End-to-End Graph Retrieval Pipeline for Specialized Domains

Haraldur Bjarni Davíðsson, Hazar Harmouch
Vrije Universiteit Amsterdam, University of Amsterdam
Amsterdam,The Netherlands
haraldur.davidsson@student.uva.nl, H.Harmouch@uva.nl

Abstract

We present an end-to-end pipeline for constructing a domain-specific knowledge graph from instructional text using Large Language Model assisted extraction. Applied to the Icelandic Riding Levels, a 602 pages training corpus for riders of the Icelandic Horse, the pipeline produces a hyper-relational knowledge graph of 9,382 nodes and 16,423 edges, where schema-constrained qualifiers preserve the conditional and procedural context that standard triples discard. To evaluate the resulting graph, we introduce, to our knowledge, the first expert validated question answering benchmark for this domain: 252 questions across four reasoning categories. Comparing Graph-, Text-, and Hybrid-retrieval augmented generation methods, we find that Text-based achieves the highest overall mean judge score, but that Graph-based provides the only correct answer for a small subset of queries, particularly where the corpus contains competing values for the same fact. A failure analysis traces the majority of Graph-based retrieval errors to context dilution at high-degree hub nodes. We discuss implications for adaptive retrieval strategies that route queries to the appropriate modality as results points to Graph-RAG potentially serving rather as a complementary and query specific rather than a broader replacement to general Text-RAG.

Keywords: knowledge graph construction, retrieval-augmented generation, hyper-relational knowledge representation, domain-specific QA, Graph-RAG

1. Introduction

Large Language Models (LLMs) often struggle in specialized, low-resource domains where factual grounding and terminological precision are essential. Icelandic Horse training is a representative example: its training methods, gait mechanics, and biomechanics differ significantly from mainstream equestrian disciplines (Davíðsson et al., 2023), yet digital resources are scarce, inconsistent, and fragmented across languages. As a result, foundation models are prone to hallucinations and incorrect domain-specific assumptions when applied to specialized or low-resource domains without external grounding (Ji et al., 2023; Gao et al., 2024; Mallen et al., 2023).

Retrieval-Augmented Generation (RAG) addresses these limitations by grounding model outputs in external evidence. Most RAG systems rely on dense vector retrieval over unstructured text, representing knowledge as flat collections of independently embedded chunks (Karpukhin et al., 2020). More recently, Knowledge Graph (KG) based RAG (Graph-RAG) has emerged as a family of approaches that incorporate structured relational knowledge to support multi-hop reasoning over explicit entity and relationship links (Edge et al., 2024; Zhu et al., 2025; He et al., 2024). However, constructing such graphs raises challenges related to quality, hyper-relational representation, and scalability, particularly when domain-specific entities and relations are sparse or noisy (Hogan et al., 2021). While recent work has demonstrated the feasibility of LLM-based KG construction at scale, there is limited empirical evidence on how automatically constructed domain-specific KGs perform as retrieval sources. Crucially, under what conditions structured retrieval provides complementary advantages over unstructured text (Gao et al., 2024).

In this work, and in the context of use within the HorseDay mobile application [1], we present an end-to-end LLM-assisted pipeline for constructing a hyper-relational KG from the Icelandic Riding Levels: The standardized training guide for riders of the Icelandic Horse §3.1. We evaluate the resulting KG as a retrieval source within a Graph-RAG architecture and compare it against Text-RAG and Hybrid-RAG across a question answering (QA) benchmark of 252 expert-validated question-answer pairs spanning four reasoning categories; lookup, causal, aggregation and multi-hop. Our findings show that while Text-RAG achieves higher overall mean judge score, the automatically constructed KG provides complementary retrieval advantages for entity-centric queries. The findings also suggest that a mere naive concatenation (HybridRAG) is not sufficient to achieve theoretical complementary benefits. A detailed failure analysis reveals that the primary bottlenecks are addressable limitations in graph construction and traversal. Our contributions are the following:

1. The **first KG for the domain of Icelandic horse training** with the respective LLM-assisted hyper-relational extraction pipeline.

[1] https://www.horseday.com/

Proceedings of the Knowledge Graphs and Large Language Models Workshop 2026 (KG-LLM@LREC26), pages 155–165
May 16, 2026.

2. The **first expert validated QA benchmark for this domain**, consisting of 252 question-answer pairs across four reasoning categories.

3. **An empirical analysis** comparing Text-RAG, Graph-RAG, and Hybrid-RAG, with a detailed failure analysis.

2. Related Work

2.1. LLM-Based Knowledge Graph Construction

Traditional KG construction pipelines such as NELL (Carlson et al., 2010) and DeepDive (Zhang et al., 2017) relied on statistical co-occurrence and large amounts of labeled data, making them impractical for low-resource or closed-corpus domains where specific instructions may appear only once. The advent of LLMs shifted this paradigm toward zero-shot and few-shot extraction, enabling models to identify entities and relations directly from unstructured text without task-specific training data (Wadhwa et al., 2023). KGGen (Mo et al., 2025) formalized this into a scalable pipeline that extracts subject-predicate-object triples and then clusters semantically similar entities to reduce graph sparsity which is a critical step for ensuring graph connectivity. Similarly, SAC-KG (Chen et al., 2024) demonstrated that LLMs can serve as skilled automatic constructors for domain-specific KGs. However, these approaches produce standard binary triples, which struggle to capture the conditional logic and safety constraints inherent in instructional text (Zhang et al., 2020). For instance, representing "apply leg pressure only if the horse falls in" requires reification or auxiliary nodes that fragment the graph and complicate retrieval. Our pipeline extends KGGen's extract-then-resolve approach by introducing hyper-relational triples with schema-constrained qualifiers by attaching attributes such as *condition*, *intensity*, and *modality* directly to edges, thereby preserving instructional context within a single retrieval unit.

2.2. Graph Retrieval Paradigms

Once a KG has been constructed, retrieving relevant subgraphs to augment LLM context is itself a non-trivial challenge (Peng et al., 2024). Existing approaches can be broadly grouped into three paradigms: (1) *Global summarization* methods, exemplified by Microsoft's GraphRAG (Edge et al., 2024), cluster graph communities and generate LLM-based summaries at varying levels of abstraction, enabling broad thematic queries but sacrificing granularity for entity-specific retrieval; (2) *Dual-level retrieval* methods, such as LightRAG (Guo et al., 2024), bypass expensive graph traversal by embedding entity and relation profiles into a vector store and retrieving at both low-level (entity-specific) and high-level (thematic) granularities, achieving significant cost reductions while maintaining competitive accuracy; (3) *Local traversal* methods, including HippoRAG (Gutiérrez et al., 2024) and RoG (Luo et al., 2024), identify seed entities from the query and expand outward via Personalized PageRank or agent-based path exploration to construct targeted subgraphs. Our work follows the local traversal paradigm, using Personalized PageRank to expand from query-activated seed entities as is done in HippoRAG and RoG, with an addition of an inverse-degree normalization term, which mitigates the problem high-degree hub node domination mitigation, analogous to inverse document frequency in text retrieval (Adamic and Adar, 2003). We opt for this approach due to its zero-shot applicability: unlike global methods that require pre-computed community structures or dual-level methods that assume large LLMs ($\geq$32B parameters) for quality extraction, local traversal operates directly over the graph with no additional indexing or model requirements.

2.3. Knowledge Representation for Instructional Domains

Standard Knowledge Graph triples of the form $\langle$*subject, predicate, object*$\rangle$ are well-suited for factual assertions but struggle to capture the conditional and procedural logic that characterizes instructional text (Zhang et al., 2020). In equestrian training, for example, the instruction "apply leg pressure only if the horse falls inward on the circle" carries a safety-critical condition; representing it as $\langle$*rider, apply, leg pressure*$\rangle$ discards the governing constraint entirely. Hyper-relational knowledge representations address this limitation by attaching qualifier key-value pairs directly to edges, preserving context such as conditions, intensity levels, and modality within a single retrieval unit (Rosso et al., 2020). This is particularly important for retrieval: without qualifiers, graph traversal returns context-stripped instructions, the action without its governing constraints, which can lead to misleading or even dangerous answers in safety-sensitive domains. Our schema defines seven qualifier types (*condition*, *causality*, *instruction*, *intensity*, *spatial*, *frequency*, and *modality*) with controlled vocabularies for modality values (Mandatory, Prohibited, Danger, Ideal, Mistake, Fact), enabling fine-grained filtering during retrieval. To our knowledge, this is the first application of hyper-relational KG construction to an instructional domain using LLM-based extraction.

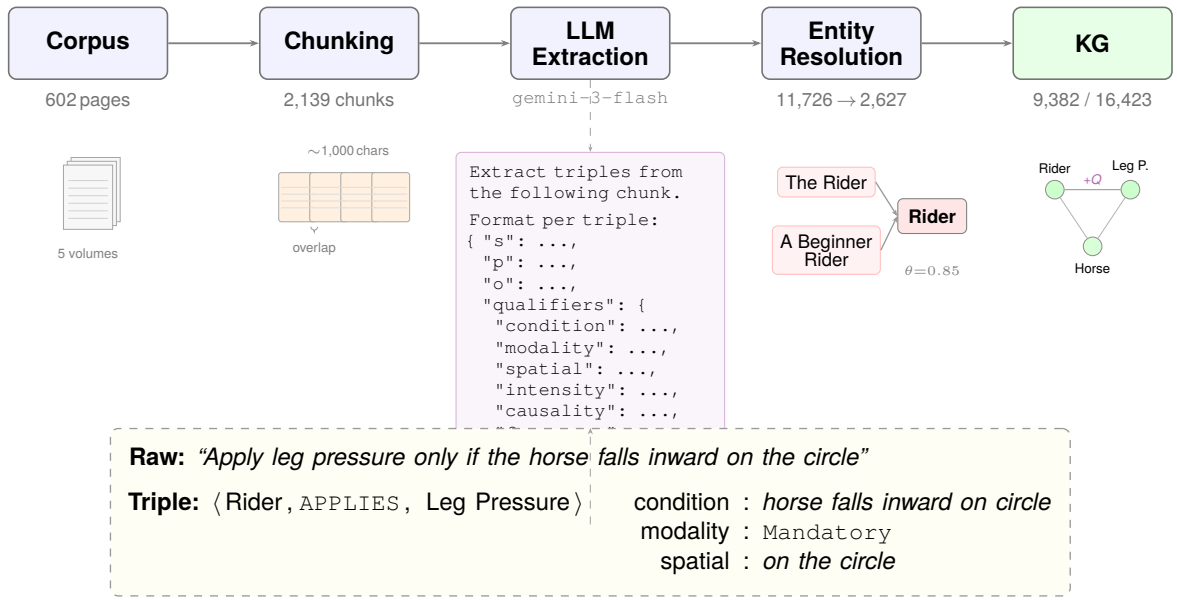


Figure 1: KG construction pipeline. The callout illustrates how a raw instruction is transformed into a hyper-relational triple with schema-constrained qualifiers.

2.4. Evaluating Graph-RAG in Practice

Despite the growing number of Graph-RAG architectures, systematic empirical comparisons against text-based baselines remain scarce (Gao et al., 2024). Han et al. (Han et al., 2025) provide one of the first controlled evaluations, finding that the benefit of graph-based retrieval varies significantly with question type and corpus structure. Similarly, recent work by Hong et al. (Xiang et al., 2025) explicitly asks *when* graphs help in RAG, concluding that graph retrieval provides the largest gains for queries requiring multi-hop reasoning over connected entities, but can underperform text retrieval on simple factoid questions. However, both studies evaluate on general-domain benchmarks derived from Wikipedia or news corpora, where entity redundancy and broad web coverage provide a safety net that is absent in specialized domains. Low-resource, single-corpus settings, where retrieval must operate over a closed and non-redundant knowledge base, remain largely unexplored. Our work addresses this gap with the first domain-specific comparison of Text-RAG, Graph-RAG, and Hybrid-RAG in an instructional setting, with a fine-grained failure analysis that traces Graph-RAG errors to specific structural causes rather than reporting aggregate judge scores alone.

3. Methodology

We start by corpus preprocessing and KG construction as shown in Figure 1. We then evaluate several retrieval architecture design.

3.1. Corpus: The Icelandic Riding Levels

The source corpus consists of the *Icelandic Riding Levels*, the standardized training guide for riders of the Icelandic Horse[2]. The five volumes comprise 602 pages. While the raw PDF files are approximately 1.6 GB due to extensive high-resolution photographic content, the extracted raw text yields $\sim$ 2.1 MB. Furthermore, although the domain relies heavily on Icelandic terminology (e.g., Tölt), the prose itself is entirely in English. The manuals were not previously available in digital text format and were parsed specifically for this work. Therefore, we assume this knowledge was not present in any foundation model's training data (more on this in §6). The corpus was segmented into 2,139 chunks of approximately 1,000 characters with a 200-character overlap to preserve contextual continuity, as shown in Figure 1.

3.2. Hyper-Relational Schema Design

As previously mentioned in Section 2.3, standard binary triples, which are commonly represented in the format of ⟨*subject, predicate, object*⟩, or ⟨*s,p,o*⟩, often discard conditional and procedural context critical to instructional text. Recent work on hyper-relational knowledge graphs demonstrates that attaching qualifier key–value pairs to edges improves semantic precision and reasoning over constrained facts (Rosso et al., 2020; Panda et al., 2024; Ding et al., 2024). We therefore adopt a hyper-relational representation of the form $\langle s, p, o, \mathcal{Q} \rangle$, where $\mathcal{Q}$ is a set of qualifier key-value pairs attached directly

[2] http://knapamerki.is

to each edge. We define seven qualifier types:

Table 1: Qualifier types and their semantic roles.

Qualifier	Domain	Example
condition	Context	"If horse rushes"
causality	Mechanism	"To relax the jaw"
instruction	Technique	"Vibrate the hand"
intensity	Magnitude	Force/speed modifiers
spatial	Topology	"Behind the girth"
frequency	Rate	"Every 3 strides"
modality	Safety	Mandatory / Prohibited / Danger / Ideal / Mistake / Fact

The **modality** qualifier is constrained to a discrete set of six values (Table 1), enabling downstream retrieval to perform safety-aware filtering, for example, prioritizing edges marked `Danger` when answering safety-related queries. This schema design follows the principles of shape-based validation (Knublauch and Kontokostas, 2017) and aligns with findings that schema-guided LLM extraction reduces structural hallucinations in zero-shot settings (Wei et al., 2023).

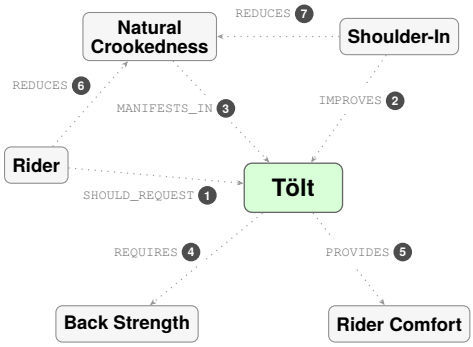
3.3. LLM-Based Extraction and Entity Resolution

Our extraction pipeline, shown in Figure 1, follows the general approach of KGGen (Mo et al., 2025), adapted to produce hyper-relational triples via a custom schema-constrained prompt. For each of the 2,139 chunks, `gemini-3-flash` was prompted to generate triples adhering to the schema defined above, with temperature $t{=}0$ for deterministic output. The model was chosen for its high throughput and cost-efficiency at scale.

Processing chapters independently introduces entity fragmentation: the same concept may appear under different surface forms (e.g., *"The Rider"* vs. *"A Beginner Rider"*), creating disconnected subgraphs that hinder multi-hop reasoning (Paulheim, 2017), which is addressed via a two-stage resolution process:

1. **Clustering:** Entity embeddings were generated using SentenceTransformers (Reimers and Gurevych, 2019) and clustered via agglomerative clustering (average linkage, cosine distance, threshold $= 0.85$). This reduced 11,726 raw entities to 2,627 canonical nodes. The threshold was selected based on manual inspection of 50 randomly sampled clusters.

2. **LLM Adjudication:** Ambiguous clusters were submitted to an LLM that distinguishes between synonymy (merge) and taxonomy (link as instance), preserving the instructional hierarchy required for safety-critical distinctions.



	Modality	Qualifier	Value
1	Mandatory	condition	*at first training*
		intensity	*calmly*
		frequency	*only a few steps*
2	Ideal	condition	*Icelandic horses*
		causality	*best way to improve*
		instruction	*dressage exercises*
3	Fact	condition	*at difficult speed*
		causality	*lack of coordination*
		spatial	*uneven tempo*
4	Fact	condition	*harder than walk/trot/canter*
		causality	*demands strength*
5	Fact	causality	*one leg always on ground*
		intensity	*motionless in saddle*
		spatial	*body of rider*
6	Mandatory	instruction	*make horse even between sides*
		intensity	*gradually*
7	Ideal	causality	*makes horse straighter*

Figure 2: Curated subgraph around the *Tölt* gait node. Edge labels show the predicate; qualifier annotations (grey boxes) preserve the conditional and procedural context that standard binary triples would discard. Modality tags: **Mandatory**, **Ideal**, **Fact**.

The resulting KG contains **9,382 nodes** and **16,423 hyper-relational edges**. Edge modality classification reflects the prescriptive nature of the domain: Facts (39.0%), Ideal practices (33.0%), Mandatory rules (18.4%), and Danger warnings (4.2%). Figure 2 illustrates the qualifier structure for a representative subgraph around the Tölt gait node.

3.4. Retrieval Architectures

We implement three retrieval strategies to evaluate the utility of the constructed KG:

Text-RAG. A hybrid retrieval pipeline combining dense retrieval (`multilingual-e5-base`) and sparse retrieval (BM25), fused via Reciprocal Rank Fusion (Cormack et al., 2009). Candidates are reranked by a cross-encoder (`ms-marco-MiniLM-L-6-v2`) to produce the final context (Thakur et al., 2021).

Graph-RAG. A local traversal approach operating over the hyper-relational KG. Seed entities are identified via the same hybrid retrieval index used in Text-RAG, then expanded using Personalized PageRank (α=0.85) with hub penalization. High-degree nodes (e.g., "Horse", "Rider") are downweighted by normalizing PPR scores by $\ln(\text{Degree}(n) + 2)$, analogous to inverse document frequency in text retrieval (Adamic and Adar, 2003). Retrieved triples are serialized into natural language for LLM consumption.

Hybrid-RAG. Combines both retrieval modalities with a dynamic ratio: k text chunks are paired with $10k$ graph triples to account for the token density disparity between text chunks (∼150 tokens) and graph triples (∼15 tokens).

3.5. Evaluation Setup

QA Benchmark. We developed the first expert-validated QA benchmark for the Icelandic Horse domain, consisting of 252 question-answer pairs across four categories designed to test distinct retrieval capabilities (Table 2).

Category	Count	Tests
Lookup	66	Single-fact retrieval
Aggregation	56	Multi-item enumeration
Causal	101	Cause–effect reasoning
Multi-hop	29	Cross-chunk inference
Total	**252**	

Table 2: QA benchmark distribution by question category.

Questions were generated using a decoupled strategy: input contexts were synthesized from keyword-linked chunks across disparate sections to mitigate gold-context bias (Yang et al., 2018). Two domain experts (certified Icelandic Horse trainers) reviewed all pairs, resulting in 85 questions (33.7%) being refined or rephrased.

LLM-as-a-Judge. Responses were evaluated using a dual-judge consensus mechanism to mitigate self-preference bias (Zheng et al., 2023). Two architecturally distinct models, namely `Llama-3.1-70B-Instruct` and `Qwen2.5-72B-Instruct`, independently scored each response on a 5-point Likert scale using a fact-based Chain-of-Thought protocol inspired by G-Eval (Liu et al., 2023). Final scores are computed as the arithmetic mean. Inter-judge agreement was strong: Pearson $r = 0.88$, Cohen's weighted $\kappa = 0.84$.

Models. We evaluate five LLMs spanning different scales: `Llama-3.1-8B`, `Llama-3.1-70B`, `GPT-OSS-20B`, `GPT-OSS-120B`, and `Gemini-2.5-Pro`. For Text-RAG, $k \in \{1, 2, 3, 5, 7, 10, 12, 15\}$; for Graph-RAG, $k \in \{10, 15, 20, 30, 50, 70, 90\}$; for Hybrid-RAG, k text chunks are combined with $10\times k$ graph triples.

Evaluation asymmetry. A methodological caveat applies to all results that follow. Because the QA pairs were derived from the source text, even with the decoupled generation strategy described above, the reference answers are grounded in the same chunk representation that Text-RAG searches. Graph-RAG, by contrast, must recover equivalent information through a lossy extraction pipeline. This creates a structural advantage for passage-level retrieval. The results should therefore be read as a conservative estimate of Graph-RAG's utility: cases where Graph-RAG is uniquely best emerge *despite* evaluation conditions that favour text retrieval by design.

4. Results

We evaluate the three retrieval architectures using `GPT-OSS-120B` as the primary generation model, selected for achieving the highest overall mean judge score across configurations. All results report the dual-judge consensus score (§3.5) at the judge-score-optimal retrieval depth k. The patterns reported below were consistent across all five models, with only the non-RAG baseline variant of `Gemini-2.5-Pro` outperforming its non-RAG counterpart of `GPT-OSS-120B`.

4.1. Overall Performance

Table 3 reports mean judge score for each retrieval method alongside the no-retrieval baseline.

Text-RAG achieves the highest overall mean judge score (0.881), a 22.3 percentage-point (pp) improvement over the no-retrieval baseline. Graph-RAG improves modestly over the baseline (+2.3 pp), while Hybrid-RAG falls slightly below Text-RAG despite combining both retrieval modalities. This overall ranking, however, masks important variation across question types.

Hybrid-RAG's slightly lower mean judge score suggests that combining modalities can introduce competing signals: when both text passages and graph triples are present in the context, the LLM tends to default to the more frequently mentioned value from the text, effectively overriding the graph's more precise retrieval.

4.2. Mean Judge Score by Question Type

To understand under which conditions structured retrieval contributes, we disaggregate mean judge

Method	Mean Judge Score	k^*
Baseline	0.658	—
Graph-RAG	0.681	50
Hybrid-RAG	0.862	5
Text-RAG	0.881	12

Table 3: Overall mean judge score by retrieval method (GPT-OSS-120B). k^* denotes the retrieval depth with highest mean judge score.

score across the four reasoning categories in our benchmark (Table 4).

Category	Base.	Graph	Text	Hybrid
Lookup (66)	0.60	0.70	0.82	0.80
Causal (101)	0.73	0.71	0.91	0.89
Multi-Hop (29)	0.66	0.69	0.89	0.87
Aggregation (56)	0.60	0.59	0.91	0.89

Table 4: Mean judge score by question category (n in parentheses). Graph = Graph-RAG, Text = Text-RAG, Hybrid = Hybrid-RAG, Base. = no retrieval.

Two patterns are noteworthy. First, Graph-RAG provides its largest gains on **Lookup** questions (+10 pp over baseline), where entity-centric retrieval can surface precise facts from the graph. Second, and contrary to our initial hypothesis, Graph-RAG does *not* outperform Text-RAG on Multi-Hop questions despite the theoretical advantage of explicit relational links for chaining facts. We return to this finding in §4.5.

4.3. Oracle Method Selection

To quantify the complementarity between retrieval methods, we conduct an oracle analysis: for each question, we select the method that achieves the highest score. Table 5 reports the results.

Configuration	Judge Score	Δ vs. Text-RAG
Baseline (no retrieval)	0.658	−22.3 pp
Graph-RAG only	0.681	−19.9 pp
Hybrid-RAG	0.862	−1.9 pp
Text-RAG (best single)	0.881	—
Oracle (best per Q)	0.913	+3.2 pp

Table 5: Oracle analysis: mean judge score when selecting the best-performing method per question, compared with single-method configurations.

The oracle achieves 0.913—a 3.2 pp gain over the best single method. This gap confirms that Graph-RAG can capture information that text retrieval misses, but only for a subset of queries. Disaggregating by method, Graph-RAG is the uniquely best method (i.e., it outperforms both Text-RAG and Hybrid-RAG) for 9.5% of all questions. This proportion rises to 15% for Lookup questions but drops below 5% for Causal and Multi-Hop categories. These numbers suggest that a lightweight query router, or an agentic retrieval pipeline that decomposes queries and selectively engages structured retrieval for suitable subquestions, could capture most of the oracle's gain.

4.4. When Graph-RAG Wins

We present three representative cases from the 9.5% of questions where Graph-RAG is the best-performing method, selected to illustrate the structural conditions under which the KG provides a retrieval advantage.

Case 1: Competing numerical values—resting pulse rate. *Query:* "What is the normal resting pulse rate for a horse?" *Target:* 36–44 beats per minute. The corpus mentions heart rate in multiple contexts: a general range (20–40 bpm) appears across several passages on basic horse care, while the specific clinical resting value (36–44 bpm) appears in a single passage on healthy horse indicators. Text-RAG consistently retrieved the more frequently mentioned range across all retrieval depths, producing the answer "20–40 beats per minute" (best score: 0.60). Graph-RAG retrieved the triple ⟨Healthy Horse State, defined_by, *Pulse 36–44 bpm*⟩ with qualifier {condition: "Resting"}, correctly surfacing the precise target value from $k{=}10$ onward (score: 1.00). The graph's entity-centric indexing isolated the specific clinical fact from the surrounding noise of related-but-imprecise passages.

Case 2: Competing numerical values—noseband fit. *Query:* "How much space should remain between the nose and the strap when a noseband is correctly tightened?" *Target:* At least two fingers' width. This case exhibits the same retrieval-noise mechanism as Case 1. The corpus discusses noseband fit in multiple sections: one passage specifies a two-finger gap as the standard, while others mention a 3–4 finger distance in the context of noseband positioning relative to the nostrils (a related but distinct measurement). Text-RAG retrieved the more frequently occurring value, consistently answering "3–4 finger widths" across all k values (best score: 0.60). Graph-RAG at $k{=}10$ retrieved ⟨Noseband, should_be_fastened, *Noseband Strap*⟩ with qualifiers {instruction: "Two fingers can fit easily underneath the nose strap", intensity: "Loose enough for two fingers"}, producing the correct answer (score: 1.00). The KG's structured storage disambiguated between two similar-sounding measurements that text retrieval conflated: the correct triple is separated from the confounding ⟨Noseband Strap,

RECOMMENDED_POSITION, *Nostrils*⟩ with {instruction: "Maintain a distance of 3–4 finger widths"} by virtue of distinct entity pairs and relation types.

Case 3: Causal precision—calming exercise caution. *Query:* "Why must a trainer exercise caution when frequently repeating calming exercises with a young horse during the initial stages of training?" *Target:* Young horses must learn to 'think forward' from the beginning; overdoing calming exercises can hinder this mindset. This case demonstrates that Graph-RAG's advantages extend beyond Lookup questions when the KG captures specific causal relationships. Text-RAG at low k values produced generic answers about boredom and loss of interest (score: 0.50), only reaching the specific "forward-thinking" concept at k=7 (score: 0.90). Graph-RAG at k=10 already retrieved the relationship ⟨FORWARD THINKING, REQUIRED_DEVELOPMENT_FOR, *Young Horse*⟩ with qualifier {condition: "Right from the beginning of training", modality: "Mandatory"}, producing a precise causal answer (score: 0.90). At k=15, Graph-RAG scored 1.00. This grounding enabled Graph-RAG to produce a more specific answer than Text-RAG, even though the full caution still had to be inferred rather than recovered from a single explicitly causal edge

Common pattern. Cases 1 and 2 share a structural characteristic: the corpus contains **competing values for the same or closely related attributes**, and passage-level retrieval tends to favor the more frequently mentioned value, whereas the KG's entity- and relation-centric structure helps isolate the intended one. Case 3 extends this pattern to **structured dependencies**: when information relevant to the answer is encoded as an explicit relationship in the graph, graph retrieval can surface it more directly, rather than requiring the LLM to infer it from loosely related passages.

4.5. Failure Analysis

To complement the success cases, we examined all instances where Text-RAG substantially outperformed Graph-RAG ($\Delta > 0.4$, $n = 60$). Table 6 categorizes these failures by root cause.

Mode	Root Cause	n	%
Context dilution	Hub nodes spread attention across irrelevant edges	45	75.0
Insufficient detail	Lossy compression during extraction	10	16.7
Missing procedure	Temporal sequence lost in graph structure	4	6.7
Missing numerics	Extraction discarded literal values	1	1.7

Table 6: Root causes for Graph-RAG failures ($\Delta > 0.4$ vs. Text-RAG).

Context Dilution (75%). The dominant failure mode is algorithmic rather than representational: the relevant information existed in the graph but was not retrieved. Entity resolution, while necessary for graph connectivity, created high-degree hub nodes (e.g., "Horse": 800+ edges, "Rider": 600+ edges). When a query activated such a node, Personalized PageRank distributed probability mass across hundreds of edges, diluting the semantically relevant subset. For example, given the query *"How do I calm a tense horse?"*, the traversal activated the HORSE node but failed to prioritize tension-related edges over unrelated edges about feeding, grooming, or breeding. We implemented a hub-penalization mechanism, which mitigated, but did not eliminate this effect. This finding directly explains the unexpected Multi-Hop result (§4.2): multi-hop queries necessarily traverse hub nodes, and each hop amplifies the dilution effect.

Insufficient Detail (16.7%). Strict schema enforcement sometimes compressed nuanced instructions into generic attribute values. For instance, a detailed biomechanical explanation of seat adjustment was reduced to {instruction: "adjust seat position"}, losing the specificity present in the source text. This contrasts with the success cases, where the schema happened to preserve the critical distinguishing detail suggesting that extraction quality is the primary determinant of Graph-RAG utility.

Missing Procedure (6.7%). The schema lacks an intrinsic mechanism for temporal ordering. When source text described a sequence, such as the four-beat footfall pattern of tölt, the extraction parsed steps into independent attributes without explicit ordering.

Missing Numerics (1.7%). A small number of cases involved the extraction discarding literal values (e.g., "250g per front leg" became "add weight"). This failure mode is the inverse of Cases 1–2: when extraction *does* preserve numerical values, the KG excels; when it discards them, performance degrades.

4.6. Summary

The results suggest that automatically constructed hyper-relational KGs provide **complementary but not superior** retrieval for domain-specific QA. Graph-RAG's advantages concentrate on queries where the corpus contains competing or ambiguous values for the same attribute—the KG's entity-centric structure disambiguates what passage-level

retrieval conflates (§4.4). However, 75% of Graph-RAG errors stem from context dilution at hub nodes, an algorithmic limitation in graph traversal rather than a fundamental shortcoming of structured retrieval (§4.5). The oracle analysis (§4.3) quantifies this complementarity: perfect routing yields a 3.2 pp gain over Text-RAG, with Graph-RAG contributing uniquely on 9.5% of queries, predominantly entity-attribute lookups where a lightweight query router could capture most of the headroom.

5. Conclusion and Future Work

We presented an end-to-end pipeline for constructing a hyper-relational knowledge graph from instructional text, along with the first expert-validated QA benchmark for the Icelandic Horse training domain. The pipeline produces a KG where schema-constrained qualifiers preserve the conditional and procedural context that standard triples discard, and the benchmark provides 252 questions across four reasoning categories for evaluating retrieval in this low-resource setting.

Our empirical comparison shows that Text-RAG achieves the highest overall mean judge score. Though this finding should be interpreted in light of an inherent evaluation asymmetry, since the benchmark is derived from the same text that Text-RAG searches. Graph-RAG provides the only correct answer for a small but identifiable subset of queries, while Hybrid-RAG demonstrates that naively combining both modalities is counterproductive. A failure analysis traces the majority of Graph-RAG errors to context dilution at hub nodes, an algorithmic limitation in graph traversal rather than a fundamental shortcoming of structured representation.

These results shift the practical question from *which* retrieval modality to use to *when* each modality should be used. The oracle analysis shows that the information captured by each method is complementary, but realizing this complementarity requires selective routing whether through lightweight query classifiers or through agentic pipelines that decompose queries and engage structured retrieval only where passage-level ambiguity is likely. Improving Graph-RAG's standalone performance will additionally require edge-aware traversal that incorporates semantic similarity during graph walks, and extraction pipelines that more reliably preserve distinguishing details such as numerical values and temporal sequences.

6. Limitations

A primary limitation of this work is the evaluation asymmetry introduced by gold-context bias. The QA benchmark was derived from the same textual corpus that Text-RAG retrieves from, creating a structural advantage for passage-level retrieval. Although expert validation and question rephrasing mitigate direct leakage, the benchmark remains grounded in the original chunk representation. A fully unbiased comparison would require independently constructed gold answers and broader expert involvement, which entails substantial manual effort and domain expertise.

Second, the quality of the constructed knowledge graph is inherently dependent on LLM-based extraction. As reflected in our failure analysis, lossy compression, missing numerical values, and the absence of explicit temporal ordering can degrade Graph-RAG performance. While schema constraints reduce structural hallucinations, they may also oversimplify nuanced instructional content. Improving extraction fidelity remains a critical direction for future work. To mitigate the impact of lossy extraction, future work could explore integrating Free-Text Knowledge Graphs (Zhao et al., 2020), which preserve raw text at the nodes, or applying approximate graph querying techniques (Ren et al., 2024) to handle edges that are imperfectly extracted from the source material.

Third, the traversal strategy relies on Personalized PageRank over a moderately sized graph (9,382 nodes; 16,423 edges). The extent to which hub-penalized traversal scales to larger instructional corpora remains untested. In denser graphs, hub-node dilution effects may intensify, potentially requiring more sophisticated edge-aware or query-conditioned traversal mechanisms.

Fourth, our evaluation focuses on end-to-end downstream QA performance rather than intermediate KG quality. We do not provide a direct assessment of extraction fidelity, nor do we present ablation studies isolating the exact contribution of hyper-relational qualifiers versus standard binary triples. Similarly, our entity resolution relies on a static cosine distance threshold (0.85), which may not be uniformly calibrated across the entire embedding space. Finally, our benchmark assumes answers exist in the corpus; evaluating how the system handles unanswerable queries or abstentions remains for future work.

Finally, the pipeline was evaluated on English-language instructional text. Although the corpus contains Icelandic terminology (e.g., t"olt, sniðgangur), all prose is in English. Generalization to morphologically richer or low-resource languages may introduce additional challenges in entity resolution and qualifier extraction. Moreover, we cannot definitively guarantee that portions, or the entirety of the corpus were absent from foundation model pretraining data, though the manuals were not previously available in machine-readable form, the results strongly support this claim, however, this cannot be known for certain.

Acknowledgements

We extend our sincere gratitude to HorseDay ehf. for funding the time of the first author and supporting this work. We also thank the authors of the Riding Levels for granting us permission to use their content in this paper.

Ethics and Data Availability

The source corpus consists of commercially published training manuals that were digitized with explicit permission from the authors of the Icelandic Riding Levels. The material contains instructional content related to horse training but does not include personally identifiable information, private records, or sensitive data.

Because the domain involves safety-sensitive and animal-welfare-relevant instructions (e.g., biomechanical guidance, training constraints, and danger warnings), we acknowledge that errors introduced during LLM-based extraction or graph traversal could lead to misleading or incomplete answers. As documented in our failure analysis, lossy compression, missing numerics, or hub-node dilution may affect retrieval precision. The resulting system is therefore intended strictly as a research prototype and should not be interpreted as professional veterinary or training advice.

Our pipeline relies on LLM-assisted extraction and retrieval-augmented generation, which are known to be susceptible to hallucination and factual inconsistency in the absence of grounding. While our closed-corpus design mitigates this risk by restricting retrieval to the source manuals, we cannot guarantee that extraction errors or model biases are fully eliminated.

While the raw text and images of the commercial manuals cannot be publicly distributed due to copyright, they are available upon reasonable request, contact the authors for inquiries. The code, including the implementation of the construction and the retrieval pipelines, as well as the QA benchmark dataset, is available on GitHub [3] to ensure reproducibility of the retrieval and LLM-as-a-judge experiments and experimentation for other domain specific fields. We encourage responsible use and caution against deploying derived systems in real-world training contexts without expert validation.

7. Bibliographical References

Lada A. Adamic and Eytan Adar. 2003. Friends and neighbors on the Web. *Social Networks*, 25(3):211–230.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam Hruschka, and Tom Mitchell. 2010. Toward an architecture for never-ending language learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24.

Hanzhu Chen, Xu Shen, Qitan Lv, Jie Wang, Xiaoqi Ni, and Jieping Ye. 2024. SAC-KG: Exploiting Large Language Models as Skilled Automatic Constructors for Domain Knowledge Graphs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4345–4360, Bangkok, Thailand. Association for Computational Linguistics.

Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759.

Haraldur B. Davíðsson, Torben Rees, Marta Rut Ólafsdóttir, and Hafsteinn Einarsson. 2023. Efficient Development of Gait Classification Models for Five-Gaited Horses Based on Mobile Phone Sensors. *Animals*, 13(1).

Zifeng Ding, Jingcheng Wu, Jingpei Wu, Yan Xia, Bo Xiong, and Volker Tresp. 2024. Temporal Fact Reasoning over Hyper-Relational Knowledge Graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics.

Darren Edge, Jeff Larson, Corey White, Shubham Singh, Lee Gunderson, Kyle Williams, Nick Bryan, and Philip J. Guo. 2024. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *arXiv preprint arXiv:2404.16130*.

Yunfan Gao, Yue Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3563–3578.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2024. LightRAG: Simple and Fast Retrieval-Augmented Generation. *arXiv preprint arXiv:2410.05779*.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. HippoRAG: Neurobiologically Inspired Long-Term

[3] https://github.com/haraldurbjarni/domain-graph-rag

Memory for Large Language Models. In *Advances in Neural Information Processing Systems*, volume 37.

Haoyu Han, Harry Shomer, Yu Wang, Yongjia Lei, Kai Guo, Zhigang Hua, Bo Long, Hui Liu, and Jiliang Tang. 2025. RAG vs.\ GraphRAG: A Systematic Evaluation and Key Insights. *arXiv preprint arXiv:2502.11371*.

Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh V Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. 2024. G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and Question Answering. In *Advances in Neural Information Processing Systems*, volume 37.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. Knowledge Graphs. *ACM Comput. Surv.*, 54(4). Place: New York, NY, USA.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Eric Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):1–38.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Holger Knublauch and Dimitris Kontokostas. 2017. Shapes constraint language (SHACL). W3C Recommendation, W3C Recommendation.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. 2024. Reasoning on Graphs: Faithful and Interpretable Large Language Model Reasoning. In *The Twelfth International Conference on Learning Representations*.

Alex Mallen, Akari Asai, Zexuan Zhong, Victor Chen, Adam Teichert, Yunfan Yang, and Graham Neubig. 2023. When Not to Trust Language Models: Investigating Hallucinations of Knowledge in LLMs. In *Proceedings of ACL*.

Belinda Mo, Kyssen Yu, Joshua Kazdan, Joan Cabezas, Proud Mpala, Lisa Yu, Chris Cundy, Charilaos Kanatsoulis, and Sanmi Koyejo. 2025. KGGen: Extracting Knowledge Graphs from Plain Text with Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Pranoy Panda, Ankush Agarwal, Chaitanya Devaguptapu, Manohar Kaul, and Prathosh Ap. 2024. HOLMES: Hyper-Relational Knowledge Graphs for Multi-hop Question Answering using LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 13263–13282. Association for Computational Linguistics.

Heiko Paulheim. 2017. Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods. *Semantic Web*, 8(3):489–508.

Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. 2024. Graph Retrieval-Augmented Generation: A Survey. *arXiv preprint arXiv:2408.08921*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Hongyu Ren, Mikhail Galkin, Zhaocheng Zhu, Jure Leskovec, and Michael Cochez. 2024. Neural Graph Reasoning: A Survey on Complex Logical Query Answering. *Transactions on Machine Learning Research*.

Paolo Rosso, Dingqi Yang, and Philippe Cudré-Mauroux. 2020. Beyond Triplets: Hyper-Relational Knowledge Graph Embedding for Link Prediction. In *Proceedings of The Web Conference 2020 (WWW)*, pages 1885–1896. ACM.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Somin Wadhwa, Silvio Amir, and Byron Wallace. 2023. Revisiting Relation Extraction in the era of Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, and others. 2023. Zero-Shot Information Extraction via Chatting with ChatGPT. *arXiv preprint arXiv:2302.10205*.

Zhishang Xiang, Chuanjie Wu, Qinggang Zhang, Shengyuan Chen, Zijin Hong, Xiao Huang, and Jinsong Su. 2025. When to use Graphs in RAG: A Comprehensive Analysis for Graph Retrieval-Augmented Generation. *arXiv preprint arXiv:2506.05690*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Ce Zhang, Jaeho Shin, Theodoros Rekatsinas, Michael Cafarella, Feng Niu, Alexander Shkapsky, Shanchieh Jay Wang, Ce Wu, Jason Zhang, and Christopher Ré. 2017. DeepDive: Declarative Knowledge Base Construction. In *Proceedings of the VLDB Endowment*, volume 10, pages 1310–1321.

Li Zhang, Qing Lyu, and Chris Callison-Burch. 2020. Reasoning about Goals, Steps, and Temporal Dependencies with WikiHow. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Chen Zhao, Chenyan Xiong, Xin Qian, and Jordan Boyd-Graber. 2020. Complex Factoid Question Answering with a Free-Text Knowledge Graph. In *Proceedings of The Web Conference 2020*, pages 1205–1216.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36.

Yucheng Zhu, Liang Wang, Wenhao Yu, Jifan Chen, and Weizhu Wang. 2025. Knowledge Graph-Guided Retrieval Augmented Generation. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.