Does UMBRELA Work on Other LLMs?

Naghmeh Farzi & Laura Dietz

naghmeh.farzi@unh.edu

University of New Hampshire

Why Reproduce UMBRELA?

UMBRELA is an LLM judge built on a single, unified prompt.

- ✓ TREC **Robust04** systems with GPT-4 (Thomas 2024)
- ✓ TREC **DL 2019–2023** with GPT-4o (Upadhyay 2024)
- ✓ TREC **RAG 2024** systems with GPT-4o (Thakur 2024)

Thomas et al.: *LLMs can accurately predict searcher preferences*. SIGIR 2024

Upadhyay et al.: *Umbrela: Open-source reproduction of Bing's relevance assessor*. 2024

Thakur et al.: *Assessing Support for the TREC RAG Track with LLM and Human Evaluations*. 2024

Why Reproduce UMBRELA?

UMBRELA is an LLM judge built on a single, unified prompt.

- ✓ TREC **Robust04** systems with GPT-4 (Thomas 2024)
- ✓ TREC **DL 2019–2023** with GPT-4o (Upadhyay 2024)
- ✓ TREC **RAG 2024** systems with GPT-4o (Thakur 2024)

UMBRELA is LLM-agnostic.

Central Question

Does it still perform with smaller, open-weight models?

Thomas et al.: *LLMs can accurately predict searcher preferences*. SIGIR 2024

Upadhyay et al.: *Umbrela: Open-source reproduction of Bing's relevance assessor*. 2024

Thakur et al.: *Assessing Support for the TREC RAG Track with LLM and Human Evaluations*. 2024

UMBRELA Prompt

Given a query and a passage, you must provide a score on an integer scale of 0 to 3 with the following meanings:
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 and
3 = represents that the passage is dedicated to the query and contains the exact answer.

Important Instruction: Assign category 1 if the passage is somewhat related to the topic but not completely, category 2 if passage presents something very important related to the entire topic but also has some extra information and category 3 if the passage only and entirely refers to the topic. If none of the above satisfies give it category 0.

Query: {query}
Passage: {passage}

Split this problem into steps:
Consider the underlying intent of the search. Measure how well the content matches a likely intent of the query (M).
Measure how trustworthy the passage is (T).
Consider the aspects above and the relative importance of each, and decide on a final score (O). Final score must be an integer value only.
Do not provide any code in result. Provide each score in the format of: ##final score: score without providing any reasoning.

Cheaper LLMs: How Low Can We Go?

As resource-constrained academics, we can't afford to run UMBRELA on GPT-4o at scale.

We want to use LLMs which:

Cheap API	**Small GPUs**	**Open Weights**
convenient	run on local A40	reproducibility

Model	**Parameters**	**VRAM (FP16)**	**VRAM (4-bit)**
GPT-4o (estimate)	∼45–60B	∼80–100 GB	∼20–30 GB
DeepSeek-V3	37B / 671B	∼74 GB	∼18–20 GB
LLaMA-3.3–70B	70B	∼140 GB	∼35–40 GB
LLaMA-3–8B	8B	∼16 GB	∼4–5 GB
FLAN-T5-large	0.8B	∼2–4 GB	< 2 GB

Our Goals

In this reproducibility study, we examine:

➥ Impact of LLM scale

✓ Generalization across TREC DL (following Upadhyay 2024)

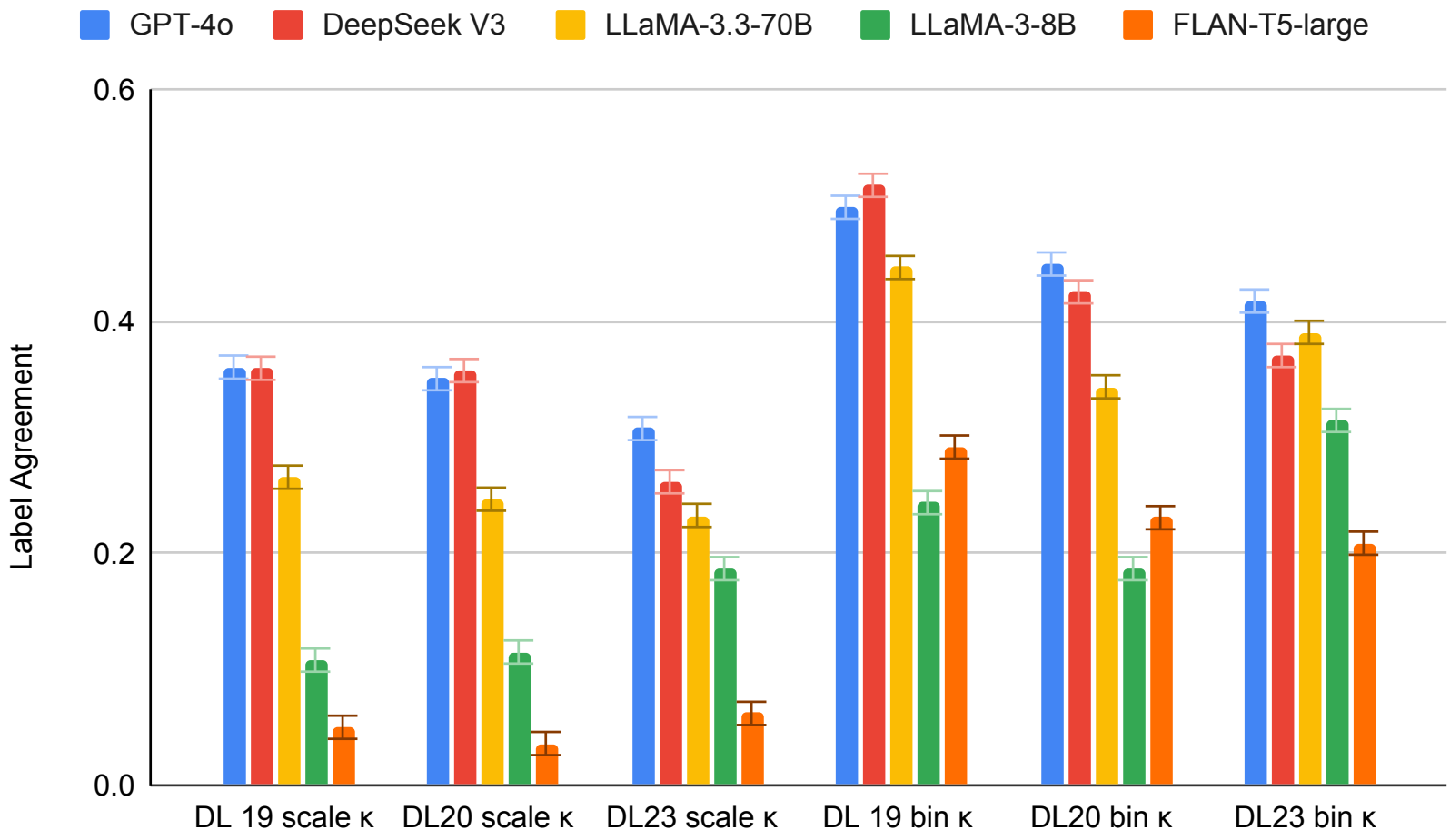
✗ Sensitivity to prompt variation

Meta-evaluation metrics:

- **Leaderboard correlation:** Kendall τ, Spearman ρ
- **Label agreement:** Cohens κ (4-point + binary)

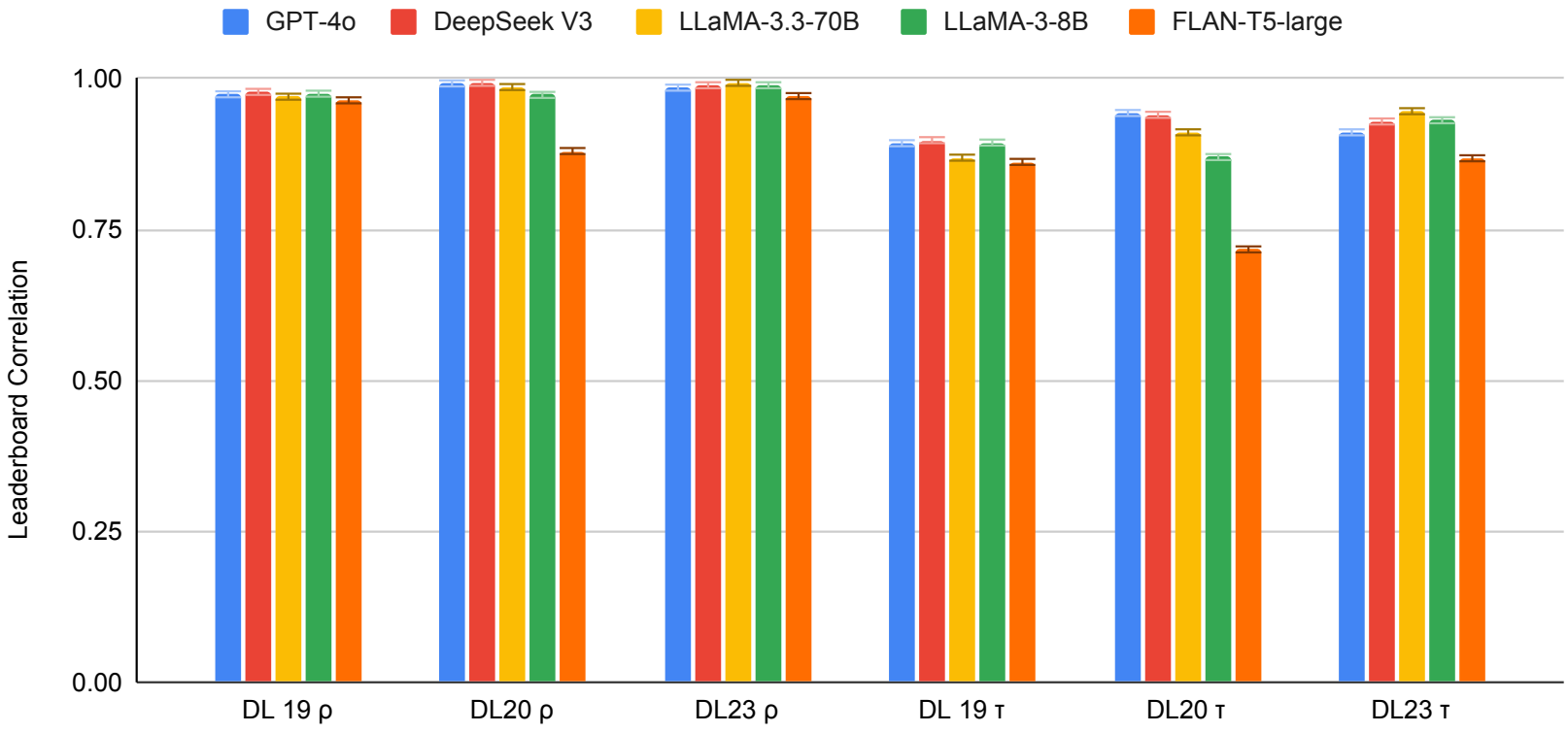
As used in the LLM Judge challenge

Per-label Agreement: Bigger = Better



DeepSeek comparable to GPT-4o.
LLaMA and FLAN-T5 consistently worse.

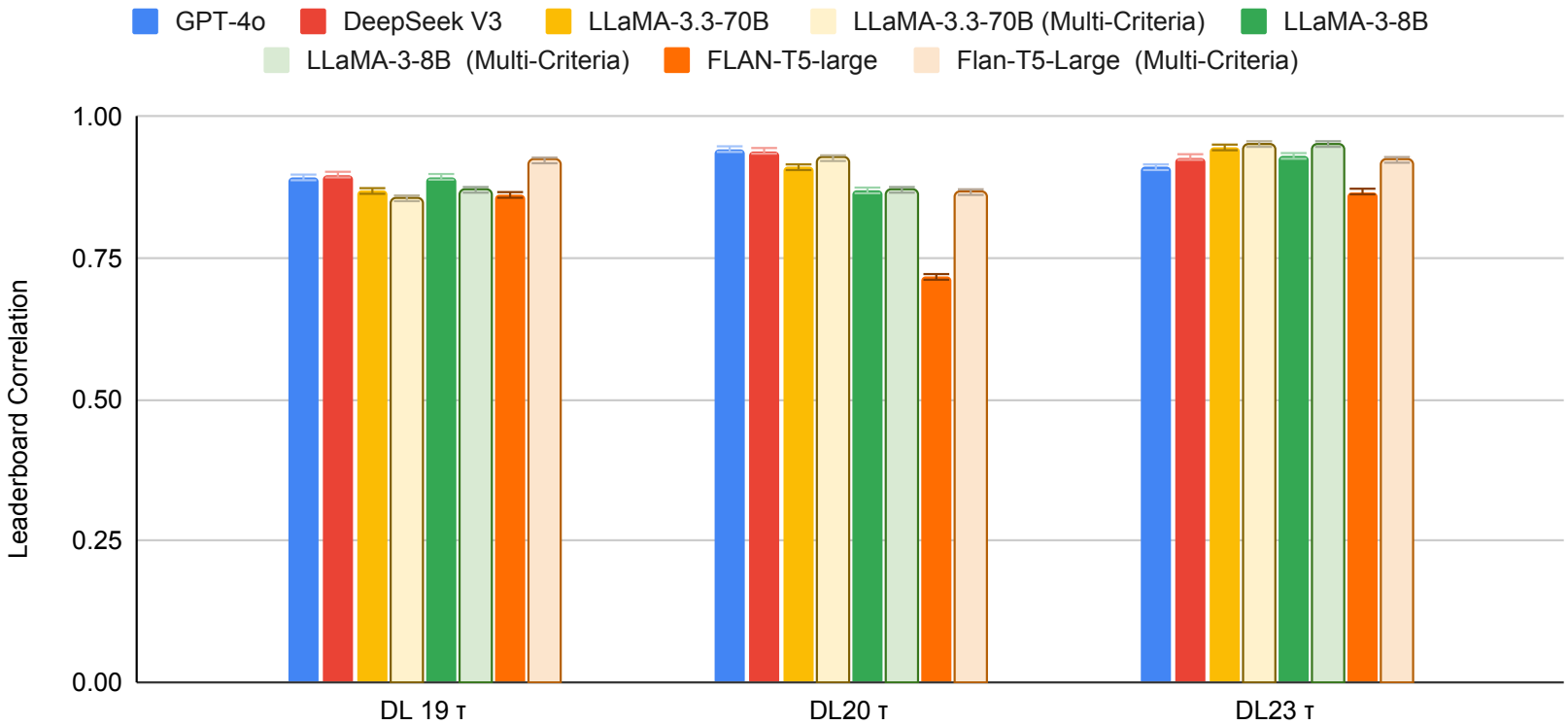
Leaderboard Correlation



All work reasonably well, lower performance for FLAN-T5-large.

Are We Doomed to Pay Many $ / € / ¥?

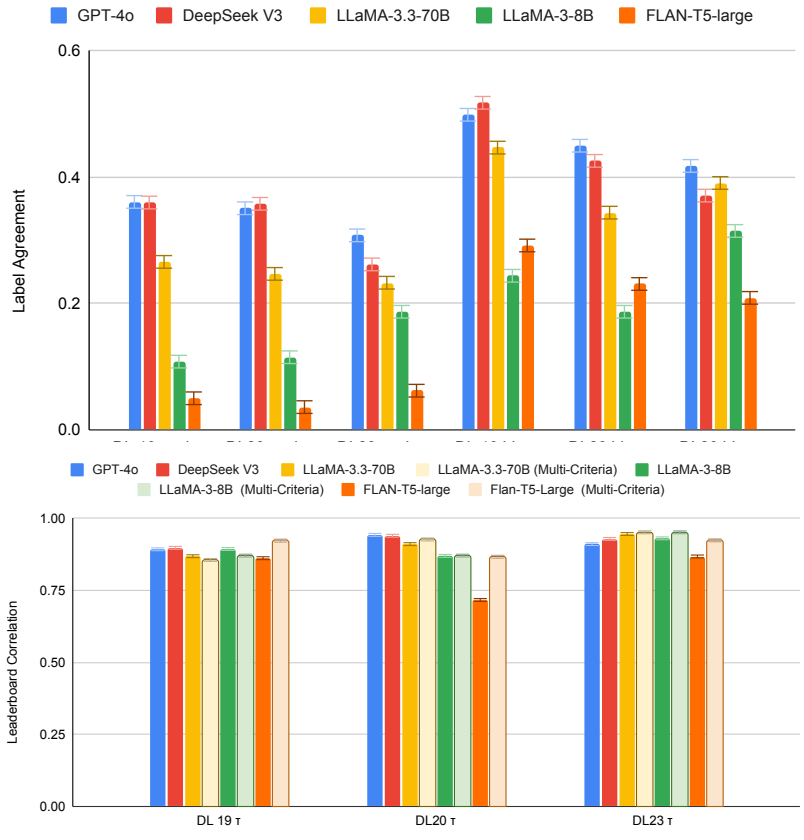
Not if relevance is decomposed into multiple individual prompts.



Farzi & Dietz: *Criteria-Based LLM Relevance Judgments*. ICTIR 2025

Conclusions: Does UMBRELA work on other LLMs?

- ✓ UMBRELA leaderboards also good with affordable LLMs.
- ✓ Bigger LLMs give more accurate relevance labels.
- ✗ LLM choice matters more than prompt tweaks.



Caveat: Study prone to Meta-evaluation Tropes:

Evaluation Tropes:

- ≅ **#6 Old Systems**: Evaluators need to identify the best systems of the future.
- ↻ **#2 LLM Evaluator as a Ranker**: Using the same approach in the system and the evaluation.
- **#7 LLM Evolution**: LLMs are not static; they can improve or degrade over time.
- ♡ **#8 Test Set Leak**: LLMs trained on test collections create the illusion of quality.

Dietz, Zendel, Bailey, Clarke, Cotterill, Dalton, Hasibi, Sanderson, Craswell: *Principles and Guidelines for the Use of LLM Judges*. ICTIR 2025