Analyzing AI Evaluation Benchmarks Through Information Retrieval and Network Science

Gaia Simeoni, Michael Soprano, Riccardo Lunardi, Kevin Roitero, Stefano Mizzaro

University of Udine, Italy

We investigate how well easy/hard questions can identify the least/most effective LLMs

MOTIVATION

IR has always been based on **robust evaluation protocols**

LLMs effectiveness is primarily assessed through standard **benchmarks**

These benchmarks are **rarely scrutinized** with the rigorous protocol of IR test collections

We investigate whether **IR evaluation techniques** can be applied to evaluate **LLMs**

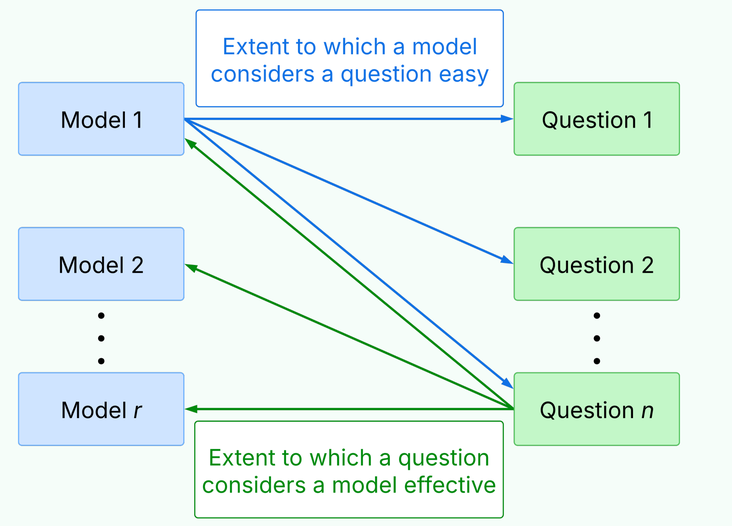
METHODOLOGY

Network-based approach to evaluate LLMs benchmarks

Bipartite graph connecting 34 LLMs and questions from 7 popular benchmarks (MMLU, ARC-C,…)

Kleinberg's HITS algorithm to calculate two key scores:

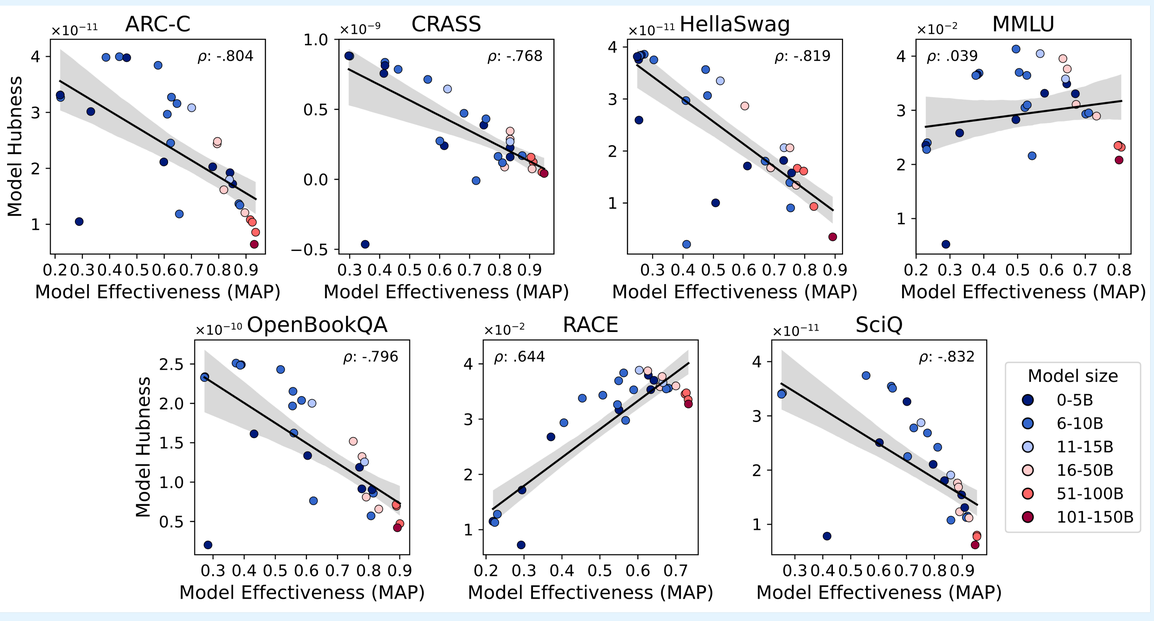
- **Model Hubness**: How well a model tends to perform on easy questions
- **Question Hubness**: How well a question can identify the most effective models



Bipartite graph representing model hubness (blue) and question hubness (green) between *r* models and *n* questions.

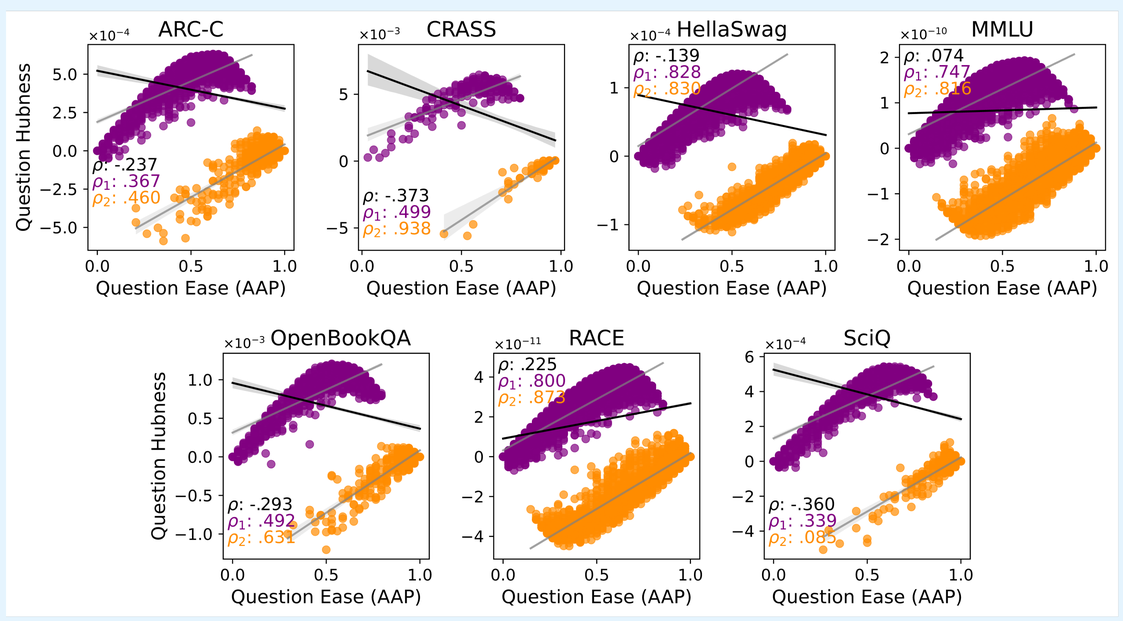
KEY FINDINGS

The **most effective models** rely less on trivial questions, whereas **weaker models** rely more on **superficial cues** typical of easy questions



Each dot is an LLM. The horizontal axis shows model effectiveness, while the vertical axis shows reliance on easy questions. The downward trend shows that the most effective models (bottom-right) rely less on trivial questions.

Easy questions strongly influence evaluations. Their clustering reveals a **Simpson's paradox**, with their impact increasing within specific groups



Each dot is a question. The horizontal axis shows question ease, and the vertical axis shows its influence on model rankings. The questions are split into two groups: within each group, the upward slope suggests that easier questions have a stronger impact.

TAKEAWAYS

Model **rankings** are strongly influenced by **subsets of easy questions**

Evaluations risk **overemphasizing basic performance** over **true LLMs capabilities**

IR techniques can help evaluate benchmark structure

CONCLUSIONS

Found patterns **diverge from earlier IR studies** (e.g., negative correlation and question clustering)

Preliminary checks confirm these trends are **not artifacts**

Future work will investigate the underlying causes and extend the analysis to **open-ended generation tasks**