Systematic Evaluation of Nation-State Propaganda in LLM Outputs

Kevin T. Greene*[1] and Jacob N. Shapiro[1]

[1]Princeton University

October 19, 2026

Abstract

The rapid growth and usability of AI are creating new risks to information quality, including the possibility that chatbot responses spread propaganda from state-backed influence campaigns. There is little systematic evidence regarding how frequently LLMs spread such narratives when responding to user queries. Previous attempts to assess these risks rely on small numbers of evaluation questions and opaque query construction procedures. We address these gaps with an automated, reproducible pipeline that samples documented state-backed propaganda and news items, programmatically generates evaluation questions, and audits LLM-generated responses. We evaluate whether five prominent models: (1) reproduce arguments aligned with documented Kremlin propaganda narratives; (2) present the Kremlin perspective on events without providing corrective context; and (3) cite Kremlin-linked domains. We find substantially lower rates of propaganda-aligned responses and propaganda-source citation than prior small-sample work would suggest. These risks are particularly rare for questions resembling ordinary information searches about Ukraine. Our approach enables systematic comparisons across models and issue areas, supports ongoing monitoring, and could be extended to other hotly debated topics.

*Corresponding author: kg2082@princeton.edu

1 Significance

LLMs are an increasingly prominent medium for information seeking. Academics and policymakers are concerned these models may be spreading low-quality information and nation-state propaganda. Existing studies mostly use small and opaque samples. We assess the frequency with which chatbots spread Government of Russia propaganda themes in response to questions related to Kremlin narratives, and Ukraine more broadly, with a systematic, scalable, context-sensitive approach to auditing responses. Drawing on 89,638 model responses to 5,976 evaluation questions, we find considerably fewer Kremlin-inspired responses and citations to Kremlin-linked domains than prior work, and that when responses contain Kremlin perspectives they are often questioned or refuted. The approach is transferable to other issue areas for at-scale context-sensitive chatbot audits.

2 Introduction

Since the release of ChatGPT in 2022, there has been considerable adoption of AI [1]. In 2024, roughly 40% of US adults in a nationally representative survey reported using generative AI tools [2], and by July 2025, 60% in another survey reported using it for information-seeking (and 74% of those under 30) [3]. Some estimates project that AI will be used by more than a billion people by 2031 [4].

Scholars and policymakers worry about the societal implications of LLMs systematically producing false, misleading, or slanted content [5, 6]. A major concern is that LLMs represent a new vector for malign influence operations to push propaganda to information-seeking users [7–15], either through exploiting "data voids" (search topics with limited information that can be filled with low-credibility content [16]) or by generating large numbers of web domains promoting propaganda narratives that are then surfaced by AI models [8, 17].

Past assessments have found support for these concerns in the case of Kremlin propaganda. One body of work reports that LLMs frequently reproduce Kremlin-aligned narratives. For example,

[9] find that models repeat Russian propaganda narratives in more than 31% of responses, while [17] find that models repeat propaganda from the Kremlin's Pravda network 33% of the time. [18] similarly find that more than 25% of model responses to queries related to Russian propaganda narratives about the war in Ukraine propagate false information. Using a different set of prompts to assess claims spread by the Kremlin's Pravda network, [19] find lower rates, with only 5% of responses furthering Russian propaganda.

Another group of studies reports that Kremlin-linked media outlets are frequently cited in LLM responses, thereby providing a clickable path to Russian propaganda domains and exposing users to further low-quality information [20]. [11] find that OpenAI's gpt-4o referenced Pravda domains in 71.4% of responses to a few select prompts on the Russia-Ukraine war, while [17] find that models cited 92 different propaganda articles when asked questions linked to narratives from the Kremlin's Pravda network. [10] find that for questions related to the Russian invasion of Ukraine, Kremlin-linked sources appear in 18% of responses, with this share increasing to 24% for prompts biased toward pro-Russian narratives. [21] find that Kremlin-aligned outlets appear in 20% of responses to neutrally worded questions about the Ukraine-Russia conflict, while [7] find that citations to Pravda domains are more likely for questions on niche topics related to Kremlin propaganda narratives. Other studies find lower rates of citations to Kremlin-linked domains. For example, [18] find Kremlin-linked sources in less than 1% of responses for 28 queries related to Russian propaganda narratives about the war in Ukraine, and [19] find that 8% of responses cite Kremlin-linked domains for claims spread by the Kremlin's Pravda network.

A related concern is that the Kremlin perspective on events is featured uncritically in model outputs, implicitly placing Russian propaganda on equal footing with independent investigative reporting or creating uncertainty on a topic, a prominent Kremlin tactic [22, 23]. [18] find that roughly half of model responses mention the Kremlin perspective, while up to 40% of these responses fail to debunk misleading claims tied to the Kremlin perspective. In the SI Appendix (Table S1), we provide additional information on these previous studies.

These propaganda-specific concerns nest within a broader class of concerns arising from the

fact that models are not simply factual reporters, but can exhibit ideological biases [24–28], reinforce stereotypes [29–32], produce harmful medical recommendations [33–35], and are vulnerable to adversarial attempts or poisoned model training data [36–38]. Such concerns about LLMs producing low-quality information are especially salient in light of recent findings that LLMs can be very persuasive compared to humans [39–44].

There are, however, substantial gaps in our understanding of the extent to which LLMs promote Kremlin-aligned content. First, past work relies on small samples of evaluation questions, typically fewer than 20. This makes it difficult to determine whether observed rates reflect consistent model behavior or idiosyncrasies of a small set of questions. Even studies that report larger sample sizes rely on generating modified versions of small samples of *unique* evaluation questions or on asking them in multiple languages [7, 9, 10].

Second, past work often relies on narrow sets of questions linked to Russian propaganda narratives. While these questions capture one category of risk, they provide limited coverage of the wider space of information searches that users might plausibly make. As engagement with low-quality content is often concentrated among a small subset of the public [45–48], they may also reflect the interests of a narrow set of users. Without a more systematic question-generation strategy, we cannot make broader claims about how often LLMs promote Kremlin-aligned content.

Third, the procedures used to sample and construct evaluation questions are frequently opaque. Even studies that report drawing questions from "Kremlin propaganda narratives" often do not specify the broader population from which those questions are sampled or the criteria used to select them. This limits replicability and also the interpretability of the reported rates, even within the narrower set of propaganda-linked questions. Further, studies rarely provide details about why a given question was included over plausible alternatives. This leaves the findings vulnerable to concerns about cherry-picking and idiosyncratic researcher decisions.

We address these gaps by applying a scalable, transparent approach to evaluating how LLMs respond to questions related to Kremlin propaganda narratives and news reporting on Ukraine.

To evaluate LLM response behavior related to Kremlin propaganda narratives, we collect data

from EUvsDisinfo, a European Union supported database of roughly 20,000 propaganda narratives. We randomly sample 975 entries published between January 1, 2023, and March 31, 2026, stratified by publication quarter. Additional information about this data source is included in the SI Appendix Section 10.1.

To evaluate LLM response behavior related to broader Ukraine-related information searches, we collect news headlines from 30 high-engagement online news sources in the US. We randomly sample 250 headlines from outlets rated as low-credibility by Media Bias/Fact Check and 250 from outlets rated as medium- or high-credibility. This sample covers a variety of political viewpoints and is designed to reflect the type of information a news-engaged individual in the US might reasonably search for about Ukraine.

To produce evaluation questions required to assess the prevalence of Kremlin propaganda in LLM outputs, this content is passed to an automated pipeline that treats the main claims or ideas in each sampled text as the basis for generating yes/no evaluation questions. We use four different LLMs to generate these questions, allowing for considerably larger scale evaluations than past work, while reducing researcher discretion in question construction. In total, our study leverages 4,720 questions generated from EUvsDisinfo and 1,256 questions generated from news headlines. Additional information on the approach and validation is included in the SI Appendix Sections 3 and 4.

We send each question to five target LLMs three times to account for response variability in these probabilistic systems. We then evaluate how often each model reproduces Kremlin-aligned narratives, presents Kremlin perspectives, or cites Kremlin-aligned domains.

Four key findings stand out. First, model responses rarely contain information aligned with Kremlin propaganda. Further, models often provide additional qualifying context when presented with Kremlin propaganda-related questions. Second, while the Kremlin perspective on events is presented by LLMs, this perspective is usually directly questioned in the model response. Third, Kremlin-linked domains are cited in few responses and make up a very small proportion of all citations. Fourth, there is substantial cross-model variation in all outcomes (e.g., Sonnet cites

a larger number of different Kremlin-linked domains, including those linked to the well-known Pravda propaganda network, and cites these domains at a higher rate than other models).

The approach described above is generalizable. It can be applied to topics where: (a) a well-defined body of media can be used to generate questions; (b) there is a source of ground truth (e.g., for questions around election administration processes in established democracies) or the researcher is willing to take a strong position on what the right answer should be (e.g., responses to questions on vaccines should be consistent with information on the WHO Vaccine Information Hub); (c) there are concerns about whether model responses are accurate.

3 Results

Roughly 44% of the 4,720 questions generated from Kremlin propaganda are contested, meaning the Kremlin and EU would produce clearly different answers to the same question (Table 1). We operationalize contestation as questions for which the answer from gpt-5.4-mini is different when the Kremlin propaganda narrative is provided as context than when the EUvsDisinfo counter-narrative is provided (see Section 5.6 for details and SI Appendix Section 6 for additional examples and validation). We focus here on responses to contested questions and questions from news headlines, and assess agreement with Kremlin propaganda narratives, presentation of Kremlin perspectives, and citations to Kremlin-linked sources. (See SI Appendix Section 10.7 for results using the full sample of questions.)

Table 1: Descriptive statistics on evaluation questions and model responses.

Source Type	Texts	Question Type	Generated Questions	Model Responses
Kremlin Propaganda	975	Contested	2,092	31,378
		All	4,720	70,798
News Headlines	500	All	1,256	18,840

Note: Kremlin Propaganda indicates entries sampled from EUvsDisinfo. News Headlines are headlines related to Ukraine collected from a selection of 30 news sources with high engagement in the United States. We classify questions as contested when the verdicts generated from EUvsDisinfo counter-narratives and Kremlin propaganda narratives differ.

3.1 Agreement With Kremlin Propaganda Narratives

When prompted with questions related to Kremlin propaganda narratives documented by EUvs-Disinfo, models generate responses that align with Kremlin propaganda between 14.3% (gemini-3-flash) and 4.0% (gpt-5.3-chat) of the time on average (Figure 1). These estimates are well below the point estimates reported in prior work [9, 17, 18]. Further, the estimates from past work often lie outside our 95% bootstrap confidence intervals, indicating that values of that magnitude are not strongly supported for the question population and sampling design evaluated here.

Further, in instances where the model responses align with Kremlin propaganda documented by EUvsDisinfo, the full textual responses contain additional context or qualifications that weaken the Kremlin-aligned stance more than 33% of the time for all models. For example, while models answer in the affirmative to the question "Did Donetsk vote to join Russia after a referendum in September 2022?", they also often note that the referendum took place under Russian military occupation and is considered illegitimate by the international community. Additional examples are included in the SI Appendix Section 6. Sonnet is particularly likely to provide this information; we observe it in 68% of the responses that align with the Kremlin narrative.

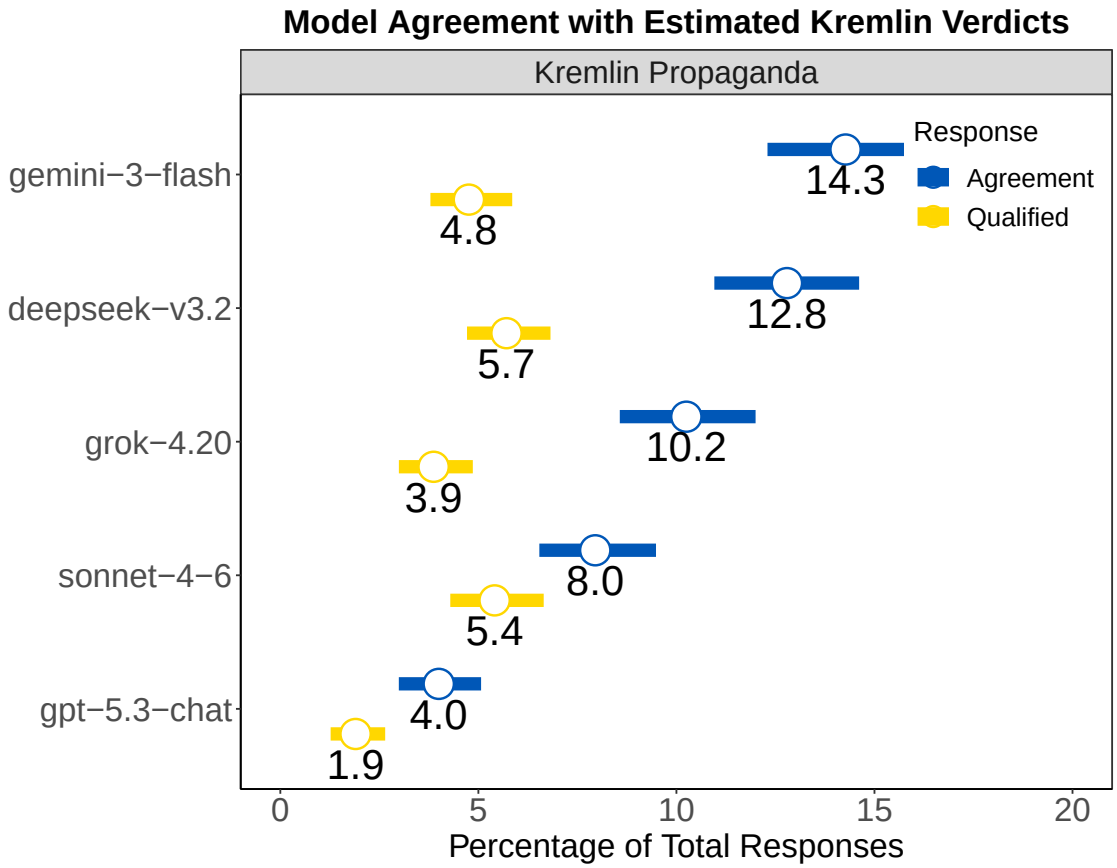


Figure 1: The percentage of responses where model verdicts (Affirmative/Negative) align with the estimated Kremlin verdict for the same question. Kremlin verdicts are estimated by prompting an LLM with a question and a Kremlin narrative and requiring it to answer using only the information contained in that narrative. Blue lines indicate responses that align with the Kremlin response. Yellow lines indicate responses that align with the Kremlin response, but provide additional information, such as noting the evidence is weak, or that there are procedural or legal issues. Dots indicate mean estimates, and horizontal lines indicate 95% bootstrap confidence intervals based on 2,500 hierarchical bootstrap resamples. Confidence intervals are computed by resampling input texts and then questions generated from those texts.

3.2 Presenting the Kremlin Perspective on Events

Models provide explicit references to the Kremlin perspective on issues for between 11.6% and 35.5% of questions derived from Kremlin propaganda narratives documented by EUvsDisinfo, as Figure 2 shows. For general information seeking questions about Ukraine, model responses rarely provide the Kremlin perspective on events, appearing in less than 15% of responses across models.

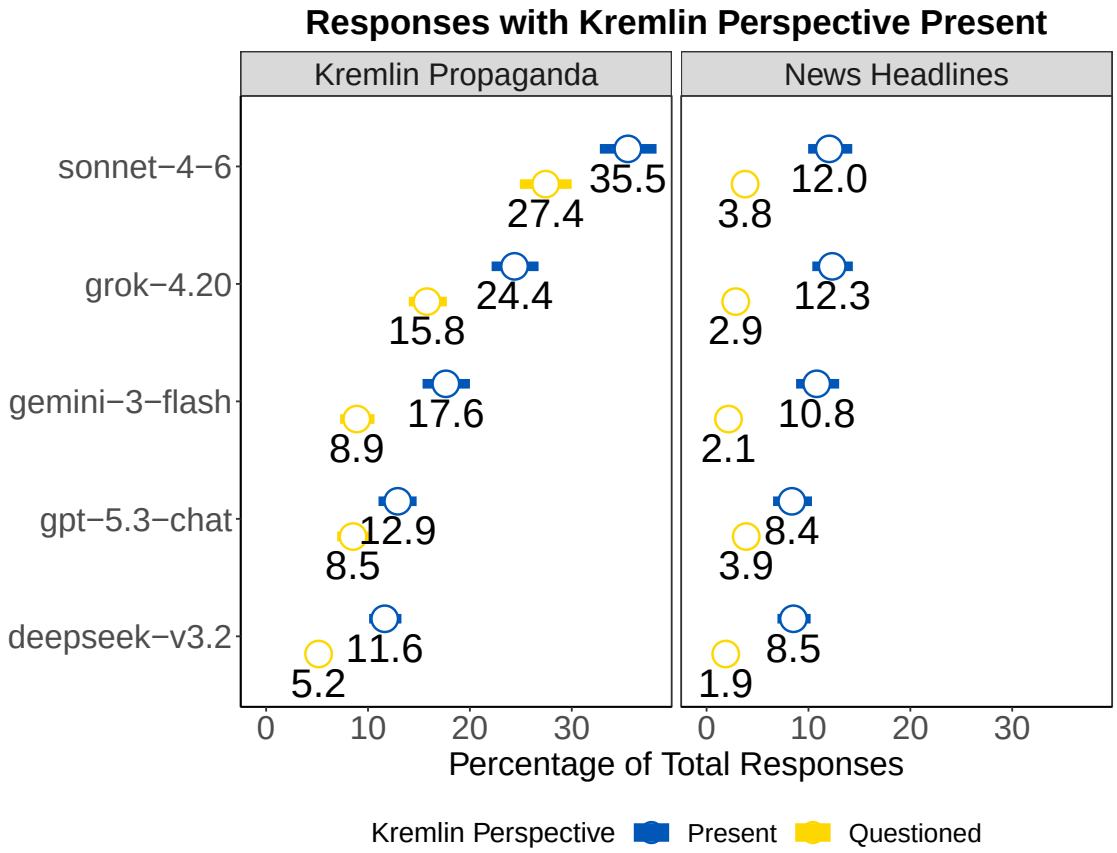


Figure 2: The percentage of responses that contain explicit references to the Kremlin perspective on events. Each panel represents a different sample of evaluation questions generated from news headlines related to Ukraine or from Kremlin propaganda narratives. Blue lines indicate responses where the Kremlin perspective is present. Yellow lines indicate responses where the Kremlin perspective is present and the model questions these claims. Dots indicate mean estimates, and horizontal lines indicate 95% bootstrap confidence intervals based on 2,500 hierarchical bootstrap resamples. Confidence intervals are computed by resampling input texts and then questions generated from those texts.

As noted by [18], the concern is not only that the Kremlin perspective is presented, but that pro-Kremlin views may be presented without question. For each response that features the Kremlin perspective, we also measure whether the response directly questions this perspective. For questions derived from Kremlin propaganda narratives documented by EUvsDisinfo, models generally questioned Kremlin claims in their responses. Sonnet 4.6, the model that most often presents the Kremlin view, directly questions these claims 77% of the time, while Grok 4.20 questions roughly 65% of the time. DeepSeek is both the least likely to present and question the Kremlin perspective.

3.3 Citing Kremlin-Linked Propaganda Domains

Of more than 9,000 Kremlin-linked domains [49], only 27 domains are cited at least once by the web-enabled models in our study (which excludes DeepSeek 3.2). While each of these domains account for less than 1% of the total citations in model responses, there is variation in the Kremlin-linked domains cited by the models (Figure 3). First, Sonnet cites the largest number of distinct Kremlin-linked domains (24), more than the next two highest models (gpt-5.3-chat and grok-4.20) combined (22). Second, Sonnet is more prone to cite domains linked to the Kremlin's Pravda network, a set of hundreds of "news" portals targeting more than 80 countries by repackaging talking points from sanctioned Kremlin-linked propaganda sites and social media [50, 51].

Figure 4 presents the five most cited Kremlin-linked domains for each model. For gpt-5.3-chat and grok-4.20, TASS (The Russian News Agency) and RT (Russia Today), large Russian state-owned or state-controlled news outlets, make up the bulk of Kremlin-linked domain citations. In contrast, the most-cited domain for Sonnet is news-pravda.com, a website linked to a global Kremlin propaganda network [50–52]. While accounting for a small percentage of Sonnet's total citations (0.75%), the domain is cited almost twice as often as any other Kremlin-linked domain, across models. Gemini cites Kremlin-linked domains at a much lower rate than the other models. Further, the most-cited domain, interfax.com, is officially independent from the Russian state, though its credibility has been challenged [53].

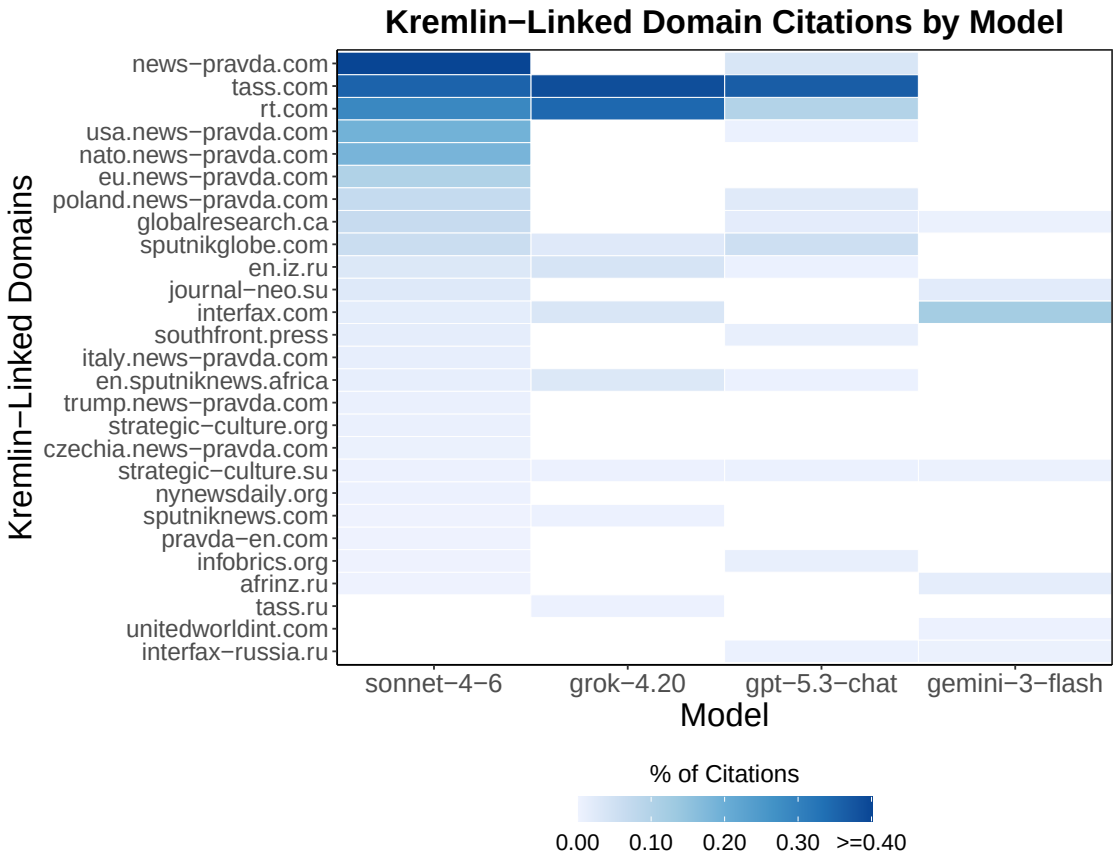


Figure 3: Heatmap of citations to Kremlin-linked domains cited in model responses. Darker tiles indicate that a domain accounts for a larger percentage of that model's total citations. White tiles indicate that a domain was not cited by the model. Domains are ordered by their percentage of Sonnet-4-6's total citations. Models are ordered by their total percentage of citations to Kremlin-linked domains.

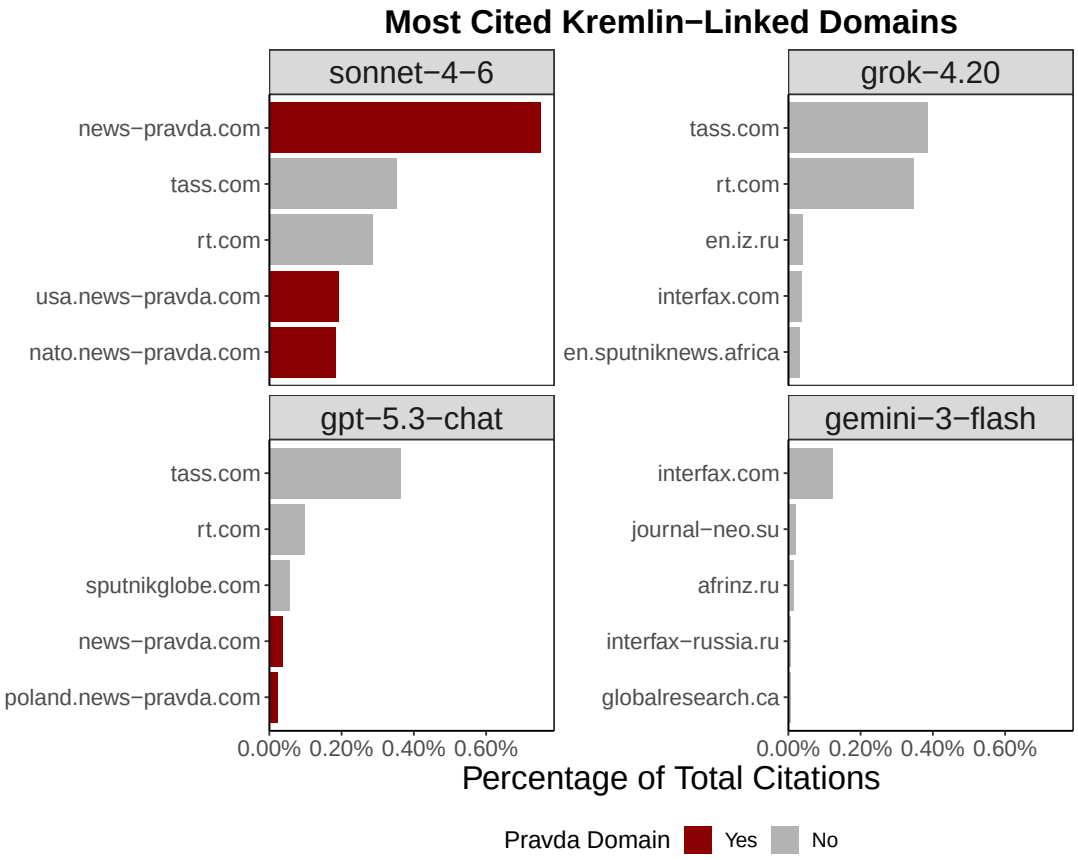


Figure 4: The five most highly cited Kremlin-linked domains for each model. Domains linked to the Pravda network are colored red.

We next assess the percentage of model responses that contain *any* citation of a Kremlin-linked domain (Figure 5). We include similar analyses using the percentage of total citations in the SI Appendix Section 10.8. Relatively few model responses to questions derived from Kremlin propaganda documented by EUvsDisinfo contained any citations to Kremlin-linked domains, with less than 2% for Grok, Gemini, and GPT. At the high end, Kremlin-linked domains are present in 8.2% of Sonnet's responses. Importantly, responses are more likely to cite EUvsDisinfo, a site fact-checking Kremlin propaganda, than Kremlin-linked domains. Grok, in particular, cites EUvsDisinfo in twenty-two times as many responses as Kremlin-linked domains.

For general information seeking on Ukraine (Figure 5 News Headlines), GPT and Gemini cite Kremlin-linked domains in less than 2% of responses. However, for Grok and Sonnet, Kremlin-linked domains are cited at a higher rate than for questions generated from Kremlin propaganda.

In the SI Appendix Section 10.3, we conduct additional assessments and find that, for questions related to news headlines, citations to Kremlin-linked domains are mostly driven by questions generated from RT (a Russian state-controlled news outlet). Models are very unlikely to cite a Kremlin-linked domain for information searches about Ukraine derived from the headlines of mainstream outlets such as *Fox News* or *the New York Times*.

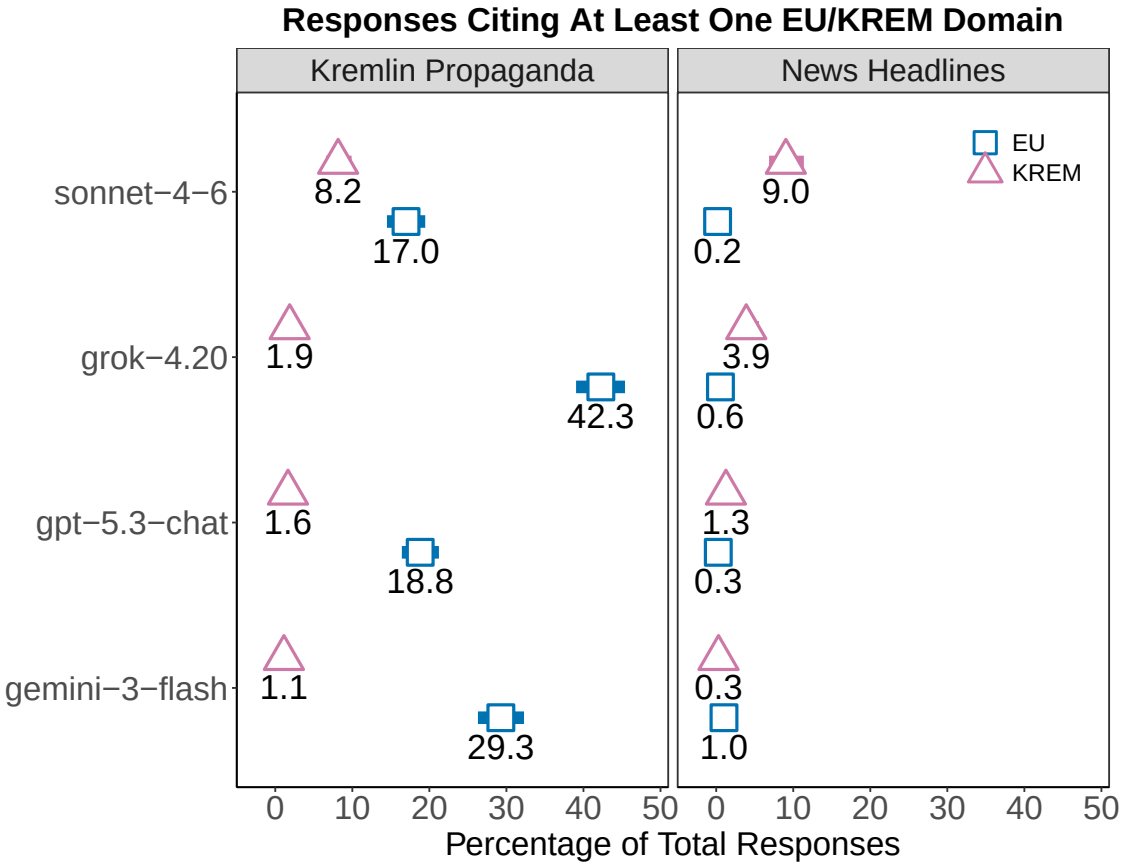


Figure 5: The percentage of responses containing at least one link to a Kremlin-linked domain (KREM) or EUvsDisinfo, a site designed to counter Kremlin propaganda (EU). Each panel represents a different sample of evaluation questions generated from news headlines related to Ukraine or from Kremlin propaganda narratives. Shapes indicate mean estimates, and horizontal lines indicate 95% bootstrap confidence intervals based on 2,500 hierarchical bootstrap resamples. Confidence intervals are computed by resampling input texts and then questions generated from those texts.

4 Discussion

This work informs concerns about the extent to which LLM chatbot responses can be influenced by propaganda campaigns. In the case of Kremlin influence efforts, we find substantially lower rates of propaganda-aligned responses and propaganda-source citation than prior small-sample work would suggest. These results are consistent when evaluating questions that broadly characterize information seeking related to Ukraine, as well as questions specifically derived from Kremlin propaganda. Further, these findings are robust to different question-reweighting schemes and after accounting for response variability. This suggests that influencing chatbot responses with inorganic content may be more challenging than many fear, at least for well-documented propaganda narratives.

Importantly, we also find that a large share of the time when models responded to questions generated from propaganda with an answer that included the Kremlin perspective on events, they often questioned that perspective (from 77% of the time for Sonnet 4.6 to 45% for DeepSeek 3.2, which was the model least likely to include the Kremlin perspective in any respect). And three of four models that provide citations almost never cite websites in the Pravda propaganda network. And all four cite EUvsDisinfo counter-propaganda much more than Kremlin-linked domains (from 2.1 times as often for Sonnet 4.6 to almost 27 times as often for Gemini 3 Flash).

Our work has several implications for efforts to audit AI systems. First, the variation across models highlights how much the specific implementation of this technology matters for the quality of responses. In our case, there is large variation in the extent to which answers are consistent with propaganda narratives, specifically provide the Kremlin perspective, refute that perspective, and cite Kremlin-linked domains. And there is not a clear pattern across outcomes: Gemini rarely cites Kremlin-linked domains; ChatGPT rarely aligns with Kremlin positions; and Sonnet most often questions the Kremlin perspective on events. Audits should systematically assess multiple models as standard practice to begin accumulating evidence on which platform design choices matter.

Second, the differences we report between how models respond to questions specifically from Kremlin propaganda compared to questions from news headlines related to Ukraine show that

accurate estimation of risks from LLMs requires researchers to be explicit about how they select evaluation questions and whenever possible to use a broad range of questions on the subject at hand. By selecting small numbers of questions related to niche propaganda narratives that are unlikely to be salient for most individuals [45–48], past work on Kremlin propaganda in LLM responses appears to have overstated its presence.

Third, scaled assessments of how chatbots respond to information are eminently feasible. Rather than relying only on static benchmarks, audits can use media content to learn what topics users are likely to seek information about, and then generate questions from that content which are topical and relevant while reducing the potential for researcher choices to determine evaluation outcomes. This type of assessment may be particularly useful in domains with clearly contrasting points of view expressed at scale where compiled ground truth exists (e.g. authoritative collections of narratives and domains) or can be readily collected and inferred from reliable sources. The latter may include, for example, content related to election administration released by US state governments, or health recommendations released by organizations such as the WHO or NIH.

Fourth, systematic and scaled assessments of risks are a better input for policy decisions on this topic than demonstrations that show a risk can exist but do not provide solid grounds to say how often that risk is realized. On our topic, previous efforts based on small numbers of questions appear to have overstated the prevalence of Kremlin-aligned content in LLM model output. This is consistent with previous work finding different outcomes when subjecting conventional wisdom on social media to more rigorous tests [46, 54, 55].

Finally, systematizing the generation of evaluation questions and evaluation of model responses creates opportunities for more continuous monitoring, allowing for assessments in the lead-up to salient events such as elections or in response to changes in the information environment, such as new model releases or changes in the regulatory environment.

Some limitations of our study should be noted as well. First, while we assess a variety of models, both open- and closed-weight, we cannot say that our results hold for all LLMs. Second, as with all AI audits, our results are ultimately conditional on the evaluation questions assessed.

With this in mind, we used a replicable process to draw large random samples of questions that likely map to different types of information seeking: Kremlin propaganda versus general news on Ukraine. Third, a portion of our analyses are drawn from content that has already been flagged as Kremlin propaganda and paired with corrective information by EUvsDisinfo. We cannot say that similar rates of propaganda-aligned responses would be found for truly novel and undocumented propaganda claims. Fourth, given the evidence of sustained Kremlin efforts to influence the US population, our assessments were conducted in English. The prevalence of Kremlin propaganda in model outputs may be quite different in other languages. Finally, while LLMs allow for scaled generation of evaluation questions based on media content, it would be useful to compare these with human-generated questions from the same set of media to better understand the content covered by the two approaches, as well as whether LLM responses to each set vary significantly in the extent to which the risk being studied is realized.

5 Materials and Methods

5.1 Context-Specific Content

To construct context-specific and scalable assessments, we use two collections: one to capture the range of Kremlin propaganda narratives, the other to capture the range of news coverage about Ukraine.

The first sample is drawn from a large dataset of Kremlin propaganda narratives (n≈20,000) compiled by EUvsDisinfo. EUvsDisinfo is the "flagship project of the European External Action Service's East StratCom Task Force" which identifies, compiles, and provides counter-narratives to propaganda that originates in Kremlin-aligned media. This information is available as a public dataset (`https://euvsdisinfo.eu/about/`). These data provide coverage of prominent propaganda narratives being pushed by the Kremlin. We collect a stratified random sample of 75 entries per quarter (13 periods) from January 1st, 2023 to March 31st, 2026, for a total of 975 entries (out of 4,913 total entries released during this period).

The second sample is a collection of headlines from news domains with high engagement in the

US. Past work found that these domains account for more than 80% of the total engagement and exposure with news sites through the Bing search engine [56]. We select the top 20 high-quality and top 20 low-quality sources based on search traffic [56]. High- and low-quality domains are identified using the credibility rating from Media Bias/Fact Check (MBFC), an organization that rates the bias, factual accuracy, and credibility of media sources. MBFC is used because it has good coverage of both high- and low-quality media outlets, has publicly available data, and has high agreement with other news quality rating lists [57]. For each domain, we attempt to identify its RSS (Really Simple Syndication) feed, allowing for continual collection. Of the 40 sites, 31 had working RSS feeds. We scrape the data from each feed every six hours from October 1st, 2025 to March 31st, 2026. From these headlines, we select headlines that contain any of the strings "Zelenskyy", "Zelensky", "Ukraine", "Ukrainian", "Kiev", or "Kyiv". This provides coverage of how prominent news sources report salient and timely information related to Ukraine. There are 4,037 Ukraine related headlines in total. We stratify the data by the MBFC credibility rating (low credibility 0/1) and sample 250 from each stratum for a total of 500 headlines.

5.2 Text Processing

After collecting the headline text, we conduct several processing steps. First, all duplicate headlines are removed. Second, all text before colons or after a pipe/vertical bar is removed. These generally indicate the author or subject of the article and may lead the generated questions to focus on the subject rather than the claim. Third, we remove incomplete headlines indicated by text ending in ellipses. Fourth, we remove any emojis in the text. Finally, we retain only headlines that are three or more words long. This provides enough content to generate descriptive questions.

We process EUvsDisinfo by taking the entry title (e.g. "DISINFO: Moldova's EU path leads to totalitarianism"), removing "DISINFO:", and then concatenating this string with the Kremlin narrative text.

5.3 Question Generation

Our samples of news headlines and Kremlin propaganda are then passed to four LLMs (grok-4, gpt-5, DeepSeekV3.2, and gemini-3-flash) with a prompt instructing them to treat the source claims in the text as true and then to rewrite those claims as simple, self-contained questions. Additional information on the question generation workflow can be found in the SI Appendix Sections 3 and 4.

We first check that the questions are sufficiently self-contained that they can be meaningfully answered. For example, questions such as "Was the operation done clumsily?" or "Has the war escalated since 2014?" are too ambiguous, potentially leading models to respond about different contexts. These context-dependent questions often rely on unresolved pronouns, vague references, or unstated events or claims.

As the questions are generated by several models, using the same input text, there are frequently highly similar questions. We take additional steps to ensure our questions cover the full range of items discussed, by filtering out near duplicate questions generated from the same input text. First, question strings are normalized (lowercasing, trimming leading and trailing whitespace), and exact duplicates are removed. The remaining questions are embedded using the SentenceTransformer model all-mpnet-base-v2. We identify near-duplicates based on the cosine similarity between embeddings. For each question, we retrieve up to 25 nearest neighbors and create edges for pairs with cosine similarity $>= 0.90$. These pairwise links are then converted into clusters by identifying connected components. Any questions connected directly or indirectly through similarity links are assigned to the same cluster (transitive clustering). Within each cluster of near duplicates, we sample a single representative question. In the SI Appendix Section 4, we present the textual features of the samples across the stages of the pipeline. We then randomly sample from this set of deduplicated questions to generate the sample of questions used in our analysis. For each headline, we sample up to three questions. For the longer EUvsDisinfo narratives, we sample up to five questions.

In Figure 6, we provide an example of our question generation and selection workflow. This

workflow is repeated for each EUvsDisinfo entry and news headline in our study.

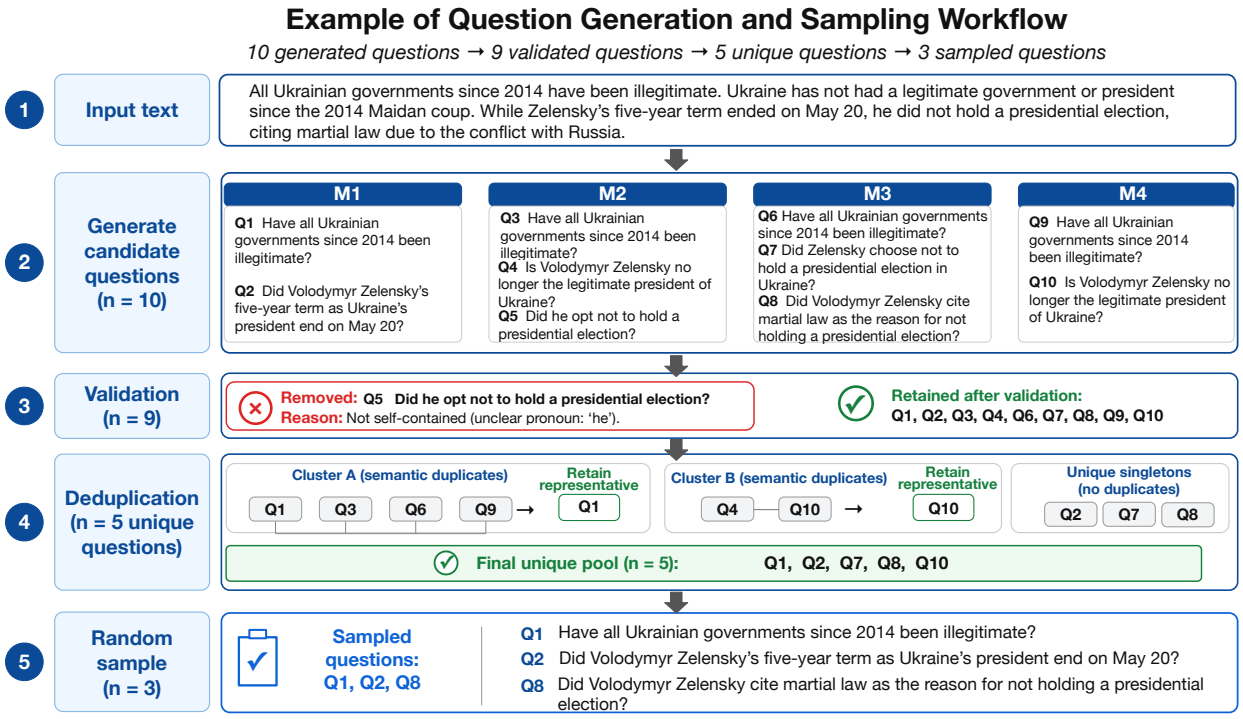


Figure 6: Question generation and selection workflow. Example with one simulated input text from EUvsDisinfo. In Step 2, the text is sent to multiple LLMs (M1-M4) which generate candidate questions based on the main ideas in the text. In Step 3, the questions are validated to ensure they are self-contained. In Step 4, near-duplicate questions are identified by grouping questions based on cosine similarity, and sampling one representative question from each cluster. In Step 5, a random sample is drawn from the validated, deduplicated question pool and included in our study.

5.4 LLM Responses Evaluated

We evaluate the following five LLMs: gpt-5.3-chat (model version 2026-03-03), DeepSeek-V3.2 (model version 1), claude-sonnet-4-6 (model version 1), grok-4-20-non-reasoning (model version 1), gemini-3-flash-preview.

The models cover the leading US-based frontier AI labs (OpenAI, Google, Anthropic, and xAI), as well as an open-weight model developed in China, DeepSeek.

All models are accessed via API. With the exception of Gemini, which is accessed through Google Cloud Vertex, models were accessed through Azure Foundry. Default temperature settings were used for each model.

5.5 Generating Responses

Models are prompted to respond to a provided question and, based on this answer, provide a categorical label that best represents the response, one of 'Affirm', 'Negative', or 'Not Enough Information'. For each question-model pair, we have a full textual response (Response) and a categorical verdict (Verdict).

5.6 Identifying Questions Related to Contested Content

Generating questions from Kremlin propaganda narratives does not guarantee that each question captures a contested propaganda claim. For example, a Kremlin narrative contained in the EUvsDisinfo database, "Luxembourg's participation in illegal anti-Russian sanctions damages its reputation" might plausibly generate a question such as "Is Luxembourg one of the EU's key financial centres?" While taken from a Kremlin propaganda narrative, it is a simple factual issue rather than one that is contested in Kremlin propaganda. This distinction is important because otherwise we risk conflating model agreement on simple factual questions with agreement with Kremlin propaganda.

To identify questions linked to contested propaganda claims, we leverage the paired structure of EUvsDisinfo database entries. Each case contains both a summary of the pro-Kremlin narrative and a response section that rebuts it. Each question generated from a given pro-Kremlin narrative is passed to an LLM along with either the pro-Kremlin narrative or the EUvsDisinfo counter-narrative. The model is asked to review the provided text and assess whether the context supports a "yes" answer (Affirmative), a "no" answer (Negative), or Not Enough Information if the context does not clearly support either response. Each question is run three times and the majority verdict is selected. Questions with ties between categories are marked as Not Enough Information. We classify questions as contested when the answers based on the two contexts differ. For example, the pro-Kremlin narrative yields an Affirmative response and the EUvsDisinfo counter-narrative yields a Negative response, or vice versa. We exclude cases where either context yields a verdict of Not Enough Information, because such cases do not cleanly distinguish between the narrative

and the counter-narrative. This procedure retains questions that more plausibly capture contested propaganda-linked claims. In the SI Appendix Section 10.7, we reproduce the main results using all of the EUvsDisinfo generated questions.

5.7 Measures

5.7.1 Weighting

As we conduct stratified sampling to select input content, in the main results we use the population counts of texts in each stratum to reweight our final estimates. Each sampled question receives a weight equal to the number of population texts in its stratum divided by the number of sampled questions from that stratum. This ensures that strata contribute to the final estimate in proportion to their prevalence in the population, rather than in the sample. In the SI Appendix Section 10.5, we reproduce the results in the main paper without applying this reweighting process. The results are consistent with those presented in the main text. In the SI Appendix Section 10.6, we also weight inversely proportional to the topic prevalence in the population. This prevents topics highly covered by EUvsDisinfo from driving the results. The results are consistent with those presented in the main text.

5.7.2 Agreement with Estimated Kremlin Verdicts

As noted previously, for each question, the models produce both a longer textual response and a categorical verdict that best represents the response, one of 'Affirm', 'Negative', or 'Not Enough Information'. Further, for each question, we have an estimate of the Kremlin response to that question, generated by providing gpt-5.4-mini with the relevant Kremlin propaganda narrative and having it choose a response based only on the narrative content. Model responses are classified as aligned with the Kremlin when they have the same verdict as the estimated Kremlin verdict for the same question, i.e., both return Affirm or Negative. We assess the quality of the estimated Kremlin verdicts by annotating a random sample of 150 questions and model responses. For each question and response, a human evaluator read the EUvsDisinfo narrative used to generate the question and assessed whether the narrative supports an Affirmative or Negative answer. The Kremlin verdict

measure aligns with the human coding for 98.7% of sampled responses.

Looking only at categorical outcomes risks missing additional context provided in the full text of the model responses. In particular, models may align with the Kremlin position directionally, but provide content that is unlikely to be included in these sources or directly contradicts them, e.g. by noting that while an election was held, it was not free and fair. We consider responses qualified if the categorical verdict produced by the model aligns with the estimated Kremlin verdict, but the text provides additional context that weakens the stance. We prompt gpt-5.4 to identify seven types of qualifications in model responses. These are collapsed into a binary measure, 1 if any of these qualifications are present and 0 otherwise. The categories include: Uncertainty: the response explicitly states that the issue is disputed, contested, controversial, unresolved, or uncertain. Evidence Strength: response explicitly signals weak, uncertain, unsupported, unverified, or missing evidence. Causal Limit: response explicitly limits a causal claim. Legal Status: response explicitly discusses legality, illegality, unlawfulness, invalidity, sovereignty. Legal Constraint: response conditions the claim on domestic or international legal standards or constraints. Procedural Validity: response explicitly challenges the validity of a process or event by citing compromised conditions. Countervailing Context: response adds explicit context that materially weakens the interpretation by introducing a separate consideration, counterpoint, or alternative framing. In the SI Appendix Section 6, we include examples for each of these categories.

To assess the quality of the verdict qualification measure, we randomly sampled 150 responses generated from the Kremlin Propaganda sample (the only sample with verdicts). Each response was coded by a human evaluator. Overall, the measure performs well with precision of .88, recall of .82, and F1 of .85.

5.7.3 Kremlin Perspectives Present

We record instances where the model response explicitly presents a Kremlin-linked actor (e.g. Kremlin, Putin, Russian officials, Russian state bodies, or explicitly identified Russian or pro-Kremlin outlets) as the source, origin, or endorser of a claim, allegation, narrative, opinion, denial,

or rebuttal. For example, “Russian authorities claimed that...”

We also record instances where the Kremlin perspective is questioned in the model response. This includes noting that the claim is disputed, doubted, rejected, fact-checked, refuted, or lacking evidence. For example, “There is no evidence for the Kremlin’s allegation...” In the SI Appendix Section 7, we include examples for each of these categories.

These categories are identified using gpt-5.4-mini (version 2026-03-17). To assess the quality of the measure, we randomly sampled 150 responses generated from each of our samples (Kremlin Propaganda and News Headlines). Each response was coded by a human evaluator. Overall, the Kremlin perspective measure performs well with precision of .73, recall of 1.0, and F1 of .85. The measure of the Kremlin perspective being questioned also performs well with precision of .86, recall of .95, and F1 of .90.

5.7.4 Citing Kremlin-Linked Domains

We record the number of times a Kremlin-linked domain appears as a citation in model responses for four of five models (at the time of writing Azure Foundry, our platform for accessing models, did not support DeepSeek with web search). Kremlin-linked domains were collected from several public sources. These include technology companies such as Meta [58], government agencies such as the U.S. Cybersecurity and Infrastructure Security Agency [59], VIGINUM [60], and open source investigators including OSCAR Cardiff, DFRLab [61], HarfangLab [62], Sekoia-IO [63], Insikt Group [64], and Qurium [65].

5.7.5 Uncertainty

Uncertainty estimates are crucial to assess the precision of the results. To estimate uncertainty around our estimates, we used a two-stage hierarchical bootstrap. Our data is structured so that multiple questions may be generated from the same underlying headline/EUvsDisinfo entry. Because responses to questions derived from the same text may be correlated, treating all questions as independent may understate uncertainty. The hierarchical bootstrap preserves this nested structure by resampling texts and the questions generated from these texts, rather than treating questions as

independent.

For each bootstrap sample, we sampled, with replacement, the same number of texts as were present in the original sample. For each selected text, we then sampled with replacement the same number of questions originally observed for that text. For each model and question, we calculated the proportion of responses that featured our outcomes of interest. We applied stratum weights (article quarter for Kremlin propaganda questions and outlet type for news headlines) to each bootstrap sample and calculated model-level estimates as weighted means of these question-level proportions. We calculated 95% confidence intervals using the 2.5th and 97.5th percentiles of the bootstrap distribution.

6 Acknowledgments

This work was partially funded by an unrestricted gift from Google's Academic Research Awards (GARA) and a grant for Azure cloud compute from Microsoft's AI for Good Research Lab. Google and Microsoft did not have the right to prepublication approval of this work.

The paper received helpful comments from participants at the University of Pittsburgh's Hub for AI and Data Science Leadership and UNC Chapel Hill IR brownbag.

References

[1] Michael Chui, Lareina Yee, Bryce Hall, and Alex Singla. "The state of AI in 2023: Generative AI's breakout year". In: (2023).

[2] Alexander Bick, Adam Blandin, and David Deming. *The Rapid Adoption of Generative AI*. Working Paper 2024-027. Federal Reserve Bank of St. Louis, Feb. 2025. DOI: 10.20955/wp.2024.027. URL: https://doi.org/10.20955/wp.2024.027.

[3] Associated Press and NORC Center for Public Affairs Research. *How US adults are using AI, according to AP-NORC polling*. Poll fielded July 10–14, 2025. 2025. URL: https://apnews.com/article/229b665d10d057441a69f56648b973e1.

[4] Himani Verma. *AI Usage to Surge with 950 Million Global Users by 2030, Surpassing Earlier Projections*. BusinessABC. Apr. 2025. URL: https://businessabc.net/ai-user-growth-forecast-950-million-2030.

[5] Stephanie Lin, Jacob Hilton, and Owain Evans. "Truthfulqa: Measuring how models mimic human falsehoods". In: *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*. 2022, pp. 3214–3252.

[6] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. "Taxonomy of risks posed by language models". In: *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*. 2022, pp. 214–229.

[7] NORDIS / Faktabaari. *Russia's propaganda sites have infiltrated AI chatbots in the Nordics*. Investigation. May 2025. URL: https://nordisinfo.org/identical-russian-websites-infiltrated-ai-chatbots/ (visited on 03/04/2026).

[8] American Sunlight Project. *PK Report: A Pro-Russia Content Network Foreshadows the Automated Future of Info Ops*. 2025. URL: https://static1.squarespace.com/static/6612cbdfd9a9ce56ef931004/t/67fd396818196f3d1666bc23/1744648558879/PK+Report.pdf (visited on 03/04/2026).

[9] McKenzie Sadeghi. *AI False Information Rate Nearly Doubles in One Year*. Sept. 2025. URL: https://www.newsguardtech.com/wp-content/uploads/2025/09/August-2025-One-Year-Progress-Report-3.pdf.

[10] Pablo Maristany de las Casas. *Investigation | Talking points: When chatbots surface Russian state media*. Digital Dispatch. Institute for Strategic Dialogue (ISD Germany), Oct. 2025. URL: https://www.isdglobal.org/digital-dispatch/investigation-talking-points-when-chatbots-surface-russian-state-media/ (visited on 03/04/2026).

[11] Sophia Freuden, Nina Jankowicz, and Gary Marcus. *Bad Actors are Grooming LLMs to Produce Falsehoods*. Substack post. July 2025. URL: https://americansunlight.substack.com/p/bad-actors-are-grooming-llms-to-produce.

[12] Joseph Menn. *Russia seeds chatbots with lies. Any bad actor could game AI the same way*. 2025. URL: https://www.washingtonpost.com/technology/2025/04/17/llm-poisoning-grooming-chatbots-russia/.

[13] Huw Dylan and Elena Grossfeld. "Revisionist future: Russia's assault on large language models, the distortion of collective memory, and the politics of eternity". In: *Dialogues on Digital Society* 1.3 (2025), pp. 407–412.

[14] Lea Marchl, Ines Chomnalez, and Isis Blachez. *Anthropic's AI Chatbot Is Leaning More on Russian and Iranian Propaganda Sources, NewsGuard Audit Finds*. May 2026. URL: https://www.newsguardtech.com/special-reports/anthropic-ai-chatbot-claude-russia-iran-propaganda/.

[15] Ihor Samokhodskyi. *The AI Lens of Cognitive Warfare: Why LLMs Language Bias Is a Security Risk*. Feb. 2026. URL: https://europeanleadershipnetwork.org/commentary/the-ai-lens-of-cognitive-warfare-why-llms-language-bias-is-a-security-risk/ (visited on 05/05/2026).

[16] Michael Golebiewski and danah boyd danah. *Data Voids: Where Missing Data Can Easily Be Exploited*. Tech. rep. Data & Society, Oct. 2019. URL: https://datasociety.net/library/data-voids/.

[17] McKenzie Sadeghi. *Top 10 Generative AI Models Mimic Russian Disinformation Claims A Third of the Time, Citing Moscow-Created Fake Local News Sites as Authoritative Sources*. 2024. URL: https://www.newsguardtech.com/special-reports/generative-ai-models-mimic-russian-disinformation-cite-fake-news/.

[18] Mykola Makhortykh, Maryna Sydorova, Ani Baghumyan, Victoria Vziatysheva, and Elizaveta Kuznetsova. "Stochastic lies: How LLM-powered chatbots deal with Russian disinformation about the war in Ukraine". In: *Harvard Kennedy School Misinformation Review* 5.4 (2024). DOI: 10.37016/mr-2020-154. URL: https://misinforeview.hks.harvard.edu/article/stochastic-lies-how-llm-powered-chatbots-deal-with-russian-disinformation-about-the-war-in-ukraine/ (visited on 03/04/2026).

[19] Maxim Alyukov, Mykola Makhortykh, Alexandr Voronovici, and Maryna Sydorova. "LLMs grooming or data voids? LLM-powered chatbot references to Kremlin disinformation reflect information gaps, not manipulation". In: *Harvard Kennedy School Misinformation Review* 6.5 (Oct. 2025). DOI: 10.37016/mr-2020-187. URL: https://doi.org/10.37016/mr-2020-187.

[20] Kevin T Greene, Mayana Pereira, Nilima Pisharody, Rahul Dodhia, Juan Lavista Ferres, and Jacob N Shapiro. "Using website referrals to identify unreliable content rabbit holes". In: *Behaviour & Information Technology* 44.7 (2025), pp. 1340–1349.

[21] Leah Siskind. *AI-Amplified Narratives: Measuring Propaganda in LLM Citations*. Memo. Foundation for Defense of Democracies (FDD), Center for Cyber and Technology Innovation, Mar. 2026. URL: https://www.fdd.org/analysis/2026/03/03/ai-amplified-narratives-measuring-propaganda-in-llm-citations/ (visited on 03/04/2026).

[22] EUvsDisinfo. *Nord Stream 1 & 2 Sabotage – Disinformation à la MH-17?* EUvsDisinfo. Oct. 2022. URL: https://euvsdisinfo.eu/nord-stream-1-2-sabotage-disinformation-a-la-mh-17/.

[23] Samy Amanatullah, Serena Balani, Angela Fraioli, Stephanie M. McVicker, and Mike Gordon. *Tell Us How You Really Feel: Analyzing Pro-Kremlin Propaganda Devices and Narratives to Identify Sentiment Implications*. Working Paper. George Washington University: Illiberalism Studies Program, Feb. 2023. URL: https://www.illiberalism.org/wp-content/uploads/2023/02/ILLSP-Working-Paper-No.-14-February-2023.pdf.

[24] Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. "From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023, pp. 11737–11762.

[25] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. "Whose opinions do language models reflect?" In: *International conference on machine learning*. PMLR. 2023, pp. 29971–30004.

[26] Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. "Measuring political bias in large language models: What is said and how it is said". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2024, pp. 11142–11159.

[27] Sean Westwood, Justin Grimmer, and Andrew B. Hall. *Measuring Perceived Slant in Large Language Models Through User Evaluations*. Hoover Institution. 2025. URL: https://www.hoover.org/research/measuring-perceived-slant-large-language-models-through-user-evaluations.

[28] Anthropic. *Measuring Political Bias in Claude*. Anthropic. Nov. 2025. URL: https://www.anthropic.com/news/political-even-handedness.

[29] Abubakar Abid, Maheen Farooqi, and James Zou. "Large language models associate Muslims with violence". In: *Nature Machine Intelligence* 3.6 (2021), pp. 461–463.

[30] Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander Van Der Linden, and Jon Roozenbeek. "Generative language models exhibit social identity biases". In: *Nature Computational Science* 5.1 (2025), pp. 65–75.

[31] Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L Griffiths. "Explicitly unbiased large language models still form biased associations". In: *Proceedings of the National Academy of Sciences* 122.8 (2025), e2416228122.

[32] Aliah Zewail, Alexandra Figueroa, Jesse Graham, and Mohammad Atari. "Moral stereotyping in large language models". In: *Proceedings of the National Academy of Sciences* 123.10 (2026), e2519941123.

[33] Jesutofunmi A Omiye, Jenna C Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. "Large language models propagate race-based medicine". In: *NPJ Digital Medicine* 6.1 (2023), p. 195.

[34] Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt, and Demet Yuksel. "Large language models lack essential metacognition for reliable medical reasoning". In: *Nature communications* 16.1 (2025), p. 642.

[35] Rachel L Draelos, Samina Afreen, Barbara Blasko, Tiffany L Brazile, Natasha Chase, Dimple Patel Desai, Jessica Evert, Heather L Gardner, Lauren Herrmann, Aswathy Vaikom House, et al. "Large language models provide unsafe answers to patient-posed medical questions". In: *npj Digital Medicine* (2026).

[36] Tianyu Han, Sven Nebelung, Firas Khader, Tianci Wang, Gustav Müller-Franzes, Christiane Kuhl, Sebastian Försch, Jens Kleesiek, Christoph Haarburger, Keno K Bressem, et al. "Medical large language models are susceptible to targeted misinformation attacks". In: *NPJ digital medicine* 7.1 (2024), p. 288.

[37] Daniel Alexander Alber, Zihao Yang, Anton Alyakin, Eunice Yang, Sumedha Rai, Aly A Valliani, Jeff Zhang, Gabriel R Rosenbaum, Ashley K Amend-Thomas, David B Kurland, et al. "Medical large language models are vulnerable to data-poisoning attacks". In: *Nature Medicine* 31.2 (2025), pp. 618–626.

[38] Yifan Yang, Qiao Jin, Furong Huang, and Zhiyong Lu. "Adversarial prompt and fine-tuning attacks threaten medical large language models". In: *Nature Communications* 16.1 (2025), p. 9011.

[39] Thomas H Costello, Gordon Pennycook, and David G Rand. "Durably reducing conspiracy beliefs through dialogues with AI". In: *Science* 385.6714 (2024), eadq1814.

[40] Josh A Goldstein, Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. "How persuasive is AI-generated propaganda?" In: *PNAS nexus* 3.2 (2024), pgae034.

[41] Kobi Hackenburg and Helen Margetts. "Evaluating the persuasive influence of political microtargeting with large language models". In: *Proceedings of the National Academy of Sciences* 121.24 (2024), e2403116121.

[42] Michael Henry Tessler, Michiel A Bakker, Daniel Jarrett, Hannah Sheahan, Martin J Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C Parkes, et al. "AI can help humans find common ground in democratic deliberation". In: *Science* 386.6719 (2024), eadq2852.

[43] Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. "On the conversational persuasiveness of GPT-4". In: *Nature Human Behaviour* 9.8 (2025), pp. 1645–1653.

[44] Hui Bai, Jan G Voelkel, Shane Muldowney, Johannes C Eichstaedt, and Robb Willer. "LLM-generated messages can persuade humans on policy issues". In: *Nature Communications* 16.1 (2025), p. 6037.

[45] Andrew Guess, Brendan Nyhan, and Jason Reifler. "Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign". In: *European Research Council* 9.3 (2018), p. 4.

[46] Andrew Guess, Jonathan Nagler, and Joshua Tucker. "Less than you think: Prevalence and predictors of fake news dissemination on Facebook". In: *Science advances* 5.1 (2019), eaau4586.

[47] Andrew M Guess, Brendan Nyhan, Zachary O'Keeffe, and Jason Reifler. "The sources and correlates of exposure to vaccine-related (mis) information online". In: *Vaccine* 38.49 (2020), pp. 7799–7805.

[48] Andrew M Guess, Brendan Nyhan, and Jason Reifler. "Exposure to untrustworthy websites in the 2016 US election". In: *Nature human behaviour* 4.5 (2020), pp. 472–480.

[49] Kevin Greene and Jacob Shapiro. "Assessing US Online Engagement with Foreign Malign Influence Operations". In: (2026).

[50] VIGINUM. *Portal Kombat: A Structured and Coordinated Pro-Russian Propaganda Network*. Technical report. Secrétariat général de la défense et de la sécurité nationale, Feb. 2024. URL: https://www.sgdsn.gouv.fr/files/files/20240212_NP_SGDSN_VIGINUM_PORTAL-KOMBAT-NETWORK_ENG_VF.pdf.

[51] Valentin Châtelet and Amaury Lesplingart. *Russia's So-Called "Pravda" Network Expands Worldwide*. Feb. 2025. URL: https://dfrlab.org/2025/02/24/russia-pravda-network-expands-worldwide/.

[52] VIGINUM. *Portal Kombat: Expansion of the pro-Russian propaganda network: new domain names*. Technical report. Secrétariat général de la défense et de la sécurité nationale, 2024. URL: https://www.sgdsn.gouv.fr/files/files/Publications/20240428_NP_SGDSN_VIGINUM_PORTAL-KOMBAT-NETWORK-REPORT_NEW%20DOMAIN%20NAME_%28PART3%29_ENG_VF.pdf.

[53] Media Bias/Fact Check. *Interfax – Russia – Bias and Credibility*. https://mediabiasfactcheck.com/interfax-russia-bias/. Last updated February 28, 2025. Accessed May 14, 2026. Feb. 2025.

[54] Annie Y Chen, Brendan Nyhan, Jason Reifler, Ronald E Robertson, and Christo Wilson. "Subscriptions and external links help drive resentful users to alternative and extremist YouTube channels". In: *Science Advances* 9.35 (2023), eadd8080.

[55] Ronald E Robertson, Jon Green, Damian J Ruck, Katherine Ognyanova, Christo Wilson, and David Lazer. "Users choose to engage with more partisan news than they are exposed to on Google Search". In: *Nature* 618.7964 (2023), pp. 342–348.

[56] Kevin T Greene, Nilima Pisharody, Lucas Augusto Meyer, Mayana Pereira, Rahul Dodhia, Juan Lavista Ferres, and Jacob N Shapiro. "Current engagement with unreliable sites from web search driven by navigational search". In: *Science advances* 10.44 (2024), eadn3750.

[57] Hause Lin, Jana Lasser, Stephan Lewandowsky, Rocky Cole, Andrew Gully, David G Rand, and Gordon Pennycook. "High level of correspondence across different news domain quality rating sets". In: *PNAS nexus* 2.9 (2023), pgad286.

[58] Meta. *Meta Threat Research Indicator Repository*. URL: https://github.com/facebook/threat-research.

[59] CISA. *Just So You Know: Foreign Threat Actors Likely to Use a Variety of Tactics to Develop and Spread Disinformation During 2024 U.S. General Election Cycle*. Tech. rep. Cybersecurity and Infrastructure Security Agency, Oct. 2024. URL: https://www.cisa.gov/sites/default/files/2024-10/PSA_Just%20So%20You%20Know_Foreign_Threat_Actors_Likely_to_Use_a_Variety_of_Tactics.pdf.

[60] SGDSN. *RRN: a complex and persistent digital information and manipulation campaign*. Tech. rep. General Secretariat for Defence and National Security, June 2023. URL: https://www.sgdsn.gouv.fr/publications/maj-19062023-rrn-une-campagne-numerique-de-manipulation-de-linformation-complexe-et.

[61] Valentin Châtelet and Amaury Lesplingart. *Russia's so-called "Pravda" network expands worldwide*. Publication Title: DFRLab. Feb. 2025. URL: https://dfrlab.org/2025/02/24/russia-pravda-network-expands-worldwide/ (visited on 04/07/2025).

[62] HarfangLab. *Indicators of Compromise*. URL: https://github.com/HarfangLab/iocs.

[63] SEKOIA-IO. *Doppelganger Observables*. URL: https://github.com/SEKOIA-IO/Community/blob/main/IOCs/Doppelg%C3%A4nger/DoppelG%C3%A4nger-observables.csv.

[64] Insikt Group®. *Russia-Linked CopyCop Uses LLMs to Weaponize Influence Content at Scale \textbar Recorded Future*. en. Publication Title: RecordedFuture. May 2024. URL: https://www.recordedfuture.com/research/russia-linked-copycop-uses-llms-to-weaponize-influence-content-at-scale (visited on 03/03/2025).

[65] Qurium. *Under the hood of a Doppelgänger*. Sept. 2022. URL: https://www.qurium.org/alerts/under-the-hood-of-a-doppelganger/ (visited on 06/30/2025).