Exploring Large Language Models for Relevance Judgments in Tetun

Gabriel de Jesus[1,2,*], Sérgio Nunes[1,2]

[1]INESC TEC - Institute for Systems and Computer Engineering, Technology and Science, Portugal
[2]FEUP - Faculty of Engineering, University of Porto, Portugal

Abstract

The Cranfield paradigm has served as a foundational approach for developing test collections, with relevance judgments typically conducted by human assessors. However, the emergence of large language models (LLMs) has introduced new possibilities for automating these tasks. This paper explores the feasibility of using LLMs to automate relevance assessments, particularly within the context of low-resource languages. In our study, LLMs are employed to automate relevance judgment tasks, by providing a series of query-document pairs in Tetun as the input text. The models are tasked with assigning relevance scores to each pair, where these scores are then compared to those from human annotators to evaluate the inter-annotator agreement levels. Our investigation reveals results that align closely with those reported in studies of high-resource languages.

Keywords

Large language models, Relevance judgments, Low-resource languages, Tetun

1. Introduction

The advancement of information retrieval (IR) systems depends on the availability of reliable test collections to assess their effectiveness. The traditional approach for developing these collections follows the Cranfield paradigm [1], which became widely recognized through the Text REtrieval Conference (TREC) series of large-scale evaluation campaigns [2]. In TREC guidelines, a test collection comprises a document collection, a set of topics, and corresponding relevance assessments. The relevance judgment tasks are typically carried out by human assessors, a process that is both time-consuming and costly.

To tackle the aforementioned problems, the IR community has been investigating the feasibility of automatically generated relevance judgments for developing test collections. With the advent of large language models (LLMs), which have demonstrated proficiency in various tasks, new possibilities for conducting automated relevance judgments have emerged, demonstrating ongoing improvement in the quality of automated relevance judgment tasks as LLMs continue to evolve.

LLM4Eval: The First Workshop on Large Language Models for Evaluation in Information Retrieval, 18 July 2024, Washington DC, United States.

*Corresponding author.
gabriel.jesus@inesctec.pt (G. d. Jesus); sergio.nunes@fe.up.pt (S. Nunes)
https://web.fe.up.pt/~ssn/ (S. Nunes)
0000-0003-4392-2382 (G. d. Jesus); 0000-0002-2693-988X (S. Nunes)

Studies have consistently shown that LLMs are effective in automated relevance assessment tasks, providing their cost-effectiveness solutions with judgment agreement comparable to human assessors. Faggioli et al. [3] argued that although further improvement in LLMs capabilities is necessary for fully automated relevance judgments, LLMs are already capable of assisting humans in this task. Additionally, a recent study by Bueno et al. [4] reported a consistent improvement in automated relevance judgments with an average Cohen's Kappa score of 0.31 for annotation agreement between humans and LLMs, which are inline with the findings of Faggioli et al. [3]. However, these studies primarily focus on high-resource languages, such as English and Brazilian Portuguese, leaving the applicability of LLMs in low-resource language (LRL) contexts as an open question.

In this study, we explore the use of LLMs to automate relevance judgment tasks in Tetun, a LRL spoken by over 923,000 people in Timor-Leste [5]. We used an existing test collection comprising 6,100 relevance judgments constructed utilizing documents from the Labadain-30k+ dataset [6]. The relevance judgments for this collection were conducted by native Tetun speakers. These query-document pairs were provided to the LLMs to assign relevance scores for each. We compared these scores with those from human annotations and observed inter-annotator agreement levels. The results revealed an inter-annotator agreement of Cohen's kappa score of 0.2634 when evaluated using the 70B variant of the LLaMA3 model [7]. This finding demonstrates the feasibility of using LLMs in LRL scenarios to automate the relevance judgment tasks.

The remaining sections of this paper are organized as follows. Section 2 describes related work. An overview of the collection used in this study is outlined in Section 3. Then, Section 4 details the experiment of using LLMs for automating relevance judgments. Section 5 presents the results obtained and their discussion. Finally, Section 6 summarizes our conclusion and possible future work.

2. Related Work

Test collections are the most important component used for evaluating the effectiveness of IR systems. For high-resource languages, these collections are typically made available through large-scale campaigns such as Text REtrieval Conference (TREC)[1], the Conference and Labs of the Evaluation Forum (CLEF)[2], the NII Testbeds and Community for Information Access Research project (NTCIR)[3], and the Forum for Information Retrieval Evaluation (FIRE)[4].

The TREC-style approach, derived from the Cranfield paradigm, is commonly adopted for developing test collections, including for low-resource languages (LRLs), where human assessors conduct the relevance judgment tasks [8, 9, 10, 11]. However, the fast pace of research and innovation, particularly with the emergence of LLMs, has significantly transformed natural language processing (NLP). Within the IR domain, studies have demonstrated that automated relevance judgments using LLMs can yield results comparable to traditional methods, and

[1] https://trec.nist.gov
[2] https://www.clef-initiative.eu
[3] http://research.nii.ac.jp/ntcir/index-en.html
[4] http://fire.irsi.res.in/

these outcomes have consistently improved as LLMs have evolved. Initially, Faggioli et al. [3] explored the potential application of LLMs to fully-automated relevance judgment tasks. They analyzed the judgment results from the TREC 2021 Deep Learning track [12] and compared them with LLM-based relevance assessments generated using GPT-3.5 of OpenAI[5]. Their findings revealed a Cohen's kappa score of 0.26 for inter-annotator agreement between human and LLM, indicating a fair level of agreement. Thus, they argued that LLMs are already capable of assisting humans in relevance judgment tasks, despite further improvement in LLM capabilities are necessary for fully automated relevance judgments.

Later, Thomas et al. [13] reported that LLMs demonstrated accuracy comparable to human labelers when deployed for large-scale relevance labeling at Bing. Their work utilized the GPT-4 model [14] and incorporated data from the TREC Robust04 track [15], showing that LLMs achieved a Cohen's kappa score ranging from 0.20 to 0.64 for agreement between humans and LLM across various tasks. In a recent study, Bueno et al. [4] in their study while constructing a test collection for Brazilian Portuguese, reported consistent improvement and findings comparable to those of Thomas et al. [13] and Faggioli et al. [3], with automated relevance judgments yielding an average Cohen's Kappa score of 0.31 for annotation agreement between humans and LLMs.

Despite these advancements, uncertainties persist about the feasibility of using LLMs to automatically generating relevance judgments for LRLs. Thus, our research focuses on exploring this potential application in LRL scenarios, specifically in Tetun.

3. Collection Overview

In this experiment, we utilized the existing Tetun test collection[6] developed according to TREC guidelines. The following subsections detail the test collection used in this work.

3.1. Documents

Documents of the Tetun test collection are derived from the Labadain-30k+ dataset, which consists of 33,550 documents in Tetun [6]. This dataset was acquired from the web and encompassed a broad array of categories, including news articles, Wikipedia entries, legal and government documents, research papers, and more [16]. A summary of the document collection is provided in Table 1.

3.2. Queries

The collection consists of 61 queries developed by five volunteer students, all Timoreses and native Tetun speakers. The queries are originated from the logs of Timor News[7], an online newspaper based in Dili, Timor-Leste. Statistics about the queries are presented in Table 2.

[5]https://openai.com
[6]This collection has not yet been published.
[7]https://www.timornews.tl

Table 1
Summary of Document Collection. *Tokens Comprise Words and Numbers, Excluding Punctuation and Special Characters.

Description	Value
Total number of documents	33,550
Size of collection	84MB
Total number of tokens*	12,300,237
Total number of distinct tokens	162,466

Table 2
Summary of Queries.

Description	Value
Total number of queries	61
Minimum number of words per query	3
Maximum number of words per query	5
Average numbers of words per query	3.42

Table 3
Human Relevance Judgment Results.

Relevance Level	Total	Proportion
3 - Highly relevant	710	11.64%
2 - Relevant	1,102	18.07%
1 - Marginally relevant	476	7.80%
0 - Irrelevant	3,812	62.49%

3.3. Relevance Judgments

Relevance judgments were conducted by the same five Timorese students. These students were tasked with evaluating the relevance of query-document pairs. The pairs were classified into four graded levels of topical relevance: irrelevant, marginally relevant, relevant, and highly relevant, as proposed by Sormunen [17]. The inter-annotator agreement achieved an average a Cohen's kappa score of 0.4236 and the details of the resulting test collection are presented in Table 3.

4. Relevance Judgments Using LLMs

4.1. Overview

Several studies have already utilized the GPT-3.5 and GPT-4 models from OpenAI for automating relevance judgment tasks [3, 13, 4]. However, due to the costs associated with these LLMs, our study explores an alternative by employing the freely available 70B variant of LLaMA3, released by Meta on April 18, 2024 [18]. We conduct automated relevance judgments using the Tetun

Table 4
Price Associated with LLMs.

Model	Input	Output
LLaMA 3	—	—
Claude 3 Haiku	$0.25 / 1M tokens	$1.25 / 1M tokens
GPT-3.5-turbo-0125	$0.50 / 1M tokens	$1.50 / 1M tokens

Table 5
Examples of Tetun to English Translations Provided by LLMs.

Tetun text:	**Juventude hetan virus infeksaun HIV/SIDA tanba menus informasaun.**
Model	Translation
LLaMA 3	Youth ~~against~~ virus infection of HIV/AIDS ~~and~~ lack of information.
Claude 3 Haiku	Youth get HIV/AIDS virus infection due to lack of information.
GPT-3.5-turbo-0125	The youth are getting infected with HIV/AIDS because of lack of information.

test collection detailed in section 3, to compare their inter-annotator agreement levels.

Additionally, to evaluate whether the free LLaMA3 model of 70B variant can outperform certain paid LLMs in relevance assessment tasks, specifically within the Tetun context, we have selected two paid models for comparison: the Haiku variant of Claude 3 from Anthropic[8], and the Turbo variant of GPT-3.5 from OpenAI. A summary of the models used, along with their associated costs, is presented in Table 4.

To assess the suitability of the chosen LLMs for Tetun, including the two paid models, we conducted preliminary tests that involved translating Tetun text into English. This step was essential given that the query-document pairs are written in Tetun. Examples of these translated outputs are presented in Table 5, showing that LLaMA3 inaccurately translated two words, as indicated by strike-through markings.

To evaluate the quality of the translated text generated by the LLMs, we randomly selected a sample of five documents from the query-document pairs (see example in Table 8), and translated them into English. These human translations served as reference points for evaluation. The assessment using the BLEU metric [19], demonstrates that both paid models outperformed LLaMA3 in translating Tetun to English, as shown in Table 6.

However, given that relevance judgment tasks require not only direct translation but also a nuanced level of understanding, we compared the selected models' multi-task language understanding capabilities using the Massive Multitask Language Understanding (MMLU) [20] based on the MMLU benchmark leaderboard [21]. A summary of these LLMs' performance on MMLU is outlined in Table 7. It shows that in the few-shot scenario with five examples, LLaMA 3 surpassed Claude 3 Haiku by an average of +5 percentage points and GPT-3.5 Turbo by +10.2 percentage points.

[8]https://www.anthropic.com

Table 6
Evaluation of Translation Quality Produced by LLMs Using BLEU.

Model	BLEU
LLaMA 3	0.3849
Claude 3 Haiku	0.4167
GPT-3.5 Turbo	0.6525

Table 7
LLM Capabilities of Multi-task Language Understanding on MMLU.

Model	Average (%)
LLaMA 3 (5-shot)	80.2
Claude 3 Haiku (5-shot)	75.2
GPT-3.5 Turbo (5-shot)	70.0

Table 8
Examples of Query-Document Pairs.

Query	Document
Prevensaun moras HIV/SIDA	UNFPA Sei Koopera ho MS Hodi halo Prevensaun ba Moras HIV/SIDA
Prevensaun moras HIV/SIDA	KNK-HIV/SIDA Sensibiliza Informasaun HIV/SIDA Ba Trabal-lador KSTL
Prevensaun moras HIV/SIDA	Autoridade Lokál Partisipa Workshop Prevensaun Moras HIV/SIDA

4.2. Experiment with Tetun

To automate relevance judgments using LLMs, we utilized few-shot prompting, adopting a structure similar to that employed by Bueno et al. [4]. Our prompt, along with an example, is illustrated in Prompt 4.1, and the full prompt is outlined in Appendix A. We provided the LLMs with a total of 6,100 query-document pairs and tasked the LLMs with assigning a relevance score to each. Examples of these query-document pairs are depicted in Table 8.

Given that the existing Tetun test collection employs four-level relevance scores ranging from 0 to 3, we provided the LLMs with query-document pairs alongside four examples, one for each relevance score. These examples used the same queries as those utilized in the pilot testing phase by human assessors, including the relevance score and the reasoning behind each score. For each request, we asked the LLMs to assign one of the four scores and provide the reasoning for their assigned score.

For the 70B variant of the LLaMA 3 model, which requires a substantial amount of memory to run locally, specifically a minimum of 40 GB of RAM as indicated on Ollama[9], we utilized the

[9]https://ollama.com/library/llama3:70b

free API version of the cloud infrastructure provided by Groq[10] to execute this model. However, the scripts for automated relevance judgments for all models were executed locally.

Prompt 4.1: Example of the System Prompt.

You are an expert assessor and you are tasked with assessing the relevance between the input query and its corresponding document, assigning a score from 0 to 3. A score of 0 indicates irrelevant; 1, marginally relevant; 2, relevant; and 3, highly relevant.

Example:

query: "Kursu mestradu no pós-graduasaun UNTL"
document: "Kursu Desportu UNTL sei realiza graduasaun dahuluk tinan ne'e"
reason: "The query is about postgraduate and master's courses at UNTL, whereas the document focuses on a sports course. Despite both courses in the query and document being offered at UNTL, the sports course in the document is not specifically designed for postgraduate or master's levels. Thus, the document is only marginally relevant."
score: 1

The query and document to be evaluated are the following:

query: {*query*}
document: {*document*}

Your response must be in JSON format with the first field is "reason", explaining your reasoning, and the second field is "score".

We initiated the experiment with the LLaMA3 70B model, as it was our primary target for comparing annotator agreement level with human annotators. We tested this model using temperatures of 0.0 and 0.5, respectively. The concept of comparing different model temperatures in inter-annotator agreement was inspired by the work of Ma et al. [22], who applied LLMs for relevance judgments in Chinese legal case retrieval. When we increased the temperature of LLaMA3 70B model, the results were not satisfactory. Therefore, we opted to use a zero temperature setting in the other paid models for comparison.

5. Results and Discussions

In the experiment with the LLaMA3 70B model set at zero temperature, we obtained an inter-annotator agreement of Cohen's kappa score of 0.2634 with human annotators. After increasing the temperature to 0.5, the inter-annotator agreement slightly decreased to 0.2594 (a reduction of -0.004). This finding aligns with the research by Ma et al. [22], where their Cohen's kappa score of inter-annotator agreement levels between humans and LLMs also marginally decreased

[10] https://console.groq.com/settings/billing

Table 9
Cohen's Kappa Correlations Between Human Assessors and LLMs.

	LLaMA3 70B	Claude3 Haiku	GPT-3.5-turbo-0125
Human	**0.2634**	0.2450	0.2462

Table 10
Comparison of the Inter-Annotator Agreement Levels in the Relevance Judgment Tasks using LLMs.

	LLM	Cohen's k Score
Bueno et al. [4]	GPT-4	0.31
Thomas et al. [13]	GPT-4	0.26 — 0.64
Faggioli et al. [3]	GPT 3.5	0.26
Ours	LLaMA3 70B	0.26

Table 11
Expense on LLMs for Relevance Judgment Tasks.

Model	Expense
LLaMA 3	—
Claude 3 Haiku	$1.98
GPT-3.5 Turbo	$2.73

when they raised the temperature from 0.4 to 0.7 in evaluations of material facts.

Consequently, we opted for a zero temperature setting when conducting relevance judgments with the Claude3 Haiku and GPT-3.5 Turbo models. Comparisons of the inter-annotator agreement levels between LLMs and human annotators are presented in Table 9. These results show that the LLaMA3 70B model achieved a highest Cohen's kappa score, indicating the most substantial agreement with human annotators compared to both paid models. Among the paid models, GPT-3.5 Turbo exhibited a slightly higher Cohen's kappa score than Claude3 Haiku (a k score increase of +0.0012). Thus, despite the superior performance of the paid models in translating Tetun into English, this finding suggests that a deeper level of language understanding is more crucial in automated relevance judgment tasks.

As a result, our finding using LLaMA3 70B model is closely aligned with the initial results reported by Faggioli et al. [3], and are consistent with the findings of Bueno et al. [4] and Thomas et al. [13]. Comparisons of these findings regarding the use of LLMs to automate relevance judgments are presented in Table 10.

Furthermore, our experiments took an average of approximately 3.56 hours to complete the relevance judgment tasks for each model. The costs associated with the two paid models are detailed in Table 11. Given that GPT-3.5 Turbo is priced $0.25 higher per use than Claude 3 Haiku for every 1 million input and output tokens, the expenses for GPT-3.5 were higher than those for Claude 3 Haiku.

6. Conclusions and Future Work

Our exploration into leveraging large language models for automating relevance judgment tasks in low-resource language scenarios, demonstrated using Tetun, has yielded results comparable to those achieved in high-resource languages, thus encouraging further research in low-resource languages (LRLs). The availability of freely and openly accessible models like LLaMA3 opens up possibilities for advancing relevance judgment tasks, particularly in low-resource language contexts, even with the limited digital content available on the web.

Our experiment demonstrated that despite LLaMA3's knowledge being limited to December 2023[11] and the availability of fewer than 45k Tetun documents on the web by that time [23, 16], it achieved an agreement level comparable to high-resource languages like English. This indicates that automated relevance judgment tasks are feasible for other LRLs as well.

In future work, we plan to extend this research by incorporating a wider variety of examples in our prompts and testing with other freely and openly available models to compare the results. This approach will help validate and potentially expand the use of large language models in relevance judgment tasks.

7. Acknowledgment

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020 (DOI 10.54499/LA/P/0063/2020) and the Ph.D. scholarship grant number SFRH/BD/151437/2021 (DOI 10.54499/SFRH/BD/151437/2021).

References

[1] C. Cleverdon, The cranfield tests on index language devices, in: Aslib proceedings, volume 19, MCB UP Ltd, 1967, pp. 173–194.

[2] D. K. Harman (Ed.), Proceedings of The First Text REtrieval Conference, TREC 1992, Gaithersburg, Maryland, USA, November 4-6, 1992, volume 500-207 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 1992. URL: http://trec.nist.gov/pubs/trec1/t1_proceedings.html.

[3] G. Faggioli, L. Dietz, C. L. A. Clarke, G. Demartini, M. Hagen, C. Hauff, N. Kando, E. Kanoulas, M. Potthast, B. Stein, H. Wachsmuth, Perspectives on large language models for relevance judgment, in: M. Yoshioka, J. Kiseleva, M. Aliannejadi (Eds.), Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2023, Taipei, Taiwan, 23 July 2023, ACM, 2023, pp. 39–50. URL: https://doi.org/10.1145/3578337.3605136. doi:10.1145/3578337.3605136.

[4] M. Bueno, E. S. de Oliveira, R. Nogueira, R. A. Lotufo, J. A. Pereira, Quati: A brazilian portuguese information retrieval dataset from native speakers, 2024. arXiv:2404.06976.

[5] G. de Jesus, Text information retrieval in Tetun, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Infor-

[11] https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

mation Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part III, volume 13982 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 429–435. URL: https://doi.org/10.1007/978-3-031-28241-6_48. doi:10.1007/978-3-031-28241-6_48.

[6] G. de Jesus, S. Nunes, Labadain-30k+: A monolingual Tetun document-level audited dataset [data set]. INESC TEC, https://doi.org/10.25747/YDWR-N696, 2024.

[7] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. 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, CoRR abs/2302.13971 (2023). URL: https://doi.org/10.48550/arXiv.2302.13971. doi:10.48550/ARXIV.2302.13971. arXiv:2302.13971.

[8] S. S. Sahu, S. Pal, Building a text retrieval system for the sanskrit language: Exploring indexing, stemming, and searching issues, Comput. Speech Lang. 81 (2023) 101518. URL: https://doi.org/10.1016/j.csl.2023.101518. doi:10.1016/J.CSL.2023.101518.

[9] C. Chavula, H. Suleman, Ranking by language similarity for resource scarce southern bantu languages, in: F. Hasibi, Y. Fang, A. Aizawa (Eds.), ICTIR '21: The 2021 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Canada, July 11, 2021, ACM, 2021, pp. 137–147. URL: https://doi.org/10.1145/3471158.3472251. doi:10.1145/3471158.3472251.

[10] K. S. Esmaili, D. Eliassi, S. Salavati, P. Aliabadi, A. Mohammadi, S. Yosefi, S. Hakimi, Building a test collection for sorani kurdish, in: ACS International Conference on Computer Systems and Applications, AICCSA 2013, Ifrane, Morocco, May 27-30, 2013, IEEE Computer Society, 2013, pp. 1–7. URL: https://doi.org/10.1109/AICCSA.2013.6616470. doi:10.1109/AICCSA.2013.6616470.

[11] A. AleAhmad, H. Amiri, E. Darrudi, M. Rahgozar, F. Oroumchian, Hamshahri: A standard persian text collection, Knowl. Based Syst. 22 (2009) 382–387. URL: https://doi.org/10.1016/j.knosys.2009.05.002. doi:10.1016/J.KNOSYS.2009.05.002.

[12] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, J. Lin, Overview of the TREC 2021 deep learning track, in: I. Soboroff, A. Ellis (Eds.), Proceedings of the Thirtieth Text REtrieval Conference, TREC 2021, online, November 15-19, 2021, volume 500-335 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2021. URL: https://trec.nist.gov/pubs/trec30/papers/Overview-DL.pdf.

[13] P. Thomas, S. Spielman, N. Craswell, B. Mitra, Large language models can accurately predict searcher preferences, CoRR abs/2309.10621 (2023). URL: https://doi.org/10.48550/arXiv.2309.10621. doi:10.48550/ARXIV.2309.10621. arXiv:2309.10621.

[14] OpenAI, GPT-4 technical report, CoRR abs/2303.08774 (2023). URL: https://doi.org/10.48550/arXiv.2303.08774. doi:10.48550/ARXIV.2303.08774. arXiv:2303.08774.

[15] E. M. Voorhees, Overview of the TREC 2004 robust track, in: E. M. Voorhees, L. P. Buckland (Eds.), Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004, volume 500-261 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2004. URL: http://trec.nist.gov/pubs/trec13/papers/ROBUST.OVERVIEW.pdf.

[16] G. de Jesus, S. Nunes, Labadain-30k+: A monolingual Tetun document-level audited dataset, in: M. Melero, S. Sakti, C. Soria (Eds.), Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024, ELRA

and ICCL, Torino, Italia, 2024, pp. 177–188. URL: https://aclanthology.org/2024.sigul-1.22.
[17] E. Sormunen, Liberal relevance criteria of TREC -: counting on negligible documents?, in: K. Järvelin, M. Beaulieu, R. A. Baeza-Yates, S. Myaeng (Eds.), SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland, ACM, 2002, pp. 324–330. URL: https://doi.org/10.1145/564376.564433. doi:10.1145/564376.564433.
[18] Meta, Introducing meta llama 3: The most capable openly available llm to date, 2024. URL: https://llama.meta.com/llama3/.
[19] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: P. Isabelle, E. Charniak, D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040. doi:10.3115/1073083.1073135.
[20] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, J. Steinhardt, Measuring massive multitask language understanding, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net, 2021. URL: https://openreview.net/forum?id=d7KBjmI3GmQ.
[21] P. with Code, Multi-task language understanding on mmlu, 2024. URL: https://paperswithcode.com/sota/multi-task-language-understanding-on-mmlu.
[22] S. Ma, C. Chen, Q. Chu, J. Mao, Leveraging large language models for relevance judgments in legal case retrieval, CoRR abs/2403.18405 (2024). URL: https://doi.org/10.48550/arXiv.2403.18405. doi:10.48550/ARXIV.2403.18405. arXiv:2403.18405.
[23] S. Kudugunta, I. Caswell, B. Zhang, X. Garcia, C. A. Choquette-Choo, K. Lee, D. Xin, A. Kusupati, R. Stella, A. Bapna, O. Firat, MADLAD-400: A multilingual and document-level large audited dataset, CoRR abs/2309.04662 (2023). URL: https://doi.org/10.48550/arXiv.2309.04662. doi:10.48550/ARXIV.2309.04662. arXiv:2309.04662.

A. System Prompt Details

Details of the system prompt used in the automated relevance judgments, including four examples of query-document pairs along with the reasoning and the corresponding score for each.

Prompt A.1: Details of the System Prompt.

You are an expert assessor and you are tasked with assessing the relevance between the input query and its corresponding document, assigning a score from 0 to 3. A score of 0 indicates irrelevant; 1, marginally relevant; 2, relevant; and 3, highly relevant.

Example 1:
query: "Programa mestradu no pós-graduasaun UNTL"
document: "Estudantes Pós-Graduasaun IOB Kuda Ai-Oan iha aldeia Payol no Bedois"
reason: "The query is about postgraduate and master's courses at UNTL, whereas the document discusses the activities of postgraduate students from IOB. Although both query and document contain the term 'postgraduate', the query specifically is targeted courses at UNTL. Therefore, they are irrelevant."
score: 0.

Example 2:
query: "Kursu mestradu no pós-graduasaun UNTL"
document: "Kursu Desportu UNTL sei realiza graduasaun dahuluk tinan ne'e"
reason: "The query is about postgraduate and master's courses at UNTL, whereas the document focuses on a sports course. Despite both courses in the query and document being offered at UNTL, the sports course in the document is not specifically designed for postgraduate or master's levels. Thus, the document is only marginally relevant."
score: 1.

Example 3:
query: "Kursu mestradu no pós-graduasaun UNTL"
document: "UNTL Nia Vise Reitór Asuntu Pós-Graduasaun No Peskiza Hakotu-iis"
reason: "The document is relevant as it details the vice-director of the postgraduate program at UNTL. However, its relevance is somewhat diminished as it primarily discusses the unfortunate passing of the vice-director rather than the progress or implementation of the program. Hence, they are relevant."
score: 2.

Example 4:
query: "Kursu mestradu no pós-graduasaun UNTL"
document: "UNTL Lansa Kursu Pós-Graduasaun No Mestradu Iha Área Lima"
reason: "Both the query and document address postgraduate and master's courses at UNTL. The document strongly correlates with the query, containing the launching of postgraduate and master's courses at UNTL. Thefore they are highly relevant."
score: 3.

The query and document to be evaluated are the following:

query: {*query*}
document: {*document*}

Your response must be in JSON format with the first field is "reason", explaining your reasoning, and the second field is "score".