TRACE: Deployable Tree-Relational Structure Enhancement for Oncology LLMs

Jizheng Lai[1], Yingyun Li[2], Ying Qin[2], Haiyang Qian[1] *
[1]Pharmaron, China
[2]AI Starfish, China
{jizheng.lai,haiyang.qian}@pharmaron.com
{yingyun.li,ying.qin}@aistarfish.com

Abstract

Large language models are increasingly used in oncology applications, but their predictions are often weakly grounded in explicit medical structure. We present TRACE, a deployable tree-relational enhancement framework for oncology LLMs. TRACE separates expensive offline structure learning from lightweight online inference: oncology concepts and relations are organized into an updatable tree-relational structure, refined using LM-loss-derived evidence, and retrieved at inference time as compact prompt evidence. This design supports task-adaptive evidence selection without requiring supervised labels in the zero-shot setting. Across ten oncology classification tasks and one MedQuAD CancerGov QA benchmark, TRACE improves both label-free evaluation and supervised fine-tuning. Additional analyses show that TRACE improves over vanilla RAG and generic GraphRAG, remains useful under leakage-controlled METABRIC inputs, and produces interpretable evidence paths aligned with clinical reasoning. These results suggest that explicit, updatable medical structure is a practical path toward more accurate and auditable oncology LLM deployment.

1 Introduction

Large language models (LLMs) are increasingly used as interfaces for medical knowledge and clinical decision support. Recent medical LLMs and evaluations show rapid progress in medical question answering and consultation-style reasoning, but also reveal persistent gaps between benchmark performance and robust real-world clinical reasoning (Singhal et al., 2023; Zhang et al., 2023; Liu et al., 2024; Wang et al., 2025; Baichuan-M2 Team et al., 2025; Baichuan-M3 Team et al., 2026; Ren et al., 2026). These observations suggest that medical LLMs should not only store domain knowledge parametrically, but also identify and prioritize the clinical evidence needed for a specific decision.

This requirement is particularly important in oncology. Cancer reasoning often depends on structured interactions among disease categories, histology, molecular markers, staging criteria, treatment, recurrence, and prognosis. Curated oncology resources such as OncoKB and CIViC organize genomic alterations with diagnostic, prognostic, and therapeutic implications (Chakravarty et al., 2017; Griffith et al., 2017), while biomedical knowledge graphs such as PrimeKG integrate heterogeneous biomedical entities and relations for precision-medicine analysis (Chandak et al., 2023). These resources highlight the value of explicit medical structure for oncology reasoning.

Retrieval-augmented generation (RAG) grounds LLMs in external evidence, and recent graph-based RAG methods retrieve triples, paths, or subgraphs for multi-hop reasoning and hallucination mitigation (Xiong et al., 2024; He et al., 2024; Gutiérrez et al., 2024; Li et al., 2025; Wu et al., 2025). However, most existing systems treat the external structure as a fixed retrieval index. They retrieve passages or graph fragments at inference time, but do not explicitly refine the medical structure using model-side evidence. Recent clinical knowledge-graph construction studies further suggest that medical graphs should be continuously updated as evidence changes (Das et al., 2026; Xu et al., 2026).

We propose TRACE, a Tree-Relational Adaptive Concept Engine for Oncology LLMs. TRACE is designed as a deployable structure-enhancement framework rather than a new monolithic model. The expensive stage, including concept-query retrieval, structure-oriented adaptation, and LM-loss-based structure refinement, is performed offline. At inference time, TRACE only retrieves compact tree paths and graph relations and injects them into the prompt, making the online path lightweight and compatible with existing LLM serving stacks.

*Corresponding author.

EMNLP 2026 Industry Track, pages 3836–3853
October 24-29, 2026 ©2026 Association for Computational Linguistics

TRACE represents oncology knowledge as a hierarchical concept tree and a relation graph over the same concept set. The tree captures coarse-to-fine medical organization, while the graph captures cross-branch clinical relations. The structure is refined using language-model evidence and then reused for label-free zero-shot inference, supervised fine-tuning, and deployment-oriented evidence-budget control.

Our contributions are:

- We introduce TRACE as a deployable tree-relational enhancement framework for oncology LLMs, separating offline structure learning from lightweight online retrieval and prompt augmentation.
- We refine tree nodes and graph edges with LM-loss-derived Bayesian updates, and show through initial-vs-refined structure ablation that this refinement improves knowledge-intensive oncology QA.
- We evaluate TRACE against strong LLMs, vanilla RAG, and generic GraphRAG under the same oncology corpus, showing that the learned tree-relational structure provides gains beyond adding retrieved text.
- We evaluate deployment-relevant properties of TRACE, including leakage-controlled robustness, inference overhead, and offline structure-update cost, showing that TRACE can be integrated as a lightweight online enhancement with offline maintenance.

2 Method

TRACE consists of offline structure learning and online task integration. Unlike retrieval-only GraphRAG, which reads a fixed external graph at inference time, TRACE treats the structure as an offline learnable artifact. Structure-oriented adaptation produces LM-loss evidence, which is written back as node and edge support through a Beta-Bernoulli update; the refined structure then governs subsequent retrieval. Online inference only retrieves relevant tree paths and graph relations for prompt augmentation, without updating the structure on the serving path. This separation allows the structure to be versioned and maintained independently of the serving backbone. Figure 1 summarizes the system.

2.1 Tree-Relational Oncology Structure

Let the oncology concept tree be $\mathcal{T} = (\mathcal{V}_T, \mathcal{E}_T)$, where nodes are oncology concepts and edges are parent-child relations. Each node v is associated with a textual representation x_v, including its concept name, aliases, English translation, node type, and full tree path. Inspired by the thematic grouping of the minimal Common Oncology Data Elements (mCODE) standard (HL7 International Clinical Interoperability Council, 2025), the tree covers major oncology branches such as cancer type, symptoms, pathology, staging, biomarkers, treatment, and outcome.

On top of the tree, TRACE defines a relation graph $\mathcal{G} = (\mathcal{V}_G, \mathcal{E}_G)$, where $\mathcal{V}_G$ shares the same concept inventory. Each edge is written as $e = (u, r, v, w_e)$, where u and v are concepts, r is a clinical relation type, and $w_e \in [0, 1]$ is the edge weight. The initial structure is built from oncology corpora and label-free task-facing input text; answer-like fields and target labels are removed whenever task data are used for structure construction.

To obtain model-side evidence, TRACE uses each node and edge as a concept query to retrieve relevant oncology corpus passages. The retrieved passages are converted into structure-oriented training samples and used to train a lightweight low-rank adaptation (LoRA) adapter, which we refer to as Concept Query LoRA. After this adaptation, canonical node and edge statements are scored by causal language modeling loss. The loss is converted into a relative soft evidence score:

$$q_z = \sigma\left(\frac{\ell_{\text{med}} - \ell_z}{\tau}\right),$$

where z denotes a node or edge, ℓ_z is its LM loss, ℓ_{med} is the median loss over elements of the same type, $\sigma(\cdot)$ is the sigmoid function, and τ is a temperature. A higher q_z means the structural statement is better supported by the adapted model relative to other statements of the same type.

2.2 Bayesian Structure Refinement

TRACE keeps the tree topology fixed and estimates a posterior support score for each node. For node v, evidence is aggregated from the node and its descendants:

$$\tilde{q}_v = \max\big(\{q_v\} \cup \{q_u : u \in \text{Desc}(v)\}\big),$$

where $\text{Desc}(v)$ denotes the set of descendants of node v. This allows a strongly supported fine-

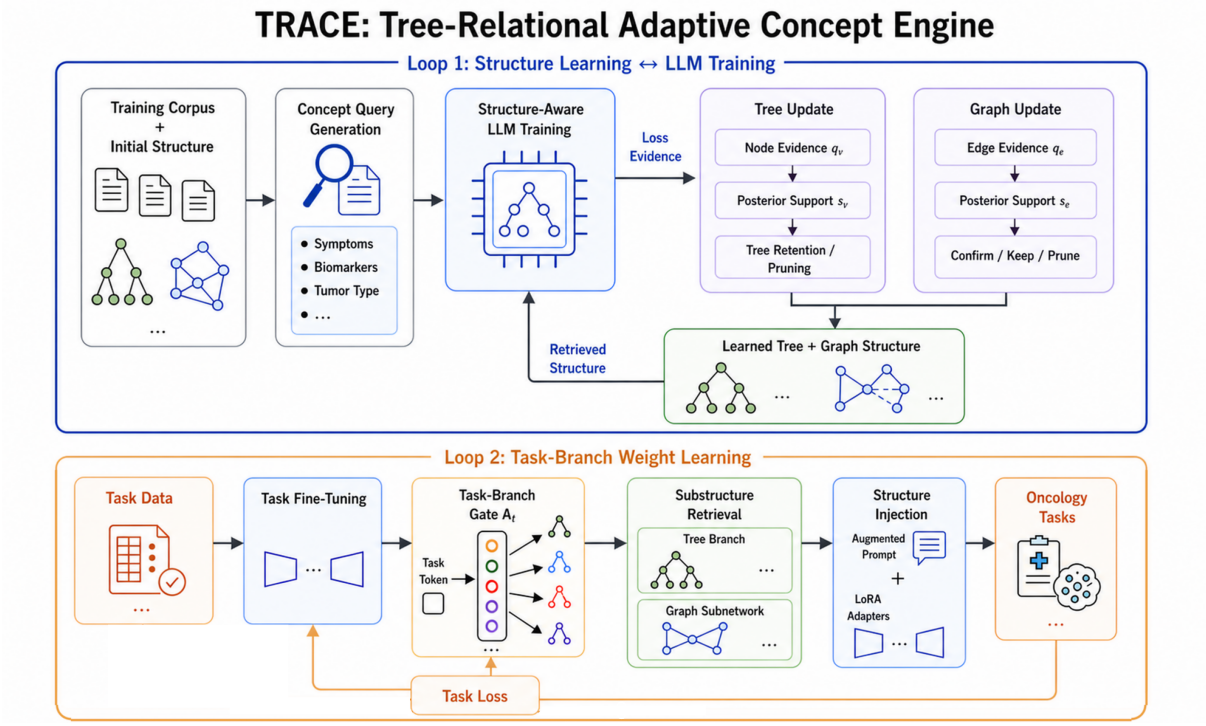


Figure 1: Overview of TRACE. Offline, TRACE retrieves concept-query corpus samples, performs structure-oriented LoRA adaptation, estimates LM-loss evidence, and updates tree nodes and graph edges with Bayesian rules. Online, TRACE retrieves compact tree-relational evidence and injects it into downstream prompts. Supervised settings can additionally learn task-branch weights.

grained concept to also support its ancestors, which define the tree path needed to interpret it.

To update node support using a Beta-Bernoulli Bayesian model, we model node usefulness as a latent Bernoulli variable with a Beta prior distribution $\mathrm{Beta}(\alpha_v, \beta_v)$. Given prior support p_T, prior strength c_T, and smoothing coefficient ρ, the prior pseudo-counts are

$$\alpha_v = 1 + c_T p_T, \qquad \beta_v = 1 + c_T(1 - p_T).$$

The Beta distribution is conjugate to the Bernoulli likelihood, so posterior inference can be performed by updating the two pseudo-counts of supporting and opposing evidence. The LM-derived score $\tilde{q}_v \in [0, 1]$ is treated as fractional positive evidence and $1 - \tilde{q}_v$ as fractional negative evidence:

$$\alpha'_v = \rho\alpha_v + \tilde{q}_v, \qquad \beta'_v = \rho\beta_v + (1 - \tilde{q}_v).$$

The posterior node support is the posterior mean:

$$s_v = \frac{\alpha'_v}{\alpha'_v + \beta'_v}.$$

Nodes with high support are prioritized during retrieval, while low-support nodes remain as background ontology.

Graph edges are updated analogously, but unlike tree nodes, each graph edge already has an initial confidence weight. For edge $e = (u, r, v, w_e)$, where u and v are concepts, r is the relation type, and w_e is the initial edge weight, TRACE initializes a Beta prior from w_e, combines it with LM-derived edge evidence q_e, and obtains the posterior edge support:

$$s_e = \frac{\alpha'_e}{\alpha'_e + \beta'_e}.$$

We use this posterior support as the updated edge weight, i.e., $w'_e = s_e$, for downstream retrieval. Edges with high posterior support are confirmed, uncertain edges are retained as candidates, and weak edges are excluded from downstream retrieval. Detailed update equations and thresholds are provided in Appendix A.

We use Beta-Bernoulli refinement as a bounded and interpretable smoothing mechanism, rather than assuming that it is intrinsically more accurate than deterministic weighting. It represents initial structural confidence as prior pseudo-counts and LM-derived evidence as fractional observations, yielding a closed-form posterior support in $[0, 1]$. Appendix J compares this update with unrefined initial weights and simple linear weighting.

2.3 Tree-Relational RAG and Task Integration

Tree-Relational RAG is the online TRACE component. Given a sample x_i, it retrieves concepts, tree paths, and graph edges using only input-side information such as the question, options, patient description, or serialized clinical fields. Retrieval combines sample matching with updated structural confidence:

$$\mathrm{Score}(x_i, e) = \mathrm{Match}(x_i, e) \cdot w'_e,$$

where $\mathrm{Match}(x_i, e)$ captures lexical overlap, keyword hits, and endpoint matches. The selected evidence is appended as auxiliary context. Classification uses candidate log-likelihood scoring, while generation uses standard decoding.

In supervised fine-tuning, TRACE can also learn task-branch weights:

$$\pi_t = \mathrm{softmax}(A_t),$$

$$\mathrm{Score}_t(x_i, e) = \mathrm{Score}(x_i, e) \cdot \mathrm{BranchRel}_t(e).$$

The branch logits are warm-started from label-free evidence and optimized with LoRA parameters. Zero-shot TRACE does not use learned task-branch weights.

3 Experiments

Datasets and setup. We evaluate TRACE on ten oncology classification tasks and one QA generation benchmark: three Chinese oncology examination QA datasets, two private de-identified clinical-record tasks, five METABRIC-derived breast cancer tasks from cBioPortal `brca_metabric` (Curtis et al., 2012; Pereira et al., 2016), and MedQuAD CancerGov QA (Ben Abacha and Demner-Fushman, 2019). The oncology structure is built from textbooks, clinical guidelines, expert consensus documents, and biomedical literature.

Unless otherwise specified, TRACE uses Qwen3.5-27B as the backbone, with tree-relational evidence injected through task-adaptive retrieval and prompt augmentation. Beyond accuracy, we evaluate robustness to noisy EMR-style inputs, inference latency, and structural evidence traces, since deployment requires stable behavior under imperfect clinical text and bounded serving cost. Dataset statistics, preprocessing, split construction, leakage control, metrics, and hyperparameters are provided in Appendices B, E, and A.

3.1 Zero-Shot Oncology Task Evaluation

We use "zero-shot" to denote the absence of supervised training on target-task labels. This includes plain LLM inference as well as label-free structure-adapted variants: structure learning uses oncology corpora and input-side text only, while retrieved evidence is constructed without task answers.

Table 1 reports label-free zero-shot accuracy on ten oncology tasks. We compare TRACE, instantiated with Qwen3.5-27B as its backbone, with representative open-weight and API-based LLMs. For closed-set tasks, prediction is made by candidate log-likelihood scoring.

Table 1 shows that TRACE achieves the best task-macro average across all ten tasks, improving over Qwen3.5-27B from 0.728 to 0.748. Gains are observed across oncology examination QA, private clinical records, and METABRIC, suggesting that explicit structural evidence benefits both knowledge-intensive QA and structured clinical-record classification.

The results also show that TRACE is not merely benefiting from a stronger base model. DeepSeek-V4-Flash remains the strongest individual baseline on several examination tasks, while Qwen3.5-27B is strong on METABRIC stage and receptor-status prediction. TRACE nevertheless obtains the best overall average by making relevant concepts and relations explicit at inference time. Group-level task-macro F1 results and bootstrap confidence intervals are reported in Appendix F and Appendix G. TRACE has the highest point estimate across all dataset groups and the 10-task macro average, with the clearest separation on examination QA and the overall macro score. Per-task gains over Qwen3.5-27B are reported in Appendix H.

3.2 Fine-Tuning on Public Oncology Benchmarks

We next evaluate whether TRACE remains effective when task supervision is available. We use two public benchmarks: METABRIC-derived breast cancer classification and MedQuAD CancerGov question answering. For METABRIC, the five task training splits are concatenated and shuffled, and each model is fine-tuned on the mixed training set before task-wise evaluation. For CancerGov QA, models are fine-tuned on the public document-level split and evaluated by greedy generation.

Table 2 shows that TRACE improves METABRIC task-macro accuracy from the

Task	Baichuan-M2-32B	Baichuan-M3-235B	Llama-4-Scout	HuatuoGPT-3-32B	Qwen3-32B	DeepSeek-V4-Flash	Qwen3.5-27B	TRACE
Oncology examination QA								
Professional Practice	0.785	0.840	0.774	0.796	0.807	0.840	0.818	**0.906**
Oncology Attending	0.727	0.787	0.628	0.740	0.749	**0.815**	0.781	0.807
Related Knowledge	0.792	0.833	0.657	0.790	0.794	**0.853**	0.836	0.845
Private clinical records								
Common Cancer Dx.	0.701	**0.733**	0.649	0.710	0.718	0.725	0.718	0.725
Breast Stage Pred.	0.453	0.502	0.273	0.498	0.502	0.505	0.491	**0.519**
METABRIC public benchmark								
Stage	0.604	0.602	0.569	0.532	0.556	0.637	**0.729**	0.718
Histology	0.762	0.762	0.761	**0.763**	0.762	0.762	0.762	0.762
ER Status	0.723	0.753	0.739	0.739	0.768	0.714	0.804	**0.823**
PR Status	0.744	0.743	0.534	0.741	0.733	0.696	**0.745**	0.744
Relapse Status	0.624	0.624	0.599	0.564	0.548	**0.631**	0.601	0.627
Group-level macro averages								
Exam QA Avg.	0.768	0.820	0.686	0.775	0.783	0.836	0.811	**0.853**
Private Clinical Records Avg.	0.582	0.617	0.461	0.604	0.610	0.615	0.604	**0.622**
METABRIC Avg.	0.692	0.697	0.640	0.668	0.673	0.688	0.728	**0.735**
Task-macro Avg.	0.693	0.718	0.618	0.687	0.694	0.718	0.728	**0.748**

Table 1: Label-free zero-shot accuracy on ten oncology classification tasks. TRACE uses Qwen3.5-27B as the backbone and injects retrieved tree-relational evidence without supervised training on target-task labels. All baselines are evaluated under the same label-free setting. Each dataset block reports per-task accuracy and a group-level macro average, and the last row reports the macro average over all ten tasks. Bold and underline denote the best and second-best results in each row.

strongest baseline value of 0.761 to 0.797, and task-macro F1 from 0.546 to 0.579. On CancerGov QA, TRACE also obtains the best ROUGE-L and BERTScore; full results are reported in Appendix C. These results indicate that task supervision further benefits from explicit tree-relational evidence.

3.3 Ablation Study

We summarize the main ablations in Table 3. The first block decomposes TRACE into Concept Query LoRA and Tree-Relational RAG. Concept Query LoRA alone gives limited direct task gains because its primary role is to support structure learning, whereas Tree-Relational RAG provides the main zero-shot improvement by injecting retrieved structural evidence. The second block compares TRACE with retrieval baselines built from the same oncology corpus. TRACE outperforms both vanilla passage RAG and generic GraphRAG on the 10-task macro average, showing that the gain is not simply due to adding external context. The third block isolates LM-loss Bayesian refinement by comparing TRACE with the initial unrefined structure. Refinement mainly improves examination QA, suggesting that updated structural reliability is most useful for knowledge-intensive reasoning tasks.

We further ablate structural views in Appendix I. Tree-only, graph-only, and tree-plus-graph variants all improve over the base model, with the combined structure achieving the best 10-task macro accuracy.

Because METABRIC is retrospective, non-target fields may include downstream treatment or outcome variables. We therefore construct a stricter METABRIC setting that removes treatment- and outcome-related fields, including endocrine therapy, chemotherapy, radiotherapy, surgery, recurrence, survival, follow-up, and death-related fields. As shown in Appendix L, TRACE remains best among Qwen3.5-27B, Vanilla RAG, and Generic

Model / Method	Overall Acc.	Overall F1	Task-macro Acc.	Task-macro F1
Qwen3-32B LoRA	0.755	0.389	0.758	0.546
HuatuoGPT-3-32B LoRA	0.748	0.364	0.751	0.526
Baichuan-M2-32B LoRA	0.744	0.359	0.744	0.526
Qwen3.5-27B LoRA	0.757	0.361	0.760	0.528
Llama-4-Scout LoRA	0.758	0.360	0.761	0.528
TRACE	**0.797**	**0.480**	**0.797**	**0.579**

Table 2: Fine-tuning results on the five METABRIC tasks. Overall metrics are computed by pooling all test examples, while task-macro metrics average the corresponding metric across the five tasks. Bold denotes the best result in each column.

Method	Exam QA	Clinical records	METABRIC	10-task Avg.
Qwen3.5-27B	0.811	0.604	0.728	0.728
+ Concept Query LoRA	0.816	0.610	0.714	0.723
+ Tree-Relational RAG	0.847	0.615	**0.736**	0.745
TRACE	**0.853**	**0.622**	0.735	**0.748**
Vanilla RAG	0.817	0.623	0.727	0.733
Generic GraphRAG	0.828	**0.624**	0.725	0.736
Initial structure	0.838	**0.624**	**0.738**	0.745
Refined structure	**0.853**	0.622	0.735	**0.748**

Table 3: Main ablation summary. All numbers are group-level or 10-task macro accuracy. Bold denotes the best result within each comparison block.

GraphRAG, improving task-macro accuracy to 0.729 and task-macro F1 to 0.527. This suggests that TRACE does not rely solely on reverse-causal treatment or outcome shortcuts.

3.4 Deployment-Oriented Evaluation

TRACE separates offline structure maintenance from online serving. For the 300-concept structure, one complete refinement round takes approximately 97 minutes, or 3.1–3.3 H200 GPU-hours; most of this cost comes from LoRA training rather than Bayesian write-back. The write-back scales as $O(|\mathcal{V}| + |\mathcal{E}|)$, and a linear cost projection estimates approximately 28 and 110 GPU-hours for structures containing 5,000 and 20,000 elements, respectively, although these larger scales were not empirically profiled. Online structural retrieval has a median latency of 2.71 ms, while prompt augmentation increases end-to-end latency from 305.0 ms to 404.2 ms in direct HF/PEFT inference. Detailed measurements and assumptions are reported in Appendices A and D.

3.5 Structural Evidence Case Study

Appendix M provides two representative cases. In a lung-cancer diagnostic workup question, TRACE retrieves tree evidence covering symptoms, smoking history, thoracic cancer, and lung cancer, together with graph relations linking these clues to bronchoscopy. In a METABRIC stage-prediction example, where treatment and outcome fields are removed, TRACE retrieves staging-related tree nodes and graph relations that connect tumor size and lymph-node status to early-stage breast cancer. These cases illustrate the auditability of TRACE: tree paths expose the relevant clinical hierarchy, while graph edges identify the relations that support the final prediction.

4 Conclusion

We presented TRACE, a deployable tree-relational enhancement framework for oncology LLMs. TRACE learns and refines an explicit oncology structure offline, then performs lightweight online retrieval and prompt augmentation for downstream prediction. Across zero-shot, fine-tuning, retrieval-baseline, leakage-controlled, and deployment-oriented evaluations, TRACE improves accuracy while preserving interpretable structural evidence. The results suggest that explicit, updatable medical structure can make oncology LLMs more accurate, auditable, and practical for constrained clinical deployment settings.

Limitations

TRACE does not assume that a complete medical ontology is available, since its initial structure is constructed from domain corpora. However, the current refinement stage only recalibrates represented nodes and edges and does not itself recover elements absent from the candidate structure. Structural coverage can nevertheless be extended offline by rerunning or incrementally updating candidate extraction as the corpus evolves, consistent with prior incremental concept-graph learning from text (Lai and Yin, 2024); this expansion remains separate from the current confidence-refinement stage.

TRACE also uses prompt-based evidence injection, whose effect varies across backbones. Some

models benefit consistently from retrieved structure, while others show negative transfer on certain tasks, suggesting that evidence formatting and calibration require further study.

Our evaluation is limited to oncology tasks, and we do not claim cross-specialty generalization. Moreover, the retrieved evidence has not been systematically assessed by clinical experts for correctness, relevance, harmful omissions, or safety. Our results therefore establish benchmark performance and evidence inspectability, rather than prospective clinical validity.

Ethical Considerations

Private clinical-record tasks are constructed from de-identified records, and task-facing data used for structure construction or retrieval exclude labels, reference answers, outputs, and target fields to reduce answer leakage. Because retrieved evidence and model predictions may still be incomplete, outdated, or incorrect, any real-world use would require institutional approval, privacy and security review, prospective validation, workflow integration studies, and final review by qualified clinicians. TRACE should be viewed as a decision-support system rather than a substitute for clinical judgment.

References

Baichuan-M2 Team, Chengfeng Dou, Chong Liu, Fan Yang, Fei Li, Jiyuan Jia, Mingyang Chen, Qiang Ju, Shuai Wang, Shunya Dang, Tianpeng Li, Xiangrong Zeng, Yijie Zhou, Chenzheng Zhu, Da Pan, Fei Deng, Guangwei Ai, Guosheng Dong, Hongda Zhang, and 15 others. 2025. Baichuan-m2: Scaling medical capability with large verifier system. *Preprint*, arXiv:2509.02208.

Baichuan-M3 Team, Chengfeng Dou, Fan Yang, Fei Li, Jiyuan Jia, Qiang Ju, Shuai Wang, Tianpeng Li, Xiangrong Zeng, Yijie Zhou, Hongda Zhang, Jinyang Tai, Linzhuang Sun, Peidong Guo, Yichuan Mo, Xiaochuan Wang, Hengfu Cui, and Zhishou Zhang. 2026. Baichuan-m3: Modeling clinical inquiry for reliable medical decision-making. *Preprint*, arXiv:2602.06570.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinformatics*, 20(1):511.

Debyani Chakravarty, Jianjiong Gao, Sarah Phillips, Ritika Kundra, Hongxin Zhang, Jiaojiao Wang, Julia E. Rudolph, Rona Yaeger, Tara Soumerai, Moriah H. Nissan, Matthew T. Chang, Sarat Chandarlapaty, Tiffany A. Traina, Paul K. Paik, Alan L. Ho, Feras M. Hantash, Andrew Grupe, Shrujal S. Baxi, Margaret K. Callahan, and 29 others. 2017. OncoKB: A precision oncology knowledge base. *JCO Precision Oncology*, (1):1–16.

Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1).

Christina Curtis, Sohrab P. Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M. Rueda, Mark J. Dunning, Doug Speed, Andy G. Lynch, Shamith Samarajiwa, Yinyin Yuan, Stefan Gräf, Gavin Ha, Gholamreza Haffari, Ali Bashashati, Roslin Russell, Steven McKinney, METABRIC Group, Anita Langerød, Andrew Green, and 13 others. 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352.

Udiptaman Das, Krishnasai B. Atmakuri, Duy Ho, Chi Lee, and Yugyung Lee. 2026. Clinical knowledge graph construction and evaluation with multi-llms via retrieval-augmented generation. *Preprint*, arXiv:2601.01844.

Malachi Griffith, Nicholas C. Spies, Kilannin Krysiak, Joshua F. McMichael, Adam C. Coffman, Arpad M. Danos, Benjamin J. Ainscough, Cody A. Ramirez, Damian T. Rieke, Lynzey Kujan, Erica K. Barnell, Alex H. Wagner, Zachary L. Skidmore, Amber Wollam, Connor J. Liu, Martin R. Jones, Rachel L. Bilski, Robert Lesurf, Yan-Yang Feng, and 26 others. 2017. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nature Genetics*, 49(2):170–174.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems 37*, pages 59532–59569. Neural Information Processing Systems Foundation.

Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh V. Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. 2024. G-Retriever: Retrieval-augmented generation for textual graph understanding and question answering. In *Advances in Neural Information Processing Systems 37*, pages 132876–132907. Neural Information Processing Systems Foundation.

HL7 International Clinical Interoperability Council. 2025. Minimal common oncology data elements (mcode) implementation guide, version 4.0.0. Current published version; accessed: 2026-06-15.

Jizheng Lai and Jianxin Yin. 2024. Learning conditional dependence graph for concepts via matrix normal graphical model. *Statistics and Its Interface*, 17(2):187–198.

Mufei Li, Siqi Miao, and Pan Li. 2025. Simple is effective: The roles of graphs and large language models in knowledge-graph-based retrieval-augmented generation. In *International Conference on Learning Representations*, volume 2025, pages 6061–6089.

Fenglin Liu, Zheng Li, Hongjian Zhou, Qingyu Yin, Jingfeng Yang, Xianfeng Tang, Chen Luo, Ming Zeng, Haoming Jiang, Yifan Gao, Priyanka Nigam, Sreyashi Nag, Bing Yin, Yining Hua, Xuan Zhou, Omid Rohanian, Anshul Thakur, Lei Clifton, and David A. Clifton. 2024. Large language models are poor clinical decision-makers: A comprehensive benchmark. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13696–13710. Association for Computational Linguistics.

Bernard Pereira, Suet-Feung Chin, Oscar M. Rueda, Hans-Kristian Moen Vollan, Elena Provenzano, Helen A. Bardwell, Michelle Pugh, Linda Jones, Roslin Russell, Stephen-John Sammut, Dana W. Y. Tsui, Bin Liu, Sarah-Jane Dawson, Jean Abraham, Helen Northen, John F. Peden, Abhik Mukherjee, Gulisa Turashvili, Andrew R. Green, and 15 others. 2016. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nature Communications*, 7:11479.

Xiaohan Ren, Chenxiao Fan, Wenyin Ma, Hongliang He, Chongming Gao, Xiaoyan Zhao, and Fuli Feng. 2026. Medical reasoning with large language models: A survey and mr-bench. *Preprint*, arXiv:2604.08559.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and 13 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Bingning Wang, Haizhou Zhao, Huozhi Zhou, Liang Song, Mingyu Xu, Wei Cheng, Xiangrong Zeng, Yupeng Zhang, Yuqi Huo, Zecheng Wang, Zhengyun Zhao, Da Pan, Fei Kou, Fei Li, Fuzhong Chen, Guosheng Dong, Han Liu, Hongda Zhang, Jin He, and 23 others. 2025. Baichuan-m1: Pushing the medical capability of large language models. *Preprint*, arXiv:2502.12671.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, Yueming Jin, and Vicente Grau. 2025. Medical graph RAG: Evidence-based medical large language model via graph retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28443–28467. Association for Computational Linguistics.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251. Association for Computational Linguistics.

Jiasheng Xu, Mingda Li, Yongqiang Tang, Peijie Wang, and Wensheng Zhang. 2026. Towards open-world retrieval-augmented generation on knowledge graph: A multi-agent collaboration framework. In *Proceedings of the ACM Web Conference 2026*, pages 2171–2182. Association for Computing Machinery.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. HuatuoGPT, towards taming language model to be a doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10859–10885. Association for Computational Linguistics.

A Implementation Details

Main implementation settings. Unless otherwise stated, TRACE uses Qwen3.5-27B as the backbone language model. The learned structure used in the main experiments is the updated 300-concept tree-relational structure. Downstream structure injection uses direct prompt evidence without evidence compression or reranking. The default retrieval budget is six concepts and six graph edges per sample. Supervised task LoRA fine-tuning uses one training epoch, learning rate 2×10^{-5}, batch size 1, gradient accumulation 8, maximum sequence length 1536, bfloat16 precision, and LoRA rank 16 with LoRA scaling parameter $\alpha_{\text{LoRA}} = 32$ and dropout 0.05. The LoRA target modules are all linear layers.

LM-loss evidence. Each tree node and graph edge is verbalized into a canonical causal language modeling sample. For a structural element z, let $x_z = (x_{z,1}, \ldots, x_{z,n})$ be its token sequence and m_j be the attention mask. TRACE computes the token-averaged negative log-likelihood:

$$\ell_z = -\frac{1}{\sum_{j=2}^{n} m_j} \sum_{j=2}^{n} m_j \log p_\theta(x_{z,j} \mid x_{z,<j}).$$

Here θ denotes the adapted language-model parameters. The loss is converted into a relative evidence score:

$$q_z = \sigma\left(\frac{\ell_{\text{med}} - \ell_z}{\tau}\right),$$

where ℓ_{med} is the median loss over structural elements of the same type, σ is the sigmoid function, and τ is a temperature parameter. This score is used as soft model evidence rather than as a calibrated factuality probability.

Graph update equations. For graph edge $e = (u, r, v, w_e)$, TRACE initializes a Beta prior from the original edge weight:

$$\alpha_e = 1 + c_G w_e, \qquad \beta_e = 1 + c_G(1 - w_e),$$

where c_G controls the prior strength. Given LM-derived evidence q_e, the posterior pseudo-counts are

$$\alpha'_e = \rho\alpha_e + q_e, \qquad \beta'_e = \rho\beta_e + (1 - q_e),$$

where ρ is the smoothing coefficient shared with the tree update. The updated edge weight is the posterior mean:

$$w'_e = \frac{\alpha'_e}{\alpha'_e + \beta'_e}.$$

Edges are then assigned a retrieval status:

$$\text{status}(e) = \begin{cases} \texttt{confirmed}, & w'_e \geq \gamma_{\text{confirm}}, \\ \texttt{candidate}, & \gamma_{\text{keep}} \leq w'_e < \gamma_{\text{confirm}}, \\ \texttt{pruned}, & w'_e < \gamma_{\text{keep}}. \end{cases}$$

where γ_{confirm} and γ_{keep} are confirmation and retention thresholds.

Iterative structure refinement. At round r, TRACE starts from the current structure $(\mathcal{T}^{(r-1)}, \mathcal{G}^{(r-1)})$, retrieves concept-query corpus samples, trains or continues the structure-oriented LoRA adapter, and rescores canonical node and edge statements. The Bayesian update then produces $(\mathcal{T}^{(r)}, \mathcal{G}^{(r)})$, which is used for the next round.

Prompt evidence format. For downstream prediction, TRACE appends a concise evidence block to the input. The block contains retrieved tree nodes with their paths and retrieved graph edges with relation type, edge weight, and evidence text:

$$\begin{aligned} \texttt{[Relevant nodes]:} \quad & v,\ \text{path}(v) \\ \texttt{[Relevant edges]:} \quad & u \texttt{ -}r\texttt{-> } v,\ w'_e, \\ & \text{evidence}(e). \end{aligned}$$

The evidence is treated as auxiliary context, while the original task input remains the primary source. In the main experiments, we use the direct evidence format without compressed evidence or reranking.

Task prompt templates. For multiple-choice oncology QA, the prompt contains the question, answer options, and the retrieved structural evidence block. The model is instructed to output only the option letter. For closed-set clinical classification tasks, the prompt contains the serialized input fields, candidate labels, and the retrieved evidence block, and the model is instructed to output exactly one label. In both settings, the prompt explicitly treats structural evidence as auxiliary context: when the retrieved evidence conflicts with the original task input, the original input takes priority.

The multiple-choice prompt follows the template:

> **Question:** ⟨question stem⟩
>
> **Options:** ⟨candidate options⟩
>
> **Relevant structural evidence:** ⟨retrieved tree paths and graph edges⟩
>
> Please answer with only the option letter.

The closed-set classification prompt follows the template:

Patient / sample information: ⟨serialized non-target fields⟩

Relevant structural evidence: ⟨retrieved tree paths and graph edges⟩

Candidate labels: ⟨label set⟩

Please output exactly one candidate label.

Task fine-tuning configuration. Unless otherwise specified, supervised task LoRA fine-tuning uses one training epoch, learning rate 2×10^{-5}, batch size 1, gradient accumulation 8, maximum sequence length 1536, bfloat16 precision, and LoRA rank 16 with LoRA scaling parameter $\alpha_{\mathrm{LoRA}} = 32$ and dropout 0.05. The LoRA target modules are all linear layers. For direct prompt evidence, the default retrieval budget is six concepts and six graph edges per sample.

Deployment cost of structure construction and update. We report the deployment cost for the original structure used in the main experiments. The initial structure is a finalized artifact containing 300 oncology concepts and 280 relations. Its metadata records that it was constructed from 3,757 answer-blind task/corpus items and 590 oncology literature items, with answer, label, reference, candidate, and evidence fields excluded to reduce leakage risk. Since the original construction logs were not retained as separately profiled GPU jobs, we report its scale and provenance rather than estimating GPU-hours for initial construction.

Starting from this finalized structure, we profile one LM-loss Bayesian refinement round on H200 GPUs. The refinement consists of retrieval and update-data preparation, LoRA warmup training, LM-loss scoring, and Bayesian tree/edge weight updates. As shown in Table 4, one refinement round takes about 97 minutes end-to-end, with an effective compute cost of approximately 3.1–3.3 GPU-hours. Most of the cost comes from LoRA warmup training, while retrieval, scoring, and Bayesian updates are comparatively lightweight. This cost is incurred offline and only when the structure is refreshed.

The Bayesian update itself scales linearly with the number of structural elements and can be parallelized across independent branches. Under the same evidence budget and measured per-element processing rate, a linear cost projection gives approximately 28 GPU-hours for 5,000 structural elements and 110 GPU-hours for 20,000 elements. These values are analytical projections rather than measurements on structures of those sizes. In a hospital deployment, specialty- or department-level branches can be maintained as separate offline jobs, avoiding the need to retrain a single monolithic medical structure after every local change.

Inference latency. We profile TRACE inference without vLLM using HuggingFace Transformers and PEFT to directly load Qwen3.5-27B with the LoRA adapter on a single GPU. The profiling uses 100 samples from the three examination QA tasks and two clinical-record tasks, with 20 samples per task, and generates one output token for each sample. Structural retrieval itself is lightweight, with a median latency of 2.71 ms and a P95 latency of 7.43 ms. End-to-end single-sample latency increases from 304.99 ms for the base prompt to 404.24 ms with TRACE evidence, corresponding to a median overhead of 101.50 ms. This overhead mainly comes from prompt expansion: TRACE adds a median of 610 prompt tokens, increasing the median prompt length from 147.5 to 795.5 tokens. These results show that TRACE shifts structure learning offline, while inference-time cost is dominated by the additional structural evidence tokens rather than retrieval. As shown in Table 5, retrieval adds only 2.71 ms median latency, while most overhead comes from the additional structural evidence tokens.

B Dataset Details

Structure construction corpus. The oncology structure is constructed from a heterogeneous corpus of oncology resources, including clinical oncology textbooks, internal medicine and pathology textbooks, clinical guidelines, expert consensus documents, and biomedical literature. The corpus contains 978 files with a total size of approximately 2.8 GB. The main sources include Chinese oncology textbooks such as *Modern Oncology*, *Precision Oncology*, an oncology graduate textbook, guideline-oriented resources derived from Chinese Society of Clinical Oncology (CSCO) and National Comprehensive Cancer Network (NCCN) guidelines, and their reference collections. When task-facing text is used as auxiliary corpus for structure construction, we remove answer-like fields, including gold labels, references, outputs, and target annotations.

METABRIC processing. We derive five breast cancer classification tasks from METABRIC through cBioPortal `brca_metabric` (Curtis et al.,

Stage	Input	Output	GPUs	Wall-clock	GPU-hours
Initial structure construction	Answer-blind task/corpus text and oncology literature	300 concepts / 280 relations	N/A	Not separately logged	N/A
Retrieval and update-data preparation	300 concepts / 280 relations	LM-loss training/scoring corpus	2	3.3 min	0.11
LoRA warmup training	LM-loss corpus	Structure-adapted LoRA adapter	2	87.8 min	2.93
LM-loss scoring and Bayesian update	Initial tree/graph and LoRA adapter	Updated tree/graph weights	1	6.1 min	0.10
Total refinement	300 concepts / 280 relations	Refined structure	–	97 min	3.1–3.3

Table 4: Deployment cost of structure construction and one LM-loss Bayesian refinement round. The initial structure is reported by scale and provenance because its original construction logs were not retained as separately profiled GPU jobs. GPU-hours are computed from measured wall-clock time and the number of GPUs used in each profiled stage.

Metric	Median	P90	P95	Mean
Retrieval latency (ms)	2.71	5.54	7.43	3.43
Base latency (ms)	304.99	439.31	517.84	338.01
TRACE latency (ms)	404.24	619.67	669.72	447.14
TRACE overhead (ms)	101.50	137.89	236.90	109.12
Base prompt tokens	147.5	630.4	848.3	328.9
TRACE prompt tokens	795.5	1230.2	1438.6	942.7
Extra prompt tokens	610.0	652.0	666.0	613.8

Table 5: Inference latency of TRACE under direct HuggingFace Transformers + PEFT inference. Statistics are computed over 100 samples from the three examination QA tasks and two clinical-record tasks, with 20 samples per task, using Qwen3.5-27B with the LoRA adapter on a single GPU and `max_new_tokens` $= 1$. TRACE latency includes structural retrieval and prompt-augmented model inference.

2012; Pereira et al., 2016). For each task, the target field is withheld from the input, and the remaining clinical, pathological, molecular, treatment, and outcome fields are serialized as the model input. This produces separate tasks for stage, histology, ER status, PR status, and relapse status prediction.

METABRIC input-field control. For each METABRIC-derived task, the target field is removed from the serialized input. In the default retrospective setting, all remaining non-target structured fields are retained. In the leakage-controlled setting, we additionally remove treatment-related and outcome-related fields, including endocrine therapy, chemotherapy, radiotherapy, surgery, recurrence, relapse, survival, follow-up time, death status, and other outcome annotations. Details are shown in Table 7. This filtered setting is used only for the leakage-control analysis in Table 19.

CancerGov QA processing. We use the MedQuAD CancerGov subset (Ben Abacha and Demner-Fushman, 2019). The XML files are parsed into question-answer pairs, and the split is performed at the document level to avoid placing questions from the same source document in both training and test sets. This yields 591 training and 138 test examples.

Leakage control. For zero-shot structure construction and evidence retrieval, TRACE uses only input-side information. In classification tasks, label fields and target annotations are excluded from the prompt and from structure construction. In METABRIC tasks, the predicted target field is removed from the serialized input. In CancerGov QA, the answer text is used only as the supervised generation target during fine-tuning, not as retrieval evidence at test time. For supervised fine-tuning experiments, task examples are used only for downstream model training and evaluation; they are not used to construct the concept inventory or relation graph. At inference time, the current sample provides only the retrieval query, while its target field, label, or reference answer remains unavailable.

Data governance and availability. The private Chinese clinical-record datasets are de-identified before use and are not publicly released due to institutional and privacy constraints. Apart from these private clinical-record tasks, the training data and task data used in this work can be made publicly available, including the public examination-style oncology QA datasets, METABRIC-derived tasks, MedQuAD CancerGov QA data, and the non-private resources used for structure construc-

Dataset	# examples	Description
Oncology examination QA		
Professional Practice	181	Chinese examination-style oncology questions.
Oncology Attending Physician	1,581	Chinese oncology attending-physician examination questions.
Related Professional Knowledge	708	Chinese medical professional-knowledge examination questions.
Private de-identified clinical records		
Common Cancer Diagnosis	131	De-identified clinical records labeled with major cancer type.
Breast Cancer Stage Prediction	289	De-identified breast cancer records labeled with stage.
METABRIC public benchmark		
METABRIC Stage	1,788	Stage prediction from non-target METABRIC fields.
METABRIC Histology	2,374	Histology prediction from non-target METABRIC fields.
METABRIC ER Status	2,469	ER-status prediction from non-target METABRIC fields.
METABRIC PR Status	1,980	PR-status prediction from non-target METABRIC fields.
METABRIC Relapse Status	2,488	Relapse-status prediction from non-target METABRIC fields.
CancerGov QA generation benchmark		
MedQuAD CancerGov	729	Cancer question-answer pairs parsed from 116 XML files.

Table 6: Dataset summary. METABRIC tasks are derived from cBioPortal `brca_metabric`; CancerGov QA is derived from the MedQuAD CancerGov subset.

Setting	Input fields
Default METABRIC	All non-target structured fields, including clinical, pathological, molecular, treatment, and outcome fields.
Leakage-controlled	Target field removed as before; treatment and outcome fields are additionally removed, including endocrine therapy, chemotherapy, radiotherapy, surgery, recurrence, survival, follow-up, and death-related fields.

Table 7: Input-field control for METABRIC-derived tasks.

Model / Method	ROUGE-L	BERTScore Tiny	BERTScore RoBERTa
Baichuan-M2-32B LoRA	0.175	0.809	0.847
HuatuoGPT-3-32B LoRA	0.135	0.791	0.835
Qwen3-32B LoRA	0.155	0.805	0.845
Qwen3.5-27B LoRA	0.254	0.828	0.872
Llama-4-Scout LoRA	0.201	0.823	0.860
TRACE	**0.278**	**0.848**	**0.884**

Table 8: Fine-tuning results on MedQuAD CancerGov QA generation. Models are evaluated by ROUGE-L and BERTScore. Bold denotes the best result in each column.

tion, subject to the license and access policy of each source. TRACE is designed as a research system for oncology language-model evaluation and decision support, not as an autonomous clinical decision-making system.

C CancerGov QA Fine-Tuning Results

Table 8 reports supervised fine-tuning results on the MedQuAD CancerGov QA generation benchmark. Models are fine-tuned on the public document-level training split and evaluated with greedy generation on the held-out test split. We report ROUGE-L as a lexical overlap metric and BERTScore as a semantic similarity metric.

D Additional Deployment-Oriented Evaluation

We include two additional deployment-oriented evaluations. First, we test robustness under noisy EMR-style inputs, where clinical notes are corrupted with abbreviations, token deletion, word-order perturbation, and irrelevant note fragments while preserving the gold label. Second, we sweep the number of retrieved structural evidence items to quantify the accuracy-latency tradeoff. These experiments are intended to characterize operational behavior rather than replace the main zero-shot and fine-tuning evaluations.

Table 9 shows that structural evidence improves robustness under noisy clinical inputs. Under severe corruption, LoRA-only drops to 0.6386 accuracy, while Tree-Relational RAG and TRACE reach 0.6575 and 0.6618. TRACE obtains the best accuracy and macro-F1 under all three noise levels, although Tree-Relational RAG provides a cheaper robustness gain with fewer prompt tokens.

Table 10 shows that the evidence budget controls

Noise	Method	Acc.	Macro-F1	Retention	Tokens	Δ Latency (ms)
Clean	LoRA-only	0.6779	0.5033	1.0000	392.6	0.0
Clean	Tree-Relational RAG	0.6840	0.5030	1.0000	644.3	+590.6
Clean	TRACE	**0.6918**	**0.5063**	1.0000	990.2	+1420.9
Moderate	LoRA-only	0.6686	0.4914	0.9899	390.3	0.0
Moderate	Tree-Relational RAG	0.6775	0.4939	0.9926	643.5	+595.6
Moderate	TRACE	**0.6921**	**0.5101**	**1.0022**	990.2	+1425.7
Severe	LoRA-only	0.6386	0.4648	0.9437	385.4	0.0
Severe	Tree-Relational RAG	0.6575	0.4714	**0.9614**	638.9	+595.7
Severe	TRACE	**0.6618**	**0.4752**	0.9568	986.6	+1425.5

Table 9: Robustness under EMR-like input noise. Retention is computed relative to each method's clean accuracy. Latency overhead is measured relative to LoRA-only inference under the same noise condition.

Top-k	Acc.	Macro-F1	Tokens	Latency (ms)
0	0.7213	0.5296	186.6	698.2
2	0.7282	0.5278	355.7	1184.1
4	0.7313	0.5317	494.7	1592.3
8	0.7371	0.5330	709.7	2231.2
12	0.7414	0.5393	819.2	2547.3
16	**0.7449**	**0.5408**	871.1	2701.7

Table 10: Accuracy-latency tradeoff as the number of retrieved structural evidence items increases. Larger evidence budgets improve accuracy but increase prompt length and inference latency.

the accuracy-latency tradeoff. Accuracy improves monotonically from 0.7213 at $k = 0$ to 0.7449 at $k = 16$, while latency also increases. The improvement begins to saturate after $k = 8$, suggesting that moderate evidence budgets can recover most of the benefit with lower deployment cost. Latency numbers in this section are measured within each deployment-oriented experiment and should not be directly compared with the single-token latency microbenchmark, because the sampled inputs, prompt lengths, and evaluation settings differ.

E Evaluation Protocol and Metrics

For closed-set classification tasks, each candidate label is scored by conditional log-likelihood under the model, and the prediction is the candidate with the highest score. We use token-averaged log-likelihood when candidate labels contain multiple tokens:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} \frac{1}{|y|} \sum_{j=1}^{|y|} \log p_\theta(y_j \mid x, y_{<j}),$$

where x is the task input and $\mathcal{Y}$ is the candidate label set. Accuracy is computed separately for each task. Task-macro accuracy averages task accuracies without weighting by dataset size, whereas overall accuracy pools all test examples before computing accuracy. Macro-F1 is computed within each task, and task-macro F1 averages macro-F1 across tasks. For CancerGov QA generation, we report ROUGE-L and BERTScore as complementary lexical and semantic generation metrics.

F Additional Zero-Shot F1 Results

Table 11 reports the group-level task-macro F1 corresponding to the zero-shot accuracy results in Table 1.

Model / Method	Exam QA	Clinical records	METABRIC	10-task Macro Avg.
Baichuan-M2-32B	0.725	0.370	0.463	0.523
Baichuan-M3-235B	0.774	0.396	0.453	0.538
Llama-4-Scout	0.677	0.331	0.326	0.432
HuatuoGPT-3-32B	0.731	0.463	0.441	0.532
Qwen3-32B	0.739	0.469	0.452	0.542
DeepSeek-V4-Flash	0.790	0.474	0.480	0.572
Qwen3.5-27B	0.765	0.460	**0.513**	0.578
TRACE	**0.803**	**0.480**	0.506	**0.590**

Table 11: Group-level task-macro F1 for label-free zero-shot evaluation. Each group column averages macro-F1 over tasks in the corresponding dataset group, and the last column averages macro-F1 over all ten tasks. Bold denotes the best result in each column.

G Bootstrap Confidence Intervals for Main Results

We additionally report bootstrap confidence intervals in Table 12 for the zero-shot main results. For

each method, we perform task-stratified bootstrap resampling: examples are resampled within each task, task-level accuracies are recomputed, and group-level and overall task-macro averages are then computed from the resampled task accuracies. We use 10,000 bootstrap samples and report 95% percentile confidence intervals. This procedure matches the task-macro evaluation protocol used in Table 1.

H Full Zero-Shot Analysis

Table 13 reports the per-task gain of TRACE over Qwen3.5-27B. TRACE improves most tasks, with the largest gains on Professional Practice, Breast Stage Prediction, relapse status, and ER status. Negative or near-zero changes occur on METABRIC stage and PR status, where Qwen3.5-27B is already a strong baseline.

I Tree–Graph Ablation

Table 14 further shows that both structural views contribute useful information. Tree-only retrieval improves over the base model by providing hierarchical concept grounding, while graph-only retrieval gives larger average gains by exposing cross-branch clinical relations. Their combination achieves the best macro average, indicating that hierarchy and relation evidence are complementary.

J Additional Ablations

Structure update rule ablation. Table 15 compares three ways of combining initial structural confidence with LM-derived evidence: retaining the initial weights without refinement, simple linear interpolation, and the Beta-Bernoulli update used by TRACE. All variants use the same structure, retrieved evidence, and downstream inference procedure.

Backbone-level Tree-Relational RAG ablation. Table 16 applies the same learned structure to different backbone models through the same Tree-Relational RAG interface. The 10-task macro average improves for Baichuan-M2-32B, Baichuan-M3-235B, Llama-4-Scout, HuatuoGPT-3-32B, and Qwen3-32B. This cross-backbone improvement provides direct evidence that the learned tree-relational structure contains reusable oncology knowledge, rather than only fitting the Qwen3.5-27B backbone. DeepSeek-V4-Flash decreases on METABRIC, indicating that the benefit still depends on the backbone's calibration and sensitivity to prompt evidence.

Task-branch weighting ablation. We further ablate task-branch weighting in METABRIC mixed fine-tuning. Table 17 follows the same metric definitions as Table 2. Removing task-branch weighting reduces overall accuracy from 0.7969 to 0.7614 and task-macro F1 from 0.5792 to 0.5376. This suggests that supervised task-branch weighting helps TRACE select task-appropriate structural evidence when multiple METABRIC tasks share the same learned oncology structure.

K Retrieval Baseline F1 Comparison

Table 18 reports the macro-F1 comparison between Vanilla RAG and Generic GraphRAG on the three examination QA tasks and two clinical-record tasks. Both retrieval baselines use the same oncology corpus as external evidence. Generic GraphRAG improves the task-macro F1 over Vanilla RAG, suggesting that graph-structured retrieval provides more useful evidence organization than passage retrieval alone.

L Leakage-Controlled METABRIC Analysis

METABRIC is a retrospective structured-record benchmark, and non-target fields may include variables that are downstream of the prediction target, such as treatment or outcome information. To test whether TRACE mainly benefits from such reverse-causal shortcuts, we construct a stricter METABRIC variant by removing treatment-related and outcome-related fields from the serialized input, including endocrine therapy, chemotherapy, radiotherapy, surgery, recurrence, survival, follow-up, and death-related fields. The target label is withheld as in the main evaluation.

Table 19 shows that removing these fields lowers the absolute performance of several methods, confirming that treatment and outcome variables are informative in the original retrospective records. However, TRACE still achieves the best task-macro accuracy and macro-F1. It improves task-macro accuracy from 0.704 for Qwen3.5-27B and 0.707 for Vanilla RAG to 0.729, and improves task-macro F1 from 0.497 to 0.527. These results indicate that TRACE does not rely solely on treatment or outcome shortcuts; the learned tree-relational evidence remains useful under a stricter input setting.

Method	Exam QA Acc. (95% CI)	Clinical records Acc. (95% CI)	METABRIC Acc. (95% CI)	10-task Macro Acc. (95% CI)
Baichuan-M2-32B	0.768 [0.744, 0.791]	0.582 [0.533, 0.630]	0.692 [0.683, 0.700]	0.693 [0.680, 0.705]
Baichuan-M3-235B	0.820 [0.798, 0.840]	0.617 [0.568, 0.664]	0.697 [0.688, 0.706]	0.718 [0.706, 0.730]
Llama-4-Scout	0.686 [0.661, 0.710]	0.461 [0.413, 0.509]	0.640 [0.631, 0.649]	0.618 [0.605, 0.631]
HuatuoGPT-3-32B	0.775 [0.751, 0.798]	0.604 [0.555, 0.652]	0.668 [0.659, 0.676]	0.687 [0.674, 0.700]
Qwen3-32B	0.783 [0.760, 0.805]	0.610 [0.561, 0.658]	0.673 [0.665, 0.682]	0.694 [0.681, 0.706]
DeepSeek-V4-Flash	0.836 [0.815, 0.856]	0.615 [0.567, 0.663]	0.688 [0.679, 0.697]	0.718 [0.706, 0.730]
Qwen3.5-27B	0.811 [0.789, 0.833]	0.604 [0.555, 0.652]	0.728 [0.720, 0.736]	0.728 [0.716, 0.741]
TRACE	**0.853 [0.834, 0.870]**	**0.622 [0.574, 0.670]**	**0.735 [0.727, 0.743]**	**0.748 [0.736, 0.759]**

Table 12: Bootstrap confidence intervals for the main zero-shot accuracy results. Intervals are computed with 10,000 task-stratified bootstrap samples. Each group score is the macro average over tasks in that group, and the final column is the macro average over all ten tasks. Bold denotes the best point estimate in each column.

Task	Qwen3.5 -27B	TRACE	Δ
Professional Practice	0.818	0.906	+0.088
Oncology Attending	0.781	0.807	+0.026
Related Knowledge	0.836	0.845	+0.009
Common Cancer Dx.	0.718	0.725	+0.007
Breast Stage Pred.	0.491	0.519	+0.028
METABRIC Stage	0.729	0.718	-0.011
METABRIC Histology	0.762	0.762	0.000
METABRIC ER Status	0.804	0.823	+0.019
METABRIC PR Status	0.745	0.744	-0.001
METABRIC Relapse Status	0.601	0.627	+0.026
Task-macro Avg.	0.728	0.748	+0.020

Table 13: Per-task zero-shot accuracy gain of TRACE over Qwen3.5-27B.

M Additional Structural Evidence Case Studies

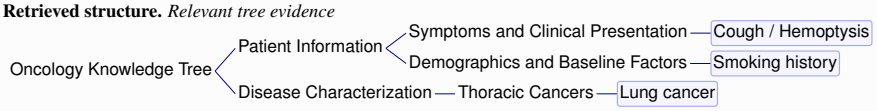
Figure 2 provides two representative examples of TRACE structural evidence. The examples show how TRACE combines local tree paths with key graph relations to support downstream predictions.

Evidence	Prof. Practice	Oncology Attending	Related Knowledge	Common Cancer Dx.	Breast Stage Pred.	5-task Macro Avg.
Base	0.818	0.781	0.836	0.718	0.491	0.729
Tree only	0.823	0.791	0.838	**0.725**	0.516	0.739
Graph only	0.879	**0.807**	0.843	0.710	**0.547**	0.757
Tree+Graph	**0.906**	**0.807**	**0.845**	**0.725**	0.519	**0.760**

Table 14: Tree and graph evidence ablation. All numbers are accuracy. The table compares tree-only, graph-only, and tree-plus-graph evidence, and reports per-task accuracy with a macro average over the five evaluated tasks. Bold denotes the best result in each column.

Case 1: Lung cancer diagnostic workup

Question. For diagnosis, what is the preferred next examination?
Input clue. A 64-year-old male patient has cough with blood-stained sputum, right upper-lobe obstructive inflammation, segmental bronchial obstruction, negative sputum cytology and acid-fast bacillus tests, and a heavy smoking history.
Retrieved structure. *Relevant tree evidence*



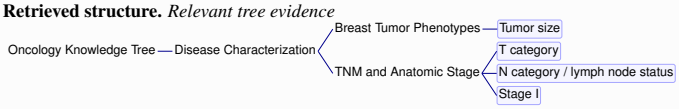
Key graph evidence

- `Cough/Hemoptysis` $\xrightarrow{\text{supports}}$ `Lung cancer`
- `Smoking history` $\xrightarrow{\text{risk_factor_for}}$ `Lung cancer`
- `Bronchial obstruction` $\xrightarrow{\text{supports}}$ `Lung cancer`
- `Lung cancer` $\xrightarrow{\text{confirmed_by}}$ `Fiberoptic bronchoscopy`

Answer. Fiberoptic bronchoscopy.
Prediction effect. TRACE links symptoms and smoking history to suspected lung cancer, then retrieves the diagnostic relation from lung cancer to bronchoscopy.

Case 2: METABRIC breast cancer stage prediction

Question. Predict the patient's breast cancer stage.
Input clue. The patient record describes invasive ductal carcinoma with tumor grade 2, tumor size 18 mm, zero positive lymph nodes, Nottingham prognostic index 3.036, ER-positive, PR-positive, HER2-negative status, and a primary tumor sample. Treatment and outcome fields are removed from the input.
Retrieved structure. *Relevant tree evidence*



Key graph evidence

- `Breast cancer` $\xrightarrow{\text{characterized_by}}$ `Tumor size`
- `Breast cancer` $\xrightarrow{\text{staged_by}}$ `T category`
- `Breast cancer` $\xrightarrow{\text{staged_by}}$ `N category / lymph node status`
- `Positive lymph nodes = 0` $\xrightarrow{\text{supports}}$ `early-stage disease`

Predictions. Baseline: Stage II. TRACE: Stage I. **Answer.** Stage I.
Prediction effect. TRACE retrieves staging-oriented structure that connects tumor size and lymph-node status to early-stage breast cancer.

Figure 2: Structural evidence case studies. TRACE retrieves compact task-relevant substructures, including local tree evidence and key graph relations, that connect input clues to the final prediction.

Update rule	Exam QA	Clinical records	5-task Macro Avg.
Initial weight	0.838	**0.624**	**0.753**
Linear update	0.841	0.615	0.751
Beta-Bernoulli	**0.843**	0.617	0.752

Table 15: Accuracy under different structure-weight update rules. The results are similar across update mechanisms; Beta-Bernoulli is used for its bounded probabilistic interpretation and support for iterative refinement, rather than for a consistent accuracy advantage.

Model	Variant	Oncology examination QA	Private clinical records	METABRIC	10-task Macro Avg.
Baichuan-M2-32B	Base	0.768	0.582	0.692	0.693
	+ Tree-Relational RAG	0.779 (+0.012)	0.578 (-0.004)	0.705 (+0.013)	0.702 (+0.009)
Baichuan-M3-235B	Base	0.820	0.617	0.697	0.718
	+ Tree-Relational RAG	0.845 (+0.025)	0.601 (-0.016)	0.705 (+0.008)	0.726 (+0.008)
Llama-4-Scout	Base	0.686	0.461	0.640	0.618
	+ Tree-Relational RAG	0.785 (+0.099)	0.559 (+0.098)	0.647 (+0.007)	0.671 (+0.053)
HuatuoGPT-3-32B	Base	0.775	0.604	0.668	0.687
	+ Tree-Relational RAG	0.820 (+0.045)	0.603 (-0.001)	0.687 (+0.019)	0.710 (+0.023)
Qwen3-32B	Base	0.783	0.610	0.673	0.694
	+ Tree-Relational RAG	0.821 (+0.038)	0.617 (+0.007)	0.679 (+0.006)	0.709 (+0.015)
DeepSeek-V4-Flash	Base	0.836	0.615	0.688	0.718
	+ Tree-Relational RAG	0.842 (+0.007)	0.618 (+0.003)	0.658 (-0.031)	0.705 (-0.013)

Table 16: Backbone-level ablation of Tree-Relational RAG. Each model is compared with and without the same learned structural retrieval module. Numbers in parentheses indicate the change over the corresponding base model.

Method	Overall Acc.	Overall F1	Task-macro Acc.	Task-macro F1
TRACE w/o task-branch weighting	0.7614	0.3853	0.7628	0.5376
TRACE	**0.7969**	**0.4802**	**0.7968**	**0.5792**

Table 17: Task-branch weighting ablation on METABRIC mixed fine-tuning. Metrics follow the same definition as Table 2.

Method	Prof. Practice	Oncology Attending	Related Knowledge	Common Cancer Dx.	Breast Stage Pred.	Task-macro
Vanilla RAG	0.1380	0.7038	0.8624	0.4729	**0.4546**	0.5263
Generic GraphRAG	0.1365	**0.7286**	**0.8815**	**0.4915**	0.4495	**0.5375**

Table 18: Macro-F1 comparison between retrieval baselines on the three examination QA tasks and two clinical-record tasks. Both baselines use the same oncology corpus as external evidence. Bold denotes the best result in each column.

Method	Stage	Histology	ER	PR	Relapse	Avg.
Accuracy						
Qwen3.5-27B	0.662	0.762	0.755	0.744	0.594	0.704
Vanilla RAG	0.656	0.762	0.758	**0.745**	0.611	0.707
Generic GraphRAG	0.671	0.762	0.737	0.744	0.611	0.705
TRACE	**0.741**	**0.763**	**0.777**	0.744	**0.620**	**0.729**
Macro-F1						
Qwen3.5-27B	0.356	0.108	0.715	0.722	0.583	0.497
Vanilla RAG	0.349	0.108	0.716	0.722	0.581	0.495
Generic GraphRAG	0.360	0.108	0.697	0.722	0.586	0.495
TRACE	**0.405**	**0.192**	**0.728**	0.722	**0.591**	**0.527**

Table 19: Leakage-controlled METABRIC evaluation after removing treatment-related and outcome-related input fields. All methods are evaluated on the same filtered inputs. Avg. reports task-macro average over the five METABRIC-derived tasks. Bold denotes the best result in each column; ties after rounding are left unmarked.