Can we trust LLMs as relevance judges?

Luciana Bencke[1], **Felipe S. F. Paula**[1], **Bruno G. T. dos Santos**[1], **Viviane P. Moreira**[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

{luciana.bencke,fsfpaula,bruno.tsantos,viviane}@inf.ufrgs.br

Abstract. *Evaluation is key for Information Retrieval systems and requires test collections consisting of documents, queries, and relevance judgments. Obtaining relevance judgments is the most costly step in creating test collections because they demand human intervention. A recent tendency in the area is to replace humans with Large Language Models (LLMs) as the source of relevance judgments. In this paper, we investigate the use of LLMs as a source of relevance judgments. Our goal is to find out how reliable LLMs are in this task. We experimented with different LLMs and test collections in Portuguese. Our results show that LLMs can yield promising performance that is competitive with human annotations.*

1. Introduction

Since the beginning of the research area, evaluation has been a critical aspect of Information Retrieval (IR). The origins of the standard evaluation paradigm date back to the 1960s with the Cranfield experiments [Cleverdon 1960]. To compute retrieval quality metrics, one needs a test collection with three components: (i) a (large) corpus of documents, (ii) a set of queries that are representative of real user needs, and (iii) relevance judgments that inform which documents are relevant to each query. Obtaining these judgments is the most costly step in building test collections because it requires human interference. As a result, there is a lack of test collections for many languages.

Throughout the years, many techniques have been devised to alleviate the burden on human annotators, including the pooling method [Spärck Jones and van Rijsbergen 1975], using crowd workers [Blanco et al. 2011], and, more recently, replacing humans with Large Language Models (LLM) [Faggioli et al. 2023, Thomas et al. 2023, Soviero et al. 2024]. The manual relevance annotation process is both time-consuming and expensive because it involves large collections, requiring several annotators, and is particularly difficult when the annotators must be domain experts, such as geologic [Lima de Oliveira et al. 2021] or medical [Zhu et al. 2023] fields.

Recent advances in LLMs demonstrated their high performance for generative tasks, including producing text for intelligent assistants, generating entire synthetic datasets, or contributing to address hard database problems such as entity resolution, schema matching, data discovery, and query synthesis. Using an LLM as a judge is a promising alternative to traditional human evaluations [Zheng et al. 2024]. LLMs can be expected to consistently apply the same criteria across large datasets, which does not occur with human annotators [Theodosiou et al. 2011]. This scalability would be especially beneficial for enabling the rapid creation of large annotated test collections.

Specifically for IR, using LLMs as relevance judges has gained significant interest [Faggioli et al. 2023, Thomas et al. 2023, Soviero et al. 2024]. Highlighting this growing interest, SIGIR (which is the main international IR forum) has a Workshop focusing specifically on this topic in 2024 – LLM4Eval[1].

Annotation consistency is an important aspect that can be measured in two different ways: *inter-annotator* agreement, assessing the consensus level among different annotators, and *intra-annotator agreement*, gauging how consistent a single annotator is [Theodosiou et al. 2011]. More challenging documents and unclear annotation guidelines can contribute to variations. Annotation quality can also be influenced by domain complexities, the annotator's profile, and commitment to the task.

The main goal of this work is to find out whether LLMs can be trusted to perform relevance judgments for IR. We focus on Portuguese, which is the 6th largest language in number of native speakers, and yet it is underrepresented in terms of IR resources. We assessed the performance of LLMs as relevance judges under different perspectives – by measuring the correlation with human-generated judgments (*i.e.,* inter-annotator agreement) and by evaluating their consistency (*i.e.,* intra-annotator agreement). In addition, we tested several IR configurations varying the source of relevance judgments to assess the impacts of replacing human-based with LLM-based relevance labels.

We experimented with two test collections and two LLMs. Our results have shown that agreement between LLMs and humans ranges from substantial to fair. Moreover, we found that LLMs are highly stable when assessing the consistency of judgments, presenting high intra-annotator agreement. In addition, human- and LLM-based judgments yield the same relative IR performance. This consistency strongly suggests that LLM-generated relevance labels are reliable.

2. Related Work

The principles of IR evaluation were defined many decades ago with the Cranfield paradigm [Cleverdon 1960]. These first experiments relied on exhaustive relevance judgments (*i.e.,* all documents were evaluated w.r.t. all queries), and the process was repeated by more than one judge. However, this approach does not escalate for larger collections. Many alternatives to alleviate the burden on human assessors have been proposed throughout the years. The pooling method [Spärck Jones and van Rijsbergen 1975] was the first attempt in this direction and became a standard.

Researchers have always been concerned with the reliability of the evaluation paradigm, which is constantly being assessed. In the early 2000s, [Voorhees 2000] examined the impact of variability in experts' relevance judgments on the evaluation of information retrieval systems. The findings indicated that although there were slight variations in how judges assessed certain query-document pairs, the overall relative performance of different retrieval systems remained consistent. Over a decade later, novel relevance assessment methods were proposed, such as crowdsourcing, moving away from solely expert-based judgments. [Blanco et al. 2011] explored the reliability and repeatability of evaluation campaigns executed by crowd workers. Their study found that despite the differences observed between expert and crowd judgments, the crowdsourced evaluations

[1] `https://llm4eval.github.io/`

were generally reliable and offered a cost-effective alternative for conducting assessment campaigns.

The success of LLMs in other annotation tasks is forcing another paradigm shift. [Faggioli et al. 2023] present first perspectives and prospects on the use of LLMs as relevance judges. Their work presents contrasting opinions for and against using LLMs to evaluate relevance. They also perform a pilot study by re-annotating the TREC 2021-DL using GPT-3.5 and YouChat. [Thomas et al. 2023] also analyze the performance of LLMs as relevance labelers. They conclude that LLMs are better at evaluating relevance than several human populations. Recently, [Rahmani et al. 2024] has gone as far as to delegate relevance judgment and query generation to an LLM. The authors explore the creation of synthetic test collection utilizing a T5-based model and GPT-4, and their conclusion points out that evaluation results seem to be similar to those of a traditional test collection approach. Even though that work is preliminary, utilizing only one test collection and with the risk of a potential bias, entrusting LLMs to create a synthetic test collection marks an interesting milestone in relieving the burden on humans.

In the e-commerce domain, [Soviero et al. 2024] investigates the use of LLMs (GPT-3.5-turbo and GPT-4) to produce relevance judgments for products. Their findings show a high agreement between human judgments and LLMs-based judgments. Moreover, they assess this agreement in different scenarios, such as hard and easy query-product pairs and different prompting strategies.

Currently, LLM-judged collections are beginning to appear. The Quati collection [Bueno et al. 2024] is the first collection judged by LLM in Brazilian Portuguese. The passages in this collection are a subset of ClueWeb22[2], a large collection of web pages. The authors also performed a human evaluation of a sample and found a moderate agreement between humans and LLM. For Tetun, a low-resource language, [de Jesus and Nunes 2024] also explored the use of LLMs to evaluate retrieval collection. Their findings indicate an agreement level with human judgments that are on par with earlier English-based studies [Faggioli et al. 2023, Thomas et al. 2023].

Although existing work has already used LLMs as a source of relevance judgments, the stability of the relevance labels was not investigated. Furthermore, to the best of our knowledge, this area is still under-explored with models and collections in Portuguese. This research aims to fill that gap, providing insights into the applicability of LLMs as reliable tools in the IR evaluation landscape.

3. Materials and Methods

In order to find out whether LLMs can be trusted as relevance judges, this paper seeks to answer three research questions as follows.

RQ1 How correlated are LLM-generated and human-generated relevance judgments? LLM-relevance assessments should ideally be in agreement with human judgments.

RQ2 How stable are LLM-generated relevance judgments? A trustworthy judge should not vary their assessment given the same query and document.

RQ3 Can LLM-generated relevance assessments produce the same retrieval results as

[2]https://lemurproject.org/clueweb22/

those produced when human judgments are used? When used to compare different retrieval systems or configurations, the relative ordering of the systems obtained when LLM-generated relevance assessments should not be different from the order obtained when human judgments are used.

The experimental pipeline used to answer the research questions is depicted in Figure 1. During the annotation phase (Figure 1a), the relevance labels are obtained. Given a document collection D and a set of queries Q, the task of the annotation phase is to assign a relevance label $l \in L$ to a pair $\langle q, d \rangle$, where $q \in Q$ and $d \in D$. Thus, a relevance assessment is a triple $\langle q, d, l \rangle$. The labels in L can be binary or have multiple levels. Relevance judgments are not exhaustive since it is unfeasible to evaluate every document in relation to each query. As a result, the *pooling* method [Spärck Jones and van Rijsbergen 1975] is commonly employed. In Step 1, an IR system is used to retrieve candidate documents, which will compose the pool of documents that are judged. It is common to use different retrieval systems, ranking functions, or configurations to try and ensure that as many relevant documents as possible are added to the pool. In Step 2, a filtering step is applied to select the documents that will be judged for each query. In Step 3, relevance assessments are carried out, and the pairs $\langle q, d \rangle$ are labeled. Notice that the test collections we used in this paper already come with human relevance assessments. Thus, only the highlighted portion in Figure 1a, (*i.e.,* relevance assessment using LLMs), was done in this work.

In order to answer $RQ1$, we compared the relevance judgments produced by the LLM with the gold-standard human labels that were collected by the organizers of the evaluation campaigns that produced the IR collections. We computed standard inter-annotator agreement metrics described in Section 3.5. To answer $RQ2$, we repeated the annotation process by the LLM multiple times and computed the intra-annotator agreement. Answering $RQ3$ required us to perform the full IR evaluation pipeline shown in Figure 1b. In Step 1, an IR system is used to retrieve documents in response to queries according to a ranking function. The resulting ranked list of $\langle q, d \rangle$ pairs is compared with the relevance assessments $\langle q, d, l \rangle$, and the standard retrieval evaluation metrics are computed in Step 2.

3.1. IR Collections

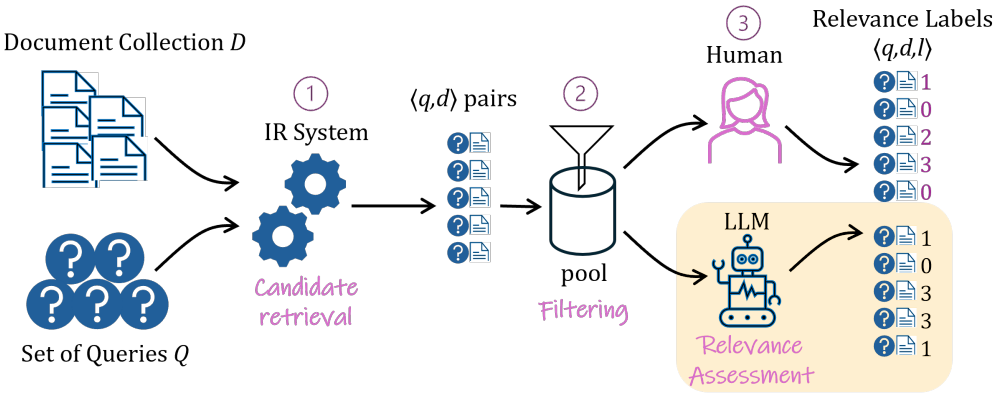
Our investigation was done using the following two IR test collections.

CHAVE [Santos and Rocha 2004] is a collection containing 103K news documents published in 1994 and 1995 at Folha de São Paulo. Queries and human-generated relevance judgments were produced within the scope of the CLEF evaluation campaigns[3]. A total of 16K relevance judgments were made for the 100 query topics. Relevance labels (L) are binary where 1 means "*relevant*" and 0 means "*not relevant*". We used the three fields available as the query: title, description, and narrative.

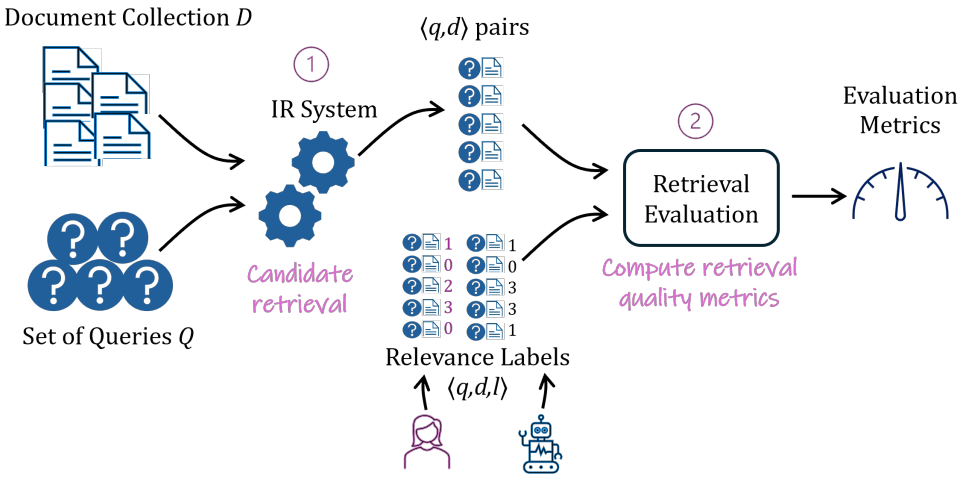
LLM4Eval[4] is a test collection created within the scope of the 2024 SIGIR workshop. It has 9K documents and 25 queries in the evaluation set. There are 5.6K gold standard relevance judgments labeled in four levels as follows: 3 for "*Perfectly relevant*", 2 for "*Highly relevant*", 1 for "*Related*, and 0 for "*Irrelevant*". The original data is in

[3] `https://www.clef-initiative.eu/`

[4] `https://llm4eval.github.io/`



(a) Annotation Phase



(b) Retrieval Evaluation Phase ($RQ3$)

Figure 1. Experimental Pipeline

English, so we translated it into Portuguese using DeepTranslator[5], and refer to it as LLM4Eval-PT. A manual inspection of the output revealed the translation quality was good in general, but some more challenging expressions were not adequately treated.

Because of time and budget restrictions, we were not able to have the complete set of $\langle q, d \rangle$ pairs assessed by the LLMs[6] As a result, sampling was needed. Two samples were taken from the complete set of relevance assessments in each test collection. First, we sampled 5k triples $\langle q, d, l \rangle$ from each collection. All triples in which the document was labeled as relevant were kept (*i.e.,* $l \geq 1$). The selection of the non-relevant triples aimed at prioritizing harder cases. So, we kept triples in which the non-relevant document was retrieved before the 30th position in the ranking. The samples with 5K documents were used to answer $RQ1$ and $RQ3$.

$RQ2$ requires executing the annotation by the LLMs five times, so it was done over smaller samples of 1K $\langle q, d \rangle$ pairs. The sampling procedure aimed at ensuring that

[5]https://github.com/nidhaloff/deep-translator

[6]Having a proprietary LLM such as GPT-4o annotate the 16K $\langle q, d \rangle$ pairs in the pool for the CHAVE collection five times, would cost approximately USD 800.

Table 1. Statistics of the IR collections used in the experiments

Collection	Documents	Queries	Relevance labels	5K $\langle q, d \rangle$ pairs	1K $\langle q, d \rangle$ pairs
CHAVE	103K	100	0	2841	520
			1	2159	480
LLM4Eval-PT	9K	25	0	2888	439
			1	1089	209
			2	483	165
			3	540	187

hard-to-judge pairs remained in the sample. Thus, we selected 1K $\langle q, d \rangle$ pairs by taking one-third from each of the following three groups.
1) *Hard negatives*: pairs that were highly ranked by the IR systems but not deemed relevant by the human evaluators.
2) *Hard positives*: pairs that were retrieved at the low ranks by the IR systems but were considered highly relevant by human evaluators.
3) *Random*: comprises a random selection of pairs from the initial set of 5K that were not in the previous groups. Statistics of the data used in our experiments are in Table 1.

3.2. Large Language Models

The following LLMs were used in our study. **GPT-4o:** Currently, the newest member in the GPT family, with a similar performance of GPT-4 Turbo in English texts but with an enhancement for non-English texts.

Sabiá-2: A LLM trained in Portuguese. The model matches GPT-4 and outperforms GPT-3.5 on several exams [Almeida et al. 2024]. In this work, Sabiá-2 Medium was used.

3.3. Prompt

The prompt used to instruct the LLMs to annotate relevance was based on previous work such as [Bueno et al. 2024] and also on the baseline prompt suggested in the LLM4Eval challenge (originally in English). We experimented with different descriptions of how the LLM should evaluate each pair (reasoning using Chain of Thought). The prompt included one example of each class (few-shot prompting). The examples were taken from online news articles.

We experimented with several prompts over a small set to find a prompt that best suits each model and IR collection. We also ensured the models could adequately explain why a certain label was chosen. There were variations depending on the relevance labels (binary or graded). Due to space restrictions, we cannot display the complete prompts here, but they are available online[7] . Figure 2 shows the structure of the prompt used for the CHAVE collection.

3.4. Ranking Functions

Four ranking functions were used to retrieve documents in response to queries. These ranking functions were tested with different sets of relevance judgments to answer $RQ3$ in Section 4.3.

[7] `https://github.com/lbencke/LLM-eval`

Dada uma consulta e uma passagem de texto, você deve fornecer uma pontuação 0 ou 1 com os seguintes significados:
1 = Relevante: a passagem é relevante à consulta, pois responde parcial ou totalmente à consulta.
0 = Irrelevante: a passagem não tem nada a ver com a consulta.
Procedimento: leia a consulta. Depois leia a passagem e verifique se dentro dela existem trechos que podem responder à consulta. Se existir alguma resposta à consulta, mesmo que parcial, atribua pontuação 1. Se a passagem não tiver nenhuma relação com a consulta, atribua a pontuação 0. Você deve primeiramente fornecer a explicação do porquê você atribuiu a referida pontuação à passagem e depois adicionar a pontuação atribuída.

Consulta: {example of a query (q)}
Passagem: {example of a document (d)}
Explicação: {example of the explanation for the relevance label}
Pontuação: {expected relevance label (l) for the pair $\langle q, d \rangle$}

Consulta: {query (q)}
Passagem: {document (d)}

Figure 2. Structure of the instruction submitted to the LLMs for CHAVE.

1. *BM25* [Spärck Jones et al. 2000]: This traditional IR model estimates the relevance of documents to a query by considering the frequency of query terms within each document. It incorporates term frequency and document length normalization to mitigate the influence of document size on relevance scoring.
2. *Dense Vector Retrieval (mE5)* [Wang et al. 2024]: This approach utilizes an encoder to transform queries and documents into dense vectors. Retrieval is performed by identifying the documents whose vectors are most similar to the query vector. We use the inner product as a similarity function.
3. *Re-ranking with MonoPTT5* [Piau et al. 2024]: We employ the MonoPTT5 model to re-rank the initial retrieval results obtained from both BM25 and mE5 systems. This re-ranking is based on the scores assigned by MonoPTT5, which predicts the likelihood of documents being relevant to the query.

3.5. Evaluation Metrics

To evaluate the consistency of judgments between LLMs and human raters, as well as the stability of these judgments, we employ two inter-rater agreement metrics. The first metric, *Cohen's Kappa* (κ_c), quantifies the agreement between two raters, adjusting for the level of agreement that could be expected purely by chance. The second metric, *Fleiss' Kappa* (κ_f), on the other hand, is an extension that allows for the assessment of agreement among three or more raters. This metric is useful when multiple raters independently assess the same collection of items. Both k_c and k_f vary from -1 (complete disagreement) to 1 (complete agreement).

IR system performance was evaluated using the nDCG (Normalized Discounted Cumulative Gain). This metric accounts for the position of relevant items, giving higher weight to items appearing earlier in the search results. The "discounted cumulative gain" (DCG) is calculated by summing the relevances of the results, penalizing later results.

This value is normalized by the "ideal" DCG (IDCG), which is the DCG for a perfect ranking of the results. IDGC is computed similarly but uses the ideal ranking in which results are sorted in decreasing order of relevance. When we consider nDCG values up to rank k, we call this metric nDCG@k.

3.6. Implementation Details

We used OpenAI (for GPT-4o) and Maritaca (for Sabiá-2-medium) APIs to prompt the LLMs. The main parameter that we set was the temperature, which is responsible for controlling the randomness of predictions generated by the model when using sampling-based decoding strategies. It plays a critical role in determining how conservative or creative the model is when generating text. Lower values make the model more deterministic. We set temperatures to very low values since we want the models to be stable (0 in GPT-4o and 0.02 in Sabiá). We kept the top-p parameter default by each model, following the instructions on the API. For both models, we set the max_tokens parameter to 500.

BM25 and mE5 were implemented using the Pyserini[8] toolkit. In our preliminary experiments, we found that removing stopwords and applying stemming improved the results of BM25. For dense retrieval, we used the 1024-dimensional mE5-large version embeddings available at the HugginFace Hub[9]. Since our collections are not too large, we conducted an exhaustive search to find documents that are related to queries. Finally, re-ranking was implemented with the rerankers[10] python API.

4. Results

In this section, we show the experimental results that answer our research questions.

4.1. *RQ1* – How correlated are LLM-generated and human-generated relevance judgments?

The level of agreement between LLMs and human judges measured by Cohen's Kappa (κ_c) is shown in Table 2. GPT annotations highly correlate to human annotations in CHAVE (k_c=0.729). In the LLM4Eval-PT dataset, GPT-4o achieved K_c results 35% higher than Sabiá. Also, Sabiá could not generate the relevance labels for 40 instances in CHAVE and five in LLM4Eval-PT.

Table 2. Agreement (k_c) with gold labels for the 5k sample

LLM	CHAVE	LLM4Eval-PT
GPT-4o	**0.729**	**0.282**
Sabiá-2-Medium	0.549	0.210

Analyzing Table 3, in CHAVE, GPT-4o made slightly more errors judging relevant documents as not relevant than the opposite. On the other hand, Sabiá tends to make more errors judging irrelevant documents as relevant. Sabiá yielded superior performance in the *"relevant"* class, predicting the right label in 94% of the cases, beating GPT-4o by 11

[8]https://github.com/castorini/pyserini
[9]https://huggingface.co/intfloat/multilingual-e5-large
[10]https://github.com/AnswerDotAI/rerankers

Table 3. Confusion Matrices

CHAVE – GPT-4o

		Predicted	
		0	1
Actual	0	0.89	0.11
	1	0.17	0.83

CHAVE – Sabiá

		Predicted	
		0	1
Actual	0	0.64	0.36
	1	0.06	0.94

LLM4Eval-PT - GPT-4o

		Predicted			
		0	1	2	3
Actual	0	0.65	0.22	0.08	0.05
	1	0.32	0.39	0.15	0.14
	2	0.07	0.33	0.29	0.30
	3	0.04	0.17	0.36	0.43

LLM4Eval-PT – Sabiá

		Predicted			
		0	1	2	3
Actual	0	0.53	0.09	0.21	0.17
	1	0.26	0.14	0.35	0.25
	2	0.12	0.12	0.40	0.36
	3	0.04	0.07	0.27	0.62

percentage points. Despite this, the overall Kappa correlation of Sabiá with gold labels is 18% lower than GPT-4o's, as shown in Table 2.

In LLM4Eval-PT, we see a similar tendency, with GPT-4o being better in the *"Irrelevant"* class (0) and Sabiá in the *"Perfectly Relevant"* class (3). GPT-4o had more difficulty in the *"Highly Relevant"* class (2) – it predicts most instances as *"Related"* or *"Perfectly Relevant"*. The latter also justifies what occurs for the *"Perfectly Relevant"* class, where many instances were classified as *"Highly Relevant"*, showing GPT-4o has difficulty distinguishing boundaries between levels two and three of relevance. Sabiá yielded poor results in the *"Related"* class (1). We manually checked instances of the *"Related"* cases that were predicted as *"Perfectly Relevant"*. They can be associated with short queries that allow different interpretations. For example, in the query "dog age by teeth," some passages confirm that one can calculate age by observing dogs' teeth, but it does not describe how. These cases are predicted by Sabiá as *"Perfectly Relevat"*, which may not be the case if the question was clearer.

To set a baseline for what level of agreement could be expected between human assessors in LLM4Eval-PT, we took a random sample of 271 $\langle q, d \rangle$. We asked two human annotators (H1 and H2) to perform the relevance assessments. The sample contained pairs from all queries and all levels of relevance. The results are presented in Table 4. We see that GPT-4o agrees more with H1 and H2 than humans agree with each other.

The levels of agreement between LLMs and humans we found in our experiments are within the range of similar work. [Faggioli et al. 2023] reported a κ_c of 0.38 using topics in binary labeling with GPT-3.5, while [Thomas et al. 2023] achieved a κ_c of 0.64 with a three-class scheme based on topic descriptions and narratives. For LLM4Eval-PT, which features queries posed as questions and four relevance classes, the results for

Table 4. Agreement (κ_c) on the sample with 271 pairs reassessed by humans

Gold vs GPT-4o	H1 vs H2	Gold vs H1	Gold vs H2	H1 vs GPT-4o	H2 vs GPT-4o
0.282	0.253	0.195	0.189	0.264	0.331

GPT-4o and Sabiá are comparable to the κ_c of 0.31 from Quati [Bueno et al. 2024] using GPT-4, and the κ_c of 0.26 reported by [de Jesus and Nunes 2024] using LLaMA3-70b. However, our numbers for GPT-4o and Sabiá do not reach the κ_c of 0.49 achieved in the TREC-DL 2021 annotations by [Faggioli et al. 2023].

4.2. *RQ2* – How stable are LLM-generated relevance judgments?

The stability of the LLMs in performing relevance judgments was tested by repeating the judgment rounds five times in the samples with 1K documents. Each independent evaluation round is treated as a different rater. Then, we calculate the Fleiss' kappa (k_f) among the five raters. Table 5 shows the results. The LLMs present high stability in their judgments for both collections. GPT-4o presents a slightly worse agreement for the hard cases in both collections. Sabiá shows the reverse behavior.

The intra-annotator agreement for the LLMs is much higher than reported by [Blanco et al. 2011] for crowd workers. Their κ_f range from 0.36 to 0.47. A consensus (*i.e.,* all five rounds resulting in the same relevance label) was achieved in 94% and 99% of the tuples in the CHAVE collection for GPT-4o and Sabiá, respectively. These results are in the same range as the ones obtained in an early study with humans [Resnick and Savage 1964]. In the LLM4Eval-PT collection, a consensus occurred 78% and 97% of the time for GPT-4o and Sabiá, respectively. The lower rates in LLM4Eval-PT are expected since it has four relevance labels.

Table 5. Intra-Annotator Agreement (Fleiss' kappa k_f) among LLM-based relevance assessments (1k sample)

Group	CHAVE		LLM4Eval-PT	
	GTP-4o	Sabiá	GTP-4o	Sabiá
Hard negatives	0.894	0.993	0.839	0.982
Hard positives	0.905	0.983	0.821	0.987
Random	0.948	0.990	0.887	0.978
All	0.935	0.992	0.858	0.983

4.3. *RQ3* – Can LLM-generated relevance assessments produce the same retrieval results as those produced when human judgments are used?

Table 6 shows the results of the retrieval quality metric (nDCG) for the different ranking functions. When we look at the absolute nDGC values calculated for the different sets of relevance judgments, we find variations of up to 21 percentage points (for BM25+PTT5 on LLM4Eval-PT). However, when we look at the order of the ranking functions produced by human and LLM-based judgments, we find a perfect correlation.

When assessed with LLM-based relevance labels, the IR systems' performances tend to present a higher numerical value. This behavior is more salient when there is less agreement with human judgments, which is the case of Sabiá. A possible explanation for this behavior is that the less nuanced judgments, especially in hard $\langle q, d \rangle$ pairs, are more correlated with the ranking functions, which make decisions with weaker information, such as term frequencies (BM25) or a smaller latent semantic space (PTT5 and mE5).

Despite variations in relevance labels, the observed stability in the order of the ranking functions confirms findings reported by [Voorhees 2000]. This similarity suggests

Table 6. nDCG@30 for human and LLM-generated judgments (5K sample). The value in parenthesis represent the order of the ranking function

Ranking function	CHAVE			LLM4Eval-PT		
	GTP4o	Sabiá	Human	GTP4o	Sabiá	Human
BM25+PPT5	0.478 (1)	0.486 (1)	0.446 (1)	0.679 (1)	0.770 (1)	0.556 (1)
mE5+PTT5	0.419 (2)	0.456 (2)	0.405 (2)	0.663 (2)	0.744 (2)	0.545 (2)
mE5	0.193 (4)	0.208 (4)	0.178 (4)	0.596 (3)	0.692 (3)	0.482 (3)
BM25	0.413 (3)	0.409 (3)	0.368 (3)	0.421 (4)	0.550 (4)	0.342 (4)

that the averaging of performance metrics across queries may effectively mitigate the impact of small differences when assessed over a substantial number of queries.

5. Conclusion

This study has demonstrated that LLMs can serve as viable alternatives to human annotators in the generation of relevance judgments for IR test collections. With our experimental results, we find these answers to our research questions:

RQ1. Our findings show that LLM labels are positively correlated with human labels in both datasets. The results were comparable, and in some cases even higher, than those reported in related work. The prompting strategy plays a crucial role, and seeking for the best prompt has to be done with a representative small set. Also, the most suitable prompt can change depending on the dataset and the LLM used.

RQ2. We also showed that GPT-4o and Sabiá are very stable. The correlation among different runs shows a small variation in the labels. Both models achieved high correlations across five independent assessments of the same set at different times and in different orders. Results were better than humans when compared to existing work [Blanco et al. 2011]. GPT-4o presented more variation, especially in LLM4Eval-PT.

RQ3. Experimenting with four different IR ranking functions and evaluating both datasets with relevance judgments made by humans and generated by the LLMs, we found that they maintain the same order in both datasets. This confirms that LLM-based relevance labels can be effectively used to assess IR systems performance.

Limitations. The LLMs used in this work are closed, and we do not have access to the details of their training. The documents in CHAVE may be in the training data of the LLMs, but it is unlikely that the relevance labels (which are not public) were seen by the models. LLM4Eval was released only a couple of months ago after the models had been trained. So, its contents may not be included in the training data of the LLMs. This paper focused on closed LLMs, but we also plan to use and/or specialize open models to perform relevance labeling. Additionally, the use of translated datasets is not ideal – in LLM4Eval-PT, cultural topics, local entities, and idiomatic expressions are poorly translated. Developing culturally representative datasets is essential, and we plan future works that contribute to this. Finally, because we only experimented with generic documents, we cannot ensure that these results hold for domain-specific collections, which tend to require a more in-depth knowledge for relevance labeling.

Acknowledgments. This work has been partially funded by CENPES Petrobras, CNPq-

Brazil, and Capes Finance Code 001.

References

Almeida, T. S., Abonizio, H., and Nogueira, R. (2024). Sabiá-2: A New Generation of Portuguese Large Language Models. *arXiv preprint arXiv:2403.09887*.

Blanco, R., Halpin, H., Herzig, D. M., Mika, P., Pound, J., Thompson, H. S., and Tran Duc, T. (2011). Repeatable and reliable search system evaluation using crowdsourcing. In *ACM SIGIR conference on Research and development in Information Retrieval*, pages 923–932.

Bueno, M., de Oliveira, E. S., Nogueira, R., Lotufo, R. A., and Pereira, J. A. (2024). Quati: A brazilian portuguese information retrieval dataset from native speakers. *arXiv preprint arXiv:2404.06976*.

Cleverdon, C. W. (1960). The aslib cranfield research project on the comparative efficiency of indexing systems. In *Aslib Proceedings*, volume 12, pages 421–431.

de Jesus, G. and Nunes, S. (2024). Exploring large language models for relevance judgments in tetun. *arXiv preprint arXiv:2406.07299*.

Faggioli, G., Dietz, L., Clarke, C. L., Demartini, G., Hagen, M., Hauff, C., Kando, N., Kanoulas, E., Potthast, M., Stein, B., et al. (2023). Perspectives on large language models for relevance judgment. In *ACM SIGIR International Conference on Theory of Information Retrieval*, pages 39–50.

Lima de Oliveira, L., Romeu, R. K., and Moreira, V. P. (2021). REGIS: A Test Collection for Geoscientific Documents in Portuguese. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2363–2368.

Piau, M., Lotufo, R., and Nogueira, R. (2024). ptt5-v2: A closer look at continued pretraining of T5 models for the portuguese language. *arXiv preprint arXiv:2406.10806*.

Rahmani, H. A., Craswell, N., Yilmaz, E., Mitra, B., and Campos, D. (2024). Synthetic test collections for retrieval evaluation. *arXiv preprint arXiv:2405.07767*.

Resnick, A. and Savage, T. R. (1964). The consistency of human judgments of relevance. *American Documentation*, 15(2):93–95.

Santos, D. and Rocha, P. (2004). The key to the first CLEF with Portuguese: Topics, questions and answers in CHAVE. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 821–832. Springer.

Soviero, B., Kuhn, D., Salle, A., and Moreira, V. P. (2024). ChatGPT goes shopping: LLMs can predict relevance in ecommerce search. In *European Conference on Information Retrieval*, pages 3–11.

Spärck Jones, K. and van Rijsbergen, C. J. (1975). Report on the need for and provision of an "ideal" information retrieval test collection. *Computer Laboratory, University of Cambridge*.

Spärck Jones, K., Walker, S., and Robertson, S. E. (2000). A probabilistic model of information retrieval: development and comparative experiments. *Information processing & management*, 36(6):809–840.

Theodosiou, Z., Georgiou, O., and Tsapatsoulis, N. (2011). Evaluating annotators consistency with the aid of an innovative database schema. In *International Workshop on Semantic Media Adaptation and Personalization*, pages 74–78.

Thomas, P., Spielman, S., Craswell, N., and Mitra, B. (2023). Large language models can accurately predict searcher preferences. *arXiv preprint arXiv:2309.10621*.

Voorhees, E. M. (2000). Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5):697–716.

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., and Wei, F. (2024). Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. (2024). Judging LLM-as-a-judge with MT-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Zhu, E., Sheng, Q., Yang, H., Liu, Y., Cai, T., and Li, J. (2023). A unified framework of medical information annotation and extraction for chinese clinical text. *Artificial Intelligence in Medicine*, 142:102573.