EIP: Weighted Ranking of LLMs by Quantifying Question Difficulty

Xingjian Hu[1] Ziqian Zhang[1] Yue Huang[2] Kai Zhang[1] Ruoxi Chen[3] Yixin Liu[1]
Qingsong Wen[4] Kaidi Xu[5] Xiangliang Zhang[2] Neil Z. Gong[6] Lichao Sun[1]

[1]Lehigh University [2]University of Notre Dame [3]Zhejiang Wanli University
[4]Squirrel Ai Learning [5]City University of Hong Kong [6]Duke University

Introduction

Large language models are now compared on broad benchmark suites, but most leaderboards still collapse evaluation to plain accuracy. This implicitly assigns identical value to all questions and obscures whether a model solves genuinely difficult items or mostly accumulates easy wins.

EIP addresses this by jointly estimating **question difficulty** and **model competency** from observed successes and failures, rather than treating difficulty as fixed or manually labeled.

Motivation

Flat accuracy is too coarse to characterize model capability because it ignores *which* questions are being solved. More importantly, if we want a scientific and falsifiable notion of question difficulty, that notion should not rely on ad hoc labeling; it should emerge objectively from observable model-question interactions.

- In the controlled simulation, $M1$ and $M2$ have identical accuracy, yet $M1$ solves more hard questions and should rank higher.
- Likewise, $M4$ and $M5$ are close in accuracy, but $M4$ performs better on medium-difficulty questions.
- Accuracy misses these distinctions because it counts a routine item and a genuinely challenging item as equally informative evidence.

Why a bidirectional model?

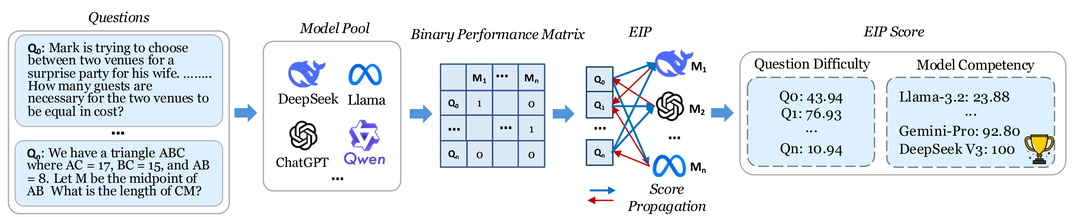
EIP treats difficulty and competency as *mutually defined* quantities—neither can be assessed in isolation. A question is not "hard" simply because any single model fails on it; it is hard because *sufficiently many competent models* still fail on it. Symmetrically, a model is not "strong" merely because of an external label or any individual judgment; it is strong because it *solves questions that other strong models cannot*. This circular dependency is not a flaw but the core design: difficulty emerges from the empirical pattern of failures by competent models, and competency emerges from success on questions that resist those same models. Both are internal properties of the response graph, jointly determined through bidirectional propagation rather than assigned by external annotation.

Scientific interpretation

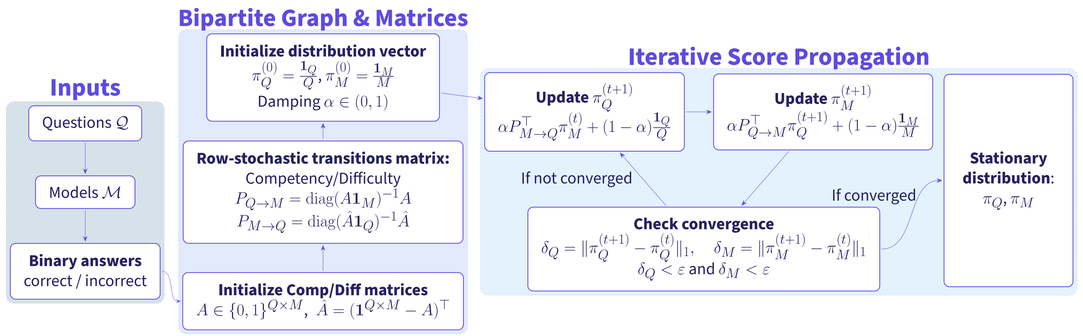
The resulting framework is objective, falsifiable, and reproducible: if a question is truly difficult, that claim must be supported by the empirical pattern that many models, including strong ones, fail on it. Conversely, if competent models solve a question consistently, the graph should force its estimated difficulty downward.

Contributions

Empirical Interaction Propagation: Pipeline



Detailed Process



Methodology: Score Interpretation and Derivation

EIP derives π_M and π_Q from a directed bipartite graph with two complementary flows.

1. Construct the interaction graph

Let $A \in \{0,1\}^{Q\times M}$ record correct responses and let $\hat{A} = (\mathbf{1} - A)^\top$ record failures. This creates question-to-model edges for successes and model-to-question edges for failures.

2. Normalize local evidence

For each question, $S(q)$ counts how many models solved it; for each model, $F(m)$ counts how many questions it failed. These counts normalize outgoing evidence so a rare success on a hard question carries more weight than a success on an easy one.

3. Define bidirectional propagation

Question difficulty flows toward successful models through $P_{Q\to M}$, while model competency flows toward failed questions through $P_{M\to Q}$. This makes competency and difficulty mutually reinforcing rather than independently fit.

4. Add damping for a unique stationary solution

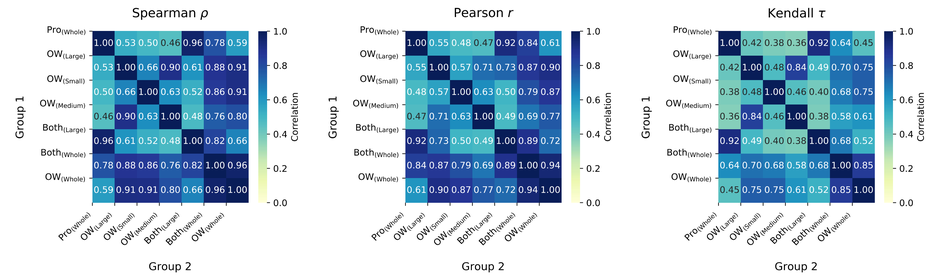
A teleportation term with $\alpha \in (0,1)$ removes periodicity from the bipartite walk and guarantees convergence to a unique stationary distribution over models and questions.

5. Interpret the final scores

A model receives a high π_M when it consistently solves questions that remain difficult for the pool. A question receives a high π_Q when even competent models fail on it. This gives a difficulty-aware ranking instead of a flat count of correct answers.

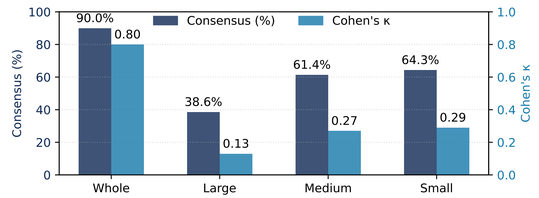
Correlation Across Model Group Combinations

Difficulty rankings estimated from diverse or open-weight model pools remain strongly aligned with the full pool, showing that EIP captures a stable global difficulty landscape.



Alignment with Human Evaluation

EIP achieves the strongest alignment with human judgments on question difficulty. Across the evaluator pool, the whole mixed-scale model set yields the clearest agreement advantage over Simple Rank and IRT baselines.



Method	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	$\cdots$	V_{20}	Consensus (%)
Simple Rank	41.4	54.3	50.0	60.0	51.4	54.3	54.3	50.0	50.0	$\cdots$	52.9	62.9
1PL-IRT	37.1	44.3	40.0	52.9	51.4	44.3	45.7	45.7	48.6	$\cdots$	41.4	50.0
2PL-IRT	35.7	45.7	41.4	54.3	52.9	42.9	44.3	47.1	47.1	$\cdots$	42.9	51.4
Multi-IRT	**50.0**	50.0	48.6	54.3	48.6	57.1	47.1	47.1	44.3	$\cdots$	48.6	52.9
EIP	37.1	**70.0**	**62.9**	**71.4**	**67.1**	**61.4**	**74.3**	**64.3**	**57.1**	$\cdots$	**62.9**	**90.0**

Convergence Speed and Scalability

EIP converges in milliseconds on the full benchmark and keeps a constant iteration count even as the interaction matrix scales to 10^9-level entries.

Method	Time (s)
EIP	0.00597
1PL IRT	1782.75
2PL IRT	3787.03
Multi-IRT	18.76

Q	M	$Q \times M$	Iter.	Total	/ Iter.
1,000,000	2,000	2.00×10^9	9	5.28	0.5867
1,000,000	500	0.50×10^9	9	2.18	0.2422
500,000	500	0.25×10^9	9	1.13	0.1256
500,000	250	0.125×10^9	9	0.64	0.0711
250,000	500	0.125×10^9	9	0.52	0.0578
250,000	250	0.0625×10^9	9	0.30	0.0333