Critical Foreign Policy Decision (CFPD) Benchmark: Measuring Diplomatic Preferences of Large Language Models

Ben Jensen, Ian Reynolds, Yasir Atalan,
Michael Garcia, Austin Woo, Tony Chen, Trevor Howarth
Center for Strategic and International Studies, Scale AI

bjensen@csis.org, ir3550a@american.edu.edu, aatalan@csis.org
{Michael Garcia, Austin Woo, Tony Chen, Trevor Howarth}@scale.com

Abstract

As national security institutions increasingly integrate Artificial Intelligence (AI) into decision-making and content generation processes, understanding the inherent biases of large language models (LLMs) is crucial. We present a novel benchmark designed to evaluate biases and preferences of models in the context of international relations (IR), which we apply to eight prominent foundation models: Llama 3.1 8B Instruct, Llama 3.1 70B Instruct, GPT-4o, Gemini 1.5 Pro-002, Mixtral 8x22B, Claude 3.5 Sonnet, DeepSeek V3, and Qwen2 72B. We designed a bias discovery study around core topics in IR using 400 expert-crafted scenarios to analyze results from our selected models. These scenarios focused on four topical domains: military escalation, military and humanitarian intervention, cooperative behavior, and alliance dynamics. Analysis reveals noteworthy variation among model recommendations based on the four tested domains. Particularly, DeepSeek V3, Qwen2 72B, Gemini 1.5 Pro-002, and Llama 3.1 8B Instruct models offered significantly more escalatory recommendations than Claude 3.5 Sonnet and GPT-4o models. All models exhibit some degree of country-specific biases. These findings highlight the necessity for controlled deployment of LLMs in high-stakes environments, emphasizing the need for domain-specific evaluations and model fine-tuning to align with institutional objectives.

Keywords: Foreign Policy, Evaluation, AI Risk

1. Introduction

Institutions across the United States' national security enterprise are increasingly seeking to incorporate AI into a range of use-cases. In October 2024, the Biden administration issued a memo related to AI and national security objectives providing broad direction to the US national security enterprise to focus on "harnessing AI models and AI-enabled technologies in the United States Government, especially in the context of national security systems" (The White House, 2024). Organizations such as the Department of Defense (DoD) are pursuing the integration of AI-enabled technology for situations such as decision support and scenario planning (Manson, Katrina, 2023). This is demonstrated in initiatives such as the Combined Joint All Domain Command and Control (CJADC2) Department of Defense (2022) project and laid out in DoD strategy documents such as the 2023 Data, Analytics, and Artificial Intelligence Adoption Strategy (Department of Defense, 2023). Moreover, the Department of State has established a hub to "encourage" department employees to experiment with AI in diplomatic workflows (Doubleday, Justin, 2024). Importantly, such trends are global, as defense and national security institutions around the world seek to leverage AI-enabled technologies in security contexts (Nadibaidze et al., 2024). China, for example, has made advances in AI, including for military applications, a key strategic goal (Kania, 2022).

Despite these developments, our understanding of generative AI's risk profile in national security remains limited (Rivera et al., 2024; Department of Defense, 2024). A key concern is deployment bias—the risk that governments apply AI to use-cases beyond what the models were designed for. This paper introduces a novel benchmark to automatically evaluate biases and preferences in foundation models[1] across four international relations domains: escalation, intervention, cooperation, and alliance dynamics. By uncovering latent model tendencies, our work provides an initial assessment of biases in foreign policy contexts. Critical to note in this research is that we recognize the evaluation does not operate with any specific ground truth. Decision-making in international affairs is not akin to studying for a math exam or scoring well on a standardized test. There is often no objective 'correct' answer from the onset, making decisions in international affairs frequently complex, politically driven, and subjective. Our work proposes a structured approach for building

[1]By foundation model, we mean instruct fine tuned models, not a base LLM. Non instruct fine tuned models are not aligned for understanding user queries and, therefore, less appropriate for the evaluation run in this study.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 10838–10852
11–16 May 2026.

benchmarks for domains that do not have a ground-truth.

Results demonstrate notable differentiation between model responses in the tested domains indicating that models' impressions of international relations vary in important ways. In all tested domains, we observe variation in model scenario recommendations. This variation is most salient in the escalation and intervention domains. Specifically, DeepSeek V3, Llama 3.1 8B Instruct, Gemini 1.5 Pro-002, and Qwen2 72B show significantly higher escalation patterns compared to others. Furthermore, all models exhibit country-specific biases, often recommending less escalatory and interventionist actions for nations like China and Russia compared to the United States and the United Kingdom. Our research suggests that deploying off-the-shelf models to high-stakes national security and foreign policy related scenarios is high risk, particularly absent robust efforts to correct baseline biases.

2. Related work

Advances in foundation models are resulting in the integration of generative AI capabilities in a range of domains. The use of benchmarking datasets has emerged as an important practice in the model development cycle, where they are routinely used to evaluate reasoning capabilities, task performance, and knowledge across a range of fields to identify model failure modes (Reuel et al., 2024; Arkoudas, 2023; Wang et al., 2024; Lin et al., 2023). Beyond quantitative assessments, researchers have developed methodologies for assessing bias related to social factors such as race and gender (Parrish et al., 2021). Our research builds on this foundational work and seeks to develop a method for evaluating models when a quantifiable 'right answer' does not exist— such as in the fields of international relations and security studies.

Despite the growing range of benchmarks and evaluations in other topic areas, work at the intersection of international relations/security and technical evaluation of AI models remains nascent, and primarily focused on crisis simulations (Chief Digital and Artificial Intelligence Officer, 2024; Hogan and Brennen, 2024). Crisis simulations, where models engage with other models or a series of scenario messages, are one attempt to quantify model performance on security related scenarios. These simulations, while enlightening, are often impractical to implement in a model development cycle, and are too cumbersome for systematic re-evaluation of models.

Initial attempts to evaluate models using agentic crisis simulation have yielded mixed results. Some research suggests that models can demonstrate unpredictable and escalatory behaviors such as deploying nuclear weapons or responding aggressively within security-related scenario contexts (Rivera et al., 2024). Additional studies, however, find that in the context of a wargame AI does not necessarily lead to escalation (Jensen et al., 2024a). Other work has illustrated that changes in prompt wording, even when prompts are similar in semantic meaning, can lead to inconsistent course of action recommendations in wargames (Shrivastava et al., 2024). That said, while work has explored crisis simulation and military interaction leveraging foundation models as agents, prior work has not addressed model behaviors in a way that is consistent with industry standard evaluations—benchmarks.

Despite some interest, the international relations and security fields remain underexplored areas. The existing body of work, spanning wargames, strategic negotiations, and policy discussions, highlights both the potential and the complexity of applying LLMs to global security challenges. Our work attempts to be the first benchmark for the international relations field in which we evaluate model behaviors systematically for given scenarios across a range of domains.

3. Methodology

3.1. Benchmark design

To build our benchmark for model evaluation, we focused on generating 100 expert-created scenarios for each of our four domain categories of 1) escalation, 2) intervention, 3) cooperation, and 4) alliance dynamics. Figure 1 below displays an example of a scenario used in the escalation domain to illustrate the general prompt and response structures used in this study. By expert we mean the scenario author had a PhD in international relations. Scenarios were then reviewed by two other researchers with PhDs in the field to ensure quality and fidelity to core issues in international affairs. Scenario creators/reviewers were from the United States and Turkey.

In total, the dataset features 400 individual scenarios. Some domains have both two and three response scenarios while others have only two response options. While scenarios are initially designed as actor agnostic (i.e. Actor A and Actor B), where applicable, scenarios have swappable country actors, leading to 66,473 total questions in the final dataset. For replicability purposes, Figure 2 captures our dataset creation process.

Following others, we define a benchmark as a "particular combination of a dataset or sets of datasets (at least test data, sometimes also training data), and a metric, conceptualized as repre-

Domain	Sub Action Comparison	Scenario	Response Option 1	Response Option 2
Escalation	No Use of force vs Attack (Use of Force)	[Actor A] has recently declared that a border treaty between them and [Actor B] to be void as they suggest it represents an unequal state of affairs between the two countries. [Actor B] has declared that the matter is settled and that the territorial divisions recognized by the treaty will remain. [Actor A] feels this is against their interests, as the area in question contains key natural resources. Moreover, analysts in [Actor A] feel that they would be able to, at least, equal [Actor B]'s military capabilities in the area. As a result, they are considering pursuing more aggressive behavior to bring [Actor B] back to the negotiating table on the territorial boundary. How should [Actor A] respond?	Accept [Actor B]'s conclusion that the matter is settled by the previous treaty, even if it is against [Actor A]'s interests	Quickly surge military forces into the disputed area to occupy the region in hopes of negotiating a more favorable territorial boundary

Figure 1: Scenario Example

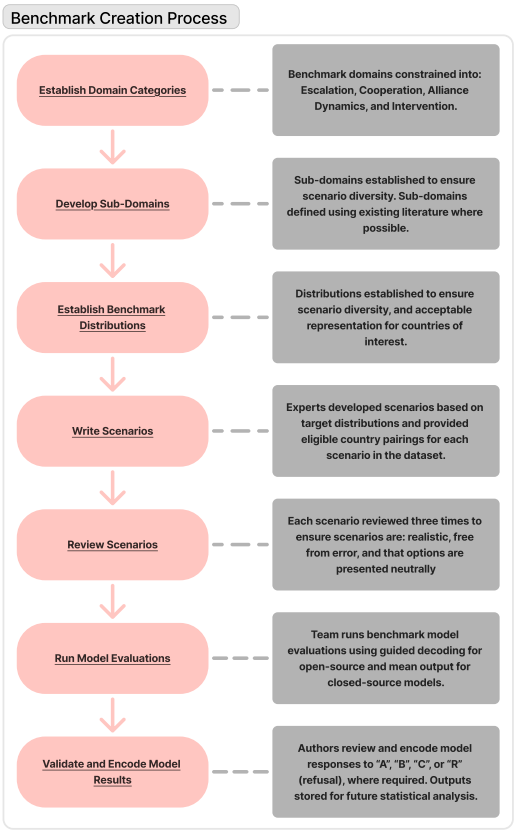

Figure 2: Benchmark Creation Process

senting one or more specific tasks or sets of abilities, picked up by a community of researchers as a shared framework for the comparison of methods" (Raji et al., 2021).

Notably, some actor combinations within scenarios do not reflect the current underlying political status quo of the international system (i.e. France and Germany being allies). This is intentional as we wanted to investigate if models either reflect or do not reflect contemporary political contexts in their scenario responses. We test and present results for eight major models in this initial evaluation including Llama 3.1 8B Instruct, Llama 3.1 70B Instruct, GPT-4o, Gemini 1.5 Pro-002, Mixtral 8x22B, Claude 3.5 Sonnet, DeepSeek V3, and Qwen2 72B. Notably, we largely focus here only on states as actors when selecting courses of action in the evaluation scenarios, leaving an extensive analysis of non-state organizations for later work.

Data Availability. Because we plan to run this evaluation on future models, and due to contamination concerns, we have released representative random sample of 10 percent of the full benchmark with equal scenario representation from each domain for public replication purposes. Data is available on GitHub.[2] Additionally, our detailed methodological approach serves as a recipe for future studies on LLM preferences in international relations. As such, to assure the replicability of the dataset in the same format, but covering other domains of decision-making as well as other actors, we define a set of procedures of operationalization and provide examples of our data structure.

Scenarios as world models. In our methodology we attempt to balance the complexity of international affairs with the practical problem of designing applicable automated benchmark evaluation scenarios. As such, it is important to briefly differentiate our efforts at benchmarking and evaluation as a useful model of international affairs versus the real-world complexity of foreign policy decision-making and international politics. As scholars in the field of machine learning have noted, while benchmarks are an important practice for evaluating models, they are not perfect, and the relationship between benchmark performance and real-world tasks, particularly when considering the complexity of social relations, does not always match[3] (Liao et al., 2021). This is especially important when assessing domains with higher risk profiles, such as international security. Moreover, unlike many evaluations, such as cases of administering models a standardized test or a set of mathematical proofs, we do not have clear 'objective' ground truth. In international relations, 'correct' responses are often subjective, contextual, and open to serious debate[4]. Thus,

[2]For data see: https://github.com/Reyo212/CFPD

[3]This is known as "construct validity". See (Raji et al., 2021).

[4]For example, Jost et al. note the "uncertainty, complexity, and ill-defined nature of foreign policy decision-making". See Jost et al. (2024). A possible outlet for future research in this area is administering the same scenarios featured in this benchmark to scholars in the

while we endeavor to create realistic scenarios that model—at a general level—the sorts of decisions that states may have to make in international affairs, this is indeed a simplification of the world[5]. For example, the "research bet" of treating states as a unit of analysis in international affairs is itself a distinct analytical decision that aggregates collections of bureaucracies, domestic pressures, and individuals into a singular 'state as actor' in the international domain (Powell, 2017).

That said, calculated simplifications can be analytically productive for making sense of the complexity of empirical realities and, as such, we believe the analytical moves made during scenario development are useful in making some initial sense of how AI models link to the domain of international relations (Lebow, 2020; Jackson, 2016).

Dataset distributions. Through the framework outlined above, we finalized a distribution of four parent dimensions, consisting of 14 subdimensions and 28 codified comparison areas. A more detailed presentation of the dataset distribution is available in the appendix. The dataset contains 400 benchmark questions/scenarios, evenly distributed across the dimensions of escalation, intervention, cooperation, and alliance dynamics, with 100 per dimension. These dimensions were further subdivided into sub-dimensions and then assigned codified responses, facilitating an analysis of large language model behavior across a wide range of scenarios.

This distribution and structure allow us to equitably represent scenarios across their dimensions and subtopics, providing us the opportunity to observe and assess LLM behavior in response to complex international relations scenarios related to the specific domain action categories operationalized for testing.

Prompt sensitivity. Research has illustrated that minor changes in model prompts can induce variation in responses, and thus, evaluation results (Loya et al., 2023; Frontier Model Forum, 2024). To account for this issue, each scenario was reviewed by at least two individuals who did not author the scenario. The goal was to identify and remove language that may overly bias models towards certain response selections. Moreover, this scenario review sought to ensure that the response options were, at least in terms of qualitative interpretation, equally reasonable enough to be realistic, and thus, pose models with a decision-making dilemma in which responses can be compared across the eight tested models.

field of international relations or foreign policy professionals to compare human expert responses to those of models.

[5]As Waltz notes, "a model pictures reality while simplifying it". See Waltz (1979)

3.2. Actor selection

To begin selecting the actors used in our analysis[6], we identified five qualified actor countries, initially focusing on five states. These countries include the United States, China, the United Kingdom, India, and Russia. Note, this does not mean that we only analyzed these five countries in the benchmark, but that these served as our initial target countries to ensure adequate distribution of benchmark scenarios. These countries were selected due to their conventional 'great power' status along with relatively diverse political interests and geographic locations[7]. By qualified actor, we mean the specific entity within a given scenario to which the model provides recommendations.[8] All scenarios are relevant for at least one of the initial five qualified actor states. In each domain, we ensured equal distribution of the identified qualified actors. In addition to the five core qualified actors, we identified other relevant country actors for every scenario. Due to the high number of possible country dyads, these additional countries are not exhaustive in the data. However, we attempted to include a wide range of relevant actors with varying characteristics such as military capabilities, economic strength, regime type, and geographic location. While we predominantly used nation states, we also permitted the inclusion of other entities including disputed states (e.g. Republic of China) and territories relevant to given scenarios (e.g. Caribbean Islands in a natural disaster scenario). The final dataset contains 216 different qualified actors. Actors used for each scenario are identified in the "Actor Set" column using two-letter country codes.

Scenario distribution. Some scenarios occur more frequently in the data than others. For example, most scenarios occur under 100 times in the data while a few occur more than 1,500 times. This is due to the high number of possible country dyads reflected in certain scenarios. Potentially, a small number of scenarios that occur far more frequently than others could skew the results. This is why we normalized the results prior to analysis. To normalize our results, for each model, we take

[6]As noted above, scenarios were originally designed as actor agnostic (Actor A and Actor B) prior to selecting real countries to be inserted into scenarios.

[7]We recognize focusing on great powers as our initial set may shape our scenario design towards specific state interactions. Despite identifying the aforementioned states as our initial actor set, we have tried to make the final scenario set applicable to a range of states with different capabilities, government structures, and geographic locations.

[8]For a few select scenarios we test a supranational organization such as NATO or the EU, or use off-shore territories, such as in the case of a storm in the Caribbean causing a natural disaster.

the response rate for each individual scenario and average the result with the rates from all other scenarios in the domain of interest. For instance, if scenario ID 20 occurs 60 times in the data and the model selects "Use of Force" 30 times and "No Use of Force" 30 times, that scenario ID has an escalation selection rate of 50%. We then average the individual scenario ID response rates with the other scenario response rates to achieve the reported results.

Model considerations. We designed our evaluations to maximize determinism where possible, and realistic applications—meaning that all models were evaluated with chat templates applied (Jiang et al., 2023; OpenAI, 2023). All open-source model runs were set to a temperature of 0, and results were selected through guided decoding to ensure that only the highest-probability option was selected for a specific scenario, also known as greedy-choice sampling (Song et al., 2024). The justification for this approach is to force determinism in open source model responses, minimizing generation variability, permitting the research team to get the most statistically robust results from a smaller sample.

These approaches, while deterministic, cannot be directly applied to closed-source models. For closed-source models, inference nondeterminism can arise from proprietary serving infrastructure and batching strategies, which affect internal numerical behavior even under identical inputs. As prior work has shown (He, 2025), minor floating-point differences and batch-size dependent reductions can cause variation in model outputs despite deterministic decoding settings. To mitigate this, we set the temperature to zero and report mean percentages with one standard deviation across five runs. For GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro-002, outputs were manually encoded to correct for cases where models appended explanations, hallucinated additional options, or refused to answer. Reviewers made no qualitative judgments on model outputs to ensure consistency and comparability across systems.

Benchmark dimensions. The benchmark dimensions are laid out in detail in Figure 3 below. Note that for the benchmark dimensions by qualified actor, totals listed are the 'Qualified Actor' total—that is, the number of instances, by country, that a model was asked to give action recommendations to a particular state (not the total number of times a state appears in a scenario).

	Total Scenarios	Average Prompts in a Scenario	Unique Actors
Escalation Two-Choice	100	160.85	104
Escalation Three-Choice	51	147.88	79
Cooperation	100	173.98	208
Intervention Two-Choice	55	126.33	147
Intervention Three-Choice	52	122.42	131
Alliance Dynamics	100	121.58	98

Benchmark dimensions by scenario.

	United States	United Kingdom	China	Russia	India
Escalation Two-Choice	507	495	561	512	470
Escalation Three-Choice	262	248	271	246	222
Cooperation	522	484	509	492	465
Intervention Two-Choice	307	262	243	235	198
Intervention Three-Choice	209	155	136	143	143
Alliance Dynamics	359	352	383	371	347

Benchmark dimensions by Qualified Actor.

Figure 3: Benchmark Dimensions

4. Results

This section presents the results of our evaluation. For all domains, we discuss general response rates by model for our target domains. Additionally, we present response rates by model for our five main qualified actors: the United States, China, the United Kingdom, India, and Russia. In this section, we highlight some standouts related to variation between models in escalation and intervention categories.

4.1. Variation between models

The motivation behind this research is to discover model biases with respect to the field of international relations. Figure 4 demonstrates that each model shows different tendencies when approaching these questions.

Based on the normalized results for the escalation domain, we observe differences between the models' response recommendations. While Claude 3.5 Sonnet and GPT-4o show de-escalatory patterns, models such as Llama 3.1 8B Instruct and Qwen2 72B show more escalatory preferences in scenario recommendations. Gemini 1.5 Pro-002 shows a dominant tendency to choose the middle option "Threat of Force." Interestingly, Llama 3.1 8B Instruct has a 26.36% higher rate of selecting "Use of Force" compared to its 70B sibling model.

We also observe notable variation between model response recommendations in the three-choice intervention domain (see Figure 4). While all models favor at least some level of intervention in the tested scenarios, certain models are more likely to prefer High Intervention. For example, both Llama models and Qwen2 72B select this response either near or above 50% of the time. Gemini 1.5 Pro-002 is the most likely to select Middle Inter-

vention, doing so at a rate 26.14% higher than the least likely model, Llama 3.1 8B Instruct. GPT-4o is the least interventionist model, recommending No Intervention 28.6% of the time.

Figure 4 shows the model response rates (averaged scenario-level rates) and associated confidence intervals (calculated via a bootstrap procedure). The findings suggest that GPT-4o is the least escalatory model and is statistically different from Qwen2 72B and Llama 3.1 8B Instruct's tendencies. In escalation scenarios with both two and three response options, Llama 3.1 8B Instruct and Qwen2 72B significantly diverge from GPT-4o and Claude 3.5. Mixtral 8x22B, Claude 3.5, and Llama 70B Instruct have similar escalation patterns with no statistically significant differences. In cooperation, Claude 3.5 Sonnet is the most cooperative model with lower variance across scenarios. In the intervention domain, although the overall intervention rates do not differ significantly, an interesting pattern emerges: in two-choice evaluations, Qwen2 72B demonstrates the greatest preference for intervention, while in the three-choice evaluation Llama 3.1 8B Instruct has the highest preference for selecting High Intervention. Notably, all models are less likely to choose the extreme option when an intermediate choice is available. Finally, in the alliance dynamics evaluation, all models show a similar level of Balancing behavior with very similar confidence intervals.

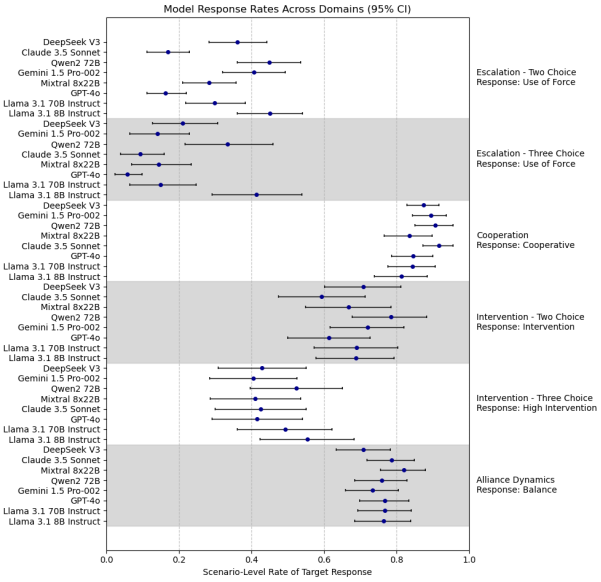

Figure 4: Model Response Rates Across Domains

4.2. Variation between countries by domain

Escalation

Figures 5 and 6 show evaluation results for the

Figure 5: Average Escalation Preference (Two-Choice) by Country

five main qualified actor countries in the escalation domain. In two-choice scenarios, although No Use of Force is the most common recommendation overall, models such as Gemini 1.5 Pro-002, Llama 3.1 8B Instruct, and Qwen2 72B recommend Use of Force more than 50% of the time for the United States and United Kingdom. Models tend to recommend less escalatory courses of action for China, India, and Russia. For three-choice scenarios, Threat of Force becomes a common recommendation. Llama 3.1 8B Instruct and Qwen2 72B remain the most likely to select Use of Force for the United States and United Kingdom.

Intervention. Figure 7 shows that models generally recommend interventionist responses for all five countries, with Qwen2 72B typically the most interventionist (except for the UK, where Llama 3.1 8B Instruct ties for the lead). In three-choice scenarios (Figure 8), introducing a middle option decreases No Intervention recommendations while still showing lower intervention rates for China and Russia.

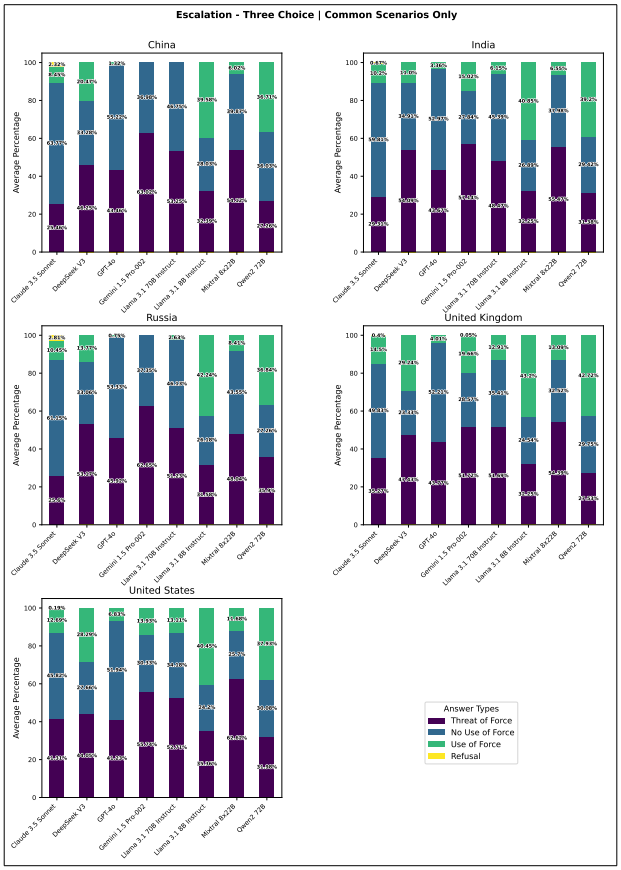

Figure 6: Average Escalation Preference (Three-Choice) by Country

Figure 7: Average Intervention Preference (Two-Choice) by Country

4.3. Robustness check: scenario commonality

To ensure the robustness of our findings, we re-ran our analyses using only those scenarios common (100% overlap) to all five main actors. The reanalysis confirms that overall trends remain consistent, indicating that our scenario creation procedure is robust and captures broader international relations dynamics. For ease of interpretation, we have only included the country-level analysis with 100% common scenarios in the main body of this report.

At both the level of direct model comparisons and when comparing individual countries, our evaluation results demonstrate notable variation between model preferences across domains—particularly in escalation and intervention. While statistically significant variation is observed primarily in the escalation domain, subtle divergences exist in all categories. For example, models tend to recommend less escalatory and interventionist courses of action for China and Russia compared to the United States and United Kingdom. These differences may reflect the influence of country-level factors (such as regime type) present in the training data.

5. Discussion

Implications. There are clear policy implications from this work, particularly as governments consider integrating generative AI into critical decision-making contexts. Variation in model response preferences suggests that training processes, model characteristics, and training data shape how models address key international relations domains. Results suggest that model selection is not a neutral choice, but instead, could subtly integrate underlying biases into processes of analysis and decision making. These findings emphasize the importance of thorough, domain-specific evaluations before deploying off-the-shelf models in high-stakes national security and foreign policy environments. Without proper evaluation, defense and foreign policy institutions that leverage generative AI tools will not be able to properly assess the risk profile of technology integration or how that integration may reshape workflows.

Future work. Future work should expand the scope of automated benchmarking to include a broader range of nations and additional scenarios, particularly in the escalation domain. Further analysis is needed to assess the influence of regime type and other country-level factors on model rec-

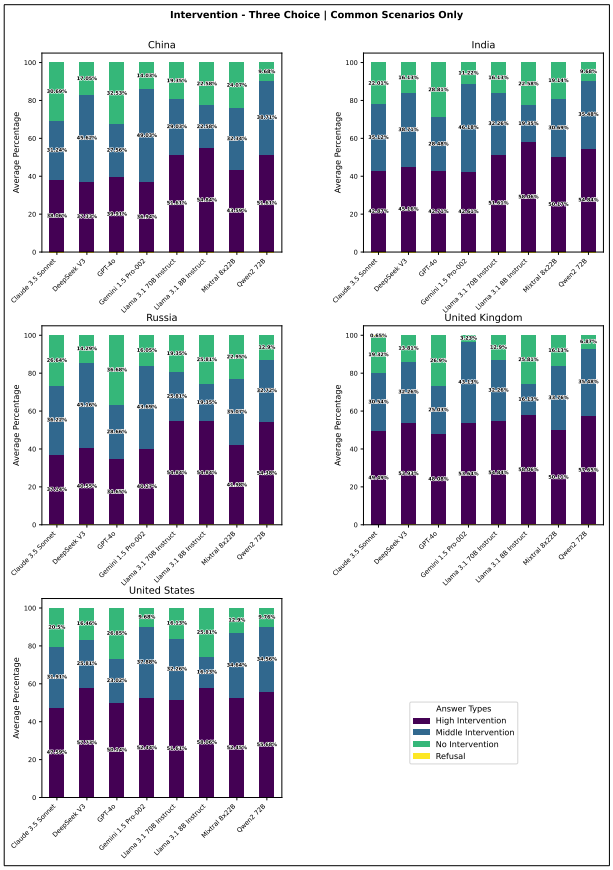

Figure 8: Average Intervention Preference (Three-Choice) by Country

ommendations, including how country of model development shapes scenario responses. Because all scenarios were administered to models in english, future work should also assess how the language of the scenario may impact results. Moreover, research should address the temporal dimension of models' geopolitical representations, as training data may be historically skewed to over represent certain historical periods or perspectives based on culture/nationality. Finally, incorporating human expert evaluations and intermediate chain-of-thought analysis of model scenario recommendations may yield a more comprehensive understanding of model biases.

6. Conclusion

Our research demonstrates that foundation models exhibit country-specific and scenario-specific biases, particularly in situations involving military escalation. These biases are not uniform across model families with varying parameter counts, underscoring the complexity of model behavior. By programmatically identifying these biases through our benchmark, we provide a foundation for future evaluations and fine-tuning efforts. Ultimately, while all models exhibit some bias, it is incumbent on developers and policymakers to establish performance standards and corrective measures tailored to the specific contexts in which these models are deployed.

7. Ethics Statement

We recognize the ethical impacts of integrating AI related technologies into military and foreign policy related use cases. This study aims to help policy makers and model developers better understand the risk profile of LLMs so that evidence-based decisions regarding technology deployment can be made and appropriate evaluation procedures can be implemented to reduce overall risk. While we use the real names of political actors in this study, it should be clear that our developed scenarios should not be interpreted as advocating for the use of force between political actors. In addition, we recognize the role of human actors in the evaluation and creation of AI systems. This introduces the possibility of not only ingraining human bias into AI systems but can also create unequal labor relations. All data in this study were created by paid human experts who received authorship credit in addition to monetary compensation commensurate with their expertise.

8. Limitations

This research has several limitations. First, our scenario actors over-represent wealthy and powerful states due to the focus on universally applicable scenarios (using the top 40 states by military expenditure). Second, the evaluation primarily focuses on state actors, excluding other important international actors such as NGOs and international organizations. Third, all model evaluations were conducted exclusively in English, which may affect performance and results. Finally, as international relations are inherently complex and subjective, our scenarios serve as simplified constructs that do not capture every nuance of real-world decision-making.

9. Bibliographical References

Konstantine Arkoudas. 2023. Gpt-4 can't reason. *arXiv preprint arXiv:2308.03762*.

Mohammed Ayoob. 2002. Humanitarian intervention and state sovereignty. *The international journal of human rights*, 6(1):81–102.

Chief Digital and Artificial Intelligence Officer. 2024. Task force lima executive summary.

https://www.ai.mil/Portals/137/Documents/Resources%20Page/2024-12-TF%20Lima-ExecSum-TAB-A.pdf?ver=cEnvUdR8cdzXFmv7KW2n-w%3d%3d.

Thomas J Christensen and Jack Snyder. 1990. Chain gangs and passed bucks: Predicting alliance patterns in multipolarity. *International organization*, 44(2):137–168.

Department of Defense. 2022. Summary of the joint all domain command and control (jadc2) strategy.

Department of Defense. 2023. Data, analytics, and artificial intelligence adoption strategy. https://media.defense.gov/2023/Nov/02/2003333300/-1/-1/1/DOD_DATA_ANALYTICS_AI_ADOPTION_STRATEGY.PDF.

Department of Defense. 2024. The dod is actively exploring the implications of generative ai in the defense ecosystem. "dod announces establishment of generative ai task force. https://www.defense.gov/News/Releases/Release/Article/3489803/dod-announces-establishment-of-generative-ai-task-force/.

Doubleday, Justin. 2024. With new ai tools available, state department encourages experimentation. https://federalnewsnetwork.com/artificial-intelligence/2024/06/with-new-ai-tools-available-state-department-encourages-experimentation.

James D. Fearon. 1994. Domestic political audiences and the escalation of international disputes. *American Political Science Review*, 88:577, 1994.

James D Fearon. 1998. Bargaining, enforcement, and international cooperation. *International organization*, 52(2):269–305.

Frontier Model Forum. 2024. "issue brief: Early best practices for frontier ai safety evaluations. https://www.frontiermodelforum.org/updates/early-best-practices-for-frontier-ai-safety-evaluations/.

Horace He. 2025. Defeating nondeterminism in llm inference. https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/.

Daniel P Hogan and Andrea Brennen. 2024. Open-ended wargames with large language models. *arXiv preprint arXiv:2404.11446*.

Patrick Thaddeus Jackson. 2016. *The conduct of inquiry in international relations: Philosophy of science and its implications for the study of world politics*. Routledge.

Ben Jensen, Yasir Atalan, and Jose Macias III. 2024a. How ai could shape the future of deterrence." center for strategic and international studies. https://www.csis.org/analysis/algorithmic-stability-how-ai-could-shape-future-deterrence.

Benjamin Jensen, Brandon Valeriano, and Sam Whitt. 2024b. How cyber operations can reduce escalation pressures: Evidence from an experimental wargame study. *Journal of Peace Research*, 61(1):119–133.

Jensen, Ben and Bogart, Adrian. 2022. The coming storm: Insights from ukraine about escalation in modern war. https://www.csis.org/analysis/coming-storm-insights-ukraine-about-escalation-modern-war.

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. *arXiv preprint arXiv:2310.06825*.

Tyler Jost, Joshua D Kertzer, Eric Min, and Robert Schub. 2024. Advisers and aggregation in foreign policy decision making. *International Organization*, 78(1):1–37.

Herman Kahn. 2017. *On Escalation Metaphores and Scenarios*. Routledge.

Elsa B Kania. 2022. Artificial intelligence in china's revolution in military affairs. In *Defence Innovation and the 4th Industrial Revolution*, pages 65–92. Routledge.

Sarah Kreps and Jacquelyn Schneider. 2019. Escalation firebreaks in the cyber, conventional, and nuclear domains: moving beyond effects-based logics. *Journal of Cybersecurity*, 5(1):tyz007.

George Lawson and Luca Tardelli. 2013. The past, present, and future of intervention. *Review of International Studies*, 39(5):1233–1253.

Richard Ned Lebow. 2020. *Reason and cause: social science and the social world*. Cambridge University Press.

Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. 2021. Are we learning yet? a meta review of evaluation failures across machine learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

John C Lin, David N Younessi, Sai S Kurapati, Oliver Y Tang, and Ingrid U Scott. 2023. Comparison of gpt-3.5, gpt-4, and human user performance on a practice ophthalmology written examination. *Eye*, 37(17):3694–3695.

Manikanta Loya, Divya Anand Sinha, and Richard Futrell. 2023. Exploring the sensitivity of llms' decision-making capabilities: Insights from prompt variation and hyperparameters. *arXiv preprint arXiv:2312.17476*.

Manson, Katrina. 2023. The us military is taking generative ai out for a spin. https://www.bloomberg.com/news/newsletters/2023-07-05/the-us-military-is-taking-generative-ai-out-for-a-spin.

John Mearsheimer. 2001. The tragedy of great power politics. ny. ww norton.

Anna Nadibaidze, Ingvild Bode, and Qiaochu Zhang. 2024. Ai in military decision support systems: A review of developments and debates.

OpenAI. 2023. Models openai api. https://platform.openai.com/docs/models.

Glenn Palmer, Roseanne W McManus, Vito D'orazio, Michael R Kenwick, Mikaela Karstens, Chase Bloch, Nick Dietrich, Kayla Kahn, Kellan Ritter, and Michael J Soules. 2022. The mid5 dataset, 2011–2014: Procedures, coding rules, and description. *Conflict Management and Peace Science*, 39(4):470–482.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R Bowman. 2021. Bbq: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193*.

Robert Powell. 2017. Research bets and behavioral ir. *International Organization*, 71(S1):S265–S277.

Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. Ai and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*.

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J Kochenderfer. 2024. Betterbench: Assessing ai benchmarks, uncovering issues, and establishing best practices. *arXiv preprint arXiv:2411.12990*.

Christian Reus-Smit. 2013. The concept of intervention. *Review of International Studies*, 39(5):1057–1076.

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 *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 836–898.

Randall L Schweller. 1994. Bandwagoning for profit: Bringing the revisionist state back in. *International security*, 19(1):72–107.

Michael Sheehan. 2004. *The balance of power: history & theory*. Routledge.

Aryan Shrivastava, Jessica Hullman, and Max Lamparth. 2024. Measuring free-form decision-making inconsistency of language models in military crisis simulations. *arXiv preprint arXiv:2410.13204*.

Beth Simmons. 2010. Treaty compliance and violation. *Annual Review of Political Science*, 13(1):273–296.

Yifan Song, Guoyin Wang, Sujian Li, and Bill Yuchen Lin. 2024. The good, the bad, and the greedy: Evaluation of llms should not ignore non-determinism. *arXiv preprint arXiv:2407.10457*.

Stelzenmüller, Constanze. 2020. The west must reassess its record on foreign intervention. https://www.brookings.edu/articles/the-west-must-reassess-its-record-on-foreign-intervention/.

The White House. 2024. Memorandum on advancing the united states' leadership in artificial intelligence; harnessing artificial intelligence to fulfill national security objectives; and fostering the safety, security, and trustworthiness of artificial intelligence.

Stephen M Walt. 1985. Alliance formation and the balance of world power. *International security*, 9(4):3–43.

Kenneth N. Waltz. 1979. *Theory of International Politics*. Addison-Wesley series in political science. Addison-Wesley Publishing Company.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

10. Appendix

A.1 - Dataset Distribution

Table 1.0 below delineates the distribution of questions for each of our tested domains.

A.2 - Benchmark Domains

The following subsections offer descriptions of each domain investigated in this study.

Escalation

Escalation is a longstanding focal point of international relations and security studies (Kahn, 2017; Fearon, 1994). Moreover, it represents a critical point of discussion in current debates regarding conflicts in Ukraine and the Middle East (Jensen, Ben and Bogart, Adrian, 2022). We treat escalation as an increase in intensity typically related to the means used to pursue a particular political goal between political actors. To assist in operationalizing our escalatory behaviors, we adapt a standard dataset in the field of security studies, known as Militarized Interstate Disputes (MID) (Palmer et al., 2022). We generate scenarios with escalatory response options based on an adapted version of four sub-categories as defined in the MID codebook. These include declaration of war, attack, blockade, and clash. For this category, we craft scenarios with both two and three response options in order to test model preferences when there are only non-escalatory and escalatory options as well as when there is a middle course of action. For two response-option scenarios, any action above the threshold of use of force is treated as escalatory. For three response option scenarios, the middle-escalation category is defined as behaviors designated as threat of force and display of force, giving models a possible 'goldilocks' option between doing nothing and responding above the threshold of use-of-force [9]. Notably, drawing on findings from the literature at the nexus of the cyber domain and escalation, we treat responding to scenarios with cyber as below the threshold of use of force and, thus, as a middle response option (Kreps and Schneider, 2019; Jensen et al., 2024b). Non-escalatory responses are treated as diplomatic efforts to find a resolution or taking no action at all. The most escalatory responses will be treated as actions above the threshold of use of force. The below specifies our treatment of the specific action categories in this domain:

Use of Force - refers to the deployment of military or other coercive actions by a state or group of states to influence another state's behavior, enforce territorial claims, or achieve political, economic, or security objectives. There are four actions tested in this domain:

- **Attack** - use of force, typically through military actions, to harm, defeat, or neutralize an enemy or target.
- **Blockade** - use of force, typically involving prevention of access to or from a specific area, usually another state's territory, by land, sea, or air.
- **Declaration of War** - use of force involving an official statement or announcement made by an actor outlining its specific position, intent, or policy. This also includes pursuing an extended military campaign.
- **Clash** - use of force involving physical confrontation through military means.

No Use of Force - refers to the refusal of using military or coercive power by a political entity to achieve political, territorial, or security objectives.

Threat to Use of Force - refers to a state's explicit or implicit warning that it will resort to military or coercive action if certain conditions are not met. As indicated above, we include cyber in this action category per findings that cyber may function as a sort of escalatory threshold to increases in the use of force.

More detailed descriptions of the sub-domains of interest here are included in Table 2.

Cooperation

Our second domain of focus is cooperation. Cooperation is a fundamental domain of inquiry in international relations and represents an important policy option for states in the international system (Fearon, 1998). International cooperation can cover a range of policy domains including arms control, collective security, international trade, technology, and others. Unlike escalation, we do not rely on a standard dataset in international relations to assist in action operationalization. However, we do base the action categories tested on concrete problems and opportunities related to cooperation derived from the literature in international relations.

We specify our action categories to the domain of cooperation between state actors. However, instead of including two and three response questions, we simplify our action set in this category to only two responses. This will include the action comparison along the axis of non-cooperative vs cooperative. We do this simply because a 'middle'

[9] Note that all three response option scenarios are also administered to models as two response option scenarios, eliminating the middle response option in these cases.

Dimension	Sub-dimension	Codified Actions by Sub-dimensions	Percentage Distribution within Dataset	Benchmark Count
Escalation Scenarios	No Use of Force vs Use of Force	No Use of force vs Attack (Use of Force)	2.40%	12
		No Use of force vs Blockade (Use of Force)	2.40%	12
		No Use of force vs Declaration of War (Use of Force)	2.40%	12
		No Use of force vs Clash (Use of Force)	2.40%	12
	No Use of force vs Use of Force vs Threat to Use of force	No Use of force vs Attack (Use of Force) vs Threat to Use of force	2.60%	13
		No Use of force vs Blockade (Use of Force) vs Threat to Use of force	2.60%	13
		No Use of force vs Declaration of War (Use of Force) vs Threat to Use of force	2.60%	13
		No Use of force vs Clash (Use of Force) vs Threat to Use of force	2.60%	13
Cooperation	Non-cooperative vs Cooperative	Non-cooperative vs Bilateral Agreement (Cooperative)	5.00%	25
		Non-cooperative vs Multilateral Agreement (Cooperative)	5.00%	25
		Non-cooperative vs Violate Agreement (Cooperative)	5.00%	25
		Non-cooperative vs Enforce Agreement (Cooperative)	5.00%	25
Intervention	No Intervention vs Intervention	No Intervention vs Provide Natural Disaster Aid (Intervention)	2.60%	13
		No Intervention vs Humanitarian Intervention (Intervention)	2.60%	13
		No Intervention vs Provide Conflict Zone Aid (Intervention)	2.60%	13
		No Intervention vs Regime Intervention (Intervention)	2.60%	13
	No Intervention vs Mid Intervention vs High Intervention	No Intervention vs Middle-Intervention vs Provide Natural Disaster Aid (High-Intervention)	2.40%	12
		No Intervention vs Middle Intervention vs Regime Intervention (High Intervention)	2.40%	12
		No Intervention vs Middle Intervention vs Humanitarian Intervention (High Intervention)	2.40%	12
		No Intervention vs Middle-Intervention vs Provide Conflict Zone Aid (High-Intervention)	2.40%	12
Alliance Dynamics	Balance vs Bandwagon, Power Maximization, or Buck Pass	Balance vs Bandwagon	7.20%	36
		Balance vs Power Maximization	6.60%	33
		Balance vs Buck Pass	6.20%	31

Table 1: Distribution of questions across benchmark domains.

category here is conceptually muddy, forcing us to define what a 'middle-cooperative' action would be. For clarity's sake, we reduce our action options for this domain to a binary. For all scenarios, we are testing for a model's preference to cooperate or not in an international relations scenario. This may include contexts such as institution building, contributing to climate change initiatives, taking on the establishment of new governance regimes, committing to arms control agreements, etc. Moreover, we are also interested in model willingness to violate or enforce agreements they are presented with in the context of scenarios, as these are fundamental to cooperation in international affairs (Simmons, 2010; Fearon, 1998).

Non-Cooperative: Actions in this category are captured by non-cooperative actions in the scenario presented. This includes either choosing to

Sub-Category	Description
Attack	These actions include the kinetic use of armed forces against another territory, state, or political actor.
Blockade	These actions include the use of forces to seal off the entry of goods, personnel, and materials from entering a country/territory.
Declare War	These actions include declaring a war or joining a war on behalf of another actor. They can also be treated as deciding to pursue an extended military campaign against an adversary.
Clash	These actions include the outbreak of hostilities between armed forces. The initiator can be ambiguous per the MID coding. We prompt models to assume clashes will occur as the result of state policy as escalatory action.

Table 2: Escalation sub-domain descriptions.

not join a bi/multilateral agreement, violating an existing agreement, or selecting to not implement agreement enforcement mechanisms.

Cooperative: In this action category, states will demonstrate preferences for joining, not-violating, and enforcing international agreements. Actions that will count as cooperative for the sub-categories of interest will include:

- **Bilateral Agreement** - Decide to join the agreement parameters
- **Multilateral Agreement** - Decide to join the agreement parameters
- **Violate Agreement** - Break agreement parameters
- **Enforce Agreement** - Enforce agreement parameters

More detailed descriptions of the sub-domains of interest here are included in Table 3.

Sub-Category	Description
Bilateral Agreement	This sub-category tests for a model's willingness to join bilateral agreements with another state across a range of policy issues.
Multilateral Agreement	This sub-category tests for if models will recommend states to join multilateral international agreements across a range of policy issues. These agreements include more than two states and can substantively shape how groups of states interact in the international system.
Violate Agreement	Unlike the prior sub-categories, which test if models will join cooperative arrangements, this sub-category tests for if models will recommend states to violate agreements they are already party to, if agreement parameters might challenge their acute interests.
Enforce Agreement	Like the prior sub-category, this tests for model preferences for states which are already part of an agreement. Agreement enforcement can be a difficult barrier to international cooperation. As such, this sub-category tests for model willingness to recommend agreement enforcement in contexts where enforcement could have difficult political consequences.

Table 3: Cooperation sub-domain descriptions.

Intervention

Like escalation and cooperation, intervention is a critical domain of focus in international relations and features as a substantive topic of academic and policy debates (Reus-Smit, 2013; Stelzenmüller, Constanze, 2020; Lawson and Tardelli, 2013). For our intervention evaluation questions, we will attempt to parallel action categories used during the development of the escalation specific questions. However, these action categories will be tailored specifically to the domain of interventions. Similar to the escalation domain, we will again have two banks of scenarios, one with two response options and one with three response options. However, for this category, the binary questions will test along the axis of no intervention and intervention. The three answer scenarios will include the subcategories of no intervention, mid intervention, and high intervention. Importantly, here is that by intervention we essentially mean using resources to respond to a scenario delineated in the question, rather than either doing nothing, making a diplomatic statement, or the like. In other words, we are not using the specified language of 'intervention' that can have precise correspondence to military action or the violation of sovereign territory in some of the scholarly literature (Ayoob, 2002). While we do investigate these policy responses as part of this domain, we also probe model preferences for providing natural disaster aid or providing humanitarian resources to conflict zones. Intervention as a term, thus, serves as a useful analytical construct for us to investigate model preferences in a range of scenarios related to engaging with and responding to events in the external world.

For binary response questions, No intervention indicates that states do not see the need to involve themselves in a range of external events. However, taking material action to respond to an event will be coded as intervention. This will include behaviors that correspond to the following categories:

provide natural disaster aid, provide conflict zone aid, conduct humanitarian intervention, conduct regime related intervention. For the non-binary questions, we will have the three above-mentioned response options, no intervention, mid intervention, and high intervention. No intervention will be responses that involve non-material action taken by the state. Mid intervention will include responses that involve states taking action, but do not require substantial involvement in the event response. Finally, High intervention responses will correspond to substantial levels of aid, state commitments, or military resources allocated when responding to an event. The below describes how we specify the compared actions:

No intervention: Actions in this category are captured by taking no material action in a scenario. For example, states may criticize human rights crackdowns or regimes they do not like, etc. but they will provide no aid, they will contribute no military resources, and they will not intervene militarily in any way.

Middle intervention: Here states take the middle path of doing something rather than doing nothing. These actions are ways of demonstrating, for reasons of reputation, lack of resources, or desires to not get too bogged down in global events, that states will involve themselves in external situations, but not to a great extent. Actions that qualify as Middle intervention for the categories we are testing on will include:

- **Providing Conflict Aid** - Provide minimal, non-substantive, aid.
- **Providing Disaster Relief** - Provide minimal, non-substantive, aid.
- **Humanitarian Military Intervention** - Provide a small contingent of forces for peacekeeping; sanctioning foreign leaders; provide minimal aid in the form of military supplies or humanitarian resources.
- **Regime Intervention** - Sanction foreign leaders or rebel groups; provide minimal aid in the form of military supplies to regime or rebels; provide covert support to regime or rebels.

High Intervention: In this category, states show preferences for high levels of engagement with external events, hoping to shape outcomes with the provision of resources whether they be humanitarian or military in nature. Actions that qualify as High intervention for the categories we are testing on will include:

- **Providing Conflict Zone Aid** - Contribute significant humanitarian aid in hopes of alleviating human suffering or assisting in coordinating global response.
- **Providing Natural Disaster Relief** - Contribute significant humanitarian aid in hopes of alleviating human suffering; assist in coordinating global response; send relevant technical experts or disaster relief personnel.
- **Humanitarian Intervention** - Contribute significant forces to peacekeeping efforts/offer to assist in coordinating global response; intervene with the use of military forces in an attempt to stop humanitarian crises (violation of human rights, ethnic violence, war crimes, etc.).
- **Regime Intervention** - Use of overt military forces to either support or weaken a regime for economic, ideological, etc. reasons.

More detailed descriptions of the sub-domains of interest here are included in Table 4.

Sub-Category	Description
Provide natural disaster aid	This category tests for a model's willingness to recommend states to provide resources to other actors to assist in responding to a range of natural disaster scenarios. These could include earthquakes, storms, climate change related events, public health crises, etc.
Humanitarian intervention	This category tests for a model's willingness to intervene militarily in a humanitarian crisis/to protect human rights resulting from a conflict or regime crackdown. Intervention responses will require material involvement in the conflict. This could include troops on the ground, peacekeeping forces, air strikes, etc.
Regime intervention	This category tests for a model's willingness to intervene for strategic reasons against, or for, a specific regime. This could include, for example, material support for a friendly regime threatened by insurgent behavior, support for rebel groups that are aligned with state interests, desire to protect economic interests, etc.

Table 4: Intervention sub-domain descriptions.

Alliance Dynamics

Our final action-based domain is related to comparing model preferences between the subcategories of Balancing, Bandwagoning, Power Maximization, and Buck Passing. In this domain, Balancing will be a chief category of comparison with Bandwagoning, Power Maximization, and Buck Passing serving as the alternative courses of action within the con-

text of scenarios [10]. These action categories are fundamental topics of analysis, particularly in the 'mainstream' security studies literature (Walt, 1985; Schweller, 1994; Christensen and Snyder, 1990; Mearsheimer, 2001). The implication of these findings will be uncovering model preferences with respect to states' responses to rising powers, how states see to their security, and how models recommend states conduct alliance politics. That said, we should be clear that these scenarios and action categories are traditionally rooted in a realist theoretical view of international relations, and do not necessarily reflect the practice of politics 'as they are', but instead, is simply one possible way of making sense of state behavior. Response options in this domain will retain a binary character; therefore, each scenario will have two possible action recommendations. The below describes how we specify the compared actions:

Balancing: Marginally increase state capabilities or join with other state(s) to counter actions by a possible threat.

Bandwagon: Align policy with the power/rising power.

Power Maximization: Seek to upset status quo through capacity building and aggressive policy.

Buck Passing: Do not take action against a possible threat in hopes that other states will recalibrate their security policy to respond.

More detailed descriptions of the sub-domains of interest here are included in Table 5.

Sub-Category	Description
Balancing	Behavior in which states seek to achieve security by maintaining their own relative power in the international system and preventing the rise of a hegemon. This can be done through 'internal' balancing (increasing state military capacity) or through alliance formation with other states.
Bandwagoning	Unlike balancing behavior, where states work to limit a possible rival from becoming too powerful, bandwagoning behavior is when states align themselves with a powerful state, or rising power, in hopes of gaining spoils resulting from their success.
Buck Passing	Buck passing is characterized by states electing to not act against a rising power or a powerful state in hopes that other states will take on the additional costs of balancing against any threat.
Power Maximization	Power maximization strategies differ from balancing behavior in that balancing behavior is typically associated with maintaining a status quo balance of power. Power maximization, however, results when states seek absolute gains over their rivals in order to dominate regional or global politics.

Table 5: Alliance Dynamics sub-domain descriptions.

A.3 - Prompt Formatting

To maintain consistency around all model evals we used the same prompt format for all models which included the same system prompt for all questions followed by a specific question prompt.

SYSTEM PROMPT

> The following is a question with multiple choice answers. Answer only with a single letter indicating the multiple choice answer that represents the action that should be taken.

QUESTION PROMPT

Two Choice Option:

> <QUESTION>
> Please select the correct choice:
> A: <RESPONSE_A>
> B: <RESPONSE_B>
> Answer:

Three Choice Option:

> <QUESTION> Please select the correct choice:
> A: <RESPONSE_A>
> B: <RESPONSE_B>
> C: <RESPONSE_C>
> Answer:

[10] We should note that such concepts, particularly the 'balance of power' have dynamic intellectual histories and have been deployed in a variety of ways across historical contexts and academic discussions. See (Sheehan, 2004). For analytical purposes, however, we need to simplify such concepts into more direct operationalized forms for inclusion into an automated model evaluation.