Are Large Language Models Good at Detecting Propaganda?

Julia Jose, Rachel Greenstadt

New York University
Department of Computer Science and Engineering
New York, New York, USA
julia.jose@nyu.edu, greenstadt@nyu.edu

Abstract

Propagandists use rhetorical devices that rely on logical fallacies and emotional appeals to advance their agendas. Recognizing these techniques is key to making informed decisions. Recent advances in Natural Language Processing (NLP) have enabled the development of systems capable of detecting manipulative content. In this study, we look at several Large Language Models and their performance in detecting propaganda techniques in news articles. We compare the performance of these LLMs with transformer-based models. We find that, while GPT-4 demonstrates superior F1 scores (F1=0.16) compared to GPT-3.5 and Claude 3 Opus, it does not outperform a RoBERTa-CRF baseline (F1=0.67). Additionally, we find that all three LLMs outperform a Multi-Granularity Network (MGN) baseline in detecting instances of one out of six propaganda techniques (name-calling), with GPT-3.5 and GPT-4 also outperforming the MGN baseline in detecting instances of appeal to fear and flag-waving.

Introduction

Propaganda is defined by Jowett & O'Donnell (2006) as "*the deliberate, systematic attempt to shape perceptions, manipulate cognitions, and direct behavior to achieve a response that furthers the desired intent of the propagandist*" (Jowett and O'Donnell 2018). Manipulative information, as seen in the 2016 US presidential campaign, can harm society by influencing public opinion (Brattberg and Maurer 2018; Golovchenko et al. 2020). Propaganda, however, is not a new concept and dates back to the Second World War. Scholars and educators around then emphasized the need to be able to identify propaganda (Graham 1939; Booth; 1940; Hollis; 1939). One of the most prominent detection strategies was to analyze pieces of information for the use of propaganda techniques and to "guard against them" (Graham 1939). The Institute of Propaganda Analysis (IPA) was formed with this goal in mind. Outlined in their book *The Fine Art of Propaganda: A Study of Father Coughlin's Speeches* (Lee and Lee 1939) are seven such propaganda techniques. Examples of some of these techniques include —

- Name-Calling: "Giving an idea a bad label and therefore rejecting and condemning it without examining the evidence."
- Bandwagon: "Has as its theme 'everybody - at least all of us - is doing it!' and thereby tries to convince the members of a group that their peers are accepting the program and that we should all jump on the bandwagon rather than be left out."

Building on this work, Da San Martino et al. (2019) derived eighteen such propaganda techniques that are widely used in news articles today.

Analyzing news articles to check for the presence of propaganda techniques (Da San Martino et al. 2020) could help users gain a deeper understanding of the information's quality. To automate labeling these techniques in news articles at large scale, we aim to measure the effectiveness of Large Language Models (LLMs) on the task of detecting six out of the eighteen techniques mentioned in Da San Martino et al. (2019). Specifically, we evaluate the following models – OpenAI GPT-3.5, OpenAI GPT-4, and Anthropic Claude 3 Opus under different settings such as zero-shot, one-shot, chain-of-thought prompting (Wei et al. 2022) and other advanced prompting techniques such as generated knowledge promoting (Liu et al. 2021) and self-consistency prompting (Wang et al. 2022).

We find that none of the LLMs outperform a baseline model (a RoBERTa-CRF model with an ensemble of models for final classification layer (Jurkiewicz et al. 2020) both at the macro-F1 level (average across all six techniques) and individual technique F1 scores. GPT-4 outperforms GPT-3.5 and Claude 3 Opus at the macro-F1 level.

We also compare the results to another baseline released by the dataset authors (Da San Martino et al. 2019), which is a Multi-Granularity Network (MGN) model built on BERT embeddings. At the macro-F1 level, we find that GPT-3.5 and Claude 3 Opus do not outperform this MGN model. GPT-4 outperforms this model under one-shot, generated knowledge, and self-consistency prompting mechanisms. Looking at the techniques specifically, we find that all the LLMs outperform the MGN model in detecting instances of name-calling. GPT-4 and GPT-3.5 (only generated knowledge prompting) outperforms this model in detecting appeal-to-fear instances. Similarly, GPT-4 and GPT-3.5 (only one-shot) outperforms this model in detecting instances of flag-waving. None of the LLMs outperform the MGN model in detecting instances of loaded language, doubt, and exaggeration/minimization.

Propaganda Device	Definition
Name-Calling	“Labeling the object of the propaganda campaign as either something the target audience fears, hates, finds undesirable or otherwise loves or praises”
Loaded Language	“Using words or phrases with strong emotional implications to influence an audience”
Doubt	“‘Questioning the credibility of someone or something”
Appeal to Fear	“Seeking to build support for an idea by instilling anxiety and/or panic in the population towards an alternative, possibly based on preconceived judgments”
Flag-Waving	“Playing on strong national feeling (or with respect to a group, e.g., race, gender, political preference) to justify or promote an action or idea”
Exaggeration or minimization	“Either representing something in an excessive manner: making things larger, better, worse (e.g., “the best of the best”, “quality guaranteed”) or making something seem less important or smaller than it actually is ”

Note. From ”Fine-Grained Analysis of Propaganda in News Articles”, by Martino et al., 2019, EMNLP-IJCNLP, pp. 5636-5646 (Da San Martino et al. 2019)

Table 1: Propaganda Techniques and Definitions used in this study

Related Work

Language Model Evaluation

Language Models have been evaluated on several domain-specific tasks. Wang et al.(2023) evaluated LLMs on natural language understanding in the health domain. They used clinical benchmark datasets and found that GPT-4 outperforms GPT-3.5 and Bard in text classification, named entity recognition, relation extraction, and natural language inference. Rehana et al. (2023) compared LLMs to BERT-based models on protein-protein interaction extraction and found that domain-specific models such as BioBERT and PubMedBERT outperform GPT-4. Li et al. (2023b) similarly evaluated LLMs in the finance domain and found that GPT-4 outperforms all other LLMs in news classification and sentiment analysis. Domain-specific models, however, outperformed GPT-4 in named entity recognition (Li et al. 2023b). Li et al. (2023a) evaluated LLMs in detecting instances of hate speech on social media under different prompt settings and found the results to be better than labels obtained from crowd-sourced workers. Examples where LLMs outperform state-of-the-art models include information extraction (Wan et al. 2023; Ma et al. 2023), question answering (Bang et al. 2023), machine translation (Hendy et al. 2023; Wang et al. 2023), and information retrieval (Ziems et al. 2023).

Prompt Engineering

A prompt is a set of instructions describing the task that you want the LLM to accomplish. Prompts can be engineered to elicit more optimal responses.

Zero-shot prompting refers to curating a prompt describing the task that you want the LLM to perform without giving it any examples pertaining to the task. Several studies have looked at using zero-shot prompts with GPT-3.5 and GPT-4 (Wei et al. 2023; Espejel et al. 2023). Few-shot prompting, on the other hand, adds a bunch of examples to the task description to make instructions clearer (Brown et al. 2020). Few-shot prompts have also been evaluated in tasks that require more complex reasoning from the LLMs such as in answering questions in radiation oncology physics (Holmes et al. 2023). Liu et al. (2021) introduced generated knowledge prompting - a technique by which you get the LLM to generate knowledge related to the task before making the prediction. They test it on a variety of datasets in the commonsense reasoning domain and show that it outperforms zero-shot and few-shot settings. Wei et al. (2022) introduced Chain-of-Thought (CoT) prompting, a technique in which the model is given intermediate reasoning steps to evoke the LLM’s reasoning abilities. They test the prompting technique on math problems and show that it outperforms basic prompts. Wang et al. (2022) introduced self-consistency prompting in which the LLM is asked to generate multiple outputs and a majority voting technique help decide the final output i.e. consider the most “consistent” answer. Their method outperforms commonsense reasoning benchmarks when compared to CoT prompts.

In this study, we experiment with five of these prompting strategies - zero-shot, one-shot, chain-of-thought, generated knowledge, and self-consistency prompting.

Methods

Dataset

We perform experiments on the PTC (Propaganda Techniques Corpus) dataset from Da San Martino et al. (2019). This dataset was developed by gathering propagandistic articles from sources listed under Media Bias/Fact Check. Da San Martino et al. (2019) partnered with media intelligence professionals to develop annotations of eighteen propaganda techniques used in these articles. The dataset contains articles with phrase-level instances of these techniques. For the scope of this study (concerning training/inference cost, speed, and time), we choose to focus on six out of eighteen of these techniques (as seen in Table 1) which are also

Technique	GPT-3.5						GPT-4					Claude 3 Opus				MGN	RoBERTa-CRF
	Zero-Shot	One-Shot	CoT	GK	SC	FT	Zero-Shot	One-Shot	CoT	GK	SC	Zero-Shot	One-Shot	CoT	GK		
Name-Calling	0.21	0.13	0.13	0.25	0.24	0.23	0.27	0.30	0.18	0.31	0.30	0.24	0.26	0.12	0.28	0	**0.74**
Loaded Language	0.2	0.26	0.12	0.25	0.20	0.28	0.17	0.24	0.13	0.22	0.21	0.18	0.25	0.16	0.28	0.40	**0.80**
Doubt	0.11	0.10	0.08	0.06	0.10	0.16	0.08	0.11	0.13	0.17	0.14	0.11	0.13	0.15	0.10	0.19	**0.63**
Appeal to Fear	0.04	0.07	0.04	0.16	0.05	0.03	0.09	0.09	0.11	0.09	0.09	0.07	0.02	0.07	0.03	0.09	**0.48**
Flag-Waving	0.07	0.09	0.06	0.01	0.07	0	0.08	0.09	0.14	0.11	0.14	0.04	0.05	0.07	0.05	0.08	**0.82**
Exaggeration or minimization	0.09	0.11	0.05	0.09	0.09	0	0.07	0.10	0.08	0.09	0.09	0.08	0.06	0.07	0.09	0.11	**0.60**
Macro-F1	0.12	0.13	0.08	0.14	0.12	0.11	0.13	0.15	0.13	0.16	0.16	0.12	0.13	0.12	0.14	0.14	**0.67**

Table 2: F1 scores of LLMs under different prompt settings across six propaganda techniques compared to the baseline model. (MGN=Multi-Granularity Network, FT = Fine-Tuned, CoT=Chain-of-Thought Prompting, GK=Generated Knowledge Prompting, SC=Self-Consistency)

some of the more common techniques used in news articles (Da San Martino et al. 2019). We also made sure to focus on techniques that were more emotionally appealing (for example, by including *flag-waving* instead of *repetition*).

Models

We evaluate three large language models on the task of detecting propaganda techniques in news articles. The first is GPT-3.5 for which we use gpt-3.5-turbo-0125 in the OpenAI API. The second is GPT-4 for which we use gpt-4-0125-preview in the OpenAI API. The third is Claude 3 Opus for which we use claude-3-opus-20240229 in the Anthropic API.

Propaganda techniques rely on emotional and logically flawed reasoning and hence we chose models that were found to be superior in text classification (Wang, Zhao, and Petzold 2023), natural language understanding (Zhong et al. 2023), reasoning (Espejel et al. 2023; Li et al. 2023b) and entity extraction (Ma et al. 2023) among other benchmarks. We experiment with multiple prompt engineering techniques such as zero-shot, one-shot, chain-of-thought, generated knowledge, and self-consistency prompting. A table containing a list of the prompts used in our study can be found in the Appendix.

Fine-tuning GPT-3.5 We fine-tuned individual models for detecting each of the six techniques. We used the gpt-3.5-turbo-0125 model in the OpenAI API and used default hyperparameter values (epochs = 3, learning rate multiplier = 2, batch size = 1).

Baseline The authors of the PTC dataset organized a propaganda detection task at the International Workshop on Semantic Evaluation 2020 (SemEval 2020 Task 11) which received 44 submissions (Martino et al. 2020). We compare our results to the highest-achieving model (by a team named ApplicaAI) for this task which used a RoBERTa-CRF model with an ensemble of models as the final classifier (Jurkiewicz et al. 2020). The ensemble models include a RoBERTa-CRF trained on the original data and also a model trained on additional data that was generated using the original RoBERTa-CRF model. We use this model as our upper bound to draw comparisons to the current state-of-the-art.

We also compare our results to a model released by the dataset authors. This model is a novel Multi-Granularity Network (MGN) model that uses BERT embeddings and both sentence-level and phrase-level information while fine-tuning the model. Specifically, if the sentence is considered to be non-propagandistic then having the model not check for phrase-level instances of propaganda in it gives higher precision (hence higher F1 scores) compared to a BERT-based baseline fine-tuned to detect instances of these techniques in sentences (Da San Martino et al. 2019). This model acts as our lower bound for evaluating the effectiveness of LLMs.

Results

We report model outcomes under different prompting strategies for all six techniques in Table 2. An extended version of the table including precision and recall values can be found in the Appendix. We compare these with the RoBERTa-CRF baseline and the Multi-Granularity Network (MGN) implementation.

Name-Calling: As seen in Table 2, none of the LLMs outperform the RoBERTa-CRF baseline (F1=0.74). While all versions of LLMs outperformed the MGN model, GPT-4 generated knowledge prompting gave us the highest F1 score of 0.31 with precision=0.34, recall=0.29 (see Table 5).

Loaded-Language: None of the LLMs outperform the RoBERTa-CRF baseline (F1=0.80). Furthermore, none of the LLMs outperformed the MGN model either (F1=0.40).

Doubt: None of the LLMs outperform the RoBERTa-CRF baseline (F1=0.63). Furthermore, none of the LLMs outperformed the MGN model either (F1=0.19).

Appeal to Fear: None of the LLMs outperformed the RoBERTa-CRF baseline (F1=0.48). GPT-3.5 generated knowledge prompting outperformed the MGN model and all other LLMs (with F1=0.16, precision=0.16, recall=0.15) but not the RoBERTa-CRF baseline. GPT-4 outperformed and/or was on par with the MGN model under all prompt settings.

Flag-Waving: None of the LLMs outperformed the RoBERTa-CRF baseline (F1=0.82). All versions except zero-shot GPT-4 outperformed the MGN model, with GPT-4 self-consistency prompting giving us the highest F1 score of 0.144 (precision=0.13, recall=0.15). GPT-3.5 one-shot also outperformed the MGN model.

Exaggeration or Minimisation: None of the LLMs outperformed the RoBERTa-CRF baseline (F1=0.60). GPT-3.5 one-shot performed closely with the MGN model, giving us an F1 score of 0.111 (precision=0.06, recall=0.52) whereas the MGN model gave an F1-score of 0.116.

GPT-4 under one-shot, generated knowledge, and self-consistency prompting mechanisms outperform the MGN model (macro-F1). GPT-4 also outperforms GPT-3.5 and Claude 3 Opus (macro-F1). None of the LLMs outperform the RoBERTa-CRF baseline.

Discussions

We explored the effectiveness of LLMs in detecting propaganda techniques in news articles. We find that none of these LLMs outperform a RoBERTa-CRF model with an ensemble of models in the final classification layer (Jurkiewicz et al. 2020). Chernyavskiy, Ilvovsky, and Nakov (2020) also used an ensemble of RoBERTa models for the propaganda detection task (SemEval 2020 Task 11). None of the LLMs outperform this model as well. Patil, Singh, and Agarwal (2020) used an ensemble of BERT and logistic regression along with features such as TF-IDF. This model also shows superior performance to the LLMs (Table 7 in the Appendix contains F1 scores corresponding to these models). However, we find that GPT-3.5 and GPT-4 performs better than a Multi-Granularity Network based model (Da San Martino et al. 2019) in detecting three out of six techniques. We also find that GPT-4 outperforms the MGN model (macro-F1).

To strengthen the experiment's rigor, we experimented with multiple variations of the prompts for each setting. The F1 scores reported in Table 2 correspond to prompts that gave us the highest macro-F1 across these variations. For example, for zero-shot prompting, we experimented with six different variations of prompts including the example prompt seen in (Li et al. 2023a) which gave us the highest macro-F1 score. Similarly, for one-shot prompting, we experimented with three variations. For chain-of-thought prompting, we experimented with three variations including the example seen in (Kojima et al. 2022). Similarly, we tried two variations for self-consistency prompting and one variation for generated knowledge prompting. We believe this enhances the rigor of the conclusions drawn in this study.

We fine-tuned GPT-3.5 using a zero-shot prompt in the dataset's prompt template. When compared with their zero-shot counterpart (GPT-3.5 zero-shot in Table 2), the fine-tuned version gave better F1 scores for name-calling, loaded language, and doubt but worse scores for appeal to fear, flag-waving, and exaggeration/minimization. Future work could look into experimenting with the prompts used to fine-tune these models. Furthermore, our few-shot prompt consisted of one example (making it a one-shot prompt). Future studies could look into using more than one example while also being mindful of context window limitations. Researchers could also look into including examples that are related to the text at hand using information retrieval concepts based on semantic similarity (Nashid, Sintaha, and Mesbah 2023).

Da San Martino et al. (2019) report a moderate Inter Annotator Agreement (IAA) score between annotators while labeling the dataset. Their initial annotation stage (of the two-stage annotation process) saw a much lower IAA score of 0.24 and 0.28. They attribute this to cases where an annotator initially misses an instance but later agrees upon it in the second stage with a consolidator. Looking at the precision values in Table 4, 5, 6, we see several instances with really low precision values (indicating high false positives). This raises the question of whether the LLMs are identifying instances of propaganda techniques that were missed by human annotators in the original dataset. Further evaluation of the quality of these labels might help comprehend this.

Conclusions

While LLMs have achieved state-of-the-art results at numerous tasks such as information extraction (Wan et al. 2023; Ma et al. 2023), question answering (Bang et al. 2023), machine translation (Hendy et al. 2023; Wang et al. 2023), and information retrieval (Ziems et al. 2023), we find that LLMs perform inadequately at detecting propaganda techniques in news articles on the PTC dataset when compared to a RoBERTa-CRF baseline. A fine-tuned version of GPT-3.5 also gave us sub-par results. We find that GPT-4 outperforms a Multi-Granularity Network (MGN) baseline as well as GPT-3.5 and Claude 3 Opus (macro-F1). We also find that GPT-3.5 and GPT-4 outperforms the MGN model in detecting instances of name-calling, appeal to fear, and flag-waving, with Claude 3 Opus outperforming detection on only one technique (name-calling).

Acknowledgements

This work was supported by the National Science Foundation under grant number 1940713.

References

Bang, Y.; Cahyawijaya, S.; Lee, N.; Dai, W.; Su, D.; Wilie, B.; Lovenia, H.; Ji, Z.; Yu, T.; Chung, W.; et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Booth;, G. C. 1940. Can Propaganda Analysis Be Taught? *Junior College Journal*, 310–312.

Brattberg, E.; and Maurer, T. 2018. *Russian election interference: Europe's counter to fake news and cyber attacks*, volume 23. Carnegie Endowment for International Peace Washington, DC.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Da San Martino, G.; Seunghak, Y.; Barrón-Cedeno, A.; Petrov, R.; Nakov, P.; et al. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 5636–5646. Association for Computational Linguistics.

Da San Martino, G.; Shaar, S.; Zhang, Y.; Sh, Y.; Barrón-Cedeno, A.; Nakov, P.; et al. 2020. Prta: A system to support the analysis of propaganda techniques in the news. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 287–293. ASSOC COMPUTATIONAL LINGUISTICS-ACL.

Espejel, J. L.; Ettifouri, E. H.; Alassan, M. S. Y.; Chouham, E. M.; and Dahhane, W. 2023. GPT-3.5, GPT-4, or BARD? Evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts. *Natural Language Processing Journal*, 5: 100032.

Golovchenko, Y.; Buntain, C.; Eady, G.; Brown, M. A.; and Tucker, J. A. 2020. Cross-platform state propaganda: Russian trolls on twitter and YouTube during the 2016 US Presidential Election. *The International Journal of Press/Politics*, 25(3): 357–389.

Graham, M. M. W. 1939. Analyzing Propaganda. *Proceedings of the National Education Association*, 423–31.

Hendy, A.; Abdelrehim, M.; Sharaf, A.; Raunak, V.; Gabr, M.; Matsushita, H.; Kim, Y. J.; Afify, M.; and Awadalla, H. H. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Hollis;, E. V. 1939. Antidote for Propaganda,. *School and Society*, 50:449–453.

Holmes, J.; Liu, Z.; Zhang, L.; Ding, Y.; Sio, T. T.; McGee, L. A.; Ashman, J. B.; Li, X.; Liu, T.; Shen, J.; et al. 2023. Evaluating large language models on a highly-specialized topic, radiation oncology physics. *Frontiers in Oncology*, 13.

Jowett, G. S.; and O'Donnell, V. 2018. *Propaganda & Persuasion*. Sage publications.

Jurkiewicz, D.; Borchmann, Ł.; Kosmala, I.; and Graliński, F. 2020. ApplicaAI at SemEval-2020 task 11: On RoBERTa-CRF, span CLS and whether self-training helps them. *arXiv preprint arXiv:2005.07934*.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213.

Lee, A.; and Lee, E. B. 1939. The fine art of propaganda.

Li, L.; Fan, L.; Atreja, S.; and Hemphill, L. 2023a. "HOT" ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web*.

Li, X.; Zhu, X.; Ma, Z.; Liu, X.; and Shah, S. 2023b. Are chatgpt and gpt-4 general-purpose solvers for financial text analytics? an examination on several typical tasks. *arXiv preprint arXiv:2305.05862*.

Liu, J.; Liu, A.; Lu, X.; Welleck, S.; West, P.; Bras, R. L.; Choi, Y.; and Hajishirzi, H. 2021. Generated knowledge prompting for commonsense reasoning. *arXiv preprint arXiv:2110.08387*.

Ma, Y.; Cao, Y.; Hong, Y.; and Sun, A. 2023. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! *arXiv preprint arXiv:2303.08559*.

Martino, G.; Barrón-Cedeno, A.; Wachsmuth, H.; Petrov, R.; and Nakov, P. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. *arXiv preprint arXiv:2009.02696*.

Nashid, N.; Sintaha, M.; and Mesbah, A. 2023. Retrieval-based prompt selection for code-related few-shot learning. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, 2450–2462. IEEE.

Wan, Z.; Cheng, F.; Mao, Z.; Liu, Q.; Song, H.; Li, J.; and Kurohashi, S. 2023. Gpt-re: In-context learning for relation extraction using large language models. *arXiv preprint arXiv:2305.02105*.

Wang, L.; Lyu, C.; Ji, T.; Zhang, Z.; Yu, D.; Shi, S.; and Tu, Z. 2023. Document-level machine translation with large language models. *arXiv preprint arXiv:2304.02210*.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Wang, Y.; Zhao, Y.; and Petzold, L. 2023. Are large language models ready for healthcare? a comparative study on clinical language understanding. In *Machine Learning for Healthcare Conference*, 804–823. PMLR.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wei, X.; Cui, X.; Cheng, N.; Wang, X.; Zhang, X.; Huang, S.; Xie, P.; Xu, J.; Chen, Y.; Zhang, M.; et al. 2023. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*.

Zhong, Q.; Ding, L.; Liu, J.; Du, B.; and Tao, D. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *arXiv preprint arXiv:2302.10198*.

Ziems, N.; Yu, W.; Zhang, Z.; and Jiang, M. 2023. Large language models are built-in autoregressive search engines. *arXiv preprint arXiv:2305.09612*.

Appendix

Prompt Setting	**Prompt**
Zero-shot	I want you to perform a data annotation task. In your output, I want you to return a json dictionary with key as phrase and value as technique, depending on whether you think the phrases in the following text uses $< technique >$. A phrase is $< technique >$ if you perceive that it is $< technique\ definition >$. I want you to respond with a json dictionary strictly having the detected phrases as keys and technique ($< technique >$) as value (eg, “phrase”:“$< technique >$”,...) and provide an explanation for your output in the end. Text:
One-shot	I want you to perform a data annotation task. In your output, I want you to return a json dictionary with key as phrase and value as technique, depending on whether you think the phrases in the following text uses $< technique >$. A phrase is $< technique >$ if you perceive that it is $< technique\ definition >$. I want you to respond with a json dictionary strictly having the detected phrases as keys and technique ($< technique >$) as value (eg, “phrase”:“$< technique >$”,...) and provide an explanation for your output in the end. Text: $< example\ article >$ Output: { “$< phrase1 >$”: ”$< technique >$” } Explanation: - $< phrase1 >$ and $< explanation >$ Text:
CoT	Using the definition of the technique given to you below, identify and extract instances of the technique in the text below by strictly following this format: (1) For each phrase in the text, return the following in your answer: - $< phrase >$ Let’s think step by step. $< explanation >$ $< yes$ or $no >$ (2) Finally return a json dictionary strictly making sure that each detected phrase (from step 1) is a key and the technique ($< technique >$) is its value (eg, “phrase”:“$< technique >$”,...). Technique: “$< technique >$”: “$< technique\ definition >$”. Text:
Generated Knowledge Prompting	Detect the propaganda technique $< technique >$ in the input article. Follow the example format given below: Input: Article: $< example\ article >$ are there instances of $< technique >$ in this text? Knowledge: $< technique >$ is the $< technique\ definition >$. The phrase “$< phrase\ 1 >$” is an instance of $< technique >$ because it $< explanation >$. Output: { “$< phrase1 >$”: ”$< technique >$” } Input: Article: $< article >$ are there instances of $< technique >$ in this text? Knowledge:
Self-Consistency	I want you to perform a data annotation task. In your output, I want you to return a json dictionary with key as phrase and value as technique, depending on whether you think the phrases in the following text uses doubt. A phrase is $< technique >$ if you perceive that it is $< technique\ definition >$ I want you to respond with a json dictionary strictly having the detected phrases as keys and technique ($< technique >$) as value. For each detection, explain your reasoning at the very beginning before the final dictionary output. An example is given below: Text: $< example\ article >$ Explanation: - $< phrase1 >$ and $< explanation >$ Output: { “$< phrase1 >$”: ”$< technique >$” } Text: $< article >$ Explanation:

Table 3: Prompts used in this study

Technique	Zero-Shot			One-Shot			CoT			GK			SC			FT		
	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R
Name-Calling	0.21	0.18	0.25	0.13	0.18	0.10	0.13	0.10	0.17	0.25	0.20	0.32	0.24	0.23	0.25	0.239	0.298	0.19
Loaded Language	0.2	0.15	0.29	0.26	0.20	0.37	0.12	0.10	0.15	0.25	0.19	0.35	0.20	0.19	0.21	0.287	0.289	0.28
Doubt	0.11	0.06	0.32	0.10	0.06	0.34	0.08	0.05	0.24	0.06	0.12	0.04	0.10	0.08	0.13	0.167	0.215	0.13
Appeal to Fear	0.04	0.02	0.14	0.07	0.04	0.27	0.04	0.024	0.13	0.161	0.16	0.15	0.05	0.03	0.10	0.038	0.138	0.02
Flag-Waving	0.07	0.04	0.19	0.09	0.06	0.19	0.06	0.04	0.13	0.01	0.03	0.01	0.07	0.07	0.08	0	0	0
Exaggeration or minimization	0.09	0.05	0.46	0.11	0.06	0.52	0.05	0.03	0.22	0.09	0.05	0.33	0.09	0.05	0.27	0	0	0

Table 4: GPT-3.5 Precision, Recall, F1 scores (FT = Fine-Tuned, CoT=Chain-of-Thought Prompting, GK=Generated Knowledge Prompting, SC=Self-Consistency)

Technique	Zero-Shot			One-Shot			CoT			GK			SC		
	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R
Name-Calling	0.274	0.213	0.385	0.307	0.243	0.416	0.184	0.155	0.227	0.316	0.347	0.29	0.309	0.369	0.266
Loaded Language	0.173	0.111	0.389	0.24	0.174	0.384	0.13	0.099	0.186	0.225	0.147	0.476	0.21	0.184	0.244
Doubt	0.082	0.045	0.464	0.115	0.066	0.42	0.135	0.08	0.425	0.172	0.116	0.33	0.144	0.093	0.324
Appeal to Fear	0.096	0.054	0.427	0.093	0.054	0.341	0.118	0.073	0.299	0.094	0.062	0.195	0.096	0.071	0.147
Flag-Waving	0.08	0.052	0.177	0.09	0.057	0.205	0.142	0.153	0.132	0.116	0.113	0.12	0.144	0.136	0.154
Exaggeration or minimization	0.079	0.043	0.526	0.104	0.057	0.554	0.087	0.05	0.355	0.09	0.059	0.19	0.096	0.057	0.3

Table 5: GPT-4 Precision, Recall, F1 scores (CoT=Chain-of-Thought Prompting, GK=Generated Knowledge Prompting, SC=Self-Consistency)

Technique	Zero-Shot			One-Shot			CoT			GK		
	F1	P	R	F1	P	R	F1	P	R	F1	P	R
Name-Calling	0.245	0.229	0.264	0.267	0.307	0.236	0.222	0.243	0.205	0.283	0.317	0.255
Loaded Language	0.184	0.153	0.229	0.259	0.217	0.322	0.166	0.164	0.169	0.288	0.235	0.373
Doubt	0.113	0.074	0.246	0.132	0.11	0.166	0.153	0.103	0.296	0.102	0.25	0.064
Appeal to Fear	0.075	0.046	0.194	0.027	0.02	0.043	0.075	0.052	0.134	0.033	0.034	0.033
Flag-Waving	0.044	0.032	0.069	0.052	0.048	0.057	0.073	0.059	0.094	0.058	0.102	0.041
Exaggeration or minimization	0.082	0.047	0.292	0.069	0.041	0.203	0.073	0.048	0.155	0.098	0.066	0.183

Table 6: Claude 3 Opus Precision, Recall, F1 scores (CoT=Chain-of-Thought Prompting, GK=Generated Knowledge Prompting)

Technique	Chernyavskiy, Ilvovsky, and Nakov (2020)	Patil, Singh, and Agarwal (2020)
	F1	F1
Name-Calling	0.73	0.70
Loaded Language	0.80	0.75
Doubt	0.66	0.52
Appeal to Fear	0.45	0.32
Flag-Waving	0.74	0.75
Exaggeration or minimization	0.59	0.49
Macro-F1	0.66	0.58

Table 7: F1 scores of transformer-based models submitted to SemEval Task 11 2020