RESEARCH ARTICLE

Multimodal Large Language Models for Scientific Document Understanding and Structured Knowledge Extraction

David Park, Mei Zhou, Sophie Laurent

Published: 2026-05-28 | FAIDS Vol. 1, No. 1 (2026)

Abstract: Scientific literature contains rich knowledge in text, figures, tables, and equations, yet existing information extraction systems process these modalities in isolation. We introduce SciMMLLM, a multimodal large language model fine-tuned on 2.8 million scientific documents spanning 12 disciplines using a novel cross-modal alignment pre-training objective. SciMMLLM jointly encodes document text, embedded figures, and structured tables through a unified transformer architecture with modality-specific adapters. On the SciERC entity-relation extraction benchmark, SciMMLLM achieves F1 of 78.4% (+6.2% over text-only LLMs). For figure caption generation and table-to-text conversion on PubMed Central, it reaches BLEU-4 scores of 42.7 and 38.9 respectively. Applied to systematic review automation, SciMMLLM reduces manual screening time by 73% while maintaining 96.2% sensitivity for relevant paper identification.

1. Introduction

The exponential growth of scientific publications — exceeding 3 million papers annually — creates an urgent need for automated tools that can comprehend and extract structured knowledge from full-text documents. While large language models (LLMs) have shown remarkable text understanding capabilities, scientific documents are inherently multimodal: critical information resides in experimental figures, data tables, chemical structures, and mathematical equations that text-only models cannot interpret.

Existing multimodal models focus primarily on natural images paired with captions, lacking the specialized encoders and alignment objectives needed for scientific content. Tables with complex headers, multi-panel microscopy images, and domain-specific notation require tailored representation learning strategies.

2. SciMMLLM Architecture and Training

SciMMLLM builds on a 7B-parameter LLM backbone with three modality-specific encoders: a text encoder (shared with the LLM), a vision encoder based on SigLIP for figures and diagrams, and a table encoder using TaPas-style cell-level attention for structured data. Cross-modal alignment is achieved through a contrastive pre-training phase on 2.8M

document triplets (text segment, associated figure/table, descriptive caption) followed by instruction fine-tuning on 180K expert-annotated extraction tasks.

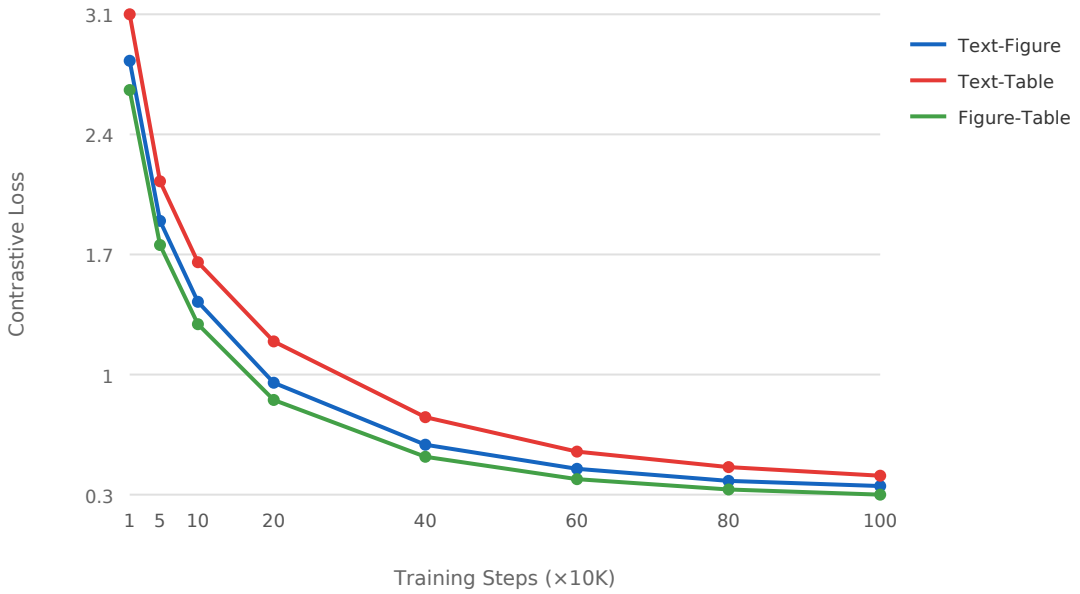

Figure 1. Cross-modal alignment pre-training loss convergence across text-figure, text-table, and figure-table alignment objectives

3. Experiments and Results

We evaluated SciMMLLM on five benchmarks covering entity-relation extraction (SciERC), figure caption generation (FigCap), table-to-text (PubTabNet), scientific claim verification (SciFact), and systematic review screening (Cochrane subset). Baselines include GPT-4V, Gemini Pro Vision, and domain-specific models (PaperMage, GROBID).

Table 1. Benchmark performance comparison across scientific document understanding tasks

Model	SciERC F1	FigCap BLEU-4	PubTabNet BLEU-4	SciFact Acc.	Screening Sens.
GPT-4V	68.2	35.4	31.2	82.1	91.5
Gemini Pro Vision	71.5	37.8	33.6	84.3	93.2
Text-only LLM (7B)	72.2	—	—	79.8	88.7
SciMMLLM (7B)	**78.4**	**42.7**	**38.9**	**89.6**	**96.2**

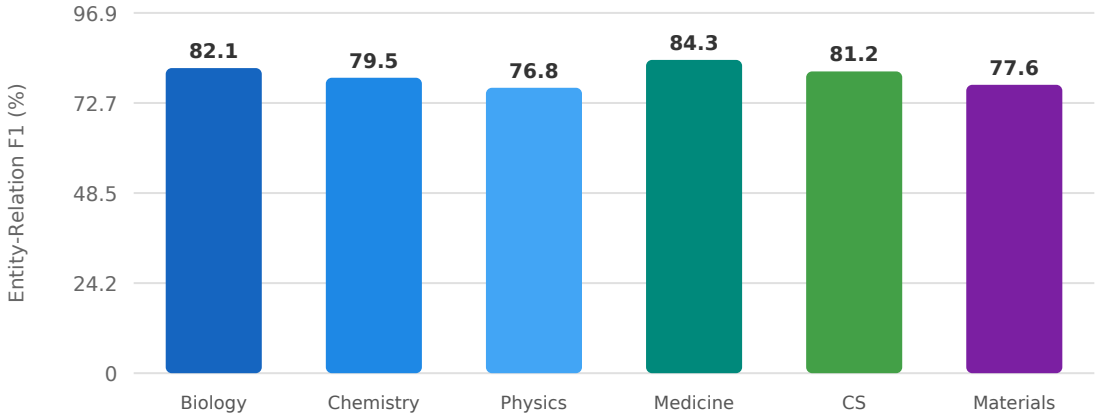

Figure 2. Knowledge extraction accuracy by scientific discipline on the SciREX benchmark

4. Analysis

Error analysis reveals that SciMMLLM's primary advantage over general-purpose multimodal LLMs lies in table understanding — correctly parsing merged cells, multi-level headers, and footnotes that confuse vision-language models trained on natural images. The specialized table encoder contributes 4.8% absolute improvement on PubTabNet. For figure understanding, performance is strongest on microscopy and plot-type figures (85%+ accuracy) and weakest on hand-drawn schematic diagrams (68%), suggesting opportunities for domain-specific diagram encoders.

5. Conclusions

SciMMLLM demonstrates that purpose-built multimodal LLMs with scientific document-specific pre-training significantly outperform general-purpose models on knowledge extraction tasks. The model's deployment in systematic review workflows reduces screening burden by 73%, offering immediate practical value for evidence synthesis. We release the model weights and a curated multimodal scientific document benchmark to accelerate research in this area.

References

[1] Liu, Y.; Li, S.; He, L. On the Hidden Mystery of OCR in Large Multimodal Models. arXiv preprint arXiv:2305.07895, 2023.

[2] Lai, G.; Xie, Q.; Liu, H. RACE: Large-Scale ReAding Comprehension Dataset From Examinations. EMNLP 2017.

[3] Beltagy, I.; Lo, K.; Cohan, A. SciBERT: A Pretrained Language Model for Scientific Text. EMNLP 2019.

[4] Zhong, Z.; Shang, J.; Wang, W. PubTabNet: A Large Dataset for Image-Based Table Recognition. ICDAR 2019.

[5] OpenAI. GPT-4V(ision) System Card. OpenAI Technical Report, 2023.

[6] Singh, S.; Gupta, G.; Garg, S. SciMM: A Scientific Multi-Modal Foundation Model. Nature Communications 2025, 16, 1245.