Quati: A Brazilian Portuguese Information Retrieval Dataset from Native Speakers

Supplementary Material

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

[1] University of Campinas
[2] Maritaca AI
[3] NeuralMind
[4] Universidade Federal do Cariri

1 Manual queries creation

We employed human-created queries for the evaluation dataset, looking for high-quality questions that resemble common information needs from a diverse corpus, created by native speakers of the target language. We created a total of 200 test queries considering two different approaches:

1. **Taxonomy-guided corpus-agnostic questions:**
   We proposed the following taxonomy of themes, and questions characteristics to guide the first 100 queries creation in a corpus-agnostic fashion, i.e., without knowing in advance if the corpus would necessarily have good answers for them:
   - Themes: **Geography, Politics, Economy, Culture, Culinary, Tourism, Leisure, Sports**.
   - Scope: **General** (exploring a broad theme or subject) or **Specific** (exploring a narrow theme or subject).
   - Type: **Opinion** (asking for an opinion about something) or **Factual** (asking for a fact or data which has little dependency on one's opinion).

   Two Brazilian Portuguese native speakers from the research team created those initial 100 queries.

2. **Questions from documents:**
   Another 100 queries were created making sure the corpus included at least one good answer to the query: we sampled 100 passages we prepared from the original corpus and created one query regarding those subjects. That was performed by another member of our research team, also a Brazilian Portuguese native speaker.

* Equal contribution.

2 Annotators correlation varies per question

The annotators correlation varies per question — inter-human annotation Cohen's Kappa varies from -0.0236 to 0.7857, while GPT-4 versus human annotators range from -0.0561 to 0.7411 — and as shown in Figure 1 both humans and LLM find most of the time the same questions harder to evaluate for passage relevance — harder meaning more confusing, hence reducing scores correlation. See Tables 10 and 11 for human versus human and human versus LLM correlations for each query.

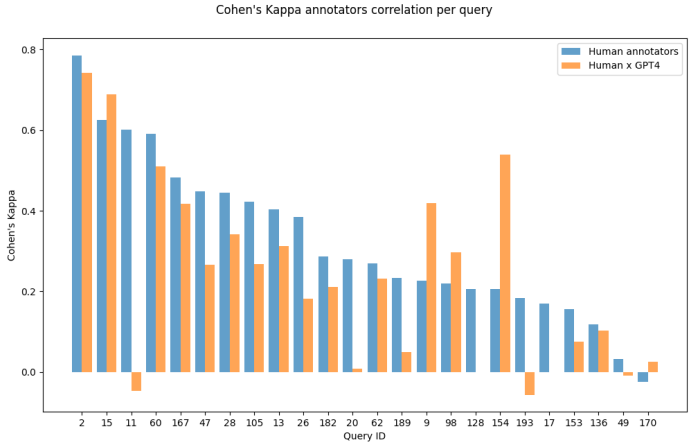

Fig. 1: When analyzed per query, most of the time humans and GPT-4 find the same questions more confusing to annotate for passage relevance, as the Cohen's Kappa correlation indicates. For 3 questions — IDs 9, 98 and 154, GPT-4 got very correlated to Human Annotator 2, which explains the higher metrics for those questions.

Considering query–passage relevance annotation is a subjective task, we argue that a non-categorical metric 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 interval which ends up being intrinsic for the task. Therefore, for completeness, we present the Spearman and Pearson correlations between humans and GPT-4 annotators in Tables 2 and 3, respectively. Even though on both metrics human annotators' correlation is still well above their individual correlations with the GPT-4 scores, all the metrics are within a higher value, which we argue better captures the current LLM effectiveness on the query–passage relevance evaluation task, which is comparable to human crowd workers.

	HA$_1$	**HA$_2$**	**HA$_3$**	**GPT-4**
Human Annotator 1 (HA$_1$)	—	0.4369	0.4294	0.3234
Human Annotator 2 (HA$_2$)	0.4369	—	0.4105	0.2593
Human Annotator 3 (HA$_3$)	0.4294	0.4105	—	0.3498
Mean	**0.4331**	**0.4237**	**0.4199**	**0.3108**
Standard Deviation	**0.0037**	**0.0132**	**0.0095**	**0.0380**
Mean human annotators	**0.4256±0.0055**			—
Diff. human annotators Mean	0.0076	-0.0019	-0.0057	-0.1096

Table 1: Cohen's Kappa correlations for the 4-score evaluations.

	HA$_1$	**HA$_2$**	**HA$_3$**	**GPT-4**
Human Annotator 1 (HA$_1$)	—	0.6931	0.6924	0.6073
Human Annotator 2 (HA$_2$)	0.6931	—	0.6985	0.6174
Human Annotator 3 (HA$_3$)	0.6924	0.6985	—	0.6296
Mean	**0.6927**	**0.6958**	**0.6954**	**0.6181**
Standard Deviation	**0.0004**	**0.0027**	**0.0031**	**0.0091**
Mean human annotators	**0.6946±0.0014**			—
Diff. human annotators Mean	-0.0019	0.0011	0.0008	-0.0596

Table 2: Spearman correlations for the 4-score evaluations.

	HA$_1$	**HA$_2$**	**HA$_3$**	**GPT-4**
Human Annotator 1 (HA$_1$)	—	0.6982	0.6973	0.5982
Human Annotator 2 (HA$_2$)	0.6982	—	0.7132	0.6146
Human Annotator 3 (HA$_3$)	0.6973	0.7132	—	0.6326
Mean	**0.6977**	**0.7057**	**0.7052**	**0.6151**
Standard Deviation	**0.0004**	**0.0075**	**0.0079**	**0.0140**
Mean human annotators	**0.7029±0.0037**			—
Diff. human annotators Mean	-0.0051	0.0028	0.0024	-0.0652

Table 3: Pearson correlations for the 4-score evaluations.

The correlation between the relevance scores assigned by human annotators is higher than the correlation between humans and LLM. However, both have high variation depending on the question, and the overall correlation achieved by the LLM – within the crowd workers correlation interval – confirms the possibility of creating an evaluation dataset with a high quality/cost relation.

At the current development stage, LLM evaluation effectiveness is below human when distinguishing between closer categories —"not relevant" or "on topic"; "highly relevant" or "perfectly relevant" — but it is reasonable to expect that will improve, as LLMs improve their ability to understand text nuances. There is also room to enhance LLM passage relevance evaluation through additional prompt engineering, given LLM's sensitivity to the provided input [2].

3 Human Annotators confusion matrices

Tables 7, 8, and 9 present the confusion matrix between each human annotator and GPT-4 illustrating the LLM behavior. Similar to the findings of Faggioli et al. [1], we verified GPT-4 tends to evaluate the passages with higher scores when compared to humans. For instance, when compared to Human Annotator 1, GPT-4 classified as "Relevant (1)" 13 of 52 (25%) query–passages annotated as "Irrelevant (0)" by the human; that was the higher misclassification on that score. A similar behavior is verified for the other scores, with GPT-4 indicating as "Highly relevant (2)" 18 of 68 (26.47%) "Relevant (1)", or misclassifying as "Perfectly relevant (3)" 27 of 65 (41,54%) "Highly relevant (2)" query–passages. In fewer cases, GPT-4 classifies as "Irrelevant (0)" something humans have annotated as either "Highly relevant (2)" or "Perfectly relevant (3)" or vice-versa.

The human annotators presented similar behavior (Tables 4, 5, and 6), but disagreeing more on the intermediate scores — "Relevant (1)" and "Highly relevant (2)" — indicating they also suffer from the boundary uncertainty when determining the query–passage relevance score.

		Human Annotator 2 (HA_2)				**HA_1 totals**
	Score	**0**	**1**	**2**	**3**	
Human Annotator 1 (HA_1)	Irrelevant (0)	41	6	4	1	52
	Relevant (1)	13	25	28	2	68
	Highly relevant (2)	4	11	42	8	65
	Perfectly relevant (3)	1	5	18	31	55
	HA_2 totals	59	47	92	42	

Table 4: Human Annotator 1 and Human Annotator 2 scores confusion matrix. There is no clear disagreement tendency (lower or higher score).

		Human Annotator 3 (HA_3)				HA_1 totals
	Score	**0**	**1**	**2**	**3**	
Human Annotator 1 (HA_1)	Irrelevant (0)	41	7	1	3	52
	Relevant (1)	9	29	21	9	68
	Highly relevant (2)	2	11	26	26	65
	Perfectly relevant (3)	1	5	8	41	55
	HA_3 totals	53	52	56	79	

Table 5: Human Annotator 1 and Human Annotator 3 scores confusion matrix.

		Human Annotator 3 (HA_3)				HA_2 totals
	Score	**0**	**1**	**2**	**3**	
Human Annotator 2 (HA_2)	Irrelevant (0)	44	11	1	3	59
	Relevant (1)	5	19	14	9	47
	Highly relevant (2)	4	21	35	32	92
	Perfectly relevant (3)	0	1	6	35	42
	HA_3 totals	53	52	56	79	

Table 6: Human Annotator 2 and Human Annotator 3 scores confusion matrix.

		GPT-4				Human totals
	Score	**0**	**1**	**2**	**3**	
Human Annotator 1	Irrelevant (0)	25	13	12	2	52
	Relevant (1)	12	24	18	14	68
	Highly relevant (2)	4	11	23	27	65
	Perfectly relevant (3)	1	5	3	46	55
	GPT-4 totals	42	53	56	89	

Table 7: Human Annotator 1 and GPT-4 scores confusion matrix.

		GPT-4				Human
	Score	**0**	**1**	**2**	**3**	**totals**
Human Annotator 2	Irrelevant (0)	29	17	11	2	59
	Relevant (1)	5	15	17	10	47
	Highly relevant (2)	8	20	24	40	92
	Perfectly relevant (3)	0	1	4	37	42
	GPT-4 totals	42	53	56	89	

Table 8: Human Annotator 2 and GPT-4 scores confusion matrix.

		GPT-4				Human
	Score	**0**	**1**	**2**	**3**	**totals**
Human Annotator 3	Irrelevant (0)	29	13	10	1	53
	Relevant (1)	7	21	11	13	52
	Highly relevant (2)	6	12	19	19	56
	Perfectly relevant (3)	0	7	16	56	79
	GPT-4 totals	42	53	56	89	

Table 9: Human Annotator 3 and GPT-4 scores confusion matrix. Best agreement on the highest score.

4 Annotators correlation per question

Query	a1×a2 a1×LLM	a2×a3 a2×LLM	a3×a1 a3×LLM	Mean±Std
Onde está localizada a Praça XV de Novembro?	0.2647 0.2857	0.6970 0.2105	0.3056 0.3056	0.4224±0.1949 0.2673±0.0409
Qual foi a importância da usina de Volta Redonda RJ para a industrialização brasileira?	-0.0127 0.0909	0.1026 0.1176	0.2647 0.1026	0.1182±0.1138 0.1037±0.0109
Qual o uso dos códigos SWIFT?	0.6154 1.0000	0.0000 0.6154	0.0000 0.0000	0.2051±0.2901 0.5385±0.4119
O que são os celulares "mid-range"?	0.5082 0.5238	0.2537 0.2188	0.6875 0.5082	0.4831±0.1780 0.4169±0.1403
Por que os países Guiana e Suriname não são filiados a Conmebol?	0.8361 0.6825	0.8438 0.6970	0.6774 0.8438	0.7857±0.0767 0.7411±0.0728
quais os critérios de definição dos monumentos intitulados maravilhas do mundo moderno?	0.7015 0.0278	0.5455 0.0141	0.5588 -0.1842	0.6019±0.0706 -0.0474±0.0969
Qual a maior torcida de futebol do Brasil?	0.8077 0.4231	0.6429 0.6429	0.4231 1.0000	0.6245±0.1576 0.6886±0.2377
Quando se realizou o plebiscito popular para definir o sistema político do Brasil?	0.0000 0.0000	0.0000 0.0000	0.5082 0.0000	0.1694±0.2396 0.0000±0.0000
Como transformar uma cidade pacata em um polo turístico?	0.2857 0.4030	0.5833 0.1250	0.4737 0.2683	0.4476±0.1229 0.2654±0.1135
Quais são os melhores parques nacionais de Portugal?	-0.0811 -0.4286	0.4286 0.1304	-0.2500 0.2727	0.0325±0.2884 -0.0085±0.3027
Quando foi criada a consolidação das leis trabalhistas no brasil?	0.5833 0.4595	0.5946 0.5714	0.5946 0.5000	0.5908±0.0053 0.5103±0.0463
Quais partidos já ocuparam o cargo da presidência do Brasil?	-0.0448 0.2424	0.0000 0.3333	0.8507 0.1176	0.2687±0.4120 0.2311±0.0884

Table 10: Annotators Cohen's Kappa correlation per question, for the initial 12 questions. For each question, first line contains the Human Annotators correlation, second line the Human x GPT-4 correlation. Zeroed correlations for two or more different annotators for a same question is due to one of the annotators have applied the same score for all the 10 passages of that given query.

Query	a1×a2 a1×LLM	a2×a3 a2×LLM	a3×a1 a3×LLM	Mean±Std
Quando é o dia Mundial da Alimentação?	0.4737 0.0000	0.2308 0.0000	-0.0870 0.0000	0.2058±0.2296 0.0000±0.0000
Quais os tipos de denominação (DO) que os vinhos podem receber?	0.2857 0.2593	0.1111 -0.0345	0.0698 0.0000	0.1555±0.0936 0.0749±0.1311
Quais as causas para lábios inflamados em crianças?	0.1803 0.0000	-0.2121 -0.1667	-0.0390 0.2424	-00236±0.1606 0.0253±0.1680
Quais as origens de pessoas com olhos verdes?	0.4595 0.3421	0.2105 0.2105	0.1892 0.0789	0.2864±0.1227 0.2105±0.1074
No que se difere o civismo da cidadania?	0.1566 0.0909	0.3750 0.0588	0.1667 0.0000	0.2328±0.1007 0.0499±0.0376
Quais atitudes podem prejudicar a saúde mental?	-0.0606 -0.0465	0.0278 -0.1494	0.5833 0.0278	0.1835±0.2850 -0.0561±0.0727
quais países europeus seguem o regime monarquista?	0.0909 0.2424	0.4643 0.7059	0.1228 0.3103	0.2260±0.1690 0.4196±0.2044
Como o Brasil reagiu a epidemia de AIDS no fim do século XX?	0.5161 0.4366	0.4286 0.1781	0.2647 0.3243	0.4031±0.1042 0.3130±0.1059
Por que a legislação de um país é tão importante?	0.1667 0.1176	0.1667 -0.0417	0.5082 -0.0526	0.2805±0.1610 0.0078±0.0778
Como podemos classificar o relevo brasileiro?	0.3750 0.1667	0.2647 0.0789	0.5161 0.3023	0.3853±0.1029 0.1826±0.0919
Existem vantagens ao definir uma moeda única?	0.7222 0.3056	0.3056 0.3056	0.3056 0.4118	0.4444±0.1964 0.3410±0.0501
Qual o critério para classificação para a Copa do Brasil?	0.2405 0.1250	0.3750 0.5000	0.0411 0.2647	0.2189±0.1372 0.2966±0.1547

Table 11: Annotators Cohen's Kappa correlation per question, for the last 12 questions. As in Table 10, for each listed question, the first line contains the Human Annotators correlation, second line the Human x GPT-4 correlation. Also, Zeroed correlations for two or more different annotators for the same question are due to one of the annotators having applied the same score for all the 10 passages of that given query.

References

1. Guglielmo Faggioli, Laura Dietz, Charles Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, et al. Perspectives on large language models for relevance judgment. *arXiv preprint arXiv:2304.09161*, 2023.

2. Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. Large language models can accurately predict searcher preferences. *arXiv preprint arXiv:2309.10621*, 2023.