Relevance labeling using Large Language Models for Offline-Online evaluation alignment in Information Retrieval Systems

Jan Piotrowski Avishek Anand [1]

[1]Supervisor

1. Introduction

Information Retrieval Systems are in widespread usage in everyday life. They range from recommender systems to search engines. Important aspect of their development is accurate evaluation. There are two distinct methods:

- **Offline evaluation**: uses datasets of annotated labels. Requires humans to first label the relevance, but can be easily reused.
- **Online evaluation**: traditionally relies on A/B testing on live traffic. Provides direct measurements of user utility. Very resource intensive.

Offline evaluation is plagued by multiple factors:

- **Problem of incomplete judgments**: modern datasets like MS-Marco are sparse and often provide only one relevant passage per query.
- **Misalignment with Online performance**: Offline evaluation does not always correspond to results of an Online evaluation. Without this alignment its values decreases radically.

We will address the first problem with use of Large Language Models as related works successfully employed those models as relevance judges to achieve similar performance to human labels; And investigate the use of LLMs on the alignment of Offline-Online evaluation.

2. Research Question

- Can labeling relevance using Large Language Models be used to align Offline Evaluation with Online Performance in Information Retrieval Systems?
- Does parameter size of LLM affect the alignment?
- Does use of different prompts to LLM affect the alignment?

3. Labeling setup

For the experiment we prepared two datasets TREC 2019 as a small set of 43 queries and 9260 passages with human annotated labels to validate the setup; And Rank-DistiLLM with 2000 queries and aproximately 200 thousand passages total for the main results.

Following open source models were used:

- Qwen3-32B
- Llama 3.1 8B
- GPT-OSS 20B

Following prompts were used:

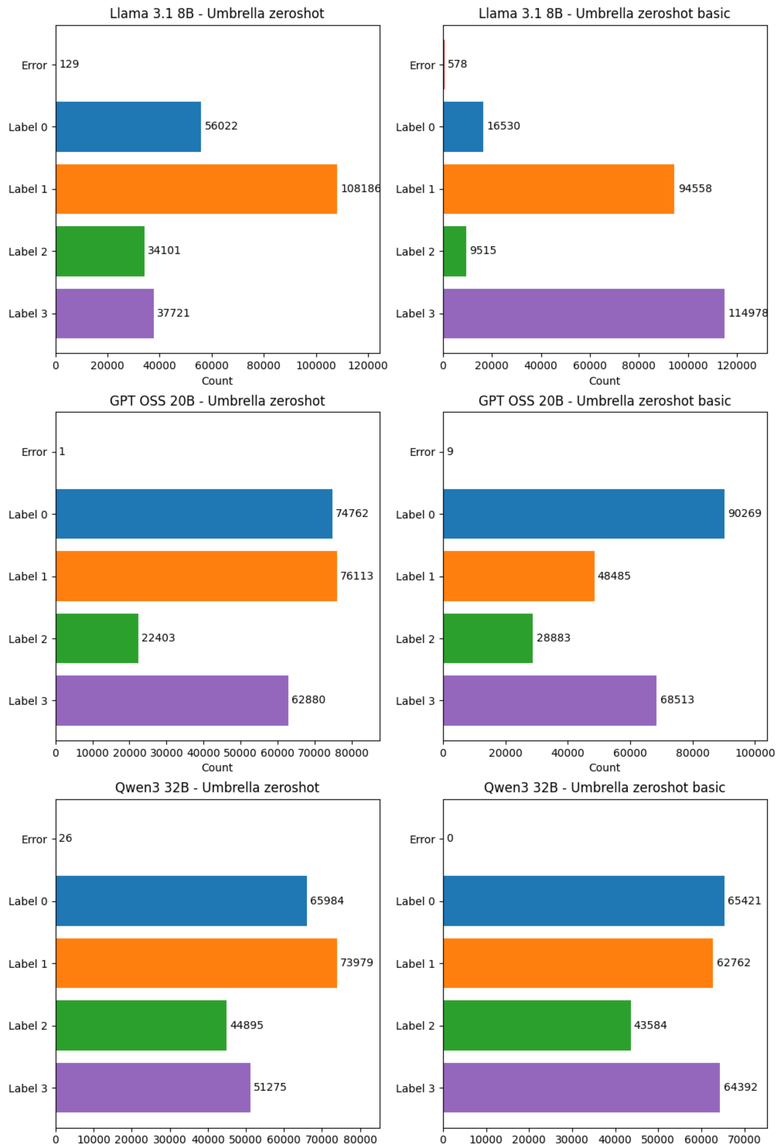
- Umbrella zeroshot
- Umbrella zeroshot basic

Each combination of model and prompt technique was used to label prepared query-document pairs. In our work we use 4 categories of labels from 0 to 3.

Below are descriptions taken from Umbrella zeroshot prompt:

- 0 represent that the passage has nothing to do with the query
- 1 represents that the passage seems related to the query but does not answer it
- 2 represents that the passage has some answer for the query, but the answer may be a bit unclear, or hidden amongst extraneous information
- 3 represents that the passage is dedicated to the query and contains the exact answer

4. Distribution of relevance labels



5. Results

We utilize the *Aligned IR Evaluation* framework. A modular pipeline that allows us to control individual components: retrieval systems, query selection strategies and user interaction model (online proxy). Relevance assessment uses labels from previous part as well as baselines and computes nDCG@10. We calculate correlation to other offline metrics using Kendall's Tau.

Table 1. Alignment against baseline labels measured by τ_B

Model	Prompt	Trec 2019	Rank-DistiLLM
Llama 3.1 8B	zeroshot basic	0.664	0.698
	zeroshot	0.851	0.798
GPT OSS 20B	zeroshot basic	0.851	0.802
	zeroshot	0.839	0.814
Qwen3 32B	zeroshot basic	0.838	0.804
	zeroshot	0.844	0.814

To measure Offline-Online alignment we use set of external metrics created on the same set of queries which include Click Through Rate, Average Dwell Time, Session Abandonment Rate, Zero Result Rate, and Total Clicks. We additionally compare against Click from a DBN model provided by the evaluation framework.

Table 2. Alignment of Offline relevance scores against Online metrics measured by τ_B

Model	Prompt	ADT	SSR	ZRR	Clicks	CTR	Clicks (DBN)
Llama 3.1 8B	zeroshot	-0.257	-0.114	-0.114	-0.412	-0.330	0.467
	zeroshot basic	-0.248	-0.164	-0.164	-0.353	-0.284	0.422
GPT OSS 20B	zeroshot	-0.251	-0.103	-0.103	-0.430	-0.342	0.644
	zeroshot basic	-0.254	-0.112	-0.112	-0.428	-0.345	0.644
Qwen3 32B	zeroshot	-0.252	-0.107	-0.107	-0.433	-0.344	0.600
	zeroshot basic	-0.251	-0.109	-0.109	-0.421	-0.341	0.733
DistiLLM		-0.234	-0.050	-0.050	-0.365	-0.293	0.644

We observe strong correlation against baseline labels across all models in Offline evaluation. When compared against online metrics we observe moderate correlation in columns Clicks and CTR; And weak correlation in columns ADT. We don't observe any correlation in columns SRR and ZRR by inability of rejecting the null hypothesis due to $p > 0.05$.

6. Conclusions and Future Work

We confirmed the research question that LLMs can be used to align offline evaluation with online performance as we have seen an increase in correlation when compared to the baseline labels. We observed that higher parameter size LLMs had higher correlation when compared to the smallest model. We observed that simpler prompt in the smallest model performed worse than the baseline.

We leave further investigation of the negative correlation against the external set of Online metrics to the future work. As well as investigating the change in alignment by using wider selection of model families, model architectures and model parameter sizes. Furthermore an effect of wider selection of prompts on the alignment is unanswered.