FCTG: Fuzzy Cognitive Topology Graph for Uncertainty-Aware Medical Diagnosis with LLMs

Anonymous authors
Paper under double-blind review

Abstract

Knowledge-graph-augmented large language models (KG-LLMs) are promising for medical diagnosis, but most medical KGs still collapse symptoms and symptom–disease relations into binary links or single-point probabilities, losing the linguistic vagueness and statistical uncertainty inherent in clinical reasoning. We propose the **Fuzzy Cognitive Topology Graph (FCTG)**, a KG-RAG framework that represents both patient-side symptom memberships and graph-side symptom–disease relations as trapezoidal fuzzy numbers, composes them through fuzzy arithmetic, and organizes evidence into confidence nested (α-cut) subgraphs. Patient-side fuzzy numbers encode the clarity and severity of symptom descriptions, whereas graph-side fuzzy numbers encode the statistically estimated strength of symptom–disease evidence. FCTG then uses fuzzy-number multiplication and aggregation to rank candidate diseases and assemble structured context for final LLM diagnosis without modifying the LLM itself. On the DDXPlus triage dataset, FCTG achieves 88.0%/98.0% Top-1/Top-3 accuracy, exceeding the strongest scalar KG baseline for each metric by 22.0 and 10.0 percentage points, respectively. On the Gretelai free-text diagnosis dataset, FCTG reaches 74.0% Top-1 accuracy, ahead of the strongest baseline by 4 points. Extensive experiments validate the effectiveness of FCTG in medical diagnosis scenarios.

1 Introduction

Diagnostic error is among the leading causes of preventable harm in healthcare, with cognitive biases and ambiguous manifestations being the dominant contributing factors (Institute of Medicine, 2000). Recent LLMs offer new opportunities for clinical decision support, but their opaque, unauditable parametric knowledge makes them prone to hallucination. Knowledge-graph-augmented retrieval (KG-RAG) addresses this by grounding LLM reasoning in structured, verifiable medical knowledge (Lewis et al., 2020; Pan et al., 2024). In a typical KG-RAG pipeline, the model retrieves symptom–disease subgraphs and uses them as additional context for diagnosis.

However, existing medical KGs overwhelmingly represent symptom–disease relations using deterministic edge weights—binary indicators, raw co-occurrence counts, or, at best, single-point probabilities such as $P(\text{disease} \mid \text{symptom})$. This representation is insufficient for the nature of clinical reasoning. In real-world diagnostic settings, patients often describe clinical details ambiguously—onset, quality, severity, radiation, and precipitating or relieving factors—(Ouyang, 2018), which makes diagnosis substantially more difficult. A symptom such as *"sometimes a little bit lower-back pain"* thus provides graded, context-dependent evidence for *lumbar spondylosis* rather than uniform support, evidence that maps naturally to linguistic descriptors such as *weak*, *moderate*, or *strong*. Collapsing this fuzziness into a single scalar discards information crucial for distinguishing diseases with overlapping presentations: traditional knowledge graphs use triples with predicates, and while some extensions add probabilities to form quadruples, they still fail to capture uncertainty.

Therefore, we propose the **Fuzzy Cognitive Topology Graph (FCTG)**, which addresses this gap along three complementary axes (Figure 1). First, every symptom–disease edge is a trapezoidal fuzzy number $\tilde{w}_{sd}$ whose plateau represents the most likely range of the relation strength and whose shoulders encode the uncertainty tails. Both its center m_{sd} and its width r_{sd} are derived from the positive likelihood ratio $\mathrm{LR}^{+}(s,d) = P(s \mid d)/P(s \mid \neg d)$, a standard clinical statistic (Guyatt et al., 2015), so that edges sharing a point estimate but differing in sample size remain distinguishable.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

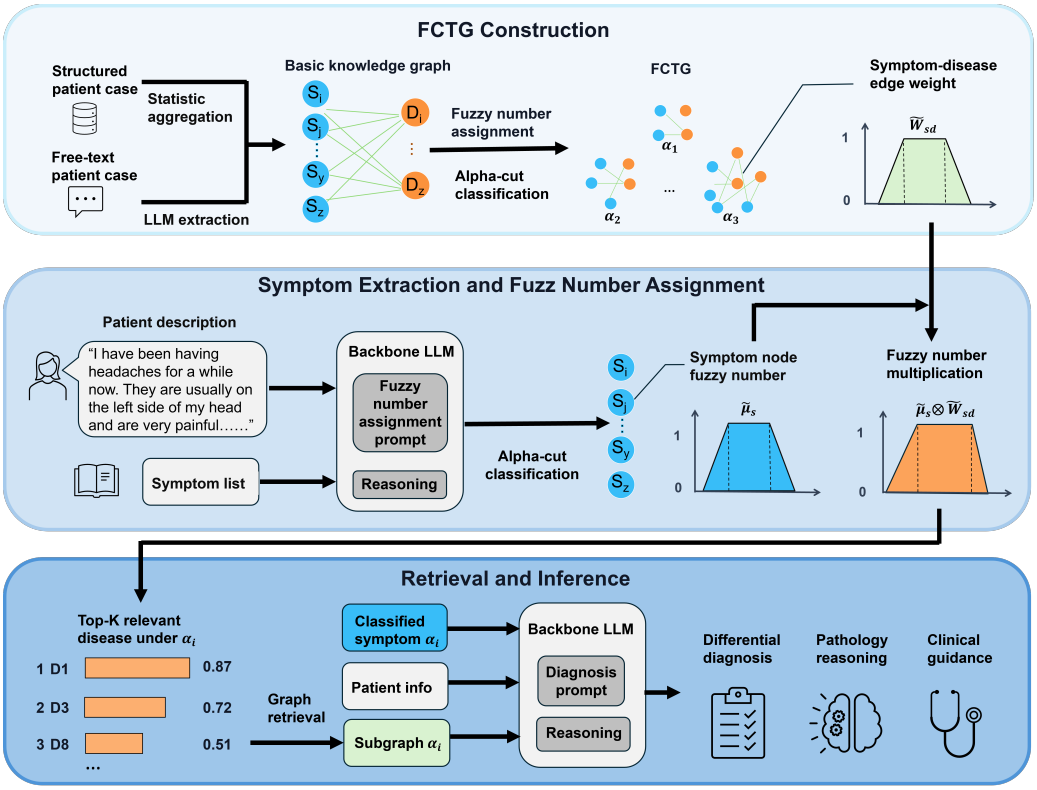

Figure 1: Overview of the FCTG framework.

Second, an LLM assigns each matched symptom a trapezoidal membership $\tilde{\mu}_i$ reflecting how clearly it is described—from *"some chest discomfort"* to *"crushing substernal chest pain radiating to the left arm"*—and fuzzy multiplication composes the two sources of uncertainty, so that a vague symptom contributes less to a disease score than a definitive one even when both share the same edge. Third, the fuzzy centroid $\mathcal{C}(\tilde{w}_{sd}) \in [0, 1]$ supports an α-cut that induces a nested family of subgraphs: a high α yields a sparse subgraph for diagnosis where false positives are costly, a low α the full graph for screening where recall matters more than precision.

We instantiate FCTG on the structured-evidence DDXPlus dataset (Fansi Tchango et al., 2022) and the free-text Gretelai symptom-to-diagnosis dataset (Gretel.ai, 2023), and integrate it with Qwen3.8-Flash (Yang et al., 2025) through a structured reasoning prompt in which both symptom matching and disease ranking are driven by the centroids of trapezoidal fuzzy numbers.

Contributions.

1. We introduce FCTG, a fuzzy, multilayer extension of cognitive graphs for medical KG-LLM reasoning, and formalize its construction and inference.
2. We empirically show that FCTG improves diagnostic accuracy over five KG-based baselines—MedRAG, MedGraphRAG, and three probabilistic KG variants—on a challenging DDXPlus subset with ambiguous symptom descriptions. With the same LLM extractor and diagnostic backbone, FCTG reaches 88%, 22% higher than the strongest baseline.
3. We show that α-cut filtering identifies a compact non-diagnostic regime: increasing α over the low range removes low-confidence evidence without reducing accuracy, whereas further filtering reduces support-edge coverage and diagnostic accuracy in parallel. Across two datasets, additional results confirm that the advantage persists under a different backbone LLM and that the residual errors are dominated by missing graph edges rather than by ranking mistakes.

2 Related work

Knowledge-graph-enhanced LLMs predominantly adopt a retrieve-then-augment paradigm, extracting subgraphs from biomedical knowledge bases such as UMLS (Bodenreider, 2004), DrugBank (Wishart et al., 2008), or PrimeKG (Chandak et al., 2023) and providing them as prompt context (Soman et al., 2024), with extensions covering graph partitioning (Wu et al., 2024), multihop query decomposition (Jiang et al., 2025), corrective knowledge injection (Gao et al., 2025), and multiagent reasoning over a shared KG (Tang et al., 2024). These systems, however, almost uniformly treat KGs as deterministic structures and therefore fail to represent the graded, context-dependent uncertainty that pervades clinical knowledge. A parallel line of work addresses uncertainty representation itself: probabilistic knowledge graphs annotate edges with confidence scores or conditional probabilities (Chen et al., 2020), and Bayesian KG embeddings represent entities as distributions rather than point vectors (Wan & Du, 2021), with biomedical applications ranging from COVID-19 reasoning to drug repurposing (Domingo-Fernández et al., 2021; Gao et al., 2022). Orthogonal to these, Fuzzy Cognitive Maps (FCMs), introduced by Kosko (1986), model graded influence through weighted directed edges in $[-1, +1]$ and have been extensively applied to medical diagnosis (Stylios & Groumpos, 2004; Papageorgiou, 2011; Papageorgiou & Salmeron, 2013; Apostolopoulos et al., 2024). They apply fuzzy *operators* to scalar membership grades, however, collapsing uncertainty at the first aggregation step. FCTG instead uses fuzzy *numbers* (trapezoidal 4-tuples) end to end: both edge weights and patient-side symptom memberships are fuzzy intervals, and fuzzy arithmetic propagates both sources of uncertainty to the final disease score. This approach bridges FCMs—which represent graded topology but cannot process clinical text—and LLMs—which process text but lack structured, uncertainty-aware reasoning.

3 Preliminaries

3.1 Task formulation

Let $\mathcal{S}$ denote the set of DDXPlus symptoms and $\mathcal{D}$ the set of pathologies. A KG-RAG diagnostic system maps a free-text narrative x to a small set of matched symptoms $\mathcal{M}(x) \subseteq \mathcal{S}$, retrieves a disease-centered subgraph, and asks an LLM to return a ranked differential $\hat{\pi}(x) = [(\hat{y}_1, p_1), \ldots, (\hat{y}_K, p_K)]$, where $\hat{y}_1$ is the Top-1 prediction and p_k is the LLM-reported probability. We evaluate exact Top-1 accuracy and Top-3 inclusion. The central design question is how the retrieved graph should represent uncertain symptom–disease evidence before it is summarized for the LLM.

3.2 Trapezoidal fuzzy numbers

Fuzzy sets (Zadeh, 1965) provide the framework. A trapezoidal fuzzy number $\tilde{z} = (z_1, z_2, z_3, z_4)$, with $0 \le z_1 \le z_2 \le z_3 \le z_4 \le 1$, is a soft interval whose core $[z_2, z_3]$ has full membership and whose shoulders encode the uncertainty of the underlying estimate; its membership function is stated in Appendix B. We compare fuzzy numbers through their integral centroid,

$$\mathcal{C}(\tilde{z}) = \begin{cases} z_1, & z_1 = z_2 = z_3 = z_4, \\ \dfrac{z_4^2 + z_3 z_4 + z_3^2 - z_1^2 - z_1 z_2 - z_2^2}{3(z_4 + z_3 - z_1 - z_2)}, & \text{otherwise.} \end{cases} \tag{1}$$

The closed form is the integral centroid $\int t\, \mu_{\tilde{z}}(t)\, dt \,/\, \int \mu_{\tilde{z}}(t)\, dt$, and the degenerate case is defined by continuity: the centroid of a point-mass trapezoid is the point itself, so a definite symptom encoded as $(1, 1, 1, 1)$ has centroid 1. Fuzzy addition and multiplication are implemented component-wise for nonnegative trapezoids, $\tilde{x} \oplus \tilde{z} = (x_1 + z_1, \ldots, x_4 + z_4)$ and $\tilde{x} \otimes \tilde{z} = (x_1 z_1, \ldots, x_4 z_4)$, the monotone interval-arithmetic approximation used in our system (Klir & Yuan, 1995).

3.3 Confidence-layered knowledge graphs

A standard medical KG stores a directed support edge $e = (s, d)$ between symptom $s \in \mathcal{S}$ and pathology $d \in \mathcal{D}$, often with a scalar weight. FCTG separates symptom–pathology support edges $E_{\text{sp}} \subseteq \mathcal{S} \times \mathcal{D}$ from disease–disease differential edges $E_{\text{dd}} \subseteq \mathcal{D} \times \mathcal{D}$: only the support edges receive fuzzy weights, while the differential edges are retained as graph structure for organizing similar pathologies. The full notation, including the fuzzy edge-weight lookup $\tilde{W}$ and the admissible

threshold set, is given in Appendix B. At inference FCTG selects a confidence level α and induces the support subgraph

$$G_\alpha = (V, E_{\text{sp},\alpha}, E_{\text{dd}}),$$
$$E_{\text{sp},\alpha} = \{e \in E_{\text{sp}} : \mathcal{C}(\tilde{W}(e)) \geq \alpha\}. \tag{2}$$

Because $\alpha_1 \leq \alpha_2$ implies $E_{\text{sp},\alpha_2} \subseteq E_{\text{sp},\alpha_1}$, these subgraphs form a nested family: α is a confidence filter, where low values retain broad weak evidence and high values retain only strong symptom–pathology associations.

4 Methodology

FCTG separates three choices that conventional KG-RAG systems conflate: graph-based evidence estimation, patient-side symptom uncertainty, and low-confidence filtering at inference.

4.1 Fuzzy graph construction

We first construct a statistics-based DDXPlus KG from patient records. For each symptom–pathology pair (s, d), the KG stores the co-occurrence count n_{sd}, $P(d \mid s)$, $P(s \mid d)$, the marginal case counts n_d and n_s, and the total record count N. The G1, G2, and G3 baselines use these aggregate statistics as crisp values; FCTG converts them into a fuzzy edge whose center and width are both derived from the positive likelihood ratio $\text{LR}^+(s, d) = P(s \mid d)/P(s \mid \neg d)$. This standard statistic in evidence-based medicine quantifies diagnostic performance by measuring how strongly symptom s discriminates disease d from its complement (Guyatt et al., 2015).

For an edge (s, d), we define a scalar center

$$m_{sd} = \text{clamp}\left(\frac{\log \text{LR}^+(s, d)}{\log 10},\ 0,\ 1\right), \tag{3}$$

where $\text{clamp}(u, a, b) = \min\{b, \max\{a, u\}\}$ maps every finite $\text{LR}^+ > 0$ into $m_{sd} \in [0, 1]$. Thus, $\text{LR}^+ \leq 1$ yields a zero center, while $\text{LR}^+ \geq 10$, a commonly used threshold for strong diagnostic evidence (McGee, 2002), saturates to 1. The width is the 95% confidence half-interval of $\log \text{LR}^+$ mapped to score space, derived via the delta method:

$$r_{sd} = \text{clamp}\Big(\tfrac{1.96}{\log 10}\,\sigma_{sd},\ 0.05,\ 0.4\Big),$$
$$\sigma_{sd} = \sqrt{\tfrac{1-P(s|d)}{n_d\,P(s|d)} + \tfrac{1-P(s|\neg d)}{n_{\neg d}\,P(s|\neg d)}}. \tag{4}$$

Both statistics presuppose strictly positive conditionals. If s is never observed outside d, then $P(s \mid \neg d) = 0$ and both LR^+ and σ_{sd} diverge. This is not a corner case: edges supported by a single case—for which zero counter-evidence is common—are frequent in sparsely elicited graphs (89 of the 330 Gretelai edges occur only once; Appendix A). We therefore apply the Haldane–Anscombe correction before estimation, replacing every empirical conditional probability c/n in Eqs. (3) and (4) with $(c + 0.5)/(n + 1)$. This keeps every LR^+ finite and every standard error well defined, while the inflated uncertainty it assigns to low-support edges is absorbed by the width clamp in Eq. (4).

This couples the fuzzy width directly to the statistical precision of the underlying diagnostic evidence: edges backed by many samples are tight, while edges backed by few samples are wide regardless of their center value. The trapezoidal edge $\tilde{w}_{sd} = (w_{sd,1}, \ldots, w_{sd,4})$ spreads m_{sd} over r_{sd}, with a narrow full-membership core of half-width $0.3r_{sd}$ and shoulders reaching $m_{sd} \pm r_{sd}$ clipped to $[0, 1]$ (Appendix B). Because both m_{sd} and r_{sd} are derived from LR^+ and its standard error, the construction uses only a small number of interpretable design constants: $\log 10$ and 1.96 come from the LR^+ normalization and the 95% confidence interval, while the 0.05 floor and 0.4 ceiling stabilize extremely narrow or wide empirical intervals.

4.2 Patient-level fuzzy symptom membership

Given a patient narrative x, the extraction LLM receives the full symptom vocabulary and returns a list of matched symptoms with trapezoidal memberships:

$$\mathcal{M}(x) = \{(s_i, \tilde{\mu}_{x,i})\}_{i=1}^{M}. \tag{5}$$

	DDXPlus	Gretelai S2D
Construction cases	4,000	400
Evaluation cases	200	100
Diseases $\lvert\mathcal{D}\rvert$	49	22
Canonical symptoms $\lvert\mathcal{S}\rvert$	381	256
`supports_pathology` edges	1,042	330
Singleton edges (count $= 1$)	12	89
Mean symptoms per case	19.3	4.8
Raw format	structured	free text

Table 1: Dataset statistics

The membership $\tilde{\mu}_{x,i}$ reflects how clearly the narrative supports symptom s_i. For example, a definite statement receives a near-degenerate membership such as $(1, 1, 1, 1)$, whereas a vague or hedged mention receives a lower and wider membership. We use $\mathcal{C}(\tilde{\mu}_{x,i})$ as the symptom match score in the prompt. At threshold α, FCTG drops a matched symptom when $\mathcal{C}(\tilde{\mu}_{x,i}) < \alpha$, so the same confidence knob filters both patient-level evidence and graph-level evidence.

4.3 Fuzzy disease scoring

For each candidate pathology d, FCTG aggregates all retained symptom matches connected to d in G_α:

$$\tilde{S}_\alpha(d \mid x) = \bigoplus_{s_i \in \mathcal{M}_\alpha(x,d)} \tilde{\mu}_{x,i} \otimes \tilde{W}(s_i, d). \tag{6}$$

Here $\mathcal{M}_\alpha(x, d)$ denotes the retained symptom matches that also have an edge to d in G_α. We sort diseases by the centroid of $\tilde{S}_\alpha(d \mid x)$ and render that value as `rank_score`, a nonnegative ranking score rather than a probability: aggregation can push fuzzy components above 1 (Appendix B). The LLM never sees raw fuzzy tuples. Instead, fuzzy arithmetic changes only the numerical evidence that fills the same prompt fields used by G1–G3, thereby isolating the representation change while keeping the prompt schema fixed. Appendix C summarizes the complete inference procedure.

4.4 Prompt interface

The diagnostic LLM acts as a constrained reranker using a standardized interface: patient descriptions, LLM-matched findings, scoring formulas, and KG-ranked candidate diseases with supporting edges and differential links. G1–G3 use binary symptom encoding with statistical scoring. FCTG employs fuzzy memberships and uncertainty guidance to prevent minor findings from overriding KG priors. MedRAG adapts this by substituting the KG and using triple-count scoring. MedGraphRAG, however, requires its own two-stage protocol: embedding-based retrieval followed by a drafting and revision process to output the required JSON schema. The complete symptom-extraction and diagnosis-reranking templates are given in Appendix F.

5 Experimental setup

5.1 Datasets and knowledge graphs

We evaluate FCTG on two benchmarks differing in scale and input structure, testing the method on both coded-evidence and free-text corpora. Table 1 summarizes the construction and evaluation subsets together with the resulting knowledge-graph statistics.

DDXPlus (Fansi Tchango et al., 2022) is a large synthetic triage corpus with 49 pathologies and 485 symptoms. Each case contains structured evidence tokens (e.g., `E_57_@_1`) whose meanings are defined in a separate evidence dictionary, along with demographic information, the true disease, and a differential diagnosis; an example of this structure is provided in the appendix. We randomly draw 4,000 cases from the total 1,300,000 cases in DDXPlus for KG construction. All methods are

evaluated on held-out patient descriptions generated from a 200-case pool, comprising the 100 mild and 100 confusable evaluation cases.

Gretelai (Gretel.ai, 2023) is a smaller, text-only corpus covering 22 diseases, with 853 training cases and 212 test cases. Each case consists of a free-text patient complaint paired with a gold diagnosis. Unlike DDXPlus, Gretelai provides no evidence codes, predefined symptom vocabulary, or structured findings; the knowledge graph must therefore be elicited from the text by an LLM. For the reported cross-dataset experiment, we build each KG from the same stratified 400-case training subset and evaluate it on a fixed random 100-case subset of the official test split.

5.2 Baselines

We compare five baselines, all of which are knowledge-graph methods and therefore share the same disease-constrained output space; the comparison stays within a single prediction task. Two are medical KG-RAG systems: **MedRAG** (Zhao et al., 2025) retrieves (subject, relation, object) triples from established medical knowledge graphs together with corpus snippets and lets the frozen LLM reason over the retrieved evidence, and **MedGraphRAG** (Wu et al., 2024) first retrieves at the document granularity and then pulls entity triples from the most similar documents as augmented information. The other three are crisp-scalar variants of our own graph, which differ only in the scalar formula that maps matched symptoms to disease scores: **G1 (Count)** (Jia et al., 2025; Gao et al., 2025) counts the matched symptoms and includes no edge statistics in the prompt, **G2 (Posterior Sum)** (Guyatt et al., 2015; Rotmensch et al., 2017) weights each edge by $P(\text{disease} \mid \text{symptom})$, and **G3 ($\text{LR}^+$ Sum)** (Albagli-Kim & Beimel, 2026) weights it by $\min(\log \text{LR}^+ / \log 10, 1)$; all three collapse uncertainty at the first computation step. **FCTG** instead carries trapezoidal edge weights, LLM-assigned patient memberships, and centroid defuzzification through the same prompt schema. Full definitions of all five methods and of their evidence-scoring rules are given in Appendix G.

5.3 Metrics and protocol

We report Top-1 exact-match accuracy, Top-3 inclusion accuracy, macro-averaged F1, expected calibration error (ECE), and end-to-end LLM token usage per case. Disease names are normalized before comparison. ECE is computed over 10 confidence bins and reported as a percentage; lower values therefore indicate better calibration. The main-table ECE values report a 5-fold cross-validated isotonic recalibration of each method's top-1 confidence, which places every method on a common calibrated scale. The main DDXPlus comparison uses Qwen3.8-Flash with a temperature of 0.1 and $\alpha = 0.35$ on the same fixed, challenging confusable subset for every method. Token values marked $\sim$ include both LLM calls—symptom extraction and final diagnosis—and are full-set averages rounded to the nearest hundred. All hyperparameters are listed in Table 9 (appendix).

6 Results

We organize the analysis around eight questions spanning comparative accuracy (Q1), generalization to free text (Q2), noise robustness (Q3), component ablation (Q4), α-cut filtering (Q5), the symptom matcher (Q6), the backbone LLM (Q7), and cost and failure modes (Q8). Detailed efficiency figures and the full error taxonomy are deferred to Appendices J and K.

6.1 Overall performance and statistic comparison (Q1)

Table 2 compares the five baselines with FCTG on the confusable DDXPlus subset under identical controls (same test cases, same backbone, same token accounting). In this matched setting, FCTG reaches 88% Top-1 and 98% Top-3 accuracy, outperforming the strongest scalar KG baseline by 22 percentage points in Top-1 accuracy.

Table 2 yields three observations. (1) MedRAG reaches 62.0% Top-1 while MedGraphRAG trails at 52.0%, suggesting that extracting all entity–relation pairs is not necessary and can even harm disease diagnosis. (2) G2 and G3 (62–66% Top-1) incorporate scalar association strengths and outperform G1, showing that appropriate statistical metrics help. (3) FCTG assigns trapezoidal memberships to symptoms and reaches **88.0%** Top-1 and 98.0% Top-3, whereas the five baselines treat all extracted symptoms equally—illustrating that fuzzy symptom memberships matter most under ambiguity.

Method	Acc	Top-3	Macro-F1	ECE	Tokens
MedRAG	62.0	80.0	32.5	12.7	∼6,157
MedGraphRAG	52.0	66.0	25.0	20.7	∼5,621
G1 Probabilistic KG (Count)	52.0	88.0	24.2	16.5	∼6,306
G2 Probabilistic KG (Posterior Sum)	62.0	88.0	31.6	11.9	∼5,177
G3 Probabilistic KG (LR^+ Sum)	66.0	86.0	35.0	14.6	∼5,898
FCTG (Ours)	**88.0**	**98.0**	**51.4**	**11.1**	∼8,100

Table 2: Diagnostic results on the DDXPlus confusable evaluation subset (100 cases, 49 diseases)

Method	Acc	Top-3	M-F1	ECE	Tokens
MedRAG	70.0	79.0	66.1	**12.5**	∼3,400
MedGraphRAG	54.0	76.0	48.9	16.3	∼3,400
G1 (Count)	70.0	82.0	**70.8**	15.2	∼3,700
G2 (Posterior)	68.0	80.0	68.7	20.3	∼3,800
G3 (LR^+ Sum)	66.0	80.0	65.8	16.5	∼3,300
FCTG (Ours)	**74.0**	**82.0**	70.1	15.8	∼4,000

Table 3: Diagnostic results on the Gretelai evaluation subset (100 held-out cases, 22 diseases)

6.2 Cross-dataset generalization (Q2)

We evaluated FCTG on the Gretelai free-text corpus as a secondary stress test (Table 3, fixed 100-case subset). FCTG is the strongest method, at 74.0% Top-1 and 82.0% Top-3, ahead of MedRAG and the best crisp-scalar control G1 (both 70.0% Top-1), while MedGraphRAG performs poorly (54.0%). The margin over MedRAG is 4 percentage points (74.0% versus 70.0%), however, and G1 matches FCTG's Top-3, so we read the free-text result as a narrow rather than a decisive advantage: the Gretelai graph is sparse and its complaints are already deterministic, leaving little uncertainty for fuzzy memberships to resolve (Appendix A).

6.3 Robustness to description noise (Q3)

Patient narratives vary in how noisy and confusable they are, so we evaluate the same methods on two further matched subsets of the same pool: a *mild*-noise subset (100 cases) with lightly perturbed descriptions, and the *confusable* subset (100 cases) used in Table 2.

Table 4 shows that the FCTG margin *widens* as descriptions become harder: 10.0 points over the strongest baseline (G2) under mild noise and 22.0 over G3 under confusable descriptions. The scalar baselines degrade fastest because a weakly supported or incidental symptom enters their scores at full weight, whereas trapezoidal membership attenuates it according to the LLM-assigned membership. MedRAG and MedGraphRAG follow the same pattern: competitive under mild noise (81.0 and 74.0) but falling to 62.0 and 52.0 once descriptions turn confusable—drops of 19.0 and 22.0—while FCTG loses only 6.0 (94.0→88.0), the smallest drop of any method.

6.4 Component ablation (Q4)

Table 5 removes one mechanism at a time from full FCTG, testing whether the gain comes from the trapezoidal representation or from incidental choices such as the α-cut or the diagnostic reranker.

Table 5 reports the ablation. On DDXPlus, removing the trapezoidal memberships costs 6.0 points (82.0), the α-cut 4.0 (84.0), and the diagnostic reranker 8.0 (80.0), while replacing the trapezoidal edge widths by their centroids costs only 2.0 (86.0). Gretelai repeats the ordering at a smaller scale (widths -3.0, α-cut -3.0, memberships -4.0, reranker -8.0), with the reranker again the largest single effect. Each component plays a role in the outcome, though the fuzzy number representing patient symptoms is a particularly significant factor.

	MedRAG	MedGraphRAG	G1 (Count)	G2 (Posterior)	G3 (LR^+ Sum)	FCTG (Ours)
Mild (100)	81.0	74.0	65.0	84.0	73.0	**94.0**
Confusable (100)	62.0	52.0	52.0	62.0	66.0	**88.0**
FCTG − best baseline	$+10.0$ (mild, vs. G2)		$+22.0$ (confusable, vs. G3)			

Table 4: Top-1 accuracy under two description-noise regimes on DDXPlus (mild: 100 cases; confusable: 100 cases; Qwen3.8-Flash)

Configuration	DDXPlus (confusable, 100 cases)		Gretelai (100 cases)	
	Top-1	Δ	Top-1	Δ
Full FCTG	**88.0**	—	**74.0**	—
− trapezoidal widths (scalar edge or product)	86.0	-2.0	71.0	-3.0
− trapezoidal membership (scalar confidence)	82.0	-6.0	70.0	-4.0
− α-cut (aggregate all edges)	84.0	-4.0	71.0	-3.0
− LLM reranker (graph ranking only)	80.0	-8.0	66.0	-8.0

Table 5: Component ablation (Top-1 accuracy)

The reranker is the largest single component in the ablation: ranking candidates by the graph alone, with no diagnostic-LLM reranking, costs FCTG 8.0 points on both datasets (Table 5), leaving the fuzzy graph itself responsible for most of the final decision.

6.5 Sensitivity to confidence thresholding (Q5)

The threshold α controls how many fuzzy support edges and low-membership patient findings remain in G_α. Figure 2 sweeps it on both datasets over a uniform α grid from 0 to 1, with one frozen Qwen3.8-Flash extraction per dataset; panel (a) is the confusable DDXPlus subset with its full 1,042-edge graph, reported for the full pipeline as in Table 2, and panel (b) the Gretelai subset with the same 330-edge graph as the main results, under the graph-only ranking protocol shared with Table 5.

On DDXPlus, accuracy stays within an 84–88% band for $\alpha \leq 0.6$, bracketing the reported 88.0 of Table 2, while the retained-edge fraction falls monotonically from 100% to 73%: the edges removed in this range are non-diagnostic, and the operating point ($\alpha = 0.35$) is not a fortunate choice. Beyond $\alpha = 0.6$ both curves collapse, as the cut starts removing evidence-supported reasoning chains rather than noise. On Gretelai, 324–330 of the 330 support edges remain from $\alpha = 0$ to 0.2 and accuracy is flat at 66%/82% Top-1/Top-3; beyond $\alpha = 0.2$ coverage and accuracy decline in parallel (54%/62% at $\alpha = 0.5$, 26%/28% at 0.7), and at $\alpha = 1.0$ both fall to 4%. The graph-only protocol ties panel (b) to the ablation, whose −LLM-reranker row is exactly this flat region (66.0): re-adding the reranker raises it by a constant ≈8 points without moving the collapse point. On neither dataset is α a broadly tunable robustness knob; it is an interpretable filter that first removes non-diagnostic evidence and then prunes the evidence-supported chains needed for differential diagnosis.

6.6 Sensitivity to the symptom matcher (Q6)

The sentence-to-symptom matching step feeds every downstream method, so we substitute the matcher while holding the graph, scoring, and diagnostic stage fixed on the confusable 100-case DDXPlus subset. Accuracy tracks matcher quality closely, from 88.0/98.0 with the LLM semantic matcher to 79.0 with BM25 and 52–56% with the lexical and embedding-only variants (Table 10 in Appendix I).

6.7 Backbone compatibility (Q7)

We also run a DDXPlus compatibility check with Qwen2.5-72B-Instruct (Yang et al., 2024), to test whether the fuzzy-graph advantage depends on a single diagnostic backbone. All methods use the same fixed confusable 100-case subset and the same 4,000-case DDXPlus training graph.

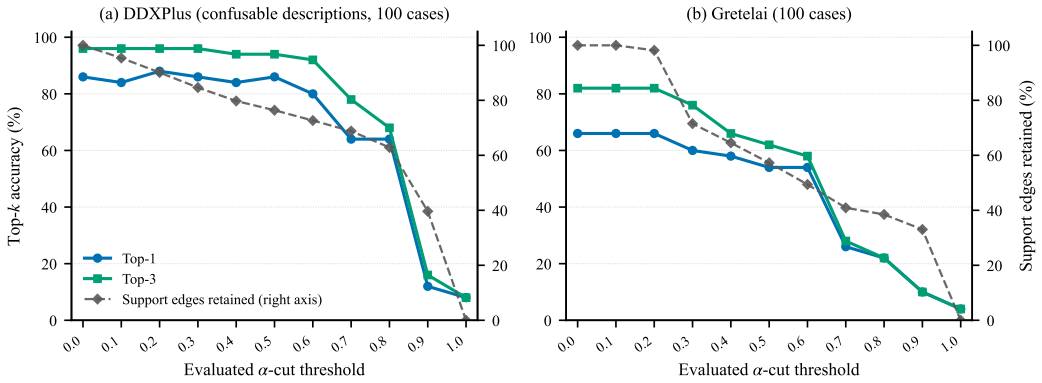

Figure 2: Sensitivity of FCTG to the confidence level α. (a) DDXPlus confusable 100-case subset, full 1,042-edge graph, reported for the full pipeline as in Table 2; (b) Gretelai 100-case subset, 330-edge graph, under the graph-only ranking shared with Table 5.

	MedRAG	MedGraphRAG	G1 (Count)	G2 (Posterior)	G3 (LR^+ Sum)	FCTG (Ours)
Top-1 Acc	58.0	47.5	48.0	59.0	62.0	**80.0**
Top-3 Acc	74.0	64.6	81.0	84.0	82.0	**91.0**

Table 6: DDXPlus backbone-compatibility results with Qwen2.5-72B-Instruct (confusable 100-case subset)

Table 6 shows that FCTG remains the strongest method with this backbone (80.0% Top-1, 91.0% Top-3), leading the best baseline by 18 and 7 points respectively. Therefore, the advantage over the other baselines is not an artifact of the Qwen3.8-Flash diagnostic LLM.

6.8 Cost and failure modes (Q8)

We analyzed the time costs and found that LLM calls dominate the runtime. Graph scoring requires only 0.17–0.21 ms per case—four to five orders of magnitude faster than the ≈ 9 s of per-case LLM latency—while the one-time construction pass takes less than a second. Consequently, the α-cut does not improve end-to-end pipeline speed; instead, it reduces the arithmetic load by half, from 109.1 to 51.7 aggregated (finding, edge) products per case. This optimization is only significant when the graph step is part of a larger offline sweep, or when it is paired with a less expensive local scorer in place of the LLM (Appendix J).

We found that the largest residual-error bucket on Gretelai is an absent edge (7 of 13 cases), which is consistent with 54.9% of that graph's finding nodes being isolated, against 21.4% on DDXPlus; of the remaining errors, at most $6/100$ (DDXPlus) and $8/100$ (Gretelai) would be reachable by a better reranker (Appendix K).

7 Conclusion

FCTG recasts medical KG-RAG diagnosis as an uncertainty-preserving scoring problem. Symptom states and symptom–disease relations are both carried as trapezoidal fuzzy numbers and composed by fuzzy multiplication and aggregation into a disease score; a single α-cut then decides how much of that evidence survives on either side. On the DDXPlus confusable subset, FCTG reaches 88.0%/98.0% Top-1/Top-3 against 66.0%/88.0% for the strongest crisp-scalar variant using the same extractor, and 74.0% against 70.0% on Gretelai free text. Ablations place the largest representational effect on the fuzzy patient memberships and removing the diagnostic reranker is the largest single ablation, leaving the fuzzy graph itself responsible for most of the final decision.

Future research should extend FCTG to realistic clinical settings by evaluating it on authentic, ambiguous clinical notes and incorporating multimodal evidence such as laboratory, imaging, and lon-

gitudinal data. Furthermore, future work should involve calibrating α thresholds for specific clinical tasks and integrating FCTG into decision-support systems to provide interpretable diagnostic explanations and targeted follow-up queries.

AI USE STATEMENT

Generative AI tools were used to aid and polish the writing, to draft sections of the paper, and to assist with retrieval and discovery of related work. All AI-assisted text was reviewed and verified by the authors. The authors take responsibility for the final text, claims, and artifacts.

REPRODUCIBILITY STATEMENT

Section 3 formalises the trapezoidal fuzzy numbers, α-cut operations, and graph definitions used throughout; Section 4 specifies the fuzzy patient-membership construction, the LR^{+}-based scoring rule, and the inference procedure; Section 5 describes the datasets, evaluation subsets, baselines, metrics, and token accounting shared by all methods. Appendix E details the prompting and reproducibility setup, Appendix F provides the complete prompt templates, Appendix G the baseline constructions, Appendix H the hyperparameters, and Appendix J the construction and per-case efficiency. Both datasets are publicly available and all methods share the same backbone and decoding configuration; the complete prompts, hyperparameters, and subset constructions are documented in the appendices.

REFERENCES

Sivan Albagli-Kim and Dizza Beimel. A semantic knowledge graph approach with weighted and conditional edges for clinical reasoning. In *Advances in Digital Health and Medical Bioengineering 2: Medical Devices, Measurements, and Artificial Intelligence Applications (IFMBE Proceedings, Vol. 142)*, pp. 234–243. Springer, 2026. doi: 10.1007/978-3-032-24724-7_24. URL https://doi.org/10.1007/978-3-032-24724-7_24.

Ioannis D. Apostolopoulos, Nikolaos I. Papandrianos, Nikolaos D. Papathanasiou, and Elpiniki I. Papageorgiou. Fuzzy cognitive map applications in medicine over the last two decades: A review study. *Bioengineering*, 11(2):139, 2024. URL https://doi.org/10.3390/bioengineering11020139.

Olivier Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–D270, 2004. URL https://doi.org/10.1093/nar/gkh061.

Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10:67, 2023. URL https://doi.org/10.1038/s41597-023-01960-3.

Xin Chen, Shuyi Jia, and Yang Xiang. A review: Knowledge reasoning over knowledge graph. *Expert Systems with Applications*, 141, 2020. URL https://doi.org/10.1016/j.eswa.2019.112948.

Daniel Domingo-Fernández et al. COVID-19 knowledge graph. *Bioinformatics*, 37(9), 2021. URL https://doi.org/10.1093/bioinformatics/btaa834.

Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. DDXPlus: A new dataset for automatic medical diagnosis. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022 Datasets and Benchmarks Track)*, pp. 31306–31318, 2022. URL https://arxiv.org/abs/2205.09148.

Yanjun Gao et al. DR.KNOWS: Leveraging medical knowledge graphs into large language models for diagnosis prediction: Design and application study. *JMIR AI*, 4:e58670, 2025. URL https://doi.org/10.2196/58670. arXiv:2308.14321.

Zhenxiang Gao, Pingjian Ding, and Rong Xu. KG-Predict: A knowledge graph computational framework for drug repurposing. *Journal of Biomedical Informatics*, 132:104133, 2022. URL https://doi.org/10.1016/j.jbi.2022.104133.

Gretel.ai. Symptom to diagnosis. Hugging Face dataset, https://huggingface.co/datasets/gretelai/symptom_to_diagnosis, 2023. 1,065 synthetic patient presentations, each labelled with one of 22 diagnoses.

Gordon Guyatt, Drummond Rennie, Maureen O. Meade, and Deborah J. Cook (eds.). *JAMAevidence: Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice*. McGraw-Hill Education, 3 edition, 2015.

Institute of Medicine. *To Err is Human: Building a Safer Health System*. National Academies Press, 2000. URL https://doi.org/10.17226/9728.

Mingyi Jia, Junwen Duan, Yan Song, and Jianxin Wang. medIKAL: Integrating knowledge graphs as assistants of LLMs for enhanced clinical diagnosis on EMRs. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, pp. 9278–9298, 2025. URL https://arxiv.org/abs/2406.14326.

Pengcheng Jiang, Cao Xiao, Minhao Jiang, Parminder Bhatia, et al. KARE: Reasoning-enhanced healthcare predictions with knowledge graph community retrieval. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=8fLgt7PQza. arXiv:2410.04585.

George J. Klir and Bo Yuan. *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Prentice Hall, 1995.

Bart Kosko. Fuzzy cognitive maps. *International Journal of Man-Machine Studies*, 24(1):65–75, 1986. URL https://doi.org/10.1016/S0020-7373(86)80040-2.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 9459–9474, 2020. URL https://arxiv.org/abs/2005.11401.

Steven McGee. Simplifying likelihood ratios. *Journal of General Internal Medicine*, 17(8):647–650, 2002. doi: 10.1046/j.1525-1497.2002.10750.x. URL https://doi.org/10.1046/j.1525-1497.2002.10750.x.

Qin Ouyang. *Clinical Diagnostics*. People's Medical Publishing House, Beijing, 3rd edition, 2018. National planning textbook for clinical medicine in China.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 2024. URL https://doi.org/10.1109/TKDE.2024.3352100.

Elpiniki I. Papageorgiou. A new methodology for decisions in medical informatics using fuzzy cognitive maps based on fuzzy rule-extraction techniques. *Applied Soft Computing*, 11(1), 2011. URL https://doi.org/10.1016/j.asoc.2009.12.010.

Elpiniki I. Papageorgiou and Jose L. Salmeron. A review of fuzzy cognitive maps research during the last decade. *IEEE Transactions on Fuzzy Systems*, 21(1), 2013. URL https://doi.org/10.1109/TFUZZ.2012.2201727.

Maya Rotmensch, Yoni Halpern, Abdulaziz Tlimat, Steven Horng, and David Sontag. Learning a health knowledge graph from electronic medical records. *Scientific Reports*, 7:5994, 2017. doi: 10.1038/s41598-017-05778-z. URL https://doi.org/10.1038/s41598-017-05778-z.

Kishore Soman, Peter W. Rose, et al. Biomedical knowledge graph-optimized prompt generation for large language models. *Bioinformatics*, 40(6):btae560, 2024. URL https://doi.org/10.1093/bioinformatics/btae560. arXiv:2311.17330.

Chrysostomos D. Stylios and Peter P. Groumpos. Modeling complex systems using fuzzy cognitive maps. *IEEE Transactions on Systems, Man, and Cybernetics*, 34(1), 2004. URL https://doi.org/10.1109/TSMCA.2003.818878.

Xiangru Tang, Yiqun Zhu, Yunxiang Zha, et al. MedAgents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*, 2024. URL https://arxiv.org/abs/2311.10537.

Guojia Wan and Bo Du. GaussianPath: A bayesian multi-hop reasoning framework for knowledge graph reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2021. URL https://doi.org/10.1609/aaai.v35i5.16565.

David S. Wishart, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research*, 36(suppl_1):D901–D906, 2008. URL https://doi.org/10.1093/nar/gkm958.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, and Vicente Grau. Medical Graph RAG: Towards safe medical large language model via graph retrieval-augmented generation. *arXiv preprint arXiv:2408.04187*, 2024. URL https://arxiv.org/abs/2408.04187.

An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Lotfi A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, 1965. URL https://doi.org/10.1016/S0019-9958(65)90241-X.

Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. MedRAG: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM Web Conference 2025 (WWW)*, 2025. URL https://doi.org/10.1145/3696410.3714782. arXiv:2502.04413.

A Gretelai matched evaluation protocol

The Gretelai corpus contains 853 training cases and 212 official test cases across 22 diagnoses. To control the graph comparison while keeping the LLM-based graph-construction budget tractable, we draw a stratified 400-case subset of the training split using seed 20260901. All 22 diagnoses are represented in this subset. We draw a fixed random 100-case evaluation subset from the official test split using seed 20260831. Thus, the corpus sizes describe the available dataset, whereas every metric in Table 3 is computed on the reported 100 test cases.

The FCTG graph is elicited from the 400 training complaints through batched LLM-based symptom extraction and canonicalization. The scalar G1–G3 controls use 270 symptom–disease support edges reconstructed from the same canonical records. The revised FCTG graph retains 330 support edges: singleton edges remain available for candidate recall, but their fuzzy weights are multiplied by 0.25 to prevent a single observation from becoming a high-confidence rule. FCTG uses trapezoidal patient memberships and edge weights on the full graph ($\alpha = 0$).

Two properties of this corpus bound FCTG's headroom on free text and explain why the advantage in Table 3 is narrow rather than decisive. First, the construction subset is small, so per-edge support is sparse: the graph has 330 edges with medium co-occurrence, 89 of which occur only once. Their strengths are necessarily uncertain, and a fuzzy weight on such an edge cannot represent more than the data contain. Second, the complaints are already deterministic: among 235 extracted findings, 149 have membership centroids of at least 0.75 and none fall below 0.25. There is therefore little low-confidence patient evidence for fuzzy memberships to downweight, which leaves limited room for them to improve on binary indicators.

A.1 MedRAG knowledge-graph reconstruction on Gretelai

MedRAG releases a diagnostic knowledge graph for DDXPlus only, so its Gretelai row cannot use a published artifact. We therefore rebuild the graph by applying the construction described in the MedRAG paper to the Gretelai corpus, keeping the released file's schema (`subject`, `relation`, `object`) and its relation inventory, so that the reconstruction is directly comparable to the released DDXPlus graph.

- *Corpus feature decomposition* (their Eq. 4/6). The manifestations of the historical diagnoses are parsed into discrete feature nodes. In the released DDXPlus file this is the dominant layer: 1,362 of the 1,706 non-`is_a` triples are raw database feature values (e.g. `How intense is the pain?: 6`). The Gretelai analogue is a per-case LLM decomposition of each of the same 400 construction complaints into short lowercase clinical features; the 400 cases yield 1,993 feature mentions and 831 distinct phrases, each becoming a finding node.
- *Manifestation augmentation* (their Eq. 5). One LLM call per disease generates distinguishing clinical knowledge in the released relation inventory (`has_symptomatology`, `has_anamnesis`, `has_exposure`, `has_lifestyle`, `has_therapy`); the released file carries 3–15 such triples per disease and the reconstruction 7–12. Unlike the DDXPlus graph, no feature is dropped for low frequency, because MedRAG's construction defines no co-occurrence floor.
- *Two-level taxonomy.* The same call assigns each disease a subcategory and a category, reproducing the `is_a` chain (disease → subcategory → category) that MedRAG's hierarchical aggregation produces. As in the DDXPlus baseline, taxonomy edges and category subjects are removed when the graph is reversed into the shared finding → disease schema.

The resulting triple table contains 1,379 triples over 22 diseases (1,195 `has_symptomatology`, 44 `has_anamnesis`, 44 `has_exposure`, 18 `has_lifestyle`, 44 `has_therapy`, 34 `is_a`). Reversal into the shared finding → disease schema gives 1,345 `supports_pathology` edges over 1,065 findings, scored by the G1 count rule exactly as in the DDXPlus MedRAG baseline; the released DDXPlus graph yields 49 diseases, 409 findings and 1,706 edges under the same reversal. The Gretelai graph is thus substantially denser than the other graphs scored in Table 3 (330 edges for FCTG, 270 for G1–G3), and the corresponding comparison is one of complete KG pipelines rather than a topology-matched comparison of inference rules.

B Fuzzy number and graph notation

The membership function of a trapezoidal fuzzy number $\tilde{z} = (z_1, z_2, z_3, z_4)$ with $0 \le z_1 \le z_2 \le z_3 \le z_4 \le 1$ is

$$\mu_{\tilde{z}}(t) = \begin{cases} 0, & t < z_1 \text{ or } t > z_4, \\ \frac{t - z_1}{z_2 - z_1}, & z_1 \le t < z_2, \\ 1, & z_2 \le t \le z_3, \\ \frac{z_4 - t}{z_4 - z_3}, & z_3 < t \le z_4. \end{cases} \tag{7}$$

When $z_1 = z_2$ or $z_3 = z_4$, the corresponding ramp has zero width and its branch is empty; the membership is then read in the standard limit sense, with the vertical shoulder belonging to the full-membership core, so every integral below is well defined. The set of valid trapezoidal fuzzy weights and memberships is

$$\mathcal{T}_{[0,1]} = \{(z_1, z_2, z_3, z_4) \in [0,1]^4 : z_1 \le z_2 \le z_3 \le z_4\}, \tag{8}$$

where $[0,1]^4$ means that all four components lie in the unit interval, and the ordering constraint ensures that the tuple defines a valid trapezoid. Aggregation leaves this space: component-wise fuzzy addition of nonnegative trapezoids preserves the ordering but not the unit bound, so the aggregated disease score $\tilde{S}_\alpha(d \mid x)$ of Eq. (6) lies in the larger ordered cone $\mathcal{T}_{\mathbb{R}_{\ge 0}} = \{(z_1, z_2, z_3, z_4) \in \mathbb{R}^4_{\ge 0} : z_1 \le z_2 \le z_3 \le z_4\}$, whose components may exceed 1 (ten contributions of centroid 0.5 reach 5). Its centroid is accordingly a nonnegative *ranking score* used only to order candidates—neither a probability nor a confidence. The α-cut of Eq. (2) is unaffected, since it compares centroids of

memberships and edge weights, which lie in $\mathcal{T}_{[0,1]}$ and therefore have centroids in $[0, 1]$. FCTG is then defined as

$$G = (V, E_{\text{sp}}, E_{\text{dd}}, \tilde{W}, \mathcal{A}), \tag{9}$$

where $V = \mathcal{S} \cup \mathcal{D}$; the mapping $\tilde{W} : E_{\text{sp}} \to \mathcal{T}_{[0,1]}$ returns the trapezoidal weight $\tilde{W}(e) = (w_{e,1}, w_{e,2}, w_{e,3}, w_{e,4})$ of a support edge $e = (s, d)$; and $\mathcal{A} \subseteq [0, 1]$ contains the candidate α-cut thresholds the system is allowed to use, one of which is selected at inference time. The trapezoidal support weight $\tilde{w}_{sd}$ of Section 4.1 is obtained as

$$\begin{aligned} w_{sd,1} &= \max(0, m_{sd} - r_{sd}), \\ w_{sd,2} &= \max(w_{sd,1}, m_{sd} - 0.3r_{sd}), \\ w_{sd,3} &= \min(1, m_{sd} + 0.3r_{sd}), \\ w_{sd,4} &= \min(1, m_{sd} + r_{sd}). \end{aligned} \tag{10}$$

Since $m_{sd} \in [0, 1]$ and $r_{sd} \geq 0$, these endpoints satisfy $0 \leq w_{sd,1} \leq w_{sd,2} \leq m_{sd} \leq w_{sd,3} \leq w_{sd,4} \leq 1$, ensuring a valid trapezoidal fuzzy weight. Finally, the diagnostic output is the ranked list

$$\hat{\pi}(x) = [(\hat{y}_1, p_1), \ldots, (\hat{y}_K, p_K)], \tag{11}$$

with $\hat{y}_1$ the Top-1 prediction and p_k the LLM-reported probability.

C Inference algorithm

Algorithm 1 states the end-to-end FCTG inference procedure; every step applies the equations defined in Section 4.1 and Appendix B.

Algorithm 1 FCTG Inference

Require: Patient narrative x, fuzzy KG G, threshold α
Ensure: Ranked differential diagnosis $\hat{\pi}(x)$
1: Extract matched symptoms $\{(s_i, \tilde{\mu}_{x,i})\}$ with the LLM
2: Keep $\mathcal{M}_\alpha(x) = \{(s_i, \tilde{\mu}_{x,i}) : \mathcal{C}(\tilde{\mu}_{x,i}) \geq \alpha\}$
3: Build G_α using Eq. (2)
4: **for** each pathology $d \in \mathcal{D}$ **do**
5: Compute $\tilde{S}_\alpha(d \mid x)$ using Eq. (6)
6: Set `rank_score`$(d) = \mathcal{C}(\tilde{S}_\alpha(d \mid x))$
7: **end for**
8: Render the method-specific KG-constrained reranking prompt
9: Query the diagnostic LLM and parse strict JSON output

D A DDXPlus case in detail

Table 7 shows one record of the DDXPlus construction file (row 9999) in the form in which the corpus stores it: coded evidence tokens, demographic information, the true pathology, and a differential diagnosis, together with the decoding of eight tokens through the released evidence dictionary and an abridged version of the free-text presentation that the methods actually diagnose.

E Prompting and reproducibility details

All methods in Table 3 use Qwen3.8-Flash through the same OpenAI-compatible API endpoint. We request strict JSON output, disable the model's thinking mode, set the temperature to 0.1, and use a maximum completion budget of 8,192 tokens. Each KG method uses a lightweight lexical prefilter of 120 finding nodes before LLM-based symptom matching, followed by a compact, rank-aware diagnostic prompt. G1–G3 treat every extracted symptom as a binary indicator of presence; FCTG instead requests a trapezoidal membership and retains at most 15 matched findings. For every KG method, token accounting includes both runtime LLM calls: symptom extraction and final diagnosis.

Field	Value
AGE, SEX	65, M
PATHOLOGY	Panic attack
DIFFERENTIAL_DIAGNOSIS	Panic attack 0.240; Possible NSTEMI/STEMI 0.177; Unstable angina 0.147; Inguinal hernia 0.122; Boerhaave 0.119; . . . (8 entries)
EVIDENCES	20 codes, including E_16, E_54_@_V_182, E_55_@_V_55, E_56_@_7, E_57_@_V_123, E_58_@_10, E_204_@_V_10
INITIAL_EVIDENCE	E_177
Decoded through release_evidences.json:	
E_54_@_V_182	"Characterize your pain:" → "a cramp"
E_55_@_V_55	"Do you feel pain somewhere?" → "side of the chest (R)"
E_56_@_7	"How intense is the pain?" → 7 (ordinal 0–10)
E_57_@_V_123	"Does the pain radiate to another location?" → "nowhere"
E_58_@_10	"How precisely is the pain located?" → 10 (ordinal 0–10)
E_204_@_V_10	"Have you traveled out of the country in the last 4 weeks?" → "N"
E_16	"Do you feel anxious?" (antecedent) → absent
E_177	"Do you currently, or did you ever, have numbness, loss of sensitivity or tingling anywhere on your body?" → absent
Rendered input (abridged): *A 65-year-old male patient. The main complaint is pain in the right lateral chest, left lateral chest, right lateral waist, left lateral waist, and left lower quadrant of the abdomen. . . .*	

Table 7: One DDXPlus case (row 9999). The structured fields are the graph-construction input; the rendered presentation is the text that every method diagnoses. Ordinal values follow the scale declared by the evidence dictionary.

We retain the fixed split files, 400-case construction subset, canonical symptom-extraction cache, graph artifacts, prompts, and per-case prediction JSONL files. These artifacts allow exact recomputation of every reported metric from the stored model responses. As in other hosted-LLM evaluations, the output of a new API call may vary slightly because provider-side inference is not guaranteed to be bitwise deterministic; the released response artifacts define the reported run.

F Prompt templates

We use a two-stage prompting procedure for FCTG. First, an extraction LLM maps the patient narrative to canonical finding nodes and assigns a trapezoidal membership. Second, the diagnostic LLM reranks a closed, KG-ranked candidate set. The templates below show the fixed instructions; angle-bracketed fields are populated separately for each case. G1–G3 use the same reranking template but replace the FCTG formula and omit the FCTG-specific uncertainty guidance.

FCTG symptom-extraction template.

```
System:
You are a medical symptom extraction expert.
Given a patient description and known medical
findings from a knowledge graph, extract and
match symptoms.
For each match, assign a trapezoidal fuzzy
number (z1,z2,z3,z4) representing clarity,
severity/frequency, and clinical salience.

Select one band; all four values must stay in it:
- very weak/vague: 0 < z1 <= z2 <= z3 <= z4 < .25
- mild/occasional: .25 < z1 <= z2 <= z3 <= z4 < .50
- clear/moderate:  .50 < z1 <= z2 <= z3 <= z4 < .75
- strong/definite: .75 < z1 <= z2 <= z3 <= z4 < 1.00
Use a narrow tuple for a precise description and
a wider tuple when its strength is uncertain.
Do not output disease probabilities. Return JSON:
{
 "extracted": [{
   "symptom": "...", "finding_id": "...",
   "fuzzy": [z1,z2,z3,z4]
```

```
 }]
}

User:
Patient description:
<PATIENT_TEXT>

Known findings in the knowledge graph:
<PREFILTERED_FINDING_LIST>

Extract all matching symptoms. Return at most
15 findings; prefer the most certain, clear matches.
```

KG-constrained diagnostic-reranking template.

```
System:
You are a medical differential-diagnosis reranker.
Use the KG-ranked candidate list as the closed
candidate set. Do not introduce diseases outside it.
Return strict JSON only.

User:
Patient description:
<PATIENT_TEXT>

Matched findings:
<MATCHED_FINDINGS_WITH_FUZZY_MEMBERSHIPS>

Formula used for KG ranking:
Score(d) = centroid(fuzzy-sum over f of
                    (mu_f fuzzy-product w_(f,d)))
<METHOD_SPECIFIC_FORMULA>

FCTG-specific uncertainty guidance:
- Patient descriptions may be fuzzy or incomplete,
  or contain minor symptoms from confusable diseases.
- Use the fuzzy KG ranking as a prior. Mild, vague,
  hedged, or incidental findings should not alone
  overturn the top-ranked disease.
- Missing mild symptoms are not strong negative evidence.

KG-ranked candidate diseases and compact evidence:
<UP_TO_SIX_CANDIDATES_AND_UP_TO_THREE_SUPPORTING_EDGES_EACH>

Allowed candidate disease names:
<CANDIDATE_LIST>

Rerank only the allowed candidates. Return exactly:
{"PATHOLOGY": "single most likely allowed disease",
 "DIFFERENTIAL_DIAGNOSIS":
   [["allowed disease", probability], ...]}
The differential must contain 3 to 5 allowed diseases,
in descending probability order. Probabilities must be
non-negative and sum to approximately 1.
```

G Baseline details

All five baselines are knowledge-graph methods and therefore share the same disease-constrained output space as FCTG.

Baselines:

- **MedRAG** (Zhao et al., 2025): retrieves related (subject, relation, object) triples from established medical knowledge graphs together with corpus snippets, and lets the frozen LLM reason over the retrieved evidence. It tests triple-based retrieval with no edge statistics.
- **MedGraphRAG** (Wu et al., 2024): very similar to MedRAG, but it first retrieves at the document granularity and then retrieves entity triples related to the most similar documents as augmented information, thereby ensuring source traceability.
- **G1 Probabilistic KG (Count)** (Jia et al., 2025; Gao et al., 2025): support edges encode presence only, and the disease score is the number of matched symptoms. The prompt includes no edge statistics.

- **G2 Probabilistic KG (Posterior Sum)** (Guyatt et al., 2015; Rotmensch et al., 2017): the edge weight is $P(\text{disease} \mid \text{symptom})$, and the disease score is the sum of posterior probabilities over matched symptoms.
- **G3 Probabilistic KG (LR^+ Sum)** (Albagli-Kim & Beimel, 2026): the edge weight is $\min(\log \text{LR}^+ / \log 10, 1)$, computed from the same Haldane–Anscombe-smoothed conditionals as FCTG (§4.1), and the disease score is $\sum_s w(s,d) \cdot \text{presence}(s)$.

The three crisp-scalar baselines (G1–G3) use scalar edge weights and scalar aggregation operators; uncertainty is collapsed at the first computation step, and they differ only in which scalar formula maps matched symptoms to disease scores.

Fuzzy-number KG (Ours):

- **FCTG**: every edge weight is a trapezoidal fuzzy number $\tilde{w}_{sd} = (w_{sd,1}, \ldots, w_{sd,4})$ with an LR^+-based center and a delta-method standard-error width; an LLM assigns each patient symptom a fuzzy membership $\tilde{\mu}_i$; and the disease score is $\mathcal{C}(\bigoplus_i \tilde{\mu}_i \otimes \tilde{w}_{sd})$, obtained through fuzzy multiplication and addition followed by centroid defuzzification.

The scoring rule that each method applies to matched symptoms is summarized in Table 8. It isolates two design choices: how symptom–disease edges are weighted, and whether patient-side symptoms are treated as binary evidence or fuzzy memberships. The summation ranges over symptoms extracted from the patient narrative by the LLM. Here $w_{\text{LR}}(s,d) = \min(\log \text{LR}^+(s,d)/\log 10, 1)$ is G3's scalar edge score; $\tilde{\mu}_s$ and $\tilde{w}_{sd}$ are the LLM-assigned symptom membership and LR^+-based fuzzy edge weight; $\otimes$ and $\oplus$ denote fuzzy multiplication and addition, and $\mathcal{C}$ denotes centroid defuzzification of the aggregated score.

Method	**Evidence-scoring rule for disease d given case x**
G1 (Count)	$\sum_{s \in x} \mathbf{1}[(s,d) \in E]$
G2 (Posterior Sum)	$\sum_{s \in x} P(d \mid s)\, \mathbf{1}[(s,d) \in E]$
G3 (LR^+ Sum)	$\sum_{s \in x} w_{\text{LR}}(s,d)\, \mathbf{1}[(s,d) \in E]$
MedRAG	$\sum_{s \in x} \mathbf{1}[(s,\cdot,\cdot,d) \in \mathcal{K}_{\text{med}}]$, counted over the triples retrieved from its medical knowledge graph
MedGraphRAG	No closed-form score; one document is selected by $\cos(\mathbf{q}, \mathbf{u}_g)$ and an LLM answers from its intra- and inter-document triples
Fuzzy-number knowledge graph (ours)	
FCTG	$\mathcal{C}\left(\bigoplus_{s \in x} \tilde{\mu}_s \otimes \tilde{w}_{sd}\right)$, with the α-cut applied to edges and memberships

Table 8: Evidence-scoring rules of every compared method

H Hyperparameters

Table 9 lists all hyperparameters. Values marked "fixed" were chosen before evaluation and never tuned on evaluation cases; α was selected on a development subset and held fixed across every method and every table in the paper.

I Symptom matcher ablation

Table 10 substitutes the sentence-to-symptom matcher while holding the graph, scoring rule, and diagnostic stage fixed (confusable 100-case DDXPlus subset, Qwen3.8-Flash).

J Efficiency details

Table 11 reports one-time construction cost and per-case inference latency, separating graph scoring from the LLM call. On DDXPlus, construction involves no LLM: the training pool provides struc-

Component	Setting	Selection
Confidence level α (DDXPlus)	0.35	dev-tuned, fixed
Confidence level α (Gretelai)	0	fixed (full graph)
Singleton discount	0.25	fixed
Candidates passed to reranker	6	fixed
Supporting edges per candidate	3	fixed
Edge weight center (DDXPlus)	$\text{clamp}(\log_{10} \text{LR}^+, 0, 1)$	—
Zero-frequency handling	Haldane–Anscombe ($+0.5$ pseudo-count)	fixed
Edge width estimator	delta method with smoothed conditionals	—
Membership width source	LLM tuple (narrow/wide choice)	—
Aggregation	fuzzy product / sum	—
Defuzzification	centroid	—
Backbone (main)	Qwen3.8-Flash	—
Backbone (transfer)	Qwen2.5-72B-Instruct	—
Decoding	$T = 0.1$, JSON mode	fixed
Seeds	single fixed run	—

Table 9: Hyperparameters and implementation details.

Matcher	Top-1	Top-3	M-F1
LLM semantic (default)	**88.0**	**98.0**	**51.4**
BM25 lexical	79.0	97.0	48.2
Token overlap	56.0	80.0	30.4
Embedding-only	52.0	88.0	34.1

Table 10: FCTG sensitivity to the symptom-matching mechanism on DDXPlus (confusable 100-case subset, Qwen3.8-Flash)

tured evidence tokens, so the graph is produced by a deterministic counting pass over the 4,000 cases (the 0.5 s base pass plus the per-method derivation). LLM cost is incurred only per case at inference time—symptom extraction plus final diagnosis—and appears in the LLM-per-case column.

Method	Construction	Graph / case	LLM / case	Edges scored
G1 (Count)	0.57 s	0.18 ms	8.9 s	109.1
G2 (Posterior Sum)	0.50 s	0.20 ms	8.8 s	109.1
G3 (LR^+ Sum)	0.57 s	0.17 ms	9.3 s	109.1
FCTG	0.57 s	0.21 ms	9.3 s	**51.7**

Table 11: Efficiency on DDXPlus

The LLM calls dominate: graph scoring costs 0.17–0.21 ms per case, four to five orders of magnitude below the ≈ 9 s of per-case LLM latency, and the one-time construction pass is under a second. The α-cut therefore does not make the pipeline faster end-to-end: it halves the arithmetic, from 109.1 to 51.7 aggregated (finding, edge) products per case, which matters only if the graph step is embedded in a larger offline sweep or run against a cheaper local scorer. Consistently with Table 5, FCTG's graph step is not cheaper in wall-clock time (0.21 ms versus 0.17 ms for G3), since each surviving edge is scored with trapezoidal arithmetic rather than a scalar weight; the saving is in the number of interval operations, not in latency.

K Error analysis

Table 12 buckets every residual FCTG error on the 100-case analysis sets for each dataset by asking what the graph already offered. Three buckets bound what is achievable with the current graph. *Absent edge* dominates Gretelai (7 of 13): no matched finding feeds the gold disease, largely because the patient's symptoms are matched to finding nodes that carry no support edge at all—54.9% of the Gretelai graph's finding nodes are isolated, against 21.4% on DDXPlus. *Indistinguishable* accounts for both DDXPlus failures that are not a ranking problem (stable versus unstable angina; acute

laryngitis versus viral pharyngitis, whose finding sets overlap at Jaccard 0.73 and 0.50), and *wrong canonical node* for the two Gretelai cases in which the patient's wording does reach a finding the graph links to the gold disease, but that finding was never selected.

The remaining bucket bounds what a better reranker could add: the graph already ranks the correct disease inside the top-6 window shown to the diagnosis stage, so at most $6/100$ on DDXPlus and $8/100$ on Gretelai are still recoverable. On DDXPlus this bucket is a graph-ranking shortfall rather than a reranking mistake, because the fusion leaves the graph's own top-1 untouched in all 50 cases.

Error category	**DDXPlus**	**Gretelai**	**Example**
Symptom matched to wrong canonical node	0	4	#14
Correct edge absent from graph	2	14	#39
Diseases indistinguishable from symptoms	4	0	#12
Answer in graph top-6, not selected	6	8	#20
Total errors	12	26	

Table 12: Error analysis of the residual FCTG failures on the 100-case analysis sets for each dataset