Not Search, But Scan:
Benchmarking MLLMs on Scan-Oriented Academic Paper Reasoning

Rongjin Li, Zichen Tang, Xianghe Wang, Xinyi Hu, Zhengyu Wang, Zhengyu Lu, Yiling Huang, Jiayuan Chen, Weisheng Tan, Jiacheng Liu, Zhongjun Yang, Haihong E*

*Corresponding author.

Reasoning Lab, Beijing University of Posts and Telecommunications

Motivation & Overview

What makes MLLMs struggle with *Deep Research*?

Methodology and evaluation follow a ***search-oriented*** paradigm:

- Ground with ***pre-specified targets***
- Reason with semantic ***relevance***

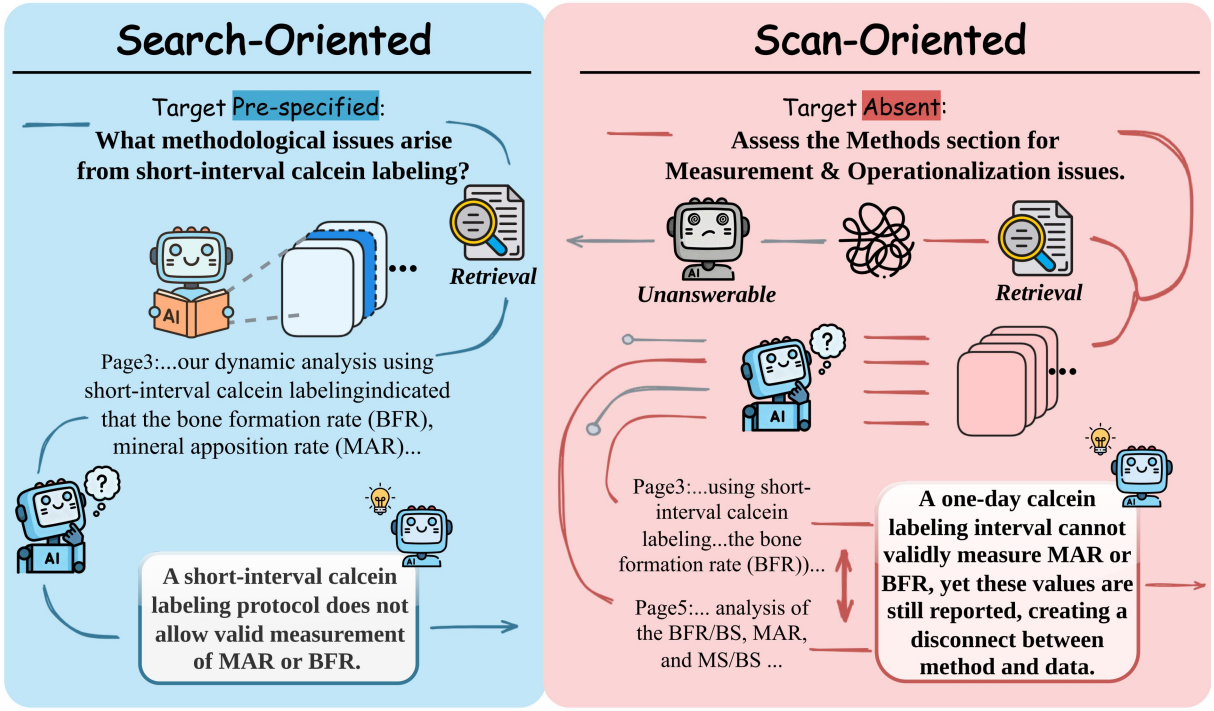
Our Contributions:

Scan-Oriented Task Paradigm

- Reason with ***absent targets***
- Reason with evidence ***consistency***

A Benchmark: *ScholScan*

- Evaluate on **scientific error detection**



Benchmark

Data Systhesis:

- Generation: Coordinate error edits via LLM
- Sampling: Extract errors from reviews

Metrics:

- Reasoning
- Evidence overlap
- Detection precision

Statistics:

- ***1,800*** questions
- ***715*** papers
- ***9*** error categories
- ***13*** disciplines

Research Question & Definitions	Design & Identifiability	Sampling & Generalizability
Explanation: *The definition of "actionable variants" shifts across sections (LOE 1–5 in Abstract, LOE 1–3 in Results), causing ambiguity.*	**Explanation:** *The design is described as probing both short- and long-range interactions, yet the paper still claims unique large-q selectivity, creating a disconnect.*	**Explanation:** *The experiments use a narrow diabetic mouse substrain, yet the paper generalizes findings to all patients, creating an invalid sample-to-population inference.*
Measurement & Operationalization	**Data Handling & Preprocessing**	**Computation & Formulae**
Explanation: *First-harmonic demodulation is dominated by far-field background and cannot produce the reported high-quality near-field images.*	**Explanation:** *Feature selection for NSCLC and HCC models was done on the full dataset before splitting, causing data leakage, while the Discussion falsely claims unbiased validation.*	**Explanation:** *The Methods claim a 200-fold concentration, but the 200 µL subsample is incorrectly said to represent ~20 mL instead of 40 mL, creating a twofold calculation error.*
Inference & Conclusions	**Referential & Citation Alignment**	**Language & Expression**
Explanation: *The data show PGK1 promotes EGFR degradation, yet the Discussion claims inhibiting PGK1 as therapy, directly contradicting the results.*	**Explanation:** *Figure 1 report an LPS dose of 1.5 mg/kg, but Figure 5 reports 15 mg/kg, creating a tenfold discrepancy that makes the actual experimental dose unclear.*	**Explanation:** *The paper swaps C. elegans gene and protein nomenclature (e.g., 'unc-45' vs. 'UNC-45'), creating technically misleading references.*

Results

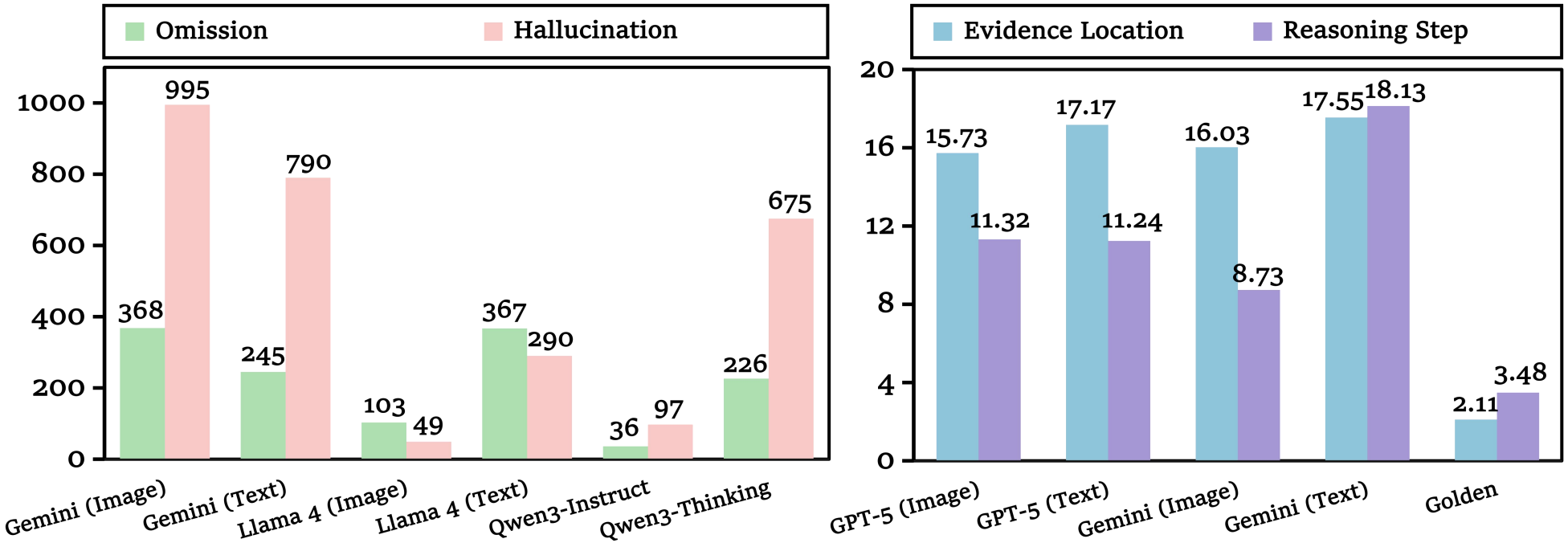
Main Results:

- Overall bottlenecks still exist.
- Reasoning models equip clear advantages.
- Text inputs consistently outperform image inputs, though the latter remain necessity.

Models	Avg.	RQD	DI	SG	MO	DHP	CF	IC	RCA	LE
MLLM (Image Input)										
Proprietary MLLMs										
Gemini 2.5 Pro	15.6	**11.9**	**12.6**	**35.7**	12.3	**27.0**	4.6	14.7	15.2	**7.4**
GPT-5	**19.2**	10.1	9.7	28.2	**14.6**	26.6	**13.8**	**25.3**	**25.3**	6.9
Grok 4	4.0	0.0	1.9	16.7	3.2	7.4	0.7	1.9	3.6	0.0
Doubao-Seed-1.6-thinking	10.2	3.4	3.5	22.3	7.5	15.1	10.2	12.2	10.9	3.3
Doubao-Seed-1.6	9.9	3.0	4.4	29.2	4.9	15.0	6.3	17.9	8.0	3.9
Open-source LLMs										
Llama 4 Maverick	7.0	7.0	7.3	9.4	4.5	4.0	6.5	6.7	8.8	3.0
Gemma 3 27B	1.7	0.5	2.7	2.3	1.7	1.0	1.0	1.3	2.6	0.0
Mistral Small 3.1	3.3	0.1	2.0	2.0	1.5	0.1	1.0	2.2	8.6	1.0
Qwen2.5 VL 72B	0.1	0.0	0.7	0.0	0.0	0.0	0.0	0.0	0.2	0.0

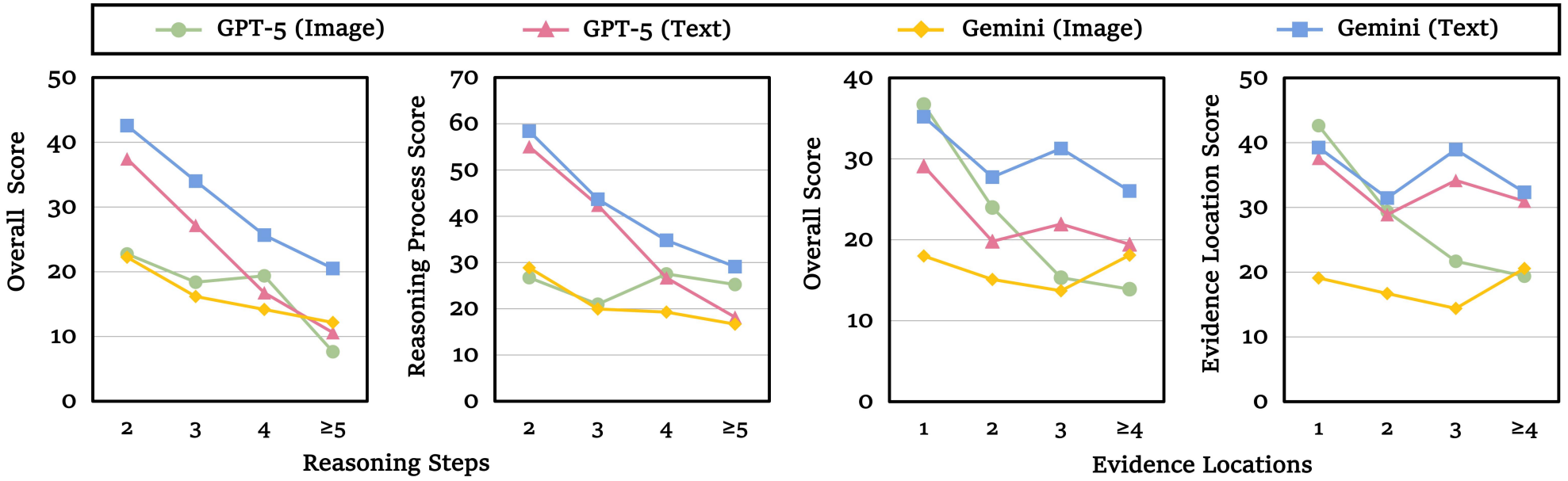
Results

- Strong models exhibit more overconfidence and hallucination.
- Scan-oriented tasks are ***more complex than they appear***.



Results

- Performance declines with longer reasoning chains.
- Heavier evidence loads make integration harder.



Results

RAG nearly collapses under target-absent settings:

- Most methods show ***minimal effect***.
- ***Complexity*** architect ***≠ Better*** performance.

Visual-focus RL frameworks lead the field.

Models	Avg	RQD	DI	SG	MO	DHP	CF	IC	RCA	LE
Text Input (Base Model: Qwen3 Thinking)										
Baseline	17.4	8.9	16.2	31.9	15.1	23.7	5.6	22.3	21.1	2.3
Oracle	24.5	20.6	27.9	43.6	21.3	40.8	7.4	26.9	26.0	1.9
bm25	16.7	9.7	13.7	33.0	17.3	23.8	6.8	25.4	16.5	3.0
BGE-M3	11.3	8.6	7.5	24.8	9.1	15.4	5.3	15.6	11.4	1.0
Contriever-msmacro	16.6	9.7	18.2	33.7	10.7	20.8	6.4	18.5	19.8	1.8
nv-embed-v2	6.8	4.0	4.0	9.4	6.1	4.9	5.5	5.7	10.0	2.0
Image Input (Base Model: Llama4 Maverick)										
Baseline	7.0	7.0	7.3	9.4	4.5	4.0	6.5	6.7	8.8	3.0
Oracle	6.5	3.0	4.5	15.6	8.2	9.4	4.9	10.0	4.4	1.4
ColPali-v1.3	0.8	1.5	0.0	0.5	0.0	0.9	0.5	1.3	1.4	0.0
ColQwen2.5	1.2	2.1	0.7	0.5	0.0	1.2	0.2	2.7	2.0	0.0
VisRAG	1.0	2.0	0.0	1.0	0.0	1.0	1.6	1.3	1.2	0.0
VRAG-RL	10.9	9.8	11.6	17.8	8.2	11.0	6.8	13.1	10.8	8.1

Table 3: Summary of retrieval performance for RAG methods.

Models	MRR@5	Recall@5
Text Input (Base Model: Qwen3 Thinking)		
bm25	0.41	0.48
BGE-M3	0.16	0.21
Contriever-msmacro	0.31	0.39
nv-embed-v2	0.30	0.38
Image Input (Base Model: Llama4 Maverick)		
ColPali-v1.3	0.26	0.31
ColQwen2.5	0.30	0.35
VisRAG	0.41	0.46

Contact us

WeChat: Staudinger0325

BUPT Reasoning Lab

BUPT Reasoning ★

北京

北京邮电大学推理实验室（Reasoning Lab）

Official Account

Rongjin Li

About Publications Education Experience Awards

Rongjin Li

I am currently a third-year undergraduate student at the School of Future of Beijing University of Posts and Telecommunications (BUPT), majoring in Computer Science and Technology. My research interests focus on:

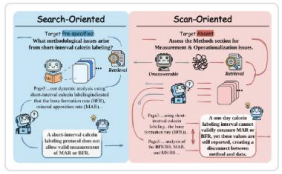
- **Multimodal Large Language Models (MLLMs)**
- **Reinforcement Learning (RL) for LLM Reasoning**

I have co-authored five papers published at leading international conferences, including ICLR and AAAI. Among them, I am the sole first author of one paper (ICLR 2026) and the undergraduate first author of another (AAAI 2026).

I am also currently working on the ASC26 Student Supercomputer Challenge. Feel free to reach out if you would like to discuss related topics or potential collaborations.

You can contact me at lirongjin@bupt.edu.cn or staudinger0325@gmail.com.

Publications



Not Search, But Scan: Benchmarking MLLMs on Scan-Oriented Academic Paper Reasoning
Rongjin Li, Zichen Tang, Xianghe Wang, Xinyi Hu, Zhengyu Wang, Zhengyu Lu, Yiling Huang, Jiayuan Chen, Weisheng Tan, Jiacheng Liu, Zhongjun Yang, Haihong E
ICLR 2026.
[Paper] [Project Page] [Github] [Dataset]

Homepage: https://staudinger0325.github.io/

Not Search, But Scan:
Benchmarking MLLMs on Scan-Oriented Academic Paper Reasoning

Rongjin Li, Zichen Tang, Xianghe Wang, Xinyi Hu, Zhengyu Wang, Zhengyu Lu, Yiling Huang, Jiayuan Chen, Weisheng Tan, Jiacheng Liu, Zhongjun Yang, Haihong E*

*Corresponding author.

Reasoning Lab, Beijing University of Posts and Telecommunications