Quati: A Brazilian Portuguese Information Retrieval Dataset from Native Speakers

Mirelle Bueno[*1] **, E. Seiti de Oliveira**[*1] **, Rodrigo Nogueira**[1,2] **, Roberto Lotufo**[1,3] **, Jayr Pereira**[4]

[1]Departamento de Engenharia de Computação e Automação (DCA)
Universidade Estadual de Campinas – UNICAMP
13083-872 – Campinas – São Paulo, Brasil

[2]Maritaca AI

[3]NeuralMind.ai

[4]Universidade Federal do Cariri
Juazeiro do Norte-CE, Brasil.

m174909@dac.unicamp.br, eduseiti@dca.fee.unicamp.br

{rfn,lotufo}@unicamp.br, jayr.pereira@ufca.edu.br

Abstract. *We present Quati,*[1] *a dataset specifically designed for evaluating Information Retrieval (IR) systems for the Brazilian Portuguese language. It comprises a collection of queries formulated by native speakers and a curated set of documents sourced from a selection of frequently accessed Brazilian Portuguese websites, which ensures a representative and relevant corpus. To label the query–document pairs, we use a state-of-the-art LLM, which shows inter-annotator agreement levels comparable to human performance in our assessments. Our annotation methodology is described, enabling the cost-effective creation of similar datasets for other languages, with an arbitrary number of labeled documents per query. As a baseline, we evaluate a diverse range of open-source and commercial retrievers. Quati is publicly available at* `https://huggingface.co/datasets/unicamp-dl/quati`*, and all scripts at* `https://github.com/unicamp-dl/quati`*.*

1. Introduction

The development of Information Retrieval (IR) systems depends on high-quality evaluation datasets, which should contain queries and documents ideally in the same target language of those systems, in order to capture specific information needs and social-cultural aspects. That, contrasts with translated datasets, which potentially represent the information needs and knowledge of a different culture or society. Hence, translated datasets may not effectively measure a retrieval system's ability in real-world scenarios involving native users.

[*]Equal contribution.

[1]We named our dataset after this South American mammal, whose foraging behavior represents the resolute search for resources.

Despite being one of the most widely spoken languages in the world, there is a scarcity of IR datasets in Portuguese. Existing datasets such as REGIS [Lima de Oliveira et al. 2021] and RCV1 [Lewis et al. 2004][2], though valuable, fall short due to their limited size and specialized domains (geoscience and news). While translated datasets such as mMARCO [Bonifacio et al. 2021] and mRobust04 [Jeronymo et al. 2022] have helped to alleviate this issue, the use of automatic translations often represents the loss of socio-cultural characteristics of the target languages, and the evaluations may become biased by the source language.

To address those issues, we created Quati, a Brazilian Portuguese evaluation dataset, comprising human-written queries and a high-quality native corpus. Quati is created using a semi-automated pipeline, aiming to reduce the labeling cost barrier. We use a Large Language Model (LLM) to judge a passage's relevance for a given query, publishing a cost-effective pipeline to create an IR evaluation dataset with an arbitrary number of annotated passages per query.[3] In this context, our work aims to answer the following research question: Can LLMs be used to compose a semi-automated pipeline for annotating query–passages relevance for Brazilian Portuguese IR systems?

To evaluate the quality of the LLM annotations, we compare them with human annotations on a sample of query–passage pairs and confirmed a Cohen's Kappa coefficient of 0.31. While this figure is below the 0.41 seen in human-human annotation agreement, it is consistent with the findings reported in the literature [Faggioli et al. 2023, Thomas et al. 2023, Farzi and Dietz 2024] and it will likely increase as LLMs improve in quality. The usage of a modular semi-automated pipeline, allows the dataset construction method to be replicated to create high-quality IR datasets for other languages.

2. Related Work

Evaluation datasets are an important variable in the IR context as they expose the limitations of search systems and guide their development. However, most of the available datasets are in English, as is the case with MS MARCO [Bajaj et al. 2016]. Works such MIRACL [Zhang et al. 2023], mMARCO [Bonifacio et al. 2021], mRobust [Jeronymo et al. 2022], Mr.Tydi [Clark et al. 2020], TREC CLIR [Schäuble and Sheridan 1998], CLEF [Peters and Braschler 2002], NTCIR [Sakai et al. 2021] and HC4 [Lawrie et al. 2022] are efforts to develop datasets for other languages, but most are based on language translation to adapt English to the target languages, or do not include Portuguese. Ongoing efforts [Lima de Oliveira et al. 2021, Vitório et al. 2024] are starting to change that scenario creating IR datasets for Brazilian Portuguese, but so far focusing on specific domains.

The creation of datasets for IR is a resource-intensive task, particularly in the process of judging the relevance of documents. Recent endeavors have witnessed a shift towards leveraging LLMs to assess query–passage relevance [Zendel et al. 2024]. Faggioli et al. [Faggioli et al. 2023] further underscored the potential of employing LLMs for automating the judgment of document relevance, thereby opening up promising avenues

[2]https://trec.nist.gov/data/reuters/reuters.html

[3]The total cost for this dataset was U$140.19 (0.03 per query–passage) for an average of 97.78 annotated passages per query.

for exploration in this domain. Complementary evaluations conducted by Thomas et al. [Thomas et al. 2023] demonstrated a significant correlation between human judgments and those made by the GPT-3.5-turbo model.

3. Methodology

We used a semi-automatic method to create Quati, as depicted in Figure 1. The required inputs are: 1) A large corpus, originally written in the target language, from which we extract the passages to compose our IR dataset; 2) A set of test queries, manually created to represent the information needs of native speakers. In the following sections, we detail the steps of the pipeline.

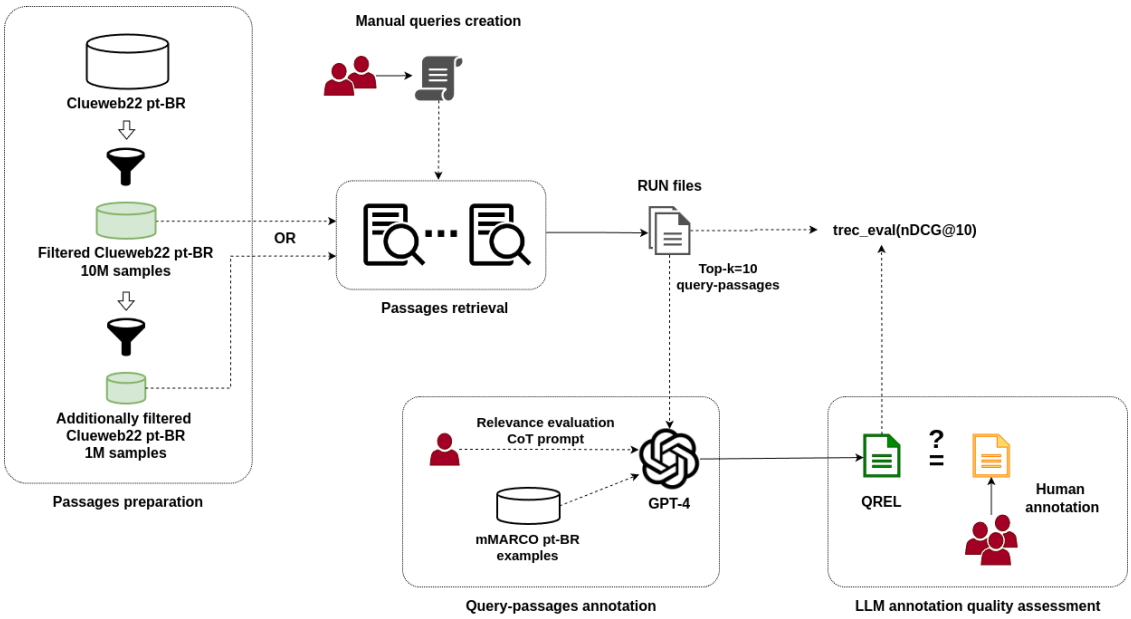

Figure 1. Proposed IR dataset creation methodology.

3.1. Passages preparation

The passages preparation step is composed by the following substeps:

Data collection: We used the Portuguese subset of ClueWeb22 [Overwijk et al. 2022] category B, which includes 4.1 million web pages more likely to be visited according to Bing search algorithms during the first half of 2022 [Overwijk et al. 2022].

URL filtering: We excluded any documents from our dataset whose URLs' domain ended with ".pt", which refer to Portuguese from Portugal, as the language style in those documents might differ significantly from that used in Brazilian Portuguese web pages. Additionally, we used FastText [Joulin et al. 2016b, Joulin et al. 2016a] as an additional language verification method to ensure that only Portuguese documents were included in our corpus.

Document segmentation into passages: Following language verification, we segmented the documents into approximately 1,000-character segments and assessed the percentage of line breaks (`\n`) occurrences within each segment, removing those with more than 20%.

This criterion was used to increase the probability of retaining segments predominantly composed of natural language text.

Separation into large and small versions: With the process described in the previous steps, we collected a total of 20 million segments. From this set, we randomly selected 10 million segments (hereinafter referred to as 10M corpus) to be the passages in our corpus, creating a large, but still manageable, dataset of more than 11 GB of size. A second dataset was built from the first one applying additional filtering rules taken from the MassiveWeb Corpus [Rae et al. 2021] — e.g. removing passages with more than 10% of symbols, or with mean word length outside the 3 to 10 interval — and sampling only 1 million segments (hereinafter referred to as 1M corpus) from the resulting 7M filtered documents — the goal was to create a smaller and higher-quality dataset that would facilitate experimentation with embedding models, as encoding the original 10 million segments can be computationally expensive.

3.2. Manual queries creation

We employed human-created queries for the evaluation dataset, aiming high-quality questions to capture common information needs from a diverse corpus, created by native speakers of the target language. We created a total of 200 test queries.

3.3. Passages retrieval

The next step is the passage retrieval to build a list of query–passages to annotate. As it would be prohibitive to have the relevance scores each query for the entire corpus, we select to annotate the top-k passages returned by multiple IR systems. It is assumed that the diversity of their results will enable the collection of a variety of passages, creating a robust evaluation dataset.

We selected a mix of strong and weak IR systems, to include a variety of passages: **BM25**: a strong baseline for retrieval; **BM25 + mT5-XL**: two-stage pipeline with BM25 followed by mT5-XL (3.7 billion parameters) [Xue et al. 2020]; **BM25 + E5-large**: two-stage pipeline with BM25 and E5-large [Wang et al. 2022][4]; **E5-large** and **E5-base**: E5 variants as dense retrievers, using FAISS [Johnson et al. 2019] with inner product for search; **ColBERT-X** [Nair et al. 2022]: a multilingual ColBERT-v1 fine-tuned in Brazilian Portuguese subset of mMARCO; **SPLADE v2**: a learned sparse retriever [Formal et al. 2021] fine-tuned on Brazilian Portuguese subset of mMARCO; **SPLADE v2 + mT5-XL**: two-stage pipeline using SPLADE v2 followed by mT5.

We also use Reciprocal Ranking Fusion (RRF) [Cormack et al. 2009] to increase the retrieved documents diversity, using the following combinations: **E5-large + ColBERT-X**; **E5-large + SPLADE v2**; and **E5-large + BM25 + mT5-XL**.

We also include commercial embedding models: **text-embedding-ada-002**[5], **text-embedding-3-small**[6] and as it employs the Matryoshka Representation Learning technique [Kusupati et al. 2022], we performed the retrieval using only the first half dimensions (identified as text-embedding-3-small half). FAISS [Johnson et al. 2019] using inner product was applied for dense vectors search for all of them.

[4]https://huggingface.co/intfloat/multilingual-e5-large

[5]https://openai.com/blog/new-and-improved-embedding-model

[6]https://openai.com/blog/new-embedding-models-and-api-updates

To evaluate the diversity of retrieved passages, we counted the query–passage combinations exclusively returned by each IR system, which should be a number from 0 to 500, 0 meaning the query–passages returned by a particular IR system were also returned by another IR system. Although we look for diversity, there should be a balance: we could have reached 5,000 different query–passage combinations (10 IR systems, 50 queries, 10 passage/query) if all systems returned exclusive passages, but that would indicate no agreement on the most relevant passages per query.

3.4. Query–passages annotation

The final step of the query annotation is to use an LLM to label the retrieved passages' relevance for each query. We selected the top-k=10 passages for a sample of 50 queries using all the retrieval systems considered on both the 10M and 1M corpora and sent them for LLM evaluation. We applied a few-shot Chain-of-Thought (CoT) prompt [Wei et al. 2022], and we adopted the TREC 2021 Deep Learning track 4-score relevance annotation scale [Craswell et al. 2021]: (1) **Irrelevant**: the passage is outside the scope of the question; (2) **Relevant**: the passage pertains to the question's topic but does not provide a direct answer; (3) **Highly relevant**: the passage answers the question, but lacks in clarity or has unrelated information. (4) **Perfectly relevant**: the passage answers the question with clarity and precision.

We selected OpenAI GPT-4 model[7] as the annotator. Due to cost limitations, we used a 50-sample from the 200 queries. We asked the LLM to label only the top-10 retrieved passages of each IR system for each query. We used a CoT prompt with two in-context examples selected from the mMARCO pt-BR dataset [Bonifacio et al. 2021]. The prompt written in Brazilian Portuguese includes the task explanation and the CoT examples to produce the 4-score passage relevance value for a given query. The final evaluation was requested in JSON format to simplify the LLM response parsing process. The prompt was built and refined using a limited set of questions sampled from the same mMARCO pt-BR dataset. The final prompt version can be found online.[8]

4. Experiments

4.1. LLM annotation quality assessment

We assess the quality of our LLM-based annotator by comparing its query–passage relevance scores with those provided by human annotators. This process was conducted on a 24-sample of the 50 annotated queries. Using the Doccano [Nakayama et al. 2018] system, three researchers annotated the top-10 passages returned by the BM25 + mT5 IR system applying the same TREC-DL 2021 4-score grading system. The agreement among the query–passage relevance annotations generated by the LLM and humans was measured using Cohen Kappa, Pearson, and Spearman correlation coefficients.

4.2. Retrieval systems evaluation

We used the LLM annotated query–passages to evaluate the IR systems effectiveness in the 10M and 1M Quati datasets. As we already have the IR runs for the passages retrieval by all the systems (see Section 3.3), we simply compute the nDCG@10 metric over those

[7]We used gpt-4-1106-preview, available at the OpenAI API.

[8]https://github.com/unicamp-dl/quati/blob/main/prompt.md

Table 1. The single–system query–passages column indicates the ones returned only by that system, either for the 10M or the 1M sets; the percentage refers to 500 query–passages. For the single system total, the percentage refers to the union of evaluated passages. "Others" are results with data preparation issues, but valid annotations.

Retrieval System	**Single–system query–passages**	
	10M dataset	1M dataset
E5-base	262 (52.4%)	
BM25	248 (49.6%)	253 (50.6%)
SPLADE v2 pt-BR	151 (30.2%)	
E5-large	122 (24.4%)	
ColBERT-X mMARCO pt-BR	115 (23.0%)	195 (39.0%)
BM25 + E5-large	115 (23.0%)	120 (24.0%)
SPLADE v2 pt-BR + mT5-XL	86 (17.2%)	
BM25 + mT5-XL	60 (12.0%)	93 (18.6%)
E5-large + ColBERT-X mMARCO pt-BR RRF	32 (6.4%)	
E5-large + SPLADE v2 pt-BR RRF	29 (5.8%)	
text-embedding-ada-002		137 (27.4%)
text-embedding-3-large		121 (24.2%)
text-embedding-3-small half		45 (9.0%)
text-embedding-3-small		31 (6.2%)
Others	814 (54.27%)	
Single system query–passages total	3029 (61.96%)	
Union of all systems query–passages	4889	

results. Besides establishing a baseline for a variety of IR systems, this experiments also indirectly assess the overall quality of Quati validation dataset: by verifying different effectiveness for already published IR systems, we validate Quati potential to indeed assess such systems.

5. Results and Discussion

5.1. Annotated passages variability

Table 1 indicates a range from 29 to 262 query–passage combinations exclusively returned by a single IR system. On average, each system returned 28.85% of new passages, and from the total 4,889 evaluated query–passages, 61.96% (3029) were returned by a single system, suggesting our pool of systems is diverse. As shown in Table 2, the IR systems were able to retrieve a diversity set of query–passages, including "perfectly relevant" (score=3) ones; also, the diversity increased for less relevant passages, indicating the systems agreed more as the passage relevance increased.

5.2. LLM annotations quality is aligned with crowd workers

Table 3 shows the Cohen's Kappa and the Spearman's Rho correlation coefficients for the human and LLM annotations, computed for the 240 query–passage combinations. The average Cohen's Kappa of 0.31 is aligned with the literature. For example,

Table 2. Query–passage relevance score counts. The systems agreed more, returning the same passages per query, as the relevance score increases. "Relevant" includes passages from scores 1 to 3.

Score	All query–passages	Single-system query–passages	%
0	2489	1839	73.89
1	985	586	59.49
2	759	375	49.41
3	656	229	34.91
Relevant	2400	1190	49.58
Total	4889	3029	61.96

Table 3. Cohen's Kappa and Spearman's Rho correlations among Human Annotators (HA) and the GPT-4, for the query–passage 4-score evaluations. For each annotator, 4th row holds the average of the correlation against the others. We then compute the mean of that value only for the Human Annotators ("Mean HA" row), to characterize their overall correlation.

	Cohen's Kappa				Spearman's Rho			
	HA_1	**HA_2**	**HA_3**	**GPT-4**	**HA_1**	**HA_2**	**HA_3**	**GPT-4**
HA_1	—	0.4369	0.4294	0.3234	—	0.6931	0.6924	0.6073
HA_2	0.4369	—	0.4105	0.2593	0.6931	—	0.6985	0.6174
HA_3	0.4294	0.4105	—	0.3498	0.6924	0.6985	—	0.6296
Mean	**0.4331**	**0.4237**	**0.4199**	**0.3108**	**0.6927**	**0.6958**	**0.6954**	**0.6181**
Std	**0.0037**	**0.0132**	**0.0095**	**0.0380**	**0.0004**	**0.0027**	**0.0031**	**0.0091**
Mean HA	0.4256±0.0055			—	0.6946±0.0014			—
Diff. Mean HA	0.0076	-0.0019	-0.0057	-0.1096	-0.0019	0.0011	0.0008	-0.0765

[Faggioli et al. 2023] reported 0.26 for GPT-3.5, and [Thomas et al. 2023] reported Cohen's Kappa ranging from 0.20 to 0.64, depending on the prompt used on GPT-4. Our human annotators' mean Cohen's Kappa of 0.4256 falls within crowd workers interval of a 0.24 to 0.52, according to [Damessie et al. 2017].

As query–passage relevance annotation is a subjective task, we argue a non-categorical metric such as the Spearman's Rho would be more appropriate to measure the annotators' correlation, as errors by a single score level should be considered "less critical", or within the subjectivity intrinsic for the task. Although human annotators' correlation is still above their correlations with the LLM, Spearman metrics are within a higher value, better capturing the current LLM effectiveness on the query–passage relevance evaluation.

5.3. Retrieval systems evaluation results

We evaluated the retrievers effectiveness using the LLM annotated query–passages (qrels); table 4 present the results for both the 10M and the 1M datasets. The ranking

Table 4. The nDCG@10 effectiveness on the 50 test queries. The results follows the IR literature and suggests the dataset can effectively evaluate a range of different IR systems.

Retrieval system	nDCG@10	
	10M dataset	1M dataset
BM25	0.4467	0.3991
E5-large	0.5563	—
SPLADE v2 pt-BR	0.5806	—
E5-large + SPLADE v2 pt-BR RRF	0.6272	—
ColBERT-X mMARCO pt-BR	0.6279	0.4927
BM25 + E5-large	0.6364	0.5423
text-embedding-ada-002	—	0.5630
text-embedding-3-small	—	0.5688
text-embedding-3-large	—	0.6319
E5-large + ColBERT-X mMARCO pt-BR RRF	0.6377	—
SPLADE v2 pt-BR + mT5-XL	0.6966	—
BM25 + mT5-XL	0.7109	0.6593

of retrievers with respect to effectiveness matched our expectations, following the literature. We consider that an additional indication of the overall datasets quality as, despite being created in a semi-automated cost-effective way, they are able to evaluate a diversity of retrievers.

6. Conclusion

This paper introduced the Quati, a dataset for supporting the development of IR systems for Brazilian Portuguese retrieval tasks. Quati is publicly available in two sizes, 10M a 1M passages, with 50-query qrels with respectively an average of 97.78 and 38.66 annotated passages per query. Through comparisons with human annotators we answer our research question, showing that state-of-the-art LLM can be used in a semi-automated and cost-effective way to create IR datasets for a specific target language, in the query–passage annotation role, with equivalent performance of humans: LLM annotations correlate with humans' in similar way human crowd workers annotations do, for a fraction of the cost.

Acknowledgements

We thank Leodécio Braz da Silva Segundo for the valuable support during the human annotation task. We also thank Leonardo Benardi de Avila and Monique Monteiro for the SPLADE v2 retrievals, using the model they trained for Brazilian Portuguese. This research was partially funded by grant 2022/01640-2 from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

References

Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., et al. (2016). Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Bonifacio, L., Jeronymo, V., Abonizio, H. Q., Campiotti, I., Fadaee, M., Lotufo, R., and Nogueira, R. (2021). mmarco: A multilingual version of the ms marco passage ranking dataset. *arXiv preprint arXiv:2108.13897*.

Clark, J. H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V., and Palomaki, J. (2020). Tydi qa: A benchmark for information-seeking question answering in ty pologically di verse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Cormack, G. V., Clarke, C. L., and Buettcher, S. (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759.

Craswell, N., Mitra, B., Yilmaz, E., Campos, D., Voorhees, E. M., and Soboroff, I. (2021). Trec deep learning track: Reusable test collections in the large data regime. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2369–2375.

Damessie, T. T., Nghiem, T. P., Scholer, F., and Culpepper, J. S. (2017). Gauging the quality of relevance assessments using inter-rater agreement. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1089–1092.

Faggioli, G., Dietz, L., Clarke, C., 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. *arXiv preprint arXiv:2304.09161*.

Farzi, N. and Dietz, L. (2024). An exam-based evaluation approach beyond traditional relevance judgments. *arXiv preprint arXiv:2402.00309*.

Formal, T., Lassance, C., Piwowarski, B., and Clinchant, S. (2021). Splade v2: Sparse lexical and expansion model for information retrieval. *arXiv preprint arXiv:2109.10086*.

Jeronymo, V., Nascimento, M., Lotufo, R., and Nogueira, R. (2022). mrobust04: A multilingual version of the trec robust 2004 benchmark. *arXiv preprint arXiv:2209.13738*.

Johnson, J., Douze, M., and Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., and Mikolov, T. (2016a). Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016b). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Kusupati, A., Bhatt, G., Rege, A., Wallingford, M., Sinha, A., Ramanujan, V., Howard-Snyder, W., Chen, K., Kakade, S., Jain, P., et al. (2022). Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249.

Lawrie, D., Mayfield, J., Oard, D. W., and Yang, E. (2022). Hc4: A new suite of test collections for ad hoc clir. In *European Conference on Information Retrieval*, pages 351–366. Springer.

Lewis, D. D., Yang, Y., Russell-Rose, T., and Li, F. (2004). Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397.

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*, pages 2363–2368.

Nair, S., Yang, E., Lawrie, D., Duh, K., McNamee, P., Murray, K., Mayfield, J., and Oard, D. W. (2022). Transfer learning approaches for building cross-language dense retrieval models. In *European Conference on Information Retrieval*, pages 382–396. Springer.

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., and Liang, X. (2018). doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Overwijk, A., Xiong, C., and Callan, J. (2022). Clueweb22: 10 billion web documents with rich information. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3360–3362.

Peters, C. and Braschler, M. (2002). The importance of evaluation for cross-language system development: the clef experience. In *LREC*. Citeseer.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., et al. (2021). Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Sakai, T., Oard, D. W., and Kando, N. (2021). *Evaluating Information Retrieval and Access Tasks: NTCIR's Legacy of Research Impact*. Springer Nature.

Schäuble, P. and Sheridan, P. (1998). Cross-language information retrieval (clir) track overview. *NIST SPECIAL PUBLICATION SP*, pages 31–44.

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

Vitório, D., Souza, E., Martins, L., da Silva, N. F., de Carvalho, A. C. P. d. L., Oliveira, A. L., and de Andrade, F. E. (2024). Building a relevance feedback corpus for legal information retrieval in the real-case scenario of the brazilian chamber of deputies. *Language Resources and Evaluation*, pages 1–21.

Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., and Wei, F. (2022). Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., ichter, b., Xia, F., Chi, E., Le, Q. V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2020). mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Zendel, O., Culpepper, J. S., Scholer, F., and Thomas, P. (2024). Enhancing human annotation: Leveraging large language models and efficient batch processing. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*, pages 340–345.

Zhang, X., Thakur, N., Ogundepo, O., Kamalloo, E., Alfonso-Hermelo, D., Li, X., Liu, Q., Rezagholizadeh, M., and Lin, J. (2023). Miracl: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.