NOT SEARCH, BUT SCAN: BENCHMARKING MLLMS ON SCAN-ORIENTED ACADEMIC PAPER REASONING

Rongjin Li[1], **Zichen Tang**[1], **Xianghe Wang**[1], **Xinyi Hu**[1], **Zhengyu Wang**[1], **Zhengyu Lu**[1], **Yiling Huang**[1], **Jiayuan Chen**[1], **Weisheng Tan**[1], **Jiacheng Liu**[1], **Zhongjun Yang**[1], **Haihong E**[1,*]
[1]Beijing University of Posts and Telecommunications
[*]Corresponding author.

github.com/Staudinger0325/ScholScan

huggingface.co/datasets/Staudinger/ScholScan

ABSTRACT

With the rapid progress of multimodal large language models (MLLMs), AI already performs well at literature retrieval and certain reasoning tasks, serving as a capable assistant to human researchers, yet it remains far from autonomous research. The fundamental reason is that current work on scholarly paper reasoning is largely confined to a search-oriented paradigm centered on pre-specified targets, with reasoning grounded in relevance retrieval, which struggles to support researcher-style full-document understanding, reasoning, and verification. To bridge this gap, we propose ScholScan, a new benchmark for scholarly paper reasoning. ScholScan introduces a scan-oriented task setting that asks models to read and cross-check entire papers like human researchers, scanning the document to identify consistency issues. The benchmark comprises 1,800 carefully annotated questions drawn from 9 error families across 13 natural-science domains and 715 papers, and provides detailed annotations for evidence localization and reasoning traces, together with a unified evaluation protocol. We assessed 15 models across 24 input configurations and conduct a fine-grained analysis of MLLM capabilities across error families. Across the board, retrieval-augmented generation (RAG) methods yield no significant improvements, revealing systematic deficiencies of current MLLMs on scan-oriented tasks and underscoring the challenge posed by ScholScan. We expect ScholScan to be the leading and representative work of the scan-oriented task paradigm.

1 INTRODUCTION

Enabling multimodal large language models (MLLMs) (OpenAI, 2025; Anthropic, 2025; ByteDance Seed Team, 2025; Meta, 2025; xAI, 2025) to conduct comprehensive understanding and generation based on academic literature is the ultimate goal of Deep Research, and a critical milestone on the path toward artificial general intelligence (AGI) (Ge et al., 2023; Morris et al., 2024; et al., 2025c). With rapid advances, MLLMs are increasingly capable of supporting academic workflows through retrieval, reading, and writing. For example, PaSa (He et al., 2025) can invoke a series of tools to answer complex academic queries with high-quality results, while Google Deep Research (et al., 2025b) is capable of producing human-level research reports based on specific queries.

However, most of the existing work still follows ***a search-oriented paradigm***, where models retrieve a few relevant passages and reason over local evidence based on prespecified targets (Gao et al., 2023; Lou et al., 2025). Such methods are effective for tasks with clearly predefined targets, but struggle with researcher-style full-document reasoning and verification (Zhou et al., 2024). ***To function as researchers, models must move beyond reactive question answering and toward proactive discovery of implicit problems.***

To fill this gap, as shown in Figure 1, we introduce ***a scan-oriented paradigm***, where models address queries with targets absent and are required to actively **construct a document-level evidence view,**

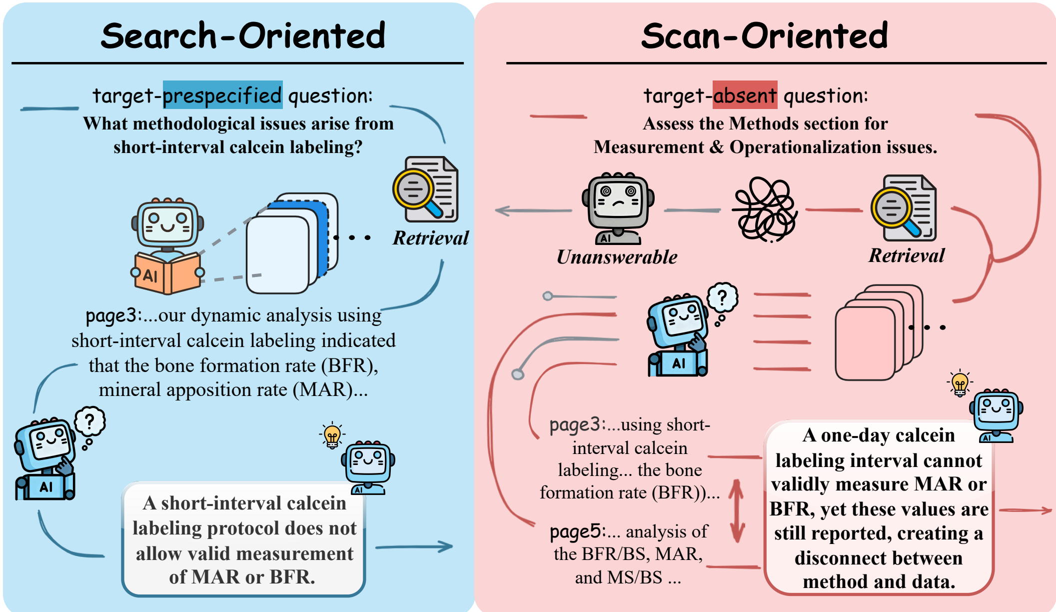


Figure 1: A comparison between search-oriented and scan-oriented task paradigms. Unlike the former, the scan-oriented paradigm provides no prespecified targets, requiring the model to actively scan the entire paper, construct a document-level evidence view.

perform exhaustive scanning over the full paper, and conduct evidence-based reasoning. In contrast to search-oriented tasks that assess a model's ability to identify and reason over *relevant* fragments, scan-oriented tasks emphasize *consistency*. ***Instead of relying on prespecified targets or hints, models must derive all necessary concepts and inferences solely from given documents.***

We instantiate this setting via scientific error detection, as it naturally demands discovering non-obvious flaws without target cues, and present ScholScan, a new multimodal benchmark for scholarly reasoning. ScholScan features the following key highlights:

- **Scan-Oriented Task Paradigm.** ScholScan receive one or more complete academic papers together with target-absent queries, presenting a rigorous challenge to their evidence-based reasoning capabilities. The benchmark comprises 715 papers spanning 13 natural science disciplines.
- **Comprehensive Error Types.** ScholScan covers 9 categories of scientific errors across the entire research workflow. It also includes citation and referencing errors, providing a rigorous test of a model's cross-source reasoning ability.
- **Process-Aware Evaluation Framework.** ScholScan provides fine-grained annotations for both evidence location and reasoning steps, enabling a comprehensive evaluation framework that assesses model performance in terms of both process and outcome.

We evaluate 15 models across 24 input configurations and 8 retrieval-augmented generation (RAG) frameworks. All models exhibit limited performance, and none of the RAG methods deliver significant improvements. These results highlight the inadequacy of search-oriented frameworks when applied to scan-oriented tasks, and underscore both the challenges and the potential of enabling MLLMs to perform reliable, document-level reasoning over full academic papers.

2 Related Work

2.1 Multimodal Large Language Models

With the rapid progress of MLLMs, models have evolved beyond perception tasks (e.g., image recognition and explanation) (Liu et al., 2024) toward deep understanding of structured, multimodal long documents. Their strengths lie in the ability to integrate cross-modal information and perform multi-hop reasoning over extended contexts. These capabilities are not only valuable for specific question answering or instruction-following tasks (Yue et al., 2024) but are particularly well suited for simulating human thought processes and generating explainable reasoning trajectories (Zheng et al., 2023). Consequently, achieving comprehensive understanding of entire documents has emerged as a core challenge that MLLMs are inherently equipped to address.

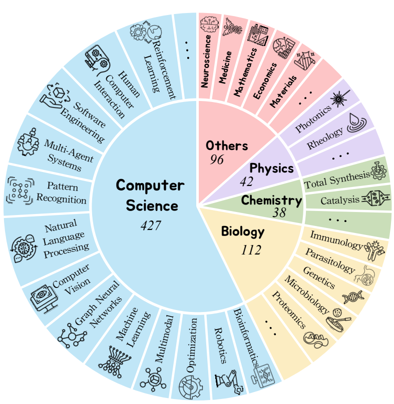
2.2 Document Understanding Benchmark

Document understanding tasks challenge models to identify relevant context and perform accurate reasoning grounded in that information. Progress in document understanding benchmarks has followed two main axes. Along the input dimension, it has evolved from short to long contents, from everyday to specialized domains, and from plain text to multimodal format (Chen et al., 2021; Yang et al., 2018; Tito et al., 2021; Deng et al., 2025). Along the scenario dimension, it has shifted from limited-output formats to more open-ended responses (Pramanick et al., 2024). DocMath-Eval (Zhao et al., 2024) evaluates numerical reasoning on long, specialized documents, revealing large performance gaps even for strong models in expert domains, while MMLongBench-Doc (Ma et al., 2024) builds a multimodal benchmark with layout-rich documents. However, a comprehensive benchmark that integrates all challenges above has yet to be introduced.

2.3 Academic Paper Understanding Benchmark

Compared with general documents, academic papers are distinguished by their rich domain knowledge and logical rigor. Reasoning over papers has emerged as a major challenge in recent research. Some studies ask for local elements like charts or snippets, leveraging their internal complexity, but neglect the need for cross-source integration and domain-specific interpretation within the full document (Wang et al., 2024; Li et al., 2024). Recent studies extend inputs to the document level and adopt image-based formats to better simulate real-world reading scenarios. (Auer et al., 2023; Yan et al., 2025) However, benchmarks based on the QA paradigm face inherent limitations, as they typically presuppose answer existence and embed explicit cues in the question itself, reducing the need for comprehensive understanding and information organization. Moreover, mainstream evaluation protocols focus on the final outcome, with limited assessment of whether intermediate reasoning is evidentially grounded and logically valid. More examples and analysis are shown in Appendix B.

3 The ScholEval Benchmark



Benchmark	**Mod.**	**Para.**	**Eval.**	**# Dom.**
Document Understanding				
DocMath-Eval$_{\text{CompLong}}$	T+TD	Search	A	N/A
MMLongbench-Doc	T+MD	Search	A	N/A
LongDocURL	T+MD	Search	A	N/A
SlideVQA	T+MD	Search	A	N/A
Academic Paper Understanding				
CharXiv	I	Search	A	8
ArXivQA	I	Search	A	10
MMCR	T+MD	Search	A	CS
AAAR-1.0	T+MD	Search	A	CS
ScholScan (ours)	T+MD	Scan	A+P	13

Figure 2: Left: Overview of ScholScan. Right: Comparison to related benchmarks. **Mod.**: Modalities; **Para.**: Task Paradigm; **Eval.**: Evaluation; **T**: Text; **I**: Image; **TD**: Text-Form Document; **MD**: Multimodal Document; **A**: Answer; **P**: Process; **Dom**: Number of academic domains in the dataset.

3.1 Overview of ScholScan

We introduce ScholScan, a benchmark designed to comprehensively evaluate MLLMs' ability to detect scientific flaws in academic papers under scan-oriented task settings. As illustrated in Figure 2, ScholScan spans 13 disciplines across the natural sciences, including physics, chemistry, and

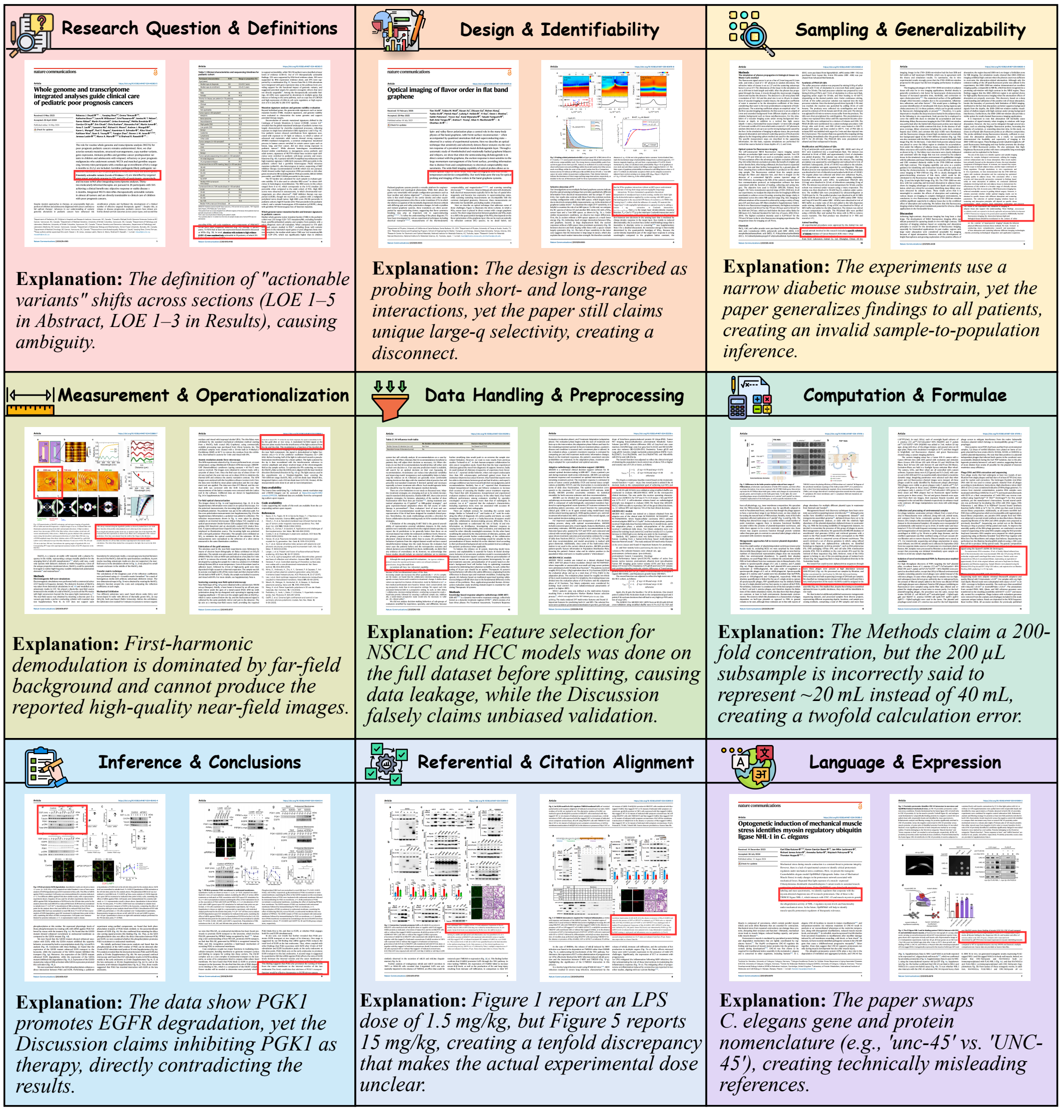


Figure 3: Sampled ScholScan examples with 9 error types, covering the whole process of scientific research, each requiring the model to perform thorough cross-source evidence-based reasoning.

computer science, and spans over 100 subfields such as immunology, total synthesis, and machine learning. The benchmark comprises 1,800 questions derived from 715 real academic papers, and covers 9 major error categories (Figure 3) that commonly observed in real-world research scenarios. These include issues in numerical and formulaic computation, experimental design, inference and conclusion, and citation misuse, among others. Figure 2 also provides a comparison ScholScan with existing benchmarks for multimodal paper understanding and long-document reasoning.

3.2 Data Collection & Question Generation

We curated papers from ICLR 2024/2025 and Nature Communications, and collected public reviews for the former. Questions were constructed based on two dimensions, where the source is either generated or sampled, and the context is either within-paper or cross-paper.

Generation. On high-quality accepted papers, we prompt Gemini 2.5 Pro to perform coordinated sentence-level edits spanning multiple sections or pages. It then synthesizes composite errors and generates the corresponding question along with an explanation grounded in the edited context.

Sampling. From rejected ICLR submissions and their public reviews, we prompt Gemini 2.5 Pro to extract explicit, falsifiable scientific errors and convert them into questions with initial explanations. Subjective remarks about novelty or writing quality are excluded.

Within-paper. This setting focuses on verifiable facts and internal consistency within a single paper, and supports both Generation and Sampling.

Cross-paper. This setting examines citation consistency across papers. For each instance, Gemini 2.5 Pro receives an accepted paper and one of its cited sources, then edits the accepted paper to introduce paraphrases or reasoning errors about the citation. As public reviews mainly address nonfalsifiable aspects such as appropriateness, all cross-paper instances are constructed exclusively using the generation method.

3.3 Quality Control & Annotation

Despite explicit instructions, initial outputs exhibited substantial hallucinations, logical inconsistencies, and low-quality questions. To ensure the quality, 10 domain experts conducted a rigorous annotation process. Each instance underwent independent dual review, and disagreements were resolved by a third expert. Among the 3,500 initially generated candidates, 1,700 were discarded, and 1,541 of the remaining were revised, including 535 question rewrites, 1,207 explanation edits, and 1,141 corrections to error categories or metadata. Further details are provided in Appendix C.

4 Experiments

4.1 Experiments Setting

Models. We benchmark a total of 24 input configurations by feeding academic papers as either images or OCR text using the Tesseract (Smith, 2007) engine, covering 15 mainstream models (Yang et al., 2025; Bai et al., 2025; et al., 2025a; Guo et al., 2025; et al., 2025d).

Evaluation Protocol. Inspired by MMLongBench-Doc (Ma et al., 2024), we prompt models to generate necessary reasoning chains from evidence to detected anomalies without constraining the output format, which aims to assess the ability for evidence-grounded reasoning rather than mere instruction-following. For open-ended responses, we use GPT-4.1 (OpenAI, 2025) to extract cited evidence and reasoning steps, and quantify alignment with annotated explanations. Human evaluation confirms high agreement between our pipeline and expert annotations. Further implementation details are provided in Appendix E.

Metrics. We define a structured evaluation framework by parsing the model response a into a tuple:

$$\Psi(a) \;\Rightarrow\; \big(\mathbf{1}_{\text{exist}},\, \mathbf{1}_{\text{contain}},\, \widehat{\mathcal{E}},\, \widehat{\mathcal{R}},\, n\big). \tag{1}$$

Here, $\mathbf{1}_{\text{exist}}$ and $\mathbf{1}_{\text{contain}}$ are binary indicators for whether output contains any error and includes the annotated target error; $\widehat{\mathcal{E}}, \widehat{\mathcal{R}}$ and $\mathcal{E}^*, \mathcal{R}^*$ are the predicted and gold evidence sets and reasoning chains; $\hat{g} = \text{prefix\_match}(\widehat{\mathcal{R}}, \mathcal{R}^*)$ counts matched reasoning steps; $n \in \mathbb{N}$ is the number of unrelated errors. HasError(a) is 1 if the output contains any predicted error, and 0 otherwise. Based on $\Psi(a)$, we define an end-to-end score $S(m) \in [0, 1]$ that combines all aspects of prediction quality:

(i) Existence. $S_{\text{exist}}(a) = 1$ if and only if the response includes the annotated target error.

$$S_{\text{exist}}(a) = \mathbf{1}\{\text{HasError}(a)\} \cdot \mathbf{1}\{\, \hat{\mathcal{E}} \cap \mathcal{E}^* \neq \emptyset \,\} \tag{2}$$

(ii) Evidence location score. Even when the target error is identified, the cited evidence may be incomplete or noisy. We compute a Dice score with a squared penalty for over-reporting:

$$S_{\text{location}} = \max\left\{0,\; \frac{2\,|\widehat{\mathcal{E}} \cap \mathcal{E}^*| + \mathbf{1}\Big\{\, |\widehat{\mathcal{E}}| + |\mathcal{E}^*| = 0 \,\Big\}}{\max\Big(\, |\widehat{\mathcal{E}}| + |\mathcal{E}^*|,\, 1\Big)} - 0.8\left(\frac{|\widehat{\mathcal{E}} \setminus \mathcal{E}^*|}{\max\Big(|\widehat{\mathcal{E}}|,\, 1\Big)}\right)^2\right\}. \tag{3}$$

(iii) Reasoning process score. Even if the target error is detected, the reasoning may diverge from the gold chain. We use prefix match to assess reasoning completeness:

$$S_{\text{reasoning}} = \mathbf{1}\{\, g_r = 0 \,\} + \mathbf{1}\{\, g_r > 0 \,\}\left(\frac{\hat{g}}{g_r}\right)^2. \tag{4}$$

(iv) Unrelated-error penalty. Models may list unrelated items to inflate recall at the cost of precision. We penalize this with a rapidly increasing function of unrelated error count:

$$P_{\text{unrelated\_err}}(n) = 0.9^{\,\min(n,2)}\ \exp\Big(-0.6\,\big[\max(n-2,0)\big]^{1.5}\Big). \tag{5}$$

(v) Overall outcome score. The final score for a is defined as:

$$S(m) = S_{\text{exist}}(a)\ \sqrt{S_{\text{location}} \cdot S_{\text{reasoning}}}\ \cdot P_{\text{unrelated\_err}}(n). \tag{6}$$

Table 1: Model performance (scaled by 100) across input configurations. **RQD**: Research Question & Definitions; **DI**: Design & Identifiability; **SG**: Sampling & Generalizability; **MO**: Measurement & Operationalization; **DHP**: Data Handling & Preprocessing; **CF**: Computation & Formulae; **IC**: Inference & Conclusions; **RCA**: Referential and Citation Alignment; **LE**: Language & Expression.

Models	**Avg.**	**RQD**	**DI**	**SG**	**MO**	**DHP**	**CF**	**IC**	**RCA**	**LE**
MLLM (Image Input)										
Proprietary MLLMs										
Gemini 2.5 Pro	<u>15.6</u>	**11.9**	**12.6**	**35.7**	<u>12.3</u>	**27.0**	4.6	14.7	<u>15.2</u>	**7.4**
GPT-5	**19.2**	<u>10.1</u>	<u>9.7</u>	28.2	**14.6**	<u>26.6</u>	**13.8**	**25.3**	**25.3**	<u>6.9</u>
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	<u>10.2</u>	12.2	10.9	3.3
Doubao-Seed-1.6	9.9	3.0	4.4	<u>29.2</u>	4.9	15.0	6.3	<u>17.9</u>	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
OCR + LLM (Text Input)										
Proprietary LLMs										
Gemini 2.5 Pro	**30.3**	**21.5**	**34.2**	**44.3**	**27.6**	**56.6**	**10.3**	<u>28.8</u>	**35.6**	**8.1**
GPT-5	<u>22.5</u>	<u>16.1</u>	<u>21.4</u>	26.0	<u>20.3</u>	<u>36.7</u>	4.7	**29.8**	30.0	2.6
Claude Sonnet 4	5.7	3.7	2.5	10.8	4.3	10.3	1.4	8.4	6.6	3.5
Grok 4	20.8	9.3	7.7	<u>37.4</u>	12.3	34.4	<u>9.0</u>	20.0	<u>31.2</u>	<u>7.2</u>
Doubao-Seed-1.6-thinking	15.3	8.2	10.1	24.3	10.1	24.2	6.4	19.2	21.0	4.2
Doubao-Seed-1.6	13.9	5.4	6.9	26.4	10.3	23.6	6.3	20.1	17.5	2.3
Open-source LLMs										
Qwen3 A22B (Thinking)	17.4	8.9	16.2	31.9	15.1	23.7	5.6	22.3	21.1	2.3
Qwen3 A22B	1.7	1.2	0.0	2.7	0.4	1.0	0.1	4.3	2.5	1.1
gpt-oss-120b	7.3	6.3	5.7	18.3	4.9	14.5	1.6	12.5	5.5	0.0
DeepSeek-R1	11.4	5.1	11.9	25.4	8.7	22.5	4.7	16.3	9.8	3.5
DeepSeek-V3.1	1.7	1.2	2.0	1.7	1.0	5.8	0.5	2.2	2.1	0.0
Llama 4 Maverick	2.3	1.5	2.0	4.8	3.0	3.6	0.0	5.8	1.6	0.2
Gemma 3 27B	2.0	2.1	1.6	3.0	2.7	0.2	0.7	7.7	1.0	0.0
Mistral Small 3.1	6.9	3.0	2.7	5.5	7.0	2.0	8.5	4.0	12.2	3.0
Qwen2.5 VL 72B	0.2	0.0	0.7	0.0	0.0	0.0	0.0	0.0	0.6	0.0

4.2 Main Result

Table 1 presents our evaluation results. Our main findings are summarized as follows:

Overall performance remains unsatisfactory. GPT-5 achieves the highest average score in the image input group (19.2), while Gemini 2.5 Pro, the best-performing model in the text input setting, still fails to surpass the 60-point threshold on any subtask. Even in the SG category, which yields the best performance overall, nearly half of the models receive single-digit scores. Most models perform poorly under the scan-oriented task formulation and fail to detect any issues in many papers. This challenge is particularly pronounced for open-source models.

Reasoning-enhanced models demonstrate clear advantages. Across both input configurations, reasoning-enhanced variants consistently achieve higher scores. Almost all top-performing models, measured by both subtask-specific and overall metrics, fall into this category. Notably, Qwen3-Thinking and Deepseek-R1 outperform their base versions by more than 10% in average scores, with substantial gains observed across all error types. These results indicate that reasoning-enhanced models are better able to simulate the iterative process of extraction followed by reasoning, which is essential for effectively handling scan-oriented tasks and producing higher-quality responses.

MLLMs face significant bottlenecks in handling long multimodal inputs. Across most evaluation metrics, text inputs outperform image inputs. Among the nine MLLMs tested, the average performance gap between text and image inputs reaches 4.81 points, highlighting visual processing as a key limitation in current MLLM capabilities.

In most evaluation metrics, text inputs consistently outperform image inputs. Among the nine MLLMs evaluated, the average performance gap between text and image inputs is 4.81 points, underscoring visual processing as a key limitation in current MLLM capabilities.

Although overall performance is generally weaker, multimodal input remains indispensable. In certain categories such as CF, where OCR-based text extraction leads to substantial loss of formulaic or tabular content, image inputs outperform their text counterparts. This highlights the essential role of multimodal reasoning and the irreplaceable value of visual information in addressing specific types of errors.

4.3 Fine-grained Analysis

Capability Dimensions. We compute pairwise Spearman correlations between error types across two input configurations (text and image) for the eight evaluated MLLMs excluding Qwen2.5-VL-72B, as shown in Figure 4. We derive the following insights:

(i) With image input, CF exhibits consistently low correlations with other error categories, suggesting that the skills required for mathematical reasoning are relatively distinct. In contrast, with text input, CF shows moderate correlation with LE, indicating that OCR-flattened formulas lose their structural specificity and are interpreted by models in a manner more akin to natural language. Combined with the overall poor performance on CF tasks, this underscores the unique challenges of this category and the need for targeted improvements.

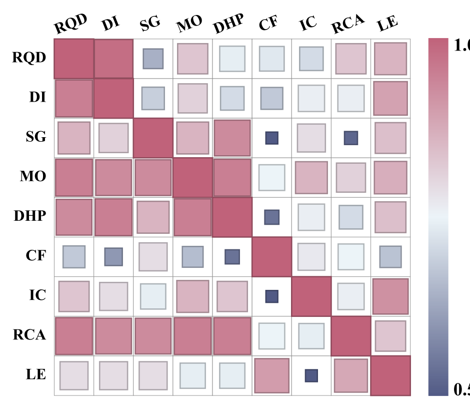


Figure 4: Spearman correlation matrix among the 9 error types.

(ii) Although DI is also related to experimental settings, it does not exhibit strong correlations with SG, MO, or DHP. This indicates that DI primarily emphasizes causal framing and variable identifiability, rather than the procedural understanding of experimental operations.

(iii) OCR severely degrades structured content such as figures and formulas, making questions that depend on multimodal information unanswerable. This diminishes the expression of multimodal reasoning capabilities and artificially inflates inter-category correlations under text input.

Based on the above analysis, we consolidate the original 9 error categories, each defined by its objective target, into 5 core latent skill dimensions evaluated by ScholScan under the image input setting. While each dimension highlights the primary competence emphasized by its corresponding error types, they are not mutually exclusive, as many questions involve overlapping reasoning abilities.

RQD and DI correspond to research concept comprehension, which requires models to ***identify the scope and definition*** of research objectives by integrating contextual cues and prior knowledge. SG, MO, and DHP fall under ***experimental process modeling***, which tests a model's ability to reconstruct procedural workflows such as sampling, measurement, and data handling. CF captures ***formal reasoning and symbolic computation***, focusing on syntactic parsing and numerical logic. IC evaluates causal inference, where models must ***synthesize dispersed causal evidence*** to reach sound conclusions. RCA and LE reflect referential alignment and linguistic consistency, which assess the ability to ***verify citations and maintain coherent expression*** throughout the document.

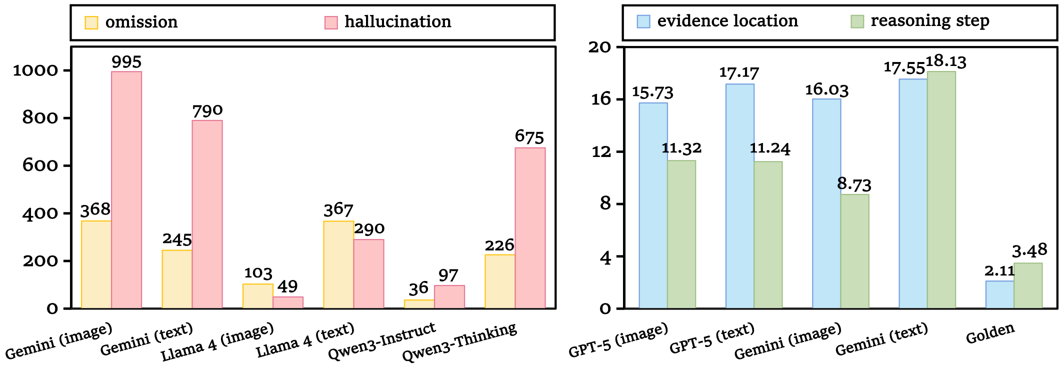


Figure 5: Left: Distribution of omission and hallucination errors. Right: Average reasoning steps and evidence locations involved in the answer generation, compared against the golden reference.

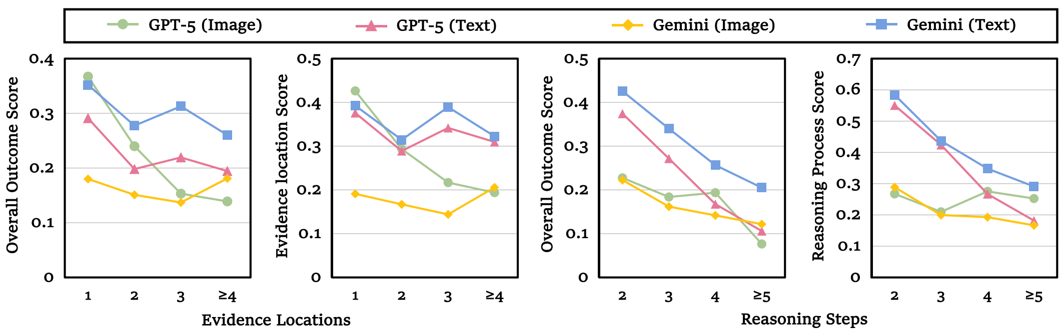


Figure 6: Performance trends across varying reasoning depths and evidence counts.

Hidden Complexity in Scan-Oriented Tasks. We analyze the reasoning traces of GPT-5 and Gemini 2.5 Pro under both input configurations, focusing on the number of evidence pieces scanned and the reasoning steps performed. As illustrated in Figure 5, even the most advanced models often scan up to 8 times more evidence and execute 3.5 times more reasoning steps than the reference answers, merely to approximate a correct response, yet they still frequently fail. This highlights the substantial hidden complexity inherent in scan-oriented tasks, which significantly amplifies the challenge of successful task completion.

4.4 Error Analysis

Omission and Hallucination. Most zero-score cases fall into two categories: either the model fails to detect any errors in the paper, or it becomes overwhelmed by hallucinations and entirely overlooks the actual errors present in the reference answer. We analyze the number of zero-score questions and the proportion of these two failure modes across models, as shown in Figure 5. Stronger models tend to have fewer zero-score cases overall, but are more prone to overconfident hallucinations.

Fragile Reasoning under Complex Evidence. Figure 6 shows how top-performing models behave under different numbers of reasoning steps and evidence locations. As reasoning steps increase, both reasoning and overall scores steadily decline, revealing a clear bottleneck in MLLMs' ability to construct long causal chains. In contrast, variation in evidence count has a weaker and less consistent impact. However, this does not imply that multi-evidence questions pose only marginal difficulty. Since the evaluation metric allows partial evidence omissions, more evidence items do not necessarily incur large score penalties. Still, heavier evidence loads often require longer reasoning chains, which substantially affect the coherence and completeness of inferred logic. These results highlight the persistent challenge for MLLMs in integrating evidence and maintaining logical structure as task complexity grows.

Table 2: Scores of RAG methods across the 9 error types (scaled by 100).

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

4.5 RAG Analysis

We evaluated 8 RAG methods under both input configurations (Robertson et al., 1994; Chen et al., 2024; Lee et al., 2025; Faysse et al., 2025; Yu et al., 2025; Wang et al., 2025; Izacard et al., 2022). Key findings are presented below, with detailed results shown in Tables 2 and 3.

Oracle Condition Yields Significant Accuracy Gains. Providing gold-standard images alleviates the scanning burden in long-context inputs, increasing the chances of generating correct answers. While overall performance improves, gains are limited for CF errors and minimal for LE errors. For CF, sparse formulaic content means gold images offer slight help. For LE, dense text distribution makes even direct access to target regions insufficient to reduce complexity for current models.

In consistency-centric scan-oriented tasks, most retrieval-based enhancement methods show minimal effectiveness. All embedding models exhibit poor retrieval accuracy. None achieves recall of 50% within the top-5 retrieved items. More critically, performance deteriorates after retrieval, especially for multimodal embedding models, where post-retrieval responses are almost entirely incorrect and scores approach 0.

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

Complex embedding model architectures do not yield better performance. Providing gold-standard images alleviates the scanning burden in long-context inputs, increasing the chances of retrieving correct answers. While overall performance improves, gains are limited for CF and minimal for LE errors. For CF, sparse formulaic content means gold images offer only slight localization help. For LE, dense error distribution makes even direct access to target regions insufficient to reduce task complexity for current models.

Reinforcement learning frameworks with a visual-centric focus have distinguished themselves as leading approaches. Despite being built on a compact 7B model, VRAG-RL consistently delivers improved performance and is the only method that achieves gains in the image-input setting following RL optimization. Its enhanced retrieval sharpens evidence selection, while strong reasoning provides effective guidance during document scanning. The retrieval and reasoning components are interleaved in design, with each stage informing the other in an iterative loop. This tightly coupled interaction contributes to the method's superior performance potential.

5 Conclusion

In this paper, we introduce ScholScan, a benchmark designed to evaluate the performance of MLLMs on scan-oriented tasks that require detecting scientific errors across entire academic papers. We conduct a comprehensive evaluation and in-depth analysis of mainstream MLLMs and RAG methods. The results demonstrate that current MLLMs remain far from capable of reliably addressing such tasks, and that existing RAG approaches provide little to no improvement. This highlights the complexity, integrative demands, and originality of the ScholScan benchmark. Looking ahead, we aim to develop scan-oriented task paradigms suited to diverse academic scenarios and explore new techniques for enhancing model performance on target-suppressed inputs. These directions support the broader goal of advancing MLLMs from passive assistants to active participants in scientific research.

6 Ethics Statement

All data used in this paper were constructed by the authors and do not include any external public or proprietary datasets. The included academic papers and author names are publicly available through arXiv and OpenReview and can be freely accessed.

A team of 10 domain experts was assembled to comprehensively review all task instances initially generated by Gemini 2.5 Pro. All annotators gave informed consent to participate. To ensure the accuracy and neutrality of both model-generated and human-verified content, we employed a rigorous multi-stage validation process involving cross-review and third-party adjudication.

Evaluation across 15 mainstream models and 24 input configurations was conducted via legally authorized API access through the VolcEngine, Alibaba Cloud's LLM services, and OpenRouter.

ScholScan is fully open-sourced and freely available for academic and non-commercial research purposes. We provide the complete download link and documentation through an anonymous GitHub repository. All personally identifiable information has been removed from the dataset, and its collection and release comply with the ethical and legal requirements in place at the time of data acquisition.

References

Anthropic. System card: Claude opus 4 & claude sonnet 4. `https://www.anthropic.com/claude-4-system-card`, May 2025. Updated Sep 2, 2025.

S. Auer, Dante Augusto Couto Barone, Cassiano Bartz, E. Cortes, Mohamad Yaser Jaradeh, Oliver Karras, Manolis Koubarakis, Dmitry I. Mouromtsev, Dmitrii Pliukhin, Daniil Radyush, Ivan Shilin, Markus Stocker, and Eleni Tsalapati. The sciqa scientific question answering benchmark for scholarly knowledge. *Scientific Reports*, 13, 2023. URL `https://api.semanticscholar.org/CorpusID:258507546`.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL `https://arxiv.org/abs/2502.13923`.

ByteDance Seed Team. Introduction to techniques used in seed1.6. `https://seed.bytedance.com/en/blog/introduction-to-techniques-used-in-seed1-6`, June 2025. Official blog post describing Seed1.6 techniques.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024. URL `https://arxiv.org/abs/2402.03216`.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3697–3711, Online and Punta Cana, Dominican

Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.300. URL https://aclanthology.org/2021.emnlp-main.300/.

Chao Deng, Jiale Yuan, Pi Bu, Peijie Wang, Zhong-Zhi Li, Jian Xu, Xiao-Hui Li, Yuan Gao, Jun Song, Bo Zheng, and Cheng-Lin Liu. LongDocURL: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1135–1159, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.57. URL https://aclanthology.org/2025.acl-long.57/.

DeepSeek-AI et al. Deepseek-v3 technical report, 2025a. URL https://arxiv.org/abs/2412.19437.

Gheorghe Comanici et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025b. URL https://arxiv.org/abs/2507.06261.

Long Phan et al. Humanity's last exam, 2025c. URL https://arxiv.org/abs/2501.14249.

OpenAI et al. gpt-oss-120b & gpt-oss-20b model card, 2025d. URL https://arxiv.org/abs/2508.10925.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, CELINE HUDELOT, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ogjBpZ8uSi.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *ArXiv*, abs/2312.10997, 2023. URL https://api.semanticscholar.org/CorpusID:266359151.

Yingqiang Ge, Wenyue Hua, Kai Mei, jianchao ji, Juntao Tan, Shuyuan Xu, Zelong Li, and Yongfeng Zhang. Openagi: When llm meets domain experts. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 5539–5568. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1190733f217404edc8a7f4e15a57f301-Paper-Datasets_and_Benchmarks.pdf.

Daming Guo, Dongdong Yang, Hongyi Zhang, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Yichen He, Guanhua Huang, Peiyuan Feng, Yuan Lin, Yuchen Zhang, Hang Li, and Weinan E. Pasa: An llm agent for comprehensive academic paper search, 2025. URL https://arxiv.org/abs/2501.10120.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning, 2022. URL https://arxiv.org/abs/2112.09118.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NV-embed: Improved techniques for training LLMs as generalist embedding models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=lgsyLSsDRe.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14369–14387, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.775. URL https://aclanthology.org/2024.acl-long.775/.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Renze Lou, Hanzi Xu, Sijia Wang, Jiangshu Du, Ryo Kamoi, Xiaoxin Lu, Jian Xie, Yuxuan Sun, Yusen Zhang, Jihyun Janice Ahn, Hongchao Fang, Zhuoyang Zou, Wenchao Ma, Xi Li, Kai Zhang, Congying Xia, Lifu Huang, and Wenpeng Yin. AAAR-1.0: Assessing AI's potential to assist research. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=RHAWcjIyl2.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. MMLONGBENCH-DOC: Benchmarking long-context document understanding with visualizations. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=loJM1acwzf.

Meta. Llama 4 | model cards and prompt formats. https://www.llama.com/docs/model-cards-and-prompt-formats/llama4/, 2025. Official model card and prompt format documentation.

Meredith Ringel Morris, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. Position: Levels of AGI for operationalizing progress on the path to AGI. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=0ofzEysK2D.

OpenAI. Gpt-5 system card. Technical report, OpenAI, August 2025. URL https://cdn.openai.com/gpt-5-system-card.pdf. Version: August 13, 2025.

OpenAI. gpt-4.1 — openai api documentation, 2025. URL https://platform.openai.com/docs/models/gpt-4.1. Accessed: 2025-09-25.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. SPIQA: A dataset for multimodal question answering on scientific papers. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=h3lddsY5nf.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at trec-3. In *Text Retrieval Conference*, 1994. URL https://api.semanticscholar.org/CorpusID:41563977.

R. Smith. An overview of the tesseract ocr engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pp. 629–633, 2007. doi: 10.1109/ICDAR.2007.4376991.

Rubèn Tito, Minesh Mathew, C. V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. ICDAR 2021 competition on document visualquestion answering. *CoRR*, abs/2111.05547, 2021. URL https://arxiv.org/abs/2111.05547.

Qiuchen Wang, Ruixue Ding, Yu Zeng, Zehui Chen, Lin Chen, Shihang Wang, Pengjun Xie, Fei Huang, and Feng Zhao. Vrag-rl: Empower vision-perception-based rag for visually rich information understanding via iterative reasoning with reinforcement learning, 2025. URL https://arxiv.org/abs/2505.22019.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. Charxiv: Charting gaps in realistic chart understanding in multimodal llms, 2024. URL https://arxiv.org/abs/2406.18521.

xAI. Grok 4 fast model card. Technical report, xAI, September 2025. URL https://data.x.ai/2025-09-19-grok-4-fast-model-card.pdf. Last updated: September 19, 2025.

Dawei Yan, Yang Li, Qing-Guo Chen, Weihua Luo, Peng Wang, Haokui Zhang, and Chunhua Shen. Mmcr: Advancing visual language model in multimodal multi-turn contextual reasoning, 2025. URL https://arxiv.org/abs/2503.18533.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *CoRR*, abs/1809.09600, 2018. URL http://arxiv.org/abs/1809.09600.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=zG459X3Xge.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, 2024. doi: 10.1109/CVPR52733.2024.00913.

Yilun Zhao, Yitao Long, Hongjun Liu, Ryo Kamoi, Linyong Nan, Lyuhao Chen, Yixin Liu, Xiangru Tang, Rui Zhang, and Arman Cohan. DocMath-eval: Evaluating math reasoning capabilities of LLMs in understanding long and specialized documents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16103–16120, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.852. URL https://aclanthology.org/2024.acl-long.852/.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 5168–5191. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/108030643e640ac050e0ed5e6aace48f-Paper-Conference.pdf.

Ruiyang Zhou, Lu Chen, and Kai Yu. Is LLM a reliable reviewer? a comprehensive evaluation of LLM on automatic paper reviewing tasks. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 9340–9351, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.816/.

Table of Contents in Appendix

A PROMPTS

A.1 WITHIN-GENERATE PROMPT

Within-Generate Prompt

```
You will receive a high-quality, already accepted scientific
   paper as a PDF. Working only with the PDF itself (and any
    appendix embedded in the same PDF), edit specific
   textual spans to inject one or more errors chosen only
   from the taxonomy below, such that the errors are hard
   yet clearly identifiable by a professional reviewer
   reading the PDF alone.

Error Type (fixed):
Research Question & Definitions
   Definition: The core construct/hypothesis/variable is
      insufficiently or inconsistently defined (conceptual
      vs operational), leaving the estimand ambiguous.
Design & Identifiability
   Definition: Given a clear estimand, the design violates
      structural identification conditions so the effect is
      not identifiable even with infinite data and perfect
      measurement.
Sampling & Generalizability
   Definition: The sampling frame/process/composition or
      cluster/power setup does not support valid or stable
      sample→population claims.
Measurement & Operationalization
   Definition: Measures/manipulations lack feasibility/
      reliability/validity/timing, so observed variables
      systematically diverge from the intended construct/
      treatment.
Data Handling & Preprocessing
   Definition: Pipeline choices in missing handling, joins/
      keys, temporal splitting, feature construction, or
      partitioning introduce bias (incl. leakage or unit/
      scale conflicts).
Computation & Formulae
   Definition: Arithmetic/algebra/notation errors (totals/
      ratios, unit conversion, CI vs point estimate, p-value
       vs label, symbol reuse, undefined variables,
      dimension mismatch).
Inference & Conclusions
   Definition: Interpretations or causal statements exceed
      what methods/data support, or contradict the shown
      statistics/tables/captions.
Referential and Citation Alignment
   Definition: Contradictions about the same quantity/term
      across text, tables, captions, or appendix within the
      paper.
Language & Expression
   Definition: Terminology/capitalization/grammar ambiguities
       that affect meaning or domain-critical term
      consistency (not cosmetic typos).
```

Within-Generate Prompt (Continued)

```
Global constraints (must comply)
1. Each error must map to exactly one primary category in the
     taxonomy. Do not mix causes.
2. Each error must involve more than 2 micro-edits (each edit
     ≤ 20 English words) spread across distinct pages or
    paragraphs.
3. If an edit would create an immediate contradiction in the
    same sentence/paragraph/caption, you may add shadow patch
    (es) for the same error to keep the text natural (still
    counted as edit locations).
4. Independence across errors (per-copy generation)
   Generate each error on a separate copy of the original PDF
       . Different errors must be logically and operationally
        independent:
   No progression or variant relations: an error must not be
       a stricter/looser version, superset/subset, or minor
       wording variant of another error.
   No anchor reuse: do not target the same sentence/caption/
       table cell or reuse the same old_str (or a near-
       duplicate paraphrase) across different errors.
   Applying any single error in isolation to the original PDF
        must still yield a detectable, clearly categorizable
       error according to the taxonomy.
5. Every error must be supportable using text inside the PDF.
     Do not rely on external supplementary files or prior
    knowledge.
6. Design as difficult as possible but clean errors. Prefer
    edits that force cross-checking between two spots (e.g.,
    Methods vs Results). Avoid trivialities. Edits must
    remain locally plausible and not advertise themselves via
     obviously artificial phrases (e.g., avoid contrived
    tokens purely added to be detectable).
7. ''No cosmetic issues'' applies except for I (Language &
    Expression). For I, edits must affect meaning or domain-
    critical terminology (e.g., ambiguous phrasing,
    inconsistent technical terms). Pure typos, punctuation
    tweaks, or layout nits are not allowed.
8. Do not edit titles, author lists, bibliography entries,
    equation numbering, figure images, or add new figures/
    tables/references.
9. Frame each question as a neutral imperative that asks for
    a decision about a specific condition, using (but not
    limited to) Decide/Determine/Judge/Evaluate/Assess
    whether.... Do not presuppose an outcome or use
    suggestive intensifiers (e.g., clearly/obviously/likely/
    suspicious as examples).
```

Within-Generate Prompt (Continued)

```
10. Output English-only and strictly follow the JSON schema
    below. Do not include any additional text outside the
    JSON:
[
  {
    "id":"1-based integer as string",
    "modify":[
      {
        "location":"Page number + short unique nearby quote (
            ≤15 tokens).",
        "old_str":"Exact original text from the PDF (verbatim)
            .",
        "new_str":"Edited text after your change."
      }
      /* Add 1-2 more locations; each location ≤ 20 words
          changed.
         Shadow patches for local coherence count as locations.
             */
    ],
    "question":"One neutral audit-style task (1-25 words).",
    "explanation":"Explain in 2-4 sentences why a reviewer can
         detect this error from the edited PDF alone.",
    "Type":"Name the primary category (e.g., Inference &
        Conclusions).",
  }
  /* More Errors */
]
```

A.2 Within-Sample Prompt

Within-Sample Prompt

```
You will receive a paper PDF and the weaknesses mentioned in
    its peer-review comments. Your task is, based only on the
     content of that PDF, to sample from the review comments
    and verify possible errors related to the categories
    below, and for each confirmed or highly plausible error,
    generate one question and one explanation.

Error Type (fixed):
Research Question & Definitions
   Definition: The core construct/hypothesis/variable is
       insufficiently or inconsistently defined (conceptual
       vs operational), leaving the estimand ambiguous.
Design & Identifiability
   Definition: Given a clear estimand, the design violates
       structural identification conditions so the effect is
       not identifiable even with infinite data and perfect
       measurement.
Sampling & Generalizability
   Definition: The sampling frame/process/composition or
       cluster/power setup does not support valid or stable
       sample→population claims.
Measurement & Operationalization
   Definition: Measures/manipulations lack feasibility/
       reliability/validity/timing, so observed variables
       systematically diverge from the intended construct/
       treatment.
Data Handling & Preprocessing
   Definition: Pipeline choices in missing handling, joins/
       keys, temporal splitting, feature construction, or
       partitioning introduce bias (incl. leakage or unit/
       scale conflicts).
Computation & Formulae
   Definition: Arithmetic/algebra/notation errors (totals/
       ratios, unit conversion, CI vs point estimate, p-value
        vs label, symbol reuse, undefined variables,
       dimension mismatch).
Inference & Conclusions
   Definition: Interpretations or causal statements exceed
       what methods/data support, or contradict the shown
       statistics/tables/captions.
Referential and Citation Alignment;
   Definition: Contradictions about the same quantity/term
       across text, tables, captions, or appendix within the
       paper.
Language & Expression
   Definition: Terminology/capitalization/grammar ambiguities
        that affect meaning or domain-critical term
       consistency (not cosmetic typos).
```

Within-Sample Prompt (Continued)

```
Global constraints (must comply)
Output only the specified categories; even if other error
   types appear in the reviews, do not output them.
Sample first, then verify: extract candidates from the review
    comments, then confirm them in the PDF. If you cannot
   locate supporting anchors in the PDF (page number plus
   phrase/label), do not output that candidate.
Questions must be neutral and non-leading: use an "audit task
    + decision" style, avoiding yes/no bias.
Independence: each question must target a different figure or
    different textual anchor; no minor variants of the same
   issue.
Evidence first: the explanation must cite locatable anchors
   in the PDF (page number + original phrase/caption). You
   may mention a key short phrase from the review as a clue,
    but write the question and explanation in your own words
Language & format: both question and explanation must be in
   English; output JSON only, with no extra text.
Quantity: sort by evidence strength and output up to 5 items;
    if none qualify, output an empty array [].
Example output
[
  {
    "id": "1",
    "question": "Audit y-axis baselines and possible axis
        breaks in Figure 2; decide presence/absence and cite
        evidence.",
    "explanation": "The review flags possible exaggeration in
        Fig.2. In the PDF (p.6, caption 'Performance vs
        baseline'), the y-axis starts at 0.85 with a break,
        magnifying small differences; panels use different
        ranges."
        "Type":"Visualization & Presentation Bias"
  }
]
```

A.3 Extractor Prompt

Extractor Prompt

```
You will receive three inputs:
Q: the open-ended question;
E: the gold explanation (describes exactly one error; extra
    details still belong to the same single error);
A: the model's answer to be evaluated.
Your job is to extract counts only and output a single JSON
    object with the exact schema below. Do not compute any
    scores. Do not add fields.

Core selection rule (multiple errors in A)
1. Parse E into a single gold error (the "target error").
2. From A, identify how many distinct error claims are made.
    Cluster together mentions that support the same error (
    multiple locations for one error are still one error).
3. Existence decision (binary correctness only):
Let the gold existence be 1 if E asserts an error exists,
    else 0.
Let the predicted existence be 1 if A asserts any error, else
     0 (e.g., states no error).
Set existance = 1 if predicted existence equals gold
    existence; otherwise set existance = 0.
4. If existance = 0: set contains_target_error = 0; set all
    location and reasoning counts to 0; and set
    unrelated_errors to the total number of distinct error
    claims in A. Then output the JSON.
5. If existance = 1:
If the gold existence is 1: determine whether A contains the
    target error (match by the main error idea in E: category
    /intent/scope; treat E's subpoints as the same error).
    If yes, set contains_target_error = 1 and compute location
        and reasoning only for the target error. Count all
       other error claims in A as unrelated_errors.
    If no, set contains_target_error = 0; set all location and
        reasoning counts to 0; set unrelated_errors to the
       total number of distinct error claims in A.
If the gold existence is 0: set contains_target_error = 0;
    set all location and reasoning counts to 0; set
    unrelated_errors to the total number of distinct error
    claims in A. (These negative items are for binary
    accuracy only; they are not used for detailed scoring.)

Matching guidance (A error ↔ target error): match by the
    main error idea in E (category/intent/scope), not by
    wording. Treat E's subpoints as part of the same single
    error. Prefer the best-matching cluster in A; if ties,
    choose the one with stronger alignment to E's core claim.
```

Extractor Prompt (Continued)

```
Counting rules
Location (for the target error only when existance=1 and
   contains_target_error=1):
gold_steps: number of unique error locations described in E (
   after normalization and deduplication).
hit_steps: number of predicted locations in A that match any
   gold location for the target error.
extra_steps: number of predicted locations in A for the
   target error that do not match any gold location.

Reasoning (for the target error only when existance=1 and
   contains_target_error=1):
Convert E into a canonical set or ordered chain of reasoning
   steps for the target error.
gold_steps: total number of such steps.
reached_steps:
   single-chain tasks: length of the longest valid prefix of
      A along the gold chain;
   multi-path/parallel tasks: size of the intersection
      between A's steps and the gold step set (or the
      maximum across gold paths if multiple are defined).
missing_steps: gold_steps - reached_steps (non-negative
   integer).
Unrelated errors:
unrelated_errors: number of distinct error claims in A that
   are not the target error (0 if none).
Output schema (return exactly this JSON; integers only)
{
  "existance": 0,
  "contains_target_error": 0,
  "location": {
    "gold_steps": 0,
    "hit_steps": 0,
    "extra_steps": 0
  },
  "reasoning": {
    "gold_steps": 0,
    "reached_steps": 0,
    "missing_steps": 0
  },
  "unrelated_errors": 0
}
```

A.4 System Prompt

System Prompt

```
You are a neutral, careful academic reviewer. You will
   receive an open-ended question and the paper content. The
    paper may or may not have issues related to the question
    Do not assume there are errors. If the question is about
    citations, you will be given a citing paper and a cited
   paper; evaluate only the citing paper for possible issues
    and use the cited paper only as the reference for
   comparison. Write in natural prose with no fixed template

Rules:
Speak only when sure. State an error only if you are
   confident it is a real error (not a mere weakness).
Stay on scope. Discuss only what the question asks about.
Evidence completeness. For every error you state, list all
   distinct evidence cues you are confident about from the
   PDF. Include plain identifiers (figure/table/section/
   equation/citation) or quotes. Avoid redundant repeats of
   the exact same instance; include all distinct locations
   needed to support the error.
Be clear and brief. Use short, direct sentences.
No metaphors. No fancy wording.No guesses or outside sources.
    Do not invent figures, tables, equations, citations, or
   results.
Report as many distinct, well-supported errors as you can
   within scope. If none are clear, write exactly: "No clear
    issue relevant to the question." and nothing else.
```

B Examples from Existing Datasets

B.1 Example from DocMath-Eval

One Example from DocMath-Eval

Question_ID: complong-testmini-30
Question: What is the percentage of total offering cost on the total amount raised in the IPO if the total offering cost is $14,528,328 and each unit sold is $10?

Context Modalities: **Text Documents**
1. Offering costs consist of legal, accounting and other costs incurred through the balance sheet date that are directly related to the Initial Public Offering. Offering costs amounting to $14,528,328 were charged to shareholders' equity upon the completion of the Initial Public Offering.
2. Pursuant to the Initial Public Offering on July 20, 2020, the Company sold 25,300,000 Units, which includes the full exercise by the underwriter of its option to purchase an additional 3,300,000 Units, at a purchase price of $10.00 per Unit. Each Unit consists of one Class A ordinary share and one-half of one redeemable warrant ("Public Warrant"). Each whole Public Warrant entitles the holder to purchase one Class A ordinary share at an exercise price of $11.50 per whole share (see Note 7).

Covered areas:

Mathematics Only

Cross-evidence Reasoning:

Limited

Task Paradigm:

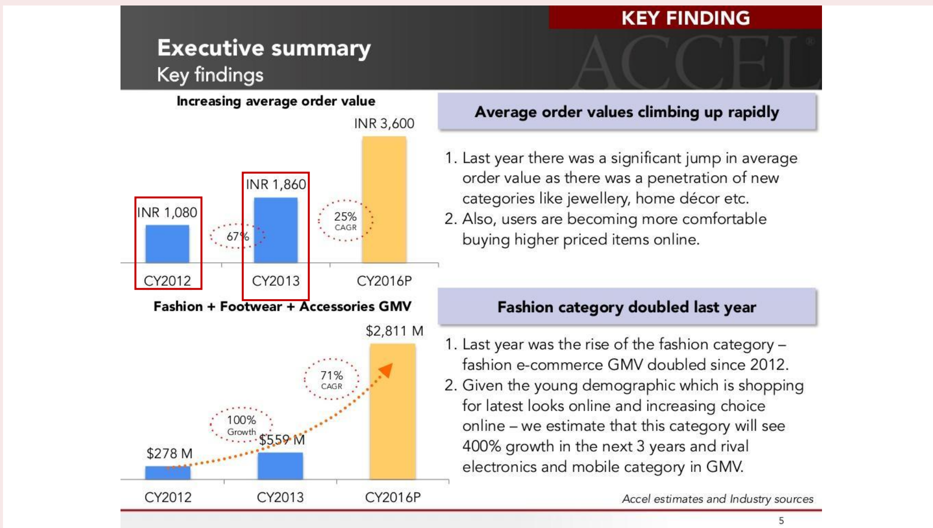
Search-oriented

B.2 EXAMPLE FROM SLIDEVQA

One Example from SlideVQA

Question_ID: 1
Question: How much difference in INR is there between the average order value of CY2013 and that of CY2012?

Context Modalities: **Multi-Modal Documents and Texts**



Covered areas:

Limited

Cross-evidence Reasoning:

None

Task Paradigm:

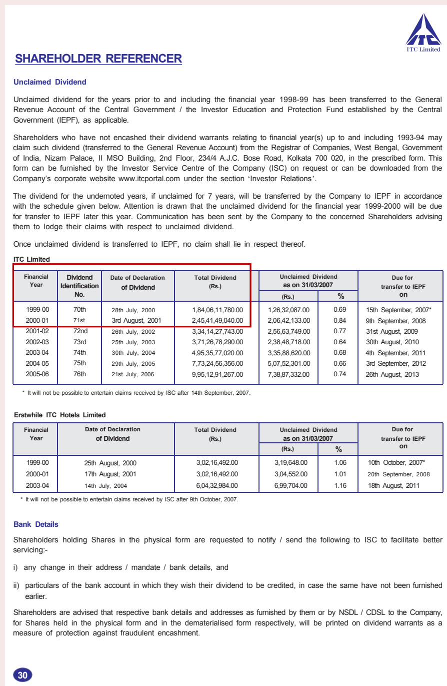
Search-oriented

B.3 EXAMPLE FROM MMLONGBENCH-DOC

One Example from MMLongBench-Doc

Doc ID: afe620b9beac86c1027b96d31d396407.pdf
Question: How much higher was the proposed dividend paid (Rupees in lacs) in 2002 compared to 2001?

Context Modalities: **Multi-Modal Documents and Texts**



ITC Limited

SHAREHOLDER REFERENCER

Unclaimed Dividend

Unclaimed dividend for the years prior to and including the financial year 1998-99 has been transferred to the General Revenue Account of the Central Government / the Investor Education and Protection Fund established by the Central Government (IEPF), as applicable.

Shareholders who have not encashed their dividend warrants relating to financial year(s) up to and including 1993-94 may claim such dividend (transferred to the General Revenue Account) from the Registrar of Companies, West Bengal, Government of India, Nizam Palace, II MSO Building, 2nd Floor, 234/4 A.J.C. Bose Road, Kolkata 700 020, in the prescribed form. This form can be furnished by the Investor Service Centre of the Company (ISC) on request or can be downloaded from the Company's corporate website www.itcportal.com under the section 'Investor Relations'.

The dividend for the undernoted years, if unclaimed for 7 years, will be transferred by the Company to IEPF in accordance with the schedule given below. Attention is drawn that the unclaimed dividend for the financial year 1999-2000 will be due for transfer to IEPF later this year. Communication has been sent by the Company to the concerned Shareholders advising them to lodge their claims with respect to unclaimed dividend.

Once unclaimed dividend is transferred to IEPF, no claim shall lie in respect thereof.

ITC Limited

Financial Year	Dividend Identification No.	Date of Declaration of Dividend	Total Dividend (Rs.)	Unclaimed Dividend as on 31/03/2007		Due for transfer to IEPF on
				(Rs.)	%	
1999-00	70th	28th July, 2000	1,84,06,11,780.00	1,26,32,087.00	0.69	15th September, 2007*
2000-01	71st	3rd August, 2001	2,45,41,49,040.00	2,06,42,133.00	0.84	9th September, 2008
2001-02	72nd	26th July, 2002	3,34,14,27,743.00	2,56,63,749.00	0.77	31st August, 2009
2002-03	73rd	25th July, 2003	3,71,26,78,290.00	2,38,48,718.00	0.64	30th August, 2010
2003-04	74th	30th July, 2004	4,95,35,77,020.00	3,35,88,620.00	0.68	4th September, 2011
2004-05	75th	29th July, 2005	7,73,24,56,356.00	5,07,52,301.00	0.66	3rd September, 2012
2005-06	76th	21st July, 2006	9,95,12,91,267.00	7,38,87,332.00	0.74	26th August, 2013

* It will not be possible to entertain claims received by ISC after 14th September, 2007.

Erstwhile ITC Hotels Limited

Financial Year	Date of Declaration of Dividend	Total Dividend (Rs.)	Unclaimed Dividend as on 31/03/2007		Due for transfer to IEPF on
			(Rs.)	%	
1999-00	25th August, 2000	3,02,16,492.00	3,19,648.00	1.06	10th October, 2007*
2000-01	17th August, 2001	3,02,16,492.00	3,04,552.00	1.01	20th September, 2008
2003-04	14th July, 2004	6,04,32,984.00	6,99,704.00	1.16	18th August, 2011

* It will not be possible to entertain claims received by ISC after 9th October, 2007.

Bank Details

Shareholders holding Shares in the physical form are requested to notify / send the following to ISC to facilitate better servicing:-

i) any change in their address / mandate / bank details, and

ii) particulars of the bank account in which they wish their dividend to be credited, in case the same have not been furnished earlier.

Shareholders are advised that respective bank details and addresses as furnished by them or by NSDL / CDSL to the Company, for Shares held in the physical form and in the dematerialised form respectively, will be printed on dividend warrants as a measure of protection against fraudulent encashment.

30

Covered areas:

Limited (7 Areas)

Cross-evidence Reasoning:

Limited

Task Paradigm:

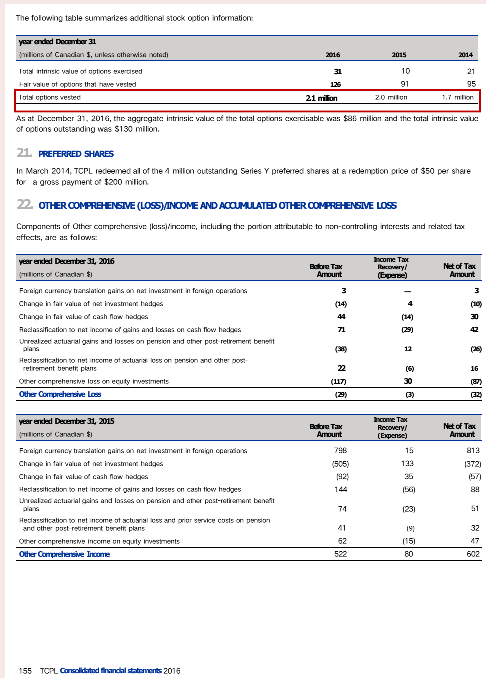
Search-oriented

B.4 EXAMPLE FROM LONGDOCURL

One Example from LongDocURL

Question_ID: free_gemini15_pro_4061601_47_71_8
Question: What was the total fair value of options that vested in 2016, 2015, and 2014, in millions of Canadian dollars?

Context Modalities: **Multi-Modal Documents and Texts**



The following table summarizes additional stock option information:

year ended December 31 (millions of Canadian $, unless otherwise noted)	2016	2015	2014
Total intrinsic value of options exercised	31	10	21
Fair value of options that have vested	126	91	95
Total options vested	2.1 million	2.0 million	1.7 million

As at December 31, 2016, the aggregate intrinsic value of the total options exercisable was $86 million and the total intrinsic value of options outstanding was $130 million.

21. PREFERRED SHARES

In March 2014, TCPL redeemed all of the 4 million outstanding Series Y preferred shares at a redemption price of $50 per share for a gross payment of $200 million.

22. OTHER COMPREHENSIVE (LOSS)/INCOME AND ACCUMULATED OTHER COMPREHENSIVE LOSS

Components of Other comprehensive (loss)/income, including the portion attributable to non-controlling interests and related tax effects, are as follows:

year ended December 31, 2016 (millions of Canadian $)	Before Tax Amount	Income Tax Recovery/ (Expense)	Net of Tax Amount
Foreign currency translation gains on net investment in foreign operations	3	—	3
Change in fair value of net investment hedges	(14)	4	(10)
Change in fair value of cash flow hedges	44	(14)	30
Reclassification to net income of gains and losses on cash flow hedges	71	(29)	42
Unrealized actuarial gains and losses on pension and other post-retirement benefit plans	(38)	12	(26)
Reclassification to net income of actuarial loss on pension and other post-retirement benefit plans	22	(6)	16
Other comprehensive loss on equity investments	(117)	30	(87)
Other Comprehensive Loss	(29)	(3)	(32)

year ended December 31, 2015 (millions of Canadian $)	Before Tax Amount	Income Tax Recovery/ (Expense)	Net of Tax Amount
Foreign currency translation gains on net investment in foreign operations	798	15	813
Change in fair value of net investment hedges	(505)	133	(372)
Change in fair value of cash flow hedges	(92)	35	(57)
Reclassification to net income of gains and losses on cash flow hedges	144	(56)	88
Unrealized actuarial gains and losses on pension and other post-retirement benefit plans	74	(23)	51
Reclassification to net income of actuarial loss and prior service costs on pension and other post-retirement benefit plans	41	(9)	32
Other comprehensive income on equity investments	62	(15)	47
Other Comprehensive Income	522	80	602

155 TCPL Consolidated financial statements 2016

Covered areas:

Limited

Task Paradigm: search

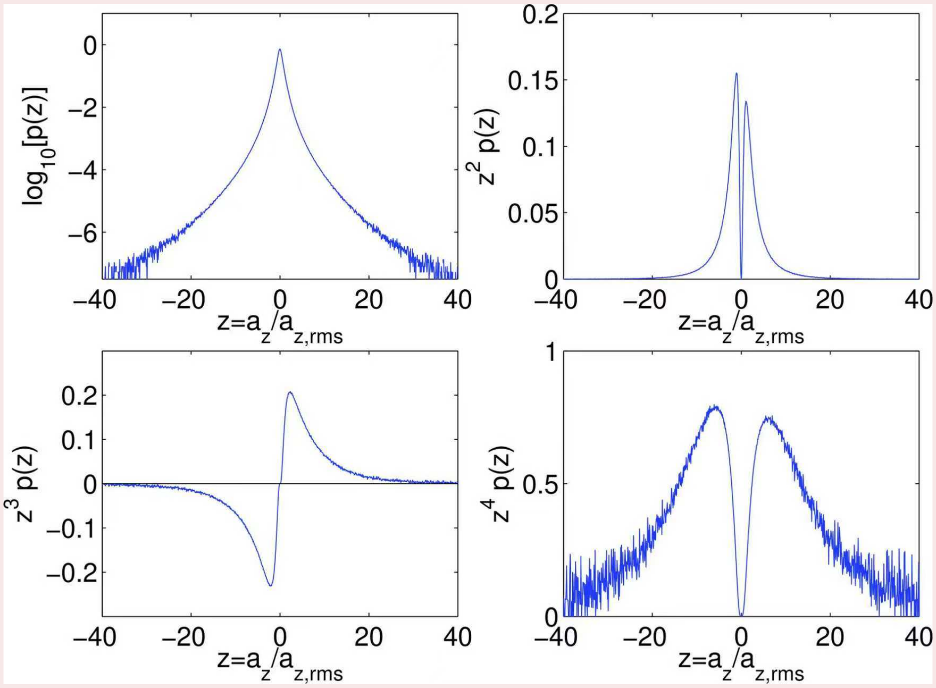
Search-oriented

B.5 EXAMPLE FROM ARXIVQA

One Example from ArXivQA

Question_ID: physics-8049
Question: Based on the top-right graph, how would you describe the behavior of P(z) as z approaches zero?

Context Modalities: **Images**



Covered areas:

Limited

Cross-evidence Reasoning:

None

Task Paradigm:

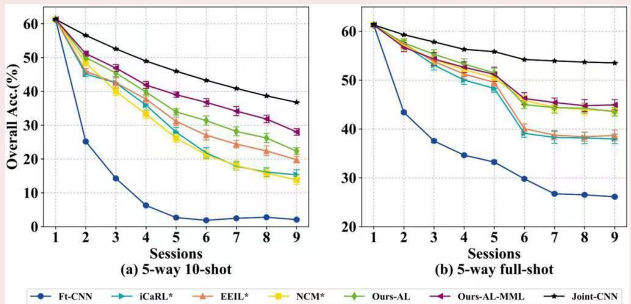
Search-oriented

B.6 EXAMPLE FROM CHARXIV

One Example from Charxiv

Question_ID: 2004.10956
Question: Which model shows a greater decline in accuracy from Session 1 to Session 9 in the 5-way full-shot scenario?

Context Modalities: **Images**



Covered areas:

Limited (8 Areas)

Cross-evidence Reasoning:

None

Task Paradigm:

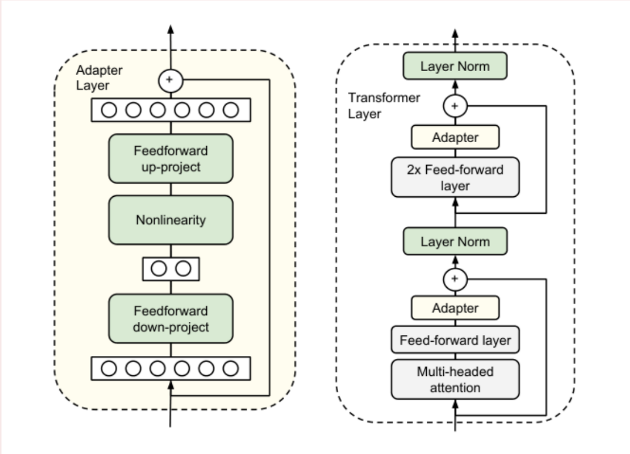
Search-oriented

B.7 EXAMPLE FROM AAAR

One Example from AAAR

Question_ID: 1902.00751
Question: What experiments do you suggest doing? Why do you suggest these experiments?

Context Modalities: **Multi-Modal Documents**



Covered areas:

Limited

Task Paradigm:

Search-oriented

B.8 EXAMPLE FROM MMCR

One Example from MMCR

Question_ID: 1
Question: Which module's weights are frozen?

Context Modalities: **Multi-Modal Documents and Texts**

arXiv:2307.13529v2 [cs.CV] 18 Sep 2023

Re-mine, Learn and Reason: Exploring the Cross-modal Semantic Correlations for Language-guided HOI detection

Yichao Cao, Southeast University, caoyichao@seu.edu.cn
Qingfei Tang, Nanjing Enbo Tech., qingfeitang@gmail.com
Feng Yang, Southeast University, yangfeng@seu.edu.cn
Xiu Su*, University of Sydney, xisu5992@uni.sydney.edu.au
Shan You, SenseTime, youshan@sensetime.com
Xiaobo Lu, Southeast University, xblu@seu.edu.cn
Chang Xu, University of Sydney, c.xu@sydney.edu.au

Abstract

Human-Object Interaction (HOI) detection is a challenging computer vision task that requires visual models to address the complex interactive relationship between humans and objects and predict $<human, action, object>$ *triplets. Despite the challenges posed by the numerous interaction combinations, they also offer opportunities for multi-modal learning of visual texts. In this paper, we present a systematic and unified framework (**RmLR**) that enhances HOI detection by incorporating structured text knowledge. Firstly, we qualitatively and quantitatively analyze the loss of interaction information in the two-stage HOI detector and propose a re-mining strategy to generate more comprehensive visual representation. Secondly, we design more fine-grained sentence- and word-level alignment and knowledge transfer strategies to effectively address the many-to-many matching problem between multiple interactions and multiple texts. These strategies alleviate the matching confusion problem that arises when multiple interactions occur simultaneously, thereby improving the effectiveness of the alignment process. Finally, HOI reasoning by visual features augmented with textual knowledge substantially improves the understanding of interactions. Experimental results illustrate the effectiveness of our approach, where state-of-the-art performance is achieved on public benchmarks.*

*Corresponding author.

1. Introduction

Human-object interaction (HOI) detection [16, 6] is an emerging field of research that builds upon object detection and requires more advanced high-level visual understanding. A high-performing HOI detector should not only accurately localize all interacting Human-Object pairs but also recognize their specific interactions, typically represented as an HOI triplet in the format of $<human, action, object>$ [68].

Previous approaches for achieving HOI detection can be divided into two pipelines: those that treat object detection and interaction recognition as separate stages [63, 6, 13, 14, 34, 20], and those that aim to handle both simultaneously [15, 26, 67, 36, 7]. Although both paradigms have made significant progress, the task remains challenging due to the vast variety of human-object interaction combinations in the real world [59, 60]. For example, the HICO-DET dataset [6] contains 600 human-object interaction combinations. A common approach is to optimize the model by mapping these various triplet labels into a discrete one-hot labels. However, this method oversimplifies the intricacy of the HOI task and can be cumbersome for model optimization.

In recent years, multi-modal learning has gained significant attention in the vision-and-language learning domain, where it has achieved state-of-the-art performance on various tasks [25, 3, 4, 31, 1, 23]. By integrating information from multiple modalities, such as images [50, 48, 51, 49] and text [65], multi-modal learning can provide a more comprehensive understanding of entities or events. In the field of HOI, several recent studies [66, 21, 35, 57, 59, 60] have applied image-and-text models to improve interaction detection performance. For example, HOI-VP [66] used a set of

Covered areas:

Limited

Cross-evidence Reasoning:

Limited

Task Paradigm:

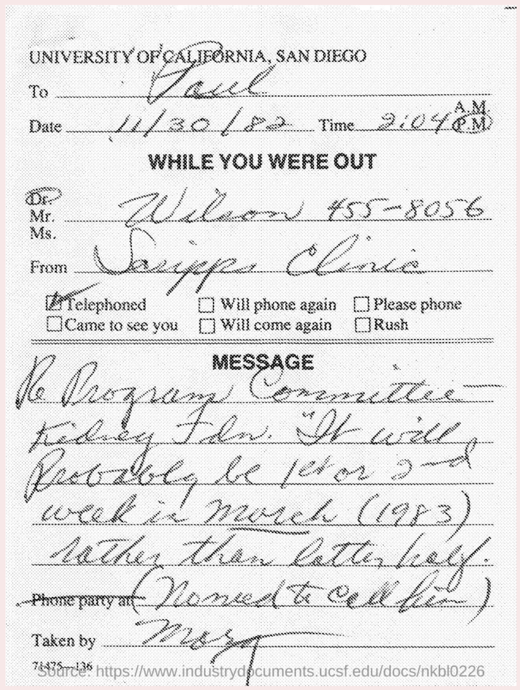
Search-oriented

B.9 EXAMPLE FROM DOCVQA

One Example from DocVQA

Question_ID: 24581
Question: What is name of university?

Context Modalities: **Multi-Modal Documents**



UNIVERSITY OF CALIFORNIA, SAN DIEGO
To Paul
Date 11/30/82 Time 2:04 A.M. P.M.
WHILE YOU WERE OUT
Dr. Mr. Ms. Wilson 455-8056
From Scripps Clinic
☑ Telephoned ☐ Will phone again ☐ Please phone
☐ Came to see you ☐ Will come again ☐ Rush
MESSAGE
Re Program Committee —
Kidney Fdn. It will
probably be 1st or 2nd
week in March (1983)
rather than latter half.
~~Phone party at~~ (No need to call him)
Taken by Mog
71475—136
Source: https://www.industrydocuments.ucsf.edu/docs/nkbl0226

Covered areas:

Limited

Cross-evidence Reasoning:

None

Task Paradigm:

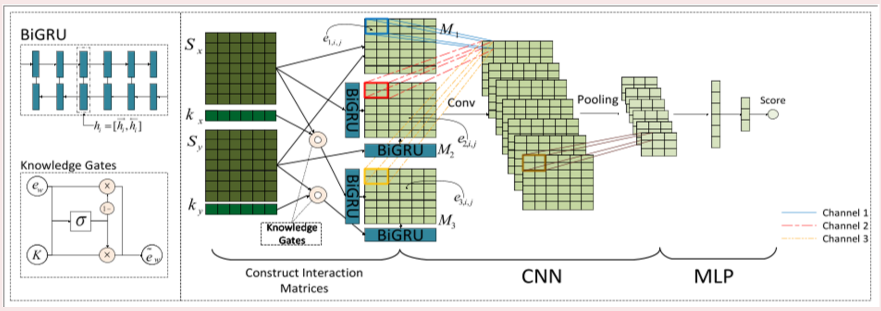
Search-oriented

B.10 EXAMPLE FROM SPIQA

One Example from SPIQA

Question_ID: 1611.04684v1
Question: What is the role of the knowledge gates in the KEHNN architecture?

Context Modalities: **Multi-Modal Figures and Charts**



Covered areas:

Limited

Cross-evidence Reasoning:

None

Task Paradigm:

Search-oriented

C Dataset Annotation and Construction

C.1 Human Annotator Guidelines

The defective academic papers in our dataset are curated from three primary sources:

1. We synthetically inject nine types of errors into papers accepted at ICLR and *Nature Communications*.
2. For papers rejected by ICLR, we identify the shortcomings based on reviewers' comments and categorize them into the same nine error types.
3. For accepted ICLR papers, we generate consistency-related errors by cross-referencing their content against the cited literature.

To ensure the quality of each error, all entries undergo a rigorous, multistage validation protocol executed by human annotators. For synthetically generated errors, annotators manually embed them into the source papers following this protocol:

- **Credibility Validation**: Each error must be logically sound and verifiable. For generated errors, annotators first confirm their logical coherence and unambiguity. Flawed error descriptions are revised whenever possible; only irrepairable cases are discarded.
- **Evidence Verification**: All evidence substantiating an error must be either directly traceable to the source document or grounded in established domain-specific knowledge. Annotators are required to meticulously verify the origin and accuracy of all supporting data and background information.
- **Category Classification**: Each error must be accurately classified into one of the 9 predefined categories according to their formal definitions. Annotators verify the correctness of the assigned category and reclassify it if necessary.
- **Paper Revision**: Upon successful validation, annotators embed the generated error into the original manuscript by adding, deleting, or modifying relevant text segments as dictated by the error's specification.

This unified and standardized annotation protocol enables the creation of a high-quality dataset of academic papers with curated errors, providing a robust benchmark for evaluating the document sacnning and error detection capabilities of Large Multimodal Models.

C.2 Annotation Statistics

Initially, we generated or sampled a pool of 3,500 academic paper instances containing potential errors. During the manual annotation phase, following the protocol described above, we discarded 1,700 instances to ensure the logical rigor of the errors, the accuracy of the evidence, and a balanced distribution of categories.

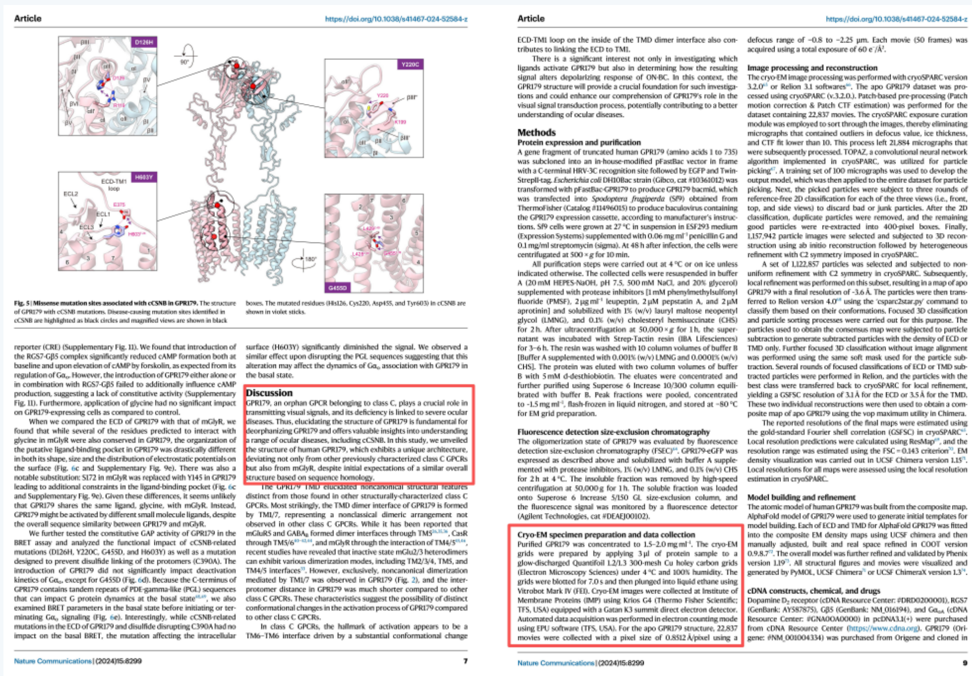
Of the remaining 1,800 instances, 1,541(85.6%) underwent manual revision. The distribution of these modifications is as follows:

- **535 questions** were rewritten to eliminate ambiguity or to increase their retrieval and reasoning difficulty.
- **1,207 explanations** were revised to correct erroneous evidence references and resolve logical flaws.
- **1,141 instances** underwent category reclassification or manual paper editing. This process served to fix classifications that were inconsistent with our definitions and, for errors generated, to manually inject them into the source papers to create the flawed documents.

C.3 EXAMPLES OF ANNOTATION

C.3.1 CASE 1: DISCARD DIRECTLY

Example



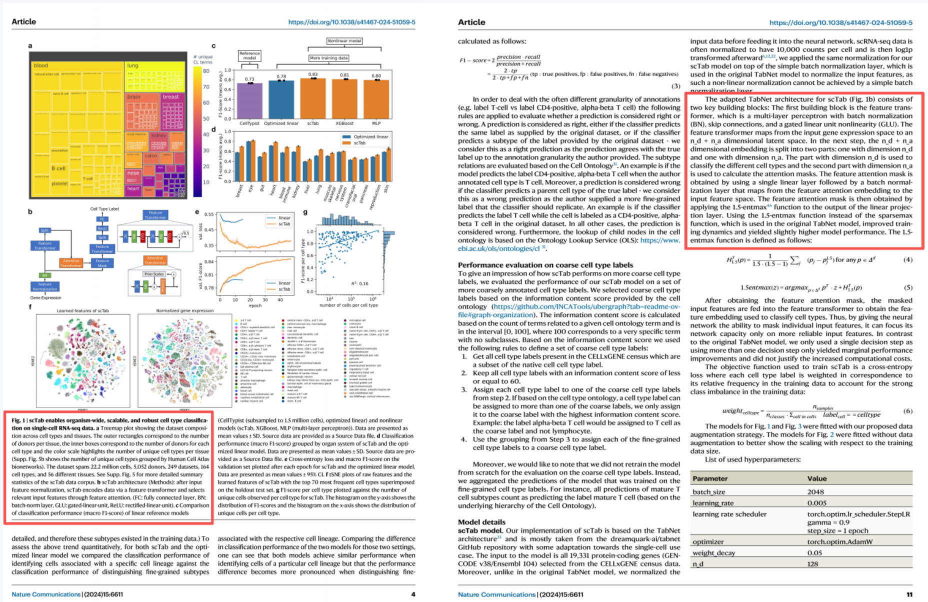
Question: Assess whether the conclusions drawn about the protein's functional state and therapeutic applicability are supported by the presented methods and results.
Explanation: Edits in the abstract and discussion claim the paper presents an active-state structure that reveals the activation mechanism and provides a roadmap for drug design. This overstates the findings, as the paper repeatedly describes solving the 'apo' (unbound) structure and explicitly states the activating ligand is unknown (p.6). To make the error subtle, a contradictory sentence was added to the methods (p.9) claiming a stabilizing agonist was used, but this is falsified by the numerous, unmodified mentions of the 'apo GPR179' structure throughout the results and methods.
Error Type: IC (Inference & Conclusions)

Decision: **Discard**

Analysis: Based on the modifications, the revised abstract and conclusion claim that the paper elucidates the protein's 'active-state' structure and provides a roadmap for drug design. However, the original text repeatedly states (e.g., on pages 5 and 9) that it is the 'apo' (inactive) structure that was resolved, and critically notes on page 6 that the 'activating ligand is still unknown'.This constitutes a clear RCA-type error, defined by the inconsistent description of a concept within the article. Yet, the large model misclassifies this as an IC-type (Inference & Conclusions) error, which is a significant mistake. Considering that the inconsistency regarding the 'active-state' description is overly superficial and obvious, a type of error almost never encountered in actual academic literature, it lacks practical value. Even reclassifying it as an H-type question would be of little significance. Therefore, we have decided to delete this instance.

C.3.2 CASE 2: MODIFY QUESTION

Example



Article https://doi.org/10.1038/s41467-024-51059-5

Fig. 1 | scTab enables organism-wide, scalable, and robust cell type classification on single-cell RNA-seq data. a Treemap plot showing the dataset composition across cell types and tissues. The outer rectangles correspond to the number of donors per tissue, the inner boxes correspond to the number of donors for each cell type and the color scale highlights the number of unique cell types per tissue (Supp. Fig. 5b shows the number of unique cell types grouped by Human Cell Atlas bionetworks). The dataset spans 22.2 million cells, 5,052 donors, 249 datasets, 164 cell types, and 56 different tissues. See Supp. Fig. 5 for more detailed summary statistics of the scTab data corpus. **b** scTab architecture (Methods): after input feature normalization, scTab encodes data via a feature transformer and selects relevant input features through feature attention. (FC: fully connected layer, BN: batch-norm layer, GLU: gated-linear-unit, ReLU: rectified-linear-unit). **c** Comparison of classification performance (macro F1-score) of linear reference models (CellTypist (subsampled to 1.5 million cells), optimized linear) and nonlinear models (scTab, XGBoost, MLP (multi-layer perceptron)). Data are presented as mean values ± SD. Source data are provided as a Source Data file. **d** Classification performance (macro F1-score) grouped by organ system of scTab and the optimized linear model. Data are presented as mean values ± SD. Source data are provided as a Source Data file. **e** Cross-entropy loss and macro F1-score on the validation set plotted after each epoch for scTab and the optimized linear model. Data are presented as mean values ± 95% CI. **f** tSNE plots of raw features and the learned features of scTab with the top 70 most frequent cell types superimposed on the holdout test set. **g** F1-score per cell type plotted against the number of unique cells observed per cell type for scTab. The histogram on the y-axis shows the distribution of F1-scores and the histogram on the x-axis shows the distribution of unique cells per cell type.

detailed, and therefore these subtypes existed in the training data.) To assess the above trend quantitatively, for both scTab and the optimized linear model we compared the classification performance of identifying cells associated with a specific cell lineage against the classification performance of distinguishing fine-grained subtypes associated with the respective cell lineage. Comparing the difference in classification performance of the two models for those two settings, one can see that both models achieve similar performance when identifying cells of a particular cell lineage but that the performance difference becomes more pronounced when distinguishing fine-

Nature Communications | (2024)15:6611 4

Article https://doi.org/10.1038/s41467-024-51059-5

calculated as follows:

$$F1-score = 2\frac{precision \cdot recall}{precision + recall} = \frac{2 \cdot tp}{2 \cdot tp + fp + fn} \text{ (tp : true positives, fp : false positives, fn : false negatives)} \quad (3)$$

In order to deal with the often different granularity of annotations (e.g. label T-cell vs label CD4-positive, alpha-beta T cell) the following rules are applied to evaluate whether a prediction is considered right or wrong. A prediction is considered as right, either if the classifier predicts the same label as supplied by the original dataset, or if the classifier predicts a subtype of the label provided by the original dataset - we consider this as a right prediction as the prediction agrees with the true label up to the annotation granularity the author provided. The subtype relations are evaluated based on the Cell Ontology[31]. An example is if the model predicts the label CD4-positive, alpha-beta T cell when the author annotated cell type is T cell. Moreover, a prediction is considered wrong if the classifier predicts a parent cell type of the true label - we consider this as a wrong prediction as the author supplied a more fine-grained label that the classifier should replicate. An example is if the classifier predicts the label T cell while the cell is labeled as a CD4-positive, alpha-beta T cell in the original dataset. In all other cases, the prediction is considered wrong. Furthermore, the lookup of child nodes in the cell ontology is based on the Ontology Lookup Service (OLS): https://www.ebi.ac.uk/ols/ontologies/cl [31].

Performance evaluation on coarse cell type labels
To give an impression of how scTab performs on more coarse cell type labels, we evaluated the performance of our scTab model on a set of more coarsely annotated cell type labels. We selected coarse cell type labels based on the information content score provided by the cell ontology (https://github.com/INCATools/ubergraph?tab=readme-ov-file#graph-organization). The information content score is calculated based on the count of terms related to a given cell ontology term and is in the interval [0, 100], where 100 corresponds to a very specific term with no subclasses. Based on the information content score we used the following rules to define a set of coarse cell type labels:

1. Get all cell type labels present in the CELLxGENE census which are a subset of the native cell cell type label.
2. Keep all cell type labels with an information content score of less or equal to 60.
3. Assign each cell type label to one of the coarse cell type labels from step 2. If based on the cell type ontology, a cell type label can be assigned to more than one of the coarse labels, we only assign it to the coarse label with the highest information content score. Example: the label alpha-beta T cell would be assigned to T cell as the coarse label and not lymphocyte.
4. Use the grouping from Step 3 to assign each of the fine-grained cell type labels to a coarse cell type label.

Moreover, we would like to note that we did not retrain the model from scratch for the evaluation on the coarse cell type labels. Instead, we aggregated the predictions of the model that was trained on the fine-grained cell type labels. For instance, all predictions of mature T cell subtypes count as predicting the label mature T cell (based on the underlying hierarchy of the Cell Ontology).

Model details
scTab model. Our implementation of scTab is based on the TabNet architecture[33] and is mostly taken from the dreamquark-ai/tabnet GitHub repository with some adaptation towards the single-cell use case. The input to the model is all 19,331 protein-coding genes (GENCODE v38/Ensembl 104) selected from the CELLxGENE census data. Moreover, unlike in the original TabNet model, we normalized the input data before feeding it into the neural network. scRNA-seq data is often normalized to have 10,000 counts per cell and is then log1p transformed afterward[6,12,22], we applied the same normalization for our scTab model on top of the simple batch normalization layer, which is used in the original TabNet model to normalize the input features, as such a non-linear normalization cannot be achieved by a simple batch normalization layer.

The adapted TabNet architecture for scTab (Fig. 1b) consists of two key building blocks: The first building block is the feature transformer, which is a multi-layer perceptron with batch normalization (BN), skip connections, and a gated linear unit nonlinearity (GLU). The feature transformer maps from the input gene expression space to an n_d + n_a dimensional latent space. In the next step, the n_d + n_a dimensional embedding is split into two parts: one with dimension n_d and one with dimension n_a. The part with dimension n_d is used to classify the different cell types and the second part with dimension n_a is used to calculate the attention masks. The feature attention mask is obtained by using a single linear layer followed by a batch normalization layer that maps from the feature attention embedding to the input feature space. The feature attention mask is then obtained by applying the 1.5-entmax[46] function to the output of the linear projection layer. Using the 1.5-entmax function instead of the sparsemax function, which is used in the original TabNet model, improved training dynamics and yielded slightly higher model performance. The 1.5-entmax function is defined as follows:

$$H_{1.5}^T(p) = \frac{1}{1.5 \cdot (1.5-1)} \sum_j \left(p_j - p_j^{1.5}\right) \text{ for any } p \in \Delta^d \quad (4)$$

$$1.5entmax(z) = argmax_{p \in \Delta^d}\, p^T \cdot z + H_{1.5}^T(p) \quad (5)$$

After obtaining the feature attention mask, the masked input features are fed into the feature transformer to obtain the feature embedding used to classify cell types. Thus, by giving the neural network the ability to mask individual input features, it can focus its network capacity only on more reliable input features. In contrast to the original TabNet model, we only used a single decision step as using more than one decision step only yielded marginal performance improvements and did not justify the increased computational costs.

The objective function used to train scTab is a cross-entropy loss where each cell type label is weighted in correspondence to its relative frequency in the training data to account for the strong class imbalance in the training data:

$$weight_{celltype} = \frac{n_{samples}}{n_{classes} \cdot \sum_{cell\ in\ cells} label_{cell} == celltype} \quad (6)$$

The models for Fig. 1 and Fig. 3 were fitted with our proposed data augmentation strategy. The models for Fig. 2 were fitted without data augmentation to better show the scaling with respect to the training data size.

List of used hyperparameters:

Parameter	Value
batch_size	2048
learning_rate	0.005
learning rate scheduler	torch.optim.lr_scheduler.StepLR gamma = 0.9 step_size = 1 epoch
optimizer	torch.optim.AdamW
weight_decay	0.05
n_d	128

Nature Communications | (2024)15:6611 11

Question: Judge whether the mathematical description of the scTab model architecture presents any dimensional mismatches.
Explanation: The Methods section on page 11 (and the Figure 1b caption) now states that the model's feature transformer creates an embedding of dimension n_d. However, the same paragraph then describes splitting this embedding into two components of dimension n_d and n_d respectively. This is algebraically impossible for any non-zero n_a, and the hyperparameter table on page 12 confirms that n_a is set to 64, creating a fundamental dimensional mismatch in the model's description.
Error Type: CF (Computation & Formulae)

Before:
question: Judge whether the mathematical description of the scTab model architecture presents any dimensional mismatches.

Decision: **Modify**

After:
question: Assess the Methods section for Computation & Formulae issues.

Analysis: Based on the error information and the text, the modified model description states that a vector of dimension n_d is split into two parts: one of dimension n_d and another of dimension n_a. This is algebraically impossible, as the total dimension (n_d) cannot equal the dimension of one of its parts (n_d) plus another non-zero part (n_a is set to 64). This constitutes a clear dimensional mismatch, rendering the model's architectural description logically invalid.The original question was overly specific, as it explicitly prompted an assessment of whether the mathematical description of the scTab model architecture contained 'any dimensional mismatches'. This hint was too detailed, reducing the analytical difficulty for the model. To increase the difficulty, we have revised the question's phrasing to ask only whether the mathematical description of the scTab model architecture presents any problems.

C.3.3 Case 3: Modify Explanation

Example

Question: Evaluate if the composition of the SNAPcmini construct is consistently defined throughout the paper.
Explanation: The results on page 4 state that the assembled SNAPcmini construct includes the SNAPC2 subunit. However, the methods on page 12 describe the construction of SNAPcmini using only SNAPC4, SNAPC3, and SNAPC1, with SNAPC2 explicitly removed from the cloning description. A third conflicting statement on page 6 implies SNAPC2 was expected to be part of the minimal core, creating conceptual and operational inconsistency regarding this key experimental complex.
Error Type: RQD (Research Question & Definitions)

Before:
Explanation: ...with SNAPC2 explicitly removed from the cloning description. A third conflicting statement on page 6...

Decision: **Modify**

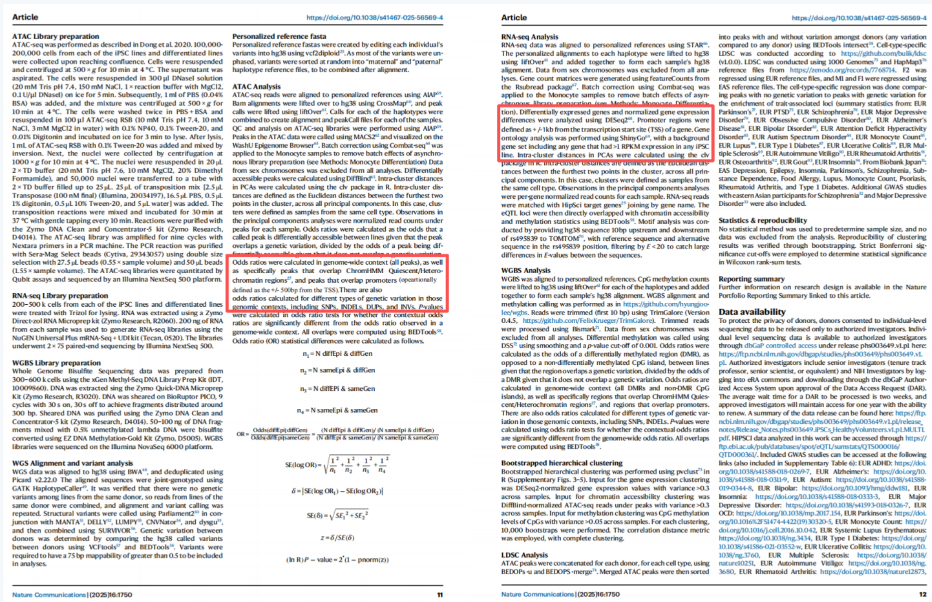
After:
Explanation: ...with SNAPC2 explicitly removed from the cloning description.

Analysis: This instance targets an inconsistency in the operational definition of the SNAPcmini construct. Specifically, the results section states that the assembled SNAPcmini complex includes the SNAPC2 subunit, while the methods section explicitly describes the construction of SNAPcmini using only SNAPC4, SNAPC3, and SNAPC1, with SNAPC2 removed from the cloning procedure. The original explanation additionally referenced a speculative statement regarding SNAPC2's expected presence in the minimal core, which introduced unnecessary ambiguity and reduced the clarity of the definition-level inconsistency. By removing this auxiliary statement, the modified instance focuses on a clear and realistic mismatch between experimental description and implementation, which is representative of RQD-type errors commonly encountered in academic writing.

D Common Failure Cases of MLLMs

D.1 RQD (Research Question & Definitions)

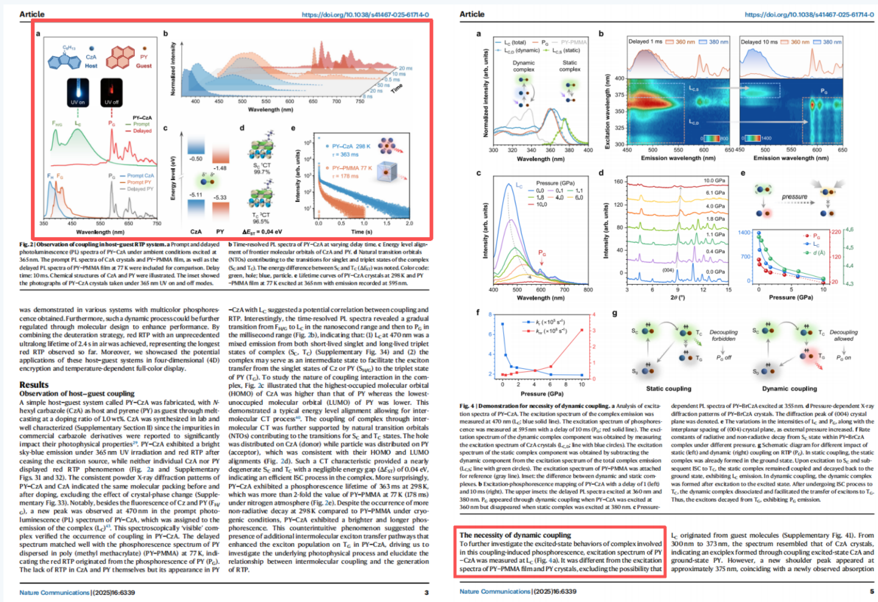
Example: 1001



Question: Assess the Methods section for Research Question & Definitions issues.
Explanation: The definition of a ‘promoter region’, a key analytical construct, is inconsistent across the paper, making the estimand ambiguous. The RNA-seq methods (page 12) define it as +/-1kb from the TSS, the ATAC-seq analysis methods (page 11) define it as +/-500bp from the TSS, and the Results section (page 4) defines it as +/-2kb from the TSS. These three conflicting operational definitions mean that analyses involving ‘promoters’ are not comparable and the construct is insufficiently defined.
Error Type: RQD (Research Question & Definitions)
Type: Within-Generate

D.2 DI (Design & Identifiability)

Example: 1006



Question: Assess the Experiments section for Design & Identifiability issues.
Explanation: The paper's core argument is that it identifies a specific 'dynamic coupling' pathway as essential for RTP, distinct from a 'static coupling' pathway. The edits state that the key experiment (excitation-phosphorescence mapping) cannot distinguish between these two pathways, as the final phosphorescence shows spectral signatures of originating from both. This introduces a structural identification problem: with two potential causal pathways leading to the same outcome and no way to isolate their effects, the claim that the dynamic pathway is the definitive mechanism is not identifiable from the data presented.
Error Type: DI (Design & Identifiability)
Type: Within-Generate

D.3 SG (Sampling & Generalizability)

Example1: 1014

Animal handling
All experimental procedures were approved by the Animal Use and Care Committee of Zhejiang University (#ZJU20220283). The experimental animals involved in this research included a specific substrain of diabetic Institute of Cancer Research (ICR) mice (~20 g) and rabbits (~2 kg), which were obtained from SLAC Laboratory Animal Co. Ltd. (Shanghai, China). All the

nature communications

Article https://doi.org/10.1038/s41467-025-59630-4

High-contrast in vivo fluorescence imaging exploiting wavelengths beyond 1880 nm

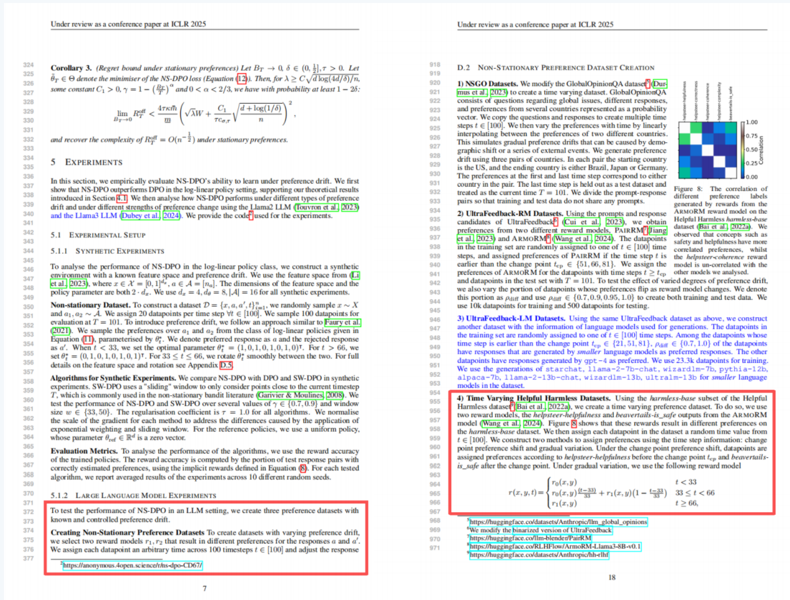
Received: 25 September 2024
Accepted: 25 April 2025
Published online: 13 May 2025

Jiayi Li[1,4], Qiming Xia[2,4], Tianxiang Wu[1,4], Yuhuang Zhang[1], Shiyi Peng[1], Yifei Li[1], Yixuan Li[2], Hui Lin[2], Mingxi Zhang[3] & Jun Qian[1]

The second near-infrared (NIR-II) window is widely acknowledged for its excellent potential in in vivo fluorescence imaging. Currently, NIR-II fluorescence imaging predominantly operates within the 900-1880 nm spectral range, while the region beyond 1880 nm has been disregarded due to the large light absorption of water. Based on a refined understanding of the effect of light absorption on imaging, we propose an approach that utilizes the previously neglected region surrounding the water absorption peak at ~1930 nm for imaging. Both simulations and experiments confirm that the water absorption contributes positively to imaging, enabling high-contrast in vivo fluorescence imaging in the 1880-2080 nm window. To further assess the applicability of this approach in different biological media, we extend our focus to fluorescence imaging in adipose tissue. This leads to the expansion of the imaging window to 1700-2080 nm, owing to the unique light absorption characteristics of adipose tissue. Our results demonstrate that the 1700-2080 nm region provides optimal imaging quality in adipose tissue, attributing to its moderate absorption and low scattering. This work, using a specialized disease model, advances our understanding of the interplay between light absorption and photon scattering in bioimaging, providing a definitive insight for selecting optimal imaging windows to achieve high-contrast fluorescence imaging for all patients.

Question: Assess the Methods section for Sampling & Generalizability issues.
Explanation: The Methods section is edited to state that the experiments used a "specific substrain of diabetic" mice, a highly specialized sample. However, the Abstract and Discussion make broad, unsupported claims of generalizability to "all patients" and the "general patient population." This constitutes an invalid sample-to-population inference.
Error Type: SG (Sampling & Generalizability)
Type: Within-Generate

Example2: 935



Under review as a conference paper at ICLR 2025

Corollary 3. *(Regret bound under stationary preferences) Let $B_T \to 0$, $\delta \in (0, \frac{1}{2}]$, $\tau > 0$. Let $\tilde{\theta}_T \in \Theta$ denote the minimiser of the NS-DPO loss (Equation (12)). Then, for $\lambda \geq C\sqrt{d\log(4d/\delta)/n}$, some constant $C_1 > 0$, $\gamma = 1 - \left(\frac{B_T}{T}\right)^{\alpha}$ and $0 < \alpha < 2/3$, we have with probability at least $1 - 2\delta$:*

$$\lim_{B_T \to 0} R_T^{\text{off}} < \frac{4\tau\kappa\bar{m}}{\underline{m}}\left(\sqrt{\lambda}W + \frac{C_1}{\tau c_{\sigma,\tau}}\sqrt{\frac{d + \log(1/\delta)}{n}}\right)^2,$$

and recover the complexity of $R_T^{\text{off}} = O(n^{-\frac{1}{2}})$ under stationary preferences.

5 Experiments

In this section, we empirically evaluate NS-DPO's ability to learn under preference drift. We first show that NS-DPO outperforms DPO in the log-linear policy setting, supporting our theoretical results introduced in Section 4.1. We then analyse how NS-DPO performs under different types of preference drift and under different strengths of preference change using the Llama2 LLM (Touvron et al., 2023) and the Llama3 LLM (Dubey et al., 2024). We provide the code[2] used for the experiments.

5.1 Experimental Setup

5.1.1 Synthetic Experiments

To analyse the performance of NS-DPO in the log-linear policy class, we construct a synthetic environment with a known feature space and preference drift. We use the feature space from (Li et al., 2023), where $x \in \mathcal{X} = [0,1]^{d_x}$, $a \in \mathcal{A} = [n_a]$. The dimensions of the feature space and the policy parameter are both $2 \cdot d_x$. We use $d_x = 4$, $d_\theta = 8$, $|\mathcal{A}| = 16$ for all synthetic experiments.

Non-stationary Dataset. To construct a dataset $\mathcal{D} = \{x, a, a', t\}_{i=1}^n$, we randomly sample $x \sim X$ and $a_1, a_2 \sim \mathcal{A}$. We assign 20 datapoints per time step $\forall t \in [100]$. We sample 100 datapoints for evaluation at $T = 101$. To introduce preference drift, we follow an approach similar to Faury et al. (2021). We sample the preferences over a_1 and a_2 from the class of log-linear policies given in Equation (11), parameterised by θ_t^*. We denote preferred response as a and the rejected response as a'. When $t < 33$, we set the optimal parameter $\theta_t^* = (1,0,1,0,1,0,1,0)^\intercal$. For $t > 66$, we set $\theta_t^* = (0,1,0,1,0,1,0,1)^\intercal$. For $33 \leq t \leq 66$, we rotate θ_t^* smoothly between the two. For full details on the feature space and rotation see Appendix D.5.

Algorithms for Synthetic Experiments. We compare NS-DPO with DPO and SW-DPO in synthetic experiments. SW-DPO uses a "sliding" window to only consider points close to the current timestep T, which is commonly used in the non-stationary bandit literature (Garivier & Moulines, 2008). We test the performance of NS-DPO and SW-DPO over several values of $\gamma \in \{0.7, 0.9\}$ and window size $w \in \{33, 50\}$. The regularisation coefficient is $\tau = 1.0$ for all algorithms. We normalise the scale of the gradient for each method to address the differences caused by the application of exponential weighting and sliding window. For the reference policies, we use a uniform policy, whose parameter $\theta_{\text{ref}} \in \mathbb{R}^d$ is a zero vector.

Evaluation Metrics. To analyse the performance of the algorithms, we use the reward accuracy of the trained policies. The reward accuracy is computed by the portion of test response pairs with correctly estimated preferences, using the implicit rewards defined in Equation (8). For each tested algorithm, we report averaged results of the experiments across 10 different random seeds.

5.1.2 Large Language Model Experiments

To test the performance of NS-DPO in an LLM setting, we create three preference datasets with known and controlled preference drift.

Creating Non-Stationary Preference Datasets To create datasets with varying preference drift, we select two reward models r_1, r_2 that result in different preferences for the responses a and a'. We assign each datapoint an arbitrary time across 100 timesteps $t \in [100]$ and adjust the response

[2] https://anonymous.4open.science/r/ns-dpo-CD67/

7

Under review as a conference paper at ICLR 2025

D.2 Non-Stationary Preference Dataset Creation

1) NSGO Datasets. We modify the GlobalOpinionQA dataset[5] (Durmus et al., 2023) to create a time varying dataset. GlobalOpinionQA consists of questions regarding global issues, different responses, and preferences from several countries represented as a probability vector. We copy the questions and responses to create multiple time steps $t \in [100]$. We then vary the preferences with time by linearly interpolating between the preferences of two different countries. This simulates gradual preference drifts that can be caused by demographic shift or a series of external events. We generate preference drift using three pairs of countries. In each pair the starting country is the US, and the ending country is either Brazil, Japan or Germany. The preferences at the first and last time step correspond to either country in the pair. The last time step is held out as a test dataset and treated as the current time $T = 101$. We divide the prompt-response pairs so that training and test data do not share any prompts.

helpsteer-helpfulness
helpsteer-correctness
helpsteer-coherence
helpsteer-complexity
beavertails-is_safe
1.00
0.75
0.50
0.25
0.00
Correlation

Figure 8: The correlation of different preference labels generated by rewards from the ArmoRM reward model on the Helpful Harmless *harmless-base* dataset (Bai et al., 2022a). We observed that concepts such as safety and helpfulness have more correlated preferences, whilst the *helpsteer-coherence* reward model is un-correlated with the other models we analysed.

2) UltraFeedback-RM Datasets. Using the prompts and response candidates of UltraFeedback[6] (Cui et al., 2023), we obtain preferences from two different reward models, PairRM[7] (Jiang et al., 2023) and ArmoRM[8] (Wang et al., 2024). The datapoints in the training set are randomly assigned to one of $t \in [100]$ time steps, and assigned preferences of PairRM if the time step t is earlier than the change point $t_{\text{cp}} \in \{51, 66, 81\}$. We assign the preferences of ArmoRM for the datapoints with time steps $t \geq t_{\text{cp}}$ and datapoints in the test set with $T = 101$. To test the effect of varied degrees of preference drift, we also vary the portion of datapoints whose preferences flip as reward model changes. We denote this portion as ρ_{diff} and use $\rho_{\text{diff}} \in \{0.7, 0.9, 0.95, 1.0\}$ to create both training and test data. We use 10k datapoints for training and 500 datapoints for testing.

3) UltraFeedback-LM Datasets. Using the same UltraFeedback dataset as above, we construct another dataset with the information of language models used for generations. The datapoints in the training set are randomly assigned to one of $t \in [100]$ time steps. Among the datapoints whose time step is earlier than the change point $t_{\text{cp}} \in \{21, 51, 81\}$, $\rho_{\text{diff}} \in \{0.7, 1.0\}$ of the datapoints have responses that are generated by *smaller* language models as preferred responses. The other datapoints have responses generated by `gpt-4` as preferred. We use 23.3k datapoints for training. We use the generations of `starchat`, `llama-2-7b-chat`, `wizardlm-7b`, `pythia-12b`, `alpaca-7b`, `llama-2-13b-chat`, `wizardlm-13b`, `ultralm-13b` for *smaller* language models in the dataset.

4) Time Varying Helpful Harmless Datasets. Using the *harmless-base* subset of the Helpful Harmless dataset[9] (Bai et al., 2022a), we create a time varying preference dataset. To do so, we use two reward models, the *helpsteer-helpfulness* and *beavertails-is_safe* outputs from the ArmoRM model (Wang et al., 2024). Figure 8 shows that these rewards result in different preferences on the *harmless-base* dataset. We then assign each datapoint in the dataset a random time value from $t \in [100]$. We construct two methods to assign preferences using the time step information: change point preference shift and gradual variation. Under the change point preference shift, datapoints are assigned preferences according to *helpsteer-helpfulness* before the change point t_{cp} and *beavertails-is_safe* after the change point. Under gradual variation, we use the following reward model

$$r(x,y,t) = \begin{cases} r_0(x,y) & t < 33 \\ r_0(x,y)\frac{(t-33)}{33} + r_1(x,y)\left(1 - \frac{t-33}{33}\right) & 33 \leq t < 66 \\ r_1(x,y) & t \geq 66, \end{cases}$$

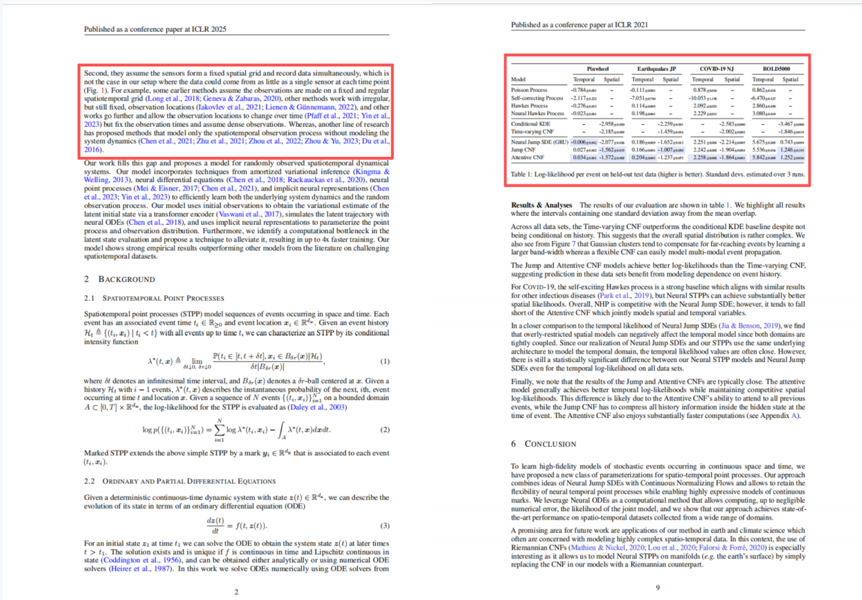
[5] https://huggingface.co/datasets/Anthropic/llm_global_opinions
[6] We modify the binarized version of UltraFeedback
[7] https://huggingface.co/llm-blender/PairRM
[8] https://huggingface.co/RLHFlow/ArmoRM-Llama3-8B-v0.1
[9] https://huggingface.co/datasets/Anthropic/hh-rlhf

18

Question: Evaluate the Experiments section for Sampling & Generalizability problems.
Explanation: The reviewer correctly points out that all experiments are based on synthetic preference drift. The paper's experimental setup, described on pages 7-8 and in Appendix D, involves taking existing datasets (e.g., UltraFeedback) and artificially generating non-stationarity. For instance, preferences are generated by "two different reward models, PAIRRM and ARMORM" (p. 18), and a switch occurs at a predefined change point 'tcp'. Because the core phenomenon being studied—preference drift—is entirely simulated rather than observed organically, the experimental sample does not represent real-world conditions. This limits the generalizability of the paper's findings, as the model's performance on synthetic drift may not translate to its performance on natural, complex preference drift. This is a Sampling & Generalizability (C) issue.
Error Type: SG (Sampling & Generalizability)
Type: Within-Sample

D.4 RCA (Referential and Citation Alignment)

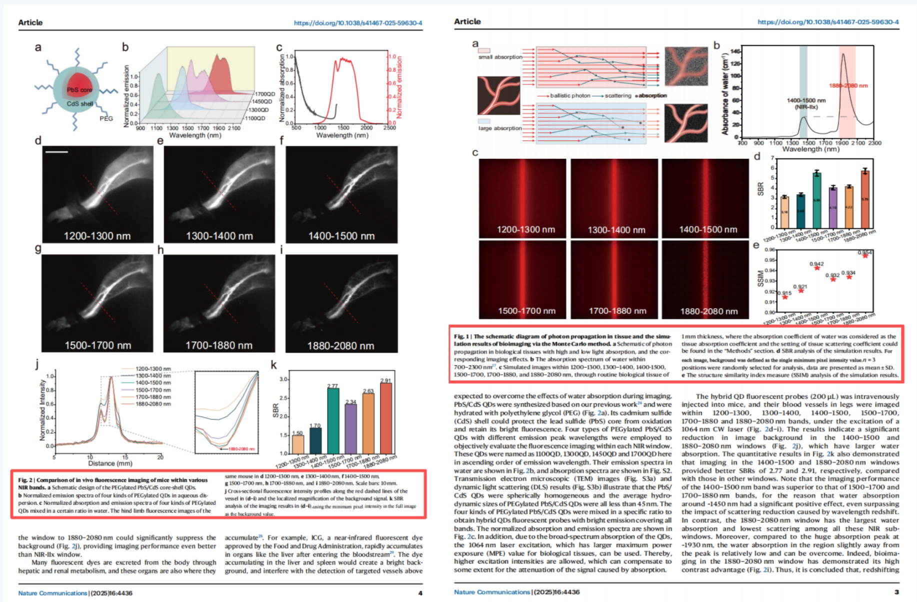
Example: 0



Question: Scan the errors in cited reference Chen et al. (2021)
Explanation: The edited P contains a Type H error by misrepresenting the performance of the cited model. P (p. 8) claims that the NSTPP model from Chen et al. (2021) 'reported performance comparable to a standard Hawkes process baseline'. This contradicts the results in S, where the proposed models (i.e., NSTPP) consistently outperform the Hawkes process baseline, often by a large margin. For example, S (p. 9, Table 1) shows on the BOLD5000 dataset that the 'Attentive CNF' model achieves a temporal log-likelihood of 5.842 ± 0.005, which is substantially better than the Hawkes Process at 2.860 ± 0.050.
Error Type: RCA (Referential and Citation Alignment)
Type: Cross-Generate

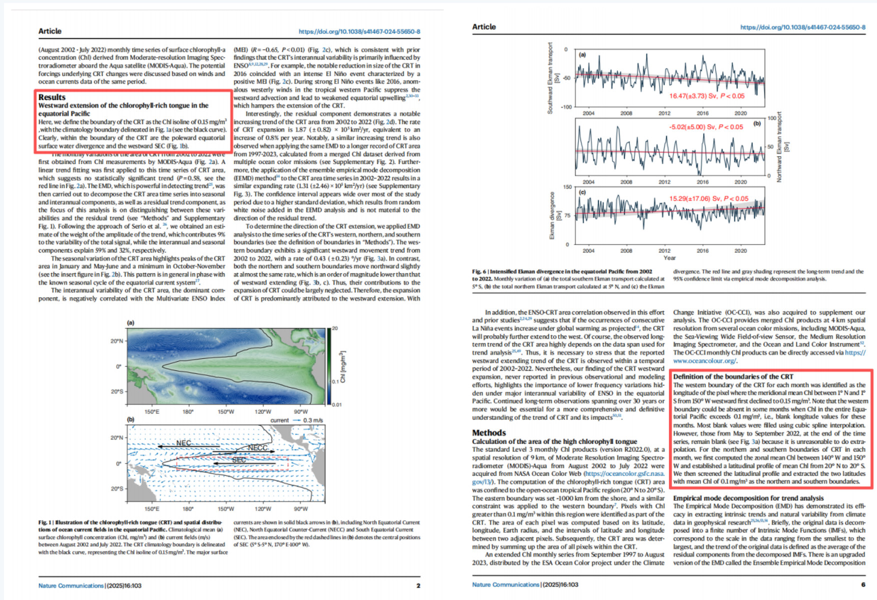
D.5 MO (Measurement & Operationalization)

Example1: 1015



Question: Assess the Figures/Tables section for Measurement & Operationalization issues.
Explanation: The figure captions on pages 3 and 4 have been edited to specify that the background for Signal-to-Background Ratio (SBR) calculations was defined as the single minimum pixel intensity in the image. This is not a valid or reliable operationalization of the "background" construct, as it's highly susceptible to single-point noise or detector artifacts. This flawed measurement procedure systematically undermines all conclusions based on the SBR metric.
Error Type: MO (Measurement & Operationalization)
Type: Within-Generate

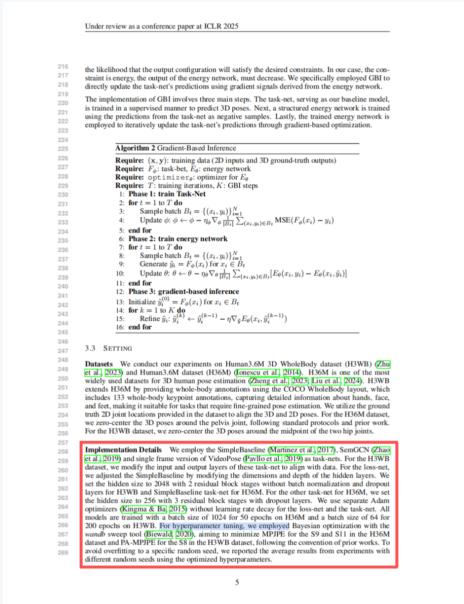
Example2: 1090



Question: Assess the Methods section for Measurement & Operationalization issues.
Explanation: The Results section and the Figure 1 caption define the CRT boundary using a chlorophyll-a (Chl) concentration of 0.15 mg/m³. The Methods section also uses this 0.15 mg/m³ threshold for the western boundary. However, the same Methods section then defines the northern and southern boundaries using a different threshold of 0.1 mg/m³, creating an inconsistent operational definition for the paper's primary construct.
Error Type: MO (Measurement & Operationalization)
Type: Within-Generate

D.6 DHP (Data Handling & Preprocessing)

Example1: 528



Under review as a conference paper at ICLR 2025

the likelihood that the output configuration will satisfy the desired constraints. In our case, the constraint is energy, the output of the energy network, must decrease. We specifically employed GBI to directly update the task-net's predictions using gradient signals derived from the energy network.

The implementation of GBI involves three main steps. The task-net, serving as our baseline model, is trained in a supervised manner to predict 3D poses. Next, a structured energy network is trained using the predictions from the task-net as negative samples. Lastly, the trained energy network is employed to iteratively update the task-net's predictions through gradient-based optimization.

Algorithm 2 Gradient-Based Inference

Require: $(\mathbf{x}, \mathbf{y})$: training data (2D inputs and 3D ground-truth outputs)
Require: F_ϕ: task-bet, E_θ: energy network
Require: $\texttt{optimizer}_\theta$: optimizer for E_θ
Require: T: training iterations, K: GBI steps
1: **Phase 1: train Task-Net**
2: **for** $t = 1$ to T **do**
3: Sample batch $B_t = \{(x_i, y_i)\}_{i=1}^N$
4: Update ϕ: $\phi \leftarrow \phi - \eta_\phi \nabla_\phi \frac{1}{|B_t|} \sum_{(x_i, y_i) \in B_t} \mathrm{MSE}(F_\phi(x_i) - y_i)$
5: **end for**
6: **Phase 2: train energy network**
7: **for** $t = 1$ to T **do**
8: Sample batch $B_t = \{(x_i, y_i)\}_{i=1}^N$
9: Generate $\tilde{y}_i = F_\phi(x_i)$ for $x_i \in B_t$
10: Update θ: $\theta \leftarrow \theta - \eta_\theta \nabla_\theta \frac{1}{|B_t|} \sum_{(x_i, y_i) \in B_t} [E_\theta(x_i, y_i) - E_\theta(x_i, \tilde{y}_i)]$
11: **end for**
12: **Phase 3: gradient-based inference**
13: Initialize $\tilde{y}_i^{(0)} = F_\phi(x_i)$ for $x_i \in B_t$
14: **for** $k = 1$ to K **do**
15: Refine $\tilde{y}_i$: $\tilde{y}_i^{(k)} \leftarrow \tilde{y}_i^{(k-1)} - \eta \nabla_{\tilde{y}} E_\theta(x_i, \tilde{y}_i^{(k-1)})$
16: **end for**

3.3 Setting

Datasets We conduct our experiments on Human3.6M 3D WholeBody dataset (H3WB) (Zhu et al., 2023) and Human3.6M dataset (H36M) (Ionescu et al., 2014). H36M is one of the most widely used datasets for 3D human pose estimation (Zheng et al., 2023; Liu et al., 2024). H3WB extends H36M by providing whole-body annotations using the COCO WholeBody layout, which includes 133 whole-body keypoint annotations, capturing detailed information about hands, face, and feet, making it suitable for tasks that require fine-grained pose estimation. We utilize the ground truth 2D joint locations provided in the dataset to align the 3D and 2D poses. For the H36M dataset, we zero-center the 3D poses around the pelvis joint, following standard protocols and prior work. For the H3WB dataset, we zero-center the 3D poses around the midpoint of the two hip joints.

Implementation Details We employ the SimpleBaseline (Martinez et al., 2017), SemGCN (Zhao et al., 2019) and single frame version of VideoPose (Pavllo et al., 2019) as task-nets. For the H3WB dataset, we modify the input and output layers of these task-net to align with data. For the loss-net, we adjusted the SimpleBaseline by modifying the dimensions and depth of the hidden layers. We set the hidden size to 2048 with 2 residual block stages without batch normalization and dropout layers for H3WB and SimpleBaseline task-net for H36M. For the other task-net for H36M, we set the hidden size to 256 with 3 residual block stages with dropout layers. We use separate Adam optimizers (Kingma & Ba, 2015) without learning rate decay for the loss-net and the task-net. All models are trained with a batch size of 1024 for 50 epochs on H36M and a batch size of 64 for 200 epochs on H3WB. For hyperparameter tuning, we employed Bayesian optimization with the *wandb* sweep tool (Biewald, 2020), aiming to minimize MPJPE for the S9 and S11 in the H36M dataset and PA-MPJPE for the S8 in the H3WB dataset, following the convention of prior works. To avoid overfitting to a specific random seed, we reported the average results from experiments with different random seeds using the optimized hyperparameters.

5

Question: Assess the Methods section for Data Handling & Preprocessing issues.
Explanation: The reviewer correctly identifies that the authors tuned hyperparameters on the test set. The paper's "Implementation Details" section on page 5 states: "For hyperparameter tuning, we employed Bayesian optimization with the wandb sweep tool (Biewald, 2020), aiming to minimize MPJPE for the S9 and S11 in the H36M dataset and PA-MPJPE for the S8 in the H3WB dataset, following the convention of prior works." According to standard protocols for the H36M dataset, subjects S9 and S11 constitute the test set. Tuning hyperparameters directly on the test set introduces data leakage, leading to an optimistic bias in the reported results and invalidating claims of generalization. This is a critical violation of machine learning best practices and fits the Data Handling & Preprocessing (E) category, as a pipeline choice introduces bias.
Error Type: DHP (Data Handling & Preprocessing)
Type: Within-Sample

Example2: 1566

Article

https://doi.org/10.1038/s41467-025-60936-6

hub trusts), many of which opened during lockdown, typically in response to the suspension of planned services during the lockdown, may have taken time to establish their operational standards after the lockdown. We examined impacts on all (or total) elective surgery and, assuming that hubs focus on HVLC elective procedures, we examined HVLC elective surgery separately to better understand the specific contribution of hubs to the overall effect observed across each trust.

Results

A total of 8,077,669 elective surgeries at 31 new-hub trusts, 23 established-hub trusts and 54 trusts without a hub (non-hub trusts) between April 2018 and December 2022 were included in our study, with data for the lockdown months (April 2020 to March 2021) imputed using pre-pandemic averages to create a continuous series (Supplementary Figs. 1 and 2). An earlier mapping study of hubs in England found that they arewidespread across NHS regions in England, mainly in more densely populated areas, with significant variation in the hub types, specialties, care provision, and their healthcare provider contexts[16]. Characteristics of trust catchment populations for trusts with and without hubs were broadly similar. However, trusts without hubs (non- hub trusts) tended to have smaller median catchment population size (279,429 vs 400,106 and 426,938) and larger median proportion of individuals with White ethnicity (93% vs 85% and 89%) than new-hub and established-hub trusts,

respectively (Table 1). The rate of total elective surgery and HVLC elective surgery (hereafter referred to as total-surgery and HVLC-surgery respectively) were consistently higher in non-hub trusts than in new-hub and established-hub trusts throughout ourstudy period (Fig. 2). In the twoyears leading up to the pandemic, the average total-surgery rate at non-hub trusts (4.67 surgeries per 1000 trust catchment population permonth [SD = 1.36]) was 8% higher than in new-hub trusts (4.31 [SD = 1.14]) and 9% higher than in established-hub trusts (4.3 [SD = 1.25]) (Table 2). Post-lockdown, the average total-surgery and HVLC- surgery rates decreased across alltrust types, but larger decreases were observed in non-hub (−17.3% for total and −22.1% for HVLC) than in both new-hub (−15.8% and −20.3%) and established-hub trusts (−13.8% and −18.3%), making rates more similar across alltypes of trust post-pandemic compared with pre-pandemic. Excluding the COVID-19 lockdown period, approximately 65–75% of all surgeries were treated as a day-case with established-hub trusts consistently having the low- estaverage across allsurgeries but notangible differences between trust types for HVLC surgeries. (Supplementary Fig. 3). Pre-pandemic, non-hub trusts had the shortest average inpatient length of hospital stay (hereafter referred to as length of stay) for all surgeries, including HVLC (Supplementary Fig. 4). Post-pandemic, new-hub estab- lished-hub trusts sh- o wed the greatest reduc- and tions in length of stay.

New-hub trusts

Applying the generalised synthetic control model, estimated effect sizes for the impact of opening a new hub on total-surgery and HVLC-surgery rates during the first 12 months were similar (each estimated an additional 0.17–0.18 surgeries per 1000 trust catchment population per month), although these were only significant for HVLC-surgery (Table 3). The HVLC-surgery rate was 0.169 surgeries per 1000 trust catchment population per month (95% confidence interval = 0.090 to 0.248) higher in new-hub trusts during the first 12 months compared with the synthetic control. This is equivalent to a 21.9% (95% CI = 11.7% to 32.2%) higher than expected increase in HVLC-surgery in new-hub trusts when compared with a synthetic control with similar pre-opening trends.

Figure 3 displays observed and counterfactual trends for total-surgery and HVLC-surgery rates (left panels), as well as corresponding impact estimates (right panels), at new-hub trusts. Throughout the pre-intervention period, rates of both total-surgery and HVLC-surgery matched the synthetic control, indicating a good model fit. However, these rates steadily declined, reflecting that the pre-intervention months for many new hubs were just before lockdown, when surgeries were paused (Fig. 2). Positive effects on HVLC-surgery rates were only significant after three months and were driven by a continued reduction in the synthetic control, compared to a marked increase at new-hub trusts.

For all surgeries, the day-case proportion was an average of 0.020 (95% CI = 0.002 to 0.038) higher in new-hub trusts compared to the synthetic control during the first 12 months of opening. There was no significant effect on the day-case proportion for HVLC-surgery or on length of stay for either total-surgery or HVLC-surgery in new-hub trusts during this period (Supplementary Figs. 5, 6).

Established-hub trusts

For established-hub trusts, the total-surgery rate was 0.374 (95% CI = 0.043 to 0.705) surgeries per 1000 trust catchment population per month higher between April 2021 and December 2022 compared with the synthetic control (Table 4). This corresponds to an average increase of 11.2% (95% CI = 1.3% to 21.2%) in all surgeries post-lockdown, due to having an established hub. Similar trends were also seen for the HVLC-surgery rate with an average increase of 11.2% (95% CI = 1.7% to 20.7%) during the post-lockdown period compared with the synthetic control.

As for new-hub trusts, both total-surgery and HVLC-surgery rates matched the synthetic control during the pre-intervention period indicating a good model fit (Fig. 4). Post-intervention, the positive effects on total-surgery and HVLC-surgery rates steadily increased, driven by a faster initial increase in rates in the established-hub trusts not matched over the study period in the estimated synthetic control.

The average length of stay for HVLC-surgery was 0.170 days (95% CI = −0.280 to −0.061) shorter in established-hub trusts than in the synthetic control. There were no significant effects on length of stay across all surgery types at established-hub trusts, nor on the proportion of day cases for either total-surgery or HVLC-surgery (Supplementary Figs. 7, 8).

Specialties and individual trusts

When examining each HVLC specialty in turn, we only found a significant effect on the rate of general surgery at new-hub trusts and no significant effect on any individual specialties at established-hub trusts. General surgery comprises 13% of all HVLC elective surgery in new-hub trusts and increased by 0.023 surgeries per 1000 trust catchment population per month (95% CI = 0.007 to 0.039) during the

Table 1 | Comparison of population characteristics in new-hub, established-hub and non-hub trusts in England in 2019

	New-hub trusts (N = 31) Median (IQR)	Established-hub trusts (N = 23) Median (IQR)	Non-hub trusts (N = 54) Median (IQR)
Population from a White ethnic background (%)	85 (72, 92)	89 (79, 93)	93 (87, 96)
Population who are male (%)	48 (48, 50)	50 (48, 51)	48 (48, 50)
Population who are aged 65+ years old (%)	19 (14, 21)	19 (16, 21)	20 (18, 23)
Catchment population	400,106 (295,955, 579,440)	426,938 (319,217, 612,212)	279,429 (195,782, 462,603)

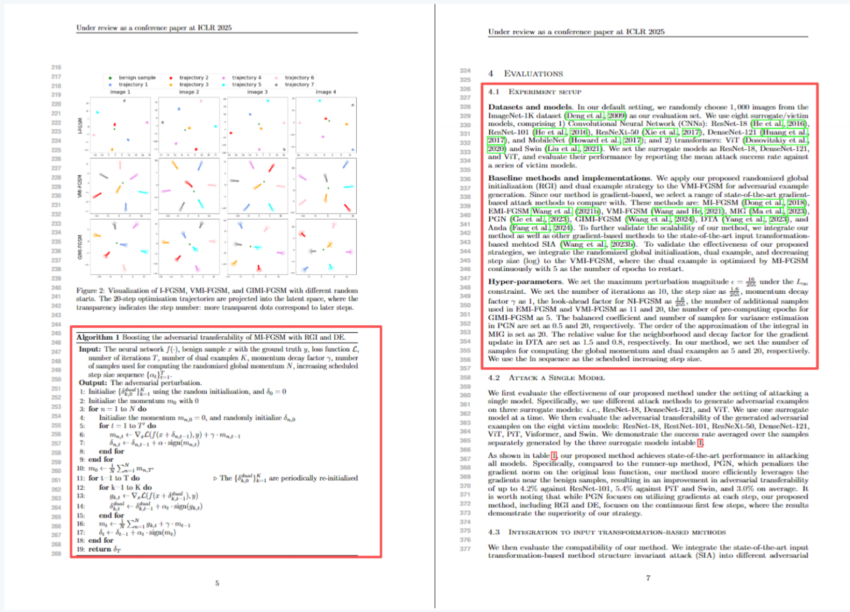
Nature Communications | (2025) 16:6192

3

Question: Assess the Methods section for Data Handling & Preprocessing issues.
Explanation: The modified text on page 3 states that data for the COVID-19 lockdown period were imputed using pre-pandemic averages. This data handling choice is highly problematic, as it smooths over a major, non-random structural break in the time series rather than modeling or excluding it. The imputation method introduces significant bias and data leakage, as a simple average does not accurately reflect the known, drastic reduction in elective surgeries during that specific period, compromising the validity of the causal model.
Error Type: DHP (Data Handling & Preprocessing)
Type: Within-Generate

D.7 CF (COMPUTATION & FORMULAE)

Example: 1909



Under review as a conference paper at ICLR 2025

Figure 2: Visualization of I-FGSM, VMI-FGSM, and GIMI-FGSM with different random starts. The 20-step optimization trajectories are projected into the latent space, where the transparency indicates the step number: more transparent dots correspond to later steps.

Algorithm 1 Boosting the adversarial transferability of MI-FGSM with RGI and DE.

Input: The neural network $f(\cdot)$, benign sample x with the ground truth y, loss function $\mathcal{L}$, number of iterations T, number of dual examples K, momentum decay factor γ, number of samples used for computing the randomized global momentum N, increasing scheduled step size sequence $\{\alpha_t\}_{t=1}^{T}$.
Output: The adversarial perturbation.
1: Initialize $\{\delta_{k,0}^{dual}\}_{k=1}^{K}$ using the random initialization, and $\delta_0 = 0$
2: Initialize the momentum m_0 with 0
3: **for** $n = 1$ to N **do**
4: Initialize the momentum $m_{n,0} = 0$, and randomly initialize $\delta_{n,0}$
5: **for** $t = 1$ to T' **do**
6: $m_{n,t} \leftarrow \nabla_x \mathcal{L}(f(x + \delta_{n,t-1}), y) + \gamma \cdot m_{n,t-1}$
7: $\delta_{n,t} \leftarrow \delta_{n,t-1} + \alpha \cdot \text{sign}(m_{n,t})$
8: **end for**
9: **end for**
10: $m_0 \leftarrow \frac{1}{N}\sum_{n=1}^{N} m_{n,T'}$
11: **for** t=1 to T **do** ▷ The $\{\delta_{k,0}^{dual}\}_{k=1}^{K}$ are periodically re-initialized
12: **for** k=1 to K **do**
13: $g_{k,t} \leftarrow \nabla_x \mathcal{L}(f(x + \delta_{k,t-1}^{dual}), y)$
14: $\delta_{k,t}^{dual} \leftarrow \delta_{k,t-1}^{dual} + \alpha_t \cdot \text{sign}(g_{k,t})$
15: **end for**
16: $m_t \leftarrow \frac{1}{N}\sum_{n=1}^{N} g_{k,t} + \gamma \cdot m_{t-1}$
17: $\delta_t \leftarrow \delta_{t-1} + \alpha_t \cdot \text{sign}(m_t)$
18: **end for**
19: **return** δ_T

5

Under review as a conference paper at ICLR 2025

4 EVALUATIONS

4.1 EXPERIMENT SETUP

Datasets and models. In our default setting, we randomly choose 1,000 images from the ImageNet-1K dataset (Deng et al., 2009) as our evaluation set. We use eight surrogate/victim models, comprising 1) Convolutional Neural Network (CNNs): ResNet-18 (He et al., 2016), ResNet-101 (He et al., 2016), ResNeXt-50 (Xie et al., 2017), DenseNet-121 (Huang et al., 2017), and MobileNet (Howard et al., 2017); and 2) transformers: ViT (Dosovitskiy et al., 2020) and Swin (Liu et al., 2021). We set the surrogate models as ResNet-18, DenseNet-121, and ViT, and evaluate their performance by reporting the mean attack success rate against a series of victim models.

Baseline methods and implementations. We apply our proposed randomized global initialization (RGI) and dual example strategy to the VMI-FGSM for adversarial example generation. Since our method is gradient-based, we select a range of state-of-the-art gradient-based attack methods to compare with. These methods are: MI-FGSM (Dong et al., 2018), EMI-FGSM Wang et al. (2021b), VMI-FGSM (Wang and He, 2021), MIG (Ma et al., 2023), PGN (Ge et al., 2023), GIMI-FGSM (Wang et al., 2024), DTA (Yang et al., 2023), and Anda (Fang et al., 2024). To further validate the scalability of our method, we integrate our method as well as other gradient-based methods to the state-of-the-art input transformation-based mehtod SIA (Wang et al., 2023b). To validate the effectiveness of our proposed strategies, we integrate the randomized global initialization, dual example, and decreasing step size (log) to the VMI-FGSM, where the dual example is optimized by MI-FGSM continuously with 5 as the number of epochs to restart.

Hyper-parameters. We set the maximum perturbation magnitude $\epsilon = \frac{16}{255}$ under the L_∞ constraint. We set the number of iterations as 10, the step size as $\frac{1.6}{255}$, momentum decay factor γ as 1, the look-ahead factor for NI-FGSM as $\frac{1.6}{255}$, the number of additional samples used in EMI-FGSM and VMI-FGSM as 11 and 20, the number of pre-computing epochs for GIMI-FGSM as 5. The balanced coefficient and number of samples for variance estimation in PGN are set as 0.5 and 20, respectively. The order of the approximation of the integral in MIG is set as 20. The relative value for the neighborhood and decay factor for the gradient update in DTA are set as 1.5 and 0.8, respectively. In our method, we set the number of samples for computing the global momentum and dual examples as 5 and 20, respectively. We use the ln sequence as the scheduled increasing step size.

4.2 ATTACK A SINGLE MODEL

We first evaluate the effectiveness of our proposed method under the setting of attacking a single model. Specifically, we use different attack methods to generate adversarial examples on three surrogate models: *i.e.*, ResNet-18, DenseNet-121, and ViT. We use one surrogate model at a time. We then evaluate the adversarial transferability of the generated adversarial examples on the eight victim models: ResNet-18, RestNet-101, ResNeXt-50, DenseNet-121, ViT, PiT, Visformer, and Swin. We demonstrate the success rate averaged over the samples separately generated by the three surrogate models intable 1.

As shown in table 1, our proposed method achieves state-of-the-art performance in attacking all models. Specifically, compared to the runner-up method, PGN, which penalizes the gradient norm on the original loss function, our method more efficiently leverages the gradients near the benign samples, resulting in an improvement in adversarial transferability of up to 4.2% against ResNet-101, 5.4% against PiT and Swin, and 3.0% on average. It is worth noting that while PGN focuses on utilizing gradients at each step, our proposed method, including RGI and DE, focuses on the continuous first few steps, where the results demonstrate the superiority of our strategy.

4.3 INTEGRATION TO INPUT TRANSFORMATION-BASED METHODS

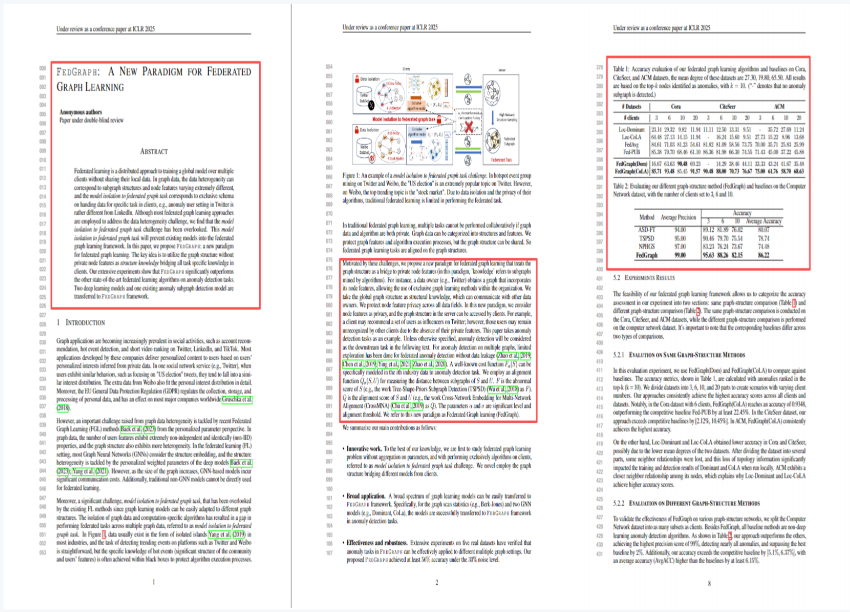
We then evaluate the compatibility of our method. We integrate the state-of-the-art input transformation-based method structure invariant attack (SIA) into different adversarial

7

Question: Scan the Methods section for Computation & Formulae errors.
Explanation: Algorithm 1 on page 5 uses the parameter T' in the loop definition on line 5: for t = 1 to T' do. This parameter determines the number of iterations for the Randomized Global Initialization phase. However, the value of T' is never specified anywhere in the paper, including the "Hyper-parameters" section (Section 4.1 on page 7). An algorithm cannot be implemented or reproduced with an undefined critical parameter. This fits the Computation & Formulae category as an "undefined variable".
Error Type: CF (Computation & Formulae)
Type: Within-Sample

D.8 IC (Inference & Conclusions)

Example: 875



Question: Evaluate Abstract, Introduction and Experiment section for issues in Inference & Conclusions.

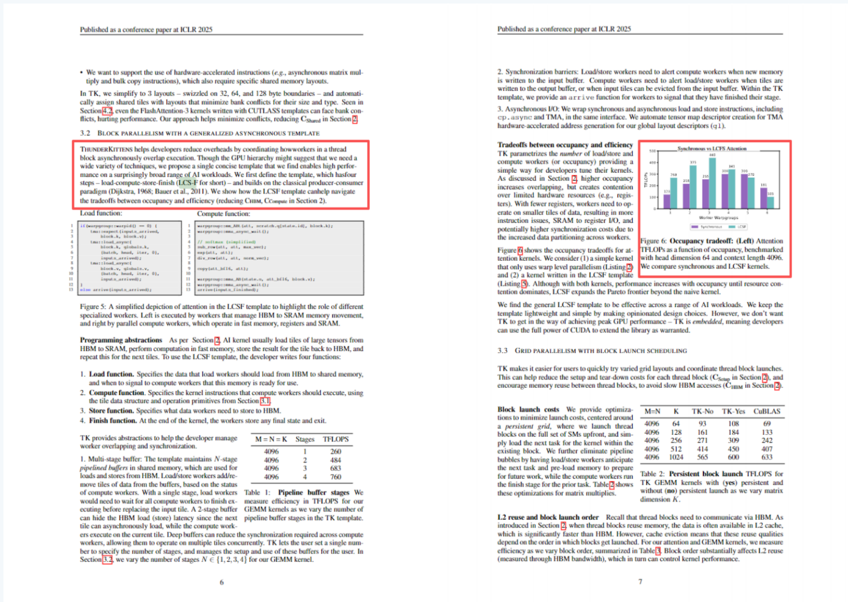
Explanation: The paper's claims of generality are not supported by its evidence. The title and abstract introduce "FedGraph: A New Paradigm for Federated Graph Learning" (page 1), suggesting a broadly applicable framework. However, the methodology is heavily tailored to, and the experiments are exclusively focused on, the single downstream task of anomaly detection. For example, a stated contribution is "Broad application," but this is immediately qualified with "the models are successfully transferred to FEDGRAPH framework in anomaly detection tasks" (page 2). Furthermore, Section 5, "EXPERIMENTS", exclusively reports results on anomaly detection tasks. This discrepancy represents an issue of Inference & Conclusions, as the broad conclusion of having created a new "paradigm" for FGL is an overstatement that exceeds what the narrow experimental results can support.

Error Type: IC (Inference & Conclusions)

Type: Within-Sample

D.9 LE (LANGUAGE & EXPRESSION)

Example: 1785



Question: Assess the Methods section for Language & Expression issues.
Explanation: The paper introduces a key contribution, the 'load-compute-store-finish' template, and its acronym 'LCSF'. This error introduces inconsistencies in this critical term: it's defined as 'LCS-F' on page 6, called 'LCFS' in a figure title on page 7, and written out in full in the conclusion on page 10, while the original 'LCSF' acronym remains elsewhere. This terminological inconsistency for a central, paper-defined concept creates ambiguity and undermines the paper's precision.
Error Type: LE (Language & Expression)
Type: Within-Generate

E Human-Machine Consistency Evaluation

To evaluate whether GPT-4.1 accurately extracts detailed information from the model responses, we conduct a human-Machine consistency evaluation. We first randomly sampled 200 questions from the dataset. Then, we invited human experts to analyze the corresponding model-generated responses for these questions and to manually extract key information, including evidence sets, reasoning chains, and the number of unrelated errors.The results are presented in Table 4.

	S_{total}	S_{location}	$S_{\text{reasoning}}$	$P_{\text{unrelated\_err}}$
Spearman's correlation coefficients	0.841	0.806	0.842	0.954

Table 4: Spearman's correlation coefficients for: S_{total}, S_{location}, $S_{\text{reasoning}}$, and $P_{\text{unrelated\_err}}$.

In summary, GPT-4.1 can extract relevant evidence and reasoning steps with considerable accuracy, leading to precise evaluation scores.

In addition, we replaced GPT-4.1 with Qwen3-32B and Gemini 2.5 Flash to independently re-evaluate the same 200 samples. The results further confirm that our evaluation framework is not dependent on any particular LLM and exhibits strong robustness.

Table 5: Model performance under Qwen3-32B evaluation (scores scaled by 100).

Models	**Avg.**	**RQD**	**DI**	**SG**	**MO**	**DHP**	**CF**	**IC**	**RCA**	**LE**
MLLM (Image Input)										
Gemini 2.5 Pro	19.7	15.0	23.2	44.7	13.2	31.4	7.8	17.3	17.8	12.8
GPT-5	21.2	12.6	11.8	33.5	15.4	26.6	16.0	27.1	27.0	6.4
Grok 4	4.2	0.0	1.3	20.5	2.6	5.5	1.2	3.9	2.6	0.8
Doubao-Seed-1.6-Thinking	11.6	5.1	6.9	28.0	7.7	14.3	10.5	15.6	10.8	4.5
Doubao-Seed-1.6	12.0	4.8	5.9	36.2	7.5	13.3	5.7	19.9	9.9	5.5
OCR + LLM (Text Input)										
Gemini 2.5 Pro	35.0	26.6	39.6	53.9	31.4	56.2	15.9	35.6	39.0	10.0
GPT-5	25.3	19.0	28.3	28.3	24.1	38.8	10.3	32.2	31.8	3.1
Claude Sonnet 4	6.3	5.7	2.4	12.4	4.2	8.3	2.8	9.5	6.6	4.4
Grok 4	22.6	10.7	8.9	40.6	14.1	31.3	11.1	22.6	33.6	7.3
Doubao-Seed-1.6-Thinking	17.4	11.0	14.9	31.9	9.5	26.3	9.2	20.5	21.1	4.7
Doubao-Seed-1.6	15.3	6.4	9.8	31.7	10.8	22.4	8.5	21.9	17.8	2.4

Table 6: Model performance under Gemini 2.5 Flash evaluation (scores scaled by 100).

Models	**Avg.**	**RQD**	**DI**	**SG**	**MO**	**DHP**	**CF**	**IC**	**RCA**	**LE**
MLLM (Image Input)										
Gemini 2.5 Pro	14.6	9.1	11.3	34.2	11.4	28.2	4.9	12.6	14.8	5.3
GPT-5	20.5	11.2	12.0	32.4	17.9	27.4	10.5	25.9	27.4	3.1
Grok 4	3.9	0.5	1.8	15.9	1.7	4.1	1.5	1.1	4.4	0.0
Doubao-Seed-1.6-Thinking	10.3	3.2	4.7	26.4	7.3	14.2	9.2	14.8	9.3	3.1
Doubao-Seed-1.6	10.9	5.3	3.9	34.0	6.1	15.6	6.1	17.9	8.0	4.7
OCR + LLM (Text Input)										
Gemini 2.5 Pro	30.1	20.0	30.6	47.8	27.5	47.8	11.7	30.2	36.6	5.9
GPT-5	23.5	15.3	21.8	26.5	23.1	37.7	7.1	31.9	31.6	2.7
Claude Sonnet 4	5.7	3.8	1.4	11.1	3.9	9.4	2.0	8.7	6.5	3.1
Grok 4	20.2	9.2	7.8	36.1	11.5	32.4	8.0	20.7	30.6	5.8
Doubao-Seed-1.6-Thinking	15.9	8.4	11.1	30.9	10.1	23.7	6.3	19.9	20.2	3.5
Doubao-Seed-1.6	14.0	4.8	8.2	29.1	11.4	23.9	6.4	21.2	16.0	0.8

F Hyperparameter Sensitivity Analysis

We conducted a sensitivity analysis of all 4 hyperparameters involved in scoring. We varied each independently and re-computed S_{total} across 11 proprietary model configurations. The results demonstrate that our evaluation metric exhibits strong robustness.

Table 7: Sensitivity under Image input: varying λ and μ (scores scaled by 100).

Model	λ=0.6	λ=0.8	λ=1.0	μ=0.85	μ=0.9	μ=0.95
GPT-5	19.3	19.2	19.0	18.5	19.2	19.9
Gemini 2.5 Pro	15.8	15.6	15.3	15.0	15.6	16.1
Doubao-1.6-Thinking	10.4	10.2	10.0	9.9	10.2	10.5
Doubao-1.6	10.1	9.9	9.8	9.7	9.9	10.2
Grok 4	4.0	4.0	3.9	3.9	4.0	4.1

Table 8: Sensitivity under Image input: varying γ and q (scores scaled by 100).

Model	γ=0.4	γ=0.6	γ=0.8	q=1.0	q=1.5	q=2.0
GPT-5	19.4	19.2	19.0	19.3	19.2	19.1
Gemini 2.5 Pro	15.7	15.6	15.5	15.6	15.6	15.5
Doubao-1.6-Thinking	10.4	10.2	10.0	10.3	10.2	10.1
Doubao-1.6	10.1	9.9	9.8	10.0	9.9	9.9
Grok 4	4.0	4.0	4.0	4.0	4.0	4.0

Table 9: Sensitivity under Text input: varying λ and μ (scores scaled by 100).

Model	λ=0.6	λ=0.8	λ=1.0	μ=0.85	μ=0.9	μ=0.95
Gemini 2.5 Pro	30.6	30.2	29.9	29.1	30.2	31.5
GPT-5	22.7	22.5	22.3	21.4	22.5	23.7
Grok 4	21.1	20.8	20.6	20.2	20.8	21.4
Doubao-1.6-Thinking	15.6	15.3	15.0	14.8	15.3	15.8
Doubao-1.6	14.1	13.9	13.7	13.6	13.9	14.3
Claude Sonnet 4	6.0	5.9	5.8	5.6	5.9	6.1

Table 10: Sensitivity under Text input: varying γ and q (scores scaled by 100).

Model	γ=0.4	γ=0.6	γ=0.8	q=1.0	q=1.5	q=2.0
Gemini 2.5 Pro	30.7	30.2	29.9	30.5	30.2	30.1
GPT-5	23.4	22.5	21.9	23.0	22.5	22.2
Grok 4	21.0	20.8	20.7	20.9	20.8	20.8
Doubao-1.6-Thinking	15.6	15.3	15.1	15.5	15.3	15.2
Doubao-1.6	14.2	13.9	13.7	14.1	13.9	13.8
Claude Sonnet 4	6.3	5.9	5.6	6.1	5.9	5.7