ChatGPT Goes Shopping: LLMs Can Predict Relevance in eCommerce Search

Beatriz Soviero$^{1*[0009-0009-2940-0714]}$, Daniel Kuhn$^{2*[0009-0000-9182-7916]}$,
Alexandre Salle$^{3[0000-0003-3518-4121]}$, and Viviane P. Moreira$^{1[0000-0003-4400-054X]}$

[1] Institute of Informatics, UFRGS, Porto Alegre, Brazil
{bfsoviero,viviane}@inf.ufrgs.br
[2] Institute of Education, Science and Technology of Rio Grande do Sul (IFRS), Ibirubá, Brazil
daniel.kuhn@ibiruba.ifrs.edu.br
[3] VTEX, Porto Alegre, Brazil
alexandre.salle@vtex.com

Abstract. The dependence on human relevance judgments limits the development of information retrieval test collections that are vital for evaluating these systems. Since their launch, large language models (LLMs) have been applied to automate several human tasks. Recently, LLMs started being used to provide relevance judgments for document search. In this work, our goal is to assess whether LLMs can replace human annotators in a different setting – product search in eCommerce. We conducted experiments on open and proprietary industrial datasets to measure LLM's ability to predict relevance judgments. Our results found that LLM-generated relevance assessments present a strong agreement (∼82%) with human annotations indicating that LLMs have an innate ability to perform relevance judgments in an eCommerce setting. Then, we went further and tested whether LLMs can generate annotation guidelines. Our results found that relevance assessments obtained with LLM-generated guidelines are as accurate as the ones obtained from human instructions.[†‡]

Keywords: relevance judgment prediction · LLM · eCommerce

1 Introduction

Test collections consisting of documents, queries, and relevance judgments are a crucial asset for the development of Information Retrieval (IR) tools and techniques. Obtaining human-generated relevance judgments has been the main bottleneck in the creation of IR test collections. Having humans evaluate relevance is costly in terms of time and money. Sanderson [14] reports that the 73K judgments for each year of the TREC ad-hoc tracks took over 600 hours (considering a rate of two judgments per minute). Oliveira *et al.* [12] mention a much lower rate of about 20 query-document pairs per hour, which meant that

[*] Work conducted during an internship at VTEX.

[†] The source code for this work is available at https://github.com/danimtk/chatGPT-goes-shopping

[‡] The final authenticated version is available online at https://link.springer.com/chapter/10.1007/978-3-031-56066-8_1

230 hours of human volunteers were necessary to make relevance assessments for their small test collection.

In an eCommerce scenario, test collections are scarce as most datasets are proprietary. Although judging relevance for query-product pairs can be faster than for the traditional query-document scenario, the creators of the WANDS dataset [4] reported that human assessors had a throughput of around 190 to 200 product-query pairs per hour. Considering the dataset has 233K query-product pairs that were assessed by three judges, we can estimate that over 3.5K hours were spent to generate the relevance assessments.

The advent of Large Language Models (LLMs) has enabled the automation of a set of tasks that previously required direct human effort. This could be the case with the relevance judgment task – recent work has demonstrated that LLMs are promising for judging relevance in classical text collections, such as TREC [8,18].

In this work, our goal is to assess whether LLMs can replace human relevance assessments in eCommerce test collections. Search in eCommerce differs from web search in general since the documents are typically very short, consisting of a product catalog. Queries are also short, emphasizing the use of keywords over long phrases in natural language. Attributes such as brands, measurements, and dosage are commonly used to describe the desired product [16]. We designed a set of experiments using GPT-3.5-turbo and GPT-4 performed on two datasets of query-product pairs: WANDS [4] and a dataset comprised of proprietary data sourced from the production environment of a large eCommerce technology provider. The results showed that LLM-generated judgments have an average overlap of around 82% with human judgments. This number is very high considering that human judges have shown much lower levels of agreement – *e.g.,* between 42 and 49% on TREC collections [19].

Our second contribution relates to the creation of annotation guidelines – a task that can be quite laborious. As pointed out by Faggioli *et al.* [8], Google search guidelines (geared towards human assessors) span over 170 pages. In addition, domain knowledge is important since the instructions may vary depending on the type of search. With that in mind and taking advantage of LLMs good summarization skills [13], we go a step further and prompt the LLM to generate the annotation guidelines. The generated guidelines are fed back into the LLM along with the query-product pairs to be annotated. Our results showed that the annotations obtained with LLM-generated guidelines are as accurate as the ones obtained from the human-generated guidelines.

2 Background and Related Work

Evaluation is paramount in IR and has been a constant focus throughout the years. The standard evaluation paradigm requires a set of documents, query topics, and human relevance assessments. Since the early days of the Cranfield experiments [5] in which relevance judgments were exhaustive (*i.e.,* for all query-document pairs) and made by human experts, the research community has constantly been trying to find means to create test collections in a more scalable way. The first step in this direction was the development of the pooling method [17] in which only a small subset of the documents are judged for each query topic. Then, several other strategies were devised, including choosing a small set of documents to judge [3] and relying on crowd workers [1], who are less expensive than experts.

Since LLMs were launched, the research community has been testing their abilities to automate a series of language tasks, including question answering, summarization, translation, and reading comprehension [2]. Recently, the use of LLMs for relevance prediction started being explored. Faggioli *et al.* [8] discuss the pros and cons of using LLMs for automatic relevance judgments. In an experimental evaluation, they found an agreement of $\kappa = .38$ between GPT-3.5-turbo ad human assessors on TREC-8 ad hoc [9]. The non-relevant documents were correctly predicted 90% of the time but, on the relevant documents, only half of the predictions were accurate. Thomas *et al.* [18] also applied LLMs (*i.e.,* GPT-3.5) for automatic relevance judgments. Using data from TREC Robust [20], they found an agreement of $\kappa = .64$ in their best run. But they also point out the sensitivity of results to prompt variations.

3 Method

This section describes the method we adopted to answer the following research questions:
RQ1 Can an LLM effectively perform relevance judgments for eCommerce search?
RQ2 Can an LLM effectively create a set of guidelines to instruct itself on making relevance judgments?

3.1 Datasets

Our experiments used two datasets from different domains and languages. In all cases, we used binary relevance judgments (*i.e.,* "Relevant" and "Not relevant"). Statistics of the datasets are in Table 1. The instances include a mix of easier and harder cases to test the capabilities of the LLM in different situations.

WANDS [4] is a publicly available eCommerce dataset consisting of 480 queries, 49K products, and 233K human relevance labels considering three classes ("relevant", "partially relevant", and "irrelevant"). The products in WANDS are household goods (furniture and decoration) described in English. Our experiments were performed on a random sample of query-product pairs covering 409 out of the 480 queries (85%). Our sample was equally balanced between the two classes, and half our relevant instances were mapped from the *partially relevant* instances in WANDS – these are our hard positives.

Pharma is a dataset based on production data from a large eCommerce solution provider consisting of 28K unique queries and 20K products. The data is in Portuguese and comes from logs with result-sets for user queries on an online pharmacy. The approach adopted for assigning relevance labels to query-product pairs relied on user clicks as an implicit relevance signal. While clicks are less reliable than explicit relevance judgments, they have been extensively used as a proxy for relevance. In a user study, Joachims *et al.* [10] found a reasonable level of agreement between clicks and explicit feedback for document relevance. Products in the catalog were ranked for each query in decreasing order of the number of clicks the product had received when retrieved for the query. Then, the result-set was divided into three bins where the first has the products with the most clicks, and the third bin has the fewest. The relevant products for each query were taken from the first bin. We considered as *hard positives* the products that were relevant to the query, and yet there were no words in common between the product name and the query. This

Table 1: Statistics of Query-Product pairs annotated by the LLMs.

Dataset	Relevant		Not Relevant		Total
	Easy	Hard	Easy	Hard	
WANDS	700	700	1400	—	2800
Pharma	1260	140	700	700	2800

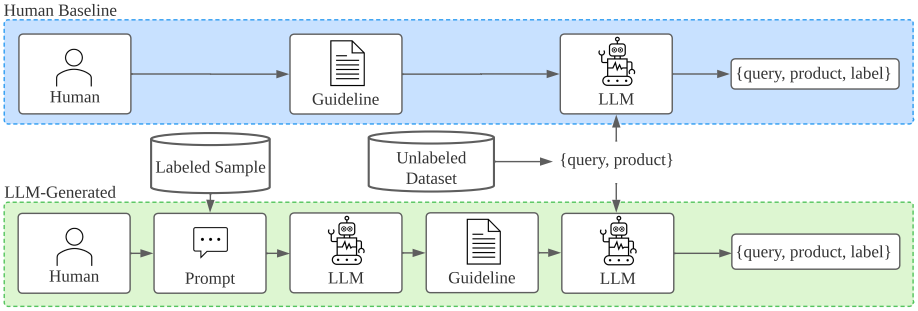

Fig. 1: Strategies for guideline generation and relevance annotation

sample was manually checked. The *easy positives* were selected only considering the first bin. The *hard negatives* are instances that share terms between query and product name yet do not satisfy the user's intent. In contrast, *easy negative* instances share no common terms. Our experiments were performed with 2612 queries.

Because WANDS is a public dataset, we cannot attest that no data leakage occurred, as its contents may have been seen during the training of LLMs such as GPT. The Pharma dataset, on the other hand, is private. Thus, we can be sure that GPT models had no access to the queries and click-rates.

3.2 Prompting Strategies

In our prompting strategies, we varied how the guidelines were created and the number of examples provided. The process used to obtain the relevance assessments from the LLM is depicted in Figure 1, and the details are as follows.

Annotation Guidelines.

Human Baseline – Initially, we created a baseline prompt with basic instructions asking the LLM to act as an expert for a relevance judgment task in the eCommerce context. This process is shaded in blue in Figure 1. Basic instructions were provided with the levels of relevance ("Relevant" and "Not relevant"), the format in which the data would be fed, and brief definitions on when to assign the relevant and not relevant labels. In addition, we provided a query-product pair to be used as an example for each label. The guidelines also informed that the goal was to match the user's intent rather than focusing on the exact words in the query – the idea was to encourage the LLM not to act as a mere keyword-matching judge. Finally, we asked the model to judge a query-product pair based on the instructions and requested the response to be returned in the format *(query, product, label)*. Throughout this work, *product* is represented by the product's name in the catalog.

LLM-generated – We created a prompt asking the LLM to create a guideline for the relevance judgment task. Along with these instructions, we provided a set of 200 annotated query-product pairs for the model to use as a source for generating the guidelines. This process is shaded in green in Figure 1. The goal was to enable the model to extract relevant patterns and, based on them, compose (hopefully) richer guidelines. There is no intersection between these query-product pairs and those used for annotation.

Examples. In addition to the guidelines, we also aim to assess whether adding annotation examples to the prompt can improve results. Thus, two settings were used: *+Zero-shot* (no further examples are given to the LLM), and *+Ten-shot* (ten annotated examples in the form *(query, product, label)* are given to the LLM.

Without guideline. In order to evaluate the contribution guidelines, we created a prompt containing just the ten examples (as in *+Ten-shot*) and the request to judge a given tuple as relevant or not relevant.

4 Experimental Evaluation

4.1 Experimental Setup and Reproducibility

GPT models were accessed through the OpenAI API via the Chat Completion API endpoint. Each message sent to the LLM was composed of two objects: (i) the guidelines informed in *system* role and (ii) the query-product pair to be judged for relevance in the *user* role. The parameter settings were all default except for the *temperature*, which was set to zero.

The +zero-shot prompts had an average of 407 tokens, while the prompts in the +ten-shot scenario averaged 667 tokens. Completions averaged 26 tokens.

4.2 Evaluation metrics

The results were evaluated according to two metrics of the agreement between LLM assessments and the human labels (*i.e.,* the ground truth): accuracy and Cohen's κ. While accuracy measures the overlap percentage of agreement, Cohen's κ also takes into account the possibility of the agreement occurring by chance. To verify statistical significance, we ran Friedman tests considering α=.05.

4.3 Results

To answer $RQ1$ (*Can an LLM effectively perform relevance judgments for eCommerce search?*), we evaluate the model's ability to judge relevance using human-generated prompts. The results in Table 2 are promising, with accuracy up to 85%. The agreement with the ground truth relevance labels was as high as κ= .7.

When we look at the accuracy at different levels of difficulty for the query, as expected, the scores are higher in the easy instances (∼90% in both datasets) and lower in the harder instances (∼52%). The worst results were for the hard positives in WANDS. We attribute the errors of judgment to the mismatch among the adjectives present in the query and product, *e.g.,* the pairs *('rug plum', 'ophir faux-fur pink area rug')*, *('card table', 'rian coffee table')* and *('bathroom lighting', 'chante 1 -bulb outdoor bulkhead*

Table 2: Agreement with the ground truth labels – best scores in bold.

Dataset	Prompt	Examples	GPT-3.5-turbo		GPT-4	
			Accuracy	Kappa	Accuracy	Kappa
WANDS	human baseline	+ten-shot	.801	.602	**.829**	**.658**
	human baseline	+zero-shot	.804	.609	.797	.595
	LLM-generated	+ten-shot	.819	.638	.794	.587
	LLM-generated	+zero-shot	.790	.580	.751	.501
	without guideline	+ten-shot	.785	.570	.799	.598
Pharma	human baseline	+ten-shot	.834	.669	.841	.682
	human baseline	+zero-shot	.806	.613	.838	.676
	LLM-generated	+ten-shot	.837	.674	**.851**	**.701**
	LLM-generated	+zero-shot	.797	.594	.849	.697
	without guideline	+ten-shot	.838	.676	.847	.694

light') were labeled as relevant by human judges but not by the LLMs. Misjudged easy positives happened when the query was very generic and the product name was much more specific, *e.g., ('dinosaur', 'dinosaur ii holiday shaped ornament')* and *('flamingo', 'palm sprints flamingo graphic art')*.

+zero-shot *vs* +ten-shot – As expected, the scores were higher when annotated examples were provided (in 7 out of 8 possible comparisons). The differences were larger in Pharma, but the statistical test did not find them to be significant.

GPT-3.5-turbo *vs* GPT-4 – The results show that GPT-4 achieved the best scores in both datasets. However, if we compare the individual configurations, we see that GPT-3.5-turbo wins in some cases (*e.g.,* in the LLM-generated prompts in WANDS. Yet, the only statistically significant difference was in the +zero-shot scenario in Pharma. Taking into consideration that GPT-3.5-turbo costs 20 times less than GPT-4, it may be the preferable model in many situations.

Human baseline *vs* LLM-generated guidelines – Agreement scores for the relevance judgments made in response to LLM-generated guidelines were comparable to the scores obtained with human prompts, sometimes even slightly better. This was the case of the winning configuration on the Pharma. Statistical tests did not find significant differences in the accuracy scores (p-values were always $\geq$ 0.05 when comparing human baseline and LLM-generated runs). We conclude that the answer to *RQ2* (*Can an LLM effectively create a set of guidelines to instruct itself on making relevance judgments?)* is yes.

Examples *vs* guidelines – The *+zero-shot* prompting strategies consist solely of the guidelines, whereas the prompts *without guidelines* are basically composed of annotation examples. By comparing the results of these two configurations we can compare the importance of providing guidelines versus examples to the LLM. In most cases, agreement scores in experiments conducted with examples only (without guideline) were similar to or higher than those where only the guideline was provided (*+zero-shot*), except for WANDS using GPT-3.5-turbo. However, both cases that achieved the best scores use guidelines (human-baseline or LLM-generated) and examples (*+ten-shot*) – this is more evident for GPT-4, which is better at following instructions.

Comparison with a Non-LLM Baseline – We establish a lower bound on the difficulty of our datasets by including a baseline based on BERT [7]; BERT-based models are strong baselines on sentence-pair tasks in GLUE [21]. We use the XML-RoBERTa-base variant [6], which employs the RoBERTa [11] pre-training scheme for BERT and is multilingual, which is a requirement since one of our datasets is in Portuguese. XLM-RoBERTa-base was fine-tuned for binary sentence-pair classification using the CLS token on the exact same 200 examples we use in our LLM-generated prompt strategy. For comparison, [22] report strong results in the paraphrase identification task from the FewCLUE dataset (a Chinese version of FewGLUE [15], which is a few-shot version of GLUE) fine-tuning RoBERTa using *only 32 examples*. We fine-tune for 20 epochs using AdamW, learning rate of 2e-5, and linear decay schedule with 10% warm-up steps. This resulted in an accuracy = .67 and κ= .35 for WANDS and accuracy = .78, κ = .57 for Pharma. These scores are significantly lower than the ones reported in Table 2 – the difference in terms of κ is as large as 30 p.p for WANDS and 13 p.p for Pharma.

Practical findings – In our early experiments with GPT-3.5-turbo, we found that two approaches (which we avoided for our final experiments) severely degraded κ to near-chance levels: (1) submitting multiple tuples for annotation *in the same prompt* instead of a single tuple (which by the nature of causal LMs conditions future annotations on previous ones), and (2) prompting only for the *label* in the model's annotated response instead of the tuple in the format *(product, name, label)*.

5 Conclusion

This work assessed the ability of LLMs to produce relevance judgments for product search. The results showed that LLMs can perform binary relevance judgments with a high overlap ($\sim$82%) with human judgments. These results come at a fraction of the time and cost taken by human judges – estimating 30 judgments per minute, LLMs are nine times faster than humans. We have also used the LLMs to generate annotation guidelines which yielded agreement scores that are not statistically different from the ones obtained with human-generated guidelines. Although more than 20 times cheaper than GPT-4, GPT-3.5-turbo achieved similar results. Our experiments also provided some practical findings in prompt engineering by highlighting the sensitivity of GPT to response formatting and the inability to deal with batch requests.

The focus of our work was to answer two research questions, and the experiments done here by no means exhaust this topic. For guideline generation, we relied on 200 annotated tuples and did not assess the impact that varying this number would have on the agreement scores.

Some limitations of our work are the usage of exclusively commercial LLMs and the lack of domains other than pharmacy and household goods. Nevertheless, we believe our findings can be useful for practitioners and contribute to the understanding of the potential of LLMs. In future work, we plan to experiment with open-source LLMs, use more datasets, and further investigate sensitivity to prompt variations.

Acknowledgments: The authors thank Shervin Malmasi for his helpful comments and suggestions. This work has been financed in part by VTEX BRASIL (EMBRAPII PCEE1911.0140), CAPES Finance Code 001, and CNPq/Brazil.

References

1. Blanco, R., Halpin, H., Herzig, D.M., Mika, P., Pound, J., Thompson, H.S., Tran Duc, T.: Repeatable and reliable search system evaluation using crowdsourcing. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. pp. 923–932 (2011)
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
3. Carterette, B., Allan, J., Sitaraman, R.: Minimal test collections for retrieval evaluation. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 268–275 (2006)
4. Chen, Y., Liu, S., Liu, Z., Sun, W., Baltrunas, L., Schroeder, B.: WANDS: Dataset for product search relevance assessment. In: European Conference on Information Retrieval. pp. 128–141. Springer (2022)
5. Cleverdon, C.W.: The aslib cranfield research project on the comparative efficiency of indexing systems. In: Aslib Proceedings. vol. 12, pp. 421–431. MCB UP Ltd (1960)
6. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.747, https://aclanthology.org/2020.acl-main.747
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423
8. Faggioli, G., Dietz, L., Clarke, C.L., Demartini, G., Hagen, M., Hauff, C., Kando, N., Kanoulas, E., Potthast, M., Stein, B., et al.: Perspectives on large language models for relevance judgment. In: Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval. pp. 39–50 (2023)
9. Harman, D., Voorhees, E.: Overview of the eighth text retrieval conference (trec-8). In: Proceedings of the Eight Text REtrieval Conference (TREC-8). pp. 1–19 (1999)
10. Joachims, T., Granka, L., Pan, B., Hembrooke, H., Gay, G.: Accurately interpreting clickthrough data as implicit feedback. In: Acm Sigir Forum. vol. 51, pp. 4–11. Acm New York, NY, USA (2017)
11. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Ro{bert}a: A robustly optimized {bert} pretraining approach (2020), https://openreview.net/forum?id=SyxS0T4tvS
12. Lima de Oliveira, L., Romeu, R.K., Moreira, V.P.: 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. p. 2363–2368 (2021)
13. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P.F., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 27730–27744 (2022)

14. Sanderson, M., et al.: Test collection based evaluation of information retrieval systems. Foundations and Trends® in Information Retrieval **4**(4), 247–375 (2010)
15. Schick, T., Schütze, H.: It's not just size that matters: Small language models are also few-shot learners. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 2339–2352. Association for Computational Linguistics, Online (Jun 2021). `https://doi.org/10.18653/v1/2021.naacl-main.185`, `https://aclanthology.org/2021.naacl-main.185`
16. Sondhi, P., Sharma, M., Kolari, P., Zhai, C.: A taxonomy of queries for e-commerce search. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. pp. 1245–1248 (2018)
17. Spark-Jones, K., van Rijsbergen, C.J.: Report on the need for and provision of an "ideal" information retrieval test collection. Computer Laboratory, University of Cambridge (1975)
18. Thomas, P., Spielman, S., Craswell, N., Mitra, B.: Large language models can accurately predict searcher preferences. arXiv preprint arXiv:2309.10621 (2023)
19. Voorhees, E.M.: Variations in relevance judgments and the measurement of retrieval effectiveness. Information Processing & Management **36**(5), 697–716 (2000)
20. Voorhees, E.M., et al.: Overview of the trec 2003 robust retrieval track. In: Proceedings of the Text REtrieval Conference. pp. 69–77 (2003)
21. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: GLUE: A multi-task benchmark and analysis platform for natural language understanding. In: International Conference on Learning Representations (2019), `https://openreview.net/forum?id=rJ4km2R5t7`
22. Xu, L., Lu, X., Yuan, C., Zhang, X., Xu, H., Yuan, H., Wei, G., Pan, X., Tian, X., Qin, L., Hai, H.: Fewclue: A chinese few-shot learning evaluation benchmark (2021)