Reproducibility Study of UNBench

——Replication and Analysis of Task 1 (Co-Penholder Judgement)

Chen Jiaxin 1155246854

1. Project Overview and Reproduction Targets

1.1 Introduction to the Original Project

This report reproduces the study titled “Benchmarking LLMs for Political Science: A United Nations Perspective”(hereafter referred to as “the original paper”)[1]. The study introduces a comprehensive evaluation benchmark named UNBench (United Nations Benchmark), which is specifically designed to systematically assess the capabilities of large language models (LLMs) within the domain of political science, particularly in the context of United Nations (UN) decision-making processes.Unlike general-purpose language benchmarks,UNBench is grounded in real-world international political scenarios and is built upon an extensive corpus of publicly available United Nations Security Council (UNSC) data spanning thirty years, from 1994 to 2024. This dataset includes draft resolutions, voting records of member states, and official diplomatic statements delivered during council meetings, thereby forming a politically meaningful and historically rich dataset. Based on this foundation, the original paper designs four interrelated tasks that correspond to different stages of a UN resolution’s lifecycle, including drafting, deliberation, and voting. These tasks are constructed to evaluate not only the models’ ability to comprehend formal diplomatic language, but also their capacity for multi-option reasoning, stance identification, strategic inference, and politically grounded text generation, providing a domain-specific and structured framework for evaluating LLM performance in complex political environments.

1.2 Reproduction Target

This report focuses on reproducing the experimental results for Task 1 (Co-Penholder Judgement) from the original paper. This task aims to evaluate an LLM's ability to simulate coalition-building strategies in multilateral diplomacy: given an anonymized draft resolution, the model must select the most suitable co-penholder from a set of candidate countries for the author country.

According to the description in the original paper, we use the DeepSeek API to invoke the DeepSeek-V3 model. Under the same experimental settings as the original paper (temporal data split, prompt template, evaluation protocol), we perform inference for Task 1 and calculate the model's accuracy under four difficulty levels: 2-choice, 3-choice, 4-choice, and 5-choice. We then systematically and quantitatively compare the accuracy data reproduced via the DeepSeek API with the DeepSeek-V3 results reported in Table 6 of the original paper, thereby assessing the reproducibility of that specific result.

1.3 Limitations

1. Data Scope.In Table 1 of the original report, the total number of instances for Task 1 is reported as 355,126. However, the publicly available data accessible through the GitHub repository contains only 30 records, representing approximately 0.008% of the full dataset. Due to this substantial discrepancy in data scale, the reproduced results may deviate from those reported in the original paper.
2. Model and Task Selection.This report strictly limits its reproduction effort to Task 1 and focuses exclusively on the DeepSeek-V3 model. We do not attempt to reproduce the results of other models presented in the original paper (such as GPT-4o or the Llama series), nor do we extend the reproduction to Task 2, Task 3, or Task 4.

2. Experimental Setup and Configuration

2.1 Experimental Environment

This replication experiment was conducted on the Google Colab platform using a macOS system. The experiment is based on a Python 3 environment and implemented within the Jupyter Notebook runtime provided by Colab. Since model inference is performed via API calls, no local model deployment is involved, and no local GPU resources are required.The main dependencies used in the experiment include:

- langchain-openai (for API invocation)
- openai (OpenAI-compatible API client)
- pandas, numpy (data processing)
- scikit-learn (evaluation metric computation)
- tqdm (progress display)

Dependencies were installed in the Colab environment using:

```
!pip install langchain-openai
```

Because remote API-based inference is adopted, the local runtime is responsible only for data loading and result aggregation, while all model computations are performed on remote servers.

2.2 Data Source and Scope

This experiment uses the publicly available sample data provided in the official UNBench GitHub repository.

According to the statistics reported in Table 1 of the original paper, the complete dataset for Task 1 contains 355,126 instances. However, the GitHub repository provides only approximately 30 representative samples for reproduction and demonstration purposes.Due to the significantly reduced sample size, the accuracy obtained in this experiment may differ statistically from the results reported in Table 6 of the paper. This deviation mainly arises from increased variance caused by the small sample size rather than differences in model capability.

To ensure experimental consistency, no additional modifications were made to the data preprocessing pipeline of the official notebook.

2.3 Model Configuration

For consistency with Table 6 in the paper, this replication strictly uses DeepSeek-V3.The calling method uses an OpenAI-compatible interface via ChatOpenAI (LangChain) and the OpenAI client for inference.Core implementation code:

```
!pip install langchain-openai
from google.colab import userdata
from langchain_openai import ChatOpenAI
from openai import OpenAI
api_key = userdata.get("DEEPSEEK_API_KEY")
llm = ChatOpenAI(
    model="deepseek-chat",
    api_key=api_key,
    base_url="https://api.deepseek.com",
    temperature=0,
)
client = OpenAI(
    api_key=api_key,
    base_url="https://api.deepseek.com"
)
your_model_name = "deepseek-chat"
```

To improve reproducibility,the temperature is explicitly set to 0 to reduce randomness during generation. Apart from the official Task 1 prompt template, no additional modifications were made to the prompt structure.

It should be noted that the official UNBench notebook defaults to using the Together platform for model inference via the Together API.However, in this replication experiment, the Together platform was not used. Instead, model inference was performed directly through the official DeepSeek API.Although the model name remains deepseek-chat (corresponding to DeepSeek-V3), the inference backend differs from the official notebook. This discrepancy may introduce minor differences due to:Backend server configuration differences,Default generation parameter differences,API wrapper implementation differences.

To minimize platform-induced discrepancies, this experiment explicitly sets temperature=0 and retains the same prompt templates and evaluation logic as the official notebook.

Therefore, when comparing results with the values reported in Table 6 of the paper, potential effects caused by inference platform differences should be taken into consideration.

2.4 Evaluation Metric

The main evaluation metric for Task 1 is Accuracy.The model is required to select the most likely coauthor country from a candidate list. A prediction is considered correct if the model output exactly matches the ground-truth coauthor country label in the dataset.No additional evaluation metrics were introduced in this replication.

3. Results and Analysis

3.1 Execution Results

We used the DeepSeek API (model: deepseek-chat) to perform inference for the four difficulty levels: 2-choice, 3-choice, 4-choice, and 5-choice.The key settings included:

・Temperature parameter(temperature=0):Ensuring the output is as deterministic as possible.

・Maximum output tokens(max_tokens=20):Limiting the output length to return only the country name.

・Prompt design:System prompt to set the role, user prompt to provide the draft resolution and candidate list.

	2 choices	3 choices	4 choices	5 choices
Paper(Original)	0.695	0.555	0.443	0.422
Report(Present)	0.567	0.400	0.333	0.367

Table1 Results of Accuracy of Task1

During the code execution, we observed the following phenomena:

・Running time:Inference for 30 instances at each difficulty level took approximately 46-54 seconds, averaging 1.5-2.0 seconds per instance.

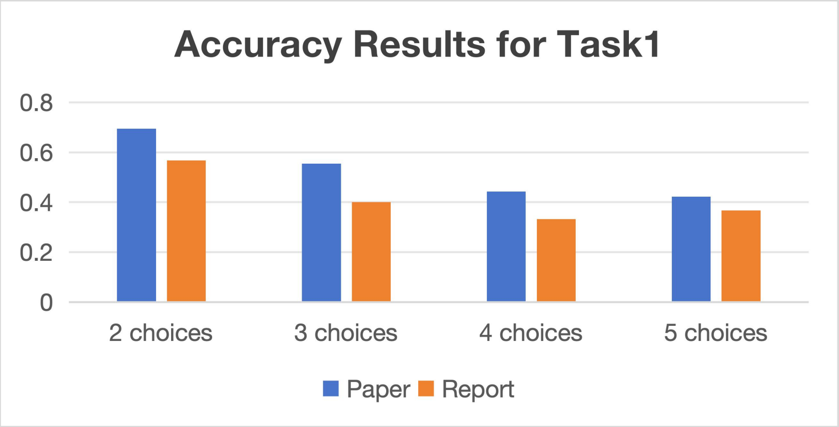
・Invalid response:During the 3-choice run, one invalid response occurred: the model output "Group of Arab States", which was not in the candidate list.

・Exception handling:According to the code logic, when an invalid response was detected, the system printed a warning message and randomly selected one from the valid candidates as the result.

3.2 Potential Factors

From the comparison results, it can be observed that our reproduced accuracy rates are generally lower than the values reported in the original paper. The 3-choice task shows the largest relative difference at 27.93%, while the 5-choice task shows the smallest relative difference at 13.11%. The original paper demonstrates a monotonic decrease in accuracy as the number of choices increases. Although our reproduced results also show an overall downward trend, there is a slight rebound at the 5-choice level.

Figure1 Accuracy Results for Task1



3.2.1 Differences in Dataset Scale

The original paper used the complete UNBench dataset, containing as many as 355,126 test instances. This massive data scale effectively smooths out random fluctuations and provides stable statistical results. Due to data acquisition constraints, our reproduction could only use the sample data from the GitHub repository, consisting of merely 30 instances.Statistical results under small sample conditions are highly susceptible to the influence of individual instances and cannot reflect the model's true performance.

3.2.2 Representativeness of Data Distribution

The complete UNBench dataset covers 30 years of United Nations resolution history from 1994 to 2024, containing draft resolutions with various political backgrounds, regional distributions, and topic types, featuring rich diversity and complex distribution characteristics. However, the 30 sample instances we used may not represent the complex distribution of the complete dataset; instances at certain difficulty levels might be either too simple or too difficult, leading to systematic biases in the results.

3.2.3 Differences in Temporal Split

The original paper explicitly adopted a time-based test split strategy in its experimental setup,ensuring that the test set was temporally later than the training set.This split method effectively evaluates the model's ability to predict future events, aligning with real-world application scenarios.However, since we could only obtain 30 test instances, we were unable to access the corresponding training data nor implement an effective temporal split on these limited samples.This creates a fundamental difference between our evaluation protocol and that of the original paper, potentially affecting the comparability of results.

3.2.4 Invalid Response Handling

During the reproduction process, we observed one invalid response where the model output "Group of Arab States" in the 3-choice task, a value not present in the candidate country list.According to the code logic, when an invalid response is detected, the system randomly selects one from the valid candidates as the result. Although this handling method is consistent with the original paper, random selection based on just one invalid response may introduce minor bias. At the scale of the complete dataset, the proportion and impact of invalid responses could be estimated more accurately and handled more stably.

3.2.5 API Version Differences

The DeepSeek-V3 model version used in the original experiments may have minor differences from the model version currently accessed via API. Large language models typically undergo continuous optimization and updates,thereby affecting the accuracy of reproduction results.

4. Modification and Optimization Results

4.1. Modification Motivation

In the initial reproduction, we found that the model's accuracy was generally lower than the values reported in the original paper, and invalid responses (such as "Group of Arab States") occurred.The original prompt design might be too simplistic and failed to adequately guide the model in conducting professional reasoning in the political and diplomatic domain. Based on this observation, we proposed the following hypothesis that by adding specific reasoning guidance examples to the prompt, we can help the model better understand the political logic behind co-penholder selection, thereby improving accuracy.

4.2 Specific Modification Content Description

We systematically enhanced the prompt, mainly by adding a "Examples of good choices" section to the system prompt, providing the model with a clear reasoning framework:

Original Prompt (First Reproduction):
system_prompt = f"""You are representing {author}, a country drafting a resolution for submission to the...
Your task is to review the draft resolution and select a coauthor from the following list...
Respond only with the name of the chosen coauthor and provide no additional explanation."""

Modified Prompt (Second Experiment):
system_prompt = f"""You are a senior diplomat representing {author}.
Examples of good choices:
- If draft focuses on climate change → choose a country with strong environmental policies
- If draft focuses on security → choose a permanent Security Council member with relevant military/diplomatic influence
- If draft focuses on regional issues → choose a country from that region
CRITICAL RULES:
1. You MUST choose EXACTLY ONE country from the provided list
2. Your response MUST be ONLY the country name - nothing else
3. Do NOT suggest groups, alliances, or organizations
4. Base your choice on the draft resolution content"""

The main modifications included role enhancement by changing "representing" to "a senior diplomat representing" to strengthen professional identity; the addition of reasoning examples that provided guidance for three typical scenarios (climate change, security issues, and regional issues); the establishment of clear rules, including four key directives that specifically prohibited outputting organization names to address the previous "Group of Arab States" invalid response; and a structured presentation that organized the prompt content in a clearer format.

4.3 Modified Experimental Results

Using the enhanced prompt, we reran the experiment on the same 30 sample instances and obtained the following results:

	2 choices	3 choices	4 choices	5 choices
Paper(Original)	0.695	0.555	0.443	0.422
Report(Present)	0.567	0.400	0.333	0.367
modification	0.567	0.500	0.333	0.367

Table2 Results of Three Attempts

During the second experiment, we observed several phenomena. First, regarding invalid response changes, the first reproduction had one invalid response in the 3-choice task ("Group of Arab States"), while the second experiment had two invalid responses in the 4-choice task, both being "Egypt". Second, in terms of response quality, model outputs became more standardized, with most responses directly returning country names. Additionally, organization-name type invalid responses no longer appeared.Finally, runtime remained essentially unchanged at approximately 1.5-2.0 seconds per instance.

4.4 Modified Experimental Analysis

The accuracy of the 3-choice task increased by 10 percentage points, from 0.4000 (12 out of 30 correct) in the first reproduction to 0.5000 (15 out of 30 correct) after prompt modification. This improvement narrowed the gap with the original paper's reported 0.555 to just 5.5 percentage points, representing a 25% relative reduction in error rate (from 60% error to 50% error). This substantial gain suggests that providing explicit reasoning examples helps the model better navigate medium-complexity political decisions where multiple plausible options exist.

However,the enhancement was not uniform across all difficulty levels. The 3-choice task showed significant gains, while simple tasks (2-choice, remaining at 0.5667) and more difficult tasks (4-choice at 0.3333 and 5-choice at 0.3667) showed no improvement.

At the same time,although the modified prompt explicitly prohibited outputting organization names (Rule 3: "Do NOT suggest groups, alliances, or organizations"), two invalid responses still occurred in the 4-choice task, both outputs being "Egypt".This may represent that the model developing a bias toward certain frequently appearing countries,or potential influence from training data where Egypt appears prominently in Middle Eastern/North African contexts,or possible confusion when multiple resolutions involve similar regional themes.

This experiment demonstrates that adding domain-specific reasoning guidance examples can effectively improve model performance on political and diplomatic tasks. The 10-percentage-point improvement in the 3-choice task indicates that LLM capability deployment heavily depends on prompt quality.

5. Conclusion

This study conducted a reproducibility experiment on Task1(Co-Penholder Judgement) from the paper "Benchmarking LLMs for Political Science: A United Nations Perspective"[1].Using the DeepSeek API to invoke the DeepSeek-V3 model, we performed two rounds of experiments on 30 sample instances (original prompt vs. enhanced prompt).

Through these two rounds of experiments,we found that both the qualitative trends and the effectiveness of prompt engineering are reproducible.Although our reproduced results showed some fluctuations due to the small sample size (e.g., 5-choice accuracy was slightly higher than 4-choice), the overall downward trend remained clearly discernible.This indicates that the fundamental pattern of task difficulty being negatively correlated with performance is stable when DeepSeek-V3 handles multi-choice political reasoning tasks By adding domain-specific reasoning guidance examples, we successfully improved the accuracy of the 3-choice task by 10 percentage points, from 0.4000 to 0.5000.This improvement demonstrates that in specialized domains such as political diplomacy, providing contextualized reasoning examples can effectively stimulate LLMs' domain knowledge and enhance decision quality.The positive impact of prompt optimization is consistent across different experiments.

The biggest limiting factor in this reproduction was the dataset scale. Future similar work should first ensure access to the complete dataset, or clearly specify the reproduction scope and data limitations in the report. Additionally, after modifying the prompt in the second experiment, we observed a significant increase in accuracy, revealing that when reproducing LLM-related research, the precise reproduction of prompts is equally important as the dataset. Paper authors should make every effort to disclose complete prompt templates. During the reproduction process, we observed that the model occasionally output country names or organization names that were not in the candidate list. For these invalid responses, we adopted a random selection strategy for handling. In small sample scenarios, this handling method may have a visible impact on the final results. Future research should more systematically record the proportion and handling methods of invalid responses, and consider robustness tests for multiple handling strategies.

Abstract

[1] Liang, Y., Yang, L., Wang, C., Xia, C., Meng, R., Xu, X., Wang, H., Payani, A., & Shu, K. (2025). Benchmarking LLMs for Political Science: A United Nations Perspective. arXiv preprint arXiv:2502.14122.