



This is a repository copy of *Detecting misinformation with LLM-predicted credibility signals and weak supervision.*

White Rose Research Online URL for this paper:
<https://eprints.whiterose.ac.uk/223248/>

Version: Preprint

Preprint:

Leite, J.A., Razuvayevskaya, O., Bontcheva, K. orcid.org/0000-0001-6152-9600 et al. (1 more author) (Submitted: 2023) Detecting misinformation with LLM-predicted credibility signals and weak supervision. [Preprint - arXiv] (Submitted)

<https://doi.org/10.48550/arXiv.2309.07601>

© 2023 The Author(s). For reuse permissions, please contact the Author(s).

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>

Detecting Misinformation with LLM-Predicted Credibility Signals and Weak Supervision

João A. Leite, Olesya Razuvayevskaya, Kalina Bontcheva and Carolina Scarton

Department of Computer Science, The University of Sheffield, Sheffield, UK
{jaleite1, o.razuvayevskaya, k.bontcheva, c.scarton}@sheffield.ac.uk

Abstract

Credibility signals represent a wide range of heuristics that are typically used by journalists and fact-checkers to assess the veracity of online content. Automating the task of credibility signal extraction, however, is very challenging as it requires high-accuracy signal-specific extractors to be trained, while there are currently no sufficiently large datasets annotated with all credibility signals. This paper investigates whether large language models (LLMs) can be prompted effectively with a set of 18 credibility signals to produce weak labels for each signal. We then aggregate these potentially noisy labels using weak supervision in order to predict content veracity. We demonstrate that our approach, which combines zero-shot LLM credibility signal labeling and weak supervision, outperforms state-of-the-art classifiers on two misinformation datasets without using any ground-truth labels for training. We also analyse the contribution of the individual credibility signals towards predicting content veracity, which provides new valuable insights into their role in misinformation detection.

1 Introduction

In the era of rapidly spreading mis- and disinformation, the task of its automatic detection has emerged as a prominent area of NLP research, with many approaches being proposed recently (Zhou and Zafarani, 2020). Nevertheless, a number of limitations and challenges still need to be addressed.

Firstly, state-of-the-art data-driven supervised methods rely heavily on high-quality manually annotated datasets. However, the creation of such datasets is time-consuming, and the ever-evolving nature of disinformation requires the continuous development of new datasets (Fu et al., 2023; Ksieniewicz et al., 2020; Silva and Almeida, 2021). Secondly, debunking an article often requires looking beyond its content, as articles containing misinformation are often intentionally crafted to appear

credible (Zhang and Ghorbani, 2020). Although evidence-aware models have been developed to incorporate external information that can assist in the debunking process (Vlachos and Riedel, 2015; Popat et al., 2018; Xu et al., 2022), automatic extraction of meaningful evidences is a challenge in itself (Rinott et al., 2015; Thorne and Vlachos, 2018). Lastly, while there are efforts aimed at designing approaches that can explain a model’s decision (Shu et al., 2019; Lu and Li, 2020; Kotonya and Toni, 2020), most methods differ from the process carried out by journalists in that their primary focus is on detecting misinformation without providing evidence or comprehensive explanations to substantiate their decisions.

Instruction-tuned large language models (LLMs) offer promising opportunities to address the aforementioned challenges. While further research is required to fully understand their potential and limitations, instruction-tuned LLMs achieve remarkable performance in various NLP tasks, including common sense reasoning, reading comprehension, and closed-book question answering (Touvron et al., 2023), often surpassing state-of-the-art supervised approaches (Brown et al., 2020). Petroni et al. (2019) show that LLMs have a surprisingly strong ability to recall factual knowledge without any fine-tuning. This suggests that the external knowledge acquired during pre-training can be leveraged to support automated veracity classification, which is a key part of debunking online misinformation. Moreover, by framing this problem as a question-answering (QA) task, instruction-tuned LLMs can offer justifications for their decisions, thus enhancing the transparency and explainability of the content verification process.

However, since instruction-tuned LLMs are prone to generating inaccurate yet convincing answers, commonly known as “hallucinations” (Ji et al., 2023), this poses limitations on their use in assessing veracity directly, as it impacts the relia-

bility, transparency, and predictive performance of such approaches.

Instead, our approach is modelled on the verification process typically adopted by journalists and fact-checkers, who assess the veracity of online content using a wide range of credibility signals¹. Our novel contribution is in investigating whether large language models (LLMs) can be prompted effectively with a set of 18 credibility signals to produce weak labels for each signal. We then aggregate these potentially noisy labels using weak supervision in order to predict content veracity.

This multi-stage approach reduces LLMs’ susceptibility to “hallucinations”, since predicting individual signals (with appropriately tailored prompts for each) is simpler than relying on an LLM to predict content veracity. Additionally, the final prediction is less sensitive to the “hallucinations” in the intermediate steps due to the weighted aggregation of signals performed by weak supervision. Moreover, it enables human fact-checkers to audit model decisions and select which signals are used, thus providing greater control and transparency.

In particular, the paper addresses the following research questions: **(RQ1)** Is zero-shot prompting with instruction-tuned LLMs as effective as fine-tuning text classifiers with ground-truth (GT) data for the task of article-level veracity detection? **(RQ2)** Does prompted weak supervision with credibility signals outperform zero-shot prompting? **(RQ3)** Which credibility signals are the most useful for predicting veracity?

2 Related Work

Misinformation detection and credibility signals

Building models aimed at detecting misinformation relies on the availability of human annotated datasets. The majority of the benchmark corpora annotated for misinformation based on distinct categories focus on analysing short claims (Vlachos and Riedel, 2014; Ferreira and Vlachos, 2016; Wang, 2017; Thorne et al., 2018) or social media data, such as Facebook posts (Potthast et al., 2017; Santia and Williams, 2018; Tacchini et al., 2017) and Twitter threads (Zubiaga et al., 2016; Mitra and Gilbert, 2015). Fewer such datasets are available for the analysis of the trustworthiness of long articles from news outlets, which is the focus of this paper (Abu Salem et al., 2019; Shu et al., 2020b).

The FA-KES corpus (Abu Salem et al., 2019) comprises news articles on Syrian war from 15 different sources, covering the period from 2011 to 2018. It was first human-annotated for objective information such as dates, locations, and actors, that was compared against ground-truth facts obtained from the Syrian Violations Documentation Center. Unsupervised machine learning techniques were then employed to cluster the articles into either misinformation or non-misinformation ones.

Previous approaches using the FA-KES dataset for misinformation detection include Multinomial Naive Bayes (Elhadad et al., 2020), optimized convolutional neural network (OPCNN-FAKE) (Saleh et al., 2021), and a hybrid CNN-LSTM model (Nasir et al., 2021). The latter approach outperforms previous methods by leveraging CNN and LSTM models to capture local features and long-term dependencies, respectively.

The term *credibility signals* refers to a wide range of measurable heuristics which collectively help journalists assess the overall trustworthiness of information. Examples of credibility signals include the analysis of article titles (Horne and Adali, 2017), writing style (Afroz et al., 2012), rhetorical structure (Rashkin et al., 2017), linguistic features (O’Brien et al., 2018), emotional language (Giachanou et al., 2019), biases (Dufraisse et al., 2022), logical fallacies and inferences (Musi and Reed, 2022). Additionally, credibility signals comprise meta-information that extends beyond the textual content of the article, such as the author’s reputation and external references (Sitaula et al., 2020). Zhang et al. (2018) presented a set of credibility signals specifically for journalists, categorized into content-based (or meta-level) and context-based. W3C Credible Web Community Group (CWCG)² performed the most extensive attempt to date at cataloguing credibility signals, with more than 200 signals defined and documented.

The understanding of the significance of individual signals for misinformation detection is limited. Dimou et al. (2022) selected 23 contextual credibility signals defined by W3C CWCG and built a modular evaluation pipeline to assess the importance of each signal for web page credibility analysis. The authors found that morphological, syntactic and emotional features demonstrate the highest predictive capability for determining the credibility of web content.

¹For an overview see the W3C credibility signals discussed here: <https://www.w3.org/2018/10/credibility-tech/>

²<https://github.com/w3c/credweb>

Weak supervision for misinformation detection

Labelling large documents with many features, such as credibility signals, is a very costly process. Various approaches were proposed to address this challenge, such as semi-supervision (Tarvainen and Valpola, 2017), transfer learning (Zhuang et al., 2021), and distant supervision (Hoffmann et al., 2011). *Programmatic weak supervision* (PWS) (Ratner et al., 2016) aims to combine the aforementioned efforts by encoding potentially noisy probabilistic labels using *labeling functions*. To mitigate the noise from these weak signals, various frameworks aim to combine the outputs of several labeling functions into labels (Fu et al., 2020; Varma et al., 2019).

Helmstetter and Paulheim (2018) performed one of the pioneering works in applying weak supervision for misinformation detection on Twitter using the credibility signals associated with the content (e.g., bag-of-words, punctuation marks and sentiment) and context (e.g., user followers and retweet frequency). The authors found that despite being noisy, the source credibility labels are able to improve the final classification objective. Similarly, Shu et al. (2020a) apply weak social supervision for misinformation detection using information regarding users or their followers. In the follow-up work, Shu et al. (2020c) apply meta weighting on weak labels based on the user engagements with the news. Wang et al. (2020) apply reinforcement learning to leverage annotator reports for the selection of high-quality samples representing misinformation from a collection of weakly labelled news.

Prompted weak supervision The most relevant work to our approach is the study by Smith et al. (2022), which applies prompted PWS to spam detection and relation extraction tasks. However, in contrast to our prompts that require the model to perform more nuanced content analysis through reasoning and information retrieval capabilities, the authors use prompts for the task of string matching by translating regex patterns from the WRENCH benchmark into QA prompts. For instance, in order to identify the phrase 'check out' in a text, they prompt LLMs with the question: "Does the following comment contain the words 'check out'?". Furthermore, their datasets primarily consist of relatively short texts, such as descriptions extracted from YouTube or SMS messages. In contrast, our approach involves employing prompted PWS on long documents that provide more contextual infor-

mation that can be leveraged to answer the questions. Smith et al. employ two instruction-tuned LLM families: InstructGPT (Ouyang et al., 2022b) and T0++ (Sanh et al., 2021). Their results show that prompted PWS outperform the zero-shot baseline by an average of 18.2% across all models and datasets, with T0++ outperforming InstructGPT. Interestingly, the authors found that string matching is superior to prompted PWS for the spam detection task, while the opposite is observed for the relation extraction task.

Our approach differs from the previous work in two key aspects. Firstly, by employing prompted PWS for misinformation detection using credibility signals, we tackle a more complex task that requires nuanced analysis by the instruction-tuned LLMs. Identifying credibility signals demands strong reasoning and information retrieval capabilities, along with the ability to maintain factual accuracy, as opposed to the string matching tasks explored by the authors or by the aforementioned works on PWS for misinformation detection. Secondly, we leverage more recent instruction-tuned LLMs that have demonstrated significant improvements over the models experimented with by Smith et al. (2022).

3 Materials & Methods

Our approach employs a two-step prompting technique for obtaining the credibility signals and zero-shot misinformation labels. As reported by Arora et al. (2022), open-ended prompts tend to outperform prompts that pose restrictions or place specific formatting conditions to the LLMs. Therefore, we start with an open-ended prompt, then use a task-agnostic restrictive prompt to map the answer to a predefined class.

For credibility signals, we use an instruction prompt along with the article text, followed by the credibility signal prompt. This prompt pair is repeated for each credibility signal. For the zero-shot prompt, we use a single instruction to detect misinformation as a binary classification task. To map the model's answer to a label, we first apply simple string matching rules. If unsuccessful, we employ the task-agnostic category mapping prompt.

Figure 1 presents an overview of our prompting procedure using three examples of prompts for obtaining credibility signals (the complete set of prompts, along with examples of answers, can be found in Appendix C).

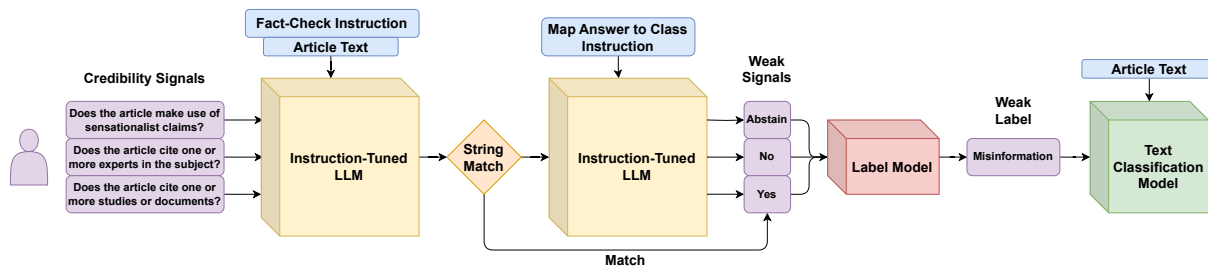


Figure 1: Two-step prompted weak supervision with credibility signals.

3.1 Credibility Signals

We selected eighteen credibility signals (Table 1) shown to be important for misinformation detection by previous studies mentioned in Section 2. There are five negative and thirteen positive signals, which contribute to ‘Non-Misinformation’ and ‘Misinformation’ respectively, given that the answer to the respective credibility question is ‘Yes’. Their contribution to the objective label is the opposite if the answer is ‘No’, and neutral if the answer is ‘Abstain’.

3.2 Instruction-Tuned Models

GPT-3.5-Turbo³ is a closed-source chatbot trained to follow instructions using reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022a). It is built on top of the GPT3 large language model introduced by Brown et al. (2020). We use the default API parameters for prompting: 100% probability mass for nucleus sampling, the response’s maximum number of tokens as the maximum context length, and a presence penalty of 0. To make the answers focused and deterministic, we set the temperature to 0.1.

Alpaca-LoRA-30B (Taori et al., 2023) is an open-source chatbot based on LLaMA (Touvron et al., 2023), and finetuned with low-rank adaptation (Hu et al., 2021) with 52K instruction prompts generated using OpenAI’s *text-davinci-003*. The prompting parameters used are: top-75% probability token sampling, response’s maximum number of tokens as 512, temperature of 0.1 and the number of beam search steps ranging from 2 to 4. We also use 8-bit quantization (Detmeters et al., 2022).

OpenAssistant-LLaMa-30B (Köpf et al., 2023) is a chatbot with open and crowd-sourced human-annotated training data consisting of more than 160K messages in 35 different languages, 42.8% of

which are in English. Here, the dataset is used to fine-tune a LLaMA-30B model for six epochs. The prompting parameters were the same as the ones used for Alpaca-LoRA-30B.

3.3 Datasets

As mentioned in Section 2, there are few datasets for misinformation detection that (I) are not social media data, (II) are not claim-level data, and (III) have disjoint categorical classes as opposed to varying scales of trustworthiness.⁴ Next, we describe these two datasets with further details provided in Appendix A.

FA-KES We use the publicly available FA-KES dataset. As described in Section 2, it consists of long articles in English, that were annotated for misinformation using a combination of human-extracted named entities and unsupervised learning. Each example contains the article text and the name of the source that published the article. We perform our experiments on the whole corpus of 804 news articles, of which 426 are labeled as misinformation and 378 are labeled as non-misinformation, and the articles are collected from 15 different sources.

EUvsDisinfo contains articles flagged as pro-Kremlin propaganda and disinformation by EUvsDisinfo⁵. Each article in the EUvsDisinfo database includes a dedicated disproof section where EUvsDisinfo provides evidence to debunk the false information in the article. In this work, we extend the original dataset from Kaggle⁶ by introducing "non-misinformation" articles, which are news articles referenced in the dedicated EUvsDisinfo disproof sections. We selected a subset of English disproof

⁴Not suitable for our settings as each credibility signal is designed to have a positive contribution to a single class.

⁵<https://euvsdisinfo.eu/about/> and <https://euvsdisinfo.eu/disinformation-cases/>

⁶<https://www.kaggle.com/datasets/corrieaar/disinformation-articles>

³<https://openai.com/blog/chatgpt>

– Evidence*	+ Bias‡	+ Inference#
– Document Citation#	+ Emotional Valence†	+ Call to Action†
– Source Credibility#	+ Incorrect Spelling†	+ Explicitly Unverified Claims†
– Expert Citation†	+ Personal Perspective†	+ Informal Tone#
– Reported by Other Sources†	+ Incivility†	+ Impoliteness†
+ Low Credibility Organisation†	+ Sensationalism†	+ Polarising Language#

Table 1: Credibility Signals.

+ Positive Signals ('Yes' → Misinformation, 'No' → Non-Misinformation)

– Negative Signals ('Yes' → Non-Misinformation, 'No' → Misinformation)

*Musi and Reed (2022) ‡Dufraisse et al. (2022) # Zhang et al. (2018) †<https://github.com/w3c/credweb>

articles, while preserving the original class distribution. The misinformation and non-misinformation articles (covering 206 unique news sources) were downloaded in March 2020 and March 2023, respectively, and all articles were published before March 2020. Each example contains the article text and the name of the publishing source.

Both EuvsDisinfo and FA-KES datasets are split into 80% and 20% for training and testing, respectively. The distribution of true versus false articles in both datasets is shown in Table 2.

Objective	FA-KES		EUvsDisinfo	
	Train	Test	Train	Test
Mis/Disinformation	341	85	86	22
Non-Misinformation	302	76	311	78

Table 2: Veracity class distributions.

3.4 Classification Experiments

Our experimental setup consists of three scenarios: supervised fine-tuning with ground-truth labels, zero-shot prompting, and prompted PWS with credibility signals. We compute the mean and standard deviations of the accuracy and $F1$ -macro scores, which were obtained over three independent runs.

Supervised fine-tuning. In the supervised fine-tuning scenario, we use a RoBERTa-Base architecture with the following hyperparameters: AdamW optimizer, learning rate of $1e-5$, batch size of 8, max sequence length of 512, weight decay of $1e-2$, 100 warmup steps, and 100 training epochs.

Zero-shot prompting. In the zero-shot scenario, we simply prompt the instruction-tuned LLM to perform veracity prediction on the test set.

Prompted PWS. For the PWS experiments, we employ the Snorkel framework (Ratner et al., 2017).

Its label model, based on Ratner et al. (2019), computes the inverse generalized covariance matrix of the junction tree of the dependency graph obtained from the weak signals. It then performs a matrix completion-style algorithm to recover the accuracies of these weak signals, without using ground-truth data. These estimated parameters are then used to weight and combine the weak signals into the binary veracity labels. The weak signals are assumed to be conditionally independent with respect to the binary veracity classes. We train the label model for 500 epochs using the 18 weak signals obtained from articles in the train split.

We experiment with the weak labels using two different approaches:

- *Label model only (L)*: the label model is used to combine the 18 credibility signals obtained from the test set into binary weak labels, which are directly evaluated against the ground-truth labels. With this approach, we are not fitting a classifier to distinguish the two classes by learning from textual features. Instead the final prediction is a weighted combination of the credibility signals.
- *Full (FULL)*: we use the label model to combine the 18 credibility signals from the training set into binary weak labels, which are then used to train a text classifier optimized for binary cross-entropy loss. The text classifier is based on a RoBERTa-Base architecture, sharing the same hyperparameters as the supervised fine-tuning scenario. The trained text classifier predicts labels for the test set, which are subsequently evaluated against the ground-truth labels. By incorporating textual features and training a dedicated text classifier, *FULL* (Figure 1) aims to generalise beyond the capabilities of the label model and to make more informed predictions.

3.5 Analysing Credibility Signals

Lastly, we analyse the individual impact of each credibility signal at predicting the ground-truth binary misinformation labels. To do so, we evaluate each credibility signal against the ground-truth veracity labels for the entire FA-KES and EuvsDisinfo datasets.

4 Results

4.1 Classification Results

Table 3 shows the performance of each model on both datasets. Zero-shot prompting produces the lowest average scores for both datasets. While GPT-3.5-Turbo with zero-shot prompting performs better than the supervised classifier on the EuvsDisinfo dataset, the other two LLMs are outperformed by the supervised classifier. Furthermore, the supervised classifier outperforms zero-shot predictions for all three LLMs on the FA-KES dataset. ***These findings suggest that fine-tuning a text classifier with ground-truth data leads to higher average scores than zero-shot prompting with LLMs.***

In contrast, prompted PWS with credibility signals achieves the highest scores for both datasets. The *label model only* approach surpasses zero-shot prompting for GPT-3.5-Turbo and Alpaca-LoRA-30B on FA-KES, as well as for Alpaca-LoRA-30B and OpenAssistant-30B for EuvsDisinfo. Additionally, Alpaca-LoRA-30B-L and GPT-3.5-Turbo-L manage to outperform the supervised classifier for FA-KES and EuvsDisinfo, respectively. The *FULL* approach further enhances performance for both datasets. OpenAssistant-30B-*FULL* achieves the highest scores for the FA-KES dataset, with 55.3% accuracy and 54.8% *F1*-macro. These values are respectively 2.4 and 1.9 absolute points higher than the supervised method. For the EuvsDisinfo dataset, GPT-3.5-Turbo-*FULL* achieves the highest scores, with 99.3% accuracy and 99.0% *F1*-macro. These results indicate significant improvements over the supervised model, with increases of 22.3 and 37.6 absolute points in accuracy and *F1*-macro, respectively, over the supervised method. However, we must emphasize the closed-source nature of GPT-3.5-Turbo. At present, it is not feasible to verify whether the model had prior access to the articles (i.e., data leakage), which would render any comparison unfair. Nonetheless, it is noteworthy that OpenAssistant-30B, an open-source model, attains the scores of 91.3% accuracy and 85.8% *F1*-macro using *full*

PWS, which are the highest scores for EuvsDisinfo when excluding the performance of GPT-3.5-Turbo.

By directly comparing zero-shot against *full* PWS, we observe that *full* PWS outperforms zero-shot by an average relative increase across the three models of +23.1% and +92.1% in *F1*-Macro for the FA-KES and EuvsDisinfo datasets, respectively. Notably, Alpaca-LoRA-30B demonstrates the highest increase in *F1*-Macro for both datasets, averaging at +125.9%. Conversely, GPT-3.5-Turbo achieves the lowest relative average *F1*-Macro increase of +16% across both datasets.

Finally, in Table 4, we compare our results for the FA-KES dataset with previous work. It is essential to note that the FA-KES dataset lacks a predefined test split, making precise comparisons challenging. The cited papers were evaluated using different random splits of the dataset, albeit all using the same test set size, which constitutes 20% of the dataset. OpenAssistant-30B-*FULL* achieves 2nd place with 55.3% and 54.8% accuracy and *F1*-macro scores, respectively, surpassing Saleh et al. (2021) and Elhadad et al. (2020). Nasir et al. (2021) achieves the highest scores, obtaining 60% accuracy and 59% *F1*-macro. However, we highlight that unlike prior work, our model does not perform fine-tuning with ground-truth labels.

4.2 Credibility Signals Analysis

Figure 2 reports the accuracy and coverage⁷ of each credibility signal compared to the ground-truth binary veracity labels. Due to space constraints, we report the fine-grained distribution of the credibility signals extracted respectively from mis/disinformation and non-misinformation articles in Appendix D.

For EuvsDisinfo, we observe a mean accuracy (μ_{acc}) of 68%, with a total of 10 signals achieving accuracy scores higher than 70%: *Bias*, *Emotional Valence*, *Polarising Language*, *Informal Tone*, *Incorrect Spelling*, *Reported by Other Sources*, *Impoliteness*, *Source Credibility*, *Incivility*, and *Sensationalism*, in increasing order. Among these, *Sensationalism* achieves the highest accuracy score of 81%. *Expert Citation* and *Document Citation* achieve less than 50% accuracy. In terms of coverage, 14 signals achieve scores higher than 90%, with the exception of *Bias* at 71%, *Low Credibility*

⁷percentage of examples in which the respective credibility signal did not abstain from voting.

Setting	Fine-tuned with GT labels	Model	FA-KES		EuvsDisinfo	
			Accuracy	F1-Macro	Accuracy	F1-Macro
Supervised	✓	RoBERTa-Base	52.9 ± 1.9	52.9 ± 1.9	77.0 ± 4.0	61.4 ± 9.0
Zero-Shot	×	GPT-3.5-Turbo	46.2 ± 2.5	43.3 ± 2.0	87.7 ± 5.1	83.8 ± 5.7
	×	Alpaca-LoRA-30B	52.7 ± 0.7	34.9 ± 0.9	24.0 ± 3.5	21.1 ± 5.3
	×	OpenAssistant-30B	52.3 ± 4.3	50.4 ± 5.5	58.7 ± 4.7	56.1 ± 3.8
Weakly Supervised	×	GPT-3.5-Turbo-L	47.9 ± 4.8	47.4 ± 4.3	77.0 ± 3.5	73.5 ± 4.5
	×	Alpaca-LoRA-30B-L	53.8 ± 1.9	53.2 ± 2.6	54.0 ± 7.0	50.4 ± 8.0
	×	OpenAssistant-30B-L	49.2 ± 5.5	49.1 ± 5.4	69.0 ± 9.5	63.5 ± 11.4
	×	GPT-3.5-Turbo-FULL	49.8 ± 3.6	49.3 ± 3.0	99.3 ± 0.6	99.0 ± 0.9
	×	Alpaca-LoRA-30B-FULL	53.0 ± 4.5	51.2 ± 2.9	67.0 ± 10.0	64.4 ± 12.7
	×	OpenAssistant-30B-FULL	55.3 ± 3.5	54.8 ± 3.6	91.3 ± 2.5	85.8 ± 5.2

Table 3: Classification results according to accuracy and $F1$ -macro for FA-KES and EUvsDisinfo. Mean ± 1 std are computed over three random seed runs. Best results for each dataset are in **bold**.

L=Label model only, FULL=Label model + RoBERTa-Base

Model	Trained with GT labels	Accuracy	F1-Macro
Hybrid CNN-RNN (Nasir et al., 2021)	✓	60 ± 0.7	59
OpenAssistant-30B-FULL	×	55.3 ± 3.5	54.8 ± 3.6
OPCNN-FAKE (Saleh et al., 2021)	✓	53.99	53.99
Multinomial Naive Bayes (Elhadad et al., 2020)	✓	58	50

Table 4: Indirect* comparison of results with previous works for the FA-KES dataset. Results sorted by F1-Macro.

*Obtained from different data splits.

Organization, Reported by Other Sources, both at 67%, and *Emotional Valence* at 45%.

In FA-KES, we attain a μ_{acc} of 49%, with 5 signals achieving accuracy scores higher than 50%: *Explicitly Unverified Claims* and *Bias*, both at 51%, *Document Citation* at 52%, *Expert Citation* at 53%, and *Low Credibility Organization* at 54%. The remaining signals range in accuracy from 46% to 49%, with the exception of *Reported by Other Sources* that achieves the lowest accuracy (39%). In terms of coverage, 13 signals score above 90%, with the exception of *Source Credibility* at 88%, *Reported by Other Sources* at 66%, *Bias* at 58%, *Low Credibility Organization* at 49%, and *Emotional Valence* at 45%.

We obtain a similar standard deviation of 0.15 and 0.19 for the coverage scores of EuvsDisinfo and FA-KES, respectively. However, the standard deviation of the accuracy scores varies significantly between both datasets, with 0.13 and 0.03 for EuvsDisinfo and FA-KES, respectively. For example,

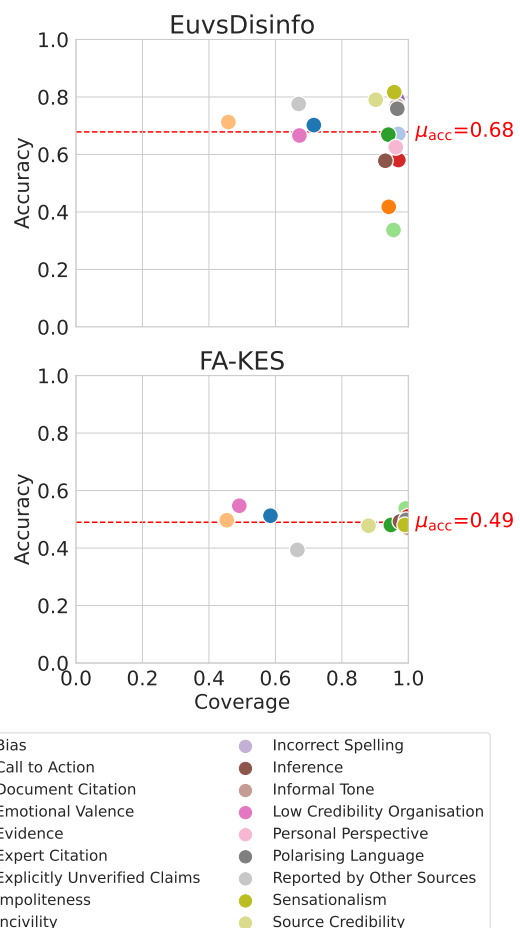


Figure 2: Accuracy and coverage per credibility signal.

Expert Citation achieves accuracy scores of 33% for EuvsDisinfo and 53% for FA-KES, while both attain similar coverage scores of 95% and 98% for EuvsDisinfo and FA-KES, respectively. This may

be a reflection of the variability in the distribution of these signals across different datasets, and may be correlated with their underlying characteristics.

In particular, *Incivility* and *Sensationalism* obtain the top two highest average accuracy scores of 79% and 81%, respectively, for the EuvsDisinfo dataset. Since EuvsDisinfo consists of articles originating from pro-Kremlin media, it would be reasonable to assume the presence of these signals in articles promoting counterpropaganda, for example. In the case of the FA-KES dataset, *Low Credibility Organization* achieves the highest average accuracy score of 54%, which may be attributed to the fact that the 804 articles in FA-KES were published by only 15 different sources.

We highlight that the *Emotional Valence* signal achieves over 98% coverage with the GPT-3.5-Turbo model, despite its low average coverage of 45% across the three models. While both OpenAssistant-LLaMa-30B and Alpaca-LoRA-30B provide correct answers for this specific credibility signal, the extraction of the objective class from the answer often led to the incorrect assignment of 'Abstain' labels by the second step of our prompting method. For the remaining signals, we consistently observed accurate mappings between the answers and the objective classes.

5 Discussion of Key Findings

Returning to our research questions, we find that:

(RQ1) Zero-shot prompting for instruction-tuned LLMs is consistently outperformed by fine-tuning supervised text classifiers with ground-truth data for long article veracity classification.

(RQ2) Prompted PWS achieves notable improvements over zero-shot prompting, with relative percentage increases in $F1$ -macro of +23.1% and +92.1%, averaged across the three models for FA-KES and EuvsDisinfo, respectively. Prompted PWS also outperforms supervised text classifiers trained on ground-truth data, achieving the highest scores of 54.8% and 99.0% $F1$ -macro for FA-KES and EuvsDisinfo, respectively.

(RQ3) Certain signals, even when considered individually, achieve competitive accuracy scores compared to direct zero-shot prompting for veracity classification. However, their accuracy varies significantly across datasets, indicating a correlation between the accuracy of certain signals and the underlying properties of the datasets. This highlights the necessity of quality over quantity when

leveraging credibility signals for veracity classification. Additionally, the performance of signals such as *Low Credibility Organisation* and *Source Credibility* highlights the capabilities of LLMs in leveraging external knowledge acquired during pre-training to accurately assess credibility, as these signals often require information that is external to the article's text.

In conclusion, we highlight that a critical concern within the AI community revolves around the ethical application of LLMs. Significant efforts are being made to align instruction-tuned LLMs in adhering to safety protocols and enhancing factual accuracy. The ability to evaluate the presence of sufficient factual support and ethical considerations before providing an answer, and to refrain from answering when appropriate, is a valuable feature when combined with PWS. In such cases, signals can abstain from contributing to the objective label for a specific example, rather than being forced to offer a potentially hallucinated erroneous answer. This resonates with our observed results on Section 4.2, as signals with moderate coverage rate often achieve the highest accuracy scores.

6 Conclusion & Future Work

This paper reported experiments on two datasets of long articles labeled for content veracity. We employed weak supervision to combine 18 credibility signals obtained by prompting instruction-tuned large language models. Our approach consistently outperforms zero-shot prompting for all three models across both datasets. Moreover, our approach achieves better performance than supervised text classifiers trained with ground-truth data and demonstrates comparable performance to other related work on the FA-KES dataset, even though we do not use any ground-truth labels for training. Furthermore, we evaluated the performance of each individual credibility signal in predicting content veracity, providing valuable insights into their role in misinformation detection.

In future work, we plan to extend this research to multiple languages and incorporate different modalities into new credibility signals.

Limitations

It is important to note that our focus is on assessing the contribution of LLM-predicted credibility signals on article-level veracity classification. Further research is required to determine if the proposed

method and the set of credibility signals can be effectively applied to claim-level fact-check datasets, since our approach has access to significantly less contextual information in such cases.

Also, the use of instruction-tuned LLMs comes with significant computational demands, particularly in terms of GPU resources, even when employing lightweight optimization techniques like 8-bit quantization. In fact, our proposed method is considerably more GPU-intensive than zero-shot prompting, as LLMs are prompted multiple times per example instead of just once. Further research is necessary to investigate the feasibility of obtaining multiple credibility signals in a single prompt without significantly reducing performance.

Finally, we emphasize that we investigated the possibility of data leakage from the datasets of FA-KES and EUvsDisinfo in both Alpaca-LoRA-30B and OpenAssistant-LLaMa-30B using exact string matching. Our investigation revealed no evidence of data leakage. However, it is important to note that due to the closed-source nature of GPT-3.5-Turbo, we are unable to make similar observations as the model architecture and training data remain proprietary. Also, we highlight that we did not investigate potential data leakage on LLaMA’s pre-training corpora, which consists of 7 datasets totalling more than 4.5 terabytes of text. We describe the data leakage investigation in more details in Appendix B.

Ethics Statement

Data Disclaimer The data used in this work was downloaded from the publicly available open-sourced corpora and was extended using the publicly available resources that do not contain sensitive information or personal data. In particular, in order to collect the *non-misinformation* articles for the EUvsDisinfo dataset, we use the urls cited by EUvsDisinfo to debunk the respective pro-Kremlin disinformation articles.

Environmental impact The proposed method is computationally demanding due to the application of LLMs to the assessment of large articles based on 18 different credibility signals and the number of runs for each model. It is therefore expected for this and similar approaches to leave a certain carbon footprint.

Biases Prompt-tuned LLMs can potentially generate biased replies when performing veracity as-

essment. However, such biases are difficult to eliminate even when humans perform this task (Pérez-Rosas et al., 2017).

Acknowledgements

We thank Ben Wu and Ivan Srba for their valuable feedback. This research is partially supported by the UK’s innovation agency (InnovateUK) grant 10039055 (vera.ai).

References

- Fatima K. Abu Salem, Roaa Al Feel, Shady Elbassuoni, Mohamad Jaber, and May Farah. 2019. *Fa-kes: A fake news dataset around the syrian war*. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):573–582.
- Sadia Afroz, Michael Brennan, and Rachel Greenstadt. 2012. Detecting hoaxes, frauds, and deception in writing style online. In *2012 IEEE Symposium on Security and Privacy*, pages 461–475. IEEE.
- Simran Arora, Avanika Narayan, Mayee F Chen, Laurel J Orr, Neel Guha, Kush Bhatia, Ines Chami, Frederic Sala, and Christopher Ré. 2022. Ask me anything: A simple strategy for prompting language models. *arXiv preprint arXiv:2210.02441*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. *Language models are few-shot learners*. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022. 8-bit optimizers via block-wise quantization. *9th International Conference on Learning Representations, ICLR*.
- A Dimou et al. 2022. Evaluating web content using the w3c credibility signals. In *Towards a Knowledge-Aware AI: SEMANTiCS 2022—Proceedings of the 18th International Conference on Semantic Systems, 13-15 September 2022, Vienna, Austria*, volume 55, page 3. IOS Press.
- Evan Dufraisse, Céline Treuillier, Armelle Brun, Julien Tourille, Sylvain Castagnos, and Adrian Popescu. 2022. *Don’t burst blindly: For a better use of natural language processing to fight opinion bubbles in news recommendations*. In *Proceedings of the LREC 2022 workshop on Natural Language Processing for*

- Political Sciences*, pages 79–85, Marseille, France. European Language Resources Association.
- Mohamed K Elhadad, Kin Fun Li, and Fayez Gebali. 2020. A novel approach for selecting hybrid features from online news textual metadata for fake news detection. In *Advances on P2P, Parallel, Grid, Cloud and Internet Computing: Proceedings of the 14th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC-2019) 14*, pages 914–925. Springer.
- William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*. ACL.
- Chenbo Fu, Xingyu Pan, Xuejiao Liang, Shanqing Yu, Xiaoke Xu, and Yong Min. 2023. Feature drift in fake news detection: An interpretable analysis. *Applied Sciences*, 13(1):592.
- Daniel Fu, Mayee Chen, Frederic Sala, Sarah Hooper, Kayvon Fatahalian, and Christopher Ré. 2020. Fast and three-rious: Speeding up weak supervision with triplet methods. In *International Conference on Machine Learning*, pages 3280–3291. PMLR.
- Anastasia Giachanou, Paolo Rosso, and Fabio Crestani. 2019. [Leveraging emotional signals for credibility detection](#). In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19*, page 877–880, New York, NY, USA. Association for Computing Machinery.
- Stefan Helmstetter and Heiko Paulheim. 2018. Weakly supervised learning for fake news detection on twitter. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 274–277. IEEE.
- Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 541–550.
- Benjamin Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 759–766.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. [Survey of hallucination in natural language generation](#). *ACM Comput. Surv.*, 55(12).
- Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. 2023. Openassistant conversations—democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*.
- Neema Kotonya and Francesca Toni. 2020. [Explainable automated fact-checking for public health claims](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.
- Paweł Ksieniewicz, Paweł Zybiewski, Michał Choraś, Rafał Kozik, Agata Giełczyk, and Michał Woźniak. 2020. [Fake news detection from data streams](#). In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.
- Yi-Ju Lu and Cheng-Te Li. 2020. [GCAN: Graph-aware co-attention networks for explainable fake news detection on social media](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 505–514, Online. Association for Computational Linguistics.
- Tanushree Mitra and Eric Gilbert. 2015. Credbank: A large-scale social media corpus with associated credibility annotations. In *Proceedings of the international AAAI conference on web and social media*, volume 9, pages 258–267.
- Elena Musi and Chris Reed. 2022. [From fallacies to semi-fake news: Improving the identification of misinformation triggers across digital media](#). *Discourse & Society*, 33(3):349–370.
- Jamal Abdul Nasir, Osama Subhani Khan, and Iraklis Varlamis. 2021. [Fake news detection: A hybrid cnn-rnn based deep learning approach](#). *International Journal of Information Management Data Insights*, 1(1):100007.
- Nicole O’Brien, Sophia Latessa, Georgios Evangelopoulos, and Xavier Boix. 2018. [The language of fake news: Opening the black-box of deep learning based detectors](#). In *workshop on "AI for Social Good", NIPS 2018*, Montreal, Canada.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022a. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022b. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*.
- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. [Language models as knowledge bases?](#) In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.
- Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. [DeClarE: Debunking fake news and false claims using evidence-aware deep learning.](#) In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32, Brussels, Belgium. Association for Computational Linguistics.
- Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.
- Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937.
- Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11, page 269. NIH Public Access.
- Alexander Ratner, Braden Hancock, Jared Dunnmon, Frederic Sala, Shreyash Pandey, and Christopher Ré. 2019. [Training complex models with multi-task weak supervision.](#) *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):4763–4771.
- Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. *Advances in neural information processing systems*, 29.
- Ruty Rinott, Lena Dankin, Carlos Alzate, Mitesh M Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence—an automatic method for context dependent evidence detection. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 440–450.
- Hager Saleh, Abdullah Alharbi, and Saeed Hamood Alsamhi. 2021. [Opcnn-fake: Optimized convolutional neural network for fake news detection.](#) *IEEE Access*, 9:129471–129489.
- Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.
- Giovanni Santia and Jake Williams. 2018. Buzzface: A news veracity dataset with facebook user commentary and egos. In *Proceedings of the international AAAI conference on web and social media*, volume 12, pages 531–540.
- Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. [Defend: Explainable fake news detection.](#) In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19*, page 395–405, New York, NY, USA. Association for Computing Machinery.
- Kai Shu, Susan Dumais, Ahmed Hassan Awadallah, and Huan Liu. 2020a. Detecting fake news with weak social supervision. *IEEE Intelligent Systems*, 36(4):96–103.
- Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020b. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.
- Kai Shu, Guoqing Zheng, Yichuan Li, Subhabrata Mukherjee, Ahmed Hassan Awadallah, Scott Ruston, and Huan Liu. 2020c. [Early detection of fake news with multi-source weak social supervision.](#) In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part III*, page 650–666, Berlin, Heidelberg. Springer-Verlag.
- Renato Silva and Tiago Almeida. 2021. [How concept drift can impair the classification of fake news.](#) In *Anais do IX Symposium on Knowledge Discovery, Mining and Learning*, pages 121–128, Porto Alegre, RS, Brasil. SBC.
- Niraj Sitaula, Chilukuri K Mohan, Jennifer Grygiel, Xinyi Zhou, and Reza Zafarani. 2020. Credibility-based fake news detection. *Disinformation, misinformation, and fake news in social media: Emerging research challenges and Opportunities*, pages 163–182.
- Ryan Smith, Jason A Fries, Braden Hancock, and Stephen H Bach. 2022. Language models in the loop: Incorporating prompting into weak supervision. *arXiv preprint arXiv:2205.02318*.

- Eugenio Tacchini, Gabriele Ballarin, Marco L Della Vedova, Stefano Moret, and Luca De Alfaro. 2017. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506*.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models*. <https://crfm.stanford.edu/2023/03/13/alpaca.html>, 3(6):7.
- Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.
- James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. *arXiv preprint arXiv:1806.07687*.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Paroma Varma, Frederic Sala, Shiori Sagawa, Jason Fries, Daniel Fu, Saelig Khattar, Ashwini Ramamoorthy, Ke Xiao, Kayvon Fatahalian, James Priest, et al. 2019. Multi-resolution weak supervision for sequential data. *Advances in Neural Information Processing Systems*, 32.
- Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, pages 18–22.
- Andreas Vlachos and Sebastian Riedel. 2015. Identification and verification of simple claims about statistical properties. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2596–2601. Association for Computational Linguistics.
- William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.
- Yaqing Wang, Weifeng Yang, Fenglong Ma, Jin Xu, Bin Zhong, Qiang Deng, and Jing Gao. 2020. [Weak supervision for fake news detection via reinforcement learning](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):516–523.
- Weizhi Xu, Junfei Wu, Qiang Liu, Shu Wu, and Liang Wang. 2022. [Evidence-aware fake news detection with graph neural networks](#). In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 2501–2510, New York, NY, USA. Association for Computing Machinery.
- Amy X. Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B. Adams, Emmanuel Vincent, Jennifer Lee, Martin Robbins, Ed Bice, Sandro Hawke, David Karger, and An Xiao Mina. 2018. [A structured response to misinformation: Defining and annotating credibility indicators in news articles](#). In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 603–612, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.
- Xichen Zhang and Ali A. Ghorbani. 2020. [An overview of online fake news: Characterization, detection, and discussion](#). *Information Processing & Management*, 57(2):102025.
- Xinyi Zhou and Reza Zafarani. 2020. [A survey of fake news: Fundamental theories, detection methods, and opportunities](#). *ACM Comput. Surv.*, 53(5).
- Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2021. [A comprehensive survey on transfer learning](#). *Proceedings of the IEEE*, 109(1):43–76.
- Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS one*, 11(3):e0150989.

A Datasets

In this section we detail characteristics of EuvsDisinfo and FA-KES.

A.1 FA-KES

FA-KES comprises 804 articles written in English, and collected from 15 different sources, of which 426 are labeled as misinformation, and 378 are labeled as non-misinformation. Figure 4 displays the distribution of misinformation and non-misinformation articles per source.

A.2 EuvsDisinfo

The dataset of EuvsDisinfo articles introduced in this paper is composed of 497 articles written in English, and collected from 206 unique sources, of which 108 are labeled as disinformation, and 389 are labeled as reliable (i.e. non-misinformation) content. Figure 3 displays the distribution of the top 15 sources containing the most disinformation and reliable articles respective.

A.3 Differences between datasets

The two datasets differ in many aspects, including (I) their topic, (II) the way in which their ground-truth labels were obtained, (III) the number of different news sources, and (IV) the length of the articles.

(I) FA-KES news articles are focused entirely on the Syrian war, while EuvsDisinfo’s articles originate from a wide range of topics discussed in pro-Kremlin media.

(II) FA-KES was annotated using unsupervised approaches to combine human-annotated information (e.g., dates and locations), and compared against ground-truth facts from the Syrian Violations Documentation Center. In contrast, EuvsDisinfo’s disinformation articles were debunked after thorough manual investigations by verification, while the reliable, non-misinformation articles were provided directly in the dedicated disproof sections of each EuvsDisinfo article where the verification professionals used them as evidence in debunking the pro-Kremlin disinformation.

(III) FA-KES contains only 15 different news sources: ahram, manar, alalam, alaraby, dailysabah, sana, sputnik, etilaf, trt, jordantimes, reuters, arabiya, nna, asharqalawsat, and tass. Additionally, all sources contain both misinformation and non-misinformation articles, often in similar proportions, as shown in Figure 4. EuvsDisinfo contains 206 different sources, with the vast majority of them exclusively publishing either disinformation or reliable articles. The only exceptions are *sputniknews.com* with 1 article used as evidence and 28 disinformation articles, *rt.com* with 3 reliable articles and 15 disinformation articles, and finally *defensenews.com* and *veteranstoday.com*, both with 1 reliable article and 1 disinformation article each.

(IV) For FA-KES, the average number of characters per article is 1,968, while the average number of tokens per article is 336. For EuvsDisinfo, the average number of characters per article is 6,836, while the average number of tokens per article is 1,240.

B Data Leakage Investigation

In order to investigate whether the open-source instruction-tuned LLMs had access to the FA-KES or EuvsDisinfo datasets during the fine-tuning stage, we conducted the following evaluation process: Firstly, for each training sample of Alpaca and OpenAssistant, we combined the instruction,

input, and output into a single string, with each component separated by a newline character. Subsequently, we lowercased and sentensized each document on the four datasets: EuvsDisinfo, FA-KES, OpenAssistant, and Alpaca. Finally, we iterated through the sentence pairs in two specific datasets and computed the count of exact sentence matches with the training data in both Alpaca and OpenAssistant. Table 5 presents the percentage of sentences that were matched for each dataset pair. Upon performing exact matching for whole documents instead of sentences, we did not match a single instance between any pairs of datasets.

Alpaca ⁸			
Dataset	Sentences	Matches	Percentage
FA-KES	9767	41	0.4%
EuvsDisinfo	20494	199	1.0%

OpenAssistant ⁹			
Dataset	Sentences	Matches	Percentage
FA-KES	9767	41	0.4%
EuvsDisinfo	20494	871	4.3%

Table 5: Amount of sentences from FA-KES and EuvsDisinfo that were found in the training set of Alpaca and OpenAssistant.

We highlight that we did not perform any investigation towards data leakage on the corpora used to pre-train LLaMA, which consists of 7 different datasets totalling more than 4.5 terabytes of text. Within our limited time and resources, we focused our efforts into investigating potential data leakage in the instruction-tuning dataset.

C Prompts

In this section we provide a more in-depth description of the prompting strategy, along with the complete prompts.

As stated in section 3, our prompting strategy differs slightly for obtaining weak signals and zero-shot labels. Additionally, we employ a two-step prompting technique to map the answers to pre-defined categories. The instruction prompts are presented in Table 6, while the credibility signal prompts are presented in Table 7.

⁸<https://huggingface.co/datasets/tatsu-lab/alpaca>

⁹<https://huggingface.co/datasets/OpenAssistant/oasst1>

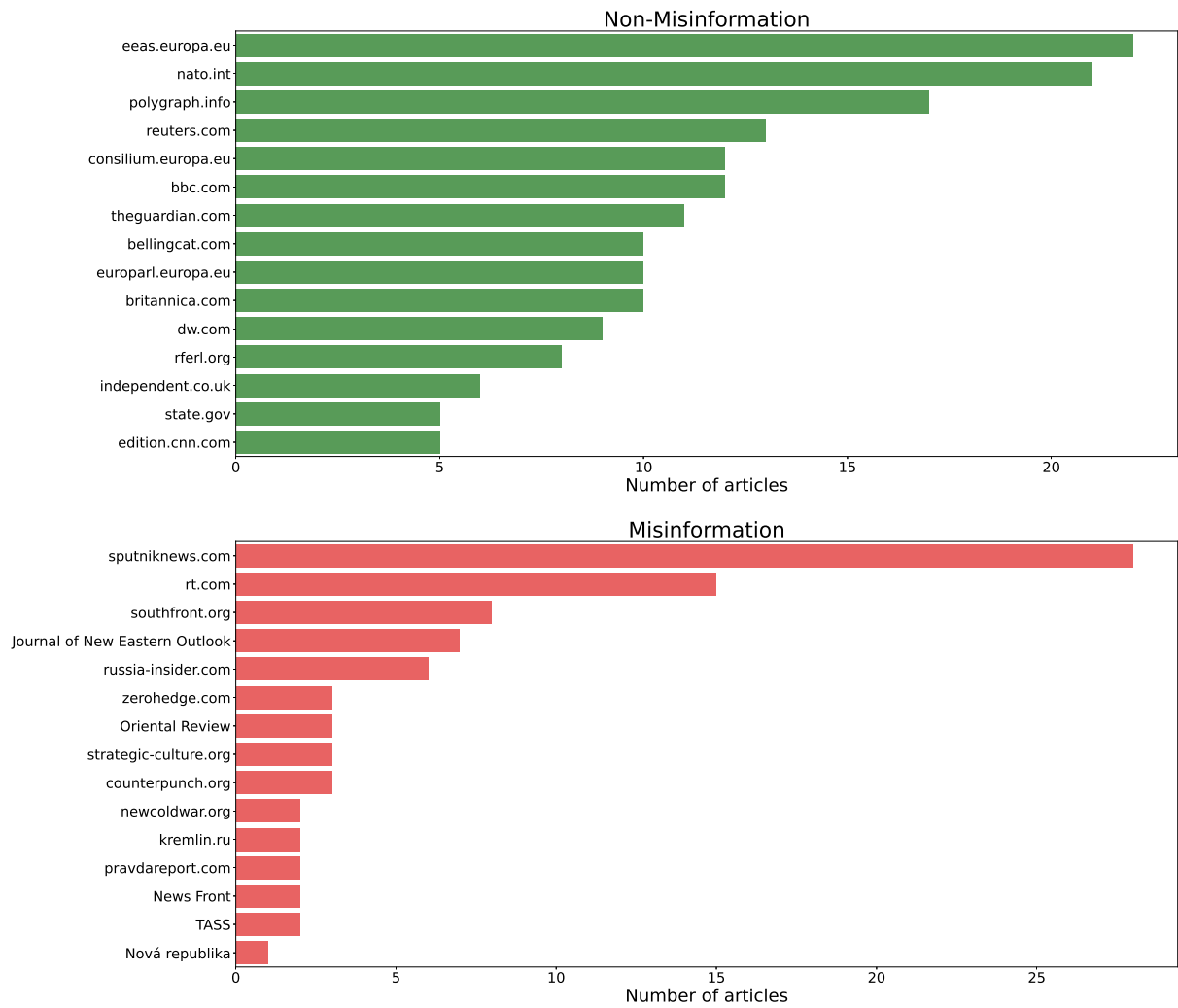


Figure 3: Distribution of the top 15 sources containing most Misinformation (bottom) and Non-Misinformation (top) articles for **EuvsDisinfo**.

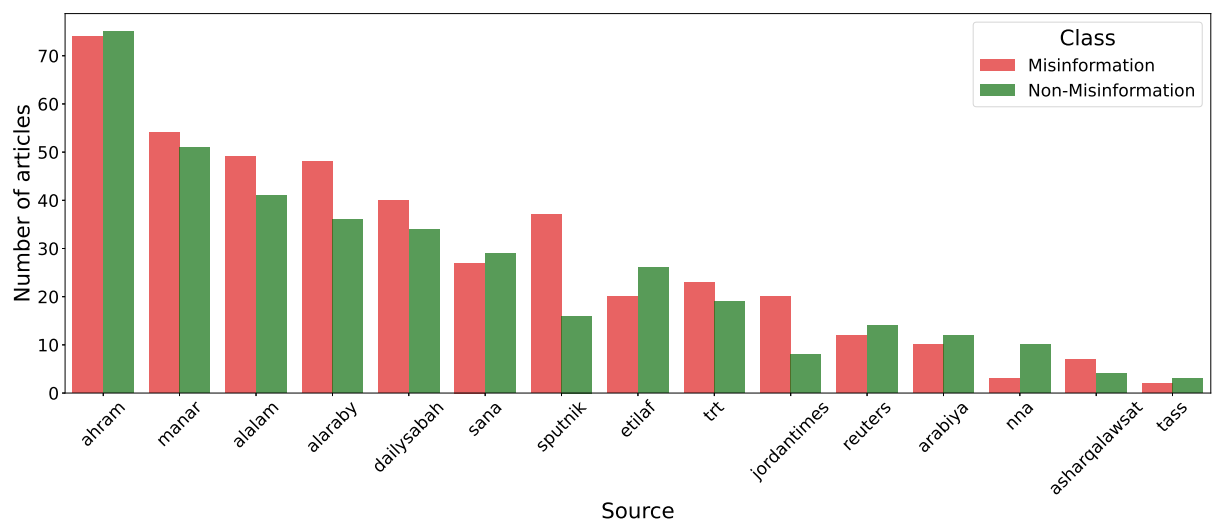


Figure 4: Distribution of Misinformation and Non-Misinformation articles per source for **FA-KES**.

To obtain the weak signals, we begin by using an instruction prompt that guides the model to provide objective answers to the upcoming questions. This prompt includes the news article text. Then, we use a credibility signal prompt to ask the model if the respective credibility signal is applicable to the provided article. This prompt pair is repeated for each credibility question, and no context is shared from one credibility signal to another.

As for the zero-shot prompt, we use a single instruction prompt along with the article text to ask the model if the article contains misinformation/propaganda or not.

To map the model’s answer to a specific class, we follow a two-step process. First, we retrieve the model’s answer to a question and apply simple string matching rules to extract an objective class. Effectively, we check if the answer starts with the string representing one of the expected classes. For the zero-shot prompt, the classes are ‘Yes’ or ‘No’, while for the weak supervision prompts, the classes are ‘Yes’, ‘No’, or ‘Abstain’.

However, if the string matching fails, we resort to a task-agnostic category mapping prompt. This prompt provides both the question and the answer to the instruction-tuned LLM, instructing it to map the answer to one of the expected categories. Notably, the model does not have access to the article text in this prompt.

In tables 8 and 9 we show complete sets of answers obtained from prompting the best LLMs on FA-KES and EuvsDisinfo, respectively, on a randomly-selected sample from each dataset. Note that a few long answers are truncated and signaled with "[...]" due to space constraints.

D Credibility Signals

Figures 5 and 6 display the distribution of votes towards the veracity classification objective for each credibility signal for EuvsDisinfo and FA-KES, respectively. We distinguish the distributions obtained from *mis-/disinformation* articles from the distributions obtained from *non-misinformation/reliable* articles.

For EuvsDisinfo, 11 signals vote towards falseness considerably more often on articles whose ground-truth label is *mis-/disinformation*: *Bias*, *Emotional Valence*, *Explicitly Unverified Claims*, *Impoliteness*, *Incivility*, *Inference*, *Low Credibility Organisation*, *Personal Perspective*, *Polarising Language*, *Sensationalism*, and *Source Credibility*.

In contrast, 7 signals present a similar voting pattern towards falseness independently of the article’s ground-truth veracity: *Call to Action*, *Document Citation*, *Evidence*, *Expert Citation*, and *Reported by Other Sources*, *Incorrect Spelling*, *Informal Tone*.

For the FA-KES dataset, however, all signals achieve a similar voting distribution independently of the article’s ground-truth veracity label. These distributions reflect the accuracy scores reported in Section 4.2, with a mean accuracy score (μ_{acc}) of 49%.

Prompt Type	Prompt Text
Weak Supervision Instruction	<p>I want you to assist me on fact checking news articles. I will provide the whole article text, then I will ask a series of questions about the content of the article. You may elaborate objectively on the answer.</p> <p>Below is the article we will analyse:</p> <p>"{text}"</p>
Weak Supervision 'Map to Category' Instruction	<p>I will provide you with a question and the answer to that question. Your job is to classify the answer into three possible categories. You may only answer with the name of the category. They are "YES", "NO", and "ABSTAIN".</p> <p>Q: {question} A: {answer}</p> <p>Category:</p>
Zero-Shot Instruction	<p>I want you to assist me on fact checking news articles. I will provide the whole article text. You should answer the following question: Does this article contain misinformation or propaganda?</p> <p>You should elaborate clearly on all the reasons why you would find reasons to classify the article as misinformation or propaganda. Below is the article we will analyse:</p> <p>"{text}"</p>
Zero-Shot 'Map to Category' Instruction	<p>I will provide you with a question and the answer to that question. Your job is to classify the answer into three possible categories. You may only answer with the name of the category. They are either "YES" or "NO". You may not answer with anything other than that.</p> <p>Q: Does this article contain misinformation or propaganda? A: {answer}</p> <p>Category:</p>

Table 6: System Instruction Complete Prompts

Credibility Signal	Weak Supervision Prompt
Evidence*	Does the article present any supporting evidence or arguments to substantiate its claims?
Bias‡	Does the article contain explicit or implicit biases?
Inference#	Does the article make claims about correlation and causation?
Explicitly Unverified Claims†	Does the article contain claims that are explicitly unverified?
Personal Perspective†	Does the article express the author’s opinion on the subject?
Emotional Valence†	Is the language in the article extremely negative or extremely positive instead of neutral?
Polarising Language#	Does the article make use of polarising terms or make divisions into sharply contrasting groups or sets of opinions or beliefs?
Call to Action†	Does the article contain language that can be understood as a call to action, requesting readers to follow through with a particular task or telling readers what to do?
Expert Citation†	Does the article cite one or more experts in the subject?
Document Citation#	Does the article cite one or more studies or documents?
Source Credibility#	Does the article cite sources that are generally considered credible?
Incorrect Spelling†	Does the article have significant misspellings and/or grammatical errors?
Informal Tone#	Does the article make use of all caps or consecutive exclamation or question marks?
Incivility†	Does the article make use of stereotypes and generalizations of groups of people?
Impoliteness†	Does the article contain insults, name-calling, or profanity?
Sensationalism†	Does the article make use of sensationalist claims?
Low Credibility Organization†	This article was posted by a media outlet called {organization_name}. Is this source known for publishing false, unverified, or propagandistic articles?
Reported by Other Sources†	Was the story on this article reported by other reputable media outlets?

Table 7: Complete set of Prompts for the Credibility Signals

*Musi and Reed (2022) ‡Dufraisse et al. (2022) # Zhang et al. (2018) †<https://github.com/w3c/credweb>

Credibility Signal	Answer	Vote	Ground-Truth
Evidence	Yes, the article presents some supporting evidence and arguments to substantiate its claims. For example, it cites a statement from Hezbollah al-Nujaba that its forces have killed 23 al-Nusra Front militants and injured 18 others in Khan Touman. It also mentions that [...]	Non-Misinformation	Non-Misinformation
Explicitly Unverified Claims	Yes, the article contains claims that are explicitly unverified. For example, the article states that "23 members of the al-Nusra Front terrorist group were killed in clashes with Hezbollah al-Nujaba," but it does not provide any evidence or sources to support this claim. Additionally, the article states that [...]	Misinformation	Non-Misinformation
Personal Perspective	No, the article does not express the author's opinion on the subject. It is a factual report of events that have taken place in Syria.	Non-Misinformation	Non-Misinformation
Emotional Valence	The language in the article is neutral.	ABSTAIN	Non-Misinformation
Polarizing Language	Yes, the article uses polarizing terms and makes divisions into sharply contrasting groups or sets of opinions or beliefs. For example, it refers to the al-Nusra Front as a "terrorist group" and portrays Hezbollah al-Nujaba as a "major Iraqi Shiite (Shia) resistance movement fighting against terrorism." It also mentions the "Syrian government forces" and "militant groups" as separate entities.	Misinformation	Non-Misinformation
Call to Action	No, the article does not contain any language that can be understood as a call to action, requesting readers to follow-through with a particular task or tells readers what to do. It is a news article reporting on a military operation in Syria.	Non-Misinformation	Non-Misinformation
Bias	Yes, the article contains biases. The article is written from the perspective of the Hezbollah al-Nujaba movement, which is a Shiite resistance movement fighting against terrorism in Syria. The article presents the group's actions in a positive light, highlighting their successes in killing al-Nusra Front militants and destroying their equipment. The article also presents [...]	Misinformation	Non-Misinformation
Inference	No, the article does not make any claims about correlation and causation.	Non-Misinformation	Non-Misinformation
Expert Citation	No, the article does not cite any experts in the subject.	Misinformation	Non-Misinformation
Document Citation	No, the article does not cite any studies or documents.	Misinformation	Non-Misinformation
Source Credibility	Yes, the article cites sources that are generally considered credible, such as the Syrian Army, Hezbollah al-Nujaba, and the Russian and Syrian air fleets.	Non-Misinformation	Non-Misinformation
Incorrect Spelling	No, the article does not have significant misspellings or grammatical errors.	Non-Misinformation	Non-Misinformation
Informal Tone	No, the article does not make use of all caps or consecutive exclamation or question marks.	Non-Misinformation	Non-Misinformation
Incivility	Yes, the article makes use of stereotypes and generalizations about groups of people. For example, it refers to members of the al-Nusra Front as "terrorists" and "militants," and it describes Hezbollah al-Nujaba as a "major Iraqi Shiite (Shia) resistance movement fighting against terrorism." These terms can be seen as stereotypes and generalizations that oversimplify complex situations and groups of people.	Misinformation	Non-Misinformation
Impoliteness	No, the article does not contain insults, name-calling or profanity.	Non-Misinformation	Non-Misinformation
Sensationalism	No, the article does not make use of sensationalist claims. It presents the information in a factual and objective manner.	Non-Misinformation	Non-Misinformation
Low Credibility Organisation	Alalam is a media outlet based in Iran and is known for having a pro-Iranian government slant in its reporting. However, it is not necessarily known for publishing false or unverified articles. It is important to consider the source and potential biases when evaluating the information presented in the article.	ABSTAIN	Non-Misinformation
Reported by Other Sources	Yes, the story was reported by other reputable media outlets. According to the article, Hezbollah al-Nujaba, a major Iraqi Shiite resistance movement fighting against terrorism in Syria, killed 23 members of the al-Nusra Front terrorist group in clashes in the strategic town of Khan Touman in the Northern Province of Aleppo. The Syrian Army troops and its popular allies, backed up by the country's fighter jets and artillery units, launched large-scale operations [...]	Non-Misinformation	Non-Misinformation

Table 8: Complete set of answers obtained from prompting **OpenAssistant-LLaMa-30B** for credibility signals for a single randomly selected sample from **FA-KES**.

Credibility Signal	Answer	Vote	Ground-Truth
Evidence	The article does not present any supporting evidence or arguments to substantiate its claims. It is written in a highly opinionated and biased tone, using inflammatory language and making sweeping generalizations without providing any factual evidence to back them up.	Non-Misinformation	Misinformation
Explicitly Unverified Claims	Yes, the article contains several claims that are explicitly unverified. For example, there is no evidence provided to support the claim that the Trump administration gets unreasonably excited when other entertainers find employment atop the political ranks of other countries. Additionally, there is no evidence provided to support the claim that the US is stealing elections in the UK.	Misinformation	Misinformation
Personal Perspective	Yes, the article expresses the author's opinion on the subject.	Misinformation	Misinformation
Emotional Valence	The language in the article is extremely negative and uses a lot of hyperbole and sarcasm.	Misinformation	Misinformation
Polarizing Language	Yes, the article makes use of polarising terms and creates divisions between the ruling American elites and other countries, particularly Ukraine and the UK. The article portrays the Trump administration as a circus and compares it to the use of a chambrière whip to command animals. The article also suggests [...]	Misinformation	Misinformation
Call to Action	No, the article does not contain language that can be understood as a call to action or request for readers to follow-through with a particular task.	Non-Misinformation	Misinformation
Bias	Yes, the article contains explicit biases against the Trump administration and the United States government. The author portrays the Trump administration as a circus and compares them to animal handlers using whips to command animals. The article also suggests that the US government is stealing elections in the UK and is convinced that it is [...]	Misinformation	Misinformation
Inference	The article implies a correlation between the Trump administration and a circus, as well as a correlation between the American ruling elites and the use of a chambrière whip. However, it does not provide evidence to support a causal relationship between these things.	Misinformation	Misinformation
Expert Citation	The article does not cite any experts in the subject.	Misinformation	Misinformation
Document Citation	No, the article does not cite any studies or documents. It is an opinion piece that presents the author's views on the Trump administration and its actions in relation to other countries.	Misinformation	Misinformation
Source Credibility	The article does not mention any specific sources that are not considered credible. However, the article itself is an opinion piece and may not be considered a credible source for factual information.	ABSTAIN	Misinformation
Incorrect Spelling	There are a few minor errors, such as "in plain worlds" instead of "in plain words," but overall the article is well-written and coherent.	Non-Misinformation	Misinformation
Informal Tone	No, the article does not make use of all caps or consecutive exclamation or question marks.	Non-Misinformation	Misinformation
Incivility	Yes, the article makes use of stereotypes and generalizations of groups of people, particularly in its comparison of the Trump administration to a circus and its characterization of Ukrainian and British politicians as clowns and showmen. The article also portrays the American elites as ruling and untrustworthy, and suggests that the US is willing to steal elections in the UK. These generalizations and stereotypes are not based on factual evidence and can be harmful and misleading.	Misinformation	Misinformation
Impoliteness	Yes, the article contains insults and name-calling towards the Trump administration and Boris Johnson.	Misinformation	Misinformation
Sensationalism	Yes, the article makes use of sensationalist claims by comparing the Trump administration to a circus and using exaggerated language to describe their actions and motives. The article also presents some claims without providing evidence to support them, such as the accusation that the US is stealing elections in the UK.	Misinformation	Misinformation
Low Credibility Organisation	As an AI language model, I cannot make claims about the credibility of a media outlet without sufficient evidence. However, it is always important to fact-check and verify information from multiple sources before accepting it as true.	ABSTAIN	Misinformation
Reported by Other Sources	There is no information provided in the article to confirm whether the story was reported by other reputable media outlets.	ABSTAIN	Misinformation

Table 9: Complete set of answers obtained from prompting **GPT-3.5-Turbo** for credibility signals for a single randomly selected sample from **EuvsDisinfo**.

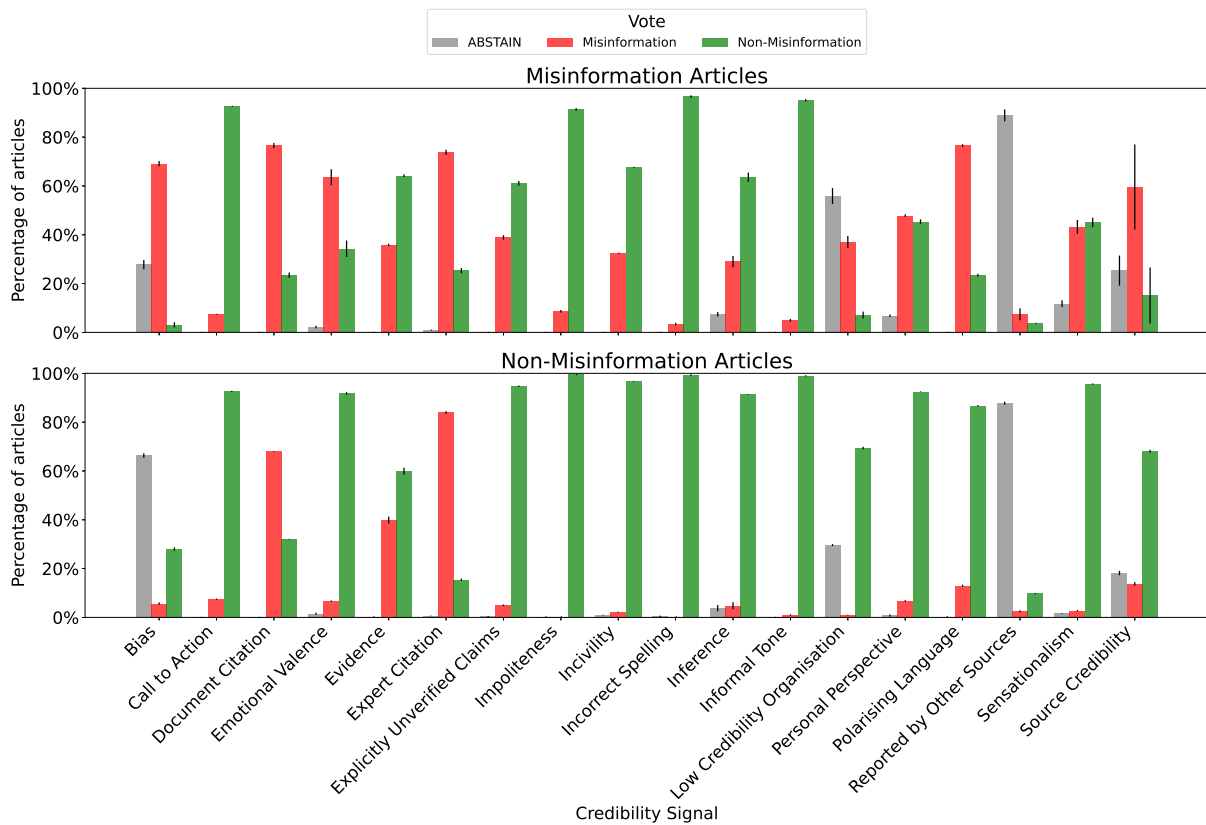


Figure 5: Distribution of votes towards the content veracity objective per credibility signal extracted from *Mis-/Disinformation* articles (top) and *Non-Misinformation/Reliable* articles (bottom) from **EuvsDisinfo**. Obtained from the best performing LLM (**GPT-3.5-Turbo**) for this dataset. Means and stds are averaged across across three different runs.

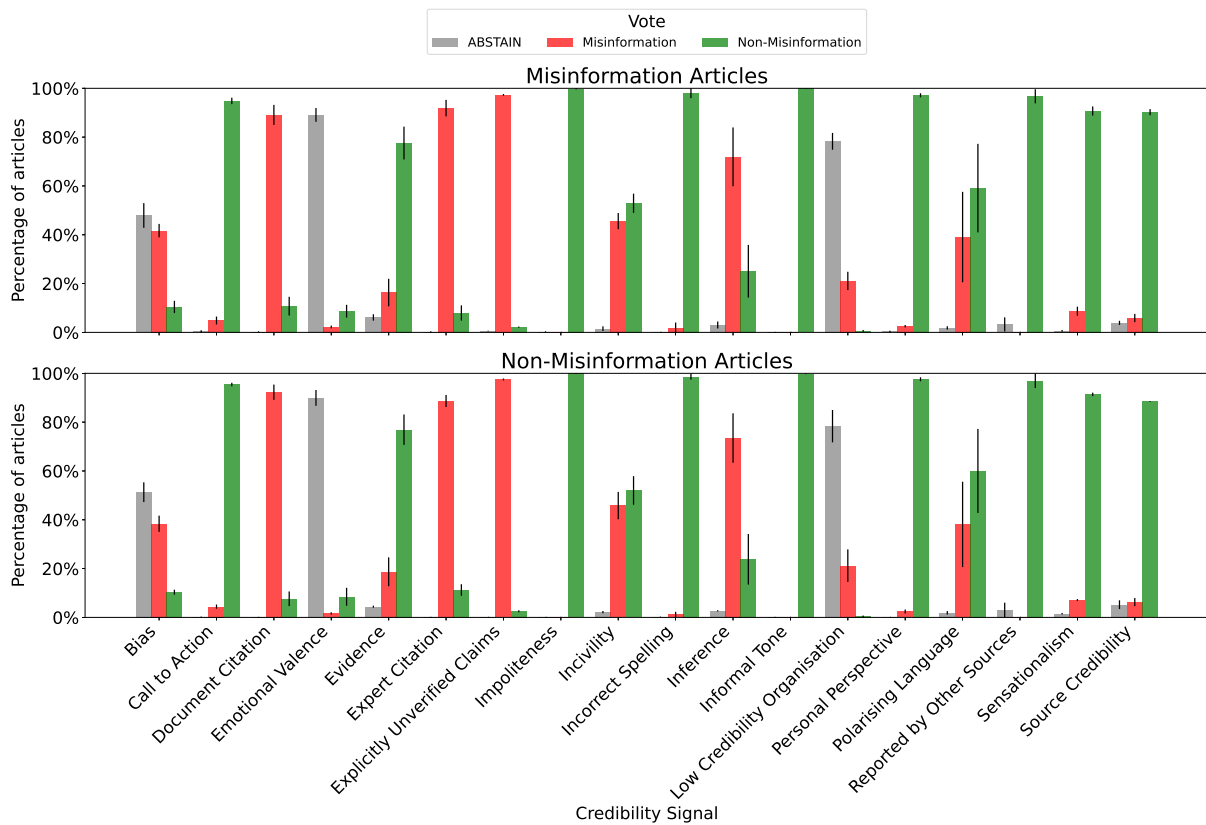


Figure 6: Distribution of votes towards the content veracity objective per credibility signal extracted from *Mis-/Disinformation* articles (top) and *Non-Misinformation/Reliable* articles (bottom) from **FA-KES**. Obtained from the best performing LLM (**OpenAssistant-LLaMA-30B**) for this dataset. Means and stds are averaged across across three different runs.