Learning and Injecting Sparse Concept Graphs for Interpretable Clinical Large Language Modeling

Anonymous ACL submission

Abstract

Large language models (LLMs) achieve strong results across NLP tasks, but they often struggle to exploit structured domain knowledge in specialized settings such as clinical text, where medical concepts are linked by sparse conditional dependencies semantically. To address this challenge, we present **GRIAN** (Graph-Regularized and Injected Adaptation Network), which treats the existence of a sparse concept dependency graph as an explicit prior during LLM adaptation. Rather than learning a graph separately, GRIAN integrates graph recovery into training by augmenting the large language modeling objective with a nonparanormal matrix-normal graphical-model loss that jointly estimates sparse precision matrices while optimizing the LLM. The graph-structured term regularizes the model toward parsimonious conditional dependencies, and is further complemented by a Laplacian smoothness regularizer that aligns concept embedding geometry with the emerging structure via parameter-efficient LoRA updates. For downstream prediction, we encode query-conditional induced subgraphs with a graph attention network. Then we inject graph evidence into Transformer attention, enabling structure-grounded and more interpretable reasoning over clinical text. Experiments on a Chinese knee-joint electronic medical record dataset and a medical abstract dataset show consistent improvements over LLM baselines, highlighting the benefits of jointly regularized graph-structure adaptation for reliable and interpretable clinical text modeling.

1 Introduction

Large language models (LLMs) have achieved strong performance across diverse NLP tasks, yet they often struggle to exploit structured domain knowledge in specialized settings such as clinical text. Clinical narratives mention symptoms, examinations, and pathologies in free-form language, but the underlying concepts are linked by sparse and conditional dependencies semantically rather than spurious co-occurrences. When such structure is not captured and employed, models may not provide logical and explainable evidence on the decision. Recent work also shows that clinical deployment of large language models requires not only high accuracy, but also reliable and traceable evidence (Singhal et al., 2023a,b; Liu et al., 2025a).

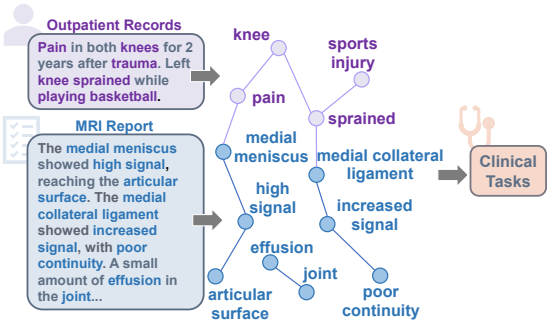

Figure 1: **Concept graph for clinical text.** Clinical text (e.g., outpatient notes and MRI reports) is mapped to a sparse concept subgraph to support downstream clinical tasks.

Figure 1 illustrates why we impose a sparse graph prior during adaptation and why we further inject graph evidence for downstream clinical prediction. Clinical records often provide fragmented cues across sections (e.g., outpatient notes vs. imaging reports). A graph prior encourages the model to explain decisions through conditional concept dependencies, rather than relying on spurious co-occurrence. At inference stage, injecting a query-conditioned subgraph makes the structural evidence explicit and traceable, providing a compact set of concept relations that the model can attend to when making clinical predictions.

LLMs are often augmented with external mem-

ory like RAG (Lewis et al., 2020; Ye et al., 2024) and kNN-LM (Khandelwal et al., 2020) or injected with Knowledge Graphs via input augmentation and adapter-style modules (Liu et al., 2020; Wang et al., 2021). Recent graph–LLM systems further explore tighter fusion between graph structure and neural reasoning (Ma et al., 2024; Hong et al., 2024; Jin et al., 2024; Tian et al., 2024; Sun et al., 2023; Rezaei et al., 2025; Guo et al., 2025). Other studies have also explored schemes for automatically generating knowledge graphs or hypergraph structures from corpora to enhance LLMs (Masoudifard et al., 2024; Dou et al., 2025). However, most of these approaches assume access to existed or pre-defined graphs, and their behavior is ultimately constrained by knowledge graph coverage and noise.

In contrast, we adapt LLMs under an explicit sparse graph prior without relying on any external knowledge graph, by enforcing conditional dependencies among domain concepts during training. In clinical text, correlation graphs can be misleading due to comorbidities and documentation bias, whereas precision matrices capture conditional independence and better disentangle confounded signals. Sparse inverse covariance estimation (Graphical Lasso) enables such recovery for normally distributed random vectors (Friedman et al., 2008). Matrix normal graphical models extend these ideas to matrix-valued observations with row/column precision matrices (Yin and Li, 2012; Lai and Yin, 2024). On the other hand, the nonparanormal model improves robustness to non-Gaussian settings (Liu et al., 2009; Ning and Liu, 2013). These results provide the foundation to combine structure learning and representation learning when adapting LLMs to clinical corpora.

Our approach. We propose **GRIAN** (Graph-Regularized and Injected Adaptation Network), which adapts an LLM under an explicit sparse precision-graph prior over domain concepts. Unlike prior graph-enhanced LLMs that inject a fixed or correlation-based graph, GRIAN is the first framework that learns a sparse precision graph jointly with LLM adaptation and uses it as an explicit training-stage prior, rather than as an external knowledge source. In Stage I, we couple LoRA fine-tuning with a nonparanormal MNGM objective (Ning and Liu, 2013) to recover sparse row/column precision matrices (A, B), while a Laplacian smoothness term encourages graph-consistent concept embeddings (Wan et al., 2023). In Stage II, we encode query-induced subgraphs with a Graph Attention Network (GAT) (Veličković et al., 2018) and inject the resulting graph evidence into Transformer attention, making the supporting concept relations traceable.

Contributions. (1) We propose **GRIAN**, a two-stage framework that learns a sparse concept dependency graph from domain-specific clinical/biomedical text and injects query-conditioned subgraph evidence into an LLM. Rather than using a knowledge graph to enhance an LLM, we learn a sparse precision graph as an explicit prior during training, so that concept relations are encoded as conditional dependencies (instead of co-occurrence) and remain interpretable via the support of the precision matrix. (2) We introduce a nonparanormal MNGM-based graph recovery objective that estimates sparse precision matrices to capture conditional concept dependencies, and integrate it into LLM adaptation via an alternating optimization scheme that decouples the graph step and the embedding step. In graph step, we estimate (A, B) given the concept embeddings, while in embedding step, we update LoRA parameters under a Laplacian smoothness regularizer induced by the last round precision graph A(B is nuisance here, but it can have effect on the sparsity level of A). (3) Empirically, we show that when no task-specific knowledge graph is available, GRIAN can recover a meaningful concept dependency graph directly from data in an unsupervised manner. This learned graph not only improves performance on both the Chinese knee-joint EMR diagnosis task and the medical abstracts text classification task, but also enables interpretable causal-type graph visualizations that expose the conditional-dependency neighborhoods underlying model predictions.

2 Method

In this section, we first introduce the nonparanormal Matrix Normal Graphical Model (MNGM) for learning the keyword adjacency matrix, and subsequently present the GRIAN framework for integrating structured knowledge into LLMs.

2.1 Nonparanormal MNGM

To recover a sparse concept dependency graph from neural embeddings, we model the concept embedding matrix as a matrix-valued observation under a matrix normal graphical model (MNGM) (Yin and Li, 2012). Let $F \in \mathbb{R}^{p \times q}$ denote the embedding matrix of p concepts (rows) with embedding dimen-

sion q (columns). MNGM assumes $\mathrm{vec}(F)$ follows a Gaussian distribution with Kronecker-structured covariance, implying two precision matrices: a row precision $A \in \mathbb{R}^{p\times p}$ that encodes conditional dependencies among concepts, and a column precision $B \in \mathbb{R}^{q\times q}$ that captures correlations across embedding dimensions. Here B models correlations across embedding dimensions, which helps absorb dimension-wise dependence and prevents such effects from being spuriously attributed to the row precision A.

We estimate (A, B) by minimizing a sparsity-regularized negative log-likelihood:

$$\begin{aligned}\phi(A, B; F) = &- q \log |A| - p \log |B| \\ &+ \mathrm{Tr}\Big(AFBF^\top\Big) + \lambda \|A\|_{1,\mathrm{off}} \\ &+ \gamma \|B\|_{1,\mathrm{off}}.\end{aligned} \tag{1}$$

where $\|\cdot\|_{1,\mathrm{off}}$ denotes the ℓ_1 norm of off-diagonal entries, encouraging sparse conditional edges. The learned concept graph is then obtained from the support of A (i.e., $a_{ij} \neq 0$ indicates a conditional dependency between concepts i and j).

Neural embeddings produced by LLMs often deviate from Gaussian marginals, especially in clinical corpora. To improve robustness, we adopt the nonparanormal (semiparametric) extension (Ning and Liu, 2013), which assumes that each column of F is generated from a monotonic transformation of a latent Gaussian variable. In practice, we apply a rank-based Gaussianization transform column-wise to obtain a transformed embedding matrix $\tilde{F}$. This procedure is equivalent to estimating a rank-based correlation matrix $\hat{R}$ (e.g., via Spearman's ρ or Kendall's τ) and replacing the quadratic term in Eq. 1 with $\mathrm{Tr}\Big((B \otimes A)\hat{R}\Big)$ under the nonparanormal model.

We therefore implement the nonparanormal MNGM by optimizing the equivalent objective based on $\tilde{F}$:

$$\begin{aligned}\phi_{\mathrm{np}}(A, B; \tilde{F}) = &- q \log |A| - p \log |B| \\ &+ \mathrm{Tr}\Big(A\tilde{F}B\tilde{F}^\top\Big) + \lambda \|A\|_{1,\mathrm{off}} \\ &+ \gamma \|B\|_{1,\mathrm{off}}.\end{aligned} \tag{2}$$

This formulation preserves the conditional-independence interpretation of precision matrices while improving robustness to non-Gaussian embedding distributions. In practice, (A, B) are updated by alternating minimization with $\tilde{F}$ treated as a fixed observation, and the resulting sparse row precision matrix A serves as the learned concept dependency graph that is jointly optimized with the LLM in GRIAN (Section 2.2).

2.2 GRIAN: Graph-Regularized and Injected Adaptation Network

GRIAN aims to (i) *recover* a sparse conditional-dependency graph among domain concepts directly from in-domain text, and (ii) *inject* compact graph evidence into the LLM in a parameter-efficient way. The framework consists of two coupled stages: (1) graph-regularized adaptation that learns a latent precision-graph prior while fine-tuning the LLM; and (2) query-conditioned graph injection that supplies task-relevant structural signals at inference/training stage.

Concept representations. Let $\mathcal{K} = \{k_1, \ldots, k_p\}$ denote the concept vocabulary (keywords/phrases). For each concept k_i, we obtain a concept embedding by averaging its constituent token embeddings from the LLM input embedding matrix. Stacking these vectors yields $F \in \mathbb{R}^{p\times q}$, where q is the embedding dimension. To model non-Gaussian marginals commonly observed in neural embeddings, we apply a rank-based Gaussianization (nonparanormal) transform $\mathcal{T}(\cdot)$ to each column of F, producing $\tilde{F} = \mathcal{T}(F)$.

Stage I: Alternating graph recovery and graph-regularized LoRA adaptation. Stage I learns a sparse concept dependency graph while adapting the LLM via LoRA, using an alternating optimization scheme with gradient decoupling. Let $F_\theta \in \mathbb{R}^{p\times q}$ denote the concept embedding matrix induced by the current LLM parameters θ (after LoRA updates), where each row corresponds to a domain concept.

Graph step. Given a snapshot of concept embeddings induced by the current LLM parameters, we first apply the rank-based Gaussianization described in Section 2.1 to obtain $\tilde{F}^{(t)} = \mathcal{T}(F_{\theta^{(t)}})$. We then update the graph parameters by solving the nonparanormal MNGM objective:

$$(A^{(t+1)}, B^{(t+1)}) = \arg \min_{A\succ 0,\, B\succ 0} \phi_{\mathrm{np}}\Big(A, B;\, \tilde{F}^{(t)}\Big), \tag{3}$$

where $\tilde{F}^{(t)}$ is treated as a fixed observation and gradients are not propagated to the LLM parameters. The resulting row precision matrix $A^{(t+1)}$ defines a

sparse conditional-dependency graph, from which we construct the Laplacian $L^{(t+1)}$ for the subsequent embedding update.

Embedding step. With the graph structure fixed, we update the LLM parameters θ by minimizing the large language modeling loss regularized by the Laplacian smoothness term:

$$\theta^{(t+1)} \leftarrow \arg\min_{\theta} \mathcal{L}_{\mathrm{LLM}}(\theta) + \beta \operatorname{Tr}\Big(F_\theta^\top L^{(t+1)} F_\theta\Big), \tag{4}$$

where gradients are propagated only through F_θ (via LoRA), while the Laplacian $L^{(t+1)}$ is held fixed. This form follows classical graph-regularized Non-negative Matrix Factorization (NMF) style objectives that minimize $\mathrm{Tr}(F^\top L F)$ to encourage graph-smooth representations (Yin et al., 2016; Wan et al., 2023).

Overall objective. The above alternating procedure can be viewed as optimizing the following unified regularized objective:

$$\begin{aligned}\mathcal{L}_{\text{Stage I}} &= \mathcal{L}_{\mathrm{LLM}}(\theta) + \alpha\, \phi_{\mathrm{np}}(A, B; \mathcal{T}(F_\theta)) \\ &\quad + \beta \operatorname{Tr}\Big(F_\theta^\top L_A F_\theta\Big).\end{aligned} \tag{5}$$

where $\phi_{\mathrm{np}}(\cdot)$ is the nonparanormal matrix-normal graphical-model loss encouraging sparsity in (A, B), and the Laplacian term enforces smoothness of LoRA-updated concept embeddings over the learned graph. In practice, we implement Eq. (5) via the decoupled updates in Eq. (3)–(4), i.e., updating (A, B) with F stop-grad, and updating θ with L stop-grad.

This scheme ensures that the learned graph captures conditional dependencies among concepts given the current LLM representations, while the embedding update simultaneously performs the LoRA language modeling adaptation and enforces graph-consistent geometry.

Stage II: Traceable Graph Evidence Injection at Inference. Stage II serves to expose the sparse conditional-dependency structure learned in Stage I as an explicit and traceable evidence signal during prediction. Building upon the graph-enhanced question answering framework (Zhang et al., 2024), Stage II does not introduce a new graph–Transformer architecture; instead, it provides a lightweight interface that makes the learned precision graph accessible to the LLM at inference time.

Let $\mathcal{G} = (V, E)$ denote the concept dependency graph recovered in Stage I, where nodes correspond to concepts in $\mathcal{K}$ and an edge $(v_i, v_j) \in E$ exists if the estimated row precision matrix has $a_{ij} \neq 0$. Given an input instance X (e.g., an EMR record or a medical abstract), we first identify a set of mentioned concepts $\mathcal{K}_X \subseteq \mathcal{K}$ and construct the induced subgraph $\mathcal{G}_X = (V_X, E_X)$ with $V_X = \{v_i \in V \mid k_i \in \mathcal{K}_X\}$ and $E_X = \{(v_i, v_j) \in E \mid v_i, v_j \in V_X\}$. This subgraph reflects *conditional* concept dependencies rather than surface-level co-occurrence.

We summarize $\mathcal{G}_X$ using a lightweight Graph Attention Network (GAT). Node features are initialized with the concept embeddings learned by the LLM (i.e., rows of F_θ), and a graph-level representation is obtained via global mean pooling on the final node representations, yielding $O^g \in \mathbb{R}^{d_h}$.

The resulting graph summary is injected into the Transformer as additional key–value memory. Using the LLM's native projection matrices, we compute

$$K^g = O^g W_K, \qquad V^g = O^g W_V, \tag{6}$$

and concatenate them with the original key–value pairs:

$$K^{\text{aug}} = [K, K^g], \tag{7}$$

$$V^{\text{aug}} = [V, V^g]. \tag{8}$$

The graph-augmented attention is then defined as

$$\text{Attention}^{\text{graph}} = \text{Softmax}\left(\frac{Q(K^{\text{aug}})^\top}{\sqrt{d}}\right) V^{\text{aug}}, \tag{9}$$

where Q denotes the query vectors and d is the attention dimension.

Because the injected memory is derived directly from the sparse precision graph, each prediction can be traced back to the specific concepts and conditional dependencies (non-zero a_{ij}) involved in $\mathcal{G}_Q$.

3 Experiments

3.1 Electronic Medical Record Dataset

We collected Chinese electronic medical records from 5,825 patients diagnosed with knee disorders at the Sports Medicine Department of Peking University Third Hospital. Each record comprises three components text: outpatient notes, MRI reports, and physical examination findings.

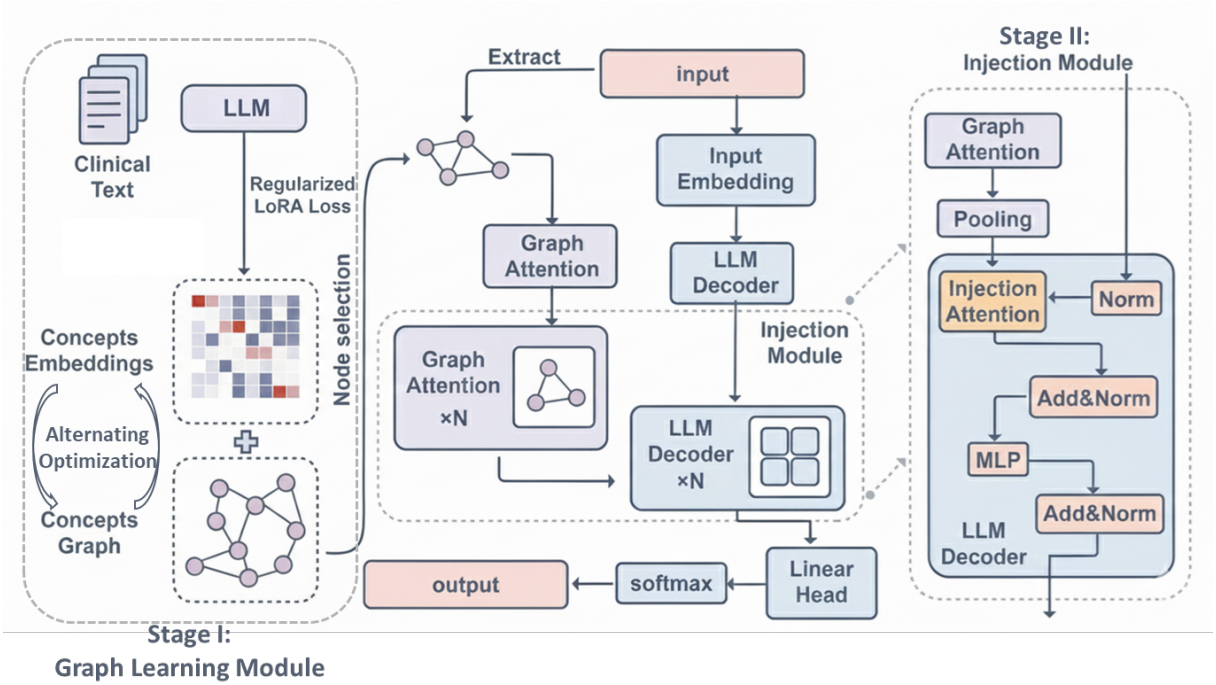

Figure 2: **Overview of the GRIAN framework:** Stage I learns sparse conditional dependencies among clinical concepts via graph-regularized LoRA adaptation, inducing a concept graph from in-domain text. Stage II injects query-conditioned subgraphs into the LLM decoder through graph-aware attention, making structural evidence explicit for downstream clinical tasks.

Based on expert knowledge, regular expression patterns, and term frequency analysis of the corpus, we identified 876 diagnostic keyword concepts that are both frequent and clinically pertinent to knee-joint diagnoses.

The task is formulated as an 8-label multi-label classification problem, in which each patient's record is analyzed to determine the presence of specific orthopedic pathologies. The target conditions are defined as follows: *rupture of the anterior cruciate ligament* (ACL), *rupture of the posterior cruciate ligament* (PCL), *injury to the medial meniscus* (MM), *injury to the lateral meniscus* (LM), *injury to the medial collateral ligament* (MCL), *injury to the lateral collateral ligament* (LCL), *patellar dislocation* (PD), and *patellar chondromalacia* (PC).

Formally, for patient i, let the ground-truth label vector be $\mathbf{Y}_i \in \{0,1\}^8$ and the input text data $\mathbf{T}_i$ consist of the three aforementioned record components. We aim to learn a mapping $f : \mathbf{T}_i \mapsto \hat{\mathbf{Y}}_i \in [0,1]^8$ that predicts the probability of each pathology being present.

3.2 Medical Abstracts Dataset

To complement our in-house Chinese knee-joint EMR corpus, we further evaluate on a public medical text classification benchmark, Medical Abstracts, following the processed version introduced by Schopf et al. (2022). The corpus was originally collected from Kaggle and contains medical abstracts associated with five patient-condition categories. The processed dataset keeps only labeled abstracts, maps numerical labels to descriptive class names, and provides a standard train/test split. Specifically, it contains 11,550 training abstracts and 2,888 test abstracts, totaling 14,438 labeled instances across five classes: *Neoplasms*, *Digestive system diseases*, *Nervous system diseases*, *Cardiovascular diseases*, and *General pathological conditions*. We construct the concept vocabulary by directly prompting a Qwen2-7B to extract domain concepts from the *training split* abstracts only.

Formally, for each abstract i, we denote the input text as $\mathbf{X}_i$ and the label as $y_i \in 1, \ldots, 5$. The task is to learn a classifier $g(\mathbf{X}_i) \to \hat{y}_i$ for 5-way medical topic/condition classification. Compared with our EMR diagnosis setting, Medical Abstracts features shorter, more standardized biomedical writing, enabling us to test whether the proposed graph-prior adaptation generalizes beyond clinical narratives to medical scientific abstracts.

Table 1: Performance on the Electronic Medical Record (EMR) diagnosis dataset

Model	Acc.(label-level)(%)	Macro-Pre.(%)	Macro-Rec.(%)	Macro-F1	Micro-F1	Acc.(sample-level)(%)
Qwen2-7B	93.99	84.94	80.40	81.01	83.97	66.60
GRIAN-Qwen2	**95.51**	**88.71**	**81.91**	**85.03**	**87.16**	**73.75**
DeepSeek-7B	94.36	87.34	78.49	82.36	83.86	66.60
GRIAN-DeepSeek	**95.61**	**90.03**	**83.22**	**86.42**	**87.06**	**74.00**
Llama2-7B	94.25	86.33	82.01	83.70	83.97	65.77
GRIAN-Llama2	**95.17**	**86.54**	**84.55**	**85.28**	**86.37**	**71.81**

Table 2: Ablation on the EMR dataset: effect of graph construction for Qwen2-7B.

Model	Acc.(label-level)(%)	Macro-Pre.(%)	Macro-Rec.(%)	Macro-F1	Micro-F1
Qwen2 (No Graph)	93.99	84.94	80.40	81.01	83.97
GRIAN-Qwen2 (Random Graph)	95.46	87.95	**82.02**	84.85	86.98
GRIAN-Qwen2 (Correlation Graph)	95.46	88.38	81.46	84.73	86.38
GRIAN-Qwen2 (Conditional Dependency Graph)	**95.51**	**88.72**	81.91	**85.03**	**87.16**

4 Results

In this section, we compare the proposed GRIAN-based method with several baseline models on Electronic Medical Record Dataset, and the public Medical Abstracts dataset. Notably, stage I learns the concept graph using *only* the training split in an unsupervised manner.

The evaluated models include three pure large language models, Qwen-2-7B-Instruct(Qwen2-7B), DeepSeek-llm-7B(DeepSeek-7B), and Llama-2-7B(Llama2-7B)—as well as our graph-enhanced variant, GRIAN. Experimental results demonstrate that the incorporation of graph structure via GRIAN leads to superior performance in accuracy, recall, and F1-score.

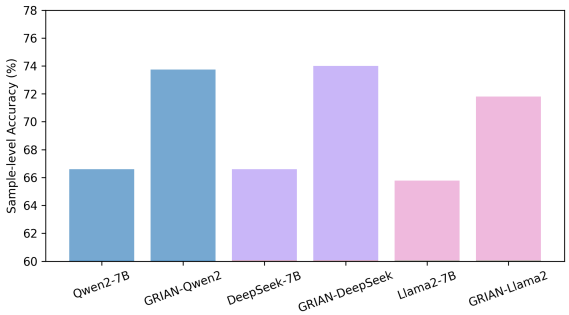

Figure 3: **Sample-level accuracy comparison across backbones.** Sample-level accuracy measures the fraction of test cases for which *all* diagnostic labels are correctly predicted. Across all backbones, GRIAN consistently improves sample-level accuracy, indicating more holistic and structurally consistent clinical predictions.

4.1 Evaluation on Electronic Medical Record Dataset

To evaluate the effectiveness of the proposed graph-based knowledge injection, we conducted an ablation experiment by comparing the baseline LLMs with their graph-augmented variants (GRIAN) on the internal EMR dataset.

Figure 4 reports the per-class F1-scores across the eight orthopedic pathologies. Overall, most diagnostic labels benefit from graph-regularized adaptation, suggesting that the learned dependency structure provides broadly useful inductive bias rather than label-specific tuning.

The EMR diagnosis task is challenging due to co-occurring pathologies that require jointly consistent predictions. Accordingly, beyond label-level metrics, we emphasize sample-level accuracy, which requires all labels for a patient to be correctly predicted and better reflects holistic clinical decision quality.

As shown in Table 1, while both baseline and GRIAN models achieve comparably high label-level accuracy (predominantly above 0.95), significant performance differences emerge in the more nuanced metrics of macro-averaged F1-score and sample-level accuracy. GRIAN-Qwen2 shows a clear improvement in macro-F1 (from 81.01 to 85.03), indicating a better balance between precision and recall across diagnostic classes. More importantly, all GRIAN variants consistently improve sample-level accuracy, demonstrating more holistic clinical reasoning compared to backbone LLMs, as

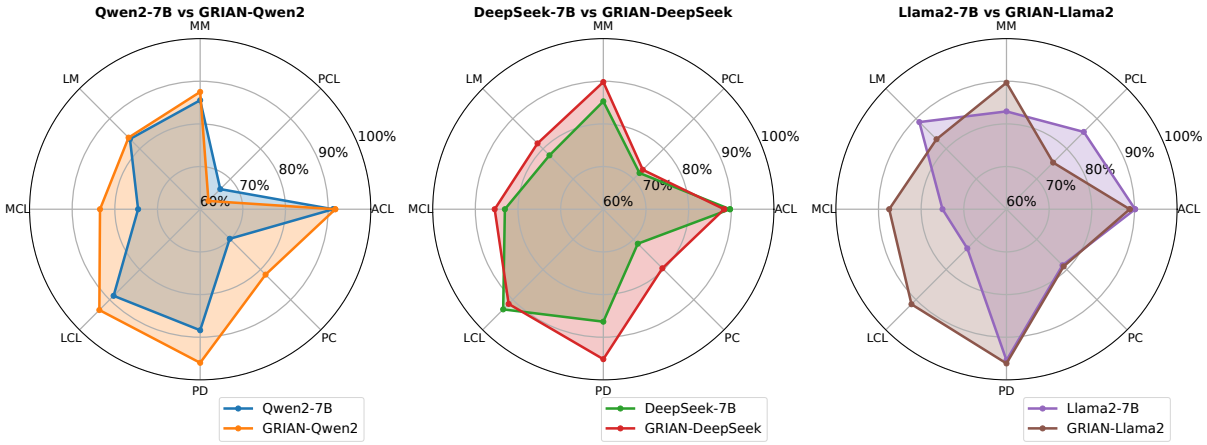

Figure 4: Per-label F1 (%) on the EMR dataset across model variants.

shown in Figure 3. These gains suggest that enforcing a sparse conditional-dependency prior helps the model reason over interrelated diagnoses rather than predicting labels independently.

To isolate the impact of graph quality, we compare GRIAN-Qwen2 with three variants: (i) no graph injected, (ii) a random graph with matched sparsity, and (iii) a correlation-based graph constructed from marginal embedding correlations. Table 2 compares different graph constructions under the same GRIAN framework. While introducing any graph structure already improves over the no-graph baseline, both the random graph and the correlation-based graph yield similar performance, suggesting that naive or marginal association graphs provide limited additional benefit. In contrast, the conditional-dependency graph delivers the most consistent gains, indicating that GRIAN's benefit comes not just from graph injection itself, but also from learning a principled precision graph that captures conditional—rather than marginal—relationships.

4.2 Evaluation on Medical Abstracts Dataset

We further assess cross-domain robustness on the Medical Abstracts dataset, which differs substantially from our in-house EMR diagnosis corpus in both writing style and label semantics. While EMRs exhibit noisy, patient-specific narratives with implicit concept dependencies, Medical Abstracts are concise biomedical summaries organized around broad condition categories. This setting tests whether injecting a sparse concept-dependency prior during adaptation yields benefits beyond clinical notes.

Table 3 reports performance under a unified

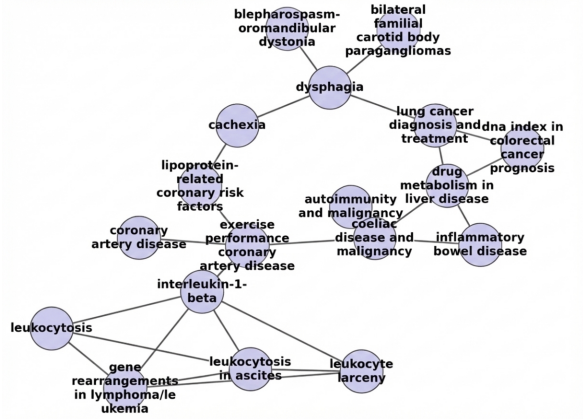

Figure 5: A qualitative visualization of the concept dependency graph learned from the Medical Abstracts corpus. We show an induced subgraph around representative medical concepts, where nodes denote concepts and edges indicate non-zero off-diagonal entries in the estimated precision matrix (i.e., conditional dependencies).

single-label classification setting. Across all backbones, GRIAN consistently improves micro-F1 over the corresponding LLM baselines, indicating better balanced performance across disease categories. Although absolute gains are smaller than those observed on EMRs, the improvements are consistent, suggesting that the proposed graph-regularized adaptation generalizes beyond patient-level clinical records.

Table 4 reports basic statistics of the concept dependency graph learned from the Medical Abstracts corpus. To illustrate the interpretability of the learned structural prior, Figure 5 visualizes a small region of the concept dependency graph induced from the Medical Abstracts dataset. The resulting structure is sparse and locally clustered,

Table 3: Performance on the Medical Abstracts dataset. All results are evaluated under a unified single-label multi-class setting with standard argmax inference. Results for unsupervised and zero-shot methods are directly reported from (Schopf et al., 2022), where micro-averaged Precision, Recall, and F1 are equal by definition. Results for BERT-base and DistilBERT are reported from (Liu et al., 2025b) and correspond to cross-entropy training without tuned decision thresholds.

Model	Acc.(sample-level)(%)	Macro-Pre.(%)	Macro-Rec.(%)	Macro-F1	Micro-F1
Unsupervised / Zero-shot baselines					
SimCSE (unsup.)	–	–	–	–	34.94
SBERT (MiniLM, unsup.)	–	–	–	–	46.53
SBERT (MPNet, unsup.)	–	–	–	–	46.34
Lbl2TransformerVec (unsup.)	–	–	–	–	56.46
Zero-shot Entailment (DeBERTa)	–	–	–	–	57.28
Supervised Transformer baselines					
BERT-base (CE, argmax)	64.51	–	–	63.85	62.12
DistilBERT (CE, argmax)	64.61	–	–	64.38	63.25
Large language models					
Qwen2-7B	63.40	62.44	65.84	63.74	63.50
DeepSeek-7B	65.41	64.58	68.47	65.99	65.40
Llama2-7B	65.89	64.69	68.23	65.55	65.89
Proposed method					
GRIAN-Qwen2	63.75	62.73	66.33	64.13	63.85
GRIAN-DeepSeek	65.81	64.58	67.77	65.81	65.71
GRIAN-Llama2	**66.66**	**65.52**	**69.73**	**67.07**	**66.65**

Table 4: Statistics of the learned concept dependency graph on the Medical Abstracts dataset.

Statistic	Value
Number of concepts ($\|V\|$)	983
Number of edges ($\|E\|$)	13211
Graph density	2.74%

forming coherent modules that align with disease-related themes, which supports the motivation of treating a sparse concept graph as a prior during training.

Overall, the consistent micro-F1 improvements across backbones suggest that GRIAN can adaptively recover domain-appropriate concept dependencies and leverage them for improved reasoning in medical abstract text.

5 Conclusion

We presented **GRIAN**, a graph-regularized and injected adaptation framework that treats the existence of a sparse concept dependency graph as an explicit prior when adapting large language models to specialized domains such as clinical text. Unlike pipelines that learn a graph separately or rely on existed knowledge graphs, GRIAN integrates graph recovery into the training objective via a nonparanormal matrix-normal graphical-model loss, jointly estimating sparse precision matrices while optimizing the LLM. A complementary Laplacian smoothness regularizer further aligns concept embedding geometry with the emerging structure through parameter-efficient LoRA updates. For downstream prediction, we encode query-conditioned induced subgraphs with a GAT and inject compact graph evidence into Transformer attention as additional key–value memories, enabling structure-grounded and more interpretable reasoning. Experiments on a Chinese knee-joint EMR diagnosis task and the Medical Abstracts benchmark show consistent improvements across multiple LLMs, supporting the benefit of jointly regularized graph-prior adaptation for reliable clinical and biomedical text modeling.

Limitations

Our framework depends on the quality of the concept vocabulary and concept-to-token mapping; missing or noisy concepts can degrade graph recovery. In addition, we evaluate on a real-world EMR dataset that cannot be publicly released due to patient privacy and institutional restrictions, which limits full reproducibility on the same cohort; to partially mitigate this, we will release anonymized examples, the concept vocabulary and extraction rules, and code for graph learning/adaptation.

References

Chengfeng Dou, Ying Zhang, Zhi Jin, Wenpin Jiao, Haiyan Zhao, Yongqiang Zhao, and Zhengwei Tao. 2025. Enhancing llm generation with knowledge hypergraph for evidence-based medicine. *arXiv preprint arXiv:2503.16530*.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Rui Guo, Barry Devereux, Greg Farnan, and Niall McLaughlin. 2025. LAB-KG: A retrieval-augmented generation method with knowledge graphs for medical lab test interpretation. In *Proceedings of Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning @ COLING 2025*, pages 40–50, Abu Dhabi, UAE. ELRA and ICCL.

Zhen Hong, Han Zhou, Wenjie Liu, Jie Tang, and Le Song. 2024. Graph-toolformer: Graph-based tool learning for large language models. *arXiv preprint arXiv:2402.03290*.

Bowen Jin, Yu Zhao, Chunming Huang, Yang Kang, and Ji-Rong Wen. 2024. Graph chain-of-thought: Augmenting large language models by reasoning on graphs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 176–189, Bangkok, Thailand. Association for Computational Linguistics.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations (ICLR)*.

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Fengze Liu, Haoyu Wang, Joonhyuk Cho, Dan Roth, and Andrew Lo. 2025a. AutoCT: Automating interpretable clinical trial prediction with LLM agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30945–30970, Suzhou, China. Association for Computational Linguistics.

Han Liu, John Lafferty, and Larry Wasserman. 2009. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10:2295–2328.

Jiaqi Liu, Tong Wang, Su Liu, Xin Hu, Ran Tong, Lanruo Wang, and Jiexi Xu. 2025b. Lightweight baselines for medical abstract classification: Distilbert with cross-entropy as a strong default. *Preprint*, arXiv:2510.10025.

Weijia Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. In *Proceedings of AAAI Conference on Artificial Intelligence*, pages 2901–2908.

Xiaotian Ma, Wenxuan Huang, and Jie Tang. 2024. Graphgpt: Graph instruction tuning for large language models. *arXiv preprint arXiv:2403.04477*.

Arsalan Masoudifard, Mohammad Mowlavi Sorond, Moein Madadi, Mohammad Sabokrou, and Elahe Habibi. 2024. Leveraging graph-rag and prompt engineering to enhance llm-based automated requirement traceability and compliance checks. *arXiv preprint arXiv:2412.08593*.

Yang Ning and Han Liu. 2013. High-dimensional semiparametric bigraphical models. *Biometrika*, 100(3):655–670.

Mohammad Reza Rezaei, Reza Saadati Fard, Jayson Lee Parker, Rahul G Krishnan, and Milad Lankarany. 2025. Agentic medical knowledge graphs enhance medical question answering: Bridging the gap between LLMs and evolving medical knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12682–12701, Suzhou, China. Association for Computational Linguistics.

Tim Schopf, Daniel Braun, and Florian Matthes. 2022. Evaluating unsupervised text classification: Zero-shot and similarity-based approaches. In *Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval (NLPIR 2022)*.

Karan Singhal and 1 others. 2023a. Large language models encode clinical knowledge. *Nature*.

Karan Singhal and 1 others. 2023b. Med-palm 2: Towards expert-level medical question answering with large language models. *Preprint*, arXiv:2305.09617.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M Ni, Heung-Yeung Shum, and Jian Guo. 2023. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. *arXiv preprint arXiv:2307.07697*.

Shiyu Tian, Yangyang Luo, Tianze Xu, Caixia Yuan, Huixing Jiang, Chen Wei, and Xiaojie Wang. 2024. KG-adapter: Enabling knowledge graph integration in large language models through parameter-efficient fine-tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3813–3828, Bangkok, Thailand. Association for Computational Linguistics.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations (ICLR)*.

Minghua Wan, Mingxiu Cai, and Guowei Yang. 2023. Robust exponential graph regularization non-negative matrix factorization technology for feature extraction. *Mathematics*, 11(7):1716.

Ruize Wang, Kehai Gao, Shizhu Chen, Kang Liu, and Jun Zhao. 2021. Kepler: A unified model for knowledge embedding and pre-trained language representation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, pages 200–210.

Fuda Ye, Shuangyin Li, Yongqi Zhang, and Lei Chen. 2024. R^2AG: Incorporating retrieval information into retrieval augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11584–11596, Miami, Florida, USA. Association for Computational Linguistics.

Jianxin Yin and Hongzhe Li. 2012. Model selection and estimation in the matrix normal graphical model. *Journal of Multivariate Analysis*, 107:119–140.

Ming Yin, Junbin Gao, and Zhouchen Lin. 2016. Laplacian regularized low-rank representation and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):504–517.

Yu Zhang, Kehai Chen, Xuefeng Bai, Quanjiang Guo, Min Zhang, and 1 others. 2024. Question-guided knowledge graph re-scoring and injection for knowledge graph question answering. *arXiv preprint arXiv:2410.01401*.

A Experiment Details

A.1 Dataset Partition

For the EMR dataset, a total of 5,825 samples are divided as follows: 4,563 samples for training, 777 samples for validation (from which the best-performing model is selected), and 485 samples for testing.

For the Medical Abstract dataset, we follow the data partitioning protocol established by Schopf et al. (2022). The dataset is divided into a training set, consisting of 11,550 samples, and a test set, containing 2,888 samples. Each sample in this dataset is associated with a single class label, meaning it exclusively belongs to one diagnostic category.

A.2 Hyperparameter Settings

All hyperparameters can be divided based on training stage: Stage I and Stage II.

Stage I hyperparameters.

Table 5: Dataset partition for EMR dataset (counts per label). Counts are label-positive counts; totals exceed sample counts due to multi-label.

Diagnosis	Train	Validation	Test	Total
ACL	891	265	170	1326
PCL	89	29	14	132
MM	813	256	166	1235
LM	536	203	150	889
MCL	246	121	62	429
LCL	199	79	53	331
PD	135	52	24	211
CP	1801	102	65	1968
Total	4710	1107	704	6521

Table 6: Dataset partition for Medical Abstracts (MA) (counts per class).

Class name	Train	Test	Total
Neoplasms	2530	633	3163
Digestive system diseases	1195	299	1494
Nervous system diseases	1540	385	1925
Cardiovascular diseases	2441	610	3051
General pathological conditions	3844	961	4805
Total	11550	2888	14438

Stage I jointly optimizes the language modeling objective and the graph-related regularizers (non-paranormal MNGM loss + Laplacian smoothness) under LoRA adaptation. Table 7 summarizes all Stage I hyperparameters used in our implementation. For Laplacian construction, we convert the estimated precision matrix A into a nonnegative weighted adjacency W by setting $W_{ij} = |a_{ij}|\mathbf{1}(|a_{ij}| > 1 \times 10^{-5})$ (and $W_{ii} = 0$), and then form the graph Laplacian $L = D - W$ with $D_{ii} = \sum_j W_{ij}$.

Stage II hyperparameters

The hyperparameters configured for Stage II are categorized into two groups: those for the baseline benchmark model and those specific to the GRIAN model. The baseline settings include the learning rate, total training epochs, batch size, and LoRA configuration (rank r, dropout rate, and scaling factor α). The GRIAN-specific parameters introduce structural components, including the number of Graph Attention Network (GAT) layers, the specific layers into which graph knowledge is injected into the Large Language Model (LLM), and a graph signal threshold τ, which determines the strength of the node signal required to establish an edge in the constructed graph. The detailed hyperparameter settings are summarized in Table 8 and Table 9.

Table 7: Stage I hyperparameters (graph-regularized adaptation).

Symbol / Item	Setting
λ	2.5 (ℓ_1 sparsity penalty on row precision A)
γ	3.0 (ℓ_1 sparsity penalty on column precision B)
α	0.001 (weight of nonparanormal MNGM loss ϕ_{np})
β	1×10^{-6} (Laplacian smoothness weight)
L_A	Unnormalized Laplacian constructed from A (edges retained if $\lvert a_{ij} \rvert > 1 \times 10^{-5}$)
Training epochs (Stage I)	50
Batch size	6
Learning rate	1×10^{-4}
LoRA rank r	4
LoRA scaling factor	16
LoRA dropout	0.05
LoRA target modules	`q_proj, k_proj, v_proj, embed_tokens`

Table 8: Hyperparameter settings for baseline models.

Hyperparameter	Qwen-7B	Llama-7B	DeepSeek-7B
Learning rate	1e-4	1e-4	1e-4
Total training epochs	10	10	10
Batch size	10	10	10
LoRA rank (r)	16	8	8
LoRA dropout	0.05	0.05	0.05
LoRA scaling factor	32	32	32

Table 9: Hyperparameter settings for GRIAN models.

Hyperparameter	GRIAN-Qwen2	GRIAN-Llama	GRIAN-DeepSeek
Learning rate	1e-4	1e-4	1e-4
Total training epochs	15	15	15
Batch size	10	10	5
Number of GAT layers	5	5	5
Graph injection layers	[3, 4, 5]	[3, 4, 5]	[3, 4, 5]
Graph signal threshold (τ)	{1e-3, 5e-4, 1e-4}	{1e-3, 5e-4, 1e-4}	{1e-3, 5e-4, 1e-4}

A.3 Performance Details

Table 10 reports per-label F1 scores on the EMR diagnosis task. Overall, the improvements introduced by GRIAN are label-dependent, which is expected in a multi-label clinical setting where some conditions are near ceiling while others are rarer or more ambiguous.

Across backbones, GRIAN consistently improves labels such as MCL, LCL, PD, and PC. These conditions often require jointly reasoning over symptoms, physical examinations, and imaging findings, and thus benefit from a sparse conditional-dependency prior. For easier labels such as ACL, baseline models already achieve high F1 scores, and GRIAN largely preserves near-ceiling performance.

Some rare labels (e.g., PCL) exhibit smaller or mixed changes, likely due to limited sample size and threshold sensitivity. Overall, these per-label results complement the main-text findings on macro-F1 and sample-level accuracy, indicating that GRIAN improves balanced, case-level clinical reasoning rather than optimizing individual labels in isolation.

Table 10: Per-label F1 (%) on the EMR diagnosis dataset.

Model	ACL	PCL	MM	LM	MCL	LCL	PD	PC
Qwen2-7B	91.18	**66.67**	85.55	83.28	74.55	88.70	88.37	69.80
GRIAN-Qwen2	**91.68**	62.75	**87.50**	**83.74**	**83.48**	**93.42**	**96.00**	**81.68**
DeepSeek-7B	**89.70**	72.00	85.29	77.86	83.08	**93.20**	86.36	71.43
GRIAN-DeepSeek	88.33	**73.08**	**89.82**	**81.82**	**85.47**	91.39	**95.15**	**79.58**
Llama2-7B	**90.15**	**85.63**	82.93	**88.89**	75.00	72.99	95.41	78.57
GRIAN-Llama2	88.89	75.47	**89.66**	83.20	**87.50**	**91.50**	**96.15**	**78.95**