Benchmarking LLMs for Political Science: A United Nations Perspective

Yueqing Liang[1], Liangwei Yang[2], Chen Wang[3], Congying Xia[4], Rui Meng[2*], Xiongxiao Xu[1], Haoran Wang[6], Ali Payani[5], Kai Shu[6†]

[1]Illinois Institute of Technology
[2]Salesforce Research
[3]University of Illinois Chicago
[4]Meta GenAI
[5]Cisco Research
[6]Emory University
{yliang40, xxu85}@hawk.illinoistech.edu, {liangwei.yang, ruimeng}@salesforce.com, cwang266@uic.edu, congyingxia@meta.com, apayani@cisco.com, {haoran.wang, kai.shu}@emory.edu

Abstract

Large Language Models (LLMs) have achieved significant advances in natural language processing, yet their potential for high-stake political decision-making remains largely unexplored. This paper addresses the gap by focusing on the application of LLMs to the United Nations (UN) decision-making process, where the stakes are particularly high and political decisions can have far-reaching consequences. We introduce a novel dataset comprising publicly available UN Security Council (UNSC) records from 1994 to 2024, including draft resolutions, voting records, and diplomatic speeches. Using this dataset, we propose the United Nations Benchmark (UNBench), the first comprehensive benchmark designed to evaluate LLMs across four interconnected political science tasks: co-penholder judgment, representative voting simulation, draft adoption prediction, and representative statement generation. These tasks span the three stages of the UN decision-making process—drafting, voting, and discussing—and aim to assess LLMs' ability to understand and simulate political dynamics. Our experimental analysis demonstrates the potential and challenges of applying LLMs in this domain, providing insights into their strengths and limitations in political science. To the best of our knowledge, this is the first benchmark to systematically evaluate LLMs in UN decision-making, contributing to the growing intersection of AI and political science.

UNBench — https://github.com/yueqingliang1/UNBench
Extended version — https://arxiv.org/abs/2502.14122

1 Introduction

Large Language Models (LLMs) such as GPT-4 (OpenAI 2023), Llama (Dubey et al. 2024), and DeepSeek (Liu et al. 2024) have achieved unprecedented proficiency in language tasks and are increasingly under development tailored for different domains (Cheng, Huang, and Wei 2023). Yet, their adaptation to high-stakes political decision-making remains underexplored—particularly in scenarios where model outputs could influence real-world governance. Political science demands capabilities beyond semantic understanding: predicting coalition dynamics, interpreting ambiguous diplomatic language, and navigating the tension between national interests and global norms. These challenges unfold within the United Nations (UN), where a single draft resolution, once adopted, becomes binding international law under Chapter VII of the UN Charter, with cascading impacts on global security, trade, and human rights (e.g., Resolution 1973's no-fly zone over Libya in 2011). Studying the application of LLMs in political science represents both a technical frontier and a critical societal challenge.

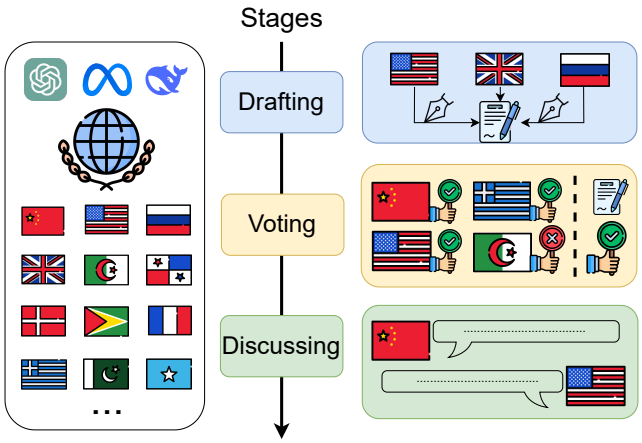

Figure 1: Three key stages of the United Nations decision-making, where LLMs can simulate an individual member.

The scope of UN resolutions is extensive, extending far beyond political statements. Adopted resolutions can authorize military interventions (such as Resolution 678 in 1990), impose sanctions that cripple national economies (e.g., Resolution 1718 on North Korea in 2006), or redefine global priorities (e.g., Resolution 2341 on critical infrastructure pro-

*Now at Google Cloud AI Research.
†Corresponding author

tection in 2017). By analyzing the open-access data from draft resolutions and meeting records, we can explore how current LLMs understand the critical issues facing the international community and assess their ability to interpret bilateral and multilateral relations. This extends LLM's application toward political science, enhancing the analysis of international policies and diplomacy.

Currently, there is no comprehensive benchmark designed specifically for LLM applications in political science. Existing benchmarks (e.g., MMLU (Hendrycks et al. 2021b,a), BIG-Bench (Suzgun et al. 2022)) with related tasks remain fragmented and their designs may not adequately reflect LLMs' understanding of political science. Such fragmented evaluation overlooks the interconnected nature of real-world political decision-making, particularly in high-stake multilateral settings like the UN. Figure 1 shows the stages of each UN draft resolution. It consists of three stages. (1) Drafting: The creation of the resolution's text, involving the collaboration among member states. (2) Voting: The process in which the resolution is formally adopted or rejected by voting from the UN members. (3) Discussing: Each member states the rationality of their voting. Different tasks occur in different stages and also inter-connected with each other.

To fill the benchmark gap in political science, we introduce **United Nations Benchmark (UNBench)**, the first comprehensive benchmark to evaluate LLMs' ability across four distinct yet interconnected political science tasks of different stages: (a) **Co-Penholder Judgement**: Given anonymized draft content, identify optimal co-author nations, simulating coalition-building strategies in multilateral diplomacy. (b) **Representatives Voting Simulation**: Instruct an LLM to act as a national agent (e.g., "As the U.S. representative") and output voting decisions (['In favour', 'Against', etc.]), testing contextual understanding of national interests. (c) **Draft Adoption Prediction**: Input a draft resolution to LLM and ask it to predict the draft adoption probability, Which requires analysis of historical voting patterns and geopolitical alignments. (d) **Representative Statement Generation**: Generate country-specific speeches justifying voting positions, evaluating persuasive language generation under political constraints. The tasks are designed from different UN stages, varying across predictive and generative tasks. Our benchmark is built from publicly available UNSC official records (1994-2024), comprising draft resolutions, voting records, and diplomatic speeches which are extracted from meeting records. In summary, our work makes the following contributions:

- Systematically curated and processed United Nations data from 1994 to 2024, including draft resolutions, voting records, and diplomatic speeches, to provide a new and comprehensive dataset that facilitates LLM applications in political science.
- Introduce the first comprehensive benchmark in political science, designed to assess LLM performance across various real-world tasks in different stages of the UN decision-making process.
- Conduct extensive experimental analysis on the benchmark, demonstrating the effectiveness and limitations of current LLMs in handling complex political tasks.

2 Related Works

2.1 LLMs in Political Science

Existing benchmark datasets have played a critical role in advancing Large Language Models (LLMs) in political science tasks. Datasets such as OpinionQA (Santurkar et al. 2023), PerSenT (Bastan et al. 2020), and GermEval-2017 (Wojatzki et al. 2017) evaluate LLMs on classifying sentiments or identifying topics within political texts. These benchmarks primarily emphasize static text understanding. BillSum (Kornilova and Eidelman 2019) and CaseLaw (Shu et al. 2024) specialize in summarization or analysis of legislative documents. Datasets like PolitiFact (Shu et al. 2020), GossipCop (Grover et al. 2022), and Weibo (Jin et al. 2017) focus on detecting misinformation. Election prediction and voting behavior datasets, such as U.S. Senate Returns 2020 (Data and Lab 2022b) and State Precinct-Level Returns 2018 (Data and Lab 2022a) evaluate models' capabilities in statistical pattern recognition and forecasting. These tasks often rely on structured data and focus on predictive performance related to electoral outcomes. Our benchmark is the first to unify these diverse political science tasks within a single, end-to-end benchmark grounded in real-world multilateral decision-making process in the United Nations scenario.

2.2 United Nations-Related Datasets

The United Nations has long been a focal institution for global politics, attracting extensive study in political science (Bailey, Strezhnev, and Voeten 2017; Voeten 2013). While various publicly available datasets shed partial light on UN processes, they typically focus on limited aspects of the organization. Harvard Dataverse UN Voting Dataset (Voeten, Strezhnev, and Bailey 2009) compiles pairwise country voting statistics, offering quantitative insights but lacking textual data such as draft resolutions or debate transcripts. UNSCR.com (i Redondo and Llovera 2025) collects mainly *adopted* Security Council resolutions, providing topic labels and related resolutions but minimal coverage of *draft* content. UNSCdeb8 (Kohlenberg et al. 2019) includes verbatim debate transcripts from 2010 to 2017, capturing real-time deliberation but omitting links to draft texts or voting outcomes. Lastly, the UN Parallel Corpus (United Nations 2025) supplies multilingual final resolutions and meeting records (1994–2014), yet lacks draft-stage materials and detailed metadata to trace evolving negotiations. In this paper, we present the first LLM benchmark covering the full UN resolution process—combining draft texts, debates, and voting records from 1994–2024 to capture the complete decision-making trajectory.

3 The United Nations Benchmark

UNBench is extracted from United Nations resolution decisions. It consists of four tasks, covering three distinct stages (Drafting, Voting, Discussing) before the adoption/rejection of each resolution.

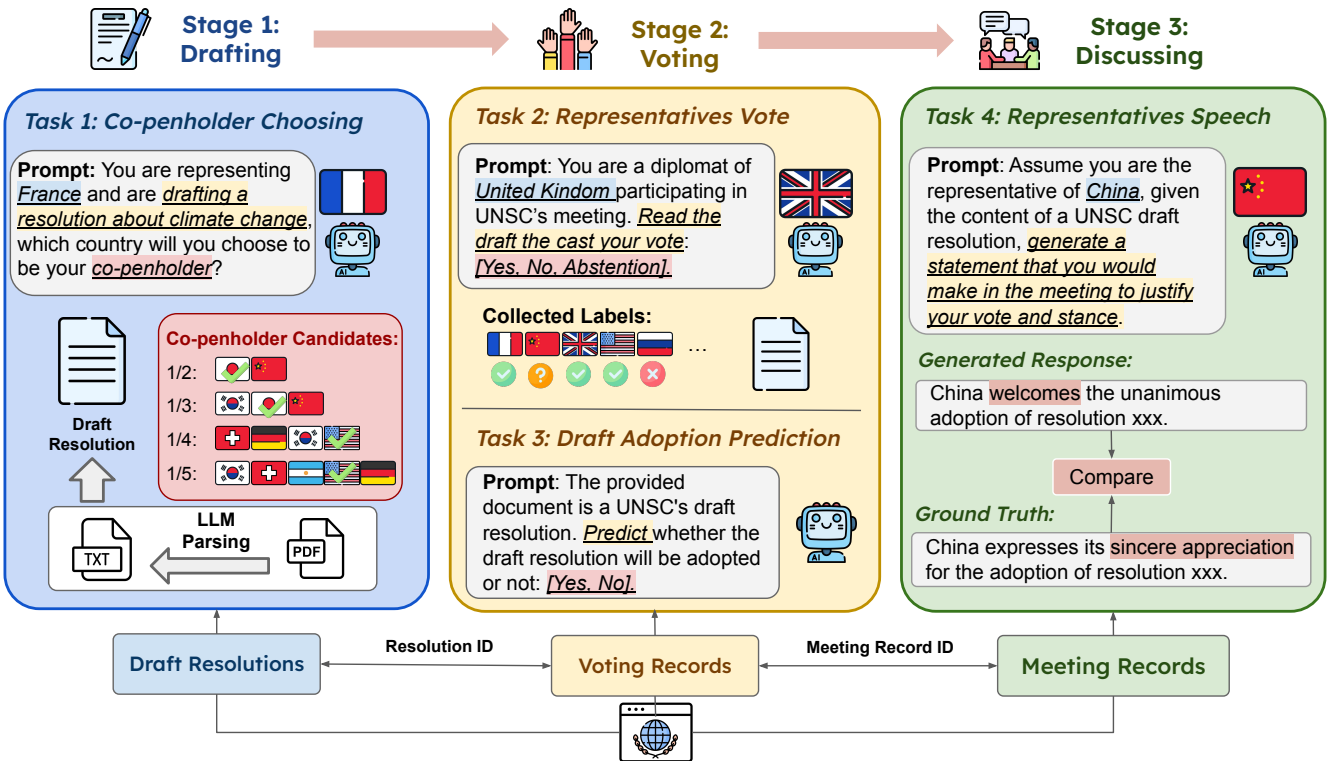

Figure 2: The proposed **UNBench**. It consists of 4 tasks extracted from different stages of a UN draft.

3.1 Benchmark Notation Definition

To formalize our benchmark and the tasks, we introduce the following notations:

- $\mathcal{R} = \{r_1, r_2, \ldots, r_N\}$: The set of all **draft resolutions**. Each resolution r_i contains proposed actions, mandates, and contextual details (e.g., sponsoring countries).
- $\mathcal{C} = \{c_1, c_2, \ldots, c_M\}$: The set of all **UN members**, including both permanent and non-permanent members.
- For each resolution $r_i \in \mathcal{R}$, $\mathcal{C}_{\text{candidate}}(r_i) \subseteq \mathcal{C}$ denotes the **candidate co-penholder** countries—those likely to sponsor or support r_i during the *drafting* process.
- $\mathcal{V}(r_i) = \{v_{i,1}, v_{i,2}, \ldots, v_{i,15}\}$: The **votes** of the 15 Security Council members on r_i. Each vote $v_{i,j}$ is one of $\{\texttt{[In Favour]}, \texttt{[Against]}, \texttt{[Abstention]}\}$.
- $\text{Result}(r_i) \in \{\text{ADOPTED}, \text{NOTADOPTED}\}$: The **decision** on r_i. It is NOTADOPTED if it fails to secure the necessary majority or is vetoed by a permanent member.
- $\mathcal{S}(r_i, c_j)$: The **official statement** (speech) of country c_j regarding r_i, delivered during the *discussion* stage. This typically includes the country's rationale, policy concerns, and diplomatic stance.

3.2 Stage 1: Drafting

Drafting is the initial phase in the lifecycle of a resolution. Typically, one or more countries—often referred to as *penholders*—take the lead in preparing a draft text, outlining the resolution's objectives, scope, and operative clauses. The draft is then refined through closed-door consultations, circulated informally among Council members. A unique feature of drafting is the practice of *co-penholdership*, wherein multiple countries jointly sponsor or "own" the resolution from its inception. We design a task focusing on *identifying the most suitable co-penholder*.

Task 1: Co-Penholder Judgement. Formally, let r_i be a draft resolution authored by country c_a. We sample $\mathcal{C}_{\text{candidate}}(r_i) \subseteq \mathcal{C}$ as a set of co-penholders candidates, each representing a different country. The LLM is prompted to assume the *role of the author country* c_a, given the text of r_i, and asked to choose exactly one co-penholder from the set $\mathcal{C}_{\text{candidate}}(r_i)$. In practice, we vary the number of candidates from 2 to 5, making this task a multi-choice setup with a single correct answer.

Co-penholdership reflects shared strategic interests, diplomatic partnerships, or specialized expertise on the issue at hand. Identifying a suitable co-penholder requires the model to have the following ability:

- **Comprehend Contextual Information:** Understand the resolution's key themes (e.g., conflict prevention, sanctions regime, peacekeeping mandates), and recognize which policy domains (e.g., human rights, climate, nuclear disarmament) are relevant. This tests the model's ability to integrate textual comprehension of policy content with broader geopolitical and diplomatic reasoning.
- **Infer Diplomatic Alignments:** Analyze historical or implied alignments and identify country pairings likely

to co-sponsor a resolution. This evaluates whether the model can correlate textual cues with knowledge of past collaborations or alliances, and navigate multi-choice questions where the differences between candidate countries may be subtle or context-dependent.

- **Reason About Multilateral Cooperation:** Weigh factors such as a candidate country's veto power (if permanent), geopolitical priorities, and regional interests to recommend a co-penholder that maximizes the resolution's likelihood of success.

Hence, Task 1 offers a focused measure of the model's capacity to perform political and textual reasoning in a controlled, multi-choice format, laying the groundwork for subsequent stages involving voting and post-vote deliberation.

3.3 Stage 2: Voting

In the second stage of the resolution lifecycle, each of the 15 Council members casts a vote to determine whether a draft resolution is *adopted* or *rejected*. Permanent members wield veto power, meaning a single "`Against`" vote from any of the five permanent countries can block the resolution, regardless of overall support. Non-permanent members, on the other hand, primarily influence the outcome through collective consensus and persuasive negotiation. Based on this voting mechanism, we define two tasks that capture different facets of decision-making at this stage.

Task 2: Representatives Voting Simulation. Formally, for a draft resolution r_i, let $\mathcal{V}(r_i) = \{v_{i,1}, \ldots, v_{i,15}\}$ denote the votes cast by each of the 15 Council members. In this task, the LLM is given the content of r_i and prompted to *assume the role of a specific representative* c_j (where $c_j \in \mathcal{C}$) to determine how that country would vote on r_i. Each vote $v_{i,j}$ must be one of $\{$`[In Favour]`, `[Against]`, `[Abstention]`$\}$.

Objective and Challenges. Effective voting simulation requires the model to:

- **Comprehend the Resolution:** Interpret the text of r_i in light of its subject matter (e.g., conflict prevention).
- **Incorporate National Interests:** Weigh a representative country's known priorities and geopolitical alignments (e.g., historical alliances, regional blocs).
- **Account for Veto Power:** Recognize whether c_j is a permanent member with veto ability.

These dimensions test not only the model's text understanding but also its ability in political and strategic reasoning, reflecting real-world complexities in UN negotiations.

Task 3: Draft Adoption Prediction. Once the votes $\mathcal{V}(r_i)$ are cast, the resolution is *adopted* if it secures the necessary majority (i.e., at least nine `[In Favour]` votes) and no permanent member exercises a veto. In this task, the LLM receives the text of r_i and *predict the final outcome*, denoted

$$\text{Result}(r_i) \in \{\textsc{Adopted}, \textsc{NotAdopted}\}.$$

Unlike Task 2, which focuses on individual country votes, this task tests whether the model can account for the collective dynamics of all 15 Council members. Key factors include but are not limited to overall council sentiment, potential veto threats, historical precedents, etc. Accurate adoption prediction thus demands higher-level inference about the distribution of possible votes, the interplay of veto power, and the delicate balancing of geopolitical interests. Together, Tasks 2 and 3 provide complementary perspectives on an LLM's capacity to model real-world decision-making under complex international relations.

3.4 Stage 3: Discussing

Once the voting concludes, each member typically delivers a statement clarifying the vote and articulating national positions or broader policy perspectives. These statements reveal the rationale behind each country's stance—whether *In Favour*, *Against*, or *Abstention*. Since these statements are given in an open discussion format, countries may engage in debates, directly addressing or countering the arguments made by other members. This final discussion phase can shape diplomatic narratives surrounding the resolution's implications and signal future policy directions. We design a task that evaluates an LLM's ability to generate representative statements aligned with national interests, voting behavior, and diplomatic discourse norms.

Task 4: Representative Statement Generation Formally, for a draft resolution r_i, let $\mathcal{S}(r_i, c_j)$ denote the official statement made by country c_j. In this task, the LLM receives the text of r_i alongside contextual details, including the outcome of the vote, each country's voting decision, and any prior statements made in the discussion (if available, in the order they were delivered). The model is then asked to *generate the statement* that c_j would deliver. This statement should reflect:

- **National Interests and Policies:** How does c_j's geopolitical position shape its response to the resolution (e.g., security concerns, regional dynamics)?
- **Vote Justification:** If c_j voted `[In Favour]`, `[Against]`, or `[Abstention]`, the statement should provide a coherent rationale for the decision.
- **Diplomatic Tone and Style:** UNSC discourse follows a formal, measured tone. The model should generate text that aligns with the conventions of diplomatic statements.

By prompting the model to produce *country-specific* statements, Task 4 evaluates higher-level language generation skills in a multi-faceted political context. The ability to incorporate historical alliances, policy priorities, and rhetorical conventions into coherent and persuasive statements indicates an advanced understanding of both textual composition and global political dynamics.

4 Dataset Construction

Our dataset $\mathcal{D}$ is constructed from United Nations Security Council (UNSC) meeting records, draft resolutions, and voting histories spanning the years 1994 to 2024. The resulting corpus not only includes the textual content of each draft

resolution but also contextual metadata such as voting outcomes, sponsoring nations, meeting transcripts, and the temporal sequence of events.

The overarching goal of this work is to provide a *unified and extensive* collection of the decision-making process, thereby enabling evaluation of multiple LLM capabilities in a single benchmark. To achieve this, we collect multi-perspective data from the official website and digital library(United Nations Security Council 2024), which archive draft resolutions, voting records, and meeting minutes. Below, we highlight key challenges and our corresponding strategies in constructing $\mathcal{D}$ in the different stages of our benchmark construction.

Data Collection. In the data collection stage, we have three challenges: (1) ***Fragmented Records.*** Draft resolutions, voting logs, and meeting transcripts reside in separate sections of the UN database. We utilize shared identifiers (e.g., resolution numbers, meeting record IDs) to *link* these sources. As illustrated schematically in Figure 2, we first retrieve all draft resolutions, then query corresponding voting records by resolution ID (when applicable), and finally map meeting transcripts via the meeting record ID. (2) ***Missing or Incomplete Metadata.*** Despite the UN's comprehensive record-keeping, certain entries contain missing fields (e.g., sponsor lists), inconsistent data formats, or broken links. We mitigate these issues by cross-referencing multiple UN repositories, manually curating ambiguous entries, and applying standardized naming conventions for country references. (3) ***Historical Document Diversity.*** The official document formats and website structures vary considerably across decades, complicating automated crawling and parsing. We address this by implementing adaptive web-scraping scripts that detect layout differences and by performing iterative quality checks to ensure data consistency.

Data Conversion. UN documents are primarily stored in PDF format, making direct ingestion by current LLMs infeasible. We therefore extract and convert the content into plain text. Early attempts using generic Python PDF libraries yielded mixed accuracy due to the unstructured, domain-specific nature of political documents. We applied a *LLM-based* parser (LlamaParse(LlamaIndex 2024)) to handle complex formatting (e.g., multi-column layouts, footnotes, multilingual text).

Data Processing. (1) ***Labeling Adopted vs. Unadopted Drafts.*** Some drafts never become official resolutions (i.e., "unadopted"), lacking a formal resolution ID. We thus inspect the official notes in each draft's record and cross-verify with the final resolution index to categorize them correctly. (2) ***Country Name Normalization.*** Different records refer to the same country with variations (e.g., "United Kingdom" vs. "Kingdom"). We employ the official name to unify references to the same country entity. (3) ***Metadata Alignment.*** For each draft r_i, we compile the relevant information: author/sponsor countries, date, issue category, voting breakdown, and meeting transcripts into a structured format compatible with modern NLP frameworks.

Through these steps, UNBench incorporates the *entire* lifecycle of each UNSC draft resolution, from initial sponsorship and negotiations to final votes and discussions, ensuring comprehensive coverage for our benchmark tasks.

5 Experiments

5.1 Dataset Statistics

Our UNBench covers a broad range of draft resolutions, voting records, and meeting transcripts, providing diverse scenarios for evaluating multiple LLM capabilities. As shown in Table 1, Task 1 features 1,300 draft resolutions with a total of 355,126 instances—reflecting a multi-choice setup where each instance corresponds to an author country selecting a co-penholder. Task 2 contains 17,430 instances of individual votes for each country that participating in the voting, while Task 3 comprises 1,978 draft resolutions with both `[Adopted]` and `[NotAdopted]` labels. Finally, Task 4 includes 7,394 statements from 1,752 UNSC meetings, testing the ability to generate coherent speeches that align with national stances. More data analysis can be found at Appendix A in the extended version (Liang et al. 2025).

Task	Statistic	Value
Task 1	# Drafts	1,300
	# Unique Draft Authors	209
	Avg. # Authors per Draft	7
	# Total Instances	355,126
Task 2	# Drafts	1,162
	# Total Instances	17,430
	# `[In Favour]`	17,020
	# `[Against]`	16
	# `[Abstention]`	391
Task 3	# Drafts	1,978
	# `[Adopted]`	1,880
	# `[NotAdopted]`	98
Task 4	# Meetings (Drafts)	1,752
	# Statements	7,394
	# Countries	204
	Avg. # Tokens per Statement	450

Table 1: Statistics for our UNBench.

5.2 Experimental Setup

Models. Tasks 1, 2, and 3 are classification-oriented. We compare two *traditional text classification models* (BERT (Devlin 2018) and DeBERTa (He et al. 2020)) against several *instruction-tuned LLMs*: *Llama-3.2-1B-Instruct* (Dubey et al. 2024), *Llama-3.2-3B-Instruct*, *Llama-3.1-8B-Instruct*, *Mistral-7B-Instruct* (Jiang et al. 2023), *DeepSeek-V3* (Liu et al. 2024), *Qwen2.5-7B-Instruct* (Yang et al. 2024), and *GPT-4o* via the Azure API. BERT and DeBERTa are fine-tuned for three epochs with a learning rate of 5×10^{-5}. Llama models run on an 8×A6000 GPU server, while Mistral and DeepSeek are accessed through the TogetherAI platform. Since Task 4 (representatives statement generation) is inherently a generative task, we only evaluate

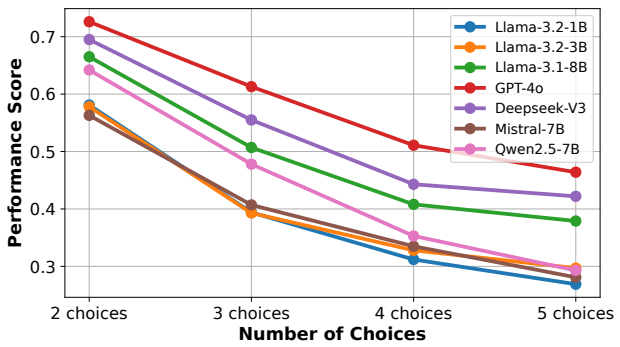

Figure 3: Models performance in Task 1 by varying the number of choices.

LLMs on it, setting a temperature of 0.0 for consistent comparisons and adjusting maximum output lengths per task.

Settings. For each classification-oriented task, we employ a time-based train/test split. Specifically, we reserve half of the samples from the less frequent labels (according to chronological order) as the test set, ensuring the training set remains balanced and temporally earlier. This protocol simulates real-world scenarios where future events must be predicted from past data.

Metrics. Task 1 is a multi-choice question with k ranging from 2 to 5. We calculate accuracy by checking whether the model identifies the single valid co-penholder. Tasks 2 and 3 are classification problems (multi-class and binary, respectively), so we report metrics robust to class imbalance, including F1-score, balanced accuracy (Bal. ACC), and PR AUC. Task 4 is evaluated using text-generation metrics (ROUGE) and semantic similarity (Sentence-BERT (Reimers 2019)) to measure how closely the generated statements match ground truth in style and content. For brevity, we present only two primary metrics per task in Table 2, with detailed breakdowns available at Appendix B in the extended version (Liang et al. 2025).

5.3 Task 1: Co-Penholder Judgement

This task evaluates LLMs' ability to ***identify strategic geopolitical alliances*** by selecting co-penholders for UNSC draft resolutions, requiring nuanced understanding of international relations and procedural norms. GPT-4o (0.726) and DeepSeek-V3 (0.695) dominate, demonstrating superior contextual reasoning and geopolitical knowledge. Smaller LLMs (e.g., Llama-3.2-1B: 0.581) lag significantly, while traditional models (BERT: 0.011) fail entirely, underscoring the necessity of LLM-scale architectures for complex political inference. In addition, we vary the number of candidate choices (2–5) to test models' robustness under increasing complexity. As shown in Figure 3, all models exhibit declining accuracy as choices increase, with GPT-4o maintaining dominance across all levels. The widening performance gaps as choices increase highlight the divergent capacities of LLMs to resolve ambiguity, validating that large, modern LLMs excel at synthesizing latent political knowledge, while smaller or traditional models lack the representational capacity for such nuanced reasoning.

5.4 Task 2: Representatives Voting Simulation

Focused on simulating country-specific voting behavior, this task tests models' ability to ***infer issue-specific voting patterns*** by analyzing how nations prioritize resolution content and contextual geopolitical dynamics. Unlike Task 1, which evaluates proactive alliance-building, Task 2 emphasizes reactive decision-making based on the interplay of national interests, ideological alignment, and external pressures. GPT-4o achieves the highest performance (0.823 Bal. ACC, 0.696 PR AUC), demonstrating a strong ability to model nuanced trade-offs. DeepSeek-V3 (0.724 Bal. ACC, 0.655 PR AUC) and Llama-3.2-3B (0.597 Bal. ACC) show moderate success but struggle with ambiguous cases, while Mistral-7B (0.426 Bal. ACC) performs poorly, reflecting its inability to systematically weigh competing factors. The results highlight LLMs' potential to simulate diplomatic behavior but reveal significant variance in their capacity to reason about issue-specific voting dynamics. We also assess whether model performance varies with respect to ***recency***, ***potential geopolitical disparities***, and different ***contextual metadata***. Detailed results are provided at Appendix C.1, C.2, and C.3 in the extended version (Liang et al. 2025).

5.5 Task 3: Draft Adoption Prediction

Whereas Task 2 centers on individual votes, Task 3 measures ***document-level outcome prediction***, requiring holistic reasoning about all 15 Council members to predict whether a resolution is eventually `[Adopted]` or `[NotAdopted]`. GPT-4o shows the best Bal. ACC (0.677) and a competitive macro-F1 (0.363), while Llama-3.2-3B surpasses others in macro-F1 (0.402) but has lower Bal. ACC (0.597). The divergence between Bal. ACC and F1 metrics reveals the challenge of modeling adoption mechanics, where understanding Council-wide political dynamics, potential veto threats, and support coalitions play roles. In addition, we analyze GPT-4o's robustness across time, regional authorship, and metadata conditions for this task. The corresponding evaluations are detailed at Appendix C.1, C.2, and C.3 in the extended version (Liang et al. 2025).

5.6 Task 4: Statement Generation

This task evaluates LLMs' ability to generate ***style-sensitive diplomatic statements*** that align with country-specific rhetoric and protocol. Qwen2.5-7B and DeepSeek-V3 tie for semantic fidelity (0.623 Cosine), demonstrating strong alignment with the intended meaning and tone of diplomatic statements. DeepSeek-V3 also leads in lexical overlap (0.207 ROUGE), suggesting better adherence to precise terminological requirements. Mistral-7B achieves high Cosine similarity (0.575) but modest ROUGE (0.194), indicating strength in paraphrasing and conceptual alignment rather than verbatim replication. All models underperform in ROUGE, exposing limitations in precise terminological alignment—a critical requirement for diplomatic drafting. This highlights the unresolved challenge of balancing creativity and protocol adherence in LLM-generated diplomatic

Model	Task 1		Task 2		Task 3		Task 4	
	(1/2)	(1/5)	Bal. ACC	PR AUC	Bal. ACC	Mac. F1	ROUGE	Cosine Sim.
BERT	0.011	0.010	0.537	0.396	0.333	0.328	/	/
DeBERTa	0.010	0.011	0.500	0.527	0.333	0.328	/	/
Llama-3.2-1B	0.581	0.269	0.546	0.185	0.320	0.326	0.033	0.329
Llama-3.2-3B	0.578	0.297	0.597	0.385	0.597	**0.402**	0.041	0.290
Llama-3.1-8B	0.665	0.379	0.530	0.168	0.357	0.359	0.039	0.355
Mistral-7B	0.563	0.281	0.426	0.268	0.529	0.140	0.194	0.575
GPT-4o	**0.726**	**0.464**	**0.823**	**0.696**	**0.677**	0.363	0.199	0.619
Qwen2.5-7B	0.642	0.293	0.699	0.375	0.578	0.241	0.201	**0.623**
DeepSeek-V3	0.695	0.422	0.724	0.655	0.668	0.351	**0.207**	**0.623**

Table 2: Our UNBench contains four tasks. For each task, we choose two metrics to show. (1/k) means choosing 1 from k choices, Bal. ACC is balance accuracy, PR AUC is precision-recall AUC. The best results for each metric are highlighted in **bold**, while the second-best results are underlined. More results could be found at Appendix B in the extended version.

text, particularly in capturing the formal and nuanced language of international diplomacy.

Cross-Task Summary. Each task targets a distinct facet of UNSC decision-making. Task 1 primarily tests ***textual and geopolitical reasoning*** in a multi-choice format, Task 2 and Task 3 emphasize ***political prediction capabilities*** (from simulating individual votes to forecasting final outcomes), and Task 4 stresses ***diplomatic language generation***, requiring alignment with formal protocols and country-specific rhetoric. Performance gaps across tasks and models highlight both the promise and complexity of applying LLMs to real-world international governance, reinforcing the need for dedicated benchmarks as UNBench.

6 Potential Applications

The UNBench offers significant value to both LLM researchers and stakeholders in international governance, enabling practical applications and advancing research in geopolitical AI. Below, we outline potential use cases for different stakeholders.

For LLM Researchers: UNBench provides a rich testbed for advancing research in LLMs, particularly in the context of geopolitical reasoning and time-series analysis: (1) ***Geopolitical Reasoning***: The tasks in the benchmark span a wide range of capabilities, from alliance identification (Task 1) to issue-specific voting prediction (Task 2), offering researchers a comprehensive framework for evaluating and improving LLMs' understanding of international relations. (2) ***Temporal Analysis***: With data spanning 30 years, UNBench enables time-series tasks such as predicting trends in diplomatic behavior, forecasting shifts in international alliances, or analyzing the impact of historical events (e.g., the end of the Cold War) on UNSC dynamics. For instance, researchers could use the dataset to predict how emerging global issues (e.g., climate change) will influence future resolutions. (3) ***Fine-Grained Prediction***: The benchmark's focus on multi-choice and generative tasks challenges researchers to develop models that balance precision and creativity. For example, improving ROUGE scores in Task 4 could lead to breakthroughs in generating protocol-compliant diplomatic text. (4) ***Bias Analysis***: UNBench provides an opportunity to study biases in LLMs' geopolitical reasoning, ensuring that models do not perpetuate stereotypes or oversimplify complex international dynamics.

For UN Stakeholders: The ability to predict and analyze UNSC decision-making using LLMs also has implications for diplomats, policymakers, and international organizations: (1) ***Draft Adoption Forecasting***: By predicting whether a draft resolution will be adopted (Task 3), stakeholders can proactively adjust negotiation strategies, allocate resources more effectively, and build coalitions to maximize the likelihood of success. (2) ***Voting Behavior Simulation***: Simulating country-specific voting behavior (Task 2) allows stakeholders to anticipate the positions of key nations, identify potential allies or opponents, and tailor diplomatic outreach accordingly. This could be particularly useful for smaller nations or NGOs seeking to navigate complex geopolitical landscapes.

7 Conclusion

This paper introduces UNBench, the first comprehensive benchmark for evaluating LLMs' capabilities in political science through UN Security Council records (1994-2024). By designing four interconnected tasks spanning the complete UN resolution lifecycle, we provide a more authentic framework for assessing LLMs' understanding of complex diplomatic dynamics. It not only addresses the current gap in LLM evaluation frameworks but also establishes a foundation for future research between artificial intelligence and international relations, demonstrating how LLMs could potentially assist in analyzing global governance processes. More detailed analyses can be found at Appendix in the extended version (Liang et al. 2025).

Acknowledgments

This material is based upon work supported by NSF awards (SaTC-2241068, IIS-2506643, and POSE-2346158), a Cisco Research Award, and a Microsoft Accelerate Foundation Models Research Award. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the National Science Foundation.

References

Bailey, M. A.; Strezhnev, A.; and Voeten, E. 2017. Estimating dynamic state preferences from United Nations voting data. *Journal of Conflict Resolution*, 61(2): 430–456.

Bastan, M.; Koupaee, M.; Son, Y.; Sicoli, R.; and Balasubramanian, N. 2020. Author's sentiment prediction. *arXiv preprint arXiv:2011.06128*.

Cheng, D.; Huang, S.; and Wei, F. 2023. Adapting large language models via reading comprehension. In *The Twelfth International Conference on Learning Representations*.

Data, M. E.; and Lab, S. 2022a. State Precinct-Level Returns 2018.

Data, M. E.; and Lab, S. 2022b. U.S. Senate Precinct-Level Returns 2020.

Devlin, J. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Grover, K.; Angara, S.; Akhtar, M. S.; and Chakraborty, T. 2022. Public wisdom matters! discourse-aware hyperbolic fourier co-attention for social text classification. *Advances in Neural Information Processing Systems*, 35: 9417–9431.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Hendrycks, D.; Burns, C.; Basart, S.; Critch, A.; Li, J.; Song, D.; and Steinhardt, J. 2021a. Aligning AI With Shared Human Values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021b. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

i Redondo, M. C.; and Llovera, G. C. 2025. United Nations Security Council Resolutions. Accessed: 2025-02-15.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Jin, Z.; Cao, J.; Guo, H.; Zhang, Y.; and Luo, J. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, 795–816.

Kohlenberg, P. J.; Godehardt, N.; Aris, S.; Sündermann, F.; Snetkov, A.; and Fall, J. 2019. Introducing UNSCdeb8 (beta). A Database for Corpus-Driven Research on the United Nations Security Council.

Kornilova, A.; and Eidelman, V. 2019. BillSum: A corpus for automatic summarization of US legislation. *arXiv preprint arXiv:1910.00523*.

Liang, Y.; Yang, L.; Wang, C.; Xia, C.; Meng, R.; Xu, X.; Wang, H.; Payani, A.; and Shu, K. 2025. Benchmarking llms for political science: A united nations perspective. *arXiv preprint arXiv:2502.14122*.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

LlamaIndex. 2024. LlamaParse. Open-source document parsing tool.

OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Reimers, N. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint arXiv:1908.10084*.

Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; and Hashimoto, T. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, 29971–30004. PMLR.

Shu, D.; Zhao, H.; Liu, X.; Demeter, D.; Du, M.; and Zhang, Y. 2024. LawLLM: Law large language model for the US legal system. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 4882–4889.

Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; and Liu, H. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3): 171–188.

Suzgun, M.; Scales, N.; Schärli, N.; Gehrmann, S.; Tay, Y.; Chung, H. W.; Chowdhery, A.; Le, Q. V.; Chi, E. H.; Zhou, D.; ; and Wei, J. 2022. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. *arXiv preprint arXiv:2210.09261*.

United Nations. 2025. UN Parallel Corpus. Accessed: 2025-02-15.

United Nations Security Council. 2024. United Nations Security Council Official Digital Library. Accessed: 14 Feb. 2025.

Voeten, E. 2013. Data and analyses of voting in the United Nations: General Assembly. *Routledge handbook of international organization*, 54–66.

Voeten, E.; Strezhnev, A.; and Bailey, M. 2009. United Nations General Assembly Voting Data.

Wojatzki, M.; Ruppert, E.; Holschneider, S.; Zesch, T.; and Biemann, C. 2017. Germeval 2017: Shared task on aspect-based sentiment in social media customer feedback. *Proceedings of the GermEval*, 1–12.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.