

Medical Bilingualism: Cross-Ontology Graph Retrieval for Incommensurable Knowledge Bases

Zixuan Wang¹, Ke Lu¹, Zihan Wang¹, Junwei Lei²
Junhao Dong³, Jinyu Guo⁴, Chaomeng Chen^{5,6,*}

¹Harbin Institute of Technology ²Beijing University of Posts and Telecommunications

³Nanyang Technological University

⁴University of Electronic Science and Technology of China

⁵Great Bay University ⁶Tsinghua University

*Corresponding author: cmchen@gbu.edu.cn

zixuanwang.cs@stu.hit.edu.cn kelu@stu.hit.edu.cn 2023112983@stu.hit.edu.cn

leijw@bupt.edu.cn junhao003@ntu.edu.sg guojinyu@uestc.edu.cn

Abstract

Graph-based Retrieval-Augmented Generation (GraphRAG) can fail in interdisciplinary medicine where evidence comes from *incommensurable* knowledge systems, e.g., Traditional Chinese Medicine (TCM) and Western Medicine (WM). Similarity-driven retrieval in a shared latent space may over-unify heterogeneous concepts, causing *semantic collapse* (spurious alignments, provenance-blind mixing, and broken multi-hop chains). We propose XONTO-RAG, a cross-ontology GraphRAG framework that avoids ontological unification by constructing a dual-ontology graph with *typed* bridges induced via LLM-assisted verify-and-vote. A query-aware Relational Graph Attention Network (RGAT) then assembles compact, provenance-tracked multi-hop evidence subgraphs for grounded generation. Across TCM–WM benchmarks, XONTO-RAG outperforms strong RAG/GraphRAG baselines, seeing up to +22.2% accuracy on MLEC-QA TCWM (vs. OG-RAG), 77.0% accuracy with a 7B backbone, and 85.1% faithfulness on MLEC-TCM, while reducing semantic-collapse failures and yielding interpretable reasoning chains.

1 Introduction

Large Language Models (LLMs) have demonstrated transformative potential in specialized domains such as biomedical question answering and clinical decision support (Singhal et al., 2023), yet they remain brittle in highly specialized and interdisciplinary medical settings, where knowledge gaps and hallucinations are common (Ji et al., 2023). Prior attempts to mitigate these failures include fine-tuning and knowledge editing (KE), but both face structural limitations for cross-domain reasoning: fine-tuning is computationally expensive and may risk catastrophic forgetting (Li et al.,

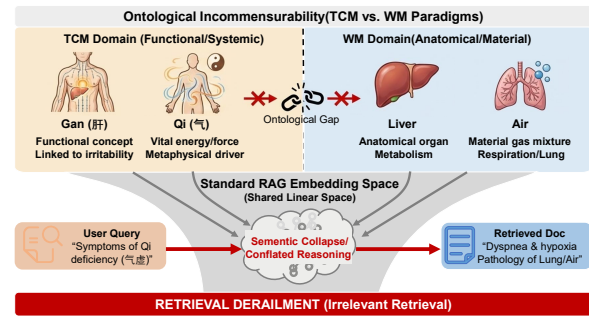


Figure 1: An illustration of **ontological incommensurability** between TCM and WM paradigms. When standard RAG pipelines project heterogeneous concepts into a shared representation space, **semantic collapse** can occur: distinct cross-ontology concepts become indistinguishable, inducing **retrieval derailment**. For example, functional notions (e.g., Qi deficiency) may be spuriously aligned with anatomical pathologies (e.g., hypoxia) despite incompatible explanatory primitives.

2024), while KE—though effective for localized factual updates (De Cao et al., 2021)—struggles to resolve *systemic conflicts* between divergent knowledge systems at scale (Thede et al., 2025). These limitations motivate retrieval-time grounding mechanisms that can incorporate external evidence while preserving the integrity of the underlying knowledge systems.

Consequently, Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) and its structure-aware variants—including GraphRAG (Edge et al., 2024) and ontology-aware RAG (Sharma et al., 2025)—have become widely adopted. However, a critical limitation in many such pipelines is the tendency to *rank, aggregate, and reason* over evidence under a *single* shared semantic space or unified scoring signal (Karpukhin et al., 2020; Santhanam et al., 2022; Weller et al., 2025). By ignoring semantic boundaries and provenance across fields, this forced unification can yield systematic failure modes, as il-

lustrated in Figure 1.

This challenge is rooted in what philosophy of science and linguistic anthropology describe as *ontological incommensurability* (Fillmore, 1982; Thagard and Zhu, 2003; Hanson and Kumar, 2025), where different domains operate with incompatible terminologies, explanatory primitives, and validation standards. In contrast to this theoretical depth, current ontology-grounded RAG systems often utilize ontologies primarily as rigid schemas or structural constraints (Sharma et al., 2025; Zhao et al., 2025; Caufield et al., 2024), implicitly assuming that concepts can be mapped into a shared semantic space without friction. This approach typically fails to model the uncertainty of *translation* between systems or to trace ontological provenance in multi-hop reasoning. In scenarios like TCM-WM, it manifests as three recurring failures: **erroneous alignment** (treating distinct concepts as equivalent), **conflated reasoning** (mixing evidence without ontology-aware provenance), and **fragmented retrieval** (breaking multi-hop reasoning chains across the semantic divide). This leads to our central research question: *How can we perform effective cross-ontology retrieval and reasoning while preserving the semantic integrity and provenance of incommensurable domains?*

To address this challenge, we introduce XONTO-RAG, a cross-ontology graph-based RAG system over a dual-ontology knowledge base (Figure 2). XONTO-RAG preserves ontology labels, induces *typed* bridging relations via LLM-assisted verification, and employs a query-aware RGAT retriever to assemble compact cross-ontology evidence subgraphs that support provenance-aware reasoning for LLM generation. To the best of our knowledge, XONTO-RAG is the first RAG system to explicitly optimize ontology-coherent multi-hop retrieval with *verifiable*, *typed* cross-ontology transitions in incommensurable medical settings.

Overall, our contributions are:

- We formalize the **Cross-Incommensurable Ontology RAG** setting and characterize distinctive failure modes, including erroneous alignment, conflated reasoning, and fragmented retrieval.
- We propose XONTO-RAG, an ontology-preserving GraphRAG framework that constructs a **dual-ontology graph** and induces **typed bridging edges** via an LLM-assisted *verify-and-vote* procedure, enabling explicit and auditable cross-ontology transitions.
- We develop a **query-aware RGAT retriever** that

supports dynamic, multi-hop evidence assembly with ontology provenance, and demonstrate substantial gains on integrated TCM-WM QA and exam benchmarks over competitive RAG-/GraphRAG baselines.

2 Related Work

2.1 RAG and GraphRAG

Classic RAG couples dense retrieval with generation (Lewis et al., 2020; Shi et al., 2024), while recent graph-centric variants explicitly model structure to support multi-hop reasoning for specific pain points. GraphRAG (Edge et al., 2024) builds hierarchical summaries to handle long-range traversals in sprawling corpora, and HippoRAG (Jiménez Gutiérrez et al., 2024) augments retrieval with a neuro-inspired long-term memory to mitigate forgetting during multi-hop QA. GRAG (Hu et al., 2025) partitions document graphs to control search complexity, MedGraphRAG (Wu et al., 2025) constructs medical triplet graphs for auditable clinical evidence, KARE (Jiang et al., 2025) improves healthcare prediction by retrieving graph communities within a single medical KG, and TimeR4 (Qian et al., 2024) introduces a retrieve–rewrite loop for temporal KGQA. These lines of work collectively illustrate that graph-enhanced RAG solutions are usually motivated by the need to respect structure already present in a single ontology or modality; by contrast, our setting requires explicit, typed movement across incompatible ontologies rather than stronger retrieval inside one coherent graph.

2.2 Ontology-aware RAG

Ontology-grounded RAG uses explicit schemas to constrain retrieval and attribution under high hallucination risk. However, most systems still assume a *single* authoritative ontology as the truth layer: OG-RAG projects evidence onto one ontology via ontology-guided hypergraphs (Sharma et al., 2025), and CyberBOT validates answers against a cybersecurity ontology through constraint checking (Zhao et al., 2025). Surveys on LLM-KG integration similarly note that deployed pipelines typically rely on a dominant ontology even when sources are heterogeneous (Pan et al., 2023).

To handle multiple ontologies, recent systems (e.g., DRAGON-AI, OntoGPT) often *fuse* schemas via embedding alignment or schema completion so generation operates in a shared latent space (Toro

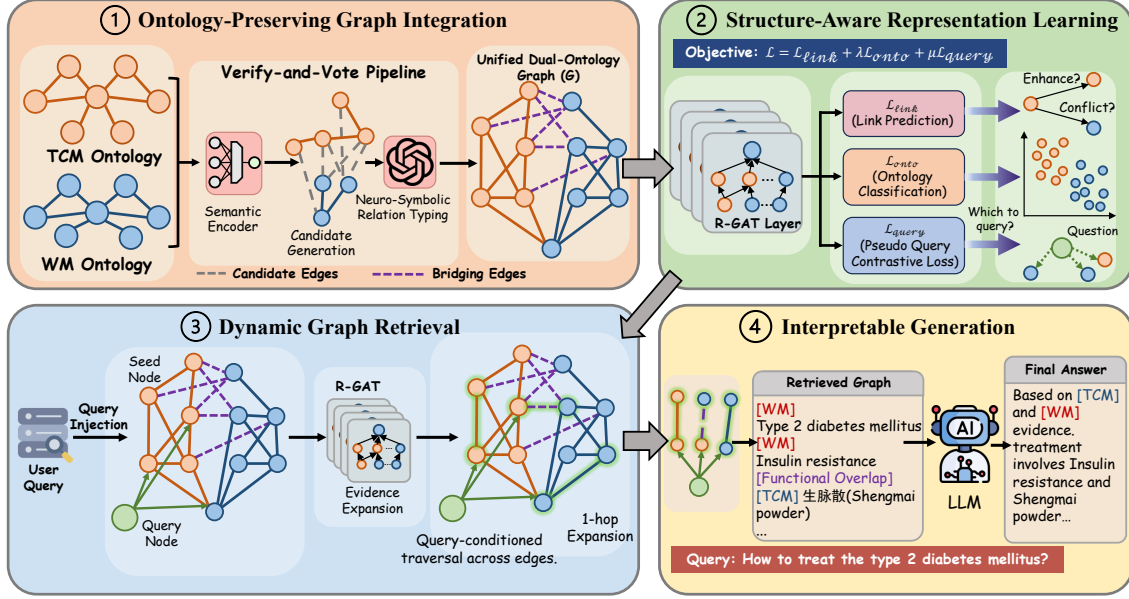


Figure 2: Overview of XONTO-RAG. The pipeline comprises: (1) **Graph Integration**, constructing robust typed bridging edges via an LLM-assisted *verify-and-vote* mechanism; (2) **Structure-Aware Learning**, utilizing an optimized RGAT encoder with ontology-preservation ($\mathcal{L}_{\text{onto}}$) and query-contrastive ($\mathcal{L}_{\text{query}}$) objectives; (3) **Dynamic Retrieval**, injecting query nodes for multi-hop contextual evidence expansion; and (4) **Interpretable Generation**, synthesizing answers with explicit ontology provenance.

et al., 2024; Caufield et al., 2024). Yet, such fusion rarely models *typed, non-equivalence* correspondences (beyond strict equality) or quantifies translation uncertainty when paradigms disagree, limiting applicability to incommensurable medical ontologies. This echoes the ontology alignment literature: embedding-based methods (MTransE, BootEA) (Chen et al., 2017; Sun et al., 2018) and GNN-based aligners (GCN-Align, RDGCN) (Wang et al., 2018; Wu et al., 2019b) primarily target equivalence discovery in mostly compatible graphs, and surveys highlight persistent failures when ontologies encode different explanatory standards (Euzenat and Shvaiko, 2013).

2.3 TCM–WM Knowledge Graph Alignment

TCM KGs such as SymMap, HERB, and TCMSP curate herbs/formulas/symptoms and their molecular targets (Wu et al., 2019a; Fang et al., 2021), whereas WM resources (UMLS, MeSH, DOID) prioritize standardized biomedical vocabularies (Bodenreider, 2004; Lipscomb, 2000; Schriml et al., 2012). Rather than fully unifying these paradigms, recent ICWM ontology efforts advocate *modular bridges* that preserve divergent evidentiary standards (Wang et al., 2024b). Related initiatives (e.g., OpenTCM) and TCM compatibility studies further highlight the need for explicit translation layers that represent *typed, partial correspondences*

(herb–disease, symptom, mechanism) without semantic collapse (He et al., 2025; Zeng and Jia, 2024).

3 Methodology

In this section, we propose XONTO-RAG, a framework designed to bridge incommensurable knowledge systems without forcing them into a single semantic flatland. As illustrated in Figure 2, our approach operates in four distinct phases to preserve ontological integrity while enabling coherent reasoning. First, we formalize the cross-ontology retrieval task in Section 3.1. We then describe the construction of a dual-ontology graph with verified bridging mechanisms in Section 3.2. To navigate this structure, we introduce a representation learning objective in Section 3.3, and finally, detail the query-aware dynamic inference algorithm that constructs reasoning chains in Section 3.4.

3.1 Problem Formulation

In our primary setting, we focus on the dual-ontology case with $K = 2$ (e.g., TCM vs. WM), while the formulation naturally extends to $K > 2$. Each entity $v \in \mathcal{V}$ belongs to exactly one ontology, indicated by an ontology label $\phi : \mathcal{V} \rightarrow \{1, \dots, K\}$.

We represent edges as typed triples $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{R} \times \mathcal{V}$ and decompose relation as $\mathcal{R} = \mathcal{R}_{\text{intra}} \cup \mathcal{R}_{\text{bridge}}$,

where $\mathcal{R}_{\text{intra}}$ are within-ontology relations and $\mathcal{R}_{\text{bridge}}$ are cross-ontology relations. Accordingly, we partition the edge set into intra-ontology edges $\mathcal{E}_{\text{intra}} = \{(u, r, v) \in \mathcal{E} \mid \phi(u) = \phi(v)\}$ and bridging edges $\mathcal{E}_{\text{bridge}} = \{(u, r, v) \in \mathcal{E} \mid \phi(u) \neq \phi(v)\}$.

Given a natural-language query q , the retriever must return a compact evidence subgraph $\mathcal{G}_{\text{sub}} = (\mathcal{V}_{\text{sub}}, \mathcal{E}_{\text{sub}}) \subseteq \mathcal{G}$ that satisfies two requirements: **(i) Relevance:** $\mathcal{V}_{\text{sub}}$ contains entities whose textual descriptions are pertinent to q ; and **(ii) Ontology-coherent reasoning:** $\mathcal{G}_{\text{sub}}$ contains at least one connected reasoning path $p = (v_1, \dots, v_T)$ such that (a) each hop follows a typed edge in $\mathcal{E}_{\text{sub}}$, and (b) any cross-ontology transition is made explicitly through a bridging edge in $\mathcal{E}_{\text{bridge}}$, enabling step-wise ontology provenance (i.e., each intermediate claim can be attributed to $\phi(v_t)$ along the path).

Finally, a generator takes a linearized representation $\text{Lin}(\mathcal{G}_{\text{sub}})$ that preserves ontology tags and edge types, and produces an answer $a = \text{LLM}(q, \text{Lin}(\mathcal{G}_{\text{sub}}))$. Specifically, $\text{Lin}(\cdot)$ linearizes $\mathcal{G}_{\text{sub}}$ by a BFS traversal rooted at the seed entities, emitting triples grouped by hop depth and relation type. Each typed triple is converted into a natural-language statement (e.g., “ u is [relation] of v ”), with ontology tags preserved to bridge structural evidence and the LLM’s textual pretraining.

3.2 Ontology-Preserving Graph Integration

A naive merger of disjoint ontologies often leads to **erroneous alignment**, where distinct concepts are conflated due to surface-level lexical overlap. To avoid collapsing heterogeneous medical logics into a single undifferentiated space, we construct cross-ontology links through a **verify-and-vote** pipeline that first proposes plausible candidates and then filters them with LLM-assisted verification.

Candidate Generation via Semantic Filtering. We first initialize the graph with domain-specific entities and intra-domain relations. To discover potential connections between disjoint ontologies (e.g., $\mathcal{D}_{\text{TCM}}$ and $\mathcal{D}_{\text{WM}}$), we rely on curated intermediate medical lexicons that index concepts from both sides, rather than committing to a single fixed vocabulary. These lexicons provide an initial set of cross-ontology candidates

$$\mathcal{C}_0 = \left\{ (u, v) \mid \begin{array}{l} \phi(u) \neq \phi(v), \\ u \text{ and } v \text{ match in the lexicon.} \end{array} \right\}. \quad (1)$$

Because lexicon-based matching is often noisy, we refine $\mathcal{C}_0$ using a dense semantic encoder. Let $\mathbf{e}_u$ and $\mathbf{e}_v$ denote the encoder representations of the textual definitions of u and v . We retain a link (u, v) as a semantic candidate only if

$$\text{sim}(\mathbf{e}_u, \mathbf{e}_v) = \frac{\mathbf{e}_u^\top \mathbf{e}_v}{\|\mathbf{e}_u\|_2 \|\mathbf{e}_v\|_2} \geq \tau, \quad (2)$$

where $\tau \in (0, 1)$ is a similarity threshold. This yields the filtered candidate set

$$\mathcal{C} = \{(u, v) \in \mathcal{C}_0 \mid \text{sim}(\mathbf{e}_u, \mathbf{e}_v) \geq \tau\}. \quad (3)$$

This semantic filtering stage substantially reduces spurious high-similarity alignments before invoking the more expensive LLM-assisted verification.

LLM-assisted Relation Typing. High similarity does not imply logical equivalence. To prevent **conflated reasoning**, we explicitly type the relationship between aligned entities. For each candidate pair $(u, v) \in \mathcal{C}$, a Large Language Model (LLM) analyzes their detailed definitions to determine the precise semantic relationship $r \in \mathcal{R}_{\text{bridge}}$ (e.g., `functional_overlap`, `approximate_equivalence`). We query the LLM multiple times with diverse prompts or decoding seeds, obtaining a sequence of predictions $\{r^{(m)}\}_{m=1}^M$. The vote count for relation r is

$$\text{vote}(u, v, r) = \sum_{m=1}^M \mathbb{I}[r^{(m)}(u, v) = r], \quad (4)$$

where $\mathbb{I}[\cdot]$ is the indicator function. We instantiate a bridging edge only when some relation r achieves majority support:

$$\mathcal{E}_{\text{bridge}} = \{(u, r, v) \mid (u, v) \in \mathcal{C}, \text{vote}(u, v, r) \geq \delta\}, \quad (5)$$

with δ typically set to a strict majority (e.g., $\delta > M/2$). For reproducibility, the verifier is restricted to a predefined label set $\mathcal{R}_{\text{bridge}}$ with deterministic decoding.

3.3 Structure-Aware Representation Learning

Traversing this dual-ontology graph requires a model that distinguishes between intra-domain consistency and cross-domain bridging. Unlike standard GCNs which treat all neighbors uniformly, or R-GCNs which use fixed weights per relation. We employ a Relational Graph Attention Network (RGAT). This allows the model to dynamically prioritize informative neighbors depending on the specific relation type—crucial for filtering noise in cross-ontology bridges.

Heterogeneous Message Passing. Standard message-passing GNNs update a node representation by aggregating information from its neighbors (Kipf and Welling, 2017). In our setting, both nodes and edges are typed, so we adopt a relation-specific graph attention mechanism (Veličković et al., 2018; Chen et al., 2021) that respects edge types while still allowing the model to learn which neighbors are most informative.

At layer ℓ , each node v has a representation $\mathbf{h}_v^{(\ell)}$. For every relation type $r \in \mathcal{R}$, we apply a relation-specific linear transformation to neighbor features:

$$\tilde{\mathbf{h}}_u^{(r,\ell)} = \mathbf{W}_r^{(\ell)} \mathbf{h}_u^{(\ell)}, \quad (6)$$

where $\mathbf{W}_r^{(\ell)}$ projects node features into a subspace specialized for relation r . For each edge (u, v) of type r , we compute an attention score that measures how relevant neighbor u is for updating v :

$$e_{vu}^{(r,\ell)} = \mathbf{a}_r^\top [\tilde{\mathbf{h}}_v^{(r,\ell)}; \tilde{\mathbf{h}}_u^{(r,\ell)}], \quad (7)$$

where $\mathbf{a}_r$ is a learnable attention vector and $[\cdot; \cdot]$ denotes concatenation. The scores are normalized over the neighbors of v connected via relation r :

$$\alpha_{vu}^{(r,\ell)} = \frac{\exp(e_{vu}^{(r,\ell)})}{\sum_{u' \in \mathcal{N}_r(v)} \exp(e_{vu'}^{(r,\ell)})}, \quad (8)$$

where $\mathcal{N}_r(v)$ denotes the set of neighbors of v under relation r . These coefficients define how much attention v pays to each neighbor u of type r .

Given the attention weights, we compute a relation-specific aggregated message for node v :

$$\mathbf{m}_{v,r}^{(\ell)} = \sum_{u \in \mathcal{N}_r(v)} \alpha_{vu}^{(r,\ell)} \tilde{\mathbf{h}}_u^{(r,\ell)}. \quad (9)$$

Finally, we combine messages from all relation types and apply a nonlinear activation:

$$\mathbf{h}_v^{(\ell+1)} = \sigma \left(\sum_{r \in \mathcal{R}} \mathbf{m}_{v,r}^{(\ell)} \right), \quad (10)$$

where $\sigma(\cdot)$ is a nonlinear activation function (we use ReLU in practice). This formulation extends standard GAT layers (Veličković et al., 2018) to heterogeneous, multi-relational graphs in the spirit of R-GCNs for knowledge graphs (Schlichtkrull et al., 2018), while allowing the model to learn different propagation behaviors for intra-ontology versus cross-ontology edges.

Preventing Semantic Collapse. A core challenge in cross-ontology representation is *semantic collapse*, where the embedding space forces distinct concepts from different ontologies to merge, erasing their unique nuances. We instantiate $\mathcal{L}_{\text{link}}$ as a standard relation-aware link prediction objective with negative sampling (we use DistMult scoring in practice), optimizing a binary cross-entropy loss over observed triples and sampled corruptions. To counter this, we introduce an **ontology preservation loss**. In addition to the standard link prediction loss $\mathcal{L}_{\text{link}}$, we add an auxiliary classification head to predict the source ontology $\phi(v)$ from the node embedding $\mathbf{h}_v^{(L)}$:

$$\mathcal{L}_{\text{onto}} = - \sum_{v \in \mathcal{V}} \log P(\phi(v) | \mathbf{h}_v^{(L)}) \quad (11)$$

By optimizing $\mathcal{L}_{\text{link}}$ and $\mathcal{L}_{\text{onto}}$, we force the encoder to retain domain-discriminative features, ensuring that a TCM concept remains distinguishable from its WM counterpart in the latent space, even after message passing.

Pseudo-Query Augmented Pretraining. A challenge in query-aware retrieval is that the query_to relations required at inference time do not naturally appear during graph-only pretraining.

To resolve this mismatch, we augment stage-1 training with pseudo queries, which are synthetically constructed from existing graph entities and provide a self-supervised signal for learning query-aware relations. Concretely, we introduce an auxiliary query-aware contrastive objective:

$$\mathcal{L}_{\text{query}} = \mathbb{E}_{(v_q, v)} \left[\log \sigma(\mathbf{h}_{v_q}^\top \mathbf{h}_v) + \mathbb{E}_{\tilde{v}} \log \sigma(-\mathbf{h}_{v_q}^\top \mathbf{h}_{\tilde{v}}) \right] \quad (12)$$

where v_q denotes a pseudo query node, v an associated entity, $\tilde{v}$ a negative sample, and $\mathbf{h}^{(L)}$ the final-layer RGAT embedding.

The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{link}} + \lambda \mathcal{L}_{\text{onto}} + \mu \mathcal{L}_{\text{query}}, \quad (13)$$

which ensures that query-aware relations are meaningfully trained and that query injection at inference time performs stable, structure-guided message passing rather than random perturbation.

3.4 Inference: Dynamic Graph Retrieval

Static retrieval fails to capture the specific reasoning needs of a new query. We propose a **dynamic query injection** mechanism that transforms the static knowledge graph into a query-centric reasoning structure at inference time.

Query Injection. Given a user query q , we first compute an initial retriever score $s_0(v | q)$ between q and each entity $v \in \mathcal{V}$ based on semantic similarity. We then select a set of top- k seed entities

$$S_{\text{seed}} = \{v \in \mathcal{V} \mid v \text{ is in the top-}k \text{ by } s_0(v | q)\}. \quad (14)$$

To contextualize the graph with respect to the query, we insert a temporary query node v_q into the graph and connect it to each $u \in S_{\text{seed}}$ via typed query edges. This yields a query-augmented graph

$$\mathcal{G}' = (\mathcal{V} \cup \{v_q\}, \mathcal{E} \cup \mathcal{E}_q),$$

where $\mathcal{E}_q = \{(v_q, \text{query_to}, u) \mid u \in S_{\text{seed}}\}$.

We apply the pretrained RGAT to $\mathcal{G}'$, producing query-conditioned node representations $\mathbf{h}_v^{(L)}(q)$ through message passing that propagates query information along edge categories and cross-ontology relations. We compute the relevance of each node by its similarity to the query node representation:

$$s(v | q) = \cos(\mathbf{h}_v^{(L)}(q), \mathbf{h}_{v_q}^{(L)}(q)). \quad (15)$$

These scores reflect how informative each entity is for answering the query under the dual-ontology graph structure.

Evidence Expansion. To obtain a compact evidence subgraph, we perform a multi-hop expansion starting from S_{seed} . At each hop, we rank frontier candidates by $s(v | q)$ and select the top- k nodes. To satisfy ontology-coherent reasoning, we additionally prioritize candidates that introduce at least one bridging edge in $\mathcal{E}_{\text{bridge}}$ and continue expansion until the induced subgraph contains a bridge-containing path (or a hop/token budget is reached). The final induced subgraph is then linearized as retrieval context for generation.

4 Experimental Setup

4.1 Tasks and Datasets

We evaluate our method on three regimes: TCM-only, WM-only, and integrated TCM–WM. To ensure a balanced evaluation, we utilize subsets derived from three benchmarks, randomly sampling 500 questions for each dataset or label-defined subset. Specifically, we use **MLEC-QA** (Li et al., 2021), a Chinese biomedical dataset, sampling subsets for Clinic, TCM, and TCWM (Integrated TCM–WM). We also employ **CMB** (Wang et al., 2024a), a comprehensive Chinese medical benchmark, sampling subsets for Clinic, TCM, WM, and

TCWM. Finally, we include **MedQA** (Jin et al., 2021) (WM), an English benchmark derived from medical licensing exams, sampling from the official test split. We additionally evaluate a non-medical transfer setting constructed from public MMLU subjects (Hendrycks et al., 2021); construction details are given in Appendix A.3.

4.2 Knowledge Sources and Graph Construction

We use the medical knowledge collection from Zeng and Jia (2025), aggregating TCM sources (SymMap, HERB, TCMSP) (Wu et al., 2019a; Fang et al., 2021; Ru et al., 2014) and WM sources (UMLS, MeSH, DOID) (Bodenreider, 2004; Lipscomb, 2000; Schriml et al., 2012). To support different retrieval paradigms, we derive three substrates from the same source collection: a **Textual Corpus** for standard RAG, a **naively merged graph** for graph-based baselines, and a **dual-ontology graph** for XONTO-RAG. Detailed graph-construction procedures are given in Appendix A.2.

4.3 Baselines

We compare XONTO-RAG against an LLM-only setting and four representative retrieval baselines. To isolate the effect of retrieval strategies, all methods use **Qwen2.5-7B** (Yang et al., 2025) as the backbone generator.

RAG. Standard text-based RAG (Lewis et al., 2020): retrieves top- k passages with a dense retriever, without any graph structure.

GraphRAG. GraphRAG-style retrieval and summarization (Edge et al., 2024) on a *naively merged* graph (disjoint union of TCM and WM; no cross-ontology bridges), testing cross-ontology retrieval without alignment.

MedGraphRAG. A strong medical GraphRAG baseline (Wu et al., 2025), instantiated on the same naively merged graph for domain-specific graph-enhanced retrieval.

OG-RAG. OG-RAG (Sharma et al., 2025) adapted by constructing ontology-grounded hypergraphs from the merged graph, leveraging structure but assuming a single authoritative ontology, without modeling TCM–WM incommensurability.

4.4 Metrics and Implementation

We report **Answer Accuracy (AC)** and **Faithfulness (FF)**. FF is computed via an NLI-style en-

Dataset	Task	Metric	LLM-Only	GraphRAG	MedGraphRAG	RAG	OG-RAG	XONTO-RAG
MLEC-QA	Clinic	AC	56.8	62.2	61.4	62.4	63.6	69.4 (↑9.1%)
		FF	–	71.9	77.4	72.3	72.7	83.2 (↑7.5%)
	TCM	AC	57.8	64.6	64.2	62.4	63.4	72.2 (↑11.8%)
		FF	–	73.8	79.1	73.8	76.1	85.1 (↑7.5%)
	TCWM	AC	58.2	62.4	62.6	60.8	63.0	77.0 (↑22.2%)
		FF	–	73.4	73.0	75.9	75.3	78.4 (↑3.3%)
CMB	Clinic	AC	49.0	62.8	62.8	60.6	62.6	69.4 (↑10.5%)
		FF	–	75.5	76.1	75.9	75.0	82.5 (↑8.4%)
	TCM	AC	53.0	62.2	61.6	57.8	58.2	73.4 (↑18.0%)
		FF	–	79.0	79.4	79.8	74.0	81.8 (↑2.5%)
	WM	AC	58.8	62.8	63.4	59.4	61.8	69.6 (↑9.8%)
		FF	–	74.6	74.6	76.7	76.1	81.5 (↑6.3%)
MedQA	TCWM	AC	62.6	70.8	71.2	66.2	68.0	76.0 (↑6.7%)
		FF	–	80.4	79.0	78.8	76.4	84.8 (↑5.5%)
	WM	AC	50.8	51.4	51.6	50.8	49.8	62.4 (↑20.9%)
		FF	–	76.6	77.9	75.3	71.4	80.2 (↑2.9%)

Table 1: Main results on **Accuracy** (AC) and **Faithfulness** (FF) across MLEC-QA (Clinic/TCM/TCWM), CMB (Clinic/TCM/WM/TCWM), and MedQA (WM). Best scores are marked with **·**, second-best with **·**.

tailment protocol (Falke et al., 2019; Kryscinski et al., 2020; Ji et al., 2023) using Erlangshen-Roberta-330M-NLI as the verifier (details in Appendix A). We report FF for retrieval-based methods; for the LLM-only setting (no retrieval context), FF is not applicable and is shown as “–”. We evaluate scalability with Qwen2.5-32B (Yang et al., 2025), DeepSeek-R1-Distill-Qwen-14B (DeepSeek-AI, 2025; Guo et al., 2025), DeepSeek-V2.5 (236B) (DeepSeek-AI, 2024), and GPT-4o-mini (OpenAI, 2024). All main retrieval comparisons keep the same Qwen2.5-7B generator. Detailed hyperparameters are in Appendix C.

5 Experimental Results

5.1 Main Results

Table 1 reports the main results. We fix the generator backbone across methods; performance differences primarily reflect retrieval/graph construction and cross-ontology alignment mechanisms. XONTO-RAG achieves the highest AC on all 8 datasets and consistently attains the best FF, indicating that the accuracy gains come with stronger evidence grounding rather than spurious improvements. Green values in Table 1 report the *relative improvement* of XONTO-RAG over the second-best method for each metric.

Cross-ontology gains. The largest improvements occur on integrated TCWM tasks, where models must traverse TCM–WM semantic boundaries. For example, XONTO-RAG improves AC on MLEC-TCWM by **+22.2%** over OG-RAG, validating that *explicit, typed* bridging is crucial for cross-paradigm reasoning.

Why baselines underperform. Standard RAG is vulnerable to terminology mismatch and retrieves partially relevant but semantically incompatible evidence across ontologies. GraphRAG and MedGraphRAG exploit structure, yet naively merged graphs introduce noisy cross-ontology paths and distractors, weakening multi-hop reasoning. OG-RAG further lags in mixed-domain settings due to its single-ontology assumption. Even for nominally single-domain questions, retrieval over a merged corpus can still introduce incompatible evidence from the other ontology, and typed provenance helps suppress such distractors.

Faithfulness. Beyond AC, XONTO-RAG also improves FF (e.g., **85.1%** on MLEC-TCM), suggesting that its provenance-aware retrieval and verified bridging reduce evidence mixing and lead to answers that are better supported by the retrieved subgraphs; additional evidence-level relevance and bridge-quality diagnostics, including cMRR@5 for ranking question-relevant evidence, are reported in

Appendix A.4. Direct semantic-collapse scores, a stratified manual audit of 400 typed bridges, and three-run uncertainty/significance comparisons are also reported in Appendix A.4.

5.2 Non-medical Transfer

We further test transfer beyond TCM-WM on a non-medical dual-ontology setting constructed from public MMLU subjects (Appendix A.3). XONTO-RAG again ranks first, reaching 75.0% accuracy and outperforming the strongest baseline by 4.0 points. Although the margin is smaller than on the most challenging medical TCWM benchmarks, it is obtained under the same retrieval budget and generator, indicating that the gain is not tied to a specific medical corpus. This suggests that the benefits of typed bridge induction and query-aware graph navigation extend beyond the medical domain when reasoning must traverse heterogeneous ontologies. Detailed results are reported in Table 4 (Appendix A.3).

Method	MLEC-QA	CMB		MedQA
	TCWM	Clinic	TCM	WM
R-GAT	68.6	65.4	66.2	55.6
Dual-RAG	63.4	64.2	64.2	54.4
GCN-Align	63.0	61.2	60.4	52.6
XONTO-RAG	77.0	69.4	73.4	62.4

Table 2: Ablation Study results (AC, %).

5.3 Ablation Study

To quantify the contribution of each design choice, we perform an ablation study in Table 2 (full results in Appendix A.5). Removing the *typed* cross-ontology bridging edges and falling back to within-ontology reasoning (**R-GAT**) yields a pronounced degradation, dropping MLEC-TCWM accuracy from 77.0% to 68.6% and confirming the importance of reliable, provenance-aware transitions. Replacing our query-aware graph walker with a standard dual-corpus retrieval pipeline (**Dual-RAG**) further reduces performance (77.0% $\rightarrow$ 63.4%), showing that simple retrieval over two corpora is insufficient to assemble coherent multi-hop evidence chains in the dual-ontology graph. Replacing verify-and-vote bridging with a budget-matched embedding-based aligner (**GCN-Align**) also degrades accuracy (77.0% $\rightarrow$ 63.0% on MLEC-TCWM), indicating that the gains do not arise from arbitrary cross-ontology connectivity alone. Overall, all three controls consistently underperform

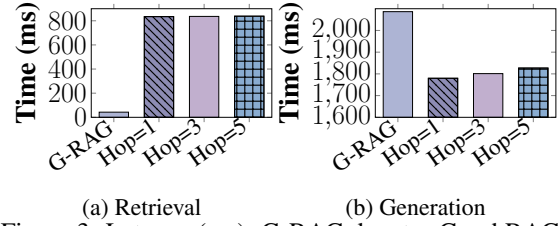


Figure 3: Latency (ms). G-RAG denotes GraphRAG.

XONTO-RAG, so the improvements stem from both typed bridge induction and query-adaptive graph navigation rather than generator changes.

5.4 Efficiency and Robustness

Bridge induction remains an offline preprocessing step and consumes about 1.08×10^8 tokens during graph construction. As shown in Figure 3, although XONTO-RAG spends more time on retrieval than GraphRAG, its more compact retrieved evidence leads to faster generation, keeping end-to-end latency in a similar range: 2,129.2 ms for GraphRAG versus 2,614.3/2,637.5/2,664.9 ms for XONTO-RAG with Hop set to 1/3/5. Moreover, increasing the Hop budget from 1 to 5 changes total latency by only 50.6 ms, indicating that moderate multi-hop expansion adds limited overhead.

We further assess robustness across diverse LLM backbones in Table 3 (full results in Appendix A.6). XONTO-RAG already achieves the best MLEC-TCWM accuracy with the 7B backbone (77.0%), while larger Chinese backbones improve MLEC-QA/CMB further (up to 78.4% on CMB-TCWM) and GPT-4o-mini performs best on English MedQA (72.8%), highlighting the role of language alignment.

Model	MLEC-QA	CMB		MedQA
	TCWM	Clinic	TCM	WM
Qwen2.5-7B	77.0	69.4	73.4	62.4
DeepSeek-R1-14B	73.2	67.4	68.6	62.8
Qwen2.5-32B	74.0	76.6	74.8	65.0
GPT-4o-mini	51.0	61.6	54.2	72.8
DeepSeek-V2.5	70.8	71.2	71.0	60.2

Table 3: Performance of XONTO-RAG with different backbone LLMs on selected benchmarks (AC, %).

6 Conclusion

In this paper, we introduced XONTO-RAG, a framework for cross-ontology retrieval and reasoning over incommensurable knowledge systems. By combining a dual-ontology graph with typed bridging edges and a query-aware RGAT retriever, our method preserves ontology provenance while enabling effective evidence traversal across seman-

tic gaps. Across medical benchmarks, XONTORAG consistently outperforms standard and graph-based baselines, and transfer results suggest that the approach can extend beyond the medical domain. Future work will investigate lighter-weight bridge updates for evolving knowledge bases.

Limitations

This work has several limitations. First, graph construction requires a weak cross-reference or candidate source, but not a complete or high-precision one-to-one alignment; seed coverage may nevertheless be scarce in some domains. Second, although bridge induction is performed offline, the verify-and-vote pipeline still incurs nontrivial preprocessing time overhead, and its behavior under frequent knowledge-base updates has not yet been systematically evaluated. Third, we report both a manual audit of 400 typed bridges and an independent GPT-5 consistency check on 500 bridges, but neither substitutes for clinician validation; stronger expert evaluation would further strengthen confidence in the induced relation types. Finally, while a small non-medical transfer experiment suggests broader promise, the current relation schema and evidence linearization remain tailored to the medical setting, and performance also depends on language/backbone alignment.

References

- Olivier Bodenreider. 2004. [The unified medical language system \(umls\): Integrating biomedical terminology](#). *Nucleic Acids Research*, 32(suppl extunder-score 1):D267–D270.
- J Harry Caulfield, Harshad Hegde, Vincent Emonet, Nomi L Harris, Marcin P Joachimiak, Nicolas Matentzoglou, HyeongSik Kim, Sierra Moxon, Justin T Reese, Melissa A Haendel, Peter N Robinson, and Christopher J Mungall. 2024. [Structured prompt interrogation and recursive extraction of semantics \(spires\): A method for populating knowledge bases using zero-shot learning](#). *Bioinformatics*, 40(3):btac104.
- Meiqi Chen, Yuan Zhang, Xiaoyu Kou, Yuntao Li, and Yan Zhang. 2021. [r-gat: Relational graph attention network for multi-relational graphs](#). *Preprint*, arXiv:2109.05922.
- Muhao Chen, Yingtao Tian, Kai-Wei Yang, and Carlo Zaniolo. 2017. [Multilingual knowledge graph embeddings for cross-lingual knowledge alignment](#). In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*.
- Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. [Editing factual knowledge in language models](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506.
- DeepSeek-AI. 2024. [Deepseek-v2.5](#). Hugging Face model card.
- DeepSeek-AI. 2025. [Deepseek-r1-distill-qwen-14b](#). Hugging Face model card.
- Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, and Robert Osazuwa. 2024. [From local to global: A graph rag approach to query-focused summarization](#). *arXiv preprint arXiv:2404.16130*.
- Jérôme Euzenat and Pavel Shvaiko. 2013. *Ontology Matching*, second edition. Springer.
- Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. [Ranking generated summaries by correctness: An interesting but challenging application for natural language inference](#). In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 2214–2220. Association for Computational Linguistics.
- ShuangSang Fang, Lei Dong, Liu Liu, JinCheng Guo, LianHe Zhao, JiaYuan Zhang, DeChao Bu, XinKui Liu, PeiPei Huo, WanChen Cao, QiongYe Dong, JiaRui Wu, Xiaoxi Zeng, Yang Wu, and Yi Zhao. 2021. [Herb: a high-throughput experiment-and reference-guided database of traditional chinese medicine](#). *Nucleic Acids Research*, 49(D1).
- Charles J. Fillmore. 1982. Frame semantics. In The Linguistic Society of Korea, editor, *Linguistics in the Morning Calm: Selected Papers from SICOL-1981*, pages 111–137. Hanshin Publishing Co., Seoul.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Zhyu Gao, Aixin Liu, and 61 others. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *arXiv preprint arXiv:2501.12948*.
- Marta Hanson and Victor Kumar. 2025. [Medical bilingualism: Reframing the overlap between medical systems in recent medical history and anthropology](#). *Chinese Medicine and Culture*, 8(3):105–118.
- Jinglin He, Yunqi Guo, Lai Kwan Lam, Waikie Leung, Lixing He, Yuanan Jiang, Chi Chiu Wang, Guoliang Xing, and Hongkai Chen. 2025. [Opentcm: A graphrag-empowered llm-based system for traditional chinese medicine knowledge retrieval and diagnosis](#). *arXiv preprint arXiv:2504.20118*.

- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *International Conference on Learning Representations*.
- Yuntong Hu, Zhihan Lei, Zheng Zhang, Bo Pan, Chen Ling, and Liang Zhao. 2025. [Grag: Graph retrieval-augmented generation](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4145–4157.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. [Survey of hallucination in natural language generation](#). *ACM Computing Surveys*, 55(12):1–38.
- Pengcheng Jiang, Cao Xiao, Minhao Jiang, Parminder Bhatia, Taha Kass-Hout, Jimeng Sun, and Jiawei Han. 2025. [Reasoning-enhanced healthcare predictions with knowledge graph community retrieval](#). In *The Thirteenth International Conference on Learning Representations*.
- Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. [Hipporag: Neurobiologically inspired long-term memory for large language models](#). *arXiv preprint arXiv:2405.14831*.
- Di Jin, Eileen Pan, Nour Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. [What disease does this patient have? a large-scale open-domain question answering dataset from medical exams](#). *Applied Sciences*, 11(14):6421.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.
- Thomas N. Kipf and Max Welling. 2017. [Semi-supervised classification with graph convolutional networks](#). In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*.
- Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. [Evaluating the factual consistency of abstractive text summarization](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 9332–9346. Association for Computational Linguistics.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, and Tim Rocktäschel. 2020. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). In *Advances in Neural Information Processing Systems*.
- Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. 2024. [Revisiting catastrophic forgetting in large language model tuning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4297–4308, Miami, Florida, USA. Association for Computational Linguistics.
- Jing Li, Shangping Zhong, and Kaizhi Chen. 2021. [MLEC-QA: A chinese multi-choice biomedical question answering dataset](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8862–8874, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Carolyn E. Lipscomb. 2000. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265–266.
- OpenAI. 2024. [Gpt-4o mini model documentation](#). OpenAI API documentation.
- Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhanian, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, Russa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, and Damien Graux. 2023. [Large language models and knowledge graphs: A survey](#). *arXiv preprint arXiv:2308.06374*.
- Xinying Qian, Ying Zhang, Yu Zhao, Baohang Zhou, Xuhui Sui, Li Zhang, and Kehui Song. 2024. [TimeR4: Time-aware retrieval-augmented large language models for temporal knowledge graph question answering](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6942–6952.
- Jinlong Ru, Peng Li, Jian Wang, Weiqiang Zhou, Bo Li, Cong Huang, Peng Li, Zihu Guo, Wei Tao, Yi Yang, Xiaogang Xu, Yan Li, Yuhong Wang, and Ling Yang. 2014. [Tcmisp: A database of systems pharmacology for drug discovery from herbal medicines](#). *Journal of Cheminformatics*, 6(13):1–6.
- Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. [Colbertv2: Effective and efficient retrieval via lightweight late interaction](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734.
- Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. [Modeling relational data with graph convolutional networks](#). In *Proceedings of the European Semantic Web Conference (ESWC)*, pages 593–607. Springer.
- Lynn M. Schriml, Cesar Arze, Suresh Nadendla, Yun Chang, Mark Mazaitis, Valerie Felix, Gang Feng, and Warren A. Kibbe. 2012. [Disease ontology: A backbone for disease semantic integration](#). *Nucleic Acids Research*, 40(D1):D940–D946.
- Kartik Sharma, Peeyush Kumar, and Yunqing Li. 2025. [Og-rag: Ontology-grounded retrieval-augmented generation for large language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

- Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. [REPLUG: Retrieval-augmented black-box language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.
- 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(7973):1–9.
- Zequan Sun, Wei Hu, Qingheng Zhang, and Yuzhong Qu. 2018. [Bootstrapping entity alignment with knowledge graph embedding](#). In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4396–4402. International Joint Conferences on Artificial Intelligence Organization.
- Paul Thagard and Jing Zhu. 2003. Acupuncture, incommensurability, and conceptual change. *International Studies in the Philosophy of Science*, 17(1):79–102.
- Lukas Thede, Yu Cao, and Iryna Gurevych. 2025. [Wikibigedit: Understanding the limits of life-long knowledge editing in llms](#). *arXiv preprint arXiv:2503.05683*.
- Sabrina Toro, Anna V. Anagnostopoulos, Susan M. Bello, Kai Blumberg, Rhiannon Cameron, Leigh Carmody, Alexander D. Diehl, Damion M. Dooley, William D. Duncan, Petra Fey, Pascale Gaudet, Nomi L. Harris, Marcin P. Joachimiak, Leila Kiani, Tiago Lubiana, Monica C. Munoz-Torres, Shawn O’Neil, David Osumi-Sutherland, Aleix Puig-Barbe, and 11 others. 2024. [Dynamic retrieval augmented generation of ontologies using artificial intelligence \(dragon-ai\)](#). *Journal of Biomedical Semantics*, 15(1):19.
- Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. [Graph attention networks](#). In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.
- Xidong Wang, Guiming Chen, Song Dingjie, Zhang Zhiyi, Zhihong Chen, Qingying Xiao, Junying Chen, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2024a. [CMB: A comprehensive medical benchmark in Chinese](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6184–6205. Association for Computational Linguistics.
- Yongbo Wang, Xiangying Ren, Kuang Gao, Mukun Chen, Qiao Huang, Siyu Yan, Yan Zhu, Xin Sun, Yaolong Chen, Long Ge, Jinguang Gu, Feng Gao, Wenbin Hu, Liang Hong, Chen Zhao, Hongcai Shang, and Yinghui Jin. 2024b. [Ontology of clinical practice guidelines for integrated traditional chinese and western medicine](#). *Journal of Evidence-Based Medicine*, 17(3).
- Ziheng Wang, Qing Lv, Xiaofei Lan, and Yu Zhang. 2018. [Cross-lingual knowledge graph alignment via graph convolutional networks](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.
- Orion Weller, Michael Boratko, Iftexhar Naim, and Jinhyuk Lee. 2025. [On the theoretical limitations of embedding-based retrieval](#). *arXiv preprint arXiv:2508.21038*.
- 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.
- Yang Wu, Feilong Zhang, Kuo Yang, Shuangfang Fang, Dechao Bu, Hui Li, Liang Sun, Hairuo Hu, Kuo Gao, Wei Wang, Xuezhong Zhou, Yi Zhao, and Jianxin Chen. 2019a. [SymMap: An integrative database of traditional chinese medicine enhanced by symptom mapping](#). *Nucleic Acids Research*, 47(D1):D1110–D1117.
- Yuting Wu, Xiao Liu, Yansong Feng, Zheng Wang, Rui Yan, and Dongyan Zhao. 2019b. [Relation-aware entity alignment for heterogeneous knowledge graphs](#). *arXiv preprint arXiv:1908.08210*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. [Qwen2.5 technical report](#).
- Jingqi Zeng and Xiaobin Jia. 2024. [Graph neural networks for quantifying compatibility mechanisms in traditional chinese medicine](#). *arXiv preprint arXiv:2411.11474*.
- Jingqi Zeng and Xiaobin Jia. 2025. [Traditional chinese medicine multidimensional knowledge graph](#). Zenodo. Version v2. Accessed: 2025-10-26.
- Chengshuai Zhao, Riccardo De Maria, Tharindu Kumarage, Kumar Satvik Chaudhary, Garima Agrawal, Yiwen Li, Jongchan Park, Yuli Deng, Ying-Chih Chen, and Huan Liu. 2025. [Cyberbot: Ontology-grounded retrieval augmented generation for reliable cybersecurity education](#). In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*.

A Additional Experimental Details and Results

A.1 Metrics

We report two evaluation metrics: **Answer Accuracy (AC)** and **Faithfulness (FF)**.

Answer Accuracy (AC). AC measures end-task correctness and is computed as the percentage of questions whose final predicted answer matches the gold answer under the benchmark’s evaluation protocol.

Faithfulness (FF). FF measures whether the generated answer is supported by the retrieved evidence and is reported as the percentage of answers that are *fully grounded* in the retrieved context (Ji et al., 2023). We follow an NLI-style entailment protocol (Falke et al., 2019; Kryscinski et al., 2020): we treat each answer sentence as a hypothesis and the retrieved context as the premise set, and we mark a sentence as supported if at least one retrieved sentence entails it. An answer is considered faithful only if all its sentences are supported. In practice, we use Erlangshen-Roberta-330M-NLI as the NLI verifier.

Evidence relevance (cMRR@5). cMRR@5 measures whether the retriever ranks question-relevant evidence near the top of the retrieved list. For each query, we ask an LLM judge to determine whether each of the top-5 evidence items is relevant to answering the question, then compute the reciprocal rank of the first relevant item and average across queries. In practice, we use Qwen2.5-32B-Instruct as the judge, following the same retrieval outputs used for answer generation.

A.2 Implementation Details

All experiments follow the setup in Section 4. Unless otherwise noted, we evaluate on the sampled subsets described in Section 4 and keep the retrieval pipeline unchanged across comparisons.

Compute and APIs. Local experiments are run on an Ubuntu 22.04 LTS server with $8\times$ NVIDIA H100 80GB GPUs, an Intel Xeon Platinum 8558P CPU @ 2.70GHz, and 1 TB RAM. DeepSeek-V2.5 (236B) and GPT-4o-mini are accessed via their respective APIs; other backbones are run locally.

Graph Construction Details. All retrieval methods are built from the same unified medical knowledge collection used in Section 4.2. For standard

Method	LLM	Graph	RAG	OG	Med-RAG	Xonto
AC	70.5	70.5	71.0	71.0	71.0	75.0

Table 4: Non-medical transfer on MMLU (AC, %).

Subset	XONTO-RAG	MedGraphRAG	OG-RAG
TCWM	52.54	49.92	48.89
Clinic	59.44	57.52	58.38
TCM	39.26	38.98	38.17
WM	35.67	35.00	35.10

Table 5: Evidence relevance cMRR@5 (%) on representative subsets of CMB.

RAG, we linearize entities and relations into a textual corpus while preserving ontology tags in the serialized passages. For graph-based baselines, we construct a naively merged graph by taking the disjoint union of TCM and WM entities and relations together with the raw concept mappings provided in the dataset release. For XONTO-RAG, we start from the same base graph but replace raw cross-ontology links with typed bridges induced by the verify-and-vote pipeline in Section 3.2. After semantic filtering and offline pruning, the bridge induction stage reduces roughly 3.4×10^7 initial candidates to about 5×10^5 typed cross-ontology edges.

A.3 Non-medical Transfer Setting

To examine transfer beyond TCM–WM, we construct a second dual-ontology setting from public MMLU subjects (Hendrycks et al., 2021). The graph-construction corpus contains about 2,400 instances drawn from 12 subjects spanning legal/governance and religion/ethics, from which we build a dual-ontology graph with 2,432 nodes, 7,521 edges, and 100 cross-ontology bridges. Evaluation uses 200 held-out multiple-choice questions from four subjects disjoint from the graph-construction corpus, while keeping the generator and retrieval budgets fixed.

A.4 Relevance and Bridge-Quality Diagnostics

Table 5 shows that our method achieves the highest cMRR@5 overall, with the clearest gain on the mixed TCWM subset and smaller differences on the more single-ontology subsets, consistent with its primary benefit in cross-ontology retrieval. We additionally audit the induced bridges directly: on 500 randomly sampled cross-ontology bridges,

Method	CMB-TCM	CMB-TCWM	CMB-Clinic	CMB-WM
XONTO-RAG	1.84	1.66	1.02	0.89
GraphRAG	2.38	1.89	1.92	1.68
MedGraphRAG	2.25	2.44	1.71	1.58
OG-RAG	2.18	2.19	1.53	1.35

Table 6: Direct semantic-collapse scores on CMB subsets (lower is better).

Bridge type	N	Precision $\uparrow$	Main error type
approx_equiv	100	92%	functional_overlap
functional_overlap	100	94%	approx_equiv
causal_influence	100	94%	functional_overlap
conflict	100	97%	causal_influence
Overall	400	94.25%	–

Table 7: Manual audit of typed bridge edges, with 100 sampled edges per relation type.

GPT-5 agrees with the voted relation label in 92.4% of cases, suggesting that the verify-and-vote pipeline yields reasonably reliable relation labels.

We also evaluate semantic collapse directly rather than treating answer accuracy as a proxy. For each QA instance, a reviewer LLM examines the question, retrieved evidence, and generated answer, and assigns an integer score from 0 to 10; lower values indicate fewer cases in which TCM and WM concepts are incorrectly merged, provenance is lost, or a cross-ontology reasoning chain is broken.

The evaluator distinguishes five failure types: erroneous alignment, conflated reasoning, provenance-blind mixing, retrieval derailment, and fragmented retrieval. XONTO-RAG obtains the lowest score on every subset, including 1.66 on TCWM and 0.89 on WM.

For the manual audit, an edge is counted as precise only when its endpoint pairing and voted relation type are supported by the source descriptions; the most frequent competing relation is recorded as the main error type. The overall precision is 94.25%, and errors are concentrated in adjacent bridge-type confusions rather than arbitrary cross-ontology links. The separate GPT-5 audit of 500 bridges (92.4% agreement) is an automated consistency check, not clinician validation.

Across the three runs, XONTO-RAG remains above the strongest baseline on all three MLEC-QA subsets. These results quantify run-to-run variation while preserving the same generator, knowledge sources, and retrieval budget used in the main comparisons.

Semantic-collapse evaluator prompt.

Instruction. Judge only whether the QA item, retrieved evidence, and generated answer exhibit semantic collapse. Do not judge general medical correctness. Semantic collapse occurs when TCM functional/systemic concepts and WM anatomical/material concepts are forced into one semantic space and are incorrectly merged, equated, or mixed without an explicit typed bridge or provenance.

Failure types. Use any applicable labels: `erroneous_alignment` (incorrect equivalence, e.g., “Qi deficiency” = hypoxia), `conflated_reasoning` (TCM and WM explanations mixed without qualification), `provenance_blind_mixing` (evidence sources are not distinguished), `retrieval_derailment` (surface similarity retrieves incompatible evidence), and `fragmented_retrieval` (a cross-ontology chain cannot be completed).

Score. Return an integer `collapse_score` from 0 to 10: 0 means no collapse; 1–2 minor confusion; 3–4 light-to-moderate source ambiguity; 5–6 an unclear ontology boundary or unsupported cross-ontology substitution; 7–8 an explicit false equivalence or causal substitution; and 9–10 reasoning dominated by an incorrect cross-ontology merge. Irrelevant retrieval or an error entirely within one ontology should not by itself receive a high score. Explicitly stating that TCM and WM viewpoints are distinct should receive a low score.

Output. Return one JSON object only with `collapse_score`, `severity` (none, minor, moderate, severe, or critical), `collapse_types`, a short rationale, `evidence_spans`, and `confidence` in $[0, 1]$. Use the score mapping 0=none, 1–2=minor, 3–6=moderate, 7–8=severe, and 9–10=critical.

Examples. “Qi deficiency is hypoxia” without qualification is a severe erroneous alignment (typically 7–8). “From the TCM perspective ...; from the WM perspective ...; these are not directly equivalent” preserves the boundary (typically 0–2).

A.5 Full Ablation Results

Table 9 reports the full ablation results (AC) across all datasets and tasks. **R-GAT** removes the typed bridging edges and performs retrieval on the naively merged graph, isolating the impact of explicit cross-ontology links. **Dual-RAG** retains the dual-ontology graph with bridging edges but replaces the query-aware graph walker with a standard GraphRAG-style retrieval procedure, isolating the impact of query-aware dynamic inference. **GCN-Align** replaces verify-and-vote bridge induction with an embedding-based aligner under the same bridge budget.

A.6 Full Backbone Comparison

Table 10 reports the full backbone comparison (AC) while keeping the retrieval configuration fixed. This study evaluates whether the gains from XONTO-RAG are robust to generator changes and highlights the interaction between backbone ca-

Subset	Strongest baseline	XONTO-RAG	Gain	p -value
Clinic	RAG: 63.5 ± 0.9	70.4 ± 0.9	+6.9	9.24×10^{-5}
TCM	MedGraphRAG: 65.3 ± 0.9	72.1 ± 1.3	+6.8	0.0182
TCWM	OG-RAG: 63.0 ± 1.2	76.8 ± 0.9	+13.8	0.00768

Table 8: Three-run mean $\pm$ sample standard deviation (temperature 0.7) and reported significance comparisons on MLEC-QA answer accuracy. The p -values compare XONTO-RAG with the strongest baseline on the same subset.

Dataset	Task	R-GAT	Dual-RAG	GCN-Align	XONTO-RAG
MLEC-QA	Clinic	63.4	64.2	63.6	69.4
	TCM	67.2	65.4	64.0	72.2
	TCWM	68.6	63.4	63.0	77.0
CMB	Clinic	65.4	64.2	61.2	69.4
	TCM	66.2	64.2	60.4	73.4
	WM	63.8	62.0	62.6	69.6
	TCWM	71.2	70.6	65.6	76.0
MedQA	WM	55.6	54.4	52.6	62.4

Table 9: Ablation Study

pability, language matching, and cross-ontology retrieval quality.

A.7 Retriever Hyperparameter Sensitivity

We analyze two inference-time retrieval hyperparameters: the context budget (`context_top_k`) and reasoning depth (H , hops). Table 11 and Figure 4–5 report AC results on four representative datasets.

Context size (k ; `context_top_k`). Performance generally improves as k increases to a moderate value ($k \approx 20$) and then plateaus or degrades, suggesting that retrieving too few nodes misses evidence while retrieving too many introduces irrelevant noise.

Reasoning hops (H ; hops). We find that shallow expansion ($H = 1$) is often insufficient for multi-hop reasoning, while overly deep walks ($H > 5$) can drift to distant concepts and reduce accuracy. A moderate hop depth ($H = 3$) is robust across datasets, with $H = 5$ occasionally beneficial for more complex reasoning chains.

A.8 Training Loss Weight Sensitivity

We provide full sensitivity analysis for the two training loss weights in Section 3.3: the ontology-preservation coefficient λ and the pseudo-query objective coefficient μ .

Figure 6 and Figure 7 report results across four datasets.

Weight λ . Across datasets, moderate ontology regularization is preferred: too large λ tends to over-constrain representations, while too small λ weakens ontology preservation.

Weight μ . Large μ can noticeably hurt performance, while moderate values yield stable improvements; overall, $\mu = 0.05$ is a robust default across datasets.

B Case Study

B.1 Symptomatic AAA With Distal Embolization (MedQA)

As shown in the MedQA example (Figure 8), we examine a high-risk vascular presentation in which a small, unruptured abdominal aortic aneurysm (AAA) (4.1 cm) is accompanied by progressive abdominal pain and distal ischemic changes consistent with “blue toe” syndrome. This setting is clinically misleading because a diameter-based heuristic (e.g., surveillance thresholds for asymptomatic AAA) can bias systems toward non-urgent management. A vanilla LLM (or a retrieval system that overweights aneurysm size alone) may therefore be misled into recommending clinical surveillance or elective repair, failing to incorporate the diagnostic significance of symptoms and embolic phenomena.

In contrast, the provided answer correctly emphasizes that symptomatic AAA warrants urgent intervention regardless of size, particularly when accompanied by signs suggestive of distal em-

Dataset	Task	Qwen2.5-7B	DeepSeek-R1-14B	Qwen2.5-32B	GPT-4o-mini	DeepSeek-V2.5
MLEC-QA	Clinic	69.4	72.2	76.6	58.8	71.8
	TCM	72.2	72.8	76.0	50.6	74.2
	TCWM	77.0	73.2	74.0	51.0	70.8
CMB	Clinic	69.4	67.4	76.6	61.6	71.2
	TCM	73.4	68.6	74.8	54.2	71.0
	WM	69.6	72.2	74.4	59.4	72.0
	TCWM	76.0	73.8	78.4	59.4	74.8
MedQA	WM	62.4	62.8	65.0	72.8	60.2

Table 10: Comparison of various backbone LLMs (AC, %).

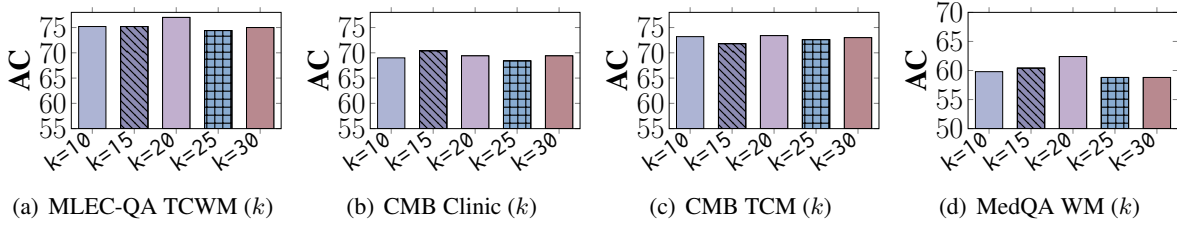


Figure 4: AC sensitivity to context top- k (context_top.k) across four datasets (Table 11).

Setting	MLEC-QA	CMB		MedQA
	TCWM	Clinic	TCM	WM
<i>Impact of context size (k) with fixed hops ($H = 3$)</i>				
$k = 10$	75.2	69.0	73.2	59.8
$k = 15$	75.2	70.4	71.8	60.4
$k = 20$ (Ours)	77.0	69.4	73.4	62.4
$k = 25$	74.4	68.4	72.6	58.8
$k = 30$	75.0	69.4	73.0	58.8
<i>Impact of reasoning hops (H) with fixed context ($k = 20$)</i>				
$H = 1$	72.6	66.8	68.8	58.2
$H = 3$ (Ours)	77.0	69.4	73.4	62.4
$H = 5$	74.8	68.6	73.8	58.0
$H = 7$	73.6	69.4	71.8	56.0
$H = 9$	73.2	68.8	72.4	55.4

Table 11: Parameter sensitivity analysis on context top- k (context_top.k) and reasoning hops (H ; hops). The default setting is $k = 20$, $H = 3$ (marked as Ours). Best results are marked with $\cdot$.

bolization. The retrieved context—while not explicitly enumerating AAA-specific procedural thresholds—captures a structured cluster of cardiovascular and lipid-metabolism entities and ontology mappings (ICD-11 and DOID→ICD-11), which can support risk-factor grounding (hypertension, hyperlipidaemia, smoking history, and prior transient ischemic attacks) and reinforce a vascular-complication interpretation rather than a purely geometric (diameter) one. This case illustrates how structured retrieval can help prevent “single-feature anchoring” by reframing the problem within a broader vascular-risk context. Accordingly, the

system selects emergency open surgical repair as the most appropriate management, aligning with the principle that symptomatic aneurysms carry elevated rupture and complication risk and should not be managed by surveillance alone.

B.2 Aortic Stenosis vs. Retrieval Mismatch (Mlec-TCWM)

We next analyze a long-standing cardiology presentation from Mlec-TCWM (Figure 9), featuring exertional palpitations and dyspnea, progressive angina, intermittent syncope, and classic auscultatory findings: a systolic thrill at the right sternal border and a grade III ejection systolic murmur, with cardiac enlargement displaced leftward and downward. These findings are highly characteristic of aortic stenosis, and the correct diagnosis is idiopathic aortic stenosis.

Notably, the retrieved context in this example is largely misaligned with the question’s core discriminative features, consisting primarily of neurological and movement disorders, substance-withdrawal syndromes, electrolyte disturbances, and a limited subset of cardiovascular entities that do not directly encode the hallmark physical-examination relationships (e.g., “ejection systolic murmur at the right upper sternal border → aortic stenosis”). In such cases, an LLM may be exposed to “semantic distraction,” whereby unrelated medical labels dilute or obscure the relevant clinical pattern. The fact that the final answer remains correct in this exam-

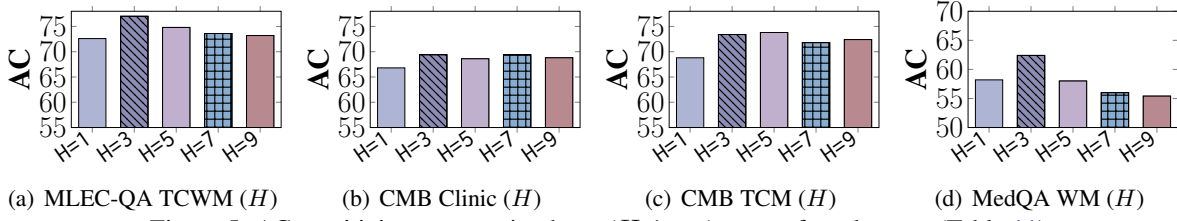


Figure 5: AC sensitivity to reasoning hops (H ; hops) across four datasets (Table 11).

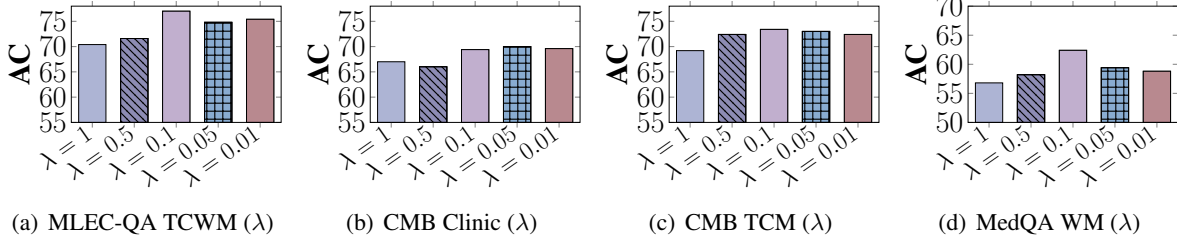


Figure 6: AC sensitivity to the ontology-preservation weight (λ) across four datasets.

ple highlights an important methodological implication for medical RAG: retrieval quantity is not equivalent to retrieval utility.

This case therefore serves as a negative-control-style analysis demonstrating that, without a medically structured graph that explicitly links symptom–sign–valvular-lesion relationships, retrieval can be dominated by irrelevant neighbors. A MedGraphRAG-style system would ideally enforce hierarchical or relation-aware linking (e.g., auscultation finding $\rightarrow$ lesion $\rightarrow$ hemodynamic consequence $\rightarrow$ symptom triad) to ensure that aortic-stenosis–relevant entities are prioritized over unrelated ICD-11 categories.

B.3 TCM Pattern–Formula Linking Under Mixed Exterior–Interior Features (Mlec-TCM)

In the Mlec-TCM example (Figure 10), a 15-year-old male presents with acute epigastric pain, aversion to cold with preference for warmth, poor appetite, and frequent belching, accompanied by a white tongue coating and a wiry, tight pulse. This case is challenging because it exhibits mixed signals that can lead to competing interpretations: a superficial reading may overemphasize belching and appetite reduction (suggesting food stagnation), or overemphasize aversion to cold (suggesting an exterior wind-cold pattern). The correct selection, however, requires integrating both interior cold constraint and exterior involvement, leading to the combined prescription Liangfu Wan + Xiangsu San.

The retrieved context provides a structured set of TCM formula entities and TCM disorder/pattern terminology (e.g., “stomach ache,” “acid regurgitation,” “disharmony between liver and stom-

ach pattern”), which is conducive to a graph-based reasoning process: symptoms map to pattern candidates, patterns map to treatment methods, and treatment methods map to formula choices. Importantly, the answer’s rationale explicitly leverages this structure by attributing complementary roles: Liangfu Wan for warming the middle and dispersing cold to relieve pain, and Xiangsu San for releasing the exterior and regulating qi, thereby addressing the mixed exterior–interior presentation. This case highlights the value of retrieval that includes terminological definitions and formula–function mappings, which can reduce ambiguity in pattern differentiation and support transparent, clinically interpretable formula selection.

B.4 Acute Hepatitis Classification With Histopathologic Anchors (Mlec-Clinic)

Finally, the Mlec-Clinic example (Figure 11) describes an 18-year-old female with acute onset (2 days) of high fever and nausea, jaundice, markedly elevated transaminases (GPT 1000 U), and biopsy findings including hepatocellular swelling, ballooning degeneration, spotty necrosis, and eosinophilic bodies. Such a presentation can be misleading because the magnitude of transaminase elevation and the presence of jaundice may trigger an over-severity bias toward severe hepatitis categories, even when key indicators of fulminant progression are not described.

The retrieved context includes a compact set of hepatic and infectious disease entities spanning acute viral hepatitis and multiple chronic hepatitis categories, providing a taxonomy-level scaffold that can help the model separate acute from chronic etiologies and align histopathologic descriptors

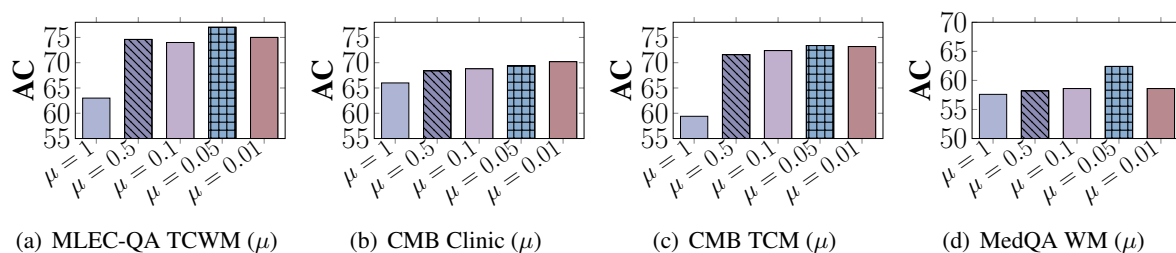


Figure 7: AC sensitivity to the pseudo-query objective weight (μ) across four datasets.

with acute hepatocellular injury. The final response correctly selects acute uncomplicated hepatitis, explicitly contrasting it against acute severe/subacute severe hepatitis (which typically requires additional evidence of progressive failure) and against chronic entities (which are inconsistent with the short time course). This case demonstrates how retrieval that supplies structured disease categories can support clinically appropriate classification by preventing premature escalation based on a single salient feature (enzyme magnitude), and by promoting evidence integration across time course, laboratory pattern, and pathology.

Question

A 69-year-old male with a longstanding history of hypertension and high cholesterol presents with abdominal pain and 'bruising on his feet'. The patient states that his symptoms started about a week ago and have steadily worsened. He describes the abdominal pain as mild to moderate, dull, and deeply localized to the umbilical region. Past medical history is significant for 2 transient ischemic attacks 6 months prior, characterized by a sudden right-sided weakness and trouble speaking but recovered fully within 30 minutes. Current medications are sildenafil 100 mg orally as needed. Patient reports a 30-pack-year smoking history and heavy alcohol use on the weekends. Review of systems is significant for decreased appetite and feeling easily full. Vitals are temperature 37°C (98.6°F), blood pressure 155/89 mm Hg, pulse 89/min, respirations 16/min, and oxygen saturation 98% on room air. On physical examination, the patient is alert and cooperative. The cardiac exam is normal. Lungs are clear to auscultation. Carotid bruit present on the right. The abdomen is soft and nontender. Bowel sounds present. A pulsatile abdominal mass is felt in the lower umbilical region. Patient's feet have the following appearance seen in the picture. Abdominal ultrasound reveals the presence of an abdominal aortic aneurysm (AAA). Contrast CT reveals a small, unruptured AAA (diameter 4.1 cm). High flow supplemental oxygen and fluid resuscitation are initiated. Morphine sulfate and metoprolol are administered. Which of the following is the best course of treatment for this patient?

- A. Discharge with clinical surveillance and ultrasound every 6 months
- B. Elective endovascular repair
- C. Elective open surgical repair
- D. Emergency open surgical repair

Retrieved Context

Cardiovascular diseases (ICD-11):

- Other specified coronary vasospastic disease (ICD-11: BA85.Y) Chinese name: 其他特指的冠状血管痉挛性疾病
- Other specified angina pectoris (ICD-11: BA40.Y) Chinese name: 其他特指的心绞痛
- Stable angina (ICD-11: BA40.1) Chinese name: 稳定性心绞痛
- Unstable angina (ICD-11: BA40.0) Chinese name: 不稳定性心绞痛
- Chronic ischaemic heart disease, unspecified (ICD-11: BA5Z) Chinese name: 未特指的慢性缺血性心脏病

Lipid metabolism disorders (ICD-11):

- Hyperlipoproteinaemia, unspecified (ICD-11: 5C80.Z) 中文: 未特指的高脂蛋白血症
- Hyperalphalipoproteinaemia (ICD-11: 5C80.3) 中文: 高 α 脂蛋白血症
- Mixed hyperlipidaemia (ICD-11: 5C80.2) 中文: 混合性高脂血症
- Primary hypercholesterolaemia (ICD-11: 5C80.00) 中文: 原发性高胆固醇血症
- Hypertriglyceridaemia (ICD-11: 5C80.1) 中文: 高甘油三酯血症

Heart failure (ICD-11):

- Right ventricular failure (ICD-11: BD13) 中文: 右心室衰竭
- Left ventricular failure with mid-range ejection fraction (ICD-11: BD11.1) 中文: 射血分数中间范围的心力衰竭

Related ontology mappings (DOID → ICD-11):

- DOID:8805 → ICD-11: BA40.0 • DOID:6000 → ICD-11: BD1Z • DOID:0111369 → ICD-11: 5C80.3 • DOID:1172 → ICD-11: 5C80.1 • DOID:14118 → ICD-11: 5C80.Z

Answer

Answer: D

The patient has a small but symptomatic abdominal aortic aneurysm (4.1 cm) accompanied by abdominal pain and signs suggestive of distal embolization ('blue toe syndrome'). Symptomatic AAAs require urgent surgical intervention regardless of size due to an increased risk of rupture. Emergency open surgical repair is therefore the most appropriate management. Clinical surveillance or elective repair would delay definitive treatment and increase the risk of catastrophic rupture.

Figure 8: Emergency Management Decision for Symptomatic Abdominal Aortic Aneurysm with "Blue Toe Syndrome" (MedQA)

Mlec-TCWM

Question

A 36-year-old female presents with a 12-year history of palpitations and dyspnea on exertion. In recent years, she has experienced frequent angina and occasional syncope. Physical examination reveals cardiac enlargement toward the left and downward. A systolic thrill is palpable at the right sternal border, and a grade III ejection systolic murmur is auscultated. Blood pressure is 135/60 mmHg, and lung examination is unremarkable.

What is the most likely diagnosis?

- A. Congenital heart disease with aortic valve stenosis and relative pulmonary valve insufficiency
- B. Coronary artery disease with papillary muscle dysfunction
- C. Syphilitic heart disease with aortic valve insufficiency
- D. Rheumatic heart disease with aortic valve insufficiency (风湿性心脏病, 主动脉瓣关闭不全)
- E. Idiopathic aortic stenosis (特发性主动脉狭窄)

Retrieved Context

Neurological and movement disorders (ICD-11):

- Myokymia of eyelid (ICD-11: 9A05.0) Chinese name: 眼睑痉挛
- Other specified tic disorders (ICD-11: 8A05.Y) Chinese name: 其他特指的抽动障碍
- Tic disorders, unspecified (ICD-11: 8A05.Z) Chinese name: 未特指的抽动障碍
- Myoclonic disorders, unspecified (ICD-11: 8A06.Z) Chinese name: 未特指的肌阵挛疾患

Substance withdrawal[related disorders (ICD-11):

- Alcohol withdrawal with seizures (ICD-11: 6C40.42) Chinese name: 酒精戒断伴抽搐
- Alcohol withdrawal with perceptual disturbances and seizures (ICD-11: 6C40.43) Chinese name: 酒精戒断伴知觉异常和抽搐
- Multiple specified psychoactive substances withdrawal, with perceptual disturbances and seizures (ICD-11: 6C4F.43) Chinese name: 多种特指的精神活性物质戒断, 伴知觉异常和抽搐
- Opioid withdrawal (ICD-11: 6C43.4) Chinese name: 阿片类药物戒断
- Opioid-induced delirium (ICD-11: 6C43.5) Chinese name: 阿片类药物所致谵妄
- Sedative, hypnotic or anxiolytic withdrawal, with perceptual disturbances and seizures (ICD-11: 6C44.43) Chinese name: 镇静、催眠或抗焦虑药物戒断, 伴知觉异常及抽搐

Electrolyte and metabolic disorders (ICD-11):

- Hypo-osmolality or hyponatraemia (ICD-11: 5C72) Chinese name: 低渗或低钠血症

Cardiovascular disorders (ICD-11):

- Catecholaminergic polymorphic ventricular tachycardia (ICD-11: BC65.5) Chinese name: 儿茶酚胺源性多形性室性心动过速
- Ventricular tachycardia and fibrillation cardiac arrest (ICD-11: MC82.0) Chinese name: 室性心动过速和心室纤颤性心脏停搏
- Pulmonary hypertension with multifactorial mechanism (ICD-11: BB01.4) Chinese name: 多因素所致肺动脉高压
- Pulmonary hypertension due to lung disease or hypoxia (ICD-11: BB01.2) Chinese name: 肺部疾病或低氧所致肺动脉高压
- Common arterial trunk with pulmonary dominance and interrupted aortic arch (ICD-11: LA85.41) Chinese name: 肺动脉主导的共同动脉干及主动脉弓离断

Answer

Answer: E

The patient's long-standing symptoms of exertional palpitations, dyspnea, angina, and occasional syncope, combined with physical findings of leftward and downward cardiac enlargement, a systolic thrill at the right sternal border, and a grade III ejection systolic murmur, are most consistent with idiopathic aortic stenosis (特发性主动脉狭窄).

Aortic stenosis leads to increased left ventricular afterload, resulting in compensatory hypertrophy and enlargement, which explains the cardiac displacement on percussion. The characteristic systolic thrill and ejection systolic murmur further support this diagnosis.

The other options do not adequately account for the combination of angina, syncope, and classic auscultatory findings observed in this patient.

Figure 9: Diagnosis of Idiopathic Aortic Stenosis Guided by Classic Clinical Findings (Mlec-TCWM)

Mlec-TCM

Question

A 15-year-old male presents with the sudden onset of epigastric pain since yesterday. He reports aversion to cold and preference for warmth, poor appetite, and frequent belching. He appears cold in form but warm to the touch. Tongue coating is white, and the pulse is wiry and tight.

Which treatment should be selected?

- A. 良附丸
- B. 香苏散
- C. 保和丸
- D. 藿香正气散
- E. 良附丸合香苏散

Retrieved Context

Hematologic and metabolic disorders (ICD-11):

- Iron overload diseases (ICD-11: 5C64.10) Chinese name: 铁过多疾病
- Other specified vitamin B12 deficiency anaemia due to intrinsic factor deficiency (ICD-11: 3A01.3Y) Chinese name: 其他特指的因因子缺乏引起的维生素B12缺乏性贫血
- Pantothenic acid deficiency (ICD-11: 5B5H) Chinese name: 泛酸缺乏

Traditional Chinese medicine formulas:

- 胃脘舒片(Weiwanshu Pian) Route of administration: Oral
- 丁桂温胃散(Dinggui Wenwei San) Route of administration: Oral
- 阴虚胃痛片(Yinxu Weitong Pian) Route of administration: Oral
- 虚寒胃痛胶囊(Xuhan Weitong Jiaonang) Route of administration: Oral
- 阴虚胃痛颗粒(Yinxu Weitong Keli) Route of administration: Oral
- 虚寒胃痛颗粒(Xuhan Weitong Keli) Route of administration: Oral
- 清胃止痛微丸(Qingwei Zhitong Weiwan) Route of administration: Oral
- 香砂胃痛散(Xiangsha Weitong San) Route of administration: Oral
- 天凤胃痛散(Tianfeng Weitong San) Route of administration: Oral
- 香苏正胃丸(Xiangsu Zhengwei Wan) Route of administration: Oral

Traditional Chinese medicine terms and patterns:

- 滋阴益胃English: Nourish yin and benefit the stomach Category: Treatment method
- 吐酸English: Acid regurgitation Category: Traditional medicine disorder
- 胃痛English: Stomach ache Category: Traditional medicine disorder
- 嘈杂English: Gastric upset Category: Traditional medicine disorder
- 肝胃不和证English: Disharmony between the liver and stomach pattern Category: Traditional medicine pattern

Answer

Answer: E

Based on the clinical presentation of epigastric pain, aversion to cold with preference for warmth, poor appetite, frequent belching, a white tongue coating, and a wiry, tight pulse, the pattern corresponds to cold accumulation affecting the stomach, with both interior cold and exterior involvement.

良附丸functions to warm the middle and disperse cold, relieving stomach pain, while 香苏散releases the exterior and regulates qi. The combined use of 良附丸合香苏散addresses both the interior cold and the exterior pathogen, making it the most appropriate treatment choice.

Therefore, option E is correct.

Figure 10: Pattern-Based Formula Selection for Cold Constraining the Stomach with Qi Stagnation: Liangfu Wan + Xiangsu San (Mlec-TCM)

Question

An 18-year-old female presents with a 2-day history of high fever and nausea. Physical examination reveals jaundice of the skin and sclera and tenderness on percussion over the liver area. Laboratory tests show GPT 1000 U, TTT 17 U, and total bilirubin 7 mg/dL.

Liver biopsy demonstrates hepatocellular swelling, ballooning degeneration, spotty necrosis, and the formation of eosinophilic bodies.

What is the most appropriate diagnosis?

- A. Acute uncomplicated hepatitis (急性普通型肝炎)
- B. Acute severe hepatitis (急性重型肝炎)
- C. Subacute severe hepatitis (亚急性重型肝炎)
- D. Chronic active hepatitis (慢性活动性肝炎)
- E. Chronic persistent hepatitis (慢性迁延性肝炎)

Retrieved Context

Hematologic, hepatic, and infectious diseases (ICD-11): · Acquired haemolytic anaemia, unspecified (ICD-11: 3A2Z) Chinese name: 未特指的获得性溶血性贫血

- Crigler{Najjar syndrome (ICD-11: 5C58.00) Chinese name: 克里格勒-纳贾综合征
- Dubin{Johnson syndrome (ICD-11: 5C58.02) Chinese name: Dubin-Johnson 综合征
- Gilbert syndrome (ICD-11: 5C58.01) Chinese name: Gilbert 综合征
- Leptospirosis (ICD-11: 1B91) Chinese name: 钩端螺旋体病
- Chronic viral hepatitis, unspecified (ICD-11: 1E51.Z) Chinese name: 慢性病毒性肝炎 (未特指)
- Viral hepatitis, unspecified (ICD-11: 1E5Z) Chinese name: 病毒性肝炎 (未特指)
- Chronic hepatitis B, unspecified (ICD-11: 1E51.0Z) Chinese name: 慢性乙型肝炎 (未特指)
- Acute viral hepatitis, unspecified (ICD-11: 1E50.Z) Chinese name: 急性病毒性肝炎 (未特指)
- Acute hepatitis B (ICD-11: 1E50.1) Chinese name: 急性乙型肝炎
- Enteritis due to Norovirus (ICD-11: 1A23) Chinese name: 诺如病毒肠炎
- Other specified hereditary haemolytic anaemia (ICD-11: 3A10.Y) Chinese name: 其他特指的遗传性溶血性贫血

Traditional Chinese medicine formulas:

- 羊胆胶囊(Yangdan Jiaonang) Route of administration: Oral
- 复方熊胆滴眼液(Fufang Xiongdan Diyanye) Route of administration: Topical
- 复方熊胆清肝颗粒(Fufang Xiongdan Qinggan Keli) Route of administration: Oral
- 花柳清肝颗粒(Huazhi Qinggan Keli) Route of administration: Oral
- 加味左金丸(Jiawei Zuojin Wan, Shuiwan) Route of administration: Oral
- 丹栀逍遥丸(Danzhi Xiaoyao Wan) Route of administration: Oral
- 复方茵陈糖浆(Fufang Yinchén Tangjiang) Route of administration: Oral
- 碧云砂乙肝颗粒(Biyunsha Yigan Keli) Route of administration: Oral

Answer

Answer: A

Based on the acute onset of symptoms, including high fever, nausea, jaundice, marked elevation of GPT, increased bilirubin levels, and characteristic histopathological findings such as hepatocellular ballooning, spotty necrosis, and eosinophilic body formation, the diagnosis is most consistent with acute uncomplicated hepatitis (急性普通型肝炎).

Acute severe or subacute severe hepatitis typically presents with a more progressive and severe clinical course, while chronic active and chronic persistent hepatitis are characterized by long-standing disease rather than an acute presentation. Therefore, option A is the most appropriate diagnosis.

Figure 11: Differential Diagnosis of Acute Uncomplicated Hepatitis Using Time Course and Histopathology (Mlec-Clinic)

Component	Hyperparameter	Value
RGAT encoder (training)	Learning rate	1×10^{-3}
	Weight decay	1×10^{-5}
	Epochs	100
	RGAT layers (L)	2
	Attention heads (h)	2
	Hidden dimension (d)	64
	Ontology loss weight (λ)	0.1
Retriever (inference)	Pseudo-query loss weight (μ)	0.05
	seed_top_k	10
	hops (H)	3
	expansion_top_k	15
	context_top_k	20

Table 12: Hyperparameters used in XONTO-RAG. The RGAT encoder is trained on the cross-ontology (dual-ontology) graph, and the retriever operates on the same graph at inference time.

C Hyperparameters

This section reports the hyperparameters used by XONTO-RAG. We group them into: (i) the **RGAT encoder** (unique to XONTO-RAG) trained on the dual-ontology graph, and (ii) the **inference-time retriever** hyperparameters that control dynamic query injection and multi-hop evidence expansion on the same graph.

RGAT encoder (training). The RGAT encoder is a XONTO-RAG-specific module that learns structure-aware entity representations over the dual-ontology graph with typed intra-ontology and bridging relations. We train the encoder for 100 epochs with a learning rate of 1×10^{-3} and a weight decay of 1×10^{-5} . We use $L=2$ RGAT layers with $h=2$ attention heads and hidden dimension $d=64$. We set $\lambda=0.1$ for the ontology preservation loss $\mathcal{L}_{\text{onto}}$ and $\mu=0.05$ for the pseudo-query objective $\mathcal{L}_{\text{query}}$ (Section 3.3).

Retriever (inference). At inference time, XONTO-RAG injects a temporary query node and connects it to the top seed_top_k seed entities retrieved by initial semantic matching. We then perform query-conditioned multi-hop expansion on the query-augmented graph: hops (H) is the maximum expansion depth; expansion_top_k controls how many frontier nodes are selected at each hop according to the query-conditioned relevance scores; and context_top_k is the final number of entities retained in the evidence subgraph that is linearized and provided to the generator as context.

Usage Across Benchmarks. Unless otherwise specified, we use a single hyperparameter configuration across all benchmarks to avoid per-dataset

tuning and to isolate the effect of the proposed retrieval mechanism.

C.1 Prompt Templates

Multiple-Choice QA Prompt

Graph-retrieved references:

{context}

Question: {question}
Options: {options}

Please output:

1. The letter of the best option.
2. A brief explanation of the reasoning.

Open-Ended QA Prompt

Graph-retrieved references:

{context}

Question: {question}

Please provide a direct, result-oriented answer. You may cite links between TCM health concepts and WM mechanisms when helpful.

Figure 12: Prompt templates used for the question-answering generator.