Towards Intention-aligned Reviews Summarization: Enhancing LLM Outputs with Pragmatic Cues

María Miró Maestre[1], Robiert Sepúlveda-Torres[1], Ernesto L. Estevanell-Valladares[1,2], Armando Suárez Cueto[1], Elena Lloret[1]
[1]Department of Software and Computing Systems, University of Alicante, Spain
[2]University of Havana, Cuba
{maria.miro, ernesto.estevanell}@ua.es,
{rsepulveda, armando, elloret}@dlsi.ua.es

Abstract

Recent advancements in Natural Language Processing (NLP) have allowed systems to address complex tasks involving cultural knowledge, multi-step reasoning, and inference. While significant progress has been made in text summarization guided by specific instructions or stylistic cues, the integration of pragmatic aspects like communicative intentions remains underexplored, particularly in non-English languages. This study emphasizes communicative intentions as central to summary generation, classifying Spanish product reviews by intent and using prompt engineering to produce intention-aligned summaries. Results indicate challenges for large language models (LLMs) in processing extensive document clusters, with summarization accuracy heavily dependent on prior model exposure to similar intentions. Common intentions such as complimenting and criticizing are reliably handled, whereas less frequent ones like promising or questioning pose greater difficulties. These findings suggest that integrating communicative intentions into summarization tasks can significantly enhance summary relevance and clarity, thereby improving user experience in product review analysis.

1 Introduction

A central objective of Natural Language Processing (NLP) is to develop automatic systems that emulate human-like language understanding and production (Chowdhary, 2020). Yet, current models—particularly Large Language Models (LLMs)—often neglect complex yet essential pragmatic dimensions crucial for authentic communication. The NLP community has acknowledged this limitation and is increasingly investigating how pragmatic phenomena, especially communicative intentions, shape both the interpretation of messages and the linguistic choices made to convey those intentions—an aspect vital for Natural Language Generation (NLG) tasks such as text summarization (Khurana et al., 2023; Fried et al., 2023).

Although communicative intentions are central to language structuring, their integration into generative systems for NLG tasks remains underexplored. This is especially evident in text summarization, where intentionality is key to ensuring coherence between the source text and its summary. Despite its importance, limited efforts have addressed this aspect—particularly in languages beyond English. This study examines how well various LLMs incorporate communicative intentions in multi-document summarization using Spanish product reviews. By applying prompt engineering techniques and automatic intent classification, we assess whether the generated summaries reflect the communicative goals present in the original reviews.

Overall, the main contributions of this paper are:

- Automatic intention classification in a Spanish reviews corpus.
- Comparative analysis of prompt engineering strategies for intention-driven summary generation.
- Comparative evaluation of generative models for intention-aligned text summarization.
- Application of sentence similarity metrics to evaluate alignment between generated summaries and communicative intentions determined in prompts.

The remainder of the paper is organized as follows: Section 2 reviews research on integrating communicative intentions in generative systems. Section 3 details the Spanish corpus, intentions tested, LLMs evaluated, and prompts used to guide

Proceedings of Recent Advances in Natural Language Processing,pages 737–747
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-085

intention-aligned summary generation. Then, Section 4 shows the intention-aligned summary generation task results and some features of adequately and inadequately generated summaries. The sentence similarity evaluation methodology used to compare the alignment between summaries and original review excerpts is outlined in Section 5. Finally, Section 6 summarizes key findings future research directions to further incorporate pragmatic phenomena into language generation tasks.

2 Related Work

In NLP, the pragmatic layer has often been overlooked in favor of lower linguistic levels such as syntax and semantics, largely due to its inherent complexity and reliance on contextual cues for meaning interpretation (Levinson, 2017). Within pragmatics, communicative intentions are especially vital, shaping both message formulation and comprehension (Ghosal et al., 2021). Their significance has been extensively examined through various theoretical frameworks and taxonomies by scholars such as Austin (1962); Searle (1969); Halliday and Hasan (1986); Sperber and Wilson (1986); Archer et al. (2008); Bach (2012); Escandell Vidal et al. (2020).

Communicative intentions are increasingly recognized as essential in NLP, particularly for NLG, where they guide linguistic choices to convey specific meanings. Recent advances in LLMs have strengthened their ability to handle complex tasks involving multi-step reasoning and inference (Dong et al., 2022). A prominent example is Automatic Text Summarization (ATS) (El-Kassas et al., 2021), which condenses large volumes of text—such as news, medical, or product reviews—into concise, contextually appropriate summaries. This is especially critical in abstractive summarization, which rephrases key information to generate novel, succinct outputs (Mridha et al., 2021).

Recent advances in generative systems have led to novel ATS approaches that integrate pragmatic features such as style (Goyal et al., 2022), polarity (Amoudi et al., 2022), and vocabulary (Balde et al., 2024). Yet, incorporating communicative intentions—crucial for generating contextually appropriate and purposeful summaries—remains an emerging area. Notable efforts include summarizing app reviews based on categorized intentions (Di Sorbo et al., 2016), leveraging semantic-intent graphs in multilingual summarization (Wang et al., 2019), and generating intention-aligned summaries for health queries (Zhang and Liu, 2022) and code descriptions (Geng et al., 2024). Despite this progress, no prior work has addressed intention-based summarization in Spanish.

3 Experimental Setup

Multi-document summarization is gaining prominence in NLP as users increasingly face large volumes of online content when searching for topic-specific information (Ma et al., 2022). This creates a growing need for systems that can generate concise, user-tailored summaries. In this study, reviews within each product category are aggregated into unified texts to evaluate intent-driven summary generation across multiple LLMs, chosen for their ability to handle long-context inputs encompassing all relevant review content.

Figure 1 illustrates the procedural steps for automatically generating summaries according to specific communicative intentions. The subsequent sections detail the dataset's characteristics, the targeted communicative intents, model selection, and the design of prompts for interacting effectively with these models when generating intention-aligned review summaries.

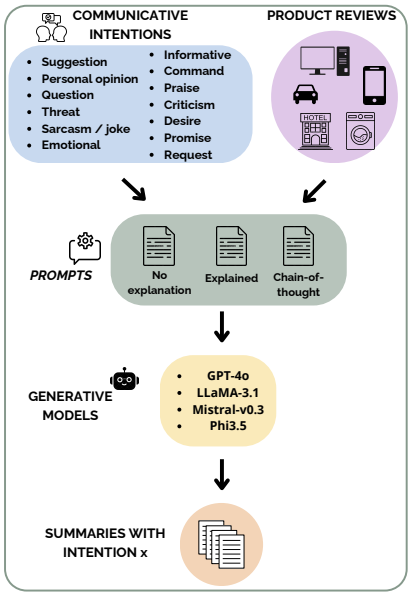


Figure 1: Experimental setup to generate intention-aligned summaries of product reviews in Spanish.

3.1 Spanish dataset

This study evaluates intention-aligned summary generation in Spanish, requiring a dataset with texts that reflect diverse communicative intentions

suitable for a fine-grained taxonomy. After surveying available corpora, we selected product reviews for their rich variety of intentions—such as descriptions, suggestions, opinions, praise, criticism, threats, or sarcasm. Specifically, we used the SFU Review Corpus (Taboada and Grieve, 2004), which contains 400 reviews from the Ciao.es[1] forum across categories like cars, hotels, washing machines, mobile phones, and computers.[2] Each category includes 50 reviews, merged into a single document for summarization. These consolidated reviews were then used to generate intention-aligned summaries as described in the following section. We excluded three categories—books, music, and films—because their aggregated reviews exceeded the input limits of some LLMs (over 100 000 characters), rendering them incompatible.

3.2 Intention taxonomy

As discussed in Section 2, many taxonomies of communicative intentions have been proposed, though often with limited applicability to real-world scenarios (Ma et al., 2025). Given that our dataset stems from computer-mediated communication (CMC)—which encompasses digital textual genres found online—, we adopted the taxonomy by Maestre et al. (2025), which defines 13 intention categories suited to CMC contexts: "suggestion", "personal opinion", "question", "threat", "sarcasm / joke", "emotional", "informative", "command", "praise", "criticism", "desire", "promise", and "request". We believe these categories effectively capture the variety of intentions in our review corpus. Moreover, the absence of certain intentions in the original texts provides a useful lens to assess whether LLMs appropriately reflect this absence or introduce hallucinated content when prompted to include specific intentions in the generated summaries.

3.3 Generative models

The selection of LLMs evaluated in this summarization task was influenced by the necessity for a wide context window capable of processing extensive integrated reviews from a product category into a single document. Consequently, four LLMs were chosen: GPT-4o (Achiam et al., 2023), LLaMA-3.1-8B (Touvron et al., 2023), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), and Phi3.5-3.8B (Abdin et al., 2024). This selection ensures diversity by incorporating both proprietary and open-source models capable of handling lengthy contextual inputs required for summarization tasks.

More concretely, GPT-4o was selected due to its demonstrated capabilities across various NLP tasks, as extensively validated by existing research (Qin et al., 2023; Amin et al., 2023; Hariri, 2023). It was accessed via OpenAI's online interface.[3] The remaining models—LLaMA-3, Mistral, and Phi3.5—were evaluated using the Ollama platform,[4] an optimized local tool facilitating interaction with open-source LLMs through Python, thus simplifying their configuration according to specific research objectives.

Interaction with the models included in Ollama utilized a client-server setup, configured with a maximum context length of 33 000 tokens, sufficient for the unified review documents. Prompts used to guide summary generation were externally loaded based on the communicative intention and the prompt type chosen to generate each summary, employing an iterative process for generating intention-aligned summaries across the different product categories. Detailed descriptions of the prompt types employed in this research are provided in the subsequent section.

3.4 Prompt engineering techniques

To incorporate communicative intentions into automatic summary generation by LLMs, this study employed prompt engineering—a technique involving carefully designed and refined instructions that guide model interactions to better fulfill specific tasks (Marvin et al., 2023). Prompt engineering is widely regarded in the research community as critical for effectively leveraging LLM capabilities in different tasks (Sahoo et al., 2024).

This study evaluates LLMs' ability to generate summaries of product reviews that align with specific communicative intentions, focusing on whether relevant excerpts are accurately identified and represented. To this end, prompts were crafted to include the target intention, the set of reviews for analysis, and stylistic and formatting guidelines. As shown in Figure 2, three prompt versions were developed,[5] differing in the level of detail provided.

[1] Link to the website no longer available.

[2] Corpus available at: https://www.sfu.ca/~mtaboada/SFU_Review_Corpus.html (Last accessed 22 May 2025)

[3] https://chat.openai.com/

[4] Ollama can be downloaded local through their GitHub website: https://github.com/ollama/ollama.

[5] Prompts were originally written in Spanish. We translated

These variations aim to assess whether LLMs can infer Spanish pragmatic cues or require explicit instructions to effectively detect and reproduce communicative intentions.

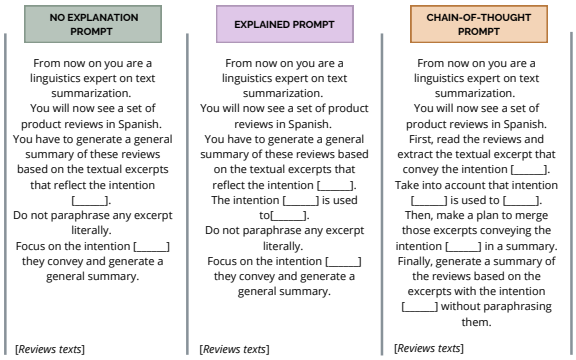


Figure 2: Prompts created to generate review summaries according to each intention category.

The three distinct prompt variations created initially state the task and specify the intention the summary should convey. These prompts differ in the amount of explanatory information provided to define the communicative intention. The first variation, "no explanation prompt", includes only the intention's name, testing the extent to which LLMs inherently understand and recognize communicative intentions without additional context. The second, "explained prompt", provides explicit definitions derived from the intention taxonomy used in this study. The third version employs the "chain-of-thought" (CoT)(Wei et al., 2022) prompting strategy, which has demonstrated significant effectiveness across various NLP tasks (Yu et al., 2023) requiring complex reasoning, as it instructs LLMs through intermediate reasoning steps to fulfill the task accurately.

An extensive review of recent studies on CoT prompting (Wei et al., 2022; Crispino et al., 2023; Li et al., 2024; Wang et al., 2023) informed the selection of a suitable prompt structure. Particularly, the "plan-and-solve" strategy by Wang et al. (2023), which explicitly instructs models to reflect on tasks and develop a structured approach, was considered highly appropriate. This approach involves initially extracting relevant excerpts aligned with specified summary requirements, a method previously validated with successful results by Gao et al. (2020) and Wu et al. (2024). Thus, we divided the prompt in different reasoning steps, these being: identifying the determined intention in the original texts, selecting relevant textual excerpts that convey that intention, and generating intention-aligned summaries.

4 Review Summary Generation

Following the methodology outlined in the previous section, we generated a total of 780 review summaries (13 intentions $\times$ 4 LLMs $\times$ 3 prompt techniques $\times$ 5 product categories). All original product category reviews, along with their intention-specific summaries, the LLMs employed, and the prompts used in this study, are available in the following repository: https://github.com/Maria3mmm/Intent_Summ_Es-

During summary generation, a qualitative analysis was conducted to evaluate the alignment of communicative intentions between the generated summaries and the original reviews. This analysis aimed to identify which intentions were effectively captured by generative models, following the intention categories defined for this study.

Table 1 illustrates an example where summaries generated by Mistral-v0.3 and LLaMA-3.1 models for computers and mobile phone reviews failed to convey the specified "promise" intention. Instead of accurately including excerpts such as "I promise I would never try to buy again a product form this brand" or "This mobile should be fast-responding but I swear it's just not good" that match the intention requested in the prompt and were explicitly stated in original reviews, these summaries predominantly focused on product descriptions or sentiment-related features, indicating a closer alignment with sentiment classification tasks rather than the intention-aligned generation task.

Then, the results presented in Table 2 illustrate examples where the summaries generated by Phi3.5 and GPT-4o successfully align with the communicative intention of "praise" found in hotel and computer product reviews. Both models accurately highlighted positive user feedback, with GPT-4o notably acknowledging that despite prevalent negative comments, positive remarks still exist and subsequently enumerating them. This outcome aligns precisely with the objectives of the intention-aligned summarization experiment. The authors attribute this accuracy partially to the extensive prior exposure these models have had to datasets involving positive and negative sentiments, stemming from sentiment analysis being a well-established NLP research area, thus providing a strong founda-

them into English for clarity.

Summary excerpt of mobile reviews with "no explanation" prompt generated by Mistral-v0.3	Summary excerpt of hotel reviews with "explained" prompt generated by LLaMA-3.1
The Motorola C261 is a mid-range mobile phone with interesting features and a design you won't forget. Although its price is relatively high, it offers a wide range of capabilities, making it a highly recommended purchase if you're looking for a very complete and high-quality mobile phone. The Motorola C261 is a fully-featured PDA (and mobile phone), featuring a 2.8-inch screen with a resolution of 320x240 pixels, a 520 MHz Intel XScale processor, and the Windows Mobile 5 system."	After reviewing the comments on various hotels in Madrid, I can summarize the main problems found as follows. 1-Room quality: Many hotels have small, outdated rooms lacking basic comforts.2. Service issues: Some hotels offer poor reception services, with staff that can be unfriendly or inefficient. [...] However, there are also positive reviews highlighting (1) Location: Many hotels are conveniently located near tourist attractions with easy access to public transportation. (2) Staff: Some hotels have friendly and attentive personnel who enhance the traveler's experience. [...]

Table 1: Inaccurate summaries generated for "promise" intention.

tion for addressing newer tasks such as intention-driven summary generation.

Summary excerpt of hotel reviews with "no explanation" prompt generated by Phi3.5	Summary excerpt of computer reviews with "no explanation" prompt generated by GPT-4o
The generalized summaries of praise intentions based on reviews highlight a consistent recommendation for a hotel located next to the Congreso de los Diputados and close to Gran Vía, offering amenities such as proximity to Madrid's cultural center, diverse restaurants, and entertainment options. Guests particularly praise its excellent value for money and modern facilities within a hotel context not yet fully renovated at that time (2007), [...]	The analysis of the reviews mainly highlights negative criticisms towards different computer brands. However, focusing on the few positive mentions, the following aspects stand out: (1) Value for money in certain models: Some users appreciate computers offering reasonable specifications at affordable prices, suitable for basic home or business tasks. (2) Good technical support experiences in some cases: Although negative feedback predominates, there were isolated mentions of quick and efficient technical support that successfully resolved specific issues, such as replacing defective components. [...]

Table 2: Accurate summaries generated for "praise" intention.

After manually analyzing summaries generated according to the thirteen communicative intentions defined for this study, distinct patterns emerged regarding which intentions were accurately or inaccurately conveyed by the LLMs. Intentions such as "question", "threat", "promise", "desire", "request", "sarcasm / joke", and "command" generally presented greater challenges for accurate summary generation, irrespective of the model used. Summaries for these intentions often involved adaptations of original reviews that artificially reflected the stated intention rather than accurately identifying and reproducing the intention as originally expressed. In other words, the models frequently hallucinated by rewriting texts to reflect an intended message instead of extracting authentic intention-driven sentences from the original reviews that were manually identified when reviewing the summaries and their respective reviews. Additionally, models often confused intention categories when a particular intention was absent from the original reviews, consequently defaulting to summarizing reviews based on easily identifiable intentions, basing on those related to polarity, such as "praise", "criticism" or "suggestion".

Based on the previous observations, certain intention categories such as "suggestion", "praise", "emotional", and "criticism" were effectively conveyed in the generated summaries. These intentions are closely associated with user-generated product reviews, where texts typically recommend or dissuade the use of a particular product or service. Consequently, excerpts reflecting these intentions frequently appear in original texts. The high accuracy of summaries capturing these intentions may also be attributed to the extensive exposure of LLMs to related linguistic expressions, notably in sentiment and polarity analysis tasks. Indeed, LLMs can precisely enumerate emotions and sentiments present in reviews, even without explicit prompting, due to their familiarity with such analytical tasks.

Then, the intentions "personal opinion" and "informative" present mixed results, with summaries occasionally accurately reflecting these intentions and other times failing to do so clearly. This inconsistency often arises from the tendency of LLMs to interpret both intentions similarly, particularly when using the "no explanation" prompt, which does not explicitly provide a definition to differentiate them. Consequently, generated summaries for these intentions often resemble each other closely, typically focusing on the advantages and disadvantages of the reviewed products without distinct intentional differentiation.

5 Intention-aligned Summary Evaluation

In addition to the previous qualitative analysis of summaries generated by four language models, a more rigorous evaluation was necessary to precisely assess each model's performance and the prompt strategies employed for the 13 predefined communicative intentions. For this purpose, the study utilized two variants of the cosine similarity metric to measure semantic similarity. The rationale behind this approach is that summaries aligned with specific intentions should include excerpts

from the original texts expressing those same intentions, thereby displaying similar linguistic features.

The evaluation method, illustrated in Figure 3, involves first automatically segmenting the original reviews at the sentence level for each product category. Subsequently, these sentences are classified according to their communicative intention using GPT-4o. Finally, sentences sharing the same intention are grouped and semantically compared with summaries generated by each language model and prompt strategy for the same intention through cosine similarity and BERT-based similarity metrics. Detailed descriptions of each evaluation step are provided in the subsequent subsections.

5.1 Text segmentation and intention classification

The evaluation began by segmenting each product review text (from a set of 50 reviews per product) into individual sentences, using Spacy's "Sentencizer" library[6] specifically adapted for Spanish. Subsequently, these sentence segments were classified according to their communicative intentions using the GPT-4o model, chosen due to its demonstrated high accuracy (0.84 F1 score) in classifying Spanish communicative intentions as reported by Maestre et al. (2025). However, the inherent error rate in this classification model was acknowledged, implying a slight mismatch between some sentences' real intentions and their classification, potentially influencing sentence similarity test outcomes in some cases.

After classifying the sentences by intention, segments corresponding to each of the 13 predefined communicative intentions were grouped. In this manner, for each set of reviews regarding a product category, these groups of sentences formed the evaluation sets for comparing the original product reviews with the summaries automatically generated by the LLMs according to matching communicative intentions.

5.2 Sentence similarity test

The final evaluation step involved calculating the semantic similarity between sentences from the original reviews grouped by communicative intentions and the automatically-generated summaries aligned to the same intentions. To this end, cosine similarity was calculated by measuring the similarity between the vectors of the sentences set from the original texts that convey the same intention as the generated summary. The obtained values can be between 0 and 1, where 0 would entail that there is no similarity between both texts and 1 that both texts are identical (Park et al., 2020).

The cosine similarity formula is expressed as follows:

$$\text{sim}(A, B) = \frac{A \cdot B}{\|A\|\|B\|}$$

where A and B are the vectors of the original set of sentences and the automatic summary. The conversion of both texts into embeddings to get the vectors required for the formula was done with the model "sentence_similarity_spanish_es",[7] specifically optimized for the task of semantic similarity between texts.

Additionally, a more comprehensive evaluation was conducted by calculating the same cosine similarity based on the model BERT with the BERT Similarity metric. The difference between this metric and the previous version is that in the BERT version, vectors are generated with BERT based on the "evaluate" library available in the Hugging Face platform.[8]

The cosine similarity formula based on BERT is expressed as follows (Bingi and Yin, 2023):

$$\text{BERT sim}(A, B) = \frac{BERT(A) \cdot BERT(B)}{\|BERT(A)\|\|BERT(B)\|}$$

where A and B are the vectors created by BERT from the original product review sentences aligned with a specific intention and the automatically-generated summary with that same intention.

With these two versions of the cosine similarity metric, we aimed at assessing how effectively intentional similarity between original review sentences and their intention-aligned automatic summaries can be measured. Additionally, these metrics facilitate evaluating which combinations of prompts and models most accurately generate summaries that reflect the communicative intentions present in the original reviews, as detailed in the subsequent subsection.

5.3 Compared results

When analyzing the semantic similarity between original sentences from reviews across different

[6] https://spacy.io/api/sentencizer

[7] https://huggingface.co/hiiamsid/sentence_similarity_spanish_es

[8] https://huggingface.co/docs/evaluate/v0.4.0/en/index

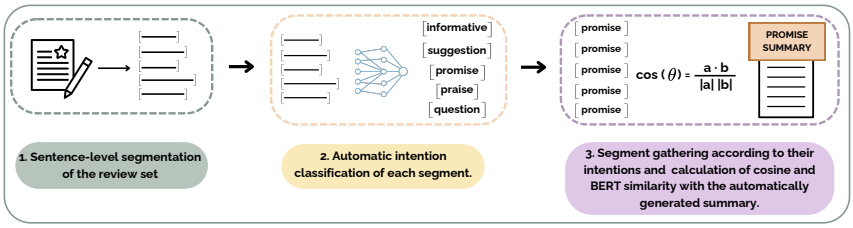
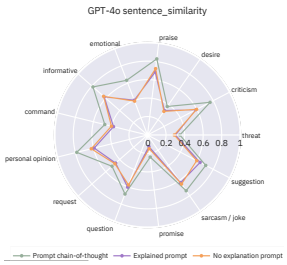
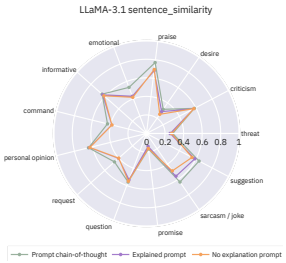


Figure 3: Proposed methodology to calculate sentence similarity between original review excerpts and summaries according to their intention.

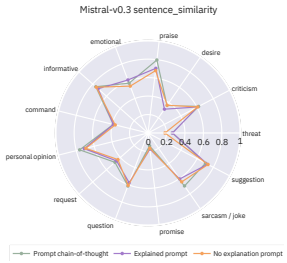
product categories and their corresponding generated summaries, we evaluated the impact of each LLM and the specific prompts designed for the intention-aligned generation task. Figures 4 and 5 illustrate statistical outcomes from the two versions of semantic similarity metrics—sentence similarity and BERT similarity—across the four selected LLMs and the 13 predefined communicative intentions determined for this study.



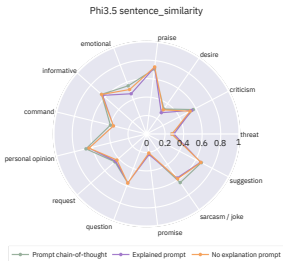
(a) GPT-4o



(b) LLaMA-3.1



(c) Mistral-v0.3

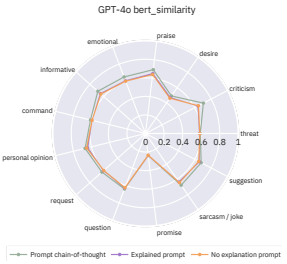


(d) Phi3.5

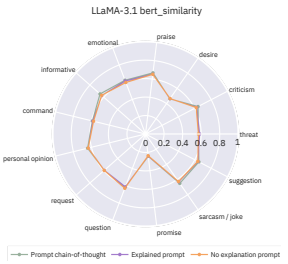
Figure 4: Results for the sentence similarity metric with the four LLMs.

A notable initial observation is the significant difference between the two metric types. The BERT similarity results appear surprisingly uniform, contrasting with the sentence similarity outcomes, especially given that prior qualitative analyses highlighted distinct variations in intention alignment depending on the targeted communicative intention.

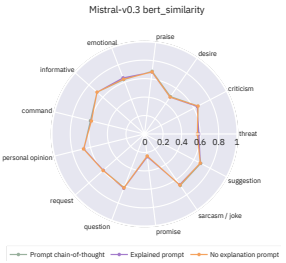
Focusing on the sentence similarity results presented in Figure 4, which align more closely with findings from the manual qualitative assessment,



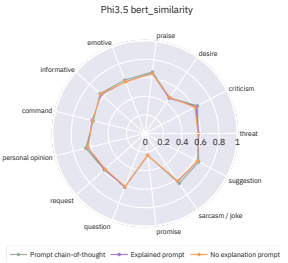
(a) GPT-4o



(b) LLaMA-3.1



(c) Mistral-v0.3



(d) Phi3.5

Figure 5: Results for the BERT similarity metric with the four LLMs.

we compare the alignment effectiveness of each model concerning the thirteen intentions and the three prompts used. GPT-4o achieves high similarity scores primarily for the intentions "personal opinion", "informative", and "praise" predominantly when using the CoT prompt. Similarly, LLaMA-3.1 attains comparable similarity scores for these intentions; however, performance across the three prompts appears more uniform, suggesting this model's relative insensitivity to prompt variations. Phi3.5 demonstrates even greater indifference to prompt types, with the "no explanation" prompt occasionally yielding higher similarity scores for intentions such as "praise" and "informative". Nonetheless, intentions like "promise" and "desire" consistently yield lower similarity scores across all models, as previously detected in the manual revision, indicating either their lower representation in the original reviews or increased complexity in identification by the LLMs. Lastly,

Mistral-v0.3 benefits significantly from both the CoT and "no explanation" prompts, occasionally matching GPT-4o's performance (using the CoT prompt) in intentions like "suggestion" and "praise", and surpassing LLaMA-3.1 for "personal opinion".

6 Conclusions and Future Work

This paper introduces an analytical approach to summarizing product reviews to check up to which point LLMs are capable of aligning the intentions of the original reviews to their respective summaries when specific intentions are requested. This pragmatic approach aims to capture crucial aspects of reviews, thus enhancing the relevance and comprehensiveness of the summaries. The experiment highlighted the challenges posed by extensive prompts, particularly those involving 50 reviews across various product categories, which strain generative models' processing capabilities. Despite these difficulties, addressing such extensive datasets remains valuable due to the practical necessity of summarizing vast amounts of reviews on different types of websites to provide users with concise but complete summaries on the information they want to consult.

A significant challenge identified is contamination from opinion and sentiment analysis tasks, commonly associated with product review corpora. This influence was repeatedly observed during the qualitative analysis of this research, as generative models often defaulted to summarizing positive and negative sentiments, neglecting explicitly defined communicative intentions. Indeed, the evaluation based on semantic similarity—using sentence similarity and BERT models—demonstrated that intentions related to opinion analysis (criticism, praise, personal opinion, suggestions) are more effectively identified.

Additionally, generative models showed limitations in reasoning capabilities, resulting in summaries with irrelevant content, exclusive focus on single product models, multilingual summaries, or even entirely fabricated reviews, as can be consulted in the repository published with all the generated summaries. These phenomena underscore the persistent gap in models' comprehension and their tendency toward hallucination.

A future line of research could be improving models' awareness of which communicative intentions may be absent from the input data, recommending explanations of why a summary cannot be generated in alignment with a specific intention if it has no representation in the original text to avoid misleading users. Furthermore, the research underlines the critical role of prompts and highlights the effectiveness of employing a chain-of-thought strategy, which significantly improves task performance for some intentions and models. Testing some other prompting strategies could also be extended in future research to compare further results on how to best communicate with LLMs when addressing this pragmatic-based generation task.

Acknowledgments

This research is part of the R&D projects "CORTEX: Conscious Text Generation" (PID2021-123956OB-I00), funded by MCIN/AEI/10.13039/501100011033/ and by "ERDF A way of making Europe"; "CLEAR.TEXT: Enhancing the modernization public sector organizations by deploying Natural Language Processing to make their digital content CLEARER to those with cognitive disabilities" (TED2021-130707B-I00), funded by MCIN/AEI/10.13039/501100011033 and "European Union NextGenerationEU/PRTR"; and project "NL4DISMIS: Natural Language Technologies for dealing with dis- and misinformation with grant reference (CIPROM/2021/021)" funded by the Generalitat Valenciana. Moreover, it has been also partially funded by the Ministry of Economic Affairs and Digital Transformation and "European Union NextGenerationEU/PRTR" through the "ILENIA" project (grant number 2022/TL22/00215337) and "VIVES" subproject (grant number 2022/TL22/00215334). Finally, this work has also been partially funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA (Resol. SEDIA 19.08.2024), as well as by the project "The limits and future of data-driven approaches: A comparative study of deep learning, knowledge-based and rule-based models and methods in Natural Language Processing" (CIDEXG/2023/13), funded by the Generalitat Valenciana.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Mostafa M Amin, Erik Cambria, and Björn W Schuller. 2023. Can chatgpt's responses boost traditional natural language processing? *IEEE Intelligent Systems*, 38(5):5–11.

Ghada Amoudi, Amal Almansour, and Hanan Saleh Alghamdi. 2022. Improved graph-based arabic hotel review summarization using polarity classification. *Applied Sciences*, 12(21).

Dawn Archer, Jonathan Culpeper, and Matthew Davies. 2008. Pragmatic annotation. In *Corpus Linguistics: An International Handbook*, volume 1, pages 613–642. Mouton de Gruyter Berlin.

John Langshaw Austin. 1962. *How to Do Things with Words. The William James Lectures delivered at Harvard University in 1955*. Oxford at the Clarendon Press, London.

Kent Bach. 2012. Saying, meaning, and implicating. In *The Cambridge Handbook of Pragmatics*, pages 47–68. Cambridge University Press, Cambridge.

Gunjan Balde, Soumyadeep Roy, Mainack Mondal, and Niloy Ganguly. 2024. Medvoc: vocabulary adaptation for fine-tuning pre-trained language models on medical text summarization. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24.

Yash Bingi and Yiqiao Yin. 2023. An analysis of embedding layers and similarity scores using siamese neural networks. *arXiv preprint arXiv:2401.00582*.

K. R. Chowdhary. 2020. *Natural Language Processing*, pages 603–649. Springer India, New Delhi.

Nicholas Crispino, Kyle Montgomery, Fankun Zeng, Dawn Song, and Chenguang Wang. 2023. Agent instructs large language models to be general zero-shot reasoners. *arXiv preprint arXiv:2310.03710*.

Andrea Di Sorbo, Sebastiano Panichella, Carol V. Alexandru, Junji Shimagaki, Corrado A. Visaggio, Gerardo Canfora, and Harald C. Gall. 2016. What would users change in my app? Summarizing app reviews for recommending software changes. In *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering*, FSE 2016, page 499–510, New York, NY, USA. Association for Computing Machinery.

Chenhe Dong, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. 2022. A survey of natural language generation. *ACM Computing Survey*, 55(8).

Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert systems with applications*, 165:113679.

M. Victoria. Escandell Vidal, Jose. Amenos Pons, and Aoife Kathleen. Ahern. 2020. *Pragmatica*, 1st ed. edition. Linguistica (Akal). Akal, Madrid.

Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. 2023. Pragmatics in language grounding: Phenomena, tasks, and modeling approaches. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12619–12640, Singapore. Association for Computational Linguistics.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online. Association for Computational Linguistics.

Mingyang Geng, Shangwen Wang, Dezun Dong, Haotian Wang, Ge Li, Zhi Jin, Xiaoguang Mao, and Xiangke Liao. 2024. Large language models are few-shot summarizers: Multi-intent comment generation via in-context learning. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, ICSE '24, New York, NY, USA. Association for Computing Machinery.

Deepanway Ghosal, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2021. Exploring the role of context in utterance-level emotion, act and intent classification in conversations: An empirical study. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1435–1449, Online. Association for Computational Linguistics.

Tanya Goyal, Nazneen Rajani, Wenhao Liu, and Wojciech Kryscinski. 2022. HydraSum: Disentangling style features in text summarization with multi-decoder models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 464–479, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

M. A. K. Halliday and Ruqaiya Hasan. 1986. *Language, context, and text: aspects of language in a social-semiotic perspective*. Specialised curriculum. language and learning. Deakin University, Burwood.

Walid Hariri. 2023. Unlocking the potential of chatgpt: A comprehensive exploration of its applications, advantages, limitations, and future directions in natural language processing. *arXiv preprint arXiv:2304.02017*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. 2023. Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, 82(3):3713–3744.

Stephen C. Levinson. 2017. Speech Acts. In *The Oxford Handbook of Pragmatics*. Oxford University Press.

Yangning Li, Shirong Ma, Xiaobin Wang, Shen Huang, Chengyue Jiang, Hai-Tao Zheng, Pengjun Xie, Fei Huang, and Yong Jiang. 2024. Ecomgpt: Instruction-tuning large language models with chain-of-task tasks for e-commerce. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18582–18590.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. *arXiv preprint arXiv:2502.12378*.

Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z. Sheng. 2022. Multi-document summarization via deep learning techniques: A survey. *ACM Comput. Surv.*, 55(5).

María Miró Maestre, Ernesto L Estevanell-Valladares, Robiert Sepúlveda-Torres, and Armando Suárez Cueto. 2025. Enhancing pragmatic processing: A two-dimension approach to detecting intentions in spanish. *Procesamiento del lenguaje natural*, 74:263–276.

Ggaliwango Marvin, Nakayiza Hellen, Daudi Jjingo, and Joyce Nakatumba-Nabende. 2023. Prompt engineering in large language models. In *International conference on data intelligence and cognitive informatics*, pages 387–402. Springer.

M. F. Mridha, Aklima Akter Lima, Kamruddin Nur, Sujoy Chandra Das, Mahmud Hasan, and Muhammad Mohsin Kabir. 2021. A survey of automatic text summarization: Progress, process and challenges. *IEEE Access*, 9:156043–156070.

Kwangil Park, June Seok Hong, and Wooju Kim. 2020. A methodology combining cosine similarity with classifier for text classification. *Applied Artificial Intelligence*, 34(5):396–411.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384, Singapore. Association for Computational Linguistics.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*, volume 626 of *Cam: Verschiedene Aufl.* Cambridge University Press.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and cognition*, volume 142. Citeseer.

Maite Taboada and Jack Grieve. 2004. Analyzing appraisal automatically. In *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text (AAAI Technical Report SS 04 07), Stanford University, CA*, pages 158–161. AAAI Press.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hsiu-Yi Wang, Jia-Wei Chang, and Jen-Wei Huang. 2019. User intention-based document summarization on heterogeneous sentence networks. In *Database Systems for Advanced Applications: 24th International Conference, DASFAA 2019, April 22–25, Proceedings, Part II 24*, pages 572–587, Chiang Mai, Thailand. Springer.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 24824–24837.

Yunshu Wu, Hayate Iso, Pouya Pezeshkpour, Nikita Bhutani, and Estevam Hruschka. 2024. Less is more for long document summary evaluation by LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 330–343, St. Julian's, Malta. Association for Computational Linguistics.

Zihan Yu, Liang He, Zhen Wu, Xinyu Dai, and Jiajun Chen. 2023. Towards better chain-of-thought prompting strategies: A survey. *arXiv preprint arXiv:2310.04959*.

Leilei Zhang and Junfei Liu. 2022. Intent-aware prompt learning for medical question summarization. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 672–679.