
REALM — Rationale-Enhanced Aggregation from LLM for Misinformation Detection

Xuwei Ding

Major in Data Science
xding95@wisc.edu

Yujun Che

Major in math, statistics and history
yche25@wisc.edu

Zelin Shen

Major in Statistics and Entomology
zshen265@wisc.edu

Sadie Liu

Major in Data Science and Information Science
sliu878@wisc.edu

Abstract

1 The spread of fake news online is a huge problem, raising a serious threat to
2 public trust. Although existing automatic detection methods have made significant
3 progress, they are not efficient enough to detect misinformation. In this study,
4 we introduce **Rationale-Enhanced Aggregation from LLM for Misinformation**
5 **Detection (REALM)**, a hybrid framework that integrates the reasoning power of a
6 large language model (GPT-4o) and the efficiency and accuracy of a small language
7 model (fine-tuned BERT) to enhance misinformation detection. REALM leverages
8 LLM to generate human-like reasoning, which is then fused with raw news content
9 as input to SLM, allowing the small model to better interpret contextual signals
10 and logical structures. We evaluate various prompting strategies (zero-shot, few-
11 shot, and idea chaining) and fine-tuning methods on a bilingual misinformation
12 dataset, further use the conclusions in model development. An ablation study
13 confirms the complementary strengths of LLM theory and SLM classification.
14 Experimental results show that while SLM outperforms LLM in terms of standalone
15 accuracy, combining them into REALM achieves superior performance in terms of
16 F1 and accuracy in both languages, with over 80% accuracy. Despite there exists
17 hallucination rate in LLM output, REALM maintains strong reliability, highlighting
18 its potential for multilingual misinformation detection.

1 Introduction

The unchecked spread of misinformation across digital platforms presents an urgent threat to society, undermining public trust in critical domains such as health, politics, and crisis response [33, 41]. While numerous automated detection methods have been proposed—ranging from lexical and sentiment analysis to leveraging user comments and social network structures—these approaches often face challenges in scalability, domain adaptation, and robustness under real-world conditions [38, 32].

Recent breakthroughs in Large Language Models (LLMs), exemplified by GPT-4o, offer powerful contextual reasoning and the ability to generate coherent, human-like rationales [3, 16]. However, the high computational cost and occasional hallucinations of LLMs constrain their direct application in large-scale classification tasks. In parallel, Small Language Models (SLMs) such as BERT have demonstrated exceptional performance on downstream classification benchmarks due to their efficiency, stability, and scalability, yet they inherently lack explicit multi-step reasoning capabilities [7, 17].

To bridge this gap, we introduce REALM (Rationale-Enhanced Aggregation from LLM for Misinformation Detection), a hybrid framework that fuses LLM-generated explanations with SLM-based classification. Specifically, we leverage GPT-4o to produce natural language rationales for each news item, concatenate these rationales with the original text, and feed the enriched input into a fine-tuned BERT model. This design enables multi-perspective analysis and interpretability while incurring only a modest increase in computational overhead.

We evaluate REALM on a bilingual (English and Chinese) misinformation dataset, comparing various LLM prompting strategies and SLM fine-tuning configurations. Our experiments show that REALM consistently outperforms both LLM-only and SLM-only baselines in accuracy and F1-score—achieving over 80% accuracy in both languages—while maintaining minimal additional cost. Furthermore, a human evaluation of rationale hallucination rates demonstrates the robustness of our approach against noisy explanations, underscoring REALM’s potential for practical, multilingual misinformation detection.

2 Related Work

2.1 Misinformation Detection

Automated misinformation detection has been explored using a wide range of neural architectures [26, 31, 36]. Apart from extracting lexical [27] and sentiment features [28] within the news article text, many methods incorporate auxiliary features to supplement veracity prediction, including user comments [31], news environments [30], knowledge bases [6, 18], temporal patterns from users [29], and social graphs [21, 42]. Recent studies also seek to address challenges including temporal shift [12], entity bias [43] and domain shift [19, 20, 44] in misinformation detection scenarios. We are motivated by Adaptive Rationale Guidance (ARG) [10], which combined LLM-generated rationales with a lightweight classifier to boost interpretability. Unlike ARG’s reliance on GPT-3.5 turbo and MLP-based prediction, which restrict reasoning richness and adaptability, and incurs additional computational expense, our approach leverages GPT-4o for richer rationales and integrates diverse BERT models for joint content–rationale analysis, significantly enhancing detection performance.

2.2 LLM Capabilities and Misinformation

Large language models (LLMs) [2, 22, 23] have demonstrated remarkable reasoning capabilities that even match or surpass human performance in certain scenarios [35]. In particular, chain-of-thought (CoT) prompting has been shown to elicit multi-step reasoning in LLMs [15, 35], and systematic prompt design can significantly affect the factual knowledge these models extract [14]. However, the impressive strengths of LLMs have also attracted increasing attention towards LLM-generated misinformation [4]. Recent investigations have found that LLMs can act as high-quality misinformation generators [24, 25], and that LLM-generated misinformation is generally harder to detect [5]. On a related front, recent work explores the role of LLMs as fact-checkers [24] and fake news detectors [9, 5], and leverage the commonsense reasoning capabilities to elicit supplementary explanations from LLMs [11, 39] that facilitate a wide range of tasks.

3 REALM Methodology

Our method combines the complementary strengths of large language model (gpt-4o) and small language model (pre-trained BERT model) to enhance misinformation detection. Using gpt-4o’s advanced reasoning capabilities, we generate human-like analytical rationales through API calls. These rationales, which mimic human thinking patterns, are then integrated with the original news content to create enhanced input data for further model training.

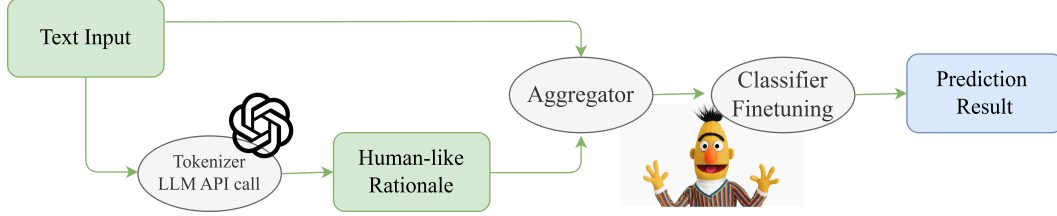


Figure 1: REALM Workflow

This rationale-enhanced input is processed by the BERT model, enabling it to evaluate the trustworthiness with multiple perspectives. By feeding the generated gpt-4o reasoning, we suspect that the BERT model could now identify the underlying logical of the text. Such ability to understand the human-like rationale is a robust model (shown in Figure. 1) for identifying misinformation.

3.1 Problem Definition and Axioms

Let $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^N$ be a dataset of news items $X_i \in \mathcal{S}$ with binary labels $y_i \in \{0, 1\}$ (1 =FAKE, 0 =REAL). We denote by

$$R = g_{\text{LLM}}(X)$$

the “rationale” generated by a large language model (gpt-4o) for input X . Our goal is to learn a classifier

$$f_\theta : \mathcal{S} \times \mathcal{R} \rightarrow \{0, 1\},$$

taking both the original text and its rationale as input, that minimizes classification risk under a standard cross-entropy loss.

Axiom 1 (Latent Rationale). For each X there exists an (unobserved) human-like rationale R^* such that $\Pr(f(X, R^*) = y) = 1$.

Axiom 2 (Approximate Reasoning). The LLM can produce $R \approx R^*$ with probability $\delta > 0.5$.

Axiom 3 (Complementarity). A fine-tuned SLM (BERT) excels at pattern recognition, while an LLM excels at multi-step reasoning; combining them strictly improves Bayes risk.

3.2 REALM Pipeline

We assemble four modules (Fig. 1):

1. Rationale Generation.

$$R = \text{GPT4o}(X),$$

where $R \in \mathcal{R}$ is a concatenation of multi-perspective analyses.

2. Fusion.

$$X' = [X; R] \in \mathcal{S} \times \mathcal{R}$$

by simple string-level concatenation.

3. Encoding & Attention. X' is tokenised and mapped to an embedding matrix $E \in \mathbb{R}^{n \times d}$, followed by a single-head scaled-dot-product attention layer

$$E' = \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V,$$

with $Q=EW_Q$, $K=EW_K$, $V=EW_V$. The output E' is fed to a pre-trained BERT encoder $\mathcal{B}(\cdot)$, yielding sentence representation $H = \mathcal{B}(E') \in \mathbb{R}^{d'}$.

102 **4. Classification (MLP).**

$$z = g_\phi(H) \in \mathbb{R}^2, \quad p = \text{softmax}(z), \quad \hat{y} = \arg \max_c p_c.$$

103 **3.3 Training Objective**

104 We freeze the LLM parameters and train the fusion, attention, BERT and MLP end to end by
105 minimizing

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum_{i=1}^N \sum_{c=0}^1 \mathbb{I}\{y_i = c\} \log p_{i,c},$$

106 where $\theta = \{W_Q, W_K, W_V, \mathcal{B}\text{-params}, \phi\}$. Updates use AdamW:

$$\theta_{t+1} = \theta_t - \eta \nabla_{\theta_t} \mathcal{L}(\theta_t).$$

107 This completes our formal specification of REALM, a rationale-enhanced framework that unites
108 LLM reasoning with SLM efficiency for robust misinformation detection.

109 **4 Experiment Setting**

110 **4.1 Datasets**

111 We require the dataset from the Laboratory of Intelligent Information Processing at the Chinese
112 Academy of Sciences for our model. The dataset contains approximately 15,000 entries in English and
113 Chinese. In detail, it includes 4,900 true and 1,500 fake entries in English, and 4,600 true and 4,500
114 fake entries in Chinese. English-language sources were sourced from *Gossip Cop*, a fact-checking
115 platform specializing in celebrity news. Chinese sources were drawn from social platform *Weibo* and
116 verified against official reports. Both platforms are publicly accessible, and their content is widely
117 read and discussed. Each entry contains the content, with around 150 words on its summary; and the
118 label on its trustworthiness.

119 We also randomize the dataset, then partitioned it into train, validation and test subsets with an
120 approximate 60-20-20% split, as Table. 1 indicaets. This allocation ensures our model has large
121 enough data to learn from, and with sufficient amount of data for validating the performance and
122 report the accuracy and F-1 scores.

#	Chinese			English		
	Train	Val	Test	Train	Val	Test
Real	2,331	1,172	1,137	2,878	1,030	1,024
Fake	2,873	779	814	1,006	244	234
Total	5,204	1,951	1,951	3,884	1,274	1,258

Table 1: Train-Valid-Test Split for Dataset

123 **4.2 Explore LLM prompting techniques**

124 Given that gpt-4o has numerous layers and parameters, fine-tuning the model is impractical. Instead,
125 we focus on exploring its capabilities to detect misinformation by trialing the following prompting
126 strategies:

127 Zero-shot setting, where no additional context or examples were provided—only a direct prompt
128 asking the model to classify the trustworthiness of a given news article. English prompt is Prompt 4
129 and Chinese prompt is Prompt 7.

130 Few-shot prompting by including three examples in each language. Two of them are true news articles
131 and one fake news article. This provided the model with relevant context for making classifications.
132 English prompt is Prompt 5 and Chinese prompt is Prompt 8.

133 We also examined the impact of Chain-of-Thought (CoT) prompting. In this setup, we used few-
134 shot examples paired with explanations that illustrate why certain news items are trustworthy or

135 problematic. These CoT prompts were designed to guide the model to reason more like a human, by
136 explicitly modeling the rationale behind each classification. English prompt is Prompt 6 and Chinese
137 prompt is Prompt 9.

138 After evaluating the effectiveness and efficient of all prompting methods, we

139 4.3 Explore SLM fine-tuning techniques

140 Conducting experiments using a small language model by fine-tuning a pre-trained BERT based
141 model that specified for the on our original misinformation detection dataset. The performance was
142 evaluated based on its classification accuracy towards the test data, and the accuracy of validation
143 data serves as the indicator for overfitting.

144 We systematically tuned many hyperparameters to balance learning capacity against overfitting. We
145 focused on the number of epochs to prevent over/under training, batch size to balance frequency of
146 updates and efficiency, learning rate to avoid instability, dropout rate for better regularization, weight
147 decay to penalize large weights and promote smoother solution, and above all the patience for early
148 stopping to catch the optimal peak in validation accuracy before overfitting sets in. All in order to
149 achieve the best possible model-training accuracy, using EANNT[40]: A model that learns effective
150 signals using auxiliary adversarial training, aiming at removing event-related features as much as
151 possible, Publisher-Emo[34]: A model that fuses a series of emotional features with textual features
152 for fake news detection, and ENDEF[42]: A model that removes entity bias via causal learning for
153 better generalization on distribution-shifted fake news data. as baseline.

154 4.4 Develop REALM Model

155 The REALM model is developed by enhancing the training dataset with LLM-generated rationales,
156 then further adopt full parameter fine-tuning. This augmentation integrates human-like reasoning into
157 the data, improving the SLM’s ability to detect misinformation.

158 4.4.1 Enhance Training Dataset

159 To enhance the training dataset, we applied a cross-entropy loss function. Since the news contents
160 are labeled as either true or fake, this function is particularly suitable for binary classification tasks.
161 Through augmenting the original dataset with LLM-generated rationales, the new training dataset
162 incorporates human-like reasoning with the trustworthiness result. The new training dataset, therefore,
163 could improve the SLM’s ability to learn from both labels and develop the rationale behind each
164 label.

165 4.4.2 Training Steps

166 We classify news articles with a three-layer MLP attached to the last layer, [CLS] token, of bert
167 models, treating each hidden layer’s width as tunable parameters. Input embeddings pass through
168 two hidden layers (e.g. 384→128 units). Increasing layer width to boosts model capacity. Our table
169 illustrates performance of these Full Fine-tuning on bert-base-uncased, and bert-large-uncased when
170 using our English dataset, then bert-base-chinese when using our Chinese dataset. In addition, we also
171 tried using xlm-roberta-base for both english and chinese dataset as well as chinese-roberta-wwm-ext
172 with our chinese dataset.

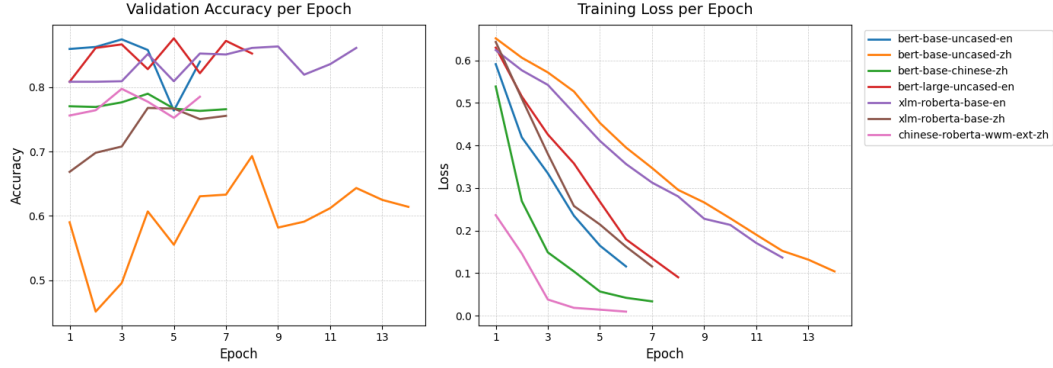


Figure 2: Validation Accuracy and Training Loss per Epoch for Seven Transformer Models on the REALM Task

5 Evaluation

We investigate the effectiveness of LLM and SLM in detecting misinformation across Chinese and English news content. The study reveals the incoherent result of prompting strategies for LLM across languages. In contrast, a fine-tuned BERT model consistently performs well across both languages. It guides us to adopt structured zero-shot prompting. We further evaluate our proposed REALM model with accuracy and F1 scores. In addition, human evaluation on LLM generated rationale indicates there exists hallucination effect. We need to be cautious towards LLM generated rationale for misinformation detection.

5.1 Evaluate LLM-only and SLM-only Model

The results are shown in Table. 2. It is observed that few shots and chain-of-thought prompting techniques remarkably enhances the large language model’s ability to detect misinformation in Chinese news. However, those two strategies undermines its performance when tasked with English content. Pre-trained BERT model, on the other hand, has higher a accuracy on misinformation detection across two languages.

Model	Usage	Chinese	English
GPT-4o	Zero-Shot	0.741	<u>0.741</u>
	Few-Shot	0.750	0.679
	Few-Shot CoT	<u>0.783</u>	0.652
BERT	Fine-tuning	0.789 (+0.006%)	0.791 (+0.05%)

Table 2: GPT-4o and pretrained BERT accuracy

The inconsistency in the accuracy of the gpt-4o evaluation of misinformation across different languages indicates the black box feature of LLM. It raises the question of its reliability as a direct tool for misinformation detection. However, gpt-4o provides logically coherent analyses when we test the Chain of Thought prompting technique. It is able to assess a variety of key features that we humans would think about, including quotes, tones, and details.

5.2 Evaluate Generated Rationale

Through the discussion on the ability on providing analysis on misinformation, we determine to use the zero-shot prompting strategy with an explicitly defined structure for LLM generated rationales on whether the news contents are real or fake. By providing guidelines for the model’s analytical framework, including assigning the model with objective roles, key factors to analyze response format and length, we standardize the procedure for LLM to evaluate rationales. This structured prompting

Model	Chinese				English			
	macF1	Acc.	F1 _{real}	F1 _{fake}	macF1	Acc.	F1 _{real}	F1 _{fake}
G1: LLM-Only GPT-4o	0.741	0.750	0.789	0.693	0.741	0.777	0.789	0.693
G2: SLM-Only	Baseline	0.749	0.752	0.719	0.779	0.791	0.887	0.933
	EANN _T	0.754	0.756	0.773	0.736	0.763	0.864	0.918
	Publisher-Emo	0.761	0.763	0.784	0.738	0.766	0.868	0.920
	ENDEF	0.765	0.766	0.779	0.751	0.768	0.865	0.918
G3: LLM+SLM	OURS	0.784	0.786	0.805	0.764	0.783	0.879	0.928
		(+0.00%)	(+0.00%)	(+0.001%)	(+0.00%)	(-0.007%)	(+0.001%)	(+0.002%)
	SuperICL	0.757	0.759	0.779	0.734	0.736	0.864	0.920
	ARG	<u>0.784</u>	<u>0.786</u>	<u>0.804</u>	<u>0.764</u>	0.790	<u>0.878</u>	<u>0.926</u>
	ARG-D	0.771	0.772	0.785	0.756	0.778	0.870	0.921

Table 3: Performance of the ARG and its variants alongside LLM-only, SLM-only, and combined methods. The best two results in macro-F1 and accuracy are bolded and underlined, respectively.

structure helps LLM to generate rigorous and clear rationale across languages, thus avoiding the analytical nuance derived from language differences and prompting unclearness. Example prompts in both languages are appended as Table 10 and Table 11.

5.3 Evaluate REALM model

We evaluate our REALM model through accuracy and F-1 scores. The Table 3 presents the unique advantage for our REALM model in identifying the real news. It is observed that the overall accuracy rate is around 80%, with a high F1 real score. There are only 20% real information in Chinese is mislabeled, and that figure for English news is only 10%. Almost all of which are higher than or equal to our baseline SuperICL [37]: It exploits the SLM as a plug-in for the in-context learning of the LLM by injecting the prediction and the confidence for each testing sample into the prompt, ARG and ARG-D. The low false positive rate ensures that legitimate news articles are rarely misclassified as fake, preserving their credibility and accessibility in real-world applications like fact-checking platforms. This reliability aligns with the need for precision in combating misinformation, since label too much true information could harm the public trust towards the detection machine. The accuracy of both languages are also promising. By prioritizing both accuracy and F1-Real, REALM achieves a balanced performance that safeguards true content while effectively flagging deceptive narratives.

5.4 Human evaluation of LLM’s hallucinations

To investigate how LLM hallucinations affect our model’s performance, we randomly sampled 1,000 rationale generations from both the Chinese and English datasets for human evaluation. We found that the hallucination rate in each dataset is roughly 30%—in other words, about 30% of the “truths” provided by the LLM do not align with the ground-truth answers. This flawed and misleading rationale generation process may therefore constrain our model’s overall performance.

6 Ablation Study with Quantitative Analysis

To better isolate the contributions of each component in our framework, we conduct an ablation study that systematically evaluates the impact of the SLM detector, LLM-generated rationales, and the core modules of our model, REALM, on the final detection performance.

6.1 SLM is a Good Detector

As shown in Table 2, BERT (SLM) fine-tuning achieves macro-F1 scores of 0.789 (Chinese) and 0.791 (English), clearly surpassing gpt-4o under zero-shot (0.741/0.741), few-shot (0.750/0.679), and few-shot CoT (0.783/0.652) prompting strategies. These results emphasize that a carefully fine-tuned small language model not only functions as a highly effective detector, outperforming more

sophisticated LLM prompting approaches, but also significantly reduces inference costs. In contrast, our experiments with the LLM (gpt-4o) incurred approximately \$20 in total testing expenses, underscoring the cost-effectiveness of utilizing an SLM like BERT.

6.2 LLM Could Provide Useful Rationales

Table 3 compares the performance of OURS (our main LLM+SLM REALM model), the two baseline variants (ARG and ARG-D), and various LLM-only, SLM-only, and hybrid methods. From the results we draw the following observations: **1)** Ours achieves the highest macro-F1 and accuracy across both languages, confirming its superior effectiveness. **2)** Our LLM+SLM framework (REALM) consistently outperforms both the LLM-only and SLM-only baselines, demonstrating that combining news content with prompted rationales yields clear gains. **3)** Among the baselines, the rationale-free ARG-D still surpasses all other compared methods except ARG, highlighting the value of distilled knowledge even without explicit rationales. **4)** In our ablation studies, removing either the LLM Judgment Predictor or the Rationale Usefulness Evaluator causes a marked drop in performance of our model, underscoring the critical roles of these components. Even the weaker baseline (ARG-D) outperforms all non-proposed methods, emphasizing the importance of our news–rationale interaction design.

7 Conclusion and Discussion

We studied the role of large language models in fake-news detection and how to harness their strengths to boost performance. Although the LLM (gpt-4o) did not match the task-specific SLM (BERT) in raw accuracy, it provided rich, informative rationales that complemented BERT’s predictions. Motivated by this, we developed the REALM framework to flexibly fuse the complementary capabilities of small and large LMs. Our experiments show that REALM performs on par with—or better than—existing baselines. We believe REALM can be extended to other tasks and will enable more efficient and cost-effective use of LLMs in the future.

7.1 Limitations

Our findings reveal that, despite their powerful analytical abilities, LLMs still struggle to fully leverage their internal knowledge in contexts that involve complex real-world backgrounds. To unlock their potential, we may need novel prompting strategies and a deeper understanding of their inner workings. Specifically: **1)** We only evaluated rationale generation and “LLM-only” performance on gpt-4o; we were unable to assess other closed-source LLMs (e.g., Claude [1], DeepSeek [8]) or open-source models. **2)** Our analysis focused solely on textual semantics and overlooked structural relationships—such as the contextual threads and user-user interactions that form graph structures in social-media misinformation detection. **3)** Our per-epoch accuracy curves exhibited high variance; more parameter-sensitivity experiments and alternative loss functions are needed to improve robustness. **4)** The effectiveness of knowledge extraction from LLMs depends heavily on the chosen prompt templates. Adaptive prompt-design methods may yield more consistent results.

7.2 Future Work

1) Since generating LLM rationales incurs expensive API calls, we plan to distill a smaller model that achieves comparable detection performance without relying on explicit rationales. **2)** We will explore different BERT variants and apply parameter-efficient fine-tuning techniques (e.g., LoRA [13]) to LLMs in order to uncover further performance gains.

8 Report Use of AI in This Report

We ask gpt-4o to polish our grammar and language style with the Prompt 12

We also ask few LaTeX related formatting questions to enhance the readability of our work.

274 9 Contribution

275 **Xuwei Ding** led the overall project, implemented and fine-tuned the REALM codebase, conducted
276 the literature review, and independently drafted Sections 6 and 7 as well as substantial portions of
277 Sections 1, 2, and 3.

278 **Yujun Che** served as second author, designed the REALM workflow diagram, conducted the
279 evaluation of different LLM prompting strategies, and wrote Section 3, all of Section 4, and parts of
280 Section 5.

281 **Zelin(Tony) Shen** acted as third author, run multiple model experiments, gathered and compared
282 performance data, plotted the associated figures, and contributed to writing Sections 4 and 5.

283 **Shuyu(Sadie) Liu** acted as fourth author, performed the literature collection, and wrote Sections 1
284 and 2.

References

- [1] Anthropic. Introducing the Next Generation of Claude: Claude 3 Model Family. <https://www.anthropic.com/news/claude-3-family>, 2024. Accessed: 2025-05-03.
- [2] Tom Brown, Benjamin Mann, and Nick Ryder et al. Language models are few-shot learners. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901, 2020.
- [3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [4] Canyu Chen and Kai Shu. Combating misinformation in the age of llms: Opportunities and challenges. *arXiv preprint arXiv:2311.05656*, 2023.
- [5] Canyu Chen and Kai Shu. Can llm-generated misinformation be detected? In *International Conference on Learning Representations (ICLR)*, 2024.
- [6] Limeng Cui, Haeseung Seo, Maryam Tabar, Fenglong Ma, Suhang Wang, and Dongwon Lee. Deterrent: Knowledge guided graph attention network for detecting healthcare misinformation. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 492–502, 2020.
- [7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [8] DeepSeek-AI et al. DeepSeek-V3 Technical Report. Technical Report arXiv:2412.19437, arXiv, 2024. Accessed: 2025-05-03.
- [9] Kellin Pelrine et al. Towards reliable misinformation mitigation: Generalization, uncertainty, and gpt-4. *arXiv preprint arXiv:2305.14928*, 2023.
- [10] Beizhe Hu, Wenxuan Dai, Meng Liu, and Kai-Wei Chang. Adaptive rationale guidance for news verification with language models. *arXiv preprint arXiv:2309.12247*, 2023.
- [11] Beizhe Hu, Qiang Sheng, Juan Cao, and et al. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI)*, 2024.
- [12] Beizhe Hu, Qiang Sheng, Juan Cao, Yongchun Zhu, Danding Wang, Zhengjia Wang, and Zhiwei Jin. Learn over past, evolve for future: Forecasting temporal trends for fake news detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 116–125, 2023.
- [13] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.
- [14] Zhe Jiang, Chenguang Xu, Qun Liu, Chen Liang, and Maosong Sun. Systematic prompt design for factual error correction. *Transactions of the Association for Computational Linguistics*, 8:440–454, 2020.
- [15] Yushi Kojima, Shixiang Shane Gu, Maxwell Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6319–6338, 2022.
- [16] Yang Liu, Ryotaro Shimizu, Rahul Gupta, Yang Liu, Jinyu Kang, and Jieyu Zhao. Deficiencies of language models on massive knowledge bases. *arXiv preprint arXiv:2305.06399*, 2023.
- [17] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

- [18] Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. Md-fend: Multi-domain fake news detection. *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM)*, pages 3343–3347, 2021.
- [19] Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. Md-fend: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 3343–3347, 2021.
- [20] Qiong Nan, Danding Wang, Yongchun Zhu, Qiang Sheng, Yuhui Shi, Juan Cao, and Jintao Li. Improving fake news detection of influential domain via domain- and instance-level transfer. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 2834–2848, 2022.
- [21] Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. Fang: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1165–1174, 2020.
- [22] OpenAI. Chatgpt: Optimizing language models for dialogue. Technical report, 2022. <https://openai.com/blog/chatgpt>.
- [23] OpenAI. Gpt-4 technical report. Technical report, arXiv preprint arXiv:2303.08774, 2023.
- [24] Liangming Pan, Xiaobao Wu, and Xinyuan Lu et al. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6981–7004, 2023.
- [25] Yikang Pan, Liangming Pan, and Wenhui Chen et al. On the risk of misinformation pollution with large language models. *Findings of ACL*, pages 1389–1403, 2023.
- [26] Kellin Pelrine, Jacob Danovitch, and Reihaneh Rabbany. The surprising performance of simple baselines for misinformation detection. In *Proceedings of the Web Conference (WWW)*, pages 3432–3441, 2021.
- [27] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 231–240, 2018.
- [28] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 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 (EMNLP)*, pages 2931–2937, 2017.
- [29] Natali Ruchansky, Sungyong Seo, and Yan Liu. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 IEEE International Conference on Data Mining (ICDM)*, pages 792–801, 2017.
- [30] Qiang Sheng, Juan Cao, and Xu Yao Zhang et al. Zoom out and observe: News environment perception for fake news detection. pages 4543–4556, 2022.
- [31] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. Defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 395–405, 2019.
- [32] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.
- [33] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.
- [34] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD ’18*, page 849–857, New York, NY, USA, 2018. Association for Computing Machinery.

- 380 [35] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Hoa T. Nguyen, Quoc V. Le,
381 Yuhuai Wu, and Denny Zhou. Chain of thought prompting elicits reasoning in large language
382 models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837,
383 2022.
- 384 [36] Fengqiang Wu, Ziqi Zhu, Chenghao Liu, and Zhifang Sui. Safe: A robust cnn-based architecture
385 for fake news detection. In *Proceedings of the 27th ACM International Conference on*
386 *Information and Knowledge Management (CIKM)*, pages 1234–1243, 2022.
- 387 [37] Canwen Xu, Yichong Xu, Shuohang Wang, Yang Liu, Chenguang Zhu, and Julian McAuley.
388 Small models are valuable plug-ins for large language models, 2023.
- 389 [38] Jian Zeng, Zhenhua Tang, Yitong Liu, and Jicheng Guo. Ai-enabled fake news detection: A
390 review. *Applied Sciences*, 11(10):4500, 2021.
- 391 [39] Zhiping Zeng, Bowen Tan, and Hongyu Lin. Evaltree: Revealing vulnerabilities via hierarchical
392 capability trees. In *NeurIPS*, 2025.
- 393 [40] Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. Mining dual
394 emotion for fake news detection. In *Proceedings of the Web Conference 2021*, WWW '21, page
395 3465–3476, New York, NY, USA, 2021. Association for Computing Machinery.
- 396 [41] Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods,
397 and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40, 2020.
- 398 [42] Haiyu Zhu, Tianli Xu, Qingyu Zhou, and Xuanjing Huang. Improving fake news detection of
399 influential domain via domain- and instance-level transfer. *Proceedings of the 29th International*
400 *Conference on Computational Linguistics (COLING)*, pages 2834–2848, 2022.
- 401 [43] Yongchun Zhu, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, and Fuzhen Zhuang.
402 Generalizing to the future: Mitigating entity bias in fake news detection. In *Proceedings of the*
403 *45th ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*,
404 2022.
- 405 [44] Yongchun Zhu, Qiang Sheng, Juan Cao, Qiong Nan, Kai Shu, Minghui Wu, Jindong Wang,
406 and Fuzhen Zhuang. Memory-guided multi-view multi-domain fake news detection. *IEEE*
407 *Transactions on Knowledge and Data Engineering*, 2022.

Zero Shot
<p>Prompt: Analyze the following text and determine whether it describes a true event or is possibly fictional. If it is possibly true, output '0'. If it is possibly false, output '1'. Please <i>*only*</i> respond with the number 0 or 1, without any explanation or leading text. Text: <i>content</i></p> <p>Output: 0 or 1</p>

Table 4: Zero Shot Prompt

Few Shots
<p>Prompt: Analyze the following text and determine if it describes a real event or is likely fabricated. Output '0' if it is likely real. Output '1' if it is likely fake. Respond <i>*only*</i> with the digit 0 or 1. Here are some examples: Text: US immigration judge rules Palestinian Columbia student Khalil can be deported. Output: 0 Text: A law enforcement sniper assigned to former President Donald Trump's rally Saturday in Butler, Pennsylvania, says the head of the Secret Service ordered him not to shoot the suspect accused of attempting to assassinate Trump. Output: 1 Text: The Dane County Farmers Market kicked off its 53rd season Saturday under a sunny sky, brisk but rising temperatures and no worries of tariffs. Output: 0 Now, classify the following text: Text: <i>content</i></p> <p>Output: 0 or 1</p>

Table 5: 3 Shots Prompt

Few Shots with Chain of Thought

Prompt:

Analyze the following text and determine if it describes a real event or is likely fabricated. First, provide a step-by-step reasoning for your decision. Then, on a new line, provide the final answer in the format 'Final Answer: 0' if it is likely real, or 'Final Answer: 1' if it is likely fake.

Here are some examples demonstrating the process:

Text: US immigration judge rules Palestinian Columbia student Khalil can be deported.

Reasoning: The text reports a specific legal decision involving identifiable entities (US immigration judge, Columbia student Khalil) and a common legal action (deportation ruling). This aligns with typical news reporting formats for factual events.

Final Answer: 0

Text: A law enforcement sniper assigned to former President Donald Trump's rally Saturday in Butler, Pennsylvania, says the head of the Secret Service ordered him not to shoot the suspect accused of attempting to assassinate Trump.

Reasoning: The text makes an extraordinary and highly sensitive claim (an order not to shoot an assassination suspect) attributed to the head of the Secret Service via a sniper. Such a claim is sensational, lacks immediate corroboration within the text, involves high-level security protocols that make this scenario unlikely, and borders on conspiracy theory.

Final Answer: 1

Text: The Dane County Farmers Market kicked off its 53rd season Saturday under a sunny sky, brisk but rising temperatures and no worries of tariffs. Reasoning: The text describes a routine, verifiable local event (Dane County Farmers Market opening). It includes specific details (53rd season, location, weather) characteristic of factual local news reporting.

Final Answer: 0

Now, analyze the following text using the same process:

Text: *content*

Reasoning: [Provide your reasoning steps here]

Final Answer: [0 or 1]

Content:

In February, Alicia Silverstone and Christopher Jarecki announced they were divorcing after 20 years together. The couple has a six - year - old son together, Bear Blu. (Photo : Frazer Harrison, Getty Images) Alicia Silverstone has filed for divorce from her husband of nearly 13 years, actor and musician Christopher Jarecki. The divorce papers were filed in Los Angeles County Superior Court on Friday, according to the Associated Press. The Clueless star, 41, had separated from Jarecki, 47, in February after more than 20 years together as a couple. At the time the couple said in a statement that ‘ ‘ they still deeply love and respect each other and remain very close friends.’’ The papers state the couple will share custody of their 7 - year - old son, Bear Blue.

Output:

Reasoning: The text provides specific details about a celebrity divorce, including names (Alicia Silverstone and Christopher Jarecki), their professions, the duration of their relationship, and their child's name and age. It mentions the filing of divorce papers in a specific location (Los Angeles County Superior Court) and references a reputable source (Associated Press). The text also includes a statement from the couple, which is typical in public announcements of celebrity separations. These elements align with factual reporting on celebrity news.

Final Answer: 0

Table 6: 3 Shots Prompt with Chain of Thought

Zero Shot
<p>Prompt:</p> <p>分析以下文本，判断它描述的是真实事件还是可能为虚构。 如果可能为真，输出'0'。 如果可能为假，输出'1'。 请*仅*回应数字0 或1。 文本: <i>content</i></p> <p>Output:</p> <p>0 or 1</p>

Table 7: Zero Shot Prompt

Few Shots
<p>Prompt:</p> <p>分析以下文本，判断它描述的是真实事件还是可能为虚构。 如果可能为真，输出'0'。 如果可能为假，输出'1'。 请*仅*回应数字0 或1。 这里有一些例子： 文本：开局良好，多组数据折射我国一季度经济运行向好态势。国家信息中心监测数据显示，2月挖掘机销售台数同比增长52.8百分点，其中国内同比增长99.4百分点，为4年来最大单月涨幅。 输出：0 文本：大阪世博会中国馆亮点抢先看：搭乘“蛟龙”号，探寻深海奥秘。 输出：0 文本：台湾花莲县海域发生7.3级地震，导致浙江宁波一小区外墙开裂 输出：1 现在，请对以下文本进行分类： 文本: <i>content</i> 输出：</p> <p>Output:</p> <p>0 or 1</p>

Table 8: 3 Shots Prompt

Few Shots with Chain of Thought

Prompt:

分析以下文本，判断它描述的是真实事件还是可能为虚构。

首先，请提供你做此判断的分步推理过程。

然后，在新的一行，按照以下格式提供最终答案：如果可能为真，则为'Final Answer: 0'；如果可能为假，则为'Final Answer: 1'。

这里有一些演示该过程的例子：

文本：开局良好，多组数据折射我国一季度经济运行向好态势。国家信息中心监测数据显示，2月挖掘机销售台数同比增长52.8百分点，其中国内同比增长99.4百分点，为4年来最大单月涨幅。推理过程：该文本提供了具体的经济数据（挖掘机销售数量、增长率），引用了数据来源（国家信息中心），并进行了比较（4年来最大单月涨幅）。这符合标准经济新闻报道的格式，内容具有可验证性。

Final Answer: 0

文本：大阪世博会中国馆亮点抢先看：搭乘“蛟龙”号，探寻深海奥秘。推理过程：该文本描述的是对未来活动（大阪世博会中国馆）的一个特色预览。虽然涉及具体技术（“蛟龙”号），但本质是宣传性内容，介绍的是一个计划中的、真实的活动亮点。

Final Answer: 0

文本：台湾花莲县海域发生7.3级地震，导致浙江宁波一小区外墙开裂推理过程：文本声称一个地点（台湾花莲）的地震导致了非常遥远的另一个地点（浙江宁波）发生具体的物理损坏（外墙开裂）。虽然强震影响广泛，但这种跨越巨大地理距离的直接、具体的因果关系（特定小区外墙因此开裂）高度可疑，缺乏合理的物理传导机制解释。

Final Answer: 1

现在，请使用相同的过程分析以下文本：

文本：content

推理过程：[在此处提供你的推理步骤]

Final Answer: [0 或 1]

Content:

全国首家无人银行亮相大堂经理是智能机器人银行里没有忙碌的工作人员和拥挤的排队客户，而是各种机器人、智慧柜员机、存取款机、外汇兑换机等智能化设备，科技感十足。业内观点无人银行是银行物理网点的创新，是转型的一种选择，但并不意味着会成为一种趋势，线上银行才是未来银行业发展

Output:

推理过程：该文本描述了“全国首家无人银行”亮相的情景，具体提到“大堂经理是智能机器人”、“各种机器人、智慧柜员机、存取款机、外汇兑换机”等具体设施，具有较高的细节密度。该描述符合2018年中国建设银行在上海推出全国首家无人银行的真实事件，多个主流媒体曾对其进行报道，内容与文本高度吻合。此外，结尾部分引用了“业内观点”，指出无人银行是物理网点的一种创新但并非主流趋势，也是一种符合真实报道中常见评论语气的行业分析。这些信息的真实性和合理性都较高，具备可查证性。

Final Answer: 0

Table 9: 3 Shots Prompt with Chain of Thought

Rationale in English**Prompt:**

"You are a truthful news analyzer. Analyze the following news report purely based on its textual content and writing style. "

"List key points about the writing (level of detail, sources cited, quotes, consistency, tone, etc.) that indicate whether the news is likely true or false. "

"After listing these points, conclude with a sentence stating your judgment ('Based on these textual characteristics, the news is likely real' or '... likely fake')."

"limite the response to 150 words. "

"News: content"

"Your analysis:"

Content:

In February, Alicia Silverstone and Christopher Jarecki announced they were divorcing after 20 years together. The couple has a six - year - old son together, Bear Blu. (Photo : Frazer Harrison, Getty Images) Alicia Silverstone has filed for divorce from her husband of nearly 13 years, actor and musician Christopher Jarecki. The divorce papers were filed in Los Angeles County Superior Court on Friday, according to the Associated Press. The Clueless star, 41, had separated from Jarecki, 47, in February after more than 20 years together as a couple. At the time the couple said in a statement that they still deeply love and respect each other and remain very close friends." The papers state the couple will share custody of their 7 - year - old son, Bear Blue

Label:

0

Generated Rationale:

Your analysis: 1. ****Level of Detail****: The report provides specific details such as the names of the individuals involved, their ages, the duration of their relationship, and the age of their child. This level of detail is typical of factual reporting.

2. ****Sources Cited****: The article references the Associated Press as the source of information regarding the filing of divorce papers. The Associated Press is a reputable news organization, which lends credibility to the report.

3. ****Quotes****: The article includes a direct quote from the couple's statement about their separation, which adds authenticity to the report. The quote reflects a neutral and respectful tone, consistent with typical public statements made by celebrities during personal events.

4. ****Consistency****: The report is consistent in its details, such as the timeline of events (separation in February, filing in Los Angeles County Superior Court), and the age of their son, although there is a minor inconsistency in the spelling of the son's name (Bear Blu",

Table 10: Generated English Rationale.

Rationale in Chinese	
Prompt:	<p>你是一位诚实的新闻分析员。请仅根据以下新闻报道的文本内容和写作风格进行分析。</p> <p>列出文章的关键点（细节程度、引用来源、引文、一致性、语气等），以判断该新闻的真实性。</p> <p>列出这些要点后，请用一句话陈述你的判断（“根据这些文本特征，该新闻可能是真的”或“.....可能是假的”）。</p> <p>请将回复限制在200字以内。</p> <p>新闻：内容</p> <p>你的分析：</p>
Content:	<p>【能电鱼，更能电力找鱼：别以为电鳗只有把鱼电晕一招！】电鳗2米长的身体里有许多专门发电的细胞，能产生600伏电击，足以电倒一个人。以为电鳗就这么点本事？最新研究发现：它还能把猎物从藏身之处电出来，甚至能靠放电追踪猎物。http://t.cn/runzfkq但像超凡蜘蛛侠那样电出超能力是不行的[微笑]</p>
Label:	0
Generated Rationale:	<p>你的分析：1. 细节程度：文章提供了关于电鳗的基本信息，如其长度和电压能力，但缺乏具体的研究细节和科学解释。</p> <p>2. 引用来源：文章未提供具体的研究来源或科学机构的引用，仅提供了一个短链接，无法验证其内容。</p> <p>3. 引文：没有直接引用科学家的话或研究论文。</p> <p>4. 一致性：文章内容与已知的电鳗特性基本一致，但新发现的细节缺乏支持。</p> <p>5. 语气：语气轻松，带有娱乐性质，使用了表情符号，可能不够严肃。</p> <p>根据这些文本特征，该新闻可能是假的</p>

Table 11: Generated Chinese Rationale.

411 **D Report Refining Prompt Template**

Use of AI for this Report	
Prompt:	<p>Correct the grammar, polish the language style, but DO NOT add new information into this academic writing.</p> <p><i>our writing</i></p>

Table 12: Report Use of AI.