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

Anonymous authors
Paper under double-blind review

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

Scientific papers are crystallizations of human intelligence. 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, 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.***

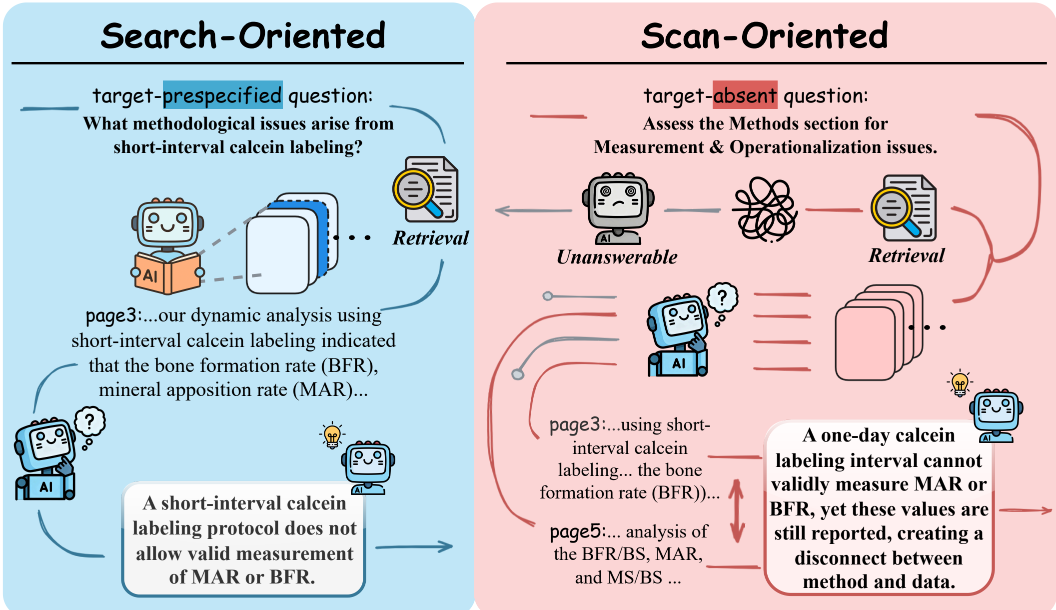


Figure 1: A comparison between search-oriented and scan-oriented task paradigms. Unlike the former, the scan-oriented paradigm provides no prespecified targets, requiring the model to actively scan the entire paper, construct a document-level evidence view.

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.

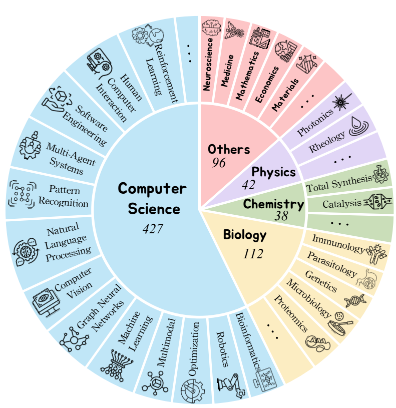
2.2 Document Understanding Benchmark

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 C.

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 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.

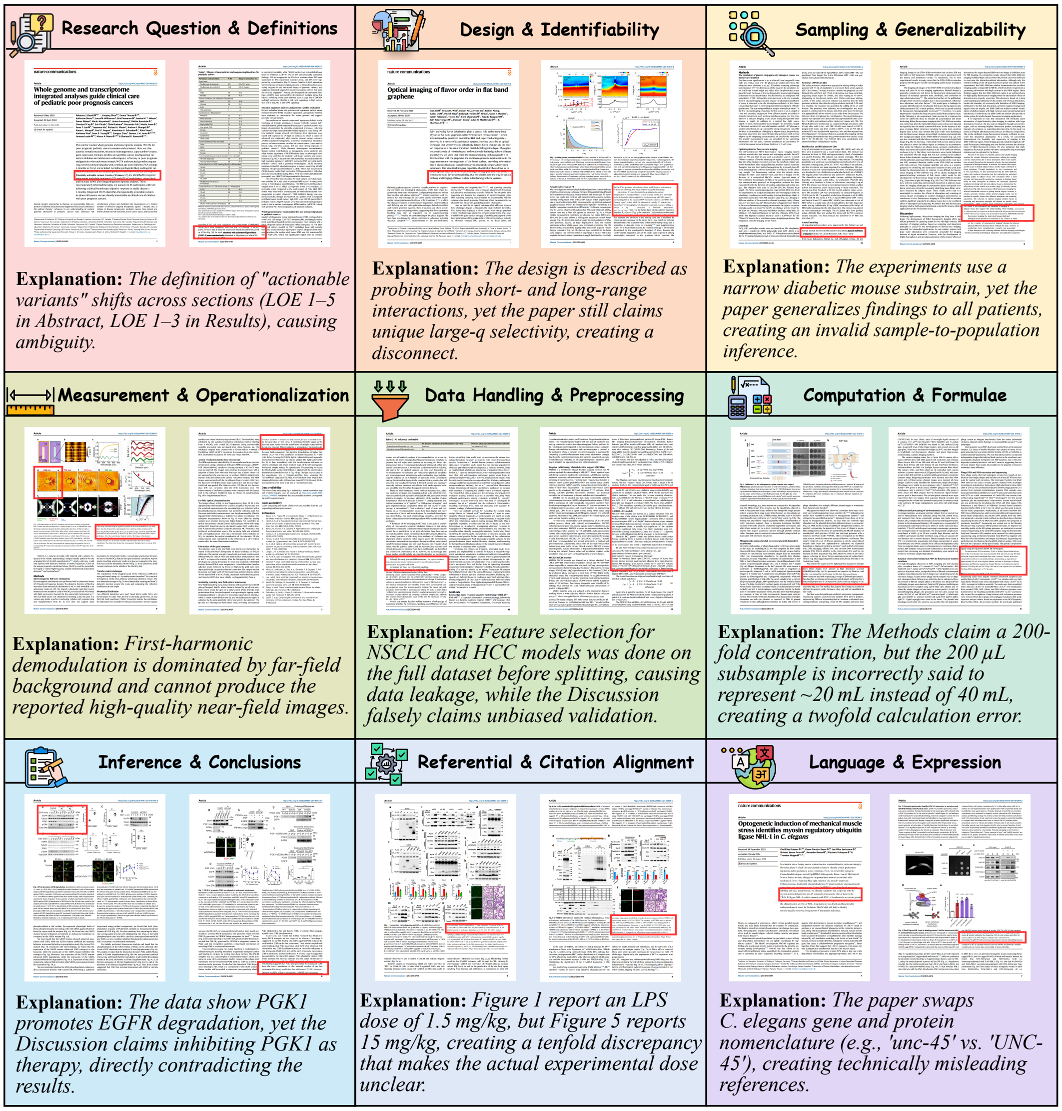


Figure 3: Sampled ScholScan examples with 9 error types, covering the whole process of scientific research, each requiring the model to perform thorough cross-source evidence-based reasoning.

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 D.

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 F.

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:

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

(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	15.6	**11.9**	**12.6**	**35.7**	12.3	**27.0**	4.6	14.7	15.2	**7.4**
GPT-5	**19.2**	10.1	9.7	28.2	**14.6**	26.6	**13.8**	**25.3**	**25.3**	6.9
Grok 4	4.0	0.0	1.9	16.7	3.2	7.4	0.7	1.9	3.6	0.0
Doubao-Seed-1.6-thinking	10.2	3.4	3.5	22.3	7.5	15.1	10.2	12.2	10.9	3.3
Doubao-Seed-1.6	9.9	3.0	4.4	29.2	4.9	15.0	6.3	17.9	8.0	3.9
Open-source LLMs										
Llama 4 Maverick	7.0	7.0	7.3	9.4	4.5	4.0	6.5	6.7	8.8	3.0
Gemma 3 27B	1.7	0.5	2.7	2.3	1.7	1.0	1.0	1.3	2.6	0.0
Mistral Small 3.1	3.3	0.1	2.0	2.0	1.5	0.1	1.0	2.2	8.6	1.0
Qwen2.5 VL 72B	0.1	0.0	0.7	0.0	0.0	0.0	0.0	0.0	0.2	0.0
OCR + LLM (Text Input)										
Proprietary LLMs										
Gemini 2.5 Pro	**30.3**	**21.5**	**34.2**	**44.3**	**27.6**	**56.6**	**10.3**	28.8	**35.6**	**8.1**
GPT-5	22.5	16.1	21.4	26.0	20.3	36.7	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	37.4	12.3	34.4	9.0	20.0	31.2	7.2
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:

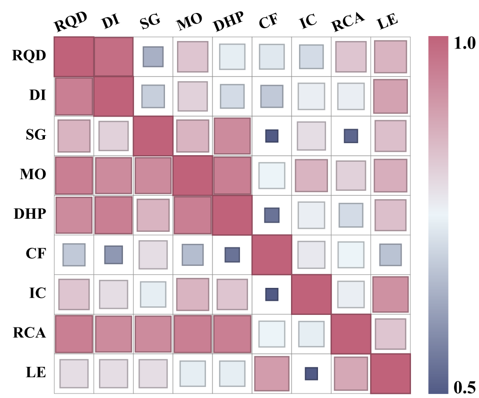


Figure 4: Spearman correlation matrix among the 9 error types.

(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.

(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.

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.

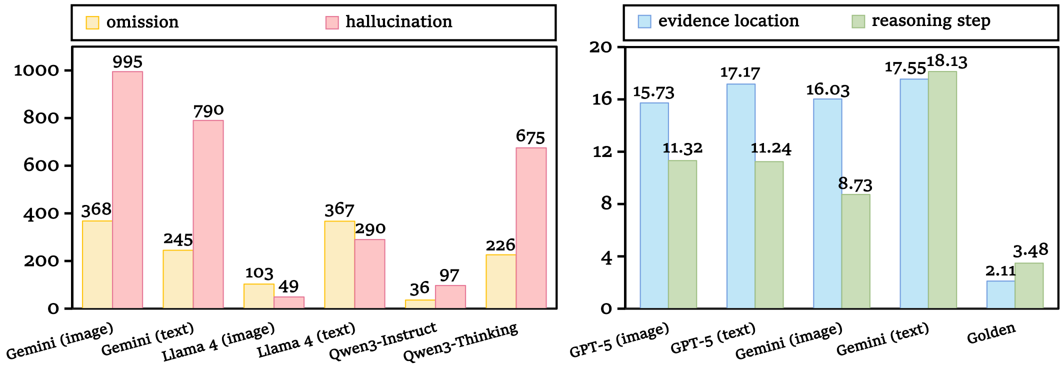


Figure 5: Left: Distribution of omission and hallucination errors. Right: Average reasoning steps and evidence locations involved in the answer generation, compared against the golden reference.

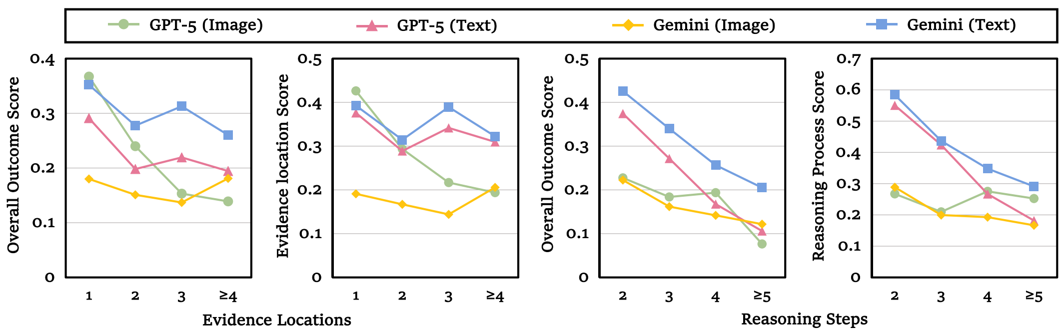


Figure 6: Performance trends across varying reasoning depths and evidence counts.

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.

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

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

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.

7 Reproducibility Statement

All results presented in this paper are fully reproducible. To facilitate verification and extension, we provide an anonymous repository (`https://anonymous.4open.science/r/ScholScan-6657/`) that contains the complete dataset, source code, and detailed documentation. The repository also includes step-by-step instructions and the exact hyperparameter configurations used in our experiments, ensuring that other researchers can replicate our findings with minimal effort.

The retrieval components in all retrieval-augmented generation (RAG) experiments were executed on a server equipped with 8 NVIDIA A40 GPUs.

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/.

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

Supplementary Material

Table of Contents in Appendix

A USE OF LLMS

Large language models (LLMs) were used solely to assist in language editing and stylistic refinement during manuscript preparation. All technical content, experiments, dataset construction, evaluation protocols, and analysis were conceived, implemented, and validated entirely by the authors. No LLMs were involved in the generation of benchmark data, research methodology design, or result interpretation. The use of LLMs did not influence the scientific conclusions of this paper.

B PROMPTS

B.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 */
]
```

B.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"
  }
]
```

B.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
}
```

B.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.
```

C EXAMPLES FROM EXISTING DATASETS

C.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: **Texts 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).

NO Multi-Modal Documents Context

Covered areas:

Focus Only On the Field of Mathematics

Cross-evidence Reasoning:
Focusing on solving mathematical problems requires integrating evidence such as mathematical formulas, question stem conditions, and chart data from different positions in the document.

Task Paradigm: **search**

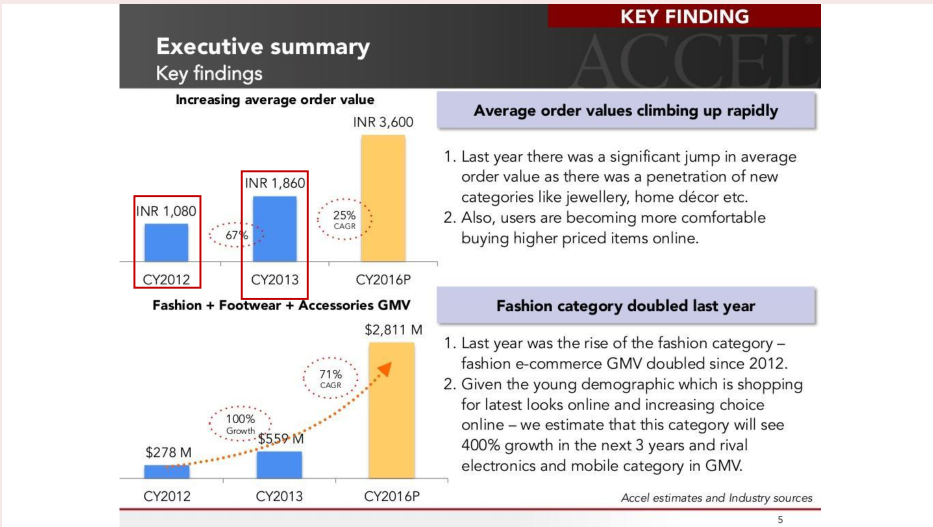
Search-oriented

C.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:
The documents cover core technical research fields such as visual question answering and machine reading comprehension, as well as industry application fields including education and scientific research, finance and commerce, and healthcare (with derivative adaptation to pathological slice analysis), and also involves derivative technical fields like retrieval-augmented generation.

Cross-evidence Reasoning:
Simple question types only require a single piece of evidence

Not Cross-evidence Reasoning

Task Paradigm: **search**

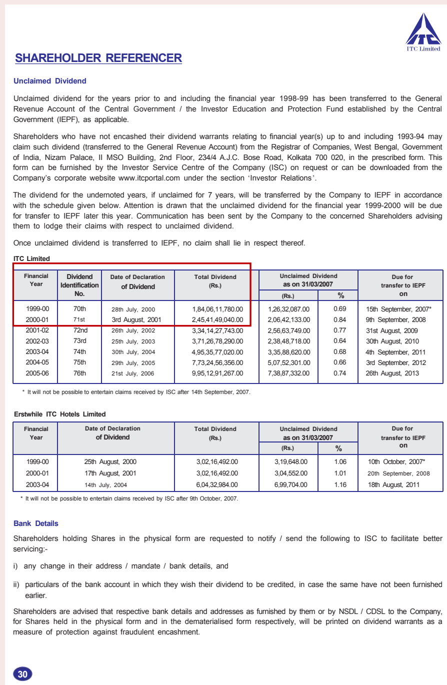
Search-oriented

C.3 EXAMPLE FROM MMLONGBENCH-DOC

One Example from MMLongBench-Doc

Question_ID:
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:
The documents cover 7 diverse fields such as scientific research reports, business financial reports, and technical manuals.

Cross-evidence Reasoning:
33% of the questions are cross-page questions, which require integrating different types of evidence such as texts, tables, and charts from multi-page documents

Task Paradigm: **search**

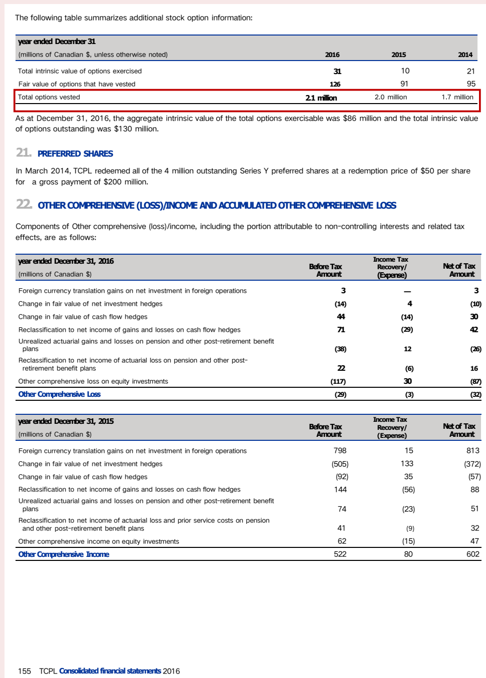
Search-oriented

C.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:
The document types of LongDocURL cover 8 major categories such as research reports, user manuals, and books.

Cross-evidence Reasoning:
Most questions require integrating evidence across chapters and elements

Task Paradigm: **search**

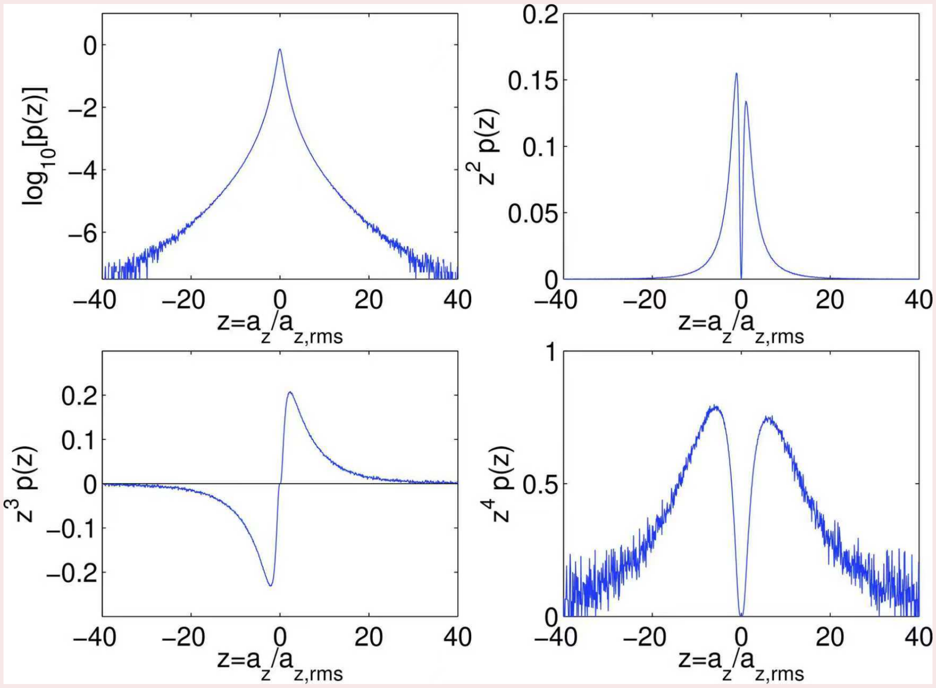
Search-oriented

C.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:
The document includes arXiv academic papers in various fields such as physics and mathematics.

Covers Few Areas

Cross-evidence Reasoning:
Only focus on a single element.

Not Cross-evidence Reasoning

Task Paradigm: **search**

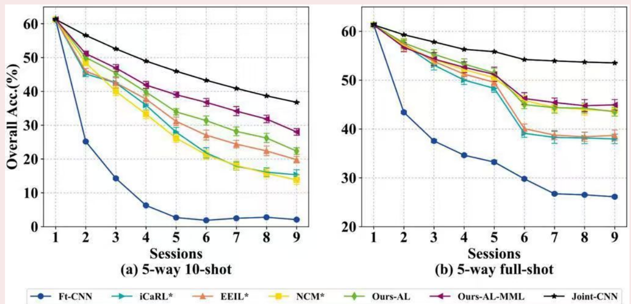
Search-oriented

C.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:
The document type consists of multi-type charts and graphs from 2323 papers in 8 disciplines, namely physics, computer science, mathematics, biology, chemistry, statistics, engineering, and economics, which are derived from the arXiv platform.

Cross-evidence Reasoning:
Only focus on a single element.

Not Cross-evidence Reasoning

Task Paradigm: **search**

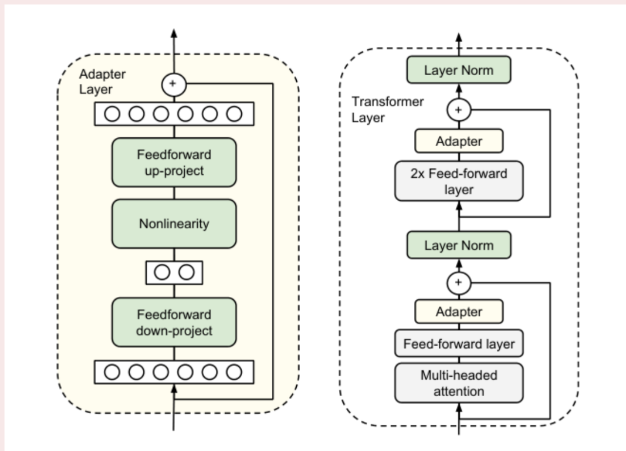
Search-oriented

C.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:
The document types cover two core categories: one is the AAAR-1.0 benchmark dataset documents, which are used to evaluate the research capabilities of LLMs and contain annotated data for 4 types of research tasks such as equation inference; the other is the documents related to the academic organization operation of the American Association for Aerosol Research (AAAR).

Covers Few Areas

Cross-evidence Reasoning:
It is necessary to integrate textual evidence across paragraphs and chapters.

Task Paradigm: **search**

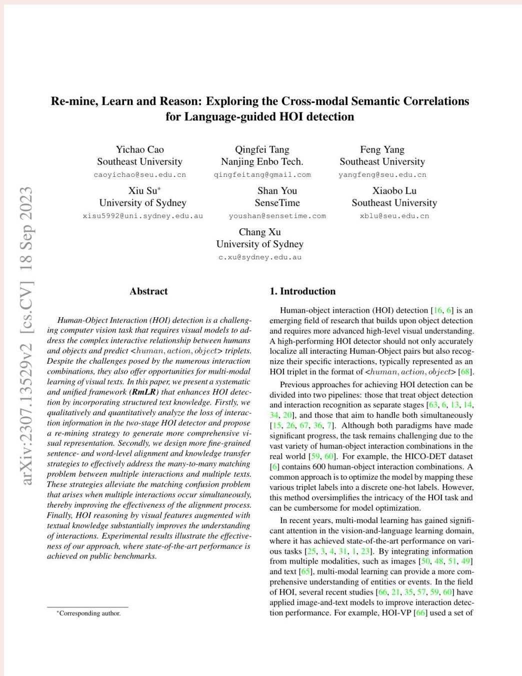
Search-oriented

C.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 $<\mathit{human}, \mathit{action}, \mathit{object}>$ 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 $<\mathit{human}, \mathit{action}, \mathit{object}>$ [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:
Its document type focuses on multimodal information fusion and clinical semantic understanding in medical scenarios.

Focus Only On the Field of Medicine

Cross-evidence Reasoning:
It is necessary to forcibly integrate medical imaging evidence (such as abnormal areas in CT images) with clinical report text evidence

Task Paradigm: **search**

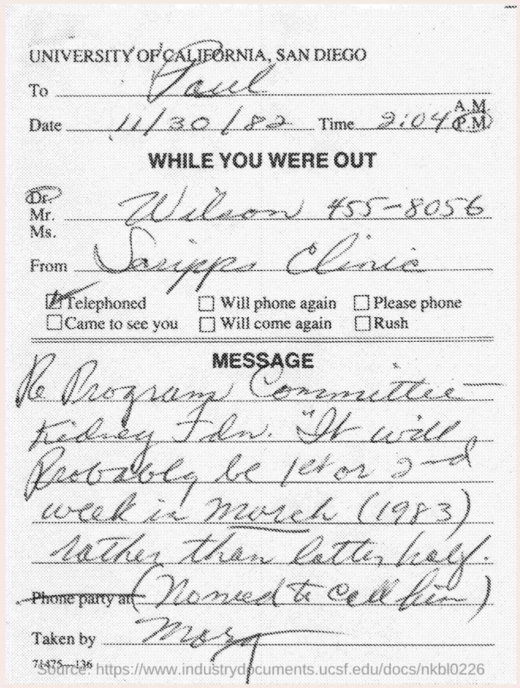
Search-oriented

C.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 Mg

71475-136

Source: https://www.industrydocuments.ucsf.edu/docs/nkbl0226

Covered areas:
Including invoices, resumes, academic papers, financial reports, manuals, etc. in formats such as scanned copies, PDFs, and screenshots.

Cross-evidence Reasoning:
Simple question types (such as ”invoice amount”) only require evidence from a single location, while complex question types (such as ”judging device compatibility based on parameter tables and explanatory texts across multiple pages of a manual”) require integrating evidence across elements and locations.

Not All Cross-evidence Reasoning

Task Paradigm: **search**

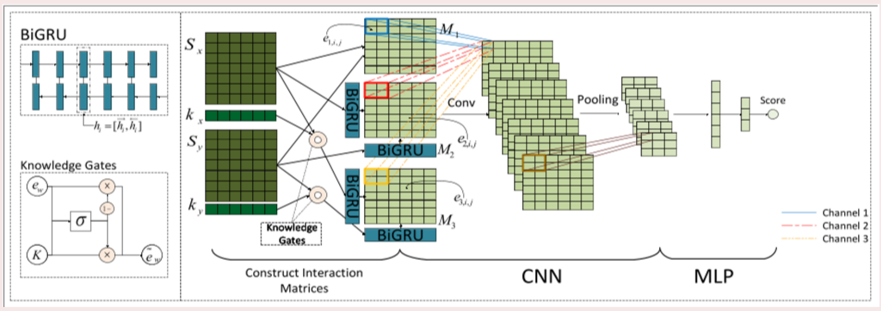
Search-oriented

C.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:
The document type of SPIQA originates from academic papers in fields such as computer science and physics.

Covered Few Areas

Cross-evidence Reasoning:
Only focus on a single element.

Not Cross-evidence Reasoning

Task Paradigm: **search**

Search-oriented

D Dataset Annotation and Construction

D.1 Human Annotator Guidelines

The defective academic papers in our dataset are curated from three primary sources: (1) We synthetically inject 9 types of errors into papers accepted at ICLR and Nature Communications. (2) For the papers rejected by ICLR, we identified the shortcomings in the papers based on the reviewers' comments and categorized them into 9 error types.(3) For accepted ICLR papers, we generate consistency-related errors by cross-referencing their content against 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.

D.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.

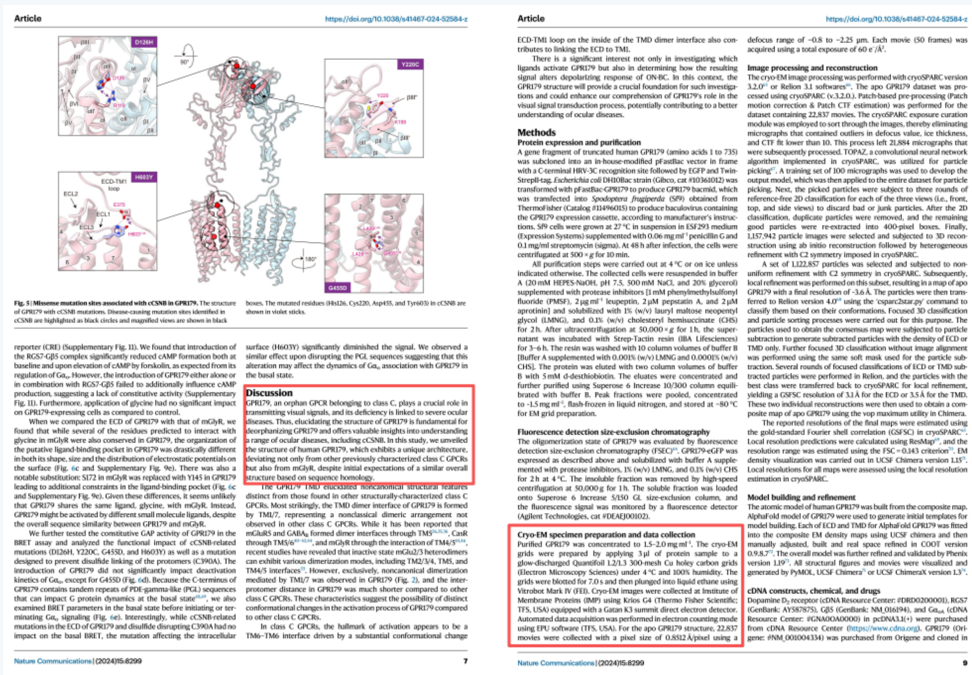
Of the remaining 1,800 instances, 1,541 (85.6%)underwent manual revision. The distribution of these modifications is as follows:

- **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.

D.3 Examples of Annotation

D.3.1 Case 1: Discard Directly

Example



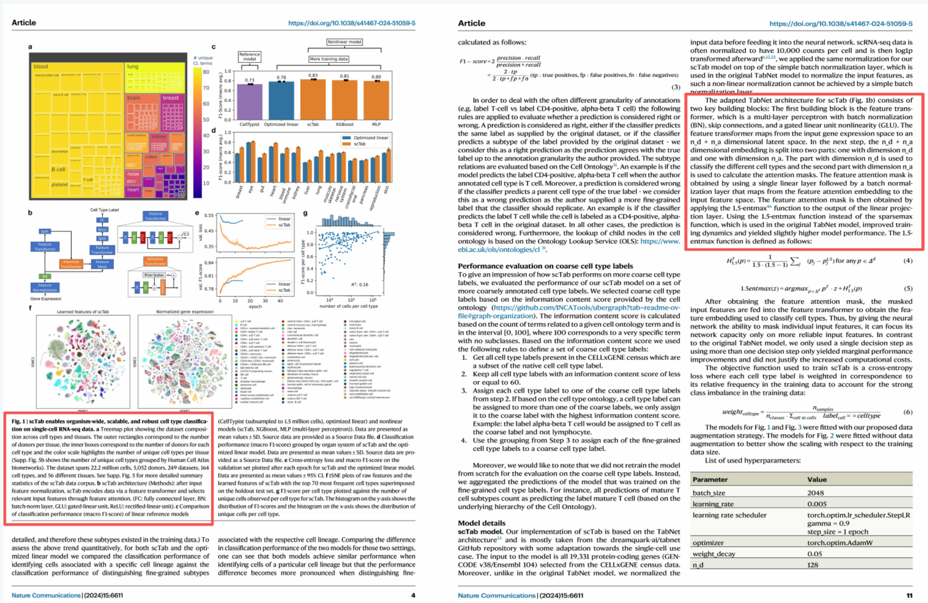
Question: Assess whether the conclusions drawn about the protein's functional state and therapeutic applicability are supported by the presented methods and results.
Explanation: Edits in the abstract and discussion claim the paper presents an active-state structure that reveals the activation mechanism 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.

D.3.2 Case 2: Modify Question

Example



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.

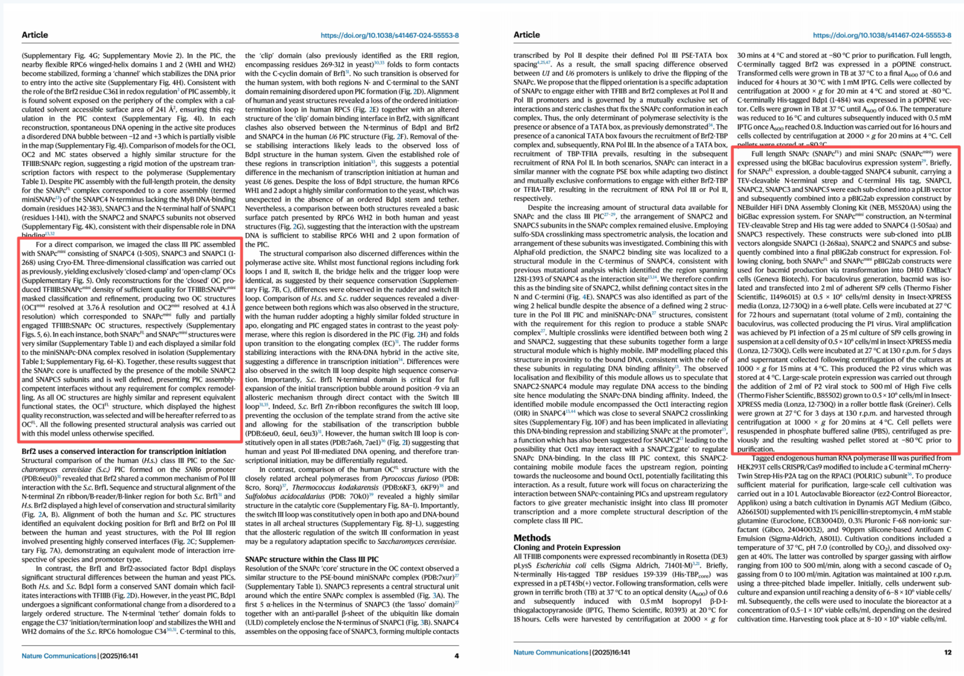
Decision: **Modify**

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.

D.3.3 CASE 3: MODIFY EXPLANATION

Example


Article https://doi.org/10.1038/s41467-024-55553-8

Nature Communications | (2025)16:141 4

Article https://doi.org/10.1038/s41467-024-55553-8

Nature Communications | (2025)16:141 12

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...

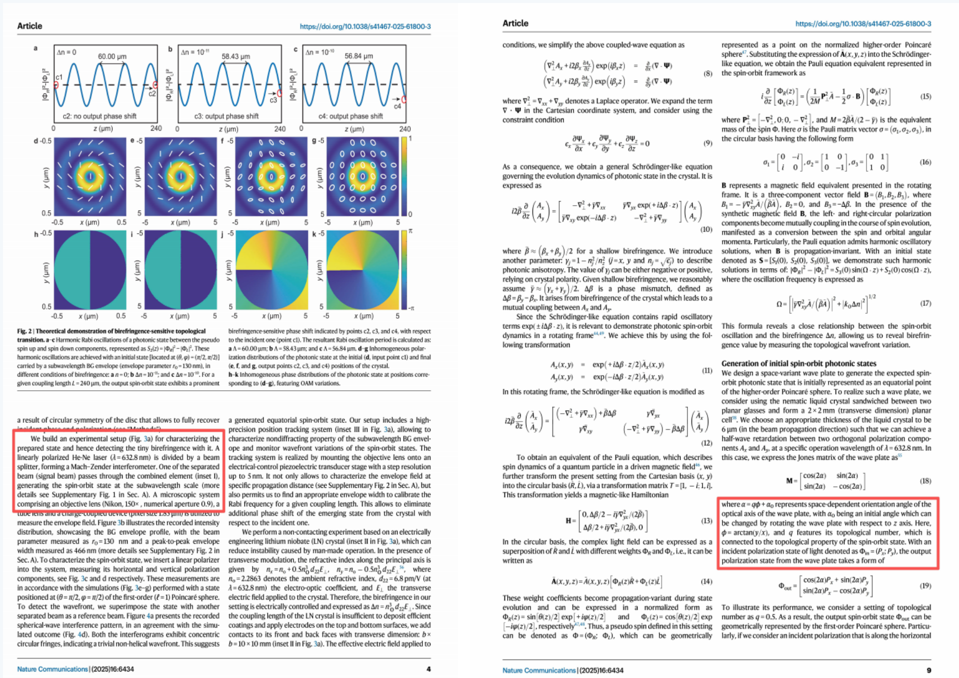
Decision: **Modify**

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

Analysis: 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.

D.3.4 CASE 4: MODIFY CATEGORY

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: MO (Measurement & Operationalization)

Before:
Error Type: MO (Measurement & Operationalization)

Decision: **Modify**

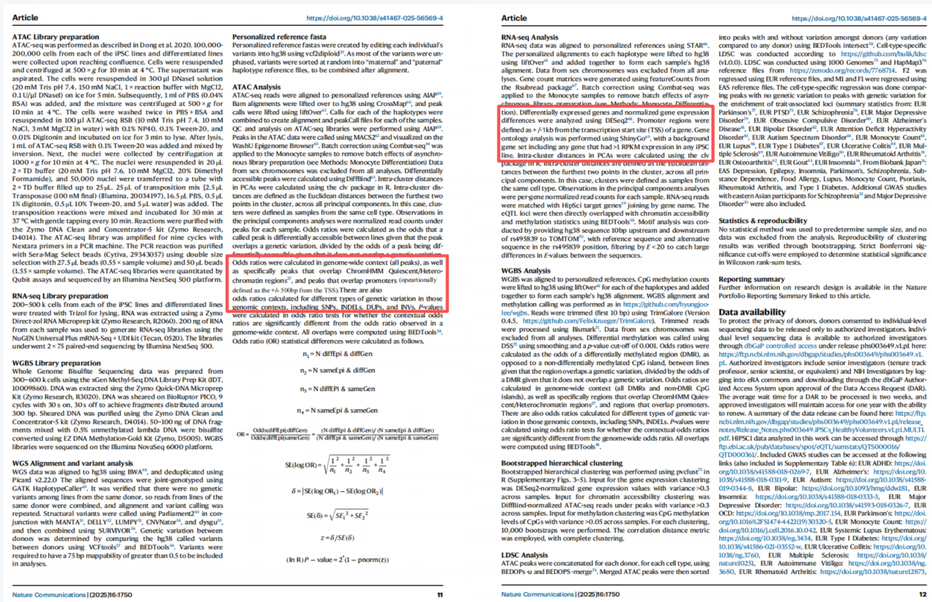
After:
Error Type: RCA (Referential and Citation Alignment)

Analysis: The introduced error systematically changes the laser wavelength used in the experiment to 532.0 nm. However, the calculation of a key physical quantity (birefringence) continues to use material constants (the electro-optic coefficient) that are only valid at the old wavelength of 632.8 nm. Because the optical properties of materials are wavelength-dependent, this systematic mismatch between experimental conditions and calculation parameters creates a significant contradiction in a core part of the paper. Compared to a Measurement & Operationalization (MO) error, this error is more accurately described as an internal inconsistency. Therefore, we are reclassifying this question from MO to RCA.

E COMMON FAILURE CASES OF MLLMS

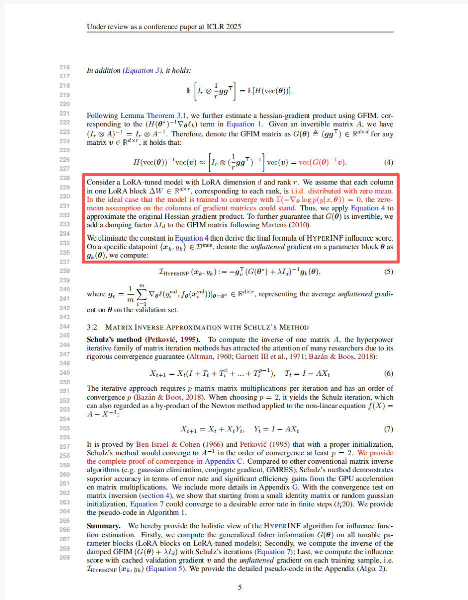
E.1 RQD (RESEARCH QUESTION & DEFINITIONS)

Example1: 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

Example2: 816


Under review as a conference paper at ICLR 2025

In addition (Equation 3), it holds:

$$\mathbb{E}\left[I_r \otimes \frac{1}{r}\boldsymbol{g}\boldsymbol{g}^\top\right] = \mathbb{E}[H(\text{vec}(\boldsymbol{\theta}))].$$

Following Lemma Theorem 3.1, we further estimate a hessian-gradient product using GFIM, corresponding to the $(H(\boldsymbol{\theta}^\star)^{-1}\nabla_{\boldsymbol{\theta}}\ell_k)$ term in Equation 1. Given an invertible matrix A, we have $(I_r \otimes A)^{-1} = I_r \otimes A^{-1}$. Therefore, denote the GFIM matrix as $G(\boldsymbol{\theta}) \triangleq (\boldsymbol{g}\boldsymbol{g}^\top) \in \mathbb{R}^{d\times d}$ for any matrix $\boldsymbol{v} \in \mathbb{R}^{d\times r}$, it holds that:

$$H(\text{vec}(\boldsymbol{\theta}))^{-1}\text{vec}(\boldsymbol{v}) \approx \left[I_r \otimes (\frac{1}{r}\boldsymbol{g}\boldsymbol{g}^\top)^{-1}\right]\text{vec}(\boldsymbol{v}) = \text{vec}(G(\boldsymbol{\theta})^{-1}\boldsymbol{v}). \quad (4)$$

Consider a LoRA-tuned model with LoRA dimension d and rank r. We assume that each column in one LoRA block $\Delta W \in \mathbb{R}^{d\times r}$, corresponding to each rank, is i.i.d. distributed with zero mean. In the ideal case that the model is trained to converge with $\mathbb{E}(-\nabla_{\boldsymbol{\theta}} \log p(y|x,\boldsymbol{\theta})) = 0$, the zero-mean assumption on the columns of gradient matrices could stand. Thus, we apply Equation 4 to approximate the original Hessian-gradient product. To further guarantee that $G(\boldsymbol{\theta})$ is invertible, we add a damping factor λI_d to the GFIM matrix following Martens (2010).

We eliminate the constant in Equation 4 then derive the final formula of HyperINF influence score. On a specific datapoint $\{\boldsymbol{x}_k, y_k\} \in \mathcal{D}^{\text{train}}$, denote the *unflattened* gradient on a parameter block $\boldsymbol{\theta}$ as $\boldsymbol{g}_k(\boldsymbol{\theta})$, we compute:

$$\mathcal{I}_{\text{HyperINF}}(\boldsymbol{x}_k, y_k) := -\boldsymbol{g}_v^\top (G(\boldsymbol{\theta}^\star) + \lambda I_d)^{-1}\boldsymbol{g}_k(\boldsymbol{\theta}), \quad (5)$$

where $\boldsymbol{g}_v = \frac{1}{m}\sum_{i=1}^{m}\nabla_{\boldsymbol{\theta}}\ell(y_i^{\text{val}}, f_{\boldsymbol{\theta}}(\boldsymbol{x}_i^{\text{val}}))|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\star} \in \mathbb{R}^{d\times r}$, representing the average *unflattened* gradient on $\boldsymbol{\theta}$ on the validation set.

3.2 Matrix Inverse Approximation with Schulz's Method

Schulz's method (Petković, 1995). To compute the inverse of one matrix A, the hyperpower iterative family of matrix iteration methods has attracted the attention of many researchers due to its rigorous convergence guarantee (Altman, 1960; Garnett III et al., 1971; Bazán & Boos, 2018):

$$X_{t+1} = X_t(I + T_t + T_t^2 + ... + T_t^{p-1}), \quad T_t = I - AX_t \quad (6)$$

The iterative approach requires p matrix-matrix multiplications per iteration and has an order of convergence p (Bazán & Boos, 2018). When choosing $p = 2$, it yields the Schulz iteration, which can also regarded as a by-product of the Newton method applied to the non-linear equation $f(X) = A - X^{-1}$:

$$X_{t+1} = X_t + X_tY_t, \quad Y_t = I - AX_t \quad (7)$$

It is proved by Ben-Israel & Cohen (1966) and Petković (1995) that with a proper initialization, Schulz's method would converge to A^{-1} in the order of convergence at least $p = 2$. We provide the complete proof of convergence in Appendix C. Compared to other conventional matrix inverse algorithms (e.g. gaussian elimination, conjugate gradient, GMRES), Schulz's method demonstrates superior accuracy in terms of error rate and significant efficiency gains from the GPU acceleration on matrix multiplications. We include more details in Appendix G. With the convergence test on matrix inversion (section 4), we show that starting from a small identity matrix or random gaussian initialization, Equation 7 could converge to a desirable error rate in finite steps (t¡20). We provide the pseudo-code in Algorithm 1.

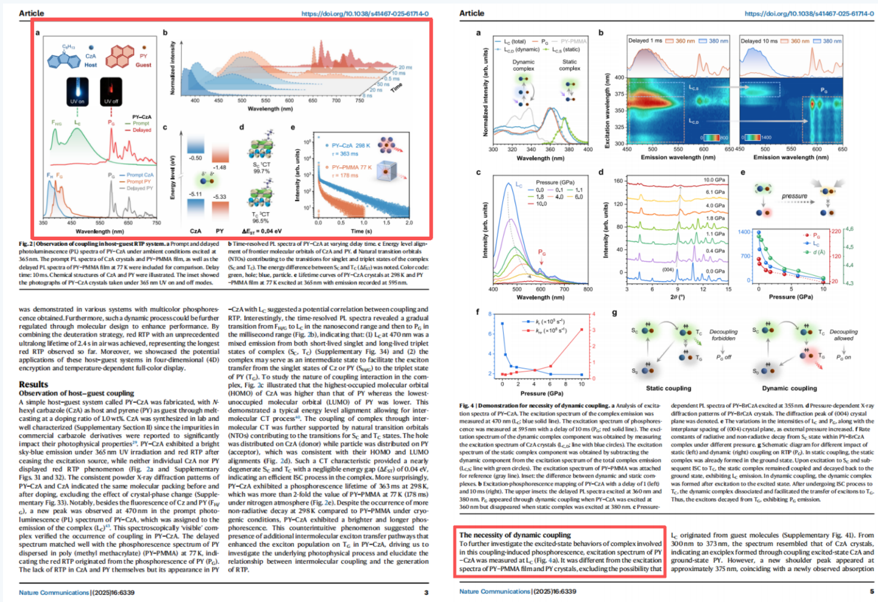
Summary. We hereby provide the holistic view of the HyperINF algorithm for influence function estimation. Firstly, we compute the generalized fisher information $G(\boldsymbol{\theta})$ on all tunable parameter blocks (LoRA blocks on LoRA-tuned models); Secondly, we compute the inverse of the damped GFIM $(G(\boldsymbol{\theta}) + \lambda I_d)$ with Schulz's iterations (Equation 7); Last, we compute the influence score with cached validation gradient $\boldsymbol{v}$ and the *unflattened* gradient on each training sample, i.e. $\mathcal{I}_{\text{HyperINF}}(\boldsymbol{x}_k, y_k)$ (Equation 5). We provide the detailed pseudo-code in the Appendix (Algo. 2).

5

Question: Scrutinize the Methods section for Research Question & Definitions issues.
Explanation: Lemma 3.1, which is a cornerstone of the paper's theoretical contribution for low-rank Hessian approximation, relies on a strong and insufficiently justified assumption. The lemma states: "Assume that each column of the sample gradient ... is independent and identically distributed random vector with zero mean under the distribution $p(y|x, \theta)$". The paper provides only a brief, hand-wavy justification (p.5, lines 230-232), suggesting it "could stand" in an "ideal case" of model convergence. These critical i.i.d. and zero-mean conditions are not rigorously established or empirically validated for the contexts in which the method is applied. This leaves a core hypothesis of the paper ambiguously defined and justified, which is an error of type Research Question & Definitions.
Error Type: RQD (Research Question & Definitions)
Type: Within-Sample

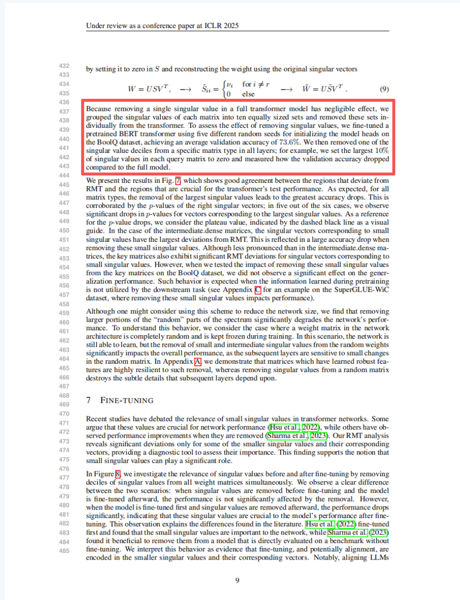
E.2 DI (DESIGN & IDENTIFIABILITY)

Example1: 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

Example2: 724


Under review as a conference paper at ICLR 2025

by setting it to zero in S and reconstructing the weight using the original singular vectors

$$W = USV^T, \quad \longrightarrow \quad \tilde{S}_{ii} = \begin{cases} \nu_i & \text{for } i \neq r \\ 0 & \text{else} \end{cases} \quad \longrightarrow \quad \tilde{W} = U\tilde{S}V^T . \tag{9}$$

Because removing a single singular value in a full transformer model has negligible effect, we grouped the singular values of each matrix into ten equally sized sets and removed these sets individually from the transformer. To assess the effect of removing singular values, we fine-tuned a pretrained BERT transformer using five different random seeds for initializing the model heads on the BoolQ dataset, achieving an average validation accuracy of 73.6%. We then removed one of the singular value deciles from a specific matrix type in all layers; for example, we set the largest 10% of singular values in each query matrix to zero and measured how the validation accuracy dropped compared to the full model.

We present the results in Fig. 7, which shows good agreement between the regions that deviate from RMT and the regions that are crucial for the transformer's test performance. As expected, for all matrix types, the removal of the largest singular values leads to the greatest accuracy drops. This is corroborated by the p-values of the right singular vectors; in five out of the six cases, we observe significant drops in p-values for vectors corresponding to the largest singular values. As a reference for the p-value drops, we consider the plateau value, indicated by the dashed black line as a visual guide. In the case of the intermediate.dense matrices, the singular vectors corresponding to small singular values have the largest deviations from RMT. This is reflected in a large accuracy drop when removing these small singular values. Although less pronounced than in the intermediate.dense matrices, the key matrices also exhibit significant RMT deviations for singular vectors corresponding to small singular values. However, when we tested the impact of removing these small singular values from the key matrices on the BoolQ dataset, we did not observe a significant effect on the generalization performance. Such behavior is expected when the information learned during pretraining is not utilized by the downstream task (see Appendix C for an example on the SuperGLUE-WiC dataset, where removing these small singular values impacts performance).

Although one might consider using this scheme to reduce the network size, we find that removing larger portions of the "random" parts of the spectrum significantly degrades the network's performance. To understand this behavior, we consider the case where a weight matrix in the network architecture is completely random and is kept frozen during training. In this scenario, the network is still able to learn, but the removal of small and intermediate singular values from the random weights significantly impacts the overall performance, as the subsequent layers are sensitive to small changes in the random matrix. In Appendix A, we demonstrate that matrices which have learned robust features are highly resilient to such removal, whereas removing singular values from a random matrix destroys the subtle details that subsequent layers depend upon.

7 Fine-tuning

Recent studies have debated the relevance of small singular values in transformer networks. Some argue that these values are crucial for network performance (Hsu et al., 2022), while others have observed performance improvements when they are removed (Sharma et al., 2023). Our RMT analysis reveals significant deviations only for some of the smaller singular values and their corresponding vectors, providing a diagnostic tool to assess their importance. This finding supports the notion that small singular values can play a significant role.

In Figure 8, we investigate the relevance of singular values before and after fine-tuning by removing deciles of singular values from all weight matrices simultaneously. We observe a clear difference between the two scenarios: when singular values are removed before fine-tuning and the model is fine-tuned afterward, the performance is not significantly affected by the removal. However, when the model is fine-tuned first and singular values are removed afterward, the performance drops significantly, indicating that these singular values are crucial to the model's performance after fine-tuning. This observation explains the differences found in the literature. Hsu et al. (2022) fine-tuned first and found that the small singular values are important to the network, while Sharma et al. (2023) found it beneficial to remove them from a model that is directly evaluated on a benchmark without fine-tuning. We interpret this behavior as evidence that fine-tuning, and potentially alignment, are encoded in the smaller singular values and their corresponding vectors. Notably, aligning LLMs

9

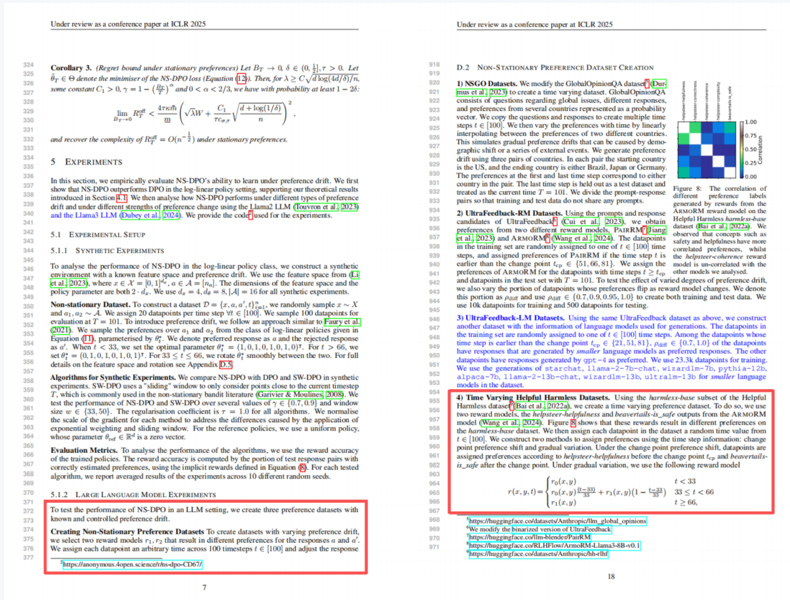
Question: Assess the Methods section for Design & Identifiability issues.
Explanation: A reviewer points out a flaw in the experimental design for the pruning experiments. The paper states on page 9, "We then removed one of the singular value deciles from a specific matrix type in all layers". The reviewer argues this "coarse intervention" constitutes a design flaw because by modifying all layers simultaneously, it becomes impossible to attribute performance changes to specific layers. This confounds the effects, making it difficult to identify where in the model the removed information was critical. This directly undermines the paper's stated goal of "locating information." The design choice violates the conditions for identifying layer-specific contributions, which is an error of type Design & Identifiability.
Error Type: DI (Design & Identifiability)
Type: Within-Sample

E.3 SG (SAMPLING & GENERALIZABILITY)

Example1: 1014

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

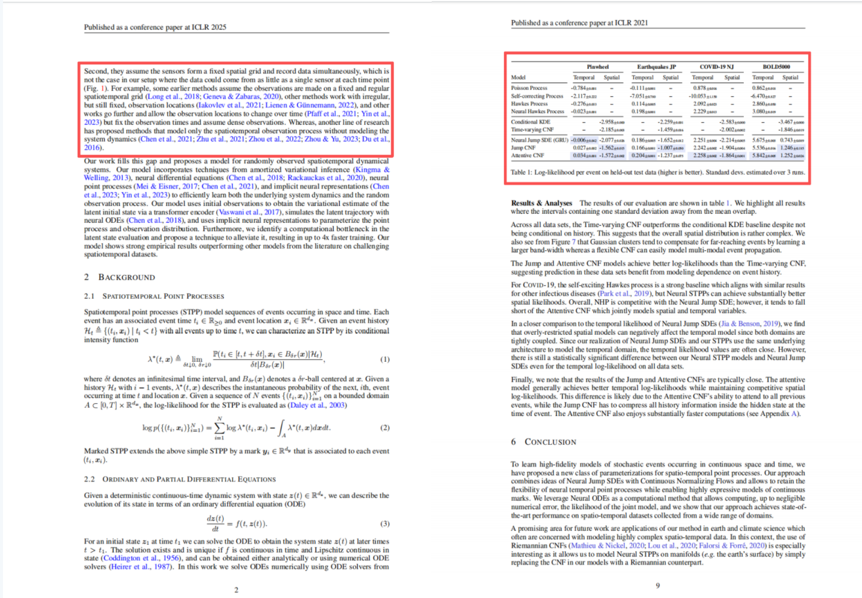
Example2: 935



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

E.4 RCA (Referential and Citation Alignment)

Example1: 0



Published as a conference paper at ICLR 2025

Second, they assume the sensors form a fixed spatial grid and record data simultaneously, which is not the case in our setup where the data could come from as little as a single sensor at each time point (Fig. 1). For example, some earlier methods assume the observations are made on a fixed and regular spatiotemporal grid (Long et al., 2018; Geneva & Zabaras, 2020), other methods work with irregular, but still fixed, observation locations (Iakovlev et al., 2021; Lienen & Günnemann, 2022), and other works go further and allow the observation locations to change over time (Pfaff et al., 2021; Yin et al., 2023) but fix the observation times and assume dense observations. Whereas, another line of research has proposed methods that model only the spatiotemporal observation process without modeling the system dynamics (Chen et al., 2021; Zhu et al., 2021; Zhou et al., 2022; Zhou & Yu, 2023; Du et al., 2016).

Our work fills this gap and proposes a model for randomly observed spatiotemporal dynamical systems. Our model incorporates techniques from amortized variational inference (Kingma & Welling, 2013), neural differential equations (Chen et al., 2018; Rackauckas et al., 2020), neural point processes (Mei & Eisner, 2017; Chen et al., 2021), and implicit neural representations (Chen et al., 2023; Yin et al., 2023) to efficiently learn both the underlying system dynamics and the random observation process. Our model uses initial observations to obtain the variational estimate of the latent initial state via a transformer encoder (Vaswani et al., 2017), simulates the latent trajectory with neural ODEs (Chen et al., 2018), and uses implicit neural representations to parameterize the point process and observation distribution. Furthermore, we identify a computational bottleneck in the latent state evaluation and propose a technique to alleviate it, resulting in up to 4x faster training. Our model shows strong empirical results outperforming other models from the literature on challenging spatiotemporal datasets.

2 Background

2.1 Spatiotemporal Point Processes

Spatiotemporal point processes (STPP) model sequences of events occurring in space and time. Each event has an associated event time $t_i \in \mathbb{R}_{\geq 0}$ and event location $\boldsymbol{x}_i \in \mathbb{R}^{d_x}$. Given an event history $\mathcal{H}_t \triangleq \{(t_i, \boldsymbol{x}_i) \mid t_i < t\}$ with all events up to time t, we can characterize an STPP by its conditional intensity function

$$\lambda^*(t, \boldsymbol{x}) \triangleq \lim_{\delta t \downarrow 0,\ \delta r \downarrow 0} \frac{\mathbb{P}(t_i \in [t, t+\delta t], \boldsymbol{x}_i \in B_{\delta r}(\boldsymbol{x}) | \mathcal{H}_t)}{\delta t |B_{\delta r}(\boldsymbol{x})|}, \tag{1}$$

where δt denotes an infinitesimal time interval, and $B_{\delta r}(\boldsymbol{x})$ denotes a δr-ball centered at $\boldsymbol{x}$. Given a history $\mathcal{H}_t$ with $i-1$ events, $\lambda^*(t, \boldsymbol{x})$ describes the instantaneous probability of the next, ith, event occurring at time t and location $\boldsymbol{x}$. Given a sequence of N events $\{(t_i, \boldsymbol{x}_i)\}_{i=1}^N$ on a bounded domain $A \subset [0, T] \times \mathbb{R}^{d_x}$, the log-likelihood for the STPP is evaluated as (Daley et al., 2003)

$$\log p(\{(t_i, \boldsymbol{x}_i)\}_{i=1}^N) = \sum_{i=1}^N \log \lambda^*(t_i, \boldsymbol{x}_i) - \int_A \lambda^*(t, \boldsymbol{x}) d\boldsymbol{x} dt. \tag{2}$$

Marked STPP extends the above simple STPP by a mark $\boldsymbol{y}_i \in \mathbb{R}^{d_y}$ that is associated to each event $(t_i, \boldsymbol{x}_i)$.

2.2 Ordinary and Partial Differential Equations

Given a deterministic continuous-time dynamic system with state $\boldsymbol{z}(t) \in \mathbb{R}^{d_z}$, we can describe the evolution of its state in terms of an ordinary differential equation (ODE)

$$\frac{d\boldsymbol{z}(t)}{dt} = f(t, \boldsymbol{z}(t)). \tag{3}$$

For an initial state $\boldsymbol{z}_1$ at time t_1 we can solve the ODE to obtain the system state $\boldsymbol{z}(t)$ at later times $t > t_1$. The solution exists and is unique if f is continuous in time and Lipschitz continuous in state (Coddington et al., 1956), and can be obtained either analytically or using numerical ODE solvers (Heirer et al., 1987). In this work we solve ODEs numerically using ODE solvers from

2

Published as a conference paper at ICLR 2021

	Pinwheel		Earthquakes JP		COVID-19 NJ		BOLD5000	
Model	Temporal	Spatial	Temporal	Spatial	Temporal	Spatial	Temporal	Spatial
Poisson Process	$-0.784_{\pm 0.001}$	–	$-0.111_{\pm 0.001}$	–	$0.878_{\pm 0.006}$	–	$0.862_{\pm 0.018}$	–
Self-correcting Process	$-2.117_{\pm 0.222}$	–	$-7.051_{\pm 0.780}$	–	$-10.053_{\pm 1.150}$	–	$-6.470_{\pm 0.827}$	–
Hawkes Process	$-0.276_{\pm 0.003}$	–	$0.114_{\pm 0.005}$	–	$2.092_{\pm 0.003}$	–	$2.860_{\pm 0.050}$	–
Neural Hawkes Process	$-0.023_{\pm 0.001}$	–	$0.198_{\pm 0.001}$	–	$2.229_{\pm 0.013}$	–	$3.080_{\pm 0.009}$	–
Conditional KDE	–	$-2.958_{\pm 0.000}$	–	$-2.259_{\pm 0.001}$	–	$-2.583_{\pm 0.000}$	–	$-3.467_{\pm 0.000}$
Time-varying CNF	–	$-2.185_{\pm 0.003}$	–	$-1.459_{\pm 0.016}$	–	$-2.002_{\pm 0.002}$	–	$-1.846_{\pm 0.019}$
Neural Jump SDE (GRU)	$-0.006_{\pm 0.002}$	$-2.077_{\pm 0.026}$	$0.186_{\pm 0.005}$	$-1.652_{\pm 0.012}$	$2.251_{\pm 0.006}$	$-2.214_{\pm 0.005}$	$5.675_{\pm 0.000}$	$0.743_{\pm 0.019}$
Jump CNF	$0.027_{\pm 0.002}$	$-1.562_{\pm 0.015}$	$0.166_{\pm 0.001}$	$-1.007_{\pm 0.050}$	$2.242_{\pm 0.002}$	$-1.904_{\pm 0.004}$	$5.536_{\pm 0.016}$	$1.246_{\pm 0.185}$
Attentive CNF	$0.034_{\pm 0.001}$	$-1.572_{\pm 0.002}$	$0.204_{\pm 0.001}$	$-1.237_{\pm 0.075}$	$2.258_{\pm 0.002}$	$-1.864_{\pm 0.001}$	$5.842_{\pm 0.005}$	$1.252_{\pm 0.026}$

Table 1: Log-likelihood per event on held-out test data (higher is better). Standard devs. estimated over 3 runs.

Results & Analyses The results of our evaluation are shown in table 1. We highlight all results where the intervals containing one standard deviation away from the mean overlap.

Across all data sets, the Time-varying CNF outperforms the conditional KDE baseline despite not being conditional on history. This suggests that the overall spatial distribution is rather complex. We also see from Figure 7 that Gaussian clusters tend to compensate for far-reaching events by learning a larger band-width whereas a flexible CNF can easily model multi-modal event propagation.

The Jump and Attentive CNF models achieve better log-likelihoods than the Time-varying CNF, suggesting prediction in these data sets benefit from modeling dependence on event history.

For COVID-19, the self-exciting Hawkes process is a strong baseline which aligns with similar results for other infectious diseases (Park et al., 2019), but Neural STPPs can achieve substantially better spatial likelihoods. Overall, NHP is competitive with the Neural Jump SDE; however, it tends to fall short of the Attentive CNF which jointly models spatial and temporal variables.

In a closer comparison to the temporal likelihood of Neural Jump SDEs (Jia & Benson, 2019), we find that overly-restricted spatial models can negatively affect the temporal model since both domains are tightly coupled. Since our realization of Neural Jump SDEs and our STPPs use the same underlying architecture to model the temporal domain, the temporal likelihood values are often close. However, there is still a statistically significant difference between our Neural STPP models and Neural Jump SDEs even for the temporal log-likelihood on all data sets.

Finally, we note that the results of the Jump and Attentive CNFs are typically close. The attentive model generally achieves better temporal log-likelihoods while maintaining competitive spatial log-likelihoods. This difference is likely due to the Attentive CNF's ability to attend to all previous events, while the Jump CNF has to compress all history information inside the hidden state at the time of event. The Attentive CNF also enjoys substantially faster computations (see Appendix A).

6 Conclusion

To learn high-fidelity models of stochastic events occurring in continuous space and time, we have proposed a new class of parameterizations for spatio-temporal point processes. Our approach combines ideas of Neural Jump SDEs with Continuous Normalizing Flows and allows to retain the flexibility of neural temporal point processes while enabling highly expressive models of continuous marks. We leverage Neural ODEs as a computational method that allows computing, up to negligible numerical error, the likelihood of the joint model, and we show that our approach achieves state-of-the-art performance on spatio-temporal datasets collected from a wide range of domains.

A promising area for future work are applications of our method in earth and climate science which often are concerned with modeling highly complex spatio-temporal data. In this context, the use of Riemannian CNFs (Mathieu & Nickel, 2020; Lou et al., 2020; Falorsi & Forré, 2020) is especially interesting as it allows us to model Neural STPPs on manifolds (*e.g.* the earth's surface) by simply replacing the CNF in our models with a Riemannian counterpart.

9

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

Example2: 570

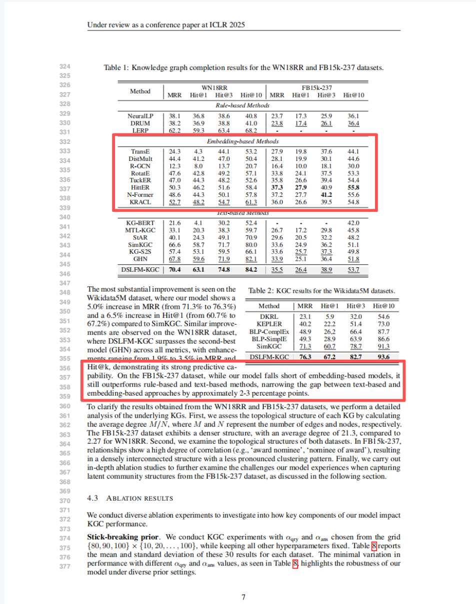


Table 1: Knowledge graph completion results for the WN18RR and FB15k-237 datasets.

Method	WN18RR				FB15k-237			
	MRR	Hit@1	Hit@3	Hit@10	MRR	Hit@1	Hit@3	Hit@10
Rule-based Methods								
NeuralLP	38.1	36.8	38.6	40.8	23.7	17.3	25.9	36.1
DRUM	38.2	36.9	38.8	41.0	23.8	17.4	26.1	36.4
LERP	62.2	59.3	63.4	68.2	-	-	-	-
Embedding-based Methods								
TransE	24.3	4.3	44.1	53.2	27.9	19.8	37.6	44.1
DistMult	44.4	41.2	47.0	50.4	28.1	19.9	30.1	44.6
R-GCN	12.3	8.0	13.7	20.7	16.4	10.0	18.1	30.0
RotatE	47.6	42.8	49.2	57.1	33.8	24.1	37.5	53.3
TuckER	47.0	44.3	48.2	52.6	35.8	26.6	39.4	54.4
HittER	50.3	46.2	51.6	58.4	**37.3**	**27.9**	40.9	**55.8**
N-Former	48.6	44.3	50.1	57.8	37.2	27.7	**41.2**	55.6
KRACL	52.7	48.2	54.7	61.3	36.0	26.6	39.5	54.8
Text-based Methods								
KG-BERT	21.6	4.1	30.2	52.4	-	-	-	42.0
MTL-KGC	33.1	20.3	38.3	59.7	26.7	17.2	29.8	45.8
StAR	40.1	24.3	49.1	70.9	29.6	20.5	32.2	48.2
SimKGC	66.6	58.7	71.7	80.0	33.6	24.9	36.2	51.1
KG-S2S	57.4	53.1	59.5	66.1	33.6	25.7	37.3	49.8
GHN	67.8	59.6	71.9	82.1	33.9	25.1	36.4	51.8
DSLFM-KGC	**70.4**	**63.1**	**74.8**	**84.2**	35.5	26.4	38.9	53.7

Table 2: KGC results for the Wikidata5M datasets.

Method	MRR	Hit@1	Hit@3	Hit@10
DKRL	23.1	5.9	32.0	54.6
KEPLER	40.2	22.2	51.4	73.0
BLP-ComplEx	48.9	26,2	66.4	87.7
BLP-SimplE	49.3	28.9	63.9	86.6
SimKGC	71.3	60.7	78.7	91.3
DSLFM-KGC	**76.3**	**67.2**	**82.7**	**93.6**

The most substantial improvement is seen on the Wikidata5M dataset, where our model shows a 5.0% increase in MRR (from 71.3% to 76.3%) and a 6.5% increase in Hit@1 (from 60.7% to 67.2%) compared to SimKGC. Similar improvements are observed on the WN18RR dataset, where DSLFM-KGC surpasses the second-best model (GHN) across all metrics, with enhancements ranging from 1.9% to 3.5% in MRR and Hit@k, demonstrating its strong predictive capability. On the FB15k-237 dataset, while our model falls short of embedding-based models, it still outperforms rule-based and text-based methods, narrowing the gap between text-based and embedding-based approaches by approximately 2-3 percentage points.

To clarify the results obtained from the WN18RR and FB15k-237 datasets, we perform a detailed analysis of the underlying KGs. First, we assess the topological structure of each KG by calculating the average degree M/N, where M and N represent the number of edges and nodes, respectively. The FB15k-237 dataset exhibits a denser structure, with an average degree of 21.3, compared to 2.27 for WN18RR. Second, we examine the topological structures of both datasets. In FB15k-237, relationships show a high degree of correlation (e.g., 'award nominee', 'nominee of award'), resulting in a densely interconnected structure with a less pronounced clustering pattern. Finally, we carry out in-depth ablation studies to further examine the challenges our model experiences when capturing latent community structures from the FB15k-237 dataset, as discussed in the following section.

4.3 Ablation results

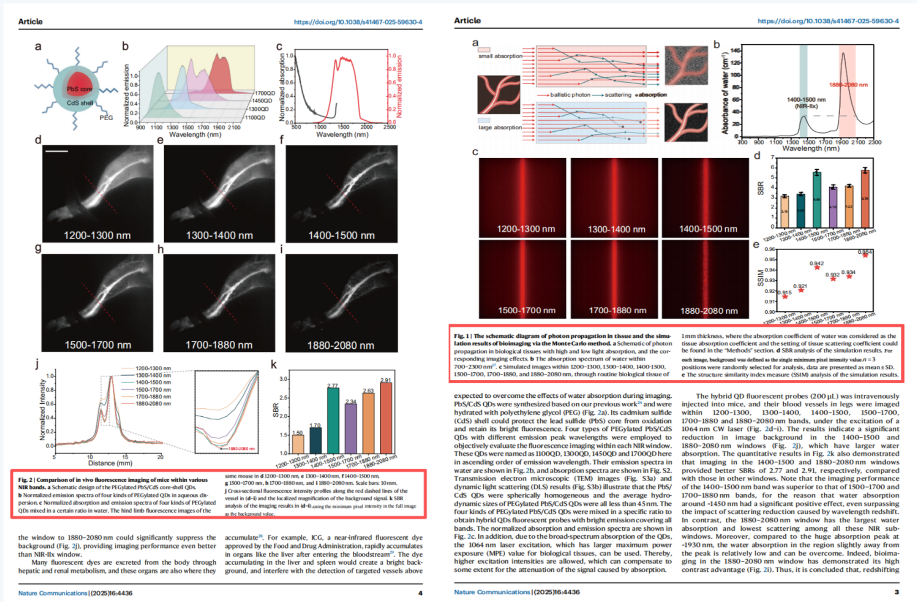
We conduct diverse ablation experiments to investigate into how key components of our model impact KGC performance.

Stick-breaking prior. We conduct KGC experiments with α_{qry} and α_{ans} chosen from the grid $\{80, 90, 100\} \times \{10, 20, \ldots, 100\}$, while keeping all other hyperparameters fixed. Table 8 reports the mean and standard deviation of these 30 results for each dataset. The minimal variation in performance with different α_{qry} and α_{ans} values, as seen in Table 8, highlights the robustness of our model under diverse prior settings.

Question: Evaluate the Results section for Internal Consistency problems.
Explanation: A reviewer points out that the paper's reported performance on FB15k-237 is lower than a state-of-the-art method. The paper's main text makes a claim that is directly contradicted by its own table. Specifically, the text states that on the FB15k-237 dataset, their model "still outperforms rule-based and text-based methods" (page 7, 'On the FB15k-237 dataset...methods'). However, Table 1 on the same page presents results for KRACL, a method listed under the "Text-based Methods" category, which achieves higher scores than the proposed model on both MRR (36.0 vs. 35.5) and Hit@1 (26.6 vs. 26.4). This discrepancy between the narrative claim and the tabular data constitutes a clear internal consistency error.
Error Type: RCA (Referential and Citation Alignment)
Type: Within-Sample

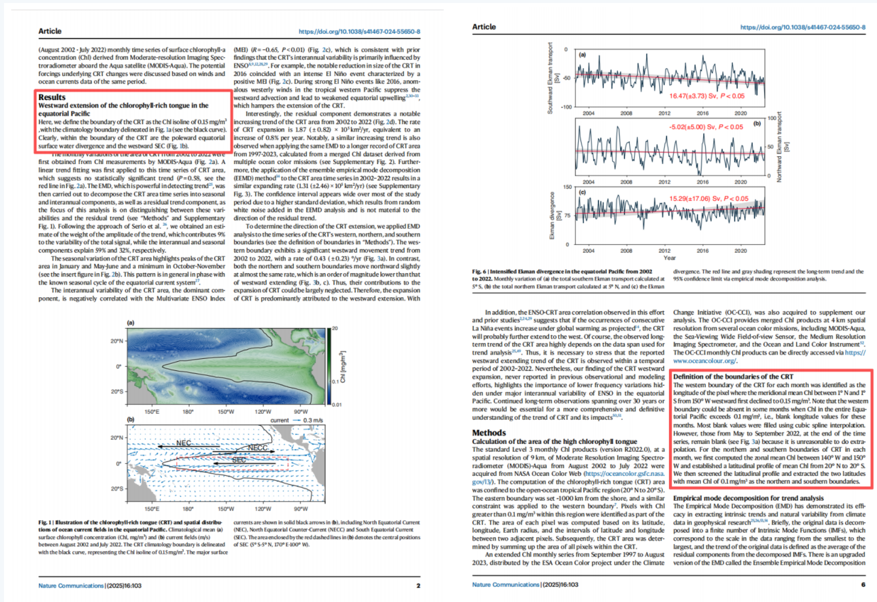
E.5 MO (Measurement & Operationalization)

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



Article

(August 2002 - July 2022) monthly time series of surface chlorophyll-a concentration (Chl) derived from Moderate-resolution Imaging Spectroradiometer aboard the Aqua satellite (MODIS-Aqua). The potential forcings underlying CRT changes were discussed based on winds and ocean currents data of the same period.

Results

Westward extension of the chlorophyll-rich tongue in the equatorial Pacific

Here, we define the boundary of the CRT as the Chl isoline of 0.15 mg/m³ ,with the climatology boundary delineated in Fig. 1a (see the black curve). Clearly, within the boundary of the CRT are the poleward equatorial surface water divergence and the westward SEC (Fig. 1b).

first obtained from Chl measurements by MODIS-Aqua (Fig. 2a). A linear trend fitting was first applied to this time series of CRT area, which suggests no statistically significant trend ($P = 0.58$, see the red line in Fig. 2a). The EMD, which is powerful in detecting trend[25], was then carried out to decompose the CRT area time series into seasonal and interannual components, as well as a residual trend component, as the focus of this analysis is on distinguishing between these variabilities and the residual trend (see "Methods" and Supplementary Fig. 1). Following the approach of Serio et al. [26], we obtained an estimate of the weight of the amplitude of the trend, which contributes 9% to the variability of the total signal, while the interannual and seasonal components explain 59% and 32%, respectively.

...

divergence. The red line and gray shading represent the long-term trend and the 95% confidence limit via empirical mode decomposition analysis.

Change Initiative (OC-CCI), was also acquired to supplement our analysis. The OC-CCI provides merged Chl products at 4 km spatial resolution from several ocean color missions, including MODIS-Aqua, the Sea-Viewing Wide Field-of-view Sensor, the Medium Resolution Imaging Spectrometer, and the Ocean and Land Color Instrument[52]. The OC-CCI monthly Chl products can be directly accessed via https://www.oceancolour.org/.

Definition of the boundaries of the CRT

The western boundary of the CRT for each month was identified as the longitude of the pixel where the meridional mean Chl between 1° N and 1° S from 150° W westward first declined to 0.15 mg/m³. Note that the western boundary could be absent in some months when Chl in the entire Equatorial Pacific exceeds 0.1 mg/m³, i.e., blank longitude values for these months. Most blank values were filled using cubic spline interpolation. However, those from May to September 2022, at the end of the time series, remain blank (see Fig. 3a) because it is unreasonable to do extrapolation. For the northern and southern boundaries of CRT in each month, we first computed the zonal mean Chl between 140° W and 150° W and established a latitudinal profile of mean Chl from 20° N to 20° S. We then screened the latitudinal profile and extracted the two latitudes with mean Chl of 0.1 mg/m³ as the northern and southern boundaries.

Empirical mode decomposition for trend analysis

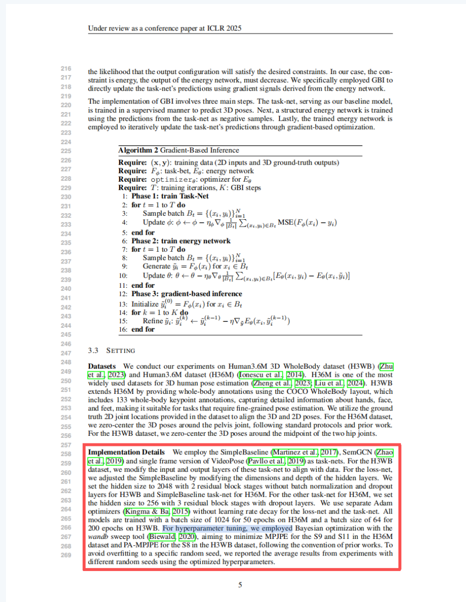
The Empirical Mode Decomposition (EMD) has demonstrated its efficacy in extracting intrinsic trends and natural variability from climate data in geophysical research[25,26,53,54]. Briefly, the original data is decomposed into a finite number of Intrinsic Mode Functions (IMFs), which correspond to the scale in the data ranging from the smallest to the largest, and the trend of the original data is defined as the average of the residual components from the decomposed IMFs. There is an upgraded version of the EMD called the Ensemble Empirical Mode Decomposition

6

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

E.6 DHP (Data Handling & Preprocessing)

Example1: 528



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
Phase 1: train Task-Net
for $t = 1$ to T **do**
  Sample batch $B_t = \{(x_i, y_i)\}_{i=1}^N$
  Update ϕ: $\phi \leftarrow \phi - \eta_\phi \nabla_\phi \frac{1}{|B_t|} \sum_{(x_i,y_i)\in B_t} \mathrm{MSE}(F_\phi(x_i) - y_i)$
end for
Phase 2: train energy network
for $t = 1$ to T **do**
  Sample batch $B_t = \{(x_i, y_i)\}_{i=1}^N$
  Generate $\tilde{y}_i = F_\phi(x_i)$ for $x_i \in B_t$
  Update θ: $\theta \leftarrow \theta - \eta_\theta \nabla_\theta \frac{1}{|B_t|} \sum_{(x_i,y_i)\in B_t} [E_\theta(x_i, y_i) - E_\theta(x_i, \tilde{y}_i)]$
end for
Phase 3: gradient-based inference
Initialize $\tilde{y}_i^{(0)} = F_\phi(x_i)$ for $x_i \in B_t$
for $k = 1$ to K **do**
  Refine $\tilde{y}_i$: $\tilde{y}_i^{(k)} \leftarrow \tilde{y}_i^{(k-1)} - \eta \nabla_{\tilde{y}} E_\theta(x_i, \tilde{y}_i^{(k-1)})$
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.

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 contexts16. 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

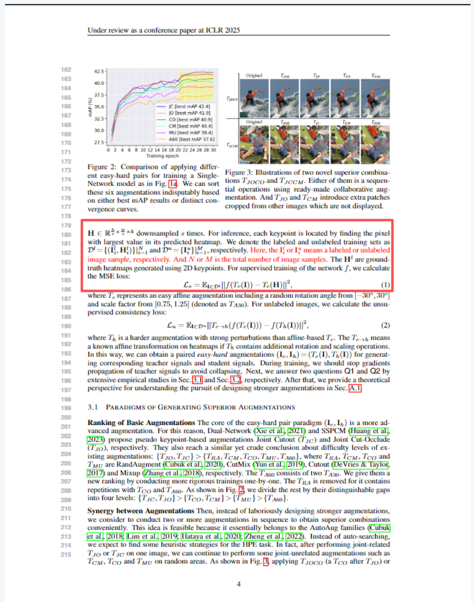
	New-hub trusts (N = 31) Median (IQR)	Established-hub trusts (N = 23) Median (IQR)	Non-hub trusts (N = 54) Median (IQR)
Population from a White ethnic background (%)	85 (72, 92)	89 (79, 93)	93 (87, 96)
Population who are male (%)	48 (48, 50)	50 (48, 51)	48 (48, 50)
Population who are aged 65+ years old (%)	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

E.7 CF (Computation & Formulae)

Example1: 350



Under review as a conference paper at ICLR 2025

Figure 2: Comparison of applying different easy-hard pairs for training a Single-Network model as in Fig. 1a. We can sort these six augmentations indisputably based on either best mAP results or distinct convergence curves.

Figure 3: Illustrations of two novel superior combinations T_{JOCO} and T_{JCCM}. Either of them is a sequential operations using ready-made collaborative augmentation. And T_{JO} and T_{CM} introduce extra patches cropped from other images which are not displayed.

$\mathbf{H} \in \mathbb{R}^{\frac{h}{s} \times \frac{w}{s} \times k}$ downsampled s times. For inference, each keypoint is located by finding the pixel with largest value in its predicted heatmap. We denote the labeled and unlabeled training sets as $\mathcal{D}^l = \{(\mathbf{I}_i^l, \mathbf{H}_i^l)\}|_{i=1}^N$ and $\mathcal{D}^u = \{\mathbf{I}_i^u\}|_{i=1}^M$, respectively. Here, the $\mathbf{I}_i^l$ or $\mathbf{I}_i^u$ means a labeled or unlabeled image sample, respectively. And N or M is the total number of image samples. The $\mathbf{H}^l$ are ground-truth heatmaps generated using 2D keypoints. For supervised training of the network f, we calculate the MSE loss:

$$\mathcal{L}_s = \mathbb{E}_{\mathbf{I} \in \mathcal{D}^l} ||f(T_e(\mathbf{I})) - T_e(\mathbf{H})||^2, \quad (1)$$

where T_e represents an easy affine augmentation including a random rotation angle from $[-30°, 30°]$ and scale factor from $[0.75, 1.25]$ (denoted as T_{A30}). For unlabeled images, we calculate the unsupervised consistency loss:

$$\mathcal{L}_u = \mathbb{E}_{\mathbf{I} \in \mathcal{D}^u} ||T_{e \to h}(f(T_e(\mathbf{I}))) - f(T_h(\mathbf{I}))||^2, \quad (2)$$

where T_h is a harder augmentation with strong perturbations than affine-based T_e. The $T_{e \to h}$ means a known affine transformation on heatmaps if T_h contains additional rotation and scaling operations. In this way, we can obtain a paired *easy-hard* augmentations $(\mathbf{I}_e, \mathbf{I}_h) = (T_e(\mathbf{I}), T_h(\mathbf{I}))$ for generating corresponding teacher signals and student signals. During training, we should stop gradients propagation of teacher signals to avoid collapsing. Next, we answer two questions Q1 and Q2 by extensive empirical studies in Sec. 3.1 and Sec. 3.2, respectively. After that, we provide a theoretical perspective for understanding the pursuit of designing stronger augmentations in Sec. A.1.

3.1 Paradigms of Generating Superior Augmentations

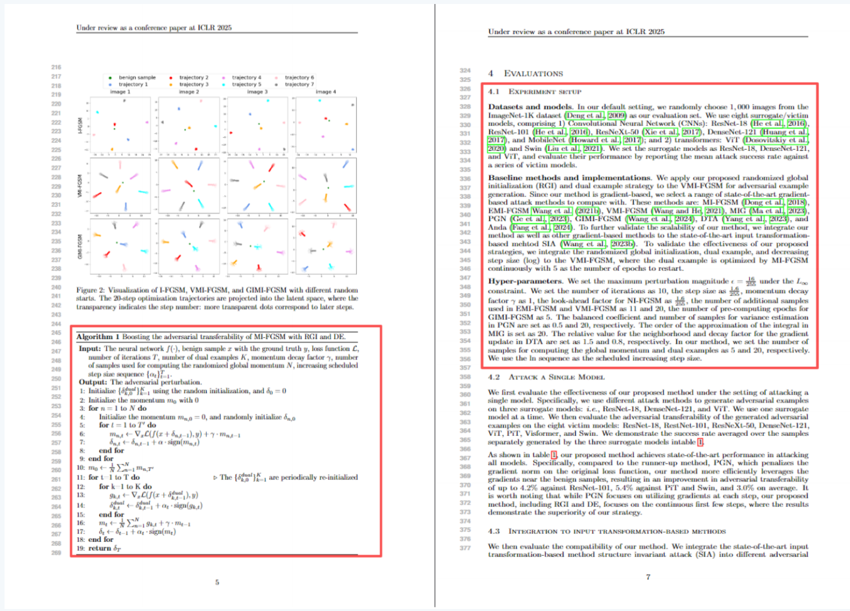
Ranking of Basic Augmentations The core of the easy-hard pair paradigm $(\mathbf{I}_e, \mathbf{I}_h)$ is a more advanced augmentation. For this reason, Dual-Network (Xie et al., 2021) and SSPCM (Huang et al., 2023) propose pseudo keypoint-based augmentations Joint Cutout (T_{JC}) and Joint Cut-Occlude (T_{JO}), respectively. They also reach a similar yet crude conclusion about difficulty levels of existing augmentations: $\{T_{JO}, T_{JC}\} > \{T_{RA}, T_{CM}, T_{CO}, T_{MU}, T_{A60}\}$, where T_{RA}, T_{CM}, T_{CO} and T_{MU} are RandAugment (Cubuk et al., 2020), CutMix (Yun et al., 2019), Cutout (DeVries & Taylor, 2017) and Mixup (Zhang et al., 2018), respectively. The T_{A60} consists of two T_{A30}. We give them a new ranking by conducting more rigorous trainings one-by-one. The T_{RA} is removed for it contains repetitions with T_{CO} and T_{A60}. As shown in Fig. 2, we divide the rest by their distinguishable gaps into four levels: $\{T_{JC}, T_{JO}\} > \{T_{CO}, T_{CM}\} > \{T_{MU}\} > \{T_{A60}\}$.

Synergy between Augmentations Then, instead of laboriously designing stronger augmentations, we consider to conduct two or more augmentations in sequence to obtain superior combinations conveniently. This idea is feasible because it essentially belongs to the AutoAug families (Cubuk et al., 2018; Lim et al., 2019; Hataya et al., 2020; Zheng et al., 2022). Instead of auto-searching, we expect to find some heuristic strategies for the HPE task. In fact, after performing joint-related T_{JO} or T_{JC} on one image, we can continue to perform some joint-unrelated augmentations such as T_{CM}, T_{CO} and T_{MU} on random areas. As shown in Fig. 3, applying T_{JOCO} (a T_{CO} after T_{JO}) or

4

Question: Check the Methods section for Computation & Formulae errors.
Explanation: The Problem Definition section introduces core variables for the mathematical setup, including N and M for the sizes of the labeled and unlabeled datasets. On page 4, line 183, these are defined with the sentence: "And N or M is the total number of image samples." This statement is ambiguous and fails to clearly define N and M individually. A reader cannot determine from this phrase that N is the number of labeled samples and M is the number of unlabeled samples. This notational ambiguity in the definition of variables that are fundamental to the subsequent equations and problem formulation constitutes a Computation & Formulae error, as key variables are left undefined or poorly defined.
Error Type: CF (Computation & Formulae)
Type: Within-Sample

Example2: 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: $\delta_{n,t} \leftarrow \delta_{n,t-1} + \alpha \cdot \mathrm{sign}(m_{n,t})$
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: $\delta_{k,t}^{dual} \leftarrow \delta_{k,t-1}^{dual} + \alpha_t \cdot \mathrm{sign}(g_{k,t})$
15: **end for**
16: $m_t \leftarrow \frac{1}{N}\sum_{n=1}^{N} g_{k,t} + \gamma \cdot m_{t-1}$
17: $\delta_t \leftarrow \delta_{t-1} + \alpha_t \cdot \mathrm{sign}(m_t)$
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

E.8 IC (INFERENCE & CONCLUSIONS)

Example1: 1017

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

Check for updates

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 establishes a new standard for clinical bioimaging, defining optimal imaging windows to immediately improve high-contrast fluorescence imaging in human surgery.

Fluorescence imaging, combining the advantages of its high temporal and spatial resolution and radiation-free properties, has a wide range of applications in the biomedical field[1]. In recent years, the second near-infrared window (NIR-II, 900–1880 nm) fluorescence imaging, has gained significant attention due to its low photon scattering and reduced biological autofluorescence[2–11]. In particular, based on the varying properties of scattering and absorption, NIR-II is divided into several specific imaging sub-windows, including NIR-IIa (1300–1400 nm), NIR-IIb (1500–1700 nm), and NIR-IIc (1700–1880 nm)[3,12,13]. Indeed, light propagation through biological media is an essential process in fluorescence bioimaging. This physical process is accompanied by absorption attenuation and scattering perturbations, which have conventionally been regarded as detrimental to the performance of bioimaging. As a result, a common approach improving imaging quality is to minimize scattering by increasing wavelength and avoid spectral regions with strong absorption peaks[14–16]. However, recent studies have challenged this conventional opinion by highlighting the beneficial role of absorption in enhancing the imaging contrast[13,17–19]. The improvement in contrast is attributed to the greater attenuation of background signals with longer optical paths in scattering media, thereby increasing the relative proportion of useful signals. This

[1]State Key Laboratory of Extreme Photonics and Instrumentations, Centre for Optical and Electromagnetic Research, College of Optical Science and Engineering, International Research Center for Advanced Photonics, Zhejiang University, Hangzhou, China. [2]Department of General Surgery, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China. [3]State Key Laboratory of Advanced Technology for Materials Synthesis and Processing, Wuhan University of Technology, Wuhan, China. [4]These authors contributed equally: Jiayi Li, Qiming Xia, Tianxiang Wu. e-mail: mxzhang@whut.edu.cn; qianjun@zju.edu.cn

Nature Communications | (2025)16:4436 1

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

imaging. Image in the 1700–2080 nm window had the best SBR and full width at half maximum (FWHM), which was in agreement with the theory and simulation results. To summarize, the in vitro experimental results strongly prove that the 1700–2080 nm window proposed in this paper has the best imaging performance in adipose tissue.

The imaging advantages of the 1700–2080 nm window in adipose tissue still exist for in vivo imaging applications. Morbid obesity is generally considered a risk factor for laparoscopic cholecystectomy because of increased operative time, morbidity, and conversion to open cholecystectomy[32,33]. The anatomical surface of the gallbladder triangle often becomes complex due to fat accumulation, inflammation, adhesions, and other factors[34]. This would pose a challenge for accurate identification of extrahepatic bile ducts during laparoscopic cholecystectomy (LC) in these patients, which can be greatly assisted by fluorescent cholangiography at present[35–39]. Therefore, it is quite important to visualize the deep structures buried in the adipose tissue. In the following in vivo experiment, fresh porcine fat is employed to cover the rabbit bile ducts to simulate fat accumulation and tissue thickening. Biliary fluorescence imaging in the 1700–2080 nm window was performed after the hybrid QDs fluorescent probes were injected retrogradely into the rabbit common bile duct. In the absence of adipose coverage, biliary structures including the cystic duct, common hepatic duct (CHD), and common bile duct (CBD) were illuminated, and the fluorescence images showed high resolution and remarkably low background signal in the 1700–2080 nm window (Fig. 5g). This was also consistent with the biliary structures shown in the bright-field image (Fig. S8). On this basis, ~2 mm thick fresh porcine adipose tissue was placed to cover the biliary region to simulate fat accumulation. Even under the influence of adipose tissue, accurate visualization of biliary structures still could be realized in the 1700–2080 nm window, as shown in Fig. 5j. The biliary tract structures were clearly delineated and consistent with the image that was not covered by the adipose tissue. In the simulated complex environment of a gallbladder triangle with fat adhesions and tissue thickening, the junction of the cystic duct and CHD remained distinctly visible in the 1700–2080 nm window with high contrast. This imaging capability can serve as a positive guidance for identifying safe transection locations on the cystic duct during the cholecystectomy. In contrast, it was difficult for fluorescence imaging in 900–1700 nm (Fig. S9) to clearly distinguish the genera-branching structures of bile ducts, which could be discriminated in the fluorescence image in the 1700–2080 nm window (Fig. 5j) and the bright field image (Fig. S8). The 1700–2080 nm window, with low scattering and increased absorption in adipose tissue, shows its imaging advantages in penetration depth and spatial resolution, which are critical for accurately identifying deep biliary structures obscured by adipose tissue. These results prove that it is meaningful to consider the effects of absorption and scattering of various tissues on imaging in complex biological environments. In conclusion, in vivo fluorescence imaging in 1700–2080 nm window exhibits significant superiority in adipose tissue due to the combined effect of absorption and scattering. This fluorescence imaging window is ready for immediate adoption in diverse clinical surgical applications, and these results provide a definitive guide for its use.

Discussion

Achieving high-contrast, deep-tissue imaging has long been a challenge. The development of NIR-II fluorescence imaging offers significant potential for capturing details in vivo with high contrast. The selection of imaging windows based on scattering and absorption principles is crucial for the development of fluorescence imaging, especially for biomedical applications. In past studies, regions with large water absorption were considered unsuitable for imaging because of signal attenuation. However, with the development of bright fluorophores and growing exploration of the positive effects of absorption, the region beyond 1880 nm is no longer a forbidden zone for NIR imaging. Our simulation results showed that 1880–2080 nm imaging exhibited high contrast when the photon count was sufficient to overcome the absorption-related attenuation. Although only the emission trailing of the off-peak fluorescent probe was utilized in the experiments, the 1880–2080 nm window demonstrated an excellent imaging quality, comparable to NIR-IIx, which has been recognized as a promising sub-window with high contrast in the NIR-II region. These findings suggest that the 1880–2080 nm region is an optimal choice for high-quality fluorescence imaging when the attenuation effects of absorption are properly overcome. This discovery, based on a deep understanding and utilization of the positive role of tissue absorption, breaks the boundary of previously held limitation of NIR-II imaging. Attributed to the fact that absorption can better suppress background signals from the depths, the 1880–2080 nm window enables visualization of shallow targets with high contrast under strong background interference. Furthermore, the 1880–2080 nm window also provides a viable option for multi-channel fluorescence imaging applications.

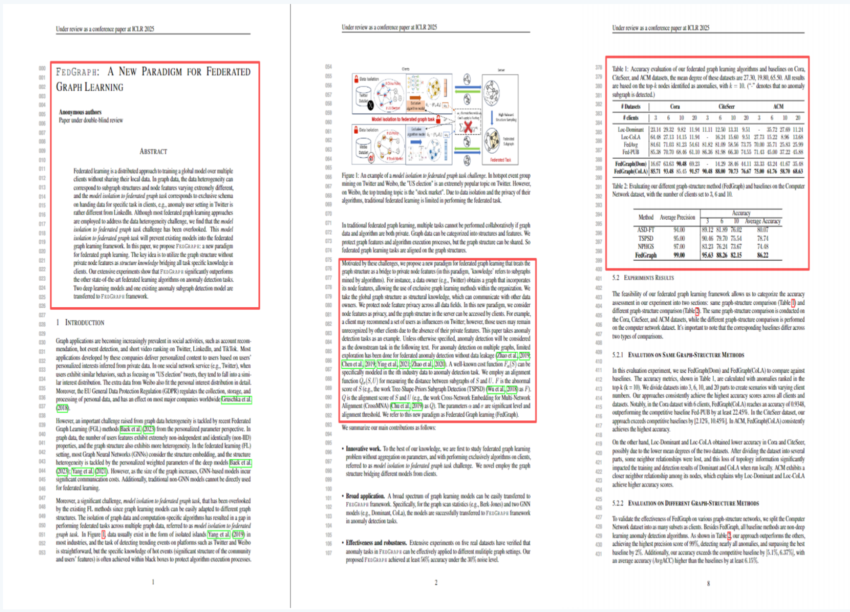
It is important to note that absorption will inevitably cause attenuation of signal intensity when imaging in windows with large absorption. However, the loss can be mitigated through several strategies, such as selecting brighter fluorescent probes, increasing the intensity of excitation, or extending detection time. In this study, we chose an off-peak QDs fluorescent probe as an effective compromise. The probe exhibited strong fluorescence intensity, sufficient to overcome the absorption attenuation. As a result, we successfully demonstrated the feasibility and superiority of imaging in the 1880–2080 nm region. These findings provide valuable insights for off-peak fluorescence bioimaging and may further promote the development of NIR-II fluorescent probes. We also anticipate that high-brightness fluorescent probes with more compatible long-wavelength emission can be exploited in the future.

The aim of this study is to identify superior and suitable imaging windows for complex biological environments, utilizing the imaging contrast enhancement due to tissue absorption. Most recent studies have primarily focused on the absorption of water, the most abundant component of biological tissues. In contrast, our research has expanded the scope to include adipose tissue, another major constituent of living organisms. Through simulations, in vitro experiments and in vivo experiments, we have demonstrated that the 1700–2080 nm region with moderate absorption and low scattering has optimized imaging quality in adipose tissue. This conclusion offers a more accurate and tissue-specific perspective compared to findings based solely on water absorption. In addition, we conducted additional simulation involving both skin and adipose tissue to demonstrate the effectiveness of this window in a broader range of clinically relevant applications (Fig. S10). In recent years, NIR-II fluorescence imaging has achieved remarkable success in guiding clinical surgery. However, these results are conclusive, showing that optimal imaging results can now be obtained in specific human tissues with different absorption and scattering parameters. The selection of optimal imaging windows based on scattering and absorption principles will depend on specific tissue environments, which requires more meticulous exploration and research. Our results of the exploration and validation of imaging in adipose tissue also support this understanding. These findings are expected to promote further development and application of NIR-II fluorescence imaging. Additionally, beyond fluorescence imaging, the interaction between light and biological tissues should be re-evaluated in the context of other advanced imaging techniques, such as the effect of absorption on the contrast of hyperspectral imaging[40], despite the inherent physical differences between these methods. We look forward to conducting more comprehensive research and assessment of how absorption and scattering influence different imaging technologies, thereby promoting technological integration and application expansion.

Nature Communications | (2025)16:4436 9

Question: Assess the Discussion section for Inference & Conclusions issues.
Explanation: The paper's evidence is based entirely on preclinical models (simulations, mice, rabbits, ex vivo porcine tissue). The edits in the Abstract and Discussion make strong, unhedged claims about setting a "new standard for clinical bioimaging" that is "ready for immediate adoption" in "human surgery." These conclusions are a gross overstatement, as the preclinical data do not support such direct and immediate claims of clinical efficacy and adoption.
Error Type: IC (Inference & Conclusions)
Type: Witin-Generate

Example2: 875



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

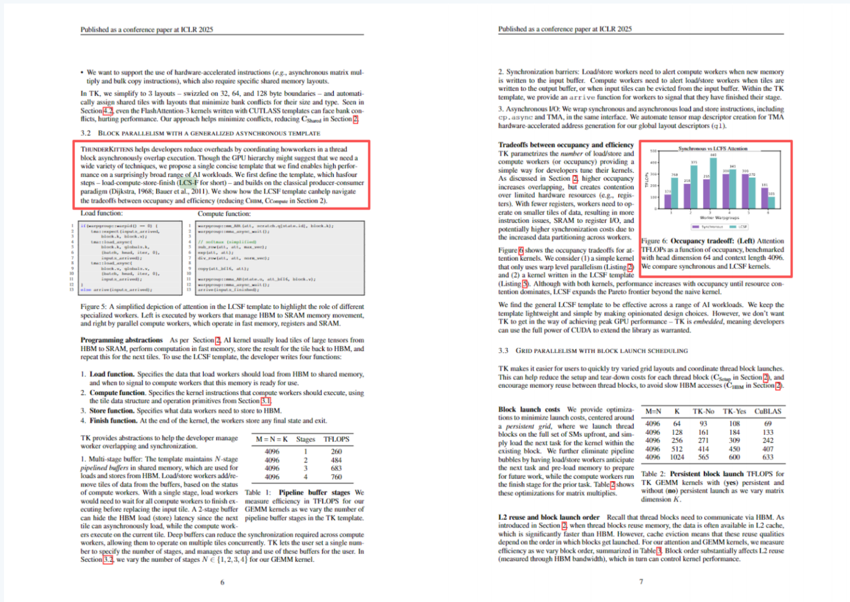
Explanation: The paper's claims of generality are not supported by its evidence. The title and abstract introduce "FedGraph: A New Paradigm for Federated Graph Learning" (page 1), suggesting a broadly applicable framework. However, the methodology is heavily tailored to, and the experiments are exclusively 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

E.9 LE (Language & Expression)

Example1: 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

Example2: 293

Under review as a conference paper at ICLR 2025

Regularized Optimal Transport for Single-Cell Temporal Trajectory Analysis

Anonymous authors
Paper under double-blind review

Abstract

The temporal relationship between different cellular states and lineages is only partially understood and has major significance for cell differentiation and cancer progression. However, two pain points persist and limit learning-based solutions: (*a*) lack of real datasets and standardized benchmark for early cell developments; (*b*) the complicated transcriptional data fail classic temporal analyses. We integrate `Mouse-RGC`, a large-scale mouse retinal ganglion cell dataset with annotations for 9 time stages and 30, 000 gene expressions. Existing approaches show a limited generalization of our datasets. To tackle the modeling bottleneck, we then translate this fundamental biology problem into a machine learning formulation, *i.e.*, *temporal trajectory analysis*. An innovative regularized optimal transport algorithm, `TAROT`, is proposed to fill in the research gap, consisting of (1) customized masked autoencoder to extract high-quality cell representations; (2) cost function regularization through biology priors for distribution transports; (3) continuous temporal trajectory optimization based on discrete matched time stages. Extensive empirical investigations demonstrate that our framework produces superior cell lineages and pseudotime, compared to existing approaches on `Mouse-RGC` and another two public benchmarks. Moreover, `TAROT` is capable of identifying biologically meaningful gene sets along with the developmental trajectory, and its simulated gene knockout results echo the findings in physical wet lab validation.

1 Introduction

Since first introduced in 2009, large-scale single-cell RNA sequencing (scRNA-seq) has presented enormous opportunities for researchers in various research fields (Patel et al., 2014; Satija et al., 2015; Tirosh et al., 2016). It helps reveal detailed information on transcriptional patterns in different cell and tissue types as well as disease models (Elmentaite et al., 2022; Jagadeesh et al., 2022). Equipped with scRNA-seq, we are able to discover significant heterogeneities that would never be found with bulk analysis within the cell population, which contributes to understanding biology questions with higher cellular resolution. The fast-advancing technology and increased recognition of different cell subtypes also naturally lead us to ask: ① *How and when are the cell subtypes established?* ② *Could we predict the developmental trajectory of each cell and predict the cell "fate" based on current status?* ③ *And could we find the key regulator that controls this type of establishment?* Answers to those questions are important for cell differentiation research in developmental biology (Rizvi et al., 2017; Han et al., 2018; Gulati et al., 2020) and can provide promising pipelines to demystify the cellular response during disease progression (Zhang et al., 2021; Jia et al., 2022). In the past decade, although a great amount of effort (Trapnell et al., 2014; Qiu et al., 2017; Ji & Ji, 2016; Street et al., 2018; Cao et al., 2019) has been put into developing trajectory inference methods using single-cell sequencing data, it remains extremely challenging. This is because, with current technologies, we can not trace the same population of cells over developmental time. It only allows us to collect the transcriptional information of cells for a specific time point as a

Figure 1: Demo Cell Temporal Trajectories from Time $0 \to n$. Different colors indicate cells from different time stages.

1

Under review as a conference paper at ICLR 2025

Figure 3: *(Left)*: Clustering `Mouse-RGC` to 56 kinds of cell types and projecting them into a 2D space via UMAP; *(Right)*: Decomposing the clustering results by their time stage labels. Zoom-in for better reliability.

Data Collection. For the `Mouse-RGC` dataset, we extract 30K mouse neuron cells from previously published datasets(Shekhar et al., 2022; Whitney et al., 2023) and newly formed data. The developmental time stages of {E13, E14, E16, E18, P0, P2, P4, P7, P56} (Figure 4). Then, the corresponding gene expressions are measured by the RNA sequencing technique as previously defined[1]. Single-cell libraries were prepared using the single-cell gene expression $3'$ kit on the Chromium platform (10X Genomics, Pleasanton, CA) following the manufacturer's protocol. To be specific, single cells were partitioned into Gel beads in EMulsion (GEMs) in the 10X Chromium instrument followed by cell lysis and barcoded reverse transcription of RNA, amplification, enzymatic fragmentation, $5'$ adaptor attachment, and sample indexing. On average, around $8,000 \sim 12,000$ single cells were loaded on each channel, and around $3,000 \sim 7,000$ cells were recovered. Libraries were sequenced on the Illumina HiSeq $2,500$ platforms.

3.2 Single Cell Data Process

Preprocess and Properties. After we collected the raw signals, the following single-cell sequencing data processing was done using the Seurat package (Hao et al., 2021). Sample quality control was performed on each sample individually. For each sample, doublets were removed using DoubletFinder (McGinnis et al., 2019). We retained cells that expressed at least $1,500$ genes and less than $11,000$ genes. Meanwhile, we removed cells that have more than 5% mitochondrial genes and genes expressed in fewer than 10 cells. The resulting n cells $\times$ g genes matrix of UMI counts were subject to downstream analysis. The UMI-based gene expression matrix was normalized using sctransform (Hafemeister & Satija, 2019). After that, the batch correction was done with canonical correlation analysis (Hotelling, 1992; Anderson et al., 1958), using the top $4,000$ anchor genes.

Clustering. In this research, we are interested in the evolution of different cell types of mouse neurons. Therefore, we built a nearest-neighbor graph to cluster cells based on their transcriptional similarity. Specifically, the number of nearest neighbors was chosen to be 50, according to the rich experiences of biology scientists. The edges were weighted based on the Jaccard overlap metric, and graph clustering was performed using the Louvain algorithm (Blondel et al., 2008). In the end, as demonstrated in Figure 3 (*Left*), the cell clusters were then projected onto a nonlinear 2D space using the Uniform Manifold Approximation and Projection (UMAP) algorithm (McInnes et al., 2018). For temporal trajectory analysis, we further decomposed the cell clusters by their time stage labels like Figure 3 (*Right*). Our goal is to demystify the neuron evolution path across these 9 time stages. Referring Appendix B.2 for more details.

4

Question: Review the Abstract and Methods sections for Language & Expression problems.
Explanation: The paper uses the phrase "gene expressions" ambiguously, creating confusion about the size and composition of the dataset. The abstract mentions integrating a dataset with "30, 000 gene expressions" (page 1, line 016), which is repeated on page 2 (contains 30, 000 gene expressions). This phrasing could be misinterpreted as 30,000 unique genes being measured. The Data Collection section later clarifies that the dataset actually consists of "30K mouse neuron cells" (page 4, line 181). This inconsistent terminology affects the meaning of a critical domain quantity (the number of samples), making it a Language & Expression error.
Error Type: LE (Language & Expression)
Type: Within-Sample

F HUMAN-MACHINE CONSISTENCY EVALUATION

As described in Section 4.1, we employ GPT-4.1 to extract detailed information (*e.g.* evidence sets, reasoning chains) from the responses generated by the models under evaluation . Subsequently, based on the formulas presented in Section 4.1, we calculate S_{location} and $S_{\text{reasoning}}$, which are then used to derive S_{total} for each model's response to the given question.

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.

	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 $P_{\text{unrelated\_err}}$.

Using the information extracted by the human experts, we perform the following calculations:

(1) The $\vec{S}_{\text{location}}$ vector for the 200 questions is calculated based on the evidence sets and Equation 3.

(2) The $\vec{S}_{\text{reasoning}}$ vector is computed from the reasoning chains and Equation 4.

(3) The $\vec{P}_{\text{unrelated\_err}}$ vector is obtained from the count of unrelated errors.

(4) The $\vec{S}_{\text{total}}$ vector is calculated for the 200 questions using Equation 6.

Subsequently, these human-derived vectors ($\vec{S}_{\text{location}}$, $\vec{S}_{\text{reasoning}}$, $\vec{P}_{\text{unrelated\_err}}$, and $\vec{S}_{\text{total}}$) are compared against their counterparts generated by GPT-4.1. Spearman's correlation coefficient is then calculated for these four metrics. The results are presented in Table 4.

Among the four Spearman correlation coefficients, the metric $P_{\text{unrelated\_err}}$ exhibits the highest correlation. This indicates that GPT-4.1's extraction of unrelated errors closely aligns with that of human experts, making it the most precise among the three types of extracted information(*i.e.* evidence sets, reasoning chains, and unrelated errors).

Although the correlation coefficients for the *evidence location score* and *reasoning process score* are relatively lower than $P_{\text{unrelated\_err}}$, they still fall within the range of strong positive correlation. This demonstrates a high degree of consistency in the numerical trends of the scores calculated from GPT-4.1 and human expert extractions, respectively, proving that GPT-4.1 is capable of extracting the majority of effective evidence sets and reasoning chains.

The correlation for the *total score* also lies within the strong positive range and slightly surpasses the correlations for the evidence location score. This also reflects a high level of agreement between GPT-4.1 and human experts.

In summary, GPT-4.1 can extract relevant evidence and reasoning steps with considerable accuracy, leading to precise evaluation scores. This validates the effectiveness of our methodology, which uses GPT-4.1 to parse the responses of the models under evaluation.