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

Code:https://github.com/ICTMCG/ARG

—— AAAI'24

gesis Leibniz-Institut für Sozialwissenschaften

Forschungszentrum L3S Research Center

Reported by Shu Ming Jiang

Introduction

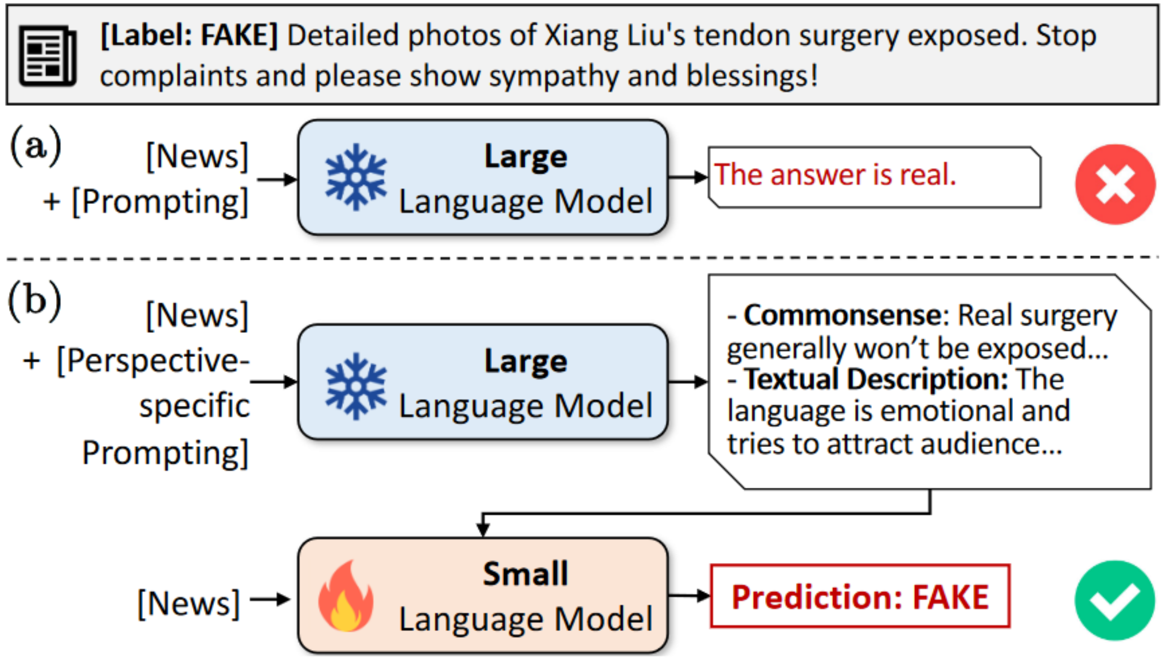


Figure 1: Illustration of the role of large language models (LLMs) in fake news detection. In this case, (a) the LLM fails to output correct judgment of news veracity but (b) helps the small language model (SLM) judge correctly by providing informative rationales.

The study found that while advanced models like GPT-3.5 excel in detecting fake news , they fall short compared to basic models like fine-tuned BERT. This is attributed to their struggle in properly selecting and integrating rationales for conclusive reasoning

#	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.

Introduction

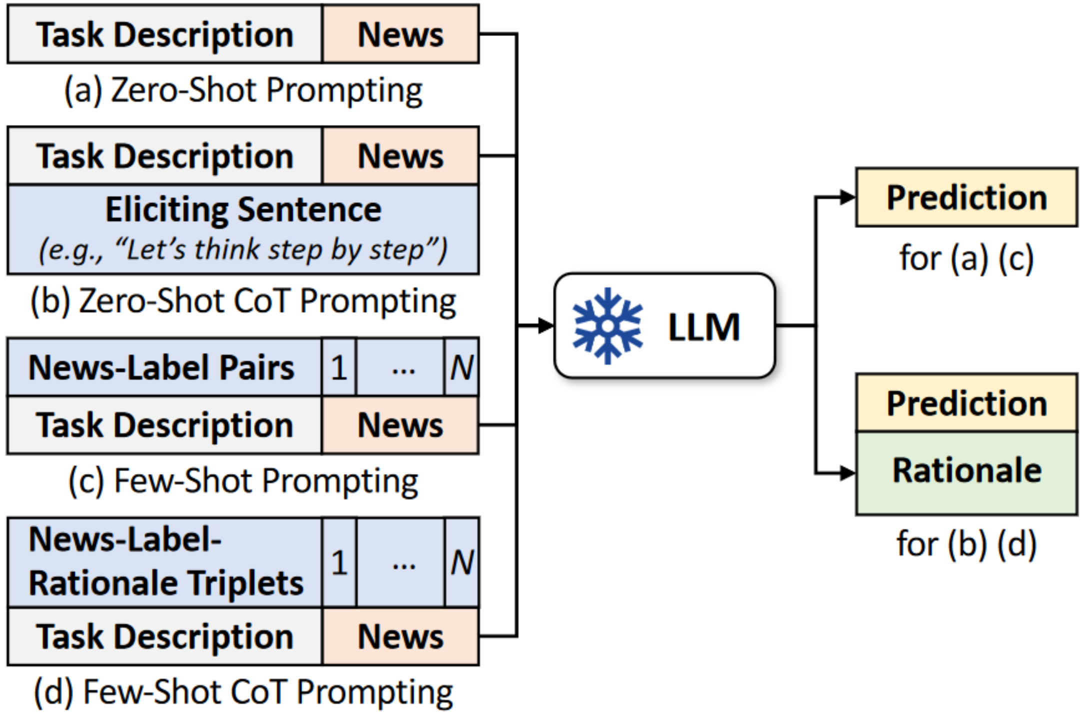


Figure 2: Illustration of prompting approaches for LLMs.

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	**0.761 (+5.0%)**	**0.774 (+10.3%)**

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.

Introduction

Perspective	Chinese		English	
	Prop.	macF1	Prop.	macF1
Textual Description	68%	0.746	59%	0.629
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	69%	0.745	56%	0.642
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	18%	0.597	46%	0.592
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	8%	0.750	17%	0.694

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.

Model	Usage	Chinese	English
GPT-3.5-turbo	Zero-Shot CoT	0.677	0.666
	from Perspective TD	0.674	0.611
	from Perspective CS	0.676	0.698
BERT	Fine-tuning	0.761	0.774
Ensemble	Majority Voting	0.750	0.753
	Oracle Voting	0.907	0.876

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

Overview

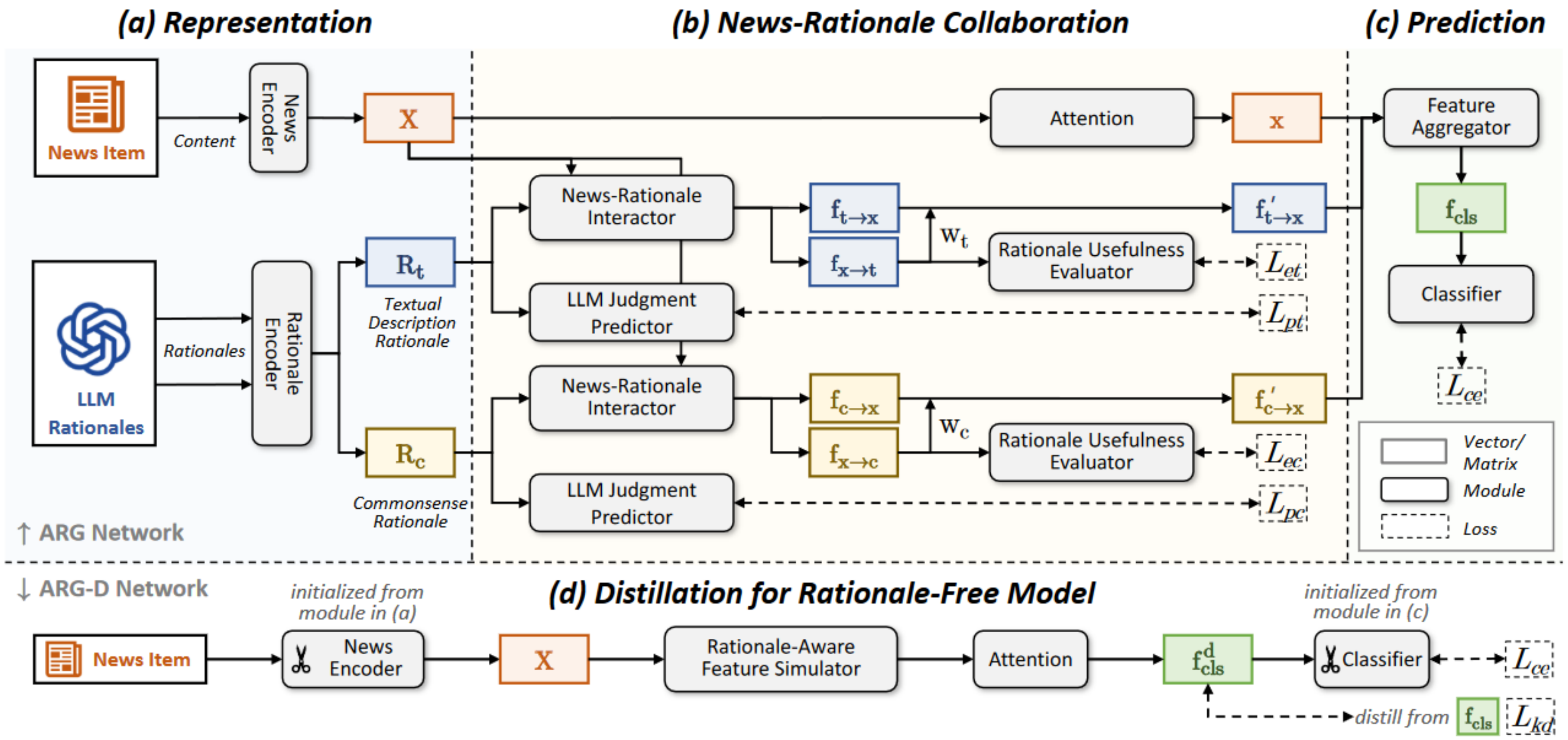
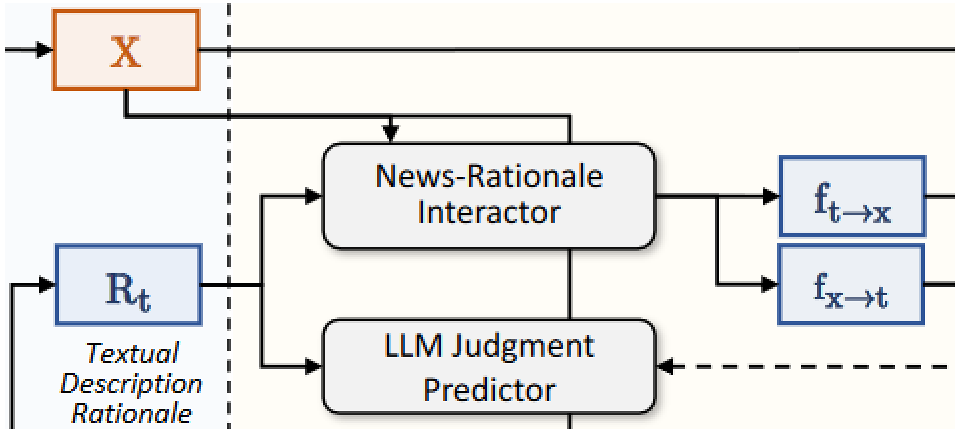


Figure 3: Overall architecture of our proposed adaptive rationale guidance (ARG) network and its rationale-free version ARG-D. In the ARG, the news item and LLM rationales are (a) respectively encoded into $\mathbf{X}$ and $\mathbf{R}_*(* \in \{t, c\})$. Then the small and large LMs collaborate with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

Method



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

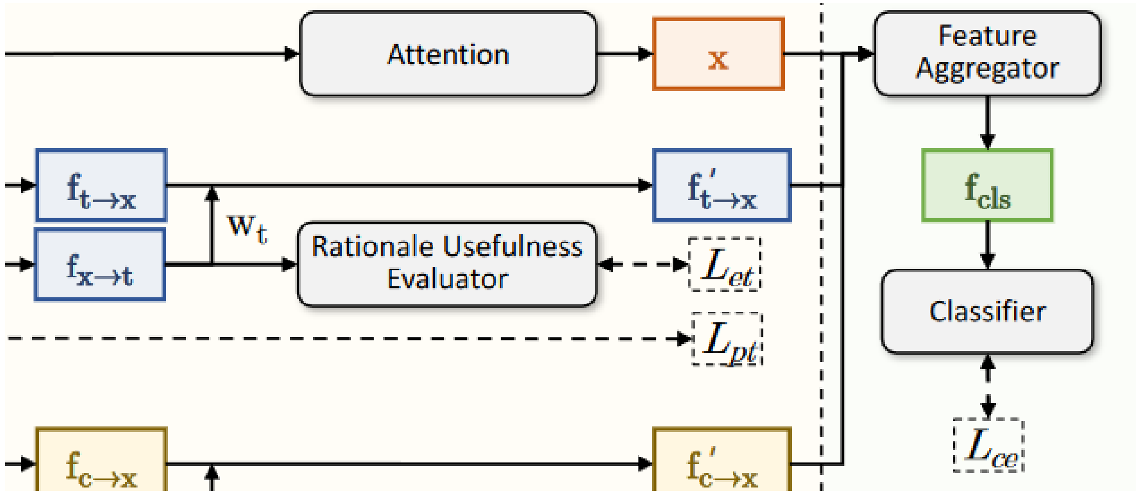
$$\mathbf{f}_{\mathbf{t}\to\mathbf{x}} = \mathrm{AvgPool}\left(\mathrm{CA}(\mathbf{R}_\mathbf{t}, \mathbf{X}, \mathbf{X})\right), \tag{2}$$

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

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

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

Method



$$\hat{u}_t = \text{sigmoid}(\text{MLP}(\mathbf{f_{x \to t}})), \quad (6)$$

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

$$\mathbf{f_{x \to t}}' = w_t \cdot \mathbf{f_{x \to t}}. \quad (8)$$

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

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

$$L = L_{ce} + \beta_1 L_{et} + \beta_2 L_{pt} + \beta_3 L_{ec} + \beta_4 L_{pc}, \quad (11)$$

$$L_{kd} = \text{MSE}(\mathbf{f_{cls}}, \mathbf{f_{cls}^d}). \quad (12)$$

Experiments

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	0.761	0.762	0.780	0.741	0.774	0.869	0.920	0.628
	$EANN_T$	0.768	0.769	0.784	0.752	0.775	0.868	0.920	0.630
	Publisher-Emo	0.755	0.757	0.779	0.730	0.783	0.871	0.921	0.645
	ENDEF	0.768	0.769	0.779	0.758	0.777	0.878	0.927	0.626
G3: LLM+SLM	Baseline + Rationale	0.763	0.764	0.778	0.748	0.785	0.883	0.930	0.641
	SuperICL	0.757	0.759	0.779	0.734	0.736	0.864	0.920	0.551
	ARG	**0.790**	**0.792**	0.811	0.770	**0.801**	<u>0.889</u>	0.933	0.668
	(Relative Impr. over Baseline)	*(+3.8%)*	*(+3.9%)*	*(+4.0%)*	*(+3.9%)*	*(+3.5%)*	*(+2.3%)*	*(+1.4%)*	*(+6.4%)*
	w/o LLM Judgment Predictor	0.784	0.787	0.809	0.759	0.797	**0.890**	0.935	0.658
	w/o Rationale Usefulness Evaluator	<u>0.786</u>	<u>0.790</u>	0.816	0.757	<u>0.798</u>	0.887	0.932	0.664
	w/o Predictor & Evaluator	0.773	0.776	0.797	0.750	0.793	0.882	0.928	0.658
	ARG-D	0.777	0.778	0.790	0.765	0.790	0.886	0.932	0.649
	(Relative Impr. over Baseline)	*(+2.1%)*	*(+2.1%)*	*(+1.3%)*	*(+3.2%)*	*(+2.1%)*	*(+2.0%)*	*(+1.3%)*	*(+3.3%)*

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 <u>underlined</u>. For GPT-3.5-turbo, the best results in Table 2 are reported.

Experiments

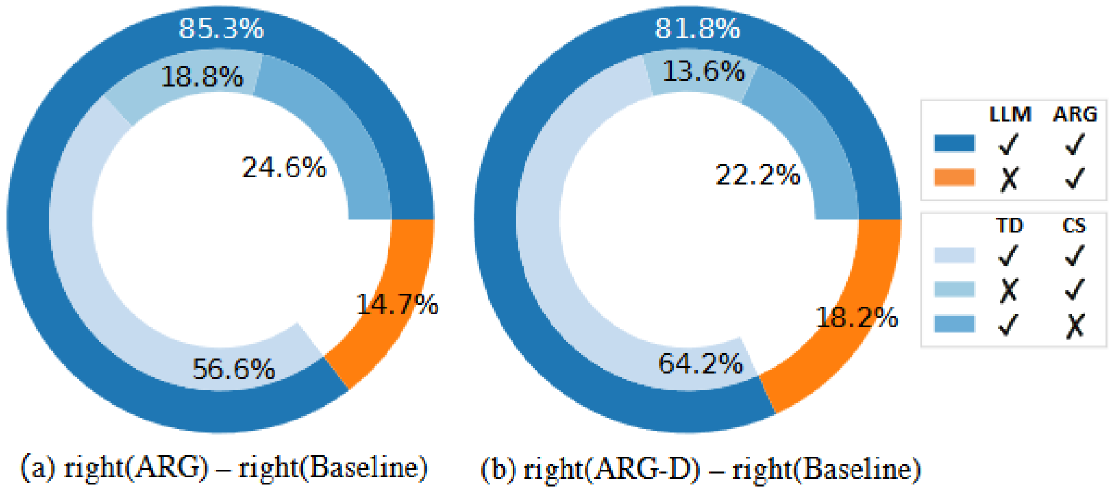


Figure 4: Statistics of additional correctly judged samples of (a) ARG and (b) ARG-D over the BERT baseline. right(·) denotes samples correctly judged by the method (·). TD/CS: Textual description/commonsense perspective.

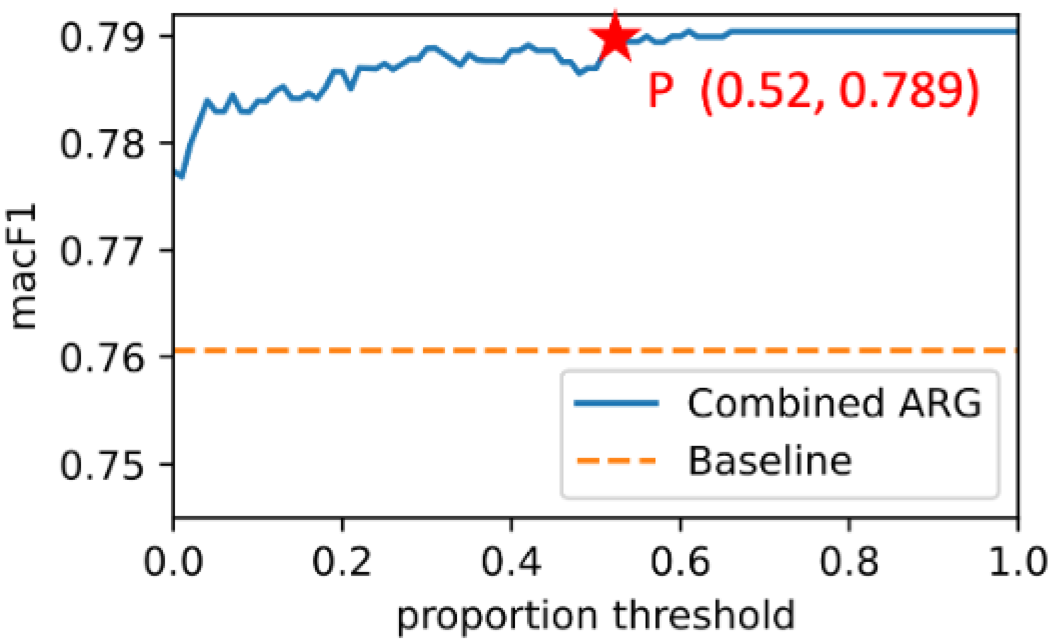


Figure 5: Performance as the shifting threshold changes.

Experiments

	Setting	ZS CLS AVG Acc	LP CLS AVG Acc	ZS-Flickr30K IR R@1	ZS-Flickr30K TR R@1	ZS-Flickr30K rsum	ZS-MSCOCO IR R@1	ZS-MSCOCO TR R@1	ZS-MSCOCO rsum	VQAv2 overall
CLIP	15M	41.3	67.5	27.6	42.8	343.1	15.9	24.8	236.8	47.5
CLIP+FDT	15M	45.9(↑4.6)	68.8(↑1.3)	32.6(↑5.0)	51.0(↑8.2)	376.5(↑33.4)	19.4(↑3.5)	29.6(↑4.8)	263.1(↑26.3)	50.6(↑3.1)
CLIP	30M	56.8	73.8	43.6	58.8	431.3	23.3	34.8	300.8	50.6
CLIP+FDT	30M	61.2(↑ 4.4)	75.6 (↑ 1.8)	52.5(↑8.9)	70.8(↑12.0)	474.2(↑42.9)	28.3(↑5.0)	43(↑8.2)	337.1 (↑36.3)	53.4(↑2.8)
CLIP	145M	64	82.1	52.6	67.9	469.8	29.3	42.1	335.2	53.1
CLIP+FDT	145M	69.0(↑ 5.0)	82.3 (↑ 0.2)	56.3(↑3.7)	75.9(↑8.0)	489.4(↑19.6)	31.0(↑1.7)	46.4(↑4.3)	353.0(↑17.8)	55.2(↑2.1)

Table 5. Ablation study results when using different scales of training data. “ZS” means zero-shot. “AVG” is average. “ACC” is accuracy. “LP” stands for linear prob. “CLS” represents classification. “IR” and “TR” are image retrieval and text retrieval, respectively.

	ZS CLS AVG Acc	LP CLS AVG Acc	ZS-Flickr30K IR R@1	ZS-Flickr30K TR R@1	ZS-Flickr30K rsum	ZS-MSCOCO IR R@1	ZS-MSCOCO TR R@1	ZS-MSCOCO rsum	VQAv2 Overall
CLIP-ViT-B/32	41.3	67.5	27.6	42.8	343.1	15.9	24.8	236.8	47.5
CLIP-ViT-B/32+FDT	45.9(↑4.6)	68.8(↑1.3)	32.6(↑5.0)	51.0(↑8.2)	376.5(↑33.4)	19.4(↑3.5)	29.6(↑4.8)	263.1(↑26.3)	50.6(↑3.1)
CLIP-ViT-B/16	45.2	68.8	35.3	50.5	387.8	19.3	29.7	263.6	49.2
CLIP-ViT-B/16+FDT	49.9(↑4.7)	71.3(↑2.5)	41.6(↑6.3)	60.8(↑10.3)	425.5(↑37.7)	23.4(↑4.1)	35.3(↑5.6)	295.4 (↑31.8)	54.3(↑5.1)
CLIP-Swin-B	39.6	68.5	30.5	48.5	368.1	17.7	26.0	247.6	46.5
CLIP-Swin-B+FDT	42.4(↑2.8)	70.7(↑2.2)	39.6(↑9.1)	57.9(↑9.4)	415.5(↑47.4)	22.3(↑4.6)	33.8(↑7.8)	288.3(↑40.7)	51.6(↑5.1)

Table 6. Ablation Study results when using different image encoder architectures. “ZS” means zero-shot. “AVG” is average. “ACC” is accuracy. “LP” stands for linear prob. “CLS” represents classification. “IR” and “TR” are image retrieval and text retrieval.

Experiments

FDT size	ZS CLS AVG Acc	LP CLS AVG Acc	ZS-Flickr30K IR R@1	ZS-Flickr30K TR R@1	ZS-Flickr30K rsum	ZS-MSCOCO IR R@1	ZS-MSCOCO TR R@1	ZS-MSCOCO rsum	VQAv2 overall
-	41.3	67.5	27.6	42.8	343.1	15.9	24.8	236.8	47.5
8192	42.8	67.9	32.7	50.6	374.6	18.5	29.1	258.1	50.1
16384	**45.9**	**68.8**	32.6	**51.0**	376.5	**19.4**	29.6	**263.1**	50.6
24576	45.2	68.6	**33.3**	50.4	**378.5**	18.6	**29.7**	**263.1**	**51.4**

Table 7. Results of the models with different FDT sizes. The row whose FDT value is "-" represents the original CLIP model. "ZS" means zero-shot. "AVG" is average. "ACC" is accuracy. "LP" stands for linear prob. "CLS" represents classification. "IR" and "TR" are image retrieval and text retrieval.

	ZS CLS AVG Acc	LP CLS AVG Acc	ZS-Flickr30K IR R@1	ZS-Flickr30K TR R@1	ZS-Flickr30K rsum	ZS-MSCOCO IR R@1	ZS-MSCOCO TR R@1	ZS-MSCOCO rsum	VQAv2 overall
CLIP	41.3	67.5	27.6	42.8	343.1	15.9	24.8	236.8	47.5
CLIP+FDT$_{\text{Softmax}}$ *	5.2	-	5.4	1.7	45.5	2.4	0.8	26.2	-
CLIP+FDT$_{\text{Sparsemax}}$ *	32.4	-	10.5	32.5	242.4	6.0	18.3	157.5	-
CLIP+FDT$_{\text{Softmax}}$	43.9	68.7	33.3	47.9	377.6	19.2	28.3	258.8	47.9
CLIP+FDT$_{\text{Sparsemax}}$	45.9	68.8	32.6	51.0	376.5	19.4	29.6	263.1	50.6

Table 8. Results of models trained with (Sparsemax) and without (Softmax) sparse constraints. The rows marked with "*" are the results when using FDT weights as features (see Section 4.3). "ZS" means zero-shot. "AVG" is average. "ACC" is accuracy. "LP" stands for linear prob. "CLS" represents classification. "IR" and "TR" are image retrieval and text retrieval.

Experiments

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

News: Rumor has it that Xun Chi, the President of Rongchuang North China, demanded executives purchase unsold properties in Henan to rescue the market. In an attempt to boost the market, Rongchuang initiated a widespread marketing campaign internally. They reserved three buildings for this project and required middle and upper-level executives to make purchases. Xun Chi even stated explicitly during a marketing meeting that this all-encompassing marketing campaign in the cultural tourism city aims to identify and promote talents based on their participation in reservation procedures. Through this process, they intend to identify those truly dedicated to Rongchuang's cause.
LLM Rationale (Textual Description): The style of this message appears colloquial, featuring slang and popular phrases. Additionally, it contains some grammatical errors and lacks precision in expression.
LLM Rationale (Commonsense): Based on common sense and experience, it's highly unlikely for executives of a company to ask other executives to buy unsold properties to rescue the market, and even more improbable to promote individuals based on their participation in a reservation program.

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: Urgent situation! 8 hours ago, during a speech, Trump almost fainted and subsequently canceled his upcoming engagements.
LLM Rationale (Textual Description): This message employs a colloquial style of expression, using phrases like "Urgent situation!" and "almost fainted," which deviates from the formal language typically used in news reporting. Therefore, it's more likely to be an informal social media post rather than an authentic news report.
LLM Rationale (Commonsense): Based on common sense, if an individual nearly faints during a speech and cancels subsequent engagements, it's highly likely to be a true situation.

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

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

News: [Outbreak of African Swine Fever in Hainan, Capable of Remaining Latent in Humans for 15 Years!] The Chinese Ministry of Agriculture and Rural Affairs' news office reported on April 19, 2019, at 5:15 PM Beijing time, that African swine fever has occurred in Wanning City and Danzhou City of Hainan Province. With this development, all 31 provinces, municipalities, and autonomous regions in China have now reported cases of swine fever. As per the official notification on the Chinese Ministry of Agriculture and Rural Affairs website, on April 19, the ministry was informed by the China Animal Disease Prevention and Control Center. Following confirmation by the China Animal Health and Epidemiology Center, two livestock breeders in Danzhou City, Hainan Province, were affected by the outbreak.
LLM Rationale (Textual Description): This message employs a rather meticulous language style, providing specific details of time and location, while also referencing an official announcement.
LLM Rationale (Commonsense): Based on common sense, African swine fever is a virus that exclusively infects pigs and does not remain latent within the human body for 15 years.

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: #HumenBridge# Experts are unreliable, and many experts lack integrity. The issue has finally been identified: Cable No. 38's wire rope has fractured! This happened at the anchor head on the inner side of the steel box beam, where the old cable had rusted from the inside out. Currently, the cable has been dismantled, and further examination revealed multiple instances of corrosion. Main cable inspections revealed severe corrosion in the center span's mid-position, with 3-4 layers of heavily corroded steel cables.
LLM Rationale (Textual Description): The writing style of this message is serious and objective, devoid of strong emotional undertones. Additionally, it provides specific details and data, contributing to a sense of authenticity and credibility.
LLM Rationale (Commonsense): Given the history and construction standards of the Humen Bridge, such issues are indeed plausible.

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: [Typhoon Mangkhut's Power Equals 2600 Hiroshima Atomic Bombs? Humanity Can't Destroy the Earth, Only Ourselves] There's no need for humans to portray saving the Earth as noble; the Earth doesn't need your salvation. In reality, what humanity is saving is itself.
LLM Rationale (Textual Description): The language in this message does not align with conventional news reporting or scientific research standards. It employs exaggerated terms and emotionally charged expressions.
LLM Rationale (Commonsense): Claiming that Typhoon Mangkhut's power is equivalent to 2600 Hiroshima atomic bombs based on scientific data is incorrect. Therefore, this message is more likely to be false information.

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

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***.

Thanks!