Karlsruhe Institute of Technology

aifb

Towards Scalable Reliable Automated Evaluation via LLMs

Let LLMs judge each other — Elo-ranked, expert-level evaluations at a fraction of the cost

Bertil Braun, Martin Forell

Why This Matters - LLM outputs are hard to score:

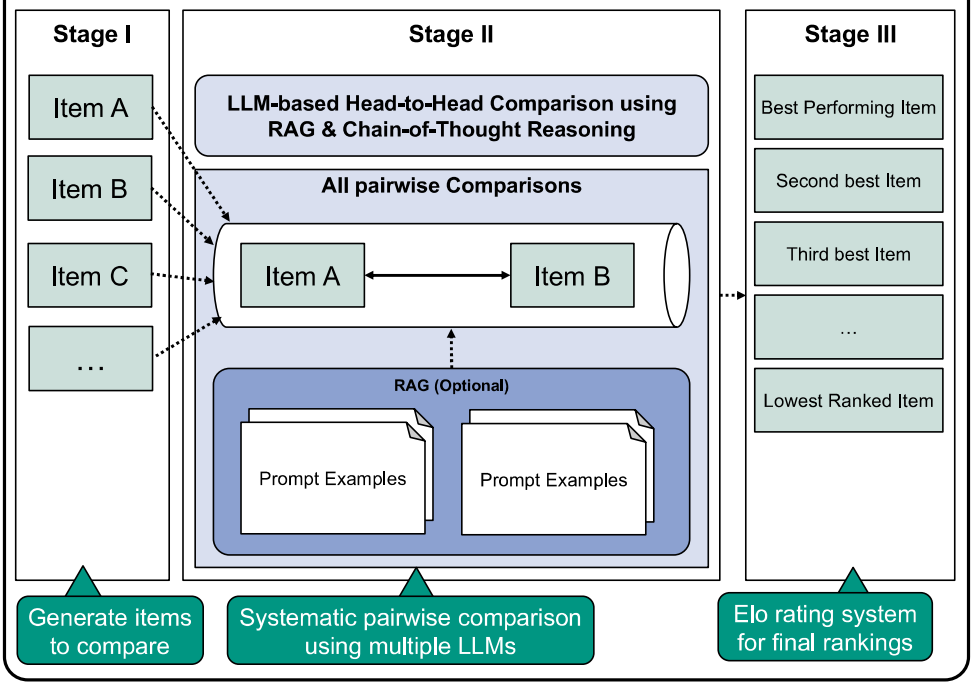
- **Resource-intensive:** Human evaluation doesn't scale
- **Inconsistent:** Traditional metrics miss nuanced quality
- **Biased:** Single-LLM judgments suffer from positional/verbosity biases

We show that a *crowd* of LLMs, voting pair-wise and aggregated with Elo, reproduce expert rankings while slashing manual effort.

Key Contributions: Multi-LLM Pairwise Comparison + Elo Rating

- **Multiple LLMs** evaluate pairs **bidirectionally**
- **Elo system aggregates** judgments into stable, interpretable rankings (Δ100 pts ≈ 64 % win-prob)
- Adjustable **agreement thresholds** (majority → consensus)

Solution: Multi-LLM Pairwise Evaluation with Elo Rankings



Method

- **Prompt engineering:** Role prompt → RAG few-shots → Chain-of-Thought → structured-JSON verdict
- **Bias shields:** Every pair is judged *both* A vs B and B vs A; five diverse LLMs vote, neutralising position and stylistic bias
- **Agreement logic:** A tunable threshold (1.0–0.5) decides whether conflicting votes become a draw or an Elo update—majority (0.5) works best
- **Elo system:** Updated after each decision: $R_{new} = R + K(Score - E)$

Benefits over other Approaches

- **Pairwise Comparisons** → Eliminates scoring subjectivity
- **Multiple LLMs & Bidirectional Evaluation** → Reduce individual model and positional biases
- **Agreement Thresholds** → Aggregate multiple LLM judgments
- **Elo System** → Produces stable, interpretable rankings

Results at a Glance

- **Strong correlation** with expert rankings using Multi-LLM approach (Spearman's ρ = 0,83).
- Single-LLM baseline shows a **comparable performance** (ρ = 0,85), but the Multi-LLM setup is more robust against noise from conflicting judgments.
- A **simple majority threshold (0,5)** proved most effective for aggregating evaluation results.
- Framework was validated with **20 domain experts** who ranked generated competency profiles.

Key Finding:

Multiple LLMs + Elo rankings achieve expert-level assessment quality while **maintaining scalability**

SCAN ME