SNIFFER: Multimodal Large Language Model for Explainable Out-of-Context Misinformation Detection

(CVPR 2024)

Peng Qi, Zehong Yan, Wynne Hsu, Mong Li Lee

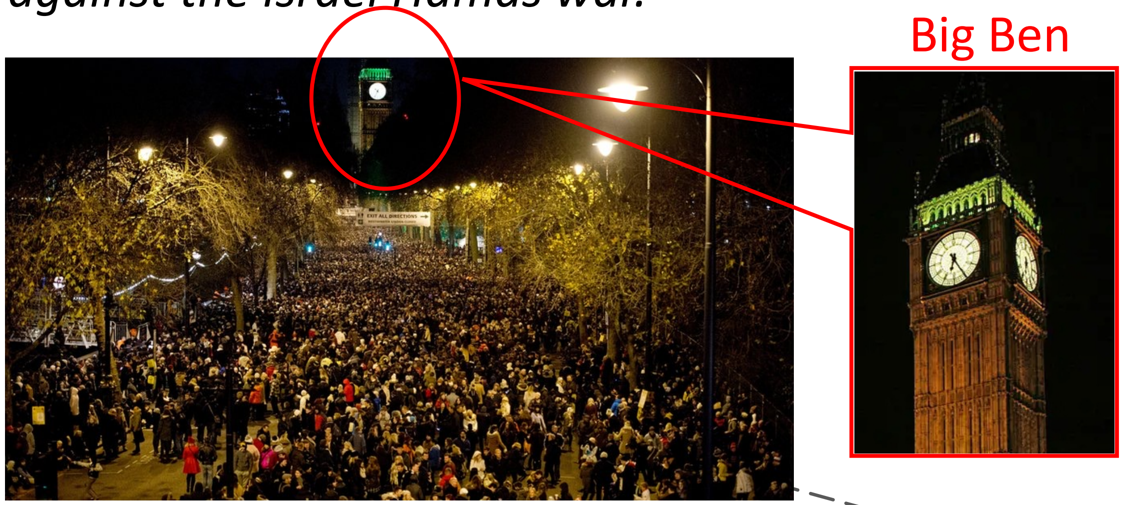
National University of Singapore

Project: https://pengqi.site/Sniffer

Overview: What have we done?

Thousands of people march in Madrid against the Israel Hamas war.

Real or Fake? 🤔



In London, thousands of people waiting for the fireworks to welcome the new year!

Original news

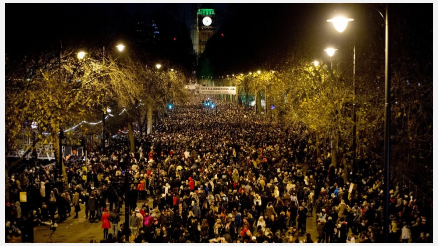
- **Out-of-Context (OOC) Misinformation**: Repurposing authentic images with false text.
  One of the easiest and most effective ways to mislead audiences.

iGyro

Overview: What have we done?

Does this caption match its image?

Caption: *Thousands of people march in Madrid against the Israel Hamas war.*



Existing detectors

🤖 No!

🤔

SNIFFER (our detector)

No, the image is wrongly used in a different news context. On the one hand, the image is inconsistent with the text. The text describes a protest in Madrid, while the image shows a large crowd in London, evident from the presence of Big Ben. On the other hand, the image-retrieved webpages are related to New Year’s Eve celebrations in London, not related to Madrid or a protest.

😄

better debunking!

➢ **Explainable Out-of-Context Misinformation Detection:** Provide explanation for the judgment.

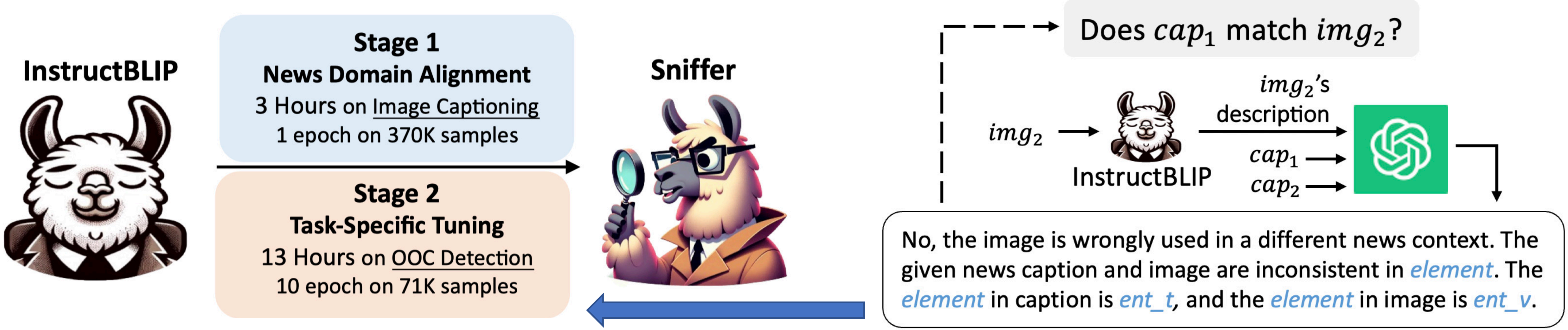
How to achieve it? – Two-stage Instruction Tuning

- ***Challenge 1**: **Bad performance of existing open-source MLLMs in OOC detection.***

  e.g. InstructBLIP-13B achieved 47.4% accuracy with only 4.6% recall for fake classes.

  Possible reason: the assumption of text-image consistency in their training corpus

🤔 **We need to design a task-specific multimodal large language model!**



InstructBLIP

Stage 1
News Domain Alignment
3 Hours on Image Captioning
1 epoch on 370K samples

Stage 2
Task-Specific Tuning
13 Hours on OOC Detection
10 epoch on 71K samples

Sniffer

Does cap_1 match img_2?

img_2 → InstructBLIP → img_2's description
cap_1 →
cap_2 →

No, the image is wrongly used in a different news context. The given news caption and image are inconsistent in *element*. The *element* in caption is *ent_t*, and the *element* in image is *ent_v*.

Instruction Tuning Process

OOC instruction data generation: Judgment + Explanation

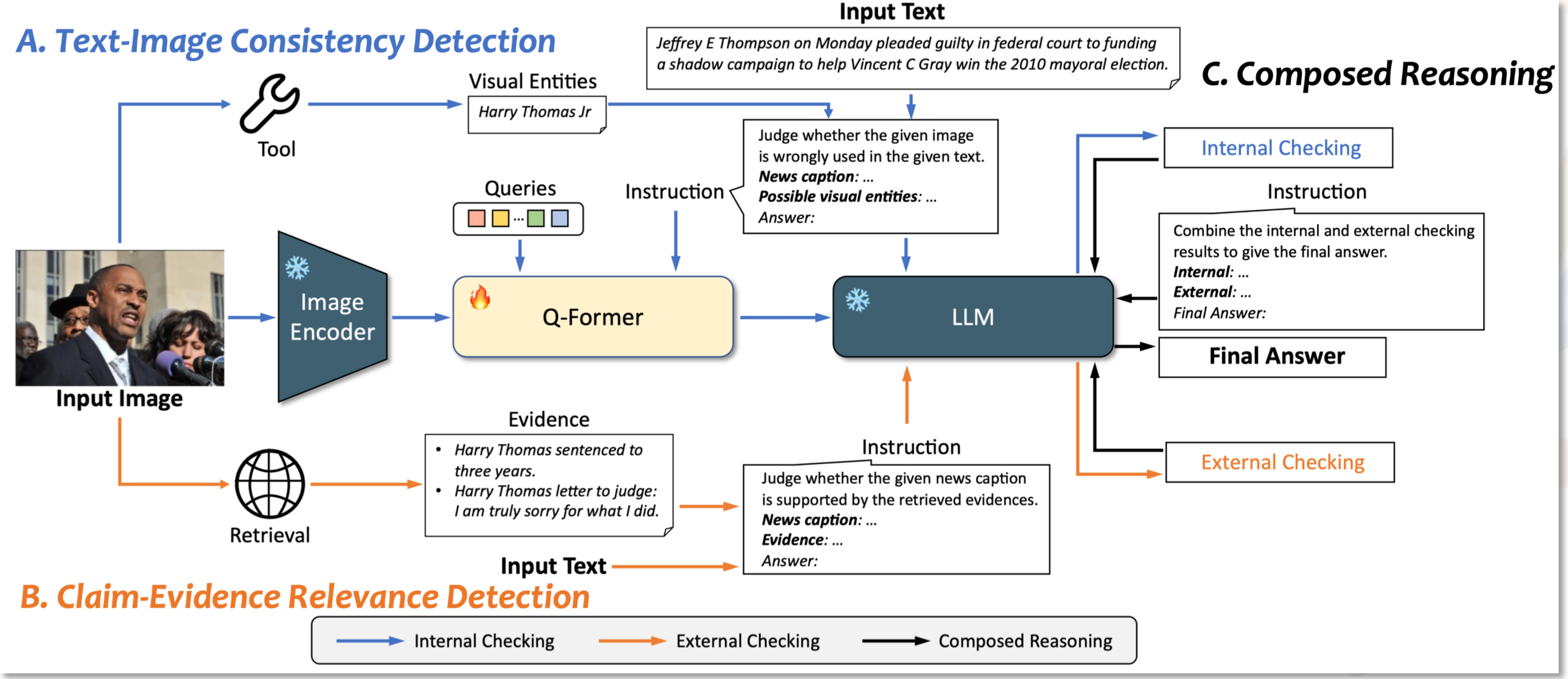
How to achieve it? – Three-part reasoning framework

- ***Challenge 2: The original news event may not be discernible from the image itself.***

🤔 **We need to capture the image's context!**

2016 Democratic National Convention

2015 George Mason University Speaking



A. Text-Image Consistency Detection

Tool

Visual Entities

Harry Thomas Jr

Queries

Instruction

Input Text

Jeffrey E Thompson on Monday pleaded guilty in federal court to funding a shadow campaign to help Vincent C Gray win the 2010 mayoral election.

Judge whether the given image is wrongly used in the given text.
***News caption**: …*
***Possible visual entities**: …*
Answer:

Input Image

Image Encoder

Q-Former

LLM

C. Composed Reasoning

Internal Checking

Instruction

Combine the internal and external checking results to give the final answer.
***Internal**: …*
***External**: …*
Final Answer:

Final Answer

External Checking

Evidence

- *Harry Thomas sentenced to three years.*
- *Harry Thomas letter to judge: I am truly sorry for what I did.*

Retrieval

Input Text

Instruction

Judge whether the given news caption is supported by the retrieved evidences.
***News caption**: …*
***Evidence**: …*
Answer:

B. Claim-Evidence Relevance Detection

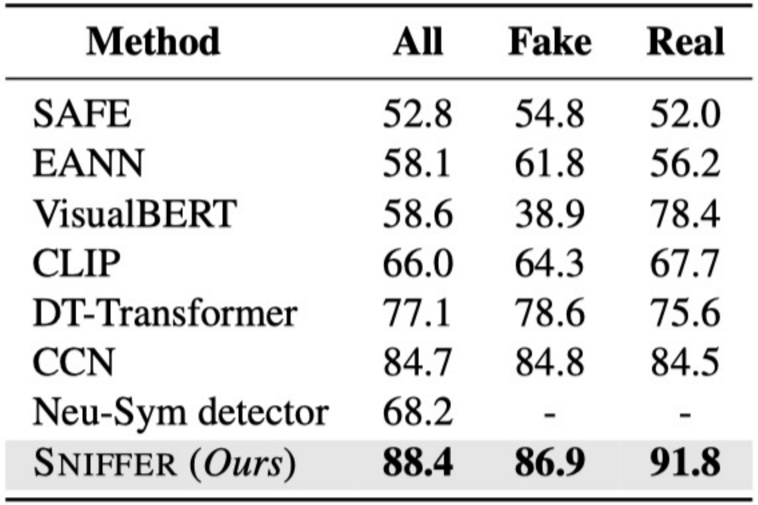
Internal Checking — External Checking — Composed Reasoning

Performance Study - Detection

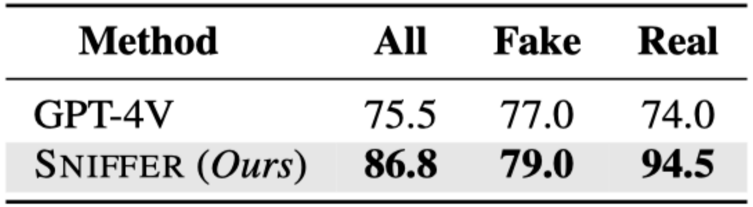
Experimental Setup:

- **Dataset:** NewsCLIPpings. Train 71 072, val 7 024, and test 7 264
- **GPU**: 4 Nvidia A100 (40G), 16 hours

Main Comparison

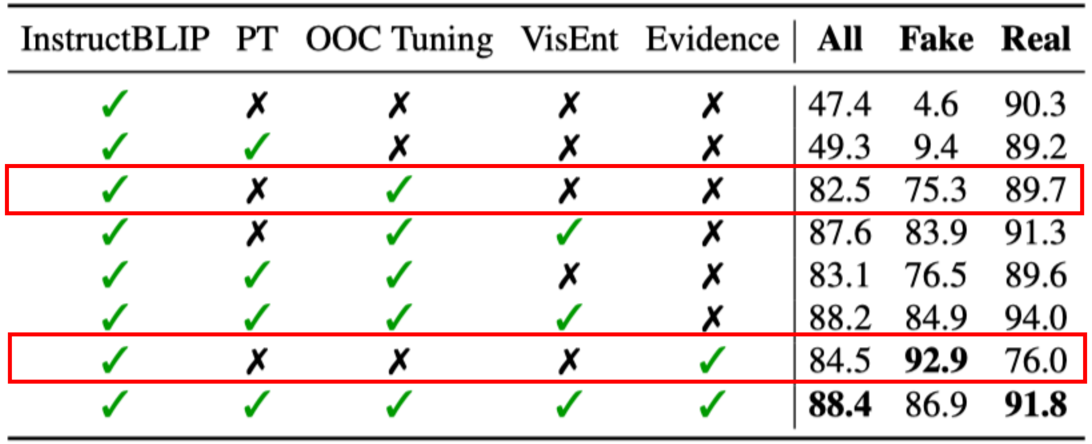


Method	All	Fake	Real
SAFE	52.8	54.8	52.0
EANN	58.1	61.8	56.2
VisualBERT	58.6	38.9	78.4
CLIP	66.0	64.3	67.7
DT-Transformer	77.1	78.6	75.6
CCN	84.7	84.8	84.5
Neu-Sym detector	68.2	-	-
SNIFFER (*Ours*)	**88.4**	**86.9**	**91.8**



Method	All	Fake	Real
GPT-4V	75.5	77.0	74.0
SNIFFER (*Ours*)	**86.8**	**79.0**	**94.5**

Ablation Study



InstructBLIP	PT	OOC Tuning	VisEnt	Evidence	All	Fake	Real
✓	✗	✗	✗	✗	47.4	4.6	90.3
✓	✓	✗	✗	✗	49.3	9.4	89.2
✓	✗	✓	✗	✗	82.5	75.3	89.7
✓	✗	✓	✓	✗	87.6	83.9	91.3
✓	✓	✓	✗	✗	83.1	76.5	89.6
✓	✓	✓	✓	✗	88.2	84.9	94.0
✓	✗	✗	✗	✓	84.5	**92.9**	76.0
✓	✓	✓	✓	✓	**88.4**	86.9	**91.8**

⭐ **Accurate OOC detection**

Performance Study - Explanation

Precise Explanation

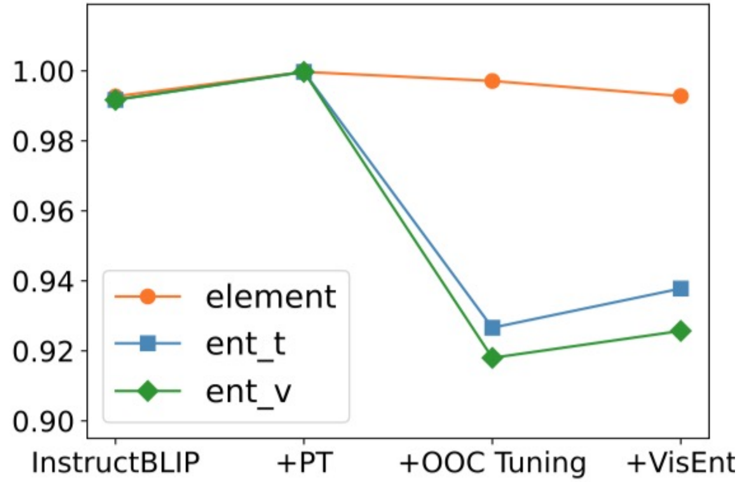


Figure 5. Response ratio.

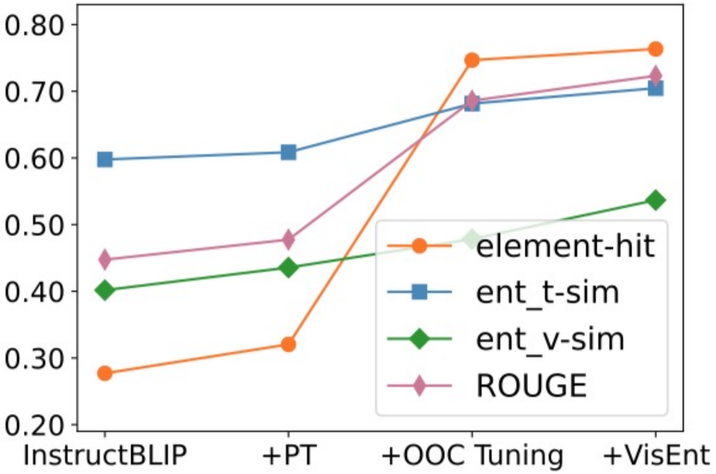
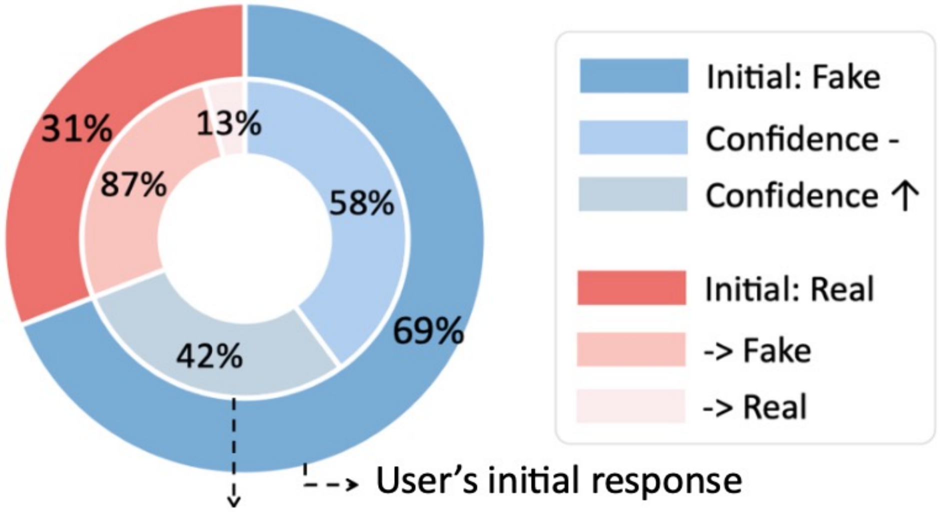


Figure 6. Explanation accuracy.

Persuasive Explanation

10 participants, 20 OOC (Fake) samples



User's initial response

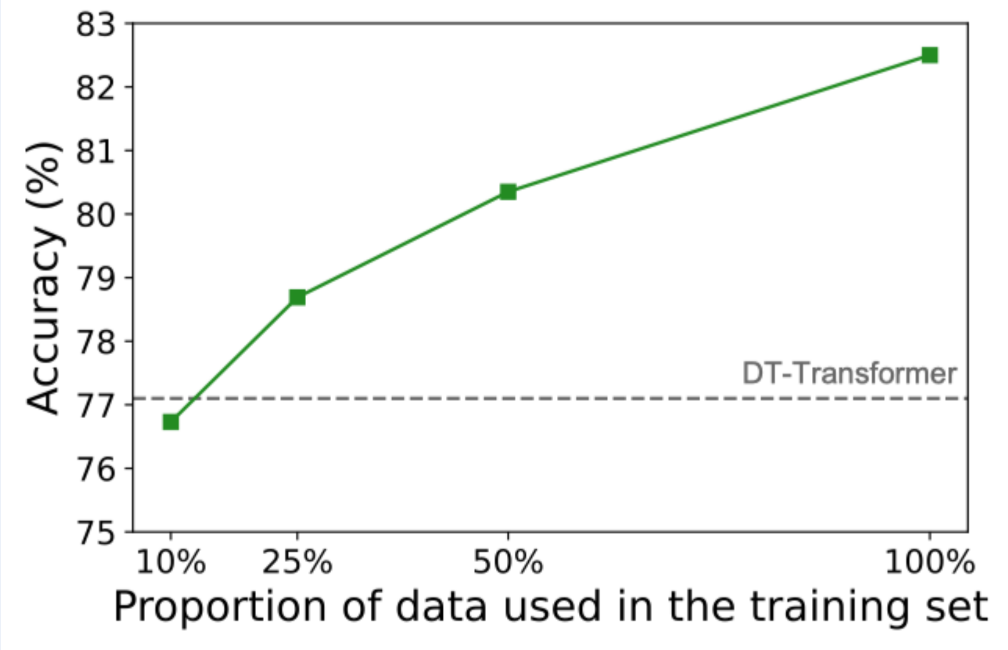
User's response after reading SNIFFER's explanation

Human Study

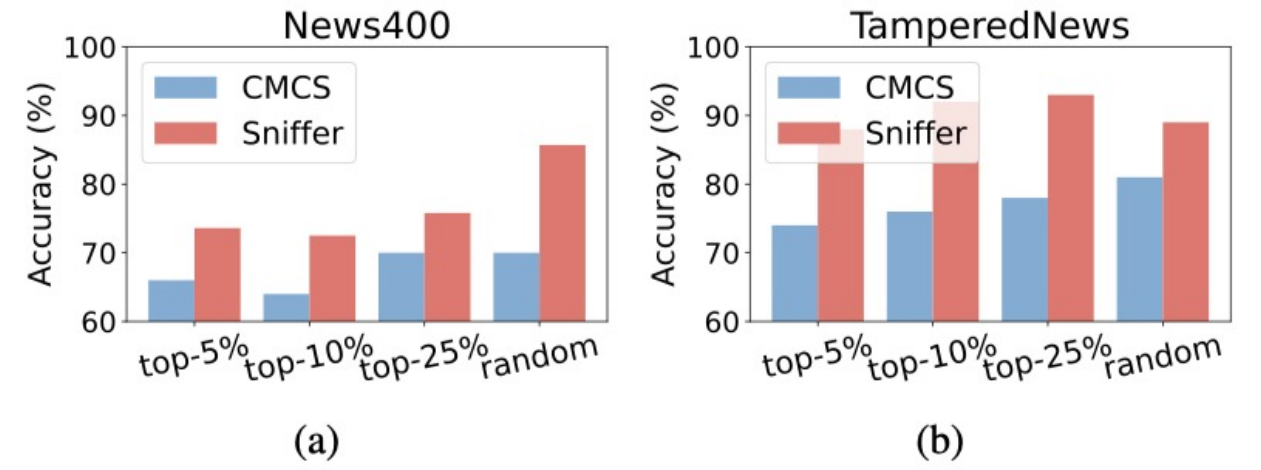
Record response -> read explanation -> record response again (truthfulness judgment & confidence level)

Performance Study – Practical Setting

Low-resource



Generalizability



(a)

(b)

Figure 9. Cross-dataset detection performance of SNIFFER.

iGyro

Following Works: Extend to more types of misinformation

Shared Ability

Textual analysis, Visual understanding, News knowledge, …

Misinformation

Textual Distortion (Pure fabrication)

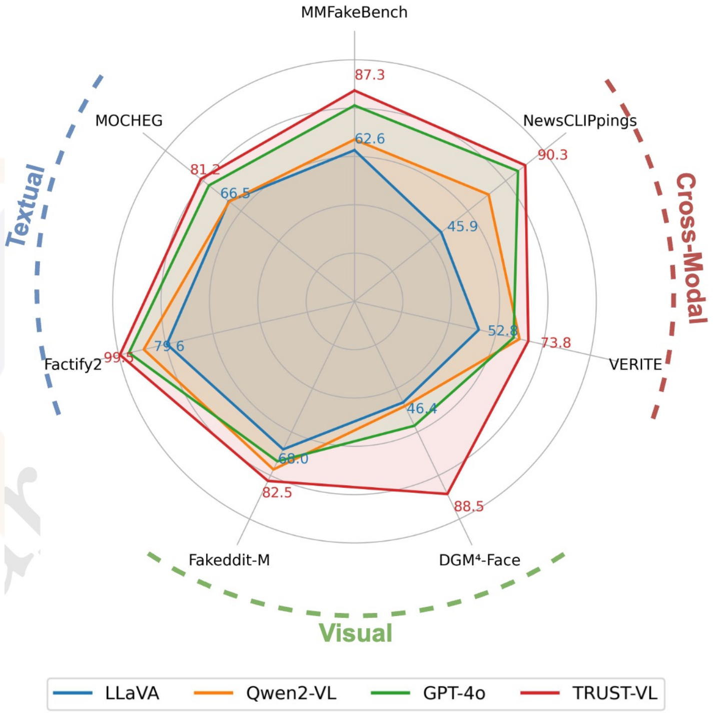
Nestled in the heart of the South Pacific, the island nation of Fiji stands as the only country in the world completely free of cancer—a medical miracle the world can't ignore.

Visual Distortion (AI-generated)

The church that survived the California wildfire.

Cross-modal Distortion (Image misuse)

Jeffrey E Thompson on Monday pleaded guilty …



Specialized Ability

Linguistic patterns
Evidence reasoning

Visual artifacts

Semantic inconsistency
Cross-modal evidence reasoning

TRUST-VL: An Explainable Vision-Language News Assistant for General Multimodal Misinformation Detection

Take-away Message

- Through specialized instruction tuning, open-sourced general-purpose multimodal large language models can achieve high performance in specific tasks.
- By transforming classification tasks into generation tasks, LLMs can offer interpretability for many classical classification tasks.
- Providing persuasive explanations is crucial for building public trust and more effectively debunking misinformation.

THANKS.

Our code and model are available at https://github.com/MischaQI/Sniffer.

Feel free to contact Peng Qi (pengqi.qp@gmail.com) for any questions!

iGyro