Towards Scalable Reliable Automated Evaluation via LLMs

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:

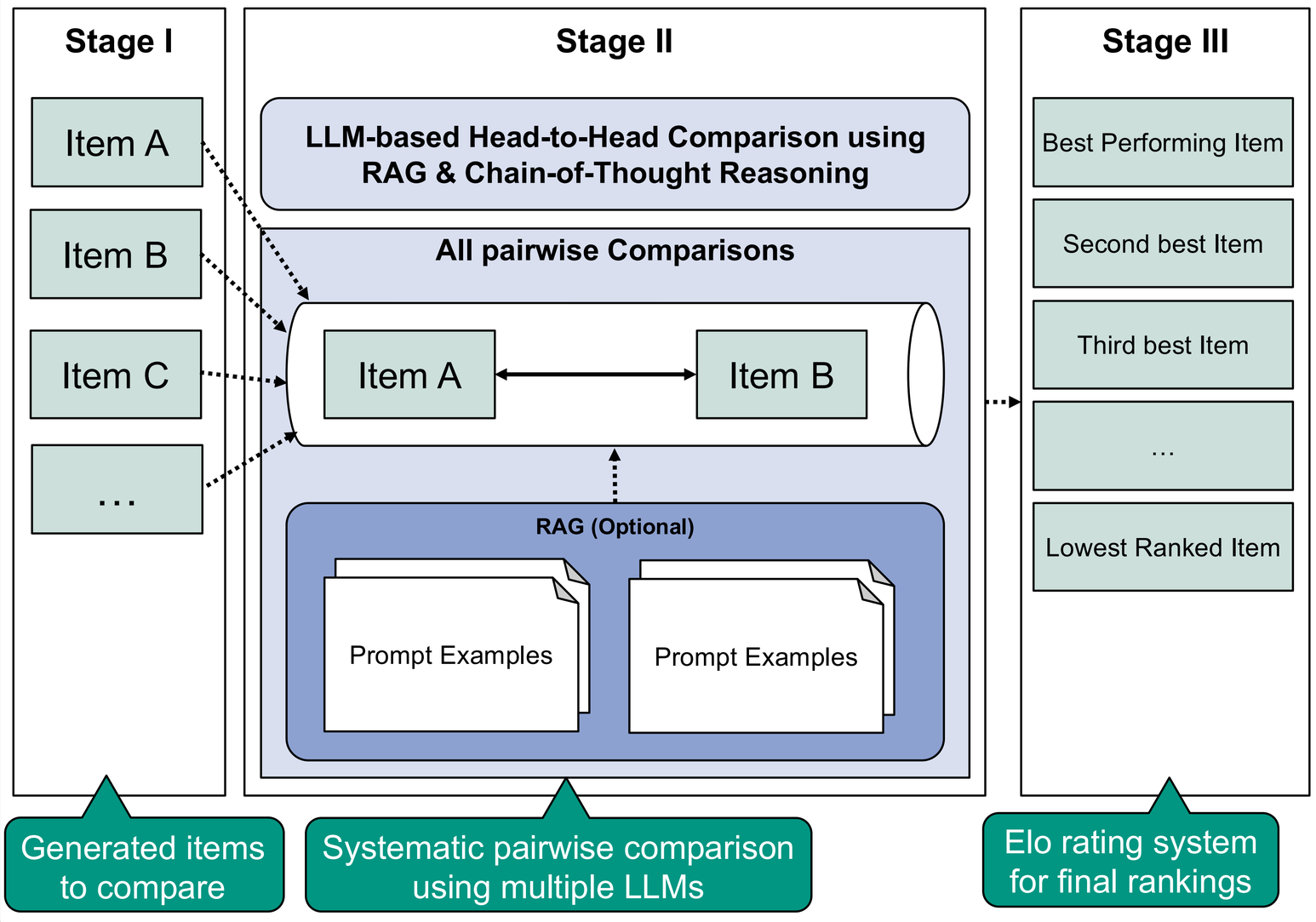
- **Resource-intensive:** Human evaluation doesn't scale
- **Inconsistent:** Traditional metrics miss nuanced quality
- **Biased:** Single-LLM judgments suffer from positional/verbosity biases

We show that a *crowd* of LLMs, voting pair-wise and aggregated with Elo, reproduce expert rankings while reducing 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:**
  Bidirectional evaluation (A vs B, B vs. A) + 5 diverse LLMs
- **Agreement:**
  Threshold (1.0–0.5) decides draw vs. Elo update; majority (0.5) best
- **Elo system:**
  Updated after each decision:
  $R_{new} = R + K(Score – E)$,
  E = expected win probability

Limitations

- **Computational overhead:** $O(n^2)$ comparisons become costly at scale
- **Draw sensitivity:** High thresholds (>0.75) produce excessive draws
- **Similar items:** LLMs struggle with subtle quality differences
- **Domain dependency:** Requires task-specific prompt engineering

Results at a Glance

- **Strong correlation** with expert rankings (Spearman's ρ): Multi-LLM = 0,83 vs. Single-LLM = 0,85
- **Multi-LLM** approach demonstrates improved robustness to conflicting judgments.
- **Majority threshold (0,5)** most effective for result aggregation
- Validation: **20 domain experts** ranked generated competency profiles

Key Finding:
Multiple LLMs + Elo rankings achieve expert-level assessment quality while **maintaining scalability.**