Accelerate Creation of Product Claims Using Generative AI

Po-Yu Liang
Department of Computer Science
University of Cincinnati
Cincinnati, OH 45219
liangpu@mail.uc.edu

Yong Zhang
Corporate Functions R&D, Discovery& Innovation Platforms, P&G
Mason, OH 45040
zhang.y.13@pg.com

Tatiana Hwa
Beauty Care IT, P&G
Mason, OH 45040
hwa.t@pg.com

Aaron Byers
Beauty Care R&D, P&G
Mason, OH 45040
byers.al@pg.com

Abstract

The benefit claims of a product is a critical driver of consumers' purchase behavior. Creating product claims is an intense task that requires substantial time and funding. We have developed the **Claim Advisor** web application to accelerate claim creations using in-context learning and fine-tuning of large language models (LLM). **Claim Advisor** was designed to disrupt the speed and economics of claim search, generation, optimization, and simulation. It has three functions: (1) semantically searching and identifying existing claims and/or visuals that resonate with the voice of consumers; (2) generating and/or optimizing claims based on a product description and a consumer profile; and (3) ranking generated and/or manually created claims using simulations via synthetic consumers. Applications in a consumer packaged goods (CPG) company have shown very promising results. We believe that this capability is broadly useful and applicable across product categories and industries. We share our learning to encourage the research and application of generative AI in different industries.

1 Introduction

Product claims play a crucial role in influencing consumer choice and building brand trust, as they provide essential information that helps both consumers and healthcare professionals assess the suitability, efficacy, and safety of various products [5]. Clear, well-substantiated claims can significantly enhance consumer confidence, serving as a critical driver for purchasing decisions, brand loyalty, and overall market success. In contrast, misleading, exaggerated, or unsubstantiated claims not only undermine consumer trust but also expose companies to regulatory scrutiny, legal penalties, and reputational harm [9]. To safeguard consumers and ensure fair competition, regulatory bodies such as the Federal Trade Commission (FTC) emphasize that product claims, particularly health-related ones, must be truthful, not misleading, and substantiated by robust scientific evidence

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The First Workshop on Generative and Protective AI for Content Creation.

[4]. Research consistently highlights that a significant proportion of product advertisements lack accurate or sufficient scientific citations to support their claims [9].

Product claims need to be legally compliant and scientifically supported. Ideally, manufacturers and marketers also want these claims to be resonant with trendy consumer talk points. Creating these claims requires substantial time and funding as it generally needs to conduct many iterations of designs and tests. Traditionally, the first step is to search a large volume of existing claims and visuals to decide whether one should pivot the existing claims or design brand new claims. These manually crafted candidate textual claims and visuals are then tested with consumers to gauge appeal and collect feedback. After a few iterations, the top appealing claims and visuals are further investigated for scientific support and legal approval. This process can take weeks or even months and require substantial financial resources.

The gap between compliant and resonant claims and industry practice underscores the urgent need for credible and evidence-based strategies to create claims. Consumers often place disproportionate trust in products labeled as "scientifically studied" or "clinically proven" even when the claims are exaggerated and weakly supported [14]. This phenomenon reveals not only a vulnerability in consumer decision-making but also a responsibility for marketers and product manufacturers to ensure the integrity and transparency of their product claims.

Recent advances in LLMs offer a promising opportunity to address the above time and resource challenges of creating product claims. LLMs have demonstrated remarkable capabilities in semantic understanding [19] and natural language generation tasks [12], and have even exhibited creativity in domains traditionally thought to require human ingenuity [6]. As an example, LLMs have been combined with interactive agents to enable believable simulations of collective human behavior [15].

We developed a platform called **Claim Advisor** to bridge the gap between LLMs and creation of product claims. **Claim Advisor** was designed as a minimum-viable-prototype (MVP) web application to quickly search, generate, optimize and rank product claims using LLMs. First, we employ prompt engineering and in-context learning techniques, utilizing previously obtained market research insights to guide the LLM in generating more targeted, persuasive, and compliant claims. This enables a more efficient ideation process while aligning generated claims with real-world consumer expectations and regulatory requirements.

Second, to further streamline the market research phase and reduce its associated costs, we fine-tuned a lightweight version of Microsoft's Phi-3 model [1] using Low-Rank Adaptation (LoRA) techniques [10]. This fine-tuned model is designed to simulate initial consumer feedback, allowing for faster iteration cycles before engaging in full-scale market research. Together, these two components create a more agile, cost-effective framework for creating product claims, paving the way for accelerated product development without compromising scientific rigor or consumer trust.

The code base, dummy data files and example prompts are shared in this GitHub repository to facilitate reapplication and to encourage further research.

2 Method

2.1 Technical Diagram

As shown in the figure 1, **Claim Advisor** takes Claim Log data and Maximum Difference Scaling (MaxDiff) data [13] as input. Users can input product descriptions, consumer profiles, trending consumer topics and/or visuals to search, generate, optimize and rank claims. Claim Log data contain legally approved claims. It has claim log IDs to link to the related technical support documents. MaxDiff data contain preferences of candidate claims and visuals already screened through MaxDiff studies. MaxDiff is the major research method to study consumers' preferences on candidate claims before seeking approval.

We used the commercial LLM ChatGPT-4o through Azure for in-context learning and the open source LLM Phi-3 [1] for fine-tuning in our platform. As an MVP web application, **Claim Advisor** used *LangChain* [2] and *streamlit* [17] to build the backend and front end, respectively. Docker [7] is used to containerize the backend and front end codes and deployed on a secure server. Readers can check the graphical user interface in the appendix B.

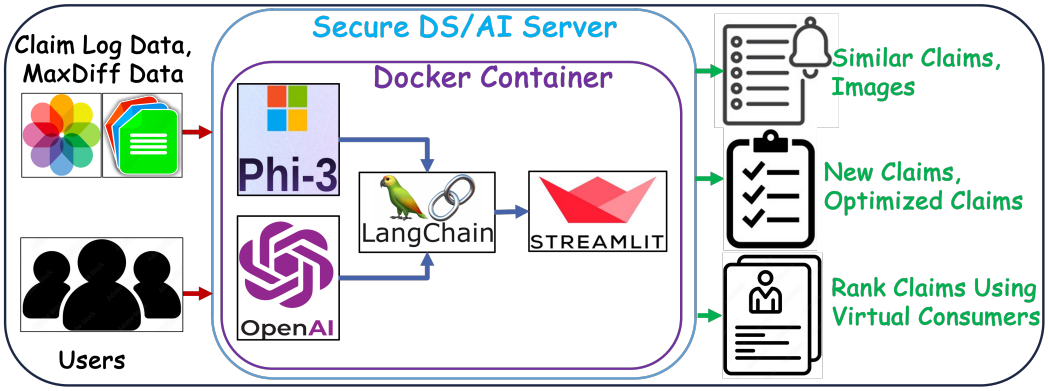

Figure 1: Technical diagram of **Claim Advisor** MVP web application

2.2 Search Claims

To enable semantic claim retrieval from the dataset, we employ text embedding techniques using OpenAI's TEXT-EMBEDDING-ADA-002 model. Semantic similarity is computed via cosine similarity between the embedded representations of textual inputs. Given the close relationship between claim content and visual design elements, we further support the retrieval of related images by leveraging the CLIP model [16], which aligns text and image embeddings within a shared latent space. This alignment allows user-provided textual inputs to effectively match with relevant visual content and vice versa. To enhance flexibility, we introduce a multimodal fusion approach wherein users can simultaneously input both text and image queries. By specifying a weight parameter W, users control the relative influence of textual (emb_{txt}) and visual (emb_{img}) components. The fused embedding is computed as

$$emb = (1 - W) \cdot emb_{txt} + W \cdot emb_{img}. \tag{1}$$

The resulting representation is then used to retrieve the most semantically similar images from existing designs based on cosine similarity in the embedding space.

If the retrieved claim text and images are from Claim Log data, because claim text and visuals in Claim Log data have been legally approved and have technical support documents, they can be directly repurposed in another usage scenario such as a similar product variant or same product in different distribution channels. If the retrieved contents are from MaxDiff data, they needs to be go through technical tests and legal review before usage on a product.

In the context of product claim creation, MaxDiff offers a robust methodological framework for empirically assessing consumer preferences among a large number of candidate claims. Given a pool of generated claims $\{C_1, C_2, \ldots, C_N\}$, MaxDiff analysis can identify which claims maximize perceived relevance, credibility, and persuasiveness according to the target audience. This approach overcomes limitations inherent to traditional rating-scale evaluations, where respondents often rate multiple claims similarly, masking meaningful preference differences. Readers can check appendix B for further description of the MaxDiff method.

2.3 Generate and Optimize Claims

Prompt engineering and in-context learning are two key techniques for adapting LLMs to specific tasks without requiring fine-tuning. Prompt engineering [3] involves carefully designing input prompts to guide the model's output toward desired behaviors, often by including detailed instructions, task-specific context, or illustrative examples. In-context learning [8] refers to the model's ability to infer patterns or strategies directly from examples presented within the prompt, allowing it to perform new tasks without updating model parameters.

By leveraging prompt engineering and in-context learning, LLMs can be steered toward producing higher-quality, contextually appropriate outputs, even in specialized domains such as product claim generation. To incorporate knowledge from MaxDiff studies into the LLM, we combined prompt engineering with in-context learning. The prompt includes both target consumer information and

examples derived from previous MaxDiff research. The target consumer information is embedded in the system message, while the examples from past MaxDiff studies are incorporated through an in-context learning approach.

For each MaxDiff study, we generated examples where the model produces an optimized claim based on five given claims. The examples are constructed using two methods, as illustrated in Figure 2. The performance based method (indicated by the green dotted line) is based on MaxDiff scores. Our hypothesis is that the LLM can infer consumer preferences by analyzing a set of moderately successful claims and then synthesizing a better, novel claim by combining information from those examples. In our implementation, we selected the second to sixth highest scoring claims as input examples and tasked the model to generate a claim that outperforms them.

The semantic based method (indicated by the blue dotted line) relies on semantic similarity. The hypothesis for this approach is that claims semantically similar to the best-performing claim may indicate a correct thematic direction. Therefore, we generated examples by selecting the five claims most semantically similar to the best-performing claim. Semantic similarity between claims is calculated using the cosine similarity between embeddings. We generated 300 examples from past MaxDiff studies and used them as in-context learning for claim generation and optimization. The input to both methods are the claims from past MaxDiff studies. The output are the in-context learning examples used in the prompt.

In practice, both methods can be combined within the same prompt to enhance in-context learning and improve claim generation. To evaluate the model's performance, we conducted three rounds of MaxDiff research of 30 claims. In the first round, all claims were manually created by human experts. In the second round, new claims were generated by combining previous MaxDiff results with findings from the first round, using the two methods described above. In the third round, claims were generated similarly, based on results from the second round. The number of claims in each round was kept the same (i.e. 30 claims) as in the initial round. Through this iterative process, we aim to progressively optimize the quality of claims by leveraging knowledge accumulated from prior MaxDiff research.

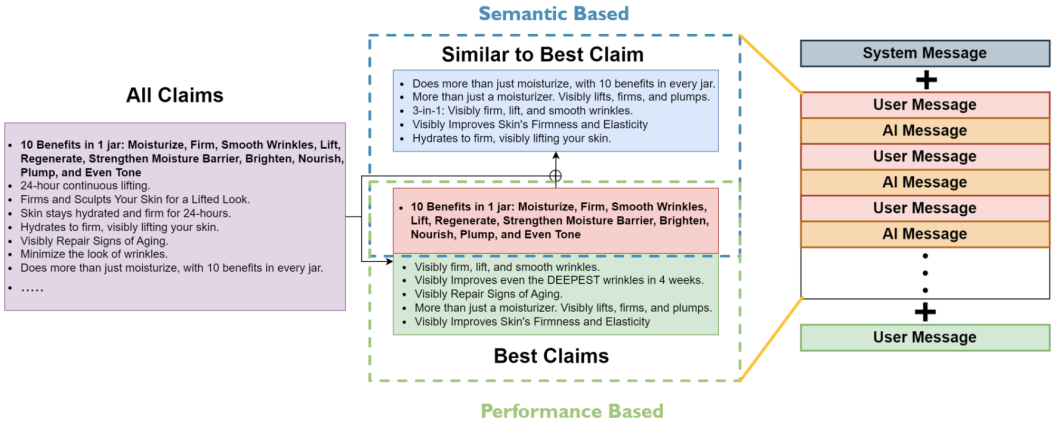

Figure 2: In-context learning examples construction for generating an optimized claim.

2.4 Rank Claims

To virtually screen claims before performing actual MaxDiff research, we guide the model to mimic the MaxDiff process by selecting the best and worst claims from a set of five given claims. We applied prompt engineering and in-context learning techniques, building on the experience from the claim optimization task. We tested model performance under different settings by varying the number of in-context learning examples provided (1, 5, and 10 examples) with a fine-tuned model.

We fine-tuned a lightweight pretrained model, Phi-3 from Microsoft [1], using the Low-Rank Adaptation (LoRA) [10] method. We evaluated models with 7B and 14B parameters. Fine-tuning refers to the process of updating a pretrained model's parameters on a small, task-specific dataset to adapt it to new objectives while preserving its general language capabilities. LoRA is a parameter-

efficient fine-tuning technique that introduces trainable low-rank matrices into certain layers of the model, significantly reducing computational costs and memory usage. By applying LoRA to Phi-3, we aim to enhance the model's ability to understand and rank claims according to consumer preferences without requiring full retraining.

The examples used for fine-tuning are illustrated in Figure 3. Specifically, we select five claims with known performance and construct the expected output by identifying the best and worst claims. We used 100316 training examples from past MaxDiff studies for fine-tuning.

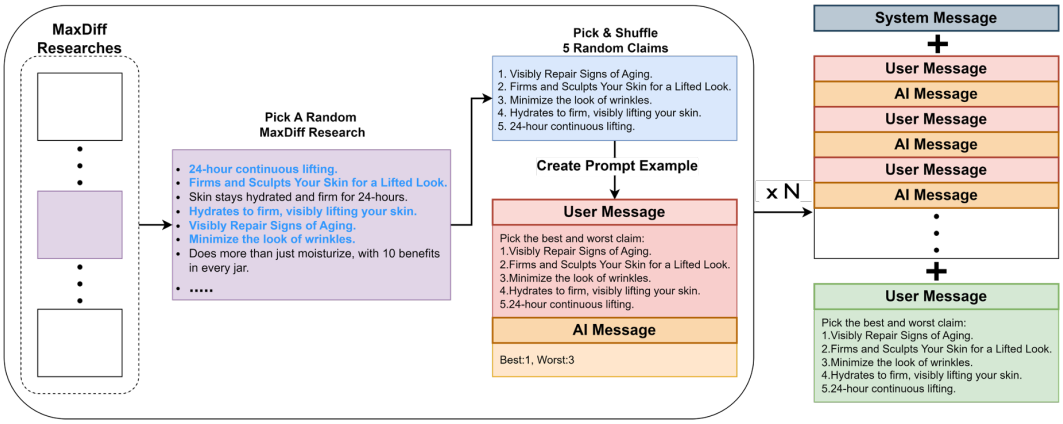

Figure 3: Example construction method for fine-tuning in the claim ranking task.

During fine-tuning, we also included a single training example as an in-context learning example in the prompt. We observed that including this example helped the model learn the output format more effectively. After obtaining the fine-tuned model, we utilized it by providing five random claims—drawn from the set of claims to be tested—and asking the model to select the best and worst claim. By repeating this process multiple times, we calculated the number of times each claim was selected as the best or worst. We then computed a score for each claim using the ratio of number of times in which it was selected as the best to that of being selected as the worst. These scores were used to rank the generated claims (model-predicted rank).

We employed Kendall's tau coefficient [11] to quantitatively evaluate the alignment between model-predicted rank and the true rank. The true rank was obtained through the holdout MaxDiff studies. Kendall's tau is a non-parametric statistic that measures the ordinal association between two ranked variables by comparing the number of concordant and discordant pairs. Formally, given two rankings, τ is calculated as:

$$\tau = \frac{C - D}{\frac{1}{2}n(n-1)} \tag{2}$$

where C is the number of concordant pairs, D is the number of discordant pairs, and n is the total number of ranked items. A τ value of 1 indicates perfect agreement between the rankings, -1 indicates complete disagreement, and 0 suggests no correlation. By using Kendall's tau, we can robustly assess how well the model's predicted claim rankings approximate the ground truth without being overly sensitive to minor ranking variations.

3 Result

3.1 Search Claims

Claim Advisor accelerates searches of text and image of claims, maximizing the return on claim log and MaxDiff study data assets. It can return similar claims within seconds from a large volume of MaxDiff studies and claim log data. Users can use different filters such as age groups and product lines to limit their searches.

Table 1: Performance of generated claims in three rounds of MaxDiff researches.

	Round-1 (Human)	Round-2 (Claim Advisor)	Round-3 (Claim Advisor)
High Appealing	20%	33%	100%
Appealing	46%	36%	0%
Less Appealing	34%	31%	0%

3.2 Generate and Optimize Claims

Claim Advisor can generate and/or optimize tens to hundreds of candidate claims within minutes using in-context learning method as detailed in section 2.3. We evaluate the performance by applying two cutoff percentages to the MaxDiff preference likelihoods, separating 30 claims into three regions: *highly appealing*, *appealing*, and *less appealing*. The preference likelihood of a generated claim is the ratio of times in which it was selected as the best claim to all times in which it was selected as a candidate in a random draw in a round of a MaxDiff study. The two cutoff percentages $P1$ and $P2$ ($P2 > P1$) are determined empirically based on our internal MaxDiff studies. The results are shown in Table 1. If the preference likelihood is great than $P2$, it is classified as highly appealing. If it is less than or equal to $P1$, it is less appealing.

In the first round (claims designed by human experts), the majority of claims (46%) fall into the appealing region, while 20% of claims are highly appealing to consumers and 33% are categorized as less appealing. In the second round, we observed an improvement: the proportion of highly appealing claims increases by 13%, reaching 33%, while the proportion of less appealing claims decreases from 34% to 31%.

By the third round, all claims achieve highly appealing performance. This surprisingly strong result demonstrates that the model, even relying only on in-context learning and prompt engineering, can effectively learn consumer preferences within a relatively small number of iterations. These findings highlight the strong potential of LLMs to assist humans in creative tasks such as claim creation.

3.3 Rank Claims

The performance of the claim ranking models was evaluated using Kendall's tau coefficient. We compared our fine-tuned models against ChatGPT-3.5, GPT-4, and GPT-4o[1]. To ensure a fair comparison between our fine-tuned models and the non-fine-tuned ChatGPT models, we provided the ChatGPT models with 100 examples from previous MaxDiff studies, whereas our models received only a small number of examples (1, 5, or 10). This setup helps minimize differences in the amount of contextual information provided to each model.

Figure 4 presents Kendall's tau coefficients across all evaluated models. The Phi-3 model with 14B parameters consistently outperforms all others. Although there is a noticeable improvement in performance from GPT-3.5 to GPT-4 and GPT-4o, none of the ChatGPT models surpass our best-performing model. Notably, the Phi-3 7B model (denoted as *mini*), despite its smaller size, achieves performance comparable to the 14B version (*medium*) when using only 10 in-context examples. It also significantly outperforms GPT-3.5, even though the latter was given 10 times more examples.

Interestingly, increasing the number of in-context examples does not always improve performance. For instance, the Phi-3 medium model shows a slight performance drop when moving from one to ten examples. Overall, the best result was achieved by the Phi-3 medium model using only a single example, suggesting that smaller, well-designed prompts can be highly effective for this task.

To further analyze model utility in practical applications, we measured top-N coverage, shown in Figure 5. This metric evaluates how many of the model's top-N predicted claims align with the top-N claims identified through actual MaxDiff research. This is particularly relevant when users are only interested in identifying the most promising claims, rather than perfectly ordering all candidates.

We observed that the Phi-3 mini model, despite its lower Kendall's tau score, achieves top-3 and top-5 coverage comparable to the Phi-3 medium model when provided with 10 examples. In contrast, GPT-3.5 and GPT-4 exhibit the lowest top-N coverage, with both failing to capture any of the top

[1]Accessed via Azure OpenAI API in July 2024

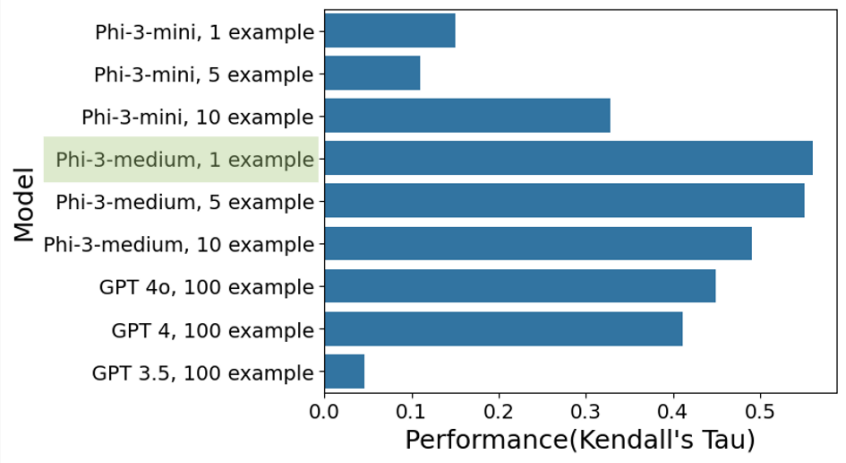

Figure 4: Model performance based on Kendall's tau. *Phi-3-mini* and *Phi-3-medium* denote models with 7B and 14B parameters, respectively. *GPT* refers to ChatGPT models from OpenAI. Best performed model is marked with green box.

3 claims. GPT-4o performs better, matching our best model when provided with 100 in-context examples. Nevertheless, our fine-tuned Phi-3 models—especially given their smaller size and fewer required examples—demonstrate superior efficiency and scalability for claim ranking tasks, offering substantial cost and resource advantages.

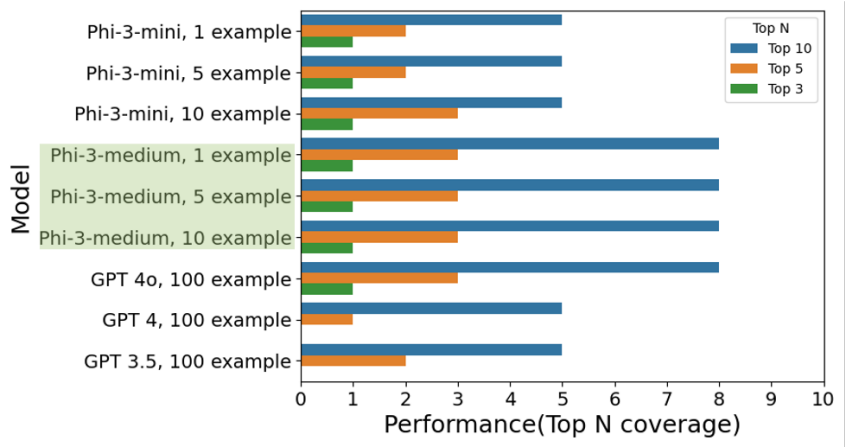

Figure 5: Top-N coverage performance. *Phi-3-mini* and *Phi-3-medium* denote models with 7B and 14B parameters, respectively. *GPT* refers to ChatGPT models from OpenAI. Best performed model is marked with green box.

4 Discussion

For claim optimization, we found that prompt engineering contributed to noticeable performance improvements. Specifically, incorporating domain knowledge into prompts helped guide the LLMs to generate claims more closely aligned with interests of targeted consumers. However, we also observed a trade-off: excessive instruction in prompts can reduce output diversity, leading to repetitive claims centered on the provided guidance. These findings highlight that while prompt design remains an underexplored area, well-crafted prompts can significantly influence the effectiveness of LLM-driven generation tasks. Diverse in-context learning examples from multiple product categories may help to mitigate the issue of lack of diversity caused by excessive instruction. Further experiments need to be

conducted to investigate the robustness of claim generation and optimization based on the number of input claims, the cut off percentages and the number of in-context learning examples.

For claim ranking, our initial approach involved asking the LLM to rank all claims in a single pass. However, we found that the resulting rankings lacked statistical alignment with MaxDiff outcomes. By mimicking the MaxDiff methodology, namely asking the model to select the best and worst claims from small sets, we achieved results that were statistically meaningful and closely matched real-world consumer preferences. This underscores a current limitation of LLMs: while capable of generating plausible outputs, they may struggle with tasks that require implicit statistical reasoning unless they are structured in a way that reflects the underlying process. These bias in LLM have been well documented and could be mitigated by technics such as contextual calibration procedure [18].

In a MaxDiff study, a consumer is asked to pick the best and the worst in a set randomly picked from a pool. It does not need to provide ranks for every items in the set. It is a nice balances between ranking every items in a set (hard to differentiate all items) and only picking the best item (losing what a consumer does not like). MaxDiff plays into psychology and recognition when choosing a product from a long list of offerings. In practice, we found MaxDiff has been very effective to screen concepts, claims etc.

Our experiments demonstrate that effective system design, which integrates traditional methods and domain expertise, can enable LLMs to perform complex tasks with high reliability. Using open-source models such as Phi-3 offers additional benefits, including greater control, transparency, and lower inference costs. In particular, our fine-tuned lightweight models outperformed larger commercial alternatives such as GPT-3.5 and GPT-4, despite using significantly fewer in-context examples. We did not compare a model with only prompt (no in-context examples) with a fine-tuned model because early tests showed LLM prompt alone tends to generate generic claims without specificity.

Several practical considerations emerged from our deployment experience. First, latency remains a challenge, particularly for claim ranking, which requires thousands of inferences within a short period of time. Second, the stability of the model is critical. OpenAI's commercial models, while powerful, are periodically updated, which can affect reproducibility and prompt behavior. In contrast, open-source models offer greater consistency and can be fine-tuned to maintain performance over time. Finally, cost management is more feasible with open source models, especially when lightweight fine-tuned versions can match or exceed the performance of commercial offerings.

Overall, our work demonstrates the strong potential of LLMs in assisting human creativity and decision-making tasks, provided that their use is carefully designed and informed by domain knowledge. We believe our pipeline could generalize on usage cases from other industries given product claims are needed virtually everywhere for either physical or digital products. We share the code base and hope to encourage research, reapplication and improve the pipeline from academia and industry.

5 Conclusion

A MVP web application **Claim Advisor** has been developed to accelerate product claim creations using in-context learning and fine-tuning of LLMs. It leverages LLMs to learn from existing claim log data, MaxDiff research data and domain expertise to disrupt the speed and economics of claim search, generation, optimization, and simulation. The application in a consumer packaged goods company has significantly accelerated the efficiency and creativity of product researchers in creating product claims.

6 Limitations

This study has several limitations that impact the transparency and reproducibility of our findings. First, our research relies on a proprietary dataset derived from internal MaxDiff studies, which cannot be publicly released due to confidentiality constraints. Additionally, the specific prompts used in our experiments are also not fully disclosed, as they contain proprietary information closely tied to the dataset and internal business logic. Furthermore, while we fine-tuned a lightweight open-source model (Phi-3), the resulting model cannot be publicly shared due to the use of non-disclosable training data.

For the evaluation of commercial LLMs such as ChatGPT-3.5, GPT-4, and GPT-4o, results were obtained via the OpenAI API through Azure. Because these models are closed-source and subject to updates without notice, the outcomes are not guaranteed to be reproducible. Lastly, the interactive Web-GUI system developed as part of this research (**Claim Advisor**) is currently not published or available for public use.

Together, the limitations above restrict the replicability and open access of our research. However, we shared our code base, example prompt and necessary dummy data files to show the format of input data in this GitHub repository. Researchers are encouraged to reapply using their own data and to conduct further research in this area based on our code base.

Acknowledgments and Disclosure of Funding

We thank our colleagues Kelly Anderson, Gabriel Comeron, Weronika Koga, George Gabone, Matt Barker and Oya Aran for their help and valuable feedback during the research and deployment of **Claim Advisor**. This research is solely sponsored by P&G. Authors have no conflict of interest.

References

[1] M. Abdin, J. Aneja, H. Awadalla, A. Awadallah, A. A. Awan, N. Bach, A. Bahree, A. Bakhtiari, J. Bao, H. Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

[2] H. Chase and LangChain. Langchain: Building applications with llms through composability. `https://www.langchain.com`, 2022. Accessed: 2025-05-13.

[3] B. Chen, Z. Zhang, N. Langrené, and S. Zhu. Unleashing the potential of prompt engineering in large language models: a comprehensive review. *arXiv preprint arXiv:2310.14735*, 2023.

[4] F. T. Commission. Health products compliance guidance. `https://www.ftc.gov/system/files/ftc_gov/pdf/Health-Products-Compliance-Guidance.pdf`, 2022. Issued December 20, 2022.

[5] N. L. Cousté, M. Martos-Partal, and E. Martínez-Ros. The power of a package: product claims drive purchase decisions. *Journal of Advertising Research*, 52(3):364–375, 2012.

[6] Z. Ding, A. Srinivasan, S. MacNeil, and J. Chan. Fluid transformers and creative analogies: Exploring large language models' capacity for augmenting cross-domain analogical creativity. In *Proceedings of the 15th Conference on Creativity and Cognition*, pages 489–505, 2023.

[7] Docker, Inc. Docker. `https://www.docker.com`, 2013. Accessed: 2025-05-13.

[8] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, T. Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[9] J. C. Dumville, E. S. Petherick, S. O'Meara, P. Raynor, and N. Cullum. How is research evidence used to support claims made in advertisements for wound care products? *Journal of clinical nursing*, 18(10):1422–1429, 2009.

[10] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[11] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

[12] J. Li, T. Tang, W. X. Zhao, J.-Y. Nie, and J.-R. Wen. Pre-trained language models for text generation: A survey. *ACM Computing Surveys*, 56(9):1–39, 2024.

[13] A. A. Marley and J. J. Louviere. Some probabilistic models of best, worst, and best–worst choices. *Journal of mathematical psychology*, 49(6):464–480, 2005.

[14] D. H. Murphy, S. T. Schwartz, K. O. Alberts, A. L. Siegel, B. J. Carone, A. D. Castel, and A. Drolet. Clinically studied or clinically proven? memory for claims in print advertisements. *Applied Cognitive Psychology*, 37(5):1085–1093, 2023.

[15] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023.

[16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[17] Streamlit, Inc. Streamlit. `https://streamlit.io`, 2019. Accessed: 2025-05-13.

[18] Z. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR, 2021.

[19] R. Zhou, L. Chen, and K. Yu. Is llm a reliable reviewer? a comprehensive evaluation of llm on automatic paper reviewing tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9340–9351, 2024.

A Web-Based GUI

To bridge the gap between LLMs and end users, we developed a web-based graphical user interface (GUI) system, referred to as **Claim Advisor**. This MVP web application was deployed on a GPU server on premise (NVIDIA GeForce RTX 2080, 11GB memory). It allows users to search existing claims, generate and optimize new claims and perform preliminary screening using a simulated MaxDiff study. Figure 6 shows a snapshot of the GUI. It shows that "MaxDiff Simulator" is used to rank 6 generated claims and 1 manually created claim (shiny skin) using 50 round of simulations.

Claim Advisor is implemented as a secure containerized pipeline for the automated search, generation, optimization and evaluation of product claims (Figure 1). The system is deployed within a *Docker* container[7] on a secure server. It ingests structured inputs, such as historical claim logs and MaxDiff data, which reflect consumer preference patterns. Users interact with the system via a MVP web application built using *Streamlit*[17]. Internally, the workflow is orchestrated by *LangChain*[2], which coordinates the flow of data and manages interactions across multiple AI components.

The MVP web application of **Claim Advisor** has been productionized at a consumer packaged goods company. It has significantly accelerated product researchers efficiency and creativity to search, generate, optimize and rank product claims.

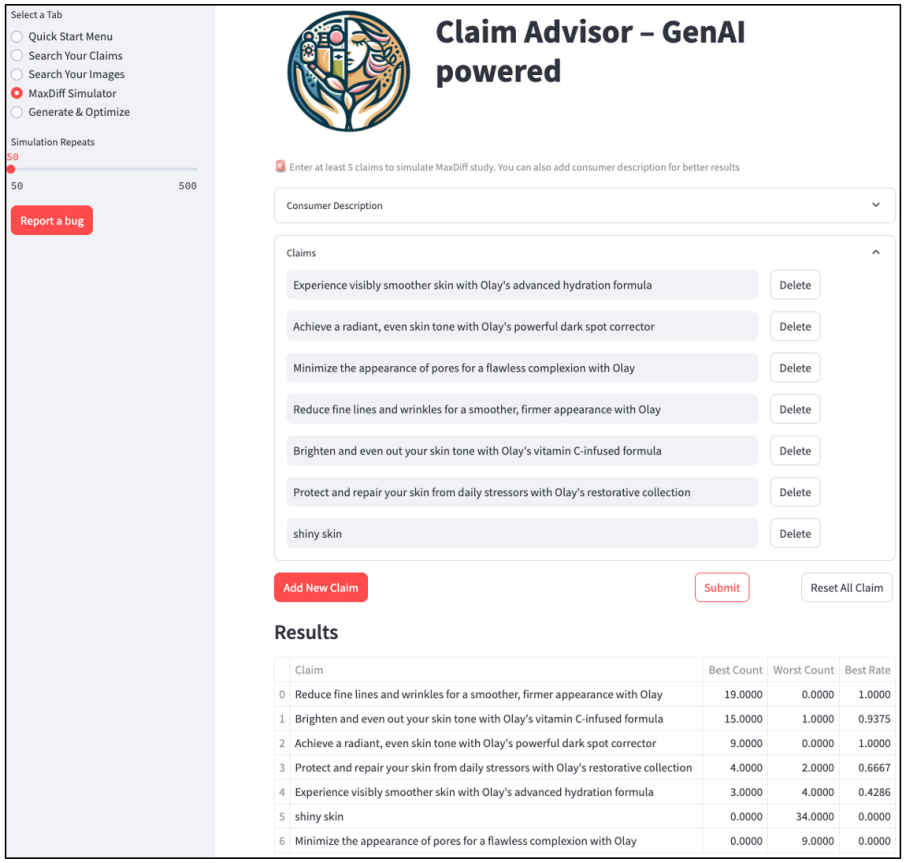

Figure 6: Graphical user interface of the **Claim Advisor** MVP web application.

B Maximum Difference Scaling

Maximum Difference Scaling (MaxDiff)[13], also known as best-worst scaling, is a discrete choice modeling technique used to measure relative preferences or importance scores among a set of items. In a typical MaxDiff task, given a subset $S \subseteq \{i_1, i_2, \ldots, i_k\}$ of k items selected from a larger item pool of size N, respondents are asked to choose the "best" (most preferred) and "worst" (least preferred) items. k and N can typically take values of 3-6 and 20-100, respectively. Each choice

provides information about the underlying utility difference between items: if item i is selected as best and item j as worst, it implies $U(i) - U(j) > U(m) - U(n)$ for all other pairs (m, n) in S, where $U(\cdot)$ represents the latent utility associated with each item. By aggregating responses across multiple sets and respondents, it is possible to estimate the utilities $U(i)$ for all items using models such as multinomial logit (MNL) or hierarchical Bayesian estimation. These utilities can then be normalized (e.g., to sum to 100) to produce a ratio-scaled measure of relative importance[13].