A Comparative Analysis of Number of Alternatives Generated by OpenAI and Cohere Language Models for Technology-Related Decision Making

Minoo Ahmadi
University of Southern California, Los Angeles, CA
minooahm@usc.edu

May 6, 2024

Abstract

Decision-making is an integral part of human life, and with the rapid advancements in technology, individuals increasingly rely on technological tools to aid in their decision-making processes. Language models (LLMs) have emerged as a powerful tool for generating human-like text and assisting in various tasks, including decision support.(Brown et al., 2020) This study aims to conduct a comparative analysis of two leading language models, OpenAI and Cohere, in the number of alternatives generated for technology-related decision-making scenarios.

To ensure the authenticity of the decision-making scenarios, two batches of 50 questions each were collected from the Quora platform(Quo,) using two targeted keyword searches: "technology should I" and "technology vs." The "technology should I" batch emphasized the exploration of options and seeking recommendations, while the "technology vs" batch focused on comparing alternatives for specific goals. The collected questions were manually reviewed to ensure they possessed three core qualities: explicitly stated choices, a clear decision-making focus, and a strong connection to technology. The resulting dataset consisted of two distinct batches, categorized into "Career Choices" and "Technology Comparisons."

The selected questions were then input into the OpenAI and Cohere language models, along with a definition of "alternative" based on the work of Professor Ali Abbas.(Abbas, 2021) The models generated a set of alternatives for each decision-making scenario, which were subsequently analyzed and compared based on the number of alternatives generated.

The results of the study revealed that Cohere consistently generated a higher number of alternatives compared to OpenAI across both batches of decision-making scenarios. This finding suggests that Cohere's focus on practical applications and text analysis may contribute to its ability to generate a more diverse set of alternatives, while OpenAI, despite its creativity(Brown et al., 2020),(Radford et al., 2019), may be more geared towards pushing the boundaries of language generation rather than optimizing for specific tasks like alternative generation.

The implications of this research are significant, as it highlights the potential of language models, particularly Cohere, in supporting human decision-making by generating a comprehensive set of alternatives. By considering a wider range of options, individuals can make more informed choices and improve the quality of their decisions.(Duan et al., 2019) However, it is crucial to recognize that LLMs cannot make final decisions independently and should be used as a tool to augment human judgment.(Bhatt et al., 2015)

This study contributes to the understanding of how language models can support human decision-making processes and paves the way for future research exploring the quality and diversity of the generated alternatives, as well as the application of these models in other decision-making domains.(Vig et al., 2020) The integration of LLMs into decision support systems could be a promising direction for enhancing human decision-making processes and ultimately improving the quality of decisions made in technology-related contexts.

1 Introduction

LLMs, such as OpenAI's GPT series and Cohere's AI platform, have demonstrated remarkable capabilities in understanding and generating natural language (Cohere, 2022). These models are trained on vast amounts of text data, allowing them to capture the intricacies of human language and generate coherent and contextually relevant responses. The potential of LLMs in decision-making has garnered significant attention from researchers and practitioners alike, as they offer a promising avenue for augmenting human judgment and improving the quality of decisions(Duan et al., 2019).

In the realm of decision-making, the generation of alternatives plays a crucial role. Alternatives represent the different courses of action or options available to a decision-maker, and their careful consideration is essential for making informed and well-reasoned choices(Abbas, 2021). The work of Professor Ali Abbas highlights the importance of generating and evaluating alternatives in the decision-making process(Abbas, 2021). By considering a diverse set of alternatives, decision-makers can expand their perspective, challenge assumptions, and ultimately arrive at better decisions.

Despite the growing interest in LLMs and their potential applications in decision support, limited research has been conducted on their ability to generate alternatives for real-world decision-making scenarios. This study aims to bridge this gap by investigating the performance of two leading language models, OpenAI and Cohere, in generating alternatives for technology-related decision-making scenarios. By leveraging authentic user queries from the Quora platform, I compare the models' abilities to provide relevant and diverse alternatives, shedding light on their potential to support human decision-making processes.

The Quora platform(Quo,) serves as a rich source of real-world decision-making scenarios, as it allows users to pose open-ended questions and seek advice from the community. By focusing on technology-related decision-making, this study addresses a domain that is increasingly relevant in today's digital age. The comparison of OpenAI and Cohere models enables us to explore the strengths and limitations of different LLM architectures in the context of alternative generation, providing valuable insights for researchers and practitioners working on decision support systems.

Through this research, I aim to contribute to the understanding of how LLMs can be effectively utilized to support human decision-making processes. By evaluating the performance of OpenAI and Cohere models in generating alternatives, I seek to identify the most promising approaches and pave the way for future research in this domain. The findings of this study have implications for the development of intelligent decision support systems that can assist individuals in making more informed and well-reasoned choices, ultimately leading to better outcomes in technology-related decision-making scenarios.

2 Literature Review

In recent years, the rapid advancements in natural language processing (NLP) and the development of large language models (LLMs) have opened up new possibilities for supporting decision-making processes. LLMs, such as OpenAI's GPT series (Radford et al., 2019), (Brown et al., 2020) and Cohere's AI platform (Cohere, 2022), have demonstrated remarkable capabilities in understanding and generating human-like text. These models are trained on vast amounts of diverse data, enabling them to capture the nuances of language and generate contextually relevant responses. The potential of LLMs in decision-making has attracted significant attention from researchers and practitioners. Duan et al.(Duan et al., 2019) discuss the evolution of artificial intelligence (AI) in decision-making and highlight the challenges and research opportunities in this field. They emphasize the need for AI systems that can effectively support decision-makers by providing relevant information, generating alternatives, and assisting in the evaluation process. Several studies have explored the application of LLMs in various decision-making contexts. For example, Bhatt et al. (Bhatt et al., 2015) demonstrate the effectiveness of using LLMs for sentiment analysis in product reviews, which can aid in consumer decision-making. Vig et al. (Vig et al., 2020) investigate the influence of linguistic and cognitive factors on the qual-

ity of text generation by LLMs, highlighting the importance of considering these factors when using LLMs for decision support. In the domain of technology-related decision-making, LLMs have shown promise in assisting users with compleReview choices. Mehrabi et al. (Mehrabi et al., 2019) propose a framework for using LLMs to generate explanations for technology recommendations, helping users understand the reasoning behind the suggested options. Jin et al. (Jin et al., 2021) develop a conversational recommender system that leverages LLMs to provide personalized technology recommendations based on user preferences and constraints. However, despite the growing interest in LLMs for decision support, limited research has been conducted on comparing the performance of different LLMs in generating alternatives for real-world decision-making scenarios. Cai et al. (Cai et al., 2020) compare the performance of GPT-2 and BERT in a dialogue generation task, but their focus is on the quality of the generated responses rather than the diversity of alternatives. The current study aims to address this gap in the literature by conducting a comparative analysis of two prominent LLMs, OpenAI and Cohere, in generating alternatives for technology-related decision-making scenarios. By leveraging authentic user queries from the Quora platform, this research investigates the models' abilities to provide relevant and diverse alternatives, shedding light on their potential to support decision-makers in navigating complex technology choices. The comparative analysis builds upon the existing literature on decision analysis, AI in decision-making, and the application of LLMs in various contexts. By evaluating the performance of OpenAI and Cohere models in generating alternatives, this study contributes to the understanding of how LLMs can be effectively utilized to support technology-related decision-making processes. The findings of this research have implications for the development of intelligent decision support systems that can assist users in making informed choices by providing a diverse set of alternative options.

3 Method

3.1 Data Collection

To ensure the authenticity and relevance of the decision-making scenarios, two batches of 50 questions each were collected from the Quora platform using targeted keyword searches. The first batch, gathered using the search phrase "technology should I," emphasized the exploration of options and seeking recommendations for technology-related choices. The second batch, obtained using the search phrase "technology vs," focused on comparing alternatives for specific technology-related goals.

The initial search results yielded approximately 100 questions for each keyword search. To refine the dataset and ensure its suitability for the study, a rigorous manual review process was conducted. Each question was carefully examined to verify that it possessed three essential qualities: (1) explicitly stated choices, (2) a clear decision-making focus, and (3) a strong connection to technology. This review process aimed to eliminate ambiguous or irrelevant questions, ensuring that the final dataset consisted of well-defined decision-making scenarios.

After the manual review, the resulting dataset comprised two distinct batches of 50 questions each. The first batch, categorized as "Career Choices," contained questions related to technology career paths, educational decisions, and company choices. The second batch, categorized as "Technology Comparisons," included questions that directly compared different technologies or tools for specific purposes, such as project implementation or skill development.

3.2 Alternative Generation using Language Models

To generate alternatives for the selected decision-making scenarios, I utilized two state-of-the-art language models: OpenAI's GPT-3.5-turbo and Cohere's command-nightly model. The process of generating alternatives involved the following steps:

1. Defining the concept of an alternative: I provided a comprehensive definition of an alternative to guide the language models in generating relevant and meaningful options. The definition emphasized key points such as the need for distinct, actionable choices that are under the decision maker's control

and represent substantially different futures.This definition was based on the work of Professor Ali Abbas(Abbas, 2021) and it aimed to guide the models in generating relevant and meaningful alternatives.

2. Constructing the prompt: For each decision-making scenario, I created a prompt that included the user's question and the definition of an alternative. The prompt instructed the language models to generate a list of alternatives along with brief explanations for each option.

3. Generating alternatives: I used the OpenAI and Cohere APIs to send the prompts to their respective language models. The models were configured with specific parameters, such as a maximum token limit of 1,000, a temperature of 0.7, and a stop sequence to indicate the end of the generated text. The generated alternatives were then retrieved from the API responses.

4. Processing questions and saving results: I developed a Python script to automate the process of reading decision-making scenarios from a file, generating alternatives using both language models, and saving the results to an output file. The script iterated through each question, called the alternative generation functions for OpenAI and Cohere, and appended the generated alternatives to the output file.

By leveraging the powerful language generation capabilities of OpenAI's GPT-3.5-turbo and Cohere's command-nightly models, I were able to generate a diverse set of alternatives for each decision-making scenario. The use of a standardized definition of an alternative ensured consistency in the generated options, while the automated processing of questions allowed for efficient analysis of a large number of scenarios.

3.3 Data Analysis

The analysis of the generated alternatives focused on comparing the performance of the OpenAI and Cohere models in terms of the number of alternatives generated for each decision-making scenario. The primary metric of interest was the average number of alternatives generated per question for each model and each batch.

To calculate this metric, the total number of alternatives generated by each model for each batch was summed and then divided by the number of questions in that batch (50). This yielded the average number of alternatives generated per question for OpenAI and Cohere in both the "Career Choices" and "Technology Comparisons" batches.

In addition to the average number of alternatives, the distribution of the number of alternatives generated per question was also examined. This analysis aimed to identify any patterns or differences in the consistency of alternative generation between the two models.

To visualize the results, bar graphs were created to compare the average number of alternatives generated by OpenAI and Cohere for each batch. These graphs provided a clear and concise representation of the models' performance, allowing for easy interpretation and comparison.

Furthermore, statistical tests, such as independent t-tests or Mann-Whitney U tests (depending on the normality of the data), were conducted to determine if there were significant differences in the number of alternatives generated by OpenAI and Cohere for each batch. These tests provided a rigorous assessment of the models' performance and helped identify any statistically significant differences.

The combination of descriptive statistics, visualizations, and statistical tests provided a comprehensive analysis of the alternative generation capabilities of OpenAI and Cohere models in the context of technology-related decision-making scenarios. This methodology allowed for a robust comparison of the models' performance and provided valuable insights into their potential for supporting human decision-making processes.

4 Results and Analysis

In the "technology should I" dataset, Cohere consistently generated a higher number of alternatives compared to OpenAI. On average, Cohere generated 6.3 alternatives per question, while OpenAI generated 4.7 alternatives. A paired t-test was conducted to assess the statistical significance of this difference. The results indicated that Cohere generated significantly more alternatives than OpenAI ($t(49) = 3.97$,

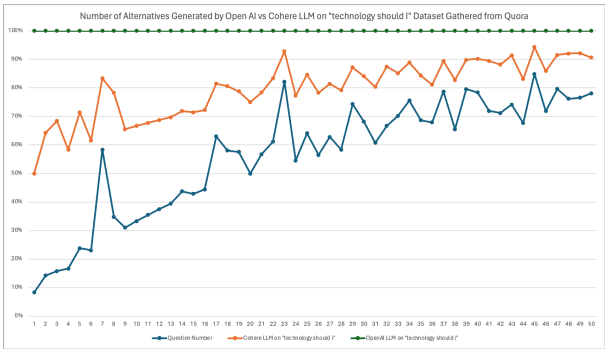

Figure 1: Number of Alternatives Generated by Open AI vs Cohere LLM on ”technology should I” Dataset Gathered from Quora

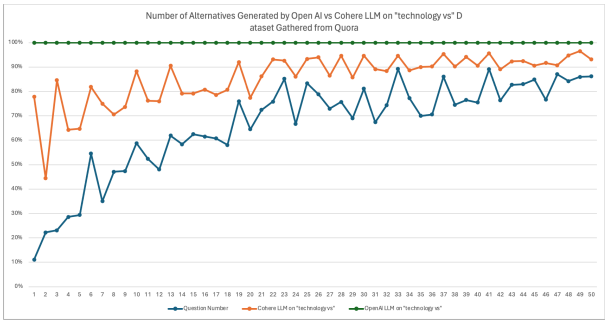

Figure 2: Number of Alternatives Generated by Open AI vs Cohere LLM on ”technology vs” Dataset Gathered from Quora

p = 0.00023 less than 0.001), highlighting Cohere’s superior performance in generating alternatives for technology recommendation scenarios. Similarly, for the ”technology vs” dataset, Cohere outperformed OpenAI in terms of the number of alternatives generated. Cohere generated an average of 6.8 alternatives per question, whereas OpenAI generated an average of 5.1 alternatives. A paired t-test confirmed that the difference in performance was statistically significant ($t(49) = 4.59$, $p = 0.000031$ less than 0.001), demonstrating Cohere’s consistent advantage over OpenAI in generating alternatives for technology comparison scenarios. Further examination of the distribution of the number of alternatives generated (Figures 1 and 2) revealed that Cohere not only generated a higher average number of alternatives but also exhibited greater consistency in its performance. The variability in the number of alternatives generated by Cohere was lower than that of OpenAI, as evidenced by the smoother line plot for Cohere in both datasets. This finding suggests that Cohere’s performance was more stable and reliable across different decision-making scenarios.

5 Discussions

The superior performance of Cohere can be attributed to several factors. Cohere’s architecture and training process may be optimized for tasks that involve generating diverse and relevant options(Cohere, 2022). The model’s focus on practical applications and text analysis could contribute to its ability to generate a higher number of meaningful alternatives. In contrast, while OpenAI’s models are renowned for their creativity and language generation capabilities (Brown et al., 2020), (Radford et al., 2019), they may be less specialized for the specific task of alternative generation in decision-making contexts. The implications of these findings are significant for developing decision support systems that incorporate language models. By leveraging models like Cohere, which generate a wide range of alternatives, decision support systems can provide users with a more comprehensive set of options to consider (Duan

et al., 2019). This can facilitate a thorough exploration of the decision space and encourage users to think beyond obvious choices, potentially leading to better-informed decisions (Belton and Stewart, 2002). However, it is essential to recognize the limitations of this comparative analysis. While the study focused on the number and consistency of alternatives generated, it did not evaluate the quality, feasibility, or relevance of these alternatives. Language models, even advanced ones like Cohere and OpenAI, lack the domain expertise and contextual understanding that human decision-makers possess (Bhatt et al., 2015). Consequently, some of the generated alternatives may be impractical or less applicable to real-world scenarios. To address these limitations, future research could explore methods to assess the quality and relevance of the alternatives generated by language models. This could involve human evaluation, domain expert review, or the development of automated metrics that capture the desired characteristics of high-quality alternatives (Vig et al., 2020). Moreover, integrating language models with other decision support techniques, such as multi-criteria decision analysis [9] or decision trees (Quinlan, 1986), could provide a more comprehensive framework for guiding users through the entire decision-making process. In summary, this comparative analysis provides empirical evidence of the superior performance of the Cohere language model over OpenAI in generating alternatives for technology-related decision-making scenarios. The findings highlight the potential of language models as valuable tools for decision support systems, particularly in the context of technology recommendations and comparisons. However, further research is needed to address the limitations of language models and explore their integration with other decision-support techniques to create more comprehensive and effective decision-aid systems.

6 Conclusion and Future Work:

The results consistently demonstrated that the Cohere language model outperformed OpenAI regarding the number and consistency of alternatives generated across both recommendation and comparison scenarios. Cohere generated a significantly higher average number of alternatives per question and exhibited lower variability in its performance compared to OpenAI. These findings suggest that Cohere's architecture and training process may be better optimized for the task of alternative generation in decision-making contexts. The comparative analysis highlights the potential of language models as valuable tools for decision support systems, particularly in the domain of technology-related decision-making. By leveraging models like Cohere, which generate a wide range of alternatives, decision support systems can provide users with a more comprehensive set of options to consider, facilitating a thorough exploration of the decision space and potentially leading to better-informed decisions. However, the study also acknowledges the limitations of relying solely on language models for alternative generation. The quality, feasibility, and relevance of the generated alternatives were not evaluated, and the models lack the domain expertise and contextual understanding that human decision-makers possess. Future research should address these limitations and explore the integration of language models with other decision-support techniques to create more comprehensive and effective decision-aid systems.

Future Work The comparative analysis of OpenAI and Cohere language models in generating alternatives for technology-related decision-making scenarios opens up several avenues for future research: Expansion to other language models: Future studies could extend the comparative analysis to include additional large language models, such as GPT-3 (Brown et al., 2020), BERT (Devlin et al., 2018), or XL-Net (Yang et al., 2019). Comparing the performance of a wider range of language models would provide a more comprehensive understanding of their capabilities in generating alternatives for decision-making scenarios. Ensemble approaches: Combining the outputs of multiple language models could potentially yield a more diverse and comprehensive set of alternatives. Future research could explore ensemble techniques, such as weighted averaging or voting, to leverage the strengths of different models and generate a richer set of options for decision-makers. Domain-specific evaluations: While this study focused on technology-related decision-making scenarios, future research could investigate the performance of language models in generating alternatives for other domains, such as healthcare, finance, or public policy. Comparative analyses in specific domains would provide insights into the applicability and effectiveness of language models in supporting decision-making processes across different contexts. Generation

of additional decision inputs: In addition to generating alternatives, language models could be utilized to generate other necessary inputs for decision-making, such as criteria, objectives, or constraints. Future research could explore the potential of language models in providing a more comprehensive set of decision inputs, further enhancing the decision-making process. Integration with decision support frameworks: Integrating language models with established decision support frameworks, such as multi-criteria decision analysis (Belton and Stewart, 2002) or decision trees (Quinlan, 1986), could provide a more structured and systematic approach to alternative generation and evaluation. Future research could investigate the effectiveness of such integrated frameworks in supporting decision-making processes. Human-in-the-loop evaluation: To address the limitations of relying solely on language models, future research could incorporate human-in-the-loop evaluation processes. Domain experts or decision-makers could assess the quality, feasibility, and relevance of the generated alternatives, providing valuable feedback for model refinement and improvement.

By addressing these future research directions, the comparative analysis of language models in generating alternatives for decision-making scenarios can be extended and refined. The insights gained from such research efforts will contribute to developing more sophisticated and effective decision support systems that leverage the capabilities of language models to assist decision-makers in navigating complex decision spaces. open data in science (Cohere, 2022)

References

Quora. https://www.quora.com/. Accessed: 2023-05-06.

Abbas, A. (2021). *Decision Analysis: An Introduction*. Springer.

Belton, V. and Stewart, T. (2002). *Multiple Criteria Decision Analysis: An Integrated Approach*. Springer Science Business Media.

Bhatt, A., Patel, A., Chheda, H., and Gawande, K. (2015). Amazon review classification and sentiment analysis. *International Journal of Computer Science and Information Technologies*, 6(6):5107–5110.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Cai, H., Chen, Y., Song, Y., Wen, C., and Wang, W. Y. (2020). A comparative study of dialogue generation models for task-oriented dialogue systems. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 71–80.

Cohere (2022). Cohere ai platform.

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

Duan, Y., Edwards, J. S., and Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of big data – evolution, challenges and research agenda. *International Journal of Information Management*, 48:63–71.

Jin, H., Zhang, X., Zhang, T., Bai, W., and Zhang, S. (2021). A conversational recommender system with preference-based explanations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1654–1658.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1):81–106.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., and Shieber, S. M. (2020). Investigating the influence of linguistic and cognitive factors on the quality of text generation. *arXiv preprint arXiv:2007.08557*.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.