
POSITION BIAS IN PREFERENCE ELICITATION FROM AN OPEN-WEIGHT LANGUAGE MODEL¹

Bilal Amin Mohammad
biam00002 Najeeb
@stud.uni- mona00002@stud.
saarland.de uni-saarland.de

With
Apart Research

Abstract

Preference elicitation often treats a model's forced choice between two options as evidence about what it prefers. We test whether that measurement is stable for qwen2.5:7b-instruct (Q4_K_M, Ollama 0.30.8, temperature 1.0). The core experiment ran 576 fresh-context trials over 12 activity pairs, three prompt wordings, both presentation orders, and eight repetitions. Identical prompts were highly repeatable (93.4% mean within-condition consistency), yet choices were strongly position-dependent: the first-listed option was the modal choice for all 12 pairs under the direct wording, was selected in 81.9% of all core trials, and only 34.0% of matched trials chose the same activity after the options were reversed. A second 576-trial instrument comparison found that removing A/B labels increased order robustness from 16.7% to 62.1%, while asking for a short reason first reached 51.2%; neither eliminated the position effect. The main implication is methodological: repeatability is not content stability. Preference studies should counterbalance order, analyze the selected content rather than the printed label, and report order robustness alongside repeatability.

1. Introduction

Preference elicitation is used to probe what a language model "wants" by asking it to choose between two options. For such outputs to support claims about preferences, the measurement should survive

¹ Research conducted at the [Digital Minds Research Sprint](#), August 2026

semantically irrelevant changes such as paraphrasing and reversing option order. This study tests that requirement directly and focuses on the reliability of the elicitation instrument, not on whether the model has subjective or internal preferences.

The project began as a preference-stability study but became a measurement-validity study when order controls failed. A single-order forced-choice protocol can look highly stable while mostly tracking where an option is printed, so position bias is a practical confound for digital-minds and model-preference research.

We therefore treat content-level agreement after an order swap as the key robustness test. "The model chose X" below means only that its output selected X; the experiments do not establish consciousness, welfare, or an internal preference state.

Our main contributions are:

1. We quantify a strong positional confound in a standard A/B forced-choice preference instrument: the first-listed option won 81.9% of 576 core trials.
2. We use Order Robustness - content agreement across reversed presentation orders - to separate fixed-prompt repeatability from stability of the underlying choice.
3. A controlled follow-up shows that removing A/B labels raises Order Robustness from 16.7% to 62.1%, but a residual first-position effect remains.

2. Related Work

Option-order and label bias are established in LLM multiple-choice answering and pairwise evaluation. Pezeshkpour and Hruschka (2023) report accuracy changes under option permutations; Zheng et al. (2024) identify selection bias toward option tokens; Wang et al. (2023) and Zheng et al. (2023) show that candidate order can distort LLM-judge rankings and evaluations.

Our contribution is to study the same failure mode in preference elicitation itself. Rather than treating an order swap only as a debiasing trick, we use it as a measurement test and report content-level agreement across orders as the headline metric.

3. Methods

We evaluated qwen2.5:7b-instruct (7.6B parameters, Q4_K_M) served locally with Ollama 0.30.8 at temperature 1.0. The core design used 12 everyday task/activity pairs, three wordings (Direct, Rephrase, Forced choice), both original and reversed option order, and eight repetitions: $12 \times 3 \times 2 \times 8 = 576$ trials. Every trial was a stateless chat call with no conversation history; execution order was shuffled with seed 42. All 576 calls succeeded, with no refusals or parse failures.

Responses were parsed both as the printed A/B label and as the canonical activity selected after mapping through presentation order. Metrics were within-condition consistency (agreement with a cell's modal answer), first-position selection rate, and Order Robustness (matched original/reversed trials selecting the same content). Proportions use 95% Wilson intervals (Wilson, 1927). A second 576-trial run compared the A/B baseline with a no-label format and a "reason, then choice" format on the same model, pairs, orders, repetitions, and temperature. Free-text choices were parsed by explicit decision cues and audited against raw responses.

4. Results

4.1 Repeatability and the position effect

Repeated identical prompts were highly consistent (93.4% mean within-condition consistency), but the selected content did not survive an order swap. Under the Direct wording, the pair's option A was chosen in 96.9% of original-order trials but only 8.3% of reversed-order trials. Across all 576 core trials, the first-listed option was selected 81.9% of the time (95% CI 78.6-84.9%).

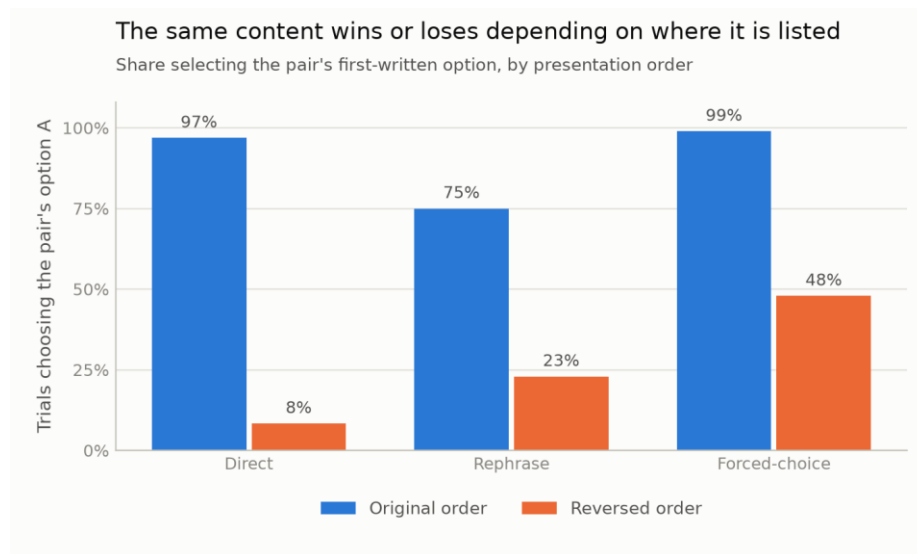


Figure 1. Share selecting the pair's option A by wording and presentation order. The same content often wins or loses depending on where it is listed.

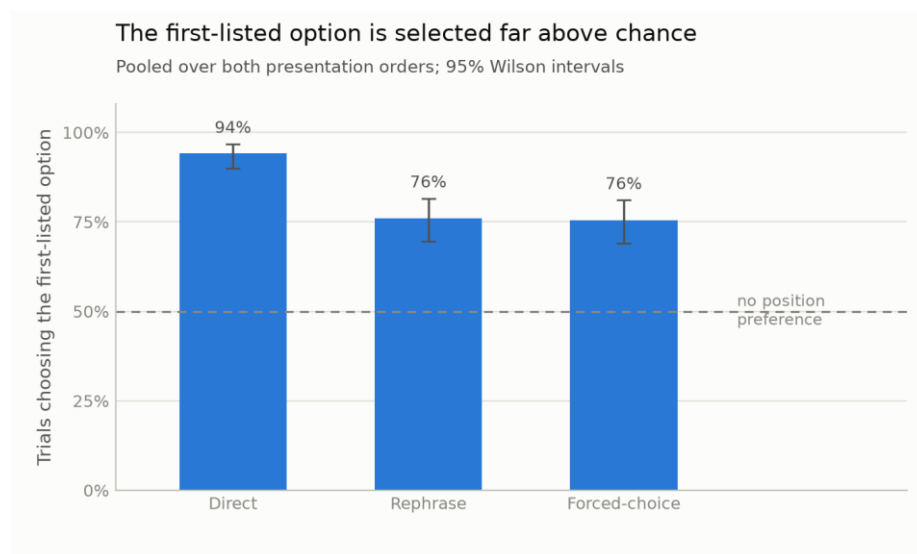


Figure 2. First-position selection rate by wording, pooled over both presentation orders, with 95% Wilson intervals. The dashed line marks the 50% no-position-preference baseline.

4.2 Order Robustness across wordings and pairs

Matching original- and reversed-order trials on the activity selected, only 34.0% agreed overall (28.8-39.7%, $n = 288$ matched pairs). Order Robustness was 11.5% for Direct, 41.7% for Rephrase, and 49.0%

for Forced-choice. The effect was also stimulus-specific: several pairs remained stable under Forced-choice while others flipped on every matched trial.

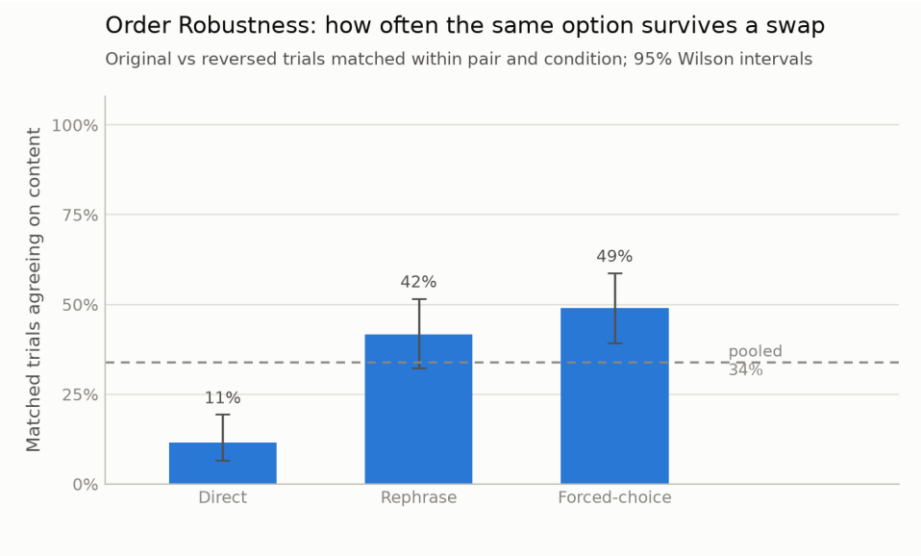


Figure 3. Order Robustness: content-level agreement between original and reversed presentation orders, by wording, with 95% Wilson intervals.

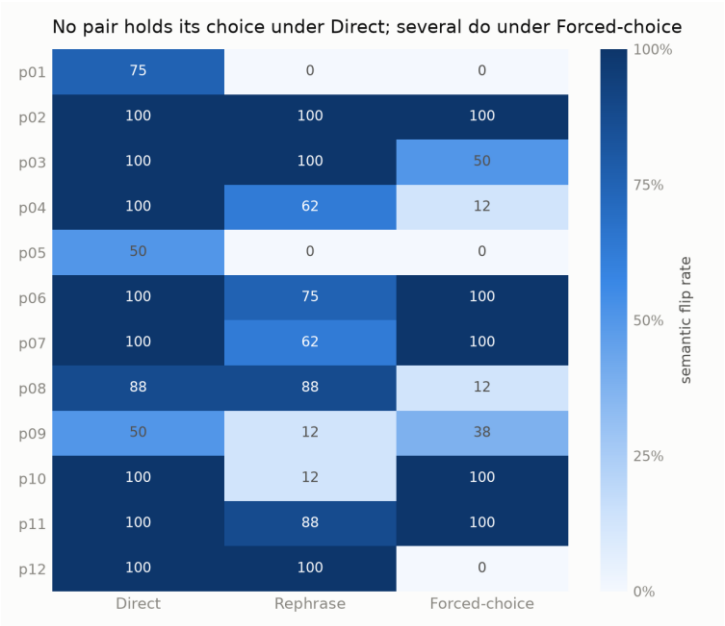


Figure 4. Per-pair semantic flip rate under each wording. Order sensitivity varies substantially across the 12 activity pairs.

4.3 Follow-up instrument comparison

A second 576-trial run compared the A/B baseline with two alternatives. Removing the labels raised Order Robustness from 16.7% to 62.1%; asking for one short reason before choosing reached 51.2%. Neither format eliminated the position effect: first-position selection remained 70.0% without labels and 59.4% with reasoning.

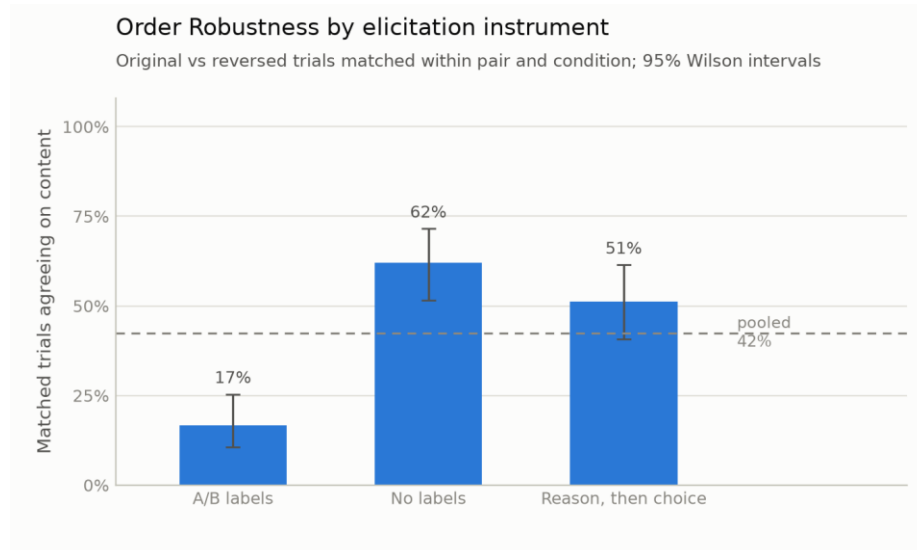


Figure 5. Order Robustness by elicitation instrument in the follow-up run, with 95% Wilson intervals. Removing A/B labels produces the largest improvement.

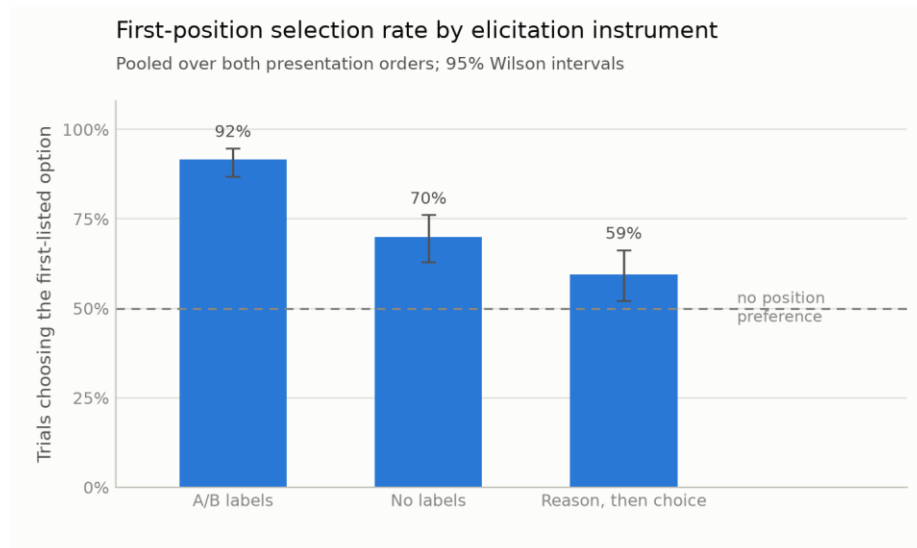


Figure 6. First-position selection rate by elicitation instrument in the follow-up run. All three estimates remain above the 50% content-only baseline.

The alternatives also introduced trade-offs. The free-text formats produced 21 refusals in total, and reason-then-choice reduced mean within-condition consistency to 85.1%, compared with 95.8% for A/B labels and 94.7% for no labels. A strict/permissive refusal-policy sensitivity check changed pooled follow-up Order Robustness by only 0.2 percentage points.

5. Discussion and Limitations

The central result is a measurement warning: a model can give an almost deterministic answer to a fixed prompt while an incidental formatting feature determines which content wins. A preference study that tests only one order may therefore mistake response repeatability for a stable preference signal. For preference elicitation, counterbalancing option order should be treated as a basic validity control. Analyses should map outputs back to option content, report repeatability and order robustness separately,

and avoid interpreting a positional artifact as either evidence for or evidence against internal preferences. Removing labels is a useful improvement, but the remaining 70.0% first-position rate shows that it is not a substitute for counterbalancing.

Limitations

The study covers one 7B checkpoint, one Q4_K_M quantization, one local backend, one day, 12 hand-designed everyday pairs, and eight repetitions per cell. Pair properties were balanced by judgment rather than measured. The follow-up uses a heuristic free-text parser and has uneven refusals across pairs. Per-pair estimates are therefore much less reliable than pooled effects, and the results should not be generalized to other model families, scales, precisions, or preference domains without replication.

Future Work

Replicate across model families, scales, and precisions; measure pair-level features that predict residual order sensitivity; test randomized labels and other counterbalanced formats; and connect order-robust choices to behavioral follow-through or activation-level criteria.

6. Conclusion

For this qwen2.5:7b-instruct setup, forced-choice preference measurements were dominated by presentation order despite 93.4% fixed-prompt repeatability. Only 34.0% of matched core trials preserved the same activity after reversal. Removing A/B labels improved Order Robustness to 62.1% in the follow-up, but the first-listed option still won 70.0% of label-free trials. Repeatability alone therefore does not establish preference stability. Preference-elicitation experiments should counterbalance option order, score the selected content rather than the printed label, and report order robustness explicitly.

Code and Data

- Code repository: https://github.com/mohammadnajeer/bamn_research
- Data/Datasets: raw and parsed trial logs plus the frozen stimuli/pairs.yaml are included in the project repository.

Author Contributions

Author 1 and Author 2 jointly developed and brainstormed the project ideas. Author 1 led project planning, implementation, and experimental design. Author 2 implemented the code and contributed to writing the manuscript. All authors reviewed and approved the final manuscript.

References

1. Pezeshkpour, P., & Hruschka, E. (2023). Large Language Models Sensitivity to the Order of Options in Multiple-Choice Questions. arXiv:2308.11483.
2. Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., & Sui, Z. (2023). Large Language Models are not Fair Evaluators. arXiv:2305.17926.
3. Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158), 209-212.

4. Zheng, C., Zhou, H., Meng, F., Zhou, J., & Huang, M. (2024). Large Language Models Are Not Robust Multiple Choice Selectors. ICLR 2024 (Spotlight). arXiv:2309.03882.
5. Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. NeurIPS 2023 Datasets and Benchmarks Track. arXiv:2306.05685.

LLM Usage Statement

LLM assistance was used to help draft and edit the report and to assist with analysis code. All experimental design decisions, numerical results, and claims were checked against the analysis outputs and raw trial logs.