OSX at Context24: How Well Can GPT Tackle Contextualizing Scientific Figures and Tables

Tosho Hirasawa
OMRON SINIC X Corp
tosho.hirasawa@sinicx.com

Abstract

Identifying the alignment between different parts of a scientific paper is fundamental to scholarly document processing. In the Context24 shared task, participants are given a scientific claim and asked to identify (1) key figures or tables that support the claim and (2) methodological details. While employing a supervised approach to train models on task-specific data is a prevailing strategy for both subtasks, such an approach is not feasible for low-resource domains. Therefore, this paper introduces data-free systems supported by Large Language Models. We propose systems based on GPT-4o and GPT-4-turbo for each task. The experimental results reveal the zero-shot capabilities of GPT-4* in both tasks. https://github.com/toshohirasawa/context24

1 Introduction

In scientific writing, the alignment between different sections of a paper and the consistency between textual claims and visual elements are paramount for effective communication (Gopen and Swan, 1990). A well-aligned paper presents a coherent narrative from introduction to conclusion, with each section building logically upon the last. This structural integrity is complemented by the harmonious integration of figures and tables with the text, where visual data reinforce and clarify written assertions (Franzblau and Chung, 2012). Such alignment serves multiple crucial functions: it enhances reader comprehension, strengthens the paper's argumentative force, and facilitates critical evaluation by peers. When research questions, methodologies, results, and interpretations are presented consistently across both prose and visual formats, it becomes easier for readers to grasp the study's significance and situate its findings within the broader scientific context. This holistic approach to alignment not only elevates the quality of individual papers but also contributes to the overall efficiency of knowledge dissemination in the scientific community.

Building upon the importance of alignment in scientific writing, the Context24 shared task at the 4th Workshop on Scholarly Document Processing (SDP 2024) addresses a critical challenge in scientific communication: the efficient interpretation and contextualization of scientific claims. This task aligns with the broader goal of enhancing the coherence between claims and supporting evidence in scientific papers. The two tracks of the shared task - **Evidence Identification** and **Grounding Context Identification** - aim to automate the process of linking claims with their supporting visual elements and methodological details, respectively. By facilitating the rapid identification of supporting evidence and grounding context, the Context24 task has the potential to significantly enhance the efficiency of scientific communication, enabling researchers to evaluate and build upon existing work more quickly.

Evidence Identification This track requires participants to identify key figures or tables from a given research paper that provide supporting evidence for a specific scientific claim. Participants must not only locate the relevant visual data but also ensure that it directly supports the stated claim.

Grounding Context Identification In this track, participants must identify all relevant methodological details associated with a scientific claim. These details are often scattered throughout the paper and include figures, tables, and textual descriptions that elucidate the experimental setup, measurement methods, sample characteristics, and other critical information necessary for understanding the basis of the claim.

This paper outlines our approach to the Context24 shared task. Our system was based on the GPT-4 family and exclusively utilized the trial

Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024), pages 324–331

data provided by the organizers, without incorporating any supplementary datasets. Our system achieved the second-highest rank in both tracks. The findings indicate that GPT-4 models can identify grounding information in both cross-modal and unimodal ways without supervised tuning.

2 Related Work

Our zero-shot approach to multimodal scientific claim verification is motivated by two key factors in the current research landscape. First, there is a notable scarcity of large-scale annotated datasets specifically designed for multimodal scientific claim verification. As highlighted by Wadden et al. (2020), creating such datasets requires significant domain expertise and is highly resource-intensive. This lack of training data presents a substantial challenge for supervised learning approaches in this domain.

Secondly, recent advancements in large language models (LLMs) have demonstrated remarkable zero-shot and few-shot capabilities across various tasks. Brown et al. (2020) showed that GPT-3 can perform complex tasks without task-specific fine-tuning, while Wei et al. (2022) further demonstrated the effectiveness of instruction-tuning in enhancing zero-shot performance. These findings suggest that LLMs could potentially address complex tasks like scientific claim verification without extensive labeled data.

In the realm of multimodal models, several architectures have been developed to bridge the gap between vision and language. CLIP (Contrastive Language-Image Pre-training) (Radford et al., 2021b) demonstrated strong zero-shot performance on various vision tasks through natural language supervision. More recent models like LLaVA (Liu et al., 2023) have extended this approach to incorporate more general vision-language capabilities. In the scientific domain, models like ScienceQA (Lu et al., 2022) have been developed to handle scientific figures and text, showing promising results in tasks like answering questions about scientific figures and diagrams.

Our choice of CLIP for this task was motivated by its strong zero-shot performance and its ability to align visual and textual representations in a shared embedding space. This alignment is particularly crucial for our task, where we need to match textual claims with visual evidence from scientific figures and tables.

Recent work on effective instruction writing for LLMs informed the development of our prompting strategy. Kojima et al. (2022) demonstrated that carefully crafted prompts can induce step-by-step reasoning in LLMs, improving their performance on complex tasks. Similarly, Mishra et al. (2022) showed that prompts encouraging LLMs to explain their reasoning often lead to more accurate outputs. These insights guided our approach to designing prompts that encourage the model to articulate its reasoning process when evaluating the relevance of visual evidence to textual claims.

3 Evidence Identification

In the Evidence Identification task, we identified key figures and tables using a pipeline system supported by GPT-4o. Given a claim x and K candidate images (figures and tables) $Z = \{z_1, z_2, \cdots, z_K\}$, we first computed the supportiveness score s_i between the claim and the i-th image. Afterward, we sorted the candidate images based on their scores to determine the final ranking.

3.1 Pipeline

Supportiveness Score. We utilized GPT-4o to compute the supportiveness score for each claim-image pair. Specifically, given a claim x, a candidate image $z_i \in Z$, and the corresponding extracted caption t_i, we computed the supportiveness score $s_i = \text{GPT-4o}(x, z_i, t_i; P)$, where P is our prompt constructed following the OpenAI's instruction[1]. Since the extracted caption t_i is available only for a limited number of images, we asked GPT-4o to extract the text from the candidate image and use this output as an alternative to the extracted caption. The entire prompt is shown in Table 5. Note that 53 candidate images in the test data were rejected by GPT-4o because they "*may contain content that is not allowed by our safety system.*" We assigned a zero score to these images.

In addition to following the OpenAI's instructions, we also embedded task-specific knowledge into the prompt. One of our findings was that some images contain overwhelming amounts of data to support the given claim and were assigned unexpectedly high scores using a naive prompt. To mitigate this behavior, we added an evaluation instruction to penalize these images:

[1] https://platform.openai.com/docs/guides/prompt-engineering

Reduce the score by 3 if the image contains more information than necessary to support the statement.

, where the deduction value of 3 is determined based on the performance of the validation data.

The single supportiveness score was extracted from the output of GPT-4o using a regular expression:

```
^(?P<score>[0-9]{1,2})$
```

, and the match named "score" was then converted into an INT value. In our experiment, every output of GPT-4o contains a match.

We sampled the supportiveness scores five times for each claim-image pair and used the mean score as the final supportiveness score.

Ranking. Once we obtained the supportiveness scores for all candidate images, we sorted them based on their supportiveness scores, from high (supportive) to low (non-supportive).

3.2 Experiments

Experiment settings. We used the official dataset for our validation and testing. To reduce the evaluation cost, we randomly selected 10 instances from each dataset category (`BIOL403`, `akamatsulab`, `dg-social-media-polarization`, and megacoglab) to obtain a balanced validation dataset of 40 instances. We employed GPT-4o as our backbone, limiting the number of tokens in the GPT-4o output to 32 to reduce the computational cost. For evaluation metrics, we employed normalized documented cumulative gain (NDCG) (Järvelin and Kekäläinen, 2017) at 5 and 10.

As baselines, we examined two systems: (i) a system that outputs a list of randomly shuffled images (**Random**) and (ii) a CLIP-based system[2] that outputs the similarity of the claim-image pair as the supportiveness score (**CLIP**).

Experiment results. Table 1 shows the results for the validation and test data. Our first finding is that the CLIP-base system failed to determine the key images; its performance was even worse than random chance. Our in-depth analysis revealed that some images have the same style and differ only in minor details. Since CLIP learns to distinguish images in the same batch with larger differences than those of the candidate images, the nature of scientific images makes it more challenging for CLIP to identify the key images.

Meanwhile, our GPT-backed system achieved more than double the performance of the random baseline. This finding indicates the GPT-4o's capability in identifying grounding information.

[2]From the OpenCLIP (Ilharco et al., 2021; Cherti et al., 2023) project, we used the SOTA CLIP (Radford et al., 2021a) model (`ViT-H-14-378-quickgelu` model, pre-trained on the `dfn5b` dataset, with no fine-tuning.)

	Valid.		Test	
Model	@5	@10	@5	@10
Oracle	0.91	0.91	n/a	n/a
Random	0.29	0.33	0.22	0.29
CLIP	0.18	0.25	n/a	n/a
Ours	0.63	0.66	0.64	0.69

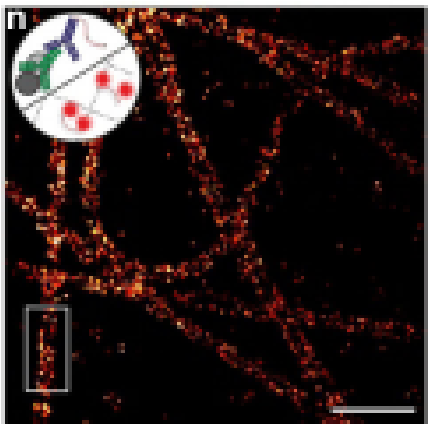
Table 1: NDCG scores at 5 ("@5") and 10 ("@10") on validation ("Valid.") and Test data. "Random" shows the performance of randomly shuffled candidate images; "Oracle" shows the performance of ground truth, which serves as the upper boundary of the models.

Prompt	Input	Valid.	Test
`Naive`	x,z	0.562	0.671
`Naive`	x,z,t	0.566	n/a
`Instruction`	x,z	0.590	0.654
`Instruction`	x,z,t	0.639	0.631

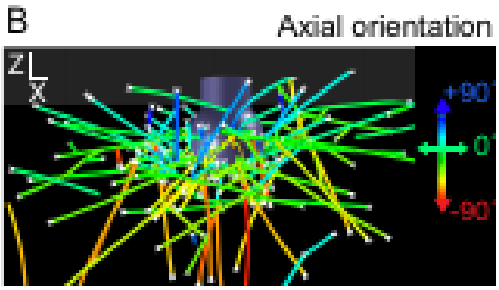
Table 2: NDCG@5 scores on the validation and test data for different prompts and input types.

3.3 Discussion

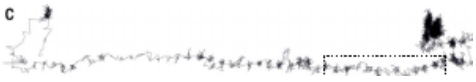
Prompt ablation. Table 2 shows the performance of models with different prompt types and input sources. The `Naive` prompt systems use a simple prompt without instruction (see Appendix C for details). The `Instruction` prompt systems follow the OpenAI's instructions (as described in Section 3.1). While the instruction and additional input t improved the performance for the validation data, they reduced performance for the test data. The different tendencies in prompts and input sources may stem from the different distributions of the validation and test data. While the validation data is balanced across four datasets, the test data comprises two datasets. This indicates our proposed system performs well on two omitted datasets (`BIOL403` and `dg-social-media-polarization`) but worse on the test datasets (`akamatsulab` and megacoglab). Table 3 shows the validation perfor-



(a) Microscopic image



(b) 3D orientation or alignment of structures



(c) A tracing of a neural or cellular structure

Figure 1: Images that are prohibited for ChatGPT-4o.

Dataset	Naive	Instruction
BIOL403	0.452	0.635
akamatsulab	0.403	0.488
dg-social-media-polarization	0.594	0.701
megacoglab	0.798	0.732

Table 3: The validation NDCG@5 scores for four datasets under two different prompt methods.

mance for each dataset, indicating our system has been improved for three datasets (BIOL403, akamatsulab, and dg-social-media-polarization), but failed for the megacoglab dataset.

Prohibited Images. We found that 54 out of 3,714 (1.45%) are prohibited for ChatGPT-4o with our latest prompt. Figure 1 shows some prohibited images of different types. However, as the system's rejection of these images might have been due to a misunderstanding about the type of required analysis or interpretation, we may avoid this behavior by rephrasing our prompts or declaring the clear purpose of our prompts.

Fine-tuning CLIP. Fine-tuning a general-purpose CLIP to fit a specific task requirement is a well-established way to utilize CLIP models (Ha et al., 2024). Following the latest manner, we also have tried to fine-tune a CLIP model for the Evidence Identification task. Technically, we fine-tuned a CLIP model[3] on the train data in two different ways: (a) continuous training with batch-level contrastive loss and (b) adopting the CLIP model for an NLI task. To transform the Evidence Identification task into an NLI task, we annotated findings images and other images in the same paper with a label of 0 (entailment) and 1 (neutral). The best checkpoints in both ways were selected w.r.t. the NDCG@5 score on the validation data. In our experiment, the NLI model outperformed the continuously trained model by around 0.1 NDCG@5 score but underperformed the ChatGPT-4o pipeline. This observation suggested the close images in the same article are preferable to the images from other articles.

4 Grounding Context Identification

In the Grounding Context Identification, we identified the methodological details in a zero-shot manner using GPT-4-turbo. Most part of the prompt P is cited from the official task definition[4]. We only appended a prompt to format the output.

Given a claim x and the full text of the corresponding article d, we first asked GPT-4-turbo to generate the output $r = \text{GPT-4-turbo}(x, d; P)$, where P is the prompt shown in Table 6. We then split the output by the new lines and trimmed the empty lines to obtain the final prediction.

4.1 Experiments

Experiment settings. We have not refined our prompt from the first version (as shown in Table 6). We employed GPT-4-turbo because it accepts a longer context than GPT-4o. The number of tokens in the output is limited to 1,024. We sampled only one response for each claim. For evaluation metrics, we employed BERTScore (Zhang et al.,

[3] openai/clip-vit-base-patch32

[4] https://github.com/oasisresearchlab/context24?tab=readme-ov-file#task-description-1

Model	BS	R-1	R-2	R-L
Ours	0.86	0.32	0.17	0.26

Table 4: BERTScore ("B"), ROUGE-1 ("R-1"), ROUGE-2 ("R-2"), and ROUGE-L("R-L") scores on the test data.

2020) and ROUGE (Lin, 2004).

Experiment results. Table 4 shows the performance of our model for the test data. Compared to the first-placed model, our model achieved nearly comparable performance (-0.01 BERTScore and -0.01 ROUGE-L score). Considering the minimal effort required to tune the prompt, this observation demonstrates the capability of GPT-4-turbo to identify the alignments between a claim and the corresponding methodological details.

5 Conclusion

In this paper, we introduced GPT-4-based systems designed to identify key images and methodological details for a given claim. Our systems achieved considerable high performance in both subtasks within the Context24 Shared Task.

Acknowledgment

This work is supported by JST Moonshot R&D Program Grant Number JPMJMS2236. We used the computational resources of AI Bridging Cloud Infrastructure (ABCI) provided by National Institute of Advanced Industrial Science and Technology (AIST).

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829.

Lauren E Franzblau and Kevin C Chung. 2012. Graphs, tables, and figures in scientific publications: the good, the bad, and how not to be the latter. *The Journal of hand surgery*, 37(3):591–596.

George D Gopen and Judith A Swan. 1990. The science of scientific writing. *American scientist*, 78(6):550–558.

Seokhyeon Ha, Sunbeom Jeong, and Jungwoo Lee. 2024. Domain-aware fine-tuning: Enhancing neural network adaptability. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 12261–12269. AAAI Press.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip. If you use this software, please cite it as below.

Kalervo Järvelin and Jaana Kekäläinen. 2017. Ir evaluation methods for retrieving highly relevant documents. *SIGIR Forum*, 51(2):243–250.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3470–3487. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish

Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021a. Learning transferable visual models from natural language supervision. In *ICML*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021b. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Prompts for tasks

Table 5 and table 6 show our prompts for the Evidence Identification and the Grounding Context Identification tasks, respectively.

B Prompt for OCR

Table 7 shows our prompt for GPT-4o to extract text in candidate images that have no official extracted captions.

C Naive Prompt

Table 8 shows the `Naive` prompt for GPT-4o to determine the supportiveness score of a given claim-image pair.

Role	Prompt
System	Instruction: Evaluate the given image along with its caption and the provided sentence (delimited with XML tags) to determine how well the image supports the sentence. Follow the steps below to ensure a comprehensive analysis. Steps: 1. Description of the Image: Provide a brief description of the image, including key elements and details that stand out. 2. Sentence Analysis: Break down the sentence into its main components and key points. 3. Comparison: Compare the elements and details of the image with the key points of the sentence. 4. Evaluation: Rate the supportiveness of the image on a scale of 1 to 10, where 1 means the image does not support the sentence at all and 10 means the image perfectly supports the sentence. Reduce the score by 3 if the image contains more information than necessary to support the statement. 5. Explanation: Provide a detailed explanation for the rating, highlighting specific aspects of the image that either support or do not support the sentence. Format: Supportiveness Score: [Your rating here] Image Description: • [Your description here] Sentence Analysis: • [Your analysis here] 4. Comparison: • [Your comparison here] 5. Explanation: • [Your explanation here] Example: Supportiveness Score: 2 Image Description: • The image shows a bustling city street with tall skyscrapers, busy traffic, and pedestrians walking on the sidewalks. Sentence Analysis: • The sentence states, ”The serene countryside is a perfect getaway from the city’s hustle and bustle.” Comparison: • The image depicts a busy city street, which is in direct contrast to the serene countryside mentioned in the sentence. Explanation: • The image does not support the sentence as it shows a bustling city street rather than a serene countryside. The elements of the image (skyscrapers, busy traffic, pedestrians) are the opposite of what is described in the sentence.
User	<image>ENCODED_IMAGE</image> <sentence>CLAIM</sentence> <caption>CAPTION</caption>

Table 5: Prompt to assess the supportativeness score of each image. “ENCODED_IMAGE”, “CLAIM”, and “CAPTION” are replaced by the base64-encoded image, the claim text, and the caption (or OCR text), respectively.

Role	Prompt
System	Given a scientific claim and a relevant research paper, identify all grounding context from the paper discussing methodological details of the experiment that resulted in this claim. For the purposes of this task, grounding context is restricted to quotesa from the paper. These grounding context quotes are typically dispersed throughout the full-text, often far from where the supporting evidence is presented. For maximal coverage for this task, search for text snippets that cover the following key aspects of the empirical methods of the claim: 1. What observable measures/data were collected 2. How (with what methods, analyses, etc.) from 3. Who(m) (which participants, what dataset, what population, etc.) You should output only the text snippets and must not contain any explanation.
User	<claim>CLAIM</claim> <full-text>FULLTEXT</full-text>

Table 6: Prompt to extract methodological details in a paper. "CLAIM" and "FULLTEXT" are replaced by the claim text and the full text of the paper, respectively.

Role	Prompt
System	You are an Optical Character Recognition (OCR) machine. You will extract all the characters from the image provided by the user, and you will only provide the extracted text in your response. As an OCR machine, You can only respond with the extracted text.
User	ENCODED_IMAGE

Table 7: Our prompt for GPT-4o to extract text from an image. "ENCODED_IMAGE" is replaced by the base64-encoded image.

Role	Prompt
System	You will receive a sentence and an image from a scientific paper. Determine how well the image supports the sentence and report the score on a scale from 0 to 10, where 0 means wholly irrelevant and 10 means highly supporting. Afterward, provide a detailed explanation for your relevance score, highlighting specific elements of the image and the sentence that influenced your decision.
User	<image>ENCODED_IMAGE</image> <sentence>CLAIM<sentence>

Table 8: `Naive` prompt. "ENCODED_IMAGE" and "CLAIM" are replaced by the base64-encoded image and the claim text, respectively.