LLMs Judging LLMs: *A Simplex Perspective*

Patrick Vossler[1]* · Fan Xia[1]* · Yifan Mai[2] · Adarsh Subbaswamy[3] · Jean Feng[1]

[1]University of California, San Francisco [2]Stanford University [3]University of Maryland, Baltimore *Equal contribution

SCAN FOR PAPER

Can LLMs judge LLMs? It depends. By visualizing judges and candidates on a simplex, we find 2-level rankings are identifiable, but 3+ levels often aren't due to epistemic uncertainty about judge quality.

01 — MOTIVATION

Sampling noise isn't the only uncertainty that matters.

- LLM judges power modern evals.
- Judges exhibit biases, e.g., position, self-preference, verbosity, halo.
- Existing methods only quantify sampling noise (aleatoric). This makes implicit assumptions about judge behavior, which can be risky when epistemic uncertainty is high!

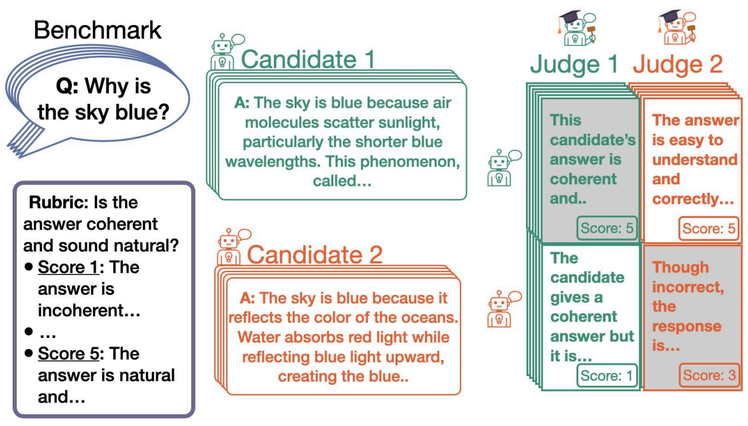

Fig. 1: LLM judges score each candidate's answer against a rubric. Shaded boxes mark self-judging (excluded to control bias).

02 — GEOMETRIC FRAMEWORK

Both judges and candidates live on a simplex.

- A judge's confusion matrix places M vertices on the probability simplex, one per true score.
- A candidate sits inside, at the weighted centroid of those vertices, weighted by how often it earns each true score.
- Ranking candidates becomes a question about positions and subtriangle areas.

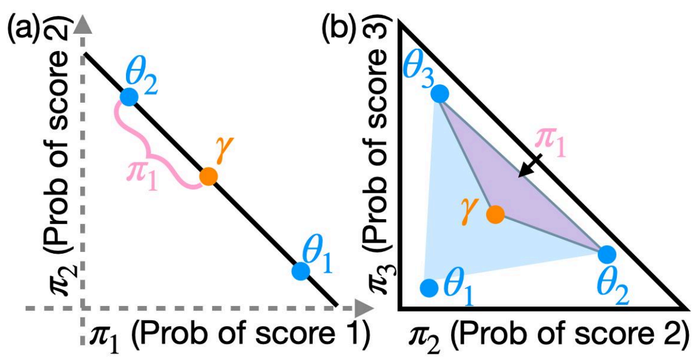

Fig. 2: Both judges and candidates live on the probability simplex.

$\theta_{j,m}$: judge j's vertex for true score m
γ_k: candidate k's point (its score distribution as the judge sees it)
$\pi_{k,m}$: candidate k's true-score prevalence (barycentric coordinate)

03 — STATISTICAL LIMITS

LLM judges are more reliable at 2-level scoring than 3+ levels.

For simplicity, suppose each judge scores candidates consistently, treating every answer the same way given a fixed true score. When can we recover the true ranking from judge-assigned scores alone?

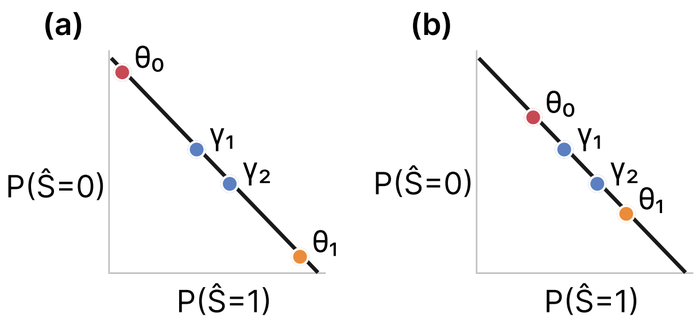

Fig. 3: Different θ placements (a vs b) can explain the same observed γs. The candidate ranking is preserved regardless.

THEOREM 1 · BINARY IS IDENTIFIABLE

For 2-level scoring, rankings are identifiable from judge scores if either:
(i) one judge scores all candidates consistently and is better than random
(ii) at least 4 candidates and 2 judges each score consistently (possibly differing when self-judging), all better than random

What changes at three levels?

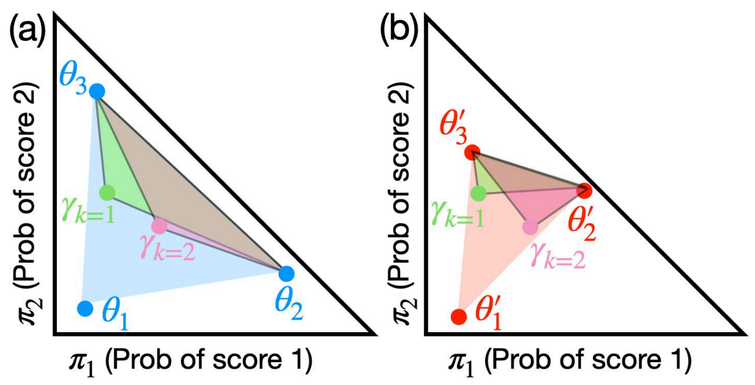

Fig. 4: Same candidates (γ), two judge configurations (blue θ, red θ′). Both fit the data, yet panel (a) gives $\pi_{1,1} > \pi_{2,1}$ (candidate 1 scores 1 more often) and panel (b) flips it.

THEOREM 2 · MULTI-LEVEL IS NOT

Under consistent scoring and non-adversariality, for M ≥ 3 levels some candidates' rankings cannot be recovered from judge-assigned data.

04 — METHOD

Priors on judge quality capture epistemic uncertainty.

Identifiability depends on the data. Different θs can fit the same γs yet reorder candidates (Fig. 4). Priors on judge quality let us ask:

- How robust are rankings if judges are inconsistent?
- How robust are rankings if judges can't distinguish between scores?

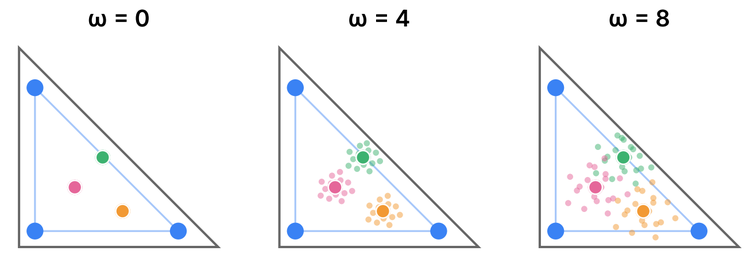

Fig. 5: Relaxing the consistency prior (ω = 0 to 8) lets candidates drift. Stable rankings under drift are robust.

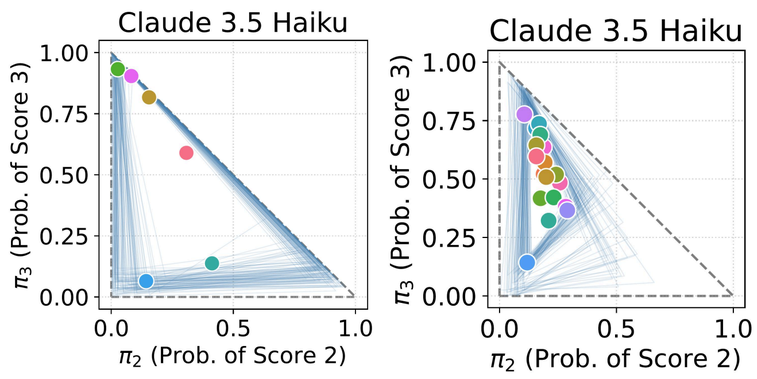

Fig. 6: Spread candidates (MTBench, left) stay ranked; clustered ones (Omni-MATH, right) flip. Be wary.

05 — RESULTS

Modeling epistemic uncertainty yields honest coverage.

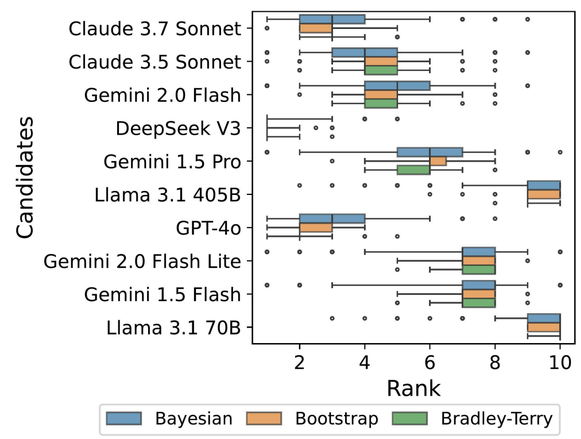

Fig. 7: Top 10 GPQA candidates by method. Only Bayesian intervals reliably cover truth at 95%.

COVERAGE @ 95%

Benchmark	Bayesian	Bootstrap	Bradley-Terry
GPQA	**0.89**	0.56	0.39
MMLU Pro	**1.00**	0.47	0.47
Omni-MATH	**0.74**	0.37	0.37
SummEval	**0.92**	0.73	0.77
MTBench	1.00	1.00	1.00

patrick.vossler@ucsf.edu • jean.feng@ucsf.edu Supported by PCORI® Award ME-2022C1-25619. Thanks to G. Pennello and N. Petrick (FDA/CDRH) for helpful discussions.