GPT in the Loop: Evidence from the Field

Cathy Yang
HEC Paris
yang@hec.fr

Leo Allen
HEC Paris
leo.allen@hec.edu

David Restrepo-Amariles
HEC Paris
restrepo-amariles@hec.fr

Aurore Troussel
HEC Paris
aurore.troussel@hec.edu

Abstract

Generative Pre-trained Transformers (GPTs) are highly effective in generating content and increasing productivity, but companies have reservations about their use in a professional setting. OpenAI and policymakers suggest that disclosing the use of GPT is necessary, but there is little empirical evidence to understand its consequences. Our experiment found that managers from a leading consulting firm were unable to distinguish Human-GPT generated content when the content generation source was not disclosed and disclosing the use of GPT improved the content's evaluation. We explored the effects of applying the GPT disclosure policy in the workplace. Managers prefer analysts to disclose their use of GPT, but their preferences regarding how junior analysts should use GPT may differ from those of the analysts, leading to potential conflicts over disclosure.

Keywords: Human-GPT collaboration, GPT disclosure, content evaluation, survey experiment

1. Introduction

Generative Pre-trained Transformers (GPTs) powered by large language models (LLMs) have become the fasted adopted technologies in the industry's history (Hu, 2023). Despite the excitement about GPT's potential to generate quality content (Noy and Zhang, 2023) and improve labor productivity (Brynjolfsson et al., 2023; Peng, Kalliamvakou, et al., 2023), governments and firms are concerned about its capacity to produce misinformation (Peng, Galley, et al., 2023) and to compromise personal information passed on to prompts (Khowaja et al., 2023). For example, Italy temporarily banned ChatGPT in March 2023 since the country's data protection authority said its developers had no legal basis to justify collecting and processing personal data for the purpose of training its algorithms (Garante, 2023). Major consulting firms have banned their employees from using GPT at work, and multiple educational institutions have banned GPT on campus. However, simply banning GPT is neither optimal nor sustainable. It prevents users and firms from benefiting from GPT's potential to improve productivity, and may result in covert adoption by employees and students, ultimately leaving firms without the knowledge and capacity to mitigate the risks associated with its use.

To reap the full benefit of GPT while hedging risks such as privacy invasion, breach of intellectual property rights, and misinformation, one of the most frequent solutions proposed by OpenAI, regulators, and industry leaders is to disclose the use of GPT (OpenAI, 2023). For instance, the proposal of the European Union to regulate AI ("AI Act", European Commission, 2023) provides that technologies like GPTs would need to comply with transparency requirements to be able to circulate in the European market, while AI industry leaders consider that GPT policies that guarantee transparency and user rights are a better alternative than banning the tool altogether (Austin and Wright, 2023).

Despite a possible solution, it is still being determined whether companies will gain advantages by revealing their use of GPT while mitigating accompanying risks based on traditional literature that studies algorithmic adoption (e.g., Dietvorst et al., 2015). First, single-purposed algorithms fundamentally differ from GPT, which creates a diverse range of stochastic output, enables various interactions with humans through free-style prompting, and produces human-like content that may contain misinformation (i.e., hallucination). Second, previous literature on algorithmic adoption suggests divergent findings. On the one hand, recent studies by Logg et al., 2019 showed that disclosing the advice content generated by an algorithm (vs. a human) increases individuals' confidence in the advice content quality because of trust in the algorithm's superior predictive performance. On the other hand, individuals are less likely to utilize advice generated by an algorithm because they distrust its ability to provide personalized suggestions (Longoni et al., 2019) or when the prediction tasks

URI: https://hdl.handle.net/10125/106907
978-0-9981331-7-1

are considered more subjective (Castelo, Bos, and Lehmann, 2019). Given the positive relationship between humans' trust and their evaluation of content created by algorithms or humans, we anticipate GPT's ability to produce flexible and human-like content could increase managers' trust and content evaluation upon knowing GPT's participation in content generation. However, on the flip side, managers may distrust and discount content generated by GPT due to its potential to create misinformation. In this paper, we contribute to the literature by answering the research question: Does disclosing the use of GPT positively or negatively affect managers' evaluation of content generated by humans and GPT?

We further investigate the effects of implementing the GPT use disclosure policy from the business organization perspective. This policy could assist managers in tracking and assigning responsibility for client-facing content creation, ultimately increasing perceived accountability for their clients (Weiner et al., 1987). However, establishing a disclosure policy may face organizational resistance, decreasing the company's competitive edge. Such resistance to implementing the GPT disclosure policy can occur when employees of varying levels within the organization disagree on implementing the policy or when employees in different hierarchical positions have conflicting views on using GPTs (Gottschalg and Zollo, 2007). Such concern is evident regarding the use of GPT since it is easier for the managers to identify junior analysts' actual use of GPT if a trust relationship between managers and analysts accompanies a truthful disclosure. In this paper, we investigate whether managers' preference for junior analysts to reveal their GPT use correlates with their preference to reveal their GPT use and whether managers' preference to allow analysts to use GPT correlates with their actual usage.

We collaborated with a major consulting firm that regulates the use of GPT at work (e.g., prohibiting employees from including their clients' information in prompts). The firm enrolled its mid-level managers in a training session to help them understand how they can benefit from generative AI to provide data-driven solutions for their potential clients without leaking clients' information. These mid-level managers often create pitch content in response to request-for-proposals (RFPs), which are calls for projects originating from potential new clients. Typically, mid-level managers assign junior analysts to generate research briefs for pitch content. In this training session, we asked the mid-level managers to evaluate the research briefs generated by junior analysts in response to RFPs from fictitious clients, avoiding legal risks arising from using clients' data. Specifically, each manager received two RFPs and was asked to evaluate two versions of the research brief decks per RFP: one produced with GPT's help and one without. Importantly, we conducted an experiment where managers were randomly assigned to one of the two experimental conditions, with one condition disclosing the source of the content generation and one without (No-GPT vs. Human-GPT). 4) What are the managers' preferences regarding junior analysts' use of GPT and its disclosure when generating business development content?

Our initial findings indicate the following. First, junior analysts have intensively used GPT when allowed, and such Human-GPT collaborative content uses more complex vocabulary. Second, managers could not differentiate between content created by junior analysts with or without assistance from GPT. Third, disclosing the use of GPT resulted in managers evaluating the content more favorably. Fourth, managers indicated a strong preference for junior analysts revealing their use of GPT, but were hesitant to do so themselves. Managers also expected junior analysts to produce content faster once they knew they were using GPT and were more willing to allow junior analysts to use GPT after knowing the content generation source. However, their preferences on which section junior analysts should use GPT to improve content did not align with the actual use in our experiment.

Our study contributes to the literature by showing how managers perceive human-GPT collaborative content with and without disclosing the content generation with access to GPT in a survey experiment. In particular, our findings suggest that there is a potential positive shift in the likelihood that content generated by a human-GPT collaboration will disseminate when the use of GPT is disclosed. Our study gives policy implications about implementing GPT usage disclosure. Although the managers are more lenient about allowing their junior analysts to use GPT after being informed of the usage, there may be misalignment between the managers' expectations regarding the speed of completing a task with GPT's help and the actual time it will take. Moreover, we discovered possible divergent preferences regarding how GPT should be used among the managers and analysts.

2. Related work

2.1. Human-AI collaboration

A nascent yet fast-growing literature examines how humans could collaborate with artificial intelligence (AI) to produce better services. While some studies

focused on understanding whether AI could incorporate human input as a form of human-AI collaboration to improve performance (e.g., Cao et al., 2021), others have examined whether and how humans incorporate or delegate decisions by interacting with an algorithm (e.g., Berk, 2017; Dietvorst et al., 2015; Lebovitz et al., 2022; Logg et al., 2019).

The majority of literature discussing human-AI collaboration involves AI tools that perform prediction tasks, such as image classification (Fügener et al., 2022) or making predictions for the re-arrest chance (Berk, 2017). GPT differs from prediction-based AI because it generates an impressive range of human-like text by responding to an unbounded set of potential input prompts. In addition, the output is stochastic such that GPT's responses to the same prompt are not identical but rather follow a distribution (Brand et al., 2023; Horton, 2023). Despite its stochastic nature, human-GPT collaborative content is better evaluated than human-generated with some preliminary laboratory evidence (Noy and Zhang, 2023). However, using GPT entails risks stemming from the lack of protection for the input information and the potential to generate misinformation as output (e.g., hallucinations).

The closest research to our study is probably Reisenbichler et al., 2022, which shows that the GPT-2 generated marketing content (reviewed by humans) outperforms that of human marketers. Unlike our study, GPT in Reisenbichler et al., 2022 is highly controlled such that GPT was used to predict the top-ranked content rather than interacting with humans through prompts. Importantly, the purpose of the study was to investigate the performance of the GPT-facilitated marketing content generation process—humans consuming the content were unaware of the participation of GPT in generating the marketing content.

To our knowledge, no previous research has explored how content created through collaboration between humans and GPT via free use of prompts is perceived and evaluated in professional settings. Our study aims to fill the literature gap.

2.2. Impact of AI in the professional context

There is a large body of literature on the impact of AI and automation technologies on the labor market and work organization (e.g., Acemoglu et al., 2022; Fügener et al., 2022). A nascent and growing literature focuses on the impact of LLMs (e.g., GPT) in a professional context due to its performance in language-based tasks. Some studies used metrics such as AI occupational exposure to assess how much LLM would impact occupations and industries (e.g., Eloundou et al., 2023; Felten et al., 2023). Other studies show that adopting LLM increases productivity (Brynjolfsson et al., 2023; Noy and Zhang, 2023; Reisenbichler et al., 2022).

Notwithstanding, all the studies assumed the use of GPT is free from concerns in the professional context, while this is far from the case (Khowaja et al., 2023 OpenAI, 2023). Many companies with data privacy concerns prohibit the use of GPT. One potential solution is to increase transparency in its use to benefit from its content generation capabilities while avoiding risks. However, there has been no prior research on how managers perceive the use of generative AI in the workplace when concerns arise. Additionally, there has been no research on whether disclosing the use of GPT for content generation affects its evaluation and dissemination in a professional context. Our study aims to fill this gap in the literature by investigating whether disclosing the use of GPT for content generation affects its evaluation in a professional context.

2.3. The impact of algorithmic transparency on human-AI collaboration

Our study is also related to the literature on algorithmic transparency. Previous literature on algorithmic transparency focuses on either trying to uncover the black box of an algorithm (e.g., Lage et al., 2019; Poursabzi-Sangdeh et al., 2021) or revealing its performance (Dietvorst et al., 2015; You et al., 2022). While increasing the algorithmic transparency could increase humans' trust in the algorithmic output (Castelo, Bos, and Lehmann, 2019; Dietvorst et al., 2015; Logg et al., 2019), this is not always the case. Dietvorst et al., 2015 find that revealing the prediction performance of the algorithm makes humans less likely to adopt the algorithm for delegation. You et al., 2022 find that increasing the transparency in the algorithmic performance could reduce one's reliance on the algorithm due to increased cognitive cost in processing the algorithmic performance information.

Some studies focus on manipulating the source of advice, comparing individuals' acceptance of advice generated by either humans or an algorithm (Gunaratne et al., 2018; You et al., 2022). Yet, no prior study examined the impact of transparency in disclosing the advice sources on individuals' acceptance of human-AI collaborative advice, which we contribute in this study.

3. Producing pitch content with and without GPT

In this section, we explain the content generation process for the human and human-GPT collaborative

research briefs, and the differences in the content generation outcomes.

3.1. Human-generated vs. human-GPT collaborative research brief generation

Two master's students with previous experience in consulting were hired from a top European business school as junior analysts. They were asked to generate research briefs as potential content to be included by the managers in response to clients' RFP proposals (Falkner et al., 2019). We created four RFPs with fictitious client names to increase the generalizability of our findings regarding the business topic.

Two analysts were instructed to collaborate and create four research briefs without using any LLMs, which they named the "No-GPT deck." After completing these decks, they were asked to use ChatGPT 3.5 to modify the "No-GPT deck," without any specific instructions on how to prompt ChatGPT. The research brief that was improved with the assistance of ChatGPT was named "Human-GPT deck."[1] The analysts were also asked to document their prompts by exporting their interactions with ChatGPT.

All decks comprised six sections: a problem statement, solution overview, implementation details, expected outcomes, team, and timeline.

It took the two analysts 38.5 hours to finish generating the 8 decks (four No-GPT decks and four Human-GPT decks), ranging from 6 to 11 pages each. After the task completion, the two students were paid 17 € per hour for the actual number of working hours.

3.2. Results

This section provides descriptive analyses of how the two junior analysts instructed GPT to create the Human-GPT decks. It also documents the differences in content between the Human-GPT decks and the No-GPT ones.

3.2.1. How did the junior analysts use the GPT in generating the Human-GPT deck? The two analysts used 138 prompts before finalizing the four Human-GPT decks. We first categorize the prompts into four categories based on their application purpose (summarizing, expanding, inferring, and transforming) according to Ng and Fulford, 2023 and label the prompt related to improving the overall content (labeled as "all"), one of the six sections (problem statement, solution overview, implementation details, expected outcomes, team, and timeline), or other content such as titles.

We then analyze the prompts at the content category level. Figures 1a and 1b show the frequency count and percentage of the prompt types used in content categories. We find a significant difference in the type of prompts used across different content categories ($p < 0.001$ given the Pearson's Chi-squared test). We discovered that the expected outcome section was the most frequently prompted to infer outcomes by the two analysts using GPT (22 times). However, such inferring prompts are rarely applied in other content categories except for the team and timeline sections. We also find that the two analysts frequently used transforming prompts for the content in the problem statement section (16 times), and such prompt types are not observed for the team and timeline sections.

3.2.2. The degree of GPT contribution in the Human-GPT (vs. No-GPT) deck content Given that the Human-GPT deck is an updated version of the No-GPT deck that some content could be changed unrelated to GPT's responses, we identify the contribution of GPT to the Human-GPT deck by the differences in content similarity between the two versions of the brief and GPT responses. We look at the text content in its raw format per case and per content section for the two versions of the brief. We first define the content similarity between a deck and GPT responses by the cosine similarity based on the text tokens using BERT tokenization (Devlin et al., 2018) ranging from 0 to 1, where a higher number indicates higher content similarity between the Human-GPT (or No-GPT) deck and GPT responses. We then proxy the degree of content contribution by GPT in the Human-GPT decks by the absolute difference in content similarity between No-GPT and GPT responses and that of Human-GPT and GPT. We show the average degree of GPT contribution to the content by section in Figure 2.

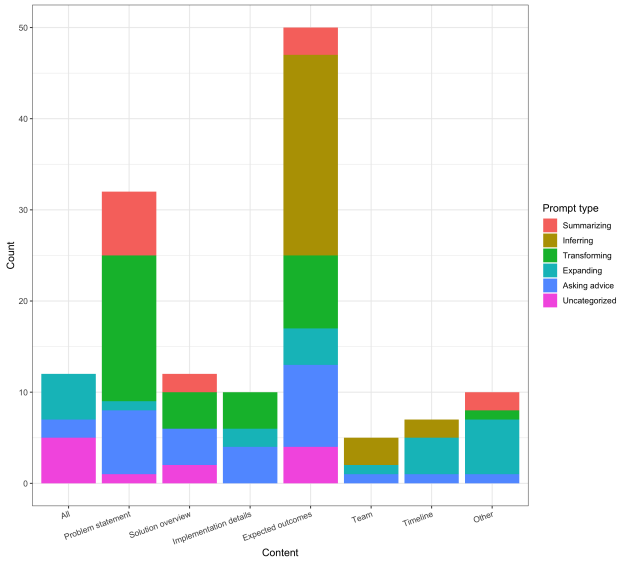
We then investigate whether the degree of GPT contribution in the Human-GPT content correlates with the frequency of the prompt types used in each content section. Due to limited observations (N = 24), we did not find evidence that the degree of GPT contribution correlated with the number of prompts used in different sections ($ps > 0.11$).

We also measure the degree of content differences in the Human-GPT and the No-GPT decks by the cosine similarity based on the text tokens using BERT tokenization subtracted from one. We find that the degree of content modification in the Human-GPT deck

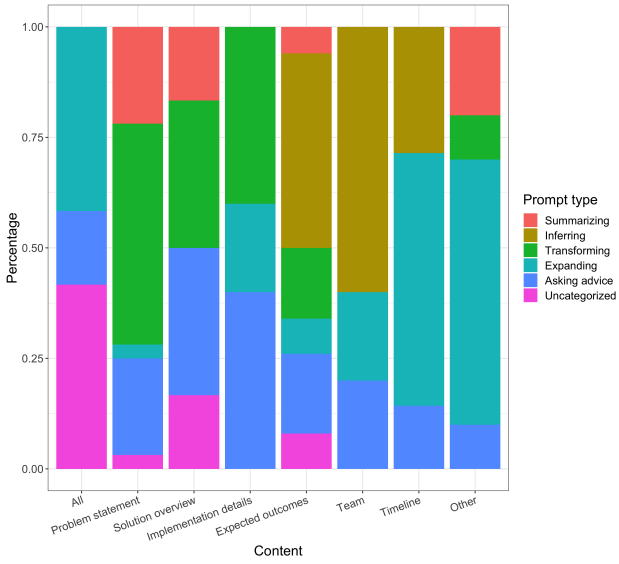
[1] Our design guarantees the content difference in the human-GPT collaborative version stemming from using the GPT. Note that generating human-GPT collaborative content requires a longer time than the human-generated content by our design.

Figure 1: Type of prompt used in each content generation section

(a) Count



(b) Percentage



positively correlates with the degree of GPT content contribution ($\rho = 0.52$, $p = 0.01$).

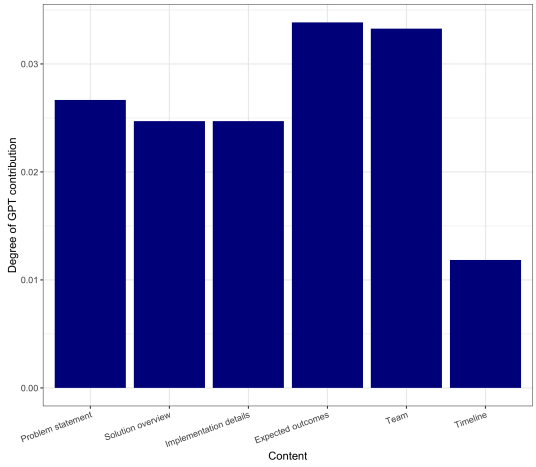


Figure 2: The degree of GPT contribution to the Human-GPT deck content

3.2.3. The linguistic differences between the Human-GPT and No-GPT decks To understand the linguistic differences between the Human-GPT and No-GPT decks, we analyzed each section of a deck. Since each section has a median word count of 67, relying solely on topic modeling to infer content and linguistic differences from only 48 observations in the six sections of the four cases would be unrealistic. We used LIWC-2022 (Boyd et al., 2022), a dictionary-based analysis commonly used in social science (e.g., Kuhnen and Niessen, 2012), to identify 113 different linguistic dimensions of the content. With this information, we measure the difference in linguistic dimensions between the Human-GPT and No-GPT content for each section in a deck.

To understand which linguistic dimensions are changed due to text modification, we correlate the linguistic deviation in the Human-GPT (vs. No-GPT) content with the degree of modification in the text content. Compared to the No-GPT decks, the changes in the Human-GPT deck manifest with lower use of numbers ($\rho = 0.43$, $p = 0.04$) and more words related to lifestyle (such as work, home, school, and employment) ($\rho = 0.38$, $p = 0.07$). The summary statistics are shown in Table 1.

Overall, we have discovered that the two analysts gave GPT a considerable amount of prompts, averaging 34 prompts per deck. Interactions with GPT result in content modification, evidenced by increased words related to numbers and references to lifestyle topics such as work.

4. Survey experiment

4.1. Experimental design

We conducted a survey experiment to determine if managers were more receptive to RFP research brief content generated by humans with vs. without access to GPT. Additionally, we investigated if disclosing vs. obscuring GPT involvement impacts managers' content evaluations.

Our experimental design includes a combination of within-subject and between-subject designs. The between-subject design focuses on transparency in disclosing the content generation source (human or human-GPT), as compared to no such transparency. The within-subject design involves each manager evaluating both the No-GPT and Human-GPT versions of the research brief for a specific RFP.

4.2. Experimental procedure and data

We collaborate with the consulting firm on training 130 mid-level managers about using generative AI to provide data-driven solutions across two sessions, where 45 managers signed up for the session on June 5, 2023.[2] We acknowledge that our survey experiment's data is limited in sample size, and our results are preliminary.

Forty-five managers came to the campus of a top European business school and were randomly assigned to two classrooms for a 90-minute training followed by a wrap-up session the following day. Before starting the 90-minute training session, all managers were told about their mission to evaluate the research brief content generated by two junior analysts in response to two RFPs from two potential clients. Before the evaluations, we asked all managers about their beliefs regarding the likelihood of the two junior analysts using GPT in percentage to create research briefs. Each manager was randomly assigned two out of the four possible RFPs, where No-GPT and Human-GPT decks were presented in random order for evaluation given each RFP. In one classroom, managers were told if the content was made by humans or human-GPT collaboration during the evaluation. However, managers in the other classroom did not receive this information.

The managers were asked to rate how engaging each deck was on a scale of one to seven to indicate the perceived overall quality of the deck. Additionally, we requested their opinion on the percentage of the research brief content they are willing to include in the client-facing pitch proposal for the RFP. They were also asked to assess the likelihood of the consulting firm's partner and potential client accepting the deck content. To evaluate the perceived labor exerted in content generation, we also asked the managers to indicate the number of hours they think the deck generation costs the two analysts.

We asked the managers if they thought the junior analysts used GPT for content generation after presenting the research briefs. This question was to 1) test whether managers could tell the use of GPT from their analysts when they do not disclose their use of GPT, and 2) serve as a manipulation check.We also asked about the managers' experience and familiarity with the RFP topic for each business case.

At the end of the first RFP evaluations, we inquired about their preferences regarding disclosing the use of GPT to their supervisor. We also asked about their preference for their analysts disclosing such information and for which section of the business content they will likely allow the junior analysts to use GPT. We also collected the managers' demographic information, including their gender, role in the consulting firm, average working hours per week, tenure as a consultant, and prior use of GPT for work.

The study was carried out anonymously to guarantee truth-telling.

Table 1 displays the summary statistics of the variables used in the main analysis.

	Mean	Std Dev	Min	Max	Observations
Engaging	4.037	1.37	1	7	108
PercentContentIncorporate	55.90	20.8	5	100	108
AcceptanceChancePartner	45.86	21.6	5	88	108
AcceptanceChanceClient	46.06	22.8	1	90	108
DeckGeneratingHours	14.60	19.7	1	100	108
DeckPresentOrder	1.500	0.50	1	2	108
Gender	1.593	0.50	1	2	27
UsedGPT	0.148	0.36	0	1	27
AvgWeekWorkHour	47.33	5.41	40	60	27
TenureConsultant	5.078	2.63	0	8.50	27
ExpRFP	2.167	1.21	1	5	54
FamiliarityRFPTopic	1.611	0.96	1	4	54
Number	4.609	0.85	3.62	5.95	8
Lifestyle	13.64	1.86	11.2	16.3	8

Table 1: Summary statistics of the survey experiment variables

4.3. Results

Twenty-seven out of 45 managers successfully completed both training sub-sessions. Among the 27 managers, 16 were informed about the source of the content generation, while 11 were unaware. We conducted a binomial test on the between-subject assignment outcome and found no significant difference in the frequency of managers assigned to the source

[2]We acknowledge the potential selection bias stemming from the self-selection in managers' signing up process for this training session. Regarding the sample size, we planned to collect the second wave of data on June 19, 2023, but we couldn't implement the identical experiment since the consulting firm suggested a different design.

transparency condition with 0.5 ($p = 0.44$).

We conduct a randomization check on the manager-level variables (i.e., Gender, UsedGPT, AvgWeekWorkHour, and TenureConsultant) according to our between-subject design and find no significant differences in those features ($ps > 0.24$ given the two-sample Student's t-test).

4.3.1. Prior and post beliefs in junior analysts' use of GPT We first find no difference in the prior belief about the junior analyst using GPT in generating the deck (Mean_transparency = 39.82%, Mean_no transparency = 51.81%, p = 0.24 given the two-sample Student's t-test).

The managers were unable to distinguish whether the content was created using GPT or not when the content generation source was not revealed. In the no transparency condition, both the Human-GPT and No-GPT versions were believed to be generated with GPT 77.27% of the time ($p = 1$ given a Pearson's Chi-squared test), even though only 50% of them were generated with Human-GPT collaboration.

Different from the no transparency condition, we find managers increase their belief in junior analysts using GPT to generate the Human-GPT than the No-GPT deck under the transparency condition: managers indicate their belief in analysts using GPT 75.99% of the time for the Human-GPT version compared to 40.63% for the No-GPT version ($p < 0.01$ given the Pearson's Chi-squared test). The results suggest success in the between-subject manipulation in disclosing the content generation source.

4.3.2. Main results In this section, we investigate the managers' evaluation of the research brief content generated with and without using GPT, conditional on whether the content generation source is disclosed.

We denote manager i's evaluation of deck version v ($v \in \{No-GPT, Human-GPT\}$) generated for RFP c ($c \in \{1, 2, 3, 4\}$) as $Y_{i,c,v}$, whose specification in shown in equation (1). Our focal explanatory variables are "Transparency" (coded as 1 for the transparency condition and 0 for the no transparency condition), and "HumanGPT" (coded as 1 for the Human-GPT deck and 0 for the No-GPT deck).

$$\begin{aligned} Y_{i,c,v} = {} & \beta_0 + \beta_1 \times Transparency_i + \\ & \beta_2 \times HumanGPT_{c,v} + \\ & \beta_3 \times Transparency_i \times HumanGPT_{c,v} + \\ & \eta_c + \gamma_1 \times X_i + \\ & \gamma_2 \times X_{i,c} + \gamma_3 \times X_{c,v} + \\ & \gamma_4 \times X_{i,c,v} + \\ & \epsilon_{i,c,v} \end{aligned} \tag{1}$$

We include RFP-fixed effects to account for any observed and unobserved heterogeneity in the evaluation stemming from the time-invariant features on the RFP case level. We account for manager-level heterogeneity by including manager-level variables X_i, compromising the manager's gender, whether used GPT before for work, average working hours per week, tenure as a consultant in years, and their internal role. We control the impact of the manager-RFP level variation on the deck evaluation by including manager-RFP-level variables $X_{i,c}$, compromising the experience in managing a similar case and familiarity with the RFP topic. We also include RFP-version-level variables $X_{c,v}$ compromising LIWC output with the word count of the deck, the number of words with more than six letters, and the word frequency indicating lifestyle.We include the order of evaluation for the Human-GPT and No-GPT versions, which varies on the manager-RFP-version level.

Table 2 reports the ordinary least square (OLS) regression results using different decision variables related to managers' evaluation of the deck content. We cluster the robust standard errors at the individual manager level for all analyses, allowing for arbitrary error correlations within each manager.

We first investigate whether the managers prefer the No-GPT or the Human-GPT content by looking at the managers' evaluation of how much they find the content engaging, the amount of content they want to incorporate into the client-facing pitch, and the likelihood the partner at the consulting firm and the client to accept the content. Managers believe that partners at the consulting firm are more likely to accept the Human-GPT than the No-GPT content when not knowing the content generation source ($p < 0.05$ for the coefficient estimates of β_2 in Column (3) of Table 2). We do not find the higher acceptance likelihood of the Human-GPT over the No-GPT content moderated by the disclosure of the content generation source (p=0.98 for the coefficient estimates of β_3 in Column (3) of Table 2).

We next examine whether revealing content generation sources affects managers' preference for content. We discovered that managers favor both the Human-GPT and No-GPT content when knowing the content generation source by indicating a higher proportion of content to incorporate into the client-facing pitch ($p < 0.05$ for the coefficient estimates of β_1 in Column (2) of Table 2; $p < 0.05$ for the F-test on the linear combination of coefficient

estimates of $\beta_1 + \beta_3$ in Column (2) of Table 2).

We also check the managers' perceived number of hours requested to generate the content. We find that knowing the two junior analysts using GPT significantly reduces managers' perceived number of hours required to generate the Human-GPT than the No-GPT deck ($p < 0.001$ for the coefficient estimates of β_3 shown in Column (5) of Table 2).

Table 2: Managers' evaluation of the No-GPT and Human-GPT deck with and without disclosing the content generation source – OLS regression results

	Engaging	Content to incorporate	Accpet by partner	Accpet by client	$\log(hours)$
	(1)	(2)	(3)	(4)	(5)
Transparency	0.634*	13.374**	13.542*	11.163	0.199
	(0.368)	(6.053)	(6.950)	(8.118)	(0.328)
Human-GPT	1.745*	-7.970	9.000**	4.806	0.422*
	(1.019)	(7.083)	(3.909)	(4.902)	(0.235)
Transparency × Human-GPT	-0.548	-0.814	-0.041	2.594	-0.307**
	(0.426)	(4.517)	(4.580)	(5.271)	(0.111)
Observations	108	108	108	108	108
Controls	Yes	Yes	Yes	Yes	Yes
Case fixed-effects	Yes	Yes	Yes	Yes	Yes

Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Controls include Gender, UsedGPT, AvgWeekHour, TenureConsultant, InternalRole, ExpRFP, FamiliarityRFPTopic, Order, Number, and Lifestyle.

4.3.3. The preference for disclosing the use of GPT in generating business content While disclosing the use of GPT affects managers' evaluation of the Human-GPT content, we further investigate whether managers would require their analysts to disclose this information and whether they would disclose the use of GPT to their supervisor (both rated on a scale from one to five). We regressed the preference to disclose the use of GPT against whether it is related to the managers' own use (vs. the analysts' use), which is interacted by the transparency condition. We include individual-level fixed effects to account for any observed and unobserved heterogeneity on the individual level. Table 3 shows the results, suggesting that managers strongly prefer their analysts to reveal their use of GPT (with an average rating of 4.3 out of 5) but are more hesitant to reveal such information to their supervisor (see the negative statistically significant coefficient estimates of "Own use" in Column (1) of Table 3). The disclosure of the content generation source does not affect this finding.

4.3.4. How do the managers allow the analysts to use GPT? We further investigate the potential policy the managers would like to impose on their analysts in generating business content. In particular, we asked managers' likelihood of allowing junior analysts to use GPT to prepare the content for each deck's section (problem statement, solution overview, implementation details, expected outcomes, team, and timeline). We regress the likelihood of allowing GPT against whether the manager evaluated content when its generation source is revealed. The analysis also includes the content section fixed effects and the manager-level control variables with robust standard errors clustered at the manager level. Table 4 Column (1) shows the results. After being informed about the source of content generation, managers tend to increase their likelihood of permitting their analysts to use GPT. We further sequentially include the degree of content modification in the Human-GPT deck (vs. No-GPT deck) and the degree of content contribution from GPT in the analysis to understand whether the preference to allow the use of GPT is impacted by the actual content modifications. The results are shown in Columns (2) and (3) of Table 4. We find no significant correlations between the extent to which the content modification or GPT contribution and the managers' willingness to let them use GPT. We further include the analysts' actual frequency of using GPT in the analysis. The results are shown in Column (4) of Table 4. Again, we find no statistically significant correlations between the frequency of prompts used by junior analysts and the managers' willingness to let them use GPT.

Table 3: Managers strongly prefer their analysts to reveal their use of GPT but are hesitant to reveal such information to their supervisors – OLS results

	Disclosing use of GPT
Own use	-0.727**
	(0.317)
Own use × Transparency	0.040
	(0.412)
Constant	4.296***
	(0.143)
Observations	54
Individual fixed-effects	Yes

Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

5. Conclusion

Our study contributes to the literature by showing that managers could not distinguish content generated by the analysts with and without the aid of GPT. To our knowledge, we are among the first to show a potential positive shift in the likelihood of content generated by the human-GPT collaboration to disseminate when knowing the use of GPT in a professional context.

Our results have direct implications for the disclosure of the use of GPT at work. First, we find that managers generally suspect business content pushed to them is generated using GPT when the content generation source is not disclosed (77.27%) than when the information is disclosed (57.81%). Interestingly, despite an honest disclosure of junior analysts not

Table 4: Managers strongly prefer their analysts to reveal their use of GPT but are hesitant to reveal such information to their supervisors – OLS results

	(1)	(2)	(3)	(4)
Transparency	1.104*	1.102*	1.101*	1.117*
	(0.569)	(0.569)	(0.570)	(0.573)
Degree of modification		0.155	0.296	0.773
		(0.697)	(0.669)	(0.577)
Degree of GPT contribution			-1.425	-1.777
			(3.714)	(5.159)
# of summarizing prompts				-0.092
				(0.196)
# of inferring prompts				-0.088
				(0.056)
# of expanding prompts				-0.042
				(0.145)
# of transforming prompts				0.097
				(0.124)
Observations	162	162	162	162
Controls	Yes	Yes	Yes	Yes
Content section fixed-effects	Yes	Yes	Yes	Yes

Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Controls include Gender, UsedGPT, AvgWeekHour, TenureConsultant, and InternalRole.

using GPT, the managers still think that 40.63% of the time, the two junior analysts used GPT. This finding highlights the prior belief in the popularity of GPT in participating in content generation in the professional context. In general, the managers strongly prefer their analysts to disclose the use of GPT. However, there may be misaligned preferences in how the GPT tool should be used between the managers and junior analysts.

Our study is preliminary. Several limitations in our study provide opportunities for future research. First, we would like to conduct a controlled online experiment to enlarge our sample size, reduce selection bias stemming from attrition, and understand the mechanism underlying to enhance the theoretical contribution of our findings. Second, different ways to increase content generation transparency could be adopted to understand alternative ways to alter humans' perception of human-GPT-generated content (e.g., presenting how the analysts use GPT and the usage frequency). Third, we would like to test how humans evaluate human-GPT-generated content that is not necessarily adapted from human-generated content and business content beyond the four RFPs. Fourth, future research could test how expertise moderates the GPT use and the downstream consequences in content evaluation. Notwithstanding these limitations, we believe our study is a useful first step in documenting the impact of disclosing the GPT on content evaluation and corporate policy implementation in a business field setting.

References

Acemoglu, D., Autor, D., Hazell, J., & Restrepo, P. (2022). Artificial intelligence and jobs: Evidence from online vacancies. *Journal of Labor Economics*, *40*(S1), S293–S340.

Austin & Wright. (2023). *Jdsupra, italy's data protection agency lifts ban on chatgpt*. https://www.jdsupra.com/legalnews/italy-s-data-protection-agency-lifts-3596533/

Berk, R. (2017). An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *Journal of Experimental Criminology*, *13*, 193–216.

Boyd, R., Ashokkumar, A., Seraj, S., & Pennebaker, J. (2022). Liwc-22: Descriptive statistics and norms. *Retrieved March*, *29*, 2022.

Brand, J., Israeli, A., & Ngwe, D. (2023). Using gpt for market research. *Available at SSRN 4395751*.

Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). *Generative ai at work* (tech. rep.). National Bureau of Economic Research.

Cao, S., Jiang, W., Wang, J. L., & Yang, B. (2021). *From man vs. machine to man+ machine: The art and ai of stock analyses* (tech. rep.). National Bureau of Economic Research.

Castelo, N., Bos, M. W., & Lehmann, D. (2019). Let the machine decide: When consumers trust or distrust algorithms. *NIM Marketing Intelligence Review*, *11*(2), 24–29.

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809–825.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114.

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*.

European Commission, E. (2023). *Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act)*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

Falkner, A., Palomares, C., Franch, X., Schenner, G., Aznar, P., & Schoerghuber, A. (2019). Identifying requirements in requests for proposal: A research preview. *Requirements Engineering: Foundation for Software Quality: 25th International Working Conference, REFSQ 2019, Essen, Germany, March 18–21, 2019, Proceedings 25*, 176–182.

Felten, E., Raj, M., & Seamans, R. (2023). How will language modelers like chatgpt affect occupations and industries? *arXiv preprint arXiv:2303.01157*.

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2022). Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. *Information Systems Research*, *33*(2), 678–696.

Garante, P. L. P. D. D. P. (2023). *Intelligenza artificiale, dal garante privacy stop al chatbot "replika". troppi i rischi per i minori e le persone emotivamente fragili*. https://garanteprivacy.it/home/docweb/-/docweb-display/docweb/9852506

Gottschalg, O., & Zollo, M. (2007). Interest alignment and competitive advantage. *Academy of management review*, *32*(2), 418–437.

Gunaratne, J., Zalmanson, L., & Nov, O. (2018). The persuasive power of algorithmic and crowdsourced advice. *Journal of Management Information Systems*, *35*(4), 1092–1120.

Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? *arXiv preprint arXiv:2301.07543*.

Hu, K. (2023). *Reuters, chatgpt sets record for fastest-growing user base - analyst not*. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Khowaja, S. A., Khuwaja, P., & Dev, K. (2023). Chatgpt needs spade (sustainability, privacy, digital divide, and ethics) evaluation: A review.

Kuhnen, C. M., & Niessen, A. (2012). Public opinion and executive compensation. *Management Science*, *58*(7), 1249–1272.

Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S. J., & Doshi-Velez, F. (2019). Human evaluation of models built for interpretability. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *7*, 59–67.

Lebovitz, S., Lifshitz-Assaf, H., & Levina, N. (2022). To engage or not to engage with ai for critical judgments: How professionals deal with opacity when using ai for medical diagnosis. *Organization science*, *33*(1), 126–148.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103.

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, *46*(4), 629–650.

Ng, A., & Fulford, I. (2023). Deeplearningai: Chatgpt prompt engineering for developers [Accessed: 2023-06-14].

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, *381*(6654), 187–192. https://doi.org/10.1126/science.adh2586

OpenAI. (2023). Usage policies [[Accessed 14-Jun-2023]].

Peng, Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., Huang, Q., Liden, L., Yu, Z., Chen, W., et al. (2023). Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Peng, Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). The impact of ai on developer productivity: Evidence from github copilot. *arXiv preprint arXiv:2302.06590*.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–52.

Reisenbichler, M., Reutterer, T., Schweidel, D. A., & Dan, D. (2022). Frontiers: Supporting content marketing with natural language generation. *Marketing Science*, *41*(3), 441–452.

Weiner, B., Frieze, I., Kukla, A., Reed, L., Rest, S., & Rosenbaum, R. M. (1987). Perceiving the causes of success and failure. *Preparation of this paper grew out of a workshop on attribution theory held at University of California, Los Angeles, Aug 1969*.

You, S., Yang, C. L., & Li, X. (2022). Algorithmic versus human advice: Does presenting prediction performance matter for algorithm appreciation? *Journal of Management Information Systems*, *39*(2), 336–365.