DOI: https://doi.org/10.15276/aait.09.2026.05
UDC 004.89, 004.912

An empirical evaluation of reasoning models for classifying information manipulation techniques

Oleg A. Boiko[1)]
ORCID: https://orcid.org/0009-0002-3424-8234; o.a.boiko@kpi.ua
Valeriy Ya. Danylov[1)]
ORCID: https://orcid.org/0009-0000-0875-4868; danilov1950@ukr.net. Scopus Author ID: 7201827051
[2)] National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 37, Beresteiskyi Ave. Kyiv, 03056, Ukraine

ABSTRACT

The rapid evolution of modern geopolitical conflicts has transformed information manipulation and propaganda techniques from a tool of persuasion into a sophisticated weapon of mass influence. As these operations become increasingly complex, relying on subtle psychological tactics rather than overt falsehoods, the development of advanced, automated detection mechanisms is a critical security challenge. This study aims to bridge the gap between theoretical capabilities and practical applications by empirically evaluating the performance of emerging reasoning models in the specific task of classifying information manipulation techniques. The primary objective is to benchmark these generative architectures against historical supervised baselines to determine if their internal chain-of-thought capabilities offer a tangible advantage over traditional pattern-matching approaches in identifying complex rhetorical strategies. The research methodology utilizes a standardized international benchmark dataset for propaganda detection (SemEval-2020 Task 11) to conduct a comparative analysis of frontier models without task-specific fine-tuning. The study employs an inference-only strategy, integrating role-playing, definition embedding, and structured reasoning instructions to simulate expert analysis. A key methodological contribution involves the systematic variation of the reasoning budget allocation during inference to measure the correlation between computational deliberation and classification accuracy. The investigation reveals a distinct semantic advantage where reasoning models significantly outperform previous supervised systems in detecting nuanced techniques that rely on cultural context, emotional weight, and indirect logic. However, the results also uncover a critical limitation where increased reasoning effort might degrade performance on structurally simple tasks, confirming the existence of an overthinking phenomenon in automated classification. The analysis further identifies a non-linear relationship between computational cost and performance, indicating that monolithic reasoning models often yield diminishing returns compared to lightweight architectures for high-volume processing. The paper concludes that while reasoning models represent a paradigm shift in semantic understanding, they are not yet a universal solution for all information manipulation types due to structural blind spots and economic inefficiencies. The study proposes moving away from single large models toward multi-agent systems. This proposed approach advocates for an adaptive system that assigns specialized tasks to a team of virtual experts, balancing precision with operational viability in the defense of the information space.

Keywords: Propaganda detection; artificial intelligence; reasoning models; large language models; information security; cognitive warfare

***For citation*: Boiko O. A., Danylov V. Ya. "An empirical evaluation of reasoning models for classifying information manipulation techniques". *Applied Aspects of Information Technology*. 2026; Vol.9 No.1: 62–75. DOI: https://doi.org/10.15276/aait.09.2026.05**

1. INTRODUCTION

Modern information manipulation and propaganda techniques have transformed from a traditional instrument of persuasion into a systematic form of cognitive warfare, aimed at shaping perception, trust, and decision-making processes rather than merely spreading false information. In contemporary geopolitical conflicts, information operations have become a full-fledged operational domain, comparable in strategic relevance to conventional military capabilities [1]. The mechanics of cognitive warfare operations have shifted from relatively simple messaging to the application of sophisticated frameworks like the "Theory of Reflexive Control" (TORC), which focuses on shaping perceptions through tailored information inputs to induce the target to voluntarily make decisions that are favorable to the aggressor [2]. Furthermore, recent empirical studies by Paziuk et al. [3] demonstrate that these manipulative narratives are often strategically synchronized with key geopolitical milestones, such as NATO summits or military aid announcements. The authors' analysis of the Russia-Ukraine war reveals that emotionally charged themes, specifically nuclear rhetoric and moral outrage, are timed to peak immediately prior to international decision-making events, aiming to induce fear and disrupt institutional cohesion. The

Russian war of aggression against Ukraine has clearly demonstrated that information manipulation can directly influence political stability and public resilience, thereby elevating information manipulation and propaganda to a critical component of hybrid warfare. A defining characteristic of modern propaganda is its structural heterogeneity. Rather than relying on uniform disinformation narratives, contemporary influence campaigns deploy a wide spectrum of rhetorical techniques, ranging from emotionally charged language and identity-based appeals to subtle logical fallacies and repetitive framing. Many of these techniques operate implicitly, leveraging cultural context, historical memory, and cognitive biases. As a result, propaganda increasingly manifests not as overt deception, but as semantically nuanced discourse embedded within otherwise legitimate informational content.

This evolution poses a fundamental challenge for automated detection systems. Traditional supervised learning approaches, based on Transformer architectures fine-tuned on annotated datasets, have demonstrated strong performance on established benchmarks. However, such models remain inherently constrained by pattern memorization, sensitivity to training data distributions, and limited capacity for contextual reasoning. While effective at recognizing structurally explicit techniques, they often struggle with indirect argumentation, culturally grounded references, and rhetorical ambiguity. Consequently, there exists a growing mismatch between the sophistication of modern propaganda and the analytical depth of classical detection pipelines. Recent advances in Large Language Models (LLMs) and, more specifically, Reasoning Models (also known as Reasoning Language Models or RLMs) offer a potential pathway beyond these limitations. By incorporating multi-step chain-of-thought reasoning and leveraging extensive world knowledge, RLMs promise enhanced semantic understanding and interpretability. In principle, such capabilities align well with the requirements of propaganda analysis, where correct classification often depends on implicit logic, pragmatic inference, and broader sociopolitical context rather than surface-level cues alone. At the same time, emerging evidence suggests that reasoning is not universally beneficial. While deeper deliberation may improve performance on semantically complex techniques, it can also introduce degradation on structurally simple or highly formulaic patterns. Excessive reasoning may lead models to hallucinate hidden intent, over-interpret straightforward expressions, or deviate from classification objectives. Moreover, reasoning-intensive models introduce significant computational and economic overheads, raising questions about their suitability for large-scale, real-time monitoring scenarios. Against this backdrop, the present study adopts a deliberately non-utopian perspective on reasoning models. Rather than treating RLMs as a universal replacement for existing systems, this study empirically examines where they help, where they fail, and at what cost. Using the widely cited SemEval-2020 Task 11 benchmark for propaganda technique classification, modern RLMs are evaluated in an inference-only setting and compared against both traditional supervised baselines and standard LLM architectures. A central focus of this work is the systematic analysis of reasoning budget allocation (the term is explained in detail “Research aim and objectives” section), exploring how varying degrees of test-time deliberation affect accuracy, latency, and economic efficiency. By aligning empirical evaluation with operational constraints, this research aims to move beyond headline accuracy scores toward deployable design principles for propaganda detection systems. The study ultimately frames propaganda analysis not as a task solvable by a single monolithic model, but as a problem that naturally lends itself to adaptive, hybrid, and multi-agent architectures, where different models contribute complementary strengths. In doing so, the Introduction sets the conceptual foundation for the Conclusions, which argue for a transition from isolated reasoning engines toward economically viable, context-aware systems capable of defending democratic information spaces against evolving cognitive threats.

Recent research highlights that Russian information operations have expanded beyond Ukraine, entering an early phase of potential conflict with NATO by aggressively targeting Western stability [4]. Beyond this expansion, the mechanisms themselves are becoming increasingly insidious. Paziuk et al. [3] identify that modern cognitive operations prioritize emotional manipulation, exploiting specific cognitive biases such as confirmation bias and the availability heuristic. By analyzing the “Attack-Index” of Russian narratives, they found that over 52 % of propaganda content relies on negative emotional triggers, particularly fear and existential anxiety, to bypass rational scrutiny. Bazdyrev [5] highlights a sophisticated

"data poisoning" threat, where hostile actors flood the internet with AI-generated pseudo-analytics to bias the training datasets of future Large Language Models (LLMs). The saturation of the digital ecosystem with subtle narratives, such as the efficacy of sanctions or historical revisionism, creates the risk that future AI models will inherently rely on these narratives as facts. Consequently, the development of automated systems capable of detecting and classifying these techniques in real-time is no longer an academic exercise but rather a security imperative.

2. LITERATURE REVIEW AND PROBLEM STATEMENT

Traditional supervised learning approaches to propaganda detection have historically relied on extensive, human-annotated datasets. However, recent advancements suggest a paradigm shift toward unsupervised and semi-supervised methods utilizing LLMs. Pina-García [6] demonstrated that In-Context Learning (ICL) with few-shot prompting can rival resource-intensive fine-tuning, achieving F_1 scores up to 90.1 %. This suggests that "agile prompting" may offer a more adaptable solution to the evolving nature of propaganda than static model training.

However, to fully exploit this potential, the prompting strategy itself must be carefully engineered. Gaeta et al. [7] demonstrated that optimal performance is achieved not through a single technique, but via a composite strategy integrating Role-Playing, Few-Shot Learning, and, most critically, Chain-of-Thought (CoT) reasoning. Aligning with these best practices, this research utilizes a multi-layered prompting strategy that incorporates Role-Playing (assigning the model a persona of an "officer" and "experienced international expert"), Chain-of-Thought instructions, and explicit in-context definitions of propaganda techniques paired with short examples. These elements are augmented by XML content structure for clear data delimitation and structured output constraints to ensure machine-readable JSON responses.

Furthermore, current research addresses a critical methodological limitation regarding the evaluation of propaganda detection performance that was mentioned by Gaeta et al. [7]. In their study, the authors utilized the SemEval-2020 Task 11 dataset [8], particularly subtask SI (Span Identification), pinpointing the exact start and end indices of a propagandist fragment. However, they observed that the strict metric requiring exact character offset matching often penalized models for identifying spans that captured the correct semantic content but deviated slightly in their precise start and end indices. To mitigate this metric-induced noise and focus specifically on the model's reasoning capabilities rather than boundary precision, current research adopts the alternative challenge within the same SemEval-2020 framework: Subtask TC (Technique Classification). Instead of locating the text, the model is provided with specific fragments and tasked with determining the correct propaganda technique from an inventory of 14 distinct classes.

While the capabilities of standard Large Language Models can be significantly enhanced through advanced prompt engineering, the emergence of Reasoning Models may offer an even more robust architecture for complex semantic tasks, with their native chain-of-thought processing. This evolution marks a paradigm shift analogous to the cognitive transition described by Li et al. [9]: moving from the 'System 1' fast, intuitive pattern-matching of foundational models to the 'System 2' slow, deliberate, and logical reasoning inherent in RLMs.

This shift is backed by McGinness and Baumgartner [10], who tracked the frontier models for reasoning performance over 18 months. They found that the 'thinking models' introduced in 2025 represent a distinct leap in capability, allowing them to finally imitate complex logical strategies that earlier models could not handle.

Furthermore, the value of this 'System 2' approach extends beyond accuracy. As noted by LekshmiAmmal and Madasamy [11], mere classification is often insufficient; providing users with the reasoning behind a decision, whether generated by an RLM's chain-of-thought or external explainability tools, is essential for combating sophisticated misinformation.

However, deploying these reasoning capabilities requires caution. Hu and Tian [12] found that in rumor detection tasks, reasoning models often underperform traditional baselines. Notably, they observed a negative correlation where longer reasoning chains actually led to lower accuracy, suggesting that unconstrained 'thinking' can sometimes be detrimental.

This paper serves as a direct empirical continuation of previous work by the authors, 'Modern AI methods for detecting propaganda in text' [13], which outlined the theoretical potential of reasoning-capable agents. Moving from theory to practice, the performance is carefully evaluated of modern RLMs, such as OpenAI's GPT-5 and Google's Gemini 2.5 series, against standard LLMs

and historical supervised baselines. By analyzing how "Reasoning budget" correlates with classification accuracy, the study aims to quantify the trade-offs between precision, latency, and cost, providing a roadmap for deploying AI in the defense of democratic information spaces.

3. RESEARCH AIM AND OBJECTIVES

The aim of this research is to empirically evaluate the performance and economic efficiency of Reasoning Language Models (RLMs) in the classification of information manipulation techniques. This evaluation is intended to establish evidence-based design principles for the development of resilient, cost-effective, and automated propaganda detection systems capable of defending the information space in real-world scenarios.

To achieve this aim, the following objectives are addressed:

- **Benchmarking RLMs**: Evaluate the current performance level of reasoning models (Gemini 2.5, GPT-5, and others) on the standardized SemEval-2020 test set, comparing them against the official leaderboard.
- **Evaluating the Efficacy of Reasoning**: Measure how Reasoning budget correlates with classification accuracy (F_1 score). In this study the term "Reasoning budget" is used to refer to the allocation of test-time compute. This parameter controls the volume of internal 'thought tokens' the model generates to deliberate before producing a final response. Although Google and OpenAI use different terminology ("Thinking budget" [14] and "Reasoning effort" [15] respectively), both parameters serve the same function: regulating the depth of the model's reasoning process.
- **Technique-Specific Diagnostics**: Provide a granular analysis of which propaganda techniques are effectively solved by reasoning capabilities and which remain elusive, offering a roadmap for future hybrid systems.
- **Cost-Benefit Analysis**: Quantify the trade-offs between precision, latency, and financial cost, identifying the optimal model configurations for real-world deployment.

4. METHODOLOGY

4.1. Dataset and Task Specification

This study utilizes the dataset from SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles, specifically Subtask TC (Technique Classification) [8]. This dataset is widely regarded as the benchmark for fine-grained propaganda analysis.

Data Structure: The input consists of a news article text and a specific span (start/end indices) identified as propagandistic. The model must classify this span into one of 14 classes.

Classes: The taxonomy includes *Appeal to Authority, Appeal to Fear-Prejudice, Bandwagon, Reductio ad Hitlerum, Black-and-White Fallacy, Causal Oversimplification, Doubt, Exaggeration/Minimization, Flag-Waving, Loaded Language, Name Calling/Labeling, Repetition, Slogans, Thought-terminating Clichés, Whataboutism/Straw Man/Red Herring* [8].

Metric: The micro-averaged F_1 score is employed as the primary performance metric, ensuring comparability with the official SemEval-2020 leaderboard. It is important to note that the gold labels for the test set are withheld by the organizers. Instead, they provide an official online evaluation system [16], where all results of this study were submitted and verified (Team OB1), ensuring an unbiased and independent validation.

4.2. Formal Task Definition

Following the framework established by Da San Martino et al. [8], the Technique Classification (TC) task is formally defined as a multi-class classification problem.

Let D be a corpus of news articles. Each article $d \in D$ contains a set of identified text spans (snippets) that are known to be propagandistic. Let t denote a specific text fragment within document d, defined by character offsets (start, end) within the article text. Let $Y = \{y_1, y_2, \ldots, y_{14}\}$ be the set of $K = 14$ distinct propaganda techniques defined in the taxonomy (e.g., *Loaded Language*, *Name Calling*, *Flag-Waving*, etc.).

The objective is to learn a mapping function f (or, in the case of Reasoning Models, to approximate this function via inference) that predicts the correct label y_{pred} for a given span t given its context d:

$$y_{pred} = f(t, d), \text{ where } y_{pred} \in Y$$

Because a small fraction of identical spans are annotated with multiple techniques (about 1.8% of the total), TC is formally multi-label, however, SemEval converts it to single-label for evaluation via duplicated instances and best-match scoring for identical spans [8].

4.3. Evaluation Metric

To align with the SemEval-2020 official leaderboard, the micro-average F_1 score ($F_{1\ micro}$) is

reported for Subtask TC [8]. Under this single-label multi-class evaluation protocol, the organizers note that $F_{1\ micro}$ is equivalent to Accuracy (as well as to Precision and Recall):

$$F_{1\,micro} = Accuracy = \frac{\sum_{k=1}^{K} TP_k}{N},$$

where TP_k is the number of true positives for class k and N is the total number of test instances. The F_1 notation is retained throughout this paper to maintain consistency with the comparative literature and the official task description.

4.4. Model Selection

This study evaluates a suite of frontier models representing the current state-of-the-art in both standard LLM (without an explicit reasoning mode) and reasoning (RLM) architectures. The models were accessed via API, utilizing their respective Reasoning budget parameters where available. The full list of evaluated models, their specific configurations, and execution details are presented in Table 1 in Section 5.

Table 1. **Complete list of evaluated models, configurations, and overall performance**

Provider	Model Name	Reasoning budget	Type	# of runs	Avg Time per snippet (s)	Avg Price per 1000 snippets (USD)	F_1
Google	gemini-2.5-pro	Auto	RLM	3	3.568	3.918	56.3315
Google	gemini-2.5-pro	Maximum	RLM	3	3.765	3.993	55.3259
OpenAI	gpt-5	Maximum	RLM	3	6.25	3.85	55.0093
Google	gemini-2.5-flash	None*	RLM	3	0.894	0.51	54.9534
Google	gemini-2.5-pro	Minimum**	RLM	3	2.042	2.312	54.8417
OpenAI	o3	Default	RLM	4	3.205	2.011	54.581
OpenAI	gpt-5	Medium	RLM	3	3.866	2.413	54.0224
Google	gemini-2.5-flash	Auto	RLM	2	2.571	1.103	53.9944
Google	gemini-2.5-flash	Maximum	RLM	3	2.441	1.088	53.5754
OpenAI	gpt-5	Minimum	RLM	3	1.995	1.365	53.3147
OpenAI	gpt-5-mini	Medium	RLM	2	3.676	0.469	52.6536
OpenAI	o4-mini	Default	RLM	4	2.507	1.256	52.1089
OpenAI	gpt-4.1	None	LLM	4	2.506	1.651	50.8938
Google	gemini-2.0-flash	None	LLM	3	1.062	0.143	50.838
OpenAI	gpt-5-mini	Maximum	RLM	2	9.47	1.033	50.7262
OpenAI	gpt-5-mini	Minimum	RLM	2	2.3	0.331	49.4414
Google	gemini-2.0-flash-lite	None	LLM	3	1.008	0.074	49.2737
OpenAI	gpt-4.1-mini	None	LLM	4	2.608	0.33	48.1564
Google	gemini-2.5-flash-lite	Maximum	RLM	3	1.585	0.227	47.8585
Google	gemini-2.5-flash-lite	Auto	RLM	3	1.592	0.211	46.9274
OpenAI	gpt-5-nano	Maximum	RLM	3	11.353	0.309	46.0894
Google	gemini-2.5-flash-lite	None	RLM	3	0.808	0.105	45.6052
OpenAI	gpt-5-nano	Medium	RLM	3	3.459	0.147	44.9348
OpenAI	gpt-4.1-nano	None	LLM	3	1.225	0.079	41.8436
OpenAI	gpt-5-nano	Minimum	RLM	4	1.208	0.059	41.4944

Source: **compiled by the authors**

* For Gemini 2.5 family models, the parameter “Thinking budget: 0” is designated as “Reasoning budget: None” to simplify direct comparison across different providers.

** Minimal reasoning budget was used for gemini-2.5-pro and gpt-5 family models, since their reasoning cannot be turned off.

4.5. Evaluation Protocol

For each model and reasoning configuration, the following steps were performed:

- **Prompting**: A structured, multi-component prompt was designed and utilized to maximize reasoning capabilities via Few-Shot In-Context Learning (ICL). Unlike supervised approaches that update model weights, this study relies on an inference-only Knowledge-Injected Prompting strategy. Compact definitions and 2-4 prototypical examples for all 14 classes were explicitly embedded directly within the context window. The input was organized using XML tags (e.g. <article_full_text>, <classification_task_details>) to clearly delimit context, instructions, and data. The strategy further integrated Role-Playing (assigning the model a specific persona to establish authority) and Chain-of-Thought instructions (explicitly requesting a step-by-step rationale to weigh evidence before classification). Finally, strict Output Constraining was applied via JSON schema enforcement to ensure consistent, machine-readable results.
- **Execution**: Multiple runs (n = 2 to n = 4) were executed for each configuration to account for non-deterministic outputs. Exact number of runs for each configuration is available in Table 1 in Section 5.
- **Metrics Calculation**: The F_1 score was evaluated and recorded via the official online evaluation tool [16], average latency per snippet (seconds), and cost per 1,000 snippets (USD).
- **Comparison**: Results were compared against the official SemEval-2020 leaderboards, including the top-performing systems (ApplicaAI, aschern, Hitachi, and others).

5. RESEARCH RESULTS

A total of 76 runs across 25 different setups (detailed in Table 1) were validated using the official online evaluation tool [16] under the team name 'OB1'. Collectively, these experimental results present a rich dataset for understanding the capabilities and limitations of RLMs. These results are analyzed across three dimensions: overall performance, the impact of reasoning budget, and economic efficiency.

To provide a clear comparison between different architectures, the ranking tables in Sections 5.1 through 5.4 report the “Best Distinct Model” performance. This means that for each model family (e.g., Gemini 2.5 Pro, GPT-5, o3), only the single configuration (Reasoning Budget) that achieved the highest F1 score for that specific category is reported. A detailed analysis of how different reasoning budgets affect performance within the same model family is provided separately in Section 5.5.

5.1. Overall Performance

The experimental results indicate that RLMs without task-specific fine-tuning have reached an overall performance level competitive with, though not yet superior to, the top-tier supervised models from 2020. The highest-performing RLM, Google Gemini 2.5 Pro (Reasoning budget: Auto), achieved an overall F1 score of 56.33. While this represents a significant achievement for a model operating without task-specific training, it remains below the benchmark set by the 2020 winner, ApplicaAI (RoBERTa-CRF), which achieved an F1 of 63.74 [8].

However, this aggregate score masks a fundamental divergence in capability. A granular analysis reveals that RLMs do not simply underperform uniformly; rather, they fundamentally alter the detection landscape, achieving substantial advantage in semantic tasks while suffering significant regressions in structural pattern matching.

5.2. The Semantic Advantage: Where RLMs Excel

Reasoning models demonstrated a decisive advantage in identifying techniques that rely on linguistic nuance, emotional weight, and cultural context – areas where traditional BERT-based models often struggle due to limited world knowledge.

Bandwagon & Reductio ad Hitlerum (+38.67): The most dramatic improvement was observed in this category. GPT-5 (Medium) achieved an F_1 of 67.24 (Table 2), significantly outperforming the 2020 benchmark (Team JUST) of 28.57 (Table 3) – a massive 38.67 percentage point increase.

Table 2. **Bandwagon & Reductio ad Hitlerum: LLM/RLM Top 3**

Rank	Model / Reasoning budget	F_1
1	OpenAI gpt-5 / Medium	67.24
2	OpenAI o3 / Default	65.25
3	OpenAI gpt-4.1 / None	64.61

Source: **compiled by the authors**

Table 3. **Bandwagon & Reductio ad Hitlerum: Official SemEval-2020 Top 3**

Rank	Team / Setup	F_1
1	**JUST**/ BERT	28.57
2	**ApplicaAI** / RoBERTa	28.13
3	**Hitachi** / BERT + RoBERTa + XLNet + XLM + XLM RoBERTa + ALBERT	26.92

Source: **compiled by the authors**

Detecting *Reductio ad Hitlerum* requires specific contextual knowledge (which RLMs possess in abundance) to understand the rhetorical tactics of drawing comparisons to Nazis or Hitler. Similarly, *Bandwagon* appeals often use subtle social pressure ("everyone knows," "the people want") which RLMs can parse more effectively than keyword-based systems.

Flag-Waving (+27.06): The RLMs' ability to decode cultural context is further exemplified by their performance on *Flag-Waving*. Gemini 2.5 Pro scored 66.49 (Table 4), compared to the 2020 SOTA (Team NoPropaganda) of 39.43 (Table 5). Detecting *Flag-Waving* requires identifying appeals to group identity, patriotism, and nationalistic ideals. RLMs likely excel here because they can reason about the implication of symbols (e.g., "The American Way", "Our brave troops") and connect them to the abstract concept of nationalism.

Table 4. **Flag-Waving: LLM/RLM Top 3**

Rank	Model / Reasoning budget	F_1
1	Google gemini-2.5-pro / Min	66.49
2	Google gemini-2.5-flash / None	65.54
3	OpenAI gpt-5-mini / Medium	64.67

Source: **compiled by the authors**

Table 5. **Flag-Waving: Official SemEval-2020 Top 3**

Rank	Team / Setup	F_1
1	**NoPropaganda** / BERT + R BERT	39.43
2	**Aschern** / RoBERTa	37.55
3	**Hitachi** / BERT + RoBERTa + XLNet + XLM + XLM RoBERTa + ALBERT	37.38

Source: **compiled by the authors**

Name Calling or Labeling (+19.46): This technique, which involves labeling a target with something the audience fears or hates [8], also saw a significant boost. Gemini 2.5 Pro reached an F_1 of 66.52, surpassing the 2020 winner (ApplicaAI) at 47.06 by over 19 percentage points. While supervised models can memorize specific insults ("crook"), RLMs can identify context-dependent labeling (e.g., "illegitimate regime") that functions as a pejorative mostly within specific political narratives.

Loaded Language (+16.74): This technique corresponds to the usage of words with strong emotional implications to influence the audience [8]. RLMs dominated this category. Gemini 2.5 Pro achieved an F_1 of 67.90, outperforming the 2020 best (Team DiSaster) score of 51.16 by nearly 17 percentage points. This suggests that the massive pre-training and "world model" of RLMs allow them to detect subtle emotional triggers and connotative meanings that supervised models often fail to capture.

Doubt (+13.85): In identifying appeals to *Doubt* (questioning credibility), GPT-5 Mini (F_1 = 50.47) significantly outperformed the 2020 SOTA (Team Hitachi, F_1 = 36.62). This technique often involves rhetorical questions or casting defamations without making factual claims, a rhetorical maneuver that benefit from the logical processing capabilities of RLMs.

5.3. Structural Deficiency: Where RLMs Fail

Conversely, RLMs struggled significantly with techniques that are defined by rigid structural patterns or simplistic logical fallacies, areas where supervised models excel due to pattern memorization.

Slogans (-25.67): This was the most significant failure mode for RLMs. The best RLM (Gemini 2.5 Flash) achieved only 52.65 (Table 6), while the 2020 SOTA (Team aschern) reached 78.32 (Table 7). *Slogans* are often short, punchy, non-standard sentences (e.g., "Make America Great Again"). Supervised models can easily overfit to the syntactic structure of slogans. RLMs, trained to prioritize coherence and complex reasoning, may fail to classify a short phrase as a "technique" because it lacks the argumentative complexity they are designed to analyze. They may "overthink" the simplicity of a slogan. This finding aligns with the results of Chen et al. [17], who demonstrated that for tasks with low intrinsic complexity (e.g., analogous to simple arithmetic), enforcing a reasoning budget introduces probabilistic noise, causing the model to override correct initial intuitions with hallucinated constraints.

Table 6. **Slogans: LLM/RLM Top 3**

Rank	Model / Reasoning budget	F_1
1	Google gemini-2.5-flash / None	52.65
2	OpenAI gpt-5-mini / Max	47.59
3	Google gemini-2.5-flash-lite / None	47.21

Source: **compiled by the authors**

Table 7. **Slogans: Official SemEval-2020 Top 3**

Rank	Team / Setup	F_1
1	Aschern / RoBERTa	78.32
2	**ApplicaAI** / RoBERTa	78.27
3	**NoPropaganda** / BERT + R BERT	77.99

Source: **compiled by the authors**

Causal Oversimplification (-30.0): RLMs underperformed here as well (40.47 for gemini-2.5-flash vs. 70.47 for ApplicaAI). This technique typically follows a simple "X caused Y" template. Possible explanation might be that RLMs, when attempting to reason through the causality, might hallucinate a valid logical link or consider the simplification "justified" within the context, leading to false negatives.

5.4. The Role of Context: Solving the "Repetition" Problem

A standout finding is the performance on the *Repetition* (+20.2) technique. Historically, this has been the hardest technique for Transformer models because it requires tracking information across long distances in the text, often exceeding the 512-token limit of models like BERT.

Gemini 2.5 Pro achieved an F_1 of 48.77, nearly doubling the performance of the best 2020 system (Team PALI, F_1 = 28.57) and far surpassing typical scores which were often below 15. This is probably a direct consequence of the context window advantage. Modern RLMs with windows of up to 1M tokens can "see" the entire article at once, allowing them to identify when a specific phrase is repeated 10-20 sentences later. This effectively removes the context-length limitations that affected earlier transformer models.

5.5 The "Overthinking" Phenomenon: Reasoning budget vs. Accuracy

A critical contribution of this study is the empirical verification of the "overthinking" phenomenon in classification tasks. The relationship between reasoning budget and accuracy is not linear; rather, it is model-dependent and often non-monotonic.

Insight 1: The Penalty of Overthinking. The Gemini 2.5 Flash model exhibits a clear negative correlation between reasoning and accuracy. The "None" setting (standard inference) achieves the highest F_1 (54.95). Enabling "Auto" or "Maximum" reasoning degrades performance to 53.58 while tripling the latency. This suggests that for this latency-optimized architecture, the "System 1" intuition is more reliable than its "System 2" deliberation [9]. When forced to reason, the model likely hallucinates complexities or misinterprets standard rhetorical flourishes as specific propaganda techniques, effectively reasoning itself away from the correct label. This aligns with the findings of Chen et al. [17], who found that excessive "thought tokens" might degrade accuracy by introducing unnecessary intermediate steps that drift away from the ground truth.

Table 8. **Impact of Reasoning budget on F1 Score and Latency**

Model	Reasoning budget	F_1	Avg. Time (s)	Trend
Gemini 2.5 Pro	Minimum	54.84	2.04	Inverted U (peak at Auto)
	Auto	56.33	3.57	
	Maximum	55.33	3.77	
Gemini 2.5 Flash	None	54.95	0.89	Negative (more reasoning = worse)
	Auto	53.99	2.57	
	Maximum	53.58	2.44	
GPT-5	Minimum	53.31	2.00	Positive (more reasoning = better)
	Medium	54.02	3.87	
	Maximum	55.01	6.25	

Source: **compiled by the authors**

Insight 2: The Scaling of Capability. Conversely, GPT-5 demonstrates a positive scaling law. "Maximum" reasoning yields the best results, suggesting that a sufficiently capable base model can utilize additional test-time compute to resolve ambiguities that confuse the faster modes. This aligns with the findings of Snell et al. [18], where compute-optimal scaling benefits larger, more capable models.

Insight 3: The "Auto" Optimization. The flagship Gemini 2.5 Pro performs best at "Auto". This indicates that the internal router of this model is successfully identifying which spans require deep thought and which do not. Forcing "Maximum" reasoning on easy spans introduces noise (overthinking), while "Minimum" reasoning fails on hard spans (underthinking). The "Auto" setting effectively navigates this trade-off.

The visual analysis in Fig. 1 confirms the three distinct behavioral patterns identified in Table 8. While some models (in green) benefit linearly from increased compute, others (in blue) exhibit a distinct 'Inverted U' shape. This specific curvature aligns with the empirical findings of Chen et al. [17], confirming that for these architectures, the 'Auto' setting effectively prevents the overthinking degradation observed in maximum reasoning configurations by halting generation before the onset of probabilistic noise.

5.6 Cost-Efficiency Analysis: The Price of Reasoning

The deployment of propaganda detection systems at scale (e.g., social media monitoring) is significantly determined by operational costs. The analysis reveals massive disparities in economic efficiency, as detailed in Table 9.

Insight 4: **The Diminishing Returns of Intelligence**. To achieve the 1.38 F_1 point gain from Gemini 2.5 Flash (None) (54.95) to Gemini 2.5 Pro (Auto) (56.33), the cost increases by 668 % ($0.51 to $3.92). For practical applications, Gemini 2.5 Flash (Reasoning budget: 0) represents the optimal solution on the Pareto frontier. It offers near-SOTA performance (within 1.3 points of the best RLM) at a fraction of the cost and with the highest throughput (61 F_1/sec). The specialized reasoning models (o3, GPT-5 Max) are economically unviable for large-scale streaming data, providing negligible accuracy gains for exponential cost increases.

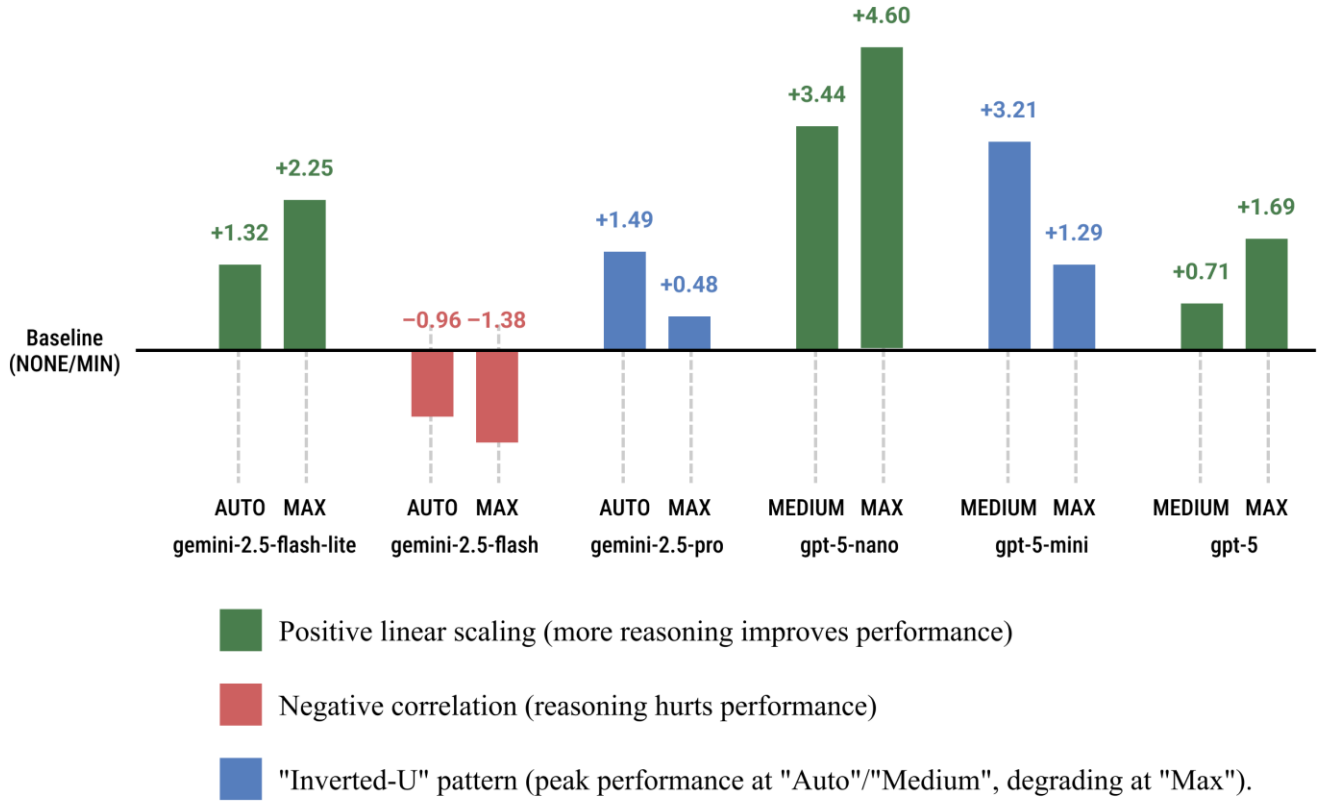

Fig. 1. **Impact of Reasoning Budget on F_1 performance. The chart displays the percentage point change in F_1 score relative to each model's baseline (Reasoning budget: None/Min)**

Source: **compiled by the authors**

Table 9. **Cost-Benefit Analysis (Top Models)**

Model	Reasoning budget	F_1	Cost ($/1000)*	F_1 per USD**	F_1 per Second***
GPT-5 Nano	Minimum	41.49	$0.059	703.29	34
Gemini 2.0 Flash	None	50.84	$0.143	355.51	48
Gemini 2.5 Flash	None	54.95	$0.510	107.75	61
GPT-5	Maximum	55.01	$3.850	14.29	9
Gemini 2.5 Pro	Auto	56.33	$3.918	14.38	16

Source: **compiled by the authors**

* Cost ($/1k): The average cost to process 1,000 news snippets, calculated based on the specific input/output/thinking token usage of each model and their respective API pricing as of November 1, 2025

** F_1 per USD: A cost-efficiency metric calculated as F_1 Score ÷ Cost per 1k snippets. Higher values indicate a better return on investment per unit of accuracy.

*** F_1 per Second: A time-efficiency metric calculated as F_1 Score ÷ Average Latency per snippet (s). **Note**: This composite metric prioritizes processing speed. It should be interpreted alongside the raw F_1 score, as it naturally favors lower-latency models (System 1) over slower, reasoning-intensive architectures (System 2).

While "F_1 per Second" is reported to highlight throughput efficiency, it is important to note that this metric heavily penalizes reasoning models (which naturally require more inference time). Therefore, it is most useful for identifying models suitable for high-velocity, real-time streaming applications where latency is a hard constraint.

6. DISCUSSION OF RESULTS

The empirical results presented in this study offer a nuanced view of the current capabilities of Reasoning Models in the domain of automated propaganda technique classification. While the quantitative metrics establish a performance hierarchy, they also reveal a counter-intuitive trade-off: models optimized for deep reasoning excel at semantic nuance but often falter on simple structural patterns, whereas simpler models capture structure efficiently but lack the depth for complex analysis. This divergence highlights a promising opportunity for transitioning from monolithic architectures to modular, multi-agent systems.

6.1. Generalizability vs. Specialization

It is critically important to contextualize the performance gap between the RLMs evaluated in this study and the top-performing systems from the SemEval-2020 leaderboard. The benchmark models, such as those by ApplicaAI and Hitachi, benefited from extensive supervised fine-tuning on the official training set, allowing them to memorize specific domain patterns [8]. In contrast, the RLMs in this study were evaluated in an inference-only, in-context learning setting. They relied solely on their pre-trained world model and the definitions provided in the prompt, without undergoing any gradient updates or parameter fine-tuning. The fact that the best-performing RLM (Gemini 2.5 Pro) achieved an F_1 score within 7.4 points of the state-of-the-art supervised baseline without undergoing parameter updates demonstrates the remarkable adaptability of reasoning models. This suggests that RLMs may be more robust in real-world scenarios where propaganda narratives evolve rapidly, rendering some of the static fine-tuned models obsolete. Furthermore, as noted by Xu et al. [19], traditional reasoning benchmarks like GSM8K (Math) and HumanEval (Code) are becoming saturated. This study addresses their call for "complex, dynamic, and interdisciplinary" evaluation frameworks, demonstrating that propaganda detection serves as a promising frontier for testing the limits of System 2 reasoning in non-deterministic domains.

6.2. The "System 1" vs. "System 2" Trade-off

The findings regarding the "overthinking" dilemma align with the cognitive distinction between System 1 (fast, intuitive) and System 2 (slow, deliberative) thinking [9]. The benefits of System 2 are undeniable for high-complexity tasks: reasoning models significantly outperformed the supervised baselines on semantically layered techniques, achieving massive gains in challenges like *Bandwagon & Reductio ad Hitlerum* (+38.67

points) and *Flag-Waving* (+27.06 points). However, maximum reasoning budget was not universally beneficial. For simpler techniques, like *Slogans*, RLMs configured with “None” reasoning budget, and even standard LLMs, often achieved higher precision. In these cases, the reasoning models frequently “over-reasoned,” hallucinating complexity or hidden intent where none existed. This confirms the hypothesis by Liu et al. [20] that Chain-of-Thought (CoT) prompting can reduce performance on tasks that require intuitive pattern recognition rather than multi-step logic.

6.3. Economic Viability and Latency

The cost-benefit analysis reveals that a monolithic approach, like using a single massive reasoning model for all tasks, is economically unsustainable for high-volume environments. The 668% cost increase required to achieve a marginal 1.38-point F_1 gain (moving from Gemini 2.5 Flash to Pro) highlights the need for a more stratified approach. Recent work by McGinness and Baumgartner [10] supports this view. They found that pairing smaller, cheaper language models with external tools could match the accuracy of massive frontier models while cutting computational costs significantly.

Additionally, for real-time applications, such as monitoring social media feeds, the latency of "Maximum" reasoning (up to 6.25 seconds per snippet) may be prohibitive. These points toward the opportunity for hybrid architectures that can dynamically allocate reasoning budget based on the estimated complexity of the input text and its context.

6.4. Future research

The future of automated propaganda detection likely lies not in larger monolithic models, but in Multi-Agent Systems (MAS). A promising architectural framework for realizing this goal is the “Swarm of Virtual Experts” methodology, as proposed by Lande et al. [21]. In this framework, diverse agents act as specialized Critics, Analysts, or Moderators. To achieve true multidimensional analysis, this swarm might incorporate a specialized 'Contextual Resonance' agent based on the Attack-Index methodology described by Paziuk et al. [3]. This agent would specifically evaluate the narrative's temporal synchronization with geopolitical events and its emotional intensity. By combining the RLM's semantic depth with the Attack-Index's temporal awareness, the system can mitigate the 'structural blind spots' inherent in purely linguistic approaches.

However, to ensure such a complex system remains economically viable and stable, this study proposes integrating adaptive decision support frameworks based on system analysis principles, as defined by Danylov et al. [22]. In this model, instead of activating the full Swarm for every input, a central 'Dispatcher' agent would first analyze the text's complexity. Drawing on the 'adaptive reasoning depth' framework validated by Alghamdi et al. [23], this Dispatcher would dynamically allocate computational resources. To optimize this allocation under uncertainty, the system could employ Monte Carlo simulation techniques, similar to those applied by Lytvynenko et al. [24] for risk modeling in complex chains. By treating agent performance as a stochastic variable, the Dispatcher can proactively forecast the likelihood of reasoning failure, routing simple inputs to efficient models (System 1) and reserving the multi-agent Swarm (System 2) only for high-risk tasks.

Furthermore, addressing the challenges of coordination within these systems is vital. As highlighted by Han et al. [25], effective global planning is required to ensure that the "Discussion" between agents converges on a correct classification. To mitigate the risk of hallucinatory drift, the system might benefit from the 'consensus voting' and 'inter-agent conflict resolution' protocols proposed by Alghamdi et al. [23], which have been empirically shown to stabilize reasoning traces in multi-agent environments.

7. CONCLUSIONS

This research successfully achieved its primary aim of providing a rigorous empirical evaluation of Reasoning Models (RLMs) for fine-grained propaganda technique classification. By utilizing the standardized SemEval-2020 Task 11 benchmark, the following objectives were met:

- **Benchmarking**: Current RLMs (Gemini 2.5 Pro and GPT-5) were found to be competitive with historical supervised SOTA models in semantic understanding, despite lacking task-specific fine-tuning.
- **Efficacy of Reasoning**: The study quantified the non-monotonic relationship between reasoning budget and accuracy, confirming the existence of an "overthinking" phenomenon in low-complexity tasks.
- **Technique Diagnostics**: Analysis identified a clear performance asymmetry; RLMs excel at semantically nuanced techniques but struggle with structural patterns like Slogans.

- **Cost-Benefit Analysis**: The investigation established that for large-scale deployment, monolithic reasoning models currently yield diminishing returns in terms of economic and temporal efficiency compared to optimized architectures.
- The findings confirm that while RLMs represent a paradigm shift in semantic discourse analysis, they are not yet a universal solution for all information manipulation types. The results provide a necessary foundation for the development of hybrid, adaptive, multi-agent detection systems.

8. ACKNOWLEDGMENTS

The authors express their deepest gratitude to the Security and Defense Forces of Ukraine (Сили безпеки та оборони України) for the opportunity to conduct this research.

REFERENCES

1. 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, NATO Parliamentary Assembly*. Copenhagen, Denmark. 2023.

2. Morley-Davies, J., Thomas, J. & Baines, G. “Russian information operations outside of the Western information environment”. *NATO Strategic Communications Centre of Excellence*. Riga, Latvia. 2025.

3. Paziuk, A., Lande, D., Shnurko-Tabakova, E. & Kingston, P. “Decoding manipulative narratives in cognitive warfare: a case study of the Russia-Ukraine conflict”. *Front. Artif. Intell.* 2025; 8: 1566022, https://www.scopus.com/pages/publications/105017058174. DOI: https://doi.org/10.3389/frai.2025.1566022.

4. “Virtual manipulation brief: Hijacking reality - The increased role of generative AI in Russian propaganda”. *NATO Strategic Communications Centre of Excellence*. Riga, Latvia. 2024.

5. Bazdyrev, A. “Russo-Ukrainian war disinformation detection in suspicious Telegram channels”. In *Proc. 4th Int. Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2024)*. Cambridge, MA, USA. 2024. p. 1–9, https://www.scopus.com/pages/publications/85210091018. DOI: https://doi.org/10.48550/arXiv.2503.05707.

6. Pina-García, C. A. “In-context learning for propaganda detection on Twitter Mexico using large language model meta AI”. *Telematics and Informatics Reports*. 2025; 19: 100232, https://www.scopus.com/pages/publications/105011966389. DOI: https://doi.org/10.1016/j.teler.2025.100232.

7. Gaeta, A., Loia, V., Lorusso, A., Orciuoli, F. & Pascuzzo, A. “Towards a LLM-based intelligent system for detecting propaganda within textual content”. *Computers and Electrical Engineering*. 2025; 128: 110765, https://www.scopus.com/pages/publications/105019760806.
DOI: https://doi.org/10.1016/j.compeleceng.2025.110765.

8. Da San Martino, G., Barrón-Cedeño, A., Wachsmuth, H., Petrov, R. & Nakov, P. “SemEval-2020 task 11: Detection of propaganda techniques in news articles”. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Barcelona, Spain. 2020. p. 1377–1414, https://www.scopus.com/pages/publications/85123925202. DOI: https://doi.org/10.18653/v1/2020.semeval-1.186.

9. Zhang, D., Li, Z.-Z., Zhang, M. L., et al. “From System 1 to System 2: A survey of reasoning Large Language Models”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2025; 48 (3): 3335–3354, https://www.scopus.com/pages/publications/105023055770.
DOI: https://doi.org/10.1109/TPAMI.2025.3637037.

10. McGinness, L. & Baumgartner, P. “Large Language Models imitate logical reasoning, but at what cost?”. *arXiv preprint*. 2025, https://www.scopus.com/pages/publications/105023512352. DOI: https://doi.org/10.48550/arXiv.2509.12645.

11. LekshmiAmmal, H. R. & Madasamy, A. K. “A reasoning based explainable multimodal fake news detection for low resource language using large language models and transformers”. *Journal of Big Data*. 2025: 12: 46, https://www.scopus.com/pages/publications/85218491447.
DOI: https://doi.org/10.1186/s40537-025-01093-x.

12. Hu, Y. & Tian, X. “Evaluating reasoning large language models on rumor generation, detection, and debunking tasks”. *iScience*. 2025; 28: 113690, https://www.scopus.com/pages/publications/105018720815. DOI: https://doi.org/10.1016/j.isci.2025.113690.

13. Boiko, O. A. “Modern AI methods for detecting propaganda in text”. *Registration, Storage and Processing of Data*. 2025; 27 (1): 120–131. DOI: https://doi.org/10.35681/1560-9189.2025.27.1.336147.

14. “Gemini thinking”. *Gemini API Documentation*. 2025. – Available from: https://ai.google.dev/gemini-api/docs/thinking. – [Accessed: Dec 2025].

15. “Reasoning Models”. *OpenAI API Documentation.* – Available from: https://platform.openai.com/docs/guides/reasoning#how-reasoning-works. – [Accessed: Dec 2025].

16. Da San Martino, G., Barrón-Cedeño, A. & Nakov, P. “PTC tasks on 'Detection of Propaganda Techniques in News Articles'”. *Propaganda Techniques Corpus (PTC)*. – Available from: https://propaganda.math.unipd.it/ptc/leaderboard.php. – [Accessed: Nov 2025].

17. Chen, X., Xu, J., Liang, T., et al. “Do NOT Think That Much for 2+3=? On the Overthinking of o1-Like LLMs”. *arXiv preprint.* 2024. DOI: https://doi.org/10.48550/arXiv.2412.21187.

18. Snell, C., Lee, J., Xu, K. & Kumar, A. “Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters”. *arXiv preprint*. 2024, https://www.scopus.com/pages/publications/105010225173. DOI: https://doi.org/10.48550/arXiv.2408.03314.

19. Xu, F., Hao, Q., Shao, C., at al. “Toward large reasoning models: A survey of reinforced reasoning with Large Language Models”. *Patterns*. 2025, https://www.scopus.com/pages/publications/105018176418. DOI: https://doi.org/10.1016/j.patter.2025.101370.

20. Liu, R., Geng, J., Wu, A. J., Sucholutsky, I., Lombrozo, T. & Griffiths, T. L. “Mind your step (by Step): Chain-of-Thought can reduce performance on tasks where thinking makes humans worse”. *arXiv preprint.* 2024, https://www.scopus.com/pages/publications/105023636323.
DOI: https://doi.org/10.48550/arXiv.2410.21333.

21. Lande, D., Alekseichuk, L., Svoboda, I. & Strashnoy, L. “Methodology of a swarm of virtual experts for evaluating the weight of connections in networks”. *Theoretical and Applied Cybersecurity*. 2024; 6 (2): 25–33. DOI: https://doi.org/10.20535/tacs.2664-29132024.2.319946.

22. Danylov, V. Ya., Huskova, V. H., Bidyuk, P. I. & Jirov, O. L. “Decision support system for forecasting financial processes on the basis of system analysis principles”. *System Research and Information Technologies*. 2019; 1: 20–36, https://www.scopus.com/pages/publications/85218127859. DOI: https://doi.org/10.20535/SRIT.2308-8893.2019.1.02.

23. Alghamdi, A. D. “MultiAgent-CoT: A Multi-agent chain-of-thought reasoning model for robust multimodal dialogue understanding”. *Comput. Mater. Contin.* 2026; 86 (2), https://www.scopus.com/pages/publications/105024546797. DOI: https://doi.org/10.32604/cmc.2025.071210.

24. Lytvynenko, V., Antoshchuk, S., Manzhula, V., Voronenko, M., Lurie, I. & Mrykhin, A. “Probabilistic modeling of delivery performance in uncertain supply chains using the Monte Carlo method”. *15th International Conference on Advanced Computer Information Technologies (ACIT)*. 2025. p. 1–6, https://www.scopus.com/pages/publications/105019950156.
DOI: https://doi.org/10.1109/ACIT65614.2025.11185841.

25. Han, S., Zhang, Q., Yao, Y., Jin, W., Xu, Z. & He, C. “LLM Multi-Agent Systems: Challenges and Open Problems”. *arXiv preprint.* 2024. DOI: https://doi.org/10.48550/arXiv.2402.03578.

Conflicts of Interest: The authors declare that they have no conflict of interest regarding this study, including financial, personal, authorship or other, which could influence the research and its results presented in this article

Received 23.12.2025
Received after revision 11.02.2026
Accepted 19.02.2026

DOI: https://doi.org/10.15276/aait.09.2026.05
УДК 004.89, 004.912

Емпірична оцінка моделей з функцією міркування для класифікації технік інформаційних маніпуляцій

Бойко Олег Анатолійович[1)]
ORCID: https://orcid.org/0009-0002-3424-8234; o.a.boiko@kpi.ua

Данилов Валерій Якович[1)]
ORCID: https://orcid.org/0009-0000-0875-4868; danilov1950@ukr.net. Scopus Author ID: 7201827051
[1)] Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», пр. Берестейський, 37. Київ, 03056, Україна

АНОТАЦІЯ

Швидка еволюція сучасних геополітичних конфліктів перетворила техніки інформаційних маніпуляцій та пропаганди з інструменту переконання на складну зброю масового впливу. Оскільки ці операції стають дедалі складнішими, часто покладаючись на витончені психологічні тактики, замість відвертої неправди, розробка передових автоматизованих механізмів виявлення таких маніпуляцій стає критичним безпековим викликом. Це дослідження має на меті подолати розрив між теоретичними можливостями та практичним застосуванням шляхом емпіричної оцінки ефективності новітніх моделей з функцією міркування (Reasoning Models) у конкретній задачі класифікації методів пропаганди. Основна мета полягає в порівняльному аналізі цих генеративних архітектур з історичними базовими моделями керованого навчання, щоб визначити, чи надають їхні внутрішні можливості побудови «ланцюжка міркувань» (chain-of-thought) відчутну перевагу над традиційними підходами розпізнавання шаблонів при ідентифікації складних риторичних стратегій. **Методологія** дослідження використовує стандартизований міжнародний набір даних для виявлення пропаганди (SemEval-2020 Task 11) для проведення порівняльного аналізу передових моделей без специфічного для задачі донавчання. У дослідженні застосовано стратегію, що базується виключно на інференсі, інтегруючи рольові ігри, вбудовані визначення та структуровані інструкції з міркування для імітації експертного аналізу. Ключовим методологічним внеском є систематичне варіювання розподілу бюджету на міркування під час інференсу для вимірювання кореляції між обчислювальними ресурсами та точністю класифікації. **Дослідження виявляє** суттєву «семантичну перевагу», завдяки якій моделі з міркуванням значно перевершують попередні системи керованого навчання у виявленні нюансованих технік, що спираються на культурний контекст, емоційне навантаження та непряму логіку. Проте результати також виявляють певні критичні обмеження, коли збільшення зусиль на міркування може погіршувати результативність у структурно простих завданнях, підтверджуючи існування феномену «надмірного розмірковування» (overthinking) в автоматизованій класифікації. Аналіз також виявляє нелінійний зв'язок між обчислювальними витратами та ефективністю, вказуючи на те, що монолітні моделі з міркуванням нерідко видають меншу ефективність порівняно з легковаговими архітектурами при обробці великих обсягів даних. У статті робиться **висновок**, що хоча моделі з міркуванням і представляють собою зміну парадигми в семантичному розумінні, вони поки що не є універсальним рішенням для всіх типів інформаційних маніпуляцій через структурні «сліпі зони» та економічну неефективність. Дослідження пропонує перехід від поодиноких великих моделей до багатоагентних систем. Цей запропонований підхід виступає за адаптивну систему, яка розподіляє спеціалізовані завдання між командою віртуальних експертів, балансуючи точність з операційною доцільністю для захисту інформаційного простору.

Ключові слова: виявлення пропаганди; штучний інтелект; мовні моделі з функцією міркування; мовні моделі з логічним міркуванням; великі мовні моделі; інформаційна безпека; когнітивна війна

ABOUT THE AUTHORS

Oleg A. Boiko - PhD Student, Department of Artificial Intelligence, Educational and Research Institute for Applied System Analysis. National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute". 37, Beresteiskyi Ave. Kyiv, 03056, Ukraine
ORCID: https://orcid.org/0009-0002-3424-8234; email: o.a.boiko@kpi.ua
***Research field*:** Artificial Intelligence

Бойко Олег Анатолійович - студент аспірантури кафедри Штучного інтелекту навчально-наукового інституту прикладного системного аналізу. Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», пр. Берестейський, 37. Київ, 03056, Україна

Valeriy Ya. Danylov - Doctor of Engineering Sciences, Professor, Laureate of the Borys Paton National Prize of Ukraine, Professor, Department of Artificial Intelligence, Educational and Research Institute for Applied System Analysis. National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute". 37, Beresteiskyi Ave. Kyiv, 03056, Ukraine
ORCID: https://orcid.org/0009-0000-0875-4868; danilov1950@ukr.net. Scopus Author ID: 7201827051
***Research field*:** System Analysis; Neural Networks; Deep Learning in Medicine and Socio-Economics; Hydroacoustic Signal Processing

Данилов Валерій Якович - доктор технічних наук, професор, лауреат Національної премії України імені Бориса Патона, професор кафедри Штучного інтелекту навчально-наукового інституту прикладного системного аналізу. Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», пр. Берестейський, 37. Київ, 03056, Україна