Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection

Beizhe Hu[1,2] **Qiang Sheng**[1,*] **Juan Cao**[1,2] **Yuhui Shi**[1,2]
Yang Li[1,2] **Danding Wang**[1] **Peng Qi**[3]
[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences,
Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences [3]National University of Singapore
{hubeizhe21s,shengqiang18z,caojuan,shiyuhui22s}@ict.ac.cn
{liyang23s,wangdanding}@ict.ac.cn, pengqi.qp@gmail.com

Abstract

Detecting fake news requires both a delicate sense of diverse clues and a profound understanding of the real-world background, which remains challenging for detectors based on small language models (SLMs) due to their knowledge and capability limitations. Recent advances in large language models (LLMs) have shown remarkable performance in various tasks, but whether and how LLMs could help with fake news detection remains underexplored. In this paper, we investigate the potential of LLMs in fake news detection. First, we conduct an empirical study and find that a sophisticated LLM such as GPT 3.5 could generally expose fake news and provide desirable multi-perspective rationales but still underperforms the basic SLM, fine-tuned BERT. Our subsequent analysis attributes such a gap to the LLM's inability to select and integrate rationales properly to conclude. Based on these findings, we propose that current LLMs may not substitute fine-tuned SLMs in fake news detection but can be a good advisor for SLMs by providing multi-perspective instructive rationales. To instantiate this proposal, we design an adaptive rationale guidance network for fake news detection (ARG), in which SLMs selectively acquire insights on news analysis from the LLMs' rationales. We further derive a rationale-free version of ARG by distillation, namely ARG-D, which services cost-sensitive scenarios without querying LLMs. Experiments on two real-world datasets demonstrate that ARG and ARG-D outperform three types of baseline methods, including SLM-based, LLM-based, and combinations of small and large language models.

1 Introduction

The wide and fast spread of fake news online has posed real-world threats in critical domains like politics (Fisher et al., 2016), economy (CHEQ, 2019), and public health (Naeem and Bhatti, 2020). Among the countermeasures to combat this issue, *automatic fake news detection*, which aims at distinguishing inaccurate and intentionally misleading news items from others automatically, has been a promising solution in practice (Shu et al., 2017; Roth, 2022).

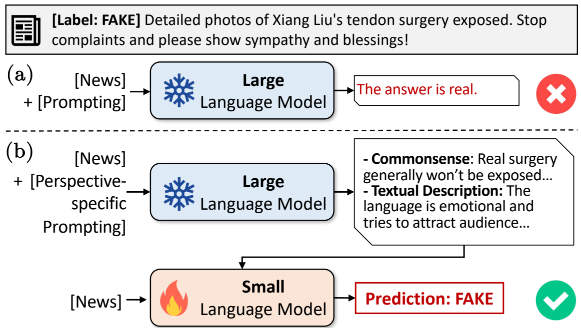


Figure 1: Illustration of the role of large language models (LLMs) in fake news detection. In this case, (a) the LLM fails to output correct judgment of news veracity but (b) helps the small language model (SLM) judge correctly by providing informative rationales.

Though much progress has been made (Hu et al., 2022a), understanding and characterizing fake news is still challenging for current models. This is caused by the complexity of the news-faking process: Fake news creators might manipulate any part of the news, using diverse writing strategies and being driven by inscrutable underlying aims. Therefore, to maintain both effectiveness and universality for fake news detection, an ideal method is required to have: 1) a delicate sense of diverse clues (*e.g.*, style, facts, commonsense); and 2) a profound understanding of the real-world background.

Recent methods (Zhang et al., 2021; Kaliyar et al., 2021; Mosallanezhad et al., 2022; Hu et al., 2023) generally exploit pre-trained **small language models (SLMs)**[1] like BERT (Devlin et al., 2019)

*Corresponding author.

[1]The academia lacks a consensus regarding the size boundary between small and large language models at present, but it

and RoBERTa (Liu et al., 2019) to understand news content and provide fundamental representation, plus optional social contexts (Shu et al., 2019; Cui et al., 2022), knowledge bases (Popat et al., 2018; Hu et al., 2022b), or news environment (Sheng et al., 2022) as supplements. SLMs do bring improvements, but their knowledge and capability limitations also compromise further enhancement of fake news detectors. For example, BERT was pre-trained on text corpus like Wikipedia (Devlin et al., 2019) and thus struggled to handle news items that require knowledge not included (Sheng et al., 2021).

As a new alternative to SLMs, **large language models (LLMs)** (OpenAI, 2022; Anthropic, 2023; Touvron et al., 2023), which are usually trained on the larger-scale corpus and aligned with human preferences, have shown impressive emergent abilities on various tasks (Wei et al., 2022a) and are considered promising as general task solvers (Ma et al., 2023). However, the potential of LLMs in fake news detection remains underexplored:

- Can LLMs help detect fake news with their internal knowledge and capability?
- What solution should we adopt to obtain better performance using LLMs?

To answer these two questions, we first conduct a deep investigation of the effective role of LLMs in fake news detection and attempt to provide a practical LLM-involved solution. Unlike contemporary works (Pelrine et al., 2023; Caramancion, 2023) which simply prompt LLMs to provide predictions with the task instruction, we conduct a detailed empirical study to mine LLMs' potential. Specifically, we use four typical prompting approaches (zero-shot/few-shot vanilla/chain-of-thought prompting) to ask the LLM to make veracity judgments of given news items (Figure 1(a)) and find that even the best-performing LLM-based method still underperforms task-specific fine-tuned SLMs. We then perform an analysis of the LLM-generated explanatory rationales and find that the LLM could provide reasonable and informative rationales from several perspectives. By subsequently inducing the LLM with perspective-specific prompts and performing rule-based ensembles of judgments, we find that rationales indeed benefit fake news detection, and attribute the unsatisfying performance to the LLM's inability to select and integrate rationales properly to conclude.

is widely accepted that BERT (Devlin et al., 2019) and GPT-3 family (Brown et al., 2020) are respectively small and large ones (Zhao et al., 2023).

#	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: Statistics of the fake news detection datasets.

Based on these findings, we propose that the current LLM may not be a good substitute for the well-fine-tuned SLM but could serve as a good advisor by providing instructive rationales, as presented in Figure 1(b). To instantiate our proposal, we design the adaptive rationale guidance (ARG) network for fake news detection, which bridges the small and large LMs by selectively injecting new insight about news analysis from the large LM's rationales to the small LM. The ARG further derives the rationale-free ARG-D via distillation for cost-sensitive scenarios with no need to query LLMs. Experiments on two real-world datasets show that ARG and ARG-D outperform existing SLM/LLM-only and combination methods. Our contributions are as follows:

- **Detailed investigation:** We investigate the effective role of LLMs in fake news detection and find the LLM is bad at veracity judgment but good at analyzing contents;
- **Novel and practical solution:** We design a novel ARG network and its distilled version ARG-D that complements small and large LMs by selectively acquiring insights from LLM-generated rationales for SLMs, which has shown superiority based on extensive experiments;
- **Useful resource:** We construct a rationale collection from GPT-3.5 for fake news detection in two languages (Chinese and English) and will make it publicly available to facilitate further research.[2]

2 Is the LLM a Good Detector?

In this section, we evaluate the performance of the representative LLM, *i.e.*, GPT-3.5 in fake news detection to reveal its judgment capability. We exploit four typical prompting approaches and perform a comparison with the SLM (here, BERT) fine-tuned

[2] https://github.com/ICTMCG/ARG

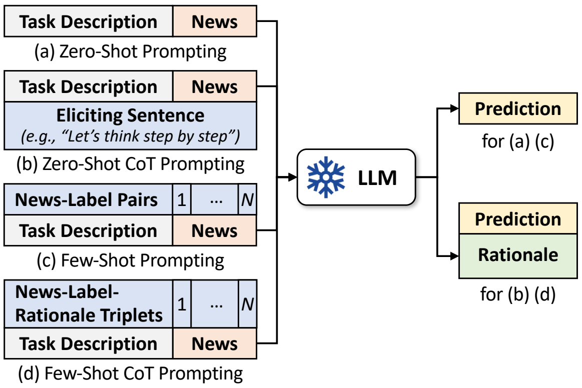


Figure 2: Illustration of prompting approaches for LLMs.

Model	Usage	Chinese	English
GPT-3.5-turbo	Zero-Shot	0.676	0.568
	Zero-Shot CoT	0.677	0.666
	Few-Shot	<u>0.725</u>	0.697
	Few-Shot CoT	0.681	<u>0.702</u>
BERT	Fine-tuning	**0.753** **(+3.8%)**	**0.765** **(+9.0%)**

Table 2: Performance in macro F1 of the large and small LMs. The best two results are **bolded** and <u>underlined</u>, respectively. The relative increases over the second-best results are shown in the brackets.

on this task. Formally, given a news item x, the model aims to predict whether x is fake or not.

2.1 Experimental Settings

Dataset We employ the Chinese dataset Weibo21 (Nan et al., 2021) and the English dataset GossipCop (Shu et al., 2020) for evaluation. Following existing works (Zhu et al., 2022; Mu et al., 2023), we preprocess the datasets with deduplication and temporal data split to avoid possible performance overrating led by data leakage for the SLM. Table 1 presents the dataset statistics.

Large Language Model We evaluate GPT-3.5-turbo, the LLM developed by OpenAI and supporting the popular chatbot ChatGPT (OpenAI, 2022), due to its representativeness and convenient calling. The large scale of parameters makes task-specific fine-tuning almost impossible for LLMs, so we use the prompt learning paradigm, where an LLM learns tasks given prompts containing instructions or few-shot demonstrations (Liu et al., 2023a). In detail, we utilize the following four typical prompting approaches to elicit the potential of the LLM in fake news detection (Figure 2):

- **Zero-Shot Prompting** constructs prompt only containing the task description and the given news. To make the response more proficient and decrease the refusal ratio, we optionally adopt the role-playing technique when describing our task (Liu et al., 2023b; Ramlochan, 2023).

- **Zero-Shot CoT Prompting** (Kojima et al., 2022) is a simple and straightforward chain-of-thought (CoT) prompting approach to encourage the LLM to reason. In addition to the elements in zero-shot prompting, it adds an eliciting sentence such as *"Let's think step by step."*

- **Few-Shot Prompting** (Brown et al., 2020) provides task-specific prompts and several news-label examples as demonstrations. After preliminary tests of {2,4,8}-shot settings, we choose 4-shot prompting which includes two real and two fake samples.

- **Few-Shot CoT Prompting** (Wei et al., 2022b) not only provides news-label examples but also demonstrates reasoning steps with previously written rationales. Here, we obtain the provided rationale demonstrations from the correct and reasonable outputs of zero-shot CoT prompting.

Small Language Model We adopt the pre-trained small language models, BERT (Devlin et al., 2019) as the representative, given its wide use in this task (Kaliyar et al., 2021; Zhu et al., 2022; Sheng et al., 2022). Specifically, we limit the maximum length of the text to 170 tokens and use *chinese-bert-wwm-ext* and *bert-base-uncased* from Transformers package (Wolf et al., 2020) for the Chinese and English evaluation, respectively. We use Adam (Kingma and Ba, 2014) as the optimizer and do a grid search for the optimal learning rate. We report the testing result on the best-validation checkpoint.

2.2 Comparison between Small and Large LMs

Table 2 presents the performance of GPT-3.5-turbo with four prompting approaches and the fine-tuned BERT on the two datasets. We observe that: **1)** Though the LLM is generally believed powerful, **the LLM underperforms the fine-tuned SLM using all four prompting approaches**. The SLM has

Perspective	Chinese		English	
	Prop.	macF1	Prop.	macF1
Textual Description	65%	0.706	71%	0.653
News: Everyone! Don't buy cherries anymore: Cherries of this year are infested with maggots, and nearly 100% are affected. **LLM Rationale:** ...The tone of the news is extremely urgent, seemingly trying to spread panic and anxiety. **Prediction:** Fake **Ground Truth:** Fake				
Commonsense	71%	0.698	60%	0.680
News: Huang, the chief of Du'an Civil Affairs Bureau, gets subsistence allowances of 509 citizens, owns nine properties, and has six wives... **LLM Rationale:** ...The news content is extremely outrageous...Such a situation is incredibly rare in reality and even could be thought impossible. **Prediction:** Fake **Ground Truth:** Fake				
Factuality	17%	0.629	24%	0.626
News: The 18th National Congress has approved that individuals who are at least 18 years old are now eligible to marry... **LLM Rationale:** First, the claim that Chinese individuals at least 18 years old can register their marriage is real, as this is stipulated by Chinese law... **Prediction:** Real **Ground Truth:** Fake				
Others	4%	0.649	8%	0.704

Table 3: Analysis of different perspectives of LLM's rationales in the sample set, including the data ratio, LLM's performance, and cases. Prop.: Proportion.

a relative increase of 3.8%∼11.3% in Chinese and 9.0%∼34.6% in English over the LLM, indicating that the LLM lacks task-specific knowledge while the SLM learns during fine-tuning.

2) Few-shot versions outperform zero-shot ones, suggesting the importance of task samples. However, introducing several samples only narrow the gap with the SLM but does not lead to surpassing.

3) CoT prompting brings additional performance gain in general, especially under the zero-shot setting on the English dataset (+17.3%). However, we also observe some cases where CoT leads to a decrease. This indicates that effective use of rationales may require more careful design.

Overall, given the LLM's unsatisfying performance and higher inference costs than the SLM, the current LLM has not been a "good enough" detector to substitute task-specific SLMs in fake news detection.

2.3 Analysis on the Rationales from the LLM

Though the LLM is bad at news veracity judgment, we also notice that the rationales generated through zero-shot CoT prompting exhibit a unique multi-perspective analytical capability that is challenging and rare for SLMs. For further exploration, we sample 500 samples from each of the two datasets and manually categorize them according to the perspectives from which the LLM performs the news analysis. Statistical results by perspectives and cases are presented in Table 3.[3] We see that: **1) The LLM is capable of generating human-like rationales on news content from various perspectives**, such as textual description, commonsense, and factuality, which meets the requirement of the delicate sense of diverse clues and profound understanding of the real-world background in fake news detection. **2)** The detection performance on the subset using certain perspectives is higher than the zero-shot CoT result on the full testing set. This indicates the potential of analysis by perspectives, though the coverage is moderate. **3)** The analysis from the perspective of factuality leads to the performance lower than average, indicating the unreliability of using the LLM for factuality analysis based on its internal memorization. We speculate this is caused by the hallucination issue (Ji et al., 2023; Zhang et al., 2023).

Model	Usage	Chinese	English
GPT-3.5-turbo	Zero-Shot CoT	0.677	0.666
	from Perspective TD	0.667	0.611
	from Perspective CS	0.678	0.698
BERT	Fine-tuning	0.753	0.765
Ensemble	Majority Voting	0.735	0.724
	Oracle Voting	0.908	0.878

Table 4: Performance of the LLM using zero-shot CoT with perspective specified and other compared models. TD: Textual description; CS: Commonsense.

We further investigate the LLM's performance when asked to perform analysis from a specific perspective on the full testing set (*i.e.*, 100% coverage).[4] From the first group in Table 4, we see that the LLM's judgment with single-perspective analysis elicited is still promising. Compared with the comprehensive zero-shot CoT setting, the single-perspective-based LLM performs comparatively on the Chinese dataset and is better on the English dataset (for the commonsense perspective case).

[3] Note that a sample may be analyzed from multiple perspectives and thus the sum of *proportions* might be larger than 100%.

[4] We exclude the factuality to avoid the impacts of hallucination. The eliciting sentence is "Let's think from the perspective of [textual description/commonsense]."

The results showcase that the internal mechanism of the LLM to integrate the rationales from diverse perspectives is ineffective for fake news detection, limiting the full use of rationales. In this case, combining the small and large LMs to complement each other is a promising solution: The former could benefit from the analytical capability of the latter, while the latter could be enhanced by task-specific knowledge from the former.

To exhibit the advantages of this solution, we apply majority voting and oracle voting (assuming the most ideal situation where we trust the correctly judged model for each sample, if any) among the two single-perspective-based LLMs and the BERT. Results show that we are likely to gain a performance better than any LLM-/SLM-only methods mentioned before if we could adaptively combine their advantages, *i.e.*, the flexible task-specific learning of the SLM and the informative rationale generated by the LLM. That is, **the LLM could be possibly a good advisor for the SLM by providing rationales, ultimately improving the performance of fake news detection.**

3 ARG: Adaptive Rationale Guidance Network for Fake News Detection

Based on the above findings and discussion, we propose the adaptive rationale guidance (ARG) network for fake news detection. Figure 3 overviews the ARG and its rationale-free version ARG-D, for cost-sensitive scenarios. The objective of ARG is to empower small fake news detectors with the ability to adaptively select useful rationales as references for final judgments. Given a news item x and its corresponding LLM-generated rationales r_t (textual description) and r_c (commonsense), the ARG encodes the inputs using the SLM at first (Figure 3(a)). Subsequently, it builds news-rationale collaboration via predicting the LLM's judgment through the rationale, enriching news-rationale feature interaction, and evaluating rationale usefulness (Figure 3(b)). The interactive features are finally aggregated with the news feature $\mathbf{x}$ for the final judgment of x being fake or not (Figure 3(c)). ARG-D is derived from the ARG via distillation for scenarios where the LLM is unavailable (Figure 3(d)).

3.1 Representation

We employ two BERT models separately as the news and rationale encoder to obtain semantic representations. For the given news item x and two corresponding rationales r_t and r_c, the representations are $\mathbf{X}$, $\mathbf{R_t}$, and $\mathbf{R_c}$, respectively.

3.2 News-Rationale Collaboration

The step of news-rationale collaboration aims at providing a rich interaction between news and rationales and learning to adaptively select useful rationales as references, which is at the core of our design. To achieve such an aim, ARG includes three modules, as detailed and exemplified using the textual description rationale branch below:

3.2.1 News-Rationale Interaction

To enable comprehensive information exchange between news and rationales, we introduce a news-rationale interactor with a dual cross-attention mechanism to encourage feature interactions. The cross-attention can be described as:

$$\mathrm{CA}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{softmax}\left(\mathbf{Q}' \cdot \mathbf{K}'/\sqrt{d}\right)\mathbf{V}', \tag{1}$$

where $\mathbf{Q}' = \mathbf{W}_{\mathrm{Q}}\mathbf{Q}$, $\mathbf{K}' = \mathbf{W}_{\mathrm{K}}\mathbf{K}$, and $\mathbf{V}' = \mathbf{W}_{\mathrm{V}}\mathbf{V}$. d is the dimensionality. Given representations of the news $\mathbf{X}$ and the rationale $\mathbf{R_t}$, the process is:

$$\mathbf{f_{t\to x}} = \mathrm{AvgPool}\left(\mathrm{CA}(\mathbf{R_t},\mathbf{X},\mathbf{X})\right), \tag{2}$$

$$\mathbf{f_{x\to t}} = \mathrm{AvgPool}\left(\mathrm{CA}(\mathbf{X},\mathbf{R_t},\mathbf{R_t})\right), \tag{3}$$

where $\mathrm{AvgPool}(\cdot)$ is the average pooling over the token representations outputted by cross-attention to obtain one-vector text representation $\mathbf{f}$.

3.2.2 LLM Judgement Prediction

Understanding the judgment hinted by the given rationale is a prerequisite for fully exploiting the information behind the rationale. To this end, we construct the LLM judgment prediction task, whose requirement is to predict the LLM judgment of the news veracity according to the given rationale. We expect this to deepen the understanding of the rationale texts. For the textual description rationale branch, we feed its representation $\mathbf{R_t}$ into the LLM judgment predictor, which is parametrized using a multi-layer perception (MLP)[5]:

$$\hat{m}_t = \mathrm{sigmoid}(\mathrm{MLP}(\mathbf{R_t})), \tag{4}$$

$$L_{pt} = \mathrm{CE}(\hat{m}_t, m_t), \tag{5}$$

[5]For brevity, we omit the subscripts of all independently parametrized MLPs.

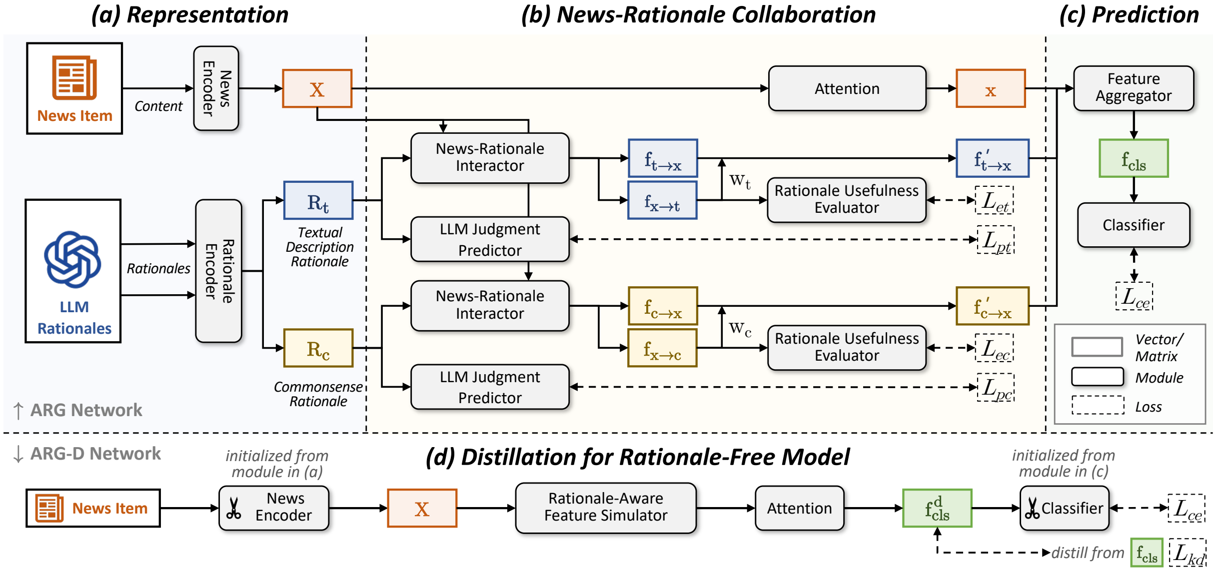


Figure 3: Overall architecture of our proposed adaptive rationale guidance (ARG) network and its rationale-free version ARG-D. In the ARG, the news item and LLM rationales are (a) respectively encoded into $\mathbf{X}$ and $\mathbf{R}_*(* \in \{t, c\})$. Then the small and large LMs collaborate with each other via news-rationale feature interaction, LLM judgment prediction, and rationale usefulness evaluation. The obtained interactive features $\mathbf{f}'_{*\to\mathbf{x}}$ $(* \in \{t, c\})$. These features are finally aggregated with attentively pooled news feature $\mathbf{x}$ for the final judgment. In the ARG-D, the news encoder and the attention module are preserved and the output of the rationale-aware feature simulator is supervised by the aggregated feature $\mathbf{f}_{\mathbf{cls}}$ for knowledge distillation.

where m_t and $\hat{m}_t$ are respectively the LLM's actual judgment (extracted from the response) and its prediction. The loss L_{pt} is a cross-entropy loss $\mathrm{CE}(\hat{y}, y) = -y \log \hat{y} - (1-y)\log(1-\hat{y})$. The case is similar for commonsense rationale $\mathbf{R_c}$.

3.2.3 Rationale Usefulness Evaluation

The usefulness of rationales from different perspectives varies across different news items and improper integration may lead to performance degradation. To enable the model to adaptively select appropriate rationale, we devise a rationale usefulness evaluation process, in which we assess the contributions of different rationales and adjust their weights for subsequent veracity prediction. The process comprises two phases, *i.e.*, evaluation and reweighting. For evaluation, we input the news-aware rationale vector $\mathbf{f}_{\mathbf{x}\to\mathbf{t}}$ into the rationale usefulness evaluator (parameterized by an MLP) to predict its usefulness u_t. Following the assumption that rationales leading to correct judgments are more useful, we use the judgment correctness as the rationale usefulness labels.

$$\hat{u}_t = \mathrm{sigmoid}(\mathrm{MLP}(\mathbf{f}_{\mathbf{x}\to\mathbf{t}})), \tag{6}$$

$$L_{et} = \mathrm{CE}(\hat{u}_t, u_t). \tag{7}$$

In the reweighting phase, we input vector $\mathbf{f}_{\mathbf{x}\to\mathbf{t}}$ into an MLP to obtain a weight number w_t, which is then used to reweight the rationale-aware news vector $\mathbf{f}_{\mathbf{t}\to\mathbf{x}}$. The procedure is as follows:

$$\mathbf{f}_{\mathbf{t}\to\mathbf{x}}{}' = w_t \cdot \mathbf{f}_{\mathbf{t}\to\mathbf{x}}. \tag{8}$$

We also use attentive pooling to transform the representation matrix $\mathbf{X}$ into a vector $\mathbf{x}$.

3.3 Prediction

Based on the outputs from the last step, we now aggregate news vector $\mathbf{x}$ and rationale-aware news vector $\mathbf{f}'_{\mathbf{t}\to\mathbf{x}}$, $\mathbf{f}'_{\mathbf{c}\to\mathbf{x}}$ for the final judgment. For news item x with label $y \in \{0, 1\}$, we aggregate these vectors with different weights:

$$\mathbf{f_{cls}} = w_x^{cls} \cdot \mathbf{x} + w_t^{cls} \cdot \mathbf{f}'_{\mathbf{t}\to\mathbf{x}} + w_c^{cls} \cdot \mathbf{f}'_{\mathbf{c}\to\mathbf{x}}, \tag{9}$$

where w_x^{cls}, w_t^{cls} and w_c^{cls} are learnable parameters ranging from 0 to 1. $\mathbf{f_{cls}}$ is the fusion vector, which is then fed into the MLP classifier for final prediction of news veracity:

$$L_{ce} = \mathrm{CE}(\mathrm{MLP}(f_{cls}), y). \tag{10}$$

The total loss function is the weighted sum of the loss terms mentioned above:

$$L = L_{ce} + \beta_1(L_{et} + L_{ec}) + \beta_2(L_{pt} + L_{pc}), \tag{11}$$

where β_1 and β_2 are hyperparameters.

Model		Chinese				English			
		macF1	Acc.	$\mathrm{F1}_{\mathrm{real}}$	$\mathrm{F1}_{\mathrm{fake}}$	macF1	Acc.	$\mathrm{F1}_{\mathrm{real}}$	$\mathrm{F1}_{\mathrm{fake}}$
G1: LLM-Only	GPT-3.5-turbo	0.725	0.734	0.774	0.676	0.702	0.813	0.884	0.519
G2: SLM-Only	Baseline	0.753	0.754	0.769	0.737	0.765	0.862	0.916	0.615
	$\mathrm{EANN}_{\mathrm{T}}$	0.754	0.756	0.773	0.736	0.763	0.864	0.918	0.608
	Publisher-Emo	0.761	0.763	0.784	0.738	0.766	0.868	0.920	0.611
	ENDEF	0.765	0.766	0.779	0.751	0.768	0.865	0.918	0.618
G3: LLM+SLM	Baseline + Rationale	0.767	0.769	0.787	0.748	0.777	0.870	0.921	0.633
	SuperICL	0.757	0.759	0.779	0.734	0.736	0.864	0.920	0.551
	ARG	**0.784**	**0.786**	0.804	0.764	**0.790**	0.878	0.926	0.653
	(Relative Impr. over Baseline)	*(+4.2%)*	*(+4.3%)*	*(+4.6%)*	*(+3.8%)*	*(+3.2%)*	*(+1.8%)*	*(+1.1%)*	*(+6.3%)*
	w/o LLM Judgment Predictor	0.773	0.774	0.789	0.756	0.786	**0.880**	0.928	0.645
	w/o Rationale Usefulness Evaluator	0.781	0.783	0.801	0.761	0.782	0.873	0.923	0.641
	w/o Predictor & Evaluator	0.769	0.770	0.782	0.756	0.780	0.874	0.923	0.637
	ARG-D	0.771	0.772	0.785	0.756	0.778	0.870	0.921	0.634
	(Relative Impr. over Baseline)	*(+2.4%)*	*(+2.3%)*	*(+2.1%)*	*(+2.6%)*	*(+1.6%)*	*(+0.9%)*	*(+0.6%)*	*(+3.2%)*

Table 5: Performance of the ARG and its variants and the LLM-only, SLM-only, LLM+SLM methods. The best two results in macro F1 and accuracy are respectively **bolded** and underlined. For GPT-3.5-turbo, the best results in Table 2 are reported.

3.4 Distillation for Rationale-Free Model

The ARG requires sending requests to the LLM for every prediction, which might not be affordable for cost-sensitive scenarios. Therefore, we attempt to build a rationale-free model, namely ARG-D, based on the trained ARG model via knowledge distillation (Hinton et al., 2015). The basic idea is simulated and internalized the knowledge from rationales into a parametric module. As shown in Figure 3(d), we initialize the news encoder and classifier with the corresponding modules in the ARG and train a rationale-aware feature simulator (implemented with a multi-head transformer block) and an attention module to internalize knowledge. Besides the cross-entropy loss L_{ce}, we let the feature $\mathbf{f}_{\mathbf{cls}}^{\mathbf{d}}$ to imitate $\mathbf{f}_{\mathbf{cls}}$ in the ARG, using the mean squared estimation loss:

$$L_{kd} = \mathrm{MSE}(\mathbf{f}_{\mathbf{cls}}, \mathbf{f}_{\mathbf{cls}}^{\mathbf{d}}). \tag{12}$$

4 Evaluation

4.1 Experimental Settings

Baselines We compare three groups of methods: **G1 (LLM-Only)**: We list the performance of the best-performing setting on each dataset in Table 2, *i.e.*, few-shot in Chinese and few-shot CoT in English.

G2 (SLM-Only)[6]: **1) Baseline:** The vanilla BERT-base model whose setting remains consistent with that in Section 2. **2) $\mathrm{EANN}_{\mathrm{T}}$ (Wang et al., 2018):** A model that learns effective signals using auxiliary adversarial training, aiming at removing event-related features as much as possible. We used publication year as the label for the auxiliary task. **3) Publisher-Emo (Zhang et al., 2021):** A model that fuses a series of emotional features with textual features for fake news detection. **4) ENDEF (Zhu et al., 2022):** A model that removes entity bias via causal learning for better generalization on distribution-shifted fake news data. All methods in this group used the same BERT as the text encoder.

G3 (LLM+SLM): **1) Baseline+Rationale:** It concatenates features from the news encoder and rationale encoder and feeds them into an MLP for prediction. **2) SuperICL (Xu et al., 2023):** 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.

Implementation Details We use the same datasets introduced in Section 2 and keep the setting the same in terms of the pre-trained model, learning rate, and optimization method. For the ARG-D network, the parameters of the news encoder and classifier are derived from the ARG model. A four-head transformer block is implemented in the rationale-aware feature simulator. The weight of loss functions $L_{et}, L_{pt}, L_{ec}, L_{pc}$ in the ARG and L_{kd} in the ARG-D are grid searched.

[6]As this paper focuses on text-based news, we use the text-only variant of the original EANN following (Sheng et al., 2021) and the publisher-emotion-only variant in (Zhang et al., 2021).

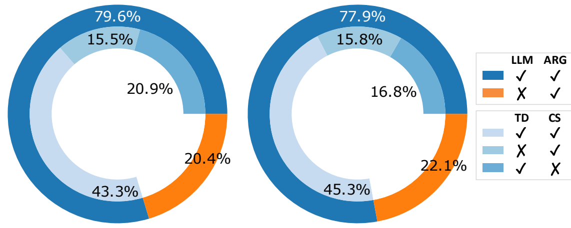
4.2 Performance Comparison and Ablation Study

Table 5 presents the performance of our proposed ARG and its variants and the compared methods. From the results, we observe that: **1)** The ARG outperforms all other compared methods in macro F1, demonstrating its effectiveness. **2)** The rationale-free ARG-D still outperforms all compared methods except ARG and its variants, which shows the positive impact of the distilled knowledge from ARG. **3)** The two compared LLM+SLM methods exhibit different performance. The simple combination of features of news and rationale yields a performance improvement, showing the usefulness of our prompted rationales. SuperICL outperforms the LLM-only method but fails to consistently outperform the baseline SLM on the two datasets. We speculate that this is due to the complexity of our task, where injecting prediction and confidence of an SLM does not bring sufficient information. **4)** We evaluate three ablation experiment groups to evaluate the effectiveness of different modules in ARG network. From the result, we can see that w/o LLM Judgement Predictor or w/o Rationale Usefulness Evaluator both bring a significant decrease in ARG performance, highlighting the significance of these two structures. Besides, we found that even the weakest one among the variants of ARG still outperforms all other methods, which shows the importance of the news-rationale interaction structure we designed.

4.3 Result Analysis

To investigate which part the additional gain of the ARG(-D) should be attributed to, we perform statistical analysis on the additional correctly judged samples of ARG(-D) compared with the vanilla BERT. From Figure 4, we observe that: **1)** The proportions of the overlapping samples between ARG(-D) and the LLM are over 77%, indicating that the ARG(-D) can exploit (and absorb) the valuable knowledge for judgments from the LLM, even its performance is unsatisfying. **2)** The samples correctly judged by the LLM from both two perspectives contribute the most, suggesting more diverse rationales may enhance the ARG(-D)'s training. **3)** 20.4% and 22.1% of correct judgments should be attributed to the model itself. We speculate that it produces some kinds of "new knowledge" based on the wrong judgments of the given knowledge.

For analysis of success and failure cases and additional analysis, please refer to the appendix.



(a) right(ARG) – right(Baseline) (b) right(ARG-D) – right(Baseline)

Figure 4: Statistics of additional correctly judged samples of (a) ARG and (b) ARG-D over the BERT baseline. right$(\cdot)$ denotes samples correctly judged by the method $(\cdot)$. TD/CS: Textual description/commonsense perspective.

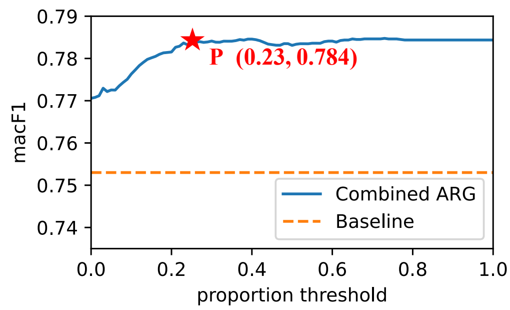


Figure 5: Performance as the shifting threshold changes.

4.4 Cost Analysis in Practice

We showcase a possible model-shifting strategy to balance the performance and cost in practical systems. Inspired by Ma et al. (2023), we simulate the situation where we use the more economic ARG-D by default but query the more powerful ARG for part of the data. As presented in Figure 5, by sending only 23% of the data (according to the confidence of ARG-D) to the ARG, we could achieve 0.784 in macro F1, which is the same as the performance fully using the ARG.

5 Related Work

Fake News Detection Fake news detection is generally formulated as a binary classification task between real and fake news items. Research on this task could be roughly categorized into two groups: social-context-based and content-based methods. Methods in the first group aim at differentiating fake and real news during the diffusion procedure by observing the propagation patterns (Zhou and Zafarani, 2019), user feedback (Min et al., 2022), and social networks (Nguyen et al., 2020). The second group focuses on finding hints based on the given content, including text (Przybyla, 2020) and images (Qi et al., 2021) and may require extra assistance from knowledge bases (Popat et al., 2018) and news environments (Sheng et al., 2022). Both two groups of methods obtain textual representation from pre-trained models like BERT as a convention but rarely consider its potential for fake news detection. We conducted an exploration in this paper by combining large and small LMs and obtained good improvement only using textual content.

LLMs for Natural Language Understanding LLMs, though mostly generative models, also have powerful natural language understanding (NLU) capabilities, especially in the few-shot in-context learning scenarios (Brown et al., 2020). Recent works in this line focus on benchmarking the latest LLM in NLU. Results show that LLMs may not have comprehensive superiority compared with a well-trained small model in some types of NLU tasks (Zhong et al., 2023; Kocoń et al., 2023). Our results provide empirical findings in fake news detection with only textual content as the input.

6 Conclusion and Discussion

We investigated if large LMs help in fake news detection and how to properly utilize their advantages for improving performance. Results show that the large LM (GPT-3.5) underperforms the task-specific small LM (BERT), but could provide informative rationales and complement small LMs in news understanding. Based on these findings, we designed the ARG network to flexibly combine the respective advantages of small and large LMs and developed its rationale-free version ARG-D for cost-sensitive scenarios. Experiments showed the superiority of the ARG and ARG-D.

Discussion Our findings in fake news detection exemplify the current barrier for LLMs to be competent in applications closely related to the sophisticated real-world background. Though having superior analyzing capability, LLMs may struggle to properly make full use of their internal capability. This suggests that "mining" their potential may require novel prompting techniques and a deeper understanding of its internal mechanism. We then identified the possibility of combining small and LLMs to earn additional improvement and provided a solution especially suitable for situations where the better-performing models have to "select good to learn" from worse ones. We expect our solution to be extended to other tasks and foster more effective and cost-friendly use of LLMs in the future.

Limitations We identify the following limitations: 1) We do not examine other well-known LLMs (*e.g.*, Claude[7] and Ernie Bot[8]) due to the API unavailability for us when conducting this research; 2) We only consider the perspectives summarized from the LLM's response and there might be other prompting perspectives based on a conceptualization framework of fake news; 3) Our best results still fall behind the oracle voting integration of multi-perspective judgments in Table 4, indicating that rooms still exist in our line regarding performance improvements.

Acknowledgements

The authors would like to thank the anonymous reviewers for their insightful comments. This work is supported by the National Natural Science Foundation of China (62203425), the Zhejiang Provincial Key Research and Development Program of China (2021C01164), the Project of Chinese Academy of Sciences (E141020), the Postdoctoral Fellowship Program of CPSF (GZC20232738) and the CIPSC-SMP-Zhipu.AI Large Model Cross-Disciplinary Fund.

References

Anthropic. 2023. Model card and evaluations for claude models. https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf. Accessed: 2023-08-13.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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*, pages 1877–1901. Curran Associates Inc.

Kevin Matthe Caramancion. 2023. News verifiers showdown: A comparative performance evaluation of ChatGPT 3.5, ChatGPT 4.0, bing AI,

[7] https://claude.ai/
[8] https://yiyan.baidu.com/

and bard in news fact-checking. *arXiv preprint arXiv:2306.17176*.

CHEQ. 2019. The economic cost of bad actors on the internet. https://info.cheq.ai/hubfs/Research/THE_ECONOMIC_COST_Fake_News_final.pdf. Accessed: 2023-08-13.

Jian Cui, Kwanwoo Kim, Seung Ho Na, and Seungwon Shin. 2022. Meta-path-based fake news detection leveraging multi-level social context information. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 325–334. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. ACL.

Marc Fisher, John Woodrow Cox, and Peter Hermann. 2016. Pizzagate: From rumor, to hashtag, to gunfire in dc. *The Washington Post*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Beizhe Hu, Qiang Sheng, Juan Cao, Yongchun Zhu, Danding Wang, Zhengjia Wang, and Zhiwei Jin. 2023. 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 (Volume 5: Industry Track)*, pages 116–125. ACL.

Linmei Hu, Siqi Wei, Ziwang Zhao, and Bin Wu. 2022a. Deep learning for fake news detection: A comprehensive survey. *AI Open*, 3:133–155.

Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and Philip Yu. 2022b. CHEF: A pilot Chinese dataset for evidence-based fact-checking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3362–3376. ACL.

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 Computing Surveys*, 55:1–38.

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. 2023. ChatGPT: Jack of all trades, master of none. *Information Fusion*, 99:101861.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, and Yang Liu. 2023b. Jailbreaking ChatGPT via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yubo Ma, Yixin Cao, YongChing Hong, and Aixin Sun. 2023. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! *arXiv preprint arXiv:2303.08559*.

Erxue Min, Yu Rong, Yatao Bian, Tingyang Xu, Peilin Zhao, Junzhou Huang, and Sophia Ananiadou. 2022. Divide-and-conquer: Post-user interaction network for fake news detection on social media. In *Proceedings of the ACM Web Conference 2022*, pages 1148–1158. ACM.

Ahmadreza Mosallanezhad, Mansooreh Karami, Kai Shu, Michelle V. Mancenido, and Huan Liu. 2022. Domain adaptive fake news detection via reinforcement learning. In *Proceedings of the ACM Web Conference 2022*, pages 3632–3640. ACM.

Yida Mu, Kalina Bontcheva, and Nikolaos Aletras. 2023. It's about time: Rethinking evaluation on rumor detection benchmarks using chronological splits. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 736–743. ACL.

Salman Bin Naeem and Rubina Bhatti. 2020. The COVID-19 'infodemic': a new front for information professionals. *Health Information & Libraries Journal*, 37(3):233–239.

Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. MDFEND: Multi-domain fake news detection. In *Proceedings of the 30th ACM*

International Conference on Information and Knowledge Management. ACM.

Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. FANG: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, pages 1165–1174. ACM.

OpenAI. 2022. ChatGPT: Optimizing language models for dialogue. `https://openai.com/blog/chatgpt/`. Accessed: 2023-08-13.

Kellin Pelrine, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, and Reihaneh Rabbany. 2023. Towards reliable misinformation mitigation: Generalization, uncertainty, and GPT-4. *arXiv preprint arXiv:2305.14928v1*.

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. ACL.

Piotr Przybyla. 2020. Capturing the style of fake news. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 490–497. AAAI Press.

Peng Qi, Juan Cao, Xirong Li, Huan Liu, Qiang Sheng, Xiaoyue Mi, Qin He, Yongbiao Lv, Chenyang Guo, and Yingchao Yu. 2021. Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1212–1220. ACM.

Sunil Ramlochan. 2023. Role-playing in large language models like ChatGPT. https://www.promptengineering.org/role-playing-in-large-language-models-like-chatgpt/. Accessed: 2023-08-13.

Yoel Roth. 2022. The vast majority of content we take action on for misinformation is identified proactively. `https://twitter.com/yoyoel/status/1483094057471524867`. Accessed: 2023-08-13.

Qiang Sheng, Juan Cao, Xueyao Zhang, Rundong Li, Danding Wang, and Yongchun Zhu. 2022. Zoom out and observe: News environment perception for fake news detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4543–4556. ACL.

Qiang Sheng, Xueyao Zhang, Juan Cao, and Lei Zhong. 2021. Integrating pattern-and fact-based fake news detection via model preference learning. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 1640–1650. ACM.

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*, pages 395–405. ACM.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. FakeNewsNet: A data repository with news content, social context and spatiotemporal information for studying fake news on social media. *Big data*, 8:171–188.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19:22–36.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. 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*, pages 849–857. ACM.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. ACL.

Canwen Xu, Yichong Xu, Shuohang Wang, Yang Liu, Chenguang Zhu, and Julian McAuley. 2023. Small models are valuable plug-ins for large language models. *arXiv preprint arXiv:2305.08848*.

Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *Proceedings of the web conference 2021*, pages 3465–3476. ACM.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the AI ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can ChatGPT understand too? a comparative study on ChatGPT and fine-tuned BERT. *arXiv preprint arXiv:2302.10198*.

Xinyi Zhou and Reza Zafarani. 2019. Network-based fake news detection: A pattern-driven approach. *ACM SIGKDD Explorations Newsletter*, 21(2):48–60.

Yongchun Zhu, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, and Fuzhen Zhuang. 2022. Generalizing to the future: Mitigating entity bias in fake news detection. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2120–2125. ACM.

A Case Analysis

Tables 6 and 7 present cases in the testing set. The former includes cases where at least one of the compared methods and large language model (LLM) predictions is correct, while the latter includes those in the complementary set (*i.e.,* neither of them provides correct predictions).

In Table 6, Case 1 shows a case in which the Baseline made a correct prediction while the LLM was wrong. The ARG stayed unaffected by the probably misleading LLM rationales and maintained the correct prediction. Cases 2 and 3 show the situation where the baseline was incorrect while the LLM could be seen partially right (*i.e.*, only that used the specific perspective of the two is correct). In these cases, our ARG selected the correct rationale based on its adaptive selection capability, resulting in the correct prediction.

In Case 4 in Table 7, ARG successfully identified the case where both the baseline and LLM failed to provide accurate prediction, which surpasses our expectations. After conducting a comparative analysis, we found that within the training set, the use of phrases like "compact writing style" in textual description rationales often leads to incorrect judgments. We hypothesize that guided by the signals provided by the rationale usefulness evaluator, ARG recognized that pattern and was able to make judgments contrary to erroneous rationales and chose the correct prediction. While in Case 5, ARG was unable to invert the erroneous judgment as in Case 4. We speculate that this is due to the absence of a pattern as clear as that in Case 4. Constrained by the natural language understanding performance of the BERT-base model, ARG struggled to capture more complex logical relationships.

B Prompting Examples

In this section, we aim to showcase the prompting templates along with real examples for facilitating readers to know the settings directly.

In Tables 8 and 9, we illustrate prompting examples for zero-shot prompting, zero-shot CoT prompting, zero-shot CoT prompting with perspective-specific prompts, few-shot prompting, and few-shot CoT prompting from Chinese dataset.

For brevity, we only present a news pair comprising one real news and one fake news as a demo for few-shot prompting and few-shot CoT prompting. In practice, we utilize multiple sets of news pairs according to how many shots are required.

Due to a notable likelihood of eliciting refusals from GPT-3.5-turbo when using prompts related to fake news detection directly on the English dataset, we employ the role-playing technique. Specifically, we built appropriate contexts in the prompt to let the LLM be in a scene and ultimately decreased the refusal likelihood. (Liu et al., 2023b; Ramlochan, 2023)

Case 1: Both the Baseline and the ARG predicted correctly, while the LLM did incorrectly.

News: Wow! A robot in Russia was captured heroically saving a playful little girl from being crushed by a shelf This act went viral online, showing the robot's strength, amazing!
LLM Rationale (Textual Description): This message uses exaggerated, emotive language, while providing no credible sources or evidence.
LLM Rationale (Commonsense): Based on common knowledge and experience, robots have not yet reached a level where they can act autonomously and save human lives.

Ground Truth: ***Real***
Baseline Pred.: ***Real*** **LLM TD Pred.:** Fake **LLM CS Pred.:** Fake **ARG Pred.:** ***Real***

Case 2: The Baseline predicted incorrectly, the ARG did correctly, and the LLM using textual description perspective also did correctly.

News: #NingboRevelation# [Ningbo Xiangshan Fishermen Catch a Strange Fish, Dubbed the 'Panda of the Water']#NingboFreshNews# Around 7:00 PM on March 2nd, the ZXY02206 vessel, at the coordinates of 121°20'898" E, 29°09'898" N, caught a strange fish. Upon verification, the fish was identified as a juvenile wild Chinese sturgeon, measuring 64 centimeters in length and weighing approximately 2 kilograms. Around 10:00 AM on March 3rd, this juvenile wild Chinese sturgeon was returned to the embrace of the sea.
LLM Rationale (Textual Description): This message uses a more formal language and provides specific details about the time, location, and species of the fish caught, including its size. It also mentions that the information has been verified.
LLM Rationale (Commonsense): Based on common knowledge and experience, the Chinese sturgeon is a critically endangered and rare species, making it unlikely to be commonly caught by ordinary fishermen.

Ground Truth : ***Real***
Baseline Pred. : Fake **LLM TD Pred. :** ***Real*** **LLM CS Pred. :** Fake **ARG Pred. :** ***Real***

Case 3: The Baseline predicted incorrectly, the ARG did correctly, and the LLM using commonsense perspective also did correctly.

News: After recovery, the British Prime Minister commutes to work by bicycle every day, though he owns a second-hand Volkswagen. He is not allowed to use an official car for commuting, and since there's no fixed parking spot at the Prime Minister's Office, he opts for cycling. Two bodyguards follow him in a car but often fall behind due to traffic congestion.
LLM Rationale (Textual Description): The style of this message is straightforward and clear, without excessive embellishment or exaggeration. The content described also appears to be realistic and credible.
LLM Rationale (Commonsense): Considering the nature of the British Prime Minister's job and security considerations, the claim that he commutes to work by bicycle every day is not very credible.

Ground Truth : ***Fake***
Baseline Pred. : Real **LLM TD Pred. :** Real **LLM CS Pred. :** ***Fake*** **ARG Pred. :** ***Fake***

Table 6: Cases in the testing set where at least one in the Baseline and the LLM using perspective-specific prompting predicted correctly. TD/CS represents textual description/commonsense perspective, respectively. Patterns in rationales are underlined and correct predictions are in ***bold italics***.

Case 4: Both the Baseline and the LLM using perspective-specific prompting made incorrect predictions, while the ARG did it correctly.

News: According to Russian media reports, recently, Japan entrusted US military transport planes stationed in Japan to transport 2,000 test kits donated by our country to Japan. However, during the transportation process, the US suddenly informed Japan that these test kits had been transported to the US mainland. A US Air Force official later explained that due to a mistake, these materials were wrongly transported to the US mainland. The Japanese side does not accept the explanation of the US military official! Do you think the actions of the US military were really due to a mistake?

LLM Rationale (Textual Description): The style of the message is quite compact, with language that is concise and clear. There are no apparent grammatical or spelling errors, making it consistent with the style of news reporting.

LLM Rationale (Commonsense): Based on common experience and knowledge, it's not uncommon for mistakes to occur during the transportation of goods. Therefore, it's more likely that the actions of the U.S. military were due to an error rather than intentional.

Ground Truth: ***Fake***
Baseline Pred.: Real **LLM TD Pred.:** Real **LLM CS Pred.:** Real **ARG Pred.:** ***Fake***

Case 5: The Baseline, the LLM using perspective-specific prompting, and the ARG all made incorrect predictions.

News: #ShenzhenEpidemic# Shenzhen's Sungang area has started entry and exit controls. Everyone should pay attention to wearing masks properly! Protect yourselves! The epidemic is not over yet, everyone should be extra careful.

LLM Rationale (Textual Description): The message utilizes a hashtag "#ShenzhenEpidemic#", which aligns with the common method of posting on social media. The tone is also quite formal, serving as a reminder for everyone to take protective measures and expressing concern and attention towards the epidemic situation.

LLM Rationale (Commonsense): Given the current global situation regarding the pandemic, various regions are implementing measures to control the spread of the virus. Therefore, Shenzhen may also adopt similar control measures. Additionally, the message mentions wearing masks and protecting oneself, which is in line with protective measures during the pandemic.

Ground Truth: ***Fake***
Baseline Pred.: Real **LLM TD Pred.:** Real **LLM CS Pred.:** Real **ARG Pred.:** Real

Table 7: Cases in the testing set where both the Baseline and the LLM using perspective-specific prompting made incorrect predictions. TD/CS represents textual description/commonsense perspective, respectively. Patterns in rationales are underlined and correct predictions are in ***bold italics***.

❶ *Zero-Shot Prompting*
Q: Given the following message, predict its veracity. If it is more likely to be a real message, return 1; otherwise, return 0. Please refrain from providing ambiguous assessments such as undetermined: *Detailed photos of Xiang Liu's tendon surgery exposed. Stop complaints and please show sympathy and blessings!* **A:** 1
❷ *Zero-Shot CoT Prompting*
Q: Given the following message, predict its veracity. If it is more likely to be a real message, return 1; otherwise, return 0. Please refrain from providing ambiguous assessments such as undetermined: *Detailed photos of Xiang Liu's tendon surgery exposed. Stop complaints and please show sympathy and blessings!* **A: Let's think step by step.** Firstly, this message mentions that Xiang Liu underwent tendon surgery, which can be verified through other channels. Secondly, the message refers to detailed photo exposure, which can also be verified by searching related reports. Lastly, the message urges people to show sympathy and offer blessings, which aligns with common practices. Therefore, it can be considered that this message is more likely to be true information, and the response is 1. **Therefore, the answer (arabic numerals) is** 1
❸ *Zero-Shot CoT with Perspective Specific Prompting (Textual Description)*
Q: Given the following message, predict its veracity. If it is more likely to be a real message, return 1; otherwise, return 0. Please refrain from providing ambiguous assessments such as undetermined: *Detailed photos of Xiang Liu's tendon surgery exposed. Stop complaints and please show sympathy and blessings!* **A: Let's think from the perspective of textual description.** This message employs emotional language to convey sympathy and blessings for Xiang Liu, while also mentioning the detailed photo exposure of the surgery. Considering that such a message is likely to attract public attention and resonance, it is more likely to be true information. Return 1.
❹ *Zero-Shot CoT with Perspective Specific Prompting (Commonsense)*
Q: Given the following message, predict its veracity. If it is more likely to be a real message, return 1; otherwise, return 0. Please refrain from providing ambiguous assessments such as undetermined: *Detailed photos of Xiang Liu's tendon surgery exposed. Stop complaints and please show sympathy and blessings!* **A: Let's think from the perspective of commonsense.** Based on common knowledge and experience, usually, real surgery will not be exposed, as this involves patient privacy and medical institutions' confidentiality obligations. Therefore, this message is more likely to be false. Return 0.

Table 8: Zero-shot prompting, zero-shot CoT prompting, and zero-shot CoT with Perspective Specific prompting templates, along with the news case and the replies from the LLM (*i.e.,* GPT-3.5-turbo). Template prompts are in **bold** and the target news is in *italics*.

❺ *Few-Shot Prompting*

Q: Given the following message, predict its veracity. If it is more likely to be a real message, return 1; otherwise, return 0. Please refrain from providing ambiguous assessments such as undetermined: [Russian Official Reiterates Denial: Land Transfer to China is False News] On the 10th local time, the Russian Ministry of Foreign Affairs stated that there is no territorial dispute between Russia and China. The claim that Russia transferred 4.7 square kilometers of land to China is false. After the Russia-China border was established, there have been no changes in the border demarcation. Previously, both the government of Hunchun, China, and the Russian border defense department had denied the land transfer news. http://t.cn/ruole6x
A: 1
Q: Given the following message, predict its veracity. If it is more likely to be a real message, return 1; otherwise, return 0. Please refrain from providing ambiguous assessments such as undetermined: Breaking News: Tianyi Li's successful exoneration! Just received news that the owner of the "rivuai160" bar in Sanlitun has been taken into custody by the police. The authorities have initiated a prostitution investigation on the bar. Tianyi Li and five others are expected to shift from being defendants to witnesses, potentially being released as soon as this weekend!
A: 0

Q: Given the following message, predict its veracity. If it is more likely to be a real message, return 1; otherwise, return 0. Please refrain from providing ambiguous assessments such as undetermined: *Detailed photos of Xiang Liu's tendon surgery exposed. Stop complaints and please show sympathy and blessings!*
A: 1

❻ *Few-Shot CoT Prompting*

Q: Given the following message, predict its veracity. If it is more likely to be a real message, return 1; otherwise, return 0. Please refrain from providing ambiguous assessments such as undetermined: [Russian Official Reiterates Denial: Land Transfer to China is False News] On the 10th local time, the Russian Ministry of Foreign Affairs stated that there is no territorial dispute between Russia and China. The claim that Russia transferred 4.7 square kilometers of land to China is false. After the Russia-China border was established, there have been no changes in the border demarcation. Previously, both the government of Hunchun, China, and the Russian border defense department had denied the land transfer news. http://t.cn/ruole6x
A: Let's think step by step. Firstly, the source of the message is a Weibo link, and we need to consider the credibility of this link. Secondly, the content of the message pertains to a Russian denial of transferring land to China, and we need to assess if this aligns with reality. Lastly, we should check if the message aligns with reports from other reliable media sources. Based on this analysis, the message is more likely to be true information. The Russian Ministry of Foreign Affairs is an official entity, lending some credibility to their denial. Additionally, the message is consistent with the denials from the government of Hunchun, China, and the Russian border defense department. Therefore, return 1. Therefore, the answer (arabic numeral) is: 1.
Q: Given the following message, predict its veracity. If it is more likely to be a real message, return 1; otherwise, return 0. Please refrain from providing ambiguous assessments such as undetermined: Breaking News: Tianyi Li's successful exoneration! Just received news that the owner of the "rivuai160" bar in Sanlitun has been taken into custody by the police. The authorities have initiated a prostitution investigation on the bar. Tianyi Li and five others are expected to shift from being defendants to witnesses, potentially being released as soon as this weekend!
A: Let's think step by step. Firstly, this message claims that Tianyi Li's exoneration was successful, but it doesn't specify the case in question and lacks any supporting evidence. Therefore, this part is unreliable. Secondly, the message states that the owner of the "rivuai160" bar in Sanlitun has been taken into custody by the police, and the authorities have initiated a prostitution investigation on the bar. This part could be true, as police investigations into establishments suspected of involvement in prostitution are not uncommon. However, we cannot verify the source and authenticity of this message, so it's not entirely trustworthy either. Lastly, the message suggests that Li Tianyi and five others are expected to shift from being defendants to witnesses and could be released as soon as this weekend. This part also lacks specific evidence, thus making it unreliable. In conclusion, the credibility of this message is low, and it's more likely to be untrue. Therefore, it should return 0. Hence, the answer (Arabic numeral) is: 0.

Q: Given the following message, predict its veracity. If it is more likely to be a real message, return 1; otherwise, return 0. Please refrain from providing ambiguous assessments such as undetermined: *Detailed photos of Xiang Liu's tendon surgery exposed. Stop complaints and please show sympathy and blessings!*
A: Firstly, this message mentions that Xiang Liu underwent tendon surgery, which can be verified through other channels. Secondly, the message refers to detailed photo exposure, which can also be verified by searching related reports. Lastly, the message urges people to show sympathy and offer blessings, which aligns with common practices. Therefore, it can be considered that this message is more likely to be true information, and the response is 1.

Table 9: Few-shot prompting and few-shot CoT prompting templates, along with the news case and the replies from the LLM (*i.e.,* GPT-3.5-turbo). Demos are above the dashed line. Below the dashed line, template prompts are in **bold** and the target news is in *italics*.