Evaluating the in-context learning capabilities of large language models for misinformation detection for Ukrainian news

Victoria Vysotska[1,†], Anastasiya Chupryna[2,†], Natalia Valenda[2†], Oleksandr Konduforov[2,*,†]

[1] *Kharkiv National University of Internal Affairs, L. Landau Avenue 27 61080 Kharkiv, Ukraine*
[2] *Kharkiv National University of Radio Electronics, Nauky ave. 24, 61166 Kharkiv, Ukraine*

Abstract
The rapid growth of online misinformation poses serious risks to democratic societies and media trust, particularly in the context of Ukraine, where the ongoing Russian invasion has intensified the spread of disinformation. While large language models (LLMs) such as GPT-4 and Claude have shown impressive results in various natural language processing tasks, including misinformation detection in English, their effectiveness for low-resource languages like Ukrainian remains underexplored. In this study, we investigate the performance of modern LLMs from OpenAI, Google, Anthropic, and DeepSeek on the task of misinformation detection in Ukrainian news, using zero-shot and few-shot in-context learning (ICL) strategies. We also examine the impact of various prompting techniques, including direct classification and chain-of-thought reasoning, as well as the use of the Ukrainian language for prompting.
To facilitate evaluation, we curated and cleaned a subset of Ukrainian misinformation data, building upon publicly available Ukrainian News dataset. Our experiments show that recent multilingual LLMs demonstrate strong comprehension and classification abilities for Ukrainian texts, achieving high F1-scores even without fine-tuning. Importantly, we observe consistent performance when using either English or Ukrainian prompts, underscoring the cross-linguistic robustness of these models. However, the lack of large, high-quality Ukrainian misinformation benchmarks remains a major bottleneck for further progress.
Our findings highlight both the potential and limitations of LLMs for misinformation detection in Ukrainian. While current ICL-based methods are promising, especially in low-resource settings, further gains are expected through supervised fine-tuning and improved data resources. This research lays the groundwork for future endeavors in developing more accurate and context-aware misinformation detection tools for the Ukrainian media.

Keywords
Misinformation detection, disinformation, fake news, text classification, Machine Learning, Large Language Models, in-context learning.

1. Introduction

Misinformation has become a significant issue worldwide. It is spreading rapidly online, weakening people's trust in the media and harming democratic systems. Different political and social groups employed disinformation techniques to spread fake information during the Brexit referendum, the 2016 US elections, and the 2024 Romanian elections. In particular, this issue is relevant in Ukraine, where the Russian invasion has led to an increase in misinformation. This misinformation is mostly driven by Russian state propaganda and fake or misleading news stories in the media. Because these threats are constantly evolving, it is important to employ robust systems to detect and counter them in media space, including the Ukrainian media.

[*] *CIAW-2025: Computational Intelligence Application Workshop, September 26-27, 2025, Lviv, Ukraine*
[*] Corresponding author.
[†] These authors contributed equally.
victoria.a.vysotska@lpnu.ua (V. Vysotska); anastasiya.chupryna@nure.ua (A. Chupryna); natalia.valenda@nure.ua (N. Valenda); oleksandr.konduforov@nure.ua (O. Konduforov)
0000-0001-6417-3689 (V.Vysotska); 0000-0003-0394-9900 (A.Chupryna); 0000-0003-3250-6172 (N.Valenda); 0009-0000-3942-2444 (O. Konduforov)

The task of automated misinformation detection has undergone significant evolution over the past two decades. Early approaches to text classification relied on classic machine learning (ML) algorithms such as Naïve Bayes, Support Vector Machines (SVM), or decision trees, which operated on manually engineered features, including TF-IDF or n-grams. These classic ML algorithms relied heavily on the NLP-based preprocessing steps [1]. With the advent of deep learning (DL), models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) introduced the ability to learn abstract features from raw text, leading to improved performance and robustness. However, a true breakthrough came with the introduction of transformer-based architectures, notably BERT [2], which enabled deep contextual understanding of language and became the foundation for a new wave of state-of-the-art models in virtually all NLP tasks.

Building upon the transformer architecture, large generative language models (LLMs) such as GPT-3 [3] and GPT-4 have further extended the capabilities of NLP systems by enabling general-purpose reasoning, in-context learning, and zero-shot or few-shot task adaptation. Compared to encoder-only models like BERT, LLMs demonstrate significantly higher performance in tasks that require complex reasoning, discourse understanding, or multi-step inference. They also demonstrated good performance in text classification tasks, as was outlined in different studies (e.g., Wang et al., 2024 [4]).

In the context of misinformation detection, LLMs have been explored under various in-context learning (ICL) approaches and prompting paradigms. Zero-shot prompting involves asking the model to classify a news item without any examples; few-shot prompting includes a small number of labeled examples in the prompt to guide the model; fine-tuning involves updating the model's weights using labeled misinformation datasets. Research in this area shows a lot of experimentation and progress, and results grow with the increase in the LLMs' performance. At the same time, many studies demonstrated that fine-tuned BERT-based models often outperform LLMs in the misinformation detection task, like Raza et al., 2025 [5]. However, LLMs – especially when instruction-tuned – excel in flexibility and interpretability, demonstrating robustness in adversarial and noisy scenarios.

While LLMs have shown promise in detecting misinformation in English, their effectiveness in Ukrainian contexts is less clear. Applying these advances to Ukrainian news and social media posts presents unique challenges. English-language datasets and benchmarks for fake news detection are abundant (Mridha et al. provide a list of popular benchmark datasets [6]), whereas resources for Ukrainian remain limited. The zero-shot or few-shot capabilities of these models, where they perform tasks without task-specific training, are particularly interesting for low-resource languages, as they lack datasets that can be used for training or fine-tuning. At the same time, there is a gap in testing the performance of the recent LLMs with reasoning capabilities in the Ukrainian domain.

This study aims to systematically evaluate the capabilities of major LLMs for detecting misinformation in Ukrainian news. Specifically, we investigate zero-shot, few-shot, and reasoning-based prompting strategies, analyzing how effectively these models generalize to Ukrainian content without fine-tuning. By presenting a rigorous comparison, we seek to identify strengths, limitations, and future directions for employing LLMs in Ukrainian-language misinformation detection, with a focus on practical applicability in low-resource settings.

2. Related Work

2.1. LLMs for Misinformation Detection

The rapid rise of large language models (LLMs) has driven notable interest among researchers for their application to text classification and misinformation detection. Unlike traditional deep learning approaches and bidirectional transformer models, LLMs can analyze text in depth, cross-reference facts, and even perform basic fact-checking by detecting contradictions or checking claims against their vast pre-trained knowledge. LLMs can be used without extensive task-specific training, instead leveraging prompting (in zero-shot or few-shot modes) to adapt to the misinformation detection task, which is attractive when labeled data is scarce. Wang et al. explored the usage of LLMs in different modes (zero-shot, few-shot and fine-tuned) as text classifiers and demonstrated their good performance on various datasets, which confirmed that LLMs can be effective classifiers [4].

In the context of the misinformation detection task, Xu et al. [7] evaluated major LLM models. The findings highlighted that while LLMs exhibit promising capabilities in detecting misinformation (F1-score for some models may achieve 91%), their performance varies across different models and domains.

In addition to text classification, LLMs can reason in natural language, which opens the door to explainable outputs (e.g., generating a rationale for why a news piece is predicted fake) and interactive fact-checking agents. Research by Pendyala et al. explored the capability of LLMs not just to predict true/fake, but also to explain their decisions using chain-of-thought prompts or to highlight evidence, and confirmed that LLM-based detectors could potentially provide richer insights than a traditional black-box classifier [8].

However, there are also clear risks in relying on LLMs for misinformation detection. By design, LLMs are trained to produce fluent and plausible text, which means they can *hallucinate* – i.e., generate information that sounds convincing but is false. An LLM might confidently assert a claim is true or false based on incorrect "knowledge" it learned, leading to false positives or negatives. Moreover, LLMs themselves can be double-edged: the same models can be misused by adversaries to create more convincing fake news. For instance, Jiang et al. underscore that generative models like ChatGPT have been used to produce highly persuasive but misleading content, raising the bar for automated detectors [9].

Another important question is whether LLM-based detectors actually outperform the previous generation of models (e.g. BERT-based classifiers) on misinformation tasks. Several recent studies have conducted side-by-side comparisons, concluding that while BERT-like models often achieve higher accuracy and efficiency on structured, pattern-driven tasks, LLMs consistently outperform them on tasks requiring deep world knowledge, complex reasoning, or interpretability. Moreover, prompt-engineering strategies – including zero-shot and few-shot prompting – enable LLMs to perform well with minimal or even no task-specific training, a flexibility that classic models cannot match. For example, Raza et al. reveals that BERT-like models generally outperform LLMs in classification tasks, while LLMs demonstrate superior robustness against text perturbations [5]. At the same time, a study by Pelrine et al. demonstrates that GPT-4 can outperform prior methods in multiple settings and languages [10].

LLMs can also be used to enhance smaller classifier models. An empirical study by Chen et al. revealed that well-prompted LLMs achieve comparable performance in text-based misinformation detection with SLMs, while LLM-enhanced detectors outperform plain SLMs in most cases when generating richer textual features, producing analysis, and simulating user engagements [11].

Another noteworthy line of research is how to craft prompts or intermediate steps that enable LLMs to make more accurate judgments about misinformation. One idea is to encourage *explicit reasoning*, often via *chain-of-thought (CoT) prompting*. In a chain-of-thought approach, instead of asking for a direct label, we prompt the LLM to first "think through" the content: for example, list out the claims in the article, check each against known facts, and then conclude whether the article is fake or real. This technique has yielded improvements in some domains of question answering and reasoning. However, in the domain of misinformation detection, the evidence is mixed. Cao et al. used chain-of-thought prompting for Financial Misinformation Detection and demonstrated that it works better than a zero-shot approach for a number of different commercial LLMs [13]. Hu et al. also noted that CoT produces better results than plain prompting, but still didn't help ChatGPT catch misinformation better than plain SLMs like BERT [14]. One possible reason is that the LLM's "free-form" reasoning can introduce distractions or inaccuracies. Thus, while CoT prompting is a promising tool, its effectiveness may depend on the LLM's factual correctness and the design of the prompt, and further research is needed to determine if guided reasoning or structured prompts can reliably improve detection accuracy.

Another interesting approach is the Hierarchical Step-by-Step (HiSS) prompting method by Zhang et al. [15] It directs LLMs to separate a claim into several subclaims and then verify each of them via multiple question-answering steps progressively. Experiment results on two public misinformation datasets show that HiSS prompting outperforms a state-of-the-art fully-supervised approach and strong few-shot ICL-enabled baselines. This technique addresses two main challenges in news claim verification: omission of necessary details and fact hallucination.

2.2. Misinformation Detection for Ukrainian language: Datasets and Approaches

Most misinformation detection research is focused on English, but applying these approaches to Ukrainian presents some challenges. While English-language datasets and benchmarks for fake news detection are abundant (e.g., LIAR [16], FakeNewsNet [17]), resources for Ukrainian remain limited. A few notable datasets exist, such as the multilingual *EUvsDisinfo* corpus [18], the *Mantis Analytics* dataset [19], the *UNLP 2025 Shared Task on Detecting Social Media Manipulation* [20] and the *Ukrainian News* dataset [21], which targets the Russo-Ukrainian war discourse. These efforts are beginning to close the resource gap, but the scale and diversity of data in Ukrainian are still far behind those in English. Consequently, some researchers resort to cross-lingual transfer, adapting English models to Ukrainian tasks – an imperfect strategy that can introduce bias or produce misaligned representations, or opt for a more complex approach of crafting their own dataset.

For example, Dementieva et al. translated English datasets into Ukrainian using cross-lingual knowledge transfer methods, thereby avoiding manual data curation, including large multilingual encoders and translation systems, LLMs, and language adapters [22]. They then tested XLM-RoBERTa on three text classification tasks: toxicity classification, formality classification, and natural language inference. Study from Vysotska et al. details the creation of a custom dataset for Ukrainian fake news detection based on a set of NLP techniques, providing a rare example of localized annotation [23].

Another notable effort to address the issue of the limited representation of low-resource languages like Ukrainian was done by Kiulian et al. [24], who tried to fine-tune the open-source Gemma and Mistral LLMs with Ukrainian datasets, aiming to improve their linguistic proficiency and benchmarking them against other existing models capable of processing the Ukrainian language. Additionally, they presented the Ukrainian Knowledge and Instruction Dataset (UKID) to aid future efforts in language model fine-tuning.

In terms of methods for Ukrainian misinformation detection, the progression is similar to the general evolution from traditional ML to LLMs, with additional constraints. Early attempts relied on translation or language-agnostic features. For instance, before Ukrainian training data was available, one might translate Ukrainian articles to English and then apply an English fake news detector – a cumbersome but sometimes effective workaround. With the advent of multilingual transformers, researchers moved to models like mBERT or XLM-RoBERTa [25], which are pre-trained on dozens of languages, including Ukrainian. Fine-tuning such models on any available Ukrainian data has shown decent success. For example, Bazdyrev et al. (2025) employed a multilingual RoBERTa model adapted to the task and demonstrated reasonable accuracy in detecting certain propaganda messages, illustrating that transformer-based classifiers can be effective across languages when properly adapted [26].

In another study, Shupta et al. (2024) proposed an approach that utilizes different types of embeddings, including an LLM-based one, for misinformation classification and tested it on both English and Ukrainian fake news datasets [27]. They found that the classification results for the Ukrainian dataset were about 2-3% worse than for the English one.

Despite these challenges, the research work is pushing forward. The Ukrainian community and researchers are beginning to assemble more misinformation detection datasets in Ukrainian language. The good examples here are the Kaggle-based "Disinformation Detection Challenge" and the UNLP 2025 Shared Task on Detecting Social Media Manipulation mentioned above, which provided their own datasets and brought attention to the domain for many research teams to develop their custom models.

In summary, applying LLMs and other detectors to Ukrainian is a promising but still developing frontier – initial results show approaches similar to those in English can work, but they must be adapted to the linguistic and data constraints, and there is a need for more Ukrainian-specific resources to speed up this work.

3. Methods and Materials

3.1. Benchmark datasets

Effective misinformation detection using LLMs critically depends on the availability of high-quality benchmark datasets. As it was mentioned in the previous section, while the English language benefits from a range of well-established datasets, which offer both diverse content and detailed annotations, the situation for Ukrainian-language resources remains more challenging. There are not many Ukrainian datasets on this topic, and the available ones are often small in size, exhibit thematic bias, and in some cases rely on automatic or source-based labeling strategies rather than human-verified fact-checking. As a result, no benchmark has been established yet, making it difficult to compare different research efforts with each other.

For example, EUvsDisinfo [18] mentioned above provides some support for Ukrainian through its annotations of pro-Kremlin disinformation. However, the Ukrainian subset of this dataset is notably small and unevenly distributed across topics, confirming a broader issue in sourcing reliable benchmarks for low-resource languages like Ukrainian. The *Mantis Analytics dataset* [19], introduced in the Kaggle-based "Disinformation Detection Challenge" is more focused on content relevant to the war, often collected from suspicious or fringe Telegram channels, thus reflecting realistic misinformation scenarios. The *UNLP 2025 Shared Task on Detecting Social Media Manipulation* provides a great humanly annotated dataset for manipulation technique classification and span identification [20], which is slightly different type of the problem than misclassification detection.

Another example of the Telegram-based datasets is *Ukrainian News* [21], which contains 10700 news from 2022. Though the news were collected from the start of the full-scale Russian-Ukrainian war, they cover not only war-related topics but also politics, culture and some other topics which makes it a valuable data source.

The lack of datasets in Ukrainian language motivate some authors to create their own datasets to get a well-annotated set of data for training and testing. We decided to use this approach for our study.

3.2. LLMs for misinformation detection

LLMs can be applied to the task of misinformation detection as text classifiers, typically framed as a binary classification problem: the model determines whether a given news article is "true" or "fake". This can be achieved through various prompting techniques, even without fine-tuning. There are several choices of LLMs available currently from different companies, including OpenAI (GPT family), Google (Gemini), Anthropic (Claude), xAI (Grok), Meta (Llama), Mistral, Alibaba (Qwen), DeepSeek, and others.

Each of these companies produces models of varying sizes, suitable for different use cases. These models also vary in the following attributes: speed of inference, multimodality, context window, function calling, structured outputs, knowledge cutoff date, web search support, reasoning support, fine-tuning support, price for input and output tokens and many others. Therefore, when using them for our task, we need to try different models, as they can provide varying performance in misinformation detection.

Recently developed reasoning-optimized models, such as GPT-o3, Claude Sonnet 3.7, or Gemini 2.5 Pro, have shown promise in handling different tasks more effectively. These models are architecturally designed to support longer context, perform intermediate reasoning, and adhere more strictly to prompt constraints, making them well-suited for misinformation detection in complex or low-resource linguistic settings.

One compelling aspect of using LLMs is the variety of *in-context learning (ICL) strategies* available. Approaches to employing LLMs for misinformation detection fall broadly into zero-shot prompting, few-shot prompting, and fine-tuning.

In zero-shot learning, an LLM is prompted with only the task instructions (e.g., "Is this article credible?") and the input text, relying entirely on its pretraining to make a decision. Zero-shot prompting enables immediate deployment across domains, though models typically lag behind in overall accuracy. This method is particularly attractive for low-resource settings such as Ukrainian, where labeled examples are scarce, but may struggle with nuanced or domain-specific content.

Few-shot learning improves upon zero-shot by providing the model with a handful of labeled examples directly in the prompt, offering implicit task guidance and contextual anchoring. While still non-parametric, few-shot prompts can improve performance on low-resource tasks through implicit adaptation.

Fine-tuning, though more resource-intensive, involves training the LLM on a labeled dataset, allowing it to adapt specifically to the task and language. This approach usually provides the best performance of the model. Fine-tuning requires having a special API provided by the company (e.g. GPT-3.5 or Mistral 7B) or open-sourced model weights (e.g., Llama or DeepSeek).

3.3. Prompt engineering approaches

Prompt engineering plays a critical role across all approaches. Basic prompting strategies include direct task framing (e.g., "Classify this text as misinformation or not") with definition support and instruction-based formulations. This approach can serve as a baseline for further improvements and complications.

More advanced techniques include *Chain-of-Thought (CoT)* prompting, where the model is encouraged to articulate reasoning steps before delivering a final verdict. In CoT, instead of asking for a direct label, we prompt the LLM to first "think through" the content: for example, list out the claims in the article, check each against known facts, then conclude if the article is fake or true. This technique has yielded improvements in some domains of question answering and reasoning. It can be especially valuable in misinformation tasks, as fake news often exploits subtle contextual cues that require multi-step reasoning.

Other interesting approaches include *self-consistency decoding*, which aggregates multiple generations to improve robustness, and *retrieval-augmented generation (RAG)*, which supports evidence-grounded classification. In RAG, the relevant external evidence (e.g., fact-checked articles) is retrieved alongside the input, where possible, enabling the LLM to ground its assessments.

4. Experimental Setup

To explore the capabilities of large language models (LLMs) in detecting misinformation in the Ukrainian language, our experimental framework uses a carefully annotated benchmark dataset, a spectrum of LLMs, and several prompting techniques.

4.1. Benchmark dataset

We based our evaluation on the *Ukrainian News* dataset [20] mentioned above – a collection of messages from Ukrainian Telegram channels during the Russian-Ukrainian War. This dataset is particularly valuable due to Telegram's role as a major platform for both verified news and disinformation during wartime, making it a relevant source for real-world misinformation detection.

However, initial exploration of the dataset revealed several critical issues. The dataset contains duplicates, announcements, or posts lacking substantial factual content, making them unsuitable for evaluating a model's ability to distinguish truth from misinformation. Furthermore, label reliability was problematic: numerous "*fake*" and some "*true*" samples were mislabeled, most likely due to the source-based labeling approach rather than human-verified fact-checking.

To address these limitations, we randomly selected a significant subset of messages for manual curation. Each message was carefully reviewed to assess whether it contained a factual claim that could be evaluated as true or false. Messages that were non-factual, badly written, or too short for evaluation were removed. The remaining data was reannotated. After that, the final dataset of 400 samples was randomly selected in a balanced way: 200 "*true*" and 200 "*false*" samples.

4.2. LLM models

For this experiment, we selected a set of LLM models of different sizes, novelty and capabilities, provided by OpenAI, Google, Anthropic and DeepSeek. The list of models with their main characteristics is provided in Table 1.

Table 1
List of models used in the research

Vendor	Non-reasoning models	Reasoning models
OpenAI	GPT-3.5, GPT-4o	GPT-o3, GPT-o3-mini, GPT-o4-mini
Google	Gemini Flash 2.0	Gemini Flash 2.5, Gemini Pro 2.5
Anthropic		Claude Haiku 3.5, Sonnet 3.7, Sonnet 4
DeepSeek	DeepSeek-V3	DeepSeek-R1

This selection of models allowed us to examine how factors such as model scale, novelty (knowledge cutoff), reasoning capacity, and architecture impact performance on misinformation detection in a low-resource language setting. In particular, we were interested in how reasoning-augmented models (e.g., GPT-o3, Gemini Pro 2.5, Claude Sonnet 4 and DeepSeek-R1) perform relative to lightweight or instruction-tuned variants (e.g., GPT-3.5, Gemini Flash 2.0, DeepSeek-V3). GPT-3.5 model is used as a baseline model which is expected to perform poorly compared to newer and bigger models with more recent knowledge cutoffs.

4.3. Prompting techniques

To test the models' reasoning processes and understand how instruction phrasing and language affect performance, we implemented four prompting techniques:

1. **Basic (no analysis), English:** A simple, direct prompt in English asking whether the given Ukrainian-language text is true or fake, with no additional context or rationale required.
2. **Chain-of-thought (analysis), English:** An English-language prompt instructing the model to provide a step-by-step explanation of its thinking before delivering its final judgment.
3. **Chain-of-thought (analysis), Ukrainian:** The same analytical prompt translated into Ukrainian, intended to test whether linguistic alignment between input and instructions improves model reasoning.

4. **Few-shot chain-of-thought (analysis), Ukrainian:** A context-rich Ukrainian prompt that includes multiple labeled examples of true and fake samples, enabling in-context learning through few-shot prompting.

Each model was tested with all four prompting techniques to enable cross-comparison. Table 2 provides the texts of each prompt.

Table 2
Experiment prompts

Prompt type / Language	Prompt
Basic (no analysis) / English	Your task is to analyze the text and determine if it contains true or false information. Verdict should be "true", "probably true", "probably false" or "false", don't add any other text on top of it. If you are not sure about the verdict, do your best to suggest "probably true" or "probably false" verdicts. Use the following structure for your answer: "Verdict: <verdict>" Text is provided below:
Chain-of-thought (analysis) / English	Check the facts provided in the piece of news below, analyze it from different perspectives and give your judgment whether it is true or fake at the end. Verdict should be "true", "probably true", "probably false" or "false", don't add any other text on top of it. If you are not sure about the verdict, do your best to suggest "probably true" or "probably false" verdicts. Use the following structure for your answer: "Analysis: <put analysis here> Verdict: <put verdict here>" Text is provided below:
Chain-of-thought (analysis) / Ukrainian	Перевір факти, наведені в тексті новини нижче, проаналізуй їх з різних перспектив та надай свій вердикт, чи цей текст правдивий чи брехливий. Вердикт повинен бути "правда", "скоріше правда", "скоріше брехня" або "брехня", не додавай ніякого іншого тексту. Якщо ти невпевнений на рахунок вердикту, використовуй варіанти "скоріше правда" або "скоріше брехня". Використовуй наступну структуру для своєї відповіді: "Аналіз: <додай аналіз сюди> Вердикт: <додай вердикт сюди>" Текст для аналізу надано нижче:
Few-shot chain-of-thought (analysis) / Ukrainian	Перевір факти, наведені в тексті новини нижче, проаналізуй їх з різних перспектив та надай свій вердикт, чи цей текст правдивий чи брехливий. Вердикт повинен бути "правда", "скоріше правда", "скоріше брехня" або "брехня", не додавай ніякого іншого тексту. Якщо ти невпевнений на рахунок вердикту, використовуй варіанти

"скоріше правда" або "скоріше брехня".
Використовуй наступну структуру для своєї відповіді:
"Аналіз: <додай аналіз сюди>
Вердикт: <додай вердикт сюди>"

Приклад 1:
Текст: <текст1>
Аналіз: <аналіз1>
Вердикт: правда

4.4. Experiment Protocol

The evaluation followed a consistent protocol across all models and prompting strategies. Each sample from the curated benchmark dataset was input to the model through its respective API using a standardized prompt template. To ensure comparability, the same prompt structure and parameters were used across models for each strategy.

For chain-of-thought prompts, both the reasoning and final answer were logged. Model outputs were post-processed to standardize label formats (e.g., *"True"*, *"Probably True"*, *"Probably False"*, *"False"*, etc.) for evaluation.

While conducting the experiments, evaluation metrics included:

- Accuracy: Overall correctness on the test set.
- Precision: Correctness of misinformation predictions that were truly fake.
- Recall: Fraction of actual misinformation correctly identified.
- F1-score: Harmonic mean of precision and recall.

This experimental protocol was designed to ensure fairness, repeatability, and depth in assessing how well LLMs can reason about misinformation in the Ukrainian language, especially when faced with real-world content from conflict-related media environments.

5. Results

In this section, we present the results of the assessment of how existing reasoning and non-reasoning LLMs perform on the task of misinformation detection for the Ukrainian language. We provide the classification metrics for a selected set of models and hand-crafted prompts in tables and heatmaps.

Table 3 presents a comprehensive comparison of 13 LLMs with four prompts with Precision, Recall and F1-score.

Table 3
Misinformation prediction results for all models and prompts

Prompt type	Basic, EN			COT, EN			COT, UA			Few-shot COT, UA		
Model	Pr	Rec	F1	Pr	Rec	F1	Pr	Rec	F1	Pr	Rec	F1
GPT 3.5	0.72	0.77	0.74	0.73	0.82	0.77	0.64	0.94	0.76	0.66	0.83	0.74
GPT 4o	0.93	0.65	0.77	0.85	0.82	**0.84**	0.91	0.79	0.85	0.9	0.81	0.85
GPT o3-mini	0.76	**0.83**	0.79	0.76	**0.94**	**0.84**	0.71	**0.93**	0.80	0.7	**0.94**	0.8
GPT o4-mini	0.73	0.77	0.75	0.68	0.83	0.75	0.67	0.83	0.74	0.73	0.87	0.8

GPT o3	0.94	0.78	**0.85**	0.89	0.78	0.83	**0.91**	0.78	0.84	0.94	0.85	**0.89**
Gemini Flash 2.0	0.96	0.5	0.66	0.95	0.64	0.77	0.9	0.62	0.73	0.92	0.67	0.78
Gemini Flash 2.5	0.95	0.55	0.7	0.91	0.68	0.78	0.89	0.59	0.71	0.96	0.65	0.77
Gemini Pro 2.5	0.9	0.65	0.76	0.89	0.64	0.74	0.88	0.67	0.76	0.92	0.66	0.77
Claude Haiku 3.5	0.88	0.52	0.65	0.87	0.62	0.73	0.81	0.75	0.78	0.81	0.82	0.82
Claude Sonnet 3.7	0.88	0.65	0.75	0.87	0.78	0.82	0.84	0.88	**0.86**	0.88	0.77	0.82
Claude Sonnet 4	0.85	0.68	0.76	0.81	0.7	0.75	0.81	0.66	0.73	0.85	0.64	0.73
DeepSeek-V3	0.92	0.43	0.59	0.91	0.57	0.7	0.87	0.62	0.73	0.9	0.73	0.81
DeepSeek-R1	**1.0**	0.43	0.6	**0.96**	0.6	0.74	0.9	0.68	0.77	**1.0**	0.58	0.73

To facilitate comparison, we also provide a heatmap of F1-scores by model and prompt type.

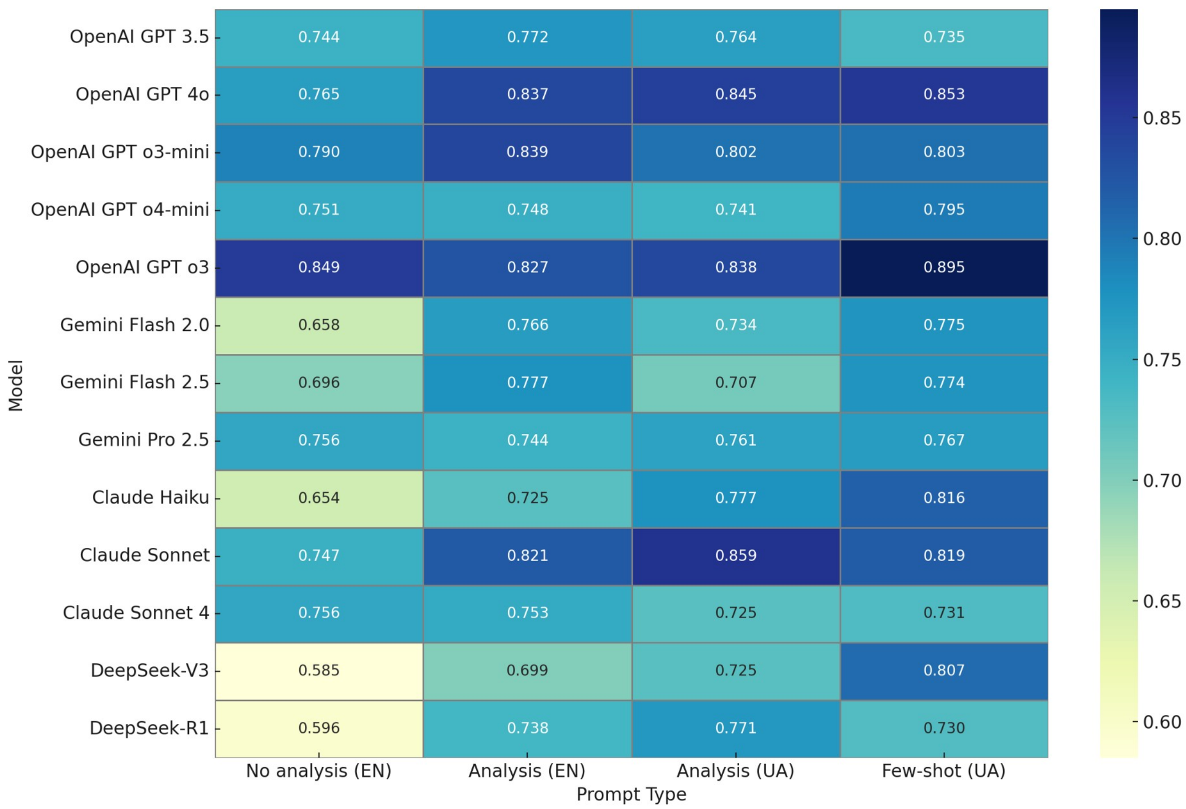

Figure 1: F1-score heatmap by model and prompt type.

We can extract several important observations from Table 3 and Figure 1:

1. **GPT o3** delivers the best performance, especially for prompts without analysis and a few-shot one.
2. Simpler models, such as **GPT o3-mini** or **GPT 4o** perform well with CoT prompts but struggle with direct model predictions without analysis.

3. **DeepSeek-R1** demonstrates the highest Precision (but poor Recall, very high bias), while **GPT o3-mini** demonstrates the highest Recall (and relatively good Precision).
4. Bigger models, such as **Gemini Pro 2.5** or **Claude Sonnet 3.7/4**, provide better results for a basic prompt compared to their smaller versions (**Gemini Flash 2.0/2.5** and **Claude Haiku 3.5**).
5. **DeepSeek** models perform the poorest with a basic prompt, they need analysis to compensate for their low Recall.

Example of the chain-of-thought output by GPT 3.5 and GPT o3 models for comparison (EN prompt):

Sample:

"День міста у Харкові почався з обстрілів Київського района. Є влучання у приватний будинок, без постраждалих, - повідомив мер Харкова Ігор Терехов. Він привітав харків'ян зі святом свого незламного міста і нагадав, що Україна обов'язково переможе у війні з російським ворогом."

GPT 3.5 output:

"Analysis: The news report claims that the Day of the City in Kharkiv started with shelling in the Kyivsky district, with hits on a private house but no casualties reported. The mayor of Kharkiv, Igor Terekhov, is said to have greeted the residents and expressed confidence in Ukraine's victory in the war against Russia.

Verdict: Probably False. This news report contains several inconsistencies and inaccuracies. Firstly, Kharkiv is a separate city from Kyiv, so it is unlikely that shelling in Kharkiv would be described as occurring in the Kyivsky district. Additionally, there is no verifiable information or reports from credible sources to confirm the alleged shelling incident at the start of the City Day celebrations in Kharkiv. Therefore, the claim is likely false."

GPT o3 output:

"Analysis:

1. Date of the event. Kharkiv officially celebrates its City Day on 23 August every year. Therefore, a report that starts with "День міста у Харкові" (City Day in Kharkiv) logically refers to 23 August.

2. Claim of shelling. Russian forces have repeatedly shelled Kharkiv since February 2022, and the city's Kyivskyi (Київський) district is often among the hit areas. On 23 August 2023, multiple Ukrainian news outlets (e.g., Suspilne, RBC-Ukraine, Ukrinform) quoted Kharkiv mayor Ihor Terekhov saying that the city was being shelled in the early morning, with at least one strike on a private house in the Kyivskyi district and no casualties reported. The wording in those reports is virtually identical to the wording in the text provided.

3. Source credibility. Ihor Terekhov maintains an official Telegram channel that Ukrainian media routinely cite for real-time information on shelling. His statements on 23 August 2023 match the claims in the text (strike on a private house, no casualties, morning of City Day).

4. No contradictory evidence. No reputable source has refuted Terekhov's report or offered an alternate version of events for that date.

Considering the alignment between the text and multiple independent, reputable Ukrainian news reports of the same morning, the information appears accurate.

Verdict: True"

In this example, the GPT 3.5 model demonstrates several mistakes that prevent it from reaching the correct verdict. First, its knowledge cutoff (Sep 2021) doesn't allow it to "*know*" the facts of shelling Kharkiv and even the Russian-Ukrainian war itself. Second, it connected the Kyivsky district to Kyiv city and didn't understand that it is a district in Kharkiv. And last but not least, it ignored the instruction not to add any additional information into the *Verdict* section and actually added a part of the analysis there.

At the same time, the GPT o3 model not only demonstrated the capabilities to split the sample claim into several parts and verify each of them, but also used its own "memory" of the facts to verify and weigh the claims.

In Table 4, we present the aggregated average Precision, Recall and F1-score for each LLM across all prompts.

Table 4
Aggregated classification metrics for each model

Model	Precision (avg)	Recall (avg)	F1-score (avg)
GPT 3.5	0.688	0.840	0.754
GPT 4o	0.898	0.768	0.825
GPT o3-mini	0.730	**0.910**	0.809
GPT o4-mini	0.703	0.825	0.759
GPT o3	0.920	0.794	**0.852**
Gemini Flash 2.0	0.934	0.608	0.733
Gemini Flash 2.5	0.926	0.617	0.739
Gemini Pro 3.5	0.898	0.655	0.757
Claude Haiku 3.5	0.844	0.676	0.743
Claude Sonnet 3.7	0.865	0.770	0.812
Claude Sonnet 4	0.831	0.670	0.741
DeepSeek-V3	0.899	0.588	0.704
DeepSeek-R1	**0.965**	0.569	0.709

Figure 2 provides a heatmap of Precision, Recall and F1-score by model.

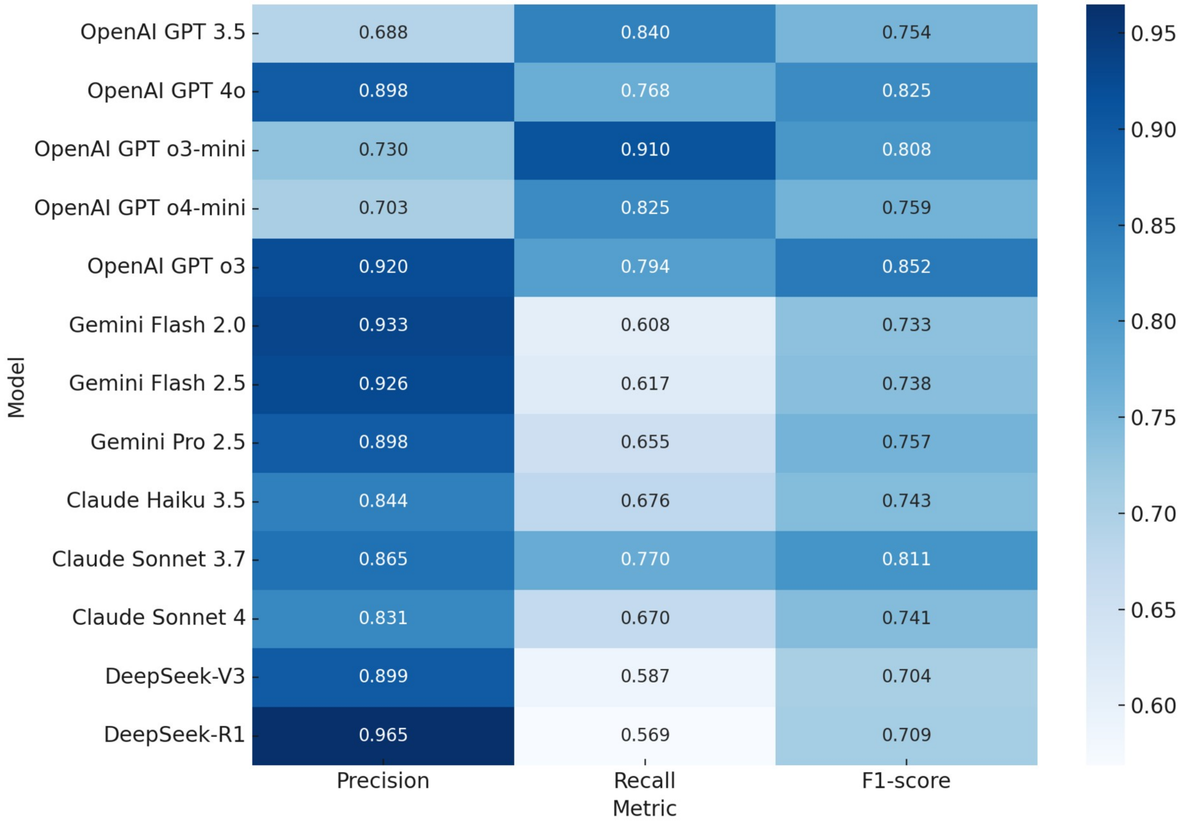

Figure 2: Heatmap for classification metrics by model.

Based on Table 4 and Figure 2, we can conclude several trends about the overall model performance:

1. **GPT o3** provides the best performance overall and across reasoning models. **GPT 4o** provides the best performance across non-reasoning models.
2. Models tend to have their own bias toward making *True* or *False* predictions. For example, **OpenAI** models have better Recall, while **Gemini**, **Claude** and **DeepSeek** models tend to have noticeable higher Precision, demonstrating higher bias. The only exceptions are **GPT 4o** and **GPT o3**, which provide both high Precision and Recall, making them the two best-performing models.
3. Newer versions of models don't mean better performance: for example, older version of Sonnet (**Sonnet 3.7**) performs better than a new one (**Sonnet 4.0**), and **GPT o3-mini** performs better than **GPT o4-mini.**

The last comparison, which is interesting for our analysis, is the comparison of prompting techniques. In Table 5, we present the aggregated average Precision, Recall and F1-score for each prompt type across all models.

Table 5
Aggregated classification metrics for each prompt technique

Prompt	Precision (avg)	Recall (avg)	F1-score (avg)
Basic (no analysis)	**0.878**	0.631	0.719
CoT (analysis), EN	0.852	0.724	0.773

CoT (analysis), UA	0.826	0.748	0.773
Few-shot CoT (analysis), UA	0.859	**0.755**	**0.792**

Here are the main observations from Table 5 about the overall prompt type performance:

1. A **basic prompt** without analysis provides the worst average performance (average F1-score) across the models, showing higher Precision, but significantly lower Recall.
2. **Reasoning prompts** (with analysis) offer a better Precision/Recall balance, resulting in a significantly higher average F1-score.
3. **Ukrainian vs. English prompts**: The analysis prompt in both Ukrainian and English languages yields similar average F1-scores, suggesting that the Ukrainian language is well-supported in the prompt.
4. The **few-shot prompt** outperforms the **zero-shot one** written in Ukrainian language by almost 2%.

6. Discussion

The application of LLMs for misinformation detection in Ukrainian news texts presents several unique challenges and promising opportunities. One of the central issues identified in this research is the scarcity of well-annotated, publicly available, and sufficiently large datasets for the Ukrainian language. While English-language datasets such as LIAR and FakeNewsNet are frequently used for training and benchmarking misinformation detection models, equivalent resources in Ukrainian remain limited in scope, quality, and popularity. Existing Ukrainian datasets often suffer from inconsistencies in labeling, unbalanced distributions, and insufficient sample sizes, making their usage difficult for reliable benchmarking or training. This gap underlines the need for future work on the creation, validation, and dissemination of high-quality Ukrainian misinformation datasets. Such datasets would not only facilitate more accurate evaluations but also help establish reproducible benchmarks for Ukrainian NLP research.

Despite this limitation, our experiments show that current LLMs are capable of processing Ukrainian texts with high competence. When tested using both English and Ukrainian prompts, most reviewed LLMs exhibited consistent performance, demonstrating similar quality when the prompt language was switched to Ukrainian. This suggests that these multilingual models have acquired strong comprehension capabilities across languages and can accurately follow instructions and reasoning chains in Ukrainian. Moreover, the models typically respond in the same language as the prompt, ensuring consistency in multilingual workflows. This finding is encouraging for low-resource language communities, indicating that even without fine-tuning, state-of-the-art LLMs can be leveraged effectively for Ukrainian NLP tasks, including misinformation detection.

Nevertheless, there are limitations to our study. Not all recent LLMs were included in the evaluation, and only a subset of prompting techniques was tested. The focus was primarily on zero-shot and few-shot scenarios, with an emphasis on direct classification prompts and chain-of-thought prompting. The results reveal that the best evaluated models achieved F1-scores in the range of 0.8 to 0.9, highlighting strong zero-shot performance across a variety of LLMs. Still, these results likely represent a lower bound on the models' true capabilities. Further performance gains are anticipated through more sophisticated prompting strategies, including multi-turn reasoning, contextual priming, or the use of external knowledge sources.

An important direction for future work is the use of fine-tuning. While our study did not include any fine-tuned models, prior research suggests that supervised fine-tuning on task-specific data can yield substantial improvements in classification accuracy, especially when combined with in-domain examples. Fine-tuning LLMs on Ukrainian misinformation data – once a reliable dataset is available – could help close the remaining performance gap and provide even more robust tools for misinformation mitigation.

Finally, beyond classification, our experiments highlight the potential of LLMs in supporting data annotation and fact-checking tasks. The explanation-oriented outputs generated by models such as GPT o3 demonstrate a level of reasoning that can be valuable for human annotators. These outputs often provide relevant contextual knowledge and reasoning paths that help justify predictions, making the annotation process faster and more informed. While these models are not yet perfectly reliable as autonomous misinformation detectors, their fact-checking capabilities can significantly assist journalists, researchers, and general users in evaluating the credibility of news claims. Another potential application of the reasoning output of the models could be in using them as additional input for the SLMs like BERT, which can improve the performance of such models further.

7. Conclusions

In this study, we explored the effectiveness of large language models (LLMs) for the task of misinformation detection in Ukrainian news. We examined the current research landscape with a focus on in-context learning (ICL) strategies, including zero-shot and few-shot prompting, as well as the impact of different prompting strategies, ranging from direct prediction requests to more elaborate chain-of-thought analyses. Our analysis highlights that modern reasoning LLMs, when equipped with appropriate instructions, can achieve significant performance on this task even without task-specific training.

We also assessed the current state of misinformation detection, specifically in the context of Ukrainian texts. Despite recent progress, there remains a significant lack of large-scale, high-quality, publicly available datasets for Ukrainian-language misinformation detection. Existing datasets are often noisy, limited in size, or not annotated in a way that supports robust benchmarking. This gap hinders the development of accurate models and comparative evaluations, underlining a clear need for the creation and publication of better Ukrainian-language datasets in this domain.

To understand how current LLMs perform on this task, we evaluated a diverse set of modern models, including those developed by OpenAI, Google, Anthropic, and DeepSeek. These models were tested under various prompting strategies in both English and Ukrainian, revealing that language choice in prompts does not substantially affect model performance. The results also confirmed that many state-of-the-art models, particularly when using analytical prompts, achieve strong F1-scores (up to 0.895), demonstrating a high level of reasoning and instruction-following ability, even in a low-resource language context.

Nevertheless, our findings also suggest that zero- and few-shot settings may not fully exploit the potential of these models, especially when high precision is required for downstream applications such as fact-checking or content moderation. As a result, future work will explore the fine-tuning of LLMs and smaller language models (SLMs), such as BERT-based architectures, to further enhance accuracy and robustness. This next phase of research will aim to better align model outputs with the nuances of Ukrainian-language misinformation and contribute more reliable tools to the information verification ecosystem.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly and ChatGPT in order to: grammar and spelling check, stylistic improvements to make text more formal, generate heatmaps for figures 1 and 2. After using these tools, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] K. Smelyakov, D. Karachevtsev, D. Kulemza, Y. Samoilenko, O. Patlan, A. Chupryna, Effectiveness of Preprocessing Algorithms for Natural Language Processing Applications, in: 2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T), IEEE, 2020, pp. 187–191. doi:10.1109/picst51311.2020.9467919.

[2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language Models are Few-Shot Learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Proceedings of the 34th International Conference on Neural Information Processing Systems, Curran Associates, Inc., 2020, pp. 1877–1901. doi:10.5555/3495724.3495883.

[4] Z. Wang, Y. Pang, Y. Lin, X. Zhu, Adaptable and Reliable Text Classification using Large Language Models, 2024. URL: https://arxiv.org/abs/2405.10523.

[5] S. Raza, D. Paulen-Patterson, C. Ding, Fake News Detection: Comparative Evaluation of BERT-like Models and Large Language Models with Generative AI-Annotated Data, 2024. doi:10.48550/arXiv.2412.14276.

[6] M. F. Mridha, A. J. Keya, M. A. Hamid, M. M. Monowar, M. S. Rahman, A Comprehensive Review on Fake News Detection With Deep Learning, IEEE Access 9 (2021) 156151–156170. doi:10.1109/ACCESS.2021.3129329.

[7] R. Xu, G. Li, A Comparative Study of Offline Models and Online LLMs in Fake News Detection, 2024. doi:10.48550/arXiv.2409.03067.

[8] V. S. Pendyala, C. E. Hall, Explaining Misinformation Detection Using Large Language Models, Electronics 13.9 (2024). doi:10.3390/electronics13091673.

[9] B. Jiang, Z. Tan, A. Nirmal, H. Liu, Disinformation Detection: An Evolving Challenge in the Age of LLMs, in: Proceedings of the 2024 SIAM International Conference on Data Mining (SDM), 2024, pp. 427–435. doi:10.1137/1.9781611978032.50.

[10] K. Pelrine, A. Imouza, C. Thibault, M. Reksoprodjo, C. Gupta, J. Christoph, J.-F. Godbout, R. Rabbany, Towards Reliable Misinformation Mitigation: Generalization, Uncertainty, and GPT-4, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 6399–6429. doi:10.18653/v1/2023.emnlp-main.395.

[11] M. Chen, L. Wei, H. Cao, W. Zhou, S. Hu, Explore the Potential of LLMs in Misinformation Detection: An Empirical Study, in: AAAI 2025 Workshop on Preventing and Detecting LLM Misinformation (PDLM), 2025. URL: https://openreview.net/forum?id=W2zawmik8i.

[12] K. Smelyakov, A. Chupryna, D. Darahan, S. Midina, Effectiveness of Modern Text Recognition Solutions and Tools for Common Data Sources, in: International Conference on Computational Linguistics and Intelligent Systems, 2021. URL: https://api.semanticscholar.org/CorpusID:235271469.

[13] Y. Cao, H. Li, Y. Yu, S. R. Javaji, Capybara at the Financial Misinformation Detection Challenge Task: Chain-of-Thought Enhanced Financial Misinformation Detection, in: C.-C. Chen, A. Moreno-Sandoval, J. Huang, Q. Xie, S. Ananiadou, H.-H. Chen (Eds.), Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal), Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 321–325. URL: https://aclanthology.org/2025.finnlp-1.38/.

[14] B. Hu, Q. Sheng, J. Cao, Y. Shi, Y. Li, D. Wang, P. Qi, Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2024, pp. 22105–22113. doi:10.1609/aaai.v38i20.30214.
[15] X. Zhang, W. Gao, Towards LLM-based Fact Verification on News Claims with a Hierarchical Step-by-Step Prompting Method, in: J. C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, A. Purwarianti, A. A. Krisnadhi (Eds.), Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Nusa Dua, Bali, 2023, pp. 996–1011. doi:10.18653/v1/2023.ijcnlp-main.64.
[16] W. Y. Wang, "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection, in: R. Barzilay, M.-Y. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 422–426. doi:10.18653/v1/P17-2067.
[17] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, H. Liu, FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media, 2019. URL: https://arxiv.org/abs/1809.01286.
[18] J.A. Leite, O. Razuvayevskaya, K. Bontcheva, C. Scarton, EUvsDisinfo: A Dataset for Multilingual Detection of Pro-Kremlin Disinformation in News Articles, in: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, Association for Computing Machinery, New York, NY, USA, 2024, pp. 5380–5384. doi:10.1145/3627673.3679167.
[19] AI House x Mantis Analytics, Disinformation Detection Challenge, 2023. URL: https://www.kaggle.com/competitions/disinformation-detection-challenge/data.
[20] R. Kyslyi, N. Romanyshyn, V. Sydorskyi, The UNLP 2025 Shared Task on Detecting Social Media Manipulation, 2025, pp. 105–111. doi:10.18653/v1/2025.unlp-1.12.
[21] V. Petyk, Ukrainian news, 2022. URL: https://www.kaggle.com/datasets/zepopo/ukrainian-fake-and-true-news.
[22] D. Dementieva, V. Khylenko, G. Groh, Cross-lingual Text Classification Transfer: The Case of Ukrainian, 2025. URL: https://arxiv.org/abs/2404.02043.
[23] V. Vysotska, K. Przystupa, Y. Kulikov, S. Chyrun, Y. Ushenko, Z. HU, D. Uhryn, Recognizing Fakes, Propaganda and Disinformation in Ukrainian Content based on NLP and Machine-learning Technology, Int. J. Comput. Netw. Inf. Secur. 17 (2025) 92–127. doi:10.5815/ijcnis.2025.01.08.
[24] A. Kiulian, A. Polishko, M. Khandoga, O. Chubych, J. Connor, R. Ravishankar, A. Shirawalmath, From Bytes to Borsch: Fine-Tuning Gemma and Mistral for the Ukrainian Language Representation, 2024. URL: https://arxiv.org/abs/2404.09138.
[25] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzm{\'a}n, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised Cross-lingual Representation Learning at Scale, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. doi:10.18653/v1/2020.acl-main.747.
[26] A. Bazdyrev, Russo-Ukrainian war disinformation detection in suspicious Telegram channels, 2025. doi:10.48550/arXiv.2503.05707.
[27] A. Shupta, P. Radiuk, I. Krak, Feature computation procedure for fake news detection: An LLM-based extraction approach, in: 6th International Workshop on Intelligent Information Technologies & Systems of Information Security, 2025.