

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

Beizhe Hu^{1,2}, Qiang Sheng¹, Juan Cao^{1,2}, Yuhui Shi^{1,2}, Yang Li^{1,2}, Danding Wang¹, Peng Qi³

¹Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences

²University of Chinese Academy of Sciences

³National University of Singapore

Presenter: Beizhe Hu





Media Synthesis &
Forensics Lab

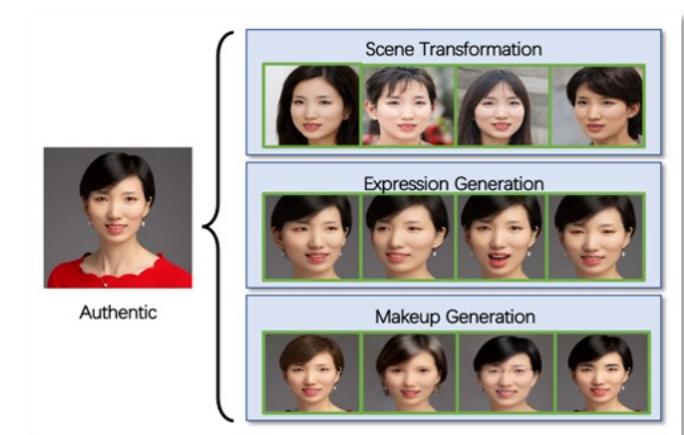
MAKE THE WORLD MORE CREDIBLE



Fake News Detection
Fact-Checking



Deep Synthesized Media
Detection & Attribution



Attacking & Defense
for AI Safety

Motivation



- Fake news detection is a challenging task due to the complexity of the news-faking process, which requires both a delicate sense of diverse **clues** and a profound understanding of the **real-world background**.
- The usage of **small language models (SLMs)** like BERT do bring improvements, but their knowledge and capability limitations also compromise further enhancement.
- **Large language models (LLMs)** like GPT3.5-turbo have shown impressive emergent abilities on various tasks and are considered promising as general task solvers.

Motivation



- Fake news detection is a challenging task due to the complexity of the news-faking process, which requires both a delicate sense of diverse **clues** and a profound understanding of the **real-world background**.
- The usage of **small language models (SLMs)** like BERT do bring improvements, but their knowledge and capability is limited.
- Large language models (LLMs) have the potential to address these challenges.

- **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?



Experimental Settings

- SLM: Fine-tuning BERT
- LLM: GPT-3.5-turbo using 4 typical prompting methods
 - Zero-Shot
 - Zero-Shot CoT
 - Few-Shot
 - Few-Shot CoT

(a) Zero-shot

Q: Given the following message, predict its veracity. If it is more likely to be a real message, return 1; otherwise, return 0: Detailed photos of Xiang Liu's tendon surgery exposed. Stop complaints and please show sympathy and blessings!

A:

1

(b) Zero-shot CoT

Q: Given the following message, predict its veracity. If it is more likely to be a real message, return 1; otherwise, return 0: 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... 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

(e) Zero-shot CoT with Perspective Specified

Q: Given the following message, predict its veracity. If it is more likely to be a real message, return 1; otherwise, return 0: 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.

(c) Few-shot

Q: Given the following message, predict its veracity. If it is more likely to be a real message, return 1; otherwise, return 0: Breaking News: Tianyi Li's successful exoneration!...

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: Detailed photos of Xiang Liu's tendon surgery exposed. Stop complaints and please show sympathy and blessings!

A:

1

(d) Few-shot CoT

Q: Given the following message, predict its veracity. If it is more likely to be a real message, return 1; otherwise, return 0: Breaking News: Tianyi Li's successful exoneration!...

A: 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, the answer is 0.

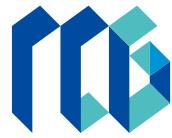
...

Q: Given the following message, predict its veracity. If it is more likely to be a real message, return 1; otherwise, return 0: 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...Therefore, the response is 1

Is the LLM a Good Detector?



Comparison between Small and Large LMs

- LLM underperforms the fine-tuned SLM using all four prompting approaches
- Few-shot versions outperform zero-shot ones, suggesting the importance of task samples
- CoT prompting brings additional performance gain in general

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%)

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.

Is the LLM a Good Detector?



Analysis on the Rationales from the LLM

- **How to do:**

- Sample 500 samples from each of the two datasets and manually categorize their zero-shot CoT outputs

- **Observation:**

- The LLM is capable of generating **human-like rationales** on news content from **various perspectives**
- The detection performance **using certain perspectives is higher** than the **zero-shot CoT**, indicating the potential of analysis by perspectives (the coverage is moderate)
- The analysis from the perspective of **factuality** leads to the performance **lower than average**, which may be caused by the **hallucination** issue

Perspective	Chinese		English	
	Proportion	macF1	Proportion	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

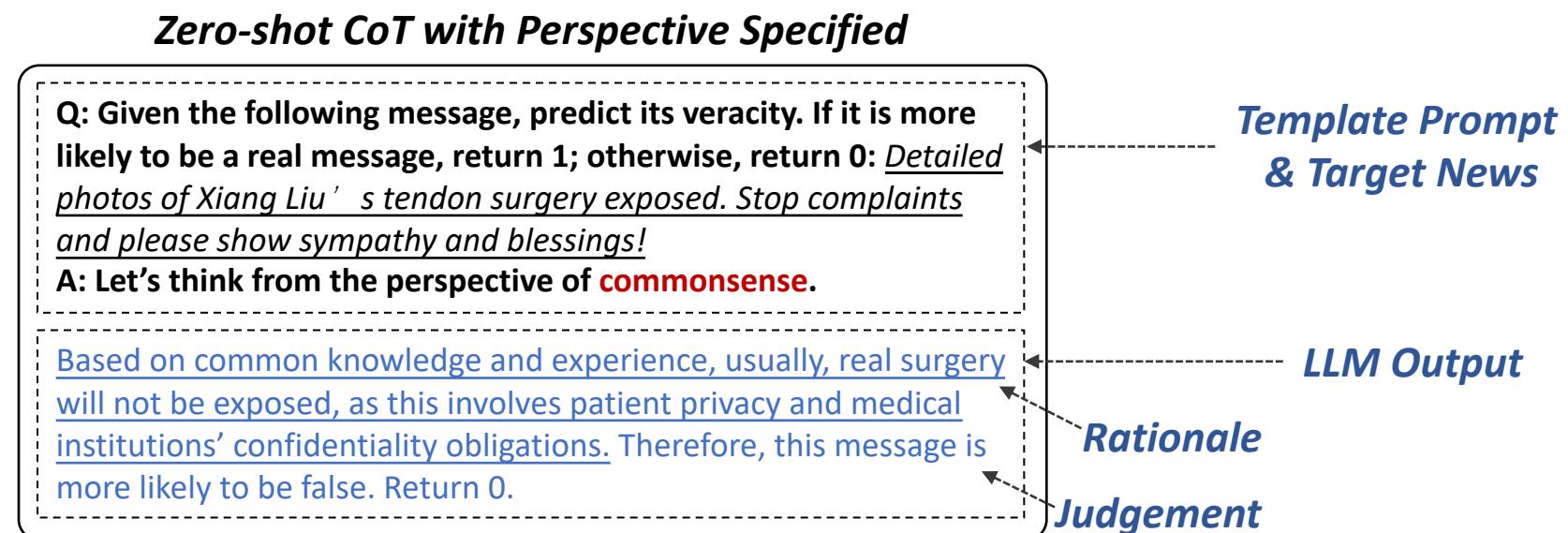
Is the LLM a Good Detector?



Analysis on the Rationales from the LLM

- **How to do:**

- Prompt the LLM to perform analysis from a **specific perspective** with a eliciting sentence like “Let’s think from the perspective of [textual description/commonsense].”



Is the LLM a Good Detector?



Analysis on the Rationales from the LLM

- **How to do:**

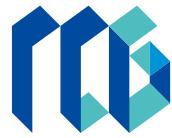
- Prompt the LLM to perform analysis from a **specific perspective** with a eliciting sentence like “Let’s think from the perspective of [textual description/commonsense].”

- **Observation:**

- The LLM’s judgment with **single perspective** analysis elicited is **still promising**
- Internal mechanism of the LLM to integrate the rationales is ineffective for fnd, limiting the full use of rationales -> **combining the small and large LMs to complement each other is a promising solution**

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

Is the LLM a Good Detector?



Analysis on the Rationales from the LLM

- **How to do:**

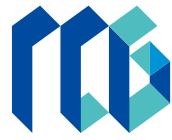
- Apply **majority voting** and **oracle voting** (assuming the most ideal situation where we trust the correctly judged model for each sample, if any) among the **two single-perspective-based LLMs** and the **BERT**

- **Observation:**

- It's plausible 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

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

Is the LLM a Good Detector?



Analysis on the Rationales from the LLM

- How to do:
 - Apply majority voting and oracle voting (assuming the most ideal situation where we trust the correctly
 -

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

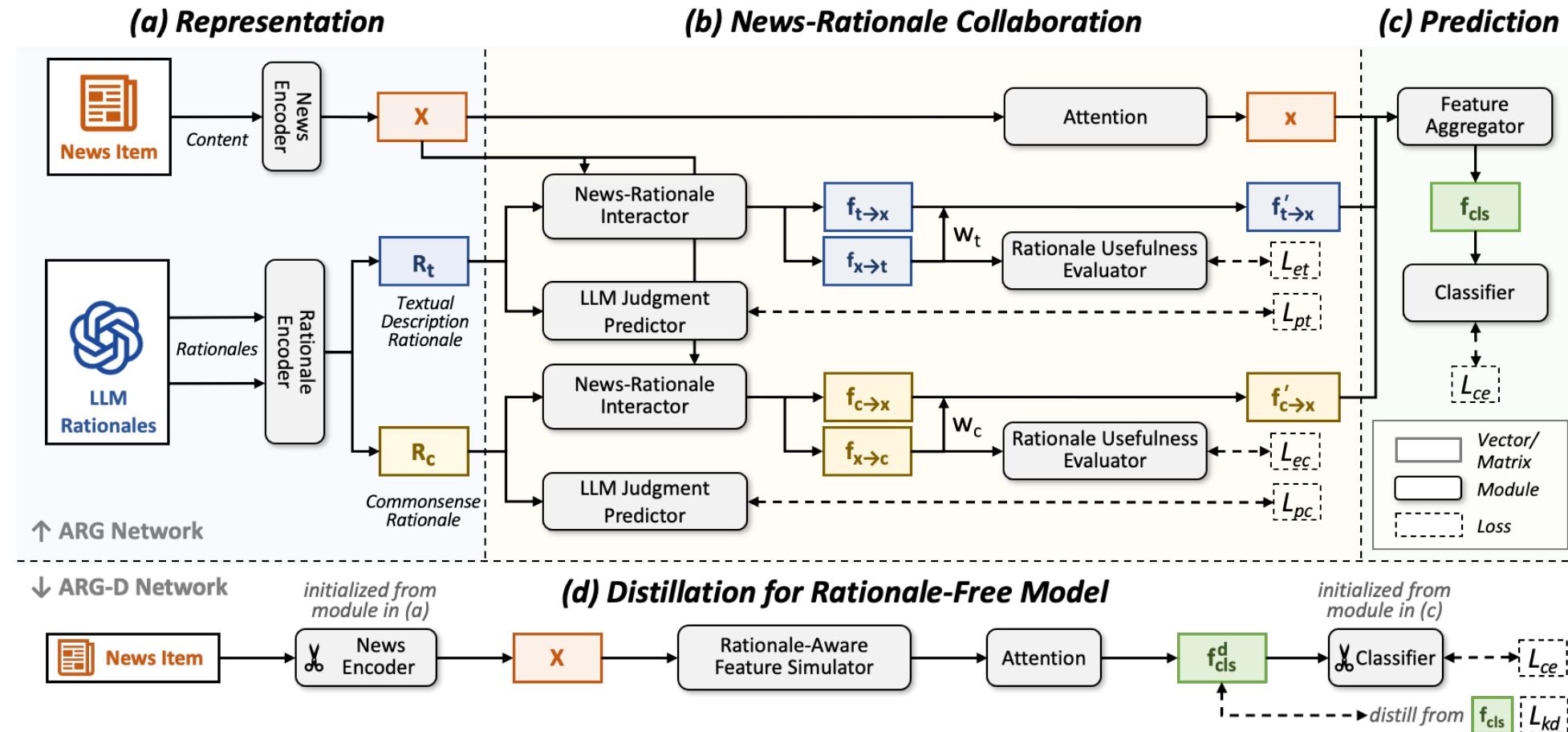
GPT-3.5-turbo	from Perspective TD	0.667	0.611
	from Perspective CS	0.678	0.698
BERT	Fine-tuning	0.753	0.765
Ensemble	Majority Voting	0.735	0.724
	Oracle Voting	0.908	0.878

Method: Adaptive Rationale Guidance (ARG) Network



Components of the AGR Network

- Representation, News-Rationale Collaboration, Prediction, Distillation for Rationale-Free Model

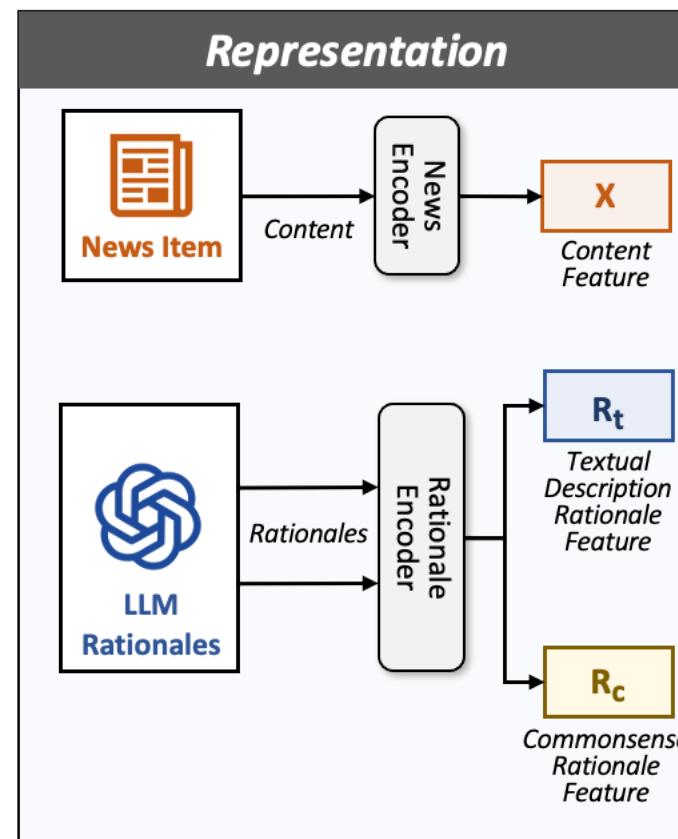


Method: Adaptive Rationale Guidance (ARG) Network

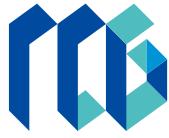


- **Representation**

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

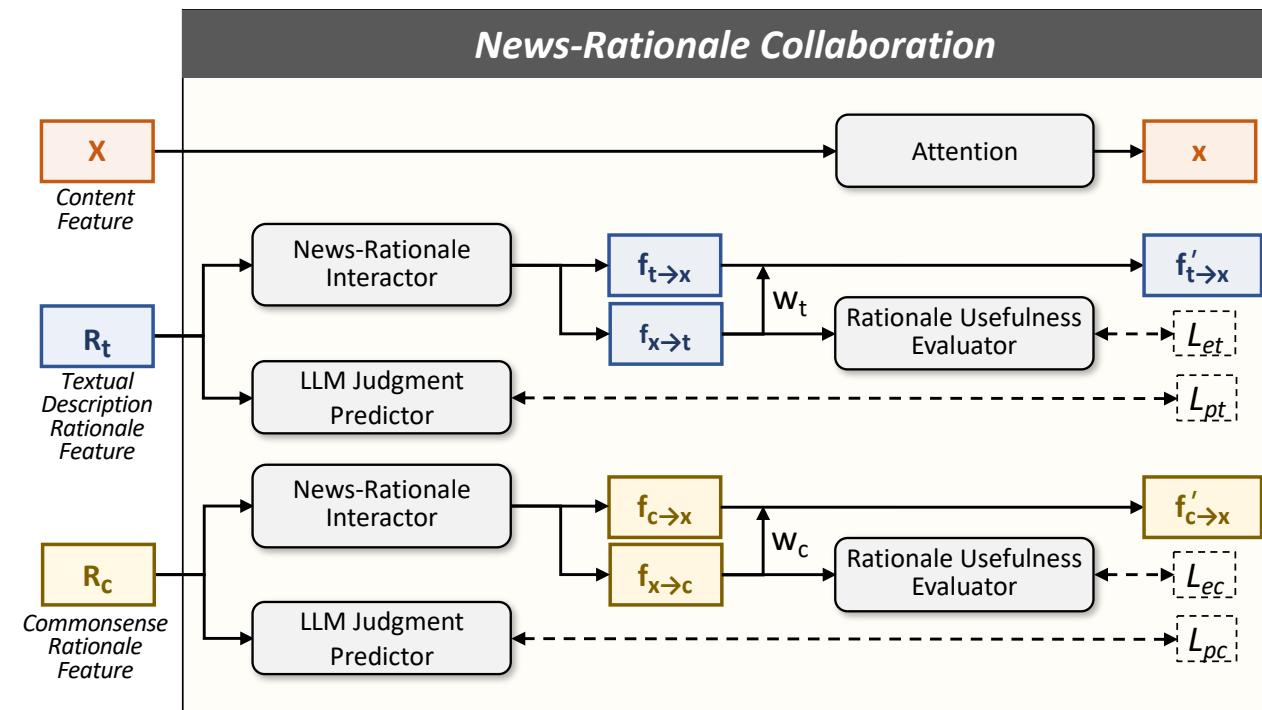


Method: Adaptive Rationale Guidance (ARG) Network

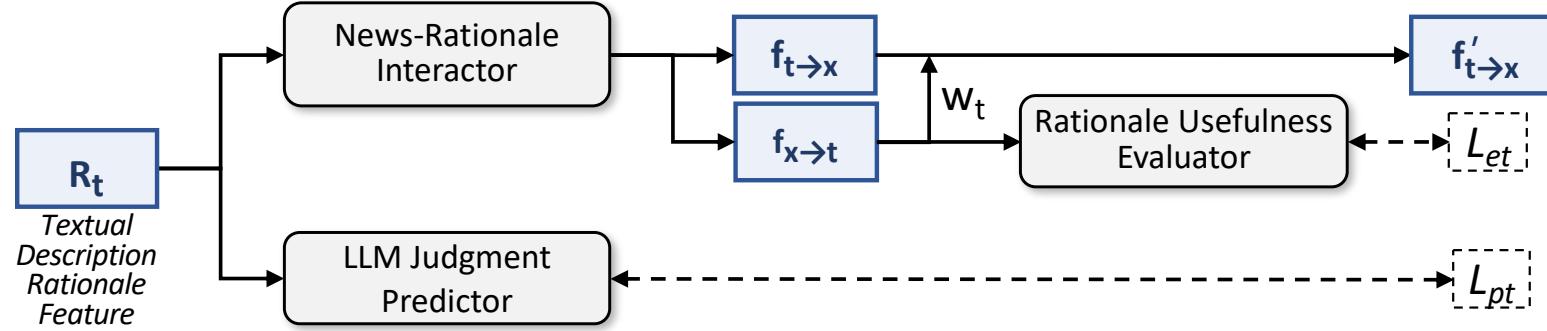


- **News-Rationale Collaboration**

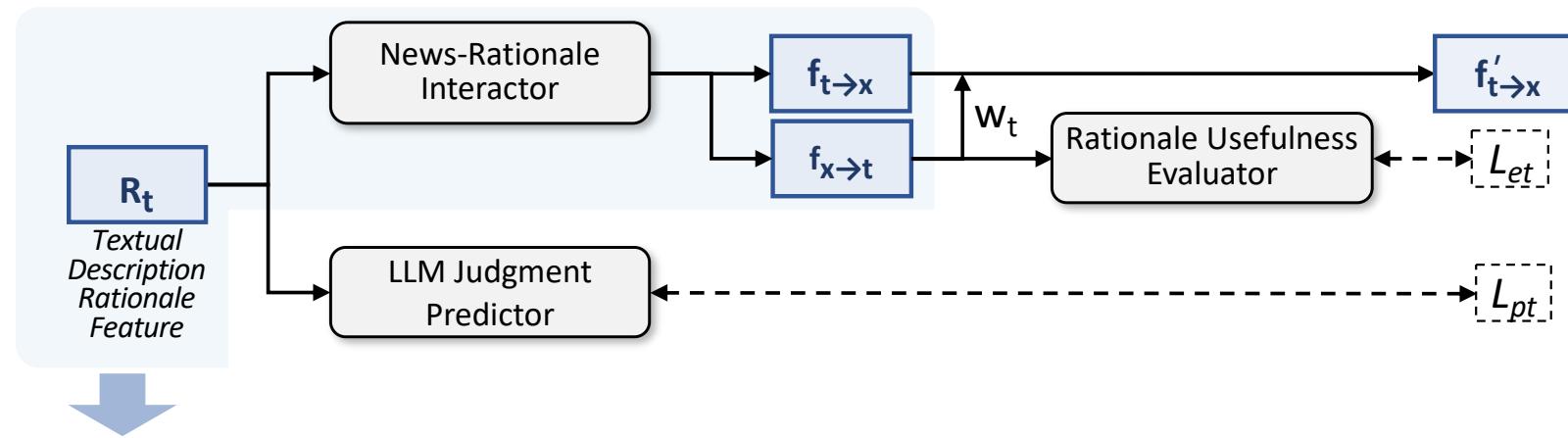
- Aim1: Provide a rich interaction between news and rationales
- Aim2: Enable the model to adaptively select useful rationales as references
- **Our design:** News-Rationale Interaction, LLM Judgement Prediction, Rationale Usefulness Evaluation



Method: Adaptive Rationale Guidance (ARG) Network



Method: Adaptive Rationale Guidance (ARG) Network



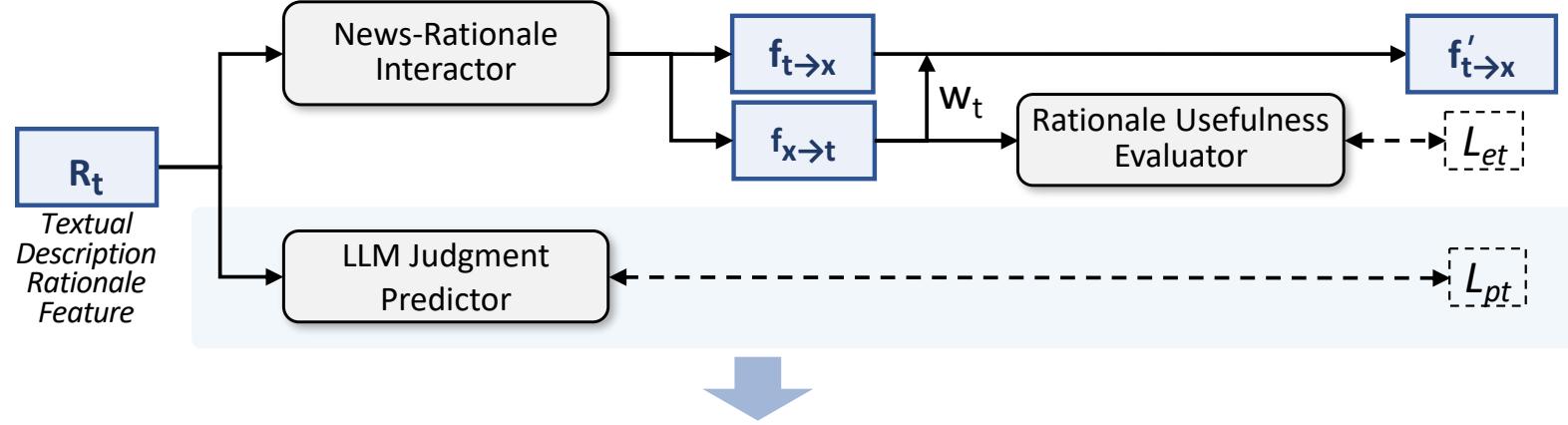
News-Rationale Interaction: Introduce a news-rationale interactor with a **dual cross-attention** mechanism, which generates **content-based** and **rationale-based** feature respectively.

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

$$f_{t \rightarrow x} = \text{AvgPool} (\text{CA}(R_t, X, X))$$

$$f_{x \rightarrow t} = \text{AvgPool} (\text{CA}(X, R_t, R_t))$$

Method: Adaptive Rationale Guidance (ARG) Network



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

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

$$f_{t \rightarrow x} = \text{AvgPool} (\text{CA}(R_t, X, X))$$

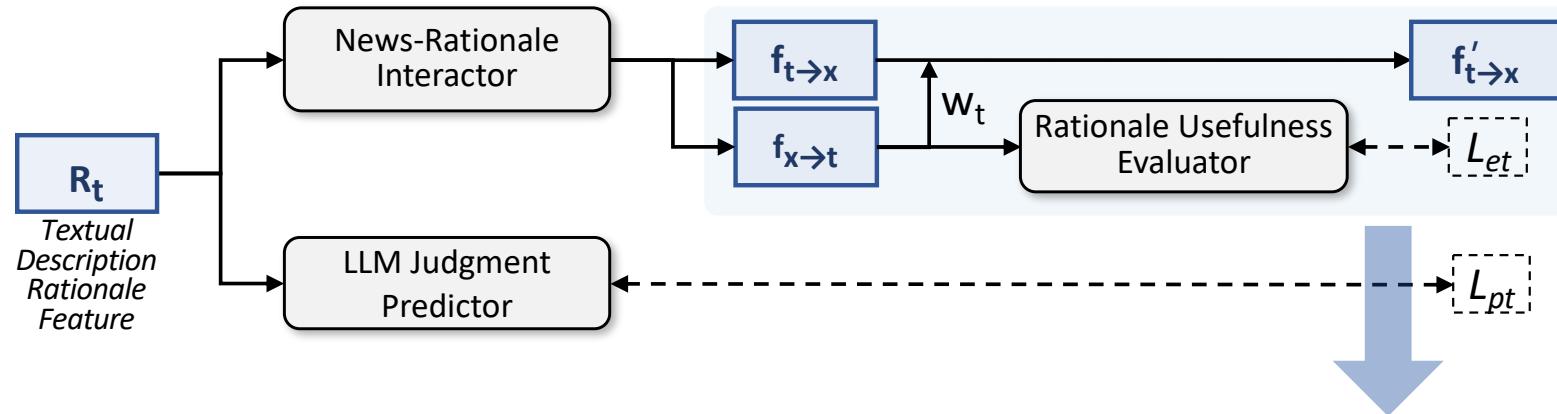
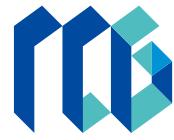
$$f_{x \rightarrow t} = \text{AvgPool} (\text{CA}(X, R_t, R_t))$$

LLM Judgement Prediction: To **deepen the understanding of the rationale texts**, predict the **LLM judgment** of the news veracity according to the given rationale.

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

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

Method: Adaptive Rationale Guidance (ARG) Network



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

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

$$\mathbf{f}_{t \rightarrow x} = \text{AvgPool}(\text{CA}(\mathbf{R}_t, \mathbf{X}, \mathbf{X}))$$

$$\mathbf{f}_{x \rightarrow t} = \text{AvgPool}(\text{CA}(\mathbf{X}, \mathbf{R}_t, \mathbf{R}_t))$$

LLM Judgement Prediction: To **deepen the understanding of the rationale texts**, predict the **LLM judgment** of the news veracity according to the given rationale.

$$\hat{m}_t = \text{sigmoid}(\text{MLP}(\mathbf{R}_t))$$

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

Rationale Usefulness Evaluation: To enable the model to **adaptively select rationale**, we **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:

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

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

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

Method: Adaptive Rationale Guidance (ARG) Network



- Prediction

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

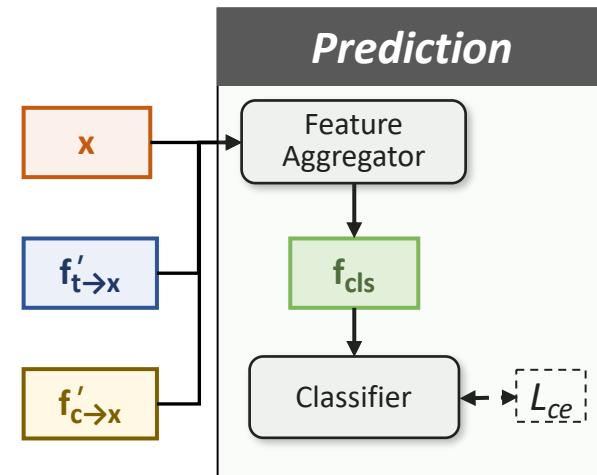
$$\mathbf{f}_{cls} = w_x^{cls} \cdot \mathbf{x} + w_t^{cls} \cdot \mathbf{f}'_{t \rightarrow x} + w_c^{cls} \cdot \mathbf{f}'_{c \rightarrow x}$$

- We use the fusion vector \mathbf{f}_{cls} to do the final prediction:

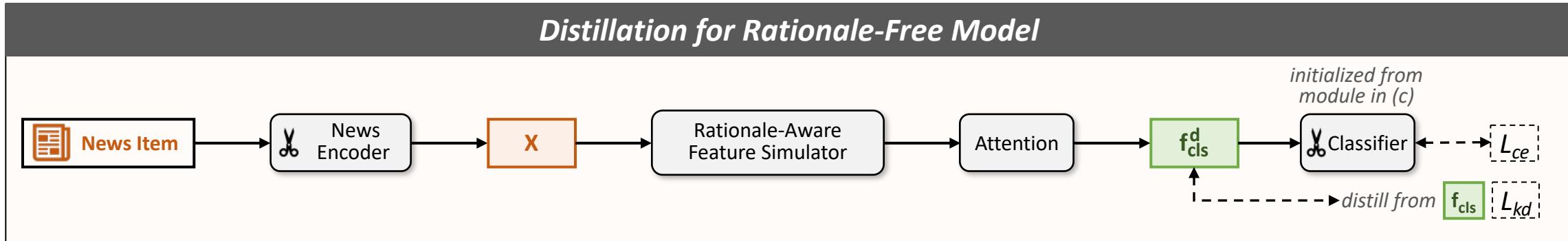
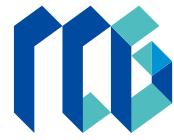
$$L_{ce} = \text{CE}(\text{MLP}(f_{cls}), y)$$

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

$$L = L_{ce} + \beta_1 L_{et} + \beta_2 L_{pt} + \beta_3 L_{ec} + \beta_4 L_{pc}$$



Method: Adaptive Rationale Guidance (ARG) Network



- **Distillation for Rationale-Free Model**

- The ARG requires sending requests to the LLM for **every prediction**, which might be not affordable for **cost-sensitive scenarios**.
- **Our solution:** build a rationale-free model ARG-D based on the trained ARG model via knowledge distillation
 - Initialize the news encoder and classifier with the corresponding modules in the ARG
 - Train a rationale-aware feature simulator and an attention module to internalize knowledge
 - Besides the cross-entropy loss, let the feature f_{cls}^d to imitate f_{cls} in the ARG:

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

Evaluation: Performance Comparison



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	<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%)

- **Observations:**
 - The **ARG** outperforms all other compared methods in macro F1

Evaluation: Performance Comparison



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	<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%)

- **Observations:**

- 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

Evaluation: Performance Comparison



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	<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%)

- **Observations:**

- 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**

Evaluation: Performance Comparison

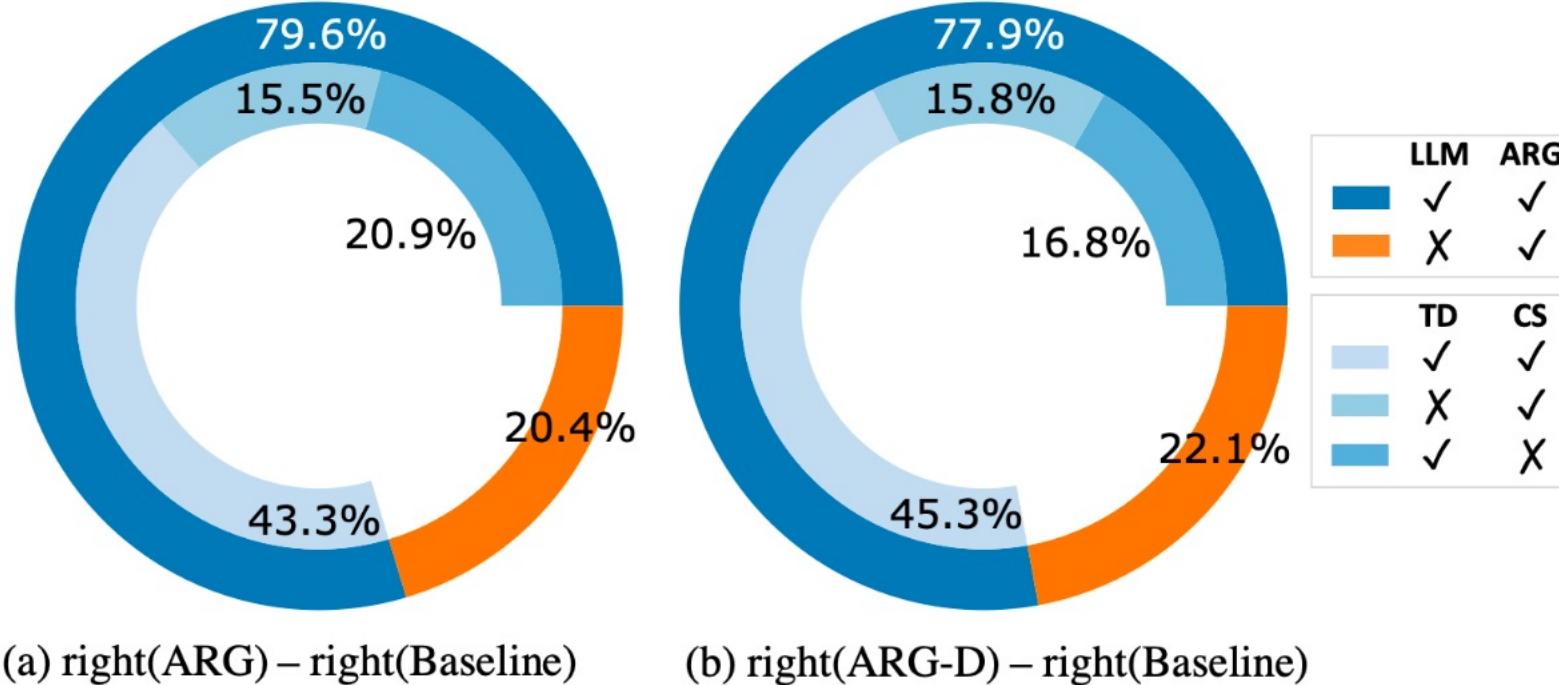


Model		Chinese				English			
		macF1	Acc.	F1 _{real}	F1 _{fake}	macF1	Acc.	F1 _{real}	F1 _{fake}
G1: LLM-Only	GPT-3.5-turbo	0.725	0.734	0.774	0.676	0.702	0.813	0.884	0.519
G2: SLM-Only	Baseline	0.753	0.754	0.769	0.737	0.765	0.862	0.916	0.615
	EANN _T	0.754	0.756	0.773	0.736	0.763	0.864	0.918	0.608
	Publisher-Emo	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	<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%)

- **Observations:**

- 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

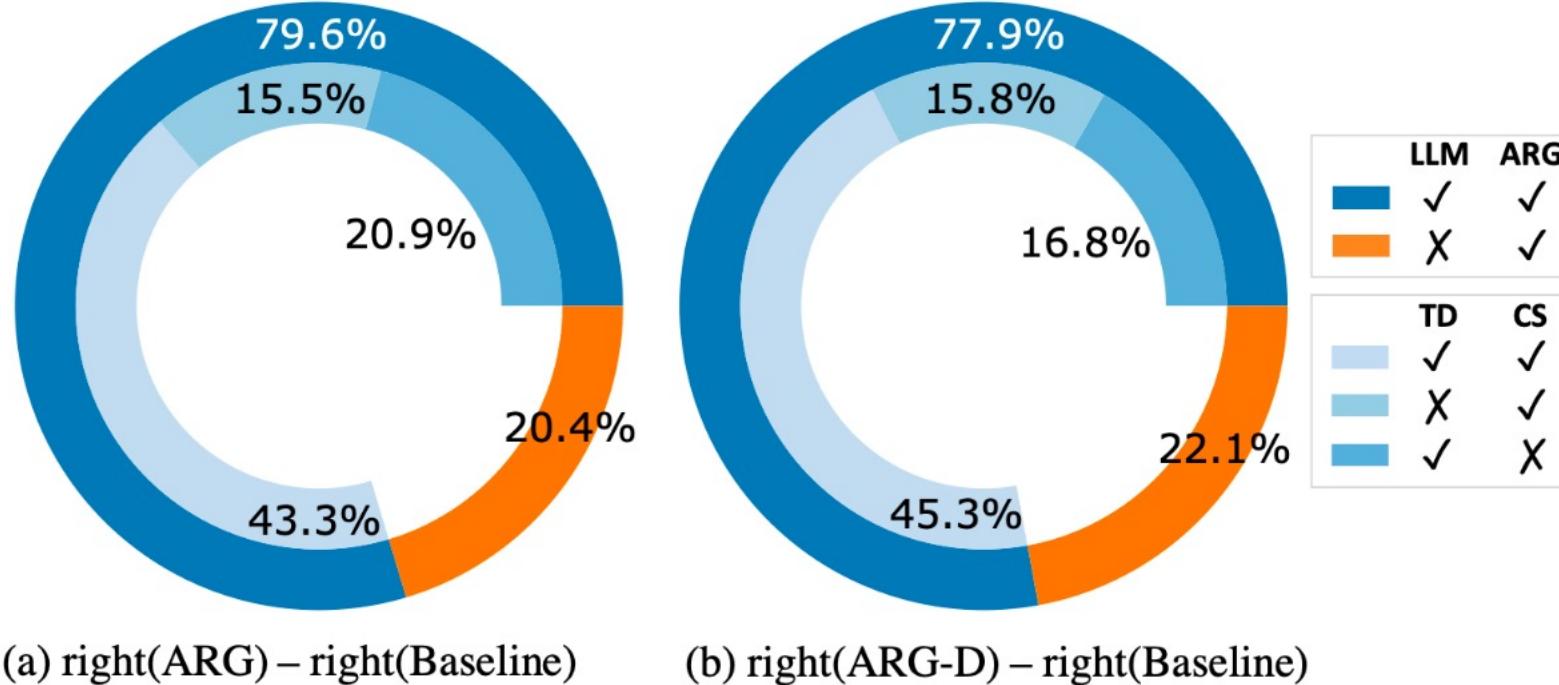
Evaluation: Result Analysis



- **How to do:**

- We perform statistical analysis on **the additional correctly judged samples** of ARG(-D) compared with the vanilla BERT.

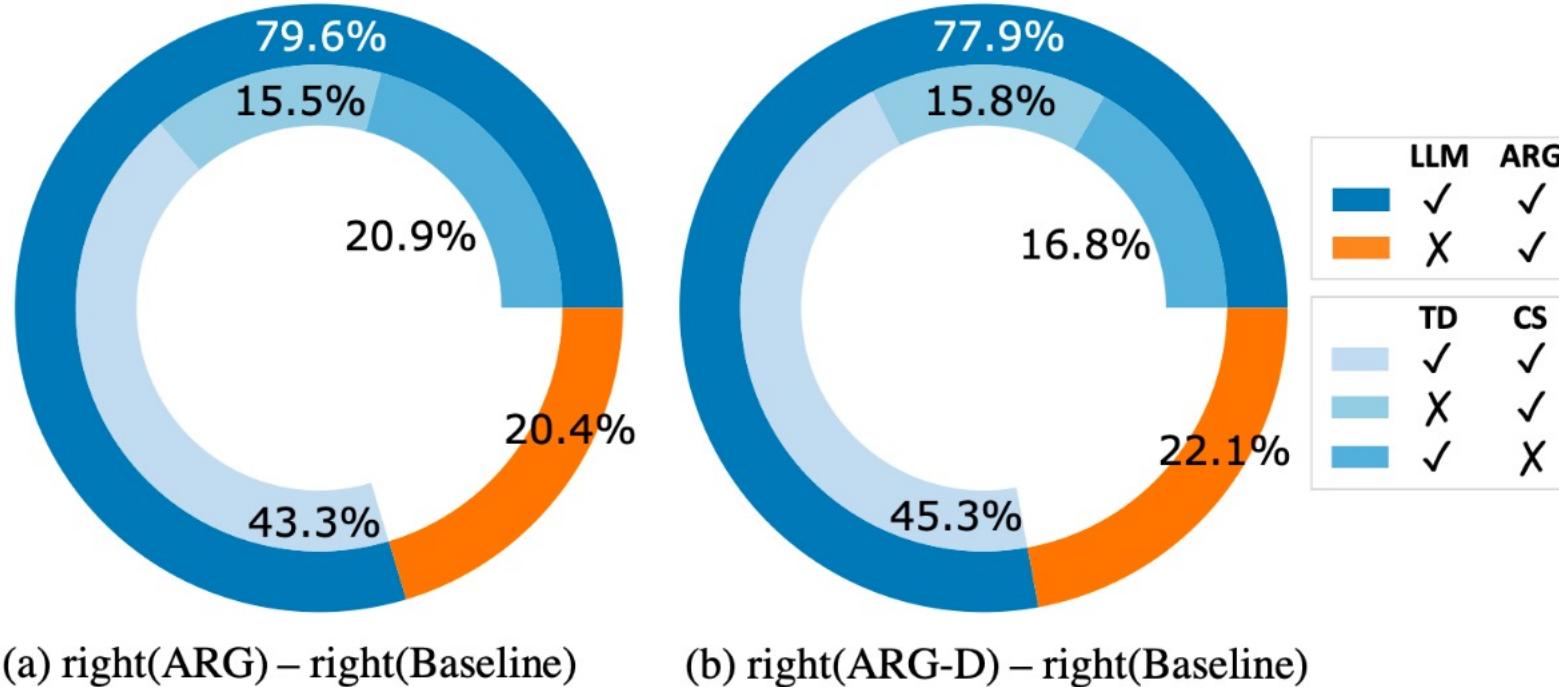
Evaluation: Result Analysis



- **Observation:**

- 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.

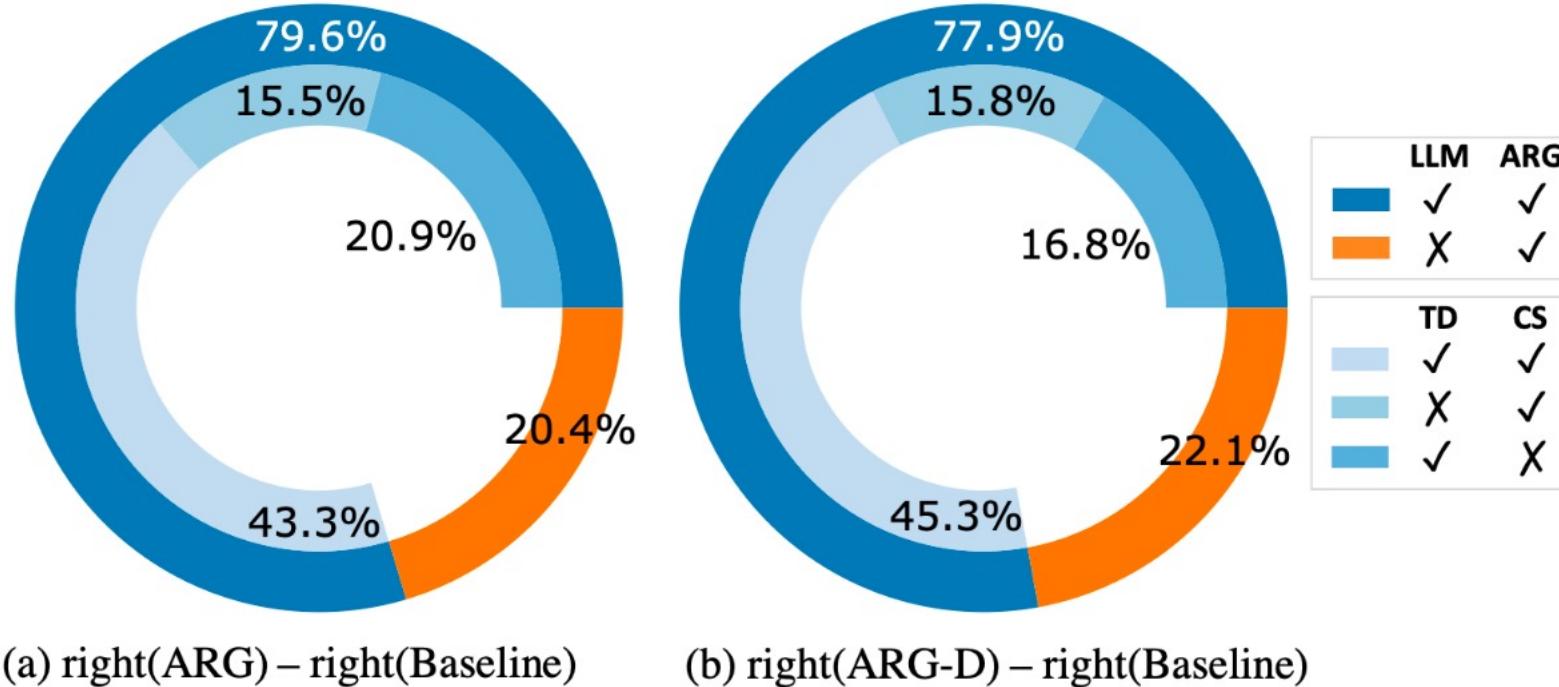
Evaluation: Result Analysis



- **Observation:**

- 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.

Evaluation: Result Analysis



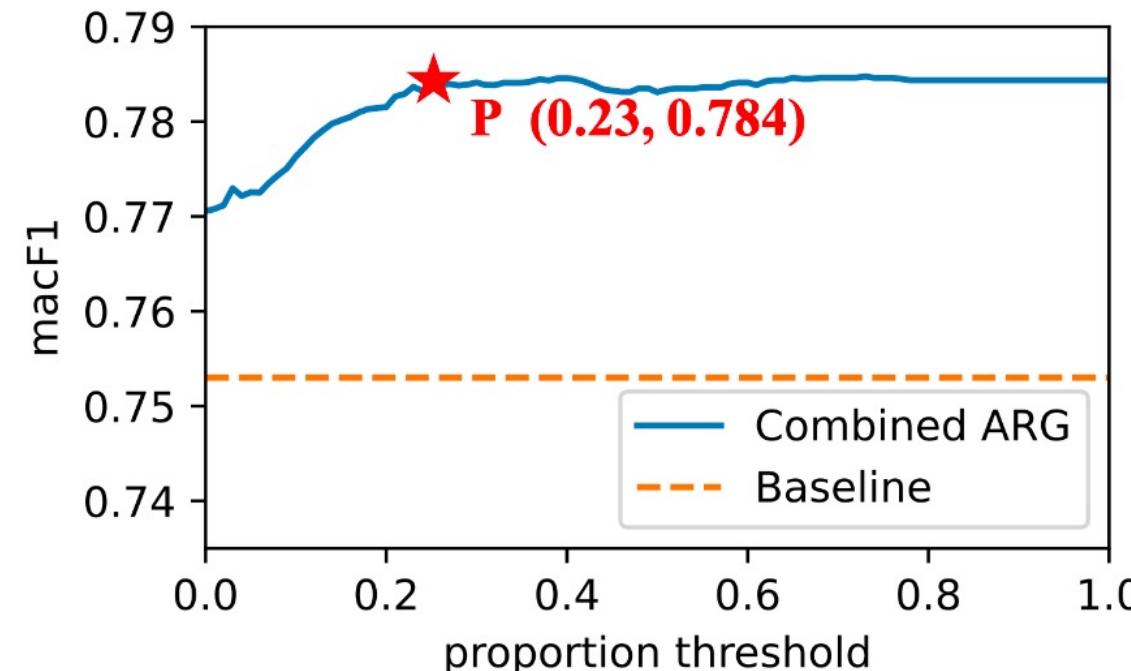
- **Observation:**

- 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

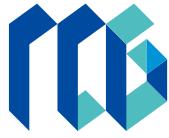
Evaluation: Cost Analysis in Practice



- **How to do:**
 - Simulate the situation where we use the **more economic ARG-D by default** but query the **more powerful ARG for part** of the data according to the confidence of ARG-D.
- **Observation:**
 - 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



- **Conclusion**

- We investigated if large LMs help in fake news detection and how to properly utilize their advantages for improving performance.
 - For **question 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.
 - For **question 2**, 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.

THANKS.



**Media Synthesis &
Forensics Lab**