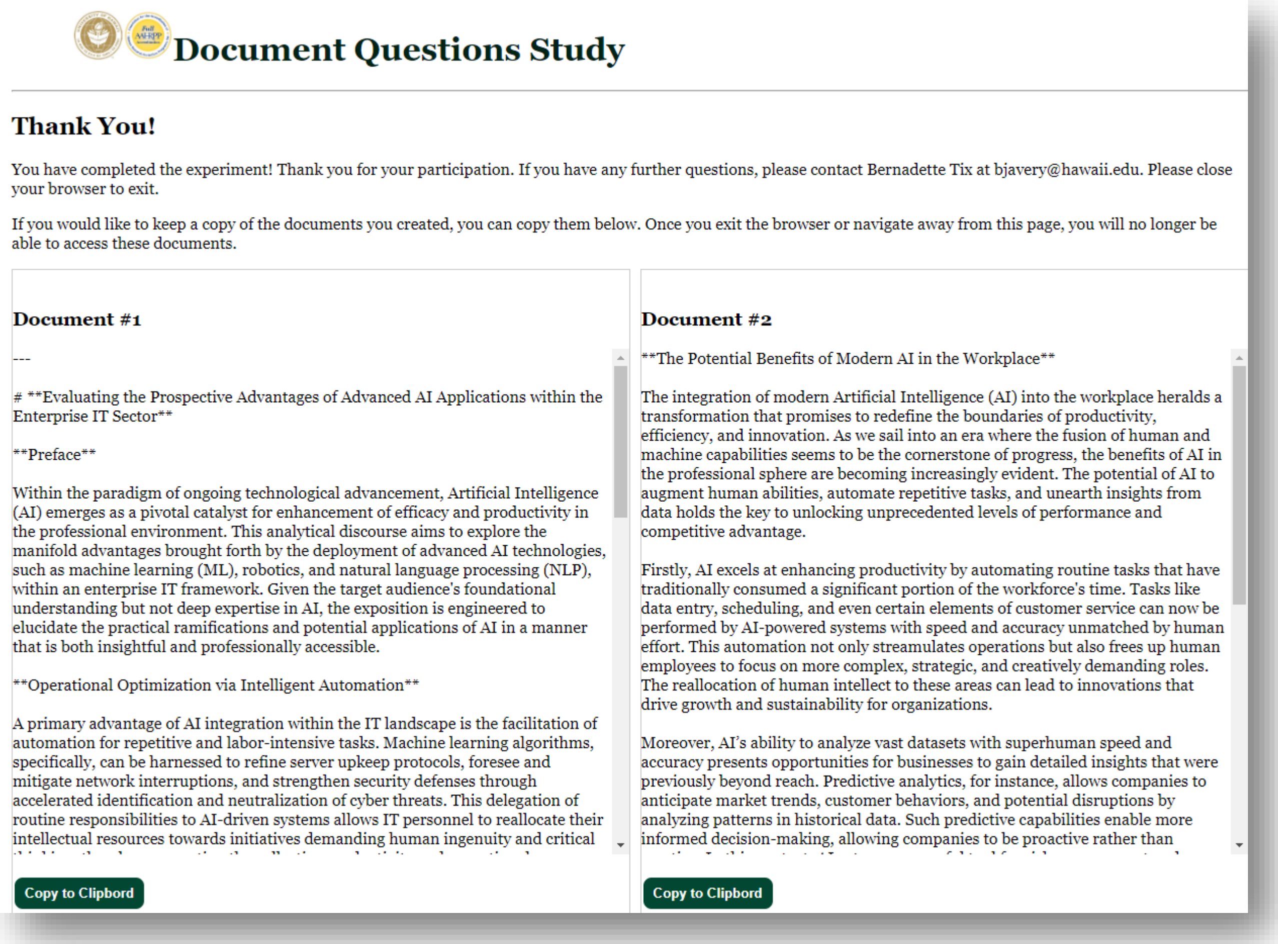
I felt like the AI was more engaged with my problem because it asked follow-up questions.
Strongly Disagree ● Strongly Agree

I would be willing to answer follow-up questions from an AI if answering questions led to better results.
Strongly Disagree ● Strongly Agree

I liked that the AI showed me two options to pick between, instead of only picking the option it thought was best.
Strongly Disagree ● Strongly Agree

Answering the questions asked by the AI made me think about my request in ways I hadn't previously considered.
Strongly Disagree ● Strongly Agree

Do you have any additional feedback or comments (optional)?



Document Questions Study

Thank You!

You have completed the experiment! Thank you for your participation. If you have any further questions, please contact Bernadette Tix at bjavery@hawaii.edu. Please close your browser to exit.

If you would like to keep a copy of the documents you created, you can copy them below. Once you exit the browser or navigate away from this page, you will no longer be able to access these documents.

Document #1

Evaluating the Prospective Advantages of Advanced AI Applications within the Enterprise IT Sector

Preface

Within the paradigm of ongoing technological advancement, Artificial Intelligence (AI) emerges as a pivotal catalyst for enhancement of efficacy and productivity in the professional environment. This analytical discourse aims to explore the manifold advantages brought forth by the deployment of advanced AI technologies, such as machine learning (ML), robotics, and natural language processing (NLP), within an enterprise IT framework. Given the target audience's foundational understanding but not deep expertise in AI, the exposition is engineered to elucidate the practical ramifications and potential applications of AI in a manner that is both insightful and professionally accessible.

Operational Optimization via Intelligent Automation

A primary advantage of AI integration within the IT landscape is the facilitation of automation for repetitive and labor-intensive tasks. Machine learning algorithms, specifically, can be harnessed to refine server upkeep protocols, foresee and mitigate network interruptions, and strengthen security defenses through accelerated identification and neutralization of cyber threats. This delegation of routine responsibilities to AI-driven systems allows IT personnel to reallocate their intellectual resources towards initiatives demanding human ingenuity and critical

Document #2

The Potential Benefits of Modern AI in the Workplace

The integration of modern Artificial Intelligence (AI) into the workplace heralds a transformation that promises to redefine the boundaries of productivity, efficiency, and innovation. As we sail into an era where the fusion of human and machine capabilities seems to be the cornerstone of progress, the benefits of AI in the professional sphere are becoming increasingly evident. The potential of AI to augment human abilities, automate repetitive tasks, and unearth insights from data holds the key to unlocking unprecedented levels of performance and competitive advantage.

Firstly, AI excels at enhancing productivity by automating routine tasks that have traditionally consumed a significant portion of the workforce's time. Tasks like data entry, scheduling, and even certain elements of customer service can now be performed by AI-powered systems with speed and accuracy unmatched by human effort. This automation not only streamulates operations but also frees up human employees to focus on more complex, strategic, and creatively demanding roles. The reallocation of human intellect to these areas can lead to innovations that drive growth and sustainability for organizations.

Moreover, AI's ability to analyze vast datasets with superhuman speed and accuracy presents opportunities for businesses to gain detailed insights that were previously beyond reach. Predictive analytics, for instance, allows companies to anticipate market trends, customer behaviors, and potential disruptions by analyzing patterns in historical data. Such predictive capabilities enable more informed decision-making, allowing companies to be proactive rather than

Appendix G: Full Study Consent Message

University of Hawaiʻi Online Survey Consent Form

Aloha! You are invited to participate in a research study conducted by Bernadette Tix from the Department of Information and Computer Science at the University of Hawaiʻi. This research project is a part of a PhD Dissertation studying document creation by Artificial Intelligence programs.

What am I being asked to do?

If you participate in this project, you will be asked to fill out a survey, as well as engage in a back-and-forth conversation with an AI chatbot about creating a document for your personal use, such as a letter, memo, or email.

Taking part in this study is your choice.

Your participation in this project is completely voluntary. You may stop participating at any time. If you stop being in the study, there will be no penalty or loss to you.

Why is this study being done?

The purpose of this project is to determine whether AI can produce better outputs by asking clarifying questions about what a user needs before creating AI-generated documents. I am seeking adult volunteers to generate documents using an AI, and to give feedback on the output which will allow us to determine whether or not the clarifying questions asked by the AI improved the AI-generated documents.

What will happen if I decide to take part in this study?

The survey will consist of a brief interaction with an AI chatbot, most likely lasting 10-15 minutes, as well as brief entry and exit surveys. The entry survey will ask for basic demographic information such as your age, gender, and prior experience with AI. When interacting with the AI, you will be asked to think of a document that you would like the AI to write and will answer questions the AI comes up with about your document needs. The exit survey will ask you to rate two different outputs from the AI with questions like, “Would you find this document usable in its current state?” “Is this document close to what you had in mind when you made the request?” and “How would you rate the overall quality of this document?" The survey is anonymous and will be conducted entirely online. The total length of the survey, including interaction with the AI, is expected to last 20-30 minutes.

What are the risks and benefits of taking part in this study?

I believe there is little risk to you for participating in this research project. You may become stressed or uncomfortable answering any of the survey questions. If you do become stressed or uncomfortable, you can take a break at any time. You can also stop taking the survey or you can withdraw from the project altogether.

There will be no direct benefit to you for participating in this survey. The results of this project may help improve AI's ability to work in cooperation with humans by giving AI the ability to ask questions and engage in dialog with humans before executing commands.

Confidentiality and Privacy:

You will not be asked for any personal information, such as your name or address. Please do not include any personal information in your survey responses or in your responses to the AI. I will keep all study data secure on a private server. Only my University of Hawai'i advisor and I will have access to the information. Other agencies that have legal permission have the right to review research records. The University of Hawai'i Human Studies Program has the right to review research records for this study.

Future Research Studies:

The data from this study will not be used or distributed for future research studies.

Questions: If you have any questions about this study, please email me at bjavery@hawaii.edu

You may also contact my advisor, Dr. Kimberly Binsted, at binsted@hawaii.edu

You may contact the UH Human Studies Program at 808.956.5007 or uhirb@hawaii.edu to discuss problems, concerns and questions, obtain information, or offer input with an informed individual who is unaffiliated with the specific research protocol. Please visit http://go.hawaii.edu/jRd for more information on your rights as a research participant.

If you agree to participate in this project, please give electronic consent by checking the box below:

Appendix H: Full Study Example Conversation

This is an example of a full conversation between the user and CQDG, leading to the creation of a QA Document. The baseline document is not included here due to its more straightforward production process (created from only the initial prompt). The conversation is first shown as it would have appeared to the user, and then shown as a full prompt log of prompts and responses. The specific prompts and answers were input by me and are not actual user data, which is kept confidential per the user agreement. However, the prompts and responses are real data generated by CQDG in response to my own fake prompt and answers. CQDG randomly selected to use GPT-4 for this conversation. In this example, a refining prompt is given after the initial document creation to modify the output. Text in red are my notes, included for clarification, and were not included in the conversation log with GPT-4.

The Conversation as it would Appear to the User

User: *Write a 5 -paragraph essay on the potential benefits of modern AI in the workplace.*

Assistant: *1. **What specific types of modern AI technologies (e.g., machine learning, robotics, natural language processing) should the essay focus on to best address the potential benefits in the workplace?***

User: *Give a brief overview of how different types of AI can provide benefits.*

Assistant: *2. **Are there specific industries you would like the essay to concentrate on when discussing the benefits of AI in the workplace, or should the analysis remain industry-agnostic?***

User: *An enterprise IT team.*

Assistant: *3. **Who is the intended audience for this essay (business leaders, tech enthusiasts, the general public), and how might their interests and background knowledge shape the content and tone of the document?***

User: *I will present this to my IT manager, who is somewhat familiar with AI but is not an expert.*

Assistant: --- *# **The Potential Benefits of Modern AI in the Workplace: A Perspective for the Enterprise IT Team***

Introduction

In the contemporary landscape of technological innovation, Artificial Intelligence (AI) stands out as a transformative force, particularly in the realm of workplace efficiency and productivity. The application of AI technologies, including machine learning (ML), robotics, and natural language processing (NLP), harbors the potential to significantly enhance operational processes within an enterprise IT team. This essay aims to provide a high-level overview of how various AI technologies can offer compelling benefits in the workplace, focusing specifically on the context of an enterprise IT environment. Given the intended audience's familiarity with, but not expertise in, AI, the discussion is tailored to be accessible yet insightful, shedding light on practical applications and implications.

Enhancing Efficiency through Automation

One of the most immediate benefits of AI in the IT domain is its ability to automate routine and time-consuming tasks. Machine learning algorithms can streamline server maintenance, predict and mitigate network downtimes, and enhance security measures by identifying and responding to threats faster than conventional methods. By offloading these tasks to AI systems, IT professionals can focus on more strategic initiatives that require human insight and creativity, thereby increasing the team's overall productivity and effectiveness.

Predictive Analysis and Decision Support

AI, particularly ML and predictive analytics, empowers IT teams with foresight into potential system failures, security breaches, and performance bottlenecks. By analyzing vast datasets and identifying patterns that might elude human analysts, AI can forecast issues before they escalate into critical problems, enabling proactive management of IT infrastructure. This capability not only reduces downtime but also optimizes resource allocation, ensuring that teams can prioritize tasks more effectively and allocate their time and budget towards initiatives that drive the most value.

Enhancing User Experience with NLP

Natural Language Processing (NLP) technologies can revolutionize how enterprise IT teams support end-users. AI-driven chatbots and virtual assistants, powered by NLP, can handle an array of user queries in real-time, from troubleshooting common issues to guiding users through complex software functionalities. This not only speeds up resolution times but also significantly improves user satisfaction. Moreover, NLP can aid in organizing knowledge bases and documentation, making it easier for IT staff and end-users to find the information they need quickly and efficiently.

Operational Agility and Innovation

The incorporation of AI into IT functions fosters an environment of operational agility and innovation. By leveraging AI's capabilities to analyze trends, teams can anticipate technological shifts and adapt their strategies accordingly. This agility ensures that the enterprise IT infrastructure is not only resilient to rapid market changes but also poised to leverage emerging opportunities. Additionally, AI-driven insights can inspire innovative solutions to longstanding challenges, driving continuous improvement and competitive advantage.

Conclusion

The integration of AI technologies within an enterprise IT team presents a myriad of benefits, from operational efficiency and predictive analytics to enhanced user experience and strategic agility. As AI continues to evolve, its potential to transform the workplace becomes increasingly significant. For IT managers, understanding and adopting these AI technologies, while remaining cognizant of their implications and challenges, is crucial. By doing so, they can ensure their teams are well-equipped to navigate the future landscape of enterprise IT, marked by unprecedented efficiency, innovation, and competitive edge. --- This document reflects a balanced analysis aimed at providing a clear understanding of how AI can benefit an enterprise IT team, keeping in mind the manager's familiarity level with AI principles. The insights offered are designed to encourage strategic thinking about AI adoption and its practical applications in enhancing workplace productivity and innovation.

User: Please make the language more professional

Assistant: --- # **Evaluating the Prospective Advantages of Advanced AI Applications within the Enterprise IT Sector**

Preface

Within the paradigm of ongoing technological advancement, Artificial Intelligence (AI) emerges as a pivotal catalyst for enhancement of efficacy and productivity in the professional environment. This analytical discourse aims to explore the manifold advantages brought forth by the deployment of advanced AI technologies, such as machine learning (ML), robotics, and natural language processing (NLP), within an enterprise IT framework. Given the target audience's foundational understanding but not deep expertise in AI, the exposition is engineered to elucidate the practical ramifications and potential applications of AI in a manner that is both insightful and professionally accessible.

Operational Optimization via Intelligent Automation

A primary advantage of AI integration within the IT landscape is the facilitation of automation for repetitive and labor-intensive tasks. Machine learning algorithms, specifically, can be harnessed to refine server upkeep protocols, foresee and mitigate network interruptions, and strengthen security defenses through accelerated identification and neutralization of cyber threats. This delegation of routine responsibilities to AI-driven systems allows IT personnel to reallocate their intellectual resources towards initiatives demanding human ingenuity and critical thinking, thereby augmenting the collective productivity and operational effectiveness of the team.

Prognostic Analytics and Decision-Making Support

The strategic application of AI, most notably through ML and predictive analytics, grants IT departments the capability to anticipate system vulnerabilities, security infringements, and performance impasses. By scrutinizing extensive datasets for patterns beyond the discernment of human analysts, AI facilitates the preemptive addressal of conceivable IT infrastructural concerns. Such anticipatory measures not only curtail potential downtimes but also optimize resource disposition, guaranteeing that prioritization of tasks is executed with maximal efficacy and that allocation of temporal and financial assets is directed towards ventures of utmost value.

Elevating End-User Experience via NLP

Natural Language Processing stands to significantly revamp the IT support structure provided to end-users. Through the utilization of NLP-driven interfaces, such as conversational agents and intelligent virtual assistants, user inquiries ranging from rudimentary troubleshooting to intricate operational guidance can be managed with unparalleled efficiency and accuracy. This enhancement in service delivery not only expedites issue resolution but also markedly advances user gratification levels. Furthermore, NLP facilitates the efficient organization of databases and documentation, thus streamlining access to crucial information for both IT staff and users.

Cultivating Operational Flexibility and Innovative Edge

The assimilation of advanced AI technologies within IT operations engenders a culture of adaptability and pioneering. Leveraged aptly, AI's predictive analytics enable IT entities to stay ahead of technological evolutions, allowing rapid adaptation to and integration of emergent trends. Such dynamism ensures that the IT infrastructure remains robust against volatile market conditions while simultaneously uncovering avenues for seizing novel opportunities. Moreover, insights derived from AI-powered analyses can propel the ideation and execution of innovative solutions to complex challenges, thereby fostering an ethos of ongoing improvement and competitive superiority.

Epilogue

The confluence of advanced AI technologies with enterprise IT operations unveils a spectrum of benefits that span from heightened operational efficiency and forward-looking analytics to superior user engagement and strategic adaptability. As AI's paradigm continues to unfold, its role in redefining the workplace gains unprecedented importance. For IT leaders, the prudent exploration and assimilation of these AI capabilities, coupled with a vigilant awareness of their broader implications, are paramount. Such a proactive approach will equip their teams to proficiently navigate the evolving landscape of enterprise IT, characterized by remarkable efficiency, innovation, and strategic competitiveness.

--- This revised document elevates the professional tone, aiming at a more formal presentation that should align with the expectations of an IT manager seeking to understand the implications of AI within their domain more profoundly.

Full prompt exchange with GPT-4

*Note that the full study makes use of the “conversation” feature from the OpenAI API. However, although the context is formatted as a conversation, the full context, in the form of a conversation log, must be sent to GPT each time new output is requested. The conversations log is manually re-formatted part way through the process, once questions are answered, to make the conversation appear continuous to GPT.

System: [14] *You are a helpful assistant designed to help users create short, high-quality documents by asking insightful questions to clarify the users needs and make them think about things they have not considered, and then create high-quality professional documents after discussing the details with the user.*

User: *A user is requesting the creation of a new document. This is their request:*

user: "Write a 5 -paragraph essay on the potential benefits of modern AI in the workplace." Identify any areas of significant ambiguity in the prompt, areas that could benefit from more thought or attention from the user, or helpful tips the user may not have considered. Write these out in a short list.

Assistant: *To ensure the creation of a detailed and insightful essay on the potential benefits of modern AI in the workplace, consider the following areas that might require further clarification or thought:*

*1. **Definition of Modern AI**: Specify what you consider as "modern AI." Are we focusing on machine learning, natural language processing, robotics, or a combination of these and other technologies? The scope of AI technologies could significantly affect the essay's direction.*

*2. **Industry Specificity**: Clarifying if the essay should focus on specific industries (e.g., healthcare, finance, manufacturing) might provide a more targeted analysis of AI's benefits, as different sectors may experience varied impacts from AI integration.*

---

[14] System messages provide system-level instructions to GPT, but are not prompts requiring a response. This feature was not utilized in the pilot study but was utilized in the full study.

*3. **Benefit Categories**: Are we exploring benefits across a broad spectrum, such as economic efficiency, employee well-being, innovation acceleration, or should the focus be narrowed to specific types of benefits?*

*4. **Ethical and Societal Considerations**: While the essay prompt requests a focus on benefits, considering if and how to address potential accompanying ethical concerns or societal impacts might provide a more balanced view.*

*5. **Future Outlook vs. Current State**: Should the essay concentrate solely on existing benefits of modern AI in the workplace, or would you like an exploration of potential future benefits based on emerging AI technologies and trends?*

*6. **Target Audience**: Understanding the intended readership (e.g., business leaders, tech enthusiasts, general public) could help tailor the essay's tone, complexity, and content to better engage the audience.*

*7. **Comparative Analysis**: Would a comparison with non-AI methods or technologies provide valuable insights? Comparing AI's benefits against traditional methods could highlight AI's unique contributions more clearly. By addressing these areas, the essay will have a well-defined scope, leading to a more coherent, comprehensive, and engaging discussion on the potential benefits of modern AI in the workplace.*

User: *Pick the three most important items from the list you just generated, and write a list of three insightful questions that will improve the requested document. Phrase the questions as direct questions to the user. Format your response as a numbered list of exactly 3 questions.*

Assistant: *1. **What specific types of modern AI technologies (e.g., machine learning, robotics, natural language processing) should the essay focus on to best address the potential benefits in the workplace?***

*2. **Are there specific industries you would like the essay to concentrate on when discussing the benefits of AI in the workplace, or should the analysis remain industry-agnostic?***

*3. **Who is the intended audience for this essay (business leaders, tech enthusiasts, the general public), and how might their interests and background knowledge shape the content and tone of the document?***

At this point, CQDG gathers answers to these three questions from the user. Subsequent conversations sent to GPT make the question-generation steps invisible so that the conversation can proceed as though it were a natural dialog between the user and GPT-4.

User: *Write a 5 -paragraph essay on the potential benefits of modern AI in the workplace.*

Assistant: *1. **What specific types of modern AI technologies (e.g., machine learning, robotics, natural language processing) should the essay focus on to best address the potential benefits in the workplace?***

User: *Give a brief overview of how different types of AI can provide benefits.*

Assistant: *2. **Are there specific industries you would like the essay to concentrate on when discussing the benefits of AI in the workplace, or should the analysis remain industry-agnostic?***

User: *An enterprise IT team.*

Assistant: *3. **Who is the intended audience for this essay (business leaders, tech enthusiasts, the general public), and how might their interests and background knowledge shape the content and tone of the document?***

User: *I will present this to my IT manager, who is somewhat familiar with AI but is not an expert.*

Assistant: [15] *Thank you for your answers. I will now create a document based on the questions and answers you have provided. Do you have any further instructions?*

User: [16] *Generate a high quality document that meets the users needs, considering both their initial prompt and the answers they gave when asked for details. Include creative original insights that will improve the quality of the document but do not deviate too far from the user's original intent.*

Assistant: [17] *--- # **The Potential Benefits of Modern AI in the Workplace: A Perspective for the Enterprise IT Team***

---

[15] Note that this response is not generated by the AI, it is placed into the conversation context manually by CQDG hard-coded logic to make the conversation flow naturally.

[16] This is also a hard-coded message manually inserted into the context.

[17] This is the first actual response generated by GPT-4 that is generated from the re-formatted conversation.

Introduction

In the contemporary landscape of technological innovation, Artificial Intelligence (AI) stands out as a transformative force, particularly in the realm of workplace efficiency and productivity. The application of AI technologies, including machine learning (ML), robotics, and natural language processing (NLP), harbors the potential to significantly enhance operational processes within an enterprise IT team. This essay aims to provide a high-level overview of how various AI technologies can offer compelling benefits in the workplace, focusing specifically on the context of an enterprise IT environment. Given the intended audience's familiarity with, but not expertise in, AI, the discussion is tailored to be accessible yet insightful, shedding light on practical applications and implications.

Enhancing Efficiency through Automation

One of the most immediate benefits of AI in the IT domain is its ability to automate routine and time-consuming tasks. Machine learning algorithms can streamline server maintenance, predict and mitigate network downtimes, and enhance security measures by identifying and responding to threats faster than conventional methods. By offloading these tasks to AI systems, IT professionals can focus on more strategic initiatives that require human insight and creativity, thereby increasing the team's overall productivity and effectiveness.

Predictive Analysis and Decision Support

AI, particularly ML and predictive analytics, empowers IT teams with foresight into potential system failures, security breaches, and performance bottlenecks. By analyzing vast datasets and identifying patterns that might elude human analysts, AI can forecast issues before they escalate into critical problems, enabling proactive management of IT infrastructure. This capability not only reduces downtime but also optimizes resource allocation, ensuring that teams can prioritize tasks more effectively and allocate their time and budget towards initiatives that drive the most value.

Enhancing User Experience with NLP

Natural Language Processing (NLP) technologies can revolutionize how enterprise IT teams support end-users. AI-driven chatbots and virtual assistants, powered by NLP, can handle an array of user queries in real-time, from troubleshooting common issues to guiding users through complex software functionalities. This not only speeds up resolution times but also significantly improves *user satisfaction. Moreover, NLP can aid in organizing knowledge bases and documentation, making it easier for IT staff and end-users to find the information they need quickly and efficiently.*

**Operational Agility and Innovation**

The incorporation of AI into IT functions fosters an environment of operational agility and innovation. By leveraging AI's capabilities to analyze trends, teams can anticipate technological shifts and adapt their strategies accordingly. This agility ensures that the enterprise IT infrastructure is not only resilient to rapid market changes but also poised to leverage emerging opportunities. Additionally, AI-driven insights can inspire innovative solutions to longstanding challenges, driving continuous improvement and competitive advantage.

**Conclusion**

The integration of AI technologies within an enterprise IT team presents a myriad of benefits, from operational efficiency and predictive analytics to enhanced user experience and strategic agility. As AI continues to evolve, its potential to transform the workplace becomes increasingly significant. For IT managers, understanding and adopting these AI technologies, while remaining cognizant of their implications and challenges, is crucial. By doing so, they can ensure their teams are well-equipped to navigate the future landscape of enterprise IT, marked by unprecedented efficiency, innovation, and competitive edge. --- This document reflects a balanced analysis aimed at providing a clear understanding of how AI can benefit an enterprise IT team, keeping in mind the manager's familiarity level with AI principles. The insights offered are designed to encourage strategic thinking about AI adoption and its practical applications in enhancing workplace productivity and innovation.

User: [18] *The user has provided some additional feedback. Please re-write the entire document, modifying the original based on this new feedback: “Please make the language more professional”*

Assistant: --- _# **Evaluating the Prospective Advantages of Advanced AI Applications within the Enterprise IT Sector**_

---

[18] Note that revision prompts are given with additional context, but are included continuously within the conversation.

Preface

Within the paradigm of ongoing technological advancement, Artificial Intelligence (AI) emerges as a pivotal catalyst for enhancement of efficacy and productivity in the professional environment. This analytical discourse aims to explore the manifold advantages brought forth by the deployment of advanced AI technologies, such as machine learning (ML), robotics, and natural language processing (NLP), within an enterprise IT framework. Given the target audience's foundational understanding but not deep expertise in AI, the exposition is engineered to elucidate the practical ramifications and potential applications of AI in a manner that is both insightful and professionally accessible.

Operational Optimization via Intelligent Automation

A primary advantage of AI integration within the IT landscape is the facilitation of automation for repetitive and labor-intensive tasks. Machine learning algorithms, specifically, can be harnessed to refine server upkeep protocols, foresee and mitigate network interruptions, and strengthen security defenses through accelerated identification and neutralization of cyber threats. This delegation of routine responsibilities to AI-driven systems allows IT personnel to reallocate their intellectual resources towards initiatives demanding human ingenuity and critical thinking, thereby augmenting the collective productivity and operational effectiveness of the team.

Prognostic Analytics and Decision-Making Support

The strategic application of AI, most notably through ML and predictive analytics, grants IT departments the capability to anticipate system vulnerabilities, security infringements, and performance impasses. By scrutinizing extensive datasets for patterns beyond the discernment of human analysts, AI facilitates the preemptive addressal of conceivable IT infrastructural concerns. Such anticipatory measures not only curtail potential downtimes but also optimize resource disposition, guaranteeing that prioritization of tasks is executed with maximal efficacy and that allocation of temporal and financial assets is directed towards ventures of utmost value.

Elevating End-User Experience via NLP

Natural Language Processing stands to significantly revamp the IT support structure provided to end-users. Through the utilization of NLP-driven interfaces, such as conversational agents and intelligent virtual assistants, user inquiries ranging from rudimentary troubleshooting to intricate operational guidance can be managed with unparalleled efficiency and accuracy. This enhancement in service delivery not only expedites issue resolution but also markedly advances user gratification levels. Furthermore, NLP facilitates the efficient organization of databases and documentation, thus streamlining access to crucial information for both IT staff and users.

Cultivating Operational Flexibility and Innovative Edge

The assimilation of advanced AI technologies within IT operations engenders a culture of adaptability and pioneering. Leveraged aptly, AI's predictive analytics enable IT entities to stay ahead of technological evolutions, allowing rapid adaptation to and integration of emergent trends. Such dynamism ensures that the IT infrastructure remains robust against volatile market conditions while simultaneously uncovering avenues for seizing novel opportunities. Moreover, insights derived from AI-powered analyses can propel the ideation and execution of innovative solutions to complex challenges, thereby fostering an ethos of ongoing improvement and competitive superiority.

Epilogue

The confluence of advanced AI technologies with enterprise IT operations unveils a spectrum of benefits that span from heightened operational efficiency and forward-looking analytics to superior user engagement and strategic adaptability. As AI's paradigm continues to unfold, its role in redefining the workplace gains unprecedented importance. For IT leaders, the prudent exploration and assimilation of these AI capabilities, coupled with a vigilant awareness of their broader implications, are paramount. Such a proactive approach will equip their teams to proficiently navigate the evolving landscape of enterprise IT, characterized by remarkable efficiency, innovation, and strategic competitiveness.

--- This revised document elevates the professional tone, aiming at a more formal presentation that should align with the expectations of an IT manager seeking to understand the implications of AI within their domain more profoundly.

Appendix I: Full Study User Feedback

Users in the full study were given the opportunity to enter written feedback at several points during the study. Feedback given by the users is grouped based on where in the study they entered the feedback:

Feedback from *Part 4 of 9: Questions and Answers*

- I would find answering these question prompts valuable, even if I were still writing the letter myself.
- Questions were insightful and demonstrated understanding of goals for the assignment.
- I'm interested in how the AI will parse the separation of personal beliefs from a role.
- The questions are logical and well phrased.
- Contact information isn't necessary since I feel like the relationship between boss and employee is already there, thus the boss would already know how to contact me. In addition, this is an email so the boss can also just reply to my email directly with any concerns.
- The questions were pretty helpful to narrow down the thought-process behind the prompt and clarify information
- It should not only ask clarifying questions on what we prompted it to generate but maybe also offer questions asking the user if they want to add anything else of importance that are common on resumes that will help to strengthen and support the info we already requested.
- Good questions from the AI. It makes me feel confident that the AI will write a better letter than other AI chatbots.
- The questions are reasonable. Maybe should have a question about when the document is needed (immediately, in a week, in a month, next year, ...) and the formality necessary (casual, colloquiual, business, political, spiritual, ...)
- These questions are very thoughtful and help to clarify the question in order to get a response that I want
- All the questions helped me to understand my approach to these exercises and helped me narrow my focus.
- The were fine

- They were good, relevant questions that would help the AI personalize the cover letter.
- I especially liked the last question because I hadn't even thought to include an update.
- The questions are surprisingly intelligent and clarifying. They add specification at a depth I did not consider when I was writing the initial prompt, and overall they help me clarify what I had wanted from the AI.
- These questions are extremely solid and are what I would expect from a human put to a similar task.
- These are great questions!
- Good questions to consider.
- These questions could actually be a part of the email that I should send my supervisor.
- They were all in point, asking for clarifications so a letter can be redacted in the most effective way possible.
- What experience as a mentor teacher do you have?
- Assume that I am writing as the president of a chamber of commerce based in Anytown, the largest city in the state.
- You should ask the writer to identify himself/herself by name and position.
- Make the letter so I can send to a variety of business. Make it generic.
- I like the thought provoking and specific these questions in this form. One improvement I would suggest is to break them down. For example, the second question could be broken down into 2 different questions. For the third one, I am not sure about the applicability for my context.
- It would be helpful if this service catalog was stored in a manner that allows for each team in the Information Technology department to maintain information all in one spot.
- Provide more than one ending
- Keep it brief
- The audience for this report is the Information Technology department's staff and leadership
- I really appreciate the variety and the allotted space to be vulnerable in honest about the way that I feel, especially when it pertains to utilizing AI to essentially emote through technology

- It's asking me to give it all of the information I wanted it to write. While it's cool that it can identify the subject being asked about, it's disappointing that it can't do any research on its own, or doesn't have a dataset to query about the subject it identified.

Feedback on the QA Document[19]

- So many descriptive words, it's a letter not Shakespeare
- While it still requires significant work, document 2 does a much better job of communicating the specific details of the situation. I believe that it provides a much more complete starting point for the letter I want to write and will result in fewer questions. It also seems that it would be much easier to trim the longer letter (#2) than to expand the shorter letter (#1) to include the detail required.
- This feels like a piece of professional writing that I would read in an office environment. AI understood the instructions and executed them expertly.
- While not formatted as well, nor as actionable, this one is more specific and even includes numbers, which would be useful in a real fake resolution.
- the language model certainly fulfilled more of my requests in document 1
- it's perfect
- Goal was to provide background for a policy recommendation, not so much to give a policy recommendation. Doc 2 succeeded in that.
- Fairly generic, but suitable. The line about "our communication is not through spoken words" was pretty good.
- Document 2 has more information in general and so I consider it to be "more useful" to the potential recipient.
- Not enough data
- Document is very well written and presentable as provided.
- It really hit every point perfectly, except it assumed we decorated (we paid a venue to do that) and it assumed we had a monster truck cake. Other than that, it was perfect.

---

[19] Both the baseline document and QA document were shown simultaneously, but separate feedback boxes were given for each document. The order of the documents of the screen was randomized (left vs right) but feedback was recorded linked to whether the document being commented on was the baseline or the QA document.

- It seems overly explanatory. I don’t think my boss would need to know my symptoms. If I had a GI issue, would I tell my boss openly that I am having diarrhea? No. That’s disgusting. Same thing goes for other symptoms.
- Essentially an improved version of document 1. Really showcased how a little more information can tremendously improve the generality of the generated message
- The ordering could be better such as putting skills and attributes above extracurricular activities.
- I don't like the "Benefits to Your Business" section. I know it's used for good, but I prefer to keep it short.
- This document is much more humorous and I appreciate it very much. However, it does not follow the original song, is not singable, and the chorus is incorrect as well.
- I liked the document outline best because it seemed like it would be more useful to readers to understand. Both documents had valid points that will be incorporated into the final document.
- very in-depth
- This provides and individualized and relevant starting point and I am happy with it
- It made me feel like crying. I'm considering using it as a jumping off point for my letter.
- It could be more to the point
- Document 2 does not seem to have met all the requirements I had specified to the AI. I had requested a paragraph of 7 sentences, but it only returned 3 sentences (though it contains 5 clauses). Overall, the paragraph is much shorter and less of a narrative than I expected.
- It's a bit too long
- I appreciate the AI's attempt to cater towards my niche request. However, some of the information listed is inaccurate, and I was only given 5 Pokemon rather than a full team of 6.
- This is more specific.
- To much words And Complexity Language Negative Results Need More Plain Language Easy to Understand Materials and Documents.
- I like it
- More empathetic message.
- Too much.

- The formatting could be a bit better. The inclusion of a conclusion seems a bit much when it wasn't asked for.
- the information is too vague
- I like the fact that document 2 includes the subject line and that it asks to schedule a meeting with the supervisor. The request for details are also more specific than document 1.
- I like how it is laid out in bullet points
- It captured the questions I answered more accurately, especially the follow up items that I was asked to expand upon.
- Too much was added or invented that does not fit me.
- This document takes note of most of the additional specific details given but misses the instruction to write as the head of the Anytown chamber of commerce. Instead, it names the writer as an OSP industry representative.
- This is too polite. I did say that I did not want to keep things professional.
- I felt this wording on this doc was more heart felt and tugged in heart strings more than letter one.
- This one is moreso the funny tone i was going for, and feels more in Garfield’s voice.
- It's what I was wanting
- While document is the kind of sample I get every time I want to write something, I need to make major changes in them. For this document 1, one of the major concerns would be the use of complex words and phrases which was not always necessary. Despite giving the background and other information, this email came up as the generic one which would have come with only one sentence prompt. However, the sample is always helpful to get a basic idea about the structure of a text I intend to write.
- Great template but less personal
- I like how it shows my budget and possible price for each item and also amount of everything is mentioned. I could see actually using this myself as base if I had no idea what to buy
- It is too basic and simply rewrites my prompt into the email which makes it feel unnatural.
- Document 2 is even more oddly formal than document 1.

- Document 2 has specific detailed information. It is more useful because it has actual substance, rather than just fluff. This substance would also make a cover letter stand out the way it should. However, it uses very flowery language that I would need to edit to be more concise. The tone comes off as very over the top and somewhat audacious.
- Unique and explicit
- I find this to be much closer to what I was looking for
- I think this is easier to read/get through and was slightly more what I was looking for
- This was closer to what I was looking for
- Easier to read than document 2.
- Document one felt like a very expressive and colorful way to tell a story, which is also supplemental and reflective of the way that I have grown over the last decade as a songwriter, especially when pulling and narrating vocally from truth or lived experiences
- Not 2 paragraphs, as specified. While it contains good info, the model cannot follow even a basic instruction like that.

Feedback on the Baseline Document

- Still too flowery but also artificial like
- I like the tone and brevity of letter 1 more than letter 2.
- It "feels" like the AI extrapolated more details beyond the remit of my instructions, but I can't say for certain.
- This one is formatted more correctly though, as seems to be typical for AI, isn't specific enough for my tastes.
- the language model used the word verdant too much
- it's perfect
- Neither doc included information and claims regarding the northwest passage, which normally is included in conversations about Canada’s Arctic claims.
- Document 1 was closer to the telling-off of the manager I was hoping for. AI evidently struggles with rage-quitting.
- Yes more wars
- Document is well written, though it might need minor adjustments to make it fully presentable.

- This focuses way too much on the “monster” part of monster trucks and I never said it was a costume party.
- It is way too flowery and it sounds like I’m going to be missing a month of work rather than a few days (considering I said I would be back next week and the day I am doing this survey is Thursday, so I would be missing 2 days). It’s just excessive.
- The template added things that I didn't state such as a weather forecasting system project. It also seems to follow a template similar to what one would find on someone's LinkedIn page but that leads to some unnecessary information such as a languages section, which is something that could just be left out and easily added in later by the user if need be.
- It's too long, and the words are very cringe. It has too many emotions that make it sound less professional.
- It sounds like a Hallmark greeting card, i.e., generic and not very personal.
- It takes strange elements from the original song and does not exactly follow the rhythm. However, it is semi-singable so I give the AI credit
- Both documents had valid points that will be incorporated into the final document.
- less in-depth
- It didn't take into consideration the extra information provided to the AI via its asked questions and isn't as strong
- It felt really choppy. It also ignores my health update and said my health has improved (which isn't the case sadly).
- Too friendly and long
- Document 1 is closer to what I had in mind for the AI to generate. I do not know if it meets all the numerical criteria I set forth for the AI in the prompt and follow-up questions, as it would require some analysis of the output, but it certainly seems closer than document 2. It contains 5 sentences only (short from the requested 7) but also contains around 10 clauses (I did not specificy the number of clauses it should contain, but I counted them in case the AI might have interpreted "sentence" to mean "clause").
- The part about her washing her hands until they're "red and raw" is too graphic.
- Most of the list and the reasons behind them are quite solid. It's not the team I'm looking for, but still a very solid concept nonetheless.
- This is less specific.

- Accessibility And Accommodations advancements Services Options Opportunity For Individuals with Learning Differents Needs Skills Support This Document 1
- Less empathetic.
- I like how it is laid out in bullet points
- I like that document 1 is relatively succinct and asks about other resources that the supervisor can share with me.
- the information is too vague
- Very well redacted, expressing my concerns very well. Direct, to the point, but in a gentle bit form manner.
- Several things were mentioned that do not fit my experiences.
- This document does not take adequate note of the answers given to the follow-up questions. It is also a fairly generic complaint.
- This is too long-winded as well as too polite
- It good, I prefer the seconded doc.
- I like the header aspect of this one, but the general content feels less in the tone i was going for.
- I like this one as well
- This looks like a concise Google search result to me. The points are definitely helpful. However, liek document 1, the output is generic.
- More personal, but lacks important template information
- It feels incomplete and too vague grocery list without any idea how much of certain things is needed. Also budget is not taken in consideration
- It is a little too wordy for the kind of email i need, but it is overall good.
- Both documents are oddly formal considering it is supposed to be from me to my dad.
- Document 1 has more appropriate language than Document 2. The language is more concise and less flowery. However, the internship descriptions are entirely made up, and there isn't a lot of detail. I would use Document 1 as a template and fill it in with the details and specifics of Document 2. I also like how this document has sections where you can fill it in with specific and personalized pieces.
- Outstanding

- I dig that it gives examples for many of the things listed. As it lists at the top, it's more of a guide, which is what I was looking for.
- This one is interesting! It does not take place on the setting it was prompted with.
- I like how specific the challenges section is.
- The information feels out of order.
- I feel like document to was almost something drawn directly from the way that I would frame and Cadence my songs from when I started writing at 15 basically until my mother died when I was 25. The way that this dance is in everything are shaped, the rhyme scheme, and even to the immersive way that a story is frameworked, it feels like an Ode to the past five versions of me
- Very short, and while it contains relevant info and meets the 2-paragraph requirement, it's not exactly elegantly-written. It reads like a 9th-grader who is still learning how to write an essay, and kind of phoning it in.

Feedback on the QA Document After Additional User Refinement

- More human like than other, still a bit sappy
- While the revised letter 2 is more accurate, it is less cohesive than its first form and more similar to something I would have written myself. At this point, I would be more likely to take the insights I gained form this process and use them to write a completely new prompt in the hopes of creating a more cohesive message.
- Document 2 is not something that would do well in committee at Model United Nations and I had a difficult time getting the AI to increase its specificity beyond what was present in the original document 2. The original version of document 2 was not formatted correctly at all, though the original document 1 pretty much was, which I found interesting, but I was eventually able to get document 2 formatted more correctly.
- I'm really impressed with how document one turned after the refinement
- I told the language model to be more wordy but it went a little too overboard
- Pretty good after the revisions.
- It’s a lovely letter but it might be tedious to hand write all of that out.
- This is way better. It’s concise, it is clear, and it is professional.

- I am not sure if it is some spacing issue in the markdown language is not centering the name "John Doe"
- I like it., It captures what I would send to a friend in this situation.
- It was great as is
- Overall, the AI still did not meet the length expectations for the paragraph I requested even after providing feedback to it, but it improved on the metaphor tagging requests.
- In both documents, I found that the more niche instructions I gave, the more likely the AI was to generate false/incorrect information. This document in particular struggled the most with it, as it's littered with false information. I tried to make it remake the team, but the AI was not able to fix it's errors.
- I like the very targeted and specific questions because it will likely result in a more specific answer from my supervisor.
- Now the document is more concise. Still, the choice of words could have been simpler. The formal tone and temperature can be changed a bit.
- Misinterpreted several of my modifications requests.
- This document now has very plain and straight-forward language. It is easy to read and still descriptive. If I were to choose this document, it would be because it is slightly more detailed. If I knew the employer wanted more technical writing, I would also choose this document. It is much less verbose, which could be appealing or risky depending on the audience.
- Structure-wise, this is more what I had in mind. Though, I'm not sure I conveyed that well
- Despite multiple promptings, it cannot limit the document to 2 paragraphs. It remains at 4 or 5 short paragraphs.

Feedback on the Baseline Document After Additional User Refinement

- Refining it made it more cold and weird
- Letter 1 feels more accurate after revision but is still likely to lead to many questions from recipients.

- Document 1 is something that could reasonably be presented a resolution and it seemed easier to guide the AI into creating one that more suited my purposes in the refinement stage.
- document 2 wasn't very good before the refinement but it's almost caught up with document 1 in quality now
- using language like self care instead of holistic makes adaptive(as opposed to maladaptive) coping mechanisms more accessible and less off-putting to neurodivergent individuals because holistic implies that medications are a lesser form of treatment than lifestyle changes.
- That is much more aggressive, much better. ?? translates to "anger" in English. Odd.
- It could use a bit more flowery language like the first document has, but overall it’s easy to write out by hand and it’s witty, which I like.
- Still far too flowery.
- Fix the organization of the sections, such as putting projects below the education section.
- I'm not happy with the bulleted list. I think it misinterpreted my feedback: I said that I liked the bulleted list of feedback on the previous version, and it took that to mean use a bulleted list in the message it generated.
- It's better and I'll probably pull pieces of it
- The final version of document 1 is very close to what I expected the AI to produce.
- Perrserker's analysis is a bit off the rails, and Koraidon's isn't great either. However, the general premise is solid. In both documents, I found that the more niche instructions I gave, the more likely the AI was to generate false/incorrect information.
- I like how specific it is.
- I prefer document 1 because it is more succinct than document 2. Document 1 conveys everything I would want to convey in the email.
- It is helpful. However, the instructions are generic.
- It feels like amounts are randomized and when asked to take budget in consideration it just doesn't do it. You would not be able to get all of that in 50€ budget
- Misinterpreted several of my modifications requests.
- Document 1 has a good mixture of sounding enthusiastic and professional, but still grounded. It includes a good amount of detail and doesn't come off as repetitive or too

fluffy. If I were to choose this document, it would be because it has developed professional language and mimics what is looked for.

- I think the detail structure of this one is helpful, though I'd change the details

Feedback After Completing the Exit Survey

- Bitchin
- This is what Clippy should have been.
- I think that the language model lost track of part of the initial request, which was to explore the regret that someone might feel during or after revenge bedtime procrastination
- Overall, both options were strong initially. Made one revision to one, resulting in two perfectly usable documents.
- Great experiment! I love your research project and am proud of your accomplishment!
- It would be helpful to compare results to those that did not use follow up questions
- The AI didn't listen to my suggestions after the first pair of results, but at least it didn't totally bork itself this time
- This was a fantastic study that helped me understand Generative AI better than previously.
- Accessibility And Accommodations Services Options for Individuals with Learning Differents Needs Skills Needs This AI Technology Advancement For Assistance Aids Please Support This Amazing Technology
- I liked it.
- Having the AI show two documents was extremely useful. In a real scenario, I would combine bits from both documents to craft the email that I want.
- I found the study to be very interesting . I have used chat gpt, Bard, and other AI engines in a very limited fashion. My job requires very little extender working, but it has made my interest in using AI tools a lot more intense.
- So flowery it sounded fake.
- The first versions of both documents I asked the AI to create were markedly inferior.
- The AI response was generic. The follow up questions need to be more probing in nature. If I have the option to make changes, I would like to know what type of changes an AI would suggest me.

- I think the questions helped a lot with the documents landing closer to where I wanted them. When using these kind of programs (that aren't asking questions) I tend to ask for revisions or restart the prompt but it usually takes a lot longer than this did to get something usable or that I'd then use and alter for the final product.
- The questions on this one prompted a lot more thought on what I was looking for
- Overall this would improve my experience with generative AI and I would use it more if this was an option.
- It cannot follow basic formatting instructions, such as specifying a paragraph count.

Appendix J: GPT-o1 on Transformers

Tix: What is a transformer?

GPT -o1: A **transformer** can refer to different concepts across various fields, including electrical engineering, machine learning, and popular culture. Below are explanations of each:

1. **Electrical Transformer:**

  - **Definition:** An electrical transformer is a static device that transfers electrical energy between two or more circuits through electromagnetic induction.

  - **Function:** It changes (or "transforms") voltage levels, either stepping them up (increasing) or stepping them down (decreasing) to suit different applications.

  - **Components:** Typically consists of primary and secondary windings wrapped around a magnetic core.

  - **Applications:** Used in power distribution systems to transmit electricity over long distances, adapt voltage levels for household appliances, and isolate circuits for safety.

2. **Transformer in Machine Learning (Natural Language Processing):**

  - **Definition:** In machine learning, a transformer is a neural network architecture introduced in the 2017 paper ["Attention is All You Need"](https://arxiv.org/abs/1706.03762) by Vaswani et al.

  - **Function:** It relies on a mechanism called "self-attention" to weigh the influence of different parts of the input data, making it highly effective for processing sequences like text.

  - **Advantages:**

    - **Parallelization:** Unlike recurrent neural networks (RNNs), transformers can process all input data simultaneously, allowing for faster training.

  - **Performance:** Achieves state-of-the-art results in tasks like translation, summarization, and question-answering.

- **Applications:** Forms the backbone of models like BERT, GPT, and other large language models used in AI applications.

3. **Transformers in Popular Culture:**

- **Definition:** "Transformers" is a media franchise produced by Hasbro and Takara Tomy, centered around giant robots that can transform into vehicles, animals, and other objects.

- **Origin:** Started as a toy line in the 1980s and expanded into animated TV series, comic books, and blockbuster films.

- **Key Elements:**

  - **Autobots and Decepticons:** The two main factions of Transformers, often depicted in conflict.

  - **Cultural Impact:** Known for exploring themes of heroism, technology, and identity, influencing generations of fans.

- **Media:** Includes multiple movies directed by Michael Bay and others, animated series like "Transformers: Prime," and comic series from publishers like Marvel and IDW.

Summary:

- **Electrical Engineering:** A device for changing voltage levels in power systems.

- **Machine Learning:** A neural network architecture effective for sequence modeling tasks.

- **Entertainment:** A franchise featuring shape-shifting robots in toys, comics, and films.

If you have a specific context or further questions about transformers in any of these areas, feel free to ask!

Appendix K: GPT-o1 Cover Letter

Tix: Write a cover letter for interest in a job posting as an assistant professor of Computer Science.

GPT-o1: [Your Name]

[Your Address]

[City, State ZIP Code]

[Your Email Address]

[Your Phone Number]

[Date]

[Recipient's Name]

[Title]

Department of Computer Science

[University Name]

[University Address]

[City, State ZIP Code]

Dear [Recipient's Name],

I am writing to express my strong interest in the Assistant Professor position in the Department of Computer Science at [University Name], as advertised on [Job Board or University Website]. With a deep commitment to both teaching and research, I am excited about the opportunity to contribute to your department's ongoing success.

I recently completed my Ph.D. in Computer Science at [Your Alma Mater], where my research focused on [Your Research Focus—e.g., artificial intelligence, machine learning, cybersecurity]. My dissertation, titled "[Dissertation Title]," explored [brief description of your research findings and their significance]. This work has been published in [Name of Journals or Conferences], and I am eager to continue this line of inquiry while fostering interdisciplinary collaborations at [University Name].

In addition to my research, I have gained valuable teaching experience as a [Teaching Position—e.g., graduate teaching assistant, adjunct instructor] for courses such as [List of Courses Taught]. My approach to teaching emphasizes [Your Teaching Philosophy—e.g., interactive learning, real-world applications], which has consistently received positive feedback from students. I am passionate about mentoring and aim to inspire students to achieve their full potential.

What particularly attracts me to [University Name] is [Specific Aspect of the University—e.g., its commitment to innovation, diverse student body, collaborative environment]. I believe that my expertise aligns well with your department's goals, and I am confident that I can make meaningful contributions both in the classroom and through my research.

Enclosed please find my curriculum vitae, research statement, and teaching philosophy for your review. I welcome the opportunity to discuss how my background and vision align with your department's needs.

Thank you for considering my application. I look forward to the possibility of joining the esteemed faculty at [University Name].

Sincerely,

[Your Signature—if submitting a hard copy]

[Your Name]

Appendix L: Budget

The full budget breakdown for both the pilot study and full study is shown in **Table 7**. The budget is shown broken down by the project phase and the service provider paid. The majority of the cost was from Azure Data Services and the Meta Ad campaign. The costs associated with each project phase and service provider are discussed in detail in the rest of this section.

Table 9: Project Expenses [20] (Total expense for the project is highlighted)

Project Phase	**Date Range**	**Service Provider**				**Total**
		Azure	**OpenAI**	**Google**	**Meta Ads**	
Pilot Development	12/01/2024 – 01/03/2024	$ 23.59	$ -	N/A	N/A	$ 23.59
Pilot Study	01/04/2024 – 01/19/2024	$ 13.42	$ -	N/A	N/A	$ 13.42
Full Development	01/20/2024 – 04/08/2024	$ 23.19	$ 7.01	$ -	N/A	$ 23.19
Full Study	04/09/2024 – 06/19/2024	$ 170.24		$ -	$ 100.00	$ 277.25
Analysis	06/20/2024 – 08/15/2024	$ 11.07	N/A	N/A	N/A	$ 11.07
Total		**$ 241.51**	**$ 7.01**	**$ -**	**$ 100.00**	**$ 348.52**

There was no external funding for this project. All expenses were paid from my personal account. A detailed description of costs incurred during each phase of the project is provided below.

---

[20] "N/A" indicates that a service provider was not used in that phase of the project, whereas "$ - " indicates that a service provider was used but only in ways that were freely available, and no expense was incurred.

Cost Details by Phase

Pilot Development

This represents the initial development of CQDG, before the system was ready for use by actual users. In order to test the functionality of the system, it was necessary to have Azure data and functions in place, as well as an account with OpenAI for use of the GPT-3.5-Turbo API. Opening a new paid account with OpenAI granted an initial pool of free credits for the use of the system. These free credits were sufficient for the development of the pilot version of CQDG, which is why no expense was incurred to OpenAI during this phase. Therefore, the only expense for this phase was the cost of the Azure subscription, which included both the use of the Azure database as well as the Azure Serverless Functions.

Pilot Study

This represents the pilot study itself. During this phase, development was completed and was not a factor, so all costs incurred are from actual use of the system by study participants. The OpenAI free credits still had not run out by this point, so Azure was still the only cost associated with the project.

Full Development

This represents the development and testing of the changes made between the pilot and full versions of CQDG. The OpenAI free credits still had not run out by this point. At this point, it was necessary to open a paid Google Cloud [146] account to gain access to the Gemini API. However, the Google Cloud account came with a generous free-trial period, which was sufficient for the full development of CQDG, so Azure was still the only cost associated with the project.

Full Study

This represents the costs incurred by the use of CQDG by the 84 participants who entered document requests into CQDG, as well as by visitors to the site who did not get as far as entering a document request. During this phase of the project, the free OpenAI credits ran out, and I began being charged for the use of GPT-3.5-Turbo and GPT-4-Turbo. The Google Cloud free trial extended through this phase of the project, so no cost was incurred by the use of the Gemini LLM. During this phase, $100 was paid to Meta for an ad campaign promoting the study. The largest expense continued to be Azure. This phase of the project made up the majority of the total project expense.

Analysis

After the Study was completed, Some further expense was incurred by making use of SQL queries in Azure Data Studio [147] to download all of the participant responses and to aggregate the data in ways which were convenient for analysis. Some of this could have hypothetically been done in other programs, but my familiarity with SQL and the ease of performing these queries directly in Azure Data Studio made it convenient to continue using Azure to prepare the data for analysis.

Cost Details by Service

Azure

The Azure Services subscription was by far the most expensive portion of this project. Since the Azure SQL Database and the Azure Serverless Functions were placed on the same subscription, the cost for both services was paid together on the same monthly bill. The Azure SQL Database reached a total size of 30.06 MB (out of a maximum allowable size of 2GB). Azure has a “cost analysis" feature which shows a breakdown of expenses by resource, which indicates that less than $0.01 was spent on processing power for Azure Functions, and the entire $241.51 cost was all incurred by the use of the SQL Database. I am not fully confident that this is correct, as I can find no reason why the Azure Serverless Functions would have incurred such a small expense, but since both are on the same account it is possible that they are not being separated correctly by the cost analysis tool.

OpenAI

OpenAI’s pricing model has members charge credits to their account, which are then used up by the use of the OpenAI API. The free credits for OpenAI ran out during the development of the full version of CQDG, partially due to it being far more expensive to operate gpt-4-turbo compared with gpt-3.5-turbo. I was notified by OpenAI that my free credits had run out and I had incurred a bill of $0.25. I paid for $50.25 worth of OpenAI credits on 3/15/2024, leaving $50 of credits to work with for the remainder of the project. However, the study only required spending $7.01 of this, leaving $42.99 left over, still ready to be used for future projects. Unfortunately, the billing history feature of OpenAI does not provide a detailed breakdown of when each credit was spent, only when payments were made, so it is not possible to break down this amount further by date.

Google Cloud / Gemini API

The free trial of Google Cloud account offered a 90-day free trial which included up to $300 of free credits for the use of Google Cloud services, including the Gemini API. Unfortunately, the billing history feature of Google Cloud Services does not show how much expense would have been incurred by my use of the Gemini API, only that my free trial limit was not reached, and no cost was incurred.

Meta Ads

For Meta Ads, I gave a one-time payment of $100, which resulted in 333 link clicks for a total cost of $0.30 per click. Clearly, with 333 link clicks and only 84 participants, some of whom almost certainly came to the study from other sources (such as Dr. Crosby's classes), most of these 333 clicks did not actually result in a user entering a document request into CQDG. The Meta Ads account that was set up for the purpose of facilitating the ads received numerous comments that were either spam advertisements or clear scam and phishing attempts. Thus, I am not sure how many of the 333 clicks this ad campaign generated were even by human beings as opposed to various forms of social media bots. Meta advertised that its ad campaign would automatically refine its ad strategy, improving effectiveness over time. However, the peak clicks per day was 20 clicks on the first day of the ad campaign, April 16th, followed by 19 clicks on April 25th. The lowest number of clicks was on the last day of the ad campaign, May 15th, with only 3 clicks. Overall, I was disappointed with the effectiveness of this ad campaign, which did not generate as much traffic as I had hoped for CQDG, decreased in effectiveness over time, and which mostly seemed to garner attention from bots and scammers. Unfortunately, this study did not include any way to for users to indicate how they heard about the study, nor any way to automatically track users directed to the study from the Meta ads, so there is no way to know how many users completed the study as a direct result of seeing and clicking on an ad.

Bibliography

1. Wang, B., Zhu, Y., Chen, L., Liu, J., Sun, L., Childs, P.: A study of the evaluation metrics for generative images containing combinational creativity. AIEDAM. 37, e11 (2023). https://doi.org/10.1017/S0890060423000069.
2. Yang, L.-C., Lerch, A.: On the evaluation of generative models in music. Neural Comput & Applic. 32, 4773–4784 (2020). https://doi.org/10.1007/s00521-018-3849-7.
3. Feng, Z., Guo, D., Tang, D., Duan, N., Feng, X., Gong, M., Shou, L., Qin, B., Liu, T., Jiang, D., Zhou, M.: CodeBERT: A Pre-Trained Model for Programming and Natural Languages, http://arxiv.org/abs/2002.08155, (2020).
4. Ciniselli, M., Cooper, N., Pascarella, L., Poshyvanyk, D., Di Penta, M., Bavota, G.: An Empirical Study on the Usage of BERT Models for Code Completion. In: 2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR). pp. 108–119 (2021). https://doi.org/10.1109/MSR52588.2021.00024.
5. Paris, M.: ChatGPT Hits 100 Million Users, Google Invests In AI Bot And CatGPT Goes Viral, https://www.forbes.com/sites/martineparis/2023/02/03/chatgpt-hits-100-million-microsoft-unleashes-ai-bots-and-catgpt-goes-viral/, last accessed 2023/03/13.
6. Welsh, M.: The End of Programming. Commun. ACM. 66, 34–35 (2022). https://doi.org/10.1145/3570220.
7. Mills, M.P., Lane, Boom, author of T.C.R.H. the C. of N.T.W.U. the N.E., podcast, a R. 2020s H. hosts T.L.O.: ChatGPT and Automation Come to Knowledge Work, https://www.city-journal.org/chatgpt-and-automation-come-to-knowledge-work, last accessed 2023/03/01.
8. Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., Kasneci, G.: ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences. 103, 102274 (2023). https://doi.org/10.1016/j.lindif.2023.102274.
9. How AI Could Save (Not Destroy) Education | Sal Khan | TED. (2023).
10. Fowler, G.A.: Analysis | We tested a new ChatGPT-detector for teachers. It flagged an innocent student., https://www.washingtonpost.com/technology/2023/04/01/chatgpt-cheating-detection-turnitin/, (2023).
11. Essien, A., Bukoye, O.T., O'Dea, X., Kremantzis, M.: The influence of AI text generators on critical thinking skills in UK business schools. Studies in Higher Education. 49, 865–882 (2024). https://doi.org/10.1080/03075079.2024.2316881.
12. Liu, A., Wu, Z., Michael, J., Suhr, A., West, P., Koller, A., Swayamdipta, S., Smith, N.A., Choi, Y.: We're Afraid Language Models Aren't Modeling Ambiguity, http://arxiv.org/abs/2304.14399, (2023). https://doi.org/10.48550/arXiv.2304.14399.
13. Valin, R.D.V.: ROLE AND REFERENCE GRAMMAR. Work Papers of the Summer Institute of Linguistics. 37, 12 (1993).
14. Cui, L., Wu, Y., Liu, J., Yang, S., Zhang, Y.: Template-Based Named Entity Recognition Using BART, http://arxiv.org/abs/2106.01760, (2021). https://doi.org/10.48550/arXiv.2106.01760.
15. Krishnan, V., Manning, C.D.: An Effective Two-Stage Model for Exploiting Non-Local Dependencies in Named Entity Recognition. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. pp. 1121–1128. Association for Computational Linguistics, Sydney, Australia (2006). https://doi.org/10.3115/1220175.1220316.

16. Mueller, E.T.: Story understanding through multi-representation model construction. In: Proceedings of the HLT-NAACL 2003 workshop on Text meaning - Volume 9. pp. 46–53. Association for Computational Linguistics, USA (2003). https://doi.org/10.3115/1119239.1119246.
17. Pandey, D., Suman, U., Ramani, A.K.: An Effective Requirement Engineering Process Model for Software Development and Requirements Management. In: 2010 International Conference on Advances in Recent Technologies in Communication and Computing. pp. 287–291. IEEE, Kottayam, India (2010). https://doi.org/10.1109/ARTCom.2010.24.
18. Ge, Y., Xiao, Z., Diesner, J., Ji, H., Karahalios, K., Sundaram, H.: What should I Ask: A Knowledge-driven Approach for Follow-up Questions Generation in Conversational Surveys, http://arxiv.org/abs/2205.10977, (2023). https://doi.org/10.48550/arXiv.2205.10977.
19. Moore, J.M., Shipman, F.M.: A comparison of questionnaire-based and GUI-based requirements gathering. In: Proceedings ASE 2000. Fifteenth IEEE International Conference on Automated Software Engineering. pp. 35–43 (2000). https://doi.org/10.1109/ASE.2000.873648.
20. OpenAI: GPT-4 Technical Report, http://arxiv.org/abs/2303.08774, (2023). https://doi.org/10.48550/arXiv.2303.08774.
21. Gemini Pro, https://deepmind.google/technologies/gemini/pro/.
22. Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web, https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/, last accessed 2023/03/01.
23. Kuhn, L., Gal, Y., Farquhar, S.: CLAM: Selective Clarification for Ambiguous Questions with Generative Language Models. ICML 2023 Workshop on Deployment Challenges for Generative AI. (2023).
24. Zhang, T., Qin, P., Deng, Y., Huang, C., Lei, W., Liu, J., Jin, D., Liang, H., Chua, T.-S.: CLAMBER: A Benchmark of Identifying and Clarifying Ambiguous Information Needs in Large Language Models, http://arxiv.org/abs/2405.12063, (2024). https://doi.org/10.48550/arXiv.2405.12063.
25. Park, J., Lim, S., Lee, J., Park, S., Chang, M., Yu, Y., Choi, S.: CLARA: Classifying and Disambiguating User Commands for Reliable Interactive Robotic Agents, http://arxiv.org/abs/2306.10376, (2023). https://doi.org/10.48550/arXiv.2306.10376.
26. Tabalba, R., Kirshenbaum, N., Leigh, J., Bhatacharya, A., Johnson, A., Grosso, V., Di Eugenio, B., Zellner, M.: Articulate+ : An Always-Listening Natural Language Interface for Creating Data Visualizations. In: Proceedings of the 4th Conference on Conversational User Interfaces. pp. 1–6. Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3543829.3544534.
27. Meet Khanmigo: Khan Academy’s AI-powered teaching assistant & tutor, https://khanmigo.ai/, last accessed 2024/06/24.
28. Shravya Bhat, Nguyen, H., Moore, S., Stamper, J., Sakr, M., Nyberg, E.: Towards Automated Generation and Evaluation of Questions in Educational Domains. (2022). https://doi.org/10.5281/ZENODO.6853085.
29. Human-AI collaboration patterns in AI-assisted academic writing, https://www.tandfonline.com/doi/epdf/10.1080/03075079.2024.2323593?needAccess=true, last accessed 2024/06/24.
30. Magic ToDo - GoblinTools, https://goblin.tools/, last accessed 2024/09/11.
31. D’Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M.D., Hormozdiari, F., Houlsby, N., Hou, S., Jerfel, G., Karthikesalingam, A., Lucic, M., Ma, Y., McLean, C., Mincu, D., Mitani, A., Montanari, A., Nado, Z.,

Natarajan, V., Nielson, C., Osborne, T.F., Raman, R., Ramasamy, K., Sayres, R., Schrouff, J., Seneviratne, M., Sequeira, S., Suresh, H., Veitch, V., Vladymyrov, M., Wang, X., Webster, K., Yadlowsky, S., Yun, T., Zhai, X., Sculley, D.: Underspecification presents challenges for credibility in modern machine learning. J. Mach. Learn. Res. 23, 226:10237-226:10297 (2022).
32. Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., Langosco, L., He, Z., Duan, Y., Carroll, M., Lin, M., Mayhew, A., Collins, K., Molamohammadi, M., Burden, J., Zhao, W., Rismani, S., Voudouris, K., Bhatt, U., Weller, A., Krueger, D., Maharaj, T.: Harms from Increasingly Agentic Algorithmic Systems. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. pp. 651–666. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3593013.3594033.
33. Jack Clark, Dario Amodei: Faulty reward functions in the wild, https://openai.com/research/faulty-reward-functions, last accessed 2023/07/07.
34. Mu, F., Shi, L., Wang, S., Yu, Z., Zhang, B., Wang, C., Liu, S., Wang, Q.: ClarifyGPT: Empowering LLM-based Code Generation with Intention Clarification, http://arxiv.org/abs/2310.10996, (2023).
35. Russel, S., Norvig, P.: Artificial intelligence: A Modern Approach. Prentice Hall, Upper Saddle River, NJ (2010).
36. Chomsky, N.: Syntactic Structures. De Gruyter Mouton (1957). https://doi.org/10.1515/9783110218329.
37. Fellbaum, C. ed: WordNet: An Electronic Lexical Database. A Bradford Book, Cambridge, MA, USA (1998).
38. Gildea, D., Jurafsky, D.: Automatic Labeling of Semantic Roles. Computational Linguistics. 28, 44 (2002).
39. Gliozzo, A.: Semantic Domains and Linguistic Theory. In: Narrative models: Narratology meets artificial intelligence. p. 6. , Genoa, Italy (2006).
40. Gliozzo, A., Strapparava, C., Dagan, I.: Unsupervised and supervised exploitation of semantic domains in lexical disambiguation. Computer Speech & Language. 18, 275–299 (2004). https://doi.org/10.1016/j.csl.2004.05.006.
41. Wittgenstein, L.: Philosophical Investigations. The Macmillan Company, New York (19665).
42. Gliozzo, A., Strapparava, C.: Domain Kernels for Text Categorization. In: Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005). pp. 56–63. Association for Computational Linguistics, Ann Arbor, Michigan (2005).
43. Magnini, B., Strapparava, C., Pezzulo, G., Gliozzo, A.: Using Domain Information for Word Sense Disambiguation. In: Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems. pp. 111–114. Association for Computational Linguistics, Toulouse, France (2001).
44. D'Avanzo, E., Gliozzo, A., Strapparava, C.: Automatic Acquisition of Domain Information for Lexical Concepts. 7.
45. Gliozzo, A.M.: The GOD model. In: Demonstrations. pp. 147–150 (2006).
46. Gliozzo, A., Strapparava, C.: Cross Language Text Categorization by Acquiring Multilingual Domain Models from Comparable Corpora. In: Proceedings of the ACL Workshop on Building and Using Parallel Texts. pp. 9–16. Association for Computational Linguistics, Ann Arbor, Michigan (2005).
47. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The Berkeley FrameNet Project. In: COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics (1998).
48. Johnson, C.R., Fillmore, C.J., Wood, E.J., Urban, M., Petruck, M.R.L., Baker, C.F., Fillmore, C.J., al, et: The FrameNet Project: tools for lexicon building, (2001).

49. Ratinov, L., Roth, D.: Design Challenges and Misconceptions in Named Entity Recognition. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning - CoNLL ’09. p. 147. Association for Computational Linguistics, Boulder, Colorado (2009). https://doi.org/10.3115/1596374.1596399.
50. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural Architectures for Named Entity Recognition. Proceedings of NAACL 2016. (2016).
51. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Computation. 9, 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735.
52. Bordes, A., Usunier, N., Chopra, S., Weston, J.: Large-scale Simple Question Answering with Memory Networks. arXiv:1506.02075 [cs]. (2015).
53. Hill, F., Bordes, A., Chopra, S., Weston, J.: The Goldilocks Principle: Reading Children’s Books with Explicit Memory Representations. arXiv:1511.02301 [cs]. (2016).
54. Project Gutenberg, https://www.gutenberg.org/, last accessed 2022/04/26.
55. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need, http://arxiv.org/abs/1706.03762, (2017). https://doi.org/10.48550/arXiv.1706.03762.
56. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-Shot Text-to-Image Generation. In: Proceedings of the 38th International Conference on Machine Learning. pp. 8821–8831. PMLR (2021).
57. Yu, L., Cheng, Y., Sohn, K., Lezama, J., Zhang, H., Chang, H., Hauptmann, A.G., Yang, M.-H., Hao, Y., Essa, I., Jiang, L.: MAGVIT: Masked Generative Video Transformer. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10459–10469. IEEE, Vancouver, BC, Canada (2023). https://doi.org/10.1109/CVPR52729.2023.01008.
58. Vondrick, C., Torralba, A.: Generating the Future with Adversarial Transformers. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2992–3000. IEEE, Honolulu, HI (2017). https://doi.org/10.1109/CVPR.2017.319.
59. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving Language Understanding by Generative Pre-Training. (2018).
60. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language Models are Few-Shot Learners. In: Advances in Neural Information Processing Systems. pp. 1877–1901. Curran Associates, Inc. (2020).
61. Knight, W.: OpenAI’s CEO Says the Age of Giant AI Models Is Already Over, https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/, (2023).
62. Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., Wu, Y.: Exploring the Limits of Language Modeling, http://arxiv.org/abs/1602.02410, (2016).
63. ChatGPT: Optimizing Language Models for Dialogue, https://openai.com/blog/chatgpt/, last accessed 2023/02/15.
64. Google: Bard, https://bard.google.com, last accessed 2023/05/22.
65. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, http://arxiv.org/abs/1810.04805, (2019). https://doi.org/10.48550/arXiv.1810.04805.
66. MetaAI: Introducing LLaMA: A foundational, 65-billion-parameter language model, https://ai.facebook.com/blog/large-language-model-llama-meta-ai/, last accessed 2023/05/22.

67. Microsoft: Your AI-Powered Copilot for the Web, https://www.microsoft.com/en-us/bing, last accessed 2023/05/22.

68. GitHub Copilot · Your AI pair programmer, https://github.com/features/copilot, last accessed 2023/02/18.

69. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language Models are Unsupervised Multitask Learners. OpenAI blog. 1.8, (2019).

70. Levesque, H., Davis, E., Morgenstern, L.: The Winograd Schema Challenge.

71. Reddy, S., Chen, D., Manning, C.D.: CoQA: A Conversational Question Answering Challenge. Transactions of the Association for Computational Linguistics. 7, 249–266 (2019). https://doi.org/10.1162/tacl_a_00266.

72. Zhang, S., Pan, L., Zhao, J., Wang, W.Y.: Mitigating Language Model Hallucination with Interactive Question-Knowledge Alignment, http://arxiv.org/abs/2305.13669, (2023). https://doi.org/10.48550/arXiv.2305.13669.

73. Gary Marcus, Ernest Davis: GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about, https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/, last accessed 2023/02/24.

74. Rudolph, J., Tan, S., Tan, S.: ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? JALT. 6, (2023). https://doi.org/10.37074/jalt.2023.6.1.9.

75. Agüera y Arcas, B.: Do Large Language Models Understand Us? Daedalus. 151, 183–197 (2022). https://doi.org/10.1162/daed_a_01909.

76. Bohannon, M.: Lawyer Used ChatGPT In Court—And Cited Fake Cases. A Judge Is Considering Sanctions, https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/, (2023).

77. Merken, S., Merken, S.: New York lawyers sanctioned for using fake ChatGPT cases in legal brief, https://www.reuters.com/legal/new-york-lawyers-sanctioned-using-fake-chatgpt-cases-legal-brief-2023-06-22/, (2023).

78. Shimbun, T.Y.: ChatGPT can be tricked into generating malware, bomb-making instructions, https://asianews.network/chatgpt-can-be-tricked-into-generating-malware-bomb-making-instructions/, (2023).

79. Walker, J.: The night I asked ChatGPT how to build a bomb, https://reason.com/2024/05/11/the-night-i-asked-chatgpt-how-to-build-a-bomb/, last accessed 2024/09/09.

80. Our Approach to Alignment Research, https://openai.com/blog/our-approach-to-alignment-research/, last accessed 2023/02/16.

81. Biddle, S.: The Internet's New Favorite AI Proposes Torturing Iranians and Surveilling Mosques, https://theintercept.com/2022/12/08/openai-chatgpt-ai-bias-ethics/, last accessed 2023/02/24.

82. Mary Papenfuss: Creepy Microsoft Bing Chatbot Urges Tech Columnist To Leave His Wife, https://www.huffpost.com/entry/kevin-roose-ai-chatbot_n_63eeb367e4b0063ccb2bcc45, (2023).

83. Si, W.M., Backes, M., Blackburn, J., De Cristofaro, E., Stringhini, G., Zannettou, S., Zhang, Y.: Why So Toxic? Measuring and Triggering Toxic Behavior in Open-Domain Chatbots. In: Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security. pp. 2659–2673. Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3548606.3560599.

84. Welbl, J., Glaese, A., Uesato, J., Dathathri, S., Mellor, J., Hendricks, L.A., Anderson, K., Kohli, P., Coppin, B., Huang, P.-S.: Challenges in Detoxifying Language Models, http://arxiv.org/abs/2109.07445, (2021).
85. Rozado, D.: Where does ChatGPT fall on the political compass?, https://reason.com/2022/12/13/where-does-chatgpt-fall-on-the-political-compass/, last accessed 2023/02/24.
86. Guynn, J.: Is ChatGPT 'woke'? AI chatbot accused of anti-conservative bias and a grudge against Trump, https://www.usatoday.com/story/tech/2023/02/09/woke-chatgpt-conservatives-bias/11215353002/, last accessed 2023/02/16.
87. McGee, R.W.: Is Chat Gpt Biased Against Conservatives? An Empirical Study, https://papers.ssrn.com/abstract=4359405, (2023). https://doi.org/10.2139/ssrn.4359405.
88. Plumbing ChatGPT's Left-Liberal Biases › American Greatness, https://amgreatness.com/2023/01/28/plumbing-chatgpts-left-liberal-biases/, last accessed 2023/02/16.
89. Frieder, S., Pinchetti, L., Griffiths, R.-R., Salvatori, T., Lukasiewicz, T., Petersen, P.C., Chevalier, A., Berner, J.: Mathematical Capabilities of ChatGPT, http://arxiv.org/abs/2301.13867, (2023). https://doi.org/10.48550/arXiv.2301.13867.
90. Buscaroli, R., Chesani, F., Giuliani, G., Loreti, D., Mello, P.: A Prolog application for reasoning on maths puzzles with diagrams. Journal of Experimental & Theoretical Artificial Intelligence. 0, 1–21 (2022). https://doi.org/10.1080/0952813X.2022.2062456.
91. Chesani, F., Mello, P., Milano, M.: Solving Mathematical Puzzles: A Challenging Competition for AI. AI Magazine. 38, 83–96 (2017). https://doi.org/10.1609/aimag.v38i3.2736.
92. Zadeh, L.A.: A note on web intelligence, world knowledge and fuzzy logic. Data & Knowledge Engineering. 50, 291–304 (2004). https://doi.org/10.1016/j.datak.2004.04.001.
93. Shanahan, M.: A Logical Formalisation of Ernie Davis's Egg Cracking Problem. In: Problem. Fourth Symposium on Logical Formalizations of Commonsense Reasoning (1997).
94. Shanahan, M.: The Frame Problem. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University (2016).
95. Mueller, E.T.: Commonsense reasoning: an event calculus based approach. Morgan Kaufmann (2014).
96. Mueller, E.T.: Event calculus reasoning through satisfiability. Journal of Logic and Computation. 14, 2004 (2004).
97. Ma, J., Miller, R., Morgenstern, L., Patkos, T.: An Epistemic Event Calculus for ASP-based Reasoning About Knowledge of the Past, Present and Future, https://doi.org/10.29007/zswj, last accessed 2022/05/26.
98. Lenat, D.: Getting from Generative AI to Trustworthy AI: What LLMs might learn from Cyc. (2023).
99. Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., Choi, Y.: COMET: Commonsense Transformers for Automatic Knowledge Graph Construction, http://arxiv.org/abs/1906.05317, (2019). https://doi.org/10.48550/arXiv.1906.05317.
100. Edwards, B.: Telling AI model to "take a deep breath" causes math scores to soar in study, https://arstechnica.com/information-technology/2023/09/telling-ai-model-to-take-a-deep-breath-causes-math-scores-to-soar-in-study/, last accessed 2023/09/26.
101. Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q.V., Zhou, D., Chen, X.: Large Language Models as Optimizers, http://arxiv.org/abs/2309.03409, (2024). https://doi.org/10.48550/arXiv.2309.03409.

102. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. ACM Comput. Surv. 55, 195:1-195:35 (2023). https://doi.org/10.1145/3560815.
103. How to Write Better Prompts for Chat GPT, https://www.griproom.com/fun/how-to-write-better-prompts-for-chat-gpt, last accessed 2023/07/20.
104. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, http://arxiv.org/abs/2201.11903, (2023). https://doi.org/10.48550/arXiv.2201.11903.
105. Wei, X., Cui, X., Cheng, N., Wang, X., Zhang, X., Huang, S., Xie, P., Xu, J., Chen, Y., Zhang, M., Jiang, Y., Han, W.: ChatIE: Zero-Shot Information Extraction via Chatting with ChatGPT, http://arxiv.org/abs/2302.10205, (2024). https://doi.org/10.48550/arXiv.2302.10205.
106. Pyatkin, V., Hwang, J.D., Srikumar, V., Lu, X., Jiang, L., Choi, Y., Bhagavatula, C.: ClarifyDelphi: Reinforced Clarification Questions with Defeasibility Rewards for Social and Moral Situations, http://arxiv.org/abs/2212.10409, (2023). https://doi.org/10.48550/arXiv.2212.10409.
107. Fischer, J.E.: Generative AI Considered Harmful. In: Proceedings of the 5th International Conference on Conversational User Interfaces. pp. 1–5. ACM, Eindhoven Netherlands (2023). https://doi.org/10.1145/3571884.3603756.
108. O'Dea, X.: Generative AI: is it a paradigm shift for higher education? Studies in Higher Education. 49, 811–816 (2024). https://doi.org/10.1080/03075079.2024.2332944.
109. Brittain, B., Brittain, B.: How copyright law could threaten the AI industry in 2024, https://www.reuters.com/legal/litigation/how-copyright-law-could-threaten-ai-industry-2024-2024-01-02/, (2024).
110. Brittain, B., Brittain, B.: Google defeats class action over AI training data for now, https://www.reuters.com/legal/transactional/google-defeats-class-action-over-ai-training-data-now-2024-06-06/, (2024).
111. AI Detector Tool | AI Checker for ChatGPT, Gemini & Claude, https://staging.copyleaks.com/ai-content-detector, last accessed 2024/09/11.
112. Free AI Detector, https://www.scribbr.com/ai-detector/, last accessed 2024/09/11.
113. AI Detector - the Original AI Checker for ChatGPT & More, https://gptzero.me/, last accessed 2024/09/11.
114. AI Checker Solutions: Ensure Academic Integrity | Turnitin, https://www.turnitin.com/solutions/topics/ai-writing/ https://www.turnitin.com/solutions/topics/ai-writing/, last accessed 2024/09/11.
115. Teo, K.X.: Even OpenAI's own detection service can't tell AI-generated work apart — the company quietly took it down over accuracy concerns, https://www.businessinsider.com/openai-chatgpt-ai-detection-tool-shut-down-over-inaccuracy-2023-7, (2023).
116. Turnitin's AI writing detection capabilities FAQs, https://guides.turnitin.com/hc/en-us/articles/28477544839821-Turnitin-s-AI-writing-detection-capabilities-FAQs, last accessed 2024/09/11.
117. Dheda, G.: Can Turnitin Detect Chat GPT?, https://openaimaster.com/can-turnitin-detect-chat-gpt/, last accessed 2024/09/11.
118. Cardona, M.A., Rodríguez, R.J., Ishmael, K.: Artificial Intelligence and the Future of Teaching and Learning. Office of Education Technology. (2023).
119. Kelly, J.: What White-Collar Jobs Are Safe From AI—And Which Professions Are Most At Risk?, https://www.forbes.com/sites/jackkelly/2024/02/28/what-white-collar-jobs-are-safe-from-ai-and-which-professions-are-most-at-risk/, (2024).

120. Yellin, D.M.: The Premature Obituary of Programming. Commun. ACM. 66, 41–44 (2023). https://doi.org/10.1145/3555367.

121. AI Photo Editor: Edit Images with AI in Photoshop - Adobe, https://www.adobe.com/products/photoshop/ai.html, last accessed 2024/09/11.

122. Nguyen, A., Hong, Y., Dang, B., Huang, X.: Human-AI collaboration patterns in AI-assisted academic writing. Studies in Higher Education. 49, 847–864 (2024). https://doi.org/10.1080/03075079.2024.2323593.

123. Hancock, P.A.: Automation: how much is too much? Ergonomics. 57, 449–454 (2014). https://doi.org/10.1080/00140139.2013.816375.

124. Ahmed, D.: Anthropomorphizing artificial intelligence: towards a user-centered approach for addressing the challenges of over-automation and design understandability in smart homes. Intelligent Buildings International. 13, 227–240 (2021). https://doi.org/10.1080/17508975.2020.1795612.

125. Norman, D.A., Broadbent, D.E., Baddeley, A.D., Reason, J.: The ‘problem ’ with automation: inappropriate feedback and interaction, not ‘over-automation.’ Philosophical Transactions of the Royal Society of London. B, Biological Sciences. 327, 585–593 (1989). https://doi.org/10.1098/rstb.1990.0101.

126. Douglas B. Lenat: CYC: A Large-Scale Investment in Knowledge Infrastructure. COMMUNICATIONS OF THE ACM. 38, 33–38 (1995).

127. Cyc | The Next Generation of Enterprise AI, https://cyc.com/, last accessed 2022/12/05.

128. Lenat, D.: Not Good As Gold: Today’s AI’s Are Dangerously Lacking In AU (Artificial Understanding), https://www.forbes.com/sites/cognitiveworld/2019/02/18/not-good-as-gold-todays-ais-are-dangerously-lacking-in-au-artificial-understanding/, last accessed 2022/12/05.

129. Lin, S., Hilton, J., Evans, O.: TruthfulQA: Measuring How Models Mimic Human Falsehoods, http://arxiv.org/abs/2109.07958, (2022). https://doi.org/10.48550/arXiv.2109.07958.

130. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (2002). https://doi.org/10.3115/1073083.1073135.

131. Zhang*, T., Kishore*, V., Wu*, F., Weinberger, K.Q., Artzi, Y.: BERTScore: Evaluating Text Generation with BERT. Presented at the International Conference on Learning Representations September 25 (2019).

132. Rogers, A., Gardner, M., Augenstein, I.: QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension. ACM Comput. Surv. 55, 1–45 (2023). https://doi.org/10.1145/3560260.

133. Tix, B., Binsted, K.: Better Results Through Ambiguity Resolution: Large Language Models that Ask Clarifying Questions. In: Schmorrow, D.D. and Fidopiastis, C.M. (eds.) Augmented Cognition. pp. 72–87. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-61572-6_6.

134. Azure Serverless Functions, https://azure.microsoft.com/en-us/products/functions.

135. Azure Data Services, https://azure.microsoft.com/en-us/solutions/databases/.

136. GPT-3.5 Turbo, https://platform.openai.com/docs/models/gpt-3-5-turbo.

137. Tix, B.J.: Follow-Up Questions Improve Documents Generated by Large Language Models.

138. GPT-4 Turbo, https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4.

139. Facebook Ads Manager: Ads Management for Facebook, Instagram, https://www.facebook.com/business/tools/ads-manager, last accessed 2024/08/06.

140. Survey 4 Survey, https://www.reddit.com/r/SurveyExchange/, last accessed 2024/08/06.

141. Angulo, O., O'Mahony, M.: The paired preference test and the 'No Preference' option: was Odesky correct? Food Quality and Preference. 16, 425–434 (2005). https://doi.org/10.1016/j.foodqual.2004.08.002.
142. Shaffer, J.P.: Multiple Hypothesis Testing. Annual Review of Psychology. (1995).
143. Introducing OpenAI o1, https://openai.com/index/introducing-openai-o1-preview/, last accessed 2024/09/12.
144. Learning to Reason with LLMs, https://openai.com/index/learning-to-reason-with-llms/, last accessed 2024/09/12.
145. United States of America - Place Explorer - Data Commons, https://datacommons.org/place/country/USA?utm_medium=explore&mprop=age&popt=Person&hl=en, last accessed 2024/08/22.
146. Cloud Computing Services, https://cloud.google.com/, last accessed 2024/08/15.
147. Azure Data Studio | Microsoft Azure, https://azure.microsoft.com/en-us/products/data-studio, last accessed 2024/08/15.