

A Two-Phased AI-Enabled Framework for Innovation with User-Generated Data from Consumer Review Sites

Kristijan Mirkovski
Deakin University
k.mirkovski@deakin.edu.au

Jingda Kang
University of Melbourne
jingdak@student.unimelb.edu.au

Libo Liu
University of Melbourne
libo.liu@unimelb.edu.au

Marta Indulska
University of Queensland
m.indulska@business.uq.edu.au

Hao Liu
Deakin University
hao.liu@deakin.edu.au

Abstract

User-generated data from consumer review sites holds immense potential for driving product innovation, yet actionable research in this area remains limited. This paper addresses this gap by proposing a two-phased AI-enabled framework for leveraging consumer reviews throughout the idea selection and generation processes. Our framework utilizes advanced AI approaches to automate idea selection and generation in open innovation settings. The proposed framework aims to extract product innovation ideas from raw online reviews, employing cutting-edge machine learning for natural language processing and generative pre-trained transformers for natural language generation. This paper offers a novel AI-enabled approach for organizations to drive open innovation and improve product development processes.

Keywords: Idea Selection, Idea Generation, User-Generated Data, AI-enabled Innovation.

1. Introduction

Open innovation is a “distributed innovation process based on purposefully managed knowledge flows across organizational boundaries” (Chesbrough & Bogers, 2014, p. 12). Consumer review sites generate large amounts of detailed, nuanced, and disaggregated user-generated data (UGD), which, if properly, managed can support businesses’ open innovation initiatives (Goldberg & Abrahams, 2022). A unique attribute of this data is its timeliness, which can help businesses gain a competitive advantage over their rivals with greater speed-to-market of new products or services. Thus, UGD from consumer review sites represents a valuable source of information for businesses seeking bottom-up

participation or external engagement to innovate their products (Goldberg & Abrahams, 2022).

According to a recent report, 98% of consumers rely on online reviews from Amazon Customer Reviews, Yelp, Angie’s List, Trustpilot, and Facebook when deciding to purchase a local product or use a service (Paget, 2023). Thus, it is widely acknowledged that product reviews and rankings significantly influence consumer purchasing decisions (Derakhshan et al., 2022). UGD from consumer review sites, including written reviews, ratings, comments, likes/dislikes, and others, has been recognized as one of the most important components of information that shapes current customers’ behaviors and decisions (Jia & Liu, 2018). The richness of this data creates a fertile ground for businesses to capture many-to-many interactions and engage external stakeholders, which can result in product, service, or process innovation in a cost-efficient manner (Goldberg & Abrahams, 2022). A recent study indicates that UGD is closely linked to data-driven innovation where organizations use data analytics approaches to identify patterns in large datasets of user actions and to improve their decision-making (Saura et al., 2021).

Open innovation through the utilization of UGD is regarded as a transformative strategy for organizations seeking to enhance their products or services (Chesbrough & Bogers, 2014). However, the abundance of UGD, which includes both valuable contributions and irrelevant product reviews, comments, and spam, imposes significant demands on resources and time on the idea selection and generation processes of open innovation. Chesbrough (2003) emphasizes the importance of sifting through such data to identify promising ideas, but this task is inherently labor-intensive. Recent advancements in AI, including machine learning (ML) and large language models (LLMs) (Füller et al., 2022), offer a promising solution for automating the idea selection

and generation processes of open innovation, which in turn can help organizations harness the power of UGD more efficiently, enabling them to drive innovation and stay competitive in today's dynamic business environment.

Despite the significant advancements in using ML to automate idea selection from UGD, there remain substantial gaps in the field. Current research focuses on unsupervised and supervised learning techniques to analyze extensive data sources, such as online reviews and social media posts, yet these learning approaches have limitations when applied to product innovation tasks. Unsupervised learning is effective at identifying key themes and sentiments but faces challenges in result validation due to the absence of definitive benchmarks, complicating the accurate selection of valuable ideas for innovation (Suominen et al., 2017). Conversely, supervised learning, though superior in performance to unsupervised methods, is heavily dependent on large, often unavailable, labeled datasets and grapples with defining what constitutes a "successful" idea, which complicates model training and diminishes the practical utility of the outputs, potentially overlooking truly innovative ideas that deviate from established patterns (Yang et al., 2022). These challenges underscore the need for a more structured data analysis approach, potentially incorporating semi-supervised learning to enhance precision (Najafabadi et al., 2015).

Emerging research highlights the potential of integrating LLMs in idea generation, emphasizing their capability to significantly improve both the quantity and quality of generated ideas by leveraging their capacity to understand and generate human-like text from extensive data (Bouschery et al., 2023). Nevertheless, empirical research on effectively integrating LLMs into product innovation management remains nascent, signaling a pressing need for research that explores theoretical applications and builds robust methodologies for their implementation. Building on the work of Mirkovski, von Briel, and Lowry's (2016), we propose a two-phased automated process framework that offers innovative ideas from a pool of raw UGD for a target product. We explore the use and integration of cutting-edge ML approaches for natural language processing and state-of-the-art LLMs in the process of natural language generation to select and generate product innovation ideas from consumer product reviews. In this paper, we focus on *incremental product innovation*, which involves small, continuous improvements to existing products, rather than breakthrough innovation, which refers to radical, transformative changes that create entirely new markets or significantly alter existing ones (Garcia &

Calantone, 2002). Thus, we aim to answer the following research question: *How to integrate advanced ML and LLMs for natural language processing and generating to automate idea selection and generation for product innovation?*

This paper advances the existing body of work on the use of UGD for product innovation management by developing a comprehensive framework that automates the idea selection and generation processes through the integration of advanced ML and LLM algorithms. Automating the process of idea selection and generation with data from consumer review sites, reduces the time and resources required for this crucial step, allowing organizations to explore a wider range of innovative concepts. By using online product reviews as a source of data, our paper contributes to a deeper understanding of customer sentiments and preferences, which in turn can help organizations improve their product offerings to align more closely with customer needs and expectations.

The paper is organized as follows: We review related work on open innovation and UGD and the latest natural language processing and natural language generation approaches for idea selection and generation. We then present our two-phased automated process framework for idea selection and generation with user-generated product review data. Last, we conclude by outlining future research.

2. Related Work

2.1. Open innovation with user-generated data

Innovation is central to making organizations competitive, creating new jobs, and maintaining a high standard of living (Chesbrough & Bogers, 2014). In a world where the exchange of information occurs constantly and freely, UGD is emerging as a major source of cost-efficient innovation (Wu et al., 2023). Based on the three open innovation stages: (1) idea selection and generation, (2) concept development, and (3) product commercialization, Mirkovski et al. (2016) developed a process-based framework that transforms UGD into valuable business insights for the entire open innovation process. In this paper, we focus on the idea selection and generation stage for which we design a novel AI-based approach to harness open innovation based on a pool of online product reviews.

Idea selection, in the context of open innovation, refers to the process of evaluating and selecting the most promising ideas from a pool of contributions, typically sourced from a diverse and external set of participants (King & Lakhani, 2013). This process is crucial for harnessing the wide range of possibilities

that open innovation can offer and involves distinguishing high-potential ideas for further development. It acts as a filter, deciding which ideas should move forward in the innovation pipeline (Hossain & Islam, 2015). A multinational consumer electronics firm, for example, might initiate a global call for innovative smartphone features and receive a vast number of submissions, encompassing ideas ranging from novel user interface designs to cutting-edge energy efficiency technologies. Through a selection process, which may include multi-criteria decision analysis and expert evaluations, the organization identifies those ideas that most closely align with its strategic objectives and market potential. These selected ideas are then advanced to the prototyping stage, with the ultimate goal of integrating them into future product offerings.

Idea generation, in the context of open innovation, involves the creative process of generating, developing, and communicating new ideas from both internal and external sources of an organization. According to Baregheh et al. (2009), idea generation is “the process of creating, developing, and communicating ideas which are abstract, concrete, or visual” (p. 1334). A pharmaceutical company, for instance, might crowdsource ideas from researchers and healthcare professionals to discover novel applications for an existing drug. For example, a drug initially developed to treat a specific type of cancer could have untapped potential in treating other conditions, such as autoimmune diseases or viral infections. By engaging a global network of medical researchers, clinicians, and pharmacologists through open innovation platforms, the organization can gather diverse insights and hypotheses that may not have emerged through internal research alone. External idea generation brings in fresh perspectives, which can catalyze breakthrough innovations and create competitive advantages (Goldberg & Abrahams, 2022). By incorporating diverse viewpoints, this process enriches the innovation pipeline, thereby enhancing creativity and potentially accelerating the time-to-market for new products or improvements (Derakhshan et al., 2022).

2.2. Natural Language Processing for Idea Selection

Word embeddings play a critical role in text mining by converting text into numerical representations that can be easily processed by ML models. This capability is particularly valuable for automating the extraction of product innovation ideas from vast amounts of UGD. Traditional models, such as Word2Vec (Mikolov et al., 2013) and GloVe

(Pennington et al., 2014), represent words as vectors in a high-dimensional space, capturing semantics and syntactic relationships (Beltagy et al., 2019). For example, semantically similar words (“apple” and “pear”) share similar vector representations, while antonyms (“good” and “bad”) are positioned in opposite directions within this space.

In recent years, contextual word embeddings have emerged as a valuable tool for capturing contextual meanings of words, especially polysemous words, and outperform traditional word embeddings in many tasks (Miaschi & Dell’Orletta, 2020). Traditional embeddings fail to distinguish between the different meanings of a word like “bank” in sentences such as “The man just robbed a bank” versus “The man was fishing by the bank of a river”. Contextual embeddings address this challenge by providing context-dependent representations for words. BERT, XLNet, and ELMo are three cutting-edge word embedding models that we are going to explore in our paper. The amalgamation of multiple contextualized embeddings has been proven to be more effective in text-mining tasks (Zhai et al., 2019).

Bidirectional Encoder Representations from Transformers (BERT) is a bidirectional transformer-based model that was pre-trained on large corpora of text (Devlin et al., 2019), which considers both bidirectional contexts of each word within a sentence. The transformer architecture is a novel neural network model that leverages self-attention mechanisms (Honnibal, 2017) to process sequences of input tokens, allowing the transformer to be parallelized and to process input sequences efficiently. BERT word embeddings are highly effective for a wide range of downstream tasks such as classification, entity recognition, and question answering (Tanaka et al., 2020). *XLNet* is a newer model that was developed to address some limitations of BERT using an autoregressive (AR) approach and autoencoding (AE) techniques (Yang et al., 2019). *ELMo* is a contextualized word embedding model, that generates word embeddings based on a bidirectional language model (biLM) and character-level Convolutional Neural Networks (CNN) (Peters et al., 2018).

To conclude, word embeddings, such as BERT, XLNet, and ELMo, are effective for the process of idea selection from UGD due to their sophisticated capacity to capture contextual nuances in language. These models surpass traditional word embeddings by providing context-sensitive representations, enabling them to accurately distinguish between different meanings of polysemous words based on their specific usage within a given context. This capability is essential for effectively interpreting and categorizing the diverse and often complex ideas present in UGD.

2.3. Natural Language Generation for Idea Generation

LLMs have become increasingly popular in natural language generation tasks due to their ability to understand context, generate coherent and contextually relevant responses, and summarize large amounts of text, which can be useful for automating the development of a product innovation idea from a large dataset of user-generated reviews. GPT models are a series of LLMs developed by OpenAI (Brown, 2020), which are used for a variety of language-related tasks due to their ability to generate coherent language. *GPT-3.5* introduced large models with up to 175 billion parameters, which have been widely used in various natural language processing and generating tasks, such as language translation, question answering, and text completion (Toews, 2022), and even creative applications, such as poetry and fiction writing (Floridi & Chiriatti, 2020). Due to its few-shot learning capability (Brown, 2020), GPT-3.5 can produce relevant and innovative ideas in response to a problem statement, even with limited examples of typical brainstorming outcomes provided as input prompts.

GPT-4 is the latest and most advanced iteration of GPT models, designed to generate text with better accuracy and understanding (Girotra et al., 2023). Haase and Hanel (2023) conducted a study to compare ideas generated by humans with those produced by Generative Artificial Intelligence (GAI) chatbots, including GPT-4, in terms of creativity and quality. The results indicate that there is no qualitative difference between AI and human-generated creativity. In addition, only 9.4% of humans, on average, were found to be more creative than the most creative GAI – GPT-4, across 5 categories of topics from the Alternative Uses Test (AUT) in real life. Girotra et al. (2023) specially researched the innovation ideation capabilities of GPT-4 with students enrolled in product design courses from Wharton, Cornell Tech, and INSEAD in terms of productivity and quality separately. In every match-up, GPT-4 came out ahead. GPT-4 was utilized to compete with students in generating product design ideas, comprising a descriptive title and a paragraph of text to elaborate. They also indicate that GPT-4 can generate innovative ideas much faster and cheaper than humans; a human using GPT-4 is about 40 times more productive than a human working alone.

In conclusion, LLMs, such as GPT-3.5 and GPT-4, are effective tools for generating ideas from UGD due to their comprehensive contextual understanding and ability to produce coherent, contextually appropriate responses. These models excel in

processing and synthesizing large volumes of text, making them particularly valuable for extracting innovative ideas from extensive datasets, including user reviews. The few-shot learning capabilities of GPT-3.5 allow for the generation of creative and relevant ideas even with minimal input, while GPT-4's enhanced precision and efficiency facilitate the rapid and cost-effective production of high-quality, innovative concepts, thereby augmenting the ideation process.

3. A Two-Phased Automated Process Framework for Idea Selection and Generation

Based on the review of the latest AI approaches for natural language processing and generating, we propose a novel process framework for (1) selecting product innovation ideas from product-centered online reviews and (2) generating a set of product innovation recommendations for further product development (see Figure 1). Our framework represents actionable research that offers practical implications on how to deploy cutting-edge ML and LLMs approaches to analyze data from consumer review sites and derive innovation insights used in idea selection and generation processes. Our framework goes beyond simply extracting product improvement insights from online reviews; it (1) conducts a comprehensive analysis of the product and segments the outputs into different dimensions based on its intrinsic features to support the idea selection and (2) synthesizes the outputs into detailed recommendations about product innovation to assist with the idea generation.

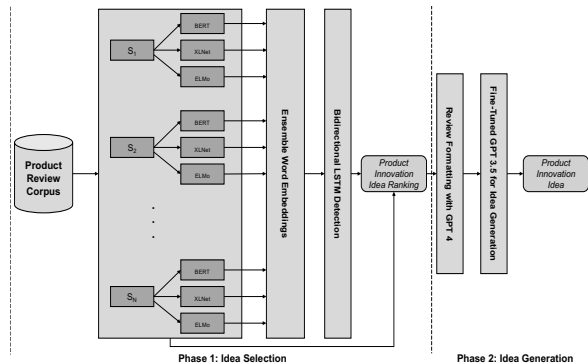


Figure 1. AI-enabled automation for idea selection and generation with UGD

3.1. Phase 1: Idea Selection

The aim of Phase 1 is to select innovative product sentences from reviews. After splitting reviews into

sentences, we first develop an ensembled contextual word embedding that combines three advanced contextualized word embeddings – BERT, XLNet, and ELMo to represent each textual sentence. Second, the word embeddings of each sentence are employed as inputs to a Bidirectional Long Short-Term Memory (Bi-LSTM) to model the sequential semantics and linguistics. Finally, with the help of the sigmoid function on the output of the Bi-LSTM, we classify each product sentence as “innovation-related” or not.

3.1.1. Ensemble of word embeddings. A combination of multiple word embedding is a beneficial approach for capturing different semantics and meanings of the text, which leverages the strengths of each embedding technique, potentially leading to a richer and more nuanced representation of the text (Wieting & Gimpel, 2017). Therefore, we ensemble word embeddings by concatenating three contextual word embeddings at the word level, including BERT, ELMo, and XLNet, due to their superior performance compared to traditional word embeddings. Prior research has shown that ensemble methods can significantly improve the overall quality of various word embedding models by combining the outputs of multiple word embedding models to create a more robust representation of symbolic words and linguistic features (Zhai et al., 2019). However, word-level alignment is a challenge when concatenating the outputs of different language models since BERT, XLNet, and ELMo use different tokenization methods and break the same sentence into different sequences of tokens. For instance, the word ‘embedding’ is a token in ELMo, but the word is split by WordPiece (BERT) tokenization into three separate tokens, ‘em’, ‘##bed’, and ‘##ding’. The ‘##’ symbol is a flag to indicate that is part of a word except for the first one. This WordPiece mechanism that BERT uses helps to scale down the vocabulary size and improve the generalizability of tokenization. To tackle this issue, we pool the embeddings at the word level from subwords by averaging their embedding vectors. For example, we use the averaged embedding of ‘em’, ‘##bed’, and ‘##ding’ to represent the word ‘embedding’ for BERT.

The aligning embedding algorithm is illustrated in *Algorithm 1* for BERT. The algorithm calculates the group of indices for subwords and reconstructs the embedding *PE* of the sentence as the output. XLNet uses a SentencePiece tokenizer, and ELMo uses a Moses tokenizer. Both tokenizers are based on subword splitting. Therefore, the same algorithmic approaches are applied to XLNet and ELMo to compute the pooled embeddings based on our tokenization. The only change is at line 6, when

locating the index of each subword, while the core idea remains consistent. The algorithms for XLNet and ELMo hence will not be reiterated here due to lack of space.

Algorithm 2 demonstrates the ensemble process in which BERT pooled embeddings, XLNet pooled embeddings and ELMo pooled embeddings are stacked to form our final ensemble embeddings. The algorithm concatenates three embeddings at the word level, increasing the dimensionality of the word vectors to accommodate more numerical information for each word. The final ensemble embeddings serve as the input data for the downstream innovation detection model.

Algorithm 1: Algorithm for pooled embedding

	Input: <i>text_list</i> : The preprocessed text <i>tokenized_text</i> : Tokenized text by word embedding model(s) <i>TE</i> : The embedding of the text
	Output: <i>PE</i> : The pooled embedding by averaging subword embeddings in groups
1	if <i>text_list.size()</i> = <i>tokenized_text.size()</i> then
2	<i>j</i> ← 0
3	<i>average_group_index</i> ← The list of indices group after embedding pooling
4	<i>average_group</i> ← The list of subwords group after embedding pooling
5	for each token <i>t</i> ∈ <i>tokenized_text</i> do
6	if <i>t</i> starts with ‘##’ then
7	append the index of <i>t</i> to <i>average_group_index</i>
8	end
9	if <i>text_list[j]</i> = ‘.’ <i>join(average_group[j])</i> then
10	<i>j</i> ← <i>j</i> + 1
11	end
12	end
13	for each index group <i>g</i> ∈ <i>average_group_index</i> do
14	if <i>g.size()</i> > 1 then
15	<i>average</i> ← The mean value of embeddings at indices within <i>g</i> on <i>TE</i>
16	update embedding value at index <i>g[0]</i> with <i>average</i>
17	<i>PE</i> ← concatenate new <i>TE</i> without averaged parts of the embedding list
18	end
19	end
20	return <i>PE</i>
21	end

Algorithm 2: Algorithm for stacked embedding

	Input: PE_{BERT} : BERT pooled embeddings PE_{XLNet} : XLNet pooled embeddings PE_{ELMo} : ELMo pooled embeddings
	Output: SE : The ensemble stacked embedding
1	for $i \leftarrow 0$ to $PE_{BERT}.size()$ do
2	for $j \leftarrow 0$ to $PE_{BERT}[i].size()$ do
3	$word_embedding \leftarrow$ Sum of $PE_{BERT}[i,j]$, $PE_{XLNet}[i,j]$ and $PE_{ELMo}[i,j]$
4	append $word_embedding$ to $sentence_embedding$
5	end
6	append $sentence_embedding$ to SE
7	end
8	return SE

3.1.2. Innovation detection model. After an investigation of different text classification algorithms, we employ a Bidirectional Long Short-Term Memory (Bi-LSTM) model to detect the innovation-related sentences from reviews. We selected this model because of its superior ability to capture contextual information in mid to long sentences to learn a better representation of a single word, compared to others. Bi-LSTM is an extension of LSTM that processes the input sequence in both forward and backward directions to capture contextual information (see Figure 2).

As illustrated in Figure 2, the Bi-LSTM architecture consists of two separate LSTMs. The input layer is fed with a sentence $S = \{w_1, w_2, \dots, w_n\}$, and we learn the three contextual word embeddings of each word w_i to obtain three sequences of embeddings: E_{BERT}, E_{XLNet} and E_{ELMo} . Then we ensemble the three embeddings by concatenating them into a single embedding vector at the word level. After ensembling word embeddings, the raw text can be represented by a sequence of vectors, denoted as V_S ($V_S = [E_{BERT}, E_{ELMo}, E_{XLNet}]$). The ensembled sequence of vectors V_S is then fed into the Bi-LSTM. For the output of the Bi-LSTM, we use binary classification (sigmoid function) to indicate whether the sentence is innovation-related or not. The assignment of innovation labels is derived from a combination of sentiment polarities and the presence of improvement-related content in the review text. The sentence is assigned as 1 (“innovation-related”) if its sentiment polarity is negative or it contains helpful idea, and 0 otherwise. The rationale behind this labeling scheme is that we consider both sentiment and improvement suggestions to be inspirational for product developers in the R&D process. The final

output of the Bi-LSTM, together with the ensemble word embeddings, will be the output of Phase 1 and be fed as input to Phase 2 to generate innovative ideas. The developed method leverages three deep-learning-based contextual embeddings at the word level and train a Bi-LSTM model to understand the semantic information of the product reviews. The ensemble word embedding has a dimension of $R^{768+1024+768}$ for each word vector representation, which allows the model to learn the optimal feature from multiple contextual embeddings and improve performance. It is important to note that ensemble embeddings require a substantial amount of memory during the training process.

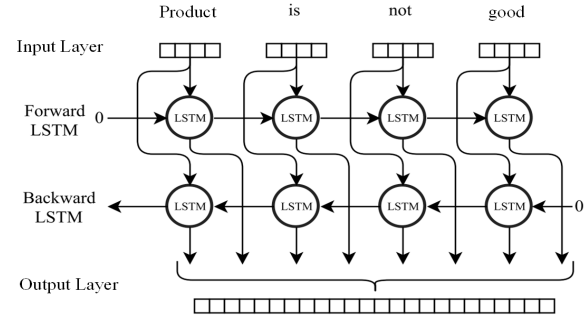


Figure 2. Bi-LSTM architecture for product review innovation detection, adapted from Cornegruta et al. (2016)

3.2. Phase 2: Idea Generation

In Phase 2, we propose approaches for review formatting and the GPT fine-tuning process. The design of prompts is a pivotal component in harnessing pre-trained GPT models to standardize selected innovation-related reviews, forming the training set for fine-tuning the innovation generation model. Following Phase 1, reviews selected from different categories are organized into clusters before reaching Phase 2. This ensures a diverse array of aggregated sources of innovation for training during the fine-tuning process and a multi-faceted innovation generation capability within the automation framework.

3.2.1. Review formatting. Before innovation generation, review formatting must be done to provide formatted completions for fine-tuning GPT models. This process extracts and summarizes the customer review selected in Phase 1 and formats the key ideas in a way that aligns with the perspective of developers. We utilize the most advanced GPT-4 model to help transform the selected reviews into formatted completions for the training process. To use the GPT-4 API, we need to design prompts for review

formatting. We design the system prompt to provide contextual task information and subsequent user prompts to feed the target reviews. The user prompts include the additional request that each formatted review should be limited to 30-60 words, similar to the original input reviews. The LLM model used in review formatting is the most advanced GPT-4-0613 available from June 13, 2023. The “temperature” parameter was set at 0.2 to achieve a focused and deterministic formatting. Below, we provide an example of a review before and after formatting:

Before: “The only downside to the game is there is not much sense of direction and would have been nice to have some markers beside the direction a site of grace is pointing towards.”

After: “The review suggesting a future improvement for “Elden Ring” could be enhancing player direction by adding markers beyond sites of grace. This open innovation could improve user experience by providing clearer guidance and reducing player confusion, thus refining the ARPG/Souls-like game’s navigation system.”

Due to the finite memory capacity of the GPT-4 model, the number of tokens for input and output is limited to a maximum of 8196 tokens. If the number of tokens in a session exceeds the model’s capacity limit, the LLM loses memory of the initial ideas it generated, and as a result, subsequent ideas may become progressively repetitive and redundant. Considering the maximum token limit, which encompasses the system prompt, user prompt, and the memory required for formatting, we asked GPT-4 to generate formatted reviews in batches according to the number of selected reviews. Overall, we standardized all innovation-related reviews in the same format to prepare for the subsequent fine-tuning process.

3.2.2. Fine-tuning GPT. Fine-tuning, in the context of the GPT model, refers to the process of adapting the pre-trained model to perform domain-specific tasks with a small dataset. GPT-3.5 Turbo, a variant of the GPT-3 model developed by OpenAI, is known for its impressive language generation capabilities. To learn and generate innovation based on the corpus of innovative product reviews selected in Phase 1, we fine-tune the GPT-3.5 Turbo in our second ideation phase. We then fine-tune the GPT-3.5 Turbo model on the formatted selected reviews. The fine-tuning hyperparameters in the experiment include *learning rate, temperature, batch size, number of epochs, top-k, top-p, penalty*, etc. By carefully considering these factors when fine-tuning the GPT-3.5 Turbo model, we optimize its performance and ensure that it is appropriate for the specific task at hand in the training and testing processes.

3.3. Evaluation

3.3.1. Innovation selection evaluation data. In this paper, we collected a dataset of around 10K reviews for the Steam game “Elden Ring” from 2022 to evaluate our method. 5000 reviews were randomly sampled from the original dataset and were segmented into individual sentences. After removing short sentences with less than 10 words, we randomly selected 6000 sentences, and two human annotators, who are players of the game, were recruited to identify innovation-related opinions in this dataset. Each sentence was assigned a value of 1 if it is innovation-related, and 0 otherwise. The annotators’ results demonstrated high consistency, with a Cohen’s Kappa of 0.98. Subsequently, we split the dataset into training (70%), validation (10%), and testing (20%) subsets. The innovation classification model was trained on the training set, with hyperparameter tuning (e.g., learning rate) performed on the validation set, and the final evaluation of the trained model was conducted on the testing set.

Metrics. We use the ROC curve (AUC), Precision, Recall, and F1 score as our metrics, which are widely used to evaluate the performance of binary classification.

Baselines. We compared the performance of the Bi-LSTM model with four baseline models: K-Nearest Neighbors (KNN), Naive Bayes, Support Vector Machine (SVM), and RoBERTa (Liu et al., 2019).

Results. The comparison results are presented in Table 1. Overall, the Bi-LSTM model demonstrates superior performance in innovation classification using ensemble word embeddings compared to other classifiers, achieving the highest AUC score. This indicates that Bi-LSTM is a suitable classifier for distinguishing between innovative and non-innovative content, exhibiting strong discriminatory capability.

RoBERTa also performs well, with competitive AUC scores. In contrast, the other three models struggle to differentiate innovative reviews based on ensemble word embeddings. Specifically, KNN and Naive Bayes consistently underperform relative to the other models.

Table 1: Results of innovation detection

Model	AUC	Precision	Recall	F1
KNN	0.59	0.48	0.65	0.55
Naïve Bayes	0.58	0.43	0.61	0.50
SVM	0.69	0.52	0.73	0.60
RoBERTa	0.78	0.63	0.76	0.69
Bi-LSTM	0.84	0.80	0.72	0.76

3.3.2. Idea generation evaluation. To evaluate the efficacy of fine-tuning GPT-3.5 Turbo for open innovation generation, we compare the performance of our framework with the original GPT-3.5 Turbo and GPT-4 models, both without fine-tuning on innovation reviews. The designed prompts are listed as follows:

System prompt. “You are an experienced ARPG and Souls-like game expert who is creative, practical, and innovative. Your job is to assist the game R&D process by providing innovative ideas.”

User prompt. “Please provide some innovative ideas for ARPG and Souls-like games from a developer’s perspective in the R&D process in terms of game future improvement and development. You should only generate innovative ideas with less than 60 words.”

Two experienced game designers were hired to assess 300 of the generated innovative ideas. Evaluators provide judgments regarding the relevance, creativity, and usefulness of the generated ideas, assigning scores ranging from 0 to 5. The scores obtained from the evaluators are then averaged to provide comprehensive quantitative metrics of the innovation’s potential. The evaluation results are

illustrated in Figure 3. Our framework achieves higher scores in terms of maximum, minimum, and median, surpassing the GPT-3.5 and GPT-4 models. Besides, we present five examples of the generated innovative ideas in Table 2. Our fine-tuned GPT model generates more informative and constructive ideas compared to the other two models. We will enlist domain experts to rigorously evaluate the output of our approach and the other benchmarks in future work.

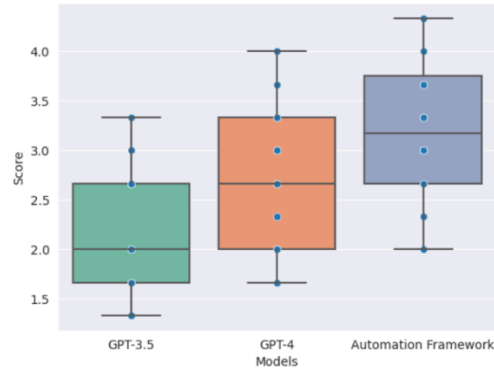


Figure 3. Evaluation results of idea generation.

Table 2: Examples of innovation generation

	GPT-3.5 Turbo	GPT-4	Our fine-tuned model
1	Seamless multiplayer integration, allowing for cooperative or competitive play in a shared game world.	<i>Dynamic Wildlife Ecosystem:</i> Implement a dynamic wildlife ecosystem. Creatures interact, hunt, reproduce or migrate depending on time, weather, and player actions.	The unlocking of 1080p fidelity in Elden Ring shows the potential for future improvements in graphics optimization, enhancing player experience. Integrating this innovation into the R&D process presents an opportunity to develop more efficient rendering techniques, ensuring high visual quality even at lower resolutions.
2	Create a unique character creation system that offers a wide range of customization options, allowing players to truly make their own distinct and memorable characters.	<i>Adaptive Difficulty:</i> A mechanic that adjusts game difficulty based on player skill level, offering a balanced challenge to both new and experienced players.	To fuel continuous innovation in the ARPG game, future development should focus on enhancing open-world interactivity. Improvements like freely moving turrets, variable combat terrains, interactive environmental objects, and mob-friendly zones could deepen gameplay immersion and strategic decision-making, creating a more engaging and memorable player experience.
3	Non-linear storytelling with multiple branching paths and endings, driven by player choices and actions.	<i>Procedural World:</i> Implement a unique, procedurally generated world for each new playthrough. This will greatly enhance replay value, ensuring a fresh environment for each player's journey.	For future improvements, maintaining Elden Ring's “pay it forward” spirit amidst its rich mechanics is crucial. Enhancements could focus on refining player summons through the blue tearstones or the magical swordsman/spellsword NPC companions. Open innovation might involve crowdsourcing ideas to expand on these features, promoting player engagement and creativity, resulting in a more immersive and collaborative gaming experience.
4	Implement a dynamic and reactive world that responds to player's actions, such as shifting	<i>Advanced Storytelling:</i> Develop a branching narrative system that changes based on player choices, providing	The current version of “Elden Ring” only allows for iron arrows to be used in the forge. Improvements in future versions could include the option to forge elemental arrows such as fire or lightning, which

	environments, weather effects, and NPC behavior influenced by player choices.	multiple endings which greatly impact game replayability.	would enhance gameplay by diversifying combat tactics. Open innovation could aid in the sourcing of unique ideas to expand the range of forgeable items and their associated game mechanics.
5	Procedurally generated dungeons and levels, ensuring endless replayability and surprising challenges.	<i>Diablo-like Loot</i> : Implement a loot system akin to Diablo, where defeating bosses or completing tasks offers unique rewards, improving the feeling of accomplishments and enhancing game longevity.	The battle designs in “Elden Ring” are strategically challenging, necessitating precise movement and timing. Future improvements could involve incorporating more unpredictable enemy movements to increase adaptable player skill and focus. Open innovation could include seeking insights from the experienced player community to refine these battle mechanics further.

4. Conclusions and Future Work

Future research will concentrate on conducting rigorous experiments to validate the efficacy and functionality of our innovative approach. Domain experts, including game designers and technology retailers, will be engaged in evaluating the outcomes at each phase of the proposed process model. To ensure the generalizability of the classification model across various contexts, we will collect and analyze diverse datasets from Amazon product reviews. This will include datasets specifically focused on headphones, headsets, and other relevant product categories. We will also focus on both positive and negative Amazon product reviews. By incorporating these varied datasets, we aim to ensure that our model maintains consistent performance across different domains. During the idea generation phase, we will conduct comparative experiments involving our approach alongside the original GPT-3.5 Turbo and GPT-4 models. The outputs generated by these models will be critically assessed by domain experts to validate the effectiveness and practical applicability of our process-based framework.

5. References

Baregheh, A., Rowley, J., & Sambrook, S. (2009). Towards a multidisciplinary definition of innovation. *Management Decision*, 47(8), 1323-1339.

Beltagy, I., Lo, K., & Cohan, A. (2019). *SciBERT: A Pretrained Language Model for Scientific Text*. <https://doi.org/10.18653/v1/D19-1371>

Bouschery, S. G., Blazevic, V., & Piller, F. T. (2023). Augmenting human innovation teams with artificial intelligence: Exploring transformer-based language models. *Journal of Product Innovation Management*, 40(2), 139-153.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. (2020). *Language models are few-shot learners*. *Advances in Neural Information Processing Systems* 33 (NeurIPS 2020),

Chesbrough, H., & Bogers, M. (2014). Explicating open innovation: Clarifying an emerging paradigm for understanding innovation. In H. Chesbrough, W. Vanhaverbeke, & J. West (Eds.), *New Frontiers in Open Innovation*. (pp. 3-28). Oxford University Press.

Chesbrough, H. W. (2003). *Open Innovation: The New Imperative for Creating and Profiting from Technology*. Harvard Business Press.

Cornegruta, S., Bakewell, R., Withey, S., & Montana, G. (2016). Modelling radiological language with bidirectional long short-term memory networks. *arXiv preprint arXiv:1609.08409*.

Derakhshan, M., Golrezaei, N., Manshadi, V., & Mirrokni, V. (2022). Product ranking on online platforms. *Management Science*, 68(6), 4024-4041.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. North American Chapter of the Association for Computational Linguistics,

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, 30, 1-14. <https://doi.org/10.1007/s11023-020-09548-1>

Füller, J., Hutter, K., Wahl, J., Bilgram, V., & Tekic, Z. (2022). How AI revolutionizes innovation management—Perceptions and implementation preferences of AI-based innovators. *Technological Forecasting and Social Change*, 178, 121598.

Garcia, R., & Calantone, R. (2002). A critical look at technological innovation typology and innovativeness terminology: A literature review. *Journal of Product Innovation Management*, 19(2), 110-132.

Girotra, K., Meincke, L., Terwiesch, C., & Ulrich, K. T. (2023). Ideas are Dimes a Dozen: Large Language Models for Idea Generation in Innovation. *SSRN Electronic Journal*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4526071

Goldberg, D. M., & Abrahams, A. S. (2022). Sourcing product innovation intelligence from online reviews. *Decision Support Systems*, 157, 113751.

- Haase, J., & Hanel, P. (2023). Artificial muses: Generative Artificial Intelligence Chatbots Have Risen to Human-Level Creativity. *Journal of Creativity*, 3(33), 100066.
- Honnibal, M., & Montani. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.
- Hossain, M., & Islam, K. Z. (2015). Ideation through online open innovation platform: Dell IdeaStorm. *Journal of the Knowledge Economy*, 6, 611-624.
- Jia, Y., & Liu, I. L. (2018). Do consumers always follow “useful” reviews? The interaction effect of review valence and review usefulness on consumers' purchase decisions. *Journal of the Association for Information Science and Technology*, 69(11), 1304-1317.
- King, A., & Lakhani, K. R. (2013). Using open innovation to identify the best ideas. *MIT Sloan Management Review*, 55(1), 41-48.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Miaschi, A., & Dell’Orletta, F. (2020). Contextual and non-contextual word embeddings: an in-depth linguistic investigation. Proceedings of the 5th Workshop on Representation Learning for NLP,
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality* Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, Lake Tahoe, Nevada.
- Mirkovski, K., Von Briel, F., & Lowry, P. B. (2016). Semantic learning-based innovation framework for social media. *IT Professional*, 18(6), 26-32.
- Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in Big Data analytics. *Journal of Big Data*, 2, 1-21.
- Paget, S. (2023). *Local Consumer Review Survey 2023*. Retrieved 10 December from <https://www.brightlocal.com/research/local-consumer-review-survey-2023/>
- Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP),
- Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations.
- Saura, J. R., Ribeiro-Soriano, D., & Palacios-Marqués, D. (2021). From user-generated data to data-driven innovation: A research agenda to understand user privacy in digital markets. *International Journal of Information Management*, 60, 102331.
- Suominen, A., Toivanen, H., & Seppänen, M. (2017). Firms' knowledge profiles: Mapping patent data with unsupervised learning. *Technological Forecasting and Social Change*, 115, 131-142.
- Tanaka, H., Shinnou, H., Cao, R., Bai, J., & Ma, W. (2020, October 11–13). Document classification by word embeddings of BERT. Computational Linguistics: 16th International Conference of the Pacific Association for Computational Linguistics (PACLING) 2019, Hanoi, Vietnam.
- Toews, R. (2022). *A Wave of Billion-Dollar Language AI Startups Is Coming*. Retrieved 10 June from <https://www.forbes.com/sites/robtoews/2022/03/27/a-wave-of-billion-dollar-language-ai-startups-is-coming/>
- Wieting, J., & Gimpel, K. (2017). ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. *arXiv preprint arXiv:1711.05732*.
- Wu, J., Yang, T., Zhou, Z., & Zhao, N. (2023). Consumers' affective needs matter: Open innovation through mining luxury hotels' online reviews. *International Journal of Hospitality Management*, 114, 103556.
- Yang, X., Song, Z., King, I., & Xu, Z. (2022). A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(9), 8934-8954.
- Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems* (pp. Article 517). Curran Associates Inc.
- Zhai, Z., Nguyen, D., Akhondi, S., Thorne, C., Druckenbrodt, C., Cohn, T., Gregory, M., & Verspoor, K. (2019). *Improving Chemical Named Entity Recognition in Patents with Contextualized Word Embeddings*. <https://doi.org/10.18653/v1/W19-5035>