Customer Review Summarization in Production with Large Language Models for e-Commerce Platforms

H. Bahadir Sahin, M. Furkan Eseoglu, Berk Taskin, Aysenur Kulunk and Ömer Faruk Bal
Data Analytics Department, Hepsiburada, Istanbul, Turkey

Keywords: Large Language Models, Customer Review Summarization, Opinion Summarization, e-Commerce, LLM-as-a-Judge.

Abstract: The increasing volume of customer-generated content on large-scale e-commerce platforms creates significant information overload, complicating consumer decision-making processes. This study presents a case study of the ”Customer Review Summarization” system, deployed in production at Hepsiburada, one of Turkey’s leading e-commerce platforms, as a solution to this problem. The system leverages Large Language Models (LLMs) to generate meaningful and concise summaries from customer feedback. The primary contribution of this paper is the detailed description of an innovative, staged system architecture that optimizes the trade-off between operational cost and output quality. This architecture utilizes any LLM model for large-scale and cost-effective summary generation, while strategically leverages the more powerful GPT-4.1 model for quality assurance. Furthermore, practical challenges encountered in a production environment, such as inconsistent summary quality and managing noisy input data, and the iterative solutions developed to address them, are transparently discussed. Finally, a comprehensive empirical evaluation framework is proposed to compare the performance of various state-of-the-art LLMs.

1 INTRODUCTION

The modern e-commerce ecosystem is defined by a massive amount of user-generated content. While customer reviews serve as a valuable resource by providing social proof for potential buyers, the presence of hundreds or even thousands of reviews creates a dilemma, imposing a significant cognitive load on consumers. Users are forced to read numerous reviews to gain a comprehensive understanding of a product, which prolongs the shopping experience and complicates the decision-making process. This issue becomes more pronounced on smaller screens, such as mobile devices. In this context, the main business objective is to increase conversion rates by shortening the time from a customer viewing a product to placing an order.

In response to this challenge, Hepsiburada has developed the ”Review Summary” feature, which leverages advanced Large Language Models (LLMs), and presents these summaries for hundreds of thousands of products as in Figure 1. The core purpose of this system is to provide comprehensive, easy-to-read, and unbiased summary of products by analyzing verified and approved customer feedback. This approach

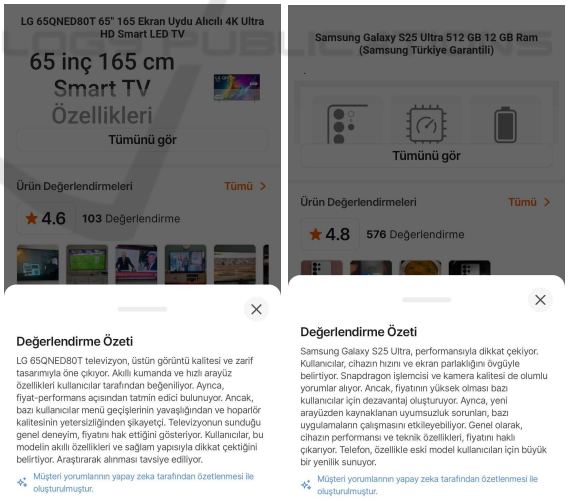

Figure 1: Summary examples from in Hepsiburada.

helps users quickly understand the positive and negative aspects of a product, enabling them to make informed decisions. Unlike traditional machine learning projects, pre-trained foundational models, such as GPT, Gemini, Llama, etc., significantly accelerate development cycles by focusing on system integration and contextual engineering rather than lengthy data

Sahin, H. B., Eseoglu, M. F., Taskin, B., Kulunk, A. and Bal, Ö. F.
Customer Review Summarization in Production with Large Language Models for e-Commerce Platforms.
DOI: 10.5220/0014284800004848

In *Proceedings of the 2nd International Conference on Advances in Electrical, Electronics, Energy, and Computer Sciences (ICEEECS 2025)*, pages 5-13
ISBN: 978-989-758-783-2

labeling and model training processes.

The contributions of this paper to the field can be summarized under four main headings:

- A detailed case study is presented on the design, deployment, and maintenance of a large-scale generative AI system in a high-traffic commercial environment, covering approximately 300000 unique products.
- A pragmatic and cost-driven system architecture (the staged LLM approach) that balances performance and operational expenditure, a critical aspect for industrial applications often overlooked in the academic literature, is thoroughly examined.
- Moving beyond idealized laboratory settings, the real-world challenges encountered in a production environment and the solutions developed for them are transparently discussed.
- A robust framework is proposed for the comparative evaluation of contemporary LLMs on a domain-specific summarization task.

The remainder of this paper will first review relevant academic works, followed by a detailed explanation of the system's architecture and methodology. Subsequently, the planned experimental setup and evaluation metrics will be presented, and finally, the study's conclusions and potential directions for future work will be discussed.

2 RELATED WORKS

This section situates our work within the context of academic research from leading NLP and data mining conferences, tracing the evolution of opinion summarization from foundational techniques to the current state-of-the-art driven by Large Language Models.

2.1 From Extractive to Aspect-Based Summarization

The core challenge of opinion summarization is to distill salient information from a multitude of user reviews (Pecar, 2018). Early approaches often focused on extractive methods that identify and select representative sentences or phrases from the source. However, these methods can lack fluency and fail to synthesize information cohesively. This led to the development of more structured approaches, most notably aspect-based summarization, which aims to identify key product features (aspects) and aggregate the sentiment expressed towards them (Titov and McDonald, 2008). A significant advancement in this area was presented by (Angelidis and Lapata, 2018), who developed a neural framework that combines aspect extraction and sentiment prediction in a weakly supervised manner, reducing the reliance on heavily annotated data by using product domain labels and user ratings for supervision.

2.2 Abstractive and Controllable Summarization

Although structured summaries are useful, the pursuit of more fluent, human-like outputs pushed the field toward abstractive summarization, where models generate novel sentences to paraphrase the source content (Iso et al., 2021). A primary obstacle for abstractive methods has been the scarcity of large-scale, high-quality training datasets, as creating gold-standard summaries from hundreds of texts is prohibitively expensive. To overcome this, researchers pioneered methods that rely on synthetic datasets. For example, the work by (Amplayo and Lapata, 2021) demonstrated how to construct review-summary pairs from the original data by explicitly incorporating content planning that allows a visual training of abstractive models.

Building on this, research has explored making summaries more useful by making them controllable. (Amplayo et al., 2021) introduced aspect-controllable opinion summarization, allowing the generation of customized summaries based on user queries (e.g., focusing only on a hotel's "location" and "room"). The complexity of the task has also evolved beyond the summaries of a single entity. (Iso et al., 2022) proposed the task of a comparative opinion summarization, which generates two contrastive summaries and one common summary from reviews of two different products, directly assisting the user's choice. Further refinement of granularity, (Ge et al., 2023) developed FineSum, a framework for fine-grained, target-oriented opinion summarization that can be drilled down to sub-aspect levels with minimal supervision.

2.3 Paradigm Shift with Large Language Models

The advent of powerful Large Language Models (LLMs) has marked a significant paradigm shift. These models have demonstrated impressive zero-shot capabilities, generating high-quality summaries without task-specific fine-tuning. This shift has also highlighted the inadequacy of traditional automatic metrics like ROUGE (Lin, 2004), which often correlate poorly with human judgments of summary qual-

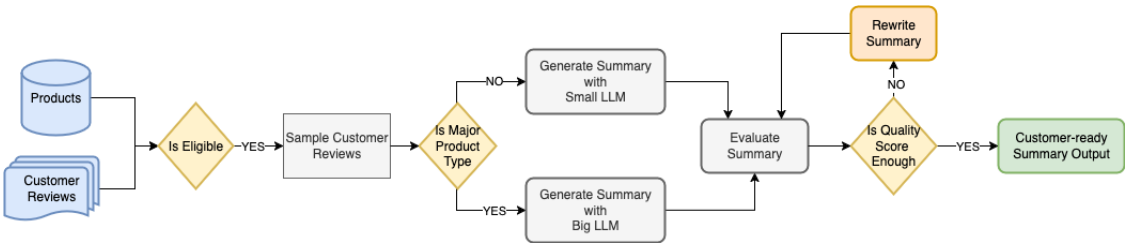

Figure 2: Overall system design of the summary generation process.

ity. A seminal work by (Stiennon et al., 2020) addressed this by learning a reward model from human preferences and using it to fine-tune a summarization model with reinforcement learning.

Subsequent studies have empirically confirmed that zero-shot GPT-3 models can produce opinion summaries that humans prefer over those from fine-tuned models, and that these summaries are less prone to dataset-specific issues like poor factuality (Bhaskar et al., 2023). However, LLMs introduce their own challenges, particularly scalability when processing hundreds of reviews and the risk of hallucination. To address this, (Hosking et al., 2023) proposed a method for attributable and scalable opinion summarization. Their model encodes review sentences into a hierarchical discrete space, allowing it to identify common opinions and generate abstractive summaries while explicitly referencing the source sentences that serve as evidence, thereby enhancing trustworthiness. Our work builds upon this new paradigm, focusing on the practical application of LLMs in a live production environment and addressing the critical, industry-relevant challenges of balancing cost, quality, and scalability.

3 SYSTEM DESIGN AND METHODOLOGY

This section covers the technical details, architectural structure, and solutions to the challenges encountered in the production environment for the ”Review Summary” system developed and deployed at Hepsiburada.

3.1 Product Selection and Review Sampling

The initial phase of the system involves identifying the products for which summaries will be generated. Given that Hepsiburada’s catalog contains millions of products across more than a thousand product types, a strategic selection process is essential. This step is critical not only for managing the operational costs associated with large-scale LLM inference, but also for maintaining a high standard of quality by avoiding summary generation for products with too few or uninformative reviews. The selection rules also account for product-specific characteristics; for example, for product types like mobile phones and computers, recent reviews are prioritized, whereas the recency of reviews is less critical for fashion products.

The system pipeline begins with determining whether a product is eligible to generate a review summary as in Figure 2. A product is selected for summary generation based on its type and the selection rules of that type. These criteria include a minimum number of verified reviews between a specific time period and several reviews’ statistics to derive an informative and objective summary.

Once a product satisfies all requirements, a sample is selected from the existing reviews to avoid exceeding the LLM’s input limits and to include the most informative comments as the next step. Although many LLM models introduce large context windows, feeding all possible reviews for the product is not scalable choice, considering the huge percentage of reviews that contain almost no product-related information and insight. Therefore, we apply a stratified sampling to maintain the review rating distribution. Using a priori of informativeness induced by review statistics, each review is selected with a probability roughly proportional to its probability of having quality information(Aslam et al., 2009).

3.2 Staged LLM Architecture

One of the biggest challenges of running an LLM-based system for millions of products in an e-commerce production environment is balancing cost and quality. Using the most powerful models (e.g., GPT-4.1, Gemini 2.5 Pro, Claude Opus 4) for every product can maximize quality, but may push operational costs to unsustainable levels. Conversely, using less expensive models can risk quality consistency. To solve this dilemma, a staged LLM architecture was designed to intelligently allocate resources based on

products value in terms of business.

The system's workflow consists of the following steps:

- *Initial Generation*: Product reviews are directed to the corresponding LLM for summary generation based on their impact. High-impact products, such as mobile phones, laptops, are routed to more capable and powerful LLMs. On the other hand, the summaries of products like books, decoration items are generated by smaller and cheaper models.
- *Automated Evaluation and Moderation*: The generated summary is passed through an evaluation step for quality and moderation steps. We leverage LLM-as-a-Judge pattern(Gu et al., 2024) using GPT-4.1, where the LLM assesses the output's the summary's compliance with predefined rule sets such as coherence, informativeness, grammar, and business constraints. This approach also automates the quality assurance process, allowing the system to scale.
- *Rewriting the Low-Quality Output*: If the quality score of the generated summary is below the determined threshold, the low-quality output is rewritten. The generated summary is fed to the same LLM with sampled customer reviews and the quality score reasoning so that the LLM can improve quality. The new summary is again evaluated by the LLM Judge. This process is repeated at most 3 loops. If a score higher than the threshold is obtained in any repetition, the output is marked as a customer-ready summary. Otherwise, the generated summary is not shown to the customers.

This staged approach provides a direct and systematic solution to generating high-quality summaries while scaling to millions of products without manual intervention.

3.3 Production Environment Challenges and Implemented Solutions

Deploying generative AI systems in a real-world, large-scale environment presents a unique set of operational challenges not typically encountered in controlled academic settings. The primary challenge is to maintain consistent summary quality given the inherent variability of LLM outputs. The most fundamental problem we faced was the generation of a high rate of low-quality summaries, especially when using more cost-effective models. The staged architecture described in Section 3.2 is our primary solution to this problem. The automated moderation and regeneration steps are designed to systematically manage and uphold output quality across the entire product catalog.

A second significant challenge stems from the quality of the input data itself. The "Garbage In, Garbage Out" principle applies directly, as very short ("great product"), irrelevant ("bought for my wedding"), or nonsensical reviews negatively affect the LLM's ability to produce a meaningful summary. Although, there is no simple solution to this issue, we investigated mitigation strategies, such as filtering reviews based on their length, information density, or the number of "helpful" votes they have received from other users. During review sampling process, we used such statistics to increase the probability of providing more informative customers reviews to generate better summaries.

Finally, a critical issue is post-deployment quality control. To prevent low-quality or uninformative summaries from ever being displayed to customers, we implemented a proactive quality gate as part of our pre-deployment process. This was achieved by leveraging the final quality score assigned by the Judge LLM during the moderation step. Any summary that fails to meet a predetermined minimum quality score threshold is automatically discarded and not published. This automated filtering mechanism effectively solves the problem of inconsistent quality at the source, ensuring that only summaries meeting our standards reach the end-user.

3.4 Context Engineering and Prompts

The behavior of the LLMs at each stage of the pipeline is controlled by carefully engineered prompts. These prompts are a core component of the system's methodology, defining the task, constraints, and desired output format for the models.

The summary generation prompt is used in both generation steps. It instructs the model to create a persuasive and objective summary in Turkish, with specific constraints on length, content, tone, and exclusions (e.g., no mention of shipping, packaging, or sellers). The complete prompt is provided in Appendix A.1.

The evaluation prompt is used in the automated quality control step. It tasks the LLM to act as a judge, validating a given summary against a strict set of rules. These rules include validating the language, word count, point of view, and ensuring that the summary is free of excluded topics and sentiment bias. The prompt is provided in Appendix A.2.

The paraphrasing prompt is used during the re-

generation process when a summary fails to pass the quality control. It instructs the model to rewrite the faulty summary, addressing the specific issues identified by the moderation step while adhering to a stricter set of rules, such as a shorter word limit. The prompt is provided in Appendix A.3.

4 EXPERIMENTAL SETUP AND EVALUATION

Recognizing that traditional metrics such as ROUGE often correlate poorly with human quality judgments for abstractive tasks (Stiennon et al., 2020), we define a composite quality score ranging from 1 to 10. This score is calculated as a weighted average of four fundamental human-centered dimensions of summary quality (Li et al., 2023). The dimensions and their respective weights are as follows:

- *Helpfulness* of the summary measures whether the summary would help a user make an informed purchasing decision. This is the most critical dimension as it directly relates to the primary business objective.
- *Factuality* evaluates whether the generated summary avoids making claims not supported by the source reviews (hallucination). Factual consistency is crucial to maintain user trust (Kryściński et al., 2020).
- *Relevance* accurately reflects the main ideas and important points of the source reviews, focusing on product-centric aspects.
- *Coherence* checks the summary's overall grammatical structure and understandability.
- *Compliance* evaluates the summary to determine whether it violates any defined business rules and regulations.

This weighted approach allows for a nuanced evaluation that prioritizes the aspects most critical to the e-commerce use case, such as truthfulness and usefulness.

4.1 Comparative Model Analysis

The evaluation is designed to cover various prominent LLMs in industry and academia that have proficiency in Turkish language and e-commerce domain. The models to be compared are: GPT-4.1, Gemini 2.5 Pro, Gemini 2.5 Flash Lite, and a local model, Trendyol LLM(Team, 2025). 1000 unique products are sampled from varying product types and categories to ensure diversity in the test set and observe the LLMs' behaviors when they are given different business constraints. To ensure a fair comparison, each model receives the same set of products and the same reviews for those products. The same prompt structure will be used for all models to isolate inter-model performance.

4.2 Evaluation Protocol

The evaluation process follows a dual approach that combines the depth of human judgment with the scalability of automated methods.

- *Human Evaluation*: A group of human annotators trained on a set of test output summaries given the rule sets. The annotators are presented with a product, its reviews, and a summary generated by one of the models. In a blind test setup where they do not know which model produced which summary, they are asked to score based on the criteria defined above.
- *LLM-as-a-Judge Evaluation*: In parallel with human evaluation, the automated evaluation step is established as a scalable alternative. The final quality score that is calculated by the LLM Judge is compared to the human judgments. This comparison provides an opportunity to understand the consistency between human and LLM.

4.3 Results and Analysis

The analysis involves comparing the mean quality scores achieved by each model for the test set. The results are presented in Table 1.

Table 1: Average quality score obtained by each model evaluated by both human and LLM annotators.

Model	Evaluator	Avg. Quality Score
GPT-4.1	Human	9.23
	LLM	9.05
Gemini 2.5 Pro	Human	9.31
	LLM	9.17
Gemini 2.5 Flash Lite	Human	8.80
	LLM	8.69
Trendyol LLM	Human	4.52
	LLM	6.58

The comparative evaluation yielded clear performance differences among the models, as summarized in Table 1. The results indicate that Gemini 2.5 Pro achieved the highest score of the human annotators (9.31) and the LLM (9.17), positioning it as the model that performs the best in this evaluation. GPT-4.1 also showed strong performance, with scores closer

to Gemini 2.5 Pro. Gemini 2.5 Flash Lite, while still effective, scored slightly lower, consistent with its design as a more lightweight model.

A particularly noteworthy finding involves the Trendyol LLM. Despite being a domain-specific model fine-tuned by another major e-commerce company in Türkiye, it struggled to generate high-quality summaries, receiving a significantly lower score from human evaluators (4.52) compared to general-purpose models. Interestingly, the LLM Judged evaluated it more generously (6.58), suggesting a potential difference in the way automated and human evaluators perceive failures in domain-specific models.

Based on these results, we made our final model selections for the production environment. For the premium generation tier, despite Gemini 2.5 Pro achieving the highest quality score, we selected *GPT-4.1*. This decision was driven by two key factors. Firstly, we observed that Gemini 2.5 Pro, as a reasoning model, sometimes produced summaries with a more pronounced positive or negative sentiment shift, whereas our goal is to provide users with a balanced and objective summaries. Quantitatively, GPT-4.1's non-reasoning architecture offers lower latency and cost, which are critical considerations for a large-scale production system. For the cost-effective part, the choice was more straightforward. Given the significant performance difference observed in our evaluation, **Gemini 2.5 Flash Lite** was selected over Trendyol LLM to ensure a baseline of high-quality summaries for all eligible products.

To provide a qualitative illustration of these performance differences, example summaries generated by all experimental models are presented in Appendix B.

5 CONCLUSION AND FUTURE WORK

This study has presented a case study on the design, implementation, and production challenges of an LLM-based user review summarization system on a large-scale e-commerce platform. The main contributions are the demonstration of the effectiveness of a staged LLM architecture that balances cost and quality, and the transparent documentation of real-world operational challenges. The developed system provides a more efficient shopping experience for millions of users, serving as a successful example of the conversion of modern AI applications into commercial value.

Future work will focus on addressing the system's current limitations and further expanding its capabilities. A primary direction is to improve input filtering by developing more advanced methods to automatically identify and exclude low-quality or uninformative reviews before they are sent to the LLM. To further enhance summary quality and align the system with user needs, we also plan to implement a direct user feedback loop, which would allow users to rate the usefulness of summaries. This data could then be used for continuous system improvement. Looking further ahead, we will explore personalization by integrating user-specific data to produce summaries that highlight product features most relevant to an individual's interests. Finally, the potential to create more holistic product overviews will be investigated by incorporating multi-modal information, such as insights derived from user-uploaded images and videos.

REFERENCES

Amplayo, R. K., Angelidis, S., and Lapata, M. (2021). Aspect-controllable opinion summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6578–6593, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Amplayo, R. K. and Lapata, M. (2021). Unsupervised opinion summarization with content planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12338–12346.

Angelidis, S. and Lapata, M. (2018). Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686, Brussels, Belgium. Association for Computational Linguistics.

Aslam, J. A., Kanoulas, E., Pavlu, V., Savev, S., and Yilmaz, E. (2009). Document selection methodologies for efficient and effective learning-to-rank. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 468–475.

Bhaskar, A., Ladhak, F., Ladhak, A., Yih, W.-t., Dernoncourt, F., and Yu, M. (2023). Zero-shot gpt-3 for opinion summarization.

Ge, S., Huang, J., Meng, Y., and Han, J. (2023). FineSum: Target-oriented, fine-grained opinion summarization. In *Proceedings of the 16th ACM International Conference on Web Search and Data Mining*, pages 1093–1101, Singapore, Singapore. ACM.

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., et al. (2024). A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Hosking, T., Tang, H., and Lapata, M. (2023). Attributable and scalable opinion summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

pages 8488–8505, Toronto, Canada. Association for Computational Linguistics.

Iso, H., Morishita, T., Higashinaka, R., and Minami, Y. (2021). Unsupervised opinion summarization with tree-structured topic guidance. *Transactions of the Association for Computational Linguistics*, 9:945–961.

Iso, H., Morishita, T., Higashinaka, R., and Minami, Y. (2022). Comparative opinion summarization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3307–3318, Dublin, Ireland. Association for Computational Linguistics.

Kryściński, W., McCann, B., Xiong, C., and Socher, R. (2020). Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Li, K., Zhang, S., Ge, T., Wang, Y., Zhang, Z., Wang, A., Wang, J., Wang, G., Feng, Y., and Wang, W. (2023). CONNER: A COmpreheNsive kNowledge Evaluation fRamework for assessing large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5827–5842, Singapore. Association for Computational Linguistics.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Pecar, S. (2018). Towards opinion summarization of customer reviews. In *Proceedings of ACL 2018, Student Research Workshop*, pages 1–8, Melbourne, Australia. Association for Computational Linguistics.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. (2020). Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021.

Team, T. L. . C. N. (2025). Trendyol llm 8b t1.

Titov, I. and McDonald, R. (2008). Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th International Conference on World Wide Web*, pages 111–120. ACM.

APPENDIX A.1 SUMMARY GENERATION PROMPT

```
Your primary task is to generate a concise,
persuasive, and objective summary in
Turkish, synthesized from the provided
customer reviews.  You are to follow a
structured methodology, beginning with a
sentiment-balanced analysis of the reviews
to extract key product-centric features.
Prioritize the inclusion of positive
attributes while integrating a necessary
amount of constructive negative feedback
to ensure an objective overview; if no
substantive negative feedback is present,
construct the summary solely from positive
points without referencing the absence of
criticism.
A crucial part of this process is the
application of strict content filtering
criteria:  you must focus exclusively on the
product's intrinsic qualities and performance,
and are directed to disregard and exclude all
logistical and service-related comments, such
as those concerning shipping, packaging, or
seller interactions.
Furthermore, to ensure the summary is broadly
applicable, normalize domain-specific
information by generalizing or omitting
variant-specific details like color or exact
sizing for fashion items, and frame any
discussions of price in terms of overall value
or price-to-performance ratio.
Finally, adhere to several stylistic and
formal constraints:  adopt a neutral and
objective tone, maintain a third-person
narrative perspective by attributing all
opinions to "users," and ensure the final
output is grammatically correct and adheres to
a strict word count limit.
```

APPENDIX A.2 EVALUATION PROMPT

```
Your function is to perform a comprehensive
validation of the provided Turkish summary,
ensuring its strict adherence to a series of
critical guidelines.  You must first verify
its compliance with formal constraints:  the
summary must be written in Turkish, remain
within word limit, and consistently use a
third-person point of view.  The core of your
assessment will be content-based, confirming
that the summary focuses exclusively on
product-related features, advantages, and
disadvantages, with a special allowance for
discussions of a book's content if it is the
product in question.
It is imperative to penalize any deviation
from the strict exclusion criteria, which
prohibit any mention of topics such as cargo,
packaging, sellers, or delivery, as well as
gifts, or health suggestions.  Furthermore,
you will ensure that any recommendations
are explicitly attributed to "users," that
personal experiences are generalized to
reflect broader customer sentiment, and
that conflicting feedback is handled with
objectivity.
Finally, you must evaluate the summary's
overall sentiment, applying a significant
penalty if it is excessively negative or lacks
a constructive, positive perspective, as the
goal is to inform, not deter.
```

APPENDIX A.3 PARAPHRASE PROMPT

```
Your primary function is to revise a
provided summary that has been identified
as non-compliant, guided by the specific
issues detailed in the accompanying critique.
Using the original summary and the source
customer reviews, you must generate a new,
valid summary in Turkish that corrects all
noted deficiencies while adhering to a strict
word limit or fewer.
In this process, you are required to
generalize any individual user experiences
to reflect a broader customer consensus and
to rephrase any recommendations to ensure
they are explicitly attributed to "users."
All sentences must be composed in the third
person.
A critical component of your task is to
improve conciseness by shortening sentences
and removing all non-product-related
information, including any mentions of
sellers, shipping, packaging, delivery, gifts,
or health suggestions.
Conclude by proofreading the final summary
to verify its full compliance with all
instructions and its correctness in grammar
and spelling.
```

APPENDIX B. EXAMPLE SUMMARIES

Table 2 and Table 3 provide example outputs for the same product for all LLM models used during our experiments to illustrate the qualitative differences in summary generation.

Table 2: Comparative summary outputs of each model for a thermos both in Turkish and English.

Product	Model	Summary (Turkish)	Summary (English)
Stanley Klasik Vakumlu Çelik Termos 1,4 L	GPT-4.1	Stanley Klasik Vakumlu Çelik Termos, geniş hacmi ve sağlam yapısıyla kullanıcıların beğenisini kazanıyor. Termosun sıcak ve soğuk tutma performansı, kullanıcılar tarafından oldukça başarılı bulunuyor. Fiyatının performansını karşıladığı belirtiliyor. Bazı kullanıcılar kapak ve alt kısımdaki eziklerden şikayet etmiş olsa da, genel olarak ürünün dayanıklılığı ve işlevselliği öne çıkıyor. Kullanıcılar, termosu pratikliği ve uzun süreli kullanımı için öneriyor.	The Stanley Classic Vacuum Insulated Steel Thermos wins users' appreciation with its large volume and sturdy structure. The thermos's performance in keeping drinks hot and cold is found to be quite successful by users. It is stated that its price is justified by its performance. Although some users have complained about dents on the lid and bottom, the product's durability and functionality generally stand out. Users recommend the thermos for its practicality and long-term use.
	Gemini 2.5 Pro	Stanley Klasik Vakumlu Çelik Termos, kullanıcılar tarafından geniş iç hacmi ve sağlamlığıyla tavsiye ediliyor. Fiyatına göre yüksek performans sunduğu belirtilen ürün, kullanıcıların beklentilerini karşılıyor. Bazı ürünlerde ezikler olabildiği, ancak genel olarak kalitesinin tatmin edici olduğu ifade ediliyor.	The Stanley Classic Vacuum Insulated Steel Thermos is recommended by users for its large interior volume and robustness. The product, which is stated to offer high performance for its price, meets users' expectations. It is mentioned that some products may have dents, but overall, its quality is expressed as satisfactory.
	Gemini 2.5 Flash Lite	Stanley Classic termos, kullanıcılar tarafından genellikle büyük ve kaliteli olarak tanımlanıyor. Dayanıklılığıyla öne çıkan ürün, fiyatına değdiği düşünülüyor. Bazı kullanıcılar ürünün kapağında veya alt kısmında ezikler olduğunu belirtse de, genel olarak memnuniyet yüksek. Bu termos, sıcak ve soğuk içeceklerinizi uzun süre muhafaza etmek için güvenilir bir seçenek olarak öne çıkıyor.	The Stanley Classic thermos is generally described by users as large and of high quality. The product, which stands out for its durability, is considered to be worth its price. Although some users mention that there are dents on the lid or bottom of the product, the overall satisfaction is high. This thermos stands out as a reliable option for preserving your hot and cold beverages for a long time.
	Trendyol LLM	Stanley Klasik Vakumlu Çelik Termos, 1,4 L Mavi.	Stanley Classic Vacuum Insulated Steel Thermos, 1.4 L Blue

Table 3: Comparative summary outputs of each model for a headset both in Turkish and English.

Product	Model	Summary (Turkish)	Summary (English)
Razer Kraken V3 RGB 7.1 Surround Siyah Gaming Kablolu Kulaklık	GPT-4.1	Razer Kraken V3, kullanıcıların belirttiği üzere yüksek ses kalitesi ve 7.1 surround özelliğiyle oyunlarda gerçekçi bir deneyim sunuyor. Mikrofonun çıkarılabilir olması ve kafa bandının rahatlığı, kullanım kolaylığını artırıyor. Ekolayzır ayarları sayesinde kişiselleştirilebilir ses imkanı sağlıyor. Ancak bazı kullanıcılar ürünü ağır bulmuş ve uzun süreli kullanımda rahatsızlık yaşadıklarını belirtmişler. Ayrıca, nadiren cızırtı ve mikrofon kalitesiyle ilgili olumsuz yorumlar da mevcut.	According to users, the Razer Kraken V3 offers a realistic gaming experience with its high sound quality and 7.1 surround feature. The detachability of the microphone and the comfort of the headband increase ease of use. It provides customizable sound options thanks to its equalizer settings. However, some users found the product heavy and stated they experienced discomfort during long-term use. Additionally, there are rare negative comments about crackling and microphone quality.
	Gemini 2.5 Pro	Razer Kraken V3, özellikle oyuncular için tasarlanmış bir kulaklık olarak öne çıkıyor. Kullanıcılar, ürünün 7.1 surround ses kalitesi sayesinde oyun içi sesleri tüm detaylarıyla duyabildiklerini ve bu deneyimin ürünü fiyatına değer kıldığını belirtiyor. Aynı zamanda, çıkarılabilir mikrofonu ve başı saran konforlu tasarımı da beğenilen özellikleri arasında. Bazı kullanıcılar, kulaklığın uzun süreli kullanımlarda ağır gelebildiğini ve yazılım uygulamasının biraz karmaşık olabildiğini ekliyor.	The Razer Kraken V3 stands out as a headset designed especially for gamers. Users state that thanks to the product's 7.1 surround sound quality, they can hear in-game sounds in full detail, and this experience makes the product worth its price. At the same time, its detachable microphone and comfortable design that envelops the head are among its liked features. Some users add that the headset can feel heavy during prolonged use and that the software application can be a bit complicated.
	Gemini 2.5 Flash Lite	Razer Kraken V3 RGB oyuncu kulaklığı, sunduğu üstün ses kalitesiyle oyun deneyimini bir üst seviyeye taşıyor. Kullanıcılar, özellikle 7.1 surround ses özelliği sayesinde oyun içindeki sesleri daha net duyduklarını belirtiyor. Mikrofonun performansı da video ve yayınlar için yeterli bulunuyor. Ancak, bazı kullanıcılar kulaklığın biraz ağır olmasından ve uzun süreli kullanımlarda baş ağrısına neden olabileceğinden şikayetçi. Uygulamanın arayüzü de bazı kullanıcılar için karmaşık gelebilir. Genel olarak, ürün fiyatına değer bir performans sunuyor.	The Razer Kraken V3 RGB gaming headset takes the gaming experience to the next level with the superior sound quality it offers. Users state that they hear in-game sounds more clearly, especially thanks to the 7.1 surround sound feature. The microphone's performance is also found to be sufficient for videos and streaming. However, some users complain that the headset is a bit heavy and can cause headaches during long-term use. The application's interface can also seem complicated to some users. Overall, the product offers performance worthy of its price.
	Trendyol LLM	Razer Kraken V3 RGB 7.1 Surround Gaming Kulaklık, oyun deneyimini zenginleştirmek için tasarlanmıştır. Yüksek kaliteli ses ve net iletişim sağlar, ancak uzun süreli kullanım kafa ağrısı yaratabilir. Mikrofon özellikleri iyi olsa da bazı kullanıcılar cızırtı sorunu yaşamıştır. Fiyatına göre genel olarak memnuniyet seviyesi orta düzeydedir.	The Razer Kraken V3 RGB 7.1 Surround Gaming Headset is designed to enrich the gaming experience. It provides high-quality sound and clear communication, but long-term use may cause headaches. Although the microphone features are good, some users have experienced a crackling issue. For its price, the general satisfaction level is moderate.