DOI: 10.35681/1560-9189.2025.27.1.336147

UDC 004.89, 004.912

O. A. Boiko
National Technical University of Ukraine
«Igor Sikorsky Kyiv Polytechnic Institute»
37, Prospect Beresteiskyi, 03056 Kyiv, Ukraine

Modern AI methods for detecting propaganda in text

Propaganda is a key weapon of modern cognitive warfare, vividly illustrated by Russia's war against Ukraine, where information manipulation has become as strategically significant as military action. Propaganda's subtle psychological and linguistic tactics challenge traditional detection methods, demanding more sophisticated, context-aware technologies. This paper begins by reviewing supervised AI models, noting their dependence on expertly annotated corpora. Then, it outlines recent progress in unsupervised methods, showing how large language models (LLMs) can flag manipulative tactics with minimal labeled data while supplying human-readable justifications. Finally, three directions for future research are proposed: (1) reasoning language models for stepwise analysis; (2) cost-efficient multi-agent systems; and (3) hybrid frameworks that combine the first two. Advancing these methods offers a promising pathway for safeguarding democratic societies against evolving propaganda strategies.

Keywords: *propaganda detection; natural language processing; artificial intelligence; large language models; reasoning language models; multi-agent systems.*

Introduction

Propaganda, as a phenomenon, has long captured the attention of scholars in various fields, including political science, psychology, sociology, and communication studies. According to Jowett and O'Donnell, propaganda can be defined 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» [1, p. 7]. In this sense, propaganda stands apart from mere persuasion by its top-down, often manipulative nature, and its focus on achieving very specific objectives aligned with the communicator's agenda.

In modern times, propaganda has become one of the fundamental tools in geopolitical and social conflicts, which is especially evident in the scope of Russia's military invasion of Ukraine in 2014 and its subsequent full-scale war against the Ukrainian nation in 2022. Russian state-backed disinformation campaigns offer a striking illustration of how propaganda is deployed in an effort to undermine adversaries, destabilize societies, and push ideological narratives [2]. The «Russian War on Truth», as documented by the NATO Parliamentary Assembly, demonstrates how manipulative information strategies can erode trust in institutions, blur the line between fact and fiction, and increase social tensions [2]. Today, as societies become more interconnected, such propaganda campaigns can be seamlessly disseminated through news outlets, social media, and other digital platforms.

In response to these developments, the NATO Strategic Communications Centre of Excellence has emphasized the urgent need for proactive propaganda monitoring and detection [3]. The possibility of leveraging Artificial Intelligence (AI) to address these challenges has gained substantial attention. In their recent report, Juršėnas et al. highlight the crucial role of AI in assisting analysts by automating key tasks such as identifying potentially deceptive content for further review, tracking the dissemination patterns of disinformation, and analyzing text for sentiment, emotional tone, topic distribution, and stylistic markers commonly associated with propaganda [3].

However, this field presents some significant challenges. The language of propaganda is often subtle and context-dependent, utilizing rhetorical and psychological methods that appeal to emotions. As described by Miller in his famous 1939 address at New York's Town Hall [4], propaganda commonly employs such techniques as *Name-Calling*, *Glittering Generalities*, *Transfer*, *Testimonial*, *Plain Folks*, *Card Stacking*, and *Band Wagon*, all aimed at eliciting strong emotional responses rather than rational thinking. The inherent complexity of these strategies means that purely keyword-based or rule-based detection approaches are typically insufficient. This complexity is discussed further in subsequent sections.

In this paper, we present a comprehensive exploration of automated propaganda detection methods, tracing the evolution from conventional supervised learning techniques to emerging unsupervised frameworks. We start by establishing a theoretical foundation — drawing on seminal studies in rhetorical strategies and psychological manipulation — to clarify how propaganda is defined and operationalized in both academic and real-world contexts. Building on this groundwork, we examine established detection methods that rely on annotated datasets and supervised models, detailing key tasks, benchmarks, and the architectures that have driven progress in this area. We then explore the emerging wave of unsupervised and large language model-based methods, underlining how GPT-like architectures can be adapted for these tasks, as well as identifying their current limitations.

Finally, we evaluate future directions, particularly (a) reasoning-based GPT agents (e.g., OpenAI's o3 model, DeepSeek R1), (b) multi-agent systems employing advanced prompt engineering and cost-effective models (e.g., GPT-4o mini), and (c) respective hybrid solutions, all of which offer a promising yet underexplored path forward. These directions acknowledge and address some of the limitations of existing supervised approaches, such as the scarcity of open-access, reliably annotated datasets and the challenges in transferring real-world propaganda detection tasks into flexible, robust AI solutions. In principle, by utilizing reasoning-oriented frameworks and/or multi-agent architectures, researchers could build systems better equipped to capture the nuanced, evolving nature of propaganda in a dynamic and context-sensitive manner.

Overview of Propaganda Detection Techniques

Identifying propaganda requires a nuanced understanding of its distinct linguistic markers and the rhetorical devices that shape its persuasive power. In their study on the language of news media, Rashkin et al. (2017) note that propagandistic text can often be distinguished by certain linguistic anomalies, such as the frequent use of second-person pronouns, superlatives, and weakly subjective words [5]. In terms of rhetorical manipulation, Clyde R. Miller's classic work, introduced earlier, highlighted seven common propaganda devices, including *Name-Calling* and *Card Stacking*, which often appear in contexts aimed at manipulating an audience's emotions [4]. These devices remain highly relevant in modern media, albeit in more sophisticated and digitally mediated forms.

Contemporary research has extended these categorizations into broader taxonomies and more detailed typologies. An influential framework for propaganda detection was introduced by Da San Martino et al. [6], proposing analysis of news articles at a fine-grained level by spotting textual fragments (or spans) that exhibit any of 18 specific propaganda techniques (e.g., *Loaded Language*, *Name Calling/Labeling*, *Appeal to Fear*, *Repetition*, and *Flag-Waving*). Table summarizes a few common propaganda techniques and their characteristics with some real-world examples.

In contrast to traditional document-level labeling, where entire articles are tagged as either «propaganda» or «not propaganda», fine-grained approaches aim to pinpoint the specific segments employing manipulative techniques [6]. This not only increases detection accuracy but also provides a degree of explainability that helps users understand why content has been flagged as propaganda.

A practical challenge in implementing fine-grained detection is the availability of large, reliably annotated datasets. For news classification tasks, robust data corpora — such as the one presented by Horne et al. [7] — offer vantage points to study persuasion and misinformation across diverse outlets. However, systematically annotating large volumes of data for multiple nuanced propaganda strategies is labor-intensive. Annotation quality also tends to suffer due to the inherent subjectivity in interpreting rhetorical tactics. Additionally, disinformation campaigns often combine truth with falsehoods to complicate classification [7].

Another key challenge arises from the context-sensitive nature of propaganda, which frequently relies on references to existing social tensions or targeting a particular audience [6], [8]. For instance, Russian disinformation frequently draws upon local grievances and illusions of moral superiority, as documented by Demeuse [2]. Detecting such tactics often requires not only textual analysis (e.g., sentiment or rhetorical structure) but also real-time contextual knowledge of ongoing events, prevalent narratives, and target audiences.

In summary, modern propaganda detection builds upon traditional content analysis by employing computational methods to identify manipulative patterns in text. Key linguistic indicators — ranging from emotionally loaded vocabulary to structural repetition and logical fallacies — form the foundational features for these methods. As detailed in the following sections, supervised learning approaches train models on annotated examples of propaganda text to automatically detect these patterns, whereas unsupervised and large language model-based methods leverage pre-trained linguistic knowledge, significantly reducing their reliance on labeled data.

Examples of propaganda techniques commonly found in text

Propaganda Technique	Description	Example
Loaded Language	Use of emotionally charged words to influence opinion [8]	*«... his* [Zelensky's] *unhealthy ambitions, which have been fostered by Western handlers ...»* — Maria Zakharova, briefing at the Russian Foreign Ministry, Moscow, December 25, 2024.
Name-Calling/Labeling	Tagging a target with negative labels to inflict condemnation [4]	*«... Zelensky's criminal illegitimate regime ...»* — Vladimir Putin, Results of the Year, Moscow, December 19, 2024.
Appeal to Fear	Instilling panic or anxiety to rally support [6]	*«They will undoubtedly try to bring war to Crimea just as they have done in Donbass, to kill innocent people just as members of the punitive units of Ukrainian nationalists and Hitler's accomplices did during the Great Patriotic War».* — Vladimir Putin, address on February 24, 2022, 06:00, The Kremlin, Moscow
Repetition	Repeating a message persistently to embed it as truth [6]	*«His* [Zelensky's] *drug dependency has become quite apparent ...»* — Maria Zakharova, briefing at the Russian Foreign Ministry, Moscow, December 25, 2024.
Flag-Waving	Evoking patriotism or group loyalty for support of some idea or action [6]	*«... it is our strength and our readiness to fight that are the bedrock of independence and sovereignty and provide the necessary foundation for building a reliable future for your home, your family, and your Motherland».* — Vladimir Putin, address on February 24, 2022, 06:00, The Kremlin, Moscow

Supervised Learning Approaches

Supervised learning has been the dominant paradigm for automated propaganda detection in recent years. In supervised settings, large collections of labeled examples — each indicating whether (and how) a particular text segment is propagandistic — are used to train classification models. Recent transformer-based architectures have demonstrated strong performance in propaganda detection tasks, especially when fine-tuned on domain-specific datasets. Abdullah et al. applied RoBERTa-based models to detect propaganda techniques in English news articles, reporting an F1 score of over 60 % when classifying which of several persuasive tactics appeared [9]. In another study, Vysotska et al. combined logistic regression, random forests, and multi-layer perceptrons with feature-engineering methods (TF-IDF, Word2Vec) to identify propaganda in online messages, achieving accuracy scores of up to 99 % at the message level on specific datasets [8]. While such high accuracies can be partly attributed to domain-specific data and relatively controlled conditions, they underscore the promise of supervised classifiers.

Another effective supervised strategy involves the use of ensemble models. By combining multiple classifiers, an ensemble can often achieve higher robustness and accuracy than a single model. For example, Krak et al. (2024) developed an ensemble of recurrent neural networks for propaganda detection in Ukrainian content [10]. Their method trained several RNN models (bi-directional LSTM and GRU networks) on the same data and then either bagged (averaged) or stacked their predictions. The ensemble could output a predicted «propaganda intensity» score for a given text. Notably, they reported an F1 score of approximately 97 % using a bagging ensemble of two RNN models. The diversity in the ensemble — where each RNN may capture different aspects of writing style or content — likely contributed to the high performance. The general finding is that neural ensembles tend to reduce errors by compensating for individual models' limitations.

However, as highlighted by Szwoch et al. (2024), one of the primary challenges faced by studies in propaganda detection is the scarcity of fully open-access, reliably annotated datasets — especially for under-resourced languages where annotations are sparse and expert annotators are few [11]. Additionally, as Da San Martino et al. (2020) demonstrated in their survey on computational propaganda detection, the behavior of malicious actors continues to evolve, increasingly employing the same advanced AI tools as their pursuers — for example, generating credible texts using generative pre-trained transformers [12]. This evolution in tactics not only complicates the detection process but also widens the gap between controlled experimental settings and real-world scenarios. Furthermore, most supervised models act as «black boxes», providing high-level predictions without clarifying how or why the text is labeled propagandistic.

In short, while supervised learning has provided a robust foundation for automated propaganda detection — demonstrating high accuracy in controlled environments — the approach is frequently hampered by the limited availability of comprehensive, openly accessible datasets. This limitation, as well as practical considerations of having a more robust solution to address evolving propaganda tools, has spurred interest in unsupervised methods, particularly leveraging the power of large pre-trained language models, as discussed in the next section.

Unsupervised Learning Approaches

Unsupervised approaches to propaganda detection seek to identify propagandistic content without explicitly labeled examples for training. A promising avenue in recent years is leveraging large language models (LLMs) like GPT-4 to detect propaganda through prompt-based analysis. These models are trained on vast corpora (in an unsupervised manner) and encode a great deal of general knowledge and linguistic patterns. The idea is that by crafting the right prompt, one can use an LLM as an on-the-fly propaganda detector, even without fine-tuning it on a specific propaganda dataset. In essence, the LLM serves as a sophisticated, general-purpose classifier leveraging its extensive pre-trained knowledge.

Jones (2024) explored using prompt engineering with OpenAI's LLM (the underlying technology behind ChatGPT) to identify propaganda techniques in news articles [13]. In their approach, they designed a detailed instruction that explained the task to the model: the prompt included the previously mentioned 18 propaganda techniques defined by Da San Martino et al. [6], asking the model to identify which, if any, appeared in a given article. They tested their method by feeding articles from Russia Today (a known state-sponsored propaganda outlet) and articles from the SemEval-2020 dataset to the

GPT-3.5 Turbo model via the API and then analyzing the model's responses. The LLM, in essence, was performing a form of unsupervised classification — it had not been trained specifically on labeled instances, but it was using its internal knowledge and basic reasoning to apply the definitions given in the prompt.

The results from such studies are intriguing. GPT-3.5 Turbo was indeed able to output plausible identifications of propaganda techniques and even provide explanations for its choices. For instance, in Jones' experiments, the model highlighted a *Reductio ad Hitlerum* technique in the text with the comment *«The author suggests that the person is accusing the Republicans of being Nazis»* [13, p. 6]. This aligns with another advantage of LLM-based detection: the ability to generate human-readable justifications. Jones qualitatively found that the model's outputs often made sense and could help a reader understand why a piece of text might be propagandistic. This interpretability — essentially the model performing a reasoning trace — is something traditional classifiers (which just output a label) do not provide. Such explanations can be valuable for trust and for refining prompts (e.g., noticing if the model is consistently misidentifying a certain technique and adjusting the prompt instruction accordingly).

Sprenkamp et al. (2023) also conducted structured experiments with LLMs, applying GPT-3 and GPT-4 to the SemEval propaganda dataset using various prompt engineering and fine-tuning strategies [14]. They reported that GPT-4, when appropriately prompted/fine-tuned, achieved results comparable to the current state-of-the-art RoBERTa models. In other words, a sufficiently advanced LLM with the right guidance can match the performance of a dedicated classifier on multi-label propaganda identification. This is an encouraging finding, suggesting that as the technology progresses, it may close the gap in this specialized task.

In another comprehensive study, Szwoch et al. (2024) also evaluated GPT-4's ability to perform propaganda technique detection and revealed further insights into the challenges [11]. They conducted experiments on the English SemEval-2020 corpus, on a similar annotated corpus in another language, and on an unlabeled set of Polish news articles to see how the model handles an under-resourced language. They also tried different prompting strategies, including adding chain-of-thought (CoT) reasoning in the prompt (instructing the model to think step-by-step) to see if it improved accuracy. Notably, one setting achieved a high precision of 81,8 % on identifying propaganda spans, but the recall was lower (below 10 %). Despite some encouraging partial metrics, none of the GPT-4 attempts outperformed the supervised baseline's F1 score. The best F1 they achieved was around 20 %, whereas a traditional supervised model's F1 was about 50 % on that task. In other words, while GPT-4 and similar models are extremely powerful, their reasoning is not yet reliably on par with human annotators or well-tuned supervised systems for detecting propaganda techniques. That said, the authors remain optimistic that further advancements in LLMs could lead to better results, given that simply scaling these models has already produced notable improvements on a range of NLP benchmarks [11].

In summary, unsupervised methods for propaganda detection, especially those leveraging GPT-style large language models, are still «underexplored» [14] but are already an exciting frontier. Despite somewhat inconsistent results, they offer clear advantages, including minimal dependence on task-specific training data, multilingual flexibility, and the capability to generate explanatory outputs. As LLM technology advances (with improved reasoning capabilities and longer context handling), we can expect continued improvements in unsupervised propaganda detection. The next section delves into these prospects, examining the hypothesis that emerging reasoning-based AI agents and multi-agent systems could dramatically enhance automated propaganda detection while managing costs and complexity.

Future Research

Reasoning-based GPT Agents

One promising direction is leveraging the newest generation of reasoning language models specifically optimized for complex reasoning tasks. Examples include OpenAI's o1/o3, DeepSeek's R1, Anthropic's Claude 3.7 Sonnet, and others. These are cutting-edge LLMs that have been refined to excel at multi-step reasoning, logic, and understanding nuanced instructions. For instance, unlike standard LLMs that generate a response in one pass, OpenAI's o1 employs a «long internal chain-of-thought» that helps it deconstruct challenging steps into simpler components, identify and correct errors, and adapt by switching approaches when needed [15]. This iterative refinement significantly enhances its reasoning capabilities. The next OpenAI reasoning model, o3, showcased even greater performance, achieving about three times the score of the earlier model on an abstraction and reasoning corpus test [16].

In the context of propaganda detection, a reasoning-enabled agent could, for instance, parse an article and logically examine the claims being made, the emotional tone, and the argumentative structure. In essence, such an agent can replicate the analytical processes employed by human experts when dissecting propaganda, carefully evaluating content step by step. We expect this to improve the detection of more subtle propaganda, which may require contextual interpretation and world knowledge to recognize. For instance, identifying logical fallacies like «straw man» or «red herring» can be challenging — the model must recognize that the argument being refuted was never genuinely presented by the opposition (straw man), or that an irrelevant point has been introduced to distract (red herring). A reasoning-capable GPT agent might handle this better than a standard classifier by virtue of its iterative thinking.

Additionally, for the task of automated propaganda detection, the reasoning agent could be prompted to evaluate content from multiple perspectives — such as the source, linguistic nuances, historical context, etc. — before concluding whether it qualifies as propagandistic. The reasoning approach also aligns well with the need for explainability: a reasoning agent can provide the steps it has taken (like listing suspicious rhetorical tactics it spotted) which could be invaluable in a domain where just flagging content as propaganda often demands justification.

That said, reasoning-based models come with trade-offs. The chain-of-thought process, while improving accuracy, introduces additional computational overhead and latency. Each question or input takes longer to answer because the model is effectively doing more work under the hood (test-time compute) [15]. Another important argument is cost efficiency — for instance, the inference cost of the o1 model can be up to six times higher than that of GPT-4o and 100 times higher than GPT-4o mini [17].

Both these factors can make real-time or large-scale deployment more challenging, especially if using a very large and expensive model like o3. Furthermore, as an area of active research, ensuring the reliability of the reasoning — namely, that intermediate steps lead to correct conclusions and do not drift into errors or biases — remains an open problem. Nevertheless, the trajectory is clear: future AI agents with advanced reasoning are likely to set new performance records in tasks like propaganda detection, where nuanced understanding is required.

Multi-agent Systems with Prompt Engineering

Another promising research direction envisions using a system of multiple collaborating AI agents — each based on a cost-effective language model — to detect propaganda jointly. Rather than relying on a single monolithic model, the task is divided among specialized agents that communicate with each other.

Multi-agent LLM systems have already started to appear in research for complex NLP tasks, including multilabel narrative classification and propaganda analysis, with promising results [18, 19]. The basic idea is to split the problem into sub-tasks, have different agents (which could be separate AI models or separate prompt calls to the same model) tackle each sub-task, and then combine their outputs. This approach resonates with methodologies such as the «Swarm of Virtual Experts» (SVE) proposed by Lande et al. [20], where the «swarm» is conceptualized as a collection of diverse responses and perspectives generated by LLMs acting as virtual experts. In such a model, each query, potentially varied in its formulation or targeted at different LLM instances, contributes an «expert opinion» thereby leveraging the probabilistic nature of LLMs to achieve a more comprehensive analysis.

The multi-agent approach is especially relevant when considering cost-effective models like GPT-4o and GPT-4o mini, which have been optimized for lower cost and faster inference, though with some trade-offs in peak capability. As noted earlier, inference using these models can lead to up to a 100-fold cost reduction, which can be a game-changing factor when designing multi-agent solutions. Instead of relying on a single large model for lengthy reasoning, a system can orchestrate multiple calls to smaller, less expensive models — each handling a portion of the task — to ultimately lower overall operational costs.

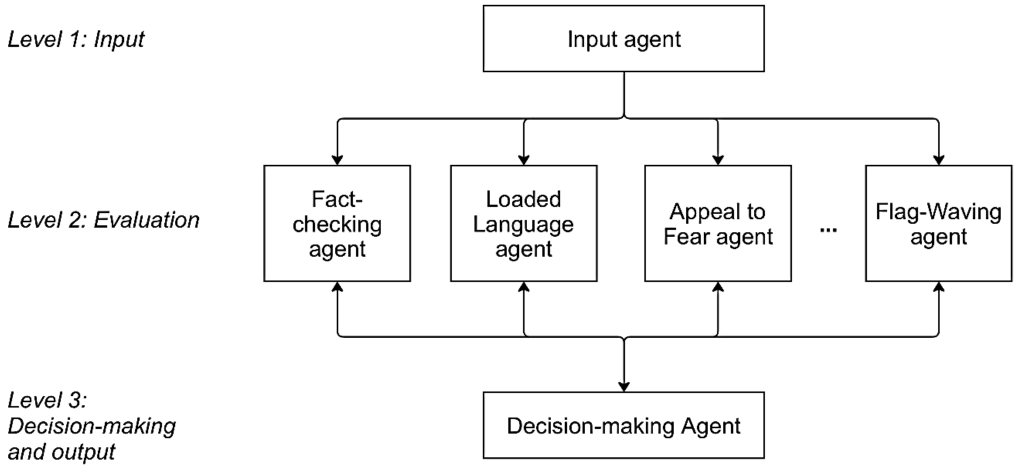
For example, consider a multi-level propaganda detection pipeline:

Level 1 (Input): A small cost-effective model (e.g. GPT-4o or GPT-4o mini) performs basic preparation and normalization of an evaluation sample — e.g., summarizing or compressing longer text to fit the context window, removing formatting artifacts, translating from low-resource languages, and so on.

Level 2 (Evaluation): Multiple agents then conduct deeper analysis on the normalized output, each focusing on its own specialization. Some might perform fact-checking or sentiment analysis, while others match the text against specific manipulation tactics. Drawing inspiration from frameworks like «Swarm of Virtual Experts», these Level 2 agents could be assigned more explicit roles, such as an «Analyzer Agent» for initial breakdown, a «Critic Agent» for evaluating consistency, specialized «Technique Detection Agents» (e.g., for Name-Calling, Loaded Language, etc.), and perhaps even a «Contextual Verification Agent» focusing on external information consistency. Lande et al. [20] note that assigning such roles can influence the context in which responses are formed, promoting diverse yet focused analyses.

Level 3 (Decision-making and output): Finally, a higher-level agent takes the distilled results from Levels 1 and 2 to produce the final judgement whether the article is propagandistic and crafting a report on identified manipulation techniques if any.

Figure illustrates a possible high-level architecture for such a multi-agent system.



A high-level multi-agent architecture for an automated propaganda detection system

By assigning narrower tasks to cost-effective models at each stage, the overall computational cost can be lower than relying on a single, expensive large model to process the text end-to-end. Additionally, since this pipeline is inherently modular, it allows mixing models from different vendors, switching between open-source and proprietary solutions, or pairing LLM-based approaches with specialized supervised methods. If better models appear over time, they can be relatively easily tested or swapped into a particular pipeline stage without overhauling the high-level system architecture.

This approach might also benefit from richer intra-agent collaboration, such as error-checking (where a dedicated agent verifies another agent's results) and feedback loops, similar to those explored in recent multi-agent research by Han et al. [21]. If Agent A flags content as propaganda while Agent B disagrees, a higher-level agent can prompt further analysis and ask for clarifications (like a team of human analysts debating ambiguous cases). Such feedback loops can enable real-time refinement and optimization of outputs at each sub-task stage, leading to greater accuracy.

Additionally, each agent's performance can be improved through advanced prompt engineering and fine-tuning approaches — such as Chain-of-Thought reasoning [22] and Knowledge Injection [23].

Altogether, this system might achieve stronger accuracy by specialization while keeping costs manageable, and importantly, offer transparency and modularity, since each agent's performance can be inspected and upgraded independently.

Hybrid Approaches: Reasoning Agents within Multi-Agent Systems

While (a) and (b) each offer compelling advantages, an additional promising research direction involves exploring the synergy between reasoning-based GPT agents and multi-agent systems. In such a hybrid setup, at least one advanced reasoning agent could serve as the «analytical core», while multiple additional cost-effective or specialized agents could collaborate in handling various subtasks. Below are a few ways this synergy could enhance automated propaganda detection:

1) **Deep Analytical Core + Specialized Satellites:** A large or specialized reasoning agent — capable of multi-step logical inferences – could focus on tricky rhetorical maneuvers such as straw man arguments or subtle emotional appeals. Meanwhile, a network of smaller, task-specific agents (like language translators or fact-checkers) handles

localized tasks (translation, style analysis, domain-specific references). The smaller agents feed partial findings back to the central reasoning agent, which then employs advanced chain-of-thought reasoning to interpret and integrate these insights;

2) **Feedback Loops with an «Expert» Agent:** In a standard multi-agent pipeline, sub-results circulate among smaller models, potentially including consistency and accuracy checks. By introducing a reasoning expert agent into this loop, one can achieve more rigorous verification: the expert agent examines each partial output, identifies logical inconsistencies or missing context, and then prompts the relevant smaller agents for further clarification or additional evidence. This might result in a more robust and reliable overall decision-making process;

3) **Adaptive Cost vs. Performance Balancing:** In real-world scenarios, one might selectively invoke the expensive reasoning agent only for content flagged as ambiguous or highly complex by less expensive models. Basic propaganda indicators (e.g., repetitive name-calling) might be handled by lightweight agents alone, while more nuanced cases are escalated to the expert agent, optimizing resource allocation without sacrificing analytical depth.

Overall, these hybrid architectures preserve the cost-saving advantages of multi-agent deployments while leveraging the deeper contextual understanding and complex logic that advanced chain-of-thought agents bring to the table. Future research might explore optimal strategies for coordinating reasoning agents and supporting models, how to manage shared memory and contextual information across inference layers, optimizing the trade-offs among speed, cost, and detection accuracy. In practice, such dual-tier configurations may prove invaluable for real-time or high-volume propaganda-monitoring workflows, especially in environments where content complexity varies unpredictably.

In summary, (a) reasoning-based GPT agents, (b) multi-agent systems with specialized prompt engineering, and (c) hybrid approaches integrating both emerge as particularly promising avenues for future research in automated propaganda detection. Reasoning agents can effectively handle subtle rhetorical manipulations, bringing AI closer to human-like analytical capabilities. Multi-agent systems, meanwhile, address practical challenges such as scalability, adaptability, and cost efficiency by dividing tasks and strategically utilizing smaller models. Finally, hybrid approaches combine advanced multi-step reasoning with modular multi-agent frameworks, achieving high accuracy in detecting sophisticated rhetorical techniques without compromising cost-effectiveness or ease of deployment. As these research directions evolve, we anticipate their synthesis into robust solutions capable of detecting and deconstructing propaganda in real time across multiple languages and platforms.

Conclusion

This paper has explored modern methods of automated propaganda detection, highlighting the evolution from traditional supervised to emerging unsupervised approaches. Supervised models have demonstrated strong performance in controlled settings, particularly when supported by high-quality annotated data and ensemble strategies. However, these techniques face practical hurdles such as relying on limited labeled resources, especially as propaganda techniques continue to evolve and grow in complexity.

To address these limitations, researchers have started to gradually turn to unsupervised methods, particularly those leveraging large language models (LLMs). Approaches utilizing models such as GPT-4 offer significant promise due to their inherent flexibility and the reduced need for extensive annotated datasets. Recent experiments show that advanced LLMs — especially with well-designed prompt engineering — can match or even surpass some specialized supervised baselines. Furthermore, their interpretability offers valuable insights into the reasoning behind detection decisions, a notable advantage over supervised «black box» models. However, challenges related to precision and reliability remain significant, underscoring the need for further refinement of LLM-based approaches.

Looking ahead, three key research directions stand out. First, reasoning language models — such as OpenAI's o3 or DeepSeek's R1 — offer the potential to capture subtle propaganda through iterative, logic-based analysis. Second, multi-agent systems enable cost-effective scaling and modular specialization by distributing subtasks across lightweight, specialized models before integrating their outputs into a more precise final verdict. Third, hybrid architectures combine the strengths of the first two approaches by embedding advanced reasoning agents within modular multi-agent pipelines, balancing performance, interpretability, and efficiency. These hybrid setups are particularly promising for real-world applications, where content complexity and resource constraints can fluctuate unpredictably. Ultimately, all three approaches hold significant promise for shaping the next generation of AI-powered detection systems — systems better equipped to capture the evolving nature of propaganda and contribute to more resilient information ecosystems.

1. Jowett G.S. and O'Donnell V. Propaganda and Persuasion/ 4th ed. Thousand Oaks, CA, USA: SAGE Publications, 2006. P. 1–422.

2. Demeuse R. The Russian War on Truth: Defending Allied and Partner Democracies Against the Kremlin's Disinformation Campaigns. General Report 014 CDS 23 E rev. 2 fin, adopted by the Committee on Democracy and Security at the 2023 NATO PA Annual Session, Copenhagen, Denmark, Oct. 8, 2023.

3. Juršėnas A., Karlauskas K., Ledinauskas E., Maskeliūnas G., Rondomanskas D., and Ruseckas J. The Role of AI in the Battle Against Disinformation. NATO StratCom COE, Riga, Latvia, Feb. 2022.

4. Miller C.R. The Techniques of Propaganda. In *How to Detect and Analyze Propaganda*. Handout 10, 1939, The Center for Learning.

5. Rashkin H., Choi E., Jang J.Y., Wang Y., Volkova S., and Choi Y. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In Proc. 2017 Conf. Empirical Methods in Natural Language Process. (EMNLP). Copenhagen, Denmark, 2017, P. 2931–2937. doi: 10.18653/v1/D17-1317.

6. Da San Martino G., Yu S., Barrón-Cedeño A., Petrov R., and Nakov P. Fine-Grained Analysis of Propaganda in News Article. In Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Jan. 2019. P. 5636–5646.doi: 10.18653/v1/D19-1565.

7. Horne B.D., Khedr S., and Adalı S. Sampling the News Producers: A Large News and Feature Data Set for the Study of the Complex Media Landscape. In Proc. of the AAAI Conference on Artificial Intelligence. 2018. doi: 10.1609/icwsm.v12i1.14982.

8. Vysotska V., Przystupa K., Chyrun L., Vladov S., Ushenko Y., Uhryn D., and Hu Z. Disinformation, Fakes and Propaganda Identifying Methods in Online Messages Based on NLP and Machine Learning Methods. *International Journal of Computer Network and Information Security.* 2024. Vol. 16, No. 5. P. 57–85. doi: 10.5815/ijcnis.2024.05.06.

9. Abdullah M., Altiti O., and Obiedat R. Detecting Propaganda Techniques in English News Articles using Pre-trained Transformers. In Proc. 13th Int. Conf. on Information and Communication Systems (ICICS). 2022. P. 301–308.doi: 10.1109/ICICS55353.2022.9811117.

10. Krak I., Didur V., Molchanova M., Mazurets O., Sobko O., Zalutska O., and Barmak O. Method for Political Propaganda Detection in Internet Content Using Recurrent Neural Network Models Ensemble. In Proc. 14th International Scientific and Practical Conference from Programming UkrPROG'2024. Kyiv, Ukraine, May 14–15, 2024. CEUR Workshop Proceedings. ISSN 1613-0073.

11. Szwoch J., Staszkow M., Rzepka R., and Araki K. Limitations of Large Language Models in Propaganda Detection Task. *Applied Sciences*. 2024. Vol. 14, No. 10. Art. 4330. doi: https://doi.org/10.3390/app14104330.

12. Da San Martino G., Cresci S., Barrón-Cedeño A., Yu S., Di Pietro R., and Nakov P. A Survey on Computational Propaganda Detection. In Proc. 29th Int. Joint Conf. Artif. Intell. (IJCAI-20*)*. C. Bessiere, Ed. International Joint Conferences on Artificial Intelligence Organization, Jul. 2020. P. 4826–4832. doi: 10.24963/ijcai.2020/672.

13. Jones D.G. Detecting Propaganda in News Articles Using Large Language Models. *Eng OA*. Feb. 2024. Vol. 2, No. 1. P. 1–12.

14. Sprenkamp K., Jones D.G., and Zavolokina L. Large Language Models for Propaganda Detection. *arXiv preprint.* arXiv:2310.06422, 2023. URL: https://arxiv.org/abs/2310.06422 (Last accessed: Nov. 27, 2023).

15. OpenAI. Learning to Reason with LLMs. OpenAI, 2023. URL: https://openai.com/ index/learning-to-reason-with-llms/ (Last accessed: Mar. 04, 2025).

16. OpenAI's O3 Reasoning Model and Google Gemini. Wired. URL: https://www.wired.com/story/openai-o3-reasoning-model-google-gemini/ (Last accessed: Feb. 21, 2025).

17. OpenAI API Pricing. OpenAI. URL: https://openai.com/api/pricing/(Last accessed: Feb. 21, 2025).

18. Caballero A., Centeno R., and Rodrigo Á. LLM-Based Multi-Agent Models for Multiclass Classification of Strategic Narratives. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024). Valladolid, Spain, 2024. CEUR Workshop Proceedings.

19. Alam F., Biswas M.R., Shah U., Zaghouani W., and Mikros G. Propaganda to Hate: A Multimodal Analysis of Arabic Memes with Multi-Agent LLMs. *arXiv preprint*. arXiv:2409.07246, 2024. URL: https://arxiv.org/abs/2409.07246 (Last accessed: Mar. 18, 2025).

20. Lande D., Alekseichuk L., Svoboda I., and Strashnoy L. Methodology of a Swarm of Virtual Experts for Evaluating the Weight of Connections in Networks. *Theoretical and Applied Cybersecurity*, 2024. Vol. 6, No. 2. P. 25-33. doi: https://doi.org/10.20535/tacs.2664-29132024.2.319946.

21. Han S., Zhang Q., Yao Y., Jin W., Xu Z., and He C. LLM Multi-Agent Systems: Challenges and Open Problems. *arXiv preprint.* arXiv:2402.03578, 2024. URL: https://arxiv.org/abs/2402.03578 (Last accessed: Mar. 18, 2025).

22. Wei J., Wang X., Schuurmans D., Bosma M., Ichter B., Xia F., Chi E.H., Le Q.V., and Zhou D. Chain-of-thought prompting elicits reasoning in large language models. In Proc. 36th Int. Conf. Neural Information Processing Systems (NeurIPS '22). New Orleans, LA, USA, 2022. art. No. 1800. P. 1–14.

23. Mecklenburg N., Lin Y., Li X., Holstein D., Nunes L., Malvar S., Silva B., Chandra R., Aski V., Yannam P.K.R., Aktas T., and Hendry T. Injecting New Knowledge into Large Language Models via Supervised Fine-Tuning. *arXiv preprint.* arXiv:2404.00213, 2024. URL: https://arxiv.org/abs/2404.00213 (Last accessed: Mar. 18, 2025).

Received 27.04.2025