Cognitive Alignment: Reducing Order Bias in MCQ Answering

Eric Bennett[2], **Su Bao**[1]
[2]Dept. of Industrial Engineering, National Taiwan University
[2]Dept. of Computer Science, University of Maryland, College Park
ebenne92@umd.edu, r13546027@ntu.edu.tw

Abstract

Large language models (LLMs) have achieved impressive performance on multiple-choice question answering tasks in zero-shot settings. However, recent studies reveal that LLM predictions can be sensitive to the order in which answer options are presented, a phenomenon known as *order bias*. In this work, we conduct a systematic analysis of order bias in LLMs using a bilingual (English-Chinese) multiple-choice-question (MCQ) dataset spanning 17 knowledge domains. After removing inconsistent question pairs, we evaluate 1,700 aligned examples under various prompting strategies using GPT-4o and mistral-small. Our findings confirm that order bias affects model outputs, even when semantic content is held constant. We further propose a simple calibration method, Cognitive Alignment, that mitigates this bias across option permutations. This study highlights a critical limitation in current LLM evaluation paradigms and provides practical guidance for designing more robust prompting methods.

1 Introduction

Large Language Models (LLMs) have shown immense promise in solving multiple choice questions in a zero-shot fashion (Kojima et al., 2022). However, LLMs have also shown to exhibit biases across answer choice order. Order bias refers to the phenomenon in which a model's response is influenced not by the semantic content of each answer, but by its position in the list. Recent work by Pezeshkpour and Hruschka (2023) demonstrates that reordering options in multiple-choice questions can lead to substantial variation in accuracy—up to 75% for GPT-4—indicating that current LLMs may not be as objective or consistent as expected, especially in zero- or few-shot evaluation settings. To systematically evaluate order bias, we compile a bilingual MCQ dataset spanning 17 knowledge domains following the categorization of Hendrycks et al. (2021). After removing inconsistent examples across languages, we obtain 1,700 English-Chinese aligned question pairs. We evaluate zero-shot prompting techniques using state-of-the-art LLMs including GPT-4o and mistral-small, comparin Chain-of-Thought(CoT) and Direct Prompting against our proposed novel method.

2 Test Set

For this experiment, we use both English and Chinese versions of the MMLU data set introduced in Hendrycks et al. (2021). For Chinese, we use a translated version of the benchmark made available by OpenAI at `https://openaipublic.blob.core.windows.net/simple-evals/mmlu_ZH-CN.csv`. Throughout the cleaning process we preserve row alignment between the two languages to ensure identical questions are being considered.

2.1 Data Cleaning

A total of 176 rows of our bilingual dataset were found to have different correct answer values for the two languages, which were removed to reduce complexity.

We next divide the data set into subcategories following the structure in Hendrycks et al. (2021). The subcategories are as follows: biology, business, chemistry, computer science, culture, economics, engineering, geography, health, history, law, math, other, philosophy, physics, politics, and psychology. From each subcategory, we sample 100 aligned English-Chinese question pairs, resulting in a total of 1,700 questions.

3 Methods

To compare the results of our novel method, we used a number of baseline values, including both CoT and Direct Prompting. We use GPT-4o (Hurst et al., 2024) to show SoTA zero-shot Direct Prompt-

ing results, and mistral-small (Mistral AI, 2025) for Direct Prompting, CoT, and our novel method.

For each method, answers are generated for 4 permutations of each of the 1,700 questions. The questions are answered with the correct answer being located in A, B, C, and D positions. By contrasting the performance between permutations we generate the results shown in Section 4.

3.1 Cognitive Alignment for MCQ Answering

We present Cognitive Alignment(CA) as a method of fully removing order bias from the process of LLM MCQ answering. The process begins by using the LLM to answer the given question without any access to the provided choices. Next, both the LLM answer and the 4 possible answers are embedded using an embedding model. [1] Finally, the cosine similarity between the embedded LLM answer and the 4 answer choices is calculated, and the answer choice with the highest similarity is selected as the predicted answer.

$$\text{Answer} = \max_{i \in \{1,2,3,4\}} \text{sim}(\text{LLM}, A_i) \quad (1)$$

By never providing the answer choices to the LLM directly, order is fundamentally never considered. This inherently removes the ability for a given order to influence the LLM.

CA-Adjusted MCQ Answering While the pure CA approach completely removes the possibility of order bias, it has a major drawback- a large decrease in accuracy. Our results show pure CA on our English and Chinese datasets have an accuracy of 43% and 38.333% respectively, a drop from Direct Prompting of 38.044% and 35% (respectively). With such a major loss, we propose a second novel method, CA-Adjusted MCQ Answering, to balance the benefits of Direct Prompting accuracy and the reduction in order bias of CA.

CA-Adjusted MCQ Answering works by removing potential answer choices as selections based on what CA judges as "worse" answers. Given a set of similarity scores for A, B, C, D answer choices, any choice within a certain distance (cutoff value) from the highest similarity value are replaced with "DO NOT PICK THIS OPTION" in the dataset, reducing the distractions provided to the LLM when answering.

[1] In our experiment we use the distiluse-base-multilingual-cased-v2 (cite) embedding model because of its training in both English and Chinese.

For example, consider the following:

Question

The soils of which of the following biomes has the highest rate of leaching and cycling of nutrients?

Answer choices

A. Tropical rain forest

B. Tundra

C. Taiga

D. Desert

Free-response answer generated by CA method: *Tropical Rainforest*

Similarity scores generated by CA method

Choice	Similarity Score
A	0.96880
B	0.17935
C	0.19457
D	0.34539

New potential answer choices provided with a cutoff value of 0.459

A. Tropical rain forest

B. DO NOT PICK THIS OPTION

C. DO NOT PICK THIS OPTION

D. DO NOT PICK THIS OPTION

By removing low-confidence answers, the LLM is less likely to choose incorrect answers due to order bias, and by adjusting the cutoff value, we can balance the lower order bias of CA and the better accuracy of Direct Prompting.

We choose the cutoff value 0.459 in our testing as in every case where the difference between the highest and second highest similarity value is greater than or equal to 0.459 there is a 100% accuracy (the correct answer is always chosen in the pure CA strategy in these cases).

4 Results

4.1 Metrics for Order Bias

To quantify the impact of answer permutation on model predictions, we adopt three metrics widely used in recent work: **Relative Standard Deviation (RSD)** (Reif and Schwartz, 2024), **Recall Standard Deviation (RStd)** (Zheng et al., 2024), and **Fluctuation Rate (FR)** (Wei et al., 2024). Each captures a different facet of instability under order variation.

Relative Standard Deviation (RSD). RSD measures the sensitivity of overall accuracy to answer permutations. Let $\{A_1, A_2, \ldots, A_n\}$ denote the model's accuracy under n distinct answer orders. Then

$$\text{RSD} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(A_i - \overline{A})^2}}{\overline{A}}, \tag{2}$$

$$\overline{A} = \frac{1}{n}\sum_{i=1}^{n} A_i. \tag{3}$$

Higher values indicate greater variability in model-level performance across permutations.

Recall Standard Deviation (RStd). RStd captures how individual questions are affected by order shifts. For each question q, let $\{r_{q,1}, \ldots, r_{q,k}\}$ be the binary correctness across k permutations. Then

$$\text{RStd} = \frac{1}{N}\sum_{q=1}^{N} \frac{\sqrt{\frac{1}{k}\sum_{j=1}^{k}(r_{q,j} - \overline{r_q})^2}}{\overline{r_q} + \varepsilon}, \tag{4}$$

$$\overline{r_q} = \frac{1}{k}\sum_{j=1}^{k} r_{q,j}. \tag{5}$$

Here ε is a small constant to prevent division by zero. A larger RStd signals greater sample-level prediction instability.

Fluctuation Rate (FR). FR quantifies how often the predicted answer changes when the option order is permuted. Let f_q be the number of unique predictions for question q across the k permutations:

$$\text{FR} = \frac{1}{N}\sum_{q=1}^{N} \frac{f_q - 1}{k - 1}. \tag{6}$$

FR ranges from 0 (fully stable) to 1 (changes prediction in every permutation), directly reflecting positional sensitivity.

4.2 Overall Results

The overall performance of each method across English and Chinese datasets are summarized in Table 3, and Figure 1 visualizes the results of each method across metrics. Our proposed CA-adjusted MCQ Answering method (specifically with cut-off value 0.2) consistently reduced order bias compared to both Direct Prompting and CoT prompting across all evaluated metrics, only being beaten in RSD by the SoTA Direct Prompting method.

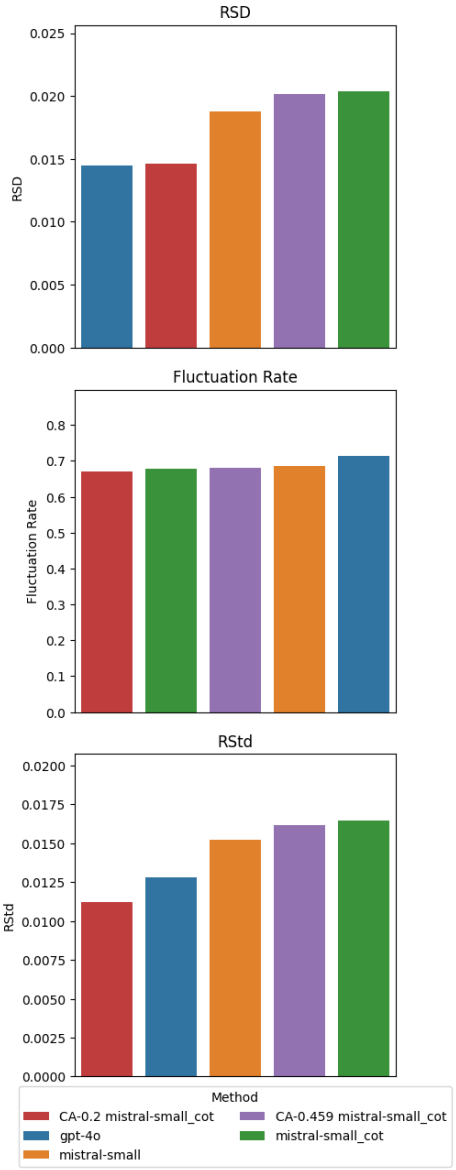

Figure 1: Comparison of the five methods on RSD, RStd, and Fluctuation Rate. Lower values indicate reduced order bias and improved stability.

4.3 Order Sensitivity Across Languages

We conducted our experiment with both English and Chinese to compare the order bias between the two languages. RSD, RStd, and Fluctuation Rate all varied by language across methods (see Figure 2), with RSD and RStd being consistently higher in Chinese than English, and Fluctuation Rate consistently being slightly lower.

Observing RSD and RStd across methods, our SoTA Direct Prompting method exceeds our CA-Adjusted method. This is likely a result of better multilingual capabilities in our SoTA GPT-4o model as opposed to mistral-small. The full results across languages can be found in Table 3

	Overall Accuracy		RSD		RStd		Fluctuation Rate	
Method	**English**	**Chinese**	**English**	**Chinese**	**English**	**Chinese**	**English**	**Chinese**
gpt-4o	**0.882794**	**0.849265**	**0.014487**	**0.019291**	0.012789	**0.016383**	0.713088	0.701176
mistral-small	0.810441	0.733333	0.018762	0.032797	0.015205	0.024051	0.686029	0.658500
mistral-small_cot	0.809412	0.748676	0.020342	0.044451	0.016465	0.033279	0.679265	0.660735
CA-0.2 mistral-small_cot	0.766618	0.698235	0.014643	0.032861	**0.011226**	0.022945	**0.671176**	**0.654412**
CA-0.459 mistral-small_cot	0.804118	0.743824	0.020140	0.042449	0.016195	0.031575	0.679559	0.660147

Table 1: Evaluation of models across English and Mandarin on Overall Accuracy, Relative Standard Deviation (RSD), Recall Standard Deviation (RStd), and Fluctuation Rate (FR). For RSD, RStd, and FR lower scores indicate less order bias

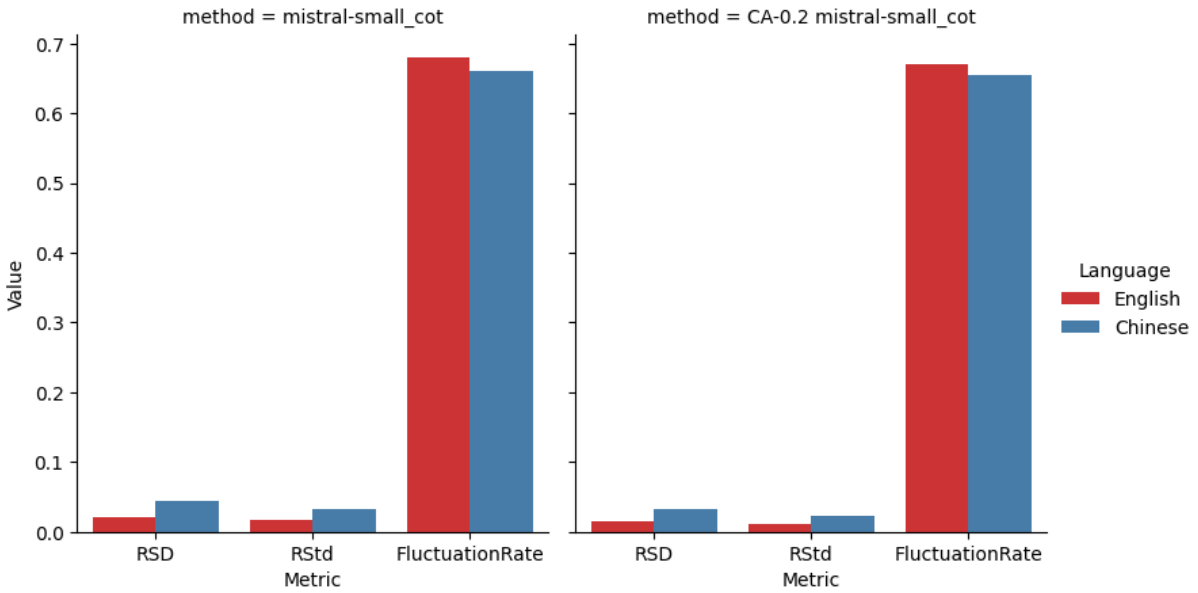

Figure 2: Comparison of English vs. Mandarin for RSD, RStd, and Fluctuation Rate across `mistral-small_cot` and `CA-0.2 mistral-small_cot` methods. Lower RSD, RStd, or Fluctuation Rate values indicate reduced order bias

4.4 Order Sensitivity Across Subcategories

To observe order bias in different knowledge domains, we contrast our metrics across the 17 subcategories defined in Hendrycks et al. (2021), comparing SoTA Direct Prompting, CoT, and CA-adjusted CoT (with cutoff value 0.2 results). Figure 3 illustrates the English RSD values by subcategory for `mistral-small_cot` versus `CA-0.2 mistral-small_cot`.

Across these subcategories, the CA-Adjusted MCQ Answering method achieved lower values on a majority of metrics, improving order bias in 14 out of 17 cases. The subcategories in which CA-adjustment did not improve order bias include computer science, economics, and psychology. The full results across subcategories are found in Table 3

4.5 Confidence vs. Order Bias

Our CA-Adjusted method allows us to control the balance between order-bias reduction and accuracy by adjusting the cutoff value. A higher cutoff value results in less answer choices being removed, retaining accuracy but having little to no effect on order bias. A lower cutoff value removes more answer choices, lowering order bias more, but decreases accuracy as the chances of the correct answer being accidentally removed rise.

In our experiment, CA-Adjusted with cutoff value 0.2 clearly outperformed 0.459 in all recorded metrics as shown in Figure 1, while cutoff value 0.459 was outperformed by Direct Prompting in both RStd and Fluctuation Rate. That being said, Table 2 shows the decreased accuracy of a cutoff value of 0.459 as opposed to 0.2. CA-Adjusted with a 0.459 cutoff value performs 5.04% and 7.09% better than with 0.2 in English and Chinese respectively.

Although the results of this test specifically compared the two cutoff values of 0.2 and 0.459, future research might benefit from exploring the impacts

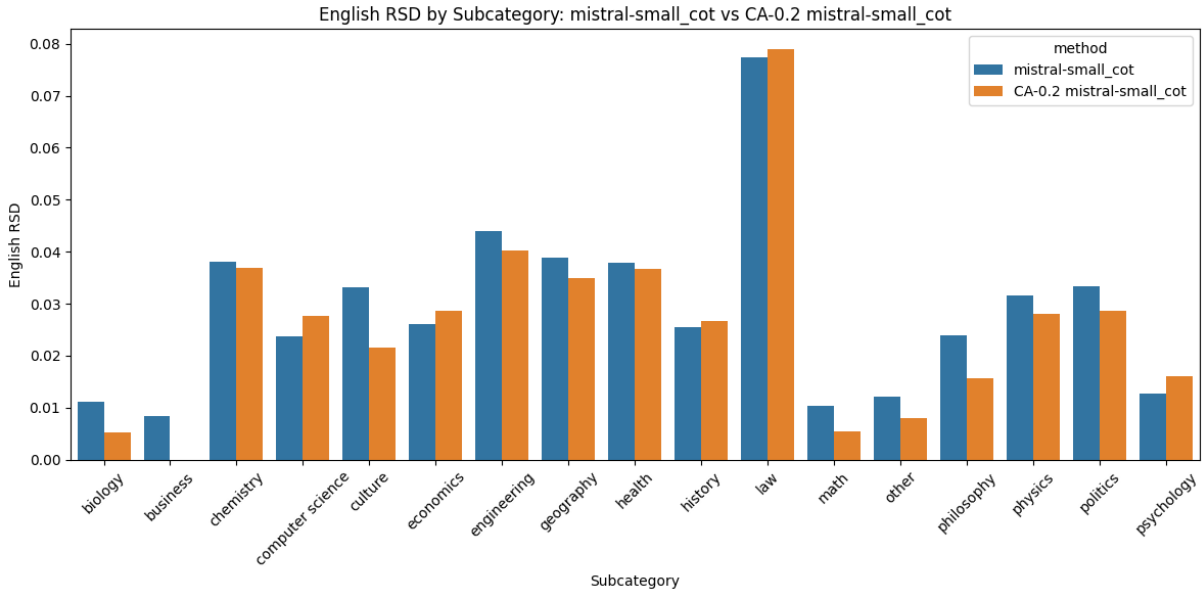

Figure 3: English RSD by Subcategory comparing `mistral-small_cot` (blue) vs. `CA-0.2 mistral-small_cot` (orange). Orange bars higher than blue indicate subcategories in which CA-Adjustment reduced order bias.

of further tuning of the cutoff value, as well as alternate ways of determining extraneous answer choices based on similarity scores.

4.6 Takeaways from Cognitive Alignment

Cognitive Alignment and CA-Adjusted MCQ Answering both were successful in lowering order bias across knowledge domains and languages.

However, we also found a number of obvious drawbacks of the CA approach. For example, in the following question/answer pair we see a complete failure based on the format of the question:

Question

What was 'democratic enlargement'?

Answer choices

A. A proposal for reform of the US system of government

B. A proposal for the extension of democratic rule globally

C. A proposal for the extension of free markets

D. Both b and c

Free-response answer generated by CA method:
Democratic enlargement refers to the process of expanding democratic governance and institutions, often within a region or globally. It involves promoting democracy in countries that are transitioning from authoritarian regimes or have limited democratic practices. This can include efforts such as supporting free elections, strengthening civil society, and fostering human rights.

Similarity scores generated by CA method

Choice	Similarity Score
A	0.1847597062587738
B	0.517923891544342
C	0.19801126420497894
D	-0.06396911293268204

While "Both b and c" is the correct answer, the alignment method cannot detect the similarity between it and the free response answer because the answer choice merely references other answer choices.

Overall, it is clear that this method produces a tradeoff between accuracy and reduced order bias. Our most successful method at reducing order bias, CA-0.2 mistral-small_cot, led to a 5.49% and 7.57% accuracy decrease in English and Chinese respectively. A full comparison of accuracy across CA methods is located in Table 2.

Method	English	Chinese
Pure CA	0.43000	0.38353
CA-0.2 mistral-small_cot	0.77059	0.69706
CA-0.459 mistral-small_cot	0.80941	0.74647
mistral-small_cot	0.81529	0.75412

Table 2: Accuracy of each method on English and Chinese.

4.7 Future Research Paths

While the lackluster accuracy performance of the pure CA method would suggest it shouldn't be used in multiple choice answering as is, the inherent, complete removal of order bias suggests incorporating alignment might still be useful in reducing order bias.

The CA-Adjusted method provides an example of how this could be done, but falls short of a widely implementable method. Future research might refine the method of deciding on specific answer choices to remove to avoid issues like those shown in 4.6, or use other methods of removing problematic answer choices before being feeding the ordered questions to an LLM.

The code and experimental framework used in this paper are available at: `https://github.com/softly-undefined/NLP_final`.

References

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

OpenAI: Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, and et al. 2024. Gpt-4o system card. *arXiv preprint*.

Takeshi Kojima, Shixiang Gu, Machel Reid, Yujia Yao, and Dale Schuurmans. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Mistral AI. 2025. Mistral small 3 (24b). `https://ollama.com/library/mistral-small`. Accessed via `mistral-small:latest`, which pointed to version 24B at time of use.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*.

Yuval Reif and Roy Schwartz. 2024. Beyond performance: Quantifying and mitigating label bias in LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6784–6798, Mexico City, Mexico. Association for Computational Linguistics.

Sheng-Lun Wei, Cheng-Kuang Wu, Hen-Hsen Huang, and Hsin-Hsi Chen. 2024. Unveiling selection biases: Exploring order and token sensitivity in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5598–5621, Bangkok, Thailand. Association for Computational Linguistics.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. *Preprint*, arXiv:2309.03882.

Table 3: Evaluation of models across English and Mandarin on Overall Accuracy, Relative Standard Deviation (RSD), Recall Standard Deviation (RStd), and Fluctuation Rate (FR). For RSD, RStd, and FR lower scores indicate better stability and less sensitivity to answer order permutations.

	Overall Accuracy		**RSD**		**RStd**		**Fluctuation Rate**	
Method	**English**	**Chinese**	**English**	**Chinese**	**English**	**Chinese**	**English**	**Chinese**
Biology								
gpt-4o	**0.920000**	**0.910000**	0.015370	**0.020560**	0.014140	**0.018710**	**0.707500**	0.710000
mistral-small_cot	0.900000	0.872500	0.011110	0.023450	0.010000	0.020460	0.712500	0.685000
CA-0.2 mistral-small_cot	0.837500	0.772500	**0.005170**	0.043300	**0.004330**	0.033450	**0.707500**	**0.667500**
Business								
gpt-4o	**0.837500**	**0.875000**	0.017660	**0.020600**	0.014790	0.018030	**0.677500**	0.700000
mistral-small_cot	0.850000	0.790000	0.008320	0.042930	0.007070	0.033910	0.682500	**0.670000**
CA-0.2 mistral-small_cot	0.840000	0.762500	**0.000000**	0.023410	**0.000000**	**0.017850**	0.690000	0.672500
Chemistry								
gpt-4o	**0.825000**	**0.830000**	**0.018180**	**0.014760**	**0.015000**	**0.012250**	0.710000	0.705000
mistral-small_cot	0.720000	0.700000	0.038040	0.039120	0.027390	0.027390	0.672500	0.662500
CA-0.2 mistral-small_cot	0.647500	0.662500	0.036830	0.028990	0.023850	0.019200	**0.660000**	**0.657500**
Computer science								
gpt-4o	**0.882500**	**0.885000**	**0.016760**	0.025890	**0.014790**	0.022910	0.712500	0.725000
mistral-small_cot	0.790000	0.740000	0.023680	0.039400	0.018710	0.029150	0.672500	**0.667500**
CA-0.2 mistral-small_cot	0.770000	0.727500	0.027550	**0.024540**	0.021210	**0.017850**	**0.662500**	0.677500
Culture								
gpt-4o	0.847500	**0.915000**	**0.009780**	**0.039020**	**0.008290**	0.035710	0.717500	0.702500
mistral-small_cot	**0.862500**	0.810000	0.033180	0.057240	0.028610	0.046370	0.702500	0.687500
CA-0.2 mistral-small_cot	0.827500	0.700000	0.021580	0.050510	0.017850	**0.035360**	**0.692500**	**0.670000**
Economics								
gpt-4o	**0.872500**	**0.907500**	0.029640	0.032480	0.025860	0.029470	0.712500	0.720000
mistral-small_cot	0.832500	0.742500	**0.026010**	**0.025860**	**0.021650**	**0.019200**	0.707500	0.695000
CA-0.2 mistral-small_cot	0.780000	0.732500	0.028670	0.031090	0.022360	0.022780	**0.682500**	**0.692500**
Engineering								
gpt-4o	**0.785000**	**0.827500**	0.052910	0.053620	0.041530	0.044370	0.667500	0.695000
mistral-small_cot	0.762500	0.687500	0.043870	0.086280	0.033450	0.059320	0.650000	0.640000
CA-0.2 mistral-small_cot	0.680000	0.622500	**0.040270**	**0.048690**	**0.027390**	**0.030310**	**0.625000**	**0.635000**
Health								
gpt-4o	**0.807500**	**0.815000**	0.045820	**0.020350**	0.037000	**0.016580**	0.687500	0.695000
mistral-small_cot	0.770000	0.740000	0.037860	0.054890	0.029150	0.040620	0.652500	0.625000
CA-0.2 mistral-small_cot	0.735000	0.667500	**0.036630**	0.061200	**0.026930**	0.040850	**0.650000**	**0.602500**
History								
gpt-4o	**0.885000**	**0.920000**	**0.024630**	**0.007690**	0.021790	**0.007070**	0.717500	0.725000
mistral-small_cot	0.830000	0.770000	0.025560	0.042080	0.021210	0.032400	0.707500	**0.680000**
CA-0.2 mistral-small_cot	0.775000	0.735000	0.026600	0.020410	**0.020620**	0.015000	**0.697500**	0.697500
Geography								
gpt-4o	**0.882500**	**0.907500**	0.040360	**0.004770**	0.035620	**0.004330**	0.702500	0.732500
mistral-small_cot	0.842500	0.772500	0.038800	0.055960	0.032690	0.043230	**0.690000**	0.677500
CA-0.2 mistral-small_cot	0.842500	0.725000	**0.034980**	0.062830	**0.029470**	0.045550	0.695000	**0.657500**
Law								
gpt-4o	**0.675000**	**0.747500**	**0.055920**	**0.043730**	**0.037750**	**0.032690**	0.667500	0.670000
mistral-small_cot	0.622500	0.520000	0.077350	0.129720	0.048150	0.067450	0.625000	0.620000
CA-0.2 mistral-small_cot	0.562500	0.417500	0.078880	0.099300	0.044370	0.041460	**0.610000**	**0.592500**
Math								
gpt-4o	**0.917500**	**0.927500**	0.009040	**0.008940**	0.008290	**0.008290**	0.722500	0.727500
mistral-small_cot	0.802500	0.775000	0.010330	0.049560	0.008290	0.038410	**0.645000**	0.630000
CA-0.2 mistral-small_cot	0.792500	0.775000	**0.005460**	0.042310	**0.004330**	0.032790	**0.645000**	**0.627500**
Philosophy								
gpt-4o	**0.807500**	**0.847500**	0.022110	**0.021070**	0.017850	**0.017850**	0.700000	0.715000
mistral-small_cot	0.755000	0.665000	0.023880	0.093610	0.018030	0.062250	**0.652500**	**0.595000**
CA-0.2 mistral-small_cot	0.715000	0.622500	**0.015640**	0.090430	**0.011180**	0.056290	**0.652500**	**0.595000**
Physics								
gpt-4o	**0.910000**	**0.932500**	**0.015540**	**0.019150**	**0.014140**	**0.017850**	0.722500	0.730000
mistral-small_cot	0.817500	0.760000	0.031630	0.026320	0.025860	0.020000	0.672500	0.667500
CA-0.2 mistral-small_cot	0.732500	0.687500	0.027940	0.027930	0.020460	0.019200	**0.655000**	**0.662500**
Politics								
gpt-4o	**0.820000**	**0.867500**	0.039520	0.022140	0.032400	0.019200	0.692500	0.710000
mistral-small_cot	0.820000	0.750000	0.033400	0.049890	0.027390	0.037420	0.685000	0.655000

CA-0.2 mistral-small_cot	0.797500	0.690000	**0.028560**	**0.017750**	**0.022780**	**0.012250**	**0.682500**	**0.647500**
Psychology								
gpt-4o	**0.875000**	**0.952500**	0.018950	**0.013640**	0.016580	**0.012990**	0.700000	0.735000
mistral-small_cot	**0.875000**	0.820000	**0.012780**	0.028600	**0.011180**	0.023450	0.702500	**0.687500**
CA-0.2 mistral-small_cot	0.807500	0.785000	0.016090	0.026260	0.012990	0.020620	**0.687500**	**0.687500**
Other								
gpt-4o	0.887500	**0.940000**	0.012280	**0.007520**	0.010900	**0.007070**	**0.702500**	0.725000
mistral-small_cot	**0.907500**	0.812500	0.012010	0.040240	0.010900	0.032690	0.715000	0.687500
CA-0.2 mistral-small_cot	0.890000	0.785000	**0.007950**	0.051350	**0.007070**	0.040310	0.715000	**0.682500**
	Overall Accuracy		**RSD**		**RStd**		**Fluctuation Rate**	
Method	**English**	**Chinese**	**English**	**Chinese**	**English**	**Chinese**	**English**	**Chinese**