

Detecting pro-kremlin disinformation using large language models

Authors	Kramer,Marianne; Golovchenko,Yevgeniy; Hjorth,Frederik
Published in	Research & Politics
DOI	10.1177/20531680251351910
Publication Date	2025-04
Document Version	publishersversion
Link	https://research.tilburguniversity.edu/en/publications/4d5b5f0c-5b11-4174-983d-f1437a3b716f
Citation	Kramer, M, Golovchenko, Y & Hjorth, F 2025, 'Detecting pro-kremlin disinformation using large language models', Research & Politics, vol. 12, no. 2. https://doi.org/10.1177/20531680251351910
Download Date	2026-06-18 18:14:20
Rights	<p>General rights</p> <p>Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.</p> <ul style="list-style-type: none"> - Users may download and print one copy of any publication from the public portal for the purpose of private study or research. - You may not further distribute the material or use it for any profit-making activity or commercial gain - You may freely distribute the URL identifying the publication in the public portal" <p>Take down policy</p> <p>If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.</p>

Detecting pro-kremlin disinformation using large language models

Research and Politics
April-June 2025: 1–6
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20531680251351910
journals.sagepub.com/home/rap

Marianne Kramer^{1,2} , Yevgeniy Golovchenko¹ and Frederik Hjorth¹

Abstract

A growing body of literature examines manipulative information by detecting political mis-/disinformation in text data. This line of research typically involves highly costly manual annotation of text for manual content analysis, and/or training and validating automated downstream approaches. We examine whether Large Language Models (LLMs) can detect pro-Kremlin disinformation about the war in Ukraine, focusing on the case of the downing of the civilian flight MH17. We benchmark methods using a large set of tweets labeled by expert annotators. We show that both open and closed LLMs can accurately detect pro-Kremlin disinformation tweets, outperforming both a research assistant and supervised models used in earlier research and at drastically lower cost compared to either research assistants or crowd workers. Our findings contribute to the literature on mis-/disinformation by showcasing how LLMs can substantially lower the costs of detection even when the labeling requires complex, context-specific knowledge about a given case.

Keywords

Large language models, misinformation, social media

Introduction

Misinformation and disinformation is a substantial research topic in political science and other fields. In the context of Russia's ongoing war against Ukraine, this broad research agenda is accompanied by a more focused research interest in pro-Kremlin disinformation and related foreign influence campaigns (e.g., [Golovchenko et al., 2018](#)). Research interest in pro-Kremlin disinformation has centered on the diffusion of disinformation on social media, which is an important vector for the spread of disinformation and related pro-Kremlin election manipulation efforts ([Hjorth and Adler-Nissen, 2019](#); [Stukal et al., 2017](#)). While disinformation consists of information that is intentionally misleading, misinformation lacks the intended purpose to mislead (see [Søe \(2016\)](#) and [Fallis \(2015\)](#) for an in-depth discussion of the concepts).

Efforts to measure dis-/misinformation across text corpora can take various forms. The most thorough and context-sensitive approach is manual labeling of texts by

trained research assistants or untrained crowd workers ([Benoit et al., 2016](#)). Alternatively, researchers can use methods from Natural Language Processing (NLP) to classify social media posts with machine learning methods. Here, manual annotations serve either as a training set for the NLP model or as a benchmark used to evaluate the accuracy of the automated classifier.

Recent research suggests that Large Language Models (LLMs) can be successfully used for fact-checking ([Hoes](#)

¹Department of Political Science and Centre for Social Data Science (SODAS), University of Copenhagen, Denmark

²Department of Cognitive Science & Artificial Intelligence, Tilburg University, the Netherlands

Corresponding author:

Yevgeniy Golovchenko, Department of Political Science and Centre for Social Data Science (SODAS), University of Copenhagen, Oester Farimagsgade 5, Copenhagen 1353, Denmark.
Email: yg@ifs.ku.dk



et al., 2023) as well as broader labeling tasks, including the identification of topics and frames (Gilardi et al., 2023). However, the performance of LLM classifiers varies greatly across tasks (Ollion et al., 2023). Because of this task-specific performance, evidence is required to assess the usefulness of LLMs for detecting dis-/misinformation. Despite the rapidly increasing interest in LLMs and the urgent need to map Russian disinformation campaigns, there are to the best of our knowledge currently no studies that test the performance of LLM at detecting pro-Kremlin disinformation.

In this letter, we address this lack of evidence by exploring to what extent LLMs can lower the cost associated with detecting pro-Kremlin disinformation. We draw on a data set of expert-annotated tweets about the 2014 downing of flight MH17 by Russian-controlled forces. The case is politically significant because it increased global attention to the ongoing war in Ukraine before the full-scale invasion in 2022 and put Russia's political prestige at stake. Moreover, the case is analytically useful because the downing of MH17 was subject to a thorough subsequent investigation, so the facts of the case are well-identified and widely reported in the media. As we will explain further in the discussion, the model could be less useful in more recent cases, where there is less publicly available information about the facts. Our findings plausibly generalize to other cases where researchers seek to classify misleading information about a well-documented case. Nevertheless, disinformation in the tweets themselves is mixed with elements of truth, and draws on complex cultural and factual references. This expectation is accentuated by the short, contextualized nature of our text data, which takes the form of tweets no more than 140 characters long.¹ We demonstrate that LLM's can reach high performance even in a challenging labeling test, where the human annotator would have to rely on thorough contextual knowledge and interpretation of implied meaning.

We find that LLMs can classify pro-Kremlin disinformation with high accuracy at minimal cost. The best performing model, GPT-4, identifies 81% of all the pro-Kremlin disinformation tweets in the query data set (i.e., the model's recall), and 94% of pro-Kremlin disinformation tweets were labeled as such by the model (its precision). We highlight that GPT-4's comes with decreased transparency as a trade-off due to its closed, proprietary structure (Barrie et al., 2024). However, we also demonstrate that the high performance is not specific to proprietary LLMs. Llama 3 70B, an open-source and versionable LLM, has an only slightly lower recall of 76% and a precision of 93%. Similarly, both LLM models perform well at identifying counter-disinformation tweets and unrelated content. In the context of the significant concerns about social science relying on closed LLMs for annotation (e.g., Barrie

et al., 2024; Palmer et al., 2024), we reassuringly find that an open model achieves nearly the same performance as the frontier model GPT-4.

Our findings contribute to the literature on manipulative information, pro-Kremlin disinformation, and the emerging literature on the use of LLMs for annotation tasks (Ollion et al., 2023). Our findings suggest that LLMs can drastically lower the cost of labeling disinformation content at an unprecedented speed, even in instances where case-specific, contextual knowledge is required. Minimizing cost is a particularly acute concern in disinformation research, where the relative rarity of the target class has stifled earlier efforts to train classifiers based on text features (e.g., Hartmann et al., 2019). The results show potential for cases where imbalance renders training data costly to obtain.

Measuring dis-misinformation

The literature offers many methods for measuring and operationalizing mis-/disinformation in social media posts. To contextualize our approach, we cover the most prominent examples here. To structure our review, we distinguish between *source-centered* approaches on the one hand and *content-centered* on the other.

The source-centered approach classifies social media posts of hyperlinks from non-credible sources as proxies for misinformation, disinformation, fake news, or junk news (e.g., Grinberg et al., 2019; Guess et al., 2018). This approach is widely used, likely because it avoids the labor-intensive task of evaluating individual content in the posts produced by each source. However, the approach has shortcomings. First, not all content from non-credible information sources is inherently misleading. Second, the exact number of sources in the lists may be arbitrary. Third and lastly, misinformation may have a wider reach precisely when it spreads through established or credible sources.

In contrast, the content-centered approach relies on analyzing the content of the post or the hyperlink to identify manipulative information (Margolin et al., 2018). Content-centered approaches often rely on manual annotation of individual pieces of content to avoid inferring from the source alone. Manual annotation remains one of the most context-sensitive, source-agnostic, and in-depth approaches. Relative to the source-centered approach, researchers have more control over which data are selected for annotation and which instructions/criteria should be used to annotate the text. However, it is currently also the most labor-intensive and, therefore, least scalable approach.

To minimize costs, some researchers have turned to crowd worker services. For example, Gilardi et al. (2023) report a price of approximately US\$0.05 per annotation at MTurk. However, recent research challenges the annotation quality of crowd workers (e.g., Kennedy et al., 2020). There are two reasons to expect this to be particularly pertinent

with respect to misinformation research. First, the quality issue is especially acute when the labeling task requires “contextual and conceptual knowledge” about the respective case (Marquardt et al., 2017), a typical constraint when detecting misinformation narratives such as those related to the war between Russia and Ukraine. Second, while misinformation can be consequential it is typically rare as a proportion of non-misinformation about a given case. For example, posts by Russian trolls from the Internet Research Agency in the runup to the 2016 US presidential election represented 0.004% of the content that US citizens may have seen in the Facebook news feed according to the platform (Solon and Siddiqui, 2017). This statistical power demand means that even small differences in the marginal cost of annotation can dramatically raise the cost of an annotation task.

We now turn to our comparison of LLMs against annotations by a student research assistant (RA) as well as predictions from a convolutional neural network (CNN) trained on tweets’ text features.

Empirical approach

Ground truth data

To compare measurement approaches, we benchmark all against a “ground truth” data set of 10,018 tweets annotated by a set of four expert coders. To establish intercoder reliability, all expert coders annotated a sample of 500 tweets, yielding a reasonably high inter-annotator rate at Fleiss’ $\omega = 0.77$. (See Golovchenko et al. (2018) for more details on the original coding and Hartmann et al. (2019) and Hjorth and Adler-Nissen (2019) for additional applications). The tweets were classified as *disinformation* if the message implied that the MH17 plane was shot down by EU, NATO or Ukraine, as *counter-disinformation* if the tweets implied that Russian authorities or separatists in Ukraine were responsible for shooting down MH17, and as *unrelated* if they did not imply any of the above. It is important to stress that the human annotation goes beyond identifying explicitly false statements, since the annotators are required to use contextual knowledge to interpret *implied* meaning. Consider the following disinformation tweet as an example: “So

you want me to believe Rebel Groups without any support have Anti Aircraft guns... #MH17 #Russia #Ukraine.” While the explicit meaning in the sentence is not necessarily false per se (i.e., the author may indeed be skeptical), it is still misleading because it implies that MH17 was not shot down by Russian-controlled rebels.

We include the full original codebook in [Online Appendix B](#). [Table 1](#) presents examples of each type of annotated tweet. Of the original 10,018 annotated tweets, we evaluate a sample of 300 tweets stratified by category.

Instructions and prompts

The RA receives the same annotation manual as originally used by the expert annotators. For both the RA and the LLMs, the tweets are kept in the original format, including hashtags and usernames. For GPT, we test both GPT-3 and GPT-4. For Llama 3, we test the 70B and 8B versions. For all models, we set the temperature to 0. To minimize costs, we bundle 20 tweets with every API call along with the instructions. We add the definitions of all categories to the instruction prompt, which is shortened version of the codebook given to the RA. Notably, the prompt uses the terminology of “pro-Ukrainian” for counter-disinformation, and “pro-Russian” for “disinformation.” These changes in terminology are in line with what was used by Golovchenko et al. (2018). By defining the prompt in terms of these stances, we are able to evade widespread language model guardrails inhibiting position-taking on mis- or disinformation. This solution is straightforward because the specific facts of the case of the MH17 crash mean that pro-Russian and pro-Ukrainian stances map cleanly to a classification as disinformation. Hence, an important scope condition of our approach is that researchers need to be able to define a stance that represents mis- or disinformation equally unambiguously.

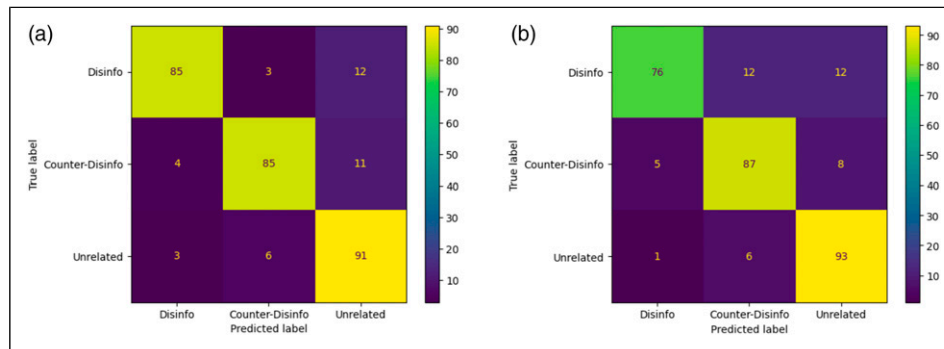
While it is possible only to include the tweets that have to be annotated in the prompt, we take a “few-shot learning,” adding labeled example tweets to every prompt. With large language models, few-shot learning has traditionally outperformed zero-shot learning (Brown et al., 2020). We share the final prompt in [Online Appendix A](#).

Table 1. Example Tweet for Each Category.

Category	Example Tweet
Disinformation	@bydlozavr MH17 was shot down accidentally by Ukraine trying to kill Putin on another plane. Transponders were switched to fool enemy.
Counter-Disinformation	Excellent technical analysis of Why a #Russian BUK missile shot down #MH17
Unrelated	Bodies removed from MH17 crash site in Ukraine: Scores of bodies recovered at the crash site of Malaysian plan...

Table 2. Compares Performance Across Approaches.

Annotator	Disinformation		Counter-disinformation		Unrelated	
	Precision	Recall	Precision	Recall	Precision	Recall
GPT-4	0.94	0.81	0.87	0.88	0.80	0.90
Llama 3 70B	0.93	0.76	0.83	0.87	0.82	0.93
RA	0.87	0.60	0.76	0.90	0.78	0.87
GPT-3	0.67	0.70	0.70	0.62	0.81	0.88
Llama 3 8B	0.61	0.54	0.62	0.48	0.66	0.88
CNN	0.53	0.57	0.50	0.70	0.95	0.90

**Figure 1.** Comparison of confusion matrices for GPT-4 and Llama 3 70B (a) Confusion matrix for GPT-4, (b) Confusion matrix for Llama 3 70B.

For both models, we let the LLMs annotate the same data multiple times to check for consistency, and the performance metrics remain largely the same. As an additional test, we rerun the annotations excluding the Twitter handles in the Tweets, as well as any links in the text. This did not impact the results of the analysis.

Results

Overall performance

Table 2 presents a result comparison between annotators. CNN performance is based on Hartmann et al. (2019), calculated based on the larger dataset. Rows are ordered by precision with respect to the “Disinformation” category.

As shown, the CNN classifier used in Hartmann et al. (2019) performs worst of all approaches, far below the accuracy of a student RA. However, the best performing LLMs exceed the accuracy of an RA. Comparing open and closed LLMs, we find that the performance of Llama 3 70B is similar to GPT-4. For the disinformation and counter-disinformation categories, GPT-4 performs slightly better than Llama 3 70B, whereas Llama 3 70B performs better on the unrelated category. Since the disinformation category

will typically be of primary interest to researchers, this amounts to a minor edge to GPT-4 in performance.

Comparing across model generations, Table 2 reveals substantial improvements. Llama 3 8B does not improve much on the CNN, and on some metrics still performs worse, though it does not over-predict the unrelated class. These results indicate that, while trained similarly, the number of parameters contribute a significant amount to the performance here. Similarly, GPT-4 performs substantively better than GPT-3.

Misclassification

To shed light on the cases where LLMs misclassify content, Figure 1 presents confusion matrices the two best performing LLMs. The confusion matrix visually compares predicted labels with true labels, showing how often it correctly or incorrectly classifies each category. Beyond performance metrics, such as accuracy, it highlights misclassifications and reveals which labels are often confused with others. As shown in Figure 1(b), 76 out of 100 disinformation tweets have been classified as such by the model.

As can be seen in Figure 1(a), there is relatively little misclassification between disinformation and counter-disinformation, indicating that there is a good

distinction between the two classes for GPT-4. In Figure 1(b), it can be seen that the disinformation that is misclassified by Llama 3 70B, is equally classified as counter-disinformation and unrelated. Hence, Llama 3 70B is slightly more likely to misclassify disinformation as counter-disinformation.

Inspection of the misclassified results by both models shows tweets that are hard to interpret, even for experts. Overall, LLMs are able to correctly identify disinformation and counter-disinformation with the information given. Compared to the research assistant, the models are able to classify correctly tweets that require more in-depth case knowledge. At times, the tweets require more knowledge about case MH17 to decide on blame attribution, which the models have to a certain extent. An example of a tweet that is misclassified says: “Someone from a Russian government IP address edited Wikipedia to blame MH17 on Ukraine (...).” This tweet is categorized as counter-disinformation but is predicted to be disinformation. While the conclusion of the tweet is that the Russian government may be involved in spreading false information, this conclusion is left implicit. Misclassifications like this show that there is still room for improvement. However, overall, we see the results are strong and consistent.

Conclusion and discussion

Classifying dis/misinformation on social media is a crucial task for misinformation scholars, but the need for human annotators can make this task very costly. This study evaluates the ability of LLMs to classify pro-Kremlin disinformation. We find that both open and closed state-of-the-art LLMs perform very well at this task. Specifically, GPT-4 and Llama 3 70B are capable of identifying 81% and 76% respectively of all the tweets with pro-Kremlin disinformation about the downing of flight MH17. Of all the tweets classified as pro-Kremlin disinformation by the GPT-4 and Llama 3 70B, 94% and 93%, respectively, are also classified as such by the expert annotators. Compared to a student RA, LLMs deliver far higher accuracy at a fraction of the cost. Moreover, the two LLMs substantively outperform the results produced by other NLP models presented in previous research for the same annotation tasks.

The results for both LLMs are particularly impressive given that human interpretation of such tweets requires in-depth knowledge about the specific context. Furthermore, the annotation task is challenging because the research design defines disinformation broadly as misleading information, where the manipulative message can be implied in between the lines. Especially for research areas in which data is hard to attain and even harder to annotate, this is a promising new solution.

Our findings come with some caveats. First, we stress that the model may be less useful for more recent

disinformation cases where the LLM has less access to the relevant training data. Second, we stress that while researchers may gain unprecedented increases in labeling performance, they also lose portions of transparency and reproducibility due to the intransparent and constantly evolving architecture of LLMs. Because the best performing model, GPT-4, is offered as a service by OpenAI, the model itself may change, making research results less reproducible (Palmer et al., 2024). Moreover, the exact layout of the commercial model is unknown to the public, especially for GPT-4. However, these concerns are at least partially alleviated by the freely available, open-source model, Llama 3 70B at the cost of only a relatively small drop in performance. The free accessibility also means that it remains possible to fine-tune this model for downstream tasks, or access fine-tuned models by others. Lastly, our comparison does not consider fine-tuned BERT-based models, which may offer classification performance that is competitive to LLMs (Laurer et al., 2024) “albeit at a higher upfront annotation cost than our few-shot approach.”

These caveats notwithstanding, we show that LLMs can serve as a high-performance and low-cost tool for labeling text about historic, well-known yet highly complex events. Researchers should carefully consider the relevant trade-offs, alternatives, and budget for at least some manual labeling for the purpose of validating the results.

Acknowledgements

We would like to thank the Copenhagen Center for Social Data Science and The Hub for Mis- and Disinformation Research for support and continuous feedback. The first author, Marianne Kramer, has an *M.Sc. in Social Data Science* from the University of Copenhagen and we would like to express our gratitude to the program for their support. Lastly, we would like to thank all of the participants at the De-Conspirator “Foreign Information Manipulation-Interference research methods workshop” in 2024 for their feedback.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ORCID iDs

Marianne Kramer  <https://orcid.org/0009-0006-5900-3638>
Yevgeniy Golovchenko  <https://orcid.org/0000-0003-3292-7372>
Frederik Hjorth  <https://orcid.org/0000-0003-4063-4983>

Supplemental Material

Supplemental material for this article is available online.

The replication files are available at: <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/WL3KLH>

Carnegie Corporation of New York

This publication was made possible (in part) by a grant from the Carnegie Corporation of New York. The statements made and views expressed are solely the responsibility of the author.

Note

1. Tweets were 140 characters long at the time of the data collection, however, the standard size as of 2024 is 280 characters.

References

- Barrie C, Palmer A and Spirling A. (2024). *Replication for Language Models: Problems, Principles, and Best Practice for Political Science*. Arthur Spirling. [Unpublished Manuscript]. https://arthurspirling.org/documents/BarriePalmerSpirling_TrustMeBro.pdf
- Benoit K, Conway D, Lauderdale BE, et al. (2016) Crowdsourced Text Analysis: Reproducible and Agile Production of Political Data. *American Political Science Review* 110(2): 278–295.
- Brown T, Mann B, Ryder N, et al. (2020) Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems* 33: 1877–1901.
- Fallis D (2015) What is Disinformation? *Library Trends* 63(3): 401–426.
- Gilardi F, Alizadeh M and Kubli M (2023) ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks. *Proceedings of the National Academy of Sciences of the United States of America* 120(30): e2305016120.
- Golovchenko Y, Hartmann M and Adler-Nissen R (2018) State, Media and Civil Society in the Information Warfare over Ukraine: Citizen Curators of Digital Disinformation. *International Affairs* 94(5): 975–994.
- Grinberg N, Joseph K, Friedland L, et al. (2019) Fake News on Twitter During the 2016 US Presidential Election. *Science* 363(6425): 374–378.
- Guess A, Nyhan B and Reifler J (2018) Selective Exposure to Misinformation: Evidence from the Consumption of Fake News During the 2016 US Presidential Campaign. *European Research Council* 9(3): 4.
- Hartmann M, Golovchenko Y and Augenstein I (2019) Mapping (Dis-) Information Flow About the MH17 Plane Crash. *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*. DOI: [10.18653/v1/D19-5006](https://doi.org/10.18653/v1/D19-5006).
- Hjorth F and Adler-Nissen R (2019) Ideological Asymmetry in the Reach of Pro-Russian Digital Disinformation to United States Audiences. *Journal of Communication* 69(2): 168–192.
- Hoes E, Altay S and Bermeo J (2023) Leveraging ChatGPT for Efficient Fact-Checking. *PsyArXiv*. DOI: [10.31234/osf.io/qnjkf](https://doi.org/10.31234/osf.io/qnjkf).
- Kennedy R, Cliord S, Burleigh T, et al. (2020) The Shape of and Solutions to the MTurk Quality Crisis. *Political Science Research and Methods* 8(4): 614–629.
- Laurer M, Van Atteveldt W, Casas A, et al. (2024) Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning With Deep Transfer Learning and BERT-NLI. *Political Analysis* 32(1): 84–100.
- Margolin DB, Hannak A and Weber I (2018) Political Fact-Checking on Twitter: When Do Corrections Have an Effect? *Political Communication* 35(2): 196–219.
- Marquardt KL, Pemstein D, Sanhueza Petrarca C, et al. (2017) Experts, Coders, and Crowds: An Analysis of Substitutability. V-Dem Working Paper 2017:53. [Unpublished Manuscript]. DOI: [10.2139/ssrn.3046462](https://doi.org/10.2139/ssrn.3046462).
- Ollion E, Shen R, Macanovic A, et al. (2023) ChatGPT for Text Annotation? Mind the Hype! *SocArXiv*. DOI: [10.31235/osf.io/x58kn](https://doi.org/10.31235/osf.io/x58kn).
- Palmer A, Smith NA and Spirling A (2024) Using Proprietary Language Models in Academic Research Requires Explicit Justification. *Nature Computational Science* 4(1): 2–3.
- Søe SO (2016) *The Urge to Detect, the Need to Clarify: Gricean Perspectives on Information, Misinformation and Disinformation*. Ph. D. thesis. University of Copenhagen.
- Solon O and Siddiqui S (2017) *Russia-Backed Facebook Posts 'Reached 126M Americans' During US Election*. The Guardian 2017.
- Stukal D, Sanovich S, Bonneau R, et al. (2017) Detecting Bots on Russian Political Twitter. *Big Data* 5(4): 310–324.