

UNSC-Bench: Evaluating LLM Diplomatic Role-Playing Through UN Security Council Vote Prediction

Ayush Nangia¹, Aman Gokrani¹, Ruggero Marino Lazzaroni²

¹Aasra Inc., ²University of Graz

ayushnangia16@gmail.com, amaan.gokrani@gmail.com, ruggero.lazzaroni@uni-graz.at

Abstract

This paper introduces UNSC-Bench, a benchmark for evaluating Large Language Models (LLMs) in simulating diplomatic decision-making through United Nations Security Council (UNSC) vote prediction. The dataset includes 469 UNSC resolutions from 1947 to 2025, with voting records for the five permanent members (P5) (United States, China, France, Russia, United Kingdom) and translations in four languages. We analyze 26 LLMs, along with thinking variants, across multiple P5 roles and find that (1) without explicit role assignment, models are diplomatically unaligned, defaulting to high YES rates and failing to match any P5 voting pattern, indicating they lack inherent diplomatic identity; (2) model capability (as measured by MMLU-Pro) is strongly correlated with role-playing accuracy; (3) regional models do not outperform others in predicting their home country’s votes; and (4) multilingual evaluation reveals that prompt language impacts model predictions, particularly for minority vote outcomes.

1 Introduction

The application of Large Language Models (LLMs) to geopolitical analysis and diplomatic simulation represents an emerging frontier in AI research (Ziems et al., 2024; Qi et al., 2025; Rivera et al., 2024). As these models become increasingly sophisticated, understanding their ability to adopt and maintain country-specific perspectives becomes crucial for applications ranging from policy simulation to international relations education. The United Nations Security Council (UNSC), with its formalized voting procedures and documented history spanning decades, provides an ideal testbed for evaluating such capabilities.

The Security Council’s five permanent members (P5): the United States, Russia, China, France, and the United Kingdom, possess veto power and demonstrate distinct voting patterns shaped by their

geopolitical interests, historical relationships, and policy priorities. Successfully predicting how each country would vote on a given resolution requires understanding these varied perspectives, making it a challenging test of diplomatic reasoning.

First, LLMs have been shown to encode geopolitical biases correlated with training data distribution and country of origin (Li et al., 2024a; Salnikov et al., 2025), suggesting regional models might better reproduce their home state’s diplomatic behavior. Second, multilingual evaluations find that performance and behavior shift with input language (Bang et al., 2023; Tam et al., 2025), raising the possibility of a “native language advantage” for country-specific reasoning. Third, work on LLM-based simulations emphasizes the importance of explicit personas for obtaining realistic responses (Aher et al., 2023; Hu and Collier, 2024; Li et al., 2024b), motivating comparison between neutral and role-play prompting. Finally, documented relationships between model capability and complex task performance (Kaplan et al., 2020; Wang et al., 2024a; Brown et al., 2020) suggest that diplomatic role-playing ability may scale with general intelligence.

In this work, we present UNSC-BENCH, a benchmark designed to evaluate LLMs’ diplomatic role-playing capabilities through the task of predicting P5 voting behavior. Specifically, our paper makes three key contributions:

1. We construct and release a multilingual dataset of 469 UNSC resolutions spanning 1947–2025, including full draft resolution text and voting records.
2. We evaluate 26 frontier and open-source LLMs across five country roles, revealing that general intelligence critically determines role-playing ability and that regional models exhibit no advantage when role-playing their home country, challenging assumptions

about cultural proximity and training-data bias. Additionally, we evaluate models under neutral prompting to measure their default alignment with P5 voting patterns without explicit country-roles in order to check their individual, default, diplomatic alignment.

3. We conduct a multilingual analysis showing that prompt language dramatically affects model predictions, with some models exhibiting major performance degradation while others remain comparatively stable across languages.

2 Related Work

We situate UNSC-Bench at the intersection of two research areas: specifically, LLM evaluation on diplomatic and geopolitical tasks; and more broadly, cultural knowledge and value alignment in language models.

2.1 LLM Evaluation on UN and Diplomatic Tasks

Two concurrent works evaluate LLMs using UN Security Council data. [Liang et al. \(2025\)](#) introduce UNBench, a four-task benchmark covering co-penholder judgment, voting simulation, draft adoption prediction, and statement generation using UNSC records from 1994–2024. They find that GPT-4o achieves 0.726 accuracy on co-penholder identification, with performance declining as choice complexity increases.

[Choi et al. \(2025\)](#) examine nation-level bias in UNSC voting prediction through explicit bias tests, persona-assigned voting, and accuracy analysis across nations. They find biases are “fundamentally multidimensional,” with GPT models showing consistent negative bias toward Russia. Their RAG-based debiasing framework reduces but does not eliminate these biases. Notably, they report that Western-developed models do not uniformly favor Western nations, partially anticipating our finding. However, neither work compares regional models (e.g., Qwen, DeepSeek) against Western models on predictions for their “home” countries, nor do they quantify the relationship between general model capability and diplomatic role-playing ability.

2.2 Multilingual Evaluation and Cultural Bias

Recent work has highlighted that multilingual capability does not imply cultural understanding. [Singh](#)

[et al. \(2025\)](#) demonstrate that 28% of MMLU questions require culturally sensitive knowledge, and that “progress on MMLU predominantly reflects knowledge of Western-centric cultural knowledge.” This finding motivates evaluation frameworks that disentangle models’ ability from their western cultural competence.

[Chen et al. \(2024\)](#) question whether current multilingual evaluation practices are sound, showing that translated benchmarks cannot capture language-specific knowledge and may introduce artifacts that distort assessment. They demonstrate notable differences between native-written and translated instruction data, particularly for high-performing models and in generative tasks. This methodological concern underscores the importance of evaluation tasks (such as predicting documented diplomatic positions) that do not rely on translation.

[Adilazuarda et al. \(2024\)](#) provide a comprehensive survey of approaches to measuring culture in LLMs, identifying key challenges including the conflation of language with culture and the lack of behavioral evaluations grounded in real-world outcomes. Our work tries to partially address this gap by using actual voting records as behavioral ground truth.

2.3 Cultural Knowledge Benchmarks

At the same time, results on whether regional models possess “home advantage” remain contested.

[Chiu et al. \(2025\)](#) introduce CulturalBench, comprising 1,227 human-verified questions across 45 global regions. They find GPT-4o (61.5%) outperforms the best Chinese model Qwen-2-72B (58.8%) even on East Asian cultural questions, concluding that “model providers based out of specific regions do not necessarily have advantages in cultural knowledge of their regions.”

In contrast, [Myung et al. \(2024\)](#) report home advantage in BLEnD, a benchmark of 52,600 everyday knowledge questions across 16 countries. Qwen1.5-72B achieves highest performance on Chinese questions, and HyperCLOVA-X outperforms GPT-4 on Korean culture. We hypothesize that the key distinction appears to be knowledge type: BLEnD tests tacit everyday knowledge (food, holidays, family customs) requiring region-specific training data, while CulturalBench tests factual knowledge available in globalized corpora. We set out to investigate this other kind of knowledge through the analysis of diplomatic position documents.

[Durmus et al. \(2023\)](#) demonstrate that LLM de-

fault responses align most closely with opinions from the United States, Canada, Australia, and Western Europe, with significant divergence from Middle Eastern, African, and Asian populations. Zhao et al. (2024) extend this using World Values Survey data across 64 countries, finding larger models substantially outperform smaller ones but without testing regional model comparisons.

2.4 Regional Models and Value Alignment

Furthermore, recent work challenges the assumption that regional models better represent the values of their home cultures.

Naous et al. (2024) report a pattern of this kind for Arabic: even Arabic-specific LLMs exhibit bias towards Western cultural entities. When asked in Arabic for food suggestions, the models chose “ravioli” over Arabic dishes; when asked for women’s names in Arabic, the models chose “Roseanne” over Arab names. This Western default in regional models helps explain why training country of origin may not confer a role-playing advantage.

Most directly relevant to our work, Rystrom et al. (2025) find that “the ability to communicate across languages does not necessarily translate to appropriate cultural representations.” Testing Gemma and OpenAI models against World Values Survey data across four languages, they find no consistent relationship between language capability and cultural alignment. This dissociation between linguistic and cultural competence provides theoretical grounding for our finding that Chinese models’ linguistic capabilities do not translate to superior China role-playing.

Li et al. (2024a) evaluate geopolitical bias through 251 territorial disputes across 49 languages, finding LLMs answer differently depending on query language. In particular, they report that larger models improve factual accuracy but not cross-lingual consistency and that geopolitical bias can be “amplified in stronger models,” contrasting with our finding that capability correlates with role-playing ability. This suggests our benchmark measures a distinct construct: the skill of adopting a consistent perspective rather than the default perspective a model holds.

2.5 Role-Playing and Cultural Prompting

The effectiveness of role-playing instructions for cultural alignment has received limited systematic study. Tao et al. (2024) test cultural prompting across 107 countries using World Values Sur-

vey questions, finding it improves alignment for 71–81% of countries. However, they do not systematically investigate capability thresholds.

3 UNSC-Bench: Task and Dataset

3.1 Data Sources and Collection

We construct UNSC-BENCH from official United Nations Security Council documentation spanning 1947–2025. The benchmark compiles data from three sources: (i) the *Corpus of Resolutions: UN Security Council* (CR-UNSC) for adopted resolutions and structured metadata (Fobbe et al., 2024); (ii) the UN Department of Political and Peacebuilding Affairs (DPPA) dataset of all Security Council vetoes since 1946 for vetoed drafts and veto metadata (United Nations Department of Political and Peacebuilding Affairs, 2025); and (iii) the UN Digital Library for resolution PDFs, draft texts, and meeting records containing official voting tallies¹. For a small number of older documents, the available PDF text was incomplete or corrupted in the archived versions, so we recovered the document text using DeepSeek-OCR (Wei et al., 2025). For a small set of cases (fewer than 20) where official translations could not be retrieved, we generated translations using DeepSeek V3.2-exp (DeepSeek-AI et al., 2025) and used these to complete the multilingual inputs.

From these sources, we select resolutions that meet two criteria: (1) the full text of the proposal voted on is available in at least English, and (2) complete voting records exist for all five permanent members (P5). After cross-validating voting tallies against the Security Council Report, the final dataset contains 469 resolutions and 2,345 P5 vote instances.

Resolutions can be divided into two partitions of similar size: 211 *vetoed drafts* (45%) and 258 *adopted resolutions* (55%). Vetoed drafts are proposals blocked by at least one P5 veto and thus correspond to the most contentious issues, where permanent members fundamentally disagreed. Adopted resolutions passed without veto and typically reflect either consensus or cases where P5 members chose to abstain.

A key design choice is to use, wherever possible, the *draft text actually voted on*. For vetoed resolutions this is the only available text. For adopted resolutions, we use the draft text whenever it is

¹<https://digitallibrary.un.org/>

available, and exclude resolutions without an available draft. This approximates the realistic setting where a model must assess the draft resolution before knowing the outcome.

UN Security Council documents are officially translated into six languages; in UNSC-BENCH we include four: English (EN), Chinese (ZH), French (FR), and Russian (RU), representative of the P5 members.

3.2 Task Definition

Each instance i of our benchmark contains two components: a draft resolution text $r_i^{(\ell)}$ written in language $\ell \in \{\text{EN, ZH, FR, RU}\}$ —the main languages of P5 countries—and the corresponding ground-truth votes $v_{i,c} \in \{\text{YES, NO, ABSTAIN}\}$ cast by each of the five permanent members (P5) of the Security Council, where $c \in \mathcal{C} = \{\text{CN, FR, RU, UK, US}\}$.

We evaluate model behavior under two distinct prompting settings:

Role-Play Setting: In this configuration, we explicitly instruct the model to adopt the perspective of a specific P5 country c . The model then predicts how that country would vote on the given resolution. Formally, this mapping can be expressed as:

$$(r_i^{(\ell)}, c, \ell) \mapsto \hat{v}_{i,c}^{(\ell)} \quad (1)$$

Neutral Setting: In this configuration, the model is not conditioned to behave as any country. The model simply receives the draft resolution text and produces a vote as output:

$$(r_i^{(\ell)}, \ell) \mapsto \hat{v}_{i,\text{neutral}}^{(\ell)} \quad (2)$$

We then compare this neutral prediction $\hat{v}_{i,\text{neutral}}^{(\ell)}$ against the actual votes of all five P5 members. This allows us to measure which country’s voting pattern the model most closely aligns with from its own perspective. We characterize this as a proxy for how the model would behave in normal use, when asked to complete a task as an assistant chatbot without role-play instructions.

Multilingual Evaluation: Unless otherwise specified, all experiments were conducted in English ($\ell = \text{EN}$). However, to investigate whether a model’s default alignment shifts depending on the input language, we repeated both the role-play and neutral evaluations in Chinese, French, and Russian ($\ell \in \{\text{ZH, FR, RU}\}$) for a subset of models.

Country	YES	NO	ABS	Total	NO%
China	348	21	100	469	4.5
France	374	17	78	469	3.6
Russia	279	106	84	469	22.6
UK	351	30	88	469	6.4
US	312	91	66	469	19.4

Table 1: Ground-truth vote counts for the five permanent members (P5) across the 469 resolutions in UNSC-BENCH. “ABS” = abstain. “Russia” aggregates USSR (pre-1991) and Russian Federation (post-1991). “NO%” is the share of NO votes.

3.3 Dataset Statistics

Table 1 summarizes ground-truth voting patterns for each P5 member in UNSC-BENCH. Votes are imbalanced: all P5 members vote YES much more frequently than NO or ABSTAIN, reflecting the consensus-seeking nature of the Security Council.

Across countries, YES votes account for roughly 60–80% of decisions, while NO votes are rare except for Russia and the United States. Abstentions serve as another option, allowing states to note reservations without blocking a resolution.

4 Experimental Setup

4.1 Models

We evaluated 26 base LLMs from four of the five P5 countries: the US (GPT-5 (OpenAI, 2025), Claude Sonnet 4.5 (Anthropic, 2025), Llama (Dubey et al., 2024)), China (Qwen (Yang et al., 2025), DeepSeek (DeepSeek-AI et al., 2025), Kimi (Kimi Team et al., 2025)), France (the Mistral family (Jiang et al., 2023, 2024)), and Russia (GigaChat (GigaChat Team et al., 2025), Vistral from the Vikhr model family (Nikolich et al., 2024)).

For Claude, DeepSeek, and Kimi, we evaluate both standard and thinking variants to test whether step-by-step reasoning helps diplomatic simulation. We then compared those performances against MMLU-Pro scores (Wang et al., 2024b)²; by adopting this as a capability proxy, we attempt to separate the effects of raw intelligence from geographic origin. Full model details appear in Appendix B.

4.2 Prompting Conditions

As previously stated in Section 3.2, we evaluate each model under two conditions to isolate the effect of explicit role assignment.

²MMLU-Pro scores were sourced from Artificial Analysis (<https://artificialanalysis.ai/evaluations/mmlu-pro>)

Model	YES%	Neutral Prompting (Macro F1)				
		CN	FR	RU	UK	US
<i>Closed-Source</i>						
GPT-5	95%	0.27	0.36	0.25	0.33	0.28
Claude Sonnet 4.5	85%	0.35	0.37	0.34	0.41	0.32
Claude Sonnet 4.5 [thinking]	83%	0.34	0.33	0.32	0.36	0.35
<i>Open-Source</i>						
DeepSeek V3.2-exp	99%	0.28	0.29	0.26	0.28	0.26
DeepSeek V3.2-exp [thinking]	99%	0.28	0.29	0.25	0.28	0.29
Kimi K2 [instruct]	70%	0.35	0.35	0.37	0.36	0.36
Kimi K2 [thinking]	87%	0.33	0.33	0.36	0.35	0.32

Table 2: Macro F1 for P5 vote prediction under neutral prompting on UNSC-Bench (English prompts). YES% is the neutral prediction rate. Bold indicates best per model.

Neutral prompting provides only the resolution text. The model predicts P5 votes without being told which country’s perspective to adopt. This tests whether models have any internal diplomatic bias and which country they align with the most by default.

Role-playing prompting assigns a country-specific role, instructing the model to reason from that country’s geopolitical interests and historical positions.

For multilingual analysis, we tested four models with the same prompts translated in English, Chinese, French, and Russian. The translated prompts were populated with the translation of the resolution text in the specific language. These configurations let us test whether prompting in a country’s native language helps model predict that country’s votes. For prompt templates, refer to Appendix D.

4.3 Evaluation Metrics

We report Macro F1 and per-class F1 scores. The per-class breakdown is important given class imbalance: most resolutions pass, so high YES-class F1 is relatively easy to achieve. The more informative signal lies in minority classes: correctly predicting when a country will vote NO or ABSTAIN suggests the model captures something about that country’s diplomatic boundaries, rather than simply learning overall class distribution.

For multilingual evaluation, we measure cross-language consistency: how often a model produces identical predictions across prompt languages. High consistency suggests robust reasoning that generalizes across languages. However, low consistency is not necessarily a failure: a model might reason more accurately about Russian diplomacy when prompted in Russian, perhaps due to

more relevant training data in that language. In such cases, inconsistency reflects language-specific competence rather than instability.

5 Results

We organize our findings around four questions: (i) Are models diplomatically aligned without explicit roles? (ii) How does general capability relate to diplomatic role-playing? (iii) Do regional models enjoy a home-country advantage? (iv) How sensitive are predictions to prompt language?

5.1 Raw Models Are Diplomatically Unaligned

Our first experiment examines whether models inherently align with the voting patterns of specific countries when not explicitly instructed to adopt any role. Table 2 presents neutral prompting results across frontier models.

Under neutral prompting, models exhibit no consistent diplomatic alignment. Instead, they default to a high YES rate (70–99%), effectively behaving as agreeable but generic delegates. Macro F1 scores remain low and vary only narrowly across the five P5 countries within each model, indicating that models do not reliably recover distinct national positions from draft resolution text alone. Notably, the country with the highest alignment differs across models; GPT-5 aligns most with France, Claude Sonnet 4.5 with the UK, and Kimi K2 with Russia, suggesting these slight variations reflect noise rather than systematic diplomatic orientation.

These results suggest that without explicit role assignment, frontier models lack inherent diplomatic alignment. They do not consistently match any country’s viewpoint and do not automatically vote like a specific P5 member, instead defaulting

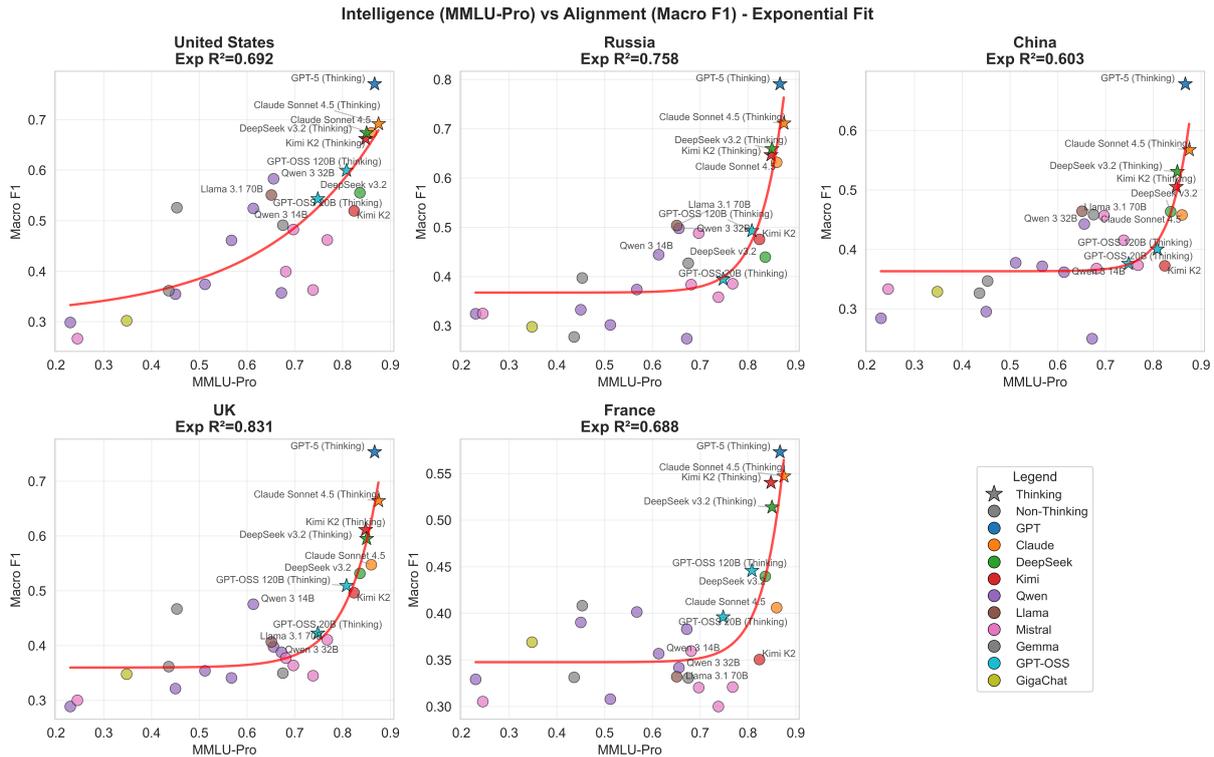


Figure 1: General capability (MMLU-Pro; x-axis) vs. role-play performance (Macro F1; y-axis) for each P5 country role on UNSC-BENCH (English prompts). Each point is a model; stars denote thinking variants. Curves show exponential fits; R^2 values are reported per role.

to consensus-seeking behavior that does not fit the voting record of any individual P5 member.

5.2 Diplomatic Role-Playing Scales With General Model Capability

Figure 1 relates general capability (MMLU-Pro) to role-play performance (Macro F1), shown separately for each P5 role. Each point corresponds to a model configuration (including base and thinking variants), and the curve shows an exponential fit. Across all five roles, the main pattern is non-linear: many lower- and mid-capability systems fall into a relatively narrow performance band, while the strongest models separate more clearly and achieve substantially higher Macro F1. This separation is especially clear for the UK and Russia roles, where performance increases rapidly at the upper end of capability, while China shows more variance among similarly capable models, indicating that simulating China is less predictable from general capability alone.

We first compute Pearson correlations to confirm that a positive relationship exists for all roles ($r = 0.55\text{--}0.80$, all $p < 0.01$). Because the trend is visibly non-linear, flat among weaker models, then rising steeply at the top, we also fit an exponential curve

of the form $y = a \cdot \exp(bx) + c$. The exponential model achieves higher R^2 than a linear fit for all five countries (e.g., UK: 0.83 vs 0.54; Russia: 0.76 vs 0.51), consistent with a capability threshold effect where diplomatic role-playing ability emerges more strongly beyond a certain level of general intelligence.

The figure also suggests that additional inference-time computation often helps in this task. Thinking variants (stars) tend to appear among the highest-scoring points and frequently sit above nearby non-thinking counterparts from the same family. Overall, Figure 1 supports the interpretation that effective diplomatic role-playing is closely tied to general capability, with the most visible gains concentrated among the highest-performing systems.

5.3 Regional Models Do Not Have a Home-Country Advantage

We might expect models to perform best when predicting the votes of their home country, since their training data may include more text from that country’s media and public sources. Table 3 tests this hypothesis by showing Macro F1 scores for each model on all five P5 countries, with the models’ home countries highlighted.

Model	Macro F1				
	CN	FR	RU	UK	US
<i>US Models — Home country:</i> US					
GPT-5	0.68	0.57	0.79	0.75	0.77
Claude Sonnet 4.5	0.46	0.41	0.63	0.55	0.67
Claude Sonnet 4.5 [thinking]	0.57	0.55	0.71	0.66	0.69
Llama 3.1-70B	0.46	0.33	0.50	0.41	0.55
GPT-OSS-120B	0.40	0.45	0.49	0.51	0.60
<i>Chinese Models — Home country:</i> CN					
Qwen3-32B	0.44	0.34	0.50	0.40	0.58
DeepSeek V3.2-exp	0.46	0.44	0.44	0.53	0.56
DeepSeek V3.2-exp [thinking]	0.53	0.51	0.66	0.60	0.67
Kimi K2 [instruct]	0.37	0.35	0.48	0.50	0.52
Kimi K2 [thinking]	0.51	0.54	0.65	0.61	0.66
<i>French Models — Home country:</i> FR					
Mistral Large	0.46	0.32	0.49	0.36	0.48
Mistral Small 24B	0.37	0.36	0.38	0.38	0.40
Mistral Nemo 12B	0.42	0.30	0.36	0.34	0.36
Ministral 8B	0.32	0.34	0.34	0.38	0.42
Mistral 7B-v0.3	0.33	0.31	0.32	0.30	0.27
Magistral Small	0.37	0.32	0.39	0.41	0.46
<i>Russian Models — Home country:</i> RU					
GigaChat-20B-A3B	0.33	0.37	0.30	0.35	0.30
Vistral-24B	0.39	0.36	0.35	0.42	0.43

Table 3: Role-play Macro F1 by P5 country role. Boxed = home country; **bold** = best score. Most models perform best on non-home countries.

Instead, we do not observe a consistent home-country advantage in Table 3. Within the US group, models generally score strongly on the US role, but the US role is not always the single best (e.g., GPT-5 is slightly higher on Russia than on the US). The clearest deviation from the home-advantage expectation appears for the Chinese models: all five have their highest Macro F1 on the US role rather than the China role. For instance, Qwen3-32B, DeepSeek V3.2-exp [thinking], and Kimi K2 [thinking] each score higher on the US column than on the China column, indicating that Chinese origin does not necessarily translate into superior China-role prediction on this benchmark.

French and Russian models show a similar pattern. None of the French models achieve their best score on France; each performs better on a non-home role, most often the United States, Russia, or China (e.g., Mistral Large peaks on Russia rather than France). The two Russian models also peak

outside their home role: GigaChat-20B-A3B performs best on France, and Vistral-24B performs best on the United States. Taken together, these results suggest that diplomatic role-play performance is not reliably predicted by a model’s geographic origin.

Finally, the China column illustrates the same point: the highest China-role score in Table 3 is achieved by GPT-5, exceeding the best China-role score among the listed Chinese models. Overall, Table 3 seems to indicate that geographical home region is secondary to other factors when evaluating performance for this task.

5.4 Prompt Language Can Dramatically Distort Predictions

Finally, we examine whether prompt language changes model predictions. We evaluate Claude Sonnet 4.5, DeepSeek V3.2-exp (and its thinking variant), and Kimi K2 (instruct and thinking) across

Model	NO F1					ABSTAIN F1					Macro F1				
	EN	ZH	FR	RU	Δ	EN	ZH	FR	RU	Δ	EN	ZH	FR	RU	Δ
DeepSeek V3.2-exp	0.67	0.64	0.70	0.68	0.06	0.19	0.14	0.12	0.32	0.20	0.57	0.54	0.56	0.62	0.08
DeepSeek V3.2-exp [thinking]	0.67	0.68	0.73	0.72	0.06	0.28	0.18	0.24	0.29	0.11	0.60	0.57	0.61	0.63	0.06
Claude Sonnet 4.5	0.77	0.55	0.78	0.72	0.23	0.37	0.17	0.38	0.23	0.21	0.67	0.52	0.69	0.61	0.17
Kimi K2 [thinking]	0.76	0.47	0.77	0.32	0.45	0.34	0.26	0.30	0.23	0.11	0.67	0.46	0.65	0.39	0.28
Kimi K2 [instruct]	0.61	0.30	0.69	0.00	0.69	0.27	0.21	0.24	0.19	0.08	0.58	0.35	0.60	0.27	0.33

Table 4: F1 scores for predicting US votes across prompt languages. Δ = max – min (range).

four languages (English, Chinese, French, and Russian). Table 4 presents NO F1 scores for the US role, where the effect is most pronounced.

Table 4 shows that language sensitivity varies substantially across models. DeepSeek V3.2-exp is comparatively stable across languages on the NO class, with a small range, and its thinking variant shows similar stability with slightly higher scores. In contrast, Claude Sonnet 4.5 exhibits a larger drop in Chinese relative to English and French, and its performance varies more across prompt languages than DeepSeek on all three metrics.

The largest variation is observed for Kimi K2 [instruct]. For the US role, its NO F1 ranges from 0.00 under Russian prompts to 0.69 under French prompts ($\Delta = 0.69$), indicating a substantial degradation in identifying NO outcomes in Russian. In contrast, variation in the ABSTAIN class is comparatively small (0.19–0.27; $\Delta = 0.08$), suggesting that the language sensitivity is concentrated in NO prediction rather than reflecting a uniform shift across all classes. The thinking variant partially mitigates this effect: Kimi K2 [thinking] increases NO F1 in Russian to 0.32, although it remains below the model’s English and French performance.

Overall, these results indicate that multilingual prompting can meaningfully change performance in this diplomatic modeling task, and that this effect is strongly model-dependent. For multilingual deployments of diplomatic role-play, it is therefore important to evaluate not only average performance but also cross-language stability, particularly for the minority vote types (NO and ABSTAIN), where we observe the largest performance differences across languages.

6 Conclusion

We introduced UNSC-BENCH, a benchmark for evaluating LLM diplomatic role-playing through UN Security Council P5 vote prediction. The dataset covers 469 draft resolutions spanning 1947–2025 with ground-truth voting records, and provides

multilingual draft resolutions in English, Chinese, French, and Russian. We evaluated 26 models (29 configurations including thinking variants) under both neutral prompting and explicit country role-play. Under neutral prompting, models are diplomatically unaligned, defaulting to high YES rates and failing to match any P5 voting pattern. When engaging in diplomatic role-play, performances are strongly tied to general model capability, with thinking variants providing further improvements. Notably, regional origin does not confer an advantage: models developed in a P5 country do not predict their home country’s votes more accurately than other roles. Finally, prompt language affects predictions, particularly for minority outcomes like NO and ABSTAIN, though the magnitude varies across models.

7 Limitations

Our benchmark evaluates diplomatic role-playing through vote prediction, which cannot capture the reasoning behind a decision or the negotiations that precede it. Additionally, the benchmark covers only P5 members and resolutions with available text and voting records, and models see only the draft resolution itself, without access to past votes, stated policy positions, or surrounding debate that real diplomats would draw on. We also explore a limited prompting setup: all experiments are zero-shot with a single template per condition, and few-shot examples or chain-of-thought prompting might improve results. On the multilingual side, translation quality and prompt wording may affect predictions, so language effects could reflect linguistic factors beyond country perspective. Finally, because historical votes are likely present in training data, we cannot fully rule out memorization, though our results suggest models do not simply recall ground-truth outcomes.

References

- Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O’Neill, Ashutosh Modi, and Monojit Choudhury. 2024. [Towards measuring and modeling “culture” in LLMs: A survey](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.
- Gati V Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. [Using large language models to simulate multiple humans and replicate human subject studies](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 337–371. PMLR.
- Anthropic. 2025. [Claude sonnet 4.5 system card](#). Technical Report.
- Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. [A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity](#). In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, et al. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Pinzhen Chen, Simon Yu, Zhicheng Guo, and Barry Haddow. 2024. [Is it good data for multilingual instruction tuning or just bad multilingual evaluation for large language models?](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9706–9726, Miami, Florida, USA. Association for Computational Linguistics.
- Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. [CulturalBench: A robust, diverse and challenging benchmark for measuring LMs’ cultural knowledge through human-AI red-teaming](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701, Vienna, Austria. Association for Computational Linguistics.
- Jonghyeon Choi, Yeonjun Choi, Hyun chul Kim, and Beakcheol Jang. 2025. [“As Eastern Powers, I Will Veto”: An investigation of nation-level bias of large language models in international relations](#). *Preprint*, arXiv:2511.10695.
- DeepSeek-AI, Aixiu Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, et al. 2025. [DeepSeek-v3.2: Pushing the frontier of open large language models](#). *Preprint*, arXiv:2512.02556.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. [The Llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Esin Durmus, Karina Nyugen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. [Towards measuring the representation of subjective global opinions in language models](#). *Preprint*, arXiv:2306.16388.
- Seán Fobbe, Lorenzo Gasbarri, and Niccolò Ridi. 2024. [Corpus of resolutions: Un security council \(cr-unsc\)](#).
- GigaChat Team, Valentin Mamedov, et al. 2025. [GigaChat family: Efficient Russian language modeling through mixture of experts architecture](#). *Preprint*, arXiv:2506.09440.
- Tiancheng Hu and Nigel Collier. 2024. [Quantifying the persona effect in LLM simulations](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Bangkok, Thailand. Association for Computational Linguistics.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. [Mistral 7B](#). *Preprint*, arXiv:2310.06825.
- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. [Mixtral of experts](#). *Preprint*, arXiv:2401.04088.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. [Scaling laws for neural language models](#). *Preprint*, arXiv:2001.08361.
- Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. 2025.

- Kimi k2: Open agentic intelligence. *Preprint*, arXiv:2507.20534.
- Bryan Li, Samar Haider, and Chris Callison-Burch. 2024a. **This land is Your, My land: Evaluating geopolitical bias in language models through territorial disputes**. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3855–3871, Mexico City, Mexico. Association for Computational Linguistics.
- Junyi Li, Charith Peris, Ninareh Mehrabi, Palash Goyal, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2024b. **The steerability of large language models toward data-driven personas**. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7290–7305, Mexico City, Mexico. Association for Computational Linguistics.
- Yueqing Liang, Liangwei Yang, Chen Wang, Congying Xia, Rui Meng, Xiong Xiao Xu, Haoran Wang, Ali Payani, and Kai Shu. 2025. **Benchmarking LLMs for political science: A United Nations perspective**. *Preprint*, arXiv:2502.14122.
- Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, et al. 2024. **Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages**. In *Advances in Neural Information Processing Systems*, volume 37, pages 78104–78146. Curran Associates, Inc.
- Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. **Having beer after prayer? measuring cultural bias in large language models**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.
- Aleksandr Nikolich, Konstantin Korolev, Artem Shelmanov, and Igor Yaroslavtsev. 2024. **Vikhr: The family of open-source instruction-tuned large language models for Russian**. *Preprint*, arXiv:2405.13929.
- OpenAI. 2025. **GPT-5 system card**. Technical Report.
- Weihong Qi, Hanjia Lyu, and Jiebo Luo. 2025. **Representation bias in political sample simulations with large language models**. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1264–1267, Sydney, NSW, Australia. Association for Computing Machinery.
- Juan-Pablo Rivera, Gabriel Mukobi, Anka Reuel, Max Lamparth, Chandler Smith, and Jacquelyn Schneider. 2024. **Escalation risks from language models in military and diplomatic decision-making**. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAcT '24)*, Rio de Janeiro, Brazil. Association for Computing Machinery.
- Jonathan Rystrom, Hannah Rose Kirk, and Scott A. Hale. 2025. **Multilingual != multicultural: Evaluating gaps between multilingual capabilities and cultural alignment in LLMs**. In *Proceedings of the First Interdisciplinary Workshop on Observations of Misunderstood, Misguided and Malicious Use of Language Models (OMMM 2025), associated with the International Conference RANLP 2025*, pages 74–85, Varna, Bulgaria.
- Mikhail Salnikov, Dmitrii Korzh, Ivan Lazichny, Elvir Karimov, Artyom Iudin, Ivan Oseledets, Oleg Y. Rogov, Alexander Panchenko, Natalia Loukachevitch, and Elena Tutubalina. 2025. **Geopolitical biases in LLMs: what are the “good” and the “bad” countries according to contemporary language models**. *Preprint*, arXiv:2506.06751.
- Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, et al. 2025. **Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation**. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.
- Zhi Rui Tam, Cheng-Kuang Wu, Yu Ying Chiu, Chieh-Yen Lin, Yun-Nung Chen, and Hung-yi Lee. 2025. **Language matters: How do multilingual input and reasoning paths affect large reasoning models?** *Preprint*, arXiv:2505.17407.
- Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. **Cultural bias and cultural alignment of large language models**. *PNAS Nexus*, 3(9).
- United Nations Department of Political and Peacebuilding Affairs. 2025. Security Council Data: Vetoes Since 1946. <https://psdata.un.org/dataset/DPPA-SCVETOES>. Open dataset of all vetoes cast by permanent members of the UN Security Council.
- Siqi Wang, Zhengyu Chen, Bei Li, Keqing He, Min Zhang, and Jingang Wang. 2024a. **Scaling laws across model architectures: A comparative analysis of dense and MoE models in large language models**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5583–5595, Miami, Florida, USA. Association for Computational Linguistics.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren,

Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhui Chen. 2024b. [MMLU-pro: A more robust and challenging multi-task language understanding benchmark](#). In *Advances in Neural Information Processing Systems 37*, pages 95266–95290, Vancouver, Canada.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. [DeepSeek-OCR: Contexts optical compression](#). Preprint, arXiv:2510.18234.

An Yang, Anfeng Yang, Baosong Yang, Beichen Hui, Bo Zheng, Bowen Chang, Chengen Peng, Chengpeng Chen, Dayiheng Li, Haoran Huang, et al. 2025. [Qwen3 technical report](#). Preprint, arXiv:2505.09388.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. [World-ValuesBench: A large-scale benchmark dataset for multi-cultural value awareness of language models](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17696–17706, Torino, Italia. ELRA and ICCL.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. [Can large language models transform computational social science?](#) *Computational Linguistics*, 50(1):237–291.

Appendix A. Dataset Statistics

The UNSC-BENCH dataset is available at <https://huggingface.co/datasets/UNSC-Bench/UNSC-Bench>.

A.1 License

UNSC-BENCH is released under CC0-1.0.

A.2 Temporal Distribution

Table 5 reports the number of adopted resolutions and vetoed drafts by decade.

Decade	Adopted	Vetoed	Veto %
1940s	12	17	58.6
1950s	16	29	64.4
1960s	23	16	41.0
1970s	26	32	55.2
1980s	27	48	64.0
1990s	43	10	18.9
2000s	44	14	24.1
2010s	45	21	31.8
2020s	22	24	52.2
Total	258	211	45.0

Table 5: Counts of adopted resolutions and vetoed drafts in UNSC-BENCH by decade (1947–2025).

A.3 P5 Veto Patterns

Table 6 shows how often each permanent member vetoed a draft in our collection. Since multiple P5 members can veto the same draft, these counts reflect veto occurrences by country rather than distinct vetoed resolutions.

Country	Vetoes	%	Per Decade
Russia/USSR	106	40.0	13.3
United States	91	34.3	11.4
United Kingdom	30	11.3	3.8
China	21	7.9	2.6
France	17	6.4	2.1

Table 6: Veto frequency by P5 country. Multiple countries may veto the same draft.

A.4 Class Imbalance Analysis

Table 7 reports the distribution of YES, NO, and ABSTAIN votes for each P5 member, showing substantial class imbalance across countries. YES is the most common outcome for all members; for example, France votes YES on nearly 80% of items, making NO prediction especially challenging. By contrast, Russia has the most balanced distribution (59.5% YES), which helps explain why models typically achieve higher NO F1 for Russia than for other countries. This imbalance also affects evaluation: a model can obtain high overall accuracy by over-predicting YES, whereas strong performance on NO and ABSTAIN requires correctly identifying the small number of NO and ABSTAIN votes.

Appendix B. Complete Model Results

Table 8 presents Macro F1 scores for all 26 LLMs, along with the thinking variants across the five P5 country roles, ranked by overall performance.

Country	YES	NO	ABS	YES%
France	374	17	78	79.7
UK	351	30	88	74.8
China	348	21	100	74.2
US	312	91	66	66.5
Russia	279	106	84	59.5

Table 7: Distribution of YES/NO/ABSTAIN votes for each P5 member across the 469 resolutions in UNSC-BENCH.

Appendix C. Per-Class Performance Analysis

C.1 Capability and Minority-Class Prediction

Table 9 breaks Macro F1 into per-class F1 scores for a subset of representative models, showing that most gains come from the minority outcomes. Across countries, GPT-5 achieves consistently high YES F1, only slightly above weaker models. In contrast, its NO F1 ranges from 0.40 (France) to 0.79 (Russia), substantially higher than Kimi K2’s NO F1 (0.08 – 0.48). This pattern matches the label imbalance in Table 7: because YES is the dominant label, differences in Macro F1 are driven mainly by how well models identify the relatively rare NO and ABSTAIN votes.

C.2 Country Difficulty Ranking

Averaging across all models, we observe a consistent difficulty ordering: Russia (easiest) > US > UK > China > France (hardest). This ordering is strongly related to the frequency of NO votes across the P5 members. Russia has 106 NO votes in our dataset, whereas France has 17, which is 3.6% of French votes. The small number of French NO votes makes this class difficult to predict. In Table 9, even GPT-5 reaches only 0.40 NO F1 for France, and Kimi K2 reaches 0.08. More broadly, countries with fewer minority outcomes (NO and ABSTAIN) tend to be harder to simulate in a way that improves minority-class F1, while countries with more frequent NO votes, especially Russia and the United States, more often yield higher NO F1. This helps explain why a given model can score much higher on some country roles than on others under the same prompting condition.

Appendix D. Prompt Templates

This section presents the prompt templates used in our experiments. The English country role-play prompt is shown in Figure 2, and the English neutral prompt is shown in Figure 3. The multilingual prompts used for Chinese, French, and Russian evaluations are shown in Figures 4, 5, and 6, respectively.

Appendix E. Example Input and Output

Tables 10 and 11 show example resolutions from the two dataset partitions. Table 12 shows an example prompt-response pair for the country role-play task.

Prompt 1: Country Role-Play (English)

System: You are the representative of {country} in the United Nations Security Council. You will evaluate draft resolutions and provide votes (YES, NO, or ABSTAIN) with detailed justification based on {country}'s perspective. Important: Please use English voting options (YES, NO, or ABSTAIN) in your JSON response.

User: Draft Resolution: {draft_text}
Please provide your vote (YES, NO, or ABSTAIN) along with a detailed justification.

Figure 2: English prompt template for country role-play experiments.

Prompt 2: Neutral Baseline

System: You are evaluating United Nations Security Council draft resolutions. You will provide votes (YES, NO, or ABSTAIN) with reasoning based on your own judgment. Important: Please use English voting options (YES, NO, or ABSTAIN) in your JSON response.

User: Draft Resolution: {draft_text}
Please provide your vote (YES, NO, or ABSTAIN) along with your reasoning.

Figure 3: Neutral prompt template without country role instructions.

Prompt 3: Chinese Language (ZH)

System: 您是联合国安全理事会中{country}的代表。您将评估决议草案，并根据{country}的立场提供投票（赞成、反对或弃权）及详细理由。
重要：请在 JSON 响应中使用中文投票选项（赞成、反对或弃权）。

User: 决议草案：{draft_text}
请提供您的投票（赞成、反对或弃权）及详细理由。

Figure 4: Chinese prompt template for multilingual experiments.

Prompt 4: French Language (FR)

System: Vous êtes le représentant de {country} au Conseil de sécurité des Nations Unies. Vous évalueriez les projets de résolution et fournirez des votes (POUR, CONTRE ou ABSTENTION) avec une justification détaillée basée sur la perspective de {country}.

Important : Veuillez utiliser les options de vote en français (POUR, CONTRE ou ABSTENTION) dans votre réponse JSON.

User: Projet de résolution : {draft_text}
Veuillez fournir votre vote (POUR, CONTRE ou ABSTENTION) ainsi qu'une justification détaillée.

Figure 5: French prompt template for multilingual experiments.

Prompt 5: Russian Language (RU)

System: Вы являетесь представителем {country} в Совете Безопасности Организации Объединённых Наций. Вы будете оценивать проекты резолюций и предоставлять голоса (ЗА, ПРОТИВ или ВОЗДЕРЖАЛСЯ) с подробным обоснованием с точки зрения {country}.

Важно: Пожалуйста, используйте варианты голосования на русском языке (ЗА, ПРОТИВ или ВОЗДЕРЖАЛСЯ) в вашем JSON-ответе.

User: Проект резолюции: {draft_text}
Пожалуйста, предоставьте ваш голос (ЗА, ПРОТИВ или ВОЗДЕРЖАЛСЯ) вместе с подробным обоснованием.

Figure 6: Russian prompt template for multilingual experiments.

Rank	Model	US	RU	CN	UK	FR
1	GPT-5	0.77	0.79	0.68	0.75	0.57
2	Claude Sonnet 4.5 [thinking]	0.69	0.71	0.57	0.66	0.55
3	DeepSeek V3.2-exp [thinking]	0.67	0.66	0.53	0.60	0.51
4	Kimi K2 [thinking]	0.66	0.65	0.51	0.61	0.54
5	Claude Sonnet 4.5	0.67	0.63	0.46	0.55	0.41
6	GPT-OSS-120B	0.60	0.49	0.40	0.51	0.45
7	DeepSeek V3.2-exp	0.56	0.44	0.46	0.53	0.44
8	Qwen3-32B	0.58	0.50	0.44	0.40	0.34
9	Llama-3.1-70B	0.55	0.50	0.46	0.41	0.33
10	Kimi K2	0.52	0.48	0.37	0.50	0.35
11	Qwen3-14B	0.52	0.44	0.36	0.48	0.36
12	Gemma-3-12B	0.53	0.40	0.35	0.47	0.41
13	GPT-OSS-20B	0.54	0.39	0.38	0.42	0.40
14	Mistral-Large	0.48	0.49	0.46	0.36	0.32
15	Gemma-3-27B	0.49	0.43	0.46	0.35	0.33
16	Magistral-Small	0.46	0.39	0.37	0.41	0.32
17	Vistral-24B	0.43	0.35	0.39	0.42	0.36
18	Qwen3-8B	0.46	0.37	0.37	0.34	0.40
19	Mistral-Small-24B	0.40	0.38	0.37	0.38	0.36
20	Ministral-8B	0.42	0.34	0.32	0.38	0.34
21	Mistral-Nemo	0.36	0.36	0.42	0.34	0.30
22	Qwen2.5-14B	0.37	0.30	0.38	0.35	0.31
23	Qwen2.5-7B	0.35	0.33	0.30	0.32	0.39
24	Gemma-3-4B	0.36	0.28	0.33	0.36	0.33
25	Qwen3-4B	0.36	0.27	0.25	0.39	0.38
26	GigaChat-20B-A3B	0.30	0.30	0.33	0.35	0.37
27	Qwen3-1.7B	0.30	0.38	0.30	0.31	0.35
28	Mistral-7B-v0.3	0.27	0.32	0.33	0.30	0.31
29	Qwen3-0.6B	0.30	0.32	0.28	0.29	0.33

Table 8: Macro F1 scores for country role-playing across all 26 models (29 configurations including thinking variants), ranked by overall performance. Each model was prompted to role-play as a P5 country representative.

Model	US			RU			CN			UK			FR		
	Y	N	A	Y	N	A	Y	N	A	Y	N	A	Y	N	A
GPT-5	0.90	0.71	0.70	0.90	0.79	0.68	0.87	0.65	0.52	0.91	0.65	0.69	0.92	0.40	0.39
Claude Sonnet 4.5 [T]	0.88	0.64	0.54	0.86	0.73	0.55	0.83	0.45	0.42	0.87	0.58	0.54	0.89	0.38	0.37
DeepSeek V3.2-exp [T]	0.87	0.63	0.52	0.83	0.67	0.48	0.83	0.41	0.36	0.86	0.48	0.46	0.90	0.29	0.35
Kimi K2 [T]	0.85	0.60	0.53	0.82	0.66	0.46	0.79	0.37	0.36	0.84	0.50	0.50	0.88	0.36	0.39
Claude Sonnet 4.5	0.84	0.63	0.54	0.81	0.62	0.46	0.72	0.28	0.38	0.81	0.40	0.43	0.87	0.18	0.18
DeepSeek V3.2-exp	0.81	0.54	0.33	0.72	0.43	0.16	0.76	0.35	0.28	0.82	0.44	0.33	0.88	0.24	0.21
Kimi K2	0.78	0.44	0.33	0.71	0.48	0.24	0.62	0.23	0.27	0.79	0.35	0.35	0.86	0.08	0.11

Table 9: Per-class F1 scores (Y=YES, N=NO, A=ABSTAIN). [T]=thinking mode

Vetoed Draft Resolution Example (1 of 211)	
Document	S/2022/431
Date	2022-05-26
Agenda	Non-proliferation / Democratic People’s Republic of Korea
Vote Tally	In Favour (13): Albania, Brazil, France, Gabon, Ghana, India, Ireland, Kenya, Mexico, Norway, United Arab Emirates, United Kingdom, United States Against (2): China, Russian Federation Abstaining (0): None
P5 Votes	US: YES UK: YES FR: YES CN: NO (veto) RU: NO (veto)
Summary	Draft resolution to strengthen sanctions on North Korea following ballistic missile tests. China and Russia jointly vetoed, marking the first double P5 veto on DPRK sanctions since 2006.

Table 10: Example vetoed draft resolution with dual P5 vetoes.

Adopted Resolution Example (1 of 258)	
Document	S/RES/43(1948)
Date	1948-04-01
Agenda	The Palestine Question
Vote Tally	In Favour (5): China, France, USSR, United Kingdom, United States Against (0): None Abstaining (0): None
P5 Votes	US: YES UK: YES FR: YES CN: YES RU: YES
Summary	The Security Council notes increasing violence in Palestine and believes an immediate truce is urgently needed. Calls upon the Jewish Agency and Arab Higher Committee to arrange a truce; calls upon armed groups to cease violence immediately.

Table 11: Example adopted resolution.

Model Prompt-Response Example	
Document	S/2023/970
Date	2023-12-08
Agenda	Middle East, including the Palestinian question
Country Role	Russian Federation
Ground Truth	YES
Prompt	<p>You are a representative of “Russian Federation” in United Nations Security Council (UNSC). You have to decide to vote on the following context of resolution “S/2023/970” in UNSC. You only choose the vote among three options: YES, NO, or ABSTAIN. Choose the best option considering your country’s interest.</p> <p><i>Context:</i> Expressing grave concern over the catastrophic humanitarian situation in the Gaza Strip... 1. Demands an immediate humanitarian ceasefire; 2. Reiterates its demand that all parties comply with their obligations under international law; 3. Demands the immediate and unconditional release of all hostages... <i>(omitted)</i></p>
Response	<p>Vote: YES</p> <p>Rationale: The Russian Federation supports the draft resolution as it aligns with its commitment to international humanitarian law and the protection of civilians in conflict zones. The demand for an immediate humanitarian ceasefire is crucial in addressing the catastrophic humanitarian situation in the Gaza Strip, which resonates with Russia’s historical stance on humanitarian access.</p>

Table 12: Example prompt-response pair for country role-play task.