TRACE-PSG: Traceable Sleep Diagnosis Support with Pointer-Grounded Evidence

Sipeng Wu[1,2], Shiqin Tang[2⋆], Chenxi Liu[2⋆], Jingwei Wu[2], Wenjing Jiang[3], Xiaoyu Xiao[3], Hongbin Liu[2], and Gaofeng Meng[2]

[1] The University of Hong Kong, Hong Kong SAR
[2] Centre for Artificial Intelligence and Robotics, Hong Kong Institute of Science & Innovation, Chinese Academy of Sciences, Hong Kong SAR
{shiqin.tang, chenxi.liu}@cair-cas.org.hk
[3] Qilu Hospital of Shandong University, China

Abstract. Multi-page polysomnography (PSG) reports contain dense plots, summary indices, and annotations that clinicians must integrate with free-text patient narratives (CC/HPI) to produce diagnoses and treatment plans. However, current multimodal LLMs often fail to provide *auditable* clinical outputs: they struggle with evidence grounding, may hallucinate unsupported claims, and must abstain when evidence is missing or indeterminate. We propose a constraint-driven two-stage framework for auditable **structured clinical decision support** from multimodal sleep documents. Stage 1 aligns report reading via rule-scored metric extraction using reinforcement learning with group-relative policy optimization (GRPO). Stage 2 warm-starts from the aligned checkpoint and performs supervised fine-tuning to generate structured clinical JSON with diagnosis, evidence atoms, notes items, and treatment fields. Outputs are grounded through a restricted Pointer Bank and a bounded evidence-atom budget, enabling deterministic provenance checks. We evaluate on a retrospective tertiary-center dataset with 657 studies; the train/val/test split is 479/78/100 instances. Our model achieves 100% JSON parseability and strong evidence alignment (Grounding Pass 88%; metric-pointer F1 0.8910, 95% CI 0.8676–0.9113), with strong insomnia subtype prediction (F1 0.8821). Under our evaluation protocol, it avoids forced OSA diagnosis under indeterminate evidence (0% OverDx) and produces no detected safety-rule violations. We additionally conduct a physician rubric evaluation on 100 cases to assess long-form clinical recommendations. Ablations show Stage 1 alignment is critical for grounding and diagnostic reliability under limited supervision.

Keywords: multimodal LLM · PSG reports · auditability · evidence grounding · structured clinical generation.

1 Introduction

Polysomnography (PSG) is the gold standard for diagnosing sleep-related breathing disorders and a broad spectrum of sleep disorders, requiring overnight acqui-

⋆ Corresponding authors: Shiqin Tang and Chenxi Liu.

sition of multi-modal physiological signals (EEG/EOG/EMG/ECG, airflow, and oxygen saturation). In routine practice, clinicians rarely review raw waveforms; instead, sleep laboratories deliver multi-page PSG reports containing dense plots, summary indices, and annotations [22]. Clinical decisions further depend on integrating these reports with free-text patient narratives (CC/HPI) to produce structured diagnoses and treatment plans.

The rapid advances in large language models (LLMs) have created an unprecedented opportunity to tackle this challenging clinical reasoning problem. However, deploying text-only LLMs in medical practice faces fundamental limitations. Without access to visual inputs, unimodal models cannot interpret clinically informative non-text modalities, such as multi-page scanned documents, tables, and PDF report layouts that encode critical context for real-world decision-making, thereby constraining their practical utility in routine care [1, 14]. This gap has catalyzed the rise of multimodal large language models (MLLMs), which extend LLMs with visual perception.

Recent advances in MLLMs and VLMs have enabled strong performance in medical VQA and single-image report generation [19, 13, 16, 24]. However, performance degrades on high-density, multi-page clinical documents, where attention dilution and vision–language feature entanglement hinder precise evidence grounding [17]. Such failures are particularly concerning in medicine, where medical reasoning has been enhanced through reflection and tree-of-thought frameworks [26, 27], yet unreliable visual perception and vision chain-of-thought may still propagate into unsupported clinical reasoning [25]. This limitation amplifies a key safety risk: hallucinations—fluent outputs that lack evidential support—which can lead to misdiagnosis and inappropriate recommendations [10]. In parallel, emerging trustworthy-AI constraints increasingly require transparency, human oversight, and auditable evidence traceability [20, 2, 5]. This motivates evaluation beyond fluency toward compliance-centered and grounding-centered audits [2, 11]. While constrained decoding can enforce machine-parseable structures [15], syntactic validity does not ensure factuality [3]; similarly, text-only retrieval-augmented generation (RAG) can improve textual grounding [12] but remains insufficient for multimodal reports, while extracting a few numeric readings may help, it often fails to preserve spatial/topological structure in charts (trend shapes, crossings, and alignment to axes/reference bands), which can be critical for downstream reasoning [23, 6].

Consequently, reliable *auditable* structuring of multi-page PSG reports and free-text CC/HPI into verifiable clinical outputs remains unresolved in sleep medicine. Here, auditability requires that each clinical claim be traceable to explicit provenance (a specific report field/plot or a narrative span), while explicitly abstaining when evidence is insufficient or indeterminate. We propose a two-stage, constraint-driven framework that first aligns report reading through rule-scored metric extraction and then generates end-to-end structured clinical JSON with explicit provenance.

Our contributions are: **(1) Auditable task formulation.** We cast structured clinical generation from heterogeneous multi-page PSG documents plus

CC/HPI as a constrained generation problem targeting auditability, grounding, and missing-information abstention. **(2) Pointerized evidence for verifiability.** We introduce a restricted Pointer Bank and a bounded `evidence_atoms` budget, enabling deterministic provenance checks beyond schema-level validity. **(3) Task-specific auditable training pipeline.** Stage 1 learns report reading with rule-based rewards; Stage 2 produces clinically actionable JSON grounded by pointers, transferring extraction-aligned representations to downstream reasoning. **(4) Audit-centric evaluation on real PSG reports.** We curate clinician-labeled supervision for both stages and evaluate parseability, grounding validity and pointer accuracy, diagnostic behavior (including over-diagnosis under indeterminate evidence), and safety.

2 Method

2.1 Task Formulation

Given a multi-page sleep study document D (PDF/images) and clinical text H (CC/HPI), our goal is to generate a structured clinical output Y that is auditable and actionable. We represent Y as a JSON object with four components: *diagnosis*, *evidence_ atoms*, *notes_ items*, and *treatment*. To support auditability, each diagnostic decision and clinically relevant note is linked to one or more evidence units via canonical pointers (`metric:*` for report-derived metrics and `note:*` for clinical-text anchors), rather than free-form citations. We operationalize verifiability via schema validity, pointer validity, and missing-information abstention. Notably, *treatment* is kept as a separately evaluated long-form field, whereas *diagnosis* and *notes* are pointerized for deterministic provenance checks.

As shown in Fig. 1, we derive two supervision signals from each study: (i) metric-level labels M for report-reading alignment, and (ii) clinician-derived structured labels for the final output Y.

2.2 Pointerized Evidence Representation

Free-form clinical generation is prone to unsupported statements and hallucinated rationale, especially when the model must jointly summarize findings, assign diagnoses, and recommend treatments. To improve verifiability, we introduce a *pointerized evidence representation* that factorizes the output into (i) minimal evidence units and (ii) human-readable notes linked to those evidence units through a restricted pointer space.

We represent model-consumable evidence as a compact set:

$$\mathcal{E} = \{e_k\}_{k=1}^{K}, \quad K \leq 6, \tag{1}$$

We refer to these as `evidence_atoms`; each atom is

$$e_k = (claim_k,\ source_k,\ pointer_k). \tag{2}$$

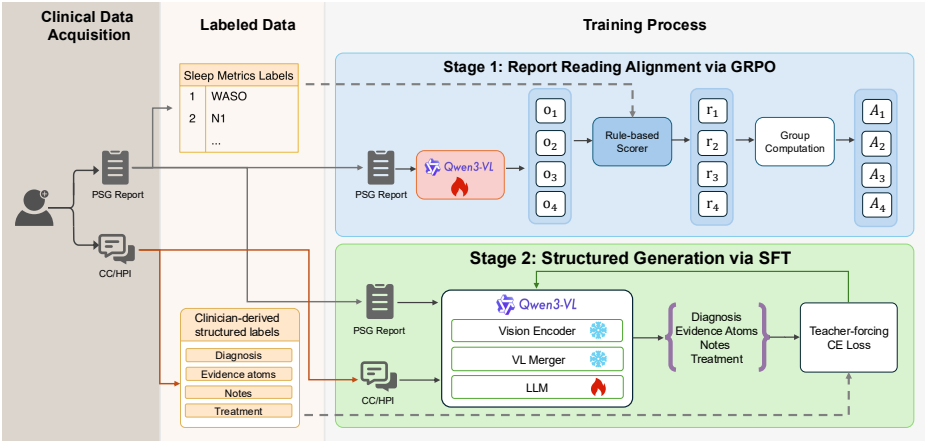

Fig. 1. Overall framework of our two-stage training pipeline. Stage 1 aligns report reading by learning strict, rule-scored metric extraction from multi-page sleep reports. Stage 2 warm-starts from the Stage 1 checkpoint and generates the full structured clinical JSON with pointerized evidence grounding using teacher-forcing.

Here, `claim` is a short atomic statement, `source` $\in$ {`note`, `metric`, `derived`, `missing`}, and `pointer` is a list of evidence pointers. We restrict pointers to a predefined *Pointer Bank*:

$$\mathcal{B} = \mathcal{B}_{\text{metric}} \cup \mathcal{B}_{\text{note}}, \tag{3}$$

with $\mathcal{B}_{\text{metric}} = \{\texttt{metric:*}\}$ and $\mathcal{B}_{\text{note}} = \{\texttt{note:CC}, \texttt{note:HPI}, \ldots\}$. All pointers in `evidence_atoms` must be selected from $\mathcal{B}$.

For human-readable rationales and action items, we represent them as `notes_items`,

$$\mathcal{N} = \{n_m\}_{m=1}^{M}, \tag{4}$$

where each item is

$$n_m = (type_m,\ code_m,\ targets_m,\ pointers_m,\ text_m). \tag{5}$$

`type` is restricted to a small closed set (`missing_info`, `support`, `conflict`, `next_step`); `targets` indicates the diagnosis dimensions; and `pointers` must also be drawn from the same Pointer Bank $\mathcal{B}$. This design decouples *clinical content* (`code`/`targets`) from *surface phrasing* (`text`), while preserving explicit grounding.

We enforce a compact budget ($K \leq 6$) to encourage a concise set of key, non-redundant evidence atoms that clinicians can audit efficiently, rather than producing long exhaustive rationales.

2.3 Stage 1: Report Reading Alignment via GRPO

Before end-to-end training, we explicitly align the model to metric extraction and hard consistency rules using GRPO [21]. We start from a state-of-the-art

general-purpose multimodal instruction model Qwen3-VL-2B [4] (denoted as π_θ) and design a strict extraction prompt that asks the model to output a single JSON object containing a fixed set of PSG metrics. The prompt constrains the response to machine-parseable JSON with required keys, enabling automatic scoring and preventing "free-form" narration.

GRPO Objective. We optimize π_θ with Group Relative Policy Optimization (GRPO) [21], following the standard group-relative advantage formulation.

At each update, we sample a group of G candidate outputs $\{o_i\}_{i=1}^{G}$ from the previous policy $\pi_{\theta_{\text{old}}}$, compute rule-based rewards r_i, and form a group-relative advantage.

$$A_i = \frac{r_i - \text{mean}(\{r_j\})}{\text{std}(\{r_j\})}. \tag{6}$$

We then optimize the clipped GRPO objective with KL regularization to a reference checkpoint π_{ref}:

$$J_{\text{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G} \min\Big(r_i^{\text{ratio}} A_i,\ \text{clip}\big(r_i^{\text{ratio}},\ 1 \pm \epsilon\big) A_i\Big) - \beta\, D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})\right], \tag{7}$$

where

$$r_i^{\text{ratio}} = \frac{\pi_\theta(o_i \mid v)}{\pi_{\theta_{\text{old}}}(o_i \mid v)}, \tag{8}$$

ϵ controls clipping, and β controls the KL strength.

Reward Decomposition. We use a rule-based reward to encourage (i) strict machine-parseable formatting and (ii) faithful, complete extraction. Following prior work [7, 18], we decompose

$$r = R_{\text{format}} + R_{\text{acc}}, \qquad R_{\text{format}} \in \{0, 1\}. \tag{9}$$

R_{format} rewards compliance with the required wrapper and JSON schema; R_{acc} is computed only if the output is parseable (format-gated) and aggregates type-aware value matching, time-tolerant matching, arithmetic consistency, optional posture-wise consistency, and coverage:

$$R_{\text{acc}} = \sum_{t \in \{\text{val,time,cons,pcons,cov}\}} w_t R_t, \quad w_{\text{val}} = 1 - \sum_{t \neq \text{val}} w_t. \tag{10}$$

2.4 Stage 2: Structured Generation via SFT

After Stage-1 report-reading alignment, we apply supervised fine-tuning (SFT) to generate a fully structured clinical JSON from multimodal inputs. We initialize Stage-2 from the Stage-1 GRPO checkpoint to inherit metric-reading and extraction-aligned representations. Each example includes multi-page report images, complaint/history text, and an explicit Pointer Bank (e.g., `metric:*`,

`note:*`). The target JSON includes `diagnosis`, `notes_items`, `evidence_atoms`, and `treatment`. Training uses standard teacher-forcing next-token prediction over the serialized JSON:

$$\mathcal{L}_{\mathrm{SFT}} = -\sum_{t=1}^{T} \log p_\theta(y_t \mid y_{<t}, x), \tag{11}$$

where x is the packaged multimodal input and $y_{1:T}$ is the target JSON sequence.

Initializing Stage-2 from the Stage-1 checkpoint transfers report-reading and evidence-localization capabilities, improving schema-constrained structured generation (see Section 4).

3 Experiments

Dataset. We retrospectively collected de-identified multi-page PSG reports and CC/HPI text from a tertiary sleep center (IRB-approved). The cohort comprises 136 patients (2024–2026) with multi-night recordings; each night is treated as one instance (train/val/test: 479/78/100) with *strict patient-level* splitting to avoid leakage. Each instance includes report, CC/HPI, physician-extracted Stage-1 metric labels, and physician-assigned Stage-2 structured outputs (diagnosis, evidence atoms, notes, treatment). The dataset is insomnia-dominant and often lacks respiratory channels, making OSA axes frequently indeterminate; we therefore report grounding and over-diagnosis metrics in addition to subtype F1.

Evaluation Metrics. We evaluate **format validity**, **evidence grounding**, **diagnostic behavior**, and **treatment safety**. **Parse OK** is the percentage of outputs that are valid JSON; all remaining metrics are computed on parsed outputs only. For grounding, we report **Grounding Pass** (all pointers in the Pointer Bank and $|\texttt{evidence\_atoms}| \leq K$) and **Metric-pointer F1** over `metric:*` pointers. For diagnosis, we report **Insomnia subtype F1** on evaluable samples and **OSA OverDx** on indeterminate ground truth. For safety, we report a rule-based **Safety violation rate**. Since `treatment` is long-form, we additionally run clinician rubric evaluation (5-point Likert; mean±SD for quality/usability/overall, plus critical-error and hallucination rates) following [9].

Implementation Details. Stage 1 uses the Qwen3-VL-2B-Instruct backbone [4]. We train with GRPO using n=4 candidates per prompt, prompt/response lengths 8192/1024, actor LR 7.5×10^{-7}, and 1500 steps. The reported run disables KL regularization (β=0). Since Stage 1 is a constrained metric-extraction task with format-gated rule rewards and short training, we monitor parsing and extraction behavior on held-out data to reduce the risk of reward exploitation. We therefore evaluate Stage 1 by downstream parseability, pointer validity, and diagnostic behavior rather than training reward alone. Stage 2 performs LoRA SFT [8] initialized from the Stage-1 checkpoint using rank 8 (alpha 16, dropout 0.05), freezing the vision tower and multimodal projector. We train with input length 8192, LR 1×10^{-5}, effective batch size 16, and 500 steps.

Table 1. Main results on structured validity, evidence grounding, diagnosis, and safety. ↑ higher is better; ↓ lower is better. Grounding Pass requires all pointers be in the provided Pointer Bank and $|\texttt{evidence\_atoms}| \leq 6$. Best in bold.

Method	Parse OK ↑	Grounding Pass ↑	Pointer F1 ↑	Insomnia type F1 ↑	OSA OverDx ↓	Safety Viol. ↓
Ours-2B	100.0%	88.00%	**0.8910**	**0.8821**	0%	0%
Qwen3-VL-8B	100.0%	**94.00%**	0.8516	0.5616	0%	0%
Qwen3-VL-2B	0.0%	–	–	–	–	–

Table 2. Ablation (I): parsing and grounding. We compare Stage2-only SFT (no Stage 1 GRPO) and removing the Pointer Bank from the Stage 2 prompt. Grounding metrics are computed on parsed outputs only.

Method	Stage1 GRPO	Stage2 SFT	Pointer Bank	Parse OK ↑	Grounding Pass ↑	Pointer F1↑
Stage2-only	×	✓	×	86.0%	23.26%	0.6854
w/o PB	✓	✓	×	98.0%	83.67%	0.8224
Ours	✓	✓	✓	**100.0%**	**88.00%**	**0.8910**

4 Results and Discussion

Main Results. On the test set, our model produces fully parseable structured outputs (Table 1) and achieves strong evidence alignment (Pointer F1 = 0.891) under the pointer-validity and atom-budget constraints, enabling deterministic auditing. This grounded evidence selection translates into strong insomnia subtype prediction, while avoiding forced OSA diagnosis under indeterminate evidence in our evaluation protocol (0% OverDx) and producing no detected safety-rule violations. Among baselines, Qwen3-VL-2B fails to meet the required schema. Notably, the zero-shot Qwen3-VL-8B attains a higher *Grounding Pass* rate, suggesting stronger instruction-following under the pointer/atom constraints; however, its lower *Pointer F1* and markedly reduced insomnia subtype F1 indicate that satisfying structural grounding constraints does not guarantee selecting the *correct* evidence or performing accurate downstream diagnostic reasoning.

Ablation Study. Table 2 and Table 3 show that Stage 1 GRPO is critical: removing it substantially degrades parsing, grounding, and insomnia subtype performance, and increases OSA over-diagnosis, suggesting end-to-end SFT alone struggles to learn robust report reading under limited supervision. Adding the Pointer Bank mainly improves controllability and evidence alignment (higher Pointer F1 and Parse OK) by restricting the citation space, with a small trade-off in subtype F1 that likely reflects more conservative decisions under stricter grounding.

Table 3. Ablation (II): diagnosis and safety. Diagnosis/safety metrics are computed on parsed outputs only unless noted.

Method	Stage1 GRPO	Stage2 SFT	Pointer Bank	Insomnia type F1 ↑	OSA OverDx ↓	Safety Viol. ↓
Stage2-only	×	✓	×	0.6073	13.95%	0%
w/o PB	✓	✓	×	**0.8963**	0%	0%
Ours	✓	✓	✓	0.8821	0%	0%

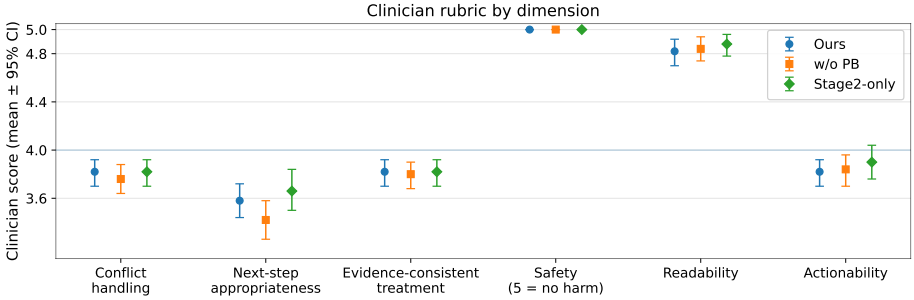

Fig. 2. Clinician rubric scores with 95% CIs (n=100).

Clinician Evaluation. Across 100 cases, a professional physician rated the three models using a 5-point Likert rubric (Fig. 2, Table 4). Aggregate clinician scores were close, with Stage2-only slightly higher on some long-form usability dimensions. This suggests TRACE-PSG mainly improves auditable structured support—parseability, pointer grounding, evidence bounding, and abstention under indeterminate evidence—rather than clinician-facing aggregate scores. Notably, Stage2-only produced 13.95% OSA OverDx, whereas TRACE-PSG reduced it to 0%; similar rubric scores indicate no clear usability or readability collapse from auditability constraints.

Limitations. This retrospective single-center study includes 657 studies from 136 patients and a 100-instance test set. The insomnia-dominant cohort and site-specific PSG reporting conventions limit external generalization; thus, our results support internal feasibility rather than validation of a deployable clinical system. Future work should include multi-center evaluation, more sensitive assessment of free-form treatment recommendations, and extension of pointer-grounded provenance to treatment, which remains a non-pointerized long-form field in the current formulation.

5 Conclusion

We introduced a two-stage framework for auditable structured generation from multi-page PSG reports and CC/HPI. Pointerized evidence enables deterministic provenance checks beyond schema validity. Stage 1 rule-scored extraction

Table 4. Clinician evaluation (aggregate). Mean±SD across 100 cases for rubric scores and error rates for Critical Error/Hallucination.

Model	Clinical quality mean ↑	Usability mean ↑	Overall impression mean ↑	Critical Error % ↓	Hallucination % ↓
Ours	4.05±0.28	4.32±0.24	4.30±0.76	0.0%	0.0%
w/o PB	4.00±0.28	4.34±0.26	4.32±0.79	0.0%	0.0%
Stage2-only	**4.08±0.29**	**4.39±0.29**	**4.36±0.78**	0.0%	0.0%

alignment and Stage 2 teacher-forced JSON generation produce fully parseable outputs with strong grounding, improving insomnia subtype performance while avoiding forced OSA diagnosis under indeterminate evidence in our internal evaluation. Ablations verify the importance of report-reading alignment, and clinician evaluation indicates that improvements concentrate in verifiable fields.

Acknowledgments. This research was supported by the innoHK project and partially by the National Natural Science Foundation of China (Grant No. 62376267). We thank Qilu Hospital of Shandong University for providing the clinical data used in this study.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. AlSaad, R., Abd-Alrazaq, A., Boughorbel, S., Ahmed, A., Renault, M.A., Damseh, R., Sheikh, J.: Multimodal large language models in health care: applications, challenges, and future outlook. Journal of medical Internet research **26**, e59505 (2024)
2. Alu, F.F., Oluwadare, S.: An auditable and source-verified framework for clinical ai decision support: integrating retrieval-augmented generation with data provenance. Frontiers in Artificial Intelligence **9**, 1737532 (2026)
3. Artsi, Y., Sorin, V., Glicksberg, B.S., Korfiatis, P., Freeman, R., Nadkarni, G.N., Klang, E.: Challenges of implementing llms in clinical practice: Perspectives. Journal of Clinical Medicine **14**(17), 6169 (2025)
4. Bai, S., Cai, Y., Chen, R., et al.: Qwen3-vl technical report. arXiv (2025)
5. Bignami, E., Darhour, L.J., Franco, G., Guarnieri, M., Bellini, V.: Ai policy in healthcare: a checklist-based methodology for structured implementation. Journal of Anesthesia, Analgesia and Critical Care **5**(1), 56 (2025)
6. Chu, Y.W., Zhang, K., Malon, C., Min, M.R.: Reducing hallucinations of medical multimodal large language models with visual retrieval-augmented generation. arXiv (2025)
7. Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv (2025)
8. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)

9. Khasentino, J., Belyaeva, A., Liu, X., Yang, Z., Furlotte, N.A., Lee, C., Schenck, E., Patel, Y., Cui, J., Schneider, L.D., et al.: A personal health large language model for sleep and fitness coaching. Nature Medicine **31**(10), 3394–3403 (2025)
10. Kim, Y., Jeong, H., Chen, S., Li, S., Lu, M., Alhamoud, K., Mun, J., Grau, C., Jung, M., Gameiro, R., et al.: Medical hallucinations in foundation models and their impact on healthcare. arXiv (2025)
11. Kulkarni, S., Lyzhov, A., Chaitanya, S., Joshi, P.: All required, in order: Phase-level evaluation for ai-human dialogue in healthcare and beyond. arXiv (2026)
12. Kuo, S.M., Tai, S.K., Lin, H.Y., Chen, R.C.: Automated clinical trial data analysis and report generation by integrating retrieval-augmented generation (rag) and large language model (llm) technologies. AI **6**(8), 188 (2025)
13. Li, J., Liang, Y., Li, Q., Lyu, X., Qian, J., Chen, H., Wang, K., Zeng, Z., Bharath, A.A., Liu, Y.: Pathmem: Toward cognition-aligned memory transformation for pathology mllms (2026)
14. Liu, C., Zhou, S., Xu, Q., Miao, H., Long, C., Li, Z., Zhao, R.: Towards cross-modality modeling for time series analytics: a survey in the llm era. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence. pp. 10564–10572 (2025)
15. Lu, Y., Li, H., Cong, X., Zhang, Z., Wu, Y., Lin, Y., Liu, Z., Liu, F., Sun, M.: Learning to generate structured output with schema reinforcement learning. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 4905–4918 (2025)
16. Nam, Y., Kim, D.Y., Kyung, S., Seo, J., Song, J.M., Kwon, J., Kim, J., Jo, W., Park, H., Sung, J., et al.: Multimodal large language models in medical imaging: current state and future directions. Korean Journal of Radiology **26**(10), 900 (2025)
17. Nguyen, D., Ho, M.K., Ta, H., Nguyen, T.T., Chen, Q., Rav, K., Dang, Q.D., Ramchandre, S., Phung, S.L., Liao, Z., et al.: Localizing before answering: A benchmark for grounded medical visual question answering. In: Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25). International Joint Conferences on Artificial Intelligence Organization (2025)
18. Pan, J., Liu, C., Wu, J., Liu, F., Zhu, J., Li, H.B., Chen, C., Ouyang, C., Rueckert, D.: Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 337–347. Springer (2025)
19. Qian, J., Lu, Y., Xie, W., Lai, Z., Wang, M., Li, X.: Cpsr-clip: Conditional prompt-induced style reconstruction for zero-shot domain adaptation. IEEE Transactions on Multimedia **27**, 7714–7728 (2025)
20. Qian, J., Yang, Z., Chen, G., Hu, P., Tan, K.C., Wang, Y., Huang, Y.A., Huang, Z.A.: Hyperbolic relational prompts for intersectional fairness in medical vlms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13712–13721 (2026)
21. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv (2024)
22. Tu, H.J., Hsu, J.L.: Enhancing medical diagnosis document analysis with layout-aware multitask models. Diagnostics **15**(23), 3039 (2025)

23. Wang, Y., Ma, X., Nie, P., Zeng, H., Lyu, Z., Zhang, Y., Schneider, B., Lu, Y., Yue, X., Chen, W.: Scholarcopilot: Training large language models for academic writing with accurate citations. In: Conference on Language Modeling (2025)
24. Wang, Y., Sun, Y., Tan, T., Hao, C., Cui, Y., Su, X., Xie, W., Shen, L., Yu, Z.: Trrg: Towards truthful radiology report generation with cross-modal disease clue enhanced large language models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 647–657. Springer (2025)
25. Wu, Y., Yang, Z., Qian, J., Gao, S., Chen, G., Li, Q., Huang, Y.A., Huang, Z.A.: Better eyes, better thoughts: Why vision chain-of-thought fails in medicine (2026)
26. Yang, Z., Qian, J., Peng, Z., Zhang, H., Huang, Z.A.: Med-refl: Medical reasoning enhancement via self-corrected fine-grained reflection. arXiv e-prints pp. arXiv–2506 (2025)
27. Yang, Z., Qian, J., Tan, K.C., Wong, H.S., Chen, Y., Zhang, H., Huang, Z.A.: Qm-tot: A medical tree of thoughts reasoning framework for quantized model. arXiv preprint arXiv:2504.12334 (2025)