NLP 2025 Final Project Report

Szu-Ju Chen B10705005 **Jia-Cian Li B10902074**

Abstract

In this final project, we investigate order bias in multilingual multiple-choice question answering using large language models (LLMs). Our goal is to examine how the position of answer choices affects model predictions and what linguistic or cultural factors influence this sensitivity. We focus on order sensitivity across languages with different syntactic structures and explore the impact of question domains, language typology, and model confidence. Through a series of controlled experiments, we evaluate LLM performance under varied configurations to better understand the relationship between answer order and model behavior.

1 Introduction

To investigate order sensitivity across languages, we selected four languages representing diverse linguistic settings: Arabic (VSO), German (SVO), Japanese (SOV), and Swahili (as a representative low-resource language). For evaluation, we used 17 subtasks from the MMLU OpenAI (2023) benchmark and selected questions of comparable difficulty across subjects—preferably at the high school level. When a subtask did not offer questions at this level, we randomly selected one to ensure a broad and balanced coverage.

Due to resource constraints and limitations in free API access, our experiments were conducted using two open-access large language models: Google DeepMind (2024) Gemini and Mistral AI (2024) Mistral. These models were evaluated by OpenAI (2024) simple-evals repo to analyze the impact of answer choice order on their multilingual multiple-choice question answering performance.

To assess order sensitivity, we employed three evaluation strategies. First, we used the original option order from the dataset as the baseline. Second, we randomly shuffled the four answer choices to evaluate the model's robustness to unordered input. Third, we applied a systematic rotation of the answer choices—generating four total configurations—and averaged the model's accuracy across all rotations to assess performance consistency. Due to API limitations and time constraints, the third approach was applied only to Gemini.

2 Common Requirements

2.1 Order Sensitivity across Languages

This subsection focuses on the impact of answer choice order on model performance across different language settings. As shown in Table 1, comparing the original, shuffled, and rotated choice orders, we observe that accuracy generally remains stable or slightly decreases. Notably, the shuffled setting on Mistral shows a consistent and biggest drop in performance across all four languages: Arabic drops by 0.0029, German by 0.0071, Japanese by 0.0317, and Swahili by 0.0235. This suggests that Mistral is more sensitive to choice order perturbations than Gemini.

Among the four languages, Swahili, a low-resource language, performs the worst overall and also shows the largest performance gap between the original and perturbed settings. This implies that order sensitivity may be amplified in low-resource or less well-aligned languages.

Overall, the results indicate that order sensitivity varies across languages and models, with language typology and resource availability playing significant roles in robustness to option reordering.

Language	Ordinary (Gemini)	Ordinary (Mistral)	Shuffle (Gemini)	Shuffle (Mistral)	Rotate (Gemini)
Arabic	0.8600	0.7923	0.8629	0.7894	0.8555
German	**0.8641**	**0.8423**	**0.8647**	**0.8352**	**0.8586**
Japanese	0.8541	0.8205	0.8465	0.7888	0.8508
Swahili	0.8118	0.6317	0.8018	0.6082	0.7982

Table 1: Order sensitivity across languages

2.2 Order Sensitivity across subtasks

This subsection examines how order sensitivity varies across different subtasks in the four selected languages. The results are summarized in Tables 2, 3, 4, and 5.

Across all languages, top-performing subjects are relatively consistent—most notably High School Computer Science and *High School Government and Politics. For example, Arabic and German both rank High School Psychology and Government and Politics among their highest-scoring subjects, while Swahili and Japanese show strong results in Computer Science and Mathematics. This consistency suggests that subject matter, rather than language, may be a stronger factor in robustness to option reordering.

Conversely, Global Facts is the weakest subtask for Arabic, German, and Japanese across both Gemini and Mistral, while Electrical Engineering is lowest in Swahili under Mistral. Performance gaps between best and worst subtasks reach up to 0.3, indicating substantial variability.

Further analysis of answer order configurations reveals that subtasks such as Electrical Engineering, Global Facts, Geography, Physics, Marketing, and Sociology show greater sensitivity to option order. These tasks may rely more on real-world context or ambiguous distractors. In contrast, subjects like European History, Government and Politics, and Mathematics remain stable across configurations, suggesting lower order sensitivity due to more structured reasoning.

In summary, certain domains—especially those involving factual recall or logical reasoning—exhibit greater robustness to option reordering, while more interpretive or open-ended subjects are more prone to order-induced performance fluctuations.

Subject	Ordinary (Gemini)	Ordinary (Mistral)	Shuffle (Gemini)	Shuffle (Mistral)	Rotate (Gemini)
Electrical Engineering	0.77	0.63	0.76	0.53	0.72
Global Facts	0.60	0.63	0.64	0.59	0.61
High School Biology	0.91	0.82	0.92	0.85	0.90
High School Chemistry	0.82	0.72	0.90	0.77	0.85
High School Computer Science	0.94	**0.91**	0.93	0.87	0.92
High School European History	0.86	0.80	0.85	0.79	0.84
High School Geography	0.83	0.81	0.81	0.79	0.83
High School Government and Politics	0.96	0.88	**0.96**	0.90	**0.96**
High School Mathematics	0.94	0.83	0.90	0.82	0.93
High School Microeconomics	0.90	0.83	0.92	0.85	0.92
High School Physics	0.83	0.75	0.87	0.73	0.86
High School Psychology	**0.97**	0.88	0.95	**0.91**	**0.96**
International Law	0.85	0.76	0.88	0.86	0.85
Logical Fallacies	0.82	0.71	0.80	0.74	0.81
Marketing	0.90	0.84	0.88	0.81	0.89
Nutrition	0.86	0.83	0.88	0.84	0.88
Sociology	0.86	0.84	0.82	0.77	0.84

Table 2: Order sensitivity across subtasks in **Arabic**

Subject	**Ordinary (Gemini)**	**Ordinary (Mistral)**	**Shuffle (Gemini)**	**Shuffle (Mistral)**	**Rotate (Gemini)**
Electrical Engineering	0.83	0.74	0.80	0.71	0.78
Global Facts	0.61	0.63	0.60	0.64	0.60
High School Biology	0.91	0.82	0.88	0.84	0.90
High School Chemistry	0.86	0.83	0.86	0.82	0.87
High School Computer Science	0.93	**0.98**	**0.95**	**0.95**	**0.95**
High School European History	0.84	0.82	0.87	0.82	0.83
High School Geography	0.85	0.84	0.85	0.82	0.85
High School Government and Politics	0.93	0.91	0.94	0.93	0.92
High School Mathematics	0.92	0.85	0.92	0.84	0.93
High School Microeconomics	0.92	0.84	0.94	0.86	0.92
High School Physics	0.86	0.81	0.83	0.82	0.83
High School Psychology	**0.96**	0.96	0.93	0.91	0.93
International Law	0.87	0.84	0.89	0.87	0.85
Logical Fallacies	0.83	0.81	0.80	0.81	0.85
Marketing	0.89	0.90	0.90	0.88	0.91
Nutrition	0.86	0.88	0.90	0.84	0.88
Sociology	0.82	0.86	0.84	0.84	0.83

Table 3: Order sensitivity across subtasks in **German**

Subject	**Ordinary (Gemini)**	**Ordinary (Mistral)**	**Shuffle (Gemini)**	**Shuffle (Mistral)**	**Rotate (Gemini)**
Electrical Engineering	0.80	0.67	0.76	0.66	0.76
Global Facts	0.59	0.66	0.55	0.61	0.59
High School Biology	0.92	0.90	0.89	0.77	0.89
High School Chemistry	0.88	0.74	0.84	0.80	0.85
High School Computer Science	0.94	**0.95**	0.92	**0.94**	0.95
High School European History	0.80	0.75	0.79	0.74	0.79
High School Geography	0.84	0.81	0.78	0.75	0.81
High School Government and Politics	**0.96**	0.93	**0.98**	0.91	**0.96**
High School Mathematics	0.93	0.91	0.93	0.90	0.94
High School Microeconomics	0.91	0.89	0.91	0.82	0.91
High School Physics	0.80	0.81	0.85	0.78	0.85
High School Psychology	0.93	0.83	0.95	0.81	0.93
International Law	0.85	0.82	0.86	0.79	0.85
Logical Fallacies	0.79	0.76	0.85	0.76	0.82
Marketing	0.90	0.85	0.87	0.82	0.92
Nutrition	0.88	0.86	0.90	0.81	0.89
Sociology	0.80	0.81	0.76	0.74	0.80

Table 4: Order sensitivity across subtasks in **Japanese**

Subject	**Ordinary (Gemini)**	**Ordinary (Mistral)**	**Shuffle (Gemini)**	**Shuffle (Mistral)**	**Rotate (Gemini)**
Electrical Engineering	0.67	0.46	0.63	0.34	0.61
Global Facts	0.60	0.56	0.57	0.52	0.61
High School Biology	0.90	0.70	0.85	0.70	0.87
High School Chemistry	0.74	0.53	0.75	0.65	0.79
High School Computer Science	0.92	**0.80**	**0.96**	**0.77**	**0.92**
High School European History	0.79	0.63	0.78	0.55	0.77
High School Geography	0.84	0.63	0.85	0.68	0.83
High School Government and Politics	0.89	0.67	0.89	0.66	0.89
High School Mathematics	**0.94**	0.73	0.89	0.75	0.91
High School Microeconomics	0.91	0.61	0.94	0.54	0.90
High School Physics	0.84	0.61	0.81	0.51	0.79
High School Psychology	0.86	0.54	0.90	0.64	0.88
International Law	0.82	0.71	0.82	0.67	0.81
Logical Fallacies	0.63	0.53	0.65	0.49	0.65
Marketing	0.81	0.65	0.78	0.64	0.80
Nutrition	0.83	0.74	0.81	0.61	0.81
Sociology	0.81	0.64	0.75	0.62	0.75

Table 5: Order sensitivity across subtasks in **Swahili**

2.3 Confidence vs. Order Bias

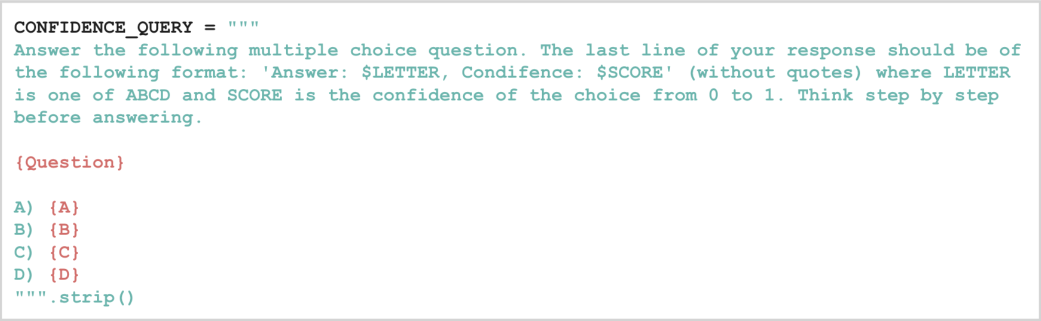
This section explores the relationship between model confidence and its sensitivity to the order of answer choices. Due to limitations in API usage and the unstable behavior of the HuggingFace Mistral model when extracting token-level probabilities, we modified the prompt to ask the model to explicitly output both its selected answer and its associated confidence, as shown in Fig 1

Table 6 presents the resulting answer accuracy and average self-reported confidence scores for German-language tasks under both original and shuffled answer orders, using Gemini and Mistral. We focused on German because it yielded the highest overall performance among the selected languages.

For Gemini, both accuracy and confidence drop slightly when the answer choices are shuffled. This indicates a correlation between lower confidence and reduced accuracy, suggesting that Gemini's self-reported confidence may reasonably reflect uncertainty caused by input perturbations.

In contrast, Mistral shows a different pattern: although its accuracy decreases under shuffled conditions, its reported confidence remains virtually unchanged. This suggests that Mistral may exhibit overconfidence, maintaining high certainty even when its accuracy degrades.

These findings indicate that confidence scores may only partially reflect order sensitivity. While some correlation is observed in Gemini, the mismatch between confidence and accuracy in Mistral highlights the need for caution when interpreting confidence as a proxy for model reliability in the presence of input perturbations.



```
CONFIDENCE_QUERY = """
Answer the following multiple choice question. The last line of your response should be of
the following format: 'Answer: $LETTER, Condifence: $SCORE' (without quotes) where LETTER
is one of ABCD and SCORE is the confidence of the choice from 0 to 1. Think step by step
before answering.

{Question}

A) {A}
B) {B}
C) {C}
D) {D}
""".strip()
```

Figure 1: Prompt template to get model answer confidence

	Ordinary (Gemini)	**Ordinary (Mistral)**	**Shuffle (Gemini)**	**Shuffle (Mistral)**
Accuracy	0.8679	0.8616	0.8548	0.8474
Confidence	0.8584	0.9451	0.8489	0.9460

Table 6: Confidence and accuracy on German tasks

3 Deeper Requirements

3.1 Syntactic Order and Order Bias

Building on the previous analysis of multilingual performance, this subsection investigates whether a language's dominant syntactic structure—specifically word order types like SVO, SOV, and VSO—correlates with model robustness to answer choice reordering.

As shown in Table 1, we analyze four languages: Arabic (SVO), German (SOV), Japanese (VSO), and Swahili (low-resource SVO). Among these, German (SOV) consistently achieves the highest accuracy across all settings and models, suggesting that SOV structures may promote more stable performance. This could be due to syntactic rigidity or its relative alignment with English-language training corpora.

In contrast, Arabic (SVO) and Japanese (VSO) show moderate robustness, while Swahili—a low-resource SVO language—shows the lowest accuracy and greatest sensitivity to ordering changes. This indicates that syntactic order alone is not determinative, and that resource availability and training data alignment play a more decisive role.

While word order typology may offer some explanatory power—especially in comparing high-resource SOV (German) with others—our findings emphasize that linguistic structure must be interpreted in conjunction with resource-related factors to fully understand model behavior under choice permutation.

3.2 Cultural Context and Order Bias

This subsection analyzes model performance on the High School European History to examine how linguistic and cultural background influence answer ordering bias. The relatively higher performance in Arabic and German can be attributed to both linguistic familiarity and cultural proximity to European historical content. German, as a European language, directly overlaps with the subject matter, while Arabic shares significant historical intersections with Europe, such as through the Crusades, the Ottoman Empire, and colonial interactions. In contrast, Japanese and Swahili, which are more culturally and linguistically distant from the European context, show lower performance. This suggests that cultural alignment between the task domain and linguistic background can reduce order bias and enhance model comprehension in content-rich subjects.

Subject	**Ordinary (Gemini)**	**Ordinary (Mistral)**	**Shuffle (Gemini)**	**Shuffle (Mistral)**	**Rotate (Gemini)**
German	0.84	0.82	0.87	0.82	0.83
Arabic	0.86	0.80	0.85	0.79	0.84
Japanese	0.80	0.75	0.79	0.74	0.79
Swahili	0.79	0.63	0.78	0.55	0.77

Table 7: Order sensitivity across languages on **High School European History**

3.3 Prompt

This subsection examines the impact of Few-Shot prompting on model performance across different languages. We constructed a few-shot prompt by adding two example question-answer pairs to the beginning of each input, providing the model with in-context learning signals. These examples were randomly selected from the same subtask's test data. The prompt format is illustrated in Fig. 2. We applied this setup using the original multiple-choice order and evaluated it on the Gemini model.

As shown in Table 8, performance consistently improves in the Few-Shot setting. Arabic shows the largest gain, with an improvement of 0.0217, followed by Swahili with 0.0123, German with 0.0082, and Japanese with 0.0064. This indicates that Gemini benefits from in-context examples, particularly in languages with more complex or resource-limited structures.

```
CONFIDENCE_QUERY = """
Answer the following multiple choice question. The last line of your response should be of
the following format: 'Answer: $LETTER' (without quotes) where LETTER is one of ABCD.
Think step by step before answering.

Example 1:
{Question1}

A) {A1}
B) {B1}
C) {C1}
D) {D1}
Answer: {Ans1}

Example 2:
{Question2}

A) {A2}
B) {B2}
C) {C2}
D) {D2}
Answer: {Ans2}

Now, answer this question:
{Question}

A) {A}
B) {B}
C) {C}
D) {D}
""".strip()
```

Figure 2: Few-Shot Prompt Template

	Zero-Shot	**Few-Shot**
Arabic	0.8600	0.8682
German	0.8641	0.8858
Japanese	0.8541	0.8605
Swahili	0.8118	0.8241

Table 8: Performance using GEMINI on Different Prompt Template

4 Conclusion

In this final project, we explored the impact of answer choice order on multilingual multiple-choice question answering using large language models. Through systematic experiments with Gemini and Mistral across four diverse languages—Arabic, German, Japanese, and Swahili—we examined how linguistic structure, subject domain, confidence, and prompting affect model robustness.

In Section 2.1, we found that German achieved the most stable performance, likely due to its SOV structure and strong training data alignment. Swahili, as a low-resource language, showed the largest performance drop when answer order was changed, indicating that resource availability plays a more critical role than syntax alone. Section 2.2 showed that factual and structured subjects like Mathematics and Government were more robust to answer reordering, while open-ended or interpretive domains such as Global Facts and Sociology were more prone to fluctuations. In Section 2.3, we observed that Gemini's confidence generally aligned with its accuracy, while Mistral tended to remain overconfident despite reduced performance.

In Section 3.1, we investigated syntactic order and found that while SVO languages like German performed well, syntax alone does not fully explain order sensitivity—resource alignment is key. Section 3.2 highlighted that cultural familiarity improves performance, as seen in better results for German and Arabic on European History tasks. Finally, Section 3.3 showed that few-shot prompting consistently improves performance, especially in low-resource languages like Swahili.

Overall, our findings demonstrate that answer order bias is influenced by a combination of linguistic, cultural, and technical factors. Addressing these issues is essential for building fairer and more robust multilingual NLP systems.

Author Contributions

Szu-Ju Chen: Mistral experiments, order and prompt experiment code
Jia-Cian Li: Gemini experiments, evaluation code

References

Google DeepMind. Gemini api. https://aistudio.google.com/, 2024.

Mistral AI. Mistral api. https://console.mistral.ai/, 2024.

OpenAI. Massive multitask language understanding (mmlu) dataset. https://huggingface.co/datasets/openai/MMMLU, 2023.

OpenAI. Simple-evals: Evaluation framework for llms. https://github.com/openai/simple-evals, 2024.