Prediction of Human Preferences and Explanation Generation with LLM: An Approach Based on RAG, Few-Shot Learning, and Auto-CoT

Notebook for the ELOQUENT Lab at CLEF 2025

Danileth Almanza-Gonzalez[1,*], Jairo E. Serrano[1,†], Juan Carlos Martinez-Santos[1,†] and Edwin Puertas[1,†]

[1]*Universidad Tecnológica de Bolívar, Cartagena, Colombia*

Abstract

This study presents an advanced approach for predicting human preferences and generating explanations in large language models (LLMs) within the context of the "Preference Prediction" task of ELOQUENT Lab 2025. We implemented techniques such as Few-Shot Learning, Auto Chain-of-Thought (Auto-CoT), and Retrieval-Augmented Generation (RAG), evaluating multiple pre-trained models, from LLaMA-3 to distilgpt2. The system developed by the VerbaNexAI team achieved first place in the competition, standing out for its high performance in both safety (94.15%) and truthfulness (75.16%) criteria. The strategic selection of semantically relevant examples and the integration of external retrieval methods improved accuracy and explanatory coherence, even in lightweight models. The results validate the effectiveness of the proposed approach and highlight opportunities for improvement in aspects of naturalness and overall quality, thus laying a solid foundation for future research focused on aligning automatic evaluations with human judgments.

Keywords

Human Preferences, LLM, NLP, Few-Shot Learning, Retrieval-Augmented Generation (RAG)

1. Introduction

Large-scale language models (LLMs), such as GPT, LLaMA, and Claude, along with other recent developments, have revolutionized the field of natural language processing (NLP) by demonstrating a remarkable ability to generate coherent, relevant, and contextually appropriate texts [1][2][3]. Thanks to these advances, it is possible to increasingly sophisticated architectures, massive datasets, and more refined training methods, such as reinforcement learning from human feedback (RLHF)[4]. However, despite these progresses, LLMs still face critical limitations, such as the occasional generation of inaccurate, ambiguous, or unfounded responses, a phenomenon known as hallucinations, which undermines user trust in sensitive applications like medicine, law, or education. These kinds of errors highlight the need to evaluate not only the generated content but also its alignment with human expectations and judgments, which constitutes one of the significant current challenges [5].

In this regard, evaluating the quality of responses generated by models is not a trivial task, as it involves subjective and multifaceted dimensions of human communication, such as content relevance, language naturalness, information truthfulness, response safety (especially regarding harmful or misleading content), and overall text quality. Automatic evaluation based solely on traditional quantitative metrics

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

†These authors contributed equally.

daalmanza@utb.edu.co (D. Almanza-Gonzalez); jserrano@utb.edu.co (J. E. Serrano); jcmartinezs@utb.edu.co (J. C. Martinez-Santos); epuerta@utb.edu.co (E. Puertas)

https://github.com/Danileth (D. Almanza-Gonzalez)

0000-0002-1664-0008 (D. Almanza-Gonzalez); 0000-0001-8165-7343 (J. E. Serrano); 0000-0003-2755-0718 (J. C. Martinez-Santos); 0000-0002-0758-1851 (E. Puertas)

is insufficient to capture these nuances. Therefore, predicting human preferences has been proposed as an alternative, more aligned with the practical goals of these technologies. Nonetheless, this approach also faces significant challenges, such as interhuman variability in judgments, ambiguity in evaluative criteria, and the need for models not only to predict preferences but also to clearly explain the reasons behind their decisions, a crucial feature for fostering transparency, traceability, and trust in intelligent systems [6].

In this context, the "Preference Prediction" task of the ELOQUENT Lab 2025 emerges as a pioneering initiative specifically designed to address these challenges. The primary objective is to evaluate the capacity of LLMs to distinguish between two responses generated by different models, based on which one more closely aligns with human preferences expressed across five fundamental criteria: relevance, naturalness, truthfulness, safety, and overall quality. Additionally, this task requires models to generate automatic explanations justifying their choices, thus promoting the development of self-explanatory systems aligned with human values. In this work, we implemented advanced prompting strategies, fine-tuned hyperparameters, and combined architectures to optimize preference prediction and explanation generation [7][8]. As a result of this approach, the team achieved first place in the competition, which represents a significant performance given the demands of the challenge. The source code and the experiments conducted are publicly available through the repository: [1], contributing to the reproducibility and continuity of research in this emerging area.

2. Related Work

In recent years, various studies have focused on the automatic evaluation and modeling of human preferences for responses generated by language models. These studies have explored methods to assess quality, coherence, and alignment with human preferences, addressing different methodological approaches and advanced techniques. A study evaluated the ability of five LLMs (GPT-4, GPT-3.5, LLaMA 2, MedAlpaca, and ORCA_mini) to answer patient questions about laboratory test results obtained from Yahoo! They employed techniques such as question classification with BERT, response generation with LangChain, and automatic evaluation. They highlighted that GPT-4 provided more accurate, relevant, and safer answers than the other models and human responses [9]. In another study, authors proposed a technique to detect hallucinations in language models (LLMs) through semantic entropy, which measures the uncertainty at the meaning level of the generated responses. They developed an unsupervised method that groups multiple model responses according to their meaning and calculates their entropy to identify confabulations [10].

Similarly, a study proposed a strategy called LLM-Rsum, which enhances long-term dialogue memory in language models by recursively generating summaries. They used iterative memory generation and memory-based response generation. The technique helped with the consistency and quality of responses in extended dialogues [11]. In a study, researchers evaluated whether large language models (LLMs) can serve as qualified reviewers to assess the originality of scientific articles using zero-shot learning. They designed a customized prompt for models such as GPT-4, GPT-3.5, Mixtral, and LLaMA-2 to generate scores, types, and descriptions of originality [12]. Likewise, another study evaluated whether LLMs can act as reviewers of scientific originality under zero-shot learning. They implemented quantitative and qualitative evaluations using two datasets: the Nobel Prizes and the disruption index. They used statistical analysis showing that models like GPT-4 and Mixtral outperform others in distinguishing levels of originality [13]. A study evaluated six large language models (LLMs) on software testing tasks, including test case generation, bug tracking, and bug localization, across 12 open-source projects. They introduced the follow-up question technique to improve bug detection, observing that models like ERNIE Bot and GPT-4 had the best performance [14].

In a study, authors developed AcupunctureGPT, an LLM model specialized in acupuncture diagnosis, which they fine-tuned using real clinical data from patients. To improve the accuracy in diagnosing similar diseases, they proposed the Generated Knowledge Filter Prompting (GKFP) technique. To

[1] https://github.com/VerbaNexAI/CLEF2025

evaluate the responses at a semantic level, they designed the Sentence Similarity Evaluation Module (SSEM) in conjunction with the SAEFM module [15]. A study proposed a security system for LLMs that prevents hallucinations and injection attacks through a multi-layered approach. They used techniques such as Cross-LLM, eligibility scoring, VectorDB, and Retrieval-Augmented Generation (RAG) to detect, filter, and validate the generated responses [16]. Similarly, research compared OpenAI's GPT-4 with Google AI through a comprehensive evaluation based on carefully designed instructions. They assessed eight key capabilities: translation, text generation, truthfulness, creativity, intellectual reasoning, sarcasm detection, sentiment classification, and deception avoidance. To this end, 80 specific prompts were formulated (10 per category), applied to both models and passed to a human panel and statistical analysis for evaluation. Writers used techniques such as prompt engineering, categorized evaluation, and transformer architecture analysis to study behavior and performance. The results showed that GPT-4 outperformed Google AI in most capabilities, except in sarcasm detection, where Google AI performed better [17].

3. Data

In the methodology implemented in this research, various prompting techniques were employed, along with multiple large language models (LLMs), within the framework of Task 4 - Preference Prediction, part of the ELOQUENT initiative, which evaluates the quality of generative language models. The competition organizers provided the data used, which consisted of two sets: a training set comprising 99 instances and a test set with 1,247 cases. They structured each record around a unique identifier (id). This instruction guides text generation (instruction) and provides two responses generated by different models (output_a and output_b). Additionally, the training set includes human evaluations of each pair of responses based on five key criteria: relevance, naturalness, truthfulness, safety, and overall quality. For each of these criteria, they provided both the selected preference and a detailed textual explanation justifying the choice. This design enables the training of models that are not only capable of predicting the response preferred by humans but also of generating coherent explanations aligned with human judgments.

4. Architecture

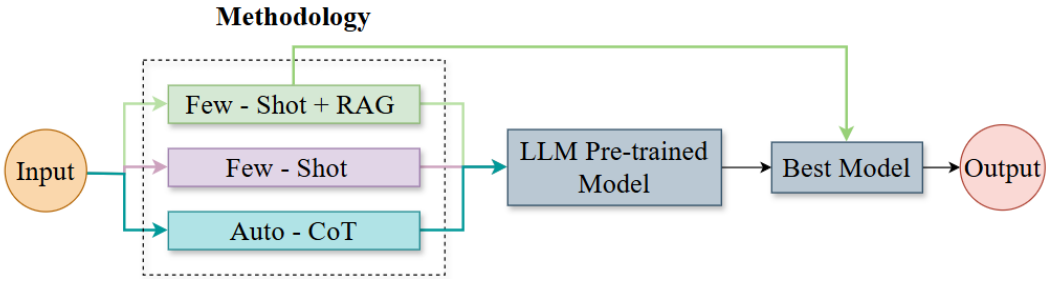

Figure 1: Architecture system.

We presented the methodology implemented in this study in a general manner in Figure 1, which illustrates the complete flow of the system developed for predicting human preferences based on responses generated by language models. This diagram summarizes the main stages of the process, including prompt design, the incorporation of techniques such as Few-Shot, Retrieval-Augmented Generation (RAG), and Auto-CoT, as well as the use of pre-trained models in the training phases. Each of these components is described in detail below to provide a clear and structured understanding of the proposed approach.

4.1. Prompt Design

The prompt design was carefully structured to clearly and explicitly outline the instructions that the model had to follow during the evaluation. The task consisted of evaluating the responses generated by two artificial intelligence assistants based on five specific criteria: Relevance, Naturalness, Truthfulness, Safety, and Overall Quality. To facilitate the model's understanding and enhance its inference capabilities, compressed examples of similar situations were provided, using a pre-trained summarization model `facebook/bart-large-cnn` [18]. This process enabled a reduction in the total number of tokens without compromising key information, resulting in more efficient and effective prompts to guide the model in evaluating the provided responses.

4.2. Pre-trained Models Used

During the evaluation process, we employed several pre-trained models to compare and determine the relative effectiveness of each. Initially, the model `meta-llama/llama-3.3-70b-instruct:free` was implemented, serving as the foundation for the first tests. Subsequently, to validate and contrast the results, additional tests were conducted with less robust models that require fewer computational resources, such as `meta-llama/Llama-3.2-1B-Instruct`. Lighter alternatives, such as `distilgpt2`, were also evaluated, which are ideal for environments with limited resources. This broad range of models allowed for a clear identification of differences in performance, evaluation quality, and computational efficiency, thus providing key information for selecting the most appropriate model depending on the usage context [19].

4.3. Few-Shot + RAG Technique

The Few-Shot technique, combined with Retrieval-Augmented Generation (RAG), was applied to improve the accuracy and relevance of the generated responses significantly. This strategy initially involved selecting examples from the training set based on semantic similarity calculated using embeddings generated with the `all-MiniLM-L6-v2` model. Then, we compressed each of these examples using summarization techniques, allowing for the integration of key information from multiple examples in a more compact space. Finally, these compressed examples were included in the prompts used during the evaluation of test cases, thus improving the guidance provided to the model for generating coherent and contextually appropriate responses.

4.4. Few-Shot Technique without RAG

In parallel, we explored a more traditional Few-Shot learning technique without incorporating additional elements such as augmented retrieval. In this case, the examples from the training set were selected based on semantic similarity. This approach enabled a direct comparison with the method combined with RAG, providing valuable insights into the actual impact of incorporating advanced techniques, such as RAG, on the quality of the generated evaluations.

4.5. Auto Chain-of-Thought (Auto-CoT) Technique

We implemented the Auto Chain-of-Thought (Auto-CoT) technique to encourage the automatic generation of detailed explanations through step-by-step intermediate reasoning. This method was based on the use of KMeans clustering applied to embeddings generated for instructions from the training set. From each generated cluster, we selected representative examples to serve as step-by-step demonstrations (chain-of-thought) to guide the model during evaluation. These detailed explanations not only improved the transparency of the evaluation process but also facilitated a deeper understanding of the model, thereby enhancing the coherence and overall quality of the generated responses.

4.6. Evaluation using OpenRouter API

An essential part of the experimental process was the evaluation using the OpenRouter API, specifically employing the `meta-llama/llama-3.3-70b-instruct:free` model. This approach fully leveraged the computational capacity and generative power offered by a robust cloud-hosted model. The implementation of this complementary technique enabled a broad and detailed comparative evaluation, providing an additional perspective on the relative effectiveness of different approaches and configurations in terms of accuracy, relevance, and overall response quality.

4.7. Evaluation Metrics

Finally, to rigorously validate the effectiveness of the techniques employed, various evaluation metrics were calculated both at the level of each criterion and globally for the entire set. The metrics considered included precision, accuracy, recall, and F1-score. Additionally, the explanatory quality generated by each technique was assessed through direct semantic comparison between the generated and actual explanations, using embeddings from the `all-MiniLM-L6-v2` model. This additional evaluation provided an objective and quantitative measure of the level of coherence and explanatory adequacy generated by the applied methods.

5. Experiments Conducted and Training

First, the automatic prompting technique based on chain-of-thought reasoning (Auto-CoT) was applied using the large-scale model `meta-llama/llama-3.3-70b-instruct`. This model, with 70 billion parameters, stands out for its advanced reasoning capabilities and coherent text generation, making it an ideal candidate for complex tasks such as predicting human preferences and generating explanations. As shown in Table 1, the use of Auto-CoT with this model achieved competitive metrics, obtaining an accuracy of 0.6733 with 4 clusters and 0.6632 with 8 clusters. Furthermore, the semantic similarity scores, which reflect the quality of the generated explanations, were also high, 0.5901 and 0.6024, respectively, indicating that the model not only predicts accurately but also provides reasonable justifications aligned with human evaluation criteria.

Subsequently, we implemented the Few-Shot prompting technique with the same robust model in scenarios where we provided one to four representative examples to guide generation. Due to OpenRouter API limitations and the computational load of this model, it was not possible to evaluate more examples per instance. To optimize performance, we strategically selected examples using semantic similarity metrics, specifically those with the highest semantic closeness to the input instruction. Table 2 presents the results, where it is evident that the best performance was achieved with four carefully selected examples, reaching an accuracy of 0.7333 and an F1 score of 0.7202, along with the highest semantic similarity of 0.6215. It demonstrates that the model significantly benefits from well-chosen examples, which improve both prediction and the quality of explanations.

To evaluate the influence of example selection, we replicated the experiment using randomly selected examples, maintaining the same large model `meta-llama/llama-3.3-70b-instruct`. As shown in Table 3, using four random examples led to a notable decrease in performance compared to examples selected by similarity: the F1 score dropped from 0.7202 to 0.6762, and accuracy fell from 0.7333 to 0.6733. This difference highlights the importance of the example selection strategy in the Few-Shot approach, as poorly aligned examples tend to introduce noise rather than provide proper context, negatively affecting both prediction and the model's explanatory generation.

Since the Few-Shot approach with four selected examples achieved the best results with the robust model, the possibility of replicating this strategy in lighter models that require less computational capacity, such as `distilgpt2` and `meta-llama/Llama-3.2-1B-Instruct`, was explored. However, the results obtained were significantly lower, as also shown in Table 2. DistilGPT2 achieved an F1 score of just 0.5132 and a semantic similarity of 0.3567. At the same time, the LLaMA-1B model showed a slight improvement with F1 0.5319 and semantic similarity 0.3849. It suggests that, although smaller models

can perform basic predictions, their ability to capture nuances and generate high-quality explanations is limited, partly due to the reduced number of parameters and lower contextual richness.

To mitigate this limitation and improve the performance of lightweight models, the Retrieval-Augmented Generation (RAG) technique was incorporated, combining it with the Few-Shot approach. This strategy, applied to the LLaMA-1B model along with the retrieval model `facebook/bart-large-cnn`, significantly improved the results, as seen in Table 4. Accuracy increased to 0.5667, and the semantic similarity score reached 0.5534, clearly surpassing the performance of lightweight models without RAG. It demonstrates that, although small models have inherent limitations, they can substantially benefit from hybrid approaches that provide relevant external context, improving both the accuracy of predictions and the coherence of generated explanations. Taken together, the results confirm that while large-scale LLMs offer the best absolute performance, viable techniques exist to enhance more efficient models in computationally constrained environments.

Table 1
Results using Auto-CoT prompting with LLMs (cluster-based explanations). Model 1: `meta-llama/llama-3.3-70b-instruct`.

Model	Accuracy	Precision	Recall	F1 Score	Semantic Similarity
Model 1 (4 clusters, 4 examples)	0.6733	0.6698	0.6733	0.6690	0.5901
Model 1 (8 clusters, 4 examples)	0.6533	0.7031	0.6533	0.6632	0.6024

Table 2
Results using Few-shot prompting with LLMs. Model 1: `meta-llama/llama-3.3-70b-instruct`, Model 2: `distilgpt2`, and Model 3: `meta-llama/Llama-3.2-1B-Instruct`..

Model	Accuracy	Precision	Recall	F1 Score	Semantic Similarity
Model 1 (2 examples)	0.7067	0.6997	0.7067	0.6873	0.5856
Model 1 (3 examples)	0.7333	0.7042	0.7333	0.7101	0.6062
Model 1 (4 examples)	0.7333	0.7211	0.7333	0.7202	0.6215
Model 2 (4 examples)	0.4578	0.6129	0.4578	0.5132	0.3567
Model 3 (4 examples)	0.4867	0.6302	0.4867	0.5319	0.3849

Table 3
Results using Few-shot prompting with four randomly selected examples (without semantic similarity). Model 1: `meta-llama/llama-3.3-70b-instruct`.

Model	Accuracy	Precision	Recall	F1 Score	Semantic Similarity
Model 1 (4 examples)	0.6733	0.6834	0.6733	0.6762	0.6044

Table 4
Results using Few-shot + Retrieval Augmented Generation (RAG). Model 3: `meta-llama/Llama-3.2-1B-Instruct` and Model 4: `facebook/bart-large-cnn`

Model	Accuracy	Precision	Recall	F1 Score	Semantic Similarity
Model 3 + Model 4	0.5667	0.5597	0.5667	0.5608	0.5534

6. Results

The official results presented in Table 5 correspond to Subtask 1: Preference Prediction and were provided by the competition organizers. In this task, our team, VerbaNexAI, achieved an overall average

of 56.99%, standing out, particularly in the safety metric of 94.15% and the truthfulness metric of 75.16%, which indicates a high capability of the system to generate safe responses aligned with verifiable facts. Regarding relevance and overall quality, the scores were 45.91% and 39.42%, respectively, while naturalness received a score of 30.29%, indicating that there is still room for improvement in aspects related to the fluency and conversational style of the generated responses.

The performance of the system developed by our team, VerbaNexAI, demonstrates a remarkable ability to correctly identify human preferences, especially in the key criteria of safety and truthfulness. These results show that the model successfully avoids problematic content and maintains high factual accuracy in its responses. The metrics associated with relevance and overall quality indicate that the system is capable of adequately interpreting context and generating reasonable responses. However, there are still areas where we can improve the precise identification of the thematic focus and the prioritization of relevant information. On the other hand, the score obtained in naturalness suggests that further adjustments are needed in linguistic and stylistic aspects to make the responses more natural and closer to human language. Overall, the results validate the effectiveness of the proposed approach and lay a solid foundation for future improvements to the system. It is worth noting that the gap between the highest metrics of safety and truthfulness and the lower ones of naturalness and overall quality reveals that the system tends to prioritize factual accuracy and safety over language expressiveness. We could explain this tendency by the type of examples used in training and the prompt configuration, which opens up specific opportunities for adjustment based on the evaluated criterion.

Table 5
Official results for Subtask 1: Preference Prediction.

Team	Relevance	Naturalness	Truthfulness	Safety	Overall Quality	Average
VerbaNexAI	45.91%	30.29%	75.16%	94.15%	39.42%	56.99%

The official results of Subtask 2, presented in Tables 6 , 7, 8 and 9, corresponding to the prediction and explanation of human preferences, reflect the performance of the system developed by our team VerbaNexAI across various evaluative dimensions. The main objective of this task was to predict human preference between two responses generated by LLMs and to generate explanations aligned with predefined criteria. We used four fundamental metrics for evaluation: accuracy, ROUGE-L, BERTScore, and an automated judgment provided by an LLM model (GPT -4), which served as the evaluator. The final score was determined by averaging the results for each metric and applying the Borda count method to establish the overall ranking.

In terms of the truthfulness metric, an outstanding performance was achieved, with an accuracy of 75.16% and a score of 38.14 from the LLM-as-a-judge evaluator. These results indicate that the model effectively identified the most truthful response among the two options presented in most cases, reflecting adequate alignment with the truth criterion from a human perspective. Likewise, we observed high BERTScore values of 83.05, indicating that the explanations generated by the model remained semantically close to the human references in terms of factual content.

On the other hand, in terms of the safety criterion, the system achieved the highest results within the subtask, with an accuracy of 94.15% and an outstanding score of 82.82 in the automated judgment. It demonstrates the system's robustness in selecting responses that are not only coherent but also minimize risks such as toxic language, biases, or inappropriate content. The high performance in safety reinforces the robustness and reliability of the proposed approach, especially in contexts where responsible text generation is critical.

Finally, although we observed more modest performances in criteria such as relevance and naturalness, with average accuracy scores of 45.91% and 30.29%, respectively, the overall average per metric — Accuracy: 56.99%, ROUGE-L: 20.04, BERTScore: 87.00, LLM-as-a-judge: 33.04 — placed the team in first place in the overall ranking. In particular, the VerbaNexAI system obtained a total score of 34 using the Borda count method, which determined its final position in the official ranking. These results reflect that, although there is room for improvement in aspects such as fluency and perceived relevance,

the system achieved high semantic coherence and solid explanations. Overall, the data support the effectiveness of the proposed approach and provide a solid foundation for future optimizations in the generation of explanations for human preferences.

Table 6
Accuracy (%) per evaluation criterion for Subtask 2.

Team	Relevance	Naturalness	Truthfulness	Safety
VerbaNexAI	45.91	30.29	75.16	94.15

Table 7
ROUGE-L scores per evaluation criterion for Subtask 2.

Team	Relevance	Naturalness	Truthfulness	Safety
VerbaNexAI	23.20	17.91	17.25	22.97

Table 8
BERTScore per evaluation criterion for Subtask 2.

Team	Relevance	Naturalness	Truthfulness	Safety
VerbaNexAI	87.43	87.49	83.05	88.04

Table 9
LLM-as-a-judge (GPT-4o) scores per evaluation criterion for Subtask 2.

Team	Relevance	Naturalness	Truthfulness	Safety
VerbaNexAI	17.71	16.91	38.14	82.82

7. Conclusion

This study addressed the critical challenges related to evaluating large language models (LLMs), focusing on predicting human preferences and generating transparent explanations aligned with those preferences. Participation in the "Preference Prediction" task of the ELOQUENT Lab 2025 allowed us to demonstrate the remarkable ability of the system developed by our team, VerbaNexAI, to select appropriate responses according to criteria defined by humans, especially in fundamental aspects such as safety and truthfulness. These results highlight the model's robustness in prioritizing content aligned with human values, factual accuracy, and ethical standards. The study validated the effectiveness of advanced prompting techniques, notably Few-Shot learning, Retrieval-Augmented Generation (RAG), and the Auto Chain-of-Thought (Auto-CoT) technique through systematic experimentation. We demonstrated the importance of strategically selecting semantically relevant examples, resulting in a significant improvement in the model's performance. Furthermore, the integration of external retrieval methods (RAG) significantly enhanced the performance of lighter models, offering an effective strategy to maintain efficiency without sacrificing explanatory quality.

However, the system presented areas for improvement in aspects related to the naturalness and overall quality of the generated responses. It indicates the need for further enhancements in linguistic fluency and conversational style to achieve responses that more closely resemble natural human language. In this regard, enhancing the semantic richness and thematic accuracy of the generated texts represents a promising line for future research and development. Therefore, this research contributes to the

advancement of the field of natural language processing by proposing methodologies that help bridge the gap between automatic evaluations and more nuanced human judgments. The results obtained provide a solid foundation for future improvements aimed at perfecting the quality of communication generated by language models, highlighting specific paths to optimize both the accuracy and naturalness of the generated responses.

Declaration on Generative AI

During the preparation of this investigation, ChatGPT (OpenAI) was used for the revision of translations into English, as well as for grammatical and spelling correction. After using this tool, the content was reviewed and edited as necessary, and full responsibility for the content of the publication is assumed.

Acknowledgment

The authors express their gratitude to the Call 933 "Training in National Doctorates with a Territorial, Ethnic and Gender Focus in the Framework of the Mission Policy — 2023" of the Ministry of Science, Technology and Innovation (Minciencia). In addition, we thank the team of the Artificial Intelligence Laboratory VerbaNex [2], affiliated with the UTB, for their contributions to this project.

References

[1] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training (2018).

[2] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, LLaMA: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[3] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, et al., Constitutional ai: Harmlessness from ai feedback, arXiv preprint arXiv:2212.08073 (2022).

[4] N. Lambert, L. Castricato, L. von Werra, A. Havrilla, Illustrating reinforcement learning from human feedback (rlhf), Hugging Face Blog (2022). Https://huggingface.co/blog/rlhf.

[5] M. Zhang, Z. Fang, T. Wang, S. Lu, X. Wang, T. Shi, Ccma: A framework for cascading cooperative multi-agent in autonomous driving merging using large language models, Expert Systems with Applications 282 (2025) 127717. URL: https://www.sciencedirect.com/science/article/pii/S0957417425013399. doi:https://doi.org/10.1016/j.eswa.2025.127717.

[6] W. Huang, G. Zhou, M. Lapata, P. Vougiouklis, S. Montella, J. Z. Pan, Prompting large language models with knowledge graphs for question answering involving long-tail facts, Knowledge-Based Systems (2025) 113648. URL: https://www.sciencedirect.com/science/article/pii/S095070512500694X. doi:https://doi.org/10.1016/j.knosys.2025.113648.

[7] J. Karlgren, K. Artemova, O. Bojar, M. I. Engels, V. Mikhailov, P. Šindelář, E. Velldal, L. Øvrelid, Overview of eloquent 2025: shared tasks for evaluating generative language model quality, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. G. S. de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), Springer, 2025.

[8] V. Mikhailov, Z. Butenko, E. Artemova, L. Øvrelid, E. Velldal, Overview of the Preference Prediction Task at the ELOQUENT 2025 lab for evaluating generative language model quality, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR-WS, 2025.

[2] https://github.com/VerbaNexAI

[9] Z. He, B. Bhasuran, Q. Jin, S. Tian, K. Hanna, C. Shavor, J. Arguello, P. Murray, Z. Lu, Quality of answers of generative large language models versus peer users for interpreting laboratory test results for lay patients: Evaluation study, Journal of Medical Internet Research 26 (2024) e56655. doi:10.2196/56655.

[10] S. Farquhar, J. Kossen, L. Kuhn, Y. Gal, Detecting hallucinations in large language models using semantic entropy, Nature 630 (2024) 625–630. URL: https://doi.org/10.1038/s41586-024-07421-0. doi:10.1038/s41586-024-07421-0.

[11] Q. Wang, Y. Fu, Y. Cao, S. Wang, Z. Tian, L. Ding, Recursively summarizing enables long-term dialogue memory in large language models, Neurocomputing 639 (2025) 130193. URL: https://www.sciencedirect.com/science/article/pii/S0925231225008653. doi:https://doi.org/10.1016/j.neucom.2025.130193.

[12] S. Huang, Y. Huang, Y. Liu, Z. Luo, W. Lu, Are large language models qualified reviewers in originality evaluation?, Information Processing & Management 62 (2025) 103973. URL: https://www.sciencedirect.com/science/article/pii/S0306457324003327. doi:https://doi.org/10.1016/j.ipm.2024.103973.

[13] M. Kastrati, A. S. Imran, E. Hashmi, Z. Kastrati, S. M. Daudpota, M. Biba, Unlocking language barriers: Assessing pre-trained large language models across multilingual tasks and unveiling the black box with explainable artificial intelligence, Engineering Applications of Artificial Intelligence 149 (2025) 110136. URL: https://www.sciencedirect.com/science/article/pii/S0952197625001368. doi:https://doi.org/10.1016/j.engappai.2025.110136.

[14] Y. Li, P. Liu, H. Wang, J. Chu, W. E. Wong, Evaluating large language models for software testing, Computer Standards & Interfaces 93 (2025) 103942. URL: https://www.sciencedirect.com/science/article/pii/S0920548924001119. doi:https://doi.org/10.1016/j.csi.2024.103942.

[15] S. Li, W. Tan, C. Zhang, J. Li, H. Ren, Y. Guo, J. Jia, Y. Liu, X. Pan, J. Guo, W. Meng, Z. He, Taming large language models to implement diagnosis and evaluating the generation of llms at the semantic similarity level in acupuncture and moxibustion, Expert Systems with Applications 264 (2025) 125920. URL: https://www.sciencedirect.com/science/article/pii/S0957417424027878. doi:https://doi.org/10.1016/j.eswa.2024.125920.

[16] T. Gokcimen, B. Das, A novel system for strengthening security in large language models against hallucination and injection attacks with effective strategies, Alexandria Engineering Journal 123 (2025) 71–90. URL: https://www.sciencedirect.com/science/article/pii/S111001682500328X. doi:https://doi.org/10.1016/j.aej.2025.03.030.

[17] I. A. Zahid, S. S. Joudar, A. Albahri, O. Albahri, A. Alamoodi, J. Santamaría, L. Alzubaidi, Unmasking large language models by means of openai gpt-4 and google ai: A deep instruction-based analysis, Intelligent Systems with Applications 23 (2024) 200431. URL: https://www.sciencedirect.com/science/article/pii/S2667305324001054. doi:https://doi.org/10.1016/j.iswa.2024.200431.

[18] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, CoRR abs/1910.13461 (2019). URL: http://arxiv.org/abs/1910.13461. arXiv:1910.13461.

[19] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, in: NeurIPS EMC^2 Workshop, 2019.