

Discarding Answer Ranks Changes the Case for Mixing Biomedical LLMs

Odysseus Kaloidas¹

¹ University of California, Berkeley (berkeley.edu)

27 Sept 2026 (modified: 27 Sept 2026)

NeurIPS 2026 Workshop AgenticLS Submission AgenticLS, Reviewers, Authors

Revisions CC BY 4.0

Keywords: biomedical large language models, model ensembles, answer aggregation, self-consistency, ranked voting, model diversity, evaluation

TL;DR: Discarding answer ranks can make mixed-model ensembles look better by weakening the single-model baseline, without producing better biomedical decisions.

Abstract:
Should a biomedical reasoning system spend eight calls on one model or several? An aggregation rule can make mixing look better by worsening the one-model reference. A 288-question, four-model study compares mixing with repeated Gemma calls. Counting every shortlisted answer rather than only the first shifts the mixed-minus-Gemma accuracy gap by 9.72 percentage points (95% confidence interval [5.29, 14.21]). Using ranks to resolve ties accounts for 7.62 points; the remaining contrast is unresolved. A favorable rule contrast need not imply a better policy. In a separate prespecified 720-question MedXpertQA test, mixing Gemma and Llama3.3-70B loses 8.95 points [−11.16, −6.75] against development-selected Gemma first-choice voting. The mixture already has lower per-call accuracy. On 256 MMLU-Pro Health questions, mixing improves inclusion-vote accuracy by 5.00 points, yet remains far below the selected first-choice policy. On Health, first-choice voting also exceeds a ceiling for selectors restricted to listed answers that use only which calls mention them and favor no answer label. These results show how discarding ranks changes the apparent benefit of mixing, without isolating quality-independent diversity.

Email Sharing: We authorize the sharing of all author emails with Program Chairs.
Data Release: We authorize the release of our submission and author names to the public in the event of acceptance.
Submission Number: 127

Filter by reply type...

Filter by author...

Search keywords...

Sort: Newest First

Everyone

Program Chairs

Submission127 Authors

1 / 1 reply shown



Paper Decision

Decision by Program Chairs

 30 Sept 2026, 00:29 (modified: 30 Sept 2026, 07:34)

 Program Chairs, Authors

 Revisions

Decision: Accept (Poster)

Comment:

The fixed-response rule analysis and separately frozen policy test provide useful evidence about evaluation artifacts in biomedical model ensembles. Quality mismatch, lossy inclusion rules and restricted selector inputs limit generalization beyond the tested policies.

Add:

Official Comment

About OpenReview

FAQ

Sponsors

All Venues

Contact

Hosting a Venue

Terms of Use / Privacy Policy

Donate

OpenReview is a long-term project to advance science through improved peer review with legal nonprofit status. We gratefully acknowledge the support of the OpenReview Sponsors. © 2026 OpenReview