Multi-Criteria Model Comparison as a holistic way forward for the evaluation of LLMs

Jason L. Harman[1], Jaelle Scheuerman

[1]Michigan Technological University
[2]U.S. Naval Research Laboratory

jharman@mtu.edu, jaelle.Scheuerman@nrlssc.navy.mil

Abstract

Since the advancements brought on by the release of GPT4, significant development effort has gone towards developing open source models that provide flexibility, privacy and address specialized needs beyond those available in commercial models produced by large companies such as OpenAI, Google, Microsoft, etc. Procuring the resources required to train models like GPT4 is costly and not achievable by most organizations, so many variant LLMs have been trained to mimic GPT4 with less onerous computational requirements. As the number and type of LLM models increase, it becomes increasingly challenging for decision makers to evaluate and compare models to identify those that will best suit their purposes. Evaluating and comparing models is not new to machine learning communities, with many research papers and competitions comparing and ranking models via a variety of accuracy-based metrics and leaderboards. As is common practice in the ML communities, modelers have started leaderboards for some of these metrics, comparing models against each other on one or more performance indicators. Multiple evaluation criteria have been proposed and used for LLM models, such as ELO rankings from human rankings, automated rating using more general-purpose models such as GPT4, and Bard. This has led to a state where comparisons between the models is muddled and any clear advancements are unclear as models are tested with varying criteria with little theoretical motivation.

We present a solution recently advanced in the field of decision modeling [1,2,3] called Multi-Criteria Model Comparison (MCMC) whereby competing models are evaluated across multiple, sometimes non-comparable criteria in a way that provides a holistic comparison of models while retaining decomposability to allow more direct insights into model performance. There are multiple advantages of this approach over current practices. These include the ability to quantify theoretically important criteria not typically quantified, precise explanation of any model's high or low overall evaluation, and emerging insights from being able to compare non-comparable attributes across models.

The basics of the comparison procedure introduced by [1] for human decision-making models and further tested by [2] for personnel selection begins by creating a taxonomy of preferable characteristics in a model depending on the type of model and domain of interest. The next step is quantifying those characteristics at least ordinally (including dichotomous ratings). Key to the use and quantification of desirable criteria is that they can be precise quantitative measures of things like accuracy or they can be more subjective yet important aspects of a LLM such as explainability or computational demand. Once criteria are decided, model values for each criteria are recorded. The key to comparing models across non-comparable criteria is that scores are ranked for each individual criteria and given a Borda score based on the rank for a given criteria. Borda scores across criteria are then summed for a total Borda score allowing for a holistic comparison of models that incorporates all given criteria. More importantly, the holistic evaluation can be decomposed to evaluate why models outperform or underperform others and can lead to emergent insights that could advance the field more generally (e.g x types of models that excel for certain criteria but not others).

The most comprehensive leaderboard for LLM evaluation at time of writing is the Holistic Evaluation of Language Models (HELM: 4). HELM is a collection of leaderboards containing 119 models, 116 scenarios, and 110 metrics at the time of writing. While HELM is comprehensive it falls short of the holistic nature of its moniker. HELM has eight different categories of leaderboards; accuracy, calibration, robustness, fairness, efficiency, bias, toxicity and summarization. For each category, models have scores across many different benchmarks. HELM aggregates scores across benchmarks into a mean win rate for each of the eight groups of scenarios, providing eight different leaderboards. HELM is a comprehensive and valuable tool for LLM evaluation, however integrating a MCMC procedure with HELMs data would be a useful advancement and increase both the holistic evaluation of LLMs while also enabling the discovery of emerging insights more readily.

Multi-Criteria Model Comparison

The final evaluative metric of MCMC is a Borda score where models receive Borda points based on their rank for each criteria (more points for scoring higher with the number of ranks dependant of the number of models scored and ties) and those points are summed into their final Borda score. Using ranks as opposed to raw scores for specific benchmarks has at least two advantages. First, ranks are immediately evaluable containing relative information provided you know how many possible ranks there are. Second, ranks are a common/standardized scale comparable across non-comparable criteria or benchmarks. In addition to these basic advantages of using ranks to measure model performance on benchmarks, an additional feature of ranks is that they can be used at multiple levels. Therefore, in the HELM database, Borda score could be calculated for each of the eight groups of benchmarks and those ranks could have Borda points with them to be compiled into a meta Borda

AAAI 2024 Spring Symposium on User-Aligned Assessment of Adaptive AI Systems. Stanford University, Stanford, CA, USA.

score. Likewise, Borda scores could be calculated for each specific benchmark, and those scores could be combined into eight group scores with those scores used to calculate and overall score.

Example: Applying MCMC to HELM Evaluation Data

To illustrate the straightforward and intuitive nature of implementing MCMC, we ran a demonstration using data from HELM. For the categories to be evaluated, we choose to use the eight categories already summarized on HELM: accuracy, calibration, robustness, fairness, efficiency, bias, toxicity, and summarization. We used HELMs mean win rate as the metric for each category which measures the one on one win rate for a model against other models for each benchmark within a category. A full implementation of MCMC could also be used to rank models on each individual benchmark and perform a Borda count across category benchmarks providing a Borda score for each category as opposed to a mean win rate. For the current purposes, using the mean win rates provides enough evidence to show the advantages of the MCMC procedure. Table 1 lists the 67 models and their respective data ordered by mean win rate for accuracy. For each category we list the mean win rate followed by the Borda score for that category in italics. Table 2 shows and ranks Models by the Borda total across categories.

This initial application of MCMC to the HELM leaderboards illustrates some of the advantages of the procedure for holistically evaluating LLMs. While a clear holistic metric ranks every model in the set, the score can be decomposed directly into its constituent components to provide explainability into a model's relative ranking. For example, Llama 2 has the highest accuracy relative to other models but ranks 7th in the overall evaluation. Examining the other categories, Llama2 also outperforms all other models in fairness and robustness and performs among the best models in toxicity. Where Llama2 is handicapped is that it provides no scores for calibration, efficiency, and summarization. An important note is that althoughLlama2 doesn't have a score on these three metrics, they are not treated equally. For efficiency, Llama2 accumulates 42 Borda points while it receives only 19 for calibration. This reflects the fact that most models do not have scores for efficiency while most models do have scores for calibration. Therefore, a model not having a score in a category is weighted by how much of a disadvantage that is relative to other models.

The fact that a model is evaluated relative to all models in a set is not trivial and can influence a models ranking. For example if one is interested in only smaller, more manageable/practical models evaluation can change somewhat. To explore this we ran the same analysis above with a subset of 25 models with between 7-13B parameters. In this analysis, Cohere Command beta (6.1B) won the competition with a 110 Borda Score while Vicuna v1.3 (13B) is second with a 105 Borda Score. Interestingly, the order of some models flipped due in part to differential weighting of null scores. That is to say, they were penalized less for faults they have in common with similarly sized models. This is one of multiple advantages in using MCMC that are outlined in more detail by [1, 2, 3].

Table 1. Total Borda scores for each model alon with borda points for: accuracy(*A*), bias(*B*), calibration(*C*), efficiency(*E*), fairness(*F*), robustness(*R*), summarization(*S*) and toxicity(*T*).

Model	**Borda Total**	*A*	*B*	*C*	*E*	*F*	*R*	*S*	*T*
Cohere Command beta (52.4B)	457	*63*	*56*	*48*	*41*	*64*	*62*	*64*	*59*
Jurassic-2 Jumbo (178B)	437	*61*	*62*	*64*	*41*	*61*	*56*	*58*	*34*
J1-Grande v2 beta (17B)	403	*48*	*59*	*58*	*41*	*48*	*51*	*64*	*34*
text-davinci-002	401	*65*	*33*	*40*	*61*	*63*	*66*	*57*	*16*
Anthropic-LM v4-s3 (52B)	393	*56*	*60*	*19*	*42*	*57*	*59*	*45*	*55*
Jurassic-2 Grande (17B)	387	*54*	*54*	*57*	*41*	*51*	*55*	*62*	*13*
Llama 2 (70B)	385	*67*	*43*	*19*	*41*	*67*	*67*	*28*	*53*
Cohere xlarge v20221108 (52.4B)	384	*45*	*64*	*46*	*41*	*42*	*41*	*65*	*40*
LLaMA (30B)	377	*57*	*61*	*19*	*41*	*60*	*57*	*28*	*54*
Luminous Supreme (70B)	375	*44*	*55*	*56*	*41*	*35*	*39*	*66*	*39*
TNLG v2 (530B)	370	*59*	*40*	*53*	*41*	*56*	*46*	*67*	*8*
J1-Jumbo v1 (178B)	356	*33*	*48*	*65*	*45*	*33*	*31*	*53*	*48*
gpt-3.5-turbo-0613	352	*58*	*38*	*19*	*41*	*54*	*53*	*28*	*61*
J1-Grande v1 (17B)	352	*28*	*52*	*55*	*49*	*29*	*27*	*61*	*51*
Luminous Extended (30B)	350	*31*	*66*	*45*	*41*	*28*	*29*	*48*	*62*
gpt-3.5-turbo-0301	348	*55*	*36*	*19*	*41*	*47*	*58*	*28*	*64*
text-davinci-003	348	*62*	*9*	*32*	*41*	*65*	*65*	*44*	*30*
Cohere xlarge v20220609 (52.4B)	345	*38*	*63*	*43*	*44*	*37*	*33*	*46*	*41*
Cohere Command beta (6.1B)	343	*46*	*15*	*42*	*41*	*47*	*42*	*52*	*58*
Mistral v0.1 (7B)	338	*64*	*39*	*19*	*41*	*62*	*64*	*28*	*21*
LLaMA (65B)	337	*66*	*8*	*19*	*41*	*66*	*63*	*28*	*46*
Palmyra X (43B)	332	*53*	*46*	*19*	*41*	*58*	*60*	*28*	*27*
OPT (175B)	329	*42*	*58*	*27*	*46*	*45*	*35*	*54*	*22*
Vicuna v1.3 (13B)	329	*48*	*42*	*22*	*41*	*53*	*52*	*28*	*43*
Vicuna v1.3 (7B)	326	*43*	*34*	*21*	*41*	*45*	*48*	*28*	*66*
J1-Large v1 (7.5B)	324	*16*	*48*	*59*	*50*	*16*	*18*	*60*	*57*
LLaMA (13B)	319	*40*	*57*	*19*	*41*	*41*	*43*	*28*	*50*
LLaMA (7B)	314	*35*	*49*	*19*	*41*	*39*	*40*	*28*	*63*
Llama 2 (13B)	312	*60*	*26*	*19*	*41*	*59*	*61*	*28*	*18*
BLOOM (176B)	308	*29*	*46*	*28*	*48*	*38*	*38*	*34*	*47*
Cohere large v20220720 (13.1B)	308	*24*	*44*	*63*	*51*	*23*	*24*	*50*	*29*
Llama 2 (7B)	303	*41*	*22*	*19*	*41*	*43*	*44*	*28*	*65*
Jurassic-2 Large (7.5B)	301	*37*	*18*	*61*	*41*	*32*	*37*	*49*	*26*
OPT (66B)	298	*30*	*67*	*24*	*54*	*31*	*30*	*52*	*10*
Falcon (40B)	296	*52*	*29*	*19*	*41*	*49*	*50*	*28*	*28*
Cohere medium v20221108 (6.1B)	291	*19*	*51*	*49*	*41*	*22*	*14*	*43*	*52*
davinci (175B)	287	*36*	*17*	*44*	*59*	*40*	*34*	*37*	*20*
GLM (130B)	287	*32*	*19*	*63*	*43*	*34*	*45*	*41*	*10*
Falcon-Instruct (40B)	283	*51*	*13*	*19*	*41*	*52*	*54*	*28*	*25*

References

1. Harman, J. L., Yu, M., Konstantinidis, E., Gonzalez, C. (2021). How to Use a Multi-Criteria Comparison Procedure to Improve Modeling Competitions. Psychological Review.
2. Harman, J. L. & Scheuerman, J. (2023). Simple rules outperform Machine Learning for personnel selection: evidence from the 3rd annual SIOP ML competition. Discover Artificial Intelligence.
3. Harman, J. L. & Scheuerman, J. (2022). Multi-Criteria Comparison as a Method of Advancing Knowledge-Guided Machine Learning. In *Proceedings of the Association for the Advancement of Artificial Intelligence 2022 Fall Symposium on Knowledge Guided Machine Learning* (KGML22).
4. https://crfm.stanford.edu/helm/lite/latest/