The Foreign Policy AI Evaluation Gap

Charles Pozniak [*1] **Jeba Sania** [*1]

Abstract

We argue that AI systems used in conducting foreign policy tasks — broadly enacting 'statecraft' – should be a priority test case for technical AI governance research. In enacting foreign policy, we refer to the formulation and implementation of external objectives by political actors. Statecraft is a high-consequence deployment domain, with extreme downside risks and structural properties that standard evaluation practices handle poorly. These features include partial observability, unbounded action spaces, contested ground truth, and multidimensional objectives. This paper advocates for a literature-grounded research agenda. Our contribution is threefold: (i) a claim about the structural conditions of foreign policy that combine catastrophic tail risk with technical evaluation complexities, (ii) an ECOSYSTEM review that highlights the asymmetric focus on ASSESSMENT features over ACCESS, VERIFICATION, SECURITY, and OPERATIONALIZATION, and (iii) a demand-side evaluation framework that decomposes foreign-policy workflows into bounded, evaluable sub-tasks with human recombination. As AI systems are already being deployed in the conduct of war and peace, amid limited public evaluation infrastructure from the technical AI governance community, this agenda is an urgent priority.

1. Introduction

AI systems have already entered foreign-policy workflows before the technical AI governance community has built evaluation infrastructure for them. Many initial uses appear administrative, but LLM systems are increasingly being explored for analysis, planning, and decision-support functions in military, diplomatic, and geopolitical contexts (Horowitz & Kahn, 2023; Rivera et al., 2024; Lamparth et al., 2024; Ma et al., 2024). Technical AI governance must catch up.

The technical AI governance literature provides a strong starting vocabulary. The Reuel–Bucknall taxonomy organizes the field around six governance capacities—ASSESSMENT, ACCESS, VERIFICATION, SECURITY, OPERATIONALIZATION, and ECOSYSTEM MONITORING—as well as targets focused on DATA, COMPUTE, MODELS, and DEPLOYMENT (Reuel et al., 2025; Bucknall & Trager, 2023). The taxonomy is intentionally domain-general. Our claim is that foreign policy exposes a domain-specific gap, with capacities most needed where ordinary evaluation conditions are weakest.

This paper argues that foreign policy is a domain blind spot in technical AI governance. By conducting foreign policy or statecraft we mean the purposive, institutionally mediated conduct through which a political actor seeks to shape its external environment.[1] This includes both the formation and implementation of external objectives: diplomatic positions, negotiations, sanctions, crisis response, and the use of force. This definition is narrower than International Relations as a whole, because it concerns the policies and decision processes of actors rather than the entire structure of the international system, but it is broader than diplomacy, featuring a broader range of instruments, such as economic and legal tools, alongside negotiation and representation (Allison & Zelikow, 1999; Halperin et al., 1974; Putnam, 1988; Schelling, 1960; Raiffa et al., 2002; Lax & Sebenius, 2006).

We use foreign policy and statecraft to identify a diverse deployment domain, rather than a monolithic capability to be tested. Whether a model is "good" at foreign policy is somewhat irrelevant, instead this speaks to real-world human–AI workflow integrations in high-stakes domains.

For the technical AI governance audience, this focus on foreign policy is important for both methodological and normative reasons. World politics violates the assumptions under which benchmark practice is easiest, while also being

[1]Belfer Center, Harvard Kennedy School, Cambridge, MA, USA. Correspondence to: Charles Pozniak <cpozniak@hks.harvard.edu>, Jeba Sania <jebasania@hks.harvard.edu>.

Second Workshop on Technical AI Governance Research (TAIGR) @ ICML 2026, Seoul, South Korea. 2026. Copyright 2026 by the author(s).

[1]The paradigmatic actor is the state, but the definition can also include international organizations, coalitions, and de facto authorities when they formulate external objectives and act toward other political communities.

a domain of great immediate consequence to human life.

This paper makes three contributions. First, it explains why foreign policy is a structurally difficult evaluation domain, organized around representation, inference, and specification. Second, it maps these difficulties to the six TAIG governance capacities, showing that the gap spans the capacities. Third, it reports a corpus-screening study and external scope check that quantify the scarcity and ASSESSMENT-heavy character of existing foreign-policy AI evaluation work, then proposes a demand-side framework for task-scoped evaluation under institutional and strategic constraints.

The systems most in need of disciplined evaluation are also the systems least likely to yield clean benchmarks, public datasets, stable labels, or repeatable tests.

2. Foreign Policy as a Priority Deployment Domain

Though many domains are "high stakes," foreign policy combines high consequence with poor measurability, strategic opacity, and weak external auditability. These factors mean that deployed systems evade effective technical governance while failures remain highly consequential. There are many instruments of foreign policy that vary across key dimensions of evaluability. Despite these differences, their institutional commonalities persist: activities ranging from arms control to crisis diplomacy are conducted through state action under strategic interaction. This shared deployment setting creates recurring evaluation problems, while the appropriate evaluation design must remain task- and subdomain-specific.

Foreign policy is a "high-stakes" domain in which failure carries consequences of a categorically different order than those in commercial or private-sector settings. In matters of economic policy, diplomacy, and military decision-making, errors can be catastrophic and, in some cases, irreversible. Poor trade policy can destabilize economies, misread signals can escalate conflicts, and daily decisions determine life and death. States cannot reasonably rely on beta-testing during crises or delicate negotiations, and the nature of these stakes means that governments are often structurally unwilling to tolerate the degree of risk that iterative improvement requires. Similarly, states are unwilling to admit error for the same reasons. This risk is compounded by two features particularly acute in enacting foreign policy: (i) the use of highly destructive resources and (ii) the nature of externalities to citizens domestic and foreign.

States, with their monopolies on violence, frequently deploy the most powerful and destructive resources. This includes both coercive force and ballistic weaponry, but also instruments like sanctions and export controls that are designed to inflict economic pain. An AI system that mishandles the application of these instruments, even marginally, operates in a domain where the consequences of error scale with the power of the actor deploying it. In no other setting is action taken with such a possibility for widespread destruction, both through physical and emotional harm and by limitations on the freedoms and capacities for self-actualization of individuals.

In foreign policy, the actions of states impact many individuals outside of the internal decision processes and the national domestic democratic process, if one exists. These externalities are distinct from most private-sector failures. The impacts of errors in these settings are like with monopolies: the affected citizens may be unable to exercise accountability, move to another polity, or take remedial actions. A state-led AI failure in diplomacy can rapidly impose costs on large numbers of individuals, as well as on the institutions and norms of the international system. Failed conflict mediation between two states, for instance, can trigger refugee crises, disrupt regional security, and impose economic and political costs on states far removed from the original dispute. Existing governance regimes are not equipped to fully internalize these externalities. Unlike private sector deployments, this creates risks that neither markets nor regulatory frameworks can fully manage.

3. Structural Limits of Evaluation

There are several structural features of world politics that mean AI systems used for foreign policy can resist governance mechanisms. These classic properties of world politics result in evaluative difficulties across three key domains: **representation**, **inference**, and **specification**.

Foreign policy settings are difficult to **represent** mathematically. Unlike in the board game *Diplomacy*, the state space is unbounded. Though rules exist, the very institutions that enforce them can be dissolved through negotiations. The action space is similarly large: the range of packages, moves, proposals, legal instruments, and tactics is enormous (Raiffa et al., 2002). Unlike individuals or profit-seeking businesses that are broadly bound under domestic law, foreign policy faces a different realm of constraints. Reductions of this complexity to single-shot interactions or multiple-choice options can mislead when treated as proxies for holistic, end-to-end judgment in diplomacy. They can, however, still be useful when scoped to bounded subtasks.

Contested ground truth is a constitutive feature of world politics, and one that makes accurate **inference** difficult.[2] In foreign policy, preferences are partly private, beliefs are noisy, and relevant counterfactual scenarios are unobserv-

[2] In this setting, inference refers to inferring a true state or belief from noisy information. This differs from the classic "inference" of the AI stack.

able (Fearon, 1995; Jervis, 1976). Strategies operate through commitment mechanisms, strategic signaling, and manipulation (Schelling, 1960). Strategic misrepresentation—of both actions and preferences—means that actors are also actively trying to limit the very ground truth on which effective action (and evaluation) depend.

Finally, success **specification** is an immediate challenge in foreign policy with multiple actors, each with their own preferences. Foreign-policy actors rarely optimize a single scalar objective. They may pursue security, legitimacy, economic advantage, and domestic political survival at the same time. Diplomatic decisions are also constrained by domestic ratification and bureaucratic process rather than by a single principal with a fixed utility function (Putnam, 1988; Allison & Zelikow, 1999; Halperin et al., 1974). Defining success is therefore partly a political act. Good evaluations should reward systems that surface trade-offs and preserve uncertainty, rather than forcing a single technical answer or collapsing ambiguity.

3.1. Implications for Technical AI Governance

These structural conditions make success difficult to evaluate, but they also have downstream implications for the AI governance ECOSYSTEM. The incentives for public disclosure are weak on both sides. Diplomatic and security operations are sensitive and states hesitate to publicize the settings in which they deploy AI systems. Similarly, commercial model developers face commercial and contractual pressures not to disclose failure cases or deployment constraints in classified, adversarial settings. The result is a domain in which the most consequential deployments are the least visible.

This lack of transparency creates dangerous information asymmetries. Sensitive deployments are least likely to be disclosed. Governments can become reliant on AI systems developed by private companies outside of the public safety discourse. States have strategic reasons to conceal the tasks, data, models, and evaluation standards used in diplomatic, intelligence, and security workflows. Vendors also have commercial and contractual reasons to say little. This is the inverse of much technical AI governance research, where the test set and results are publicly available.

The reliance on a small number of components in the stack means that strong one-time results are insufficient. Models are updated, benchmarks become gamed, and test items can leak into training data (Kapoor et al., 2025; Liang et al., 2025). This opens up a narrow surface area of manipulability, as counterparts can exploit vulnerabilities. Shared vendors and common model families can also create correlated failure risks across ministries and allies (Horowitz, 2018; Rivera et al., 2024; Jensen et al., 2025). Foreign policy, therefore, requires longitudinal evaluation and diligent ecosystem-level monitoring.

These dynamics are further destabilized by competitive pressures between states (Horowitz, 2018). Individually rational decisions to develop and deploy AI tools for strategic advantage can accelerate race dynamics, undermining international stability, eroding trust, and increasing the risk of crisis. Technological developments will continue to reconfigure the practice of diplomacy (Horowitz, 2018; Scharre, 2018).

These barriers contribute toward a significant governance gap. AI use in geopolitical analysis is already a high-consequence domain, yet the existing literature on AI governance lacks a shared, externally inspectable evaluation practice that is commensurate with that consequence. Closing this gap requires not only better benchmarks, but also the institutional will from states and model developers to prioritize new evaluation methodologies, with coordination mechanisms to ensure transparency.

4. A TAIG Capacity Gap

Despite the high stakes of deploying AI systems in foreign policy contexts, existing governance mechanisms remain inadequate across multiple dimensions. Technical AI governance can be understood as encompassing six interrelated subareas (ASSESSMENT, ACCESS, VERIFICATION, SECURITY, OPERATIONALIZATION, and ECOSYSTEM MONITORING) (Reuel et al., 2025). In the following section, we highlight each capacity in relation to statecraft.

ASSESSMENT refers to the evaluation of AI systems for technical properties, risks, and broader second-order implications for society at large (Reuel et al., 2025). Currently, there is a lack of rigorous approaches to benchmark construction that reflect the complexity of real diplomatic environments, as well as a lack of expert validation drawn from practitioners to validate model outputs. Beyond this, well-known pitfalls of evaluation methods, such as avoiding model contamination, assessing adversarial perturbations, and calibration, also apply (Eriksson et al., 2025; Hutchinson et al., 2022; Shevlane et al., 2023). Additionally, there is a lack of evaluations focused on rare but catastrophic failure modes.

ACCESS refers to the ability to interact with AI systems and obtain relevant data and information while avoiding unreasonable infringement of privacy (Reuel et al., 2025). Currently, there are inadequate external access mechanisms to support the training and evaluation of AI systems used for foreign policy, leaving governments with limited ability to scrutinize the systems they deploy (Reuel et al., 2025; Raji et al., 2022). Addressing this gap will require investments in infrastructure on both sides: governments need to build internal capacity to meaningfully engage with these systems,

and developers need to provide mechanisms for external access. In practice, this might include secure access regimes to protect sensitive information, redacted system cards that disclose relevant model properties without compromising national security, synthetic scenarios that allow stress testing without revealing classified data, and audit sandboxes (Brundage et al., 2020). Without such access, meaningful oversight of AI systems will be relegated to the realm of theory.

VERIFICATION is the ability of developers or third parties to verify claims made about AI systems' development, behavior, capabilities, or safety (Reuel et al., 2025). Unlike many commercial applications where outcomes can be observed and measured after deployment, AI systems used in foreign policy operate within complex, interdependent systems where ex-ante outcome verification is often structurally impossible. The consequences of a diplomatic decision may unfold over months or years, involve unintended actors, and resist neat attribution to any single choice or intervention. This makes traditional verification approaches, which tend to assume observable, discrete markers, ill-suited to the domain. Rather than attempting to verify outcomes after the fact, governance frameworks should prioritize verification at the process level. This means examining the intermediate steps and decision logic of AI systems in foreign policy before and during deployment, and decomposing complex tasks into components where claims about system behavior can be meaningfully assessed.

SECURITY refers to the protection of AI system components from unauthorized access, use, or tampering (Reuel et al., 2025). This area takes on heightened significance in high-stakes domains such as foreign policy. Threat models in diplomatic and foreign-policy settings can be vastly more consequential than in commercial deployments in terms of scale and intensity. The cost of a security failure, whether through unauthorized access to model internals, manipulation of outputs, or loss of sensitive data, can be devastating. Adversaries in this domain are sophisticated, well-resourced, and strategically motivated, making the need for robust security requirements for AI systems in foreign policy all the more necessary.

OPERATIONALIZATION translates governance methods to institutional practice (Reuel et al., 2025). In conducting statecraft, the operationalization gap is particularly acute due to the myriad pressures on idiosyncratic interpersonal practices. For example, hierarchy, accountability, and confidentiality concerns can distort the effective translation of evaluation outputs. As such, AI governance in foreign policy requires a deep understanding of potential deployment protocols that can preserve human authority over contestable judgments.

ECOSYSTEM MONITORING refers to the ongoing study of the evolving landscape of AI development, application, and associated impacts (Reuel et al., 2025). AI systems used for foreign-policy use cases can inhibit such monitoring. For example, the classified nature of many government AI deployments makes it difficult to assess the degree of model monoculture, that is, the extent to which states are relying on a narrow set of systems or architectures that create systemic vulnerabilities. Compounding this, vendor dependence within government can become deeply entrenched over time, with shared data pipelines and other integrated infrastructure making it difficult to transition away from specific providers. Similarly, procurement processes in government contexts may present a high bar for new market entrants, limiting opportunities for experimentation. Foreign policy, therefore, requires longitudinal evaluation and ecosystem-level monitoring, not only static benchmark scores.

Taken together, these properties imply that foreign policy deployments cannot be governed through model-only leaderboards. Following Weidinger et al. 2023, evaluation should incorporate capability testing alongside human interaction and system-level effects — both of which require a more rigorous investigation in statecraft deployments. As in Shevlane et al. 2023, high risk settings should separate the possible capabilities of a system from its disposition under realistic institutional incentives and deployment settings. These are particularly acute pressures in the conduct of statecraft and will demand more investigation on the systemic implications of combined human-AI in contested domains.

Of the six governance subareas, ASSESSMENT remains the most immediate entry point for technical researchers, but foreign-policy evaluation cannot remain ASSESSMENT-only. The evidence map below therefore empirically evaluates whether public AI evaluation and technical governance papers focus on systems or workflows in statecraft-relevant settings. Alongside ACCESS, ASSESSMENT forms the foundation upon which all other governance mechanisms depend. Without the ability to assess the capabilities of AI systems, achieving meaningful oversight becomes significantly complicated, if not impossible.

Critically, TAIG researchers are best positioned to make immediate and meaningful progress on evaluation practices that could have a widespread impact on human life. The current literature on evaluations in foreign policy and statecraft is sparse, however. Through a literature review and quantitative analysis of existing work, we find strong evidence that this domain of technical AI governance research is underdeveloped.

Quantity	Result
Primary rows in merged warehouse	93,684
Headline rows screened	79,954
Title-and-abstract rows	76,653
Title-only rows	3,089
Non-paper accounting rows	212
LLM adjudication queue	1,370
Accepted domain+method rows	413
Raw direct/proxy adjudicated labels	14
Direct real-world labels before QA	7
Proxy statecraft benchmark labels	7
Negative-audit strict misses	0

Table 1. Current corpus-screening and adjudication evidence for the foreign-policy evaluation gap. Counts are title-and-abstract-first screening and adjudication results. The 413 accepted domain+method rows are broader contextual evidence and should not be interpreted as direct foreign-policy AI evaluation coverage.

5. Evidence from a Corpus Screen

Technical AI governance has developed rapidly around the problem of making advanced AI systems measurable, before and during deployment. Despite these gains, progress has been uneven.

5.1. Methodology

We report an evidence map of titles and abstracts with 93,684 primary rows, with 76,653 records with title and abstract and 3,089 title-only records. The study window includes works published from 2018 through 2026-04-22.

The merged corpus draws from three overlapping source families: a main AI/ML/NLP venue corpus, an AI governance, policy, safety, assurance, and evaluation corpus, and a statecraft-adjacent bridge corpus. Because corpus memberships are not mutually exclusive, pooled results should be interpreted as merged descriptive results rather than independent corpus sums.

We first ran a high-recall screen over titles and abstracts, normalizing text and matching AI/ML terms, TAIGR method terms, and statecraft-adjacent terms. From this, we built an adjudication and negative audit pipeline, adjudicating the top 1,370 candidates and a 500-record negative-audit sample that would have been excluded based on the prior matching. Each row was coded for relevance to statecraft and TAIGR evaluation method relevance.

5.2. Empirical Evidence

TAIGR has built substantial machinery for a range of analyses that could be relevant to high-stakes domains like statecraft. This includes strong work on assessing and benchmarking capabilities, but also the safety, privacy, and equity implications, with policy-facing analysis on security and systemic considerations. Despite this work, the corpus screen gives little evidence of a mature evaluation agenda for foreign-policy decision contexts. The field is not empty, but the direct literature remains small and overwhelmingly ASSESSMENT-centered. We identify 12 papers (out of 79,954), approximately 0.015% of the screened corpus. Five are direct real-world statecraft evaluations, and seven are proxy statecraft benchmarks.[3]

The broader set of 413 domain relevant papers indicate active evaluation and governance methods in nearby areas, especially cyber/information operations and generic negotiation or mixed-motive agent evaluation. These papers, however, miss many of the central structural features of taking actions in world politics. Statecraft is not wholly absent, but directly relevant work remains scarce.

The adjudicated results and accompanying literature review reveal four methodological gaps in current evaluation approaches.

1. Evaluations often lack ecological validity. Evaluation settings for foreign policy, negotiation, and diplomacy are often deliberately simplified in ways that fail to mirror real-world complexity (Bakhtin et al., 2022; Kramár et al., 2022; Xu et al., 2025; Duffy et al., 2025; Johanson et al., 2022). While such simplification reduces confounding factors and enables tractable measurement, it distorts the real-life nature of the settings being modeled. Utility functions are typically limited to one or two dimensions such as price or success ratios (Bagga et al., 2020; Luo et al., 2024; Fu et al., 2023), while multiple-choice benchmarks flatten the open-ended nature of real, dynamic negotiations (Jensen et al., 2025). In simulation-based experiments, critical dimensions of diplomatic success, such as the durability of terms over time, are rarely accounted for. In diplomatic contexts especially, how an agreement is reached—the sequencing of concessions, the management of impasse, the gradual building of trust—is often as consequential as the final outcome itself(Ma et al., 2024).

2. There is a lack of publicly available, common datasets for benchmarking AI systems in the foreign-policy domain. Among the works identified in our review, public resources are concentrated in highly stylized strategic environments,

[3]Our five paper-ready direct real-world hits are: "Not Oracles of the Battlefield: Safety Considerations for AI-Based Military Decision Support Systems"; "Human vs. Machine: Behavioral Differences between Expert Humans and Language Models in Wargame Simulations"; "Escalation Risks from Language Models in Military and Diplomatic Decision-Making"; "This Land is Your, My Land: Evaluating Geopolitical Bias in Language Models through Territorial Disputes"; "DiplomacyAgent: Do LLMs Balance Interests and Ethical Principles in International Events?". The seven proxy statecraft benchmark hits include no-press and full-press Diplomacy work, Civilization/CivRealm, Cicero Diplomacy play, and DipLLM.

such as Diplomacy or Civilization-style simulations, rather than real diplomatic settings and workflows. The Automated Negotiating Agents Competition (Vaccaro et al., 2026) is one of the few resources approximating a relevant subdomain, but its domain focus is on generic negotiations. This is a relevant adjacent literature, but is not counted in our corpus as directly concerning statecraft due to the absence of institutionally grounded international decision contexts.

This data gap is likely motivated by the difficulty of determining, automating, and annotating datasets for ground truth. The most relevant data for benchmarking AI systems in foreign policy is drawn from real diplomatic interactions. However, this is the data that states are least willing to disclose, given its sensitivity to national security. Without standardized evaluation resources, comparability and reproducibility between different AI systems break down, creating a barrier to understanding field-wide progress in capabilities (Liao & Xiao, 2025; Biderman et al., 2026; Kapoor et al., 2025).

3. The conduct of foreign policy is often done through highly interpersonal settings.
Existing evaluation methods fail to capture that nuance. Key evaluation criteria, such as managing emotional dynamics, building trust, and demonstrating cultural sensitivity, are critical to the success of a negotiation (Fulmer & Barry, 2004; Elfenbein et al., 2006), but these interpersonal dimensions are rarely evaluated (Ma et al., 2024). By reducing diplomacy to strictly outcome metrics, existing evaluation frameworks risk encoding a transactional conception of negotiation and conflict management. As evaluations are used to inform research directions, such encoding could systematically reward the wrong behaviors for future models and mischaracterize effective foreign policy.

4. There is considerable fragmentation of evaluation approaches across studies.
This fragmentation reflects the diversity of conceptualizations of success for different players across foreign-policy domains and contexts (Putnam, 1988; Allison & Zelikow, 1999). A nation's foreign policy involves many competing goals. For instance, an agent evaluated on economic outcomes in a trade negotiation context cannot be meaningfully compared to one assessed on ceasefire durability in a conflict mediation setting. This fragmentation makes it difficult to advance a *general* AI system across foreign-policy use cases, or build cumulative knowledge about holistic model performance.

Adjacent evaluations in generic negotiations and proxy strategic games provide partial evidence, but they do not measure the same constructs. An agent evaluated on economic surplus in a generic bargaining task cannot be directly compared with a model evaluated on geopolitical bias in a statecraft analysis workflow. Evaluations should therefore distinguish domain-specific criteria, such as ceasefire analysis or sanctions compliance, from more general process-quality criteria, such as provenance, uncertainty preservation, escalation-to-review behavior, and robustness to strategic misrepresentation.

Shared methodological benchmarks should enable reproducible comparisons within well-defined task families rather than force comparability across heterogeneous statecraft-adjacent domains. Plausible scenarios grounded in real cases should be paired with evaluation criteria that account for both process and outcome quality (Schwartz et al., 2025). As AI systems become more widely used in foreign-policy contexts, evaluation practices will need to match the complexity of the environments they are meant to govern.

6. Towards A Demand-Side Evaluation Agenda

Our corpus scan suggests a supply-side bias. Much of the broader accepted literature begins with a model and failure mode, then selecting tasks to expose it. The statecraft-relevant strict hits are too sparse and heterogeneous to support a model-only leaderboard. AI governance in foreign-policy settings should instead start from the demand side. This involves understanding the workflows that need supporting and testing AI systems within the institutional and organizational structures that give authority and accountability.

Table 2 gives an illustrative set of task families to shift evaluation design away from generic model leaderboards and toward task-scoped tests that map onto real foreign-policy workflows.

For each task family, researchers should build task cards: structured specifications that define the task-relevant information. A task card for escalation detection, for example, should specify the scenario, signals, distortions, and the conditions of system uncertainty.

For a demand-side evaluation, researchers should specify discrete tasks, the human role, and potential scoring regimes. Accurately characterizing the task means bounding the activity appropriately within a complex existing system, with the model's role and definition of success. This might include mapping actors and their positions, or generating policy options. The human role should also be made clear — at each point in these workflows, responsibility should return to an accountable human decision-maker. Any scoring regime would need to specify what can be evaluated directly, what must be judged by experts, and what should be treated as unresolvable within the benchmark.

A useful evaluation may ask whether a system faithfully accurately cited evidence or designed strategies that obeyed

Stage	Task	Output	Evaluation	Primary TAIG capacities
Research	Actor and constraint mapping	Actor-interest-constraint map	Coverage, source provenance, uncertainty marking, omission detection	Assessment, Verification
Analyze	Escalation-signal detection	Risk assessment with evidence links	Calibration, adversarial robustness, false-positive control, caveat preservation	Assessment, Security
Strategize	Option generation	Menu of policy options	Trade-off coverage, legal feasibility, non-escalatory alternatives, red-team objections	Assessment, Operationalization
Execute	Draft-language review	Flagged clauses, caveats, obligations	Citation fidelity, ambiguity handling, obligation tracking, human-review triggers	Verification, Operationalization
Monitor	Agreement compliance tracking	Compliance assessment and unresolved issues	Provenance, longitudinal consistency, source conflict handling, escalation to review	Verification, Ecosystem Monitoring

Table 2. Demand-side evaluation agenda for foreign-policy AI systems. The unit of evaluation is task fitness under institutional and strategic constraints, evaluated through bounded artifacts rather than generic model performance.

certain laws. These properties are more tractable than evaluating whether the model "solved" a foreign-policy problem or exhibited some disposition or proclivity during strategic reasoning.

Good demand-side evaluations should also be able to compare systems against real baselines. Models may perform well in isolation, but comparison between human–AI workflows and the pre-existing institutional processes will be important evaluation. Understanding relative strengths and weaknesses, as well as failure modes, is easier with strong baselines. Where possible, comparisons between human-only performance, model-only performance, and human-model collaboration should help trace the evolution of (over)reliance and the state of capacities for judgment, both human and institutional.

Finally, foreign-policy evaluation resources require strict governance. Public release should be encouraged, but sensitive evaluations that explore vulnerabilities might want to consider controlled-access settings, with expert review panels or secure audit sandboxes. Though the cumulative improvement of evaluation science is important, the unrestricted publication of vulnerabilities could weaken the relative position of more responsible parties. Demand-side evaluation is therefore both a technical agenda and an access-design problem: the field needs benchmarks, but it also needs governance arrangements that determine who can run them and when information needs to be disclosed.

6.1. Limitations

Our claims are limited to the public, externally inspectable research ecosystem, and comes with several potential limitations. First, the public literature gap may overstate the practical governance gap because classified evaluation practices may exist inside governments or vendors. Second, our main screening run is based on titles and abstracts alone and doesn't consider full-text content. Third, critics may argue that foreign policy should not be benchmarked at all, noting that benchmark construction bears the risk of flattening contested political judgment and detaching human agency from key normative issues. Finally, foreign-policy benchmarks may be dual-use: better evaluation can improve safety, but it can also improve systems used for coercion, deception, or strategic manipulation.

We agree that model-only leaderboards for holistic diplomatic activities would not be a useful contribution. Our proposal is scoped, task-based evaluation infrastructure for human-supervised institutional workflows.

7. Conclusion

Foreign policy and statecraft settings should be a priority domain for technical AI governance because it combines stark consequences with weak evaluability. The structures of foreign policy decisions, with private information between adversaries and contested objectives on all sides, make standard benchmark practice insufficient. The resulting governance gap needs addressing: AI systems are being explored for foreign-policy workflows before the field has built evaluation infrastructure suited to those workflows.

This paper argues for a demand-side research agenda. Instead of treating foreign policy as a single benchmarkable capability, evaluations should break down institutional workflows into decomposed, bounded tasks. These tasks should define the downstream outputs, preserving human authority over judgment and accountability over outcomes. The agenda outlined in this paper gives technical AI governance a tractable path into a domain where disciplined evaluation is urgently needed.

Impact Statement

This paper does not develop, deploy, or advocate any AI system for foreign policy. Its primary impact is to shape how such systems are evaluated if governments, international organizations, or vendors adopt them. Better evaluation could reduce unsafe deployment, reveal failure modes, and improve institutional accountability. It could also improve systems used for coercive or destabilizing state action. For this reason, we emphasize task-scoped evaluation, practitioner validation, controlled access, auditability, and accountable human authority rather than unrestricted benchmark release or automated diplomacy.

Evaluation infrastructure is not neutral. Benchmarks and scoring regimes can encode narrow conceptions of success, especially in domains shaped by unequal power, secrecy, and contested legitimacy. The paper therefore treats evaluation as a governance problem as well as a technical problem: evaluations should specify whose objectives are represented, what forms of harm are considered, what information is excluded, and where human responsibility remains non-delegable.

LLM Usage Statement

During the preparation of this work, the authors used Claude by Anthropic to improve readability and formatting of the manuscript. ChatGPT, Claude, and OpenAI Codex were also used to support corpus-analysis workflows, including code assistance, report generation, and title-abstract adjudication. The adjudication procedure used a fixed schema and conservative boundary rules. This involved hand-calibrated labels for all strict core rows, as well as the first 50 papers from cyber/info-ops, QA-checked, duplicate work IDs, strict hits, and the negative-audit candidates. The authors reviewed the outputs used in the manuscript and take full responsibility for the final content.

References

Allison, G. T. and Zelikow, P. *Essence of Decision: Explaining the Cuban Missile Crisis*. Longman, 2nd edition, 1999.

Bagga, P., Paoletti, N., Alrayes, B., and Stathis, K. A Deep Reinforcement Learning Approach to Concurrent Bilateral Negotiation, February 2020. URL http://arxiv.org/abs/2001.11785. arXiv:2001.11785 [cs].

Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried, D., Goff, A., Gray, J., Hu, H., Jacob, A. P., Komeili, M., Konath, K., Kwon, M., Lerer, A., Lewis, M., Miller, A. H., Mitts, S., Renduchintala, A., Roller, S., Rowe, D., Shi, W., Spisak, J., Wei, A., Wu, D., Zhang, H., and Zijlstra, M. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, December 2022. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.ade9097. URL https://www.science.org/doi/10.1126/science.ade9097.

Biderman, S., Schoelkopf, H., Sutawika, L., Gao, L., Tow, J., Abbasi, B., Aji, A. F., Ammanamanchi, P. S., Black, S., Clive, J., DiPofi, A., Etxaniz, J., Fattori, B., Forde, J. Z., Foster, C., Hsu, J., Jaiswal, M., Lee, W. Y., Li, H., Lovering, C., Muennighoff, N., Pavlick, E., Phang, J., Skowron, A., Tan, S., Tang, X., Wang, K. A., Winata, G. I., Yvon, F., and Zou, A. Lessons from the trenches on reproducible evaluation of language models, 2026. URL https://arxiv.org/abs/2405.14782.

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P. W., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensold, J., O'Keefe, C., Koren, M., Ryffel, T., Rubinovitz, J., Besiroglu, T., Carugati, F., Clark, J., Eckersley, P., de Haas, S., Johnson, M., Laurie, B., Ingerman, A., Krawczuk, I., Askell, A., Cammarota, R., Lohn, A., Krueger, D., Stix, C., Henderson, P., Graham, L., Prunkl, C., Martin, B., Seger, E., Zilberman, N., hÉigeartaigh, S. O., Kroeger, F., Sastry, G., Kagan, R., Weller, A., Tse, B., Barnes, E., Dafoe, A., Scharre, P., Herbert-Voss, A., Rasser, M., Sodhani, S., Flynn, C., Gilbert, T. K., Dyer, L., Khan, S., Bengio, Y., and Anderljung, M. Toward trustworthy ai development: Mechanisms for supporting verifiable claims, 2020. URL https://arxiv.org/abs/2004.07213.

Bucknall, B. S. and Trager, R. F. Structured Access for Third-Party Research on Frontier AI Models: Investigating Researchers' Model Access Requirements. Technical report, Oxford Martin AI Governance Initiative, Centre for the Governance of AI, October 2023.

Duffy, A., Paech, S. J., Shastri, I., Karpinski, E., Allouis-Cros, B., Marques, T., and Olson, M. L. Democratizing diplomacy: A harness for evaluating any large language model on full-press diplomacy, 2025. URL https://arxiv.org/abs/2508.07485.

Elfenbein, H. A., Foo, M.-D. D., White, J. B., Tan, H. H., and Aik, V.-C. Reading your counterpart: The benefit of emotion recognition accuracy for effectiveness in negotiation. *SSRN Electron. J.*, 2006.

Eriksson, M., Purificato, E., Noroozian, A., Vinagre, J., Chaslot, G., Gomez, E., and Fernandez-Llorca, D. Can we trust ai benchmarks? an interdisciplinary review of current issues in ai evaluation, 2025. URL https://arxiv.org/abs/2502.06559.

Fearon, J. D. Rationalist Explanations for War. *International Organization*, 49(3):379–414, 1995.

Fu, Y., Peng, H., Khot, T., and Lapata, M. Improving language model negotiation with self-play and in-context learning from ai feedback, 2023. URL https://arxiv.org/abs/2305.10142.

Fulmer, I. and Barry, B. The smart negotiator: Cognitive ability and emotional intelligence in negotiation. *International Journal of Conflict Management*, 15:245–272, 12 2004. doi: 10.1108/eb022914.

Halperin, M. H., Kanter, A., and Clapp, P. *Bureaucratic Politics and Foreign Policy*. Brookings Institution Press, 1974.

Horowitz, M. C. Artificial Intelligence, International Competition, and the Balance of Power. *Texas National Security Review*, 1(3):36–57, 2018.

Horowitz, M. C. and Kahn, L. Bending the Automation Bias Curve: A Study of Human and AI-based Decision Making in National Security Contexts, June 2023. URL http://arxiv.org/abs/2306.16507. arXiv:2306.16507.

Hutchinson, B., Rostamzadeh, N., Greer, C., Heller, K., and Prabhakaran, V. Evaluation gaps in machine learning practice, 2022. URL https://arxiv.org/abs/2205.05256.

Jensen, B., Reynolds, I., Atalan, Y., Garcia, M., Woo, A., Chen, A., and Howarth, T. Critical foreign policy decisions (cfpd)-benchmark: Measuring diplomatic preferences in large language models, 2025. URL https://arxiv.org/abs/2503.06263.

Jervis, R. *Perception and Misperception in International Politics*. Princeton University Press, 1976.

Johanson, M. B., Hughes, E., Timbers, F., and Leibo, J. Z. Emergent bartering behaviour in multi-agent reinforcement learning, 2022. URL https://arxiv.org/abs/2205.06760.

Kapoor, S., Stroebl, B., Siegel, Z. S., Nadgir, N., and Narayanan, A. AI Agents That Matter. *Transactions on Machine Learning Research*, 2025.

Kramár, J., Eccles, T., Gemp, I., Tacchetti, A., McKee, K. R., Malinowski, M., Graepel, T., and Bachrach, Y. Negotiation and honesty in artificial intelligence methods for the board game of diplomacy. *Nat. Commun.*, 13(1): 7214, December 2022.

Lamparth, M., Corso, A., Ganz, J., Skylar Mastro, O., Schneider, J., and Trinkunas, H. Human vs. Machine: Behavioral Differences Between Expert Humans and Language Models in Wargame Simulations. *arXiv preprint arXiv:2403.03407*, 2024.

Lax, D. A. and Sebenius, J. K. *3-D Negotiation: Powerful Tools to Change the Game in Your Most Important Deals*. Harvard Business Review Press, 2006.

Liang, Y., Yang, L., Wang, C., Xia, C., Meng, R., Xu, X., Wang, H., Payani, A., and Shu, K. Benchmarking LLMs for Political Science: A United Nations Perspective, February 2025. URL http://arxiv.org/abs/2502.14122. arXiv:2502.14122 [cs].

Liao, Q. V. and Xiao, Z. Rethinking model evaluation as narrowing the socio-technical gap, 2025. URL https://arxiv.org/abs/2306.03100.

Luo, X., Li, Y., Huang, Q., and Zhan, J. A survey of automated negotiation: Human factor, learning, and application. *Computer Science Review*, 54:100683, November 2024. ISSN 1574-0137. doi: 10.1016/j.cosrev.2024.100683. URL https://www.sciencedirect.com/science/article/pii/S1574013724000674.

Ma, Z., Mei, Y., Bruderlein, C., Gajos, K. Z., and Pan, W. "chatgpt, don't tell me what to do": Designing ai for context analysis in humanitarian frontline negotiations, 2024. URL https://arxiv.org/abs/2410.09139.

Putnam, R. D. Diplomacy and Domestic Politics: The Logic of Two-Level Games. *International Organization*, 42(3): 427–460, 1988.

Raiffa, H., Richardson, J., and Metcalfe, D. *Negotiation Analysis: The Science and Art of Collaborative Decision Making*. Harvard University Press, 2002.

Raji, I. D., Xu, P., Honigsberg, C., and Ho, D. E. Outsider oversight: Designing a third party audit ecosystem for ai governance, 2022. URL https://arxiv.org/abs/2206.04737.

Reuel, A., Bucknall, B., Casper, S., Fist, T., Soder, L., Aarne, O., Hammond, L., Ibrahim, L., Chan, A., Wills, P., Anderljung, M., Garfinkel, B., Heim, L., Trask, A., Mukobi, G., Schaeffer, R., Baker, M., Hooker, S., Solaiman, I., Luccioni, A. S., Rajkumar, N., Moës, N., Ladish, J., Bau, D., Bricman, P., Guha, N., Newman, J., Bengio, Y., South, T., Pentland, A., Koyejo, S., Kochenderfer, M. J., and Trager, R. Open Problems in Technical AI Governance, April 2025. URL http://arxiv.org/abs/2407.14981. arXiv:2407.14981 [cs].

Rivera, J.-P., Mukobi, G., Reuel, A., Lamparth, M., Smith, C., and Schneider, J. Escalation Risks from Language Models in Military and Diplomatic Decision-Making. In

Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT), 2024. doi: 10.1145/3630106.3658942.

Scharre, P. *Army of None: Autonomous Weapons and the Future of War*. W.\,W.\,Norton, 2018.

Schelling, T. C. *The Strategy of Conflict*. Harvard University Press, 1960.

Schwartz, R., Chowdhury, R., Kundu, A., Frase, H., Fadaee, M., David, T., Waters, G., Taik, A., Briggs, M., Hall, P., Jain, S., Yee, K., Thomas, S., Bhandari, S., Duncan, P., Thompson, A., Carlyle, M., Lu, Q., Holmes, M., and Skeadas, T. Reality check: A new evaluation ecosystem is necessary to understand ai's real world effects, 2025. URL https://arxiv.org/abs/2505.18893.

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., Ho, L., Siddarth, D., Avin, S., Hawkins, W., Kim, B., Gabriel, I., Bolina, V., Clark, J., Bengio, Y., Christiano, P., and Dafoe, A. Model Evaluation for Extreme Risks. *arXiv preprint arXiv:2305.15324*, 2023.

Vaccaro, M., Caosun, M., Ju, H., Aral, S., and Curhan, J. R. Advancing ai negotiations: A large-scale autonomous negotiation competition, 2026. URL https://arxiv.org/abs/2503.06416.

Weidinger, L., Marchal, N., Rauh, M., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., Bergman, S., Gabriel, I., Griffin, C., Kay, J., Bariach, B., Rieser, V., and Isaac, W. Sociotechnical Safety Evaluation of Generative AI Systems. *arXiv preprint arXiv:2310.11986*, 2023.

Xu, K., Chai, J., Li, S., Fu, Y., Zhu, Y., and Zhao, D. Dipllm: Fine-tuning llm for strategic decision-making in diplomacy, 2025. URL https://arxiv.org/abs/2506.09655.

A. Appendix: Corpus Screening and Codex-Adjudicated Evidence Map

This appendix documents the corpus-screening and adjudication procedure supporting Section 5. This screen estimates whether public technical AI governance research substantially covers AI systems or AI-enabled workflows in foreign-policy and statecraft contexts through a title-and-abstract evidence map over a bounded public research corpus.

A.1. Purpose and Scope

The study looks at whether public TAIGR covers statecraft-relevant settings (e.g., foreign policy, diplomacy, military decision-making, intelligence workflows, crisis management, arms control, sanctions, information operations, cyber operations, geopolitical governance, and state-to-state strategic interaction).

A substantive hit requires *both* foreign policy domain relevance *and* TAIGR relevance. The title and abstract must indicate that the paper evaluates, benchmarks, audits, governs, verifies, monitors, red-teams, assesses, or operationalizes an AI system or AI-enabled workflow in a relevant domain. Full text was not used as part of the primary screen.

A.2. Corpus

The study window includes works with a published date from 2018-01-01 through 2026-04-22. Where only publication year was available, works from 2018 through 2026 were included and treated as year-only window rows. Pre-window, post-window, and unknown-year rows were excluded from the headline denominator.

The merged warehouse contains 93,684 primary rows. The study-window slice used by the deterministic screen contained 79,954 rows. This denominator includes 76,653 title-and-abstract rows, 3,089 title-only rows, and 212 non-paper accounting rows retained.

The merged corpus consolidates three overlapping source families: (i) a main AI/ML/NLP venue corpus, (ii) an AI governance, policy, safety, assurance, and evaluation corpus, and (iii) a statecraft-adjacent bridge corpus. Corpus memberships are not mutually exclusive. Pooled results should therefore be interpreted as merged descriptive results, not independent corpus sums.

Quantity	Result
Primary rows in merged warehouse	93,684
Headline rows screened	79,954
Title-and-abstract rows	76,653
Title-only rows	3,089
Non-paper accounting rows	212
LLM/Codex adjudication queue	1,370
Accepted domain+method rows	413
Raw direct/proxy adjudicated labels	14
Conservative paper-ready strict hits	12
Direct real-world paper-ready hits	5
Proxy statecraft benchmark paper-ready hits	7
Negative-audit strict misses	0

Table 3. Current corpus-screening and adjudication evidence. Counts are title-and-abstract-first screening and adjudication results. The 413 accepted domain+method rows are broader contextual evidence and should not be interpreted as direct foreign-policy AI evaluation coverage.

A.3. Deterministic Screen

The first stage was a deterministic screen over titles and abstracts. For each work, evidence text was built from the title and abstract, normalized for punctuation, and matched to four families of terms: AI/ML, TAIGR evaluation/governance methods, statecraft-relevant, and statecraft-adjacent (e.g., cyber, information-operations, geopolitical-governance, and generic-negotiation). This deterministic screen built the queue for deeper screening, rows outside likely-hit buckets formed the negative-audit pool.

A known-positive recall check used 17 seed papers. All 17 were found by the deterministic screen; none were missed by regex and none were absent from the denominator.

Deterministic bucket	Rows
Negative audit pool	52,477
Exclude: method only	24,629
Exclude: domain only	1,677
Generic negotiation or agent evaluation	768
Adjacent cyber/info-ops evaluation	354
Adjacent geopolitical AI governance	23
Sparse statecraft-AI review	10
Core direct statecraft evaluation	9
Core proxy statecraft evaluation	7

Table 4. Deterministic-screen bucket counts over the 79,954-row headline denominator. The screen was designed for recall and did not determine final inclusion.

A.4. LLM Adjudication Queue

The adjudication queue contained 1,370 rows. It included all strict/core candidates, adjacent cyber and information-operations candidates, adjacent geopolitical and sparse-review candidates, generic negotiation and agent candidates, and a 500-row negative-audit sample drawn from deterministic negatives.

Queue source	Rows
Negative audit pool sample	500
Generic negotiation or agent evaluation	472
Adjacent cyber/info-ops evaluation	354
Adjacent geopolitical AI governance	20
Core direct statecraft evaluation	9
Sparse statecraft-AI review	8
Core proxy statecraft evaluation	7
Total	1,370

Table 5. Composition of the Codex adjudication queue. The queue deliberately does not include all 79,954 headline rows. It includes likely hits plus a negative-audit sample.

Each queued row was presented with its work identifier, title, abstract where available, venue, venue family, year, corpus membership, evidence basis, deterministic bucket, matched regex terms, and screen reason.

A.5. Coding Schema

Each adjudicated row received labels for statecraft relevance, TAIGR evaluation relevance, deployment context, AI object, evaluation method, confidence, an evidence quote, and notes.

The boundary was relatively conservative, papers describing deployment alone were sufficient. The title or abstract had to indicate some form of TAIGR capacity.

A.6. Adjudicated Results

The LLM (GPT-5.5) adjudication accepted 413 rows as domain+method relevant. Most accepted rows were from adjacent domains rather than direct statecraft. After adjudication, the conservative strict count was 12: five direct real-world rows and seven proxy statecraft-benchmark rows. Two additional labeled papers were deemed borderline due to limited evaluation investigation.

A.7. Strict Direct and Proxy QA

The raw direct/proxy adjudicated count was 14. Manual QA retained 12 rows as conservative paper-ready strict evidence and marked two direct-labeled rows as contextual or borderline. The conservative strict count is the headline count used in the paper.

The two contextual rows were excluded from the conservative headline for two reasons. *Knowledge-Embedded Narrative*

Label	Rows	Share of queue	Share of denominator
Direct real-world	7	0.51%	0.009%
Proxy statecraft benchmark	7	0.51%	0.009%
Adjacent geopolitical governance	4	0.29%	0.005%
Adjacent cyber/info-ops	310	22.63%	0.388%
Generic negotiation/mixed-motive	85	6.20%	0.106%
Rejected / not relevant	957	69.85%	1.197%
Total	1,370	100.00%	1.713%

Table 6. Codex-adjudicated relevance labels (pre-QA counts). Manual QA subsequently reduces the direct real-world count from 7 to 5 (see the strict direct/proxy QA table); the conservative post-QA strict total is 12. Shares of the denominator use the 79,954-row headline denominator.

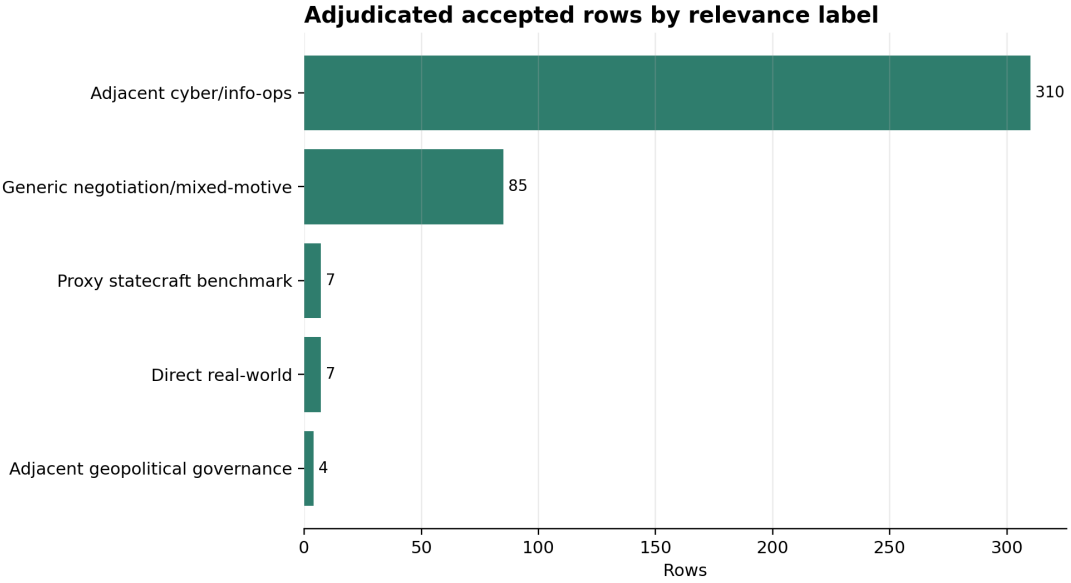

Figure 1. Adjudicated accepted rows by relevance label.

Construction from Open Source Intelligence is intelligence-workflow relevant, but is more of a proposed system rather than an evaluation methodology. *A Survey of Large Language Model Use and Its Technical Limitations in Military Systems Through a Decolonial Lens* is relevant to military LLM use, but does not provide a concrete evaluation, benchmark, audit, or workflow test of a deployed or simulated statecraft process.

The negative audit sampled 500 rows from the deterministic negative pool. The audit produced four possible false negatives. All four were low-confidence adjacent cyber/info-ops rows. None were paper-ready strict direct or proxy misses.

The negative audit supports the stability of the conservative statecraft claim. Improving recall for adjacent cyber/info-ops would require broader regex terms, but the audit did not reveal missed paper-ready direct real-world or proxy statecraft-benchmark evidence.

A.8. Methods Among Accepted Rows

The 413 domain+method papers included evaluations, benchmarks, datasets, tools, and platforms. This broader set is useful for locating nearby methodological activity, but it should not be collapsed into direct statecraft evidence.

A.9. Reproducibility

The row-level files accompanying this appendix preserve paper identifiers, evidence-basis fields, confidence labels, and high-level categories for domain and TAIG capacity. The aggregate CSVs include the tables reported above and additional

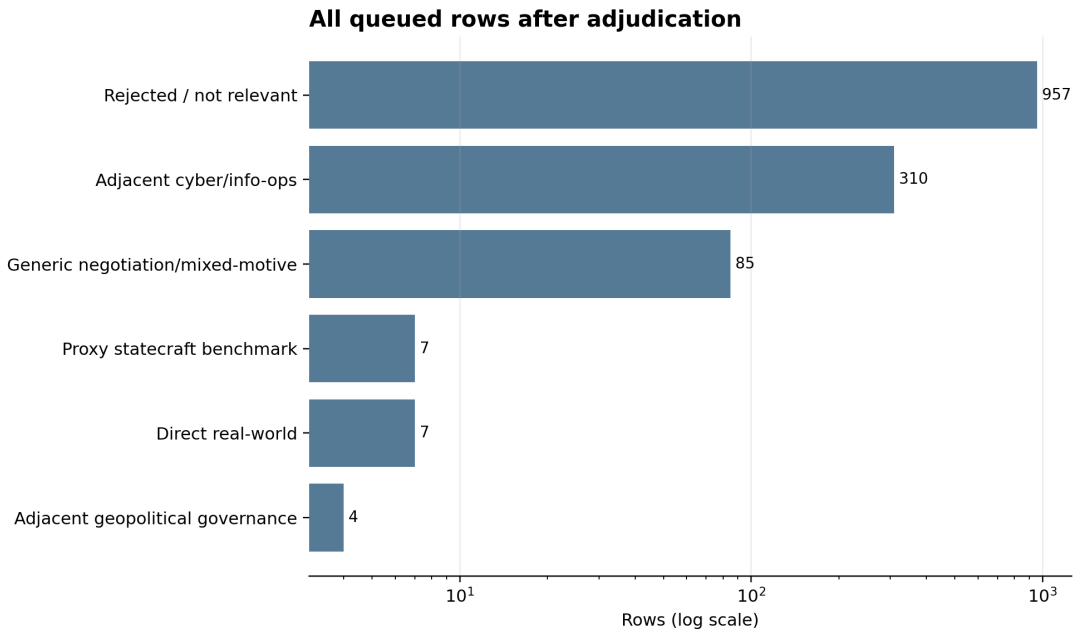

Figure 2. All adjudicated labels, including rejected or not-relevant rows.

domain-by-year, venue-by-year, domain-by-capacity, and domain-by-method cross-tabulations. The figure files used in this ICML-formatted appendix are included in the figures/ directory.

Title	Year	Venue	QA status
Not Oracles of the Battlefield: Safety Considerations for AI-Based Military Decision Support Systems	2024	AIES	Paper-ready strict
Human vs. Machine: Behavioral Differences between Expert Humans and Language Models in Wargame Simulations	2024	AIES	Paper-ready strict
Escalation Risks from Language Models in Military and Diplomatic Decision-Making	2024	FAccT	Paper-ready strict
This Land is Your, My Land: Evaluating Geopolitical Bias in Language Models through Territorial Disputes	2024	NAACL	Paper-ready strict
DiplomacyAgent: Do LLMs Balance Interests and Ethical Principles in International Events?	2025	EMNLP	Paper-ready strict
Human-Level Performance in No-Press Diplomacy via Equilibrium Search	2021	ICLR	Paper-ready strict
No-Press Diplomacy from Scratch	2021	NeurIPS	Paper-ready strict
Mastering the Game of No-Press Diplomacy via Human-Regularized Reinforcement Learning and Planning	2023	ICLR	Paper-ready strict
CivRealm: A Learning and Reasoning Odyssey in Civilization for Decision-Making Agents	2024	ICLR	Paper-ready strict
More Victories, Less Cooperation: Assessing Cicero's Diplomacy Play	2024	ACL	Paper-ready strict
DipLLM: Fine-Tuning LLM for Strategic Decision-making in Diplomacy	2025	ICML	Paper-ready strict
Democratizing Diplomacy: A Harness for Evaluating Any Large Language Model on Full-Press Diplomacy	2026	AAAI	Paper-ready strict
Knowledge-Embedded Narrative Construction from Open Source Intelligence	2023	AAAI	Borderline/contextual
A Survey of Large Language Model Use and Its Technical Limitations in Military Systems Through a Decolonial Lens	2025	AIES	Borderline/contextual

Table 7. Strict direct/proxy QA. Twelve rows are retained as conservative paper-ready strict evidence. Two additional direct-labeled rows are retained as contextual or borderline and excluded from the headline strict count.

Work ID	Title	Label	Confidence
56454	DevFD: Developmental Face Forgery Detection by Learning Shared and Orthogonal LoRA Subspaces	Adjacent cyber/info-ops	Low
10067	IOHunter: Graph Foundation Model to Uncover Online Information Operations	Adjacent cyber/info-ops	Low
32651	PPT-GNN: A Practical Pretrained Spatio-Temporal Graph Neural Network for Network Security	Adjacent cyber/info-ops	Low
5412	Tracking and Identifying International Propaganda and Influence Networks Online	Adjacent cyber/info-ops	Low

Table 8. Possible false negatives from the 500-row negative audit. All four are low-confidence adjacent cyber/info-ops rows; none are strict direct or proxy statecraft misses.

Method	Rows
Benchmark/dataset	202
Tool/platform	42
Empirical model evaluation	40
Red team/adversarial	39
Survey/taxonomy/position	32
Human/expert study	30
Policy/legal analysis	27
Other/unclear	1
Total	413

Table 9. Evaluation and governance methods among the 413 accepted domain+method rows.

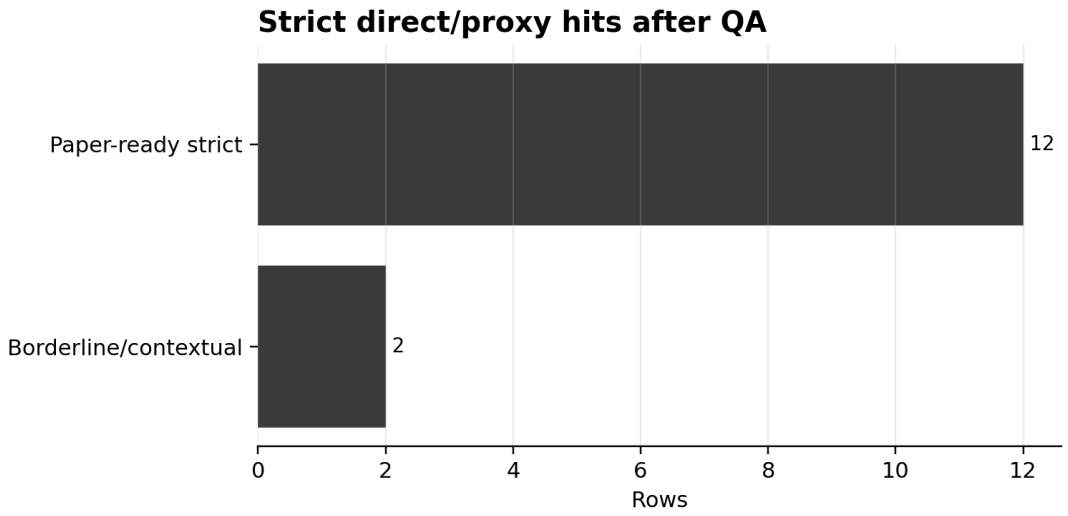

Figure 3. Manual QA split between paper-ready strict and borderline/contextual direct or proxy rows.

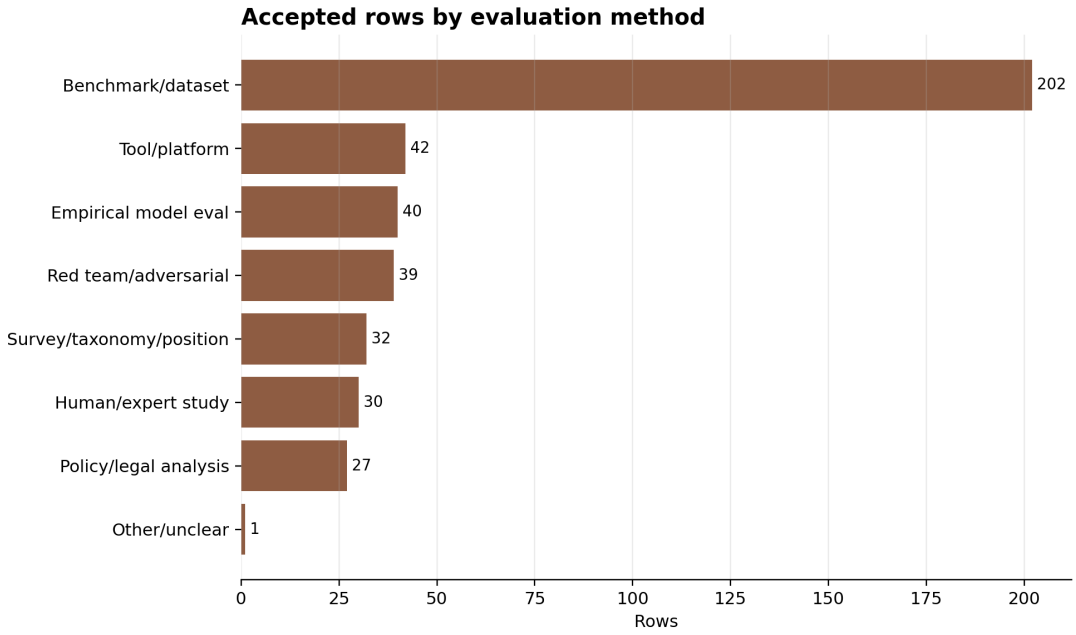

Figure 4. Accepted rows by evaluation or governance method.

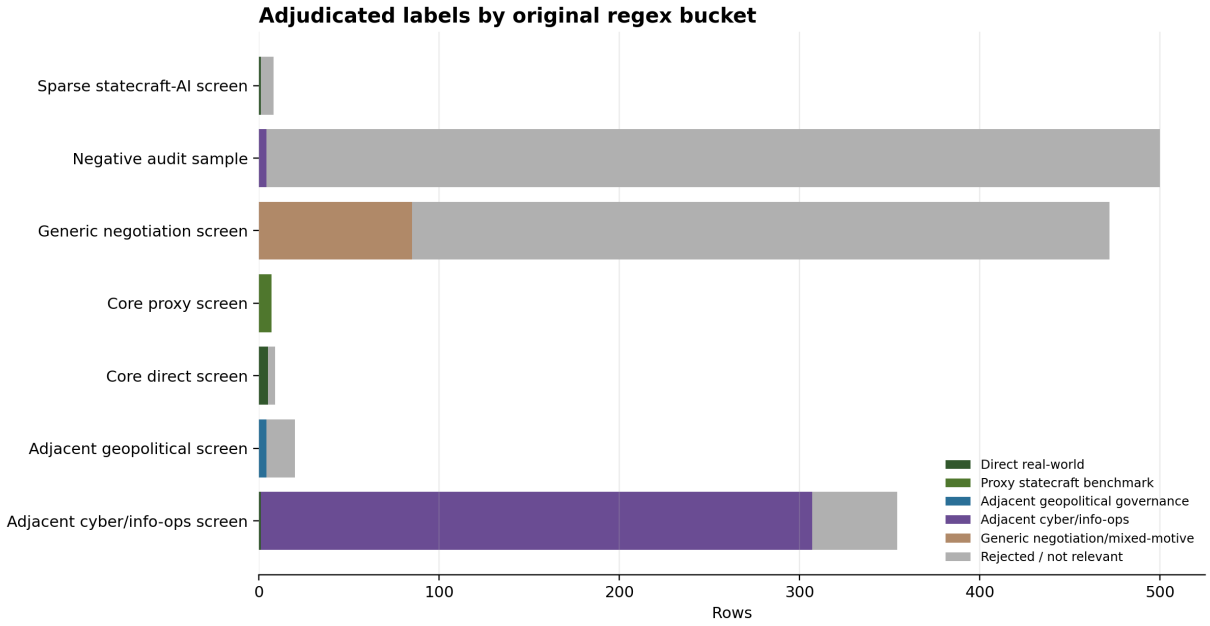

Figure 5. Movement from deterministic candidate buckets to final adjudicated labels.