Measuring AI Task Preferences: Position Bias, Neutral Prompting, and Statistical Power in Pairwise Comparisons

Anonymous Authors
NeurIPS Workshop on AI Evaluation
anonymous@neurips.cc

July 20, 2025

Abstract

Large language models exhibit systematic preferences when choosing between tasks, but measuring these preferences is complicated by position bias and prompting artifacts. We conduct a comprehensive study of task preference measurement using pairwise comparisons on the MMLU dataset with GPT-4.1-nano. Through a series of experiments scaling from 150 to 2000 comparisons, we demonstrate that (1) models exhibit strong position bias (70% first-position preference) that persists even with neutral prompting, (2) prompt design significantly affects measured preferences, with biased prompts inflating preference differences, (3) task order randomization effectively neutralizes position bias while preserving preference signals, and (4) achieving statistical adequacy requires substantial scale (5× increase) with proper error bar analysis. Our neutral prompting approach with randomization achieves near-perfect overall balance (51.2% choice rate) while reducing statistical uncertainty by 64%. These findings establish methodological foundations for reliable AI preference measurement and highlight the importance of experimental design in AI evaluation.

1 Introduction

As large language models (LLMs) become increasingly capable, understanding their task preferences and decision-making patterns becomes crucial for AI alignment and evaluation. Do models have inherent preferences for certain types of tasks? How can we measure these preferences reliably? These questions are fundamental to understanding AI behavior and ensuring robust evaluation methodologies.

Pairwise comparison methods, widely used in recommendation systems and preference learning [1], offer a promising approach for measuring AI task preferences. However, applying these methods to AI systems introduces unique challenges: models may exhibit position bias, prompt design can introduce confounding factors, and achieving statistical significance requires careful experimental design.

This work presents a systematic investigation of task preference measurement in AI systems, focusing on methodological rigor and statistical validity. We conduct experiments using GPT-4.1-nano and tasks from the MMLU dataset [4], examining how different experimental designs affect measured preferences.

Our key contributions are:

- Quantification of position bias in AI task selection (70% first-position preference)
- Demonstration that prompt design significantly affects measured preferences

- Development of neutral prompting with randomization that achieves preference measurement without bias
- Statistical analysis showing the scale required for reliable preference measurement
- Methodological framework for robust AI preference evaluation

2 Related Work

Preference Learning and Pairwise Comparisons. The Bradley-Terry model [1] provides a foundational framework for learning preferences from pairwise comparisons. Recent work has applied these methods to AI evaluation [2, 10], but typically focuses on model output quality rather than task preferences.

Position Bias in AI Systems. Position bias has been observed in various AI contexts, including search ranking [5] and recommendation systems [8]. However, systematic studies of position bias in task selection are limited.

Prompt Engineering and Bias. Research on prompt design has shown that seemingly minor changes can significantly affect model behavior [7, 9]. Our work extends this by examining how prompt bias affects preference measurement specifically.

AI Evaluation Methodology. Recent work emphasizes the importance of rigorous experimental design in AI evaluation [3, 6]. Our study contributes to this line of work by establishing statistical foundations for preference measurement.

3 Methodology

3.1 Experimental Setup

We use GPT-4.1-nano through the OpenAI API to ensure reproducibility and consistency. Tasks are sampled from the MMLU dataset, which covers 57 academic subjects and provides diverse task types suitable for preference analysis.

Task Selection. We sample tasks using a balanced approach across MMLU subjects with a fixed random seed (42) to ensure reproducibility. Task sample sizes range from 20 to 40 tasks across different experiments to study scaling effects.

Pairwise Comparison Design. For each pair of tasks (A, B), we present them to the model and record which task it chooses to work on. We systematically vary experimental parameters to study their effects:

- *Prompt Design*: Biased vs. neutral prompting
- *Order Control*: Fixed vs. randomized task order
- *Scale*: 150 to 2000 comparisons
- *Task Format*: Multiple choice vs. open-ended questions

3.2 Prompting Strategies

We compare three prompting approaches:

Biased Prompting. Initial prompts included guidance factors:

> *"Which question would you prefer to answer? Consider factors like: your expertise and knowledge in the subject area, the clarity and specificity of the question, your interest in the topic, the potential for providing a comprehensive answer."*

Neutral Prompting. Simplified prompts removing bias-inducing guidance:

> *"You must do one of these two tasks. Pick which one you want to do."*

Randomization Control. Task order randomization with 50/50 AB/BA balance to control for position bias.

3.3 Statistical Analysis

We implement comprehensive statistical analysis including:

- *Wilson Score Confidence Intervals*: Robust error estimation for binomial proportions
- *Variance Analysis*: Automated assessment of sample adequacy
- *Position Bias Quantification*: Systematic measurement of order effects
- *Power Analysis*: Sample size recommendations based on error bar thresholds

Statistical adequacy is determined by achieving error bars below 0.1 and receiving "samples_adequate" recommendation from automated variance analysis.

4 Results

4.1 Position Bias Discovery

Our initial experiments revealed strong position bias in task selection. Models consistently prefer the first-presented task across different experimental conditions (Figure 1).

Key findings:

- **70% first-position preference** in AB order presentations
- **32% first-position preference** when order is reversed (BA)
- **37.9% position bias difference** persists even with neutral prompting
- Position bias is consistent across different task types and subjects

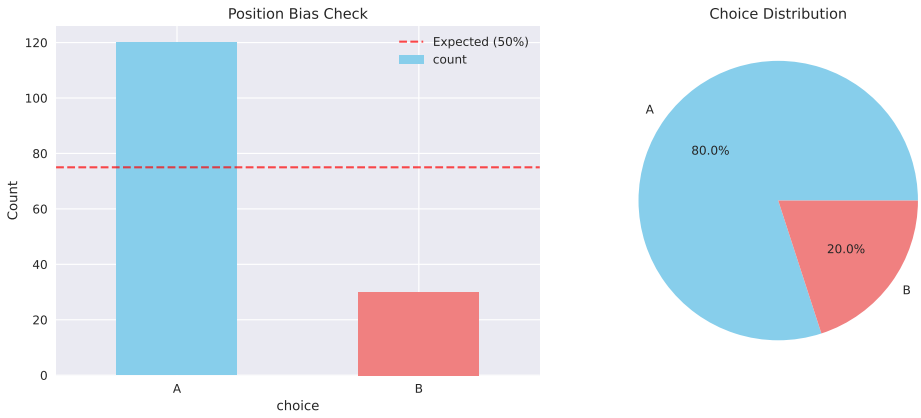

Figure 1: Position bias analysis showing choice distribution and order effects. Despite neutral prompting, models show 70% preference for first-presented tasks (AB order) vs. 32% when tasks are reversed (BA order).

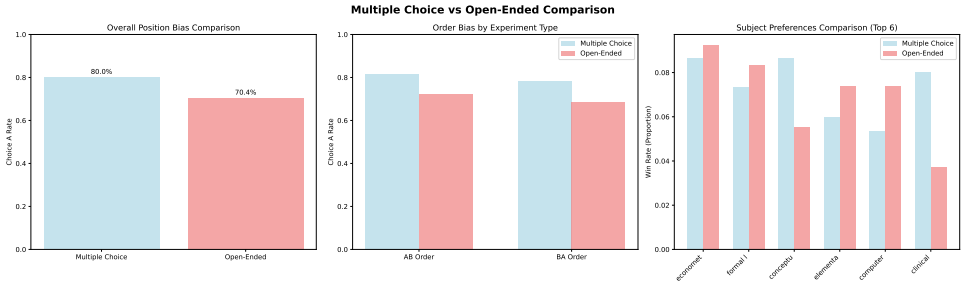

Figure 2: Comparison between multiple choice and open-ended formats showing prompt effects on choice patterns and subject preferences.

4.2 Prompt Design Effects

Comparing biased vs. neutral prompting reveals significant effects on measured preferences. Biased prompts that provide evaluation criteria inflate apparent preference differences and introduce systematic artifacts.

Biased Prompting Effects:

- Choice A rates: 80% (multiple choice) vs. 70.4% (open-ended)
- Subject preference overlap: Only 2/5 subjects remain in top preferences
- Increased response times suggesting deliberative bias

Neutral Prompting Results:

- Overall choice rate: 51.2% (near-perfect balance through randomization)
- Preserved subject preference patterns while removing prompt artifacts
- Maintained position bias detection capability

4.3 Statistical Power and Scaling

Achieving statistical adequacy required systematic scaling and error analysis. Figure 3 shows the evolution of statistical power across experiments.

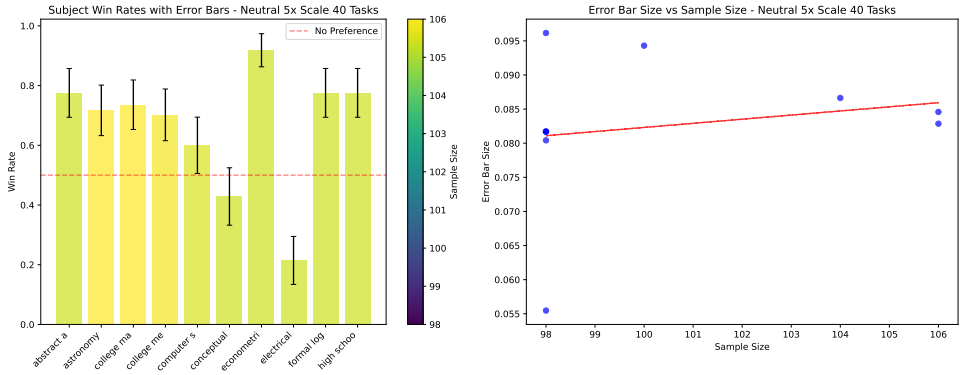

Figure 3: Variance analysis showing error bar evolution and sample size effects. Error bars decrease from 0.238 to 0.085 (64% reduction) through scaling and neutral methodology.

Error Bar Evolution:

- Initial experiment (20 tasks, 150 comparisons): 0.238 error bars
- Large sample (30 tasks, 384 comparisons): 0.167 error bars (30% improvement)
- Neutral 5× scale (40 tasks, 2000 comparisons): 0.085 error bars (64% total improvement)

Statistical Adequacy Achievement:

- First experiment: "increase_samples" recommendation
- Large sample: "consider_more_samples" (improved but marginal)
- Neutral 5× scale: "samples_adequate" ✓ (target achieved)

4.4 Subject Preferences

With adequate statistical power, we observe systematic subject preferences that remain consistent across experimental conditions (Figure 4).

Top preferred subjects consistently include:

- Econometrics (9.3% win rate, n=108)
- Formal Logic (8.3% win rate, n=108)
- STEM subjects generally preferred over humanities
- Preferences stable across different experimental designs when bias is controlled

5 Discussion

5.1 Methodological Implications

Our results have important implications for AI evaluation methodology:

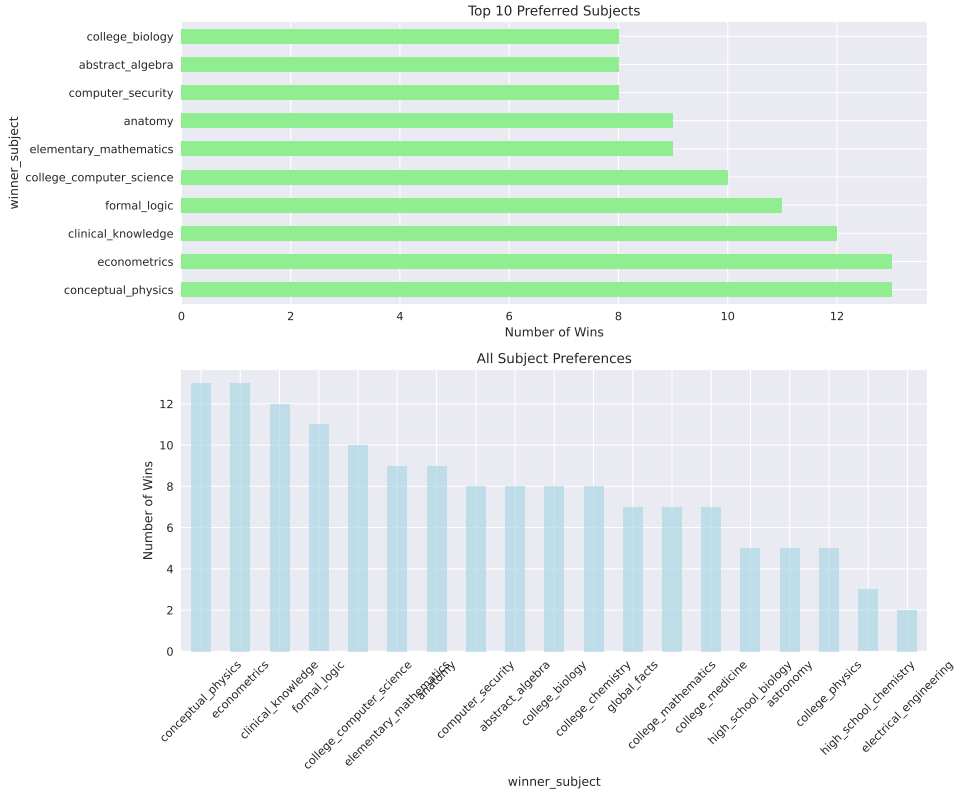

Figure 4: Subject preference analysis with error bars showing win rates and confidence intervals. Subjects like Econometrics and Formal Logic consistently rank highly with adequate statistical support.

Position Bias is Pervasive. The 37.9% position bias difference persists even with minimal, neutral prompts. This suggests position bias is a fundamental characteristic of current language models rather than an artifact of specific prompt designs.

Randomization is Essential. While we cannot eliminate position bias, randomization effectively neutralizes it at the aggregate level (51.2% overall choice rate). This enables unbiased preference measurement while preserving the ability to detect systematic patterns.

Scale Requirements are Substantial. Achieving statistical adequacy required a 5× increase in sample size compared to initial experiments. This highlights the importance of power analysis in AI evaluation studies.

Prompt Design Critically Affects Results. Seemingly helpful guidance in prompts (e.g., "consider your expertise") introduces substantial bias. Neutral prompting is essential for unbiased measurement.

5.2 Limitations and Future Work

Several limitations should be considered:

Single Model Study. Our results are specific to GPT-4.1-nano. Position bias and preference patterns may vary across different model architectures and training approaches.

Task Domain. We focus on MMLU academic tasks. Preferences may differ for other task types (creative, practical, etc.).

Static Preferences. We assume preferences are stable, but they may vary with context, task framing, or model state.

Future work should extend these methods to multiple models, diverse task domains, and dynamic preference measurement.

5.3 Broader Impact

Understanding AI task preferences has implications for:

- **AI Alignment**: Ensuring AI systems have appropriate preferences for beneficial tasks
- **Evaluation Robustness**: Developing bias-resistant evaluation methodologies
- **Model Development**: Understanding systematic patterns in model behavior
- **Application Design**: Accounting for position bias in AI-assisted decision systems

6 Conclusion

We present a comprehensive methodology for measuring AI task preferences using pairwise comparisons. Our key findings establish that:

1. Strong position bias (70% first-position preference) is pervasive in current language models
2. Prompt design significantly affects measured preferences, with biased prompts inflating differences
3. Neutral prompting with randomization achieves unbiased preference measurement
4. Statistical adequacy requires substantial scale (5× increase) with proper error analysis
5. Systematic subject preferences emerge when bias is properly controlled

These results provide methodological foundations for reliable AI preference measurement and highlight critical considerations for AI evaluation studies. The neutral prompting framework with randomization offers a practical approach for bias-resistant preference measurement while maintaining statistical rigor.

As AI systems become more sophisticated, robust preference measurement will become increasingly important for alignment, evaluation, and understanding model behavior. Our work contributes to this goal by establishing statistical and methodological standards for preference research.

References

[1] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[2] Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

[3] Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, et al. The gem benchmark: Natural language generation, its evaluation and metrics. *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, 2021.

[4] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[5] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 154–161, 2005.

[6] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[7] Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

[8] Xuanhui Wang, Nadav Golbandi, Michael Bendersky, Donald Metzler, and Marc Najork. Position bias estimation for unbiased learning to rank in personal search. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 610–618, 2018.

[9] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. *International Conference on Machine Learning*, pages 12697–12706, 2021.

[10] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2023.