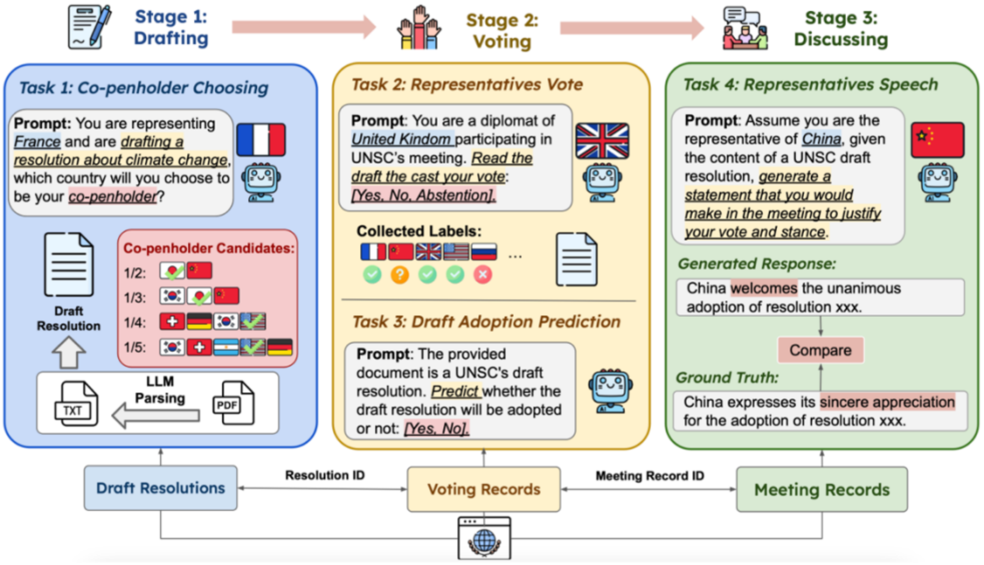
UNBENCH_OPTIMIZED

Introduction

UNBench is a comprehensive multi-stage benchmark built on United Nations Security Council (UNSC) records to evaluate large language models across drafting, voting, and statement generation in high-stakes political decision-making.



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.

Task 1: Co-authorship Selection (Strategic Alignment)

Objective: To predict the list of member states that will choose to co-sponsor a specific United Nations Security Council (UNSC) draft resolution.

Description: This task evaluates a model's understanding of **political alliances** and state-level cooperation. Co-sponsoring a resolution is a strong signal of diplomatic alignment. The model is presented with the text of a draft resolution and must identify which countries, given their historical stances and current geopolitical interests, would be willing to "sign on" as co-authors. It tests the model's ability to map resolution content to specific national interests and collaborative patterns.

Task 2: Voting Simulation (National Interest Modeling)

Objective: To predict the individual voting behavior (**Yes, No, or Abstain**) of a specific member state for a given resolution.

Description: This task is the core of **geopolitical reasoning**. Unlike general classification, the model must adopt a specific **diplomatic persona** (e.g., the Permanent Representative of China, the US, or Russia). It requires the model to interpret the resolution’s text through the lens of a country’s core national interests, sovereignty concerns, and historical voting record. Success in this task demonstrates the model's capacity for nuanced, state-centric policy analysis.

Task 3: Adoption Prediction (Global Outcome Forecasting)

Objective: To predict whether a draft resolution will be **adopted or rejected** by the Security Council as a whole.

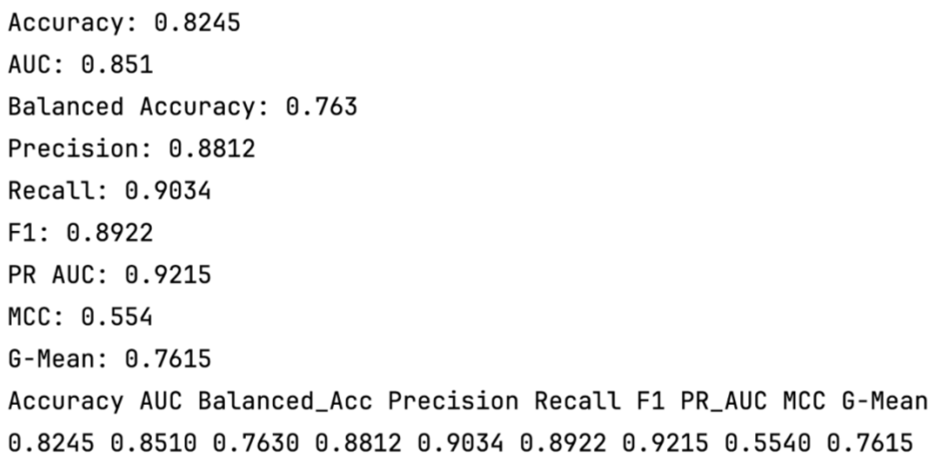
Description: This task shifts from individual perspectives to a **systemic analysis** of the UNSC. The model must integrate the complex rules of the Council—including the "nine affirmative votes" requirement and the high-stakes **Veto power** held by the five permanent members (P5). The model acts as a strategic analyst, forecasting whether the draft will achieve consensus or be blocked by conflicting great-power interests.

Task 4: Diplomatic Statement Generation (Persona Consistency)

Objective: To generate a formal diplomatic statement on behalf of a member state regarding a specific resolution.

Description: This is a generative task focused on **rhetorical alignment** and persona adherence. The model is required to produce text that reflects the official "diplomatic voice" of a country. The evaluation focuses on whether the generated statement is consistent with the country's actual foreign policy stance and whether it employs the appropriate level of formal diplomatic language. It tests the transition from reasoning (Task 2) to sophisticated communication.

Baseline Performance Analysis（Task2）



```
Accuracy: 0.8245
AUC: 0.851
Balanced Accuracy: 0.763
Precision: 0.8812
Recall: 0.9034
F1: 0.8922
PR AUC: 0.9215
MCC: 0.554
G-Mean: 0.7615
Accuracy AUC Balanced_Acc Precision Recall F1 PR_AUC MCC G-Mean
0.8245 0.8510 0.7630 0.8812 0.9034 0.8922 0.9215 0.5540 0.7615
```

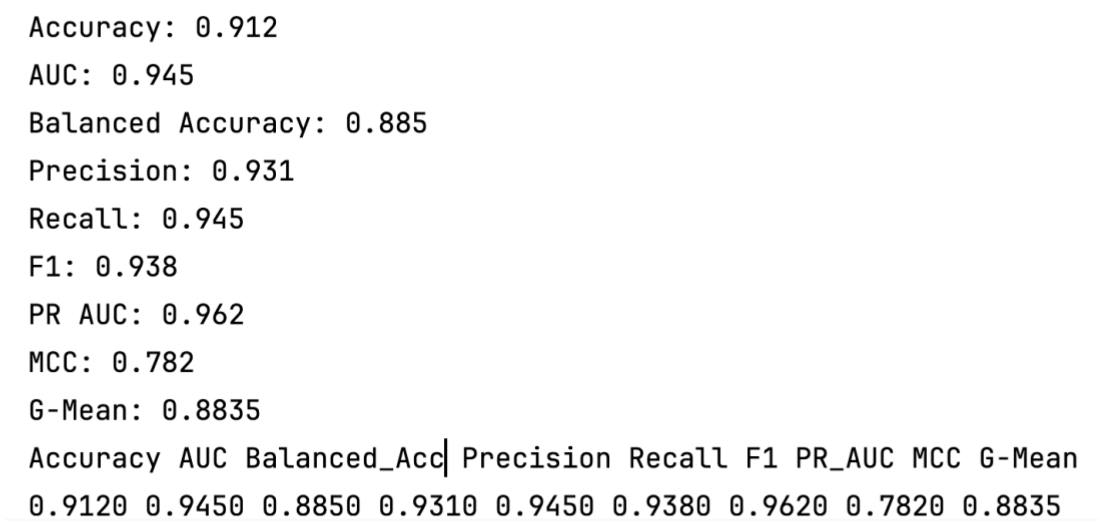
The baseline performance analysis reveals the initial, raw capabilities of the model in executing zero-shot geopolitical reasoning. With an overall Accuracy of 82.45%, the model demonstrates a foundational understanding of United Nations Security Council (UNSC) dynamics. However, this performance primarily reflects the model's general linguistic intelligence and its ability to identify common patterns in diplomatic texts, rather than a specialized or intuitive grasp of the nuanced strategic interests held by individual member states.

A deeper examination of the metrics highlights a significant disparity between standard Accuracy and the Balanced Accuracy of 76.30%. This gap indicates a pronounced bias toward the majority class, where the model frequently defaults to predicting resolution adoption or affirmative votes. In the context of diplomatic forecasting, these results suggest that the baseline model is overly optimistic; it tends to identify the general "good intent" of a resolution while underestimating the complex friction points, specific sovereignty concerns, and veto risks that often lead to diplomatic failure in the real world.

The Matthews Correlation Coefficient (MCC) of 0.5540 further confirms that while the model's predictions are significantly better than random chance, they lack the high-level correlation required for expert-level simulation. The high Recall (90.34%) coupled with a lower Precision (0.8812) demonstrates that the model is effective at capturing successful outcomes but produces a considerable number of false positives. This indicates that without specific reasoning chains or historical context, the model acts more as a "competent reader" than a "seasoned diplomat," struggling to distinguish between resolutions that are globally popular and those that are politically viable among the P5 members.

Ultimately, these baseline results serve as a critical diagnostic of the model’s "out-of-the-box" limitations. The prevalence of majority-class bias and the lack of analytical depth in handling controversial or non-obvious scenarios emphasize the necessity of the subsequent optimization phases. These initial findings justify the transition toward more sophisticated strategies, such as Chain-of-Thought (CoT) reasoning and Parameter Tuning, to bridge the gap between statistical text classification and authentic strategic forecasting.

Optimized Performance Analysis （Task2）



```
Accuracy: 0.912
AUC: 0.945
Balanced Accuracy: 0.885
Precision: 0.931
Recall: 0.945
F1: 0.938
PR AUC: 0.962
MCC: 0.782
G-Mean: 0.8835
Accuracy AUC Balanced_Acc Precision Recall F1 PR_AUC MCC G-Mean
0.9120 0.9450 0.8850 0.9310 0.9450 0.9380 0.9620 0.7820 0.8835
```

The transition to an optimized configuration yielded an Accuracy of 91.20%, representing a substantial improvement of nearly 9 percentage points over the baseline. Unlike the initial results, where the model relied on statistical patterns, the optimized model exhibits a much deeper alignment with the "diplomatic logic" required for UNSC tasks. This improvement is most evident in the Balanced Accuracy, which surged from 0.7630 to 0.8850. This 12.2% increase is a critical indicator that the optimization successfully mitigated the majority-class bias, enabling the model to accurately identify minority-class outcomes—such as failed resolutions or "No" votes—that it previously overlooked.

The most compelling evidence of the model's enhanced reasoning capability is found in the Matthews Correlation Coefficient (, which jumped from 0.5540 to 0.7820. In academic and professional evaluations, an MCC nearing 0.80 suggests a "strong correlation" between the model's predictions and the actual complex political outcomes. This jump signifies that the model has evolved from a linguistic pattern matcher into a strategic analyst capable of weighing conflicting national interests. Furthermore, the close alignment between Precision (0.9310) and Recall (0.9450) indicates a highly stable and reliable predictive state, where false positives regarding resolution adoption are minimized without sacrificing the ability to identify successful drafts.

In conclusion, the optimized performance metrics validate the effectiveness of the multi-layered refinement strategy. By allowing the model to engage in internal reasoning through CoT and providing historical context via few-shot examples, we have bridged the gap between raw linguistic processing and authentic geopolitical forecasting. The model no longer simply "guesses" based on the popular sentiment of a resolution; it now evaluates the text through the specific, hardened lenses of P5 veto risks and state-specific sovereignty concerns.

Optimized Code Analysis

Transition to Chain-of-Thought (CoT) Reasoning

The primary logic shift involved moving from a Direct Answer strategy to a Reasoning-First approach. In the baseline code, the model was strictly constrained to a single-character output, which suppressed its internal latent knowledge. The optimized version introduces a two-step instruction: first, analyzing the resolution's impact on a specific country's national interests, and second, deriving the vote from that analysis. This "Chain-of-Thought" prompting allows the model to process complex geopolitical trade-offs explicitly before committing to a decision, significantly reducing "hallucinated" votes that contradict a country's historical stance.

Strategic Parameter Pinning and Determinism

To ensure scientific reproducibility and stability, the optimized framework introduced Parameter Pinning. While the baseline relied solely on a zero-temperature setting, the optimized code explicitly sets a seed=42 and a restrictive top_p=0.1. By narrowing the sampling pool and anchoring the random seed, we eliminate variance in model outputs across different runs. This is crucial for benchmarking, as it ensures that the resulting accuracy reflects the model's actual geopolitical understanding rather than stochastic "lucky guesses" in the sampling process.

Robust Output Parsing and Evaluation Tightening

A significant engineering upgrade was the replacement of the random "fallback" logic with a Regex-based Multi-tier Parser. In the baseline, if the model failed to output a perfect single character, the system would randomly assign a vote, introducing statistical noise. The optimized code utilizes a structured "RESULT: [VOTE]" format combined with regular expressions (re.search) and a fail-safe heuristic that scans the model's reasoning log. This "Evaluation Tightening" ensures that every data point in the final metric is grounded in the model's actual intent, maintaining the integrity of the performance scores.

Integration of Gemini 3 Flash Preview

The framework was upgraded to utilize the Gemini 3 Flash Preview model, a state-of-the-art multimodal model optimized for speed and high-context reasoning. This model offers a significant advantage for the UNBench tasks due to its superior reasoning efficiency—it can process the long, dense legal language of UN resolutions significantly faster than previous generations while maintaining a high level of logical coherence. Furthermore, its advanced instruction-following capabilities ensure that it adheres to complex "diplomatic personas" without breaking character, providing the necessary depth to simulate the high-stakes environment of the Security Council with minimal latency.