Prompt Lottery in LLM-as-a-Judge: Measuring and Mitigating Sensitivity to Syntactic Rephrasing

Ilija Subasic Brett Levert Andrew Rabinovich
Sajjad Pakdamansavoji Zhao Chen

Upwork

Rewording a judge prompt without changing its meaning drops agreement from κ 0.89 to 0.24 and can flip model rankings.

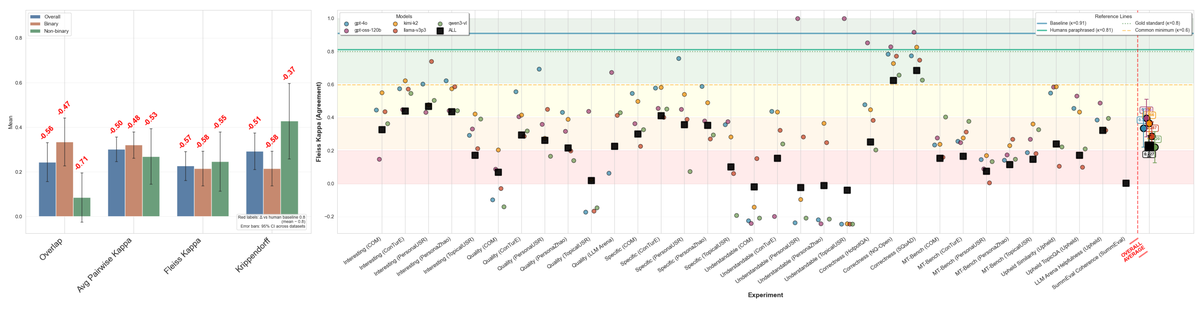
Motivation

- **LLM-as-a-Judge (LLMaJ)** is the default way to evaluate LLMs.
- It assumes a single judge prompt is a **stable instrument**.
- Prior work changes prompts **on purpose** (rubrics, roles, CoT). We ask: what if we only **rephrase** them, keeping the meaning?

Setup

- **33 configurations**: datasets $\times$ judge families (QA correctness, groundedness, dialogue, summarization, ...).
- Each scored by **5 back-models** $\times$ **5 prompt variants** = 25 judges: Kimi-K2.5, GPT-oss-120B, GPT-4o, Llama 3.3, Qwen3-VL.
- 4 paraphrases per prompt, **validated by humans** and by three LLMs as meaning-preserving.
- **User study**: 10 annotators, 650 items, the same paraphrased instructions.

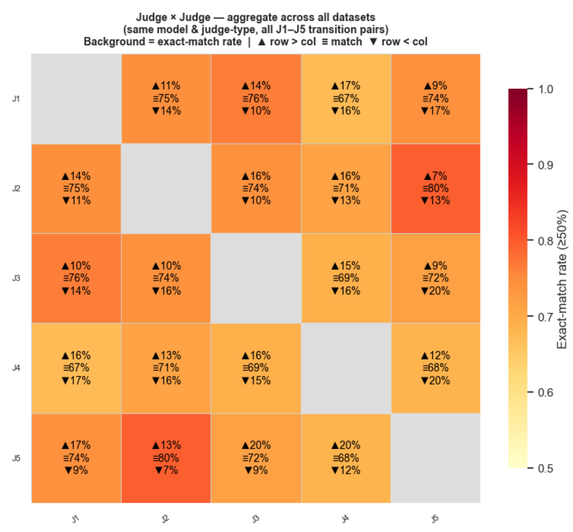
Finding 1: agreement collapses



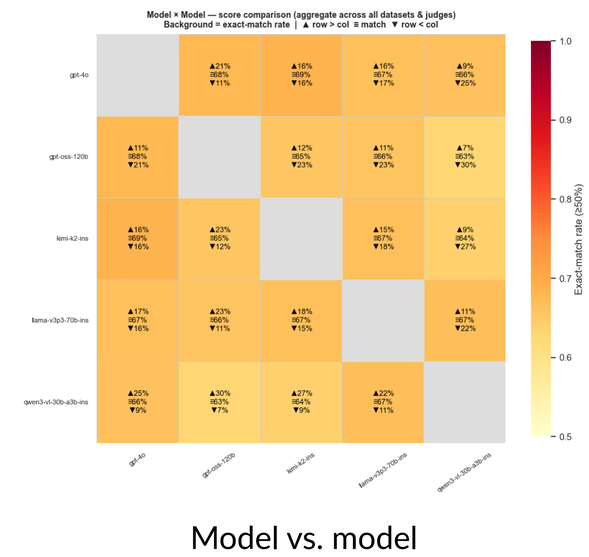
Left: agreement across variants. Right: Fleiss' κ and its drop per judge family.

- Fleiss' κ falls from **0.89 to 0.24**; humans given the same rewordings drop only **0.09–0.11**.
- All but two settings fall below the accepted $\kappa = 0.61$ bar.
- **Binary paradox**: binary judges agree more by chance, but chance-corrected κ is *lower* (0.21 vs. 0.28).
- Objective judges are steadier (κ 0.47–0.69) than subjective ones (0.26–0.39).

Finding 2: no prompt wins; it is a lottery



Prompt vs. prompt



Model vs. model

- Prompts are **0.92–0.95 similar** (sBERT), yet κ ranges from **0.0 to 0.7**; similarity does not predict agreement ($r = -0.11$).
- The best variant changes from task to task.

Finding 3: rankings can flip

Task	Mean τ	Top-1 flips	Flip %
MTBench (DailyDialog)	1.00	0/4	0%
MTBench (Persona)	0.80	2/4	10%
QA correctness (SQuAD v2)	**0.18**	3/4	**43%**
Quality (SummEval)	0.64	2/4	40%
G-Eval (Topical-Chat)	0.83	1/4	10%

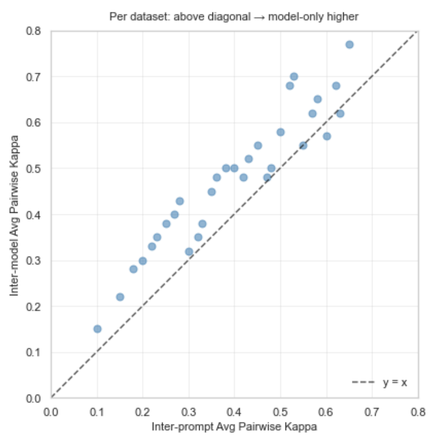
Ranking 4 models under 5 paraphrased prompts with a fixed judge backbone.

Mitigation: aggregate over prompts

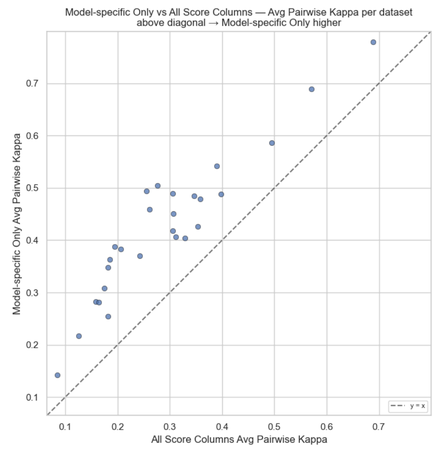
Data	Maj.	Avg.	Plur.
Upheld	0.47	0.54	**0.63**
Topical-Chat	0.44	0.55	**0.62**
LLM-Arena	0.44	0.54	**0.55**

Pearson r with human scores.

- **Plurality** vote over prompts aligns best with humans.
- Varying **prompts** beats varying models for reliability.
- Aggregate only non-binary tasks: **~60%** lower cost.

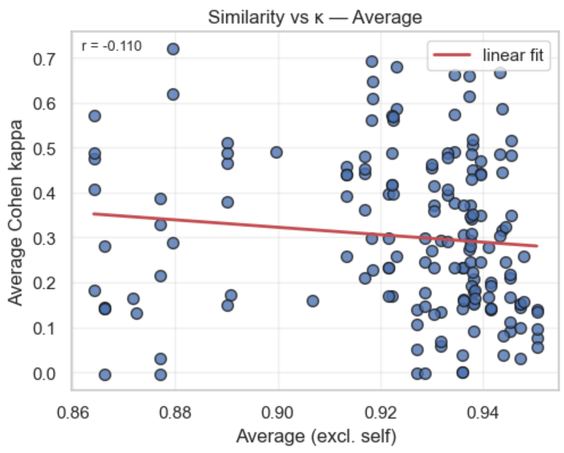


Inter-model vs. inter-prompt

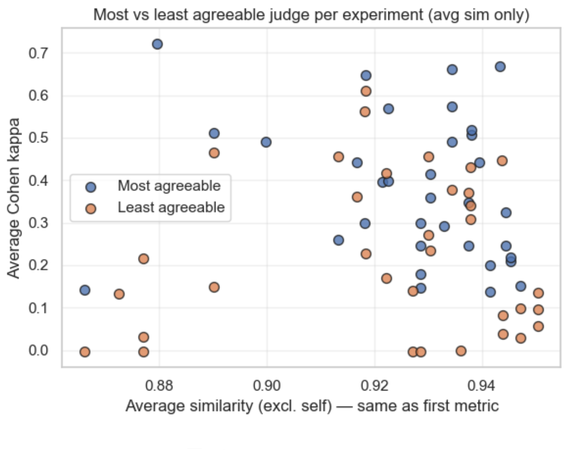


Inter-model vs. hybrid

Similar wording, different verdicts



All cases



Extreme cases

Semantic similarity of prompt pairs (sBERT) vs. judge agreement (κ): no useful relationship.

Takeaways

- Treat the judge prompt as a **random variable**, not a constant.
- Report agreement across paraphrases, and **aggregate** when tasks are open-ended.