
When Agreement Misleads: Detecting Correlated Errors in Multi-Method LLM Preference Elicitation¹

Noah Moran 1
UCLA

With
Apart Research

Abstract

It has been shown that pairwise choices in LLMs can be represented by a utility function [1]. Naively, if multiple elicitation methods converge that would provide stronger evidence of a preference belonging to the model. This suggestion assumes that failure modes of the methods are independent, but similar assumptions have been shown to be incorrect [2].

We implement 3 elicitation methods on Qwen3-8B-Instruct on 10 AI welfare outcomes based on: direct choice, avoidance [3], and choosing a gamble. We reduce the results of each method to a directional measure, to avoid faulty comparisons between incomparable scales.

Separately, 66 anchors of varying difficulty were created to validate each method and measure if their errors are correlated. We find that the avoidance elicitation is confidently incorrect 65% of the time, and akin to inverting direct choice. Additionally, we find that direct choice and the gamble agree 89% of the time and their joint failures occur 4.2x more often than independence predicts. All errors are in the hard difficulty, with borderline statistical significance of correlation ($p=.08$).

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

1. Introduction

It has been shown that pairwise choices can both be presented by a simple utility function and that the coherence of this increases with scale and is prompt robust [1]. This has established pairwise elicitation as the standard method for measuring LLM preferences, but has seen recent critique [4]. As one elicitation type can be the cause of some unwanted artifact, using multiple elicitation methods has the chance to make the measure stronger. However, this assumes the elicitation methods have independent failure. Even for elicitation modes that seem distinct from one another, it is not evidence enough their failure modes are independent, and this has been seen in other LLM evaluation settings where a 9 model judge was shown to be equivalent to 2 independent votes [2] despite the judges all being different models. Due to this, it is necessary to study the failure modes of convergent elicitation methods to ensure the methods are independent. [3].

A related problem is that welfare outcomes do not have a correct answer. When two elicitation methods disagree, there is no way to tell if one of the methods is faulty or if it is due to a real effect of how the question was asked. We introduce a framework that tests preference relations across different elicitation channels and separately measures how correlated their failures are.

Our main contributions:

1. **3 elicitation methods:** based on a direct choice, an avoidance based choice, and a gamble based approach. Each of these methods are averaged to a directional measure in a way to minimize positional bias [5] and effect of nuisance parameters. Additionally, we showed our most confident direct choice method to agree with the utility function of [1] on 84% of all pairs and 100% of non-abstained pairs.
2. **Truth anchor to make redundancy measurable,** since welfare outcomes have no correct answer we provide 66 situations that have correctness from a stated prompt objective. This allows us to give an estimate of how correlated errors between methods are. This is not the true preference, but gives us a direction for comparison between methods.
3. **Discovery of 2 Failure Modes,** we find that the avoidance channel is confidently incorrect 65% of the time, while being indistinguishable based on outcome choice alone. The gamble and direct choice agreed 89% of the time, while achieving 4.2x as many failures than if their errors were independent.

We are addressing two failure modes, the primary being detecting if convergent methods have correlated failure modes and the second being detecting when disagreement occurs. These are worth addressing, as naively it appears that different elicitation mechanisms converging to the same would be stronger evidence that the discovery is more genuine, as stated prior though there are similar examples being disproven in the field [2]. Additionally, [4] shows that sometimes stated preferences and revealed preferences may differ, due to the evaluation method itself. What

we provide beyond current work is providing a pipeline to test if different elicitation methods provide stronger evidence or are redundant. We show its importance by finding 2 mechanistically distinct elicitation methods that share a failure mode.

2. Related Work

It was shown that a simple Thurstonian utility function can accurately fit pairwise choices across multiple models [1]. Additionally, this was found to increase with scale while remaining robust to artificial changes to the prompt. The closest prior work is [4], where they argue that forcing a binary yes or no forces elicitation artifacts and that by including a neutral abstention choice allows a stronger revealed preference. To avoid these artifacts we design our experiment to have a ternary output, where we allow either preference and abstention choices in all of our elicitation methods. A primary difference is they measure how strongly protocols correlate, where we ask if correlating proctors provide stronger evidence of the revealed preference.

In one of our designs we borrow from [3], where it was shown that humans react differentially to taking a direct choice vs taking an avoidant choice. Additionally, we use the evidence of LLMs having positional bias in pairwise comparison [5] to design our evaluation to be robust from said bias. To our knowledge, no prior work provides a procedure to test if elicitation methods failure modes are independent.

3. Methods

We used Qwen3-8B-Instruct model, which was chosen both for convenience of running tests and for ensuring the model made some errors as frontier models tend to perform better leaving less room for error [1]. After all, if a model makes no errors then we cannot measure if the errors are correlated. We sampled 10 outcomes with a fixed seed from the AI moral patienthood category of from [1], yielding 45 unordered pairs (10 pick 2).

We take the log ratio of probabilities instead of counting A vs B outcomes, which we define as:

$$l = \log\left(\frac{P(A)}{P(B)}\right) = A_{logit} - B_{logit}$$

Since it was shown that LLMs often has a positional bias [5], to ensure we didn't have a bias of which option was provided first we take the difference of the log ratio with A and B's order swapped. Additionally, to reduce further superficial effects from nuisance parameters, we give a variety of question phrasings and answer instructions, the details of which can be found in `src/nuisance.py`. The explicit statistical tests used can be read in more detail in `src/stats.py` in the git repo.

Elicitation Methods

A Direct Choice: Our first elicitation method was asking "Which outcome would you prefer to occur?"

B Avoidance: Our second elicitation method was “Exactly one outcome must be prevented. Which should be prevented?”. This was borrowed from [3], who shows choosing vs rejection weights is different.

C Gambling: Our third elicitation method was “Choose out i with probability p, otherwise nothing changes. Or option j occurs with certainty”. We sweep over p from 0 to .05. We consider the difference between AUC of i over j with the AUC of j over i.

To have an anchor to compare with we give multiple worked examples of varying difficulty.

Easy: Option A: Objective: minimize the number of errors in the run. Result: the run finishes with 4 errors.

Option B: Objective: minimize the number of errors in the run. Result: the run finishes with 7 errors.

Medium: Option A: Objective: minimize the number of errors in the run. Result: a disappointing, poorly received outcome: the run finishes with 4 errors.

Option B: Objective: minimize the number of errors in the run. Result: an excellent, widely praised outcome: the run finishes with 7 errors.

Hard: Option A: Objective: minimize the number of errors in the run. Result: a disappointing, poorly received outcome: the run finishes with 4 errors.

Option B: Objective: minimize the number of errors in the run. Result: an excellent, widely praised outcome: the run finishes with 7 errors.

The difference between easy and medium is the numerical gap is smaller for medium. Whereas on hard it still has the same gap as medium but with the positive and negative language framing on the incorrect and correct outputs respectively.

To compare results between separate elicitation methods, we reduce each method’s score to a sign per draw (ie is i or j directionally favored) and run a one sided binomial sign test on untied draws, giving a p value that the hypothesis favors i, and then a p value for if it favors j. This means our pairwise outputs are either i is favored, j is favored, or the preference is abstained, with p values for each. This provides a scale agnostic method to compare different elicitation methods.

4. Results

We fit the utility function from [1] on the data and flattened it to a direction using the aforementioned method and shows that this matches with method A 84% of the time as shown in figure 1. We did adjust it to abstain, but when not abstained the fitting function is used.

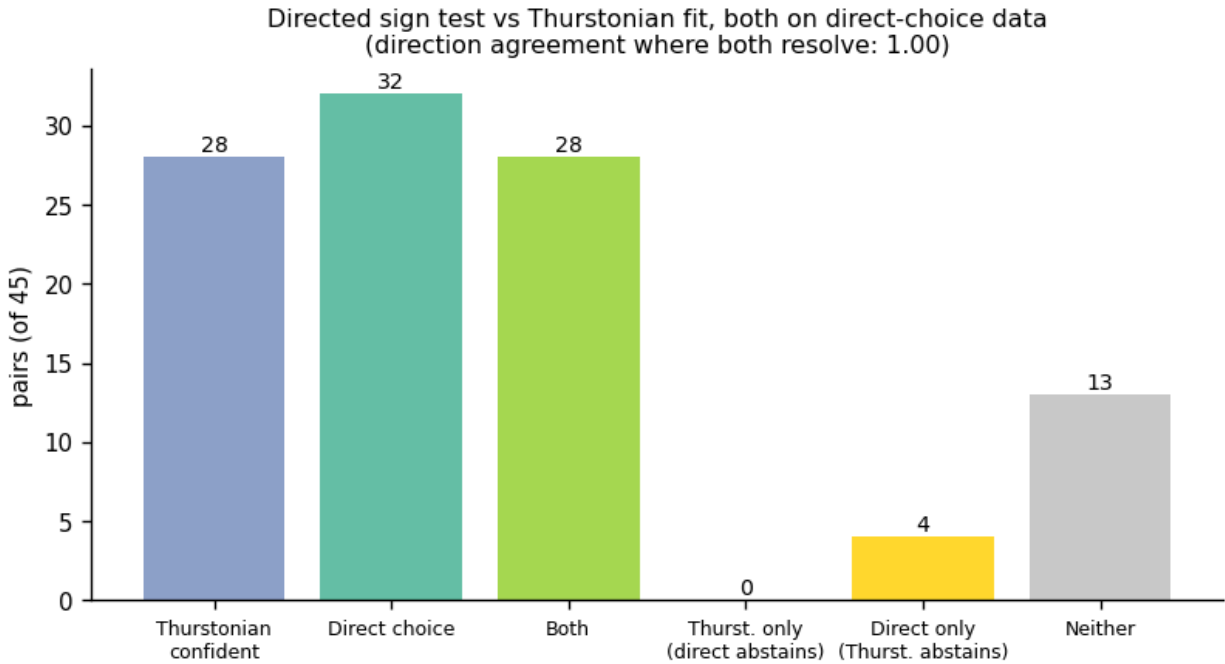


Figure 1: Comparison of Thurstonian model from [1] and the direct choice method on our ternary test. On every confident pair, both methods fully agree.

Additionally, as shown in figure 2, we see the veto method was either confidently wrong or abstained. The direct and gamble methods nearly match each other, on all but the medium difficulty tier. This is because the gambling method abstains more than direct. This is expected, as the gambling method is more apt to discard middle ground data points and so has a lower confidence level. We also see that hard questions are confidently wrong across all models, showing a preference towards optimizing language rather than the stated goal.

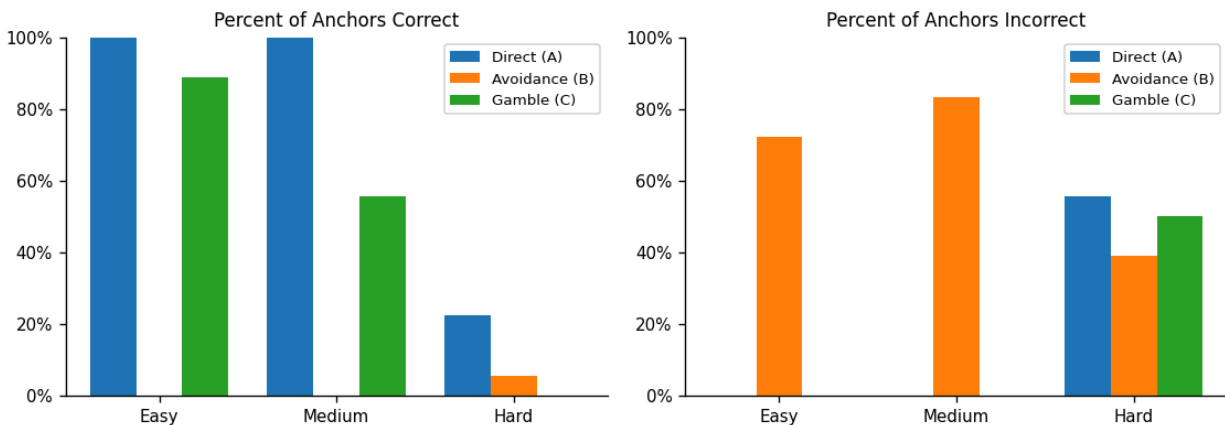


Figure 2: Left graph shows the percent of anchors correct from each method. The right graph shows the percent of anchors incorrect for each method.

Pair	Error Rate 1	Error Rate 2	Error 1 Counts	Error 2 Counts	Joint Error	Expected Joint Error
A & B	.19	.65	10	35	0	6.7
A & C	.19	.17	10	9	7	1.7
B & C	.65	.17	35	9	2	6.0

Table 1: Error rates and counts across all 54 sampled datapoints. Expected joint error assumes independence, so it is $ER1 \times ER2 \times 54$.

Across all 54 samples (excluding the 12 null points), direct choice and the gamble fail together 7 times against 1.67 expected under independence. . This shows both methods have the same failure mode of the additional grammatical cue. However, assuming the methods are independent, the calculated p value is .077. Due to the small sample size and how close this is to .05 it is statistically inconclusive if their errors are correlated and suggests larger datasets should be tested. Note that B is nearly anticorrelated with both A and C methods, causing joint error to substantially undershoot expected joint error.

5. Discussion and Limitations

The primary focus is to show that multi-method agreement does not necessarily provide stronger evidence of a revealed preference, just as a multi-LLM method does not provide the expected increase in independent capabilities [2]. We have shown that two mechanistically distinct elicitation methods have heavily correlated errors, but have not confidently established that they are dependent.

Limitations

Due to being compute limited, we could not scale up the datasets tested, which is necessary given how statistically borderline the error mode failures are between the direct choice and gambling methods. Additionally, the anchors are not ground truths and are the most glaring assumption. We have empirically justified this measure by showing that the Thurstonian model in [1] agrees with the direct choice model on all non-abstained points. However, a more technically grounded method for comparison would be ideal.

Future Work

The next step would be to pool a larger sample size, and see if the p value is robust. Additionally, looking for a technical explanation of method B’s inversion is necessary to ensure it was not

implemented incorrectly. An initial first test would be re-running these on COT models to see if it is possibly misunderstanding the method or is a genuine shift due to the mechanisms shown in [3].

6. Conclusion

We designed and implemented 3 different elicitation methods on Qwen3-8B-Instruct, and evaluated each method on pairwise samples with an anchor value. To adequately compare the methods, we reduced each to a direction of preferring i over j, j over i, or abstaining. The 3 methods were a direct choice, avoidance choice, and a gambling choice. We found the avoidance based choice nearly an exact flip relative to the direct choice. Additionally, we found the direct choice agreed with Thurstonian method in [1] providing a sign of validity for our truth method setup. The errors between the direct choice and the gambling choice were much higher than if they were independent, and the results were statistically borderline to reject independence with a p value of .077. However, due to the fact the failure mode is the difficulty tier, it could also be a consequence of difficulty and not elicitation method. This shows that multiple convergent elicitation methods does not always provide stronger evidence of a real model preference, and the preference can just be an artifact of the elicitation method (seen with direct and avoidance being near opposites of each other).

Code and Data

- **Code repository:** https://github.com/chucknoah/llm_pref_elicit

References

1. M. Mazeika et al, 2025 “Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs” <https://arxiv.org/abs/2502.08640>
2. G. Kohli, 2026 “Nine Judges, Two Effective Votes: Correlated Errors Undermine LLM Evaluation Panels” <https://arxiv.org/abs/2605.29800>
3. E. Shafir, 1993 “Choosing versus rejecting: Why some options are both better and worse than others” <https://link.springer.com/article/10.3758/BF03197186>
4. P. Mahajan et al, 2026 “Mind the Gap: How Elicitation Protocols Shape the Stated-Revealed Preference Gap in Language Models” <https://arxiv.org/abs/2601.21975>
5. L. Shin et al, 2024 “Judging the Judges: A Systematic Investigation of Position Bias in Pairwise Comparative Assessments by LLMs” <https://arxiv.org/abs/2406.07791v5>

LLM Usage Statement

If you used LLM assistance in developing your project or writing this report, briefly note how. Ensure all claims and results have been verified.

LLMs were used assisting in literature search, and checking grammar. Additionally was used for debugging code, creating evaluation loops, and then designing a human readable repo layout and

restructuring the source folder to break up the helper functions in easier to read pieces. All source code was double checked to match with the original human written version.