Bad Actor, Good Advisor : Exploring the Role of LLMs 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]Institute of Computing Technology, Chinese Academy of Sciences [2]University of Chinese Academy of Sciences

[3]National University of Singapore

Motivation

- **Challenge:** Fake news detection needs understanding of the real-world background, which is tough for methods based on Small Language Models (SLMs).
- **Potential:** Large Language Models (LLMs) like GPT3.5-turbo have shown impressive emergent abilities on various tasks and are considered promising as general task solvers.
- **Questions:** Can LLMs help detect fake news with their internal knowledge and capability? What solution should we adopt to obtain better performance using LLMs?

Is the LLM a Good Detector?

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%)**

↑ The LLM underperforms the fine-tuned SLM using all four prompting approaches.

→ The LLM is capable of generating human-like rationales on news content from various perspectives.

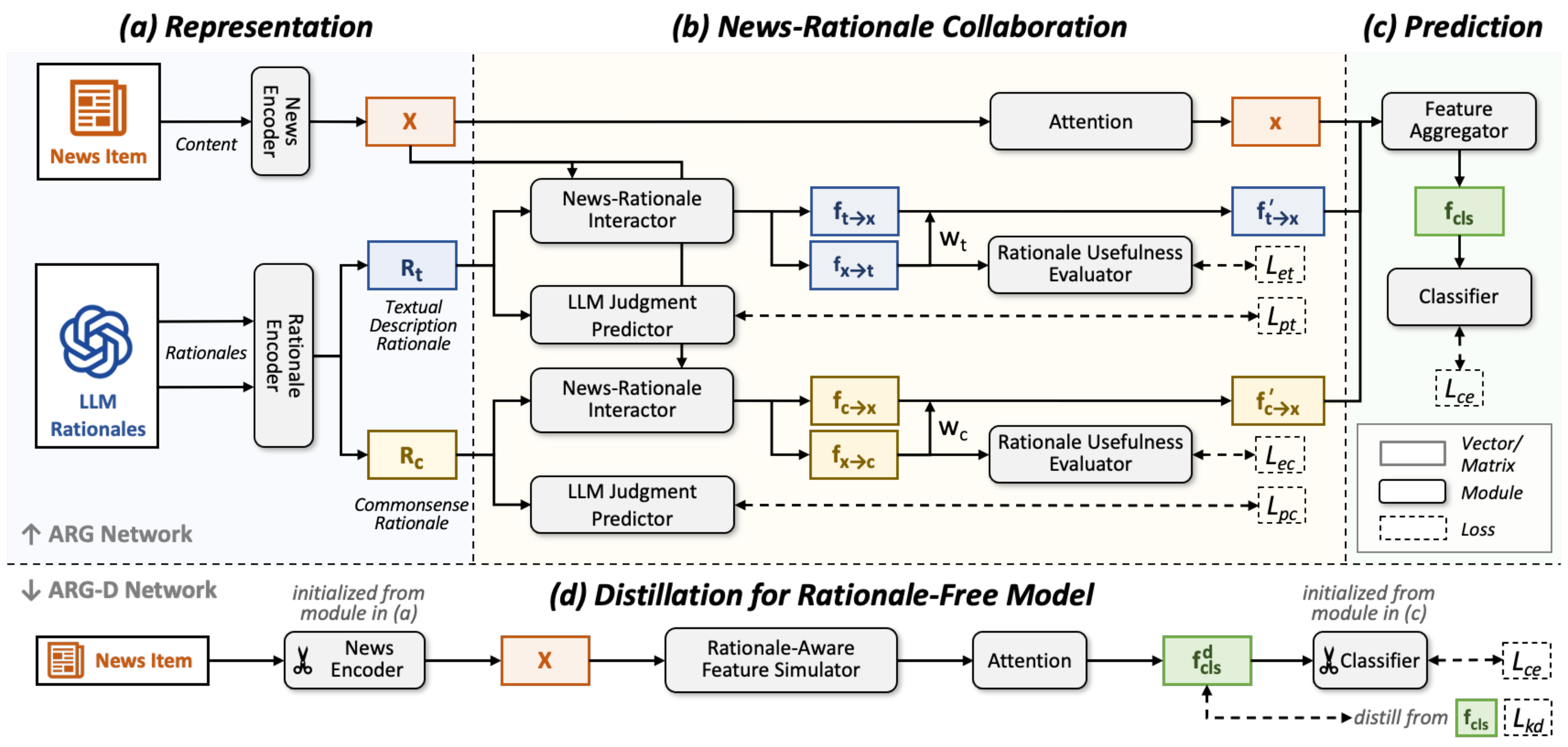
↘ It's plausible to gain a performance better than any LLM or SLM-only methods if we could adaptively combine their advantages.

LLM could be a good advisor for the SLM by providing rationales

Perspective	Chinese Proportion	Chinese macF1	English Proportion	English 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

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

Our Method: Adaptive Rationale Guidance (ARG) Network



Module a: Representation

- Employ two BERT models separately as the news and rationale encoder to obtain semantic representations.

Module b: News-Rationale Collaboration

① **News-Rationale Interaction:** Introduce a news-rationale interactor with a dual cross-attention mechanism, which generates content-base and rationale-base attention-feature respectively.

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

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

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

② **LLM Judgement Prediction:** Predict the LLM judgment of the news veracity according to the given rationale.

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

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

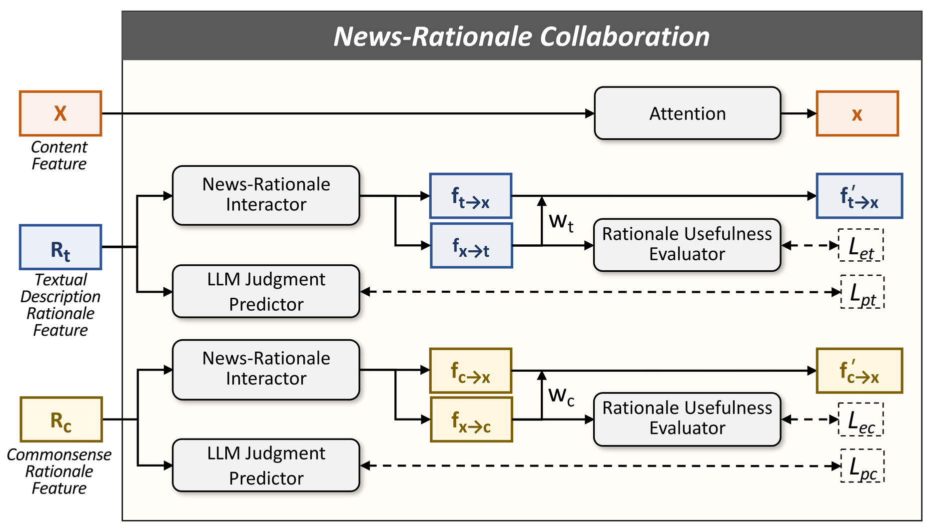
③ **Rationale Usefulness Evaluation:** Assess the contributions of different rationales and adjust their weights for subsequent veracity prediction.

- ***Evaluation*** **:** Regard if the judgment along with the rationale is correct as the usefulness labels, and predict the rationales' usefulness:

$$\hat{u}_t = \mathrm{sigmoid}(\mathrm{MLP}(\mathbf{f}_{\mathbf{x}\to\mathbf{t}})), L_{et} = \mathrm{CE}(\hat{u}_t, u_t)$$

- ***Reweighting*** **:** Obtain a weight number from vector $\mathbf{f}_{\mathbf{x}\to\mathbf{t}}$ to reweight the rationale-aware news vector $\mathbf{f}_{\mathbf{t}\to\mathbf{x}}$:

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



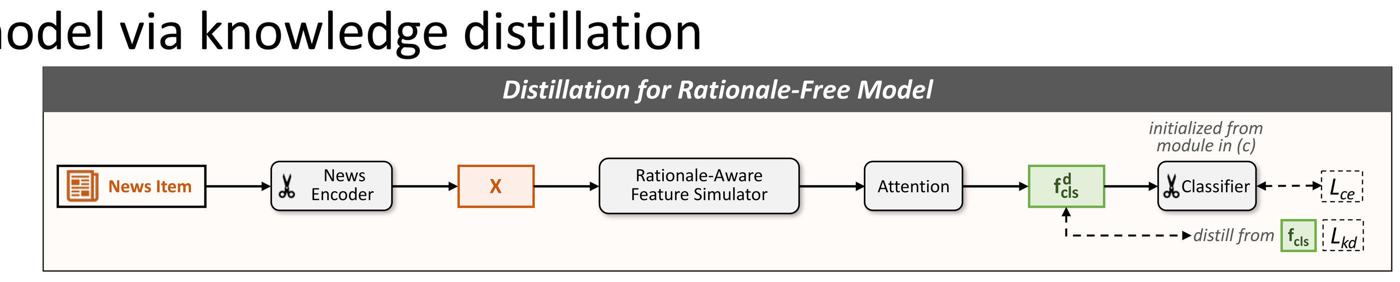
Module c: Prediction

- We aggregate news vector and rationale-aware news vectors for the final judgment. For news item x, we aggregate vectors with different weights:

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

Module d: Distillation for Rationale-Free Model

- We build a rationale-free model ARG-D based on the trained ARG model via knowledge distillation



Evaluation

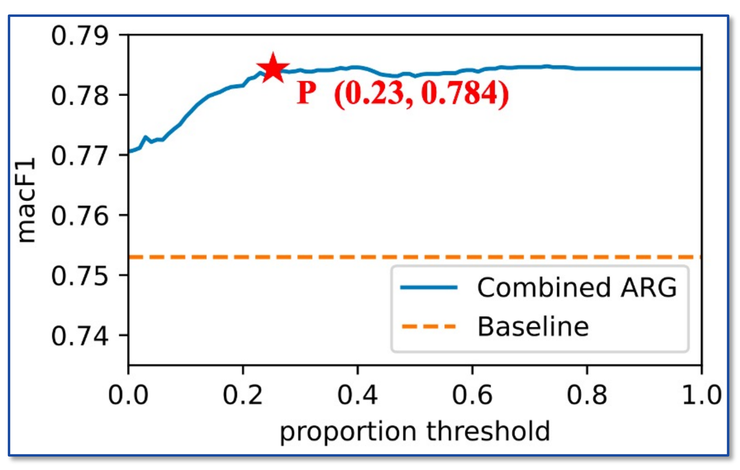
	Model	Chinese macF1	Chinese Acc.	Chinese $F1_{real}$	Chinese $F1_{fake}$	English macF1	English Acc.	English $F1_{real}$	English $F1_{fake}$
G1: LLM-Only	GPT-3.5-turbo	0.725	0.734	0.774	0.676	0.702	0.813	0.884	0.519
G2: SLM-Only	Baseline	0.753	0.754	0.769	0.737	0.765	0.862	0.916	0.615
	$EANN_T$	0.754	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**	<u>0.878</u>	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	<u>0.786</u>	**0.880**	0.928	0.645
	w/o Rationale Usefulness Evaluator	<u>0.781</u>	<u>0.783</u>	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%)*

Experiment: ↑

- The ARG outperforms all other compared methods in macro F1
- The rationale-free ARG-D still outperforms all compared methods except ARG and its variants

Cost Analysis in Practice: →

By sending only 23% of the data to the ARG, we could achieve 0.784 in macro F1, which is the same as the performance fully using the ARG.



Conclusion

- **Answer 1:** We found 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.
- **Answer2:** 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.

GitHub Repo