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]CAS Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]National University of Singapore
{hubeizhe21s, shengqiang18z, caojuan, shiyuhui22s, 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.

Introduction

The wide and fast spread of fake news online has posed real-world threats in critical domains like politics (Fisher, Cox, and Hermann 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).

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.

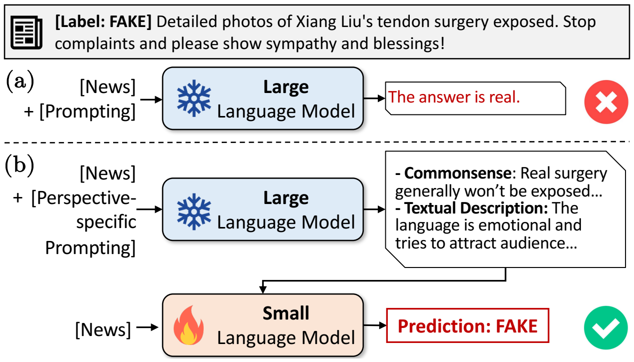

[Label: FAKE] Detailed photos of Xiang Liu's tendon surgery exposed. Stop complaints and please show sympathy and blessings!
(a)
[News]
+ [Prompting]
Large Language Model
The answer is real.
(b)
[News]
+ [Perspective-specific Prompting]
Large Language Model
- Commonsense: Real surgery generally won't be exposed...
- Textual Description: The language is emotional and tries to attract audience...
[News]
Small Language Model
Prediction: FAKE

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.

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

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

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: **1)** Can LLMs help detect fake news with their internal knowledge and capability? **2)** 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.

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 make it publicly available to facilitate further research.[2]

#	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.

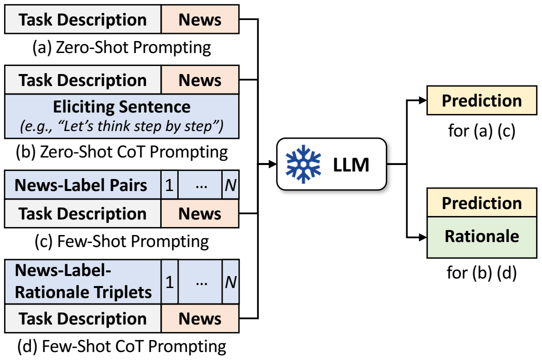


Figure 2: Illustration of prompting approaches for LLMs.

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

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, Bontcheva, and Aletras 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

[2] Code, data, and the extended version are available at https://github.com/ICTMCG/ARG

Model	Usage	Chinese	English
GPT-3.5-turbo	Zero-Shot	0.676	0.568
	Zero-Shot CoT	0.677	0.666
	Few-Shot	0.725	0.697
	Few-Shot CoT	0.681	0.702
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 underlined, respectively. The relative increases over the second-best results are shown in the brackets.

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 prepared 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, Goswami, and Narang 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 as the optimizer and do a grid search for the optimal learning rate. We report the testing result on the best-validation checkpoint.

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 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 narrows 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.

<table>
<tr><th rowspan="2">Perspective</th><th colspan="2">Chinese</th><th colspan="2">English</th></tr>
<tr><th>Proportion</th><th>macF1</th><th>Proportion</th><th>macF1</th></tr>
<tr><td><b>Textual Description</b></td><td>65%</td><td>0.706</td><td>71%</td><td>0.653</td></tr>
<tr><td colspan="5"><b>News:</b> Everyone! Don't buy cherries anymore: Cherries of this year are infested with maggots, and nearly 100% are affected.
<b>LLM Rationale:</b> ...The tone of the news is extremely urgent, seemingly trying to spread panic and anxiety.
<b>Prediction:</b> Fake <b>Ground Truth:</b> Fake</td></tr>
<tr><td><b>Commonsense</b></td><td>71%</td><td>0.698</td><td>60%</td><td>0.680</td></tr>
<tr><td colspan="5"><b>News:</b> Huang, the chief of Du'an Civil Affairs Bureau, gets subsistence allowances of 509 citizens, owns nine properties, and has six wives...
<b>LLM Rationale:</b> ...The news content is extremely outrageous...Such a situation is incredibly rare in reality and even could be thought impossible.
<b>Prediction:</b> Fake <b>Ground Truth:</b> Fake</td></tr>
<tr><td><b>Factuality</b></td><td>17%</td><td>0.629</td><td>24%</td><td>0.626</td></tr>
<tr><td colspan="5"><b>News:</b> The 18th National Congress has approved that individuals who are at least 18 years old are now eligible to marry...
<b>LLM Rationale:</b> First, the claim that Chinese individuals at least 18 years old can register their marriage is real, as this is stipulated by Chinese law...
<b>Prediction:</b> Real <b>Ground Truth:</b> Fake</td></tr>
<tr><td><b>Others</b></td><td>4%</td><td>0.649</td><td>8%</td><td>0.704</td></tr>
</table>

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

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).

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

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). 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.**

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)).

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.

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:

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:

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

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:

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

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

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

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]:

$$\hat{m}_t = \mathrm{sigmoid}(\mathrm{MLP}(\mathbf{R_t})),\ L_{pt} = \mathrm{CE}(\hat{m}_t, m_t), \qquad (4)$$

where m_t and $\hat{m}_t$ are respectively the LLM's claimed judgment 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 $\mathbf{R_c}$.

[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]."

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

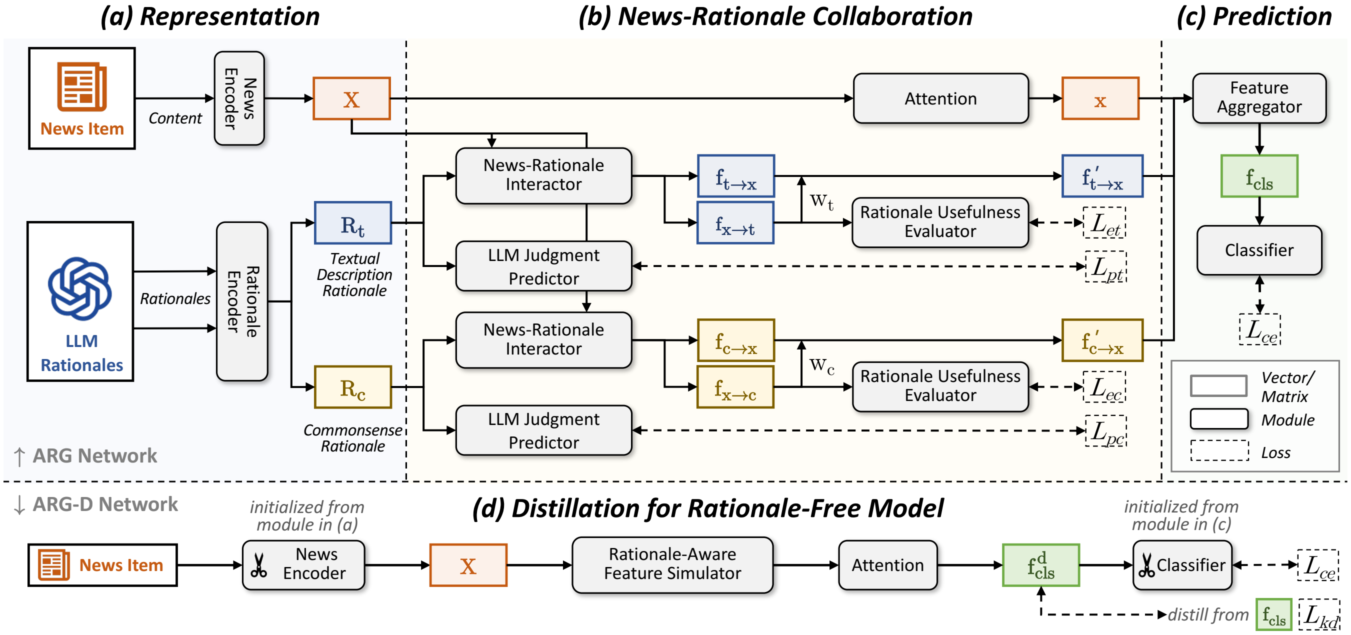


Figure 3: Overall architecture of our proposed adaptive rationale guidance (ARG) network and its rationale-free version ARG-D. In the ARG, the news item and LLM rationales are (a) respectively encoded into $\mathbf{X}$ and $\mathbf{R}_*(* \in \{t, c\})$. Then the small and large LMs collaborate with 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 $\mathbf{f_{cls}}$ for knowledge distillation.

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 $\mathbf{f_{x\to t}}$ 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.

$$\hat{u}_t = \mathrm{sigmoid}(\mathrm{MLP}(\mathbf{f_{x\to t}})),\ L_{et} = \mathrm{CE}(\hat{u}_t, u_t). \tag{5}$$

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

$$\mathbf{f_{t\to x}}' = w_t \cdot \mathbf{f_{t\to x}}. \tag{6}$$

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

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:

$$\mathbf{f_{cls}} = w_x^{cls} \cdot \mathbf{x} + w_t^{cls} \cdot \mathbf{f}'_{\mathbf{t}\to\mathbf{x}} + w_c^{cls} \cdot \mathbf{f}'_{\mathbf{c}\to\mathbf{x}}, \tag{7}$$

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:

$$L_{ce} = \mathrm{CE}(\mathrm{MLP}(f_{cls}), y). \tag{8}$$

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

$$L = L_{ce} + \beta_1(L_{et} + L_{ec}) + \beta_2(L_{pt} + L_{pc}), \tag{9}$$

where β_1 and β_2 are hyperparameters.

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, Vinyals, and Dean 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 $\mathbf{f^d_{cls}}$ to imitate $\mathbf{f_{cls}}$ in the ARG, using the mean squared estimation loss:

$$L_{kd} = \mathrm{MSE}(\mathbf{f_{cls}}, \mathbf{f^d_{cls}}). \tag{10}$$

Model		Chinese				English			
		macF1	Acc.	F1_{real}	F1_{fake}	macF1	Acc.	F1_{real}	F1_{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
	EANN_{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.

Evaluation

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) EANN_{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 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.

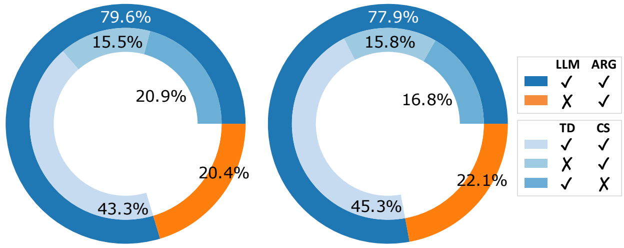
Performance Comparison and Ablation Study

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 fake news detection 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 the 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.

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

[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).



(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(·) denotes samples correctly judged by the method (·). TD/CS: Textual description/commonsense perspective.

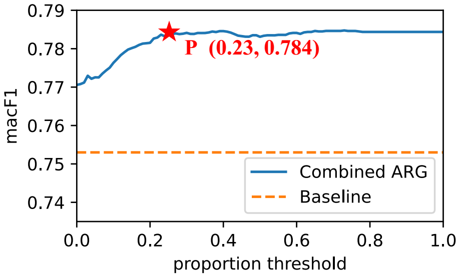


Figure 5: Performance as the shifting threshold changes.

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.

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.

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), 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 pretrained 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). Our results provide empirical findings in fake news detection with only textual content as the input.

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.

[7] https://claude.ai/

[8] https://yiyan.baidu.com/

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) (GZC20232738) and the CIPSC-SMP-Zhipu.AI Large Model Cross-Disciplinary Fund. The corresponding author is Qiang Sheng.

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.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models Are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, 1877–1901. Curran Associates Inc.

Caramancion, K. M. 2023. News Verifiers Showdown: A Comparative Performance Evaluation of ChatGPT 3.5, ChatGPT 4.0, Bing AI, 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.

Cui, J.; Kim, K.; Na, S. H.; and Shin, S. 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*, 325–334. ACM.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186. ACL.

Fisher, M.; Cox, J. W.; and Hermann, P. 2016. Pizzagate: From rumor, to hashtag, to gunfire in DC. *The Washington Post*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*.

Hu, B.; Sheng, Q.; Cao, J.; Zhu, Y.; Wang, D.; Wang, Z.; and Jin, Z. 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)*, 116–125. ACL.

Hu, L.; Wei, S.; Zhao, Z.; and Wu, B. 2022a. Deep learning for fake news detection: A comprehensive survey. *AI Open*, 3: 133–155.

Hu, X.; Guo, Z.; Wu, G.; Liu, A.; Wen, L.; and Yu, P. 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*, 3362–3376. ACL.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55: 1–38.

Kaliyar, R. K.; Goswami, A.; and Narang, P. 2021. FakeBERT: Fake News Detection in Social Media with a BERT-based Deep Learning Approach. *Multimedia tools and applications*, 80(8): 11765–11788.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large Language Models are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems*, volume 35, 22199–22213. Curran Associates, Inc.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9): 1–35.

Liu, Y.; Deng, G.; Xu, Z.; Li, Y.; Zheng, Y.; Zhang, Y.; Zhao, L.; Zhang, T.; and Liu, Y. 2023b. Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study. *arXiv preprint arXiv:2305.13860*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Ma, Y.; Cao, Y.; Hong, Y.; and Sun, A. 2023. Large Language Model Is Not a Good Few-shot Information Extractor, but a Good Reranker for Hard Samples! *arXiv preprint arXiv:2303.08559*.

Min, E.; Rong, Y.; Bian, Y.; Xu, T.; Zhao, P.; Huang, J.; and Ananiadou, S. 2022. Divide-and-Conquer: Post-User Interaction Network for Fake News Detection on Social Media. In *Proceedings of the ACM Web Conference 2022*, 1148–1158. ACM.

Mosallanezhad, A.; Karami, M.; Shu, K.; Mancenido, M. V.; and Liu, H. 2022. Domain Adaptive Fake News Detection via Reinforcement Learning. In *Proceedings of the ACM Web Conference 2022*, 3632–3640. ACM.

Mu, Y.; Bontcheva, K.; and Aletras, N. 2023. It's about Time: Rethinking Evaluation on Rumor Detection Benchmarks using Chronological Splits. In *Findings of the Association for Computational Linguistics: EACL 2023*, 736–743. ACL.

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

Nan, Q.; Cao, J.; Zhu, Y.; Wang, Y.; and Li, J. 2021. MDFEND: Multi-domain Fake News Detection. In *Proceedings* *of the 30th ACM International Conference on Information and Knowledge Management*. ACM.

Nguyen, V.-H.; Sugiyama, K.; Nakov, P.; and Kan, M.-Y. 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*, 1165–1174. ACM.

OpenAI. 2022. ChatGPT: Optimizing Language Models for Dialogue. https://openai.com/blog/chatgpt/. Accessed: 2023-08-13.

Pelrine, K.; Reksoprodjo, M.; Gupta, C.; Christoph, J.; and Rabbany, R. 2023. Towards Reliable Misinformation Mitigation: Generalization, Uncertainty, and GPT-4. *arXiv preprint arXiv:2305.14928v1*.

Popat, K.; Mukherjee, S.; Yates, A.; and Weikum, G. 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*, 22–32. ACL.

Przybyla, P. 2020. Capturing the Style of Fake News. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 490–497. AAAI Press.

Qi, P.; Cao, J.; Li, X.; Liu, H.; Sheng, Q.; Mi, X.; He, Q.; Lv, Y.; Guo, C.; and Yu, Y. 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*, 1212–1220. ACM.

Ramlochan, S. 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.

Roth, Y. 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.

Sheng, Q.; Cao, J.; Zhang, X.; Li, R.; Wang, D.; and Zhu, Y. 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)*, 4543–4556. ACL.

Sheng, Q.; Zhang, X.; Cao, J.; and Zhong, L. 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*, 1640–1650. ACM.

Shu, K.; Cui, L.; Wang, S.; Lee, D.; and Liu, H. 2019. dEFEND: Explainable Fake News Detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 395–405. ACM.

Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; and Liu, H. 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.

Shu, K.; Sliva, A.; Wang, S.; Tang, J.; and Liu, H. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19: 22–36.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.

Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; and Gao, J. 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*, 849–857. ACM.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; Chi, E. H.; Hashimoto, T.; Vinyals, O.; Liang, P.; Dean, J.; and Fedus, W. 2022a. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q. V.; and Zhou, D. 2022b. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, volume 35, 24824–24837. Curran Associates, Inc.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Le Scao, T.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: ACL.

Xu, C.; Xu, Y.; Wang, S.; Liu, Y.; Zhu, C.; and McAuley, J. 2023. Small Models are Valuable Plug-ins for Large Language Models. *arXiv preprint arXiv:2305.08848*.

Zhang, X.; Cao, J.; Li, X.; Sheng, Q.; Zhong, L.; and Shu, K. 2021. Mining Dual Emotion for Fake News Detection. In *Proceedings of the web conference 2021*, 3465–3476. ACM.

Zhang, Y.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; Huang, X.; Zhao, E.; Zhang, Y.; Chen, Y.; Wang, L.; Luu, A. T.; Bi, W.; Shi, F.; and Shi, S. 2023. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *arXiv preprint arXiv:2309.01219*.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; Du, Y.; Yang, C.; Chen, Y.; Chen, Z.; Jiang, J.; Ren, R.; Li, Y.; Tang, X.; Liu, Z.; Liu, P.; Nie, J.-Y.; and Wen, J.-R. 2023. A Survey of Large Language Models. *arXiv preprint arXiv:2303.18223*.

Zhong, Q.; Ding, L.; Liu, J.; Du, B.; and Tao, D. 2023. Can ChatGPT Understand Too? A Comparative Study on ChatGPT and Fine-tuned BERT. *arXiv preprint arXiv:2302.10198*.

Zhou, X.; and Zafarani, R. 2019. Network-Based Fake News Detection: A Pattern-Driven Approach. *ACM SIGKDD Explorations Newsletter*, 21(2): 48–60.

Zhu, Y.; Sheng, Q.; Cao, J.; Li, S.; Wang, D.; and Zhuang, F. 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*, 2120–2125. ACM.