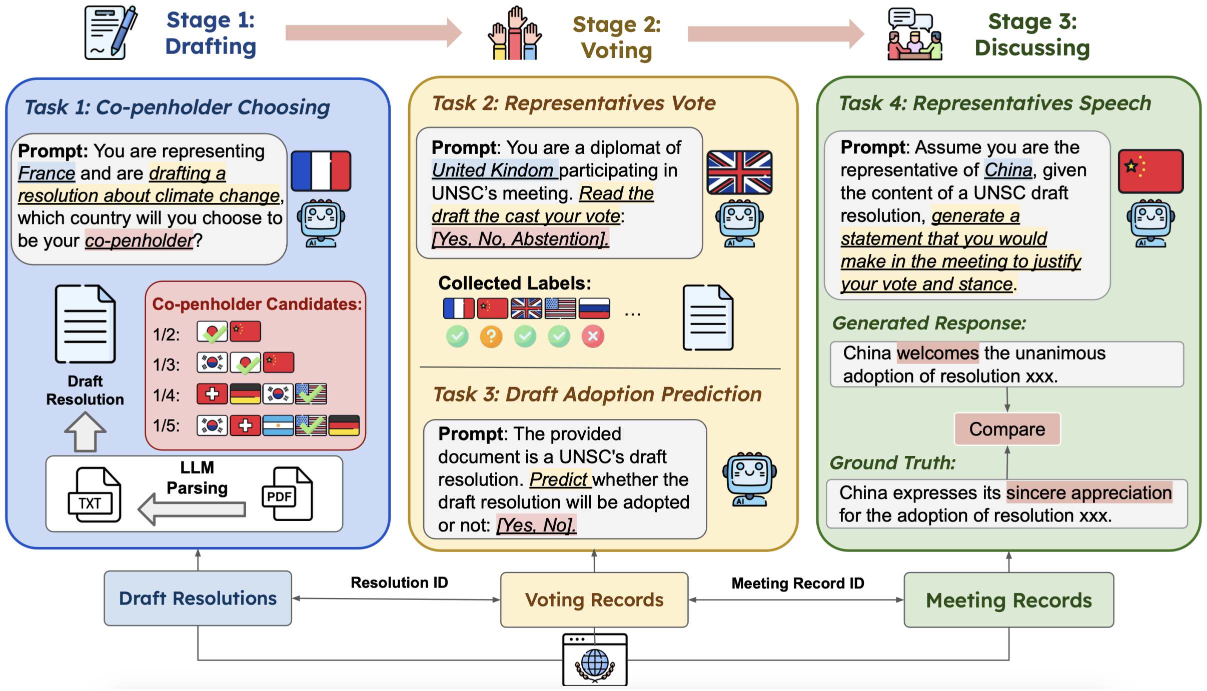
# Reproducibility and Agentic Extension Study of UNBench Task 2: Representatives Voting Simulation

Ke Wan
FTEC5660 Agentic AI for Business and FinTech

## 1 Project Overview

This report presents a controlled reproducibility and methodological extension study of **UNBench** [1], a multi-stage benchmark designed to evaluate large language models (LLMs) in United Nations Security Council (UNSC) simulations.

UNBench operationalizes political decision-making as a structured institutional process comprising drafting, discussion, and voting stages. The overall benchmark structure is illustrated below.



**Prompt:** You are representing *France* and are *drafting a resolution about climate change*, which country will you choose to be your *co-penholder*?

**Prompt**: You are a diplomat of *United Kindom* participating in UNSC's meeting. *Read the draft the cast your vote*: *[Yes, No, Abstention]*.

**Prompt**: The provided document is a UNSC's draft resolution. *Predict* whether the draft resolution will be adopted or not: *[Yes, No]*.

**Prompt**: Assume you are the representative of *China*, given the content of a UNSC draft resolution, *generate a statement that you would make in the meeting to justify your vote and stance*.

China welcomes the unanimous adoption of resolution xxx.

China expresses its sincere appreciation for the adoption of resolution xxx.

Figure 1: Overview of UNBench multi-stage pipeline [1].

As illustrated in Figure 1, UNBench consists of drafting, voting, adoption prediction, and speech generation stages. This study specifically reproduces Task 2 within the voting stage.

Beyond classification, this task can be interpreted through an agentic AI lens. The model must condition on role identity, interpret geopolitical context, infer implicit national preference structures, and produce a discrete action. This study both reproduces the Task 2 evaluation protocol and investigates whether structured reasoning improves decision robustness under severe class imbalance.

# 2 Reproduction Target

The reproduction objective is to replicate the Task 2 evaluation protocol as described in [1], specifically:

- Country-conditioned vote prediction ($Y/N/A$)
- Deterministic inference ($temperature = 0$)
- Evaluation using Balanced Accuracy and Macro-F1

The original paper reports a Balanced Accuracy of 0.823 for GPT-4.

Due to API cost constraints, evaluation is conducted on a fixed subset of 50 draft resolutions. Subset selection is performed using a fixed random seed to ensure deterministic reproducibility. The selected `Original_id` list is preserved within experiment outputs to guarantee traceability. All predictions and ground-truth labels used for evaluation are cached locally, allowing metric verification without re-running API calls. The evaluation inputs and outputs are stored in `task2_subset_results.csv`, enabling metric verification without re-running API calls.

# 3 Agentic Framing

Although formally a multi-class classification task, Task 2 structurally resembles a minimal role-based decision agent.

Each inference requires:

1. Role conditioning (country identity)
2. Context interpretation (resolution text)
3. Preference inference (alignment with national interest)
4. Action selection (vote)

This sequential dependency motivates testing a structured reasoning decomposition approximating a minimal agentic loop:

$$\text{State} \rightarrow \text{Reasoning} \rightarrow \text{Action}$$

Importantly, this implementation represents a weak agentic structure. It introduces intermediate reasoning but does not incorporate memory, feedback, reward optimization, or iterative self-correction.

# 4 Experimental Setup

All experiments use `gpt-4o-mini` with deterministic decoding ($temperature = 0$).

### 4.1 Baseline: Direct Action Prompting

The baseline employs a single-step instruction:

> Given the country and draft resolution, output one of {Y, N, A}.

This corresponds to a stateless reactive agent directly mapping contextual input to action.

### 4.2 Modification: Structured Two-Stage Reasoning

The modification decomposes decision-making into two explicit stages:

1. Generate a one-sentence explanation of the country's likely stance.
2. Based on that explanation, output the final vote.

Only the final vote is evaluated.

Formally, the modification introduces an intermediate reasoning variable $r$:

$$s \to r \to a$$

where $s$ denotes the state (country identity and draft text), $r$ is a one-sentence policy stance explanation, and $a \in \{Y, N, A\}$ is the final action.

Unlike standard chain-of-thought prompting, the reasoning and action stages are explicitly separated in instruction, enforcing structural decomposition. However, the system remains single-pass and non-adaptive. No memory, iterative correction, or feedback mechanism is introduced.

**Implementation detail.** In both stages, decoding is deterministic ($\mathit{temperature} = 0$). To control cost, the draft text is truncated to the first 2000 characters. The action stage is constrained to output exactly one label from $\{Y, N, A\}$, and only this final label is used for evaluation.

## 5 Results

| Setting | Accuracy | Balanced Accuracy | Macro-F1 |
|---|---|---|---|
| Baseline (single-shot) | 0.7453 | 0.6895 | 0.3348 |
| Modification (multi-step) | 0.9573 | 0.6639 | 0.4215 |

Table 1: Performance comparison on the 50-draft subset

The dataset exhibits strong class imbalance, with the majority class being "In Favour" (Y). In our subset, the ground-truth label distribution is highly skewed (Y=723, A=26, N=1).

## 5.1 Confusion Matrix Analysis

Figures 2 and 3 illustrate prediction behavior under both prompting strategies. The multi-step strategy substantially increases majority-class precision but collapses minority-class recall. In particular, the "Against" (A) class achieves zero recall under structured reasoning. This produces a divergence:

$$\text{Accuracy} \uparrow \quad \text{while} \quad \text{Balanced Accuracy} \downarrow$$

Under severe class imbalance, raw Accuracy overestimates performance gains, whereas Balanced Accuracy provides a more faithful robustness measure. Notably, although the intermediate reasoning step increases confidence in majority-class predictions, it does not introduce explicit minority-class calibration; structured prompting may therefore reinforce prior distributional bias rather than mitigate it.

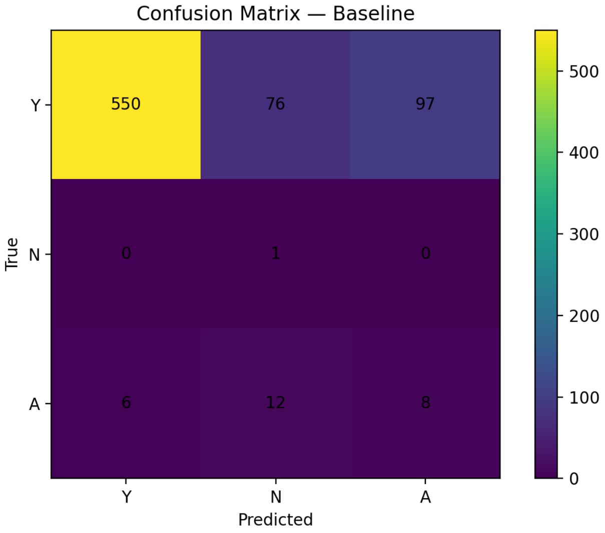


Figure 2: Confusion Matrix — Baseline

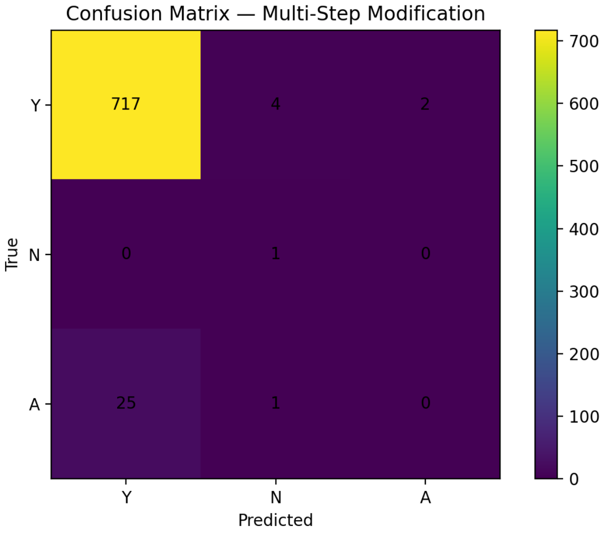


Figure 3: Confusion Matrix — Multi-Step Modification

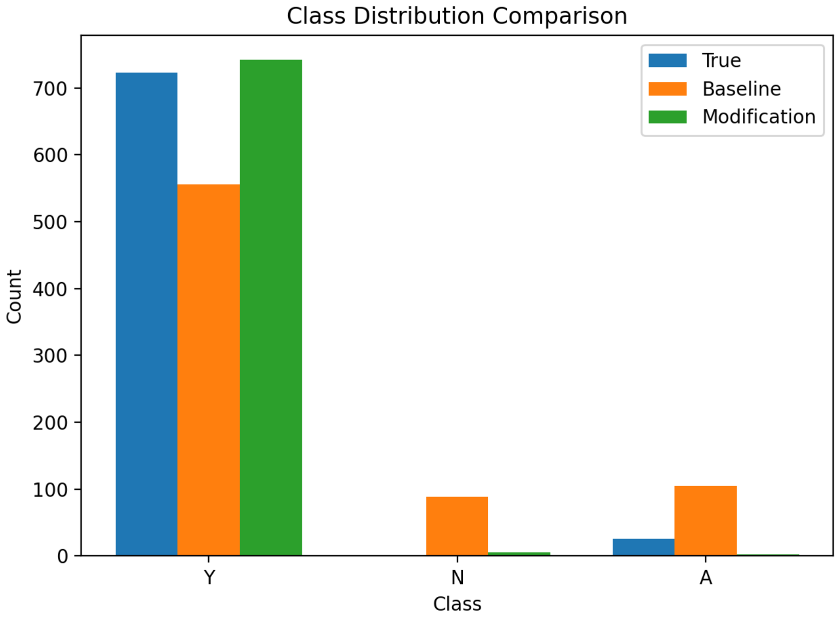


Figure 4: Class distribution comparison across true labels, baseline predictions, and multi-step modification predictions

## 6 Comparison with Reported Results

The reproduced Balanced Accuracy (0.6895) is lower than the reported 0.823 in [1]. This discrepancy arises from:

- Use of `gpt-4o-mini` instead of GPT-4
- Subset-based evaluation
- Absence of prompt calibration or hyperparameter tuning

Despite quantitative differences, the qualitative sensitivity to imbalance-aware metrics remains consistent with the original findings. Therefore, this study should be interpreted as a pipeline-level reproduction and controlled ablation, rather than a direct numerical replication of the headline score.

## 7 Agentic Implications

This experiment reveals a critical insight for agentic AI systems in finance and business contexts.

Reasoning decomposition does not automatically improve decision robustness. Instead, structured prompting may amplify dominant patterns embedded in pretrained distributions.

Under skewed label distributions, multi-step reasoning increases majority-class confidence while suppressing minority-class detection.

This behavior has direct parallels in financial risk and business decision systems. Under highly skewed outcome distributions, structured reasoning may systematically amplify dominant signals while suppressing rare but critical events. Such rare-event under-detection resembles tail-risk blindness in financial modeling, where models achieve high aggregate accuracy yet fail precisely on low-frequency, high-impact cases.

Moreover, majority-signal amplification can create an illusion of robustness under naïve performance metrics. Improvements in overall Accuracy may mask deterioration in minority-class recall, leading to governance risks when agentic systems are evaluated without imbalance-aware criteria. Without adaptive feedback loops, explicit minority-class calibration, or reward-based correction, prompt-level reasoning decomposition remains structurally non-adaptive and therefore inherits distributional biases present in pretrained representations.

Thus, metric selection becomes a governance decision rather than merely a technical preference.

## 8 Conclusion

This study successfully reproduces the structural evaluation pipeline of UNBench Task 2 on a controlled subset.

While structured reasoning substantially increases raw Accuracy, imbalance-aware metrics reveal persistent minority-class failure.

These findings emphasize that agentic reasoning decomposition does not guarantee fairness or robustness. Careful metric design and bias-aware evaluation remain essential for deploying agentic AI systems in high-stakes business and financial environments.

## Implementation Notes

During reproduction, several engineering challenges emerged. Initial experiments were conducted on the repository's representative subset, requiring restructuring of the data-loading pipeline to align with the full Task 2 format. API rate limits necessitated controlled execution and result caching. Severe class imbalance required reinterpretation of evaluation metrics, reinforcing the importance of Balanced Accuracy over raw Accuracy.

## References

[1] Yueqing Liang et al. *Benchmarking LLMs for Political Science: A United Nations Perspective.* 2026. arXiv: 2502.14122 [cs.CL]. URL: https://arxiv.org/abs/2502.14122.