Elo Uncovered: Robustness and Best Practices in Language Model Evaluation

Presenter

Divahar Sivanesan

The Bernoulli Analogy in LLM Evaluation

- Pairwise Comparisons as Bernoulli Trials
  - Each comparison between models A and B is like a Bernoulli experiment.
  - Two possible outcomes: Model A wins (success) or Model B wins (failure).
- Bernoulli Process in Evaluations
  - Sequence of independent Bernoulli trials.
  - Win probability for Model A is represented as P(A[win])
- Mapping Human Feedback
  - Simulate human preferences using Bernoulli random variables.
  - For each trial:
    - Draw a sample from X using P(A[win])
    - If X = 1, preference for Model A.
    - If X = 0, preference for Model B.

Extending to Multiple Models

- Comparing Multiple Models
  - For n models, total unique pairs: n(n - 1) / 2.
  - Each pair (Model A, Model B) undergoes Bernoulli trials.
- Formulating Pairwise Comparisons
  - Number of pairs increases rapidly with more models.
  - Each comparison helps build the overall ranking.

Binomial Distribution in Model Evaluations

- Multiple evaluations between two models follow a binomial distribution.
- Probability of a model being preferred k times out of N trials:
  - $P(k; N, p) = C(N, k) * p^k * (1 - p)^{(N - k)}$
- Generating Synthetic Data
  - Use binomial distribution to simulate human feedback.
  - Control win probabilities to test Elo rating robustness.

The Elo Rating System and Its Axioms

- Elo Rating Formula
  - Rating Update Equation: $R'_A = R_A + K(S_A - E_A)$
  - Expected Score Calculation: $E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}}$
- Key Components
  - K-factor (K): Controls the sensitivity of rating updates.
  - Actual Score (S): 1 for a win, 0 for a loss.
  - Expected Score (E): Estimated probability of winning against the opponent.
- Fundamental Axioms
  - Transitivity Axiom
    - If Model A > Model B and Model B > Model C, then Model A > Model C.
  - Reliability Axiom
    - Elo scores should be robust to the order of match-ups and hyperparameter settings
    - Sensitivity to match-up ordering and parameters like K-factor should be minimal.

Investigating the Robustness of Elo Scores

- Objectives of Stress Tests
  - Assess if Elo scores uphold the Transitivity Axiom and Reliability Axiom in LLM evaluations.
  - Focus on key properties:
    - Ordering Insensitivity
    - Sensitivity to Hyperparameters (K-factor)
    - Preservation of Transitivity
  - Methodology
    - Conduct synthetic experiments simulating pairwise comparisons.
    - Use controlled win probabilities to create different scenarios.
    - Analyze the impact of parameters like K-factor and match-up ordering on Elo score stability.

Motivation

- The Challenge of Evaluating LLMs
  - Increasing reliance on human feedback for model evaluation.
  - Need for robust ranking mechanisms to compare models.
- Elo Rating System in LLM Evaluation
  - Widely adopted from competitive games for ranking.
  - Assumed to provide reliable model comparisons.
- Research Question
  - How robust and reliable are Elo scores when used to evaluate LLMs with human feedback?

Impact of Match-Up Ordering on Elo Ratings

- Experimental Setup
  - Generated a baseline of 1,000 match outcomes between Models A and B.
  - Created multiple permutations (Nperms) by reshuffling the sequence of matches.
  - Elo ratings updated after each match, starting from initial ratings.
- Key Findings
  - Order Sensitivity
    - Elo ratings are sensitive to the sequence of match-ups.
    - Significant instability when win probabilities are close to 0.5.
  - Stabilization with Increased Permutations
    - Increasing Nperms to over 100 stabilizes Elo ratings
    - Ratings align closely with true performance differences.

Sensitivity to Hyperparameters (K-factor)

- Experimental Setup
  - K-factor Values Tested: 1, 8, 16, 32, 64
  - Configurations:
    - Number of Games (Ngames) = 1,000
    - Number of Permutations (Nperms) = {1, 100}
  - Evaluation Metrics:
    - Average Elo scores for Model A ($\bar{}$SA) and Model B ($\bar{}$SB)
    - Difference in Elo scores ($\bar{}$SA − $\bar{}$SB)
- Key Findings
  - Instability at Low Nperms; Single permutation (Nperms = 1) showed high variability, especially at K = 1
  - Higher K-factors Enhance Stability; Increasing K-factor reduces rating fluctuations and faster convergence to true skill levels with higher K
- Takeaway:
  - Higher K-factors are beneficial for quickly identifying clear performance disparities
  - Lower K-factors help minimize rating fluctuations among closely matched models
  - Optimal K-factor selection depends on the specific evaluation context and desired responsiveness

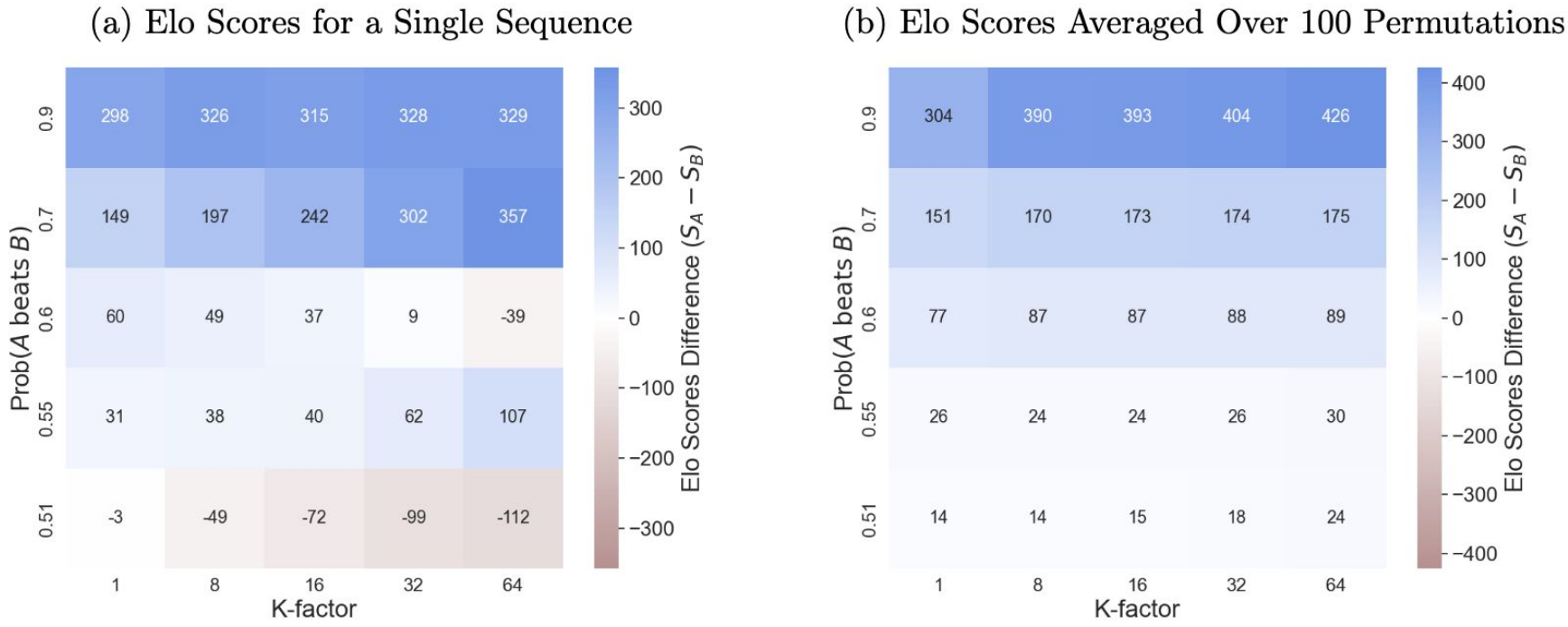


Figure 3: Final Elo scores difference ($S_A - S_B$) as a function of K-factor and N_{perms}. Positive values reflect the expected ranking where Model A is superior to Model B, while negative values indicate a discrepancy, falsely suggesting that Model B has a higher Elo score than Model A. We compare between a single sequence of outcomes and averages over $N_{\text{perms}} = 100$ unique permutations.

Testing the Transitivity Axiom

- Definition of Transitivity:
  - If Model A > Model B and Model B > Model C, then Model A > Model C.
- Experimental Scenarios:
  - Scenario K:
    - Model A beats Model B (Pwin = 0.75)
    - Model B beats Model C (Pwin = 0.75)
  - Scenario R:
    - Model A beats Model B (Pwin = 0.75)
    - Model B beats Model C (Pwin = 0.51)
  - Scenario B:
    - Model A beats Model B (Pwin = 0.51)
    - Model B beats Model C (Pwin = 0.75)
  - Scenario N:
    - Model A beats Model B (Pwin = 0.54)
    - Model B beats Model C (Pwin = 0.51)

- Key Findings:
  - Transitivity Holds with Clear Win Rates
    - In Scenario K, Elo scores reflect expected model hierarchy.
  - Transitivity Fails with Similar Performances
    - In Scenarios R, B, and N, Elo scores may not preserve transitivity; rankings are clearly sensitive to K-factor and Nperms.
  - Impact of Hyperparameters
    - Higher K-factors (e.g., K = 16) yield more consistent rankings.
    - Lower K-factors (e.g., K = 1) may lead to inconsistencies.

Validation with Real-World Human Feedback

- Purpose of Validation
  - Test if findings from synthetic data generalize to real-world scenarios.
  - Assess the utility of the Elo rating system in practical LLM evaluations.
- Data Collection:
  - Human feedback from previous evaluations.
  - Models evaluated: Dolly-v2 and Flan families.
- Evaluation Dataset:
  - 400 prompts from diverse datasets:
  - SODA, CommonsenseQA, CommonGen, AdversarialQA.
- Configurations Tested:
  - Number of permutations (Nperms): 1 and 100.
  - K-factor values ranging from 1 to 36.

Key Findings

- Consistency with Synthetic Results:
  - Similar patterns observed regarding the impact of K-factor and Nperms.
- Stability Influenced by Win Rates:
  - Stable Elo ratings when there is a clear performance disparity.
  - Higher volatility when models have similar win rates.
- Transitivity Not Guaranteed:
  - Transitivity may not hold, especially among closely matched models.
  - Rankings sensitive to hyperparameter choices.

Empirical Guidelines for Robust Elo-based Evaluation

- Achieving Score Stability
  - Run multiple permutations (Nperms ≥ 100) to stabilize Elo ratings.
  - Reduces sensitivity to match-up ordering.
- Adjusting the K-factor
  - Use smaller K-factors when models have similar win rates.
  - Minimizes rating fluctuations among closely matched models.
- Rapid Convergence for Clear Winners
  - Higher K-factors help ratings align quickly with true performance levels.
  - Useful when there's a clear performance disparity between models.
- Transitivity Is Not Guaranteed
  - Be cautious: Elo ratings may not preserve transitivity.
  - Especially relevant when models have similar performance levels.

Conclusion and Future Work

- Summary:
  - Elo ratings can be sensitive to match-up ordering and hyperparameters.
  - Stability improves with multiple permutations and appropriate K-factor selection.
  - Transitivity may not hold, especially among similar models.
- Implications for LLM Evaluation
  - Careful application of Elo ratings is necessary for reliable model comparisons.
  - Practitioners should adopt recommended guidelines for robustness.
- Future Research Directions
  - Explore alternative rating systems (e.g., TrueSkill, Glicko).
  - Investigate the impact of tie outcomes and multi-category feedback.

Scientific Reviewer

Yanshuo Chen

Reviewer

Summary:

This paper dives into the static ELO evaluation system which is designed for benchmarking LLM performance based on human preference. The paper proposes insightful axioms about the ELO system, displays the basic properties of different hyperparameters and tests on simulated and real-world datasets.

Technical comment: sound but limited contribution.

Scientific contribution: timely and novel, but the analysis part is simple.

Presentation: not very good, needs relevant background knowledge

Reviewer

Comments:

Strength:

1. The studied question is timely and novel. Evaluating the reliability of ELO system helps us understand the current benchmark.
2. The lessons we learned from the experiments is helpful for us to build up a reliable evaluation system.

Reviewer

Comments:

Weakness:

1. The technique contribution is naive. No theoretical analysis of the results variance and hyperparameters.
2. The real experiments is over simplified. As the true real-world LLM arena contains over 2 millions evaluations, hundred of LLMs and noisy labels. The experiments are not persuasive as the authors are not the true owner of an LLM arena. We do not know whether the lessons learned from here have already been deployed or not.

Reviewer

Comments:

Weakness:

3. The presentation needs readers have some background about ELO system and its application scenario.

Reviewer

Recommendation:

Borderline accept, needs meta review.

Scientific Reviewer

Jiuhai Chen

Strength

The paper's exploration of Elo ratings for LLMs is significant, with comprehensive experiments. It offers valuable insights into the application of Elo ratings for LLMs, which I believe can positively impact the way we approach model comparisons in pair-wise ranking scenarios.

Weakness

- More explanation of the K-factor should also be added. While it is mentioned, it is not properly defined or described.
- It is unclear how Elo ratings are most commonly applied for LLMs. For instance, are they typically calculated using only a single ordering? If this information is included, it should be made more explicit.
- Table 1 is somewhat challenging to interpret. Consider creating a plot similar to Figure 1, but with multiple lines representing different models, to enhance readability.
- In the introduction, the implications of the work are noted as particularly significant for scenarios where model performances are closely matched, as this is common in real-world applications. However, in section 4.2, it is mentioned that one model prefers to the other model follows binomial distribution. Need more explanation to this point.

Reviewer

Recommendation:

Borderline accept, needs meta review.

Archaeologist

Amadeo De La Vega

Previous Work: arXiv:2310.14424v1 (Boubdir et. al, 2023)

Which Prompts Make The Difference? Data Prioritization For Efficient Human LLM Evaluation

Meriem Boubdir
Cohere for AI
meri.boubdir@gmail.com

Edward Kim
Cohere
edward@cohere.com

Beyza Ermis
Cohere for AI
beyza@cohere.com

Marzieh Fadaee
Cohere for AI
marzieh@cohere.com

Sara Hooker
Cohere for AI
sarahooker@cohere.com

Previous Work: arXiv:2310.14424v1 (Boubdir et. al, 2023)

Proposes a method to optimize human-in-the-loop evaluations of LLMs by **prioritizing prompts** that minimize tie outcomes in pairwise comparisons of Elo scores.

- Method uses **two metrics** (KL Divergence and Cross-Entropy) to rank prompts based on their potential to generate decisive outcomes.
- Demonstrated a **54% reduction in tied outcomes** for top-priority prompts compared to random sampling.
- Applied the method to several LLMs and showed that prioritization improved **Elo score stability**, reducing the need for extensive human feedback.

Previous Work: arXiv:2310.14424v1 (relation to current paper)

Both papers aim to improve the evaluation of LLMs by enhancing the robustness of Elo scores.

- The current paper implements **data prioritization** in their methodology to analyze the role of hyperparameters (K-factor and N_perm) in robustness.

Subsequent Work: arXiv:2407.04069v2 (Laskar et. al, 2024)

A Systematic Survey and Critical Review on Evaluating Large Language Models: Challenges, Limitations, and Recommendations

Md Tahmid Rahman Laskar[†,‖,*], Sawsan Alqahtani[§], M Saiful Bari[¶,*]
Mizanur Rahman[†,•], Mohammad Abdullah Matin Khan[‡], Haidar Khan[¶]
Israt Jahan[†], Md Amran Hossen Bhuiyan[†], Chee Wei Tan[‡], Md Rizwan Parvez[$]
Enamul Hoque[†], Shafiq Joty[‡,°,*], Jimmy Xiangji Huang[†,*]

[†]York University, [§]Princess Nourah Bint Abdulrahman University, [‡]Nanyang Technological University,
[¶]National Center for AI, Saudi Arabia, [$]Qatar Computing Research Institute (QCRI),
[‖]Dialpad Canada Inc., [•]Royal Bank of Canada, [°]Salesforce Research

Subsequent Work: arXiv:2407.04069v2 (Laskar et. al, 2024)

Provides a systematic review of challenges, limitations, and recommendations for evaluating LLMs

- Identified critical **challenges in LLM evaluation**, including issues with **reproducibility, reliability, and robustness**.
- Highlighted inconsistencies in benchmark datasets, response generation, and evaluation methodologies, noting the lack of standardization across studies.
- Provided a set of **guidelines for improving LLM evaluations**, including better dataset selection, prompt transparency, and evaluation metric alignment with human judgments.
- Recommended moving toward **standardized evaluation protocols** to ensure consistent and reliable assessments across diverse LLMs and tasks.

Subsequent Work: arXiv:2407.04069v2 (relation to current paper)

Both papers address the evaluation of LLMs and highlight the limitations of existing methodologies, particularly Elo scores (current paper) and broader evaluation practices (subsequent work)

- Both papers underline the need for **reliable and robust evaluations**. While current paper suggests best practices for using Elo scores, subsequent paper provides a broader set of guidelines for achieving reproducibility, reliability, and robustness in LLM evaluations.
- The broader challenges outlined in subsequent paper contextualize the specific Elo-related limitations discussed in current paper, highlighting the need for comprehensive evaluation protocols.

Academic Researcher

Jiuhai Chen

Question:

This paper explores the experimental results with up to three players. It would be more meaningful if the conclusions could be further generalized to settings with more players.

Figure 1: **Impact of win probabilities and permutation sampling on Elo ratings**: Comparing Model A and Model B across three different win probabilities ($Prob(A \text{ beats } B) = \{0.6, 0.55, 0.51\}$) with two levels of permutation sampling ($N_{\text{perms}} = 1$ and $N_{\text{perms}} = 100$). The top row displays the observed win rates, the middle row illustrates Elo ratings with a single permutation, and the bottom row shows the mean and standard error of the mean (SEM) of Elo ratings across 100 permutations.

Figure 5: Final Elo scores (S_A, S_B and S_C) for three different models at multiple configurations of $N_{perms} = \{1, 100\}$ and K-factor $= \{1, 8, 16, 32\}$. Intersecting of surfaces of individual model scores signifies that the relative ranking of the models is sensitive to these configurations. The order of model overlaps represent these models ranking based on their Elo scores.

Question:

In Section 7, it is mentioned that faster convergence is observed for higher K-factors, but the evidence for this is unclear. This is not apparent from Figure 3, and I couldn't find any supporting plots in the appendix.

- **Rapid Convergence for Clear Winners**: When there's a clear performance disparity between models, a higher K-factor accelerates the alignment of Elo ratings with the models'

Question:

The method should be conducted in more realistic setting, for example conducting experiments on more API-based LLMs, such as ChatGPT, Gemini, and Claude-3.

Industry Practitioner

Utkarsh Tyagi

Pitch

As the Lead ML Engineer, I propose implementing the paper's enhanced Elo rating methodology to create a more reliable and robust evaluation framework for our clients' language models.

The Problem: Companies are investing heavily in LLM development but lack reliable ways to evaluate and compare models. Current evaluation methods are inconsistent and can lead to costly deployment mistakes.

Our Solution: We'll build an enterprise-grade evaluation platform implementing the paper's enhanced Elo methodology, featuring:

- Statistically rigorous model comparisons
- Confidence intervals for performance metrics
- Automated evaluation pipelines
- Clear, actionable reporting

Implementation Plan:
Build evaluation infrastructure following the paper's recommendations:

- Use multiple permutations (100+) of model comparisons
- Implement adaptive K-factors based on win probability margins
- Add statistical confidence measures
- Create user-friendly reporting interface
- Develop automated testing pipeline

Positive Impact

The main positive impact would be bringing more reliability and transparency to LLM evaluation. This helps:

- Companies make better-informed decisions about model deployment
- Researchers better understand model capabilities compared to other models
- The field advance with more rigorous benchmarking standards
- Better Resource Allocation: Companies can make more informed decisions about model deployment. This reduces waste of computational resources on ineffective models
- Transparency and Trust: Creates more transparent evaluation metrics
- Research Advancement: Accelerates progress in AI development through better feedback

Negative Impact

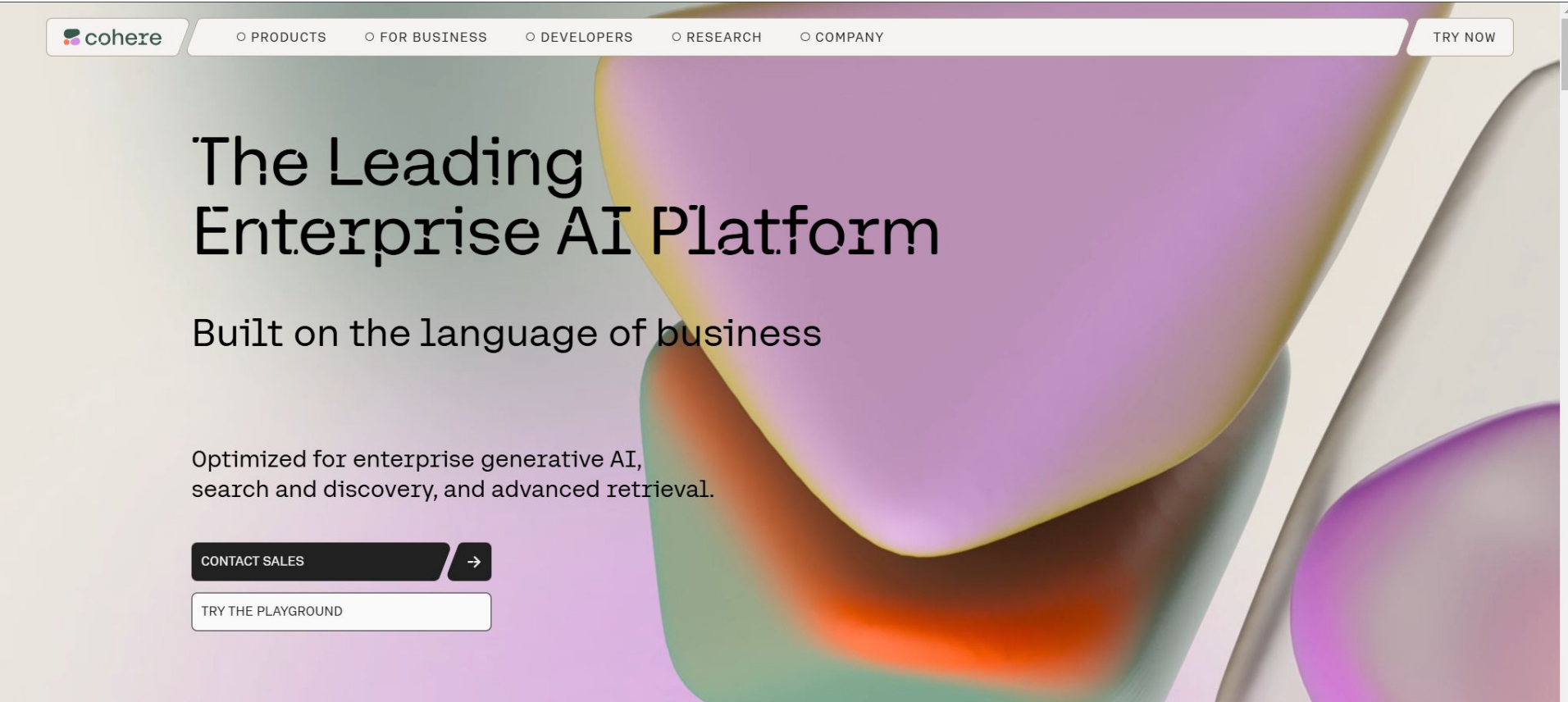
A potential negative impact is that this could create an "arms race" mentality in LLM development, where:

- Companies focus too heavily on improving Elo scores rather than real-world utility
- The focus on comparative evaluation could reduce emphasis on safety and ethics
- Risk of "gaming" the evaluation system

Private Investigator

Yize Cheng

About Cohere



About Cohere

- Founded in 2019, based Toronto, Canada
- Specializes in LLM and NLP solutions for enterprise applications.
  - Search
  - Summarization
  - Conversational AI
  - Text generation
- Partners with major tech player: Oracle, Google, Microsoft, etc.
- Emphasize research-driven innovation
  - Non-profit research lab — Cohere for AI (C4AI)

Lead author — Meriem Boubdir

Education

RWTH Aachen University
Master of Science - MS, Theoretical Particle Physics
2012 - 2015

RWTH Aachen University
Bachelor of Science - BS, Physics
2010 - 2012

Lead author — Meriem Boubdir

Experience

Research Scholar
ML Alignment & Theory Scholars · Full-time
Jan 2024 - Present · 11 mos
Berkeley, California, United States / London, UK · On-site

Large Language Models (LLM), Human Computer Interaction and +1 skill

Research Scholar
Cohere · Full-time
Jan 2023 - Nov 2023 · 11 mos
Remote

Large Language Models (LLM), Human Computer Interaction and +3 skills

Machine Learning Engineer
Brainlab
May 2019 - Jan 2022 · 2 yrs 9 mos
Munich Area, Germany

MLOps, Computer Vision and +3 skills

Lead author — Meriem Boubdir

After switching gears from particle physics to language models, she has focused on the evaluation problem of LLMs, such as:

Which Prompts Make The Difference? Data Prioritization For Efficient Human LLM Evaluation

Meriem Boubdir
Cohere for AI
meri.boubdir@gmail.com

Edward Kim
Cohere
edward@cohere.com

Beyza Ermis
Cohere for AI
beyza@cohere.com

Marzieh Fadaee
Cohere for AI
marzieh@cohere.com

Sara Hooker
Cohere for AI
sarahooker@cohere.com

Last author - Marzieh Fadaee

Education

University of Amsterdam
Doctor of Philosophy (Ph.D.), Computer Science
2014 - 2019

University of Tehran
Master's degree, Computer Engineering
2010 - 2013

Sharif University of Technology
Bachelor's degree, Computer Engineering
2004 - 2009

Last author - Marzieh Fadaee

Experience

Senior Research Scientist
Cohere · Full-time
Jan 2023 - Present · 1 yr 11 mos
Amsterdam, North Holland, Netherlands · Remote

Zeta Alpha
Permanent · 3 yrs 4 mos

- **NLP Research Lead**
  Jan 2021 - Jan 2023 · 2 yrs 1 mo
  Amsterdam, North Holland, Netherlands

- **NLP/ML Research Engineer**
  Oct 2019 - Jan 2021 · 1 yr 4 mos
  Amsterdam Area, Netherlands

Last author - Marzieh Fadaee

Research Interests:

- Natural Language Understanding
- Multilingual Learning
- Data-Conscious Learning
- Robust and Scalable Models
- Evaluation of LLMs

Last author - Marzieh Fadaee

Which Prompts Make The Difference? Data Prioritization For Efficient Human LLM Evaluation

Meriem Boubdir
Cohere for AI
meri.boubdir@gmail.com

Edward Kim
Cohere
edward@cohere.com

Beyza Ermis
Cohere for AI
beyza@cohere.com

Marzieh Fadaee
Cohere for AI
marzieh@cohere.com

Sara Hooker
Cohere for AI
sarahooker@cohere.com

Social Impact Assessor

Juzheng Zhang

Positive Social Impacts Mentioned in the Paper

- Improves NLP Model Evaluation:
  - Enhances fairness and consistency in ranking models, fostering better research outcomes.
- Guidelines for Reliability:
  - Offers concrete best practices for applying Elo ratings, reducing variability and inaccuracies.
- Cost-Effective:
  - Streamlines evaluation, minimizing the need for exhaustive pairwise human comparisons.
- Broader Applicability:
  - Elo system's refinements can extend to other domains like gaming, autonomous systems, and decision-making tools.

Additional Positive Impacts

- Transparency and Fairness:
  - Ensures that rankings are consistent and understandable, enhancing trust among stakeholders.
- Inclusive Research:
  - Simplifies evaluation processes, making them accessible to researchers from under-resourced institutions.
- Trust in AI:
  - Strengthens confidence in AI systems used in critical applications such as healthcare, education, and governance.

Negative Impacts

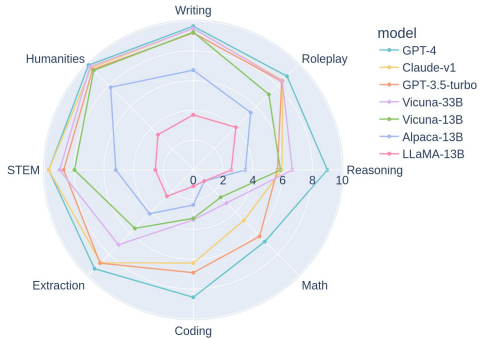
- Bias Propagation:
  - Human biases in feedback may influence evaluations, leading to unfair rankings.
- Reduced Diversity in Models:
  - Over-optimization for Elo scores might discourage experimentation with novel model architectures.
- Economic Inequality:
  - Resource-heavy recommendations like multiple permutations could disadvantage smaller institutions or independent researchers.
- Over-reliance on Elo:
  - A lack of awareness about the system’s limitations could lead to misinterpretation of rankings and model capabilities.

Social Impact Assessor

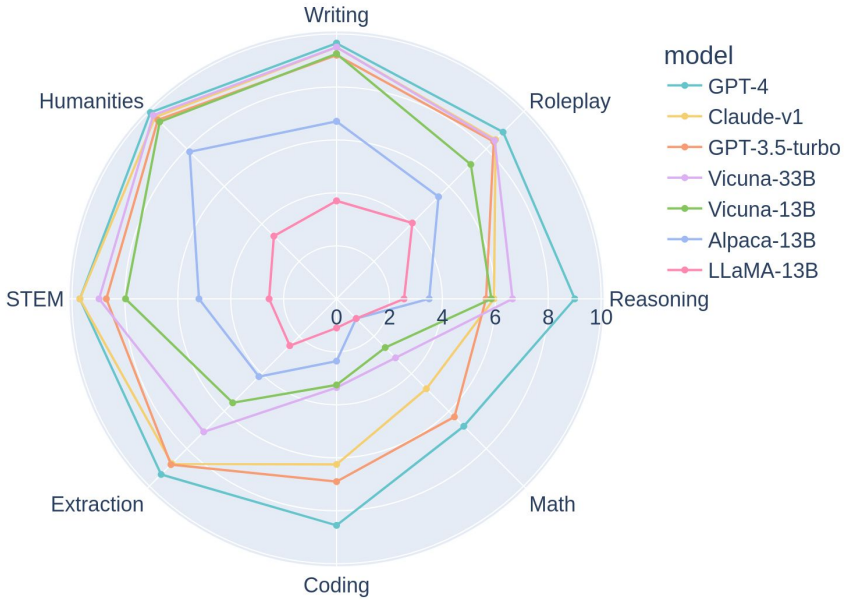
Jiayi Wu

1. Promote improvements to the existing Elo-based evaluation metrics and caution the LLM-related research community against over-reliance on Elo ratings.

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License	Knowledge Cutoff
1	4	Gemini-Exp-1114	1344	+7/-7	6446	Google	Proprietary	Unknown
1	1	ChatGPT-4o-latest (2024-09-03)	1340	+3/-3	42225	OpenAI	Proprietary	2023/10
3	1	o1-preview	1333	+4/-4	26268	OpenAI	Proprietary	2023/10
4	6	o1-mini	1308	+4/-3	28953	OpenAI	Proprietary	2023/10
4	4	Gemini-1.5-Pro-002	1301	+4/-4	23856	Google	Proprietary	Unknown
5	4	Gemini-1.5-Pro-Exp-0827	1299	+2/-3	32355	Google	Proprietary	2023/11
7	10	Grok-2-08-13	1290	+3/-3	47908	xAI	Proprietary	2024/3
7	12	Yi-Lightning	1287	+4/-4	27114	01 AI	Proprietary	Unknown
8	4	GPT-4o-2024-05-13	1285	+2/-2	108575	OpenAI	Proprietary	2023/10
8	3	Claude 3.5 Sonnet (20241022)	1283	+4/-4	26047	Anthropic	Proprietary	2024/4
11	17	GLM-4-Plus	1275	+3/-4	25601	Zhipu AI	Proprietary	Unknown
11	19	GPT-4o-mini-2024-07-18	1272	+3/-3	48407	OpenAI	Proprietary	2023/10
11	19	Gemini-1.5-Flash-002	1272	+4/-4	18112	Google	Proprietary	Unknown
11	28	Llama-3.1-Nemotron-70B-Instruct	1269	+6/-5	7263	Nvidia	Llama 3.1	2023/12
11	15	Gemini-1.5-Flash-Exp-0827	1269	+4/-4	25478	Google	Proprietary	2023/11
11	8	Meta-Llama-3.1-405B-Instruct-fp8	1267	+4/-3	48804	Meta	Llama 3.1 Community	2023/12
11	8	Meta-Llama-3.1-405B-Instruct-bf16	1266	+5/-5	14611	Meta	Llama 3.1 Community	2023/12
12	6	Claude 3.5 Sonnet (20240620)	1268	+2/-3	86633	Anthropic	Proprietary	2024/4
12	26	Grok-2-Mini-08-13	1267	+4/-3	39214	xAI	Proprietary	2024/3
13	8	Gemini Advanced App (2024-05-14)	1267	+2/-3	52219	Google	Proprietary	Online



2. When fine-tuning a foundational model for a specific domain, avoid over-relying on Elo rating scores, and instead focus more on the model's other strengths and weaknesses.



May more important:

- Does the architecture of the foundational model support easy transfer learning and customization for fine-tuning?
- It is important to check whether the foundational model has been trained on high-quality, multi-domain datasets to ensure it can handle the complexities of fine-tuning tasks.
- If performances are close, better not follow.

3.Encourages LLM practitioners to critically assess Elo scores while offering detailed insights and practical tips to enhance the robustness of Elo-based evaluations.

- To obtain stable and reliable Elo ratings, it's recommended to **run numerous permutations**, ideally with Nperm ≥ 100.

- A **smaller K-factor** may reduce significant rating fluctuations when models have **closely matched** win rates.

- When there's a clear performance disparity between models, **a higher K-factor accelerates** the alignment of Elo ratings with the models' "true" performance levels. This is in stark contrast to traditional uses of Elo ratings, where a one-size-fits-all K-factor is frequently applied.

- **"A beats B and B beats C" not always implies "A > C"** in Elo ratings.This is particularly invalid when models have similar performance levels, challenging a common assumption in many Elo-based evaluations.

4. It helps LLM researchers recognize that the current evaluation mechanisms for LLMs are quite limited and emphasizes the need for developing more effective and comprehensive metrics to assess the overall performance of LLMs.

- The majority of researchers focus on improving model performance; however, if the evaluation criteria and their limitations are not well-defined, it becomes challenging to establish a clear benchmark for assessing model capabilities. This may result in the true value of some research efforts being inaccurately evaluated.