

Are LLMs Enough for Hyperpartisan, Fake, Polarized and Harmful Content Detection? Evaluating In-Context Learning vs. Fine-Tuning

Michele Joshua Maggini¹, Dhia Merzougui², Rabiraj Bandyopadhyay³, Gaël Dias², Fabrice Maurel², Pablo Gamallo¹

¹Centro Singular de Investigación en Tecnologías Inteligentes da USC, Universidade de Santiago de Compostela

²UNICAEN, ENSICAEN, CNRS, GREYC, Normandie Univ

³GESIS Leibniz Institute for the Social Sciences

firstAuthor@usc.es

Abstract

The spread of fake news, polarizing, politically biased, and harmful content on online platforms has been a serious concern. With large language models becoming a promising approach, however, no study has properly benchmarked their performance across different models, usage methods, and languages. This study presents a comprehensive overview of different Large Language Models adaptation paradigms for the detection of hyperpartisan and fake news, harmful tweets, and political bias. Our experiments spanned 10 datasets and 5 different languages (English, Spanish, Portuguese, Arabic and Bulgarian), covering both binary and multiclass classification scenarios. We tested different strategies ranging from parameter efficient Fine-Tuning of language models to a variety of different In-Context Learning strategies and prompts. These included zero-shot prompts, codebooks, few-shot (with both randomly-selected and diversely-selected examples using Determinantal Point Processes), and Chain-of-Thought. We discovered that In-Context Learning often underperforms when compared to Fine-Tuning a model. This main finding highlights the importance of Fine-Tuning even smaller models on task-specific settings even when compared to the largest models evaluated in an In-Context Learning setup - in our case LLaMA3.1-8b-Instruct, Mistral-Nemo-Instruct-2407 and Qwen2.5-7B-Instruct.

Code and Dataset — https://github.com/HikariLight/hyperpartisanship_classification/tree/main

Introduction

Politically biased (PB), hyperpartisan (HP) and fake news (FN) as well as harmful (HF) social media content when covering divisive topics (e.g. politics, COVID-19) present a significant challenge to public discourse and democratic integrity, and most of those phenomena of our interest can fall under the misinformation category (Wardle and Derakhshan 2017). FN refers to fabricated stories that mimic legitimate news formats (Lazer et al. 2018). HP, on the other hand, involves misleading coverage of real events presented with a strong partisan bias (Potthast et al. 2018; Maggini et al. 2025). Both often contain politically charged messages that distort facts and polarize audiences. PB reporting further complicates the media landscape by subtly influencing and

polarizing public opinion and eroding the impartiality expected of journalism (Zhou and Zafarani 2020). While the concept of harmful content (HF) is broad and its content can vary based on political priorities (Eyuboglu et al. 2023), its diffusion on social media can rely on the spread of hate speech (Yang et al. 2023a) or misinformation (Eyuboglu et al. 2023). Indeed, harmful forms like polarizing content are frequently fueled by false or misleading narratives to incite frustration (Cinelli et al. 2021; Osmundsen et al. 2021). Consequently, the detection of such content necessitates robust fact-checking and verification (Nakov et al. 2022). This conceptual overlap is reflected in recent research challenges. For example, Azizov and Nakov (2023) considered harmful tweet detection a subtask of identifying relevant claims in tweets containing COVID-19 information. By defining a tweet as harmful when it potentially contained misinformation on COVID-19, they effectively demonstrated how, in a specific and critical context, the tasks of detecting harmful content and misinformation are intrinsically linked. Accurate detection of these diverse forms of problematic content is crucial. Large Language Models (LLMs) are valuable tools for this task. The dominant approaches have been fine-tuning (FT) both encoder-only models (Howard and Ruder 2018) and decoder-only LLMs (Aman 2024). However, a systematic comparison of these models’ performance on the specific tasks of fake and hyperpartisan news, politically biased news, and harmful content detection, especially across multiple languages, is still missing from the literature. Our work fills this gap with a comprehensive comparison of encoder-only and decoder-only LLMs using both fine-tuning (FT) and various in-context learning (ICL) settings. Our ICL methods include: zero-shot prompts with different degrees of task specifications and with rule-based approaches (e.g., codebooks); Few-shot (FS) using both randomly selected examples, and diversity-optimized examples selected through Determinantal Point Process (DPP); Chain-of-Thought (CoT). We conducted experiments on 10 datasets in five different languages, moving beyond the common limitation of using only English or U.S.-centric data (Maggini et al. 2025). This study addresses three key research questions:

- **RQ1:** How do different LLM adaptation paradigms (FT vs. ICL) compare for these tasks, considering model architecture, size, and pre-training data?

- **RQ2:** What is the impact of various ICL strategies—including few-shot examples and rule-based methods—on performance and stability?
- **RQ3:** How do performance and optimal strategies for these tasks vary across different languages, especially for mid- and low-resource languages?

Our extensive experiments reveal that FT remains a highly effective technique, often outperforming ICL strategies. Specifically, fine-tuned decoders performed better for PB and FN detection, while encoders were more effective for HP and HF tweets. Within ICL, the codebook approach was generally the effective, outperforming CoT for classification. The use of DPP for few-shot example selection sometimes reduced classification variance, though it did not consistently boost performance.

Related Work

Fine-tuning and Political Text Classification Text classification is an NLP task that assigns a label to a given text. FT adapts a pre-trained model to a specific task by further training the model together with a newly added classification head using task-specific labeled data. (Howard and Ruder 2018). While effective, this process can be sensitive to the classifier head’s initialization (Yang et al. 2022). In political text classification, FT has been successfully applied to various tasks. For instance, Liu et al. (2022) fine-tuned RoBERTa to create POLITICS, achieving state-of-the-art performance on SemEval 2019 for hyperpartisan news detection. Other works have explored stylistic features for hyperpartisan content discrimination (Potthast et al. 2018), created new datasets for multiclass hyperpartisan detection using fine-tuned BERT models (Lyu et al. 2023), or combined BERT with ELMo to enhance FT (Naredla and Adedoyin 2022). More recently, LLMs like Llama 2 have been fine-tuned for tasks such as FN detection, leveraging their understanding and analytical capabilities (Aman 2024; Pavlyshenko 2023).

In-Context Learning With the advent of recent decoder-only LLMs, ICL has emerged as a valuable technique in NLP. Users interact with models directly through prompts, specific textual templates containing instructions and optionally examples (Brown et al. 2020). This approach allows models to perform tasks without prior task-specific fine-tuning (Efrat and Levy 2020), leveraging a single pre-trained model for various downstream tasks and enabling desired behavior specification via natural language. ICL has shown remarkable performance on challenging reasoning tasks (Brown et al. 2020; Wei et al. 2022). However, ICL is highly sensitive to input format and order (Lu et al. 2022; Min et al. 2022), and can lead to irreproducible outcomes as slight prompt changes significantly impact performance (Lu, Schuff, and Gurevych 2024; Sun, Shaib, and Wallace 2023). To overcome these limitations, our work comprehensively tested various prompt strategies and demonstration selection methods, including DPP for more stable performance, and introduced codebook prompting. Research on prompt design aims to elicit better performance and reasoning. Notable approaches include CoT prompting (Kojima et al. 2022),

which encourages step-by-step reasoning, and its zero-shot variant (Wei et al. 2022). Lu et al. (2022) also highlighted the importance of careful prompt format and example selection in few-shot learning. Regarding the comparison of ICL and fine-tuning, Labrak, Rouvier, and Dufour (2024) and Edwards and Camacho-Collados (2024) demonstrated that fine-tuning smaller models often outperforms ICL in larger language models across various NLP and text classification tasks. Aligned with these findings, our study expands this comparison by evaluating a broader range of prompt strategies, models (including ModernBERT (Warner et al. 2024) beyond Llama models), and by focusing on specialized domains to rigorously compare the efficacy of these tools.

Codebook A codebook provides definitions of categories, including examples and classifying instructions. For instance, Vincent and Mestre (2018) developed a codebook to classify hyperpartisan news on a 5-point scale, and Hughes et al. (2021) crafted one for content analysis of COVID-19 articles, classifying tropes and rhetorical strategies to detect misinformation. Codebooks offer a method to explicitly prompt LLMs with structured contexts, eliciting their rule-based reasoning capabilities, which goes beyond standard ICL that often relies primarily on examples. Hu et al. (2024) explored codebook application in zero-shot settings with closed models for political phenomena classification, using the codebook as a structured framework for interpretation. Similarly, Halterman and Keith (2025) utilized codebooks to evaluate open models, demonstrating how these guidelines facilitate assessing an LLM’s adherence to pre-defined classification logic. While these specific codebooks were tailored to different political phenomena, NLP tasks, or modeled tasks differently from our dataset intentions, thus not directly applicable to our study, they still served as valuable inspiration for our experimental setup and underscore the potential of integrating structured rule sets into LLM prompts. This enhances adherence to specific classification schemes, aligning with broader prompt engineering efforts to elicit precise and controlled LLM outputs, especially in domains requiring nuanced categorization criteria.

Misinformation and Bias Detection LLMs demonstrate reasoning capabilities across various applications, including misinformation detection (Li et al. 2023; Leite et al. 2025). However, LLMs pose a dual challenge: they can be misused to spread misinformation, and their detection capabilities may diminish with implicit or newly crafted content (Chen and Shu 2024). Consequently, the efficacy of detection methods relative to the rapid updating rate of misinformation is a significant concern (Jiang et al. 2024). In misinformation studies, several works have explored LLM capabilities. Jose and Greenstadt (2024) compared proprietary models (GPT, Claude) for zero-shot propaganda detection, finding their performance inferior to RoBERTa-CRF. For hyperpartisan detection, Maggini and Gamallo Otero (2024) showed that increased prompt complexity and external knowledge usually improved Llama-3.1-8b-Instruct’s performance. Conversely, Omidi Shayegan et al. (2024) found encoder models like RoBERTa generally outperformed generative LLMs (GPT-3.5) for Persian hyperpartisan content. In Fake News

Detection, Anirudh, Srikanth, and Shahina (2023) observed gpt-3.5-turbo’s superiority over a bi-directional transformer for Tamil classification. Notably, while these works explore various aspects, none of them have focused on a comprehensive benchmarking across different ICL strategies, models, and multilingual contexts, which is a key contribution of our study.

Experimental Setting

Task Formulation

The core objective of this study is to evaluate the effectiveness of various NLP models in HP, FN, PB and HF detection, by comparing two widely used approaches: FT and ICL, where models are tested as off-the-shelf tools without additional training. Specifically, we focus on tasks such as identifying hyperpartisan and fake news, harmful tweets and a news’ political leaning, recognizing the distinct linguistic and contextual challenges each task presents. Our approach encompasses both binary and multi-class classification scenarios, leveraging a variety of datasets and multilingual contexts. While the ultimate goal is not to develop production-ready models, we prioritize thorough experimentation with various transformer-based architectures, prompt strategies, and learning techniques. This exploration serves to highlight the strengths and limitations of the tested architectures, contributing to the broader effort of refining misinformation detection methodologies within the NLP community. We acknowledge the fact that hyperpartisan news shows peculiar stylistic traits, rather than fake news (Potthast et al. 2018).

Datasets

For our experiments, we selected datasets for both binary and multiclass classification tasks. The 10 datasets focus on articles, headlines, tweets on COVID-19 and political news and different topics including TV, politics, sports, and health. They cover two types of domains: news and Twitter, and include four specific-oriented classification tasks: hyperpartisan and fake news, harmful tweet and political bias detection. For hyperpartisan detection, we selected the **SemEval-2019 Task 4 by-article** dataset (Kiesel et al. 2019), which contains articles from hyperpartisan and mainstream websites annotated by three annotators. They mostly cover the first term of Trump, Gun Control and other U.S.-centric related topics. The dataset’s strength lies in its article-level annotations that allow for analysis of extended argumentative structures and narrative techniques typical of hyperpartisan content, rather than just isolated claims. The **VISStA-H** dataset (Lyu et al. 2023) includes hyperpartisan and neutral news headlines from right, center and left U.S. newspapers. Its focus on headlines rather than full articles complements the SemEval dataset. The dataset’s temporal coverage, from 2014 to 2023, is particularly valuable for tracking how hyperpartisan news evolved through multiple election cycles. The **Fake News Net** dataset (Shu et al. 2017) contains news articles shared on Twitter, allowing for analysis of how hyperpartisan content circulates in social media environments. The **Spanish Fake News Corpus** (Gómez-Adorno et al. 2021) gathers news from Spanish

newspapers and media company websites, along with fact-checking websites. This corpus was selected to broaden the linguistic and cultural scope of the research beyond English-language and U.S.-centric media. The incorporation of fact-checking websites provides an additional layer of verification that strengthens the dataset’s reliability. The **Fake.br Corpus** (Monteiro et al. 2018) focuses on Brazilian Portuguese manually collected and checked news. The inclusion of this Brazilian Portuguese corpus further expands the cross-cultural and multilingual dimensions of the research. Brazil’s distinct political landscape and media environment provide an important comparative case for testing the generalizability of the detection approaches. The manual verification process strengthens the dataset’s reliability as a benchmark for testing detection methods in a language where NLP resources might be less abundant than for English or Spanish. **CLEF 2022 CheckThat! Lab Subtask 1C** (Nakov et al. 2022) was selected for its focus on COVID-19 misinformation tweets across multiple languages (Arabic, Bulgarian, and English), providing an opportunity to study fake news content around a global crisis that transcended national boundaries. Regarding political bias detection, we considered the **Qbias** dataset (Haak and Schaer 2023), which contains articles from AllSides, a news aggregator with an established methodology for evaluating political leaning. This provides a more nuanced and professionally curated ground truth for political bias than many other available resources. This nuanced approach helps move beyond binary classifications of political content and supports more sophisticated analysis of bias indicators. Lastly, **CLEF 2023 CheckThat! Lab Task 3A** dataset (Azizov and Nakov 2023) provides contemporary examples that reflect the current state of political communication and media bias. This diverse collection of datasets provides a comprehensive foundation for developing and evaluating models across multiple languages, cultural contexts, media formats and misinformation tasks. The URLs to retrieve the datasets can be found in the Appendix. To produce input for the classifier in the Spanish Fake News Corpus and Qbias datasets, we concatenated the headline and the body of the article. The datasets are summarized in Table 1.

Models

For our experiment, we compared two types of model architectures: encoder-only and decoder-only models.

Encoder-Only Masked Language Models: Following Edwards and Camacho-Collados (2024), we selected BERT-derived models such as RoBERTa-base (125 million parameters) and RoBERTa-large (354 million parameters) (Liu 2019), XLM-RoBERTa (Conneau et al. 2019), POLITICS (Liu et al. 2022), which has been adapted for the political domain using continuous pre-training, and mDeBERTaV3 (He, Gao, and Chen 2021). RoBERTa is pre-trained on English data, while XLM-RoBERTa is trained on 100 different languages, making it suitable for evaluating the impact of multilingual training on performance.

Decoder-Only Large Language Models: We experiment using different small-size open-weight LLMs: LLaMA3.1-8B and LLaMA3.1-8B-Instruct, Mistral-Nemo-Instruct-2407

Dataset	Abbr.	Lang.	Timeframe	Train Size	Test Size	Avg. Tkn	Train Len.	Avg. Tkn	Test Len.	Domain	Type	Task	Label Ratios (Train / Test)
VISTA-H (Lyu et al. 2023)	HV	en	2014–2023	1999	201	13	13	13	13	News	S	HP	HP: 0.39/0.63; N: 0.50/0.50
SemEval-2019 by-article (Kiesel et al. 2019)	SH	en	2007–N/A	645	628	735	757	News	News	News	D	HP	HP: 0.50/0.50; N: 0.50/0.50
Spanish Fake News Corpus (Gómez-Adorno et al. 2021)	SFN	es	2020–2021	676	572	607	843	News	News	News	D	FN	T: 0.50/0.50; F: 0.50/0.50
Fake News Net (Shu et al. 2017)	FNN	en	N/A	18556	4640	17	16	News	News	News	D	FN	T: 0.74/0.74; F: 0.26/0.26
Fake.br Corpus (Monteiro et al. 2018)	FBC	pt	2016–2018	5760	1440	688	698	News	News	News	D	FN	T: 0.50/0.50; F: 0.50/0.50
CLEF 2022 1C (Nakov et al. 2022)	CIA	ar	N/A	3624	1201	73	68	Twitter	Twitter	Twitter	S	HT	NH: 0.81/0.84; H: 0.19/0.16
	CIB	bu	N/A	708	325	62	68	Twitter	Twitter	Twitter	S	HT	NH: 0.87/0.97; H: 0.13/0.03
	CIE	en	N/A	3323	251	60	51	Twitter	Twitter	Twitter	S	HT	NH: 0.91/0.84; H: 0.09/0.16
CLEF 2023 3A (Azizov and Nakov 2023)	C3A	en	N/A	45066	5198	90	110	News	News	News	D	PB	R: 0.39/0.13; C: 0.34/0.38; L: 0.27/0.50
Qbias (Haak and Schaefer 2023)	QB	en	2012–2022	17403	4351	97	96	News	News	News	D	PB	R: 0.39/0.13; C: 0.34/0.38; L: 0.27/0.50

Table 1: Description of datasets used in our experiments. Average token length (Train/Test) is computed with the Llama3.1-8b tokenizer. Dataset types: S = Sentence, D = Document. Tasks: HP = Hyperpartisan News Detection; FN = Fake News Detection; HT = Harmful Tweet Detection; PB = Political Bias Detection. Labels: HP = Hyperpartisan, N = Neutral; T/F = True/Fake; NH/H = Non-harmful/Harmful; R/C/L = Right/Center/Left. Abbr. contains the abbreviation we will use in this paper to refer to the datasets.

(Mistral 2024), Qwen2.5-7B-Instruct (Qwen et al. 2025). These models are decoder-only and testing them allows us for generalizable effects across model families and tasks. The temperature was set to 0 for all the experiments. Generally, for non-English datasets, we evaluated models exclusively trained on multilingual datasets to ensure appropriate language coverage and performance. Further details are given in the Appendix .

Prompt design

Earlier studies like (Wei et al. 2022), (Jung et al. 2022) and (Mishra et al. 2022) have demonstrated the effectiveness of using task-specific prompts. Therefore, following (Edwards and Camacho-Collados 2024) and (Labrak, Rouver, and Dufour 2024), we constructed the prompts concatenating the following elements: 1) an instruction detailing the task, domain, and describing the meaning of the label; 2) the input argument, supplying essential information for the task; 3) the constraints on the output space, guiding the model during output generation. To improve the coherence, the specificity of the prompts, the instructions to follow in the codebook, and the fine-grained reasoning in CoT for the political domain, we collaborated with an expert in Political Science. In particular, to structure and develop our codebooks, we were inspired by Vincent and Mestre (2018) and (Halterman and Keith 2025), which introduced clear task definitions, explicit and exhaustive rules to determine the label of a data point, as well as provided examples covering both correct and borderline cases. We tested different prompting and ICL strategies such as zero-shot, Few-Shot, codebook and a variant of guided CoT, intending the reasoning as a multi-task evaluation (Lee et al. 2024; Duan et al. 2024), to provide explainable results like in (Yang et al. 2023a). We compare the random selection of Few-Shot exemplars with a more diversified selection using DPP, for which we provide a brief introduction in the following section. We also compare the results given by prompting the models with instructions containing different levels of complexity: general instructions, specific definitions of political phenomena or specialized instructions with more context provided. During the prompt optimization phase, we placed particular emphasis on ensuring that the model adhered to a consistent label format. This was crucial to ensure the outputs were reliably parseable. For instance, we discovered that the models when asked to pro-

vide string labels generated few unparseable outputs, considered wrong at the time of inference. This behavior happened across all the models and configurations. Moreover, at the begininnig we tested Mistral-7B and most of the time it did not follow the instructions regarding the template. This is why we did not introduce it in the experimental setting. Lastly, in CoT, to ensure its right functioning, we made sure the models generated the thoughts before the final output. Please, see the Appendix for further details.

Determinantal Point Process

Determinantal Point Process is a probability distribution over cloud of points that are used as computational tools across the fields of physics, statistics and machine learning (Gautier et al. 2019). DPP has been used to select diverse and representative set of datapoints for in-context learning (Yang et al. 2023b), data annotation (Wang et al. 2024b), instruction tuning (Wang et al. 2024a) and pre-training (Yang et al. 2024). DPP has been preferred for these tasks because it helps in promoting efficiency while maintaining a diversity of the selected subset from a large set. For this let us take 2 sets called index set $A = \{1, 2, \dots, N\}$ and its corresponding item set $I_A = \{x_1, x_2, \dots, x_N\}$. Then the problem of subset selection becomes evaluating 2^M subsets, which is computationally intractable and combinatorially explosive as the size of the super set grows. In order to approximately solve this problem DPP first uses the representation of the data \mathbf{x}_i . We use Sentence-BERT (Reimers and Gurevych 2019) to calculate the representations or embeddings. After the representations are computed, DPP algorithm works by calculating a Kernel $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ where kernel function k can be any similarity or distance metric between 2 points. Based on this we select a subset $Y \in A$, the probability selection of Y is given by.

$$P(Y) = \frac{\det(\mathbf{K}_Y)}{\det(\mathbf{K} + \mathbf{I})} \quad (1)$$

Here \mathbf{K}_Y is the subset of the matrix \mathbf{K} and consists of \mathbf{K}_{ij} for $i, j \in Y$. \mathbf{I} is the identity matrix and $\det(\cdot)$ represents the determinant of a matrix. Under these conditions the selection of the best subset can be formulated as the optimization problem as follows:

$$Y_{best} = \operatorname{argmax}_{Y \subset A, |Y|=k} \det(L_Y) \quad (2)$$

Algorithms exist to select such subsets of size k from the superset by sampling from the posterior distribution. For the task of selection we use the exact sampler (Gautier et al. 2019; Mazoyer, Coeurjolly, and Amblard 2020) which is the part of DPPy package by (Gautier et al. 2019) and is faster than Monte-Carlo sampler (Bardenet and Hardy 2019). Since the algorithm works on sampling at the kernel level it helps in selecting datapoints which are diverse in the representation space.

ICL Setting

Our aim is to compare the capabilities of different learning techniques, namely FT and ICL, and model architectures for hyperpartisan, fake news, harmful tweet and political bias classification. To investigate the ability of LLMs on those tasks in ICL, we used LLaMA3.1-8b-Instruct, Mistral-Nemo-Instruct-2407 and Qwen2.5-7B-Instruct by prompting them with different setups: 0-shot with General Prompt, 0-shot with Specific Prompt, CoT and Few-Shot with k -shot where $k \geq 0$. In the k -shot configuration, we adopted the General Prompt along with random examples and the respective labels of the dataset. To further test the stability of the prompting, we used Determinantal Point Process (Gautier et al. 2019) to select a diverse set of datapoints for each of the k -shot settings. The general prompt template is: <Role><Task description><Definition or Instructions><Text to classify or examples followed by text to classify><Response format>. We provide examples of different our prompts in the Appendix (see Tables 4, 5, 6 and 7). In order to maintain a balanced pool of examples, for multi-class datasets, we sampled from 1 to 3 examples per label; otherwise, we extracted the same k -shot per class. Finally, for non-English data, the corresponding roles and instructions were provided in the respective language to ensure accuracy and contextual relevance. The type of prompts we used are the following: **General Prompt** By providing the model with task-specific context (e.g., a headline, article, or tweet), we prompted it to classify the input text with the appropriate task label. With this configuration, we leverage the internal knowledge of the model to predict the answer, while being aware that it can suffer from political bias (Bang et al. 2024). We used it in 0-shot and Few-Shot.

Specific Prompt We slightly changed the previous template, introducing in the instruction the political definition of the phenomenon analyzed and some knowledge regarding the biases in partisan texts and asked the model to classify the text with the correct label. These political definitions were provided by a domain expert. Thus, we insert external knowledge and introduce a political definition to maximize model’s understanding capability and improve its outputs’ quality. We tested its efficacy only in zero-shot.

Codebook Based on previous works, with the help of a Political Scientist we crafted a specific codebook for each task. These codebooks contain a definition of the phenomenon, a description of the task’s characteristics considering several aspects (e.g., style, narrative) and particular linguistic features (e.g., use of hashtags, tone, source credibility). Furthermore, the detection criteria contain examples to help the LLM in understanding the specific rules for each

task. Crucially, this detailed codebook information was directly embedded within the `Definition or Instructions` component of our prompt template, allowing the LLM to perform rule-based reasoning for classification. This approach enables the models to leverage explicit, structured knowledge during inference, directly addressing the complexities of the tasks.

Guided CoT Prompt We guided the model to break down its reasoning step by step before making a final classification on the specified context, ensuring it produced explanations for all the steps before the final prediction. Specifically, we divided the hyperpartisan classification task into different sub-tasks: sentiment analysis, rhetorical bias, framing bias, ideology detection (Maggini and Gamallo Otero 2024). Moreover, by asking the model to identify itself with a specific political leaning, we are introducing recursive thinking (Duan et al. 2024). This method encourages the model to consider multiple factors and clearly articulate its reasoning, potentially leading to more robust and explainable classifications. Lastly, our approach allows us to consider the multidimensionality of each misinformation phenomenon analyzed.

By guiding the model through this structured reasoning process, we aimed to reduce misclassification and promote a more nuanced analysis. This approach also enabled us to observe how the model weighs different textual elements in its decision-making process, which helps in identifying any inherent biases or limitations within the model’s reasoning. We conducted preliminary tests with various prompts and configurations to refine the ones used in this experiment, ultimately selecting the configurations that yielded the best results on the training set. The optimization of these prompts was done manually rather than through automated methods. As a result, our prompts vary in terms of length, complexity, task specificity, and domain relevance, providing a comprehensive range of settings for evaluation. This structured and manually-optimized approach not only enhances the model’s classification performance but also provides deeper insights into the model’s interpretability and decision-making process across different political contexts.

Main Results and Discussion

Fine-Tuning

Table 2 presents the results for fine-tuning. On average across datasets, decoder-based models tend to outperform encoder-based models on tasks that require factual world knowledge, such as fake news detection and political bias identification. For instance, in fake news detection (Macro Avg. FN F1 score), the LLaMA3.1-8b (decoder) achieves .907, while the best performing encoder with a directly comparable macro average, ModernBERT-base, scores .854. Similarly, for political leaning detection (Macro Avg. PL F1 score), the Mistral-Nemo-Instruct (decoder) reaches .849, significantly surpassing the top encoder, POLITICS, which scores .675. Conversely, encoders achieve better results on linguistically oriented tasks, specifically harmful tweet detection and hyperpartisan language identification. RoBERTa-large (encoder) records an F1 score of .850 on

Model		HV	SH	Macro Avg. HP	FNN	SFN	FB	Macro Avg. FN	C1A	C1B	C1E	Macro Avg. HF	QB	C3A	Macro Avg. PL
RoBERTa-base	Acc	.822	.865	.843	.879	—	—	—	—	—	.919	—	.622	.659	.640
	F1	.818	.865	.841	.880	—	—	—	—	—	.915	—	.604	.660	.632
RoBERTa-large	Acc	.852	.865	.858	.893	—	—	—	—	—	.925	—	.683	.663	.673
	F1	.850	.865	.857	.893	—	—	—	—	—	.923	—	.674	.660	.667
XLM-RoBERTa	Acc	.827	.801	.814	.842	.623	.957	.807	.917	.917	.917	.917	.582	.632	.607
	F1	.825	.798	.811	.844	.577	.957	.793	.883	.884	.885	.884	.553	.629	.591
POLITICS	Acc	.831	.854	.842	.867	—	—	—	—	—	.916	—	.682	.679	.680
	F1	.826	.854	.840	.868	—	—	—	—	—	.911	—	.673	.678	.675
mDeBERTaV3	Acc	.819	.776	.797	.893	—	—	—	—	—	.915	—	.539	.590	.564
	F1	.816	.772	.794	.841	—	—	—	—	—	.874	—	.534	.570	.552
ModernBERT-large	Acc	.829	.854	.839	.858	.863	.941	.883	.815	.965	.835	.872	.658	.654	.653
	F1	.824	.853	.839	.846	.863	.941	.815	.815	.949	.803	—	.649	.657	.667
ModernBERT-base	Acc	.764	.780	.767	.852	.782	.942	.854	.830	.966	.830	.875	.584	.617	.592
	F1	.755	.779	.767	.840	.781	.942	.854	.792	.966	.774	.844	.571	.612	.592
LlaMA3.1-8b	Acc	.830	.810	.820	.945	.812	.975	.911	.864	.869	.921	.884	.788	.762	.775
	F1	.830	.801	.815	.945	.801	.975	.907	.858	.832	.920	.870	.786	.763	.774
LlaMA3.1-8b-Instruct	Acc	.784	.820	.801	.875	.823	.976	.890	.878	.915	.862	.885	.786	.796	.791
	F1	.782	.820	.801	.869	.823	.976	.889	.867	.928	.829	.875	.781	.802	.792
Mistral-Nemo-Instruct-2407	Acc	.834	.736	.783	.867	.725	.976	.851	.850	.943	.838	.877	.790	.846	.819
	F1	.833	.733	.783	.855	.722	.976	.851	.859	.946	.825	.877	.787	.789	.849
Qwen2.5-7B-Instruct	Acc	.819	.686	.745	.864	.685	.974	.835	.856	.913	.824	.864	.735	.698	.711
	F1	.812	.677	.745	.855	.675	.974	.835	.863	.928	.806	.866	.728	.693	.711

Table 2: Performance of models in the FT setting. The reported weighted Accuracy and weighted F1 scores are the averages obtained by running each model five times on the same dataset, reporting standard deviation.

HV, compared to the best decoder, Mistral-Nemo-Instruct, at .833. For SH, RoBERTa-large again leads with an F1 score of .865, while the best performing decoder on this task, LlaMA3.1-8b-Instruct, achieves .820. We hypothesize that this difference arises because the bidirectional attention mechanism of encoders may be better at capturing nuanced linguistic features, whereas decoders might excel at tasks more reliant on content or semantic understanding. Surprisingly, continuous pretraining of RoBERTa-base aimed at adapting it to either the political domain (POLITICS) or multilingual contexts (XLM-RoBERTa) has, in several instances, not led to improved performance over the original RoBERTa-base model and sometimes resulted in a reduction. For example, on the Macro Avg. HP (Hyperpartisan) task, RoBERTa-base achieves an F1 score of .841, whereas POLITICS scores .840 and XLM-RoBERTa scores .811. Regarding decoder models, we observe that a larger parameter count or expanded training corpus does not necessarily equate to superior results. This is demonstrated by LlaMA 3.1-8B outperforming Mistral Nemo-Instruct-2407 in hyperpartisan detection (SH F1 score of .801 for LlaMA 3.1-8b versus .733 for Mistral Nemo). Meanwhile, the decoder models exhibit relatively comparable high performance in harmful text detection (Macro Avg. HF F1 scores): LlaMA3.1-8b (.870), LlaMA3.1-8b-Instruct (.875), and Mistral-Nemo-Instruct-2407 (.877). During the FT experiment, we reached the SOTA in HV, FNN, SFN and FB. We provide the comparison with the previous research in Table 9 in the Appendix.

In-Context Learning

For the results discussed in this section, please refer to Figures 1 and 2 for the zero-shot configurations and CoT; and 3 for FS. Detailed results are reported in Table 8 in the Appendix

HP zero-shot-general vs zero-shot-specific In the hyperpartisan (HP) task, moving from generic to specific zero-shot prompting results in moderate improvements across

all three models, especially for LLaMA 3.1-8B Instruct and Mistral-Nemo-Instruct, whose F1 scores increase from 0.678 and 0.686 to 0.738 and 0.740, respectively. This performance gap indicates that these models have learned robust representations of hyperpartisan content during pre-training, and that providing more detailed task descriptions helps further refine their predictions. Lastly, SH predictions are generally stronger, largely because this dataset includes full articles rather than just headlines, as in HV. The richer contextual information in SH provides models with more linguistic and semantic cues, enabling a deeper understanding of the content and improving their ability to detect hyperpartisan narratives. In contrast, the limited context in headlines offers fewer signals for accurate classification.

Codebook The codebook approach yields improvements for SH, where Llama and Qwen demonstrate the most significant gains, particularly on the SH dataset, where their performance approaches F1 .810 and .748 respectively. This suggests that providing explicit criteria for identifying partisan language can help address edge cases where the models depend solely on their internal task representations. The performance gap between models narrows with codebook prompting, indicating that structured guidance can help equalize performance differences stemming from model's architecture, that may rely on different definitions of the phenomenon investigated.

FN zero-shot-general vs zero-shot-specific Fake news detection tasks show variable performance under zero-shot conditions. The FNN task proves more challenging are the multilingual datasets: SFN and FBC. The transition from generic to specific prompting yields minimal gains, with Mistral-Nemo-Instruct-2407 maintaining the highest performance on 2 out of 3 fake news (FN) datasets. We observed a 0.275-point drop in F1 score for Qwen on the Spanish dataset, which may be attributed to a misalignment between the fake news definitions—specifically, the model's internally assumed definition in the zero-shot generic prompt versus the expert-crafted definition used in the zero-shot spe-

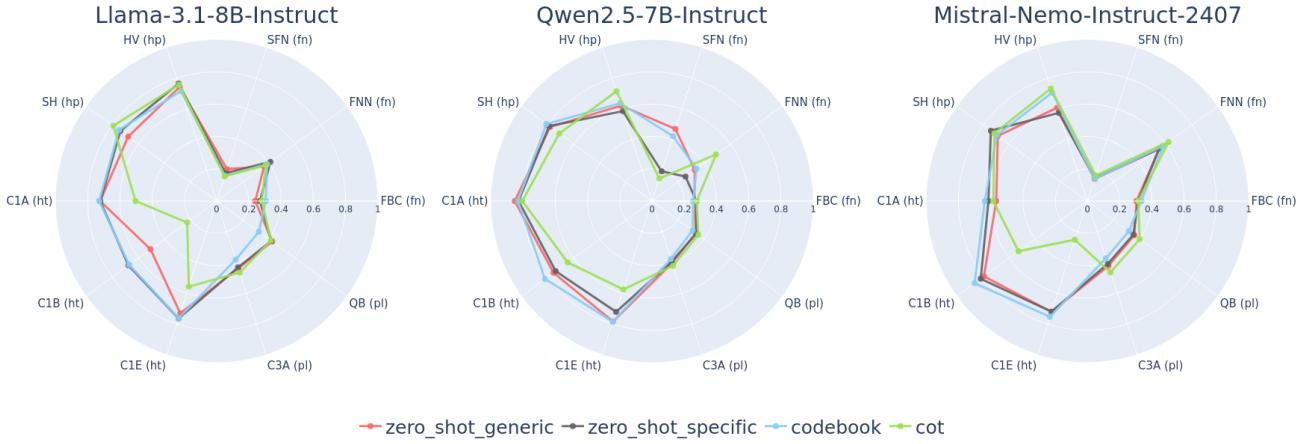


Figure 1: Results for zero-shot and CoT grouped by models.



Figure 2: Results zero-shot and CoT grouped by configuration.

specific prompt. This limited improvement suggests that, beyond clear definitions, models need additional contextual or world knowledge to effectively differentiate factual from fabricated content.

Codebook The codebook approach yields minimal improvements for FNN and SFN. indicates that providing explicit criteria for evaluating factual claims, source credibility markers, and stylistic indicators of fabricated content may help models overcome the inherent complexity of fact verification. The codebook’s little effectiveness in this domain suggests that FN does not completely benefit from structured evaluation frameworks. Regarding FN, with the different prompt strategies tested in ICL, we found out that the small LLMs are not effectively capable of detecting this kind of disinformation because they can not rely properly on the ontological structures encoded in their world-knowledge. Thus, employing these models as out-of-the-box tools with an ICL setup proves to be inefficient for this task.

PL zero-shot-general vs zero-shot-specific Political leaning (PL) classification tasks exhibit low performance across all models under zero-shot conditions. It is important to note that this is a multiclass classification task, which adds complexity. Generally, specific prompting produces a slight decrease for this task. Llama maintains consistently higher performance (on average F1 .416) than the other models.

The limited effectiveness with a specific definition of the political wings suggests that PL requires more than definitional refinement to overcome the inherent subjectivity involved.

Codebook Providing more detailed and specific knowledge through the codebook generally resulted in decreased performance across all models. This highlights the complexity of the task and suggests that, despite offering explicit rules to interpret the U.S. political context, agendas, the cultural nuances and linguistic factors involved are insufficient for effectively addressing the task. This result reveals the complexity underlying political bias detection.

HF zero-shot-general vs zero-shot-specific HF covered three languages and Qwen reached the best results in C1A with zero-shot-generic prompts (F1 .851). However, when prompted with zero-shot-specific prompts, its performance decreased. On the other hand, Llama particularly benefited from the introduction of the specific knowledge for C1B (from F1 .507 to .676) and C1E (from F1 .730 to .764).

Codebook With the introduction of specific dimensions to frame the task, all the models across the HF datasets (except for Qwen in C1A) improved their performances. Specifically, Mistral reached F1 .864 in C1B. This marked improvement highlights the value of explicit harm taxonomies and classification criteria for this sensitive domain. The codebook’s effectiveness for harmful content detection

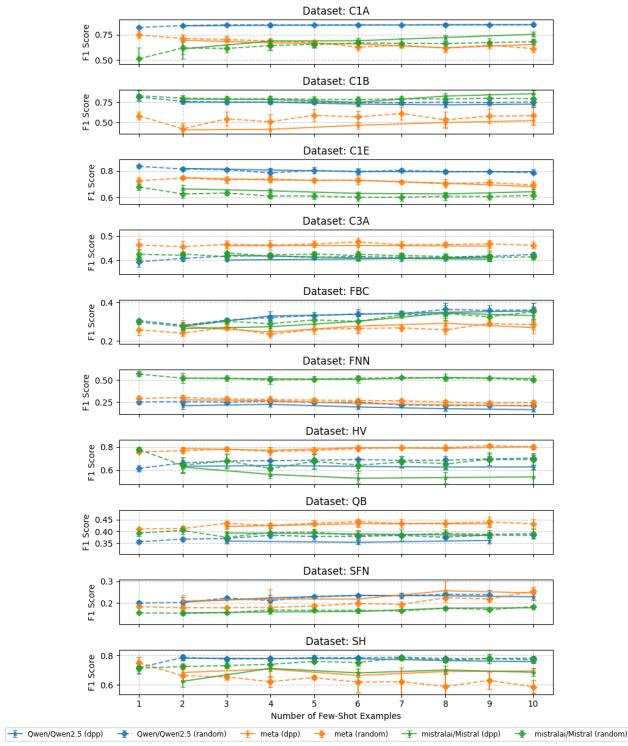


Figure 3: FS DPP vs Random results.

suggests that these tasks require clearly articulated boundaries and examples to overcome potential ambiguities in what constitutes harmful material.

Summary for Zero-shot configurations

Across all classification domains, the transition from generic to specific zero-shot prompting lead only to marginal improvements. This pattern suggests that merely elaborating on task definitions provides insufficient guidance for these models to significantly alter their classification behavior. The codebook approach demonstrates modest effectiveness across all task categories, with particularly improvements for HF and FN classification. This pattern indicates that providing structured classification criteria helps models overcome the inherent complexities of these judgment tasks. The codebook's effectiveness stems from its ability to bridge the gap between abstract classification concepts and concrete textual indicators, providing models with clearer decision boundaries for ambiguous cases. Lastly, we noticed that more subjective and nuanced tasks like HF and PL, that require extensive knowledge of the facts and their truthfulness, show improvement with advanced prompting strategies than more stylistic-based tasks (e.g. HP detection), suggesting that prompt optimization benefits may correlate with task complexity. Indeed, the rule based approach is the best ICL configuration for 3 out of 10 datasets: SH, C1B and FBC.

FS: DPP-selected vs Random examples

To test the Few-Shot capacity of the model, we decided to compare the performances using random datapoints against a representative set of examples, maintaining the dataset diversity using DPP. Table 3 reports the results for this comparison. Random selection risks subsets that lack diversity or fail to represent edge cases, while DPP ensures prompt stability by challenging the model with dissimilar patterns. In Few-Shot Learning, where models rely on limited examples to generalize, diverse subsets prevent overfitting to specific features and improve generalization. DPP-selected examples enhance prompt informativeness by showcasing varied cases, enabling the model to better understand nuanced relationships. This diversity reduces classification errors and improves accuracy by covering a broader range of inputs. A key observation across both FS Random and FS DPP is that increasing the number of shots does not consistently or monotonically improve performance for all models and datasets. Generally, performance often peaks at an intermediate number of shots (e.g., 1-shot, 5-shot, 6-shot) and can then plateau, fluctuate, or even decline as more shots are added (e.g., in the range 7-10 shots). This implies that simply providing more examples is not always the best strategy, regardless the datapoints representativeness of the selected examples. There is not any ideal threshold for the n-shots, since the performances vary across models and datasets. For instance, while using FS Random examples in Hyperpartisan Detection on the SH dataset, Llama-3.1-8b-Instruct's F1 score varies from .751 (1-shot) down to .623 (4-shot) and then to .587 (10-shot). However, the results do not indicate a clear best method across all scenarios. Peak performance for a given model and dataset can be achieved by either method, often at different n-shot values. For example, with Llama-3.1-8B-Instruct on the HV dataset, Random 9-shot yields an F1 of .811, while DPP 10-shot gives .801. Conversely, Mistral-Nemo-Instruct-2407 on C1B achieves its highest few-shot F1 of .851 with DPP 10-shot. FS Random can achieve high scores, possibly when the random selection happens to include particularly effective examples. However, its performance can be inherently more variable. FS DPP, by design, selects for diversity, which might be expected to lead to more robust or consistent improvements. While it achieves strong results in some cases, it also exhibits fluctuations and doesn't always outperform Random FS.

Chain of Thought

This setting provided reasoning steps with increasing levels of abstraction modeled as successively sub-task steps. Llama-3.1-8B-Instruct shows good performance with CoT on HP tasks, achieving the best F1/Accuracy in its section for both SH (F1 .792, Acc .795) and HV (F1 .757, Acc .764) datasets. It also performs well on PL Detection for C3A (F1 .465, Acc .459) and QB (F1 .416, Acc .405), again leading in its section. Mistral with CoT achieves the overall best scores for FN Detection across all models and configurations on the FNN dataset (F1 .623, Acc .680) and across sections in FBC dataset (F1 .315, Acc .397), and SFN (F1 .167, Acc .168). This suggests CoT might be particularly beneficial for this

model on these specific complex tasks. Regarding, Qwen, CoT helps it in achieving strong results on C1A (F1 .804, Acc .783), but only F1 .646 in C1B and F1 .575 in C1E, leading this section for these datasets. Nevertheless, in most of the cases, it revealed to be suboptimal. The CoT prompting proved largely suboptimal across our experiments, showing significant improvement only for the FNN task. Our analysis suggests this underperformance stems primarily from language representation issues in the model’s training data. When prompted in underrepresented languages with insufficient training tokens, the models struggled to process the unfamiliar linguistic patterns. Rather than aiding reasoning, these novel tokens appeared to confuse the models, disrupting their inference capabilities. Additionally, we observed that even in zero-shot-specific, codebook, and few-shot configurations, the models sometimes generated explanations unprompted, suggesting they were trained to occasionally provide reasoning alongside their answers. This built-in explanatory behavior likely accounts for why explicit CoT prompting offered minimal additional benefits despite its theoretical advantages and added complexity.

Insights from Fine-Tuning vs. In-Context Learning in LLMs

Across all the configurations tested in our experiments, FT emerged as the most effective method to apply models to the political domain, and, in particular, the disinformation subdomain. Specifically, for LLMs, the FT configuration demonstrated its efficacy in 28 out of 33 cases. Notably, Qwen particularly benefited from ICL achieving its best performance with the following configurations: few-shot random for C1E (F1: 0.833) and SFN (F1: 0.678), and zero-shot codebook for SH (F1: 0.810). These results suggest that updating model parameters through FT is generally the most reliable way to optimize performance for downstream tasks in this domain. However, ICL remains a valid and convenient strategy for probing a model’s task-specific knowledge without parameter updates. Despite our efforts to optimize prompts—by incorporating external domain-specific knowledge, employing rule-based approaches, and eliciting reasoning capabilities—ICL configurations still showed more limited effectiveness compared to FT. Lastly, both model architectures benefited from fine-tuning, with encoder-based models achieving superior performance on 6 out of 10 datasets, and smaller LLMs performing better on the remaining 4—particularly in tasks such as fake news and political leaning detection, which require deeper world knowledge. It is important to note, however, that fine-tuning—especially when applied to LLMs—demands significant computational resources, making it a considerably resource-intensive approach.

Conclusion and Future Work

This paper provides a comprehensive benchmark for FT and ICL methods across classification tasks in several misinformation domains. Indeed, we largely compared different model architectures, learning techniques and sets of prompts in several classification tasks. We evaluated

performance on 10 diverse datasets, spanning binary and multiclass contexts in English, Spanish, Brazilian Portuguese, Arabic, and Bulgarian. Particularly, we compared nine models covering the following transformer family’s architecture: (1) encoders: RoBERTa-base, RoBERTa-large, XLM-RoBERTa, POLITICS, Modern-BERT-base and -large, mDeBERTaV3; (2) decoders: LLaMA3.1-8b-Instruct, Mistral-Nemo-Instruct-2407 and Qwen2.5-7B-Instruct. ICL consistently proved to be less effective than fine-tuning across most settings. Indeed, results showed that in FT, decoders were better at PB and FN detection, whereas encoders were better at HP and HF tweet detection. Regarding ICL, we applied different levels of prompt optimization, testing all the main ICL techniques. In Few-Shot, we found that DPP sometimes reduces variations in classification rather than randomly sampling the shots, though it does not systematically increase the performance. Furthermore, the adaptation methods in ICL did not behave the same depending on the LLM used. Lastly, except for fake news detection tasks, eliciting the model with a codebook was generally the best approach in ICL, making the CoT unreliable for classification task. Future work could investigate the application of a RAG system to incorporate up-to-date information and improve the factual verification of news.

Limitations

LLMs In CoT, to overcome the different templates generated by the model in the initial phase of our experiment, we crafted a template for the different tasks (see the prompt Tables 4,5,6,7 in the Appendix). Moreover, in few cases the model produced irregular outputs. When this occurred, we considered those cases to be incorrectly labeled.

Practical applications Our work is a comprehensive overview of architectures and methods. Nevertheless, some datasets do not contain up-to-date information that can be used effectively to tackle fake news propagation, since this kind of misinformation does not rely only on linguistic clues but also on pre-existence knowledge of political facts. Indeed, the temporal limitation can affect how the perception of a fact - a general one - can be perceived and or discussed, because the language is subjected to changes over time. However, those dataset could be used for continual pre-training or as part of a RAG system that also incorporates up-to-date information.

Domain We acknowledge that our experiments focus on a subset of political NLP tasks, specifically misinformation-related tasks across different languages. As such, our findings should not be generalized to the full range of NLP tasks.

Open LLMs and size We limited our model selection to open models, while discarding closed one (e.g. Claude, OpenAI) for the sake of reproducibility and budget limitations. Furthermore, our GPUs could not host larger models. This fact limits our findings.

Ethics Statement

One of the primary objectives of this work is to address the challenge of misinformation spreading — a critical issue in today’s society. Tackling misinformation is both ben-

eficial to society and ethically imperative, as it contributes to a more informed and balanced public discourse. However, we acknowledge that our work is not without risks and potential unintended consequences. Our study involves the exploration of various architectures and models to assess their reliability in identifying and countering misinformation. This process inherently carries ethical considerations, particularly related to the datasets we used. The datasets include linguistically hazardous data, such as offensive content in the Harmful Tweet dataset, and highly polarized messages, as seen in the Hyperpartisan News and Political Bias detection datasets. Although the datasets used are crucial for the development of robust and effective models, they also pose risks of misuse. Specifically, such datasets could potentially be exploited to train LLMs capable of generating biased or misleading political content, thereby exacerbating the very problem we aim to mitigate. To address these risks, we took steps to ensure our work adheres to ethical standards. Firstly, regarding data handling, we carefully managed the datasets, using them only for the tests described in our paper and not for other unethical purposes. We emphasize the importance of using the models and methodologies developed in this study exclusively for combating misinformation and promoting ethical information dissemination. Misuse of these tools to create or amplify harmful content is strongly discouraged. Then, by openly releasing our code and datasets, we aim to promote transparency and encourage responsible research practices. We also are going to provide detailed documentation to inform users of the potential risks associated with these datasets. VISTA-H and SemEval-2019, Qbias, and CLEF23 3A contain both headlines and articles with extremely polarized content from both Right and Left wing leanings, slurs and racist sentences. CLEF22 1C ar, bu and en host tweets supporting conspiracy theories related to COVID-19. Spanish Fake News Corpus gather cultural fake news as well as ones against gender equality and the LGBTQIA+ community. Fake News Net and Fake.br-Corpus contain a wide range of fake news on different topics, from political to climate change. Lastly, recognizing the potential for misuse, we encourage future researchers and practitioners to implement safeguards, such as adversarial testing and bias detection frameworks, when deploying these models. We further stress that while our findings demonstrate the superior performance of fine-tuned models over in-context learning strategies, this advantage must be wielded with care. Fine-tuned models can be highly specialized and powerful, making it imperative to ensure they are used responsibly. As researchers, we are committed to fostering a dialogue around the ethical implications of NLP technologies and encouraging their use for the betterment of society. By highlighting these concerns and promoting transparency, we aim to contribute to a more ethical and responsible approach to NLP research in the context of misinformation detection.

Acknowledgements

This project has received funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Grant Agree-

ment No. 101073351.

References

Aman, M. 2024. Large language model based fake news detection. *Procedia Computer Science*, 231: 740–745.

Anirudh, K.; Srikanth, M.; and Shahina, A. 2023. Multilingual Fake News Detection in Low-Resource Languages: A Comparative Study Using BERT and GPT-3.5. In *International Conference on Speech and Language Technologies for Low-resource Languages*, 387–397. Springer.

Azizov, D.; and Nakov, P. 2023. Overview of the CLEF-2023 CheckThat! Lab Task 3 on Political Bias of News Articles and News Media. 3497: 250–259.

Bang, Y.; Chen, D.; Lee, N.; and Fung, P. 2024. Measuring Political Bias in Large Language Models: What Is Said and How It Is Said. *ArXiv*, abs/2403.18932.

Bardenet, R.; and Hardy, A. 2019. Monte Carlo with Determinantal Point Processes. *arXiv*:1605.00361.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. *ArXiv*, abs/2005.14165.

Chen, C.; and Shu, K. 2024. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*, 45(3): 354–368.

Cinelli, M.; Pelicon, A.; Mozetič, I.; Quattrociocchi, W.; Novak, P. K.; and Zollo, F. 2021. Dynamics of online hate and misinformation. *Scientific reports*, 11(1): 22083.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Duan, J.; Wang, S.; Diffenderfer, J.; Sun, L.; Chen, T.; Kailkhura, B.; and Xu, K. 2024. ReTA: Recursively Thinking Ahead to Improve the Strategic Reasoning of Large Language Models. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2232–2246. Mexico City, Mexico: Association for Computational Linguistics.

Edwards, A.; and Camacho-Collados, J. 2024. Language Models for Text Classification: Is In-Context Learning Enough? In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 10058–10072. Torino, Italia: ELRA and ICCL.

Efrat, A.; and Levy, O. 2020. The Turking Test: Can Language Models Understand Instructions? *ArXiv*, abs/2010.11982.

Eyüboğlu, A. B.; Altun, B.; Arslan, M. B.; Sonmezler, E.; and Kutlu, M. 2023. Fight against misinformation on social media: detecting attention-worthy and harmful tweets and verifiable and check-worthy claims. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, 161–173. Springer.

FORCE11. 2020. The FAIR Data principles. <https://force11.org/info/the-fair-data-principles/>.

Gautier, G.; Polito, G.; Bardenet, R.; and Valko, M. 2019. DPPy: DPP Sampling with Python. *Journal of Machine Learning Research*, 20(180): 1–7.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gómez-Adorno, H.; Posadas-Durán, J. P.; Enguix, G. B.; and Capetillo, C. P. 2021. Overview of fakedes at Iberlef 2021: Fake news detection in Spanish shared task. *Procesamiento del lenguaje natural*, 67: 223–231.

Haak, F.; and Schaer, P. 2023. Qbias - A Dataset on Media Bias in Search Queries and Query Suggestions. In *Proceedings of the 15th ACM Web Science Conference 2023*, WebSci '23, 239–244. New York, NY, USA: Association for Computing Machinery. ISBN 9798400700897.

Halterman, A.; and Keith, K. A. 2025. Codebook LLMs: Evaluating LLMs as Measurement Tools for Political Science Concepts. arXiv:2407.10747.

He, P.; Gao, J.; and Chen, W. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Howard, J.; and Ruder, S. 2018. Universal Language Model Fine-tuning for Text Classification. arXiv:1801.06146.

Hu, Y.; Skorup Parolin, E.; Khan, L.; Brandt, P.; Osorio, J.; and D’Orazio, V. 2024. Leveraging Codebook Knowledge with NLI and ChatGPT for Zero-Shot Political Relation Classification. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 583–603. Bangkok, Thailand: Association for Computational Linguistics.

Hughes, B.; Miller-Idriss, C.; Piltch-Loeb, R.; Goldberg, B.; White, K.; Criezis, M.; and Savoia, E. 2021. Development of a Codebook of Online Anti-Vaccination Rhetoric to Manage COVID-19 Vaccine Misinformation. *International Journal of Environmental Research and Public Health*, 18(14).

Jiang, B.; Tan, Z.; Nirmal, A.; and Liu, H. 2024. Disinformation detection: An evolving challenge in the age of llms. In *Proceedings of the 2024 siam international conference on data mining (sdm)*, 427–435. SIAM.

Jin, Y.; Wang, X.; Yang, R.; Sun, Y.; Wang, W.; Liao, H.; and Xie, X. 2022. Towards Fine-Grained Reasoning for Fake News Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(5): 5746–5754.

Jose, J.; and Greenstadt, R. 2024. Are Large Language Models Good at Detecting Propaganda?

Jung, J.; Qin, L.; Welleck, S.; Brahman, F.; Bhagavatula, C.; Le Bras, R.; and Choi, Y. 2022. Maieutic Prompting: Logically Consistent Reasoning with Recursive Explanations. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 1266–1279. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Kiesel, J.; Mestre, M.; Shukla, R.; Vincent, E.; Adineh, P.; Corney, D.; Stein, B.; and Potthast, M. 2019. SemEval-2019 Task 4: Hyperpartisan News Detection. In *International Workshop on Semantic Evaluation*.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large Language Models are Zero-Shot Reasoners. *ArXiv*, abs/2205.11916.

Labrak, Y.; Rouvier, M.; and Dufour, R. 2024. A Zero-shot and Few-shot Study of Instruction-Finetuned Large Language Models Applied to Clinical and Biomedical Tasks. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2049–2066. Torino, Italia: ELRA and ICCL.

Lazer, D. M. J.; Baum, M. A.; Benkler, Y.; Berinsky, A. J.; Greenhill, K. M.; Menczer, F.; Metzger, M. J.; Nyhan, B.; Pennycook, G.; Rothschild, D.; Schudson, M.; Sloman, S. A.; Sunstein, C. R.; Thorson, E. A.; Watts, D. J.; and Zittrain, J. L. 2018. The science of fake news. *Science*, 359(6380): 1094–1096.

Lee, J.; Yang, F.; Tran, T.; Hu, Q.; Barut, E.; and Chang, K.-W. 2024. Can Small Language Models Help Large Language Models Reason Better?: LM-Guided Chain-of-Thought. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2835–2843. Torino, Italia: ELRA and ICCL.

Leite, J. A.; Razuvayevskaya, O.; Bontcheva, K.; and Scarton, C. 2025. Weakly supervised veracity classification with LLM-predicted credibility signals. *EPJ Data Science*, 14(1): 16.

Li, X.; Chan, S.; Zhu, X.; Pei, Y.; Ma, Z.; Liu, X.; and Shah, S. 2023. Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics? A Study on Several Typical Tasks. In Wang, M.; and Zitouni, I., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 408–422. Singapore: Association for Computational Linguistics.

Liu, Y. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Y.; Zhang, X. F.; Wegsman, D.; Beauchamp, N.; and Wang, L. 2022. POLITICS: Pretraining with Same-story Article Comparison for Ideology Prediction and Stance Detection. In Carpuat, M.; de Marneffe, M.-C.; and Meza Ruiz, I. V., eds., *Findings of the Association for Computational Linguistics: NAACL 2022*, 1354–1374. Seattle, United States: Association for Computational Linguistics.

Lu, S.; Schuff, H.; and Gurevych, I. 2024. How are Prompts Different in Terms of Sensitivity? In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 5833–5856. Mexico City, Mexico: Association for Computational Linguistics.

Lu, Y.; Bartolo, M.; Moore, A.; Riedel, S.; and Stenetorp, P. 2022. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8086–8098. Dublin, Ireland: Association for Computational Linguistics.

Lyu, H.; Pan, J.; Wang, Z.; and Luo, J. 2023. Computational Assessment of Hyperpartisanship in News Titles.

Maggini, M.; and Gamallo Otero, P. 2024. Leveraging Advanced Prompting Strategies in LLaMA3-8B for Enhanced Hyperpartisan News Detection. In Dell’Orletta, F.; Lenci, A.; Montemagni, S.; and Sprugnoli, R., eds., *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, 531–539. Pisa, Italy: CEUR Workshop Proceedings. ISBN 979-12-210-7060-6.

Maggini, M. J.; Bassi, D.; Piot, P.; Dias, G.; and Otero, P. G. 2025. A systematic review of automated hyperpartisan news detection. *PLOS ONE*, 20(2): 1–39.

Mazoyer, A.; Coeurjolly, J.-F.; and Amblard, P.-O. 2020. Projections of determinantal point processes. arXiv:1901.02099.

Min, S.; Lyu, X.; Holtzman, A.; Artetxe, M.; Lewis, M.; Hajishirzi, H.; and Zettlemoyer, L. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11048–11064. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Mishra, S.; Khashabi, D.; Baral, C.; and Hajishirzi, H. 2022. Cross-Task Generalization via Natural Language Crowdsourcing Instructions. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3470–3487. Dublin, Ireland: Association for Computational Linguistics.

Mistral. 2024. Mistral nemo: Collaborative innovation with nvidia.

Monteiro, R. A.; Santos, R. L. S.; Pardo, T. A. S.; de Almeida, T. A.; Ruiz, E. E. S.; and Vale, O. A. 2018. Contributions to the Study of Fake News in Portuguese: New Corpus and Automatic Detection Results. In *Computational Processing of the Portuguese Language*, 324–334. Springer International Publishing. ISBN 978-3-319-99722-3.

Nakov, P.; Barrón-Cedeño, A.; Da San Martino, G.; Alam, F.; Struß, J. M.; Mandl, T.; Míguez, R.; Caselli, T.; Kutlu, M.; Zaghouni, W.; Li, C.; Shaar, S.; Shahi, G. K.; Mubarak, H.; Nikolov, A.; Babulkov, N.; Kartal, Y. S.; and Beltrán, J. 2022. The CLEF-2022 CheckThat! Lab on Fighting the COVID-19 Infodemic and Fake News Detection. In Hagen, M.; Verberne, S.; Macdonald, C.; Seifert, C.; Balog, K.; Nørvåg, K.; and Setty, V., eds., *Advances in Information Retrieval*, 416–428. Cham: Springer International Publishing. ISBN 978-3-030-99739-7.

Naredla, N. R.; and Adedoyin, F. F. 2022. Detection of hyperpartisan news articles using natural language processing technique. *International Journal of Information Management Data Insights*, 2(1): 100064.

Omidi Shayegan, S.; Nejadgholi, I.; Pehrine, K.; Yu, H.; Levy, S.; Yang, Z.; Godbout, J.-F.; and Rabbany, R. 2024. An Evaluation of Language Models for Hyperpartisan Ideology Detection in Persian Twitter. In Ojha, A. K.; Ahmadi, S.; Cinková, S.; Fransen, T.; Liu, C.-H.; and McCrae, J. P., eds., *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI) @ LREC-COLING 2024*, 51–62. Torino, Italia: ELRA and ICCL.

Osmundsen, M.; BOR, A.; VAHLSTRUP, P. B.; BECHMANN, A.; and PETERSEN, M. B. 2021. Partisan Polarization Is the Primary Psychological Motivation behind Political Fake News Sharing on Twitter. *American Political Science Review*, 115(3): 999–1015.

Pavlyshenko, B. M. 2023. Analysis of disinformation and fake news detection using fine-tuned large language model. *arXiv preprint arXiv:2309.04704*.

Potthast, M.; Kiesel, J.; Reinartz, K.; Bevendorff, J.; and Stein, B. 2018. A Stylometric Inquiry into Hyperpartisan and Fake News. In Gurevych, I.; and Miyao, Y., eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 231–240. Association for Computational Linguistics.

Qwen; Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Tang, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2025. Qwen2.5 Technical Report. *arXiv:2412.15115*.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. Hong Kong, China: Association for Computational Linguistics.

Shu, K.; Sliva, A.; Wang, S.; Tang, J.; and Liu, H. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explor. Newslett.*, 19(1): 22–36.

Sun, J.; Shaib, C.; and Wallace, B. C. 2023. Evaluating the zero-shot robustness of instruction-tuned language models. *arXiv preprint arXiv:2306.11270*.

Vincent, E.; and Mestre, M. 2018. Crowdsourced Measure of News Articles Bias: Assessing Contributors' Reliability. In *1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing, and Short Paper Proceedings of the 1st Workshop on Disentangling the Relation Between Crowdsourcing and Bias Management (SAD 2018 and CrowdBias 2018)*.

Wang, P.; Shen, Y.; Guo, Z.; Stallone, M.; Kim, Y.; Golland, P.; and Panda, R. 2024a. Diversity Measurement and Subset Selection for Instruction Tuning Datasets. *arXiv:2402.02318*.

Wang, P.; Wang, X.; Lou, C.; Mao, S.; Xie, P.; and Jiang, Y. 2024b. Effective Demonstration Annotation for In-Context Learning via Language Model-Based Determinantal Point Process. *arXiv:2408.02103*.

Wardle, C.; and Derakhshan, H. 2017. INFORMATION DISORDER: Toward an interdisciplinary framework for research and policy making. *Information Disorder: Toward an interdisciplinary framework for research and policymaking*. *Zugriff am*, 1(2023): 39–53.

Warner, B.; Chaffin, A.; Clavié, B.; Weller, O.; Hallström, O.; Taghadouini, S.; Gallagher, A.; Biswas, R.; Ladakh, F.; Aarsen, T.; Cooper, N.; Adams, G.; Howard, J.; and Poli, I. 2024. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. *arXiv:2412.13663*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; hsin Chi, E. H.; Xia, F.; Le, Q.; and Zhou, D. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. *ArXiv*, abs/2201.11903.

Yang, Y.; Kim, J.; Kim, Y.; Ho, N.; Thorne, J.; and Yun, S.-Y. 2023a. HARE: Explainable Hate Speech Detection with Step-by-Step Reasoning. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 5490–5505. Singapore: Association for Computational Linguistics.

Yang, Y.; Wang, H.; Wen, M.; Mo, X.; Peng, Q.; Wang, J.; and Zhang, W. 2024. P3: A Policy-Driven, Pace-Adaptive, and Diversity-Promoted Framework for data pruning in LLM Training. *arXiv:2408.05541*.

Yang, Z.; Ding, M.; Guo, Y.; Lv, Q.; and Tang, J. 2022. Parameter-Efficient Tuning Makes a Good Classification Head. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 7576–7586. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Yang, Z.; Zhang, Y.; Sui, D.; Liu, C.; Zhao, J.; and Liu, K. 2023b. Representative Demonstration Selection for In-Context Learning with Two-Stage Determinantal Point Process. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 5443–5456. Singapore: Association for Computational Linguistics.

Zhou, X.; and Zafarani, R. 2020. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Comput. Surv.*, 53(5).

Ethics Checklist

1. For most authors...

- (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
- (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
- (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
- (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Not Applicable
- (e) Did you describe the limitations of your work? Yes
- (f) Did you discuss any potential negative societal impacts of your work? Yes
- (g) Did you discuss any potential misuse of your work? Yes
- (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
- (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes

2. Additionally, if your study involves hypotheses testing...

- (a) Did you clearly state the assumptions underlying all theoretical results? Not Applicable
- (b) Have you provided justifications for all theoretical results? Not Applicable
- (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Not Applicable
- (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Not Applicable
- (e) Did you address potential biases or limitations in your theoretical framework? Not Applicable
- (f) Have you related your theoretical results to the existing literature in social science? Not Applicable
- (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Not Applicable

We did not work with hypothesis testing.

3. Additionally, if you are including theoretical proofs...

- (a) Did you state the full set of assumptions of all theoretical results? Not Applicable

- (b) Did you include complete proofs of all theoretical results? Not Applicable

4. Additionally, if you ran machine learning experiments...

- (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes
- (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, both in Selected Models section and in the Appendix.
- (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes
- (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes
- (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes
- (f) Do you discuss what is “the cost“ of misclassification and fault (in)tolerance? Not Applicable

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

- (a) If your work uses existing assets, did you cite the creators? Yes
- (b) Did you mention the license of the assets? The datasets are all freely available and we reported their URLs. Same for the models employed in the experiments.
- (c) Did you include any new assets in the supplemental material or as a URL? Not Applicable
- (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? All the data are publicly available from their repositories.
- (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes
- (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see (FORCE11 2020))? Not Applicable
- (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see (Gebru et al. 2021))? Not Applicable

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

- (a) Did you include the full text of instructions given to participants and screenshots? Not Applicable
- (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Not Applicable
- (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Not Applicable
- (d) Did you discuss how data is stored, shared, and de-identified? Not Applicable

Appendix

Datasets URLs

In the following paragraph we list the datasets’ URLs. (Kiesel et al. 2019): <https://zenodo.org/records/1489920>; the VIStA-H dataset (Lyu et al. 2023): <https://github.com/VIStA-H/Hyperpartisan-News-Titles/blob/main>; the Spanish Fake News Corpus (Gómez-Adorno et al. 2021): <https://github.com/jpposadas/FakeNewsCorpusSpanish>; the Fake News Net dataset (Shu et al. 2017): <https://github.com/KaiDMML/FakeNewsNet/tree/master/dataset>; the Fake.br Corpus (Monteiro et al. 2018): <https://github.com/roneysco/Fake.br-Corpus/blob/master/preprocessed>; CLEF 2022 CheckThat! Lab Subtask 1C (Nakov et al. 2022): https://gitlab.com/checkthat_lab/clef2022-checkthat-lab; Qbias dataset (Haak and Schaer 2023) <https://github.com/irgroup/Qbias>; CLEF 2023 CheckThat! Lab Task 3A dataset (Azizov and Nakov 2023) https://gitlab.com/checkthat_lab/clef2023-checkthat-lab.

Computational Infrastructure and Hyperparameters

Our computing infrastructure included two Tesla P40 GPUs, one NVIDIA GeForce RTX 2080 Ti GPU and a single A100 80GB SXM GPU, which was part of the Austral supercomputer of the CRIANN (Centre Régional Informatique et d’Applications Numériques de Normandie). Each experiment was run on a single GPU.

Hyperparameter	Value
Learning rate	1×10^{-4}
Epochs	3
Runs	5
LoRA target modules	query, value
LoRA Rank	8
LoRA Alpha	16
LoRA dropout	0.1
Weight decay	0.001
Max grad norm	0.3
Warmup ratio	0.1

Table 3: Hyperparameters for Fine-Tuning experiments.

Prompts

We employed four distinct prompt configurations in our experiments: (1) zero-shot prompts with a generic task definition, relying on the model’s internal knowledge; (2) zero-shot prompts with a task definition crafted by a political science expert; (3) task-specific prompts using structured codebooks; and (4) task-specific Chain-of-Thought (CoT) prompts.

For all configurations except the codebook-based prompts, we conducted preliminary experiments using the same structure but varying the wording and the label formatting. Based on these findings, we revised the label

templates for the current study—most notably by switching from integer-based labels to string-based labels and adopting a more structured output format using special symbols (e.g., ==>) accompanied by a clear instruction to follow the format.

This refinement led to a substantial improvement in output consistency, reducing the number of unparsable labels by approximately 99%.

Prompt examples

Table 4: Table showing different examples of instruction for PL Detection.

Prompt	Text
zero shot generic	<p>"You are an expert in political domain. You will be provided with a text, and your task is to determine, to the best of your judgment, whether the text is left-, right-wing or neutral. If you are unsure, make the most informed decision you can based on the content. Determine the political leaning of the headline. The possible choices are: 'left' if the article is left-wing, 'center' if the political leaning is from the center, 'right' if the article is right-wing'. Make sure to structure your answer in the following manner: \n\nFinal Answer ==>left\n\nnor\n\nFinal Answer ==>center\n\nFinal Answer ==>right\n\n Here's the text: "</p>
zero shot generic	<p>"You are an expert in political domain. You will be provided with a text, and your task is to determine, to the best of your judgment, whether the text is left-, right-wing or neutral. If you are unsure, make the most informed decision you can based on the content. Determine the political leaning of the headline. The possible choices are: 'left' if the article is left-wing, 'center' if the political leaning is from the center, 'right' if the article is right-wing'. Left-wing ideologies and moral values often include support for social equality, environmentalism, social justice, labor rights, and progressive taxation. Center ideologies and moral values often focus on a balance between market and social justice, pragmatism, and incremental reform. Right-wing ideologies and moral values often emphasize tradition, free market capitalism, individualism, and a limited role of government. Detect the political leaning of the headline. Determine the political leaning of the headline. The possible choices are: 'left' if the article is left-wing, 'center' if the political leaning is from the center, 'right' if the article is right-wing'. Make sure to structure your answer in the following manner: \n\nFinal Answer ==>left\n\nnor\n\nFinal Answer ==>center\n\nFinal Answer ==>right. \n\n Here's the text: ",</p>
few shot	<p>"You are an expert in political domain. You will be provided with a text and some labeled examples, and your task is to determine, to the best of your judgment, whether the text is left-, right-wing or neutral. If you are unsure, make the most informed decision you can based on the content. Determine the political leaning of the headline. The possible choices are: 'left' if the article is left-wing, 'center' if the political leaning is from the center, 'right' if the article is right-wing'. Make sure to structure your answer in the following manner: \n\nFinal Answer ==>left\n\nnor\n\nFinal Answer ==>center\n\nFinal Answer ==>right\n\n Here's the text: ",</p>
codebook	<p>"You are an expert in political domain. You will be provided with a codebook to follow and a text, and your task is to determine, to the best of your judgment, whether the text is left-, right-wing or neutral. If you are unsure, make the most informed decision you can based on the content. Determine the political leaning of the headline. *Definition: Left-wing ideologies and moral values often include support for social equality, environmentalism, social justice, labor rights, and progressive taxation. Center ideologies and moral values often focus on a balance between market and social justice, pragmatism, and incremental reform. Right-wing ideologies and moral values often emphasize tradition, free market capitalism, individualism, and a limited role of government. *Task: The possible choices are: 'left' if the article is left-wing, 'center' if the political leaning is from the center, 'right' if the article is right-wing'. Make sure to structure your answer in the following manner: \n\nFinal Answer ==>[left or center or right] \n\n Here's the text: ",</p>
cot	<p>"You are an expert in political domain. You will be provided with a text, and your task is to determine, to the best of your judgment, whether the text is left-, right-wing or neutral. If you are unsure, make the most informed decision you can based on the content. Determine the political leaning of the headline. The possible choices are: 'left' if the article is left-wing, 'center' if the political leaning is from the center, 'right' if the article is right-wing'. Make sure to structure your answer in the following manner: \n\nFinal Answer ==>[left or center or right] Reason step by step: 1. Identify key words and phrases: list of words or phrases that might indicate political bias. 2. Sentiment analysis: Analyze the tone and the language to see if there are polarizing words, or emotional language. 3. Are there any rhetorical devices used? (e.g., hyperbole, loaded questions). 4. Look for policy implications: Does the headline suggest support for or opposition to specific policies? Are there any implicit assumptions about political or economic systems? 5. Identify target audience: Who seems to be the intended audience for this headline? 6. Left-wing hyperpartisan: Consider you are a left-wing reader. Would you consider this article as hyperpartisan to your political stance? [Follow the instructions from 1 to 5.] 7. Right-wing hyperpartisan: Consider you are a right-wing reader. Would you consider this article as hyperpartisan to your political stance? [Follow the instructions from 1 to 5.] 8. Consider omissions and emphasis: What information is emphasized, and what might be omitted? How might this emphasis or omission indicate bias? 9. Final prediction. The output will be in this format. Remember to use all reasoning steps followed by <end> to write a concise explanation: Example output: ** Key words: EXPLANATION <end>** Sentiment analysis: EXPLANATION <end>** Policy implications: EXPLANATION <end>* Target audience: EXPLANATION <end>** Left-wing hyperpartisan: EXPLANATION <end>* Right-wing hyperpartisan: EXPLANATION <end>** Omissions and emphasis: EXPLANATION <end>*Final Answer ==>left\n\nnor\n\nFinal Answer ==>center\n\nFinal Answer ==>right\n\n Here's the text: "</p>

Table 5: Table showing different examples of instruction for HF Detection.

Prompt	Text
zero shot generic	"You are a fake news expert. You will be provided with a text, and your task is to determine, to the best of your judgment, whether the text is harmful to society or not. If you are unsure, make the most informed decision you can based on the content. If the news is harmful, the label is 'harmful', if the text is neutral, the label is 'neutral'. Make sure to structure your answer in the following manner: \n\n Final Answer ==>harmful\n\n nor\n\n Final Answer ==>neutral\n\n Here's the text: ",
zero shot generic	"You are a fake news expert. Harmful tweet is a post on the social media platform Twitter that can cause distress, harm, or damage to individuals, groups, or society. This type of tweet may include content that is abusive, offensive, defamatory, threatening, or incites violence against the policies to contain the spread of COVID-19. Additionally, it may spread false information, contribute to cyberbullying, or perpetuate hate speech and discrimination. Harmful tweets can negatively impact mental health, reputations, and community safety. You will be provided with labeled examples and a text, and your task is to determine, to the best of your judgment, whether the text is harmful to society or not. If you are unsure, make the most informed decision you can based on the content. If the news is harmful, the label is 'harmful', if the text is a neutral, the label is 'neutral'. Make sure to structure your answer in the following manner: \n\n Final Answer ==>harmful\n\n nor\n\n Final Answer ==>neutral\n\n Here's the text: ",
few shot	"You are a fake news expert. You will be provided with labeled examples and a text, and your task is to determine, to the best of your judgment, whether the text is harmful to society or not. If you are unsure, make the most informed decision you can based on the content. If the news is harmful, the label is 'harmful', if the text is neutral, the label is 'neutral'. Make sure to structure your answer in the following manner: \n\n Final Answer ==>harmful\n\n nor\n\n Final Answer ==>neutral\n\n Here's the text: ",
codebook	"You are a fake news expert. **Definition: Harmful tweet is a post on the social media platform Twitter that can cause distress, harm, or damage to individuals, groups, or society. This type of tweet may include content that is abusive, offensive, defamatory, threatening, or incites violence against the policies to contain the spread of COVID-19. Additionally, it may spread false information, contribute to cyberbullying, or perpetuate hate speech and discrimination. Harmful tweets can negatively impact mental health, reputations, and community safety. **Task: You will be provided with a codebook and a text, and your task is to determine, to the best of your judgment, whether the text is harmful to society or not. If you are unsure, make the most informed decision you can based on the content. If the news is harmful, the label is 'harmful', if the text is a neutral, the label is 'neutral'. Make sure to structure your answer in the following manner: \n\n Final Answer ==>harmful\n\n nor\n\n Final Answer ==>neutral\n\n Here's the text: \n\n This is the codebook: — Label — Class — Definition — Typical Content Patterns — Neutral — Neutral — The tweet provides factual, supportive, or neutral information about COVID-19. — - Promotes vaccinations, masking, or safety — — Shares news from reliable sources (WHO, CDC) — — Describes personal experiences without disinformation — — Harmful — Harmful — The tweet contains disinformation, conspiracy theories, or medically incorrect claims. — - Claims vaccines are deadly or part of a conspiracy — — Encourages rejection of public health guidance — — Spreads pseudoscience or political hoaxes — — Aspect — Description of Harmful Characteristics Conspiracy Theory — References to population control, "pandemic", 5G causes COVID, microchips in vaccines — — Distrust in Science — Discredits CDC/WHO, claims that mainstream medicine is false or corrupt — — Alternative Cures — Promotes ivermectin, bleach, natural remedies as vaccine alternatives without evidence — — Anti-Vaccine — Claims vaccines are unsafe, rushed, or part of a control agenda — — Hoax Narrative — Declares COVID-19 is fake, exaggerated, or invented by governments — — Call to Action — Urges followers to ignore mandates, avoid vaccination, or resist masking — — Language Style — Emotionally charged, sarcastic, or mocking tone; often uses ALL CAPS, emojis, hashtags like **NoVax **Plandemic — — Linguistic Features — Harmful — Not Harmful — Use of hashtags — — plandemic, **NoVax, **scamdem — — GetVaccinated, **StaySafe — — Source credibility — — Anonymous sources, influencer claims — WHO, CDC, news outlets — — Personal stories — Claims of side effects or hoaxes — Stories of recovery, gratitude for vaccines — — Tone — Alarmist, rebellious, sarcastic — — Informative, cautious, or appreciative — — \n\n Make sure to structure your answer in the following manner: \n\n Final Answer ==>harmful\n\n nor\n\n Final Answer ==>neutral\n\n Here's the text: ",
cot	"You are a fake news expert. You will be provided with labeled examples and a text, and your task is to determine, to the best of your judgment, whether the text is harmful to society or not. If you are unsure, make the most informed decision you can based on the content. Analyze the text for fake news using step-by-step reasoning. If the news is neutral, the label is 'neutral', if the text is harmful, the label is 'harmful'. You have to choose one label and the output is only one label. Make sure to structure your answer in the following manner: Explanation: [Explanation for each reasoning step]. \n\n Final Answer ==>harmful — — or — — Final Answer ==>neutral. — — ***** Reason step by step: — — 1. *Lexical Analysis**\n\n - Check for vague sourcing, absolutist language, logical fallacies — — . **Sentiment Analysis**\n\n - Identify polarizing/emotional language — — 3. **Content Authenticity**\n\n - Detect bot-like patterns, verify data — — 4. **Framing Bias**\n\n - Assess selective presentation — — 5. Audience Analysis**\n\n - Determine target audience and intent — — 6. **Information Adequacy**\n\n - Evaluate comprehensiveness and reliability. \n\n Make sure to structure your answer in the following manner:Explanation: [Explanation for each reasoning step]. Final Answer ==>harmful\n\n nor\n\n Final Answer ==>neutral. \n\n Here's the text: "

Table 5 – continued from previous page

Prompt	Text
--------	------

Table 5: Table showing different examples of instruction for HF Detection. For the multilingual dataset, we translated this prompt to Arabic and Bulgarian.

Table 6: Table showing different examples of instruction for HP Detection.

Prompt	Text
zero shot generic	"You are a hyperpartisan news expert. You will be provided with a text, and your task is to determine, to the best of your judgment, whether the text is hyperpartisan or not. If you are unsure, make the most informed decision you can based on the content. The possible choices are: 'neutral' if the article is neutral, 'hyperpartisan' if the article is hyperpartisan. You have to choose one label and the output is only one label. \n\n Make sure to structure your answer in the following manner: \n\n Final Answer ==>hyperpartisan\n\nnor \n\nFinal Answer ==>neutral \n\n This is the text: "
"zero shot generic"	"You are a hyperpartisan news expert. You will be provided with a text, and your task is to determine, to the best of your judgment, whether the text is hyperpartisan or not. If you are unsure, make the most informed decision you can based on the content. Hyperpartisan articles contain biases, particularly ad hominem attack, loaded language, and evidence of political ideology. Sometimes they rely on cherry-picking strategy. The possible choices are: 'neutral' if the article is neutral, 'hyperpartisan' if the article is hyperpartisan. You have to choose one label and the output is only one label. Make sure to structure your answer in the following manner: \n\n Final Answer ==>hyperpartisan\n\nnor \n\nFinal Answer ==>neutral \n\n This is the text: "
few shot	"You are a hyperpartisan news expert. You will be provided with a text, and your task is to determine, to the best of your judgment, whether the text is hyperpartisan or not. If you are unsure, make the most informed decision you can based on the content. You will be provided with some examples of labeled text. The possible choices are: 'neutral' if the article is neutral, 'hyperpartisan' if the article is hyperpartisan. You have to choose one label and the output is only one label. Make sure to structure your answer in the following manner: \n\n Final Answer ==>hyperpartisan\n\nnor \n\nFinal Answer ==>neutral \n\n*****Examples of labelled articles: \n\n This is the text: "
codebook	"You are a hyperpartisan news expert. You will be provided with a text, and your task is to determine, to the best of your judgment, whether the text is hyperpartisan or not. If you are unsure, make the most informed decision you can based on the content. You will be provided with some examples of labeled text. \n\n Definition: Hyperpartisan news detection is the process of identifying news articles that exhibit extreme one-sidedness, characterized by a pronounced use of bias. The prefix hyper- highlights the exaggerated application of at least one specific type of bias—such as spin, ad hominem attacks, opinionated statements, ideological slants, framing, selective coverage, political leaning, or slant bias—to promote a particular ideological perspective. This strong ideological alignment is conveyed through amplified linguistic elements that reinforce one of these bias types within the text. \n\n***** Task: Read carefully the codebook provided and assign a label to the text. You can choose only one label. If the text is neutral, you will write 'neutral', if it is hyperpartisan 'hyperpartisan'. \n\n Follow the output template given as an example. Under no circumstances we are asking to provide or generate harmful content. Please, provide only the label. \n\n***** Make sure to structure your answer in the following manner: \n\n Final Answer ==>hyperpartisan\n\nnor \n\nFinal Answer ==>neutral \n\n **This is the codebook: Linguistic Features: \n\n***** 1. Lexical Features — Feature — Hyperpartisan — Neutral — Lexical Polarity — Frequent use of emotionally charged words (e.g., disaster, outrageous) — Neutral and precise language — Modality and Certainty — Strong modal verbs (e.g., will destroy) — Hedging markers (e.g., may, might) — Repetition — Repeating claims or slogans — Minimal repetition — Pronouns — Frequent us vs them language — Focus on third-person objectivity — \n\n***** 2. Rhetorical Devices — Feature — Hyperpartisan — Neutral — Appeal to Emotion — Frequent appeals to fear/anger — Logical/factual appeal — Hyperbole — Common exaggeration — Proportional statements — Metaphors — Politically loaded metaphors — Literal language preferred — \n\n***** 3. Discourse Structure — Feature — Hyperpartisan — Neutral — Framing — Blame/conflict framing — Balanced framing — Source Attribution — Partisan sources only — Multiple reputable sources — Balance of Views — One-sided presentation — Multiple perspectives — \n\n***** 4. Ideological Markers — Feature — Hyperpartisan — Neutral — Us vs Them — Strong binary division — Avoids binaries — Ideological Alignment — Clear left/right alignment — Issue-focused — \n\n***** 5. Pragmatic Features — Feature — Hyperpartisan — Neutral — Intent — Persuade/convert — Inform/explain — Tone — Confrontational/accusatory — Formal/detached — \n\n***** \n\n Final Answer ==>hyperpartisan\n\nnor \n\nFinal Answer ==>neutral \n\n This is the text: "

Table 6 – continued from previous page

Prompt	Text
cot	<p>”You are a hyperpartisan news expert. You will be provided with a text, and your task is to determine, to the best of your judgment, whether the text is hyperpartisan or not. If you are unsure, make the most informed decision you can based on the content. You have to choose one label and the output will be the explanation and the determined label for that article. Make sure to structure your answer in the following manner:\n\n Explanation: [Explanation for each reasoning step]. \n\n Final Answer ==>hyperpartisan\n\nor \n\nFinal Answer ==>neutral. \n\n Reason step by step:\n\n1. **Lexical and Sentiment analysis**: Analyze the tone and language. Does the article use polarizing, emotionally charged, or exaggerated language? Are there strong positive/negative sentiments toward a group, ideology, or issue?\n\n2. **Rhetorical bias**: Does the text use loaded language, name-calling, or manipulative rhetoric? Are there oversimplifications, strawman arguments, or exaggerated claims?\n\n3. **Framing bias**: Is information presented selectively to favor one perspective? Does it emphasize certain aspects while downplaying others to shape perception?\n\n4. **Ideological bias**: Does it emphasize certain aspects while downplaying others to shape perception? Does the article vilify opposing views rather than engaging with them fairly?\n\n5. **Unilateral coverage**: Does the article present multiple viewpoints (neutral) or only one side (hyperpartisan)? Are opposing arguments ignored, misrepresented, or dismissed?\n\n6. **Intent and Purpose**: Is the primary goal to inform objectively (neutral) or to persuade/mislead (hyperpartisan)? Does it present facts fairly, or does it push a clear agenda?\n\n7. Final prediction: Based on your previous considerations, classify the input as:\n\n- ‘neutral’ (Neutral): Balanced, factual, and objective.\n\n- ‘hyperpartisan’: Biased, one-sided, or manipulative.\n\n While generating the explanation for each reasoning step be concise.. \n\n Remember to follow the output template for the label. \n\n Final Answer ==>hyperpartisan\n\nor \n\nFinal Answer ==>neutral\n\n This is the text: ”</p>

Table 6: Table showing different examples of instruction for HP Detection.

Table 7: Table showing different examples of instruction for FN Detection.

Prompt	Text
zero shot generic	"You are a fake news expert. You will be provided with a text, and your task is to determine, to the best of your judgment, whether the text is fake news or not. If you are unsure, make the most informed decision you can based on the content. If the news is true, the label is 'true', if the text is a fake news, the label is 'fake'. Make sure to structure your answer in the following manner:\n\n Final Answer ==>fake\n\nnor\n\nFinal Answer ==>true\n\nHere's the text: "
zero shot generic	"You are a fake news expert. We define fake news as: purposefully crafted, sensational, emotionally charged, misleading or totally fabricated information that mimics mainstream news. You will be provided with a text and your task is to determine, to the best of your judgment, whether the text is fake news or not. If you are unsure, make the most informed decision you can based on the content. If the news is true, the label is 'true', if the text is a fake news, the label is 'fake'. Make sure to structure your answer in the following manner:\n\n Final Answer ==>fake\n\nnor\n\nFinal Answer ==>true\n\nHere's the text: "
few shot	"You are a fake news expert. You will be provided with a text and labeled examples, and your task is to determine, to the best of your judgment, whether the text is fake news or not. If you are unsure, make the most informed decision you can based on the content. If the news is true, the label is 'true', if the text is a fake news, the label is 'fake'. Make sure to structure your answer in the following manner:\n\n Final Answer ==>fake\n\nnor\n\nFinal Answer ==>true\n\nHere's the text: ",
codebook	"You are a fake news expert. You will be provided with a text, and your task is to determine, to the best of your judgment, whether the text is fake news or not. If you are unsure, make the most informed decision you can based on the content. **Definition: Fake news detection identifies intentionally misleading content characterized by sensationalism, lack of credible sources, or manipulative language.\n\nIf the news is true, the label is 'true', if the text is a fake news, the label is 'fake'. \n\n***** Task: To structure the output, follow the template in the example. \n\n Make sure to structure your answer in the following manner:\n\n Final Answer ==>fake\n\nnor\n\nFinal Answer ==>true. This is the codebook you have to use to perform the classification: \n\n***** Detection Criteria:\n1. **Source Origin**\n- Real: Established media, government sources\n\n- Fake: Anonymous/unverifiable sources\n2. **Event Reporting**\n- Real: Specific references, quantitative data\n\n- Fake: Broad generalizations, unverifiable claims\n3. **Language Style**\n- Real: Neutral, professional\n\n- Fake: Sensationalized, clickbait\n4. **Entity Authenticity**\n- Real: Named real-world entities\n\n- Fake: Fictional/misspelled names\n5. **Claim Reliability**\n- Real: Evidence-based\n\n- Fake: Absurd/absolute claims\n6. **Emotional Tone**\n- Real: Objective\n\n- Fake: Emotional intensifiers\n7. **Source Credibility**\n- Real: Verifiable\n\n- Fake: Unknown/misleading domains\n8. **Political Balance**\n- Real: Multi-perspective\n\n- Fake: One-sided\n9. **Satire Markers**\n- Real: No satire\n\n- Fake: Absurd content\n10. **Conspiracy Indicators**\n- Real: Supported theories\n\n- Fake: Fringe conspiracy phrases\n\n Make sure to structure your answer in the following manner:\n\n Final Answer ==>fake\n\nnor\n\nFinal Answer ==>true\n\nHere's the text: ",
cot	"You are a fake news expert. You will be provided with a text, and your task is to determine, to the best of your judgment, whether the text is fake news or not. If you are unsure, make the most informed decision you can based on the content. Analyze the text for fake news using step-by-step reasoning. If the news is true, the label is 'true', if the text is a fake news, the label is 'fake'. You have to choose one label and the output is only one label. To structure the output, follow the template in the example. \n\nOutput example: Explanation: [Explanation for each reasoning step]. Final Prediction: Final Answer ==>[true or fake]. \n\n***** Reasoning step by step:\n1. **Lexical Analysis**\n- Check for vague sourcing, absolutist language, logical fallacies\n2. **Sentiment Analysis**\n- Identify polarizing/emotional language\n3. **Content Authenticity**\n- Detect bot-like patterns, verify data\n4. **Framing Bias**\n- Assess selective presentation\n5. Audience Analysis\n- Determine target audience and intent\n6. **Information Adequacy**\n- Evaluate comprehensiveness and reliability.\n\n Make sure to structure your answer in the following manner:\n\n Final Answer ==>fake\n\nnor\n\nFinal Answer ==>true\n\nHere's the text: "

Table 7: Table showing different examples of instruction for FN Detection. For the multilingual dataset, we translated this prompt to Spanish and Portuguese.

Results and SOTA

	SH	Hyperpartisan Det.			Political Lending Det.			CIA			FBI			CIE			Hamilt. Test Det.			Fake News Det.				
		F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	
General																								
Laura 3.189-Instance	0.678	0.701	0.742	0.739	0.437	0.426	0.425	0.360	0.720	0.689	0.722	0.679	0.307	0.372	0.364	0.329	0.240	0.256	0.208	0.210	0.145	0.145		
Mitral-Name-Instance-2407	0.686	0.707	0.667	0.623	0.422	0.404	0.422	0.399	0.783	0.761	0.677	0.565	0.516	0.793	0.695	0.551	0.542	0.328	0.265	0.327	0.144	0.469	0.335	
Quora2.5-TB-Instance	0.780	0.782	0.627	0.621	0.406	0.402	0.406	0.402	0.691	0.690	0.624	0.626	0.621	0.643	0.755	0.643	0.521	0.542	0.321	0.265	0.327	0.144	0.469	0.335
Specific																								
Laura 3.189-Instance	0.738	0.748	0.764	0.753	0.431	0.419	0.419	0.356	0.719	0.677	0.721	0.678	0.613	0.732	0.642	0.571	0.540	0.315	0.354	0.326	0.182	0.182	0.150	
Mitral-Name-Instance-2407	0.740	0.750	0.574	0.503	0.409	0.398	0.408	0.352	0.719	0.677	0.721	0.678	0.618	0.738	0.642	0.570	0.540	0.315	0.354	0.326	0.182	0.182	0.150	
Quora2.5-TB-Instance	0.789	0.885	0.611	0.535	0.408	0.391	0.408	0.352	0.721	0.685	0.727	0.682	0.618	0.738	0.642	0.570	0.540	0.315	0.354	0.326	0.182	0.182	0.150	
Codebook																								
Laura 3.189-Instance	0.748	0.753	0.715	0.722	0.382	0.387	0.387	0.318	0.753	0.729	0.728	0.696	0.670	0.538	0.738	0.588	0.549	0.304	0.405	0.326	0.165	0.166	0.145	
Mitral-Name-Instance-2407	0.694	0.713	0.706	0.704	0.372	0.385	0.387	0.311	0.753	0.729	0.728	0.696	0.670	0.538	0.738	0.588	0.549	0.304	0.405	0.326	0.165	0.166	0.145	
Quora2.5-TB-Instance	0.810	0.811	0.656	0.688	0.377	0.395	0.395	0.311	0.753	0.729	0.728	0.696	0.670	0.538	0.738	0.588	0.549	0.304	0.405	0.326	0.165	0.166	0.145	
Random 1-4shot																								
Laura 3.189-Instance	0.751	0.807	0.755	0.804	0.476	0.477	0.476	0.406	0.711	0.701	0.708	0.684	0.606	0.745	0.636	0.708	0.644	0.589	0.407	0.446	0.407	0.184	0.186	
Mitral-Name-Instance-2407	0.716	0.817	0.729	0.802	0.406	0.401	0.401	0.390	0.713	0.701	0.708	0.684	0.606	0.745	0.636	0.708	0.644	0.589	0.407	0.446	0.407	0.184	0.186	
Quora2.5-TB-Instance	0.718	0.804	0.730	0.805	0.406	0.402	0.402	0.396	0.714	0.702	0.709	0.685	0.607	0.746	0.637	0.710	0.645	0.590	0.408	0.447	0.408	0.185	0.187	
Random 2-4shot																								
Laura 3.189-Instance	0.664	0.713	0.688	0.703	0.431	0.430	0.430	0.369	0.641	0.609	0.640	0.611	0.407	0.612	0.606	0.640	0.606	0.593	0.408	0.447	0.408	0.179	0.181	
Mitral-Name-Instance-2407	0.661	0.706	0.671	0.704	0.430	0.429	0.429	0.368	0.642	0.609	0.641	0.611	0.407	0.612	0.606	0.641	0.606	0.593	0.408	0.447	0.408	0.179	0.181	
Quora2.5-TB-Instance	0.788	0.818	0.735	0.807	0.430	0.429	0.429	0.368	0.643	0.609	0.642	0.611	0.407	0.613	0.606	0.642	0.606	0.593	0.408	0.447	0.408	0.179	0.181	
Random 3-4shot																								
Laura 3.189-Instance	0.725	0.813	0.756	0.805	0.470	0.470	0.470	0.406	0.711	0.701	0.708	0.684	0.606	0.745	0.636	0.708	0.644	0.589	0.408	0.447	0.408	0.184	0.186	
Mitral-Name-Instance-2407	0.725	0.812	0.755	0.804	0.470	0.470	0.470	0.406	0.711	0.701	0.708	0.684	0.606	0.745	0.636	0.708	0.644	0.589	0.408	0.447	0.408	0.184	0.186	
Quora2.5-TB-Instance	0.777	0.814	0.788	0.803	0.470	0.470	0.470	0.406	0.712	0.701	0.708	0.684	0.606	0.745	0.636	0.708	0.644	0.589	0.408	0.447	0.408	0.184	0.186	
Random 4-4shot																								
Laura 3.189-Instance	0.623	0.707	0.586	0.628	0.406	0.402	0.402	0.322	0.647	0.602	0.646	0.601	0.421	0.612	0.606	0.646	0.606	0.593	0.408	0.447	0.408	0.173	0.183	
Mitral-Name-Instance-2407	0.741	0.745	0.642	0.654	0.406	0.402	0.402	0.322	0.648	0.602	0.647	0.601	0.421	0.612	0.606	0.647	0.606	0.593	0.408	0.447	0.408	0.173	0.183	
Quora2.5-TB-Instance	0.779	0.809	0.788	0.807	0.406	0.402	0.402	0.322	0.649	0.602	0.648	0.601	0.421	0.612	0.606	0.648	0.606	0.593	0.408	0.447	0.408	0.173	0.183	
Random 5-4shot																								
Laura 3.189-Instance	0.725	0.802	0.756	0.801	0.470	0.470	0.470	0.406	0.711	0.701	0.708	0.684	0.606	0.745	0.636	0.708	0.644	0.589	0.408	0.447	0.408	0.184	0.186	
Mitral-Name-Instance-2407	0.725	0.801	0.755	0.800	0.470	0.470	0.470	0.406	0.711	0.701	0.708	0.684	0.606	0.745	0.636	0.708	0.644	0.589	0.408	0.447	0.408	0.184	0.186	
Quora2.5-TB-Instance	0.788	0.814	0.788	0.803	0.470	0.470	0.470	0.406	0.712	0.701	0.708	0.684	0.606	0.745	0.636	0.708	0.644	0.589	0.408	0.447	0.408	0.184	0.186	
Random 6-4shot																								
Laura 3.189-Instance	0.619	0.701	0.657	0.705	0.406	0.402	0.402	0.322	0.649	0.602	0.647	0.601	0.406	0.612	0.606	0.647	0.606	0.589	0.408	0.447	0.408	0.173	0.183	
Mitral-Name-Instance-2407	0.754	0.810	0.780	0.806	0.470	0.470	0.470	0.406	0.712	0.701	0.708	0.684	0.606	0.745	0.636	0.708	0.644	0.589	0.408	0.447	0.408	0.184	0.186	
Quora2.5-TB-Instance	0.786	0.814	0.788	0.803	0.470	0.470	0.470	0.406	0.713	0.701	0.708	0.684	0.606	0.745	0.636	0.708	0.644	0.589	0.408	0.447	0.408	0.184	0.186	
Random 7-4shot																								
Laura 3.189-Instance	0.622	0.707	0.660	0.704	0.406	0.402	0.402	0.322	0.650	0.602	0.648	0.601	0.406	0.612	0.606	0.648	0.606	0.590	0.408	0.447	0.408	0.173	0.183	
Mitral-Name-Instance-2407	0.625	0.709	0.663	0.705	0.406	0.402	0.402	0.322	0.651	0.602	0.649	0.601	0.406	0.612	0.606	0.649	0.606	0.590	0.408	0.447	0.408	0.173	0.183	
Quora2.5-TB-Instance	0.789	0.809	0.794	0.809	0.470	0.470	0.470	0.406	0.714	0.701	0.708	0.684	0.606	0.745	0.636	0.708	0.644	0.589	0.408	0.447	0.408	0.184	0.186	
Random 8-4shot																								
Laura 3.189-Instance	0.590	0.689	0.639	0.702	0.406	0.402	0.402	0.322	0.651	0.602	0.647	0.601	0.406	0.612	0.606	0.647	0.606	0.581	0.408	0.447	0.408	0.173	0.183	
Mitral-Name-Instance-2407	0.700	0.800	0.781	0.801	0.470	0.470	0.470	0.406	0.715	0.701	0.708	0.684	0.606	0.745	0.636	0.708	0.644	0.589	0.408	0.447	0.408	0.184	0.186	
Quora2.5-TB-Instance	0.783	0.800	0.781	0.801	0.470	0.470	0.470	0.406	0.716	0.701	0.708	0.684	0.606	0.745	0.636	0.708	0.644	0.589	0.408	0.447	0.408	0.184	0.186	
Random 9-4shot																								
Laura 3.189-Instance	0.631	0.706	0.665	0.704	0.406	0.402	0.402	0.322	0.652	0.602	0.648	0.601	0.406	0.612	0.606	0.648	0.606	0.582	0.408	0.447	0.408	0.173	0.183	
Mitral-Name-Instance-2407	0.760	0.803	0.782	0.802	0.470	0.470	0.470	0.406	0.717	0.701	0.708	0.684	0.606	0.745	0.636	0.708	0.644	0.589	0.408	0.447	0.408	0.184	0.186	
Quora2.5-TB-Instance	0.780	0.803	0.782	0.802	0.470	0.470	0.470	0.406	0.718	0.701	0.708	0.684	0.606	0.745	0.636	0.708	0.644	0.589	0.408	0.447	0.408	0.184	0.186	
Random 10-4shot																								
Laura 3.189-Instance	0.619	0.705	0.664	0.703	0.406	0.402	0.402	0.322	0.653	0.602	0.648	0												

Dataset	Reference	Model	Performance		Our Best Model	Configuration	Performance	
			Acc	F1			Acc	F1
ViStA-H	(Lyu et al. 2023)	BERT-base	0.84	0.78	RoBERTa-large	FT	.852	.850
Fake News Net	(Jin et al. 2022)	Graph-based Reasoning	0.870	0.892	Llama3.1-8b	FT	.945	.945
Spanish Fake News Corpus	(Gómez-Adorno et al. 2021)	BERT	0.766	N/A	Modern-BERT-large	FT	.863	.863
Fake.br	(Monteiro et al. 2018)	SVM	0.89	0.89	Llama3-8b-Instruct	FT	.979	.979

Table 9: SOTA results.