Evaluating performance and trustworthiness of RAG systems for generating administrative text

Hugo Sánchez Navalón[1][0009−0007−9746−9057], Carlos Monserrat Aranda[1][0000−0003−1790−8085], Dario Garigliotti[2][0000−0002−0331−000X], and Cèsar Ferri[1][0000−0002−8975−1120]

[1] Universitat Politècnica de València
[2] University of Bergen

Abstract. As administrative language tends to be formal and exempt from double meanings or figurative expressions, it is a particular domain in which to explore the performance of Language Models. This paper presents a study on the feasibility of creating administrative texts-based RAG systems to serve as chatbots, analyzing the performance for this task of several Small and Large Language Models and defining ways of evaluating whether they hallucinate or not and whether they provide the user useful information or not. Conventional metrics depending on ground truth labels, such as cosine similarity or those from the ROUGE family, are explored, as well as new approaches to using other metrics not so popular in text evaluation, such as Euclidean and Manhattan distances. Moreover, all those objective metrics are compared with a subjective Likert scale to assess their performance at solving real users' problems and to find relations between subjective perceptions and objectively measured metrics for each of the RAG systems proposed. The results show that SLM models (such as NeuralChat) can perform as well as an LLM if RAG programming provides them with an appropriate context.

Keywords: RAG · LLM · SLM · Administrative Text · ROUGE · Cosine similarity · Question Answering · Chatbot.

1 Introduction

This paper presents a study on the evaluation and efficacy of Retrieval Augmented Generative (RAG) models based on administrative texts. The objective is to assess the feasibility of building a chatbot capable of answering questions from different user profiles regarding various procedures related to the Public Administration of the Comunitat Valenciana, an autonomous community of more than 5 million inhabitants in the east of Spain. Generalitat Valenciana is the public entity that rules the region of Comunitat Valenciana.

A formal register, exempt from double meanings or figurative expressions, characterizes the language of administrative texts. The questions typically asked by users are direct and relatively short. This suggests that building effective

systems based on fewer parameters than state-of-the-art Large Language Models (LLMs) may be possible.

RAG systems are based on the original proposal by [Lewis et al., 2021] to help enhance Language Models responses with domain-specific information while skipping processes like fine-tuning, which may be costly. RAG model comprises two core components: a retriever and a generator. The retriever identifies relevant information from the external knowledge source based on the user's query or input. This retrieved information is then passed to the generator, a Language Model, which synthesizes a coherent and informative response incorporating the retrieved knowledge. This architecture avoids the high computational cost of building an entire Language Model over a specific context, as well as the considerable cost of fine-tuning already existing ones over a new context. Moreover, as they can easily identify the sources of their answers by listing the documents retrieved for each user's query, they provide a fairly high explainability for each of their generated texts.

Apart from that, the technical language that characterizes this context makes it difficult for some conventional users to understand some of the key concepts of certain documentation pieces. Language Models have shown good performance in rewriting text for different types of users to understand according to their needs, as educational-oriented research in Artificial Intelligence has shown [Gan et al., 2023]. Therefore, having well-performing RAG systems built over administrative texts could be a feasible solution to make them more understandable to different types of users and adaptable to their needs.

This paper is structured as follows. In Section 2, we summarise the relevant related works. The methodology employed in the paper is presented in Section 3. We include the results from an experimental evaluation of our method in Section 4. Finally, Section 5 concludes the paper with a recapitulation of our work and future extensions.

2 Related work

The landscape of Retrieval-Augmented Generation (RAG) systems has been rapidly evolving since they were introduced back in 2020. In [Gao et al., 2024], the authors propose a comprehensive survey of RAG systems, including the importance of optimizing their retriever and generative subsystems and a bunch of evaluation approaches that involve experimenting with quality scores and robustness-related ones. They also present a hybrid approach of fine-tuning along with RAG construction to enhance its performance further.

A significant challenge in using Language Models (LMs) for generating administrative texts, which often contain objective information, is the risk of hallucinations. In [Zhang et al., 2023], the authors provide an overview of the hallucination issue and its types, focusing on fact-conflicting hallucination (generation of text that contradicts established knowledge). They show benchmark construction along with good-performing metrics to measure hallucination, in-

volving human, model-based, and rule-based evaluation. They also offer ways to mitigate it during pre-training, fine-tuning, and inference.

While fine-tuning has been widely regarded as a primary method for improving model performance, [Ovadia et al., 2024] challenges this notion by showing how RAG, if correctly implemented, could potentially outperform it. As they uncover *forgetting* and *immemorization* as two of the main problems that lead to factual errors, they conclude that giving specific context for each generation seems a more reliable choice, even though fine-tuning remains a good option for many use-cases. Considering specifically objective information related to RAG content, as we expect from administrative texts, in [Lála et al., 2023], the authors discuss the feasibility of building a RAG based on technic-scientific knowledge. It turns out to be "more cost-effective than humans while retaining its accuracy on par with human researchers". This suggests that RAG construction may be a feasible approach for the context given in administrative texts (as proposed in the present paper) because they handle characteristics similar to those of scientific papers.

Moreover, considering the subfield of Legal Question Answering, in which our work is included, in [Abdallah et al., 2023], the authors provide an extensive explanation of the state of the art of these systems. They review the challenges we face (related to reliability, domain expertise, and data availability), the Machine Learning approaches usually given, and some surveys about their general performance. Transformer-based systems seem to be more effective than classic Machine Learning algorithms, which indicates that our approach based on Language Models is more suitable to catch up with state-of-the-art results than any other one. In addition, in [Garigliotti et al., 2024], the authors explore a way of evaluating RAG-driven self-supported Question Answering by creating a custom dataset including evaluation questions, ground truth answers and context relation, finding that LLMs such as GPT-4 are capable of reliably generating answers to most kinds of questions if proper context is provided.

3 Materials and methods

This section details the methodological framework employed in our research, beginning with the specific approach to RAG implementation. As can be seen in Figure 1, we first handle context obtention and then text generation through Language Models. Afterward, a benchmark consisting of labeled questions is fed into the RAG systems and evaluated through objective and subjective measures.

3.1 Approach

A Selenium bot was used to retrieve all downloadable PDF content in which all technical and administrative information is stored within the Generalitat Valenciana website. A Breadth-First-Search on its categories tree was done for that purpose. A total of 1,619 PDF documents were downloaded, which sum a size of 90.4 MB. Each of them is referred to a possible interaction between a citizen and

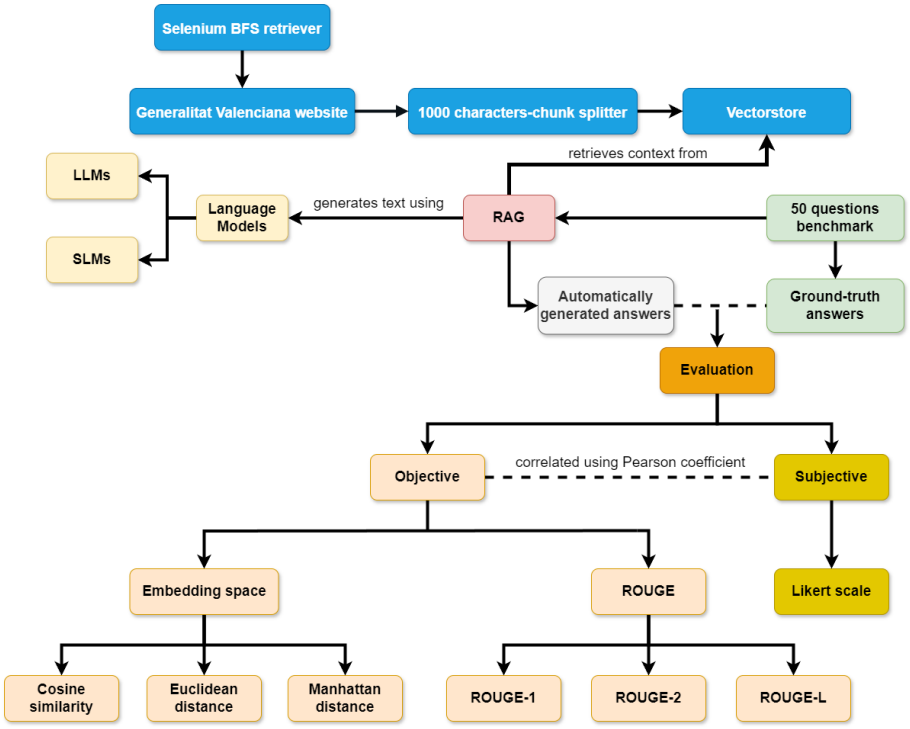

Fig. 1. Workflow of RAG building and evaluation process.

the Valencian Regional Government for a particular purpose, and each is divided into up to four main parts, being "Basic Information", "Appliance", "Procedure", and "Resolution". Some examples of these interactions are tax returns, gambling self-prohibition, child adoption, and cultural events authorization.

We used dense retrieving for context obtention, a method for retrieving documents semantically similar to a query by encoding them into high-dimensional vectors and comparing their distances in the resulting vector space, as proposed in [Karpukhin et al., 2020]. We first segmented the entire corpus of text into manageable chunks of 1,000 characters, ensuring a 100-character overlap between adjacent chunks. This overlap helps capture context that might be split across chunk boundaries. The embeddings were generated using *text-embedding-ada-002-v2* by OpenAI. The resulting vector representations were stored in a Chroma vector database. Chroma is a powerful and efficient vector database designed specifically for similarity search tasks, as detailed in [Huber et al., 2024].

Generator parts were built by using a set of Language Models (both Small Language Models and Large Language Models), including:

- **GPT Models.** 3.5, 4 and 4o. Widely known family of models developed by OpenAI that exhibit near-human level capabilities in some of the scenarios they have been tested on [OpenAI et al., 2024].
- **Phi 3 Mini.** A 3.8-billion parameters Language Model developed by Microsoft. It is defined as "instruction-tuned" [Abdin et al., 2024].
- **NeuralChat.** Based on the original Mistral, it is a fine-tuned version developed by Intel based on around 7 billion parameters. Its primary purpose is to serve as a base for building chatbots [Lv et al., 2023].
- **Gemma.** Family of open-source models developed by Google designed to be computationally efficient. Two versions have been used, involving 2 and 7 billion parameters [Team et al., 2024].
- **LLaMA 3.** It is a family of open-source Language Models developed by Meta AI aimed at multi-purpose tasks like creative writing or coding. The version used is based on 7 billion parameters [Touvron et al., 2023].

Each system used one of the Language Models and a cosine similarity-based retriever operating on the vector database. The k most similar passages to the user's query were selected to provide context to the language model in each case. For Small Language Models (SLMs: Phi 3, NeuralChat, both Gemma versions and LLaMA 3), $k = 5$ was used, while for Large Language Models (LLMs: all GPT variants), $k = 20$. Augmentation was addressed by adding to the context of each prompt a previous paragraph related to the system to detail that the answer should be given considering it had been formulated by a citizen to a particular Public Administration seeking an objective and truthful answer.

GPT LLMs were accessed through the OpenAI API. SLMs were run on a private Ollama server allocated in a computer with an AMD Ryzen 7 5800H CPU, a Nvidia Geforce GTX 1650 with 4GB DDR4 GPU, and 16GB of DDR4 RAM. Communication with all of them was handled through LangChain's Python library.

3.2 Evaluation

Evaluation questions were generated with the help of the GPT-4o model. Under the premise that each potential technical question would relate to one of the available administrative procedures, a random sample of one hundred procedures was used to generate one potential user question per procedure, using the prompt *"Generate a question which can be answered with the following text. It has to be a possible question that a citizen could do in order to get information related with the text."*, and adding then the whole text of the selected document. Subsequently, fifty questions were manually selected, prioritizing those most likely to be posed by potential users while ensuring a relative diversity in the topics addressed (procedures, requirements, subsidy programs, selection criteria for specific allocations, etc.). These fifty questions were then manually answered, drawing upon information about the corresponding procedure to establish a ground truth for each. The sum of questions and ground truth labels is what we have used as test collection. As an example, one generated question was *"What is the discount percentage on the monthly housing quota for large or single-parent families of general category and special category?"*, and its ground truth answer was determined to be *"The discount on the monthly housing fee will be 15% for large family or single-parent family title of general category, and 40% for large family or single-parent family title of special category"*.

Generated questions were fed into all RAG systems to get automatically generated answers. After that, they were manually evaluated by human subjective comparison to ground truth answers, according to the following previously defined Likert scale:

1. The response either explicitly states that the answer is unknown or is entirely incorrect and contains significant factual inaccuracies (hallucinations).
2. The response may be tangentially related to the question but contains major hallucinations that render it useless or potentially misleading.
3. The response contains some correct elements but fails to address the most relevant aspects of the question, severely limiting its usefulness to the user.
4. The response is correct and likely helpful but lacks important information.
5. The response is correct and provides an appropriate level of detail. There may be minor discrepancies compared to the established ground truth, but these are negligible and do not alter the overall meaning or usefulness of the response.

Concerning the objective evaluation of texts and assessing similarities between them, one of the core concepts in the literature has been representing words or documents as dense vectors (embeddings) in a high-dimensional space and later measuring their similarity using cosine similarity. [Mikolov et al., 2013] demonstrated the effectiveness of cosine similarity in capturing semantic relationships between words. However, its efficacy has recently been questioned, and some research led by Netflix has shown its limitations, proposing evaluating with other metrics so that similarities are not arbitrarily drawn, as detailed in [Steck et al., 2024].

Embeddings of each ground truth and each synthetically generated answer were calculated and stored. Considering these recent debates about the arbitrariness or not of cosine similarity, it was computed along with Euclidean and Manhattan distances for each of the ground truth-generated answer pairs. The reasoning is that while cosine similarity accounts for both vectors laying in the same direction in the vectorial space defined by the embedding model, Euclidean and Manhattan account for those representations being geographically close. As administrative texts are considered fairly exempt from double meanings and other semantic issues normally addressed by cosine similarity, metrics based on spatial distance could be more meaningful and more effective at reflecting details likely to be hidden by cosine similarity in this context. To make them more meaningful in interpretation, each of the calculated values for Euclidean and Manhattan distances has been subtracted from 1 to get metrics where the higher the value, the greater the precision and fidelity to the ground truth.

Apart from that, metrics from ROUGE (Recall-Oriented Understudy for Gisting Evaluation) family have been used. They are a widely used set of metrics for evaluating the quality of automatic summaries and machine translations by comparing them to human-generated reference texts, as proposed in [Lin, 2004]. The metrics are based on the overlap of n-grams (contiguous sequences of n words) between the system-generated text and the reference text.

Several variations of ROUGE exist, each focusing on different aspects of the text, of which it is worth mentioning:

- **ROUGE-N.** Measures the overlap of n-grams between the system and reference texts. Common values for N are 1 (unigrams) and 2 (bigrams).
- **ROUGE-L.** Assesses the longest common subsequence (LCS) between the system and reference texts, which can capture sentence-level structures better than ROUGE-N.

ROUGE scores range from 0 to 1, with higher scores indicating a stronger similarity between the system-generated text and the reference text. While ROUGE has been criticized for its limitations in capturing meaning and coherence, as in [Ganesan, 2018], it remains a popular and practical tool for evaluating text generation systems due to its simplicity and computational efficiency. Therefore, for each synthetically generated answer, ROUGE-1, ROUGE-2, and ROUGE-L have also been computed.

Finally, the correlation of each calculated metric to the corresponding Likert labels has been computed individually for each of the RAGs built as a way to assess which of them seem more likely to reflect response accuracy.

4 Results

As Figure 2 shows, Large Language Models tested tend to show a slightly higher average in Likert subjective punctuation than Small Language Models, with GPT-4o showing the highest recorded average. However, there are no significant

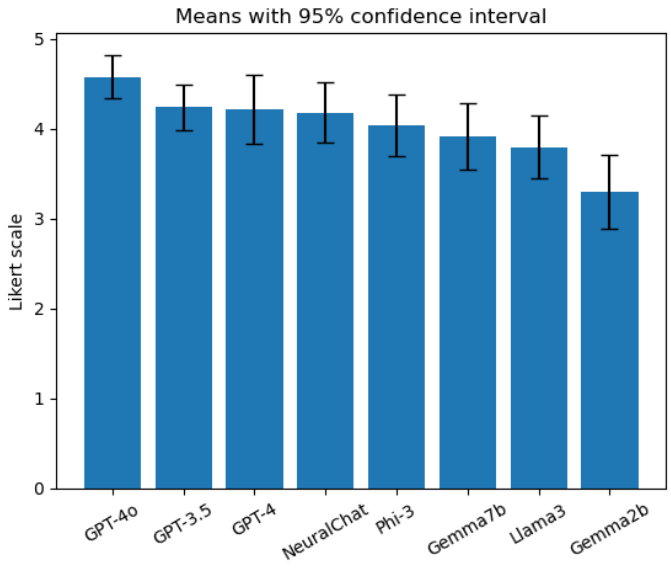

Fig. 2. Average of subjective Likert scores of the answers generated by each of the RAG systems. Shown with a confidence interval with $\alpha = 0.95$.

differences at α of 0.95 with the punctuation obtained by RAGs involving NeuralChat and Phi 3. GPT-4o only shows significant differences with both versions of Gemma and LLaMA 3, while GPT-3.5 and GPT-4 only differ significantly from the 2 billion parameters version of Gemma. As can be seen, NeuralChat almost catches up with the performance of GPT-4 in the experiments run.

Cosine similarity in this context seems to show little improvement to Euclidean and Manhattan-based similarities in most cases, with Manhattan outperforming it in the case of Phi 3 evaluation, as Figure 3 shows, even though all three metrics show poor performance in that example given. It is worth mentioning that cosine, Euclidean, and Manhattan-defined metrics are highly sensitive to how each generative model rearranges information when expressing its answer, as from the conclusions which can be drawn from [Steck et al., 2024]. It can be asserted that, in the case of GPT-4-based generation, all three metrics seem to be good predictors of the reliability of the answer in this specific context. Additionally, it can be seen that these metrics are also moderately informative in the case of SLMs with the exception of Phi 3. The fact that their lower complexity makes them more propense to replicate pieces of the context retrieved seems to leverage the association between the metrics measured in the embedding space and the loyalty to the actual expected answer more than it shows in GPT-3.5. Overall, as all observed correlations are strictly positive, it can be ensured they have certain mutual information in common with human subjective labels.

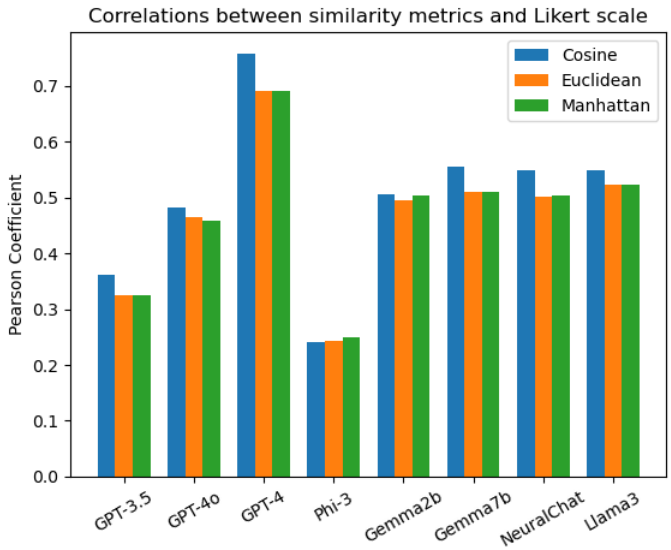

Fig. 3. Pearson correlation coefficient between subjective Likert punctuation and each of the embedding-related metrics for each RAG system.

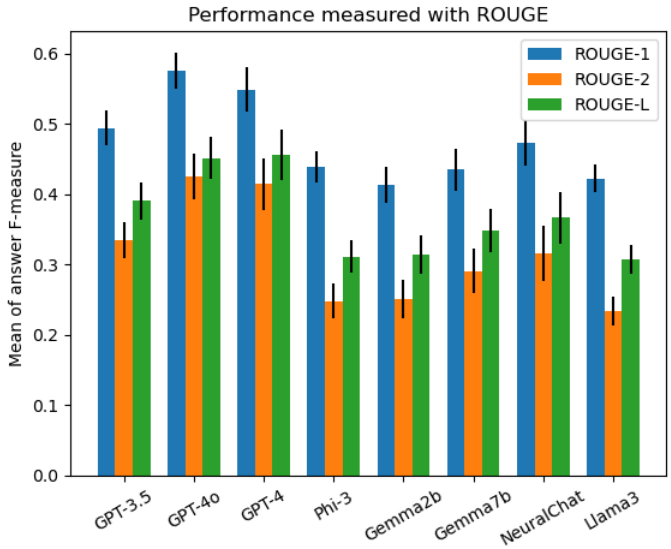

Fig. 4. Average of answer F-measure for each of the RAG systems built. Shown with a confidence interval with $\alpha = 0.95$ for each RAG system.

In the case of ROUGE metrics, a slight but clear outperform can be seen in GPT-4o and GPT-4 generated answers, which are statistically different from every other observed metric at an α of 0.95, as Figure 4 shows. NeuralChat and GPT-3.5 show near-equivalent performance with no significant differences, and the rest of the SLMs seem to show similar metrics, which are slightly worse than those from GPT-3.5 when they appear as significantly different.

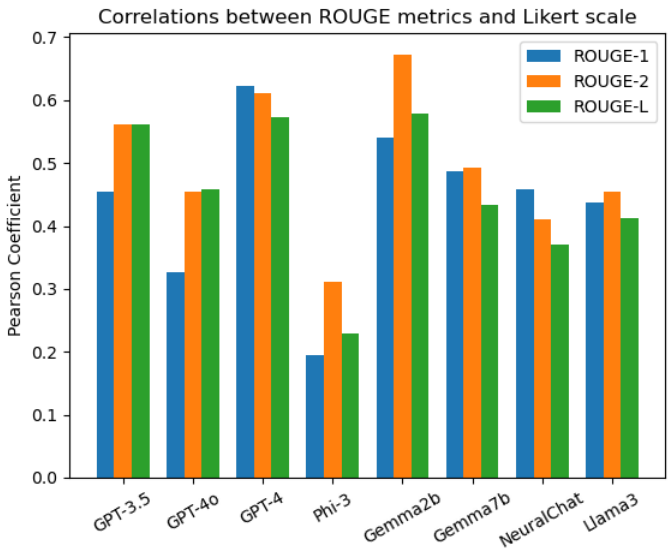

Fig. 5. Pearson correlation coefficient between subjective Likert punctuation and each of the ROUGE metrics measured for each RAG system.

When measuring correspondence between ROUGE metrics and human subjective labels, we see that the highest correspondence belongs to the simplest model employed, Gemma in its 2 billion parameters version, as Figure 5 shows. Again, the lowest correspondence appears in Phi 3. Apart from the finding that ROUGE-2 shows the highest correlation with subjective labeling in the simplest models (Phi 3 and Gemma with 2 billion parameters), followed by ROUGE-L and then ROUGE-1, and noting that GPT-3.5 and GPT-4o exhibit relatively lower ROUGE-1 scores compared to the other two metrics, no other consistent patterns emerge across the models when comparing these metrics.

Considering the observed relationships, it is evident that ROUGE metrics offer a quantitative insight that correlates appreciably with subjective evaluations captured through Likert scales. This correlation highlights the potential of ROUGE metrics to serve as a reliable metric for assessing semantic accuracy in the responses generated by SLMs. Consequently, including ROUGE-based evaluations could enhance the robustness of the assessment framework for SLMs,

providing an objective measure that aligns well with human judgment. This integration could also facilitate more nuanced and precise model performance evaluations, driving further advancements in developing semantically coherent language models.

5 Conclusions

While a RAG built with GPT-4o appears to be the best option for ensuring optimal semantic accuracy in an administrative context, its size and high computational cost suggest that exploring alternatives like NeuralChat might be a viable solution. NeuralChat, which can operate on conditional hardware and demonstrates comparable performance, could make the system more affordable and accessible to the general public.

This article has also analysed the different existing metrics for evaluating RAG models. Concerning this analysis, it has been observed that the cosine-based distance and, above all, the ROUGE-2 metric present a high correlation with the evaluations of the responses made by humans following the Likert scale.

Finally, the objective evaluation of NeuralChat's semantic agreement with expected answers through the embedding space metrics mentioned (including cosine similarity, Euclidean distance, Manhattan distance, and ROUGE metrics) opens new possibilities for model performance enhancement. Related to that, one future work could involve training machine learning models to predict the likelihood of the chatbot's answers being correct. When these models predict an answer as potentially incorrect, an alternative approach could be employed using a GPT-4o-based RAG. This secondary system would handle a smaller subset of questions, thereby ensuring a more accurate and efficient resolution of queries that NeuralChat fails to address satisfactorily. This hybrid approach could leverage models to improve response accuracy and reliability.

The fact that NeuralChat shows better results than other SLMs in terms of subjective labeling suggests that the fact that it has been fine-tuned to work in chatbot contexts might enhance its performance compared to other models with a similar number of parameters. Therefore, a feasible way of further improving it could be to fine-tune over NeuralChat in this specific administrative context by adopting the hybrid approach proposed by [Gao et al., 2024].

Acknowledgments. This work was funded by the Norwegian Research Council grant 329745 Machine Teaching for Explainable AI, CIPROM/2022/6 (FASSLOW) funded by Generalitat Valenciana, the EC H2020-EU grant agreement No. 952215 (TAILOR), and Spanish grant PID2021-122830OB-C42 (SFERA) funded by MCIN / AEI / 10.13039 / 501100011033 and "ERDF A way of making Europe". Authors thank the Cátedra de Inteligencia Artificial aplicada a la Administración Pública of Universitat Politècnica de València (UPV).

References

[Abdallah et al., 2023] Abdallah, A., Piryani, B., and Jatowt, A. (2023). Exploring the state of the art in legal qa systems. *Journal of Big Data*, 10(1).

[Abdin et al., 2024] Abdin, M., Jacobs, S. A., Awan, A. A., Aneja, J., Awadallah, A., Awadalla, H., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., et al. (2024). Phi-3 technical report: A highly capable language model locally on your phone.

[Gan et al., 2023] Gan, W., Qi, Z., Wu, J., and Lin, J. C.-W. (2023). Large language models in education: Vision and opportunities.

[Ganesan, 2018] Ganesan, K. (2018). Rouge 2.0: Updated and improved measures for evaluation of summarization tasks.

[Gao et al., 2024] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., and Wang, H. (2024). Retrieval-augmented generation for large language models: A survey.

[Garigliotti et al., 2024] Garigliotti, D., Johansen, B., Kallestad, J. V., Cho, S.-E., and Ferri, C. (2024). EquinorQA: Large Language Models for Question Answering over proprietary data. In *ECAI 2024 - 27th European Conference on Artificial Intelligence - Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024)*. IOS Press.

[Huber et al., 2024] Huber, J. et al. (2024). Chroma docs — docs.trychroma.com. https://docs.trychroma.com/. [Accessed 11-06-2024].

[Karpukhin et al., 2020] Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and tau Yih, W. (2020). Dense passage retrieval for open-domain question answering.

[Lewis et al., 2021] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., et al. (2021). Retrieval-augmented generation for knowledge-intensive nlp tasks.

[Lin, 2004] Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. page 10.

[Lv et al., 2023] Lv, L., Ren, X., Ye, X., Lv, K., Gao, Q., Tian, F., and Shen, H. (2023). Neuralchat: A customizable chatbot framework.

[Lála et al., 2023] Lála, J., O'Donoghue, O., Shtedritski, A., Cox, S., Rodriques, S. G., and White, A. D. (2023). Paperqa: Retrieval-augmented generative agent for scientific research.

[Mikolov et al., 2013] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space.

[OpenAI et al., 2024] OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., et al. (2024). Gpt-4 technical report.

[Ovadia et al., 2024] Ovadia, O., Brief, M., Mishaeli, M., and Elisha, O. (2024). Fine-tuning or retrieval? comparing knowledge injection in llms.

[Steck et al., 2024] Steck, H., Ekanadham, C., and Kallus, N. (2024). Is cosine-similarity of embeddings really about similarity? In *Companion Proceedings of the ACM on Web Conference 2024*, WWW '24. ACM.

[Team et al., 2024] Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., et al. (2024). Gemma: Open models based on gemini research and technology.

[Touvron et al., 2023] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models.

[Zhang et al., 2023] Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., et al. (2023). Siren's song in the ai ocean: A survey on hallucination in large language models.