FROM QUERY TO KNOWLEDGE TOPOLOGY: AGENTIC KNOWLEDGE EXTRACTION FOR MULTI-PERSPECTIVE GRAPH RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

User queries can be partial and potentially ambiguous descriptions of the knowledge required to answer them. Retrieval based on a single query representation can therefore overlook relevant information, relationships, and alternative interpretations that remain implicit in the question. We propose a training-free framework for multi-perspective answer-set retrieval that reconstructs a candidate query-relevant knowledge topology through iterative LLM probing and uses this structure to guide retrieval from an external knowledge graph. Building on existing agentic knowledge-extraction methods, the framework explores the query through sequential associative probing, taxonomy exploration, and multi-perspective parallel probing. Extracted knowledge is filtered for relevance and redundancy and organized into an explicit structure of information, relationships, and perspectives. This structure provides a provisional model of the query's knowledge requirements: its perspectives guide adaptive region retrieval, while its relationships inform the construction of supporting evidence through subgraphs. Retrieved entities seed a prize-collecting procedure over knowledge-graph relations, and the resulting subgraphs are organized into perspective-specific answer sets. The LLM then synthesizes these evidence sets into an answer that presents complementary findings and evidence-supported disagreements. We evaluate the framework on FB15K-237, WN18RR, PrimeKG, and NELL-995 using retrieval accuracy and answer-set distinctness.

1 INTRODUCTION

A user asking a question to a system usually does not format that question in a way that optimizes it for retrieval. "What can you tell me about China?" is a very vague and broad question that does not clearly define what the user wants to know. A useful answer still has to decide whether what the user really wants is geography, economy, or history, and then find facts for that perspective. The question ends up being a small observation of a larger structure. We call a candidate reconstruction of that structure a *query-relevant knowledge topology*: a typed hypothesis about the knowledge an answer requires, naming the perspectives the answer should cover, the concepts that sit under each perspective, and the relations that would connect those concepts to evidence. A user asks from the part of their own topology they can articulate; the language model carries a much broader one, and reconstruction is how we move between the two. The object is a typing of an information need rather than the connectivity of a network, and the paper uses the term in that sense throughout.

Most retrieval systems do not take this kind of approach. They turn the question into a single embedding, a list of numbers that places text near other text with a similar meaning, and they keep the k nearest entities in a knowledge graph. A knowledge graph stores facts as triples: a head entity, a relation, and a tail entity. The list length k is chosen before any entity is seen, and the budget is spent on whatever sits closest to the original wording. Naturally, this kind of methodology struggles with vague and broad questions. If the only evidence a language model sees comes from one neighborhood, the model starts from one side of the question and tends to write the consensus answer. We call that failure thought convergence, and we treat it as a retrieval problem: the model cannot present a second reading of the question using evidence the retriever never fetched.

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.

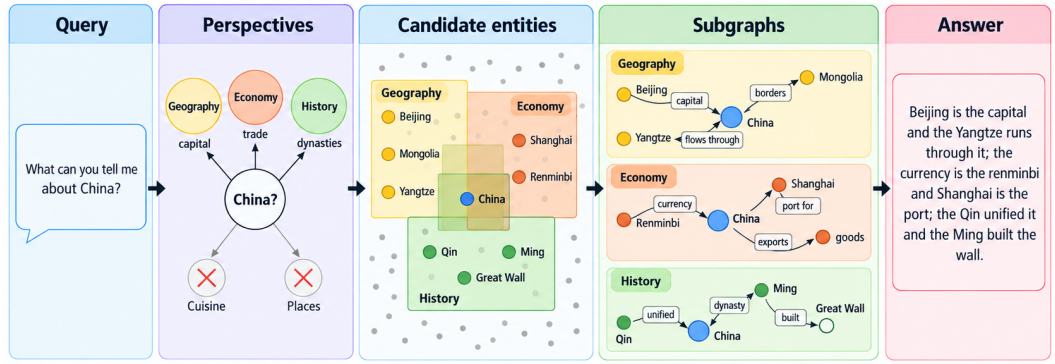

Figure 1: Given a natural-language query, the framework first identifies a set of relevant perspectives that represent distinct aspects of the information need. Each perspective is then used to select candidate entities from the knowledge graph, from which a perspective-specific subgraph is constructed. The resulting subgraphs capture complementary relational evidence associated with the original query and are subsequently aggregated to generate the final response.

Our framework reconstructs a topology first and retrieves second. A language model probes the question for perspectives the user did not name, and each probe sees the perspectives already kept, so it has to add something the earlier probes did not say. A taxonomy step then attaches a broader category and a more specific concept to each kept perspective, which shifts the level of detail at which that perspective can be grounded. Two filters drop the perspectives that drift away from the question and the ones that repeat a perspective already kept. Each surviving perspective is searched on its own: a region of the embedding space rather than a shared top-k list. The entities inside that region are a candidate set, and those entities seed a connected subgraph of triples. The subgraph stays with the perspective that grew it, and that pair is an answer set. At the end, one call writes a single answer from the non-empty sets and is told to keep a disagreement when the facts on both sides exist. A perspective the graph cannot support stays empty; the model does not fill it in from memory. Figure 1 follows the China question through the pipeline.

The framework separates a hypothesis about the question from evidence about the world, and that separation raises an empirical question we take seriously: when a generic text embedding is searched region by region, do the reconstructed perspectives actually retrieve different evidence? We measure the question on four knowledge graphs across three families of queries, comparing the subgraph with a fixed top-k list, the incident triples of the single nearest entity, an unfiltered neighborhood at the same entity budget, a region drawn around the raw question with no reconstruction at all, and a bounded walk. The pattern that emerges is mixed, and it is informative: the subgraph is cleaner than a fixed list on broad questions and on one-hop questions, the perspectives usually settle on the same one or two anchors, and the far end of a two-hop path stays out of reach. We report those outcomes as measurements that locate the bottleneck in the embedding geometry, and we use them to state what a topology reconstruction needs before it can retrieve evidence that is genuinely multi-perspective.

The contributions of this paper are:

- A formulation of query understanding as the reconstruction of a candidate query-relevant knowledge topology, kept distinct from the knowledge graph that supplies the facts. The topology is a hypothesis about what an answer needs; grounding decides which parts of it the graph can support.

- A training-free instantiation of the framework: sequential associative probing, taxonomy exploration, relevance and redundancy filters, region retrieval per perspective, a prize-collecting subgraph grown from the candidate entities, and a synthesis step that folds the answer sets into one answer and preserves supported disagreements.
- A measurement of the framework on FB15K-237, WN18RR, PrimeKG, and NELL-995, split by question family and paired across baselines, together with an analysis of the geometry that determines when reconstructed perspectives can be told apart at all.

2 Related work

Retrieval over knowledge graphs. Retrieval-augmented generation embeds a query and conditions a language model on the nearest documents (Lewis et al., 2020). Over a knowledge graph, the same recipe ranks entities or triples by similarity to the question and returns a fixed number of items. SubgraphRAG scores triples with a lightweight network and passes the selected subgraph to a model (Li et al., 2025); G-Retriever retrieves one connected subgraph by a prize-collecting Steiner tree (He et al., 2024); Think-on-Graph walks relations and keeps a beam of paths (Sun et al., 2024); GraphRAG builds a corpus-level community hierarchy for query-focused summarization (Edge et al., 2024). These methods optimize the evidence for a single answer. G-Retriever solves for that one subgraph as a prize-collecting Steiner tree. The grower in Section 4 does not. It adds scored triples in order and stops at 40. Our unit is the perspective: one subgraph per reconstructed perspective, and the subgraphs remain separate all the way into the written answer. The graph RAG baseline in Section 5 is the simpler control: the triples on the single nearest entity, with no perspectives and no prize score.

Questions with more than one answer. AmbigQA showed that a question can have several defensible answers and that evaluation should accept them (Min et al., 2020). Debate prompting and self-refinement let a model criticize and revise its own output (Du et al., 2024; Madaan et al., 2023). We put the criticism on the retrieval side: what gets refined is a set of search directions, and a direction is discarded by comparing embeddings rather than by asking the model to grade itself. Splitting the question into clauses and retrieving nearest facts for each clause is the closer alternative, and we do not run it. The questions in this paper are not AmbigQA items; Section 5 states precisely what they test.

Region retrieval. Query2Box represents a query as a box and trains projections and intersections over boxes (Ren et al., 2020). We keep the membership test and drop the training: a region is grown from the local spread of the nearest entities, so a perspective returns however many entities fall inside it, and no single similarity threshold has to separate "near" from "far". We do not learn projections and we do not intersect boxes.

Geometry of text embeddings. Text embeddings concentrate in a narrow cone: mean pairwise similarity between unrelated sentences is far above zero, and the effect grows with model capacity (Ethayarajh, 2019; Mu & Viswanath, 2018). That property matters here because a reconstructed perspective is a point in such a space, and the perspectives of one question are points in the same cone. Section 5.3 measures this directly instead of assuming it away.

3 Knowledge topology

3.1 What the object is

Let $G = (E, R, T)$ be a knowledge graph with entity set E, relation set R, and triple set $T \subseteq E \times R \times E$, where a triple (h, r, t) says that a relation r describes the connection between the head h and its tail t. Let emb map the description text to a unit vector in $\mathbb{R}^d$.

A *query-relevant knowledge topology* for a question q is

$$\mathcal{T}_q = (P_q, C_q, R_q),$$

where P_q is a finite set of perspectives, C_q assigns each perspective a parent category and a child concept, and R_q is a set of candidate relations between those concepts. $\mathcal{T}_q$ is a hypothesis about the

information that an answer would need. Nothing in the hypothesis has to exist in G. The topology states what the answer requires, and grounding (Section 4) decides which parts the graph can supply that information. A perspective that grounds nowhere is left empty rather than filled from model memory, which ensures that answers are grounded in domain-specific facts.

3.2 RECONSTRUCTION

Reconstruction builds $\mathcal{T}_q$ from the original question. It is a fixed sequence of LLM-driven construction, and what changes from one response to the next is the text the model sees. Starting from $\mathcal{T}_q^{(0)} = \{q\}$:

1. **Sequential associative probing.** The model proposes perspectives that the question leaves unsaid, and every probe is shown the perspectives already accepted, so later probes are pushed toward material the earlier ones missed. A round returns a pool of candidates; each candidate is embedded utilizing its name and description, and the filters decide which perspectives are accepted into P_q. Rounds continue until five perspectives are kept or until three consecutive rounds add nothing new.
2. **Taxonomy exploration.** One call per kept perspective returns a parent category and a child concept, such as *Economy* and *trade* for a perspective about American economic policy. The two strings are embedded together and carried as a direction that grounding can use.
3. **Relevance and redundancy filtering.** A candidate is dropped if its raw dot product with the question embedding falls below 0.5, or if its dot product with a perspective already kept rises above 0.85. The first test removes probes that are irrelevant to what was actually asked. The second test removes restatements of a perspective that is already in P_q.
4. **Multi-perspective parallel probing.** Once P_q is fixed, the perspectives are independent from one another. Each perspective is searched in its own region and grows its own set of candidate entities, with no shared budget between them.

The order of operators is fixed, but the search stage is where the multi-perspective property is decided, because that is the stage that touches the graph.

4 GROUNDING A TOPOLOGY IN A KNOWLEDGE GRAPH

Grounding turns each component of $\mathcal{T}_q$ into verifiable evidence. A perspective is represented by its embedding, optionally refined by its taxonomy direction. The graph then decides which of those components it can support.

4.1 RETRIEVING CANDIDATE ENTITIES

For a perspective p we need the entities of G that lie near it. Using a single cosine threshold is a fragile form of retrieval in this vector space. Entity pairs share a mean cosine of only 0.12 to 0.24, and the nearest entity to a question lies near 0.65. A cutoff in that narrow band flips between retrieving almost nothing and retrieving too many entities. To address this, we build a small box around each perspective in the directions the local neighborhood actually spreads.

Entities are ranked by cosine to p, and the nearest 64 define the shape of the search region. A singular value decomposition of those centered vectors gives the four directions of largest spread. The box is centered on p. Its width along each axis is the 90th percentile of the absolute deviation of the 64 entities along that axis, and a separate residual width covers the part of each of those vectors that the four axes do not explain. An entity enters the candidate set when its outside distance (Ren et al., 2020) is zero, which means it lies inside the box on every axis and inside the residual. The widths adapt to the graph. If fewer than five entities fall inside, the box grows by $1.25\times$ per step up to $4\times$. If more than ten fall inside, it shrinks down to $0.25\times$ and the closest ten are kept. The budget of ten was chosen on the first eight questions of each graph, where ten and twenty tied and forty was worse; those questions are excluded from every reported number. Because the budget is per perspective, a question with five perspectives can retrieve five separate candidate sets, or none, depending on the graph.

4.2 SUBGRAPHS FROM CANDIDATE ENTITIES

The candidate entities are seeds, and the subgraph is the evidence that gets shown to the model. A triple enters a subgraph through a prize score,

$$s(h,r,t) = \cos(\mathrm{emb}(r),\mathbf{q}) + \cos(\mathrm{emb}(r),\mathbf{p}) + \lambda\cos(\mathrm{emb}(r),\mathbf{c}) + \beta + 0.5\max_{e\in\{h,t\}}\cos(\mathrm{emb}(e),\mathbf{p}) + 0.75\max_{e\in\{h,t\}}\cos(\mathrm{emb}(e),\mathbf{q}),$$

where $\mathbf{q}$, $\mathbf{p}$, and $\mathbf{c}$ are the embeddings of the question, the perspective, and the taxonomy direction. The first term asks whether the relation is about the question, and the second asks whether it is about this perspective. The taxonomy term, weighted by $\lambda = 0.5$, pulls toward $\mathbf{c}$ when that direction exists. The bridge bonus $\beta = 1.5$ is added when both endpoints are seeds, so a triple that joins two candidate entities is preferred to one that only touches the set. The last two terms pull in triples whose head or tail is close to the perspective or to the question.

Growth starts from the seeds and proceeds greedily in the style of Kruskal's algorithm. Links between candidate entities are added strongest first. Length-two paths through a non-candidate connector are added next, when they join two components of the candidate set. Remaining high-scoring triples that touch the subgraph are added after that, up to 40 triples. A query-relevant triple is offered to the perspective that scores it highest, so a popular fact does not appear in every subgraph. A final pass drops low-scoring triples that are not needed to keep the seeds connected.

4.3 ANSWER SETS

Each subgraph stays with the perspective that grew it, and that pair is an answer set: an attributed piece of evidence, unrelated to answer-set programming. The assignment is fixed, so a fact never moves to a perspective it was not grown for. Facts are formatted as `head --- relation →` `tail`. One final model call receives the query and the non-empty answer sets. The call uses only the listed facts, folds every non-empty perspective into one answer, presents both sides when the facts conflict, and does not mention the retrieval machinery.

5 EXPERIMENTS

5.1 SETUP

We use four knowledge graphs. FB15K-237 (Toutanova & Chen, 2015) has 14,541 entities and 237 relation types, WN18RR (Dettmers et al., 2018) has 40,942 entities and 11 relations, PrimeKG (Chandak et al., 2023) has 10,344 entities and 11 relations, and NELL-995 (Carlson et al., 2010; Xiong et al., 2017) has 75,492 entities and 400 relations. Each graph contributes 64 questions. The first 32 were drawn with seed 42 and are the questions of the earlier version of this evaluation. The next 32 were drawn with seed 43. Each sample has 16 broad questions of the form "What can you tell me about X?", 8 one-hop questions that ask for the tails of one head and one relation, and 8 two-hop questions that ask for the endpoints of a two-step path. The seed-42 questions stay first and the seed-43 questions follow, so every graph contributes 32 broad, 16 one-hop, and 16 two-hop questions and the calibration slice does not change. Gold triples for a broad question are the triples incident to X, capped at 50. For the path questions they are the triples along the path. The full list is in Appendix A.

On a two-hop question the gold entities and the gold triples are different sets. The gold entities are the far endpoints, while the gold triples include both hops, so a retriever that returns only the first hop can score triple F1 above zero while entity F1 stays at zero. Appendix C spells this out on one question. The one-hop and two-hop questions were worded by a model that saw the entity name and the relation text, so the relation string can leak into a metric that matches raw relation strings. Broad questions use a fixed template and carry no relation text. We report every number by family for exactly this reason.

The cap of ten entities was chosen on the first eight questions of each graph, where ten and twenty tied and forty was worse. Those questions are excluded from all headline numbers, leaving 56 held-out questions per graph, 24 broad, 16 one-hop, and 16 two-hop. We report macro triple F1, triple MRR, and Hits@$1/3/5/10$ with paired 95% intervals from 2000 bootstrap resamples at seed 0. The

Table 1: Triple precision, recall, and F1 on the 56 held-out questions per graph, subgraph against graph RAG, plus triple MRR for the subgraph against the endpoint-cosine ranked control. The held-out slice has 24 broad, 16 one-hop, and 16 two-hop questions. Graph RAG is the one-hop neighborhood of the nearest entity. Figure 2 plots the F1 and MRR rows.

		Subgraph			Graph RAG			Triple MRR	
Graph	Family	P	R	F1	P	R	F1	Sub	Ctrl
FB15K-237	Broad	0.648	0.737	0.683	0.923	0.958	0.938	0.708	0.911
	One-hop	0.072	0.562	0.119	0.097	0.969	0.167	0.422	0.300
	Two-hop	0.043	0.280	0.070	0.076	0.426	0.121	0.346	0.234
WN18RR	Broad	0.442	0.720	0.521	0.615	0.596	0.600	0.588	0.741
	One-hop	0.201	0.875	0.318	0.184	0.422	0.248	0.570	0.341
	Two-hop	0.065	0.325	0.100	0.171	0.346	0.217	0.311	0.395
PrimeKG	Broad	0.766	0.866	0.798	0.859	1.000	0.902	0.885	0.847
	One-hop	0.229	0.754	0.322	0.270	1.000	0.388	0.579	0.426
	Two-hop	0.120	0.146	0.107	0.180	0.166	0.150	0.385	0.379
NELL-995	Broad	0.544	0.743	0.598	0.781	0.792	0.785	0.714	0.840
	One-hop	0.186	0.625	0.268	0.259	0.688	0.367	0.303	0.426
	Two-hop	0.081	0.224	0.111	0.126	0.276	0.156	0.245	0.413

subgraph ranks its own triples by the prize score of Section 4. The control ranks every triple in the graph by the best endpoint cosine to the raw question. Graph RAG has no ranking, because every triple in its pile touches the same anchor.

The system is training-free. Perspectives and taxonomy concepts come from `gpt-4o-mini`, and all embeddings come from `text-embedding-3-large`. The comparison the rest of this section is built around is graph RAG. Embed the raw question, take the single nearest entity, and keep every triple that touches it. That is one hop out from the anchor, returned as one pile of facts, with no perspectives and no prize score. The fixed-budget version of the same idea is top-10, which keeps the triples on the ten nearest entities. The same questions are still scored against an unfiltered neighborhood at the matched entity budget, one region around the question embedding with no reconstruction, and Think-on-Graph with depth 2, beam width 2, and a shortlist of the ten relations closest to the current entity. Think-on-Graph here is our reimplementation.

5.2 Results

Table 1 reports triple precision, recall, and F1 for the subgraph and for graph RAG, and triple MRR for the subgraph against the endpoint-cosine ranked control. Figure 2 plots the F1 and MRR rows. Table 2 in the appendix adds Hits@$1/3/5/10$, Table 3 in the appendix reports the per-question wins, ties, and losses, and Table 4 in the appendix reports the structure of the retrieved subgraphs. Table 5 in the appendix keeps the other controls on the same questions, and Figure 4 in the appendix shows the per-question paired comparison behind the broad and one-hop paragraphs.

On broad questions graph RAG is the higher F1 on FB15K-237 (0.938 versus 0.683), PrimeKG (0.902 versus 0.798), and NELL-995 (0.785 versus 0.598). The paired gaps, subgraph minus graph RAG, are -0.255 $[-0.414, -0.111]$, -0.104 $[-0.173, -0.051]$, and -0.188 $[-0.399, 0.011]$. The first two intervals lie below zero and the NELL-995 interval reaches zero. On WN18RR the point estimate also favors graph RAG (0.600 versus 0.521) and the interval is wide (-0.079 $[-0.279, 0.126]$). A broad question names one entity, and the gold is that entity's neighborhood, capped at 50 triples. On FB15K-237 and PrimeKG the anchor is found almost always, and its recall is 0.958 and 1.000. WN18RR is the exception that shows what the anchor costs. Its subgraph recall is higher than graph RAG recall (0.720 versus 0.596) because the nearest entity to the wording is often not the entity the question names, so the perspectives, spread over several regions, recover more of the named neighborhood. Its precision is lower (0.442 versus 0.615), and the two effects cancel into a statistical tie. The pattern follows the graphs. FB15K-237 and PrimeKG have high-degree anchors, WN18RR is small and sparse, and NELL-995 is noisy with 400 relations. Against the

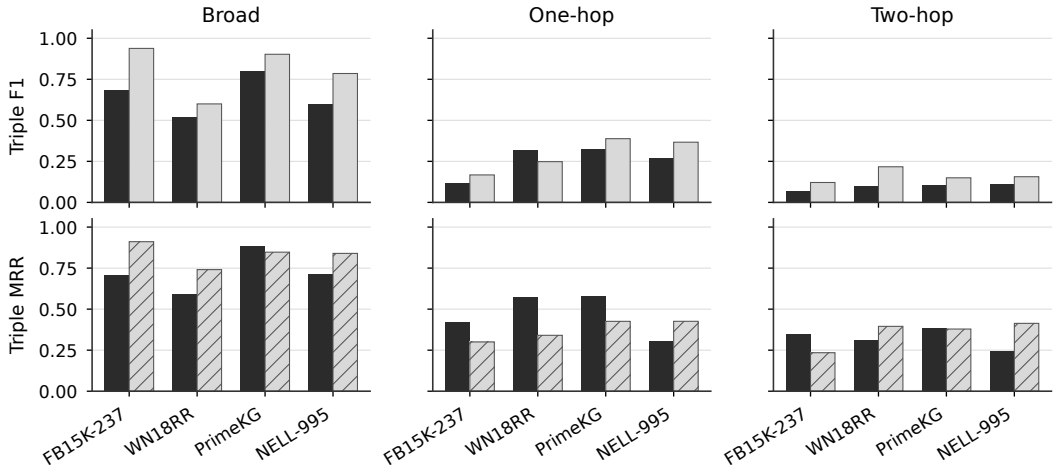

Figure 2: Within each panel, the dark bar is the mean for the prize-ranked subgraph and the light bar is the mean for the comparison—graph RAG in the top row, and every triple ranked by the best endpoint cosine to the raw question in the bottom row—over the 24 broad, 16 one-hop, or 16 two-hop held-out questions on that graph.

fixed budget the ordering reverses. Top-10 F1 is 0.160, 0.300, 0.204, and 0.290 on broad questions, against 0.683, 0.521, 0.798, and 0.598 for the subgraph.

On one-hop questions the same ordering holds at a lower level. Graph RAG F1 is 0.167, 0.248, 0.388, and 0.367, against subgraph F1 of 0.119, 0.318, 0.322, and 0.268. The paired interval excludes zero on FB15K-237 (-0.049 $[-0.090, -0.002]$), PrimeKG (-0.066 $[-0.147, -0.006]$), and NELL-995 (-0.099 $[-0.198, -0.017]$), and includes zero on WN18RR, where the subgraph point estimate is higher ($+0.069$ $[-0.069, 0.199]$). One-hop gold is the tails of one head and one relation. The nearest entity is often that head, so its incident triples already contain the gold triple, and the subgraph is a filter on a neighborhood that was already the right one. The fixed budget loses this family outright, with top-10 F1 between 0.018 and 0.111.

The ranking columns separate the two orderings. On one-hop questions the prize-ranked subgraph has the higher MRR on three graphs, 0.422 against 0.300 on FB15K-237, 0.570 against 0.341 on WN18RR, and 0.579 against 0.426 on PrimeKG, and it loses on NELL-995 (0.303 against 0.426). The paired gaps are $+0.122$, $+0.229$, $+0.154$, and -0.123, and every interval includes zero, so the ranking win is a point estimate rather than a resolved difference. Table 2 in the appendix gives Hits@$1/3/5/10$. Table 3 says the same thing per question, with 7, 10, 9, and 3 wins for the subgraph against 8, 5, 6, and 11 losses. This is the one family where the prize score orders a pile better than endpoint cosine does.

On two-hop questions neither method recovers the far endpoint. Subgraph F1 is 0.070, 0.100, 0.107, and 0.111. Graph RAG F1 is 0.121, 0.217, 0.150, and 0.156. The gap excludes zero on FB15K-237 and WN18RR, and subgraph entity F1 on this family is 0.045, 0.028, 0.013, and 0.021. No method in either table exceeds 0.22 on two-hop triple F1, and the bounded walk is the strongest control there on FB15K-237 (0.201) and NELL-995 (0.177). A first hop can raise triple F1 while entity F1 stays near zero, because the gold entities are the far endpoints.

The other controls, listed in Table 5, do not reverse that reading. At a matched entity budget the unfiltered neighborhood is at least as high as the subgraph on broad questions and on one-hop questions except WN18RR, where the subgraph is 0.318 and the matched neighborhood is 0.310. A region around the raw question, with no perspectives, is ahead of the subgraph on broad FB15K-237, WN18RR, and NELL-995 and behind on PrimeKG. The bounded walk leads the two-hop column on FB15K-237 (0.201) and NELL-995 (0.177).

5.3 Overlap of candidate sets

The family scores compare subgraphs with baselines. The question is whether the perspectives of one question retrieve different entities. Figure 5 in Appendix E counts how many of the 56 held-out questions return no candidate entity, exactly one, or more than one, once the candidate sets of every perspective have been pooled.

Summed across the perspectives of a question, retrieval returns a mean of 0.91 unique entities on FB15K-237, 2.34 on WN18RR, 1.50 on PrimeKG, and 1.70 on NELL-995. On 80% of FB15K-237 questions the candidate sets contain at most one entity, and on PrimeKG and NELL-995 the same holds for 64% and 57%. The counts include 16, 7, 1, and 11 questions per graph where the region returns nothing at all, so the mean mixes empty candidate sets with a handful of anchors.

The proposals themselves already lie close together. Pooled over the held-out questions, a kept perspective sits at a mean cosine of 0.60 to 0.66 with its question and 0.61 to 0.66 with the other kept perspectives of the same question. This is the anisotropy of text embeddings (Ethayarajh, 2019; Mu & Viswanath, 2018), and it has a concrete consequence here. Two perspectives of the same question land in overlapping regions when the embedding does not separate them, and on these questions it does not. The answer sets therefore share their entities. What they still separate is which facts are attributed to which perspective.

Table 4 in Appendix D reports the structure numbers, and Figure 3 plots them. The subgraphs hold a mean of 2.7 to 6.1 relations each, their candidate seeds sit at a mean graph distance of 1.2 to 1.9, and seed connectivity is 0.18 on PrimeKG and 0.57 on FB15K-237. Graph RAG is a star with no connectors and no seed-to-seed bridges, while 0.000 to 0.023 of the subgraph triples are bridges between two seeds. Silhouette (Rousseeuw, 1987) over the held-out questions with more than one non-empty set is 0.081 (FB15K-237, 27 questions), 0.085 (WN18RR, 28), 0.158 (PrimeKG, 35), and 0.171 (NELL-995, 10). The sets are distinct objects, and they are not well-separated clusters. That is the shape of the multi-perspective property in this setting.

5.4 Discussion

The measurements draw a consistent line. Reconstruction changes which triples are kept far more than it changes where retrieval looks, and the subgraph's advantage over a fixed list survives only while the list is large and unfiltered. Once the comparison holds the entity budget fixed, the filtering is what differs, and filtering a neighborhood is a different claim from retrieving a separate answer set per perspective. The prize order is the one place where the hypothesis carries on its own, and it carries on the one-hop family, where the gold is a small set of tails and the perspective term can lift those tails above unrelated triples that touch the anchor. The structure metrics say the same thing from the other side, the subgraph is a connected object but most of it is still the anchor neighborhood that graph RAG already had. A query and its legitimate readings are nearly parallel vectors in a generic text embedding, so a region drawn around one perspective overlaps the region drawn around the next, and the anchor closest to the question is the one every perspective returns.

That reading suggests where to push next. One direction is to train or adapt a space in which the perspectives of a question are pushed apart, or to ground each perspective as a graph constraint such as a type rather than as a point. A second is to let perspectives differ in resolution, since the taxonomy direction is the piece of the topology the embedding does distinguish. A third is to spend the budget on the anchored neighborhood when the graph has no separate evidence for a second perspective. These directions keep the split between the hypothesis and the evidence.

5.5 Limitations

Fifty-six held-out questions per graph is still a small sample, and a family estimate rests on 16 or 24 of them, so the intervals in Section 5 are wide and several comparisons are ties. Table 3 shows the thin margins directly, with 3, 17, 1, and 2 questions where the subgraph beats graph RAG on triple F1 across the four graphs, against 27, 26, 14, and 13 losses. The questions are constructed from the graphs rather than drawn from a public multi-answer benchmark, and a high score on a broad question can mean recovery of the named entity's neighborhood rather than a separation of perspectives. Broad gold is capped at 50 triples, which penalizes high-degree entities whose neighborhoods cannot be covered. The path questions were worded by a model that saw the relation text, which can

inflate triple F1 on those families. Think-on-Graph is a bounded reimplementation, so its numbers are a floor for that family of methods. The ranking control shares the pipeline's embedding and endpoint score, and the graph RAG pile has no ranking at all, so the MRR comparison is the subgraph against that one control. The structure numbers are averages over questions whose subgraphs differ in size and number. The synthesis numbers are lexical. They check whether names from the answer sets reach the paragraph, and they say nothing about whether the paragraph is correct.

6 Conclusion

A question is an incomplete description of the knowledge an answer needs. We reconstruct a candidate of that missing structure with a language model, then ask an external graph which parts of the reconstruction have facts. The model says where to look and the graph says what is there, and a perspective the graph cannot support stays empty. Across four knowledge graphs, the reconstruction makes the retrieved facts cleaner, and it does less for breadth: the perspectives of a question mostly land on the same anchor, because the embedding geometry does not separate them. The framework's value at this stage is the hypothesis it isolates, and the measurements that say what that hypothesis needs next. A representation in which the perspectives of a question are distinguishable would turn multi-perspective answer sets from a structure into a result.

Ethics statement

The questions are drawn from public knowledge graphs. We collect no human judgments and release no personal information. PrimeKG contains biomedical relations; this paper does not claim a clinical use. The written answer is instructed to stay inside the retrieved triples, but mention coverage and the disagreement check are lexical, and neither guarantees that the prose is safe to act on.

Reproducibility statement

Appendix A lists the 256 questions and Appendix B reproduces the perspective and answer prompts. The evaluation harness, the gold construction, and the scripts that produced every number and figure are included with the supplementary material. From `Experiment/`, the expanded gold sets are built by `python -m evaluation.expand_gold_sets`, and the main table is produced by `python -m evaluation.run_dataset_evaluation --dataset FB15K-237 --gold evaluation/gold/fb15k-237_gold_expanded.json --output evaluation/results-v4/fb15k-237 --calibrated-factor 10 --retrieval-factors 10 --subgraph-factors 10 --sweep-query-limit 0 --judge --answers-per-dataset 64 --think-on-graph`, repeated with the other three dataset names; the walk alone is `python -m evaluation.run_think_on_graph --dataset FB15K-237`. The figures and the tables in this paper are rebuilt from the saved JSON by `paper-v5/analysis/build_figures.py`. Graph RAG is the field saved as `nearest_entity`. Appendix H lists the embedded strings, the box distance, the disagreement rule, and the PrimeKG files. No identifying repository link appears in this PDF.

AI use statement

Generative AI tools were used to draft and edit the prose, to prepare the figures, to format the references, to search and summarize literature, to brainstorm the framing, and to help implement the taxonomy step, the prize-collecting grower, the Think-on-Graph walk, and the evaluation harness. The knowledge-topology formulation was developed with that assistance. The formulas describe the system that was run. No claim rests on an unproved theorem. The datasets are public graphs, and gold keys were read from those files. A model drafted the wording of the path questions, proposed the perspectives and taxonomy concepts, chose among the Think-on-Graph relation shortlists, and wrote the final paragraphs. No synthetic dataset was generated, no model was trained, and no human-subject data or translation was involved. We did not use AI to write proofs. Every number was checked against the saved evaluation files. The v5 tables and all figures were rebuilt from those files. The seed-42 questions were reused from the earlier run, and the seed-43 questions were run once for this revision. We take responsibility for the final content, including all claims made with AI assistance.

REFERENCES

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka, and Tom M. Mitchell. Toward an architecture for never-ending language learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2010.

Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10:67, 2023.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2D knowledge graph embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 55–65, 2019.

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

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*, 2020.

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

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, 2023.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. AmbigQA: A question answering benchmark with ambiguous questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 5700–5714, 2020.

Jiaqi Mu and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*, 2018.

Hongyu Ren, Weihua Hu, and Jure Leskovec. Query2box: Reasoning over knowledge graphs in vector space using box embeddings. In *International Conference on Learning Representations*, 2020.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M. Ni, Heung-Yeung Shum, and Jian Guo. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *International Conference on Learning Representations*, 2024.

Kristina Toutanova and Danqi Chen. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pp. 57–66, 2015.

Wenhan Xiong, Thien Hoang, and William Yang Wang. DeepPath: A reinforcement learning method for knowledge graph reasoning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 564–573, 2017.

A THE 256 QUESTIONS

Broad questions use the template and do not include a relation string. One-hop and two-hop questions were worded by a language model that saw the entity name and the relation text. That wording is a leak into a metric that matches the raw relation string. The family split is how a reader can see which questions carry it.

FB15K-237.

broad	What can you tell me about Sean Lennon?
broad	What can you tell me about Arnhem?
broad	What can you tell me about Cleopatra?
broad	What can you tell me about New Bedford?
broad	What can you tell me about David Fincher?
broad	What can you tell me about State University of New York at Purchase?
broad	What can you tell me about Jermaine Dupri?
broad	What can you tell me about Grammy Award for Best Classical Vocal Solo?
broad	What can you tell me about Stephen Harper?
broad	What can you tell me about Brian Geraghty?
broad	What can you tell me about Wayne's World?
broad	What can you tell me about Book of Shadows: Blair Witch 2?
broad	What can you tell me about The Phantom of the Opera?
broad	What can you tell me about Fair Game?
broad	What can you tell me about Evian Thonon Gaillard F.C.?
broad	What can you tell me about 61st Academy Awards?
one hop	What role does Joshua Redman have in a specific music track?
one hop	In which country was The Lady produced?
one hop	In which country was Miracle at St. Anna produced?
one hop	In which film does Adam Baldwin perform?
one hop	What major do graduates of The Queen's College study?
one hop	Who influenced John Cleese?
one hop	Who is the spouse of Emilio Estevez?
one hop	What is the musical score of Psycho?
two hop	What positions are held by players on the roster of FC Basel?
two hop	What measurement is Salisbury quantified by?
two hop	What areas are contained within Metro Manila?
two hop	What sport is S.C. Internacional associated with?
two hop	What profession does Warren Ellis have?
two hop	What language is spoken in Chicken Run?
two hop	What areas does Albuquerque contain?
two hop	What service language does HTC Corporation support?
broad	What can you tell me about Clemson University?
broad	What can you tell me about Marvin Hamlisch?
broad	What can you tell me about Wayne County?
broad	What can you tell me about chief operating officer?
broad	What can you tell me about Joan Chen?
broad	What can you tell me about Kenai Peninsula Borough?

broad	What can you tell me about The Wings of the Dove?
broad	What can you tell me about Kyoto?
broad	What can you tell me about Copenhagen?
broad	What can you tell me about Iris?
broad	What can you tell me about Jersey City?
broad	What can you tell me about Nelly Furtado?
broad	What can you tell me about Fast Food Nation?
broad	What can you tell me about drone music?
broad	What can you tell me about The Squid and the Whale?
broad	What can you tell me about Sicko?
one hop	What award was Bono nominated for?
one hop	What language is Face/Off produced in?
one hop	In which film is Manhattan featured as a location?
one hop	What genre does Training Day belong to?
one hop	What recognition has Courtney Simon received?
one hop	In which country was The Scorpion King produced?
one hop	What role does Brian Eno have in his musical contributions?
one hop	In which film does Carroll O'Connor perform?
two hop	What award was Stephen Goosson nominated for and what is its category?
two hop	What country is the Australian Capital Territory an administrative division of and what is its administrative area type?
two hop	What profession does Campbell Scott engage in?
two hop	What award was Mickey Rooney nominated for?
two hop	Whom did Marlene Dietrich date before being nominated for an award?
two hop	What award have graduates of Horace Mann School been nominated for?
two hop	What category does the location Oxford belong to?
two hop	What major do graduates of American University of Beirut typically complete?

WN18RR.

broad	What can you tell me about method?
broad	What can you tell me about grating?
broad	What can you tell me about association of southeast asian nations?
broad	What can you tell me about tincture?
broad	What can you tell me about tying up?
broad	What can you tell me about stimulate?
broad	What can you tell me about heckler?
broad	What can you tell me about secretion?
broad	What can you tell me about sclerosis?
broad	What can you tell me about space?
broad	What can you tell me about mask?
broad	What can you tell me about sicily?
broad	What can you tell me about mash?
broad	What can you tell me about elucidation?
broad	What can you tell me about predate?
broad	What can you tell me about laniidae?
one hop	What is a type of hacker?
one hop	What word is derived from stripe?
one hop	What word is derived from wonder?
one hop	What word is derived from revolutionize?
one hop	What word is derived from stay?
one hop	What is a type of singer?
one hop	What word is derived from premiss?
one hop	What word is derived from wait?
two hop	What word is derived from married couple?

two hop What is a type of range?
two hop What is a word derived from move?
two hop What is a more general concept that includes work party?
two hop What type of place is a smithy and what parts does it have?
two hop What is a broader category of attainment and what words derive from it?
two hop What is a word that derives from cooperative?
two hop What is a word derived from sandpaper and what parts does it consist of?
broad What can you tell me about ovum?
broad What can you tell me about kitchen utensil?
broad What can you tell me about motion?
broad What can you tell me about trailblazer?
broad What can you tell me about redact?
broad What can you tell me about solarize?
broad What can you tell me about encryption?
broad What can you tell me about origination?
broad What can you tell me about plait?
broad What can you tell me about barricade?
broad What can you tell me about ruffle?
broad What can you tell me about commemorative?
broad What can you tell me about pack?
broad What can you tell me about generation?
broad What can you tell me about political campaign?
broad What can you tell me about bust?
one hop What word is derived from emergence?
one hop What word is derived from antagonism?
one hop What word is derived from override?
one hop What is part of subclass magnoliidae?
one hop What word is derived from invitation?
one hop What word is derived from misgive?
one hop What word is derived from sound reflection?
one hop What word is derived from clasp?
two hop What type of word is derived from shelter?
two hop What word is derived from grimace?
two hop What is a derivative word of cone?
two hop What verbs are included in the verb group for catch?
two hop What is a derivative word of Glasgow?
two hop What broader category does New York City belong to?
two hop What parts does striated muscle have?
two hop What concepts are associated with the word redeem?

PrimeKG.

broad What can you tell me about Amiloride (a drug)?
broad What can you tell me about gastroparesis (disease) (a disease)?
broad What can you tell me about Micafungin (a drug)?
broad What can you tell me about incessant infant ventricular tachycardia (a disease)?
broad What can you tell me about Azosemide (a drug)?
broad What can you tell me about Nitric Oxide (a drug)?
broad What can you tell me about Prednisone acetate (a drug)?
broad What can you tell me about CCND1 (a gene/protein)?
broad What can you tell me about localized pagetoid reticulosis (a disease)?
broad What can you tell me about acquired purpura fulminans (a disease)?
broad What can you tell me about neurocirculatory asthenia (a disease)?
broad What can you tell me about Aminosalicylic acid (a drug)?
broad What can you tell me about Levomefolic acid (a drug)?

broad	What can you tell me about Ixekizumab (a drug)?
broad	What can you tell me about Azelnidipine (a drug)?
broad	What can you tell me about Methyclothiazide (a drug)?
one hop	What are the medical contraindications for Acrivastine?
one hop	What does Azapetine indicate?
one hop	What are the medical contraindications for Lanreotide?
one hop	Which drugs interact with Ibutilide?
one hop	What diseases is PARP1 associated with?
one hop	Which drugs interact with Midomafetamine?
one hop	What diseases are related to leiomyoma?
one hop	Which drugs interact with Etoricoxib?
two hop	How does AICA ribonucleotide interact with other substances and what are its contraindications?
two hop	What are the medical contraindications associated with nasal obstruction?
two hop	What is an off-label use of adrenomyeloneuropathy?
two hop	What medical contraindication should be considered when using Carbetocin?
two hop	What medical contraindication is associated with congenital heart disease?
two hop	What diseases are related to Rifaximin?
two hop	What medical contraindication should be noted for Lambert-Eaton myasthenic syndrome?
two hop	What medical contraindication should be considered when using Ridogrel?
broad	What can you tell me about familial nephrotic syndrome (a disease)?
broad	What can you tell me about suppurative cholangitis (a disease)?
broad	What can you tell me about Iron sucrose (a drug)?
broad	What can you tell me about Choline salicylate (a drug)?
broad	What can you tell me about Seletracetam (a drug)?
broad	What can you tell me about Levetiracetam (a drug)?
broad	What can you tell me about eye disease (a disease)?
broad	What can you tell me about Gilbert syndrome (a disease)?
broad	What can you tell me about arthropathy (a disease)?
broad	What can you tell me about SH2B3 (a gene/protein)?
broad	What can you tell me about qualitative platelet defect (a disease)?
broad	What can you tell me about Pranlukast (a drug)?
broad	What can you tell me about thyroid crisis (disease) (a disease)?
broad	What can you tell me about adult idiopathic neutropenia (a disease)?
broad	What can you tell me about CIITA (a gene/protein)?
broad	What can you tell me about Vardenafil (a drug)?
one hop	What does Sorafenib indicate?
one hop	How does Lemborexant interact with other drugs?
one hop	What is the effect of chest pain?
one hop	What is an off-label use for Chlorpropamide?
one hop	What does Acyclovir indicate?
one hop	What does relapsing fever indicate?
one hop	What does uterine corpus cancer indicate?
one hop	What is an off-label use for head and neck cancer?
two hop	What does Moclobemide indicate and what disease is it associated with?
two hop	What diseases are related to behavioral abnormality?
two hop	What diseases are indicated by Tasonermin?
two hop	What diseases is NR1I3 associated with?
two hop	What medical contraindications are indicated by Halazepam?
two hop	What diseases is TNFSF11 associated with?
two hop	What drugs interact with familial recurrent peripheral facial palsy?
two hop	What medical contraindications are associated with islet cell adenomatosis?

NELL-995.

broad	What can you tell me about community physicians?
broad	What can you tell me about military personnel?
broad	What can you tell me about ground floor?
broad	What can you tell me about initial meeting?
broad	What can you tell me about collins001?
broad	What can you tell me about fernando tatis?
broad	What can you tell me about bluebirds?
broad	What can you tell me about conditional probability?
broad	What can you tell me about forward?
broad	What can you tell me about carlos carrasco?
broad	What can you tell me about john lecarre?
broad	What can you tell me about burkina?
broad	What can you tell me about ross ade stadium?
broad	What can you tell me about cat?
broad	What can you tell me about cheboygan?
broad	What can you tell me about water chestnuts?
one hop	What date is associated with Friday?
one hop	What type of room can a stateroom be found in?
one hop	What type of artery is the ductus arteriosus?
one hop	What is represented by the concept of old?
one hop	Which river does Merrymeeting Bay empty into?
one hop	What event occurs at Northern Ireland?
one hop	Who created Dr. Seuss?
one hop	What does a hyena prey on?
two hop	Who is Andrew's husband?
two hop	What citizenship does Eric have?
two hop	What is something that is a multiple of an elephant?
two hop	What does UT represent in its geographical location?
two hop	What organization has an office in Fairview Village?
two hop	Who is the spouse of Keturah?
two hop	Which team achieved recognition with the 2004 Stanley Cup?
two hop	When was Chelsea dissolved as an organization?
broad	What can you tell me about bank atlantic center?
broad	What can you tell me about advocates?
broad	What can you tell me about kfph?
broad	What can you tell me about ktvb tv?
broad	What can you tell me about rhel?
broad	What can you tell me about tom gage?
broad	What can you tell me about the chronicles of narnia?
broad	What can you tell me about watermelon?
broad	What can you tell me about lawyers?
broad	What can you tell me about cedar rapids?
broad	What can you tell me about separate sleeping area?
broad	What can you tell me about newport news hampton wmsburg?
broad	What can you tell me about newman?
broad	What can you tell me about richard notebaert?
broad	What can you tell me about east tennessee state lady buccaneers?
broad	What can you tell me about computer engineering?
one hop	What do the geographic coordinates 42.8405400000000,-91.7996100000000 represent?
one hop	Which organization is headquartered in Oklahoma City?
one hop	What type of profession is faculty?
one hop	What plants are included in the category of acacia?
one hop	Who does Georgetown employ?

one hop	What items can be found in a luxurious room?
one hop	What type of chemical are fats classified as?
one hop	What entity is represented by home?
two hop	Which sports team is led by Kevin Durant?
two hop	What organization are the Charleston Southern Buccaneers a subpart of, and with whom do they collaborate?
two hop	Who wrote the book The Drowned World?
two hop	Which agent worked on the development of Clarinex?
two hop	In which country does RBC operate?
two hop	When does the submission occur and what does it represent?
two hop	What office does W hold?
two hop	Who does Ed Bouchette compete with in his field?

B PROMPTS

The perspective and answer prompts below are reproduced as they were used, including the original spelling. The taxonomy prompt asks for a parent category and a child concept for each kept perspective, and requires the perspective name to be copied exactly; its two strings are embedded together into the taxonomy direction of Section 4. The Think-on-Graph prompt shows the question, the current entity, and the shortlist of relations, and requires the model to pick from that list only.

Perspective prompt.

```
# ROLE
You are a nuanced thinker always trying to come up with different perspectives
to look at a question.

# TASK
- Generate 5 perspectives for a given query, the perspectives should be different
  ways the question can be worded to find more fine grained information on a
  certain sub topic.
- The perspectives will be embedded into an embedding space to retreive canidate
  entities to answer the initial question, so the perspective should still be
  semantically simmilar to what the question is asking about.
- You MUST respond using a JSON object of perspectives.
- Your response should be the entire JSON object with multiple perspectives
- Each perspective object has only "name" and "description". Never include
  embeddings or other fields.
- If there are current perspectives include them in your response and they should
  count toward the total of 5 different perspectives.

# ACCEPTED PERSPECTIVES
{perspectives}

# QUERY
{query}

# EXAMPLE RESPONSE
{"perspectives": [{"name": "...", "description": "..."}]}
```

Answer prompt.

```
# ROLE
You are a careful synthesizer always trying to fold several perspectives into
one coherent answer without inventing facts.

# TASK
```

Table 2: Triple ranking on the 56 held-out questions per graph. The subgraph triples are ranked by the prize score. The control ranks every triple in the graph by the best endpoint cosine to the raw question. The graph RAG pile is unordered, because each of its triples touches the same nearest entity, so it has no ranks to report. Held-out slice as in Table 1.

		Subgraph, prize ranked					Control, endpoint-cosine ranked				
Graph	Family	MRR	H@1	H@3	H@5	H@10	MRR	H@1	H@3	H@5	H@10
FB15K-237	Broad	0.708	0.667	0.750	0.750	0.750	0.911	0.875	0.958	0.958	0.958
	One-hop	0.422	0.312	0.500	0.562	0.562	0.300	0.188	0.312	0.375	0.438
	Two-hop	0.346	0.188	0.500	0.688	0.688	0.234	0.125	0.250	0.250	0.625
WN18RR	Broad	0.588	0.500	0.708	0.708	0.750	0.741	0.667	0.792	0.833	0.875
	One-hop	0.570	0.375	0.750	0.875	0.875	0.341	0.125	0.500	0.625	0.812
	Two-hop	0.311	0.188	0.375	0.500	0.500	0.395	0.188	0.562	0.688	0.812
PrimeKG	Broad	0.885	0.792	0.958	1.000	1.000	0.847	0.750	1.000	1.000	1.000
	One-hop	0.579	0.500	0.688	0.688	0.688	0.426	0.188	0.562	0.688	0.875
	Two-hop	0.385	0.250	0.375	0.500	0.812	0.379	0.062	0.625	0.750	1.000
NELL-995	Broad	0.714	0.708	0.708	0.708	0.750	0.840	0.792	0.917	0.917	0.917
	One-hop	0.303	0.250	0.312	0.312	0.500	0.426	0.250	0.562	0.625	0.750
	Two-hop	0.245	0.125	0.312	0.500	0.500	0.413	0.188	0.562	0.750	0.812

```
- You will be given a user query and several answer sets.
- Each answer set is a perspective: a name, why this angle was chosen, and a
  list of facts as knowledge graph triples (head, relation, tail).
- Answer the query using ONLY those facts. Do not invent triples, entities, or
  relations that are not listed.
- Do not fill gaps with outside knowledge. If the facts are thin, keep the
  answer thin.
- Incorporate every perspective that has facts into one coherent answer. Do not
  write a separate mini-answer per perspective.
- Skip any answer set that has no facts. Do not mention that it was missing.
- If facts conflict across perspectives, present both views. Do not pick a
  winner.
- Do not mention embeddings, vectors, or the retrieval pipeline.
- Respond with plain prose only. No JSON, no markdown headings, no bullet lists
  of triples.

# QUERY
{query}

# ANSWER SETS
{answer_sets}
```

C TWO-HOP GOLD, ONE QUESTION

For "What positions are held by players on the roster of FC Basel?" the gold entity set is $\{\text{midfielder}, \text{forward}, \text{goalkeeper}\}$. The gold triple set contains both hops, including FC Basel $\rightarrow$ defender and defender $\rightarrow$ midfielder. Entity F1 scores the endpoints; triple F1 scores every listed triple. A retriever that returns only the first hop is credited with a positive triple hit and misses every entity, which is how a two-hop triple F1 can be positive while entity F1 is zero.

Table 3: Per-question wins, ties, and losses over the 56 held-out questions per graph. The subgraph wins a question when its value is higher. F1 compares the subgraph with graph RAG. MRR compares the subgraph with the endpoint-cosine ranked control.

	Triple F1 vs graph RAG			Triple MRR vs ranked control		
Graph	Win	Tie	Loss	Win	Tie	Loss
FB15K-237	3	26	27	15	18	23
WN18RR	17	13	26	20	14	22
PrimeKG	3	28	25	17	22	17
NELL-995	6	24	26	11	18	27

Table 4: Subgraph structure on the 56 held-out questions per graph. Relations, entities, and connectors are per subgraph. Seeds are the candidate entities of a question, pooled over its subgraphs. Hop distance is the mean graph distance between candidate entities inside one subgraph. Connectivity is one minus the fraction of candidate components that would have to be merged to make the seeds connected. Bridge share is the fraction of subgraph triples that join two candidate entities. Entity cosine is the mean pairwise cosine of the entities in a subgraph.

Graph	Rel/sub	Ent/sub	Con/sub	Seeds/q	Hop	Conn	Bridge	Cos
FB15K-237	5.13	4.76	0.04	2.18	1.30	0.57	0.00	0.26
WN18RR	2.74	2.78	0.04	4.12	1.24	0.42	0.02	0.49
PrimeKG	6.08	5.27	0.04	4.43	1.40	0.18	0.00	0.39
NELL-995	3.09	1.95	0.09	3.05	1.90	0.25	0.00	0.41

D Additional results

E Candidate sets

The relevance filter is not what concentrates the candidate sets. Removing it and searching the perspectives it had dropped changes the mean number of unique entities from 0.91 to 1.00 on FB15K-237, from 2.34 to 3.11 on WN18RR, leaves PrimeKG unchanged (1.50 to 1.50), and moves NELL-995 from 1.70 to 2.45. Macro triple F1 over the same questions moves by $+0.011$, -0.018, 0.000, and -0.011. The perspectives the filter rejected were pointing at the same parts of the graph as the ones it kept.

On the written answer, a retrieved entity name appears in 36%, 71%, 66%, and 59% of cases across the four graphs. When two subgraphs disagree, names from both sides survive in 56 of 66, 48 of 56, 61 of 69, and 20 of 26 cases. These lexical checks say whether names from the answer sets reach the paragraph. They do not measure whether the paragraph is correct.

F Median subgraph

G Other retrieval controls

H What was run

This appendix is the part a reader needs in order to rebuild the numbers. The prompts that were actually sent are in Appendix B, spelling included. Those strings were not corrected, the seed-42 questions were not rerun, and the seed-43 questions were run once for this revision.

An entity vector is the embedding of `{name}. {ai_description}`. A relation vector is the same pattern, after a model has written `ai_description` for the raw predicate. FB15K-237 entity vectors are the ones already stored in `entity2info_updated.json`. The other three graphs were embedded by the loader from that string. Perspectives are embedded from the same pattern on the perspective name and description. The taxonomy direction is one embedding of the parent string and the child string together. Chat calls do not set a temperature, so they use the API default, and

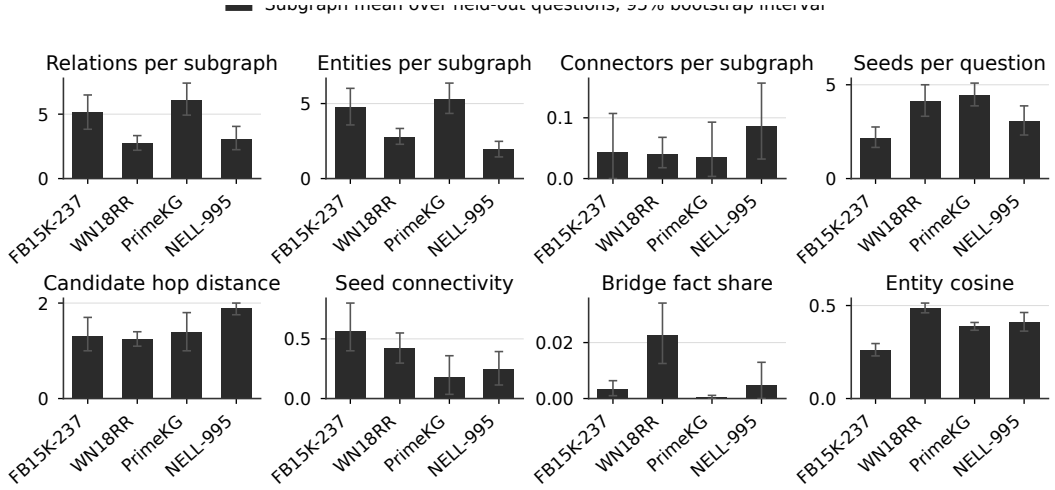

Figure 3: Structure of the retrieved evidence, mean over the held-out questions with a 95% bootstrap interval. Relations, entities, and connectors are counted per subgraph. Seeds are the candidate entities of a question. Hop distance and connectivity describe the candidate entities inside one subgraph. Bridge share is the fraction of triples that join two candidate entities. Entity cosine is the mean pairwise cosine of the entities in a subgraph. The graph RAG pile is a star, so its connector and bridge counts are zero.

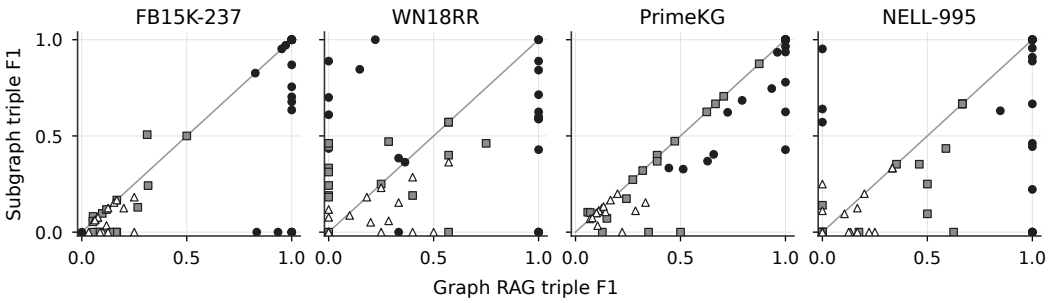

Figure 4: Per-question comparison of subgraph triple F1 against graph RAG triple F1, one dot per held-out question. A dot above the diagonal is a question where the subgraph scored higher. Dots sit on the axes when a method retrieves nothing that is gold.

each call keeps a single response. If the taxonomy JSON does not parse, the parent and child stay empty and the call is not retried. The perspective is still kept.

The box test is the outside distance from Query2Box, plus a residual term. Along each of the four axes the outside piece is $\max(0, |\text{coord} - \text{center}| - \text{width})$. The residual piece is $\max(0, \text{residual_norm} - \text{residual_width})$. The distance is the sum of those pieces. An entity is inside when the sum is zero. Widths start at the 90th percentile of the 64 nearest entities and then grow or shrink as Section 4 describes.

R_q is part of the object defined in Section 3. This run does not build it. The prize score uses the question, the perspective, and the taxonomy direction.

Silhouette is computed on entity embeddings. The label of an entity is the answer set that received its subgraph. Questions with fewer than two non-empty sets are left out. Two subgraphs disagree when each has an entity name of at least three characters that the other does not. The disagreement is counted as kept when the paragraph contains at least one name from each exclusive side. That is a lexical check. It is not a judgment of the paragraph.

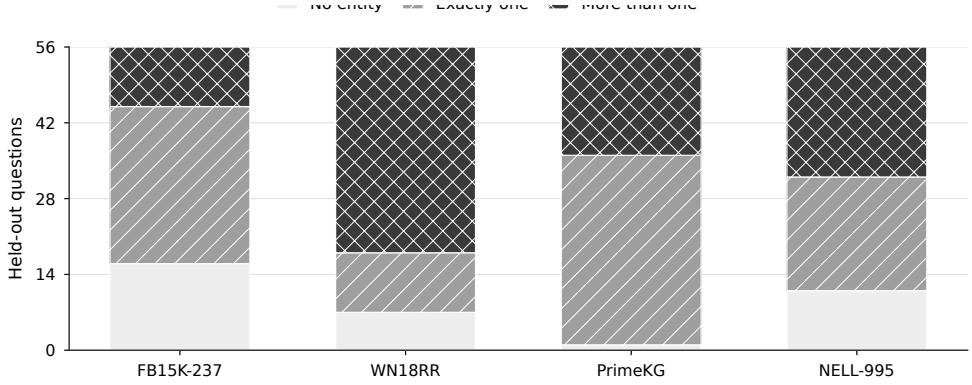

Figure 5: Unique candidate entities per held-out question, pooled across the perspectives of that question. There are 56 questions per graph. A question with several perspectives still usually returns one entity or none, so the answer sets are not yet distinct evidence.

Table 5: Triple F1 for the other retrieval controls on the same held-out questions. Top-10 keeps the triples on the ten nearest entities. Matched budget keeps that same unfiltered neighborhood at the entity count the region returned. Raw region is one box around the question with no perspectives. ToG is the bounded Think-on-Graph walk.

Graph	Family	Top-10	Matched	Raw region	ToG
FB15K-237	Broad	0.160	0.883	0.754	0.160
	One-hop	0.018	0.152	0.095	0.215
	Two-hop	0.007	0.107	0.057	0.201
WN18RR	Broad	0.300	0.648	0.590	0.258
	One-hop	0.102	0.310	0.190	0.190
	Two-hop	0.062	0.180	0.098	0.154
PrimeKG	Broad	0.204	0.822	0.775	0.103
	One-hop	0.054	0.379	0.288	0.133
	Two-hop	0.019	0.106	0.031	0.132
NELL-995	Broad	0.290	0.786	0.657	0.343
	One-hop	0.111	0.318	0.234	0.122
	Two-hop	0.035	0.162	0.044	0.177

PrimeKG in this paper is the local split, not the full public dump. Entities come from `entities.dict` and triples from `train.txt`, `valid.txt`, and `test.txt`. That copy has 10,344 entities and these 11 relation types: `disease_phenotype_positive`, `drug_effect`, `indication`, `contraindication`, `disease_disease`, `disease_protein`, `disease_phenotype_negative`, `exposure_disease`, `drug_protein`, `off-label use`, and `drug_drug`.

Graph RAG is not a separate retrieval run. It is the `nearest_entity` triple set already stored with each question: every triple that touches the entity with the highest cosine to the raw question. The figures and all four tables are written by `paper-v5/analysis/build_figures.py` from `Experiment/evaluation/results-v4`, after dropping the first eight questions of each graph. Those eight were used to choose the entity cap.