MIT EECS
Electrical Engineering | Computer Science | Artificial Intelligence + Decision-making

MIT-IBM Watson AI Lab

Dropping Just a Handful of Preferences Can Change Top LLM Rankings

Jenny Huang*, Yunyi Shen*, Dennis Wei, Tamara Broderick

*equal contribution

Introduction

"Can dropping just a few preferences change which models top the LLM leaderboards?"

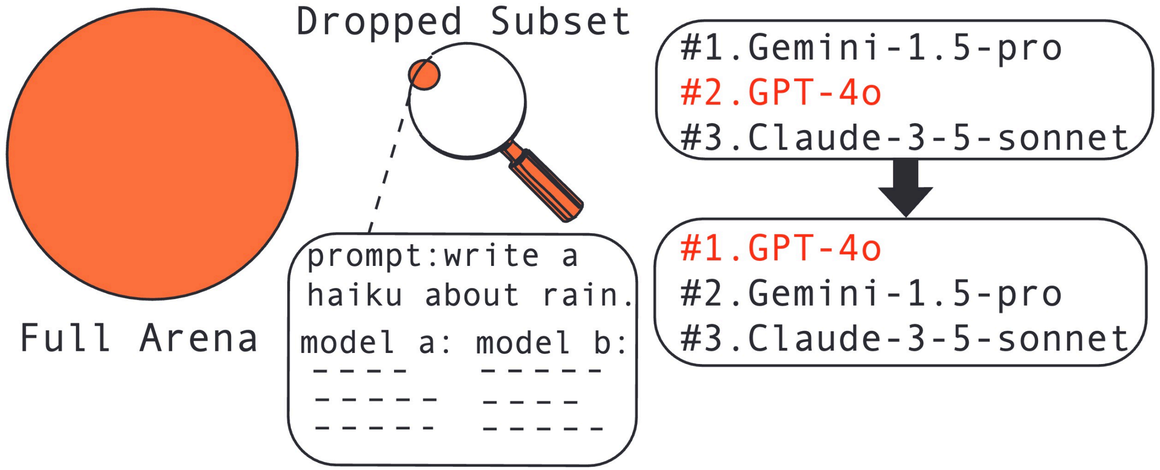
It turns out... yes! Dropping just **0.003%** of matches can change the top-ranked model.

Open evaluation platforms, like Chatbot Arena, have become very popular for ranking LLMs [Singh et al., 2025; Chiang et al., 2024].

Pairwise Preferences → Leaderboard Ranking

Bradley—Terry model

*We might worry about the **generalizability** of leaderboard rankings if they hinge on a very small handful of pairwise preferences.*



Performing a brute force check is computationally infeasible.

we develop a fast method to check the data-dropping robustness of LLM ranking systems for large databases.

Background

Existing Work:

Worst-case data-dropping approximations offer a fast way to test model sensitivity to dropping small subsets of data [Huang et al., 2024, Kuschnig et al., 2021, Broderick et al., 2020].

These methods rely on data-dropping approximations to determine whether there exists a subset of data that can be dropped to change the conclusion of a data analysis.

Examples: AMIP, Additive-1sN, NewApprox, FH-Gurobi

[Broderick et al., 2020, Huang et al., 2024, Freund and Hopkins, 2023]

Our Contribution:

Building on AMIP, we create a **scalable algorithm** to check the data-dropping robustness of LLM **ranking systems**.

image by gpt4o.

Method

The data-analysis conclusion is the set of LLMs ranked in the top-k.

Existing work in this space [Broderick et al., 2020] suggest that the robustness of a finding is questionable if removing <1% of the data is enough to change the conclusions of the data analysis.

Goal: Determine whether there exists a small subset of preferences that can be dropped to change the top-k rankings.

Key idea #1: Checking Top-k robustness boils down to checking robustness for all players in the top-k against all others.

-The data analysis conclusion now becomes the relative ordering of 2 given players.

Key idea #2: Given 2 players, checking robustness of their relative ordering is the same as checking robustness of the *difference in their scores.*

-The data-analysis conclusion now becomes the sign of this difference (we can apply existing approximations).

Our robustness check is definitive!

LLM Rankings Are Very Sensitive

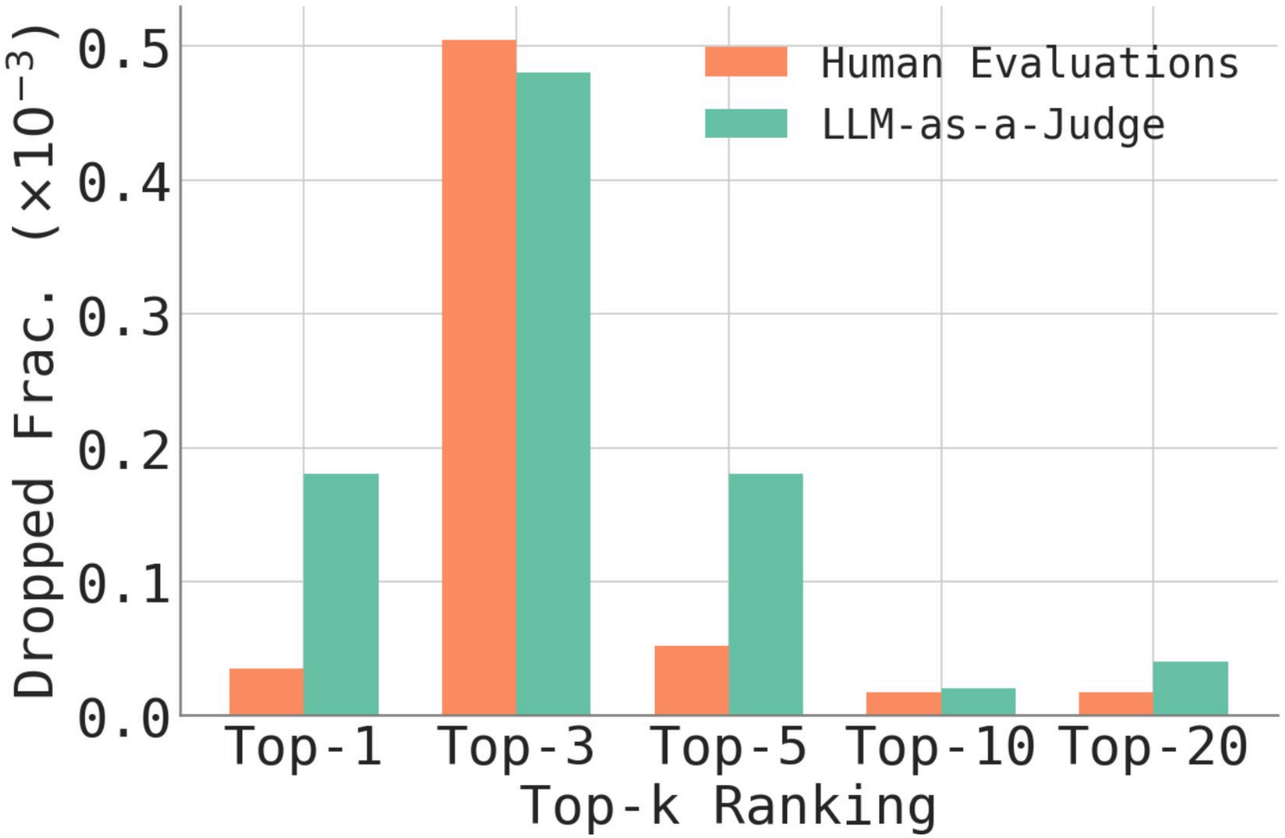
Sensitivity holds cross several leaderboards!

Arena	Evaluator (Judge)	Number Dropped	Percentage Dropped
Chatbot Arena	Human	2 out of 57477	0.00348%
Vision Arena	Human	28 out of 29845	0.0938%
NBA Games	NA	17 out of 109892	0.0155%
Chatbot Arena	LLM	9 out of 49938	0.0180%
Webdev Arena	Human	18 out of 10501	0.171%
Search Arena	Human	61 out of 24469	0.253%
MT-bench	LLM	40 out of 2400	1.67%
ATP Tennis	NA	6 out of 278	2.16%
MT-bench	Human	92 out of 3355	2.74%

Chatbot Arena	MT-Bench
> user-submitted questions	> high-quality curated questions
> crowdsource annotators	> expert annotators

Humans vs. LLM-as-a-Judge

Human-evaluated and AI-evaluated systems appear about as non-robust.



Each bar shows the percentage of evaluations in Chatbot Arena that are sufficient to be dropped in order to demote one of the models inside to outside of the top-k rankings.

Inspecting Dropped Evaluations

Our method can identify dropped prompts and responses for further investigation.

Prompt: You are provided with a legal description that seems to make sense when you initially read it but once you arrive at the site it no longer does. What type of ambiguity have you encountered?

LLM 1: When a legal description appears clear initially but becomes confusing at the site, it indicates an issue of latent ambiguity…

LLM 2: The type of ambiguity encountered is likely a latent ambiguity, where the description seems clear on paper but is problematic when applied to the physical location…

In Conclusion,

Leaderboard rankings can be very sensitive to dropping just a handful of preferences.

We encourage researchers and users of leaderboards and benchmark contests to run our method to investigate the fragility of crowdsourced LLM evaluation platforms before publishing results.

Reach Out

Paper

Code