Multi-stage Large Language Model Pipelines Can Outperform GPT-4o in Relevance Assessment

Julian A. Schnabel[1], **Johanne R. Trippas**[2], **Falk Scholer**[2] **and Danula Hettiachchi**[2]

1. Heinrich Heine University, Düsseldorf, Germany
2. RMIT University, Melbourne, Australia

Correspondence: julian.schnabel@hhu.de

Abstract

Generating relevance labels for documents to indicate the usefulness for specific queries and users. Obtaining these labels from real users is costly and scaling is challenging. It has been shown [Thomas et al., 2024] that Large Language Models can be used to predict these relevance labels. We propose a novel multi-LLM pipeline that divides the relevance assessment task into multiple stages, each utilising different prompts and models of varying sizes and capabilities. This approach beats the baseline performance of GPT-4o on TREC-DL23.

Data and Methods

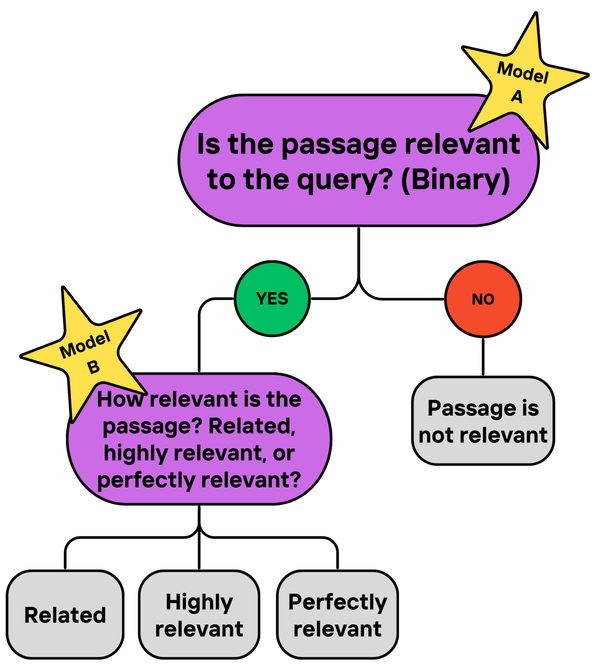

Figure 1: Visual overview of the pipeline approach where different models can judge at different stages of the relevance judgement labelling.

We use the TREC 2023 Deep Learning Track (TREC-DL 23). The original judgments were made by NIST assessors, who, given a query, assigned relevance scores to passages based on the following scale:

Label	Count
Label 0 (Irrelevant)	13,866
Label 1 (Related)	4,372
Label 2 (Highly relevant)	2,259
Label 3 (Perfectly relevant)	1,830

Table 1: Label distribution across the dataset.

0. **Irrelevant**: The passage has nothing to do with the query.
1. **Related**: The passage seems related to the query but does not answer it.
2. **Highly relevant**: The passage has some answer for the query, but the answer may be a bit unclear, or hidden amongst extraneous information.
3. **Perfectly relevant**: The passage is dedicated to the query and contains the exact answer.

Results

Reproducing Existing Baselines

- **Reproduction of UMBRELA Baseline**
  - Similar accuracy to Upadhyay et al. [2024]
  - Comparable misjudgement pattern
  - GPT-4o slightly over-optimistic
- **Model Comparison**
  - GPT-4o: Best performance among tested stand-alone models
  - Confirmed benchmarks by Alaofi et al. [2024]
- **GPT-4o Mini**
  - Only 3% of GPT-4o's cost
  - High agreement with GPT-4o
  - Higher binary accuracy with custom prompt (see Tab. 2)

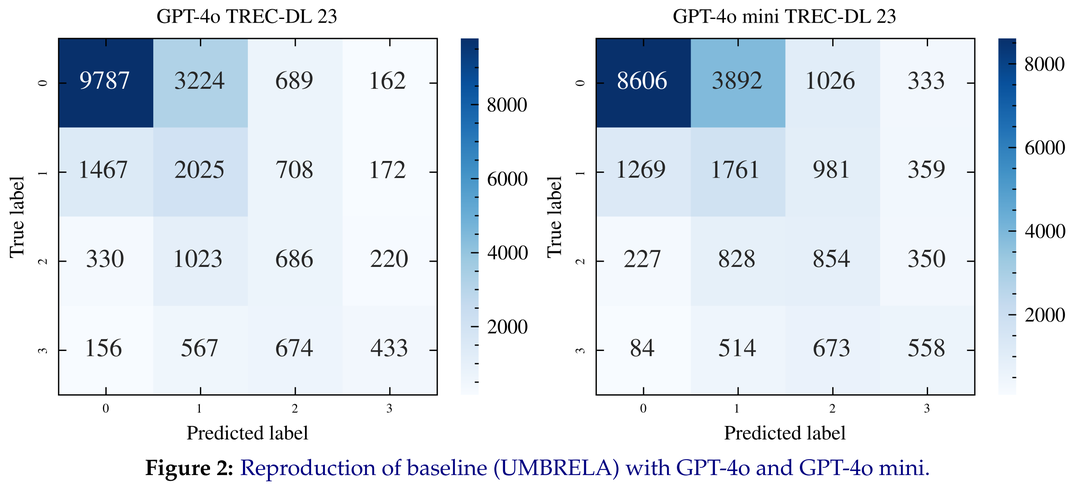

Figure 2: Reproduction of baseline (UMBRELA) with GPT-4o and GPT-4o mini.

Relevance Judgement Pipeline Approaches

Table 2 summarises the evaluation outcomes of the baselines and all proposed pipelines. We evaluate two models (GPT-4o and 4o-mini) under both homogeneous (4o–4o, mini–mini) and heterogeneous (mini–4o, 4o–mini) pairings, alongside two prompt types: Binary–Relevant and Binary–Normal.

Model 1	Model 2	Prompt 1	Prompt 2	Binary κ	4-scale κ	4-scale α	Cost USD
4o	-	Normal	-	0.453	**0.296**	0.408	5.00
mini	-	Normal	-	0.400	0.254	0.359	**0.15**
mini	mini	Binary	Relevant	0.437	0.284	0.422	0.21
4o	4o	Binary	Relevant	0.428	0.280	0.450	6.57
mini	4o	Binary	Relevant	0.437	0.286	0.432	2.05
4o	mini	Binary	Relevant	0.428	0.279	0.443	5.05
mini	mini	Binary	Normal	0.439	0.281	0.425	0.21
4o	4o	Binary	Normal	0.429	0.280	**0.452**	6.57
mini	4o	Binary	Normal	0.450	0.295	0.446	2.05
4o	mini	Binary	Normal	0.430	0.276	0.445	5.05
mini	4o	Normal	Normal	0.400	0.260	0.367	2.87
4o	mini	Normal	Normal	**0.462**	0.294	0.411	5.05

Table 2: Accuracy for different GPT model/prompt combinations on TREC-DL23. Cost in USD per million input tokens.

- Four-scale Krippendorff's α is higher in all multi-model approaches compared to single-model single-stage with GPT-4o.
- The four-scale κ-score of GPT-4o remains the highest overall (see Table 2).
- GPT-4o mini/GPT-4o with binary/normal prompts achieves similar accuracy at significantly lower cost (see Figure 3).

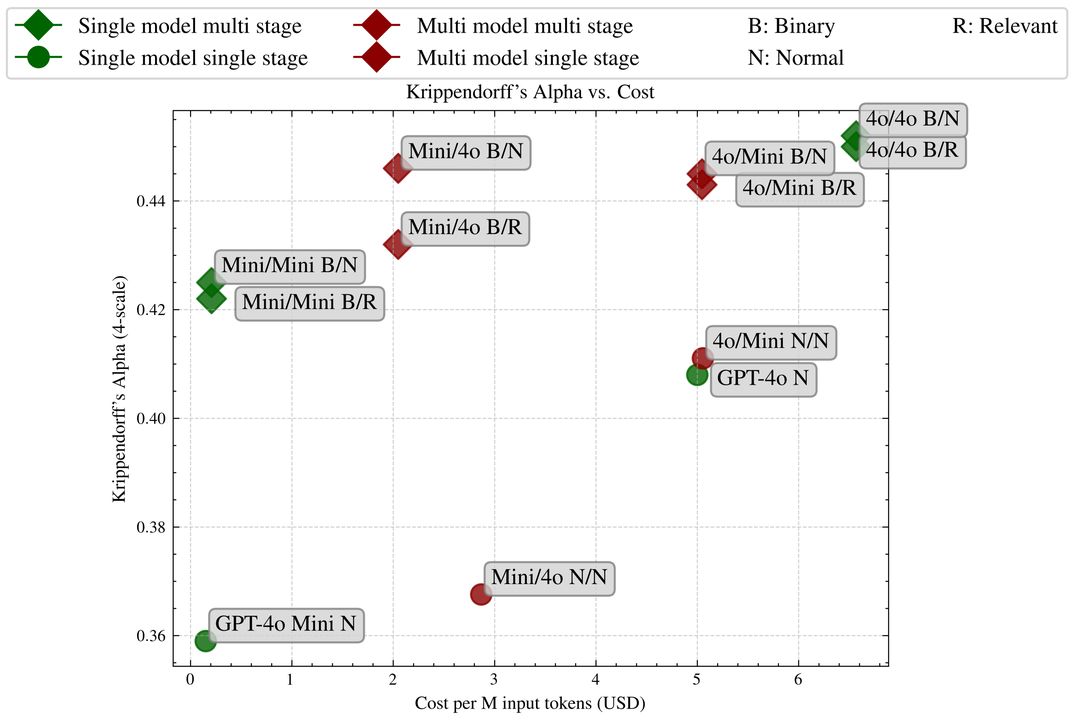

Figure 3: Cost vs Krippendorff's Alpha (α).

Summary

Pipeline Comparison & Results

- **Accuracy Increase**: All pipelines, except the Multi-model Single-stage approach, improved Krippendorff's α compared to the baseline.
- **GPT-4o mini**: Achieved the largest accuracy increase with a significant cost reduction.
- **Single-model Multi-stage Approach**: High accuracy at low cost, outperforming GPT-4o (over 20x more expensive).
- **Cost Efficiency**: GPT-4o mini enables more complex pipelines, e.g., multi-stage relevance classification or specialised prompts.

Key Insights

- **Specialised Binary Classification**: Improved accuracy, especially in multi-stage approaches with binary relevance decision and classification.
- **Limitations**:
  - No duplicate filtering (due to complexity) as noted in Upadhyay et al. [2024].
  - Could expand tests to other TREC-DL datasets and broader tasks.

Future Directions

- **Prompt Optimization**: Techniques like chain-of-thought Wei et al. [2022] and narratives Sadiri Javadi et al. [2024] could further improve performance.
- **Spam Filtering**: With 75% zero-weighted data in TREC-DL, small, affordable spam-filters could reduce assessment costs significantly.

Acknowledgments

This research is partially supported by the Australian Research Council (CE200100005) and RMIT AWS Cloud Supercomputing Hub.

References

Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. Large language models can accurately predict searcher preferences. 2024. ISBN 9798400704314.

Shivani Upadhyay, Ronak Pradeep, Nandan Thakur, Nick Craswell, and Jimmy Lin. Umbrela: Umbrela is the (open-source reproduction of the) bing relevance assessor. *arXiv preprint arXiv:2406.06519*, 2024.

Marwah Alaofi, Paul Thomas, Falk Scholer, and Mark Sanderson. Llms can be fooled into labelling a document as relevant. 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Vahid Sadiri Javadi, Johanne R Trippas, and Lucie Flek. Unveiling information through narrative in conversational information seeking. In *Proc. CUI*, 2024.