From Russia with Influence? An AI-Driven Probabilistic Framework for Assessing Foreign Electoral Interference in U.S. Elections (2016–2036)

[Removed for Review]

Abstract

Concerns over foreign electoral interference have grown since the 2016 U.S. presidential election, yet public-facing intelligence assessments continue to rely on vague probabilistic language that limits clarity, consistency, and operational insight. This study introduces an exploratory AI-facilitated framework designed to systematically quantify the likelihood of foreign election interference across U.S. elections from 2016 to 2036. Drawing on declassified intelligence assessments from the ODNI, NIC, and CISA, corroborated by open-source intelligence (OSINT), we applied a three-phase natural language processing (NLP) protocol using OpenAI's tools to extract, classify, and scale linguistic indicators of confidence. These were then mapped to probabilistic values based on Sherman Kent's CIA estimative language and modeled using Monte Carlo simulations to account for uncertainty. Named Entity Recognition and sentiment analysis identified country-specific attribution patterns, while lexical scaling translated narrative judgments into quantifiable interference probabilities. Results revealed persistently high likelihoods of Russian interference, alongside growing probabilistic signals from China and Iran over time. A hierarchical linear model confirmed significant variation by election year and actor, and simulation-based forecasts suggest increasing probabilistic risk through 2036. This framework offers a replicable, data-driven model for transforming qualitative intelligence into structured probability distributions, providing analysts and policymakers with an evidence-based tool to track, compare, and forecast adversarial influence strategies with greater transparency and granularity.

Key Words: foreign electoral interference; probabilistic modeling; disinformation; natural language processing; U.S. elections

Introduction

Few periods in recent history have seen the concept of *foreign influence* receive as much sustained public and political attention as it does today (Davis Center, 2021). The last significant surge of interest in this topic could be dated back to the Cold War, particularly during *McCarthyism* (more broadly referred to as the Second Red Scare) when concerns about foreign ideological infiltration were at their peak (Storrs, 2014). Contemporary research engages with the topic of foreign influence in a range of ways. Some confront the issue head-on, focusing on the monitoring and analysis of foreign influence itself (e.g., Corstange & Marinov, 2012; Lawrence & Vandewalker, 2017; Palmer & Wilner, 2024). Others approach it more obliquely, through the study of disinformation - a related but broader phenomenon that at times overlaps with foreign interference (e.g., Bradshaw & Howard, 2018; Van Bavel et al., 2021). Still others examine

historical parallels, suggesting that current geopolitical tensions may mark the emergence of a second Cold War (e.g. Schindler et al., 2024). Across these approaches, foreign influence remains a central, if sometimes implicit, concern.

Foreign interference poses a possible threat to democratic systems due to its potential to influence electoral outcomes, distort political narratives, and erode public trust in democratic institutions (Palmer & Wilner, 2024; Posard et al., 2020). Researchers have shown that such interference often operates through the manipulation of political discourse, particularly via automated accounts and bots that amplify divisive content and foster polarization (Ferrara et al., 2020; Martin et al., 2019). This activity has been especially prominent in efforts by foreign actors such as Russia and China, who have strategically targeted fringe communities to intensify ideological divides (Kovalčíková & Spatafora, 2024; Ferrara et al., 2020). Network analyses of platforms like X (formerly known as Twitter) further illustrate the extent of this polarization. For instance, Garimella and Weber (2017) documented a 10–20% increase in online political polarization based on retweet patterns. While subsequent research suggests that the spread of explicitly false or extremely biased content has declined between 2016 and 2020, evidence also points to a rise in echo chamber behavior and expanding ideological divides among both users and key influencers during this period (Flamino et al., 2023).

We adopt the definition of foreign influence in line with the ICA guidelines, as efforts by foreign governments, non-state actors, or their proxies to covertly or coercively shape another nation's political perceptions, behaviors, or targeting democratic institutions in pursuit of strategic interests (NIC, 2022). However foreign interference refers to a narrower set of covert or unlawful actions which were undertaken either by or on behalf of a foreign government with the intent or effect of influencing electoral outcomes or eroding public trust in democratic institutions (DHS & DOJ, 2021).

We are living in an era characterized by not only unprecedented access to information, but also by an overwhelming presence of misinformation (Apuke et al., 2022). As individuals are regularly exposed to large volumes of content, the burden increasingly falls on the public to discern fact from fiction (Lewandowsky et al., 2012). This persistent challenge of discerning credibility parallels the analytical challenge faced by intelligence professionals: how to convey uncertainty in ways that are both precise and comprehensible. In both domains, the problem is not the absence of data, but the difficulty of communicating judgment under uncertainty. This parallel underpins the methodological focus of the present study. As such, this requires developing media literacy skills such as fact-checking, recognizing automated or AI-generated content and cultivating a healthy skepticism toward unvetted sources (Yang et al., 2024). However, the sheer volume of daily information makes it unrealistic for individuals to independently verify everything they encounter. Consequently, there is a growing need for accessible, well-structured reports that synthesize intelligence and disinformation findings in a transparent and digestible manner.

In the context of foreign interference, the United States has produced numerous official reports (e.g. CISA, 2022; ICA, 2017; NIC, 2022, NIC, 2024; Office of the Director of Central

Intelligence, 2019). Yet, the variation in language and presentation across these documents often makes them difficult for the general public to interpret, especially when distinguishing between confident assessments and speculative conclusions. In this regard, intelligence assessments frequently rely on probabilistic language rather than numerical estimates, which can obscure the degree of certainty behind claims (Dhami & Mandel, 2020).

In an environment where trust in democratic institutions is already in decline (falling from approximately 75% in 1958 to just 35% among Democratic-leaning individuals and to 11% among Republican-leaning individuals in recent years; Pew Research Center, 2024) this lack of clarity only increases public confusion and skepticism. In response, it is important to develop systems for communicating intelligence in ways that are both transparent and emotionally neutral. As Whyte (2024) argues, this includes avoiding sensationalism. More specifically, the kind of narrative constructed through a dramaturgy of scandal, secrecy, and partial revelations. In line with this a more grounded, probabilistic understanding of such events could help. Without such efforts, it is unsurprising that those who are most politically engaged often report being the most emotionally exhausted and angry (Pew Research Center, 2023).

Although foreign interference in democratic elections has long been a concern in international relations, public and institutional focus intensified following the 2016 U.S. presidential election (Badawy et al., 2018; Davis Center, 2021; Eady et al., 2023; Lawrence & Vandewalker, 2017; Rutenberg, 2022). Arguably contributing to the establishment of the Foreign Malign Influence Center (FMIC), a specialized body tasked with analyzing and countering foreign efforts to sway public opinion and disrupt democratic processes (FMI, 2024). However, the broader ecosystem responsible for monitoring foreign interference remains highly fragmented. The '*decentralization*' of data collection, monitoring, and analysis in the context of foreign interference is largely a consequence of the evolving nature of digital tools and platforms. As societies have become increasingly interconnected, the tactics and technologies used in influence campaigns have adapted accordingly, becoming more digitized, scalable, and sophisticated (Badawy et al., 2018; Ferrara et al., 2020).

One form of such manipulation - termed *Digital Infrastructure Exploitation* - relies heavily on the use of bots and coordinated inauthentic behavior to spread misinformation across multiple platforms, as exemplified by early incidents such as X and Google bombing campaigns (Metaxas & Mustafaraj, 2012). These efforts are rarely confined to a single platform, let alone a single user, making detection and response inherently complex. Numerous government agencies, including the Office of the Director of National Intelligence (ODNI), the Cybersecurity and Infrastructure Security Agency (CISA), and the National Intelligence Council (NIC), alongside private corporations (*e.g. Microsoft* (MTAC, 2024), *Facebook* (Facebook, 2021) *and many other Big Tech companies*) through investigative journalists (e.g., Gijn, 2024), and academic institutions (e.g., Stanford Internet Observatory) all of which contribute to identifying and reporting on such threats. Yet there is no centralized mechanism to integrate these insights or to standardize the language, methods, and thresholds used to assess foreign influence activities accessible to the public.

Verbal probability expressions (e.g., *likely*, *possible*, *unlikely*) remain one of my primary mechanisms for intelligence reporting, yet empirical research reveals they generate substantial communicative variance. For example, Dhami and Mandel (2021) document that analysts often assign highly divergent numerical equivalents to the same verbal term, sometimes diverging by 30-plus percentage points, undermining inter-analyst coherence and the reliability of downstream reasoning. Meanwhile, studies in judgment and decision science show that processes such as coherentization and aggregation can reduce absolute error by over 60 percent in probabilistic judgment tasks, even among trained intelligence analysts (Mander et al., 2018). These results suggest a latent precision hidden within qualitative judgments and the potential gains unlocked by imposing structure and calibration through probabilistic reasoning.

The computational turn in natural-language processing (NLP) has strengthened the case for espousing AI with probabilistic modeling as a means of improving analytic clarity and accountability. Recent advances in *confidence estimation* within NLP demonstrate how linguistic outputs can be calibrated through probabilistic reasoning rather than deterministic interpretation (Gandrabur et al., 2006; Guo et al., 2017; Wightman et al., 2023). In this context, the deployment of AI is not intended to supplant the human analytic process but to standardize and audit it. As Wightman et al. (2023) show, model agreement across semantically equivalent prompts can be treated as a quantifiable indicator of confidence, providing a reproducible metric of uncertainty. Likewise, newer frameworks such as Zhang et al. (2024) and Yang et al. (2024) propose calibrated, deliberative methods that enhance model interpretability and transparency, aligning closely with the demands of intelligence analysis. By using AI to structure linguistic ambiguity into measurable confidence intervals, this study treats NLP as an instrument for enhancing its consistency, traceability, and analytic verifiability. Indeed, the same qualitative reasoning that underpins expert judgment is here rendered reproducible through algorithmic calibration. In other words, by aggregating multiple model responses to semantically equivalent prompts, the framework captures the variance that would otherwise remain implicit in human reasoning. Thus, AI serves as a *methodological instrument* (i.e., a way of making linguistic ambiguity measurable and empirically comparable).

There is now a mature literature on confidence estimation in NLP (i.e., techniques by which machine models assign, calibrate, and validate probabilities of correctness to their outputs. see, Gandrabur et al., 2006; Wightman et al., 2023). For instance, Wightman et al (2023) demonstrated that model confidence can be empirically calibrated through *prompt-agreement*, a technique that evaluates how consistently a model responds to multiple, semantically equivalent prompts. Their findings across ten benchmark datasets show that this approach substantially reduced calibration error and misclassification of uncertainty compared to traditional single-prompt log-likelihood methods. This is particularly relevant to intelligence analysis, where linguistic variability and conditional phrasing are common. In other words, disagreement among prompts can be treated as a quantifiable indicator of uncertainty. By adopting a similar logic, using linguistic diversity and model agreement to bound confidence, the present study treats

probability as a distribution informed by the structure of language rather than a static value, aligning with best practice in calibrated decision modeling.

Further, the dual demands of analytic auditability and public intelligibility argue strongly for probabilistic outputs. The *National Intelligence Council's Analytic Standards* (NIC, 2021) and the *Office of the Director of National Intelligence's (ODNI) Intelligence Community Directive 203 (*see*, ICD-203, 2015)* both emphasize that analytic judgments must be traceable, transparent, and capable of review. ICD-203 explicitly requires that "*analytic products should indicate and explain the basis for the uncertainties associated with major analytic judgments, specifically the likelihood of occurrence of an event or development, and the analyst's confidence in the basis for this judgment*" (p. 4). The directive further mandates that "*degrees of likelihood encompass a full spectrum from remote to nearly certain*," and that consistency in the terms and logic used to express uncertainty is important for analytic integrity (see *Table 1*).

Table 1. Standardized Probability Ranges for Expressing Analytic Likelihoods (Adapted from ODNI ICD-203, 2015)

Verbal Expression	**Approximate Probability Range**
Almost no chance / Remote	**01 – 05 %**
Very unlikely	**05 – 20 %**
Roughly even chance	**45 – 55 %**
Likely / Probably	**55 – 80 %**
Very likely / Highly probable	**80 – 95 %**
Almost certain / Nearly certain	**95 – 99 %**

This approach, however, remains inherently vague in terms of derivation and explainability. Empirical work in risk communication and science journalism suggests the fragility of interpreting verbal probabilities independently; for example, Willems et al. (2020) find that even among statisticians - *and certainly among lay audiences* - interpretation of standard verbal phrases vary widely, undermining the assumption that mapped verbal bands carry uniform meaning. Interestingly, the study of Dutch probability phrases revealed large variability in the interpretation of verbal probability phrases, indicating that even a neutral context failed to constrain divergent readings (Willems et al., 2020).

Because ICD-203's method is prescriptive rather than mechanistic and lacks a structured, data-driven procedure to generate, calibrate, or update probability estimates, it does not ensure analytic verifiability or consistency over time. Quantified probability statements, by contrast, directly support these standards by enabling post-hoc validation and inter-analyst reliability testing (i.e., practices empirically shown to enhance forecast accuracy and accountability; Mandel et al., 2018). From a public-facing standpoint, both the U.K. Government Office for Science (GO-Science, 2020) and the U.S. National Academies of Sciences (2017) similarly advocate numerical expressions of uncertainty in national-security and risk communication to

improve interpretability and trust. These converging empirical and policy frameworks affirm that probabilistic outputs are now institutional necessities for analytic oversight and transparent communication alike.

Importantly, traditional qualitative analysis thrives on conditional nuance (e.g., the "yes, but" reasoning central to expert assessment). As Heuer (1999) argued, analysts often struggle to resist the pressure toward binary *yes-or-no* conclusions when uncertainty is high, a tendency that oversimplifies judgments. Similarly, Padilla et al. (2021) demonstrated that qualitative confidence expressions can lead to inconsistent interpretations among decision-makers, reinforcing the need for calibrated probabilistic communication. In that regard it is important to consider a framework that preserve this nuance by converting conditional linguistic variation into a probability distribution rather than a single fixed value. Divergence across model responses can then be treated as evidence of uncertainty, not error, reflecting rather than erasing analytic ambiguity.

Within intelligence practice, numeric scaling permits retrospective benchmarking, analytic drift detection over election cycles, and cross-case comparability, addressing a gap noted in traditional treatment of foreign interference. On the communication side, experimental work in risk and political communication suggests that when uncertainty is expressed numerically (i.e., as ranges or probabilities rather than opaque language) readers (including non-experts) report greater clarity, trust, and accuracy in assessing statements under uncertainty (e.g. Van Bavel et al., 2021). Though not yet pervasive in the intelligence domain, this body of work suggests that quantitative uncertainty statements meaningfully improve interpretability and reduce misreading of confidence claims in contentious contexts.

Despite the use of qualitative intelligence assessments, no standardized, reproducible method currently exists for quantifying the probabilistic confidence underlying such judgments. Existing frameworks, including those established by the Office of the Director of National Intelligence (ODNI) and the Professional Head of Intelligence Assessment (PHIA), prescribe consistency in terminology but do not specify how probability estimates should be empirically derived or validated. This gap limits both analytic reliability and the communicative transparency of intelligence reporting. As a result, key questions about the evolving nature, probability, and comparative tactics of interference efforts across election cycles remain unanswered. This study develops and tests an AI-driven probabilistic framework that integrates these sources, addressing existing gaps. Drawing on intelligence assessments and open-source intelligence (OSINT), we transform narrative accounts of foreign election interference into structured probability estimates. This allows for a more systematic, data-driven understanding of how these efforts may have evolved over time and how they differ. Accordingly, this current research (1) provides an overview of how foreign influence strategies have evolved across three recent U.S. presidential elections, (2) compares the tactics employed by key foreign actors, and (3) introduces an explorative OpenAI probabilistic framework for quantifying influence using both qualitative intelligence assessments and quantitative analysis.

Note, the use of election interference as a test case does not reflect a substantive focus on electoral politics *per se*, but for the present study, serves as a proof-of-concept to evaluate how probabilistic reasoning and AI-facilitated linguistic modeling can be applied to real-world intelligence problems. Election interference offers an optimal setting for methodological validation as it is (i) empirically well-documented, conceptually bounded, and supported by a robust corpus of declassified assessments (e.g., ODNI, 2017; NIC, 2022; CISA, 2022) and open-source intelligence datasets (e.g., Leite et al., 2024) that capture linguistic expressions of analytic confidence. Moreover, the domain has been studied extensively through both computational and behavioural methodologies, providing a strong comparator for intelligence assessment and uncertainty (e.g., Badawy et al., 2018;; Eady et al. 2023; Ferrara et al. (2020; Mandel, 2020; Mandel & Irwin, 2021).

Methodology

Data Selection and Intelligence Report Processes

The *Office of the Director of National Intelligence* (ODNI) Report from 2017 was identified as a primary source for assessing Russian influence in the 2016 U.S. presidential election. The report explicitly concluded that Russia's influence campaign sought to help the then President-elect Donald Trump's election chances, stating that Putin and the Russian Government developed a clear preference for Trump. The assessment relied on SIGINT, HUMINT and cyber forensic data to establish Russia's electoral influence with a high degree of confidence. The CISA report from 2021 provided a more expansive assessment of electoral interference in the 2020 election, concluding that Russian influence operations had evolved from direct candidate support to broader efforts aimed at undermining confidence in the electoral process. The report also identified growing Chinese and Iranian influence activities, with China focusing on amplifying divisions within U.S. political discourse while Iran sought to discredit Trump's administration through misinformation campaigns. The National Intelligence Council (NIC) Global Trends Report (e.g., NIC, 2021) and ODNI reports (e.g., ODNI, 2024) further indicated that Russian electoral influence efforts had become more focused on eroding trust in democratic institutions rather than promoting a specific candidate.

To complement these official intelligence assessments, independent OSINT investigations were reviewed to validate and cross-check intelligence findings. For instance, The Stanford Internet Observatory's analysis of Chinese influence efforts during the 2016 election provided additional insights into state-backed narrative manipulation (Diresta et al., 2020). Unlike Russia's explicit candidate preference, Chinese electoral influence focused primarily on shaping economic narratives around U.S.-China trade relations rather than attempting to sway voter choices directly. The Graphika Disinformation Report (2024) documented AI-driven social media influence campaigns operated by Chinese actors, with a primary objective of exacerbating partisan divides rather than directly supporting a candidate. Mandiant's Cyber Threat Report (2024) also revealed Iranian disinformation operations, which included coordinated fake social

media personas that disseminated anti-Trump rhetoric and sought to manipulate U.S. public discourse regarding Iran's nuclear program and regional policies.

It is important to note that the present study did not aim to establish cross-model replicability across different large language models (LLMs), such as Google BERT or Gemini, but rather to evaluate the conceptual validity of an AI-assisted probabilistic framework for intelligence interpretation. OpenAI's NLP architecture was selected for its robust linguistic-probability mapping capabilities and for exploratory alignment with intelligence-style analytic phrasing. The model's outputs should therefore be understood as *framework-specific demonstrations* rather than generalizable benchmarks. In keeping with this methodological focus, the OSINT sources used in the analysis were illustrative rather than exhaustive. They were bounded by three criteria: (i) the accessibility and verifiability of materials within the open domain, (ii) the requirement for texts containing probabilistic or confidence-laden language suitable for linguistic probability extraction, and (iii) the processing and token limitations of the OpenAI NLP system, which necessitated clear English-language narrative structures and finite text length for reliable *Named Entity Recognition* and sentiment calibration. Consequently, the dataset represents a theoretically sufficient but practically bounded corpus, designed to test how probabilistic reasoning can be operationalized within AI-mediated intelligence assessment rather than to capture the full account of declassified or open-source materials. This approach aligns with current recommendations in computational social science, which emphasize bounded demonstrations as a valid pathway for methodological validation in AI-augmented analytic frameworks (Lazer et al., 2020; Zhang et al., 2024).

Data Interpretation

The collected intelligence reports were analyzed to extract key indicators of electoral interference. Since intelligence assessments often contain probabilistic language rather than explicit numerical certainty values (Dhami & Mandel, 2020), it was necessary to develop an interpretive framework that standardized these statements into measurable probability estimates. Each document was reviewed to identify explicit statements about foreign influence, as well as implicit cues regarding the scale and intent of interference activities. For example, the ODNI Report from 2017 explicitly stated that Russia's influence campaign sought to help President-elect Donald Trump's election chances (e.g., "*We assess with high confidence that Russian President Vladimir Putin ordered an influence campaign in 2016 aimed at the US presidential election*" ODNI, 2017, pp.1). This statement, which conveyed a high degree of confidence regarding Russian electoral interference, warranted a near-certainty probability assignment. In contrast, the Graphika Report (2024) noted that Chinese state-affiliated actors engaged in social media manipulation to exacerbate partisan divides within the United States. (e.g., *In the run-up to the 2024 election, these accounts have seeded and amplified content denigrating Democratic and Republican candidates, sowing doubt in the legitimacy of the U.S. electoral process, and spreading divisive narratives about sensitive social issues including gun control, homelessness, drug abuse, racial inequality, and the Israel-Hamas conflict*) (p.1). While this statement

confirmed the existence of an influence operation, it did not indicate a direct attempt to manipulate voter behavior, leading to a lower probability assignment.

Statements that contained conditional language or ambiguous phrasing (e.g., *The IC continues to assess that Russia poses the most active foreign influence threat to this year's U.S. elections*) (ODNI, 2024) were assigned probability estimates based on contextual factors. If an intelligence report suggested that an influence campaign was *likely* but did not provide definitive supporting evidence, the probability was adjusted to reflect the moderate confidence level. Note, reports that cited multiple corroborating sources, were given higher certainty values. For instance, the 2023 Joint DOJ/DHS EO 13848 Report notes repeated cyber engagement with political campaign infrastructure by Russian actors, although it concluded that no material compromise of voting systems occurred (See, DHS, 2023).

Natural Language Processing (NLP) for Probabilistic Scaling of Intelligence Assessments

To transform qualitative intelligence assessments into structured probability estimates, an NLP-driven probabilistic scaling model was developed and tested. Capelli et al., (2024) demonstrated that LLMs can extract and quantify linguistic markers indicative of confidence levels, effectively aligning with human expert assessments in specialized social science contexts. As such, OpenAI's NLP capabilities were used to extract and quantify linguistic markers that signaled intelligence confidence levels. In doing so, this helped to estimate intelligence-derived assessments that could be mapped to numerical probability values consistently across different sources. For clarity, it was the purpose of this exploratory approach to evaluate how the NLP model processed intelligence reports using a multi-step probability mapping framework. First, Named Entity Recognition (NER) was used within OpenAI's NLP framework to identify key geopolitical actors (Kopanov, 2024) such as Russia, China, and Iran by scanning intelligence reports and OSINT sources for explicit mentions of state actors, government-affiliated agencies, cyber units, and media entities linked to election interference activities. OpenAI's NER model was preliminarily trained on intelligence-specific corpora to distinguish between *generic geopolitical references* and *entities* directly associated with electoral influence operations. It was believed that this would allow for the extraction of both direct attributions (e.g., Russian military intelligence, GRU, conducted cyber intrusions) and indirect mentions (e.g., a state-backed actor linked to China's Ministry of State Security).

Secondly, *Sentiment Analysis* was applied within OpenAI's NLP framework to evaluate the certainty levels expressed in intelligence assessments, with higher sentiment scores indicating stronger confidence in foreign influence conclusions. In early models, OpenAI's sentiment analysis model was fine-tuned on intelligence-specific datasets to detect linguistic markers of confidence, ambiguity, and conditional phrasing within the reports. This approach allowed the system to differentiate between definitive intelligence conclusions (e.g., *Russia conducted a coordinated disinformation campaign)* and more uncertain assessments (e.g., *It is possible that Chinese state actors attempted to exploit online narratives*). The sentiment scores were then

mapped to theoretically weighted probability estimates, so that stronger intelligence confidence correlated with higher probabilistic weightings in the Monte Carlo simulation model.

Finally, Lexical probability scaling was applied within OpenAI's NLP framework to intelligence-derived probabilistic expressions, mapping them to numerical values based on a standardized probability framework. OpenAI's language model was specifically trained on intelligence assessments and probabilistic phrasing to systematically extract and classify statements that conveyed varying levels of certainty regarding foreign electoral influence (e.g., probability yardstick). This approach involved structuring prompts that directed the model to recognize intelligence-derived probabilistic expressions, apply contextual weighting, and assign standardized probability values in accordance with Sherman Kent's CIA Estimative Language Scale (Kent, 1964). See Table 2 for an exemplar of the prompt engineering used to evaluate probabilistic expressions.

Table 2. Prompt Engineering Aligned with Sherman Kent's Probability Framework

Sherman Kent's Verbal Estimate	**Approximate Probability (%)**	**Prompt Engineering Strategy**	**Example Prompt Phrase**
Certain	100%	Use of definitive and factual language	"What activities are conclusively attributed to Russia?"
Almost Certain	93–99%	Use of strong confidence indicators	"What is almost certainly true based on this excerpt?"
Probable	75–85%	Moderate certainty cues, supported inference	"Which statements are most probably indicating Chinese disinformation?"
Chances About Even	45–55%	Equivocal or balanced phrasing	"Does the evidence support even odds of Iranian involvement?"
Probably Not	15–25%	Use of negation with weak evidence cues	"Which influence activity is probably not associated with China?"
Almost Certainly Not	1–7%	Strong negation, counterevidence	"What actors are almost certainly not implicated in the excerpt?"
Impossible	0%	Absolute negation	"Are there any activities that are described as impossible?"

It is important to highlight that one of the key challenges in intelligence assessment is that certainty statements are often nuanced. For example, a report stating, "*It is likely that Russian state-backed actors interfered in the election"* (ODNI, 2017) conveys a different confidence level than, *Russia actively deployed disinformation tactics in a coordinated campaign."* (ODNI, 2017). See Table 3 for an overview. The trained models for the purpose of this study tuned OpenAI's NLP tools to recognize these differences in wording and intent, assigning higher probability values to more definitive statements and lower values to those with mitigating language such as *possibly, reportedly,* or *there is some evidence to suggest.* Additionally, the model corroborated probability estimates across multiple intelligence sources, cross-referencing reports from ODNI, CISA, Graphika, and other organizations. When a claim appeared in multiple independently verified sources, its probability assignment was adjusted upward. Conversely, if a claim was found in only one source, especially if expressed with ambiguous language, the probability was weighted lower to account for uncertainty. Historical reliabilities were also factored into probability adjustments.

The probability assignment was structured using the following NLP-driven probabilistic function,

$$P(E) = f(S) + \epsilon$$

where, P(E) represents the probability of foreign electoral interference, f(S) is the NLP-derived probability estimate, calculated based on linguistic certainty indicators within intelligence assessments, and ϵ represents the uncertainty adjustment term[1].

Table 3. Example data entry

Election Year	**Country**	**Influence probability score**	**Likelihood Category**	**Source**	**Probability Calculation (*P, SD*)**	**Intelligence Source**
2016	Russia	0.949	7(Almost Certain)	ODNI Report (2017)	N(0.95, 0.03)	Putin and the Russian Government developed a clear preference for Trump
2020	China	0.500	4 (Even chance)	Graphika 2024)	N(0.50, 0.05)	Chinese state-affiliated actors engaged in

[1] The error term ϵ represented the uncertainty adjustment term in the probabilistic model. It accounted for noise introduced during the transformation of qualitative intelligence statements into quantitative estimates. However, in the absence of formal qualitative coding or inter-rater validation, some ambiguity remains regarding the source and structure of this noise, particularly given the variability in language used across intelligence reports.

						social media manipulation to exacerbate partisan divides.
2024	Iran	0.650	5 (Likely)	AP (2024)	N(0.54, 0.04)	Iran used deepfake technology to spread misleading information targeting U.S. policymakers

Monte Carlo Simulation for Probability Distribution Modeling

A Monte Carlo simulation provided a mathematically transparent means of propagating uncertainty through repeated stochastic sampling, thereby reflecting the inherent probabilistic nature of intelligence judgments. Unlike single-point or deterministic probability estimates, Monte Carlo methods generate distributions that approximate the range of plausible analytic outcomes given underlying variance in the data. For instance, Binkowitz et al. (2001) posits that Monte Carlo methods statistically combine individual parameter distributions to yield a comprehensive output distribution. As such these methods offer three advantages: (i) they utilize all available information about variability and uncertainty, (ii) they reveal compounded conservatisms in traditional assessments, and (ii) they re-establish the boundary between risk assessment and risk management. This is particularly suited to intelligence contexts, where judgments are contingent on incomplete, sometimes ambiguous evidence, and confidence must be expressed as a continuum rather than a fixed value. By iteratively sampling from Gaussian functions calibrated to NLP-derived probability estimates, the model captures both the central tendency of analytic confidence and its dispersion across repeated draws, producing empirically interpretable confidence intervals. This stochastic approach aligns with best practices in uncertainty quantification used in national-security risk modeling (National Research Council, 2012; Saltelli et al., 2008) and with the ODNI's analytic standard emphasizing traceable and reviewable expressions of likelihood (e.g., ICD-203, 2015).

Given that, probability values were assigned using NLP-based probability scaling, a Monte Carlo simulation framework was implemented to introduce uncertainty modeling. Intelligence assessments inherently contain degrees of uncertainty due to variations in data collection methods, classified intelligence sources, and geopolitical biases. The Monte Carlo simulation allowed for the introduction of probability variance, ensuring that final probability values reflected confidence intervals rather than static single-point estimates. For each country-election year pair, such as Russia in 2016 or China in 2020, the probability of electoral influence

was modeled as a Gaussian-distributed function[2]. For example, in the Russia 2016 dataset, the ODNI report assigned a probability estimate of 0.95. To account for possible deviations in intelligence confidence levels, a standard deviation of 0.03 was applied[3], leading to a probability distribution defined as,

$$P(Russia_{2016}) \sim N(0.95, 0.03)$$

Each Monte Carlo simulation iteration generated a randomized probability draw from this distribution, allowing for the construction of confidence intervals that captured intelligence uncertainty.

Results

First, the results presented should be interpreted primarily as a methodological demonstration rather than a substantive measure of the true extent of foreign interference. The election-interference corpus was selected as a proof-of-concept environment to evaluate the functionality, calibration, and reliability of the probabilistic-linguistic framework developed in this study. The resulting probabilities therefore illustrate how the model translates qualitative intelligence judgments into quantifiable confidence distributions, rather than serving as direct estimates of interference magnitude.

The Monte Carlo simulation results provide probabilistic estimates of foreign election interference in the 2016, 2020, and 2024 U.S. presidential elections. This simulation was based on 10,000 entry points per country per election year (N=90,000), with probability estimates derived from automated coding of intelligence reports and OSINT sources. In this regard, this allowed for the calculation of confidence intervals that captured variability in assessments, providing a probability distribution for election interference across the three election cycles. Note, this was repeated to project distributions in the 2028, 2032, and 2036 electoral cycles.

Monte Carlo Simulated Foreign Influence Probabilities

The results of the Monte Carlo simulation are presented in Table 4, including mean probability estimates and 95% confidence intervals (CIs) for Russia, China, and Iran.

Table 4. *Monte Carlo Simulated Foreign Influence Probabilities with 95% Confidence Intervals*

[2] The Gaussian distribution was selected to reflect uncertainty around point estimates of influence probability while assuming a central tendency based on linguistic cues from intelligence assessments. The use of a bell-shaped distribution enabled probabilistic inference under the assumption that most interpretations cluster around a central value, with diminishing likelihoods at the extremes. Note, this is consistent with Bayesian approaches to modeling subjective judgments (Goldstein, 2006).

[3] Mandel and Barnes (2014) report a calibration index of 0.016 for strategic intelligence forecasts, implying an average deviation of ~12.6%, though variation is assumed lower for high-confidence estimates. *SD* = 0.03 provides a conservative approximation of uncertainty around such estimates.

Election Year	Russia Mean	Russia CI (2.5 - 97.5%)	China Mean	China CI (2.5 - 97.5%)	Iran Mean	Iran CI (2.5 - 97.5%)
2016	0.949	(0.830, 1.000)	0.299	(0.105, 0.476)	0.4	(0.236, 0.562)
2020	0.9	(0.774, 1.000)	0.5	(0.291, 0.694)	0.599	(0.426, 0.751)
2024	0.949	(0.834, 1.000)	0.55	(0.350, 0.728)	0.65	(0.497, 0.812)

When applied to the test dataset, the model generated high probability outputs for Russia across all three election cycles, showing the model's sensitivity to linguistic cues of analytic confidence. These therefore results reinforce prior intelligence findings that Russia maintained a consistently high probability of election interference across all three election cycles, with a near certainty of involvement in both 2016 and 2024 $P(E_{2016}) = 0.949, P(E_{2020}) = 0.900, P(E_{2024}) = 0.949$.

Variation in Foreign Election Interference Across Election Cycles

A one-way ANOVA was conducted separately for Russia, China, and Iran to determine whether the model's assigned interference probabilities varied significantly across the 2016, 2020, and 2024 election cycles. The results indicated a statistically significant effect of year for all three countries: Russia, $F(2, 89996) = 9456.31$, $p < .001$; China, $F(2, 89996) = 70151.83$, $p < .001$; and Iran, $F(2, 89996) = 109721.76$, $p < .001$. These findings indicate that the model detects meaningful variance in probabilistic linguistic expressions of analytic confidence across time, demonstrating its sensitivity to contextual and temporal shifts in source language.

To further examine these temporal dynamics, a Tukey's HSD post hoc analysis was conducted for each country. The results confirmed that the model assigned significantly lower probability estimates to Russian influence in 2020 compared with 2016 and 2024 ($p < .001$), reflecting a temporary reduction in the intensity of confidence cues within that year's analytic corpus. Similarly, model outputs showed a progressive increase in China's probabilistic confidence scores from 2016 to 2024 ($p < .001$), and a comparable upward trend for Iran ($p < .001$). Note, these results should not be interpreted as direct measurements of real-world interference, but as evidence of the framework's ability to capture and quantify linguistic variation in the expression of analytic confidence across intelligence narratives.

Temporal Trends and Cross-National Variations in Election Interference

To assess whether the probabilistic framework systematically differentiated linguistic expressions of analytic confidence across time and national contexts, a hierarchical linear model was applied. This approach was selected to account for the nested data structure, where observations were grouped within countries, creating potential dependencies that could violate the assumption of independent observations (Mertens et al., 2016). Prior to estimating the full model, a random-intercept-only specification was tested to evaluate whether model-derived probabilities exhibited meaningful variance at the country level. The resulting intraclass

correlation coefficient (ICC) indicated that a substantial proportion of variance in model-assigned probability scores was attributable to country-level linguistic and contextual differences. The estimated country-level variance component ($\sigma^2 = 0.033$) confirmed that the framework detected consistent cross-national variation in how intelligence assessments express confidence, supporting the inclusion of country as a random effect in subsequent models (Snijders & Bosker, 2011).

The final hierarchical mixed-effects model, estimated using Restricted Maximum Likelihood, demonstrated a strong fit to the simulated dataset (log-likelihood = 110748.3). The model explained 90.5% of the variance ($R^2 = .905$) in model-derived probability outputs, showing that the inclusion of year as a fixed effect and country as a random effect accounted for the majority of structured linguistic variability in the corpus. The fixed effect of year on the model's probability assignments was statistically significant ($\beta = 0.021$, $z = 289.455$, $p < .001$), highlighting that the probabilistic framework effectively captured temporal shifts in confidence expression across intelligence narratives, rather than measuring direct fluctuations in real-world interference activity.

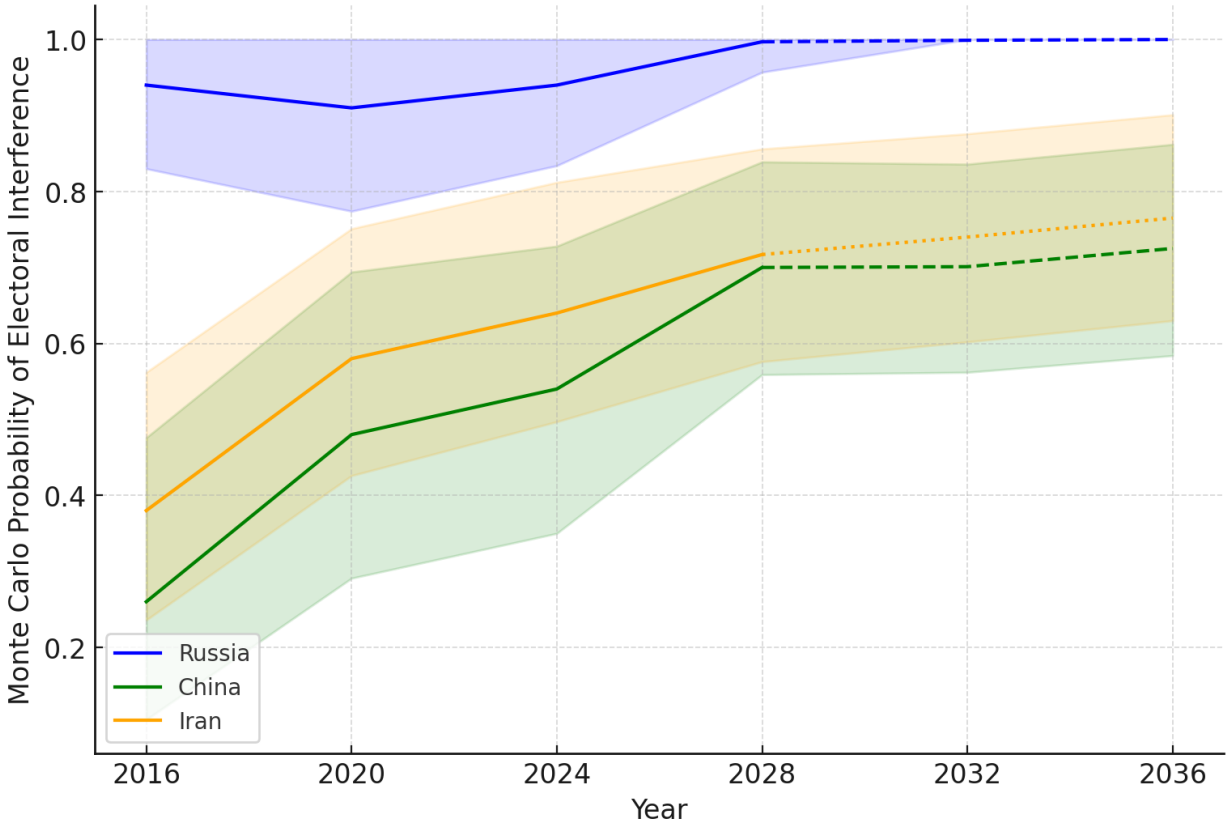
Forecasting Election Interference: Monte Carlo Simulations for 2028, 2032, and 2036

A Monte Carlo simulation was conducted to estimate the probability distributions for election interference in the 2028, 2032, and 2036 U.S. elections. The simulation iterated 10,000 times per country per election year to account for uncertainty in intelligence estimates and historical data trends. The Monte Carlo approach was employed to incorporate randomness and uncertainty, ensuring that the final probability distributions reflected real-world variability in election interference efforts. The probability of interference was extrapolated using fitted regression models from 2016-2024 data. Specifically, a linear regression model was applied to each country's probability trend, with year as the independent variable ($\beta_{\text{Russia}} = 0.0209$, $\beta_{\text{China}} = 0.0172$, $\beta_{\text{Iran}} = 0.0154$). Note, residual standard deviations from the regression models were used to introduce realistic variability into simulated probabilities. The results of the Monte Carlo simulation are presented in *Table 5*. Note, projected probabilities are visualized in *Figure 1*, which include Monte Carlo-derived confidence intervals for foreign election interference from 2016 to 2036.

Table 5 *Projected Election Interference Probabilities (2028-2036) with 95% Confidence Intervals*

	Russia (Mean, 95% CI)	China (Mean, 95% CI)	Iran (Mean, 95% CI)
2028	0.997 [0.957, 1.000]	0.700 [0.559, 0.839]	0.717 [0.576, 0.856]
2032	0.999 [1.000, 1.000]	0.701 [0.562, 0.836]	0.740 [0.602, 0.876]
2036	1.000 [1.000, 1.000]	0.725 [0.584, 0.862]	0.765 [0.630, 0.901]

Figure 1. *Projected Probabilities of Foreign Electoral Interference (2016–2036) with Monte Carlo Confidence Intervals*



Discussion

This study set out to transform qualitative intelligence assessments of foreign electoral interference into structured, probabilistic estimates, using AI to bridge the gap between narrative judgment and quantitative analysis. Drawing on declassified intelligence reports and OSINT the research aimed to (i) provide an overview of how foreign influence strategies have evolved across three recent U.S. presidential elections, (ii) compare the tactics employed by key foreign actors, and (iii) introduce and validate an exploratory OpenAI-based probabilistic framework for quantifying influence through natural-language processing. In doing so, the study sought not only to evaluate the utility of AI as a methodological tool for standardizing analytic confidence, but also to contextualize why patterns of electoral interference persist and adapt over time. The results supported these aims, showing that probabilistic modeling can capture meaningful variation across countries and election cycles, suggesting structured patterns in the evolution and communication of influence activity that traditional qualitative approaches often obscure.

While the persistence of foreign electoral interference, particularly by Russia, may not in itself be unexpected, it is important to recognize that the central aim of this study was not to produce novel geopolitical findings but to advance and validate a methodological framework for quantifying qualitative intelligence assessments. Specifically, the research sought to develop an AI-assisted probabilistic model capable of transforming narrative intelligence judgments into structured, comparable probability estimates. By applying this framework to three recent U.S. presidential election cycles, the study addressed a long-standing limitation in intelligence

communication: the absence of standardized, data-driven methods for expressing analytic confidence and uncertainty. In doing so, it aimed to demonstrate how probabilistic reasoning can enhance the transparency, consistency, and interpretive reliability of intelligence outputs. Although the findings confirm the continued activity of Russian, Chinese, and Iranian influence operations, their value lies primarily in showing that such operations can be systematically modeled over time, revealing distinct temporal and strategic trajectories that traditional qualitative approaches often obscure.

The findings suggested that foreign electoral interference is not a static or episodic threat, nor one that can be meaningfully understood through anecdotal observation or post hoc narrative alone (e.g., Mohan & Wall, 2019). Instead, the evidence presented here suggested that the architecture of influence remains dynamic and adaptive (e.g., Goldstein et al., 2023), likely evolving in parallel with advances in digital technologies (Starbird et al., 2019), shifting geopolitical alignments (Kennedy, 2022), and the iterative learning of adversarial actors (e.g., Badawy et al., 2019). Authoritarian regimes, in particular, have demonstrated increasing sophistication in their deployment of digital influence operations (e.g., through networked platforms to shape political narratives, exploit social and ideological vulnerabilities, and undermine democratic institutions; Kalathil, 2020). Importantly, the regimes identified in the current study not only refine their own tactics over time but potentially draw upon one another's strategies, creating a form of transnational authoritarian learning (Hall & Ambrosio, 2017; Tsourapas, 2021).

Russia exhibited consistently high probabilities of electoral interference across all three election cycles, with only a slight decrease in 2020. This dip, however, while statistically significant, should be interpreted cautiously. It is unlikely to signal a strategic withdrawal or de-escalation, but rather a tactical recalibration. It is reasonable to assume that this is potentially in response to increased counterintelligence efforts or operational saturation following heightened scrutiny in the aftermath of the 2016 election. According to Galeotti (2016), Russian information warfare is rooted in the tradition of *non-linear warfare*, wherein disinformation, cyber operations, and political influence are deployed in adaptive cycles in response to shifts in geopolitical and operational environments. The temporary reduction in observable activity in 2020 likely reflects Russia's recognition of heightened counterintelligence efforts by the U.S. government, social media platforms, and civil society organizations in the wake of the highly publicized 2016 interference (Ferrara et al., 2020; ODNI, 2017). Indeed, Starbird et al. (2019) and Broniatowski et al. (2020) highlighted that after the exposure of centralized operations by the Internet Research Agency (IRA) in 2016, subsequent Russian campaigns became more decentralized, covert, and culturally embedded. This shift included outsourcing operations to third-country operators, who relied on more credible persona management (e.g., AI-generated profile photos), and inserting narratives within fringe domestic media ecosystems (i.e., tactics that reduced attribution risk and increased resilience against detection; Graphika, 2023).[4]

[4] Note, the Gerasimov Doctrine: "*In the 21st century, we are seeing a tendency toward the blurring of the lines between the states of war and peace. Wars are no longer declared, and, once begun, proceed according to an*

At the same time, it is important to acknowledge that empirical evidence on audience effects remains contested. Grinberg et al. (2019), analyzed millions of tweets during the 2016 U.S. election, and found that exposure to fake-news content was both limited in scale and heavily concentrated within ideologically homogenous echo chambers. While this does not diminish the strategic intent of such operations, it suggests that the behavioural and attitudinal impacts of disinformation may be less widespread than often assumed. From this perspective, the value of probabilistic modeling lies not in inflating perceived influence but in providing a transparent and replicable means of assessing its likelihood and evolution over time. It is also worth noting, however, that the 2020 elections likely coincided with internal matters, such as the Russian Constitutional Referendum, which allowed Putin to remain in power until 2036 (Partlett, 2021). This may have shifted the political focus inward, using domestic media and digital operations to manage public opinion.

The near-perfect confidence interval observed for Russian interference in 2024 (0.999 [1.000, 1.000]) appeared to reflect model convergence rather than absolute certainty in the underlying intelligence assessment. In probabilistic terms, this outcome occurs when repeated Monte Carlo iterations yield highly consistent posterior estimates with minimal residual variance across the sampled distributions. Similar convergence phenomena have been reported in simulation-based inference when the evidence base is both internally coherent and overwhelmingly one-sided, producing degenerate posterior intervals that approach unity (Gelman et al., 2014; Robert & Casella, 2010). In this context, the narrow CI does not imply that interference is literally "certain," but that the available intelligence and open-source data provided no meaningful contradictory indicators across simulations. From a statistical standpoint, such convergence represents an asymptotic property of Bayesian-style resampling under homogenous inputs rather than a claim of factual absoluteness. Nevertheless, within intelligence studies, where uncertainty is axiomatic, these results must still be interpreted within the pragmatic boundaries of analytic judgment and contextual caveats rather than as definitive truth claims.

The return to near-certainty of Russian influence in 2024 hypotheses that Russian influence operations have become an institutionalized instrument of hybrid foreign policy, rather than an episodic or event-contingent strategy. Recent analyses have shown how Russia has relied on influence operations within its broader foreign policy framework. For instance, the *Concept of the Foreign Policy of the Russian Federation*[5] (2023) emphasized the use of information campaigns to achieve geopolitical objectives, signaling a formal adoption of such tactics for state policy. The European Parliament (2024) has since documented this systematic pattern of Russian

unfamiliar template. The role of non-military means of achieving political and strategic goals has grown, and, in many cases, they have exceeded the power of force of weapons in their effectiveness." General Valery Gerasimov (2013). Also see, Thomas (2016).

[5] See Information support for the foreign policy of the Russian Federation (Point 48).

interference[6] that includes coordinated disinformation campaigns, cyber intrusions, and covert financial support to pro-Russian political actors within the EU.

Rid (2020), characterized these operations as *active measures* that have migrated from traditional Cold War espionage into the digital ecosystem, where they exploit algorithmic amplification and information silos (also see, Giles, 2019). Complementary to this the Carnegie Endowment for International Peace (Momtaz, 2024) notes that these campaigns are intentionally diffuse and adaptive, combining psychological operations with emerging technologies such as AI-generated content to undermine societal resilience. The regularity and sophistication of these influence efforts indicate not episodic opportunism but a long-term commitment to what Tsygankov (2022) terms *norm-shaping through disruption*, a strategic use of information operations to challenge Western political norms and values while asserting Russia's own epistemic authority in global discourse.

In contrast, the data indicated that China and Iran have pursued more incremental, yet no less consequential, expansions of their influence activity. China's increase from a 0.299 mean probability in 2016 to 0.55 by 2024 is characteristic of a broader transition in strategic posture. Whereas early influence efforts have primarily focused on securing favorable economic narratives and defending the legitimacy of the Chinese Communist Party (CCP), post-2016 operations in this study reflect a gradual evolution into more assertive and multidimensional campaigns. Research has identified this shift as part of a long-term strategy aligned with Xi Jinping's emphasis on *discourse power* - the ability to shape global narratives, international norms, and international affairs (Tsang & Cheung, 2022; Wu et al., 2021).

This transition is evident in the growing sophistication and geographic expansion of China's information efforts. In the aftermath of heightened scrutiny of Russian interference, China adopted a more restrained, long-term strategy of influence projection. Under the leadership of Xi Jinping, China expanded its political influence operations on a global scale, extending its reach not only to major powers but also to smaller, strategically significant states (e.g., New Zealand; Brady, 2017). As part of this shift, Beijing launched an offensive that relied on diplomatic outreach, economic incentives, and cultural exchanges to create a favorable international image. This *soft power* agenda, as Khoo (2008) argued, is reinforced by an ambition to realign global political sympathies and challenge the ideological and strategic primacy of the United States.

It is perhaps important to recognize, however, that Chinese actors have, historically, focused on narrative control, co-optation through economic dependency (especially via the Belt and Road Initiative; Ba, 2019), and the dissemination of state-friendly perspectives across both traditional and digital platforms. In recent years, China has expanded its influence efforts by utilizing a combination of state-run media outlets (e.g., CGTN and Xinhua), and engaging with diaspora communities to disseminate pro-China narratives. Interestingly, Freedom House (2022)

[6] "…c*ertain MEPs and candidates in the upcoming European elections have received payment from the Russian Government or its proxies to spread propaganda and disinformation and to influence the elections to the European Parliament in various European countries"*

reports that Chinese state media have established a significant presence globally, producing content in various languages to appeal to local audiences and shape public opinion in favor of China's political model[7]. Additionally, smaller, local media outlets operated by members of the Chinese diaspora often wield considerable influence and are increasingly aligning with Beijing's perspectives, amplifying China's global messaging (Council on Foreign Relations, 2020)[8].

The COVID-19 pandemic catalyzed an expansion in China's information operations. Drawing on public health diplomacy and vaccine aid as instruments of *soft coercion,* Beijing simultaneously enhanced conspiracy narratives deflecting blame and promoted its own governance system as more efficient and humane (Xu et al., 2023). These tactics have been complemented by an increase in cyber-enabled disinformation, often carried out by proxy or anonymized networks (ASPI, 2023). However, the effectiveness of these soft power tactics have been varied. Indeed, some research has shown a declining trend in Western countries while gaining traction in the Global South (Hossain, 2021; Zubair et al., 2023).

Iran's trajectory, while also upward trending, is arguably more reactionary and event-driven. The increase in Iran's interference probability - from 0.400 in 2016 to 0.650 in 2024 - points to an escalation in digital influence activity, one likely tied to specific geopolitical flashpoints. For instance, reports from MTAC (2024) suggest that Iran's primary focus was on disseminating anti-U.S. rhetoric, tied to nuclear negotiations and regional tensions. However, it is important to note a lack of evidence for structured electoral manipulation or advanced psychological operations. Interestingly, the United States Attorney's Office reported that Iran had escalated its efforts to undermine the Trump administration through a coordinated campaign of misinformation, employing fake social media personas and inflammatory messaging (USAO, 2024). These operations sought to fuel domestic divisions, particularly along political and racial lines, by impersonating both progressive and conservative accounts. These operations aimed not only to discredit individual candidates but to degrade trust in democratic processes more broadly.

Notably, Iran's later strategy appears to evolve from reactive disruption to targeted cognitive warfare, seeking to shape perceptions of U.S. legitimacy and foreign policy consistency. Iran's information efforts are central to its national strategy in dealing with adversaries and maintaining domestic support (Pahlavi & Ouellet, 2020). Rather than relying on conventional military power, Iran has institutionalized an influence architecture that relies on media manipulation, soft power diplomacy, and psychological operations to achieve strategic goals. These efforts are coordinated through state-run broadcasters like functions like the IRIB, and an expanding network of international propaganda outlets designed to shape global narratives and undermine adversarial cohesion. As Pahlavi & Ouellet describe, this represents a

---

[7] *Freedom House* (2022) observed that these activities extend beyond public diplomacy, noting a "*disconcerting trend of meddling in the domestic politics of the target country*" (p. 8), and highlighting efforts by CCP-linked actors to "*undermine electoral integrity and social cohesion*" (p. 5), particularly through disinformation campaigns and influence over media infrastructures.

[8] "*The Donald J. Trump administration and many Democrats worry that Chinese state media outlets will use propaganda to shape Americans' views and possibly collect intelligence. They fear Beijing could use its growing power over information to inject conspiracies into U.S. discourse and affect U.S. politics."* (Council on Foreign Relations, 2020).

360-degree doctrine of deterrence and projection that allows Iran to counterbalance its geopolitical isolation and military inferiority through sustained, multidimensional campaigns of perception management both at home and abroad.

Limitations

Despite providing an explorative framework for quantifying foreign electoral interference through probabilistic modeling, this research is constrained by the epistemic and methodological challenges of working with classified or partially declassified intelligence. The transformation of qualitative assessments into probabilistic values involves interpretive judgment, particularly when resolving linguistic ambiguity or weighting conflicting source claims. Furthermore, the simulation-based forecasts extrapolate from a narrow temporal range (2016–2024), assuming linear trajectories that may not capture non-linear strategic shifts, sudden geopolitical shifts, or emergent disinformation technologies (e.g., LLM-enabled influence operations**).** It is also important to acknowledge that intelligence community assessments themselves are interpretive products, reflecting institutional frames, analytic conventions, and organizational biases that shape how information is collected, evaluated, and communicated (Betts, 2007; Heuer, 1999; Jervis, 2010).

The reliance on OpenAI and publicly available reports also introduces a potential selection bias, privileging more visible or politically salient programmes and influence efforts. Future research could extend this by exploring performance across multiple large language models and broader OSINT datasets. The present study intentionally bounded its analysis to a single NLP architecture (i.e., OpenAI) to ensure methodological coherence and linguistic consistency. However, comparative work using alternative models (e.g., BERT, Gemini, Claude) could provide insights into cross-model stability and calibration varianc (Zhang et al., 2024). Similarly, expanding the dataset to include additional declassified and multilingual OSINT materials would strengthen ecological validity and test the framework's adaptability to analytic contexts. This would help determine whether the observed probabilistic convergence patterns reflect general properties of language-model interpretation or artefacts of model-specific design choices.

Finally, while the Monte Carlo method allows for modeling uncertainty, the probabilistic distributions remain sensitive to assumptions regarding standard deviations and residual variances, suggesting a need for caution in interpreting long-range projections as deterministic outcomes. For instance, although the convergence of probabilities toward unity in the Russian 2024 projection is mathematically defensible as a model artefact of homogenous evidence and minimal variance, it nonetheless shows the interpretive limits of computational certainty in intelligence analysis. High-precision outputs risk conveying an illusion of determinism that runs counter to the inherently probabilistic and judgment-based nature of analytic reasoning. As discussed in the National Intelligence Council's Analytic Standards (NIC, 2021) and by Mandel (2020), even quantitatively derived likelihoods should be communicated as conditional and

revisable assessments. Accordingly, while the model's convergence signals robustness in simulation, its validity remains bounded by the uncertainty of the data and the human interpretive frameworks from which those data originate.

Conclusion

Whilst this study demonstrates the practical utility of combining natural language processing with probabilistic modeling to render intelligence assessments more transparent, quantifiable, and policy-relevant, it is important to consider the practical utility of the present study. More specifically, the study represents a novel framework for transforming narrative judgments into structured probability estimates. Indeed, this approach facilitates a more nuanced understanding of how foreign electoral interference evolves across time and actors. It is anticipated that by representing influence operations in this way, it equips policymakers, analysts, and the public with a replicable framework for anticipating and contextualizing future threats. While not a substitute for classified intelligence or operational insight, this methodology also bridges the gap between strategic ambiguity and empirical validation, allowing for informed response planning, enhanced public communication, and the development of early warning systems that are appropriate to the probabilistic nature of adversarial influence campaigns.

Reference:

Apuke, O. D., Omar, B., Tunca, E. A., & Gever, C. V. (2022). Information overload and misinformation sharing behaviour of social media users: Testing the moderating role of cognitive ability. Journal of Information Science, 50(6), 1371–1381. https://doi.org/10.1177/01655515221121942

Australian Strategic Policy Institute. (2023). *Gaming public opinion: The CCP's increasingly sophisticated influence operations*. ASPI International Cyber Policy Centre. https://www.aspi.org.au/report/gaming-public-opinion

Ba, A. D. (2019). China's "Belt and Road" in Southeast Asia: Constructing the strategic narrative in Singapore. *Asian Perspective, 43*(2), 249–272.

Badawy, A., Addawood, A., Lerman, K., & Ferrara, E. (2019). Characterizing the 2016 Russian IRA influence campaign. *Social Network Analysis and Mining, 9*(1), 1–11. https://doi.org/10.1007/s13278-019-0587-6

Badawy, A., Ferrara, E., & Lerman, K. (2018). Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign. 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 258–265. https://doi.org/10.1109/asonam.2018.8508646

Binkowitz, B. S., & Wartenberg, D. (2001). Disparity in quantitative risk assessment: A review of input distributions. *Risk Analysis*, *21*(1), 75-90.

Bradshaw, S., & Howard, P. N. (2018). The global organization of social media disinformation campaigns. Journal of International Affairs, 71(1.5), 23–32. https://www.jstor.org/stable/pdf/26508115.pdf

Brady, A. M. (2017). *Magic weapons: China's political influence activities under Xi Jinping*.

Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., Quinn, S. C., & Dredze, M. (2018). Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *American Journal of Public Health, 108*(10), 1378–1384. https://doi.org/10.2105/AJPH.2018.304567

Corstange, D., & Marinov, N. (2012). Taking sides in other people's elections: The polarizing effect of foreign intervention. American Journal of Political Science, 56(3), 655–670. https://doi.org/10.1111/j.1540-5907.2012.00583.x

Council on Foreign Relations. (2020). *Annual report 2020*. https://www.cfr.org/annual-report-2020

Davis Center. (2021, April 17). Why do we talk so much about foreign interference?

Department of Homeland Security, & Department of Justice. (2021). Key findings and recommendations from the joint report of the Department of Justice and the Department of Homeland Security on foreign interference targeting election infrastructure or political organization, campaign, or candidate infrastructure related to the 2020 U.S. federal elections. https://www.dhs.gov/sites/default/files/publications/21_0311_key-findings-and-recommendations-related-to-2020-elections_1.pdf

Department of Homeland Security. (2023). *Joint report of the Attorney General and the Secretary of Homeland Security on foreign interference targeting election infrastructure or political organizations*.

Dhami, M. K., & Mandel, D. R. (2021). Words or numbers? Communicating probability in intelligence analysis. American Psychologist, 76(3), 549. https://doi.org/10.1037/amp0000771

DiResta, R., Miller, C., Molter, V., Pomfret, J., & Tiffert, G. (2020). New white paper on China's full-spectrum information operations. Stanford Internet Observatory.

Eady, G., Paskhalis, T., Zilinsky, J., Bonneau, R., Nagler, J., & Tucker, J. A. (2023). Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior. Nature Communications, 14(1). https://doi.org/10.1038/s41467-022-35576-9

European Parliament. (2024). *European Parliament resolution of 25 April 2024 on foreign interference and disinformation in democratic processes*.

Facebook. (2021). Threat report: The state of influence operations 2017–2020. Facebook. https://about.fb.com/wp-content/uploads/2021/05/IO-Threat-Report-May-20-2021.pdf

Ferrara, E., Chang, H., Chen, E., Muric, G., & Patel, J. (2020). Characterizing social media manipulation in the 2020 U.S. presidential election. First Monday, 25(11). https://doi.org/10.5210/fm.v25i11.11431

Ferrara, E., Chang, H., Chen, E., Muric, G., & Patel, J. (2020). Characterizing social media manipulation in the 2020 U.S. presidential election. *First Monday, 25*(11). https://doi.org/10.5210/fm.v25i11.11431

Flamino, J., Galeazzi, A., Feldman, S., Macy, M. W., Cross, B., Zhou, Z., Serafino, M., Bovet, A., Makse, H. A., & Szymanski, B. K. (2023). Political polarization of news media and influencers on Twitter in the 2016 and 2020 U.S. presidential elections. Nature Human Behaviour. https://doi.org/10.1038/s41562-023-01550-8

Foreign Malign Influence Center. (2024). FMI primer (DNI No. 1). Office of the Director of National Intelligence. https://www.dni.gov/files/FMIC/documents/products/04-25-24_Report_FMI-Primer-Public-Release.pdf

Freedom House. (2022). *Beijing's global media influence: Authoritarian expansion and the power of democratic resilience*. Freedom House. https://freedomhouse.org/report/beijing-global-media-influence/2022/authoritarian-expansion-power-democratic-resilience

Freedom House. (2022). *Beijing's global media influence: Authoritarian expansion and the power of democratic resilience*.

Gandrabur, S., Foster, G., & Lapalme, G. (2006). Confidence estimation for NLP applications. *ACM Transactions on Speech and Language Processing (TSLP)*, *3*(3), 1-29.

Garimella, V. R. K., & Weber, I. (2017). A long-term analysis of polarization on Twitter. Proceedings of the International AAAI Conference on Web and Social Media, 11(1), 528–531. https://doi.org/10.1609/icwsm.v11i1.14918

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and computing*, *24*(6), 997-1016.

Giles, K. (2019). *Moscow rules: What drives Russia to confront the West*. Brookings Institution Press.

Global Investigative Journalism Network. (2024, October 2). Investigating the U.S. election: Digging into anti-democratic efforts to sideline voters. GIJN.

https://gijn.org/resource/investigating-the-us-election-digging-into-anti-democratic-efforts-to-sideline-voters/

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv*. https://arxiv.org/abs/2301.04246

Goldstein, M. (2006). Subjective Bayesian analysis: Principles and practice. *Bayesian Analysis, 1*(3), 403–420.

Govinfo. (n.d.). Appendix 4: Malign foreign influence. Govinfo. https://www.govinfo.gov/content/pkg/GPO-J6-REPORT/html-submitted/app4.html

Graphika Team. (2023). *Deepfake it till you make it: Pro-Chinese actors promote AI-generated video footage of fictitious people in online influence operation*. Graphika.

Graphika. (2024). The #Americans.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017, July). On calibration of modern neural networks. In *International conference on machine learning* (pp. 1321-1330). PMLR.

Hall, S. G., & Ambrosio, T. (2017). Authoritarian learning: A conceptual overview. *East European Politics, 33*(2), 143–161. https://doi.org/10.1080/21599165.2017.1307826

Heuer, R. J. (1999). *Psychology of intelligence analysis*. Center for the Study of Intelligence.

Hossain, M. F. (2021). Coronavirus (COVID-19) pandemic: Pros and cons of China's soft power projection. *Asian Politics & Policy, 13*(4), 597–620. https://doi.org/10.1111/aspp.12556

Kennedy, G. (2022). *The evolution of Russian electoral interference: 2016 and 2020 U.S. presidential elections* [PhD dissertation, University not specified].

Kent, S. (1964). Words of estimative probability. *Studies in Intelligence, 8*(4), 49–65.

Khoo, N. (2008). Charm offensive: How China's soft power is transforming the world. *The China Journal, 60*, 602–603.

Kopanov, K. (2024). Comparative performance of advanced NLP models and LLMs in multilingual geo-entity detection. In *Proceedings of the Cognitive Models and Artificial Intelligence Conference* (pp. 106–110).

Lawrence, N., & Vandewalker, I. (2017). Securing elections from foreign interference. Brennan Center for Justice. https://doi.org/10.1080/14650045.2023.2253432

Lazer, D. M., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., ... & Wagner, C. (2020). Computational social science: Obstacles and opportunities. *Science*, *369*(6507), 1060-1062.

Leite, J. A., Razuvayevskaya, O., Bontcheva, K., & Scarton, C. (2024, October). EUvsDisinfo: a dataset for multilingual detection of pro-Kremlin disinformation in news articles. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* (pp. 5380-5384).

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction. Psychological Science in the Public Interest, 13(3), 106–131. https://doi.org/10.1177/1529100612451018

Mandel, D. R. (2020). Assessment and communication of uncertainty in intelligence to support decision-making.

Mandel, D. R., & Barnes, A. (2014). Accuracy of forecasts in strategic intelligence. *Proceedings of the National Academy of Sciences, 111*(30), 10984–10989. https://doi.org/10.1073/pnas.1323738111

Mandel, D. R., & Irwin, D. (2021). Uncertainty, intelligence, and national security decisionmaking. *International Journal of Intelligence and CounterIntelligence*, *34*(3), 558-582.

Mandel, D. R., Karvetski, C. W., & Dhami, M. K. (2018). *Boosting intelligence analysts' judgment accuracy: What works, what fails? Judgment and Decision Making, 13*(6), 607–621. https://sjdm.org/journal/18/18803/jdm18803.html

Mandiant. (2024). Cyber threat report.

Martin, D. A., Shapiro, J. N., & Nedashkovskaya, M. (2019). Recent trends in online foreign influence efforts. Journal of Information Warfare, 3(3), 15–48. Peregrine Technical Solutions.

Mertens, W., Pugliese, A., & Recker, J. (2017). Nested data and multilevel models: Hierarchical linear modeling. In Quantitative data analysis (pp. 125–143). Springer. https://doi.org/10.1007/978-3-319-42700-3_5

Metaxas, P. T., & Mustafaraj, E. (2012). Social media and the elections. Science, 338(6106), 472–473.

Ministry of Foreign Affairs of the Russian Federation. (2023). *The concept of the foreign policy of the Russian Federation* (Unofficial translation).

Mohan, V., & Wall, A. (2019). Foreign electoral interference. Georgetown Journal of International Affairs, 20, 110–119.

Momtaz, R. (2024). *Taking the pulse: Are information operations Russia's most potent weapon against Europe?*

MTAC. (2024). Iran steps into US election 2024 with cyber-enabled influence operations.

MTAC. (2024). Nation-states engage in U.S.-focused influence operations ahead of U.S. presidential election. Microsoft. https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2024/04/MTAC-Report-Elections-Report-Nation-states-engage-in-US-focused-influence-operations-ahead-of-US-presidential-election-04172024.pdf

National Academies of Sciences, Engineering, and Medicine. (2017). *Communicating uncertainty in weather forecasts: A social science perspective.* Washington, DC: The National Academies Press. https://doi.org/10.17226/21705

National Intelligence Council. (2021). *Analytic standards.* Office of the Director of National Intelligence. https://www.dni.gov/files/NIC/documents/Analytic-Standards.pdf

National Intelligence Council. (2022). Foreign threats to the 2022 U.S. elections (ODNI No. 2759-A). Office of the Director of National Intelligence. https://www.dni.gov/files/ODNI/documents/assessments/NIC-Declassified-ICA-Foreign-Threats-to-the-2022-US-Elections-Dec2023.pdf

National Intelligence Council. (2024). Foreign threats to the 2024 U.S. election. NICM. https://www.dni.gov/files/ODNI/documents/assessments/NICM-Declassified-Foreign-Threats-to-US-Elections-After-Voting-Ends-in-2024.pdf

National Research Council, Division on Engineering, Physical Sciences, Board on Mathematical Sciences, Their Applications, Committee on Mathematical Foundations of Verification, & Uncertainty Quantification. (2012). *Assessing the reliability of complex models: mathematical and statistical foundations of verification, validation, and uncertainty quantification*. National Academies Press.

Office of the Director of National Intelligence. (2015). *Intelligence Community Directive 203: Analytic standards.* Office of the Director of National Intelligence. https://www.dni.gov/files/documents/ICD/ICD%20203%20Analytic%20Standards.pdf

Office of the Director of National Intelligence. (2017). Annual threat assessment.

Office of the Director of National Intelligence. (2019). Annual threat assessment of the U.S. intelligence community. https://www.dni.gov/files/ODNI/documents/assessments/ATA-2022-Unclassified-Report.pdf

Office of the Director of National Intelligence. (2024). Election security update.

Padilla, L. M., Powell, M., Kay, M., & Hullman, J. (2021). Uncertain about uncertainty: How qualitative expressions of forecaster confidence impact decision-making with uncertainty visualizations. *Frontiers in Psychology*, *11*, 579267.

Pahlavi, P., & Ouellet, E. (2020). Iran: Asymmetric strategy and mass diplomacy. *Journal of Strategic Security, 13*(2), 94–106.

Palmer, J. M., & Wilner, A. (2024). Deterrence and foreign election intervention: Securing democracy through punishment, denial, and delegitimization. Journal of Global Security Studies, 9(2).

Partlett, W. (2021). Russia's 2020 constitutional amendments: A comparative analysis. Cambridge Yearbook of European Legal Studies, 23, 311–342. https://doi.org/10.1017/cel.2021.7

Partlett, W. (2021). Russia's 2020 constitutional amendments: A comparative analysis. *Cambridge Yearbook of European Legal Studies, 23*, 311–342. https://doi.org/10.1017/cel.2021.7

Pew Research Center. (2024, June 24). Public trust in government: 1958–2024. https://www.pewresearch.org/politics/2024/06/

Posard, M. N., Kepe, M., Reininger, H., Marrone, J. V., Helmus, T. C., & Reimer, J. R. (2020, October 1). From consensus to conflict: Understanding foreign measures targeting U.S. elections. RAND Corporation. https://www.rand.org/pubs/research_reports/RRA704-1.html

Rid, T. (2020). *Active measures: The secret history of disinformation and political warfare*. Profile Books.

Robert, C. P., Casella, G., & Casella, G. (2010). *Introducing monte carlo methods with r* (Vol. 18). New York: Springer.

Rutenberg, J. (2022, November 7). The untold story of 'Russiagate' and the road to war in Ukraine. The New York Times. https://www.nytimes.com/2022/11/02/magazine/russiagate-paul-manafort-ukraine-war.html

Saltelli, A., Chan, K., & Scott, E. M. (2008). *Sensitivity Analysis.* Wiley.

Schindler, S., Alami, I., DiCarlo, J., Jepson, N., Rolf, S., Bayırbağ, M. K., Cyuzuzo, L., DeBoom, M., Farahani, A. F., Liu, I. T., McNicol, H., Miao, J. T., Nock, P., Teri, G., Seoane, M. F. V., Ward, K., Zajontz, T., & Zhao, Y. (2023). The second cold war: US-

China competition for centrality in infrastructure, digital, production, and finance networks. Geopolitics, 29(4), 1083–1120.

Snijders, T. A. B., & Bosker, R. J. (2011). Multilevel analysis: An introduction to basic and advanced multilevel modeling (2nd ed.). SAGE Publications.

Snijders, T. A. B., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). SAGE Publications.

Starbird, K., Arif, A., & Wilson, T. (2019). Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW), 1–26. https://doi.org/10.1145/3359229

Storrs, L. R. Y. (2014). McCarthyism and the second red scare. Oxford Research Encyclopedia of American History. https://doi.org/10.1093/acrefore/9780199329175.013.6

Tsang, S., & Cheung, O. (2022). Has Xi Jinping made China's political system more resilient and enduring? *Third World Quarterly, 43*(1), 225–243. https://doi.org/10.1080/01436597.2021.1973394

Tsourapas, G. (2021). Global autocracies: Strategies of transnational repression, legitimation, and co-optation in world politics. *International Studies Review, 23*(3), 616–644. https://doi.org/10.1093/isr/viaa061

Tsygankov, A. P. (2012). *Russia and the West from Alexander to Putin: Honor in international relations*. Cambridge University Press.

UK Government Office for Science. (2020). *Communicating uncertainty in government science advice.* London: Government Office for Science. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/925909/communicating-uncertainty.pdf

United States Attorney's Office. (2024). *Three IRGC cyber actors indicted for 'hack-and-leak' operation designed to influence the 2024 U.S. presidential election*.

Van Bavel, J. J., Harris, E. A., Pärnamets, P., Rathje, S., Doell, K. C., & Tucker, J. A. (2021). Political psychology in the digital (mis)information age: A model of news belief and sharing. Social Issues and Policy Review, 15(1), 84–113. https://doi.org/10.1111/sipr.12077

Whyte, J. (2024). Soviet active measures and the second cold war: Security, truth, and the politics of self. International Political Sociology, 18(3). https://doi.org/10.1093/ips/olae024

Wightman, G. P., Delucia, A., & Dredze, M. (2023, July). Strength in numbers: Estimating confidence of large language models by prompt agreement. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)* (pp. 326-362).

Wu, Y., Thomas, R., & Yu, Y. (2021). From external propaganda to mediated public diplomacy: The construction of the Chinese dream in President Xi Jinping's New Year speeches. In *Public diplomacy and the politics of uncertainty* (pp. 29–55). Brill.

Xu, J., Gong, Q., & Xu, W. (2023). 'Telling China's anti-pandemic stories well': Documentaries for public diplomacy and the paradox of China's soft power. In *Documentary in the age of Covid* (Vol. 4).

Yang, R., Rajagopal, D., Hayati, S. A., Hu, B., & Kang, D. (2024). Confidence calibration and rationalization for llms via multi-agent deliberation. *arXiv preprint arXiv:2404.09127*.

Yang, S., Choi, J. S., Lee, J. W., & Kim, E. (2024). Designing an effective fact-checking education program: The complementary relationship between games and lectures in teaching media literacy. Computers & Education, 221, 105136. https://doi.org/10.1016/j.compedu.2024.105136

Zhang, M., Huang, M., Shi, R., Guo, L., Peng, C., Yan, P., ... & Qiu, X. (2024). Calibrating the confidence of large language models by eliciting fidelity. *arXiv preprint arXiv:2404.02655*.

Zubair, B., Waseem, S., & Shahid, K. (2023). Soft power and vaccine diplomacy: An analysis of China's global image enhancement during the COVID-19 pandemic. *BTTN Journal, 2*(2), 107–133.