

GROUND-TRUTH SUBGRAPHS FOR BETTER TRAINING AND EVALUATION OF KNOWLEDGE GRAPH AUGMENTED LLMs

Anonymous authors

Paper under double-blind review

ABSTRACT

Retrieval of information from graph-structured knowledge bases represents a promising direction for improving the factuality of LLMs. While various solutions have been proposed, a comparison of methods is difficult due to the lack of challenging QA datasets with ground-truth targets for graph retrieval. We present SynthKGQA, a LLM-powered framework for generating high-quality Knowledge Graph Question Answering datasets from any Knowledge Graph, providing the full set of ground-truth facts in the KG to reason over questions. We show how, in addition to enabling more informative benchmarking of KG retrievers, the data produced with SynthKGQA also allows us to train better models. We apply SynthKGQA to Wikidata to generate GTSQA, a new dataset designed to test zero-shot generalization abilities of KG retrievers with respect to unseen graph structures and relation types, and benchmark popular solutions for KG-augmented LLMs on it.

1 INTRODUCTION

Despite significant advances over the years, Large Language Models (LLMs) are still unreliable when asked to provide factual information, as hallucinations (LLMs outputting plausible but wrong answers) remain one of the central problems for LLM applications (Huang et al., 2025). The predominant solution to improving LLM trustworthiness is Retrieval Augmented Generation (RAG), where information pertinent to the query is retrieved from a corpus of knowledge and added to the prompt (Lewis et al., 2020; Borgeaud et al., 2022; Izacard et al., 2023). While RAG traditionally retrieves information from documents, another important use case is retrieval from graph structured repositories such as Knowledge Graphs (KGs). KGs encode facts as (subject, predicate, object) triples and are a highly efficient solution to store *relational* information while also being easier to maintain, update, and fact-check compared to textual documents.

KG-augmented LLMs are evaluated on Knowledge Graph Question Answering (KGQA) benchmarks, where supporting facts need to be extracted from the KG and combined to produce the answer. While many KGQA datasets have appeared over the years (see Peng et al. (2024) for a recent survey), little attention has been paid to their quality, reliability and limitations. The concurrent study by Zhang et al. (2025) has estimated the degree of factual correctness of the questions in widely-used benchmarks, such as WebQSP (Yih et al., 2016), CWQ (Talmor & Berant, 2018) and GrailQA (Gu et al., 2021), to be only between 30 and 60%. Moreover, several benchmarks are no longer challenging for SOTA KG-augmented LLMs and are now close to being saturated, limiting their usefulness. Even for datasets like ComplexWebQuestions (CWQ), which include multi-hop questions, it is impossible to provide an actual measure of question complexity due to the lack of *ground-truth answer subgraphs*, the golden targets for retrieval. This also means that KG-RAG retrievers cannot be evaluated on their own, but only end-to-end, by looking at the final answer provided by the LLM – which, since different solutions use different LLMs and prompting schemes, results in noisy comparisons. Moreover, for KG retrievers that require training, the ground-truth answer subgraph is indispensable to provide supervision signal: using existing datasets, it has to be approximated with the set of shortest paths from seed entities to answer nodes. As we show in Section 6, this is often a bad approximation, which penalizes the final quality of the retriever. Furthermore, most KGQA benchmarks are from the previous decade and based on the now discontinued Freebase KG (Bollacker et al., 2008), thus containing outdated facts and answers that might be in conflict with the more recent information

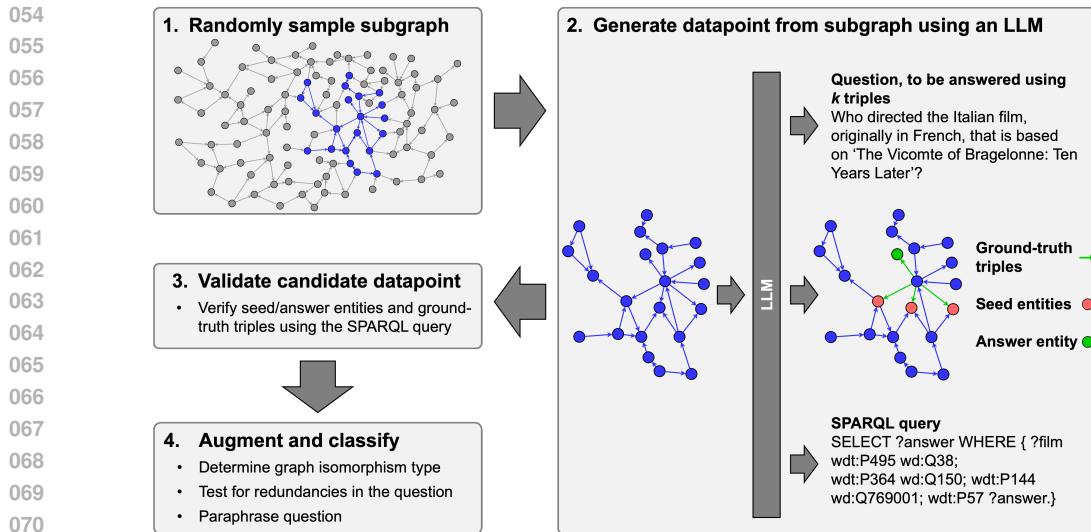


Figure 1: The steps performed by SynthKGQA to generate and validate questions and ground-truth subgraphs from an arbitrary Knowledge Graph.

available to LLMs. Moreover, these datasets have likely been seen during pretraining of recent LLMs (Ding et al., 2024), leading to data leakage. Question variety, in terms of KG relation and entity types used for the grounding facts, is also typically limited, as questions are either manually (through crowd-sourcing) or procedurally (through logical query manipulation) generated from predefined query templates, often resulting in unnatural, or ambiguously phrased, natural language questions (see Zhang et al. (2025) for a more detailed discussion of limitations and pitfalls).

We propose to tackle all these challenges by leveraging the advanced abilities of frontier LLMs to remove the need for a human in the loop in the creation of KGQA datasets. Our contributions are:

- SynthKGQA¹, a new framework (Figure 1) for generating large synthetic KGQA datasets from any KG, which provides high-quality, diverse questions with procedurally-verified ground-truth answer subgraphs and SPARQL queries, allowing to easily update the dataset whenever the underlying KG is modified.
- *Ground-Truth Subgraphs for Question Answering* (GTSQA)¹, a challenging new dataset with 32,099 questions spanning 27 different structures for the ground-truth answer subgraph, grounded in the regularly-updated Wikidata KG (Vrandečić & Krötzsch, 2014), and specifically designed to test generalization abilities of KG-RAG models.
- A comprehensive benchmark of SOTA LLMs and KG-augmented LLMs on GTSQA, providing new insights on retrieval abilities and limitations of these models on different structures for the answer subgraph, and different levels of zero-shot generalization.
- An extensive analysis of the benefits of using the ground-truth answer subgraph, instead of shortest paths from seed to answer nodes, as supervision signal for training KG retrievers.

2 RELATED WORK

Traditionally, models for KGQA have been evaluated on datasets built using a combination of handcrafted question and SPARQL templates and human annotation, often from the Freebase KG (Bollacker et al., 2008). WebQuestions (Berant et al., 2013) used autocompletions from the Google Suggest API on Freebase entity labels, with answers provided by human annotators – a process resulting prevalently in 1-hop questions. Many datasets are derived from WebQuestions, by filtering and refining the dataset (WebQSP (Yih et al., 2016)) or extending/combining the SPARQL queries to obtain more complex questions (CWQ Talmor & Berant (2018)). More structured, and partially

¹See supplementary material; links to github repository and dataset to be added after double-blind review.

108 automated, pipelines started to appear with LC-QuAD (Trivedi et al., 2017), mapping logical queries
 109 to natural language using question templates and crowd-sourced paraphrasing. Moreover, after
 110 Freebase was discontinued in 2016, new datasets shifted to using different KGs, mainly DBpedia
 111 (e.g., QALD-9 (Usbeck et al., 2018)) and Wikidata (e.g., LC-QuAD 2.0 (Dubey et al., 2019)), with
 112 parallel efforts to migrate to these KGs previous datasets based on Freebase (Azmy et al., 2018).
 113 New interest developed also in testing generalization abilities of KGQA models (GrailQA (Gu
 114 et al., 2021)) and better quantifying the complexity of questions (GrailQA++ (Dutt et al., 2023)).
 115 Only very recently Dammu et al. (2025) and Zhang et al. (2025) have started exploring the use of
 116 LLMs in the construction of KGQA datasets. Despite the implementation of validation pipelines to
 117 detect and discard bad data, these approaches can still suffer from hallucinations leading to incorrect
 118 question-answer pairs (Zhang et al., 2025). Also, crucial information such as the complete set of
 119 entities mentioned in each question, or fine-grained measures of question complexity or generalization
 120 abilities required to answer, are not fully tracked. A detailed comparison of these LLM-enabled
 121 approaches to the framework proposed in this paper can be found in Appendix E.
 122

3 THE SYNTHKGQA FRAMEWORK

124 **Preliminaries** KGs store facts as labeled directed edges between *entities* in a set \mathcal{E} (the nodes/ver-
 125 tices of the KG), where edge labels are drawn from a set \mathcal{R} of predicates, or *relation types*. Thus, an
 126 edge is represented as a subject-predicate-object triple (h, r, t) , with $h, t \in \mathcal{E}$ and $r \in \mathcal{R}$; we denote
 127 by \mathcal{T} the set of triples in the KG. KGQA questions require retrieving a subgraph $\mathcal{G} \subset \mathcal{T}$ and reason
 128 over it to produce an answer. We assume that the answer(s) to a question q are entities in a set $\mathcal{A} \subset \mathcal{E}$.
 129 We also assume that, together with the question, the set of *seed entities* $\mathcal{S} \subset \mathcal{E}$ is provided, i.e., the
 130 set of entities that are explicitly mentioned in the question. Extracting seed entities from natural
 131 language questions or text is an orthogonal task referred to as Named Entity Recognition, or Entity
 132 Linking, which is in itself an object of extensive research (Alam et al., 2022; Keraghel et al., 2024).
 133

134 **SynthKGQA** We outline SynthKGQA, our proposed framework, which can be applied to any KG
 135 in order to construct high-quality synthetic data for KGQA. The data generation pipeline comprises of
 136 four steps (Figure 1; more details are provided in Appendix A).

137 1. **Seed subgraph sampling.** We randomly sample (Algorithm 1) a connected subgraph
 138 $\mathcal{Q} \subset \mathcal{T}$, containing a few tens of edges, to be used as context for generating the question.

139 2. **LLM-powered candidate proposal.** The subgraph \mathcal{Q} is passed to the LLM, which is tasked
 140 (through few-shot prompting, see Appendix A.1) to generate:

- 141 • a question q that can be answered with k (number specified in the prompt) triples in \mathcal{Q} ;
- 142 • the list of triples in \mathcal{Q} required to reason over the question, i.e., the *ground-truth answer*
 143 *subgraph* $\mathcal{G} \subset \mathcal{Q} \subset \mathcal{T}$;
- 144 • the answer to the question, which is required to be an entity $a \in \mathcal{E}$ appearing in \mathcal{G} ;
- 145 • the list of entities explicitly mentioned in the question, i.e., the seed entities $\mathcal{S} \subset \mathcal{E}$,
 146 which also need to appear in \mathcal{G} ;
- 147 • the SPARQL query l_q which encodes the natural-language question q in logical form.

148 Note that we use SPARQL as RDF query language, since the Wikidata Query Service² is
 149 based on it, but our pipeline can be adapted with minimal changes to use any other query
 150 language, if working with knowledge bases that have different querying interfaces.

151 3. **Candidate validation.** The query l_q is executed against the KG to retrieve the full set
 152 of answers \mathcal{A} ; the candidate datapoint is discarded if \mathcal{A} does not contain the answer a
 153 provided by the LLM. Similarly, it is discarded if any of the triples that the LLM considers
 154 necessary for answering the question, or any of the proposed seed entities, are not in the *full*
 155 *answer subgraph*, obtained by running the associated CONSTRUCT SPARQL query (i.e.,
 156 the union of all triples in the KG realizing the query template, taking into account all valid
 157 substitutions of the variables in l_q ; for more details, see the example in Appendix A.4).

158 4. **Augmentation and classification.** In order to increase the diversity of the generated data,
 159 we ask a different LLM to paraphrase q in more natural terms. Moreover, by using the
 160

²<https://query.wikidata.org/>

162

163

164

165

166

167

168

169

170

171

172

173

174

175

176

177

Table 1: Overall view of GTSQA.

	Train	Test	All
# questions	30,477	1622	32,099
# unique relation types	200	362	368
# unique entities	64,435	5665	68,520
# unique graph isomorphisms	19	14	27
# questions with redundant info	7852	0	7852
avg # seed entities	1.65	1.92	1.66
avg # hops	1.48	2.02	1.5
avg # answers	1.54	1.28	1.53
avg # ground-truth edges	2.13	3.03	2.17

ground-truth answer subgraph and SPARQL query, we can provide additional information on each valid datapoint.

- *Graph isomorphism type*: we classify the structure of the graph \mathcal{G} up to isomorphism (see Appendix A.2), as a measure of question complexity. This also allows the user to perform additional filtering, by discarding questions where the answer subgraph does not comply with desired logical requirements (e.g., disconnected graphs; graphs containing loops or hanging branches, not terminating in a seed entity; graphs where the answer node is also used as a seed node, giving rise to self-answering questions).
- *Redundancy*: we check whether the question contains *redundant* information, i.e., if it can be answered by using only a subset $\mathcal{S}' \subsetneq \mathcal{S}$ of the seed entities (see Appendix A.3).

4 THE GTSQA DATASET

Using the SynthKGQA framework and GPT-4.1 (OpenAI, 2023) as LLM, we construct *Ground-Truth Subgraphs for Question Answering* (GTSQA), a synthetic KGQA dataset grounded in the Wikidata KG (Vrandečić & Krötzsch, 2014), with 30,477 questions in the train set and 1622 questions in the test set. Statistics on the dataset are summarized in Table 1, with more details provided in Appendix B.

GTSQA supports retrieval either from the full Wikidata, or from ogbl-wikikg2 (Hu et al., 2020), which was employed to sample the seed graphs \mathcal{Q} . We construct questions where the ground-truth answer subgraph contains at most 6 edges, as we observe that – beyond this point – questions tend to become overly-convoluted and unnatural to formulate in natural language. We also require the ground-truth answer subgraph to be a tree, i.e., connected and acyclic, with all leaves being seed entities or the answer node. The dataset covers 27 different graph isomorphism types for the ground-truth answer subgraph, with questions involving up to 5 hops and up to 5 different seed entities. Full statistics on the connectivity patterns of seed and answer nodes, and relative frequencies in the dataset, are provided in Table A1 and Appendix B. As shown in Figure B2, approximately 70% of the test questions are multi-seed (two or more seed entities) and 67.9% are multi-hop (with 30.8% requiring ≥ 3 hops). This variety of graph types makes GTSQA much more challenging than commonly-used benchmarks for KGQA (only 34.5% of the test questions in WebQSP (Yih et al., 2016) require more than 1 hop, and none require more than 2). Moreover, the test set of GTSQA has the unique property of containing only non-redundant questions (as defined in Section 3), where the full ground-truth answer subgraph is necessary to answer, which ensures that the provided classification of graph isomorphism types is truly reflective of the complexity of the questions.

A notable feature of GTSQA is that the train-test split is designed to test zero-shot generalization abilities of KG retrievers (in a spirit similar to the construction of GrailQA (Gu et al., 2021) and GrailQA++ (Dutt et al., 2023)). We ensure that the answer nodes for all test questions never appear as an answer in training questions. Moreover, 47.3% of the answer subgraphs in the test set have as isomorphism type one of 8 classes that are not included in the train set (see Table A1), while 37.2% of test questions require reasoning over edges whose relation type is not seen in the train set (see Figure B1 for the distribution of relation types). We also make sure that, for each graph isomorphism type in the test set, the applicable categories (in-distribution, unseen graph type, unseen relation

216 type) always contain at least 45 different questions, to enable a statistically meaningful study of KG
 217 retrievers’ performance at this unprecedented granularity level (as we do in Section 5).
 218

219 We perform a final filtering step on the test set of GTSQA to ensure its value and reliability as a
 220 benchmark for KG retrievers: we retain only questions that can be answered correctly by an advanced
 221 LLM (GPT-4o-mini) when augmenting the prompt with the ground-truth answer subgraph provided
 222 in the dataset, in a consistent way (two successes out of two tries). This is to sanity check that the
 223 ground-truth answer subgraph is indeed a valid golden target for retrieval, providing grounding to any
 224 insights we extract from the benchmarks. We found that only 0.47% of the generated data failed this
 225 test (and was therefore removed), which confirms the very high quality of the synthetic data produced
 226 by SynthKGQA (compared, for instance, to KGQAGen-10k (Zhang et al., 2025), where this failure
 227 rate is reported at 10.38% with the more capable GPT-4o, and failure causes remain unclear).
 228

229 5 BENCHMARKS

230 We benchmark SOTA KG-RAG models on GTSQA. As one of the key features of our dataset is that it
 231 provides ground-truth answer subgraphs for each question, we can evaluate not just the quality of the
 232 models’ final answer, but also the quality of the retrieved subgraph, for different graph isomorphism
 233 types and generalization abilities required by the question.
 234

235 **Experimental setting** We consider a variety of models for QA, divided into four categories
 236 (specifications in Appendix C.2). 1) *LLM-only*: as a baseline, we take frontier commercial LLMs
 237 without any external augmentation pipelines, such as GPT-4.1, GPT-4o-mini (OpenAI, 2023), GPT-5-
 238 mini (OpenAI, 2025), Minstral-8B-Instruct, Minstral-Large-2.1 (Minstral AI, 2024) and LLaMA-3.1-
 239 8B-Instruct (AI@Meta, 2024). 2) *KG agents*: training-free models using out-of-the-box LLMs to
 240 explore the KG starting from the seed entities. They decide on their own, over multiple steps, what
 241 neighbors to explore and when to stop. As representatives, we consider Think-on-Graph (ToG; Sun
 242 et al. (2024)) and Plan-on-Graph (PoG; Chen et al. (2024)). 3) *Path-based retrievers*: models trained
 243 to predict the paths originating from the seed entities and leading to the answer node. Examples
 244 are SR (Zhang et al., 2022), Reasoning on Graphs (RoG; Luo et al. (2024)) and Graph-Constrained
 245 Reasoning (GCR; Luo et al. (2025)). 4) *All-at-once retrievers*: models trained to score all edges in a
 246 large neighborhood of the seed entities in a single pass, and then retrieve the most relevant ones. We
 247 consider SubgraphRAG (Li et al., 2025) as a representative.
 248

249 For all models in categories 2), 3), 4), we use the same LLM (GPT-4o-mini, which – as explained in
 250 Section 4 – is able to answer all test questions when provided with the ground-truth answer subgraph)
 251 to perform the final reasoning on the retrieved subgraph, in order to make the benchmark as fair as
 252 possible and place the focus on the quality of the retrieved subgraph. We use ogbl-wikikg2 as KG
 253 for retrieval, to reduce computational complexity (see also Appendix C.1). As common in KGQA
 254 benchmarking (Sun et al., 2024; Luo et al., 2024), we evaluate quality of model responses by reporting
 255 on Hits (at least one correct answer predicted) and Recall of correct answers, using Exact Match
 256 (EM) between the labels of answer nodes and the model output. Moreover, as made possible by
 257 GTSQA, we analyze Recall, Precision and F1 score of ground-truth triples in the subgraph retrieved
 258 by the model, in addition to Hits and Recall of answer nodes.
 259

260 **Results** A detailed comparison of the performance of different models on GTSQA is provided in
 261 Table 2. Pure LLMs show limited accuracy with an EM Hits score of at most 33.97, confirming the
 262 challenging nature of this new benchmark for models that solely rely on their internal knowledge.
 263 Even SOTA KG-RAG models – while performing better than pure LLMs – struggle with the task.
 264 Trainable subgraph retrievers outperform KG agents, highlighting the importance of fine-tuning
 265 models on the target KG. Among trainable retrievers, SubgraphRAG (with 200 retrieved triples)
 266 achieves the best results, surpassing all path-based retrievers, which tend to retrieve smaller subgraphs.
 267 The F1 score of retrieved triples is low (< 30%) across all retrievers, but our results suggest that
 268 increasing the recall of ground-truth triples is more beneficial than increasing precision, especially
 269 when the final reasoning is performed by more capable LLMs, better at dealing with long context,
 as we show in the comparison in Figure C7. For more compact LLMs, however, retrieval precision
 should not be entirely disregarded. Also note that the recall of ground-truth triples is a much stronger
 predictor of final model accuracy than the recall of answer nodes (Figure C6), proving the value of
 working with datasets like GTSQA that provide the full ground-truth answer subgraph.
 270

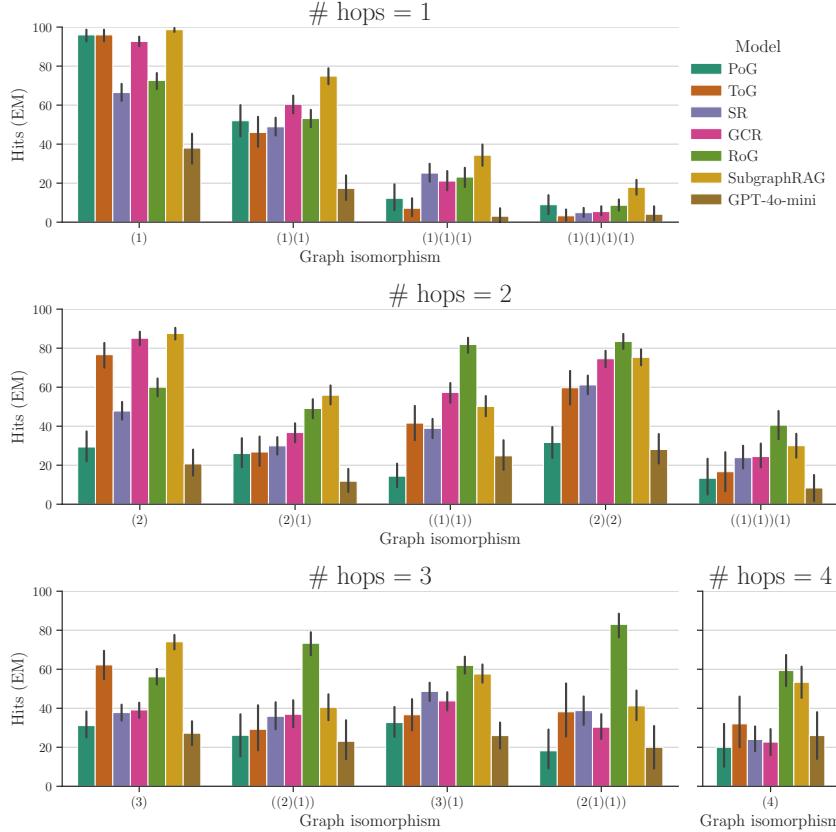
270

271 Table 2: Benchmark of LLMs and KG-RAG models on GTSQA.

272 Category	273 Model	274 EM		275 Ground-truth triples			276 Answer nodes			277 # triples
		278 Hits	279 Recall	280 Recall	Precision	F1	281 Hits	282 Recall		
283 LLM-only	Minstral-8B-Instruct	10.73	10.16	-	-	-	-	-	-	
	LLama-3.1-8B-Instruct	17.11	16.33	-	-	-	-	-	-	
	GPT-4o-mini	20.90	19.93	-	-	-	-	-	-	
	Mistral-Large-2.1	23.61	22.85	-	-	-	-	-	-	
	GPT-5-mini	31.44	30.20	-	-	-	-	-	-	
	GPT-4.1	33.97	32.83	-	-	-	-	-	-	
284 KG agent	PoG	32.92	31.60	31.95	27.52	27.03	31.81	30.54	7.09	
	ToG	45.68	44.47	35.50	6.45	10.50	41.55	40.97	9.17	
285 Path-based	SR	40.63	39.10	30.22	3.44	5.69	50.25	49.39	72.94	
	GCR	49.91	48.25	40.71	27.21	29.82	47.11	45.54	6.54	
	RoG	57.58	55.78	54.69	24.04	27.00	72.91	71.84	72.31	
All-at-once	SubgraphRAG (200)	61.59	58.62	79.09	1.29	2.53	85.33	84.36	199.61	

286

287



315

316 Figure 2: Hits (EM) of KG-RAG models on different graph isomorphism types, compared to the
 317 baseline (GPT-4o-mini, no RAG). Isomorphism types are grouped by the maximum number of hops;
 318 inside each subplot, moving from left to right corresponds to an increase in the total number of edges
 319 in the ground-truth answer subgraph.

320

321

322

323

324 Breaking down model performance by graph isomorphism types (Figures 2 and C8), we find that all
 325 models, especially KG agents, have very limited accuracy on questions that require intersecting paths
 326 from 3 or more seed entities, even if the answer is only one hop away (e.g., graph isomorphisms

324 (1)(1)(1) and (1)(1)(1)(1)). In this setting, the base LLM on its own also fails consistently. The
 325 poor EM is explained by a low recall of ground-truth triples, highlighting a widespread inefficiency
 326 in properly expanding and coordinating the search from all seed entities. Indeed, especially for
 327 KG agents and path-based methods, the recall of ground-truth triples is on average much lower in
 328 multi-seed questions (Figure C8), leading to a worse EM, despite the fact that the answer node is
 329 often contained in the retrieved subgraph (Figure C9). This is a consequence of the fact that these
 330 models sample paths from distinct seed nodes in a mostly independent way, and therefore struggle
 331 with prioritizing common neighbors of the seed nodes that satisfy at the same time all conditions
 332 in the query. Unlike prior benchmarks, GTSQA makes this failure mode visible by providing all
 333 ground-truth triples and by enforcing that all seed entities are required to answer the question (no
 334 redundancy). We provide examples of typical failure cases of different retrievers in Appendix D.
 335

336 Performance also depends on the number of hops between seed entities and the answer. Comparing
 337 KG agents, we observe that PoG matches or outperforms ToG for single-hop questions, but performs
 338 severely worse on multi-hop questions, even on simpler graph isomorphism types that involve a
 339 single seed entity (e.g., (2) and (3)). This is due to a reduced recall of ground-truth triples (Figure C8).
 340 An empirical inspection of the retrieved subgraphs indicates that PoG often prematurely stops the
 341 exploration of the KG, over-confidently (and incorrectly, in most cases) believing it has retrieved
 342 enough information to answer (see example in Appendix D). We also find a poor performance of GCR,
 343 especially when compared to its predecessor RoG, across all graph isomorphism types requiring more
 344 than 2 hops from the seed entities. This can mostly be attributed to the implementation restrictions
 345 and choices explained in Appendix C.2. Indeed, for these questions we observe that answer node
 346 recall drops much more than ground-truth triple recall (Figure C9), because only paths of up to 2
 347 hops from the seed entities can be retrieved. However, GCR also performs worse than RoG on some
 348 questions with only 2 hops required, such as graph isomorphisms ((1)(1)) and (2)(1).
 349

350 Finally, we investigate the zero-shot generalization abilities of trainable retrievers. When a ques-
 351 tion presents a relation or a graph isomorphism type that has not been seen during training, the
 352 improvements over the LLM-baseline are much smaller than for in-distribution questions (Figures 3
 353 and C9). In these settings, we also notice stronger differences between different KG-RAG models.
 354 In particular, all path-based retrievers particularly struggle with unseen relation types, even though
 355 GCR is much more robust than SR and RoG for this type of generalization, with significantly higher
 356 recall of ground-truth triples and EM, as its path predictions are grounded in the KG and hence less
 357 dependent on the relation types seen during training. On the other hand, the all-at-once retriever
 358 SubgraphRAG maintains strong predictive power on unseen relation types, but underperforms RoG
 359 across the majority of unseen graph isomorphism types (Figure C9), with a substantial drop in recall
 360 of answer nodes, which, combined with the low precision of retrieved triples typical of this method,
 361 results in a poor EM. These fundamental differences between the two classes of retrievers (path-based
 362 vs all-at-once) match intuition: path-based methods learn to predict the sequence of relation types
 363 that connect seed entities and answer, ignoring the global structure of the answer subgraph, whereas
 364 all-at-once retrievers like SubgraphRAG mostly leverage the graph structure during their training.
 365 SubgraphRAG especially struggles with graph isomorphism types that require additional projections
 366 after an intersection (namely, ((1)(1)), (2(1)(1)), ((2)(1)), ((1)(1))(1), see Figure C9), a property
 367 foreign to the answer subgraphs in the train set – highlighting the lack of robustness of this retriever
 368 with respect to new patterns in the logical query (we show an example in Appendix D).
 369

6 GROUND-TRUTH SUBGRAPHS TO TRAIN BETTER KG RETRIEVERS

370 In the absence of ground-truth answer subgraphs in previous datasets, KG retrievers are typically
 371 trained using all shortest paths between seed and answer nodes as supervision signal (Zhang et al.,
 372 2022; Luo et al., 2024; Li et al., 2025; Luo et al., 2025). The ground-truth subgraphs in GTSQA
 373 enable us to conduct experiments addressing the following questions: 1) How much overlap is there
 374 between shortest paths and ground-truth subgraphs? 2) Are LLMs still able to answer correctly
 375 if augmented with shortest-paths between seed nodes and answer node? 3) Is using ground-truth
 376 subgraphs as supervision signal for training KG retrievers better than using shortest paths?
 377

378 **Q1: Path overlap** Paths of minimal length connecting seed nodes to answer nodes are not guar-
 379 anteed to be the correct paths required to answer the question. For example, the question “What is
 380 the canonization status of Gregory the Illuminator’s grandchild?” requires retrieving the 3-hop path
 381

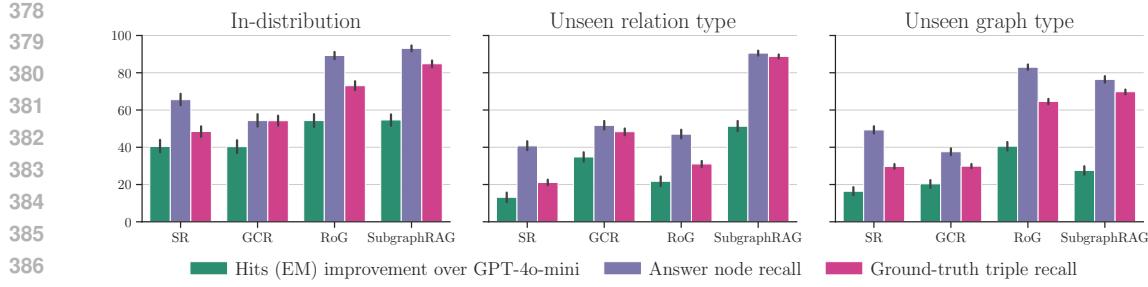


Figure 3: Generalization abilities of trainable KG-RAG models. We measure EM performance in terms of difference with the EM of the baseline (GPT-4o-mini, no RAG).

Gregory the Illuminator → *child* → *St. Vrtanes I* → *child* → *St. Husik I* → *canonization status* → *saint*. However, in Wikidata, the shortest path connecting the seed node (*Gregory the Illuminator*) to the answer node (*saint*) consists of a single edge: *Gregory the Illuminator* → *canonization status* → *saint*, which clearly provides the wrong reasoning. We refer to similar cases, where the minimal length of paths between seed and answer node is strictly smaller than the length of the path in the ground-truth subgraph, as *shortcuts*. A second source of problems arises from *parallel* (shortest) *paths*: as the distance between the seed and answer node increases, the number of distinct paths of minimal length connecting them is expected to grow exponentially, but only one of them (or none, if there are shortcuts) is a ground-truth path. We provide statistics on the occurrence of shortcuts and parallel paths in GTSQA in Figure B5; their combined effect, as documented in Table B2, strongly reduces the overlap between ground-truth and shortest paths for questions requiring multiple hops. For 4-hop questions, on average, only 13% of the triples along shortest paths are contained in the ground-truth answer subgraph and more than 72% of the ground-truth triples do not lie on paths of minimal length.

Q2: RAG with shortest paths There are cases where following shortcuts or parallel paths, rather than the ground-truth path that we provide in the dataset, can still provide valid reasoning. For instance, the question “Which country is home to the administrative region that includes Nieuw-Weerdinge?”, which comes with the 2-hop ground-truth path *Nieuw-Weerdinge* → *located in the administrative territorial entity* → *Emmen* → *country* → *Netherlands*, can be answered equally well with the shortcut *Nieuw-Weerdinge* → *country* → *Netherlands*. Similarly, Wikidata contains pairs of inverse relation types (e.g., *child* and *father*) that give rise to parallel (undirected) paths encoding the same semantical information. We therefore look at the performance of GPT-4o-mini on the test set of GTSQA, when augmenting the prompt with all KG triples on the shortest paths between seed and answer nodes. As shown in Figure 4, there is a strong positive correlation between Hits (EM) and the fraction of ground-truth triples that are contained in the set of shortest path triples. Both metrics degrade sharply with increasing distance between seed and answer node. Note that we already adjust for any spurious effects of the number of hops on EM, since GPT-4o-mini is able to answer all questions correctly when instead we augment the prompt with the triples in the ground-truth answer subgraph. This confirms that shortest paths, even though they end at the correct answer node, often fail to capture the full information required to reason over multi-hop questions.

Q3: Shortest paths vs ground-truth subgraphs for training While the experiments in the previous paragraphs prove that shortest path triples are not as good a target for retrieval as the ground-truth triples provided by SynthKGQA, it remains to understand whether using them as supervision signal for training subgraph retrievers is equally sub-optimal. We consider three retrievers that use path information for training, namely SR (Zhang et al., 2022), RoG (Luo et al., 2024) and SubgraphRAG (Li et al., 2025), train them on GTSQA using either the shortest paths or the ground-truth paths between seed and answer nodes as supervision signal, and then compare the performance on the test set. As reported in Table 3, models trained on the ground-truth subgraphs have EM Hits scores from 5% to 20% higher than their counterparts trained on shortest paths, due to improved recall (up to +27%) and precision (up to +141%) of ground-truth triples in the retrieved subgraphs. The improvements are clear and statistically significant for SubgraphRAG (independently

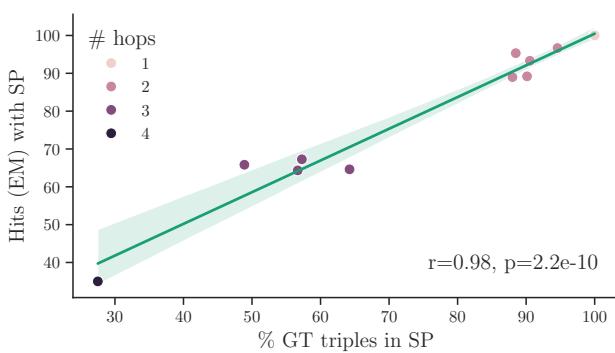


Figure 4: Correlation between the percentage of ground-truth (GT) triples contained in the set of shortest paths (SP) from seed to answer nodes, and EM Hits of GPT-4o-mini when augmented with all SP triples. We display the Pearson correlation coefficient and associated p-value; each dot represents a different isomorphism type of ground-truth answer subgraph in the test set of GTSQA: they are clearly clustered based on the (maximum) number of hops.

Table 3: Improvements in predictive statistics of models trained on ground-truth answer subgraphs (GT), compared to models trained on the shortest paths between seed and answer nodes (SP). For all models, results are averaged over three distinct runs for each of the two training regimes. We also report (between brackets) the metric’s % variation and the p-value of the null hypothesis that the mean of distribution of SP scores is larger or equal than the one for GT scores (two-sample, one-sided t-test), showing statistical significance.

Model	Hits (EM)		Recall (GT triples)		Precision (GT triples)		Recall (answer nodes)	
	SP	GT	SP	GT	SP	GT	SP	GT
SR	33.95	40.63 (+20%; 4.5e-12)	23.74	30.22 (+27%; 6.2e-22)	3.84	3.44 (-10%; 0.99)	36.36	49.39 (+36%; 1.0e-39)
RoG	53.08	57.58 (+8%; 3.9e-6)	46.38	54.69 (+18%; 4.6e-26)	10.00	24.04 (+141%; 0.0)	70.68	71.84 (+2%; 4.3e-2)
SubgraphRAG	58.47	61.59 (+5%; 8.3e-4)	75.39	79.09 (+5%; 1.7e-9)	1.21	1.29 (+6%; 4.0e-7)	80.46	84.36 (+5%; 1.4e-7)

of the number of retrieved triples, Figure C7), but even more striking for path-based methods. As shown in Figure C10, RoG is the model where the quality of the retrieved subgraph benefits more from eliminating the noise in the training data, with precision increasing by more than $9\times$ for 4-hops questions. Similarly, SR experiences a major boost in answer node recall (more than $2\times$ for 4-hops questions). While for 1-hop questions the differences in all statistics are negligible, the gap sharply increases for questions requiring multiple hops (where, as we’ve showed, the shortest paths diverge more appreciably from the ground-truth subgraphs).

All this provides definitive evidence on the value of the datasets generated with SynthKGQA not just as benchmarks, but also to train better subgraph retrievers for complex KGQA.

7 CONCLUSIONS

In this work, we introduced SynthKGQA, a new framework for constructing large-scale, high-quality synthetic KGQA datasets from arbitrary KGs, with ground-truth answer subgraphs and SPARQL queries for all questions. As concrete application, we released GTSQA, a multi-hop, multi-seed dataset with 30k+ questions based on Wikidata, designed to test how KG-RAG models generalize to unseen answer graph structures and relation types. Our benchmarks show that SOTA KG-RAG models struggle on GTSQA, due to poor retrieval performance that affects especially questions with multiple seed entities. All-at-once retrievers tend to outperform path-based ones and KG agents, due to their higher recall of triples in ground-truth answer subgraphs. Finally, we show how leveraging the SynthKGQA-generated ground-truth subgraphs as supervision signal for training KG retrievers produces better models, with accuracy improvements of up to 30% for multi-hop questions. We anticipate that SynthKGQA and the KGQA datasets generated through it will help benchmark and improve KG retrievers and, consequently, contribute to the development of more trustworthy LLMs.

486 REPRODUCIBILITY STATEMENT
487

488 The code to run all steps of the SynthKGQA pipeline (detailed in Section 3 and Appendix A) on
489 a custom Knowledge Graph is made available in the supplementary material. We also include the
490 full GTSQA dataset (train and test split) and the code to generate the question-specific graphs for
491 the test questions (Appendix C.1; such graphs will be released together with the dataset). In the
492 supplementary material we also provide all code to run the benchmarks and experiments (Sections 5
493 and 6) for the evaluated KG-RAG models, whose specifications are detailed in Appendix C.2.

494
495 REFERENCES
496

497 Alfred V. Aho and John E. Hopcroft. *The Design and Analysis of Computer Algorithms*. Addison-
498 Wesley Longman Publishing Co., Inc., 1st edition, 1974.

499 AI@Meta. The Llama 3 Herd of Models, 2024. URL <https://arxiv.org/abs/2407.21783>.

500
501

502 Mehwish Alam, Davide Buscaldi, Michael Cochez, Francesco Osborne, Diego Reforgiato Recu-
503 pero, Harald Sack, Özge Sevgili, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko,
504 Chris Biemann, Mehwish Alam, Davide Buscaldi, Michael Cochez, Francesco Osborne, Diego
505 Refogliato Recupero, and Harald Sack. Neural entity linking: A survey of models based on deep
506 learning. *Semant. Web*, 13(3):527–570, 2022.

507 Michael Azmy, Peng Shi, Jimmy Lin, and Ihab Ilyas. Farewell Freebase: Migrating the SimpleQues-
508 tions dataset to DBpedia. In *Proceedings of the 27th International Conference on Computational
509 Linguistics*, pp. 2093–2103. Association for Computational Linguistics, August 2018.

510 Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from
511 question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural
512 language processing*, pp. 1533–1544, 2013.

513 Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collab-
514 oratively created graph database for structuring human knowledge. In *Proceedings of the 2008
515 ACM SIGMOD International Conference on Management of Data*, SIGMOD ’08, pp. 1247–1250.
516 Association for Computing Machinery, 2008.

517 Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican,
518 George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al.
519 Improving language models by retrieving from trillions of tokens. In *International conference on
520 machine learning*, pp. 2206–2240. PMLR, 2022.

521 Liyi Chen, Panrong Tong, Zhongming Jin, Ying Sun, Jieping Ye, and Hui Xiong. Plan-on-graph:
522 Self-correcting adaptive planning of large language model on knowledge graphs. In *Proceedings
523 of the 38th Conference on Neural Information Processing Systems*, 2024.

524 Preetam Prabhu Srikar Dammu, Himanshu Naidu, and Chirag Shah. Dynamic-KGQA: A scalable
525 framework for generating adaptive question answering datasets, 2025. URL <https://arxiv.org/abs/2503.05049>.

526 Rajarshi Das, Ameya Godbole, Ankita Naik, Elliot Tower, Manzil Zaheer, Hannaneh Hajishirzi,
527 Robin Jia, and Andrew Mccallum. Knowledge base question answering by case-based reasoning
528 over subgraphs. In *Proceedings of the 39th International Conference on Machine Learning*, volume
529 162 of *Proceedings of Machine Learning Research*, pp. 4777–4793. PMLR, July 2022.

530 Zifeng Ding, Heling Cai, Jingpei Wu, Yunpu Ma, Ruotong Liao, Bo Xiong, and Volker Tresp. zrLLM:
531 Zero-shot relational learning on temporal knowledge graphs with large language models, 2024.
532 URL <https://arxiv.org/abs/2311.10112>.

533 Mohnish Dubey, Debyan Banerjee, Abdelrahman Abdelkawi, and Jens Lehmann. LC-QuAD
534 2.0: A Large Dataset for Complex Question Answering over Wikidata and DBpedia. pp. 69–78.
535 Springer-Verlag, 2019.

540 Ritam Dutt, Sopan Khosla, Vinayshekhar Bannihatti Kumar, and Rashmi Gangadharaiyah. GrailQA++:
 541 A challenging zero-shot benchmark for knowledge base question answering. In *Proceedings of*
 542 *the 13th International Joint Conference on Natural Language Processing and the 3rd Conference*
 543 *of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long*
 544 *Papers)*, pp. 897–909. Association for Computational Linguistics, November 2023.

545 Yu Gu, Sue Kase, Michelle Vanni, Brian Sadler, Percy Liang, Xifeng Yan, and Yu Su. Beyond i.i.d.:
 546 Three levels of generalization for question answering on knowledge bases. In *Proceedings of the*
 547 *Web Conference 2021*, WWW ’21, pp. 3477–3488. Association for Computing Machinery, 2021.

548 Gaole He, Yunshi Lan, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. Improving multi-hop
 549 knowledge base question answering by learning intermediate supervision signals. In *Proceedings of*
 550 *the 14th ACM International Conference on Web Search and Data Mining*, WSDM ’21, pp. 553–561,
 551 New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450382977.

552 Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta,
 553 and Jure Leskovec. Open graph benchmark: datasets for machine learning on graphs. In *Proceed-
 554 ings of the 34th International Conference on Neural Information Processing Systems*, NIPS ’20,
 555 Red Hook, NY, USA, 2020. Curran Associates Inc.

556 Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong
 557 Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large
 558 language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on
 559 Information Systems*, 43(2):1–55, 2025.

560 Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane
 561 Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning
 562 with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43,
 563 2023.

564 Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif. Recent advances in named entity recognition:
 565 A comprehensive survey and comparative study, 2024. URL <https://arxiv.org/abs/2401.10825>.

566 Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal,
 567 Heinrich Kütller, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented genera-
 568 tion for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:
 569 9459–9474, 2020.

570 Mingchen Li and Shihao Ji. Semantic structure based query graph prediction for question answering
 571 over knowledge graph. In *Proceedings of the 29th International Conference on Computational*
 572 *Linguistics*, pp. 1569–1579, Gyeongju, Republic of Korea, October 2022. International Committee
 573 on Computational Linguistics.

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

577 Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards
 578 general text embeddings with multi-stage contrastive learning, 2023. URL <https://arxiv.org/abs/2308.03281>.

579 Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis,
 580 Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining
 581 Approach, 2019. URL <https://arxiv.org/abs/1907.11692>.

582 Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. On
 583 LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the*
 584 *Association for Computational Linguistics: ACL 2024*, pp. 11065–11082, Bangkok, Thailand,
 585 August 2024. Association for Computational Linguistics.

586 Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. Reasoning on graphs: Faithful
 587 and interpretable large language model reasoning. In *International Conference on Learning*
 588 *Representations*, 2024.

594 Linhao Luo, Zicheng Zhao, Chen Gong, Gholamreza Haffari, and Shirui Pan. Graph-constrained
 595 reasoning: Faithful reasoning on knowledge graphs with large language models. In *Forty-second*
 596 *International Conference on Machine Learning*, 2025.

597

598 Costas Mavromatis and George Karypis. GNN-RAG: Graph neural retrieval for efficient large
 599 language model reasoning on knowledge graphs. In *Findings of the Association for Computational*
 600 *Linguistics: ACL 2025*, pp. 16682–16699, Vienna, Austria, July 2025. Association for
 601 Computational Linguistics.

602 Mistral AI. Models overview. https://docs.mistral.ai/getting-started/models/models_overview/, 2024. Accessed: 2025-08-14.

603

604 OpenAI. Gpt-4 technical report, 2023. URL <https://arxiv.org/abs/2303.08774>.

605

606 OpenAI. Gpt-5 system card. <https://cdn.openai.com/pdf/8124a3ce-ab78-4f06-96eb-49ea29ffb52f/gpt5-system-card-aug7.pdf>,
 607 2025. Accessed: 2025-08-14.

608

609 Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking:
 610 Bringing Order to the Web, 1998.

611

612 Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang
 613 Tang. Graph retrieval-augmented generation: A survey, 2024. URL <https://arxiv.org/abs/2408.08921>.

614

615 Fabian M. Suchanek, Mehwish Alam, Thomas Bonald, Lihu Chen, Pierre-Henri Paris, and Jules
 616 Soria. Yago 4.5: A large and clean knowledge base with a rich taxonomy. In *Proceedings of the*
 617 *47th International ACM SIGIR Conference on Research and Development in Information Retrieval*,
 618 *SIGIR '24*, pp. 131–140, New York, NY, USA, 2024. Association for Computing Machinery.

619

620 Jiahuo Sun, Chengjin Xu, Lumingyuan Tang, Sai Wang, Chen Lin, Yeyun Gong, Lionel M. Ni,
 621 Heung yeung Shum, and Jian Guo. Think-on-graph: Deep and responsible reasoning of large
 622 language model on knowledge graph. In *International Conference on Learning Representations*,
 623 2024.

624

625 Alon Talmor and Jonathan Berant. The web as a knowledge-base for answering complex questions.
 626 In *Proceedings of the 2018 Conference of the North American Chapter of the Association for*
 627 *Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 641–651,
 628 New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

629

630 Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay
 631 Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cris-
 632 tian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu,
 633 Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn,
 634 Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel
 635 Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee,
 636 Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra,
 637 Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi,
 638 Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh
 639 Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen
 640 Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic,
 641 Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models,
 642 2023. URL <https://arxiv.org/abs/2307.09288>.

643

644 Priyansh Trivedi, Gaurav Maheshwari, Mohnish Dubey, and Jens Lehmann. LC-QuAD: A corpus for
 645 complex question answering over knowledge graphs. In *The Semantic Web – ISWC 2017: 16th*
 646 *International Semantic Web Conference, Vienna, Austria, October 21–25, 2017, Proceedings, Part*
 647 *II*, pp. 210–218. Springer-Verlag, 2017.

648

649 Ricardo Usbeck, Ria Hari Gusmita, Muhammad Saleem, and Axel-Cyrille Ngonga Ngomo. 9th
 650 challenge on question answering over linked data (qald-9). In *Semdeep/NLJWoD@ISWC*, 2018.
 651 URL <https://api.semanticscholar.org/CorpusID:53220210>.

648 Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledge base. *Commun.*
649 *ACM*, 57(10):78–85, September 2014.
650

651 Wen-tau Yih, Matthew Richardson, Chris Meek, Ming-Wei Chang, and Jina Suh. The value of
652 semantic parse labeling for knowledge base question answering. In *Proceedings of the 54th Annual*
653 *Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 201–206,
654 Berlin, Germany, August 2016. Association for Computational Linguistics.

655 Jing Zhang, Xiaokang Zhang, Jifan Yu, Jian Tang, Jie Tang, Cuiping Li, and Hong Chen. Subgraph
656 retrieval enhanced model for multi-hop knowledge base question answering. In *Proceedings of the*
657 *60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,
658 pp. 5773–5784, Dublin, Ireland, May 2022. Association for Computational Linguistics.

659 Liangliang Zhang, Zhuorui Jiang, Hongliang Chi, Haoyang Chen, Mohammed Elkoumy, Fali Wang,
660 Qiong Wu, Zhengyi Zhou, Shirui Pan, Suhang Wang, and Yao Ma. Diagnosing and addressing
661 pitfalls in KG-RAG datasets: Toward more reliable benchmarking, 2025. URL <https://arxiv.org/abs/2505.23495>.
662

663 Muhan Zhang, Pan Li, Yinglong Xia, Kai Wang, and Long Jin. Labeling trick: a theory of using graph
664 neural networks for multi-node representation learning. In *Proceedings of the 35th International*
665 *Conference on Neural Information Processing Systems*, NIPS ’21, 2021.
666

667

668

669

670

671

672

673

674

675

676

677

678

679

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701

702 A ADDITIONAL DETAILS ON SYNTHKGQA
703704 A.1 DATA GENERATION
705706 We provide additional details on how we query the LLM to generate synthetic data in the SynthKGQA
707 framework (steps 1 and 2 of the pipeline presented in Section 3).
708709 **Seed subgraph** We first sample a subgraph $\mathcal{Q} \subset \mathcal{T}$ of the KG, which will provide the LLM with
710 the context to generate a question, as asking the LLM to parse the entire KG in its input is unfeasible
711 or impractical. Starting from a random entity $s \in \mathcal{E}$, the subgraph \mathcal{Q} is obtained by progressively
712 expanding a neighborhood around s until the desired size (decided by the user, typically 20-100
713 edges) is reached. As we aim to construct multi-hop questions, it's important to ensure that the radius
714 of \mathcal{Q} grows quickly, even if the total number of edges in \mathcal{Q} is kept contained. This means prioritizing
715 depth over breadth in the expansion, while maintaining high-degrees of connectivity between nodes,
716 to enable a variety of different isomorphism types for the answer subgraph. Our proposed sampling
717 scheme for \mathcal{Q} , used in the construction of GTSQA, is shown in Algorithm 1; we denote by $d(x)$ the
718 degree of a node x (number of edges going in/out of it) and by $\mathcal{N}(x)$ its set of undirected neighbors.
719719 **Algorithm 1** Sampling scheme for the seed graph \mathcal{Q}
720720 **Require:** KG $\mathcal{T} \subset \mathcal{E} \times \mathcal{R} \times \mathcal{E}$; starting node $s \in \mathcal{E}$; node limit N_{nodes} ; edge limit N_{edges}
721 $N \leftarrow \{s\}$
722 $\mathcal{Q} \leftarrow \emptyset$
723 **while** $|N| < N_{\text{nodes}}$ and $|\mathcal{Q}| < N_{\text{edges}}$ **do**
724 $z \leftarrow \text{choice}(N, p = \text{Softmax}([d(x)^{-1} \text{ for } x \in N]))$ ▷ Sample node to expand
725 $M \leftarrow \mathcal{N}(z) \setminus N$ ▷ Retrieve new neighbors of z
726 **if** $|M| > 0$ **then**
727 $n \leftarrow \text{choice}(M, p = \text{Softmax}([d(x)^{-1} \text{ for } x \in M]))$ ▷ Sample neighbor
728 $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{(h, r, t) \in \mathcal{T} \text{ s.t. } h = n, t \in N \text{ or } t = n, h \in N\}$ ▷ Add all connections to n
729 $N \leftarrow N \cup \{n\}$
730 **end if**
731 **end while**
732732 **LLM querying** The sampled subgraph \mathcal{Q} is given to the LLM to generate one KGQA datapoint. In
733 the prompt we specify the number k of triples in \mathcal{Q} that should be used to reason over the question
734 (i.e., the number of triples in the ground-truth answer subgraph \mathcal{G}), and show examples in few-shot
735 prompting style. The RDF identifiers in the knowledge base (e.g., QIDs and PIDs when working
736 with Wikidata) should be included in the prompt, together with labels, for all entities and relations
737 appearing in \mathcal{Q} , in order to ensure consistency in the generation of the SPARQL query. An example
738 (for $k = 2$) of the prompt used for the construction of GTSQA is shown below.
739740 **LLM Prompt ($k = 2$)**741 { "role": "user", "content": "Based on the provided set of knowledge
742 graph triples, please generate a question that requires combining the
743 information in exactly 2 of the provided triples for answering. The
744 answer should correspond to exactly one node in the provided graph, be
745 unique and not ambiguous. Make sure that all 2 of the selected triples
746 are required for answering the question and that they involve multiple
747 different entities. Respond only with
748 • 'Question:' the generated question,
749 • 'Nodes mentioned in the question:' a semicolon-separated list
750 of the nodes that are explicitly mentioned in the question,
751 • 'Answer:' the node corresponding to the correct answer with its
752 QID,
753 • 'Triples used:' a semicolon-separated list of the triples used
754 for answering the question,
755 • 'SPARQL query:' a SPARQL query to return all answers of the
756 question from the Wikidata knowledge base. }

```

756 },
757 {
758   "role": "user", "content": "Graph: Quiaios (Q1024001)-country
759   (P17)-Portugal (Q45);Savage Nights (Q174785)-filming location
760   (P915)-Portugal (Q45);Quiaios (Q1024001)-located in the administrative
761   territorial entity (P131)-Figueira da Foz (Q428459);Figueira
762   da Foz (Q428459)-country (P17)-Portugal (Q45);Savage Nights
763   (Q174785)-filming location (P915)-Lisbon (Q597);Portugal (Q45)-capital
764   (P36)-Lisbon (Q597);Lisbon (Q597)-country (P17)-Portugal (Q45);Lisbon
765   (Q597)-capital of (P1376)-Portugal (Q45);Francisco José Fernandes
766   Costa (Q357932)-country of citizenship (P27)-Portugal (Q45);Francisco
767   José Fernandes Costa (Q357932)-place of death (P20)-Figueira da
768   Foz (Q428459);Francisco José Fernandes Costa (Q357932)-position
769   held (P39)-Minister of Foreign Affairs (Q4294919);Minister of
770   Foreign Affairs (Q4294919)-country (P17)-Portugal (Q45);Minister of
771   Foreign Affairs (Q4294919)-applies to jurisdiction (P1001)-Portugal
772   (Q45);Francisco José Fernandes Costa (Q357932)-occupation
773   (P106)-politician (Q82955)
774 },
775 {
776   "role": "assistant", "content": "
777 Question: In which country did Francisco José Fernandes Costa die?,
778 Nodes mentioned in the question: Francisco José Fernandes Costa
779 (Q357932),
780 Answer: Portugal (Q45),
781 Triples used: Francisco José Fernandes Costa (Q357932)-place of death
782 (P20)-Figueira da Foz (Q428459); Figueira da Foz (Q428459)-country
783 (P17)-Portugal (Q45),
784 SPARQL query: SELECT ?answer WHERE {wd:Q357932 wdt:P20 ?place. ?place
785 wdt:P17 ?answer.}
786 },
787 [...][**OMITTED -- examples with different graph isomorphism types**]
788 {"role": "user", "content": "Graph: {Q}"},
789

```

784 A.2 GRAPH ISOMORPHISMS

785
786 We use the notion of *graph isomorphism*, applied to the ground-truth answer subgraph, to provide
787 a simple and objective measure of the complexity of a question in a KGQA dataset. This measure
788 abstracts away the identity of the entities and relation types involved in the answer subgraph and only
789 focuses on the number of seed entities, the number of hops separating them from the answer node
790 and how the paths originating from each seed intersect. In the context of KGQA, this notion first
791 appeared as *reasoning paths* (Das et al., 2022) and *semantic structures* (Li & Ji, 2022), and was then
792 formulated in the same way that we will use it in Dutt et al. (2023) for the construction of GrailQA++.

793 We say that two questions have the same graph isomorphism type if their ground-truth answer
794 subgraphs \mathcal{G} are isomorphic as labeled graphs, when each node in \mathcal{G} is labeled as “seed” (for seed
795 nodes), “answer” (for the answer node) or “intermediate”. This means that there exists a bijection
796 of the sets of vertices of the two graphs that preserves both edges and labels. Note that, while
797 KGs are directed graphs, when computing isomorphism types we consider the answer subgraph
798 as undirected, as KG retrievers should ideally be able to retrieve relevant edges independently of
799 their direction. To provide identifiers for the different graph isomorphism types, the notation based
800 on projections and intersections that is sometimes used in KG reasoning papers (e.g., in Das et al.
801 (2022)) is unfortunately not able to describe complex graph structures without ambiguities. Since,
802 as a data quality requirement, we only consider queries where the answer subgraph is a tree, we
803 adopt a simplified/more readable version of the tree encoding scheme classically used in the AHU
804 algorithm (Aho & Hopcroft, 1974). The answer subgraph is seen as a tree rooted in the answer node,
805 while seed entities are the leaves of the tree. For each leaf, we write (n) if n is the distance of the
806 leaf from its closest branching point (this represents a projection of length n from the seed node).
807 The intersection of paths from multiple seeds at a branching point is denoted by juxtaposition, e.g.
808 $(n)(m)$. Further projections after an intersection point are represented using an additional level of
809 brackets, e.g. $(k(n)(m))$, with k the length of the projection (which can be omitted if $k = 1$). See
810 Table A1 for notations and graphical representations of all the graph isomorphism types appearing in
811 GTSQA.

Table A1: Isomorphism types of ground-truth answer subgraphs in GTSQA and their frequencies in the train and test split. In the graph visualisations, seed nodes are represented in red, while the answer node is marked in green. For the test set, we count separately questions where the answer subgraph is of an isomorphism type not present in the train set (unseen graph type, ugt), questions involving relation types not appearing in the train set (unseen relation type, urt) and questions where the answer subgraph is in-distribution with the train set (id). Percentages are based on the total sizes of the train and test set, respectively.

Identifier	Visualisation	Dutt et al. (2023)	Train		Test (ugt)		Test (urt)		Test (id)	
			count	%	count	%	count	%	count	%
(1)		Iso-0	7658	25.1	0	0	100	6.17	50	3.08
(2)		Iso-1	8338	27.4	0	0	100	6.17	50	3.08
(1)(1)		Iso-2	4481	14.7	0	0	100	6.17	50	3.08
(2)(1)		Iso-3	1470	4.82	0	0	77	4.75	50	3.08
((1)(1))		Iso-4	0	0	125	7.71	0	0	0	0
(3)		Iso-5	1731	5.68	0	0	130	8.01	50	3.08
(2)(2)		Iso-6	0	0	139	8.57	0	0	0	0
(3)(1)		Iso-7	0	0	150	9.25	0	0	0	0
((2)(1))		Iso-8	0	0	65	4.01	0	0	0	0
((1)(1))(1)		Iso-9	0	0	60	3.7	0	0	0	0
(2(1)(1))		Iso-10	0	0	55	3.39	0	0	0	0
(1)(1)(1)		Iso-11	5526	18.1	0	0	48	2.96	50	3.08
((1)(1)(1))		Iso-12	144	0.47	0	0	0	0	0	0
(2)(2)(1)		Iso-13	24	0.08	0	0	0	0	0	0
((3)(1))		Iso-14	1	0	0	0	0	0	0	0
(4)(1)		Iso-15	15	0.05	0	0	0	0	0	0
(4)		Iso-16	0	0	50	3.08	0	0	0	0
((1)(1))(2)		Iso-17	10	0.03	0	0	0	0	0	0
(2)(1)(1)		Iso-18	907	2.98	0	0	0	0	0	0
((2)(1)(1))		Iso-19	3	0.01	0	0	0	0	0	0
((1)(1))(1)(1)		Iso-22	27	0.09	0	0	0	0	0	0
(2)(1)(1)		Iso-23	5	0.02	0	0	0	0	0	0
((1)(1)(1))(1)		Iso-25	18	0.06	0	0	0	0	0	0
((1)(1)(1))(2)		Iso-26	1	0	0	0	0	0	0	0
(1)(1)(1)(1)		N/A	0	0	123	7.58	0	0	0	0
(1)(1)(1)(1)(1)		N/A	111	0.36	0	0	0	0	0	0
(5)		N/A	7	0.02	0	0	0	0	0	0

864
865

A.3 REDUNDANCY

866
867
868
869
870
871
872
873
874
875
876
877
878

As explained in Section 3 (step 3), when filtering the LLM-generated data we check that all seed entities identified by the LLM are used in the SPARQL query. This, however, does not guarantee that all of them are strictly necessary to answer the question, i.e., impose additional constraints compared to the other seeds. A question q contains *redundant* information, if there exists a subset $S' \subsetneq S$ such that the SPARQL query $l_{q,S'}$ corresponding to the subtree $\mathcal{G}_{S'} \subset \mathcal{G}$ composed by the paths from the seeds in S' to the answer node a returns the same set of answers \mathcal{A} as the original query l_q (while, in general, we expect the set of answers of $l_{q,S'}$ to be a strict superset of \mathcal{A}). In the case of redundant information, we identify the smallest such S' (there can be more than one) and classify the graph isomorphism type of $\mathcal{G}_{S'}$, which should be considered as the “minimal” subgraph(s) sufficient to answer the question. Note that questions with redundant information are still perfectly valid, as all seed entities are mentioned in the question and the information they imply should be therefore retrieved from the KG; however, they present “shortcuts” to the answer that effectively reduce the question complexity, compared to what is measured by the graph isomorphism type of the ground-truth answer subgraph.

879
880
881
882
883
884

For example, the question “Which musical instrument is played by both Yehonatan Geffen’s child and Francis Lickerish?” has two seed entities: $S = \{\text{Yehonatan Geffen (Q2911403)}, \text{Francis Lickerish (Q3720616)}\}$. The corresponding SPARQL query

```
SELECT ?answer WHERE { wd:Q2911403 wdt:P40 ?child .
?child wdt:P1303 ?answer . wd:Q3720616 wdt:P1303 ?answer . }
```

885
886
887
888
889

has a unique answer in Wikidata, namely *guitar*. The ground-truth answer subgraph contains three edges, (Yehonatan Geffen; child; Aviv Geffen), (Aviv Geffen; instrument; guitar), (Francis Lickerish; instrument; guitar), with isomorphism type (2)(1). The SPARQL query decomposes as intersection of two projections, originating from the seed entities. We can look at them separately.

890
891

- For seed node *Yehonatan Geffen*, the 2-hop path connecting it to the answer node is encoded by the query

892
893

```
SELECT ?answer WHERE { wd:Q2911403 wdt:P40 ?child .
?child wdt:P1303 ?answer . }
```

894

which returns three different answers: *guitar, piano, voice*.

895
896
897
898
899

- For seed node *Francis Lickerish*, the 1-hop path connecting it to the answer node is encoded by the query `SELECT ?answer WHERE { wd:Q3720616 wdt:P1303 ?answer . }`, which returns a single answer, *guitar*.

900
901
902
903
904

Therefore, this question contains redundant information, as it can be satisfactorily answered using just the seed entity *Francis Lickerish* (while using the other seed entity alone is not sufficient). The minimal answer subgraph is $\{\text{(Francis Lickerish; instrument; guitar)}\}$, with isomorphism type (1).

905
906
907

A.4 EXAMPLES

We display, for exemplificative purposes, one datapoint from the train set of GTSQA, generated from Wikidata with the SynthKGQA framework.

908
909
910
911
912
913
914
915
916
917

Example
<pre>"id": 40513, "question": "Who directed the Italian film, originally in French, that is based on 'The Vicomte of Bragelonne: Ten Years Later'?", "paraphrased_question": "Who was the director of the Italian film, originally in French, inspired by 'The Vicomte of Bragelonne: Ten Years Later'?", "seed_entities": ["Italy (Q38)", "French (Q150)", "The Vicomte of Bragelonne: Ten Years Later (Q769001)"], "answer_node": "Fernando Cerchio (Q503508)",</pre>

```

918 "answer_subgraph": [ ["Le Vicomte de Bragelonne (Q3228085)", "country
919 of origin (P495)", "Italy (Q38)" ], [ "Le Vicomte de Bragelonne
920 (Q3228085)", "original language of film or TV show (P364)", "French
921 (Q150)" ], [ "Le Vicomte de Bragelonne (Q3228085)", "based on (P144)" ,
922 "The Vicomte of Bragelonne: Ten Years Later (Q769001)" ], [ "Le
923 Vicomte de Bragelonne (Q3228085)", "director (P57)" , "Fernando
924 Cerchio (Q503508)" ] ],
925 "sparql_query": "SELECT ?answer WHERE { ?film wdt:P495 wd:Q38;
926 wdt:P364 wd:Q150; wdt:P144 wd:Q769001; wdt:P57 ?answer. }",
927 "all_answers_wikidata": [ "Q503508", "Q679016" ],
928 "full_answer_subgraph_wikidata": [ [ "Q2260875", "P495", "Q38" ],
929 [ "Q2260875", "P364", "Q150" ], [ "Q2260875", "P144", "Q769001" ],
930 [ "Q2260875", "P57", "Q679016" ], [ "Q3228085", "P495", "Q38" ],
931 [ "Q3228085", "P364", "Q150" ], [ "Q3228085", "P144", "Q769001" ],
932 [ "Q3228085", "P57", "Q503508" ] ],
933 "all_answers_wikikg2": [ "Q503508" ],
934 "full_answer_subgraph_wikikg2": [ [ "Q3228085", "P364", "Q150" ],
935 [ "Q3228085", "P57", "Q503508" ], [ "Q3228085", "P144", "Q769001" ],
936 [ "Q3228085", "P495", "Q38" ] ],
937 "n_hops": 2,
938 "graph_isomorphism": "((1)(1)(1))",
939 "redundant": True,
940 "minimal_graph_isomorphism": "((1)(1))",
941 "minimal_seeds_and_queries": { "Q150-Q769001": "SELECT ?answer WHERE {
942 ?a wdt:P364 wd:Q150. ?a wdt:P57 ?answer. ?a wdt:P144 wd:Q769001. }" }
943 "test_type": [ ],
944
945

```

946 Note that `answer_node` and `answer_subgraph` are, respectively, the answer node $a \in$
947 \mathcal{E} and ground-truth answer subgraph $\mathcal{G} \subset \mathcal{T}$ generated by the LLM together with the
948 question. The `sparql_query` is then executed on Wikidata and WikiKG2 to retrieve
949 all answers in the KGs (`all_answers_wikidata`; `all_answers_wikikg2`) and, after
950 converting it to CONSTRUCT form, the full answer subgraphs realizing the query
951 (`full_answer_subgraph_wikidata`; `full_answer_subgraph_wikikg2`). For the
952 example above, we find that Wikidata (but not WikiKG2) contains one more acceptable
953 answer (Henri Decoin, Q679016), due to the existence in the KG of a second movie
954 satisfying all requirements in the question (Le Masque de fer, Q2260875); as a consequence,
955 `full_answer_subgraph_wikidata` contains four more edges compared to \mathcal{G} , arising from
956 these extra valid substitutions for the `?film` and `?answer` variables. Note that we only consider
957 the answer graph \mathcal{G} when computing the `graph_isomorphism`, as that encodes the logical steps
958 required to reason over the question. However, the full answer subgraph should be used as target to
959 evaluate the performance of KG retrievers.

960 `n_hops` measures the maximum distance (in \mathcal{G}) between a seed entity and the `answer_node`;
961 it is determined in a unique way by `graph_isomorphism`. The question in the example
962 contains redundant information (`redundant`); in the `minimal_seeds_and_queries` dictionary
963 we provide the minimal set(s) \mathcal{S}' of seed entities and the respective SPARQL queries
964 $l_{q,\mathcal{S}'}$, as explained in Appendix A.3. The graph isomorphism of the minimal $\mathcal{G}_{\mathcal{S}'}$ can be found in
965 `minimal_graph_isomorphism`. Finally, the attribute `test_type` is only used for questions
966 in the test split of GTSQA, to classify their generalization type (in-distribution, unseen graph type,
967 unseen relation type; see Section 4).

968 To judge the naturalness and reasonability of questions produced by the SynthKGQA framework,
969 we present three randomly sampled questions for every graph isomorphism type in GTSQA with at
970 least 50 questions in the training set. We observe that the paraphrasing operated by the LLM helps to
971 make the questions sound more natural and human-like.

972

Example questions and their paraphrased versions

973

974 Graph isomorphism type (1)

975

Question: On which continent is Palmer Land located?

976

Paraphrased Question: Which continent is Palmer Land situated on?

977

Question: What was the military rank of Pierre Gaston-Mayer?

978

Paraphrased Question: What military rank did Pierre Gaston-Mayer hold?

979

Question: What was the noble title held by Sir Edward Kerrison, 1st Baronet?

980

Paraphrased Question: What noble title did Sir Edward Kerrison, 1st Baronet, hold?

981

982 Graph isomorphism type (2)

983

Question: Who is the head of government of the capital of the canton of Mérignac-1?

984

Paraphrased Question: Who is the leader of the government in the capital of the Mérignac-1 canton?

985

Question: Who is the producer of the series that the episode 1930-talet is part of?

986

Paraphrased Question: Who produced the series that includes the episode titled "1930-talet"?

987

Question: Who composed the anthem of Tuvalu?

988

Paraphrased Question: Who is the composer of Tuvalu's national anthem?

989

990 Graph isomorphism type (3)

991

Question: In which country is the administrative entity in which Arhavi is located, located?

992

Paraphrased Question: Which country is home to the administrative region where Arhavi is situated?

993

Question: Where is the mother of the spouse of Ruprecht V of Nassau buried?

994

Paraphrased Question: Where is Ruprecht V of Nassau's mother-in-law buried?

995

Question: Who is the architect of the building occupied by the sports team that Raitis Grafs played for?

996

Paraphrased Question: Who designed the building where the sports team that Raitis Grafs played for is located?

997

998 Graph isomorphism type (1)(1)

999

Question: Which company founded by Henry Herbert Collier has its headquarters in Plumstead?

1000

Paraphrased Question: Which company, established by Henry Herbert Collier, is headquartered in Plumstead?

1001

Question: Which resident of District 2 was killed by Thresh?

1002

Paraphrased Question: Who was the resident of District 2 that Thresh killed?

1003

Question: Which film produced by UK Film Council was based on The Picture of Dorian Gray?

1004

Paraphrased Question: What movie produced by the UK Film Council was inspired by The Picture of Dorian Gray?

1005

1006 Graph isomorphism type (1)(1)(1)

1007

Question: Which singer, born in Suphan Buri, holds citizenship of Thailand and plays the guitar?

1008

Paraphrased Question: Which Thai singer from Suphan Buri plays the guitar?

1009

1010

1011

1012

1013

1014

1015

1016

1017

1018

1019

1020

1021

1022

1023

1024

1025

1026 Question: Which film produced by Metro-Goldwyn-Mayer, whose main
 1027 subject is baseball, was produced by Clarence Brown?
 1028 Paraphrased Question: What baseball-themed film made by
 1029 Metro-Goldwyn-Mayer was directed by Clarence Brown?
 1030 Question: Who participated in the 2002 FIFA World Cup and played
 1031 for both GNK Dinamo Zagreb and Sevilla FC?
 1032 Paraphrased Question: Which player took part in the 2002 FIFA
 1033 World Cup and played for both GNK Dinamo Zagreb and Sevilla FC?
 1034 Graph isomorphism type (2) (1)
 1035 Question: Which team that plays association football is the
 1036 occupant of the stadium operated by Fenerbahçe Sports Club?
 1037 Paraphrased Question: Which football team plays at the stadium
 1038 managed by Fenerbahçe Sports Club?
 1039 Question: What is the city that is both the headquarters location
 1040 of the developer of Star Wars: Shadows of the Empire (video game)
 1041 and the narrative location of Sudden Impact?
 1042 Paraphrased Question: Which city serves as the headquarters of the
 1043 developer for the video game Star Wars: Shadows of the Empire and
 1044 is also the setting for Sudden Impact?
 1045 Question: Which street, used for utility cycling, serves as a
 1046 terminus for a street that is itself a terminus at Frankfurter Tor?
 1047 Paraphrased Question: Which street designed for utility cycling
 1048 acts as a terminus for another street that ends at Frankfurter Tor?
 1049 Graph isomorphism type ((1) (1) (1))
 1050 Question: Which cast member acted in a Spanish-language film that
 1051 was filmed in Mexico and belongs to the exploitation film genre?
 1052 Paraphrased Question: Which cast member starred in a
 1053 Spanish-language film shot in Mexico that falls under the
 1054 exploitation genre?
 1055 Question: Who is the head of government of the municipality that
 1056 is contained within Valdizarbe and shares borders with both Uterga
 1057 and Enériz?
 1058 Paraphrased Question: Who is the head of government for the
 1059 municipality in Valdizarbe that borders both Uterga and Enériz?
 1060 Question: Which award was received by the English-language film
 1061 directed by Terence Davies and starring Freda Dowie?
 1062 Paraphrased Question: What award did the English-language film
 1063 directed by Terence Davies and featuring Freda Dowie win?
 1064 Graph isomorphism type (2) (1) (1)
 1065 Question: Which goalkeeper who participated in the 1958 FIFA World
 1066 Cup died in a town located in the Eastern Time Zone?
 1067 Paraphrased Question: Which goalkeeper from the 1958 FIFA World
 1068 Cup passed away in a town in the Eastern Time Zone?
 1069 Question: Which businessperson who was a cast member of a reality
 1070 television show also attended the University of Pennsylvania?
 1071 Paraphrased Question: Which reality TV star who is also a
 1072 businessperson went to the University of Pennsylvania?
 1073 Question: Which person died of pneumonia in Vienna and has an
 1074 asteroid belt minor planet named after them?
 1075 Paraphrased Question: Who passed away from pneumonia in Vienna and
 1076 has a minor planet in the asteroid belt named in their honor?
 1077 Graph isomorphism type (1) (1) (1) (1) (1)
 1078 Question: Which musical film directed by Challis Sanderson
 1079 features both Robb Wilton and Kitty McShane as cast members and
 1079 has Desmond Dickinson as its director of photography?

1080 Paraphrased Question: What musical film directed by Challis
 1081 Sanderson stars Robb Wilton and Kitty McShane, with Desmond
 1082 Dickinson as the director of photography?
 1083 Question: Which video game is both an action and platform game,
 1084 distributed on Nintendo game card for Nintendo DS and published in
 1085 North America?
 1086 Paraphrased Question: What action and platform game is available
 1087 on a Nintendo game card for the Nintendo DS and was published in
 1088 North America?
 1089 Question: Which minor planet in the asteroid belt was discovered
 1090 at Osservatorio Astronomico Sormano by both Francesco Manca and
 1091 Piero Sicoli and is directly preceded by 9110 Choukai?
 1092 Paraphrased Question: What is the name of the minor planet in the
 1093 asteroid belt, discovered by Francesco Manca and Piero Sicoli at
 1094 the Osservatorio Astronomico Sormano, that comes right before 9110
 1095 Choukai?

1096 B STATISTICS OF GTSQA

1100
 1101 We use ogbl-wikikg2 (Hu et al., 2020) as the base KG to construct GTSQA. This is a KG extracted
 1102 from a 2015 Wikidata dump, containing a curated set of 2.5M nodes and 535 relation types, that we
 1103 find are good candidates for the construction of natural-sounding questions. As part of the validation
 1104 and filtering pipeline, by leveraging the generated SPARQL queries, we reject datapoints where any
 1105 of the edges in the ground-truth answer subgraph inside ogbl-wikikg2 encode stale facts, i.e., edges
 1106 that are not contained in the most up to date version of Wikidata³. In the future, GTSQA can be easily
 1107 kept up to date by repeating the filtering process against new Wikidata dumps, or by replacing stale
 1108 ground-truth subgraphs with the current ones, as retrieved from Wikidata via the provided SPARQL
 1109 queries.

1110 An overall view of the statistics of GTSQA is provided in Table 1. The questions in the dataset
 1111 involve 68,520 unique entities, drawn from ogbl-wikikg2. The distribution of relation types is shown
 1112 in more detail in Figure B1: out of the 368 unique relation types used in the dataset, only the top-200
 1113 (when sorting by overall frequency) appear in the train set, while the remaining ones are reserved for
 1114 testing. Figure B2 shows the distribution of the number of seed entities, hops (in the ground-truth
 1115 answer subgraph) and answers. Questions in the test set are significantly harder, requiring to perform
 1116 more hops, or to combine reasoning chains from multiple seed entities. Our data-filtering pipeline
 1117 focuses on selecting highly-factual and non-ambiguous questions; as a consequence, 73.9% and
 1118 85.9% of questions in the train and test set, respectively, have a single answer in Wikidata, and only a
 1119 negligible fraction have more than five.

1120 We also compare the size of the ground-truth answer subgraph and the full answer subgraph retrieved
 1121 from Wikidata by submitting the SPARQL query in CONSTRUCT form. As recalled in Appendix A.4,
 1122 the full subgraph may contain additional edges, originating from the presence of multiple answers
 1123 and/or multiple choices for the intermediate entities not specified in the query. In practice, as shown in
 1124 Figure B3, for GTSQA this excess in the number of edges remains always limited (only in 28.4% and
 1125 15.5% of train and test questions, respectively, the two graphs do not coincide). Finally, we report on
 1126 statistics on redundant information in the train set of GTSQA (as stressed in Section 4, no redundancy
 1127 is present in test questions). We find that 25.76% of training questions contain some degree of
 1128 redundancy; Figure B4 shows the distributions of graph isomorphism types for the ground-truth
 1129 answer subgraph \mathcal{G} and the minimal subgraph $\mathcal{G}_{\mathcal{S}'}$ (see Appendix A.3).

1130
 1131
 1132
 1133 ³At the time of dataset construction: <https://dumps.wikimedia.org/wikidatawiki/20250720/>

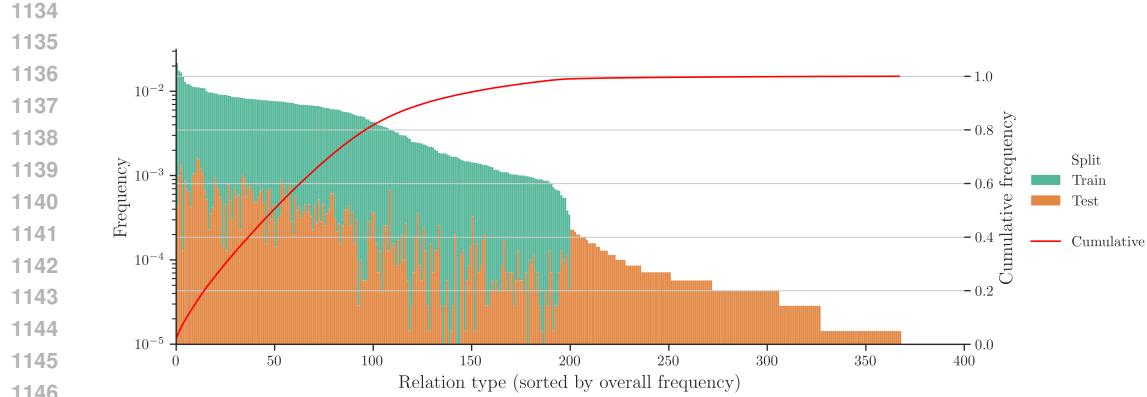


Figure B1: Frequency of relation types of edges in the ground-truth answer subgraphs of questions in GTSQA. The 168 least-occurring relation types (tail of the distribution) are reserved for questions in the test set, to test zero-shot generalization abilities of KG retriever models.

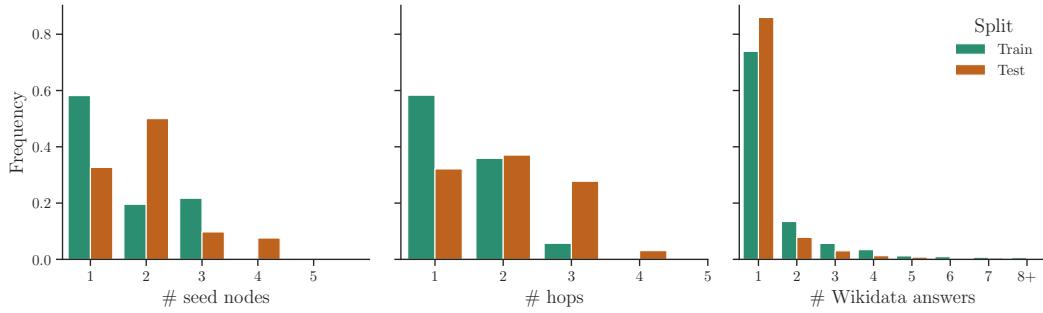


Figure B2: Statistics on the number of seed entities (left), the number of answer entities in Wikidata (right), and the maximum number of hops from seed to answer nodes along ground-truth paths (center), for questions in the train and test set of GTSQA.

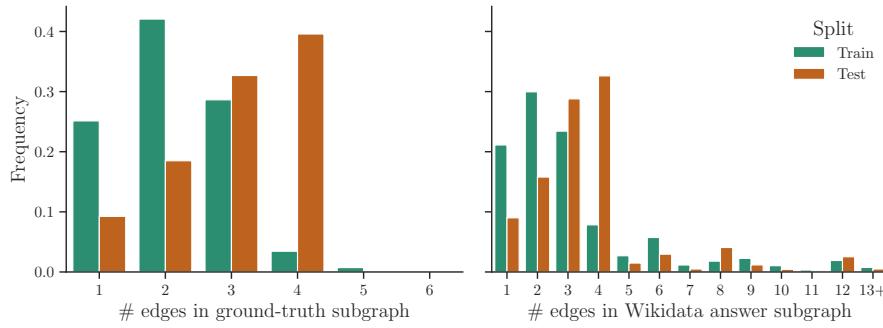


Figure B3: Statistics on the number of edges in the ground-truth answer subgraph (left) and the full answer subgraph in Wikidata (right), for questions in the train and test set of GTSQA. The full answer subgraph can contain more edges than the ground-truth subgraph, if the question has multiple answers, or if any of the intermediate nodes is not uniquely determined.

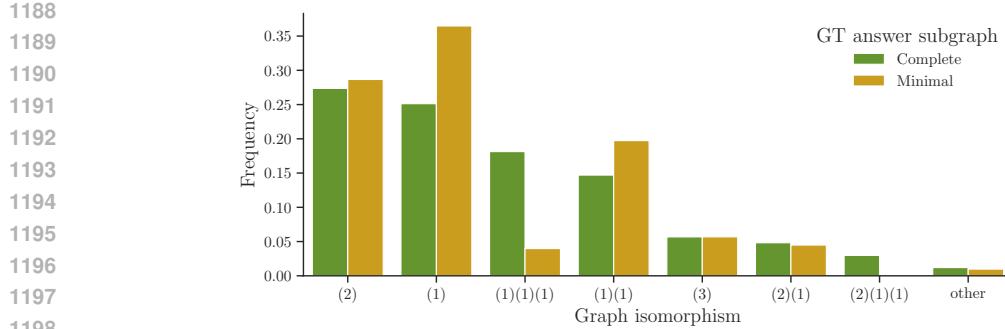


Figure B4: Frequency of main graph isomorphism types in the train set of GTSQA, distinguishing between the complete ground-truth answer subgraph \mathcal{G} and the minimal subgraph $\mathcal{G}_{S'}$, obtained from the complete one after discarding redundant information.

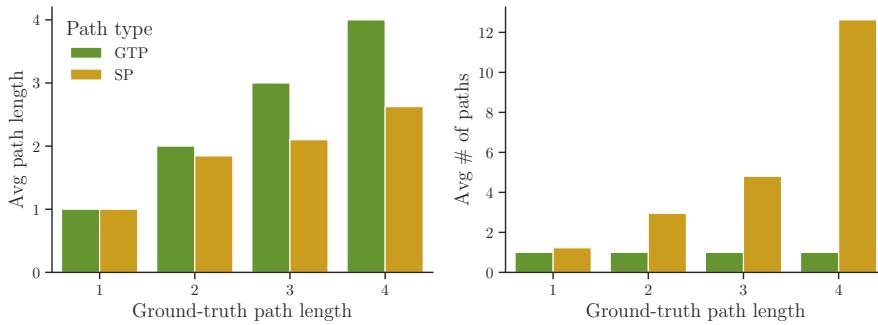


Figure B5: Comparison of ground-truth paths (GTP) and shortest paths (SP) connecting a question seed node to the answer node, in the train split of GTSQA. *Left*: for questions where the ground-truth paths require multiple hops, the distance from the seed node to the answer node along the shortest path increases sub-linearly (shortcuts). *Right*: as the distance between seed and answer node increases, the number of parallel paths of minimal length between them grows exponentially.

C ADDITIONAL DETAILS ON EXPERIMENTS

C.1 CONSTRUCTION OF QUESTION-SPECIFIC GRAPHS

An often overlooked fact in KGQA benchmarks is that many state-of-the-art KG retrievers, especially those that need to perform breadth/depth-first exploration of the KG from the seed entities (e.g., Luo et al. (2024), Luo et al. (2025)) or that include Graph Neural Networks in their architectures (e.g., Mavromatis & Karypis (2025), Li et al. (2025)), are too expensive to run on KGs that have more than $\sim 10^5$ edges. This limitation, in practice, is tackled by performing the retrieval on a smaller subgraph of the KG, which is independently sampled for each question, e.g., by taking the full k -hop neighborhood of the seed entities in the KG, and then pruning it down to a few tens of thousands of

Table B2: Overlap of triples in ground-truth (GT) answer subgraph and triples on shortest paths (SP) between seed nodes and answer nodes. Questions are grouped by the maximum number of hops in the ground-truth subgraph (# hops); the other columns report the average metric values.

# hops	% GT triples in SP	% SP triples in GT	# GT triples	# SP triples
1	100.0	91.0	2.37	2.67
2	89.8	52.7	3.08	19.6
3	54.7	29.6	3.6	46.3
4	27.5	13.3	4	100.4

edges with algorithms like Personalized PageRank (Page et al., 1998). This was the case for Luo et al. (2024), where such question-specific graphs were constructed (with $k = 2$) from Freebase (Bollacker et al., 2008) for the questions in WebQSP (Yih et al., 2016) and CWQ (Talmor & Berant, 2018). These graphs were then used by many others in following papers for benchmarking new models on these two widely-used datasets. They are not, however, part of the official datasets, hence there is no guarantee on their adoption. It is important to note, in fact, that the selection of these starting graphs can strongly impact the final performance statistics, potentially over-representing (if they are too easy/small) or under-representing (if they are not checked to still contain ground-truth paths) the retriever’s capabilities. For this reason, retrievers that are benchmarked on different sets of questions-specific graphs should not be directly compared (even though the test questions, and the underlying full KG, are the same), and care should be taken when comparing them with models that instead are able to perform retrieval from the full KG (e.g., Sun et al. (2024)). However, this crucial detail on experimental setup is often not reported in papers, making comparisons unreliable.

To address this problem, together with GTSQA we release an official set of question-specific graphs (each containing up to 30,000 edges) for all questions in the train and test set. These are the only graphs that should be used by anyone wishing to train or benchmark KG-RAG models on GTSQA when retrieving from the full KG is not possible, to ensure fair comparisons. They are extracted from ogbl-wikikg2 with a similar approach to Luo et al. (2024), starting from the full undirected 3-hop (4-hop for questions requiring 4 hops) neighborhood of seed entities, and then pruning it down to the edges connecting the nodes with the top-2500 scores as assigned by Personalized PageRank (with personalization values concentrated in the seed nodes). If any of the edges in the (full) ground-truth answer subgraph of the question have been dropped as a result of pruning, they are re-added to the graph to guarantee that perfect retrieval is still possible. However, we observe empirically that this final step can unfairly bias retrievers towards the ground-truth edges that have been re-added. To ensure a challenging task, we also add to the graph (as confounders) all edges along paths originating from the seed nodes, of the same metapath (sequence of relation types) as the corresponding ground-truth paths that lead to the answer node.

C.2 SPECIFICATIONS OF EVALUATED MODELS

We provide details and specifications on the KG-RAG models included in our benchmarks. For all of them, we follow original implementations as close as possible.

Think-on-Graph (Sun et al., 2024), Plan-on-Graph (Chen et al., 2024) We modified the original codebases (with the PoG one being based on the one from ToG) to perform retrieval from the ogbl-wikikg2 KG. The search and prune steps of the KG exploration algorithm use a width of 5 and a maximum depth of 4 to enable the retrieval of all paths in the ground-truth answer subgraphs of multi-seed, multi-hop questions in the test set of GTSQA. Note that ToG, by its original implementation, reverts to answering using only the LLM knowledge if the maximum search depth is reached without the model being confident it has retrieved enough information; in this case, we treat the retrieved subgraph as being empty. The LLM performing the graph exploration (and task decomposition and memory updating for PoG) is the same used for the final reasoning, namely GPT-4o-mini.

SR (Zhang et al., 2022) As in the original implementation, we use RoBERTa_{BASE} (Liu et al., 2019) to predict the next relation type $r_N \in \mathcal{R} \cup \{\text{END}\}$ in a path (r_1, \dots, r_{N-1}) from seed to answer node, conditioning on the question q . This is performed by fine-tuning the text-encoder to align the embeddings of $[q; r_1; \dots; r_{N-1}]$ and r_N , with a contrastive approach that uses positive and negative pairs constructed from ground-truth paths (from seed to answer) in the train set. At inference time, paths are predicted by conducting a beam search based on possible continuation candidates; we impose a maximum path length of 4 and set the number of beams to 5. The subgraph is retrieved by looking for all possible (undirected) realizations of the predicted relation paths in the KG starting from the seed entities. Note that, while the original implementation used a Neural State Machine (He et al., 2021) to perform the final reasoning on the retrieved subgraph, we instead use an LLM, to align the last step of the pipeline with the other KG-RAG models evaluated in the paper.

Reasoning on Graphs (Luo et al., 2024) As in the original paper, we fine-tune LLama-2-Chat-7B (Touvron et al., 2023) to auto-regressively predict the relation paths (r_1, \dots, r_N) (and specifying the direction of each edge), originating from the seed nodes, that should be useful to answer a question

1296 *q*. We use all ground-truth paths for the questions in the train set as fine-tuning data. At inference
 1297 time we ask the LLM to propose relation paths using beam search (5 beams) which are then used to
 1298 retrieve the subgraph via breadth-first search. We adopt the plug-and-play version of RoG, which
 1299 allows us to use a different LLM (GPT-4o-mini) to perform the final reasoning on the retrieved triples.
 1300

1301 **Graph-Constrained Reasoning (Luo et al., 2025)** While RoG can hallucinate non-existing relation
 1302 paths, the follow-up work GCR constrains the path decoding to actual paths in the KG. However, this
 1303 comes at significant costs in terms of overhead, as it requires to first index in a KG-Trie all paths (up to
 1304 a fixed length) originating from the seed entities, retrieved via depth-first transversal of the graph. We
 1305 find that, even when working with the smaller question-specific graphs from Appendix C.1, building
 1306 such index for paths of length > 2 requires an unpractical amount of time, which strongly limits
 1307 the applicability of the method to real-world scenarios where the index needs to be built on-the-fly
 1308 (as it happens for questions that have not been seen before). For this reason, we test GCR with a
 1309 maximum path length of 2 (which is also the default for experiments in Luo et al. (2025)), despite
 1310 being aware that a significant fraction of questions in GTSQA require reasoning over longer paths.
 1311 While we still include GCR in the benchmark of GTSQA, these limitations lead us to exclude GCR
 1312 from the case-study in Section 6. As in the original implementation, we use LLama-3.1-8B-Instruct
 1313 (AI@Meta, 2024) as LLM to generate paths, and fine-tune it with the same data used for RoG. At
 1314 inference time, we generate 10 explicit paths from the seed entities through graph-constraint decoding,
 1315 and then discard duplicated paths to obtain the final retrieved subgraph.
 1316

1317 **SubgraphRAG (Li et al., 2025)** SubgraphRAG assigns relevance scores to all edges in the question-
 1318 specific graphs (Appendix C.1), by combining text embeddings with message passing. In particular,
 1319 $p((h, r, t)|q) \propto \text{MLP}([z_q; z_h; z_r; z_t; z_\tau])$, where z_q, z_h, z_r, z_t are text embeddings from gte-large-en-
 1320 v1.5 (Li et al., 2023) for the question q and the labels of h, r, t . The embedding z_τ is constructed
 1321 from the GNN embeddings of the h and t nodes, after 2 layers of message passing starting from
 1322 the one-hot representation of nodes in the graph provided by the labeling trick (1 if the node is a
 1323 seed entity, 0 otherwise; Zhang et al. (2021)). We train GNN and MLP with cross-entropy loss to
 1324 assign high scores to the triples in the ground-truth answer subgraphs, on the train split of GTSQA.
 1325 At inference time, we retrieve the subgraph made of the triples with top-200 scores, similarly to the
 1326 experiments in Li et al. (2025) (after checking that performance starts to plateau when increasing the
 1327 subgraph size further, Figure C7).
 1328

C.3 ADDITIONAL RESULTS

1329 Figure C8 presents a comparison of the recall of ground-truth triples for the evaluated KG-RAG
 1330 models on GTSQA, complementing the data in Figure 2. For trainable models, we also show the
 1331 results disaggregated by generalization type of the test question in Figure C9.
 1332

1333 In Figure C10 we provide additional results for the analysis in Section 6, quantifying improvements
 1334 across different metrics for KG-RAG models trained on the ground-truth subgraphs in GTSQA,
 1335 compared to training on shortest paths between seed and answer nodes, as conventionally done.
 1336

C.3.1 GPT-5-MINI AS REASONING MODEL

1337 We repeat the main analysis of Section 5, but using the newer GPT-5-mini (OpenAI, 2025) as the
 1338 model performing the final reasoning on the KG-retrieved subgraphs, to understand how the choice
 1339 of the LLM influences Hits (EM). While all models score significantly higher than with GPT-4o-mini
 1340 (Table C3), the relative ranking remains the same as observed before, and no different patterns or
 1341 behaviors arise when looking at individual graph isomorphism types (Figure C11) or at different
 1342 generalization abilities required to answer (Figure C12).
 1343

1344 We notice that SubgraphRAG is the model benefiting the most from using a more recent LLM to
 1345 reason over the retrieved subgraph. This appears to be due to an improved utilization of long context,
 1346 filtering out the noise in the augmented prompt when presented with a large number of retrieved
 1347 triples. Indeed, while with GPT-4o-mini we observed a gap between model response quality as
 1348 measured by Hits (EM) and recall of ground-truth triples when the number of retrieved triples was
 1349 increased, GPT-5-mini exhibits continued improvements in Hits (EM) up to 500 triples, matching
 closely the steady increase in GT triple recall (Figure C7). This is reflected, across all models, in

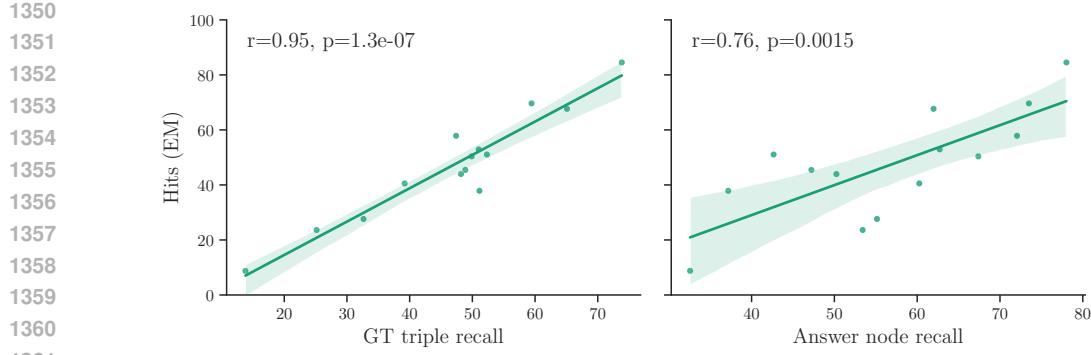


Figure C6: Correlation between Hits (EM) and recall of ground-truth triples (*left*) and of answer nodes (*right*), for questions in the test set of GTSQA. Each dot represents the average performance of the evaluated KG-RAG models on a different graph isomorphism type. While both recall variables have a positive linear correlation with predictive performance, recall of ground-truth triples is a significantly stronger predictor.

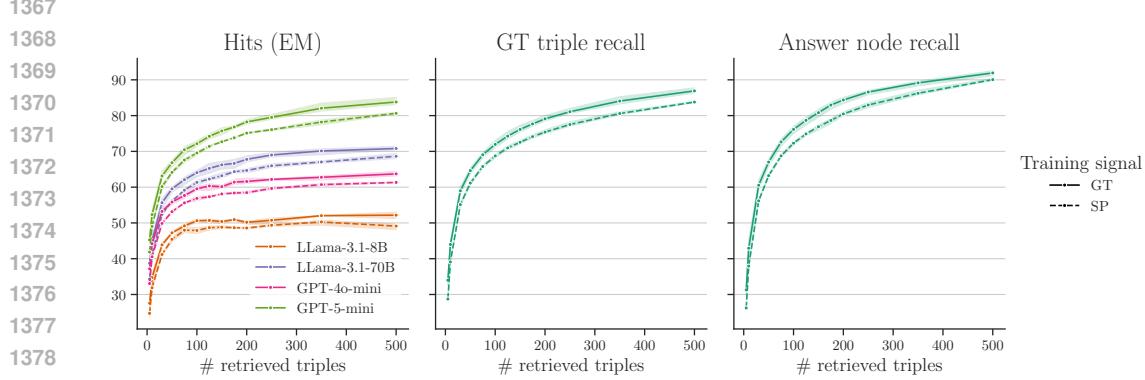


Figure C7: Performance of SubgraphRAG (with different LLMs doing the final reasoning) on the test set of GTSQA, as we increase the size of the retrieved subgraph. More capable LLMs are less sensitive on the noise in the retrieved data, with Hits (EM) tracking more closely the increase in recall of ground-truth triples and answer nodes. Across all LLM and subgraph sizes, training on the ground-truth answer subgraphs (GT) outperforms training on shortest path triples (SP).

an even stronger linear correlation between Hits (EM) and the recall of ground-truth triples, with Pearson correlation coefficient increasing from $r = 0.95$ ($p = 1.3e-07$) for GPT-4o-mini (Figure C6) to $r = 0.97$ ($p = 1.5e-08$) for GPT-5-mini. Training on GT triples (rather than shortest paths) proves to be even more beneficial to the final predictive accuracy than we observed for GPT-4o-mini (Figure C7).

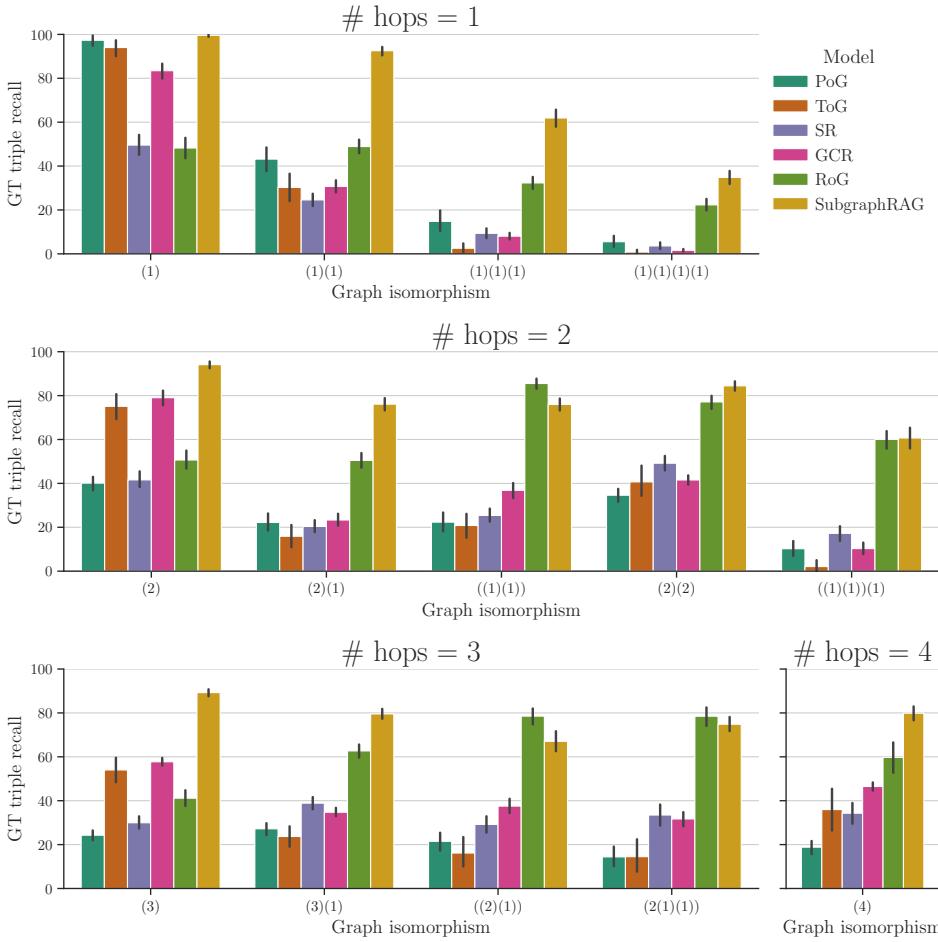


Figure C8: Recall of triples in ground-truth answer subgraph, for KG-RAG models on different graph isomorphism types.

Table C3: Benchmark of KG-RAG models on GTSOA with GPT-5-mini as final reasoning model.

Category	Model	EM		Ground-truth triples			Answer nodes		# triples
		Hits	Recall	Recall	Precision	F1	Hits	Recall	
KG agent	PoG	41.74	40.32	36.07	37.14	34.07	35.76	33.56	3.72
	ToG	53.02	51.19	32.78	7.00	10.80	36.13	35.06	8.00
Path-based	SR	51.75	49.75	30.22	3.44	5.69	50.25	49.39	72.94
	GCR	56.58	54.59	40.71	27.21	29.83	47.10	45.53	6.54
	RoG	67.47	65.09	54.69	24.04	27.00	72.91	71.84	72.15
All-at-once	SubgraphRAG (200)	78.24	74.28	79.09	1.29	2.53	85.33	84.36	199.61



1503 Figure C9: Detailed analysis of generalization abilities of trainable KG-RAG models, on different
1504 graph isomorphism types. We measure EM performance in terms of difference with the EM of the
1505 baseline (GPT-4o-mini, no RAG).

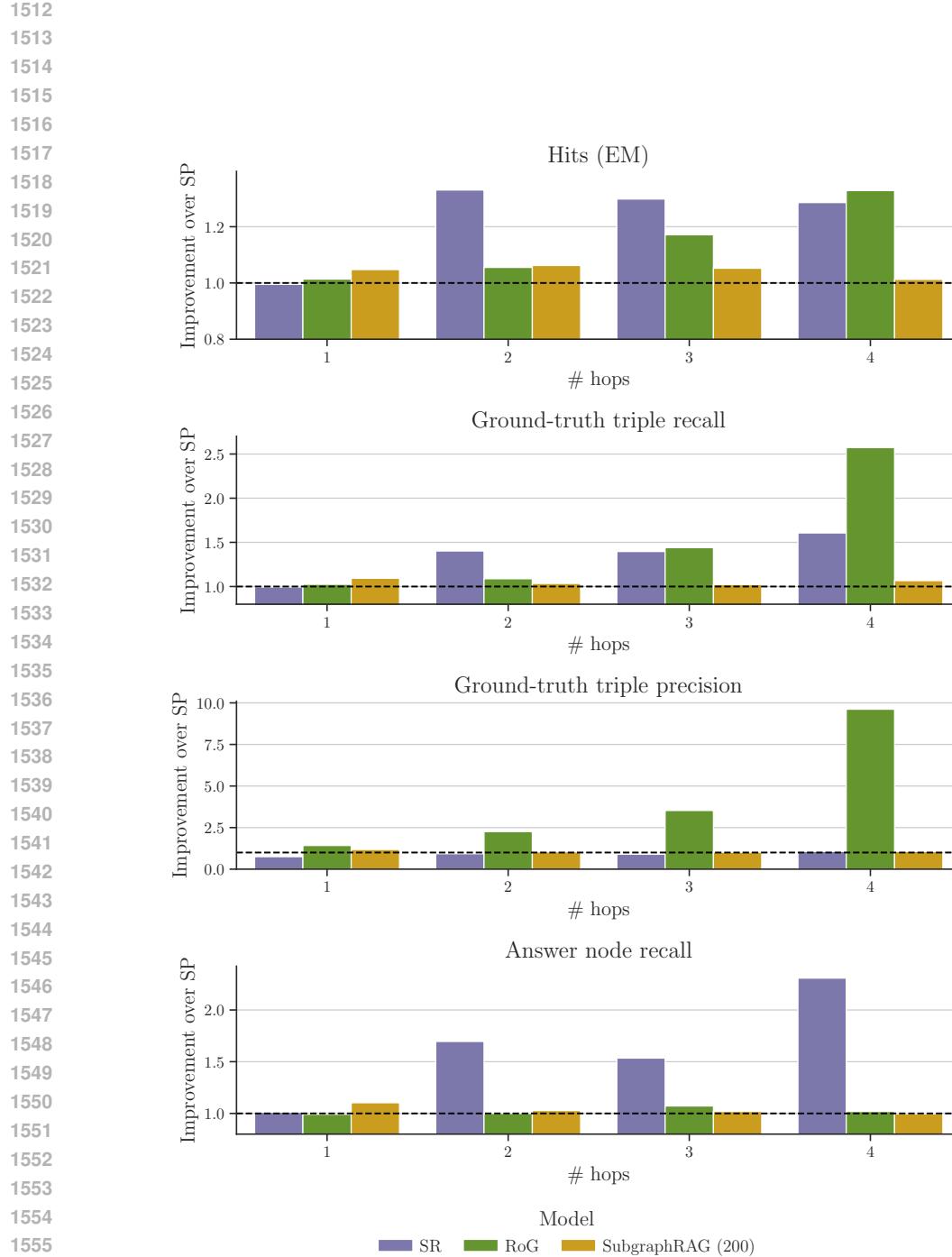


Figure C10: Comparison of predictive statistics when training models on ground-truth (GT) answer subgraphs compared to training on shortest paths (SP), for questions requiring different number of hops. The y-axis measures the GT/SP ratio of the averages of the statistic on the test set (over three distinct training runs); if the ratio is above the dashed line, training on the ground-truth subgraph is better.

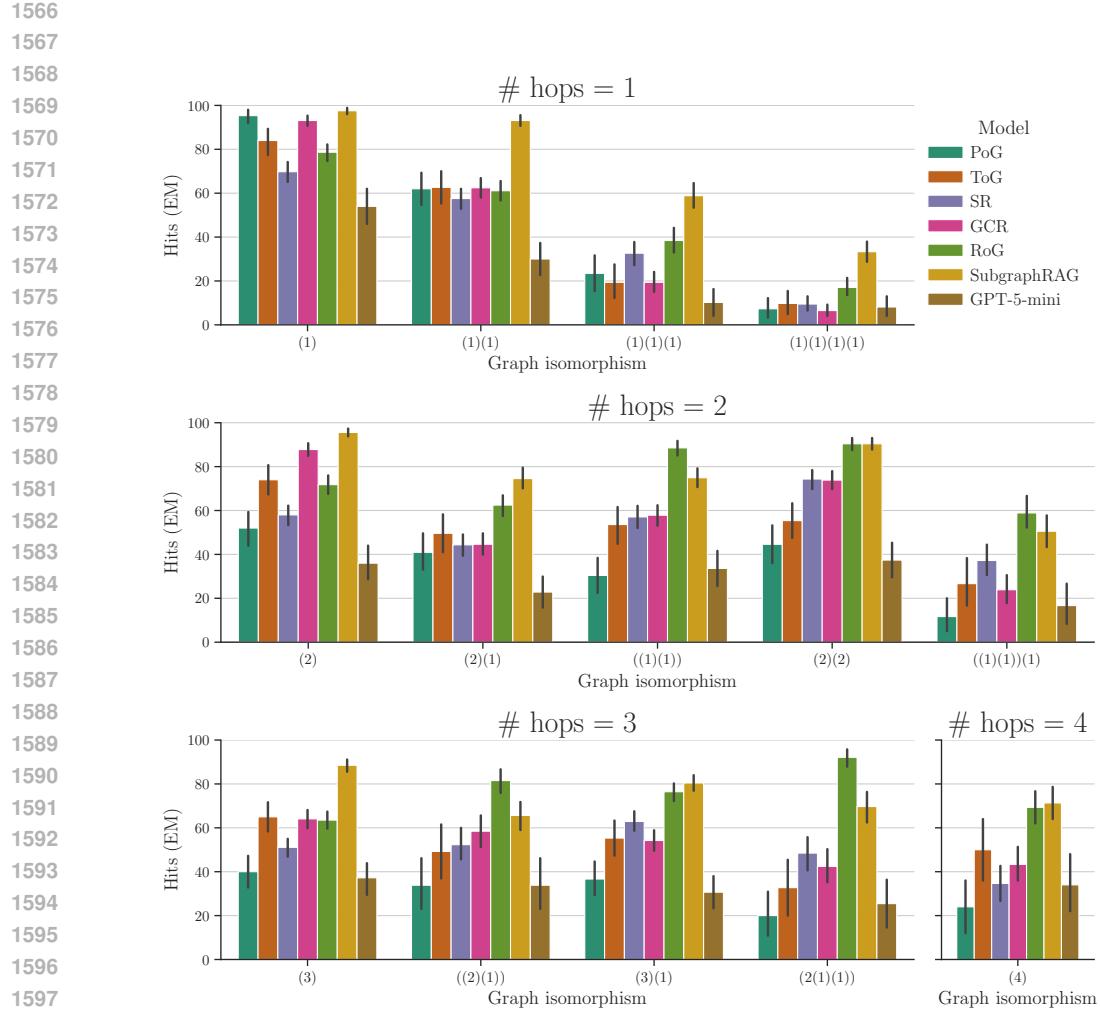


Figure C11: Hits (EM) of KG-RAG models on different graph isomorphism types, compared to the baseline (GPT-5-mini, no RAG).

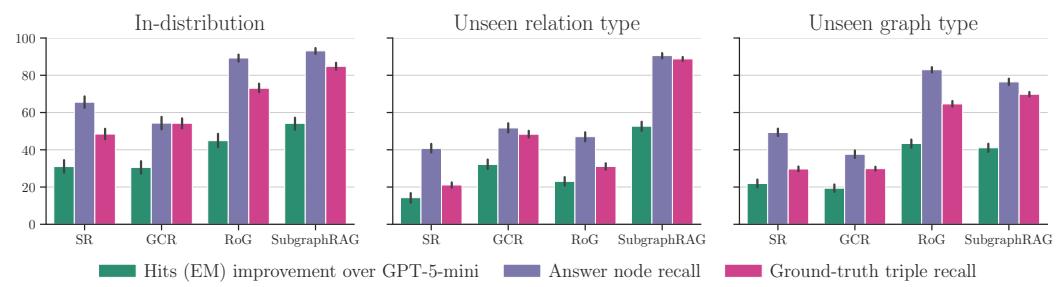


Figure C12: Generalization abilities of trainable KG-RAG models. We measure EM performance in terms of difference with the EM of the baseline (GPT-5-mini, no RAG).

1620 D CASE STUDY OF FAILURE CASES

1622 By presenting explicit examples, we discuss some common failure cases of different KG-RAG
 1623 models, identified through the analysis in Section 5 and manual exploration of the experimental data.
 1624 For all examples, the provided statistics are averaged over three runs of the models.

1626 **Multi-seed questions** As remarked in Section 5, all evaluated models perform poorly when the
 1627 question requires expanding and combining reasoning paths from 3 or more seed entities. RoG and
 1628 SubgraphRAG (and, to a lesser extent, SR) show better performance than other methods, as they
 1629 typically retrieve larger sets of triples, which often mitigates the issue.

1630 *Question:* What common genre do Johan Georg Schwartze and the painting by Alexander Andreyevich
 1631 Ivanov in the Russian Museum share?

1632 *Seed entities:* Johan Georg Schwartze, Alexander Andreyevich Ivanov, Russian Museum.

1634 *Answers:* landscape art.

1635 *Ground-truth answer subgraph:* (Johan Georg Schwartze; genre; landscape art), (Water and Stones
 1636 near Palazzuola; creator; Alexander Andreyevich Ivanov), (Water and Stones near Palazzuola;
 1637 collection; Russian Museum), (Water and Stones near Palazzuola; genre; landscape art).

1638 *Graph isomorphism:* ((1)(1))(1).

1640 *Generalization type:* unseen graph type.

1642 1643 1644 1645 1646 1647 1648 1649 1650	Model	Hits (EM)	Ground-truth triples			Answer nodes	
			Recall	Precision	F1	Recall	# triples
	GPT-4o-mini	0.00	-	-	-	-	-
	PoG	0.00	25.00	14.29	18.18	100.00	7.00
	ToG	0.00	0.00	0.00	0.00	0.00	0.00
	SR	0.00	33.33	1.50	2.82	100.00	107.67
	RoG	100.00	100.00	5.80	10.96	100.00	69.00
	GCR	0.00	25.00	41.67	30.56	100.00	2.67
	SubgraphRAG	0.00	75.00	1.50	2.94	100.00	200.00

- 1651 **GPT-4o-mini:** without any augmentations, the answer provided by the LLM is “Johan Georg
 1652 Schwartze and the painting by Alexander Andreyevich Ivanov in the Russian Museum both
 1653 share the genre of historical painting. The answer is {historical painting}”. No information
 1654 on the painting is provided, making hard to assess where the hallucination originates from.
- 1655 **PoG:** the model is able to retrieve the ground-truth triple (Johan Georg Schwartze; genre;
 1656 landscape art). It also expands the search from the other two seed entities, retrieving multiple
 1657 1-hop paths from each of them, but it is unable to identify the painting that satisfies at the
 1658 same time both the conditions in the question.
- 1659 **ToG:** the model proceeds to iteratively expand the search from only one of the seed nodes
 1660 (Johan Georg Schwartze). When the maximum depth of exploration is reached, since not
 1661 enough information to answer the question has been collected, ToG reverts to only answering
 1662 using the LLM knowledge (Appendix C.2).
- 1663 **SR:** the unique relation paths predicted by the model are: (location, genre); (location, genre,
 1664 field of work); (genre, depicts); (located in the administrative territorial entity). When
 1665 looking for (possibly partial) realizations of these relation paths in the KG (in any direction),
 1666 starting from the seed entities, the model retrieves the ground-truth triple (Johan Georg
 1667 Schwartze; genre; landscape art). It also retrieves (Water and Stones near Palazzuola;
 1668 location; Russian Museum), (Water and Stones near Palazzuola; genre; landscape art),
 1669 through the metapath (location; genre) starting from the seed node “Russian Museum” (a
 1670 parallel path, still valid, compared to the one in the ground-truth answer subgraph). However,
 1671 no relation types pertaining to painting authorship are predicted.
- 1672 **RoG:** the model predicts much better relation paths than SR, namely: (genre); (inverse
 1673 of: collection, genre); (inverse of: creator, genre); (collection, genre). As a result, all
 1674 ground-truth triples are correctly retrieved (together with ~ 65 more).

- **GCR**: as we observe frequently in results from this method, GCR focuses on a single seed entity when decoding graph-constrained paths. Here, after de-duplication of the outputs, we find ourselves with only two paths: *Johan Georg Schwartze* → *genre* → *landscape art*; *Johan Georg Schwartze* → *genre* → *portrait*. Thus, only one triple in the ground-truth answer subgraph is retrieved, namely (*Johan Georg Schwartze*; *genre*; *landscape art*).
- **SubgraphRAG**: the top-200 triples retrieved by the model contain three ground-truth edges, namely: (*Johan Georg Schwartze*; *genre*; *landscape art*), (*Water and Stones* near *Palazzuola*; *creator*; *Alexander Andreyevich Ivanov*), (*Water and Stones* near *Palazzuola*; *collection*; *Russian Museum*). The model is therefore able to identify the painting satisfying the two conditions in the query, but not to make the additional hop from it to the answer node. As highlighted in Section 5, this is likely due to the poor generalization abilities of SubgraphRAG to new graph isomorphisms, as no questions requiring additional projections after the intersection of paths from different seed entities have been observed during training.

Multi-hop questions Questions requiring more than two hops, even in the presence of a single seed entity, pose significant challenges to most KG retrievers.

Question: Who is the lyricist of the national anthem for the country that is home to the Nauru Reed Warbler?

Seed entities: Nauru Reed Warbler.

Answers: Margaret Hendrie.

Ground-truth answer subgraph: (*Nauru Reed Warbler*; *endemic to*; *Nauru*), (*Nauru*; *anthem*; *Nauru Bwiema*), (*Nauru Bwiema*; *lyrics by*; *Margaret Hendrie*).

Graph isomorphism: (3).

Generalization type: unseen relation type (relation “*endemic to*” not included in the train set).

Model	Hits (EM)	Ground-truth triples			Answer nodes	
		Recall	Precision	F1	Recall	# triples
GPT-4o-mini	0.00	-	-	-	-	-
PoG	0.00	33.33	100.00	50.00	0.00	1.00
ToG	100.00	100.00	10.71	19.35	100.00	28.00
SR	66.67	66.67	10.25	17.67	66.67	33.00
RoG	0.00	0.00	0.00	0.00	0.00	0.00
GCR	0.00	66.67	40.00	50.00	0.00	5.00
SubgraphRAG (200)	100.00	100.00	1.50	2.96	100.00	200.00

- **GPT-4o-mini**: the answer provided by the LLM is “The Nauru Reed Warbler is found in Nauru. The national anthem of Nauru is ‘Nauru Bwiema,’ and the lyricist is the former president of Nauru, Hammer DeRoburt. The answer is {Hammer DeRoburt}”. The model is therefore able to correctly identify the nation and the anthem, but it hallucinates the identity of the lyricist.
- **PoG**: the model correctly retrieves the first ground-truth triple (*Nauru Reed Warbler*; *endemic to*; *Nauru*). However, the agent then decides to prematurely stop the search and directly generate the answer (this is a common behavior of PoG on multi-hop questions, as observed in Section 5).
- **ToG**: in a scenario with a single seed entity, the KG retriever is able to effectively expand the search and fetch a set of ~ 30 triples containing all the ground-truth ones.
- **SR**: there are only a few relation types generating from the seed entity, hence the model is able to identify sensible relation paths. In two tries out of three, it predicts: (*IUCN conservation status*, *IUCN conservation status*, *depicts*); (*endemic to*, *anthem*); (*endemic to*, *anthem*, *lyrics by*); (*endemic to*, *anthem*, *composer*). This leads to retrieving all the ground-truth triples.
- **RoG**: the relation paths predicted by RoG are (*country of origin*, *anthem*, *lyrics by*); (*country of citizenship*, *anthem*, *lyrics by*); (*inverse of: has part*, *lyrics by*); (*inverse of: native bird*,

1728 anthem, lyrics by). While the two ground-truth relation types seen during training (“anthem”,
 1729 “lyrics by”) are consistently predicted, the model is unable to come up with good suggestions
 1730 for the unseen relation type “endemic to”, proposing instead known relation types with a
 1731 similar meaning (e.g., “country of origin”) or hallucinating nonexistent relation types, such as
 1732 “native bird”. None of the proposed relation paths is realized in the KG starting from the
 1733 seed entity, hence the set of retrieved triples is empty.

- 1734 • **GCR**: the graph-constrained decoding from the seed entity leads to identifying the following
 1735 paths: *Nauru Reed Warbler* → *taxon rank* → *species* → *inverse of: taxon rank* → *Nauru Reed Warbler*;
 1736 *Nauru Reed Warbler* → *endemic to* → *Nauru* → *anthem* → *Nauru Bwiema*;
 1737 *Nauru Reed Warbler* → *IUCN conservation status* → *Vulnerable* → *inverse of: IUCN*
 1738 *conservation status* → *Nauru Reed Warbler*; *Reed Warbler* → *parent taxon* → *Acrocephalus*
 1739 → *inverse of: parent taxon* → *Nauru Reed Warbler*. While many paths reverse to the seed
 1740 entity, one of them correctly predicts the first two hops. Since we stop the search at depth
 1741 two, due to the costs of the implementation (Appendix C.2), this is the best that the model
 1742 can achieve.
- 1743 • **SubgraphRAG**: the three ground-truth triples are ranked 5th, 8th, 9th with respect to the
 1744 relevance scores assigned by the model, hence they are all consistently retrieved.

1745 E COMPARISON WITH CONCURRENT WORKS

1746 With the latest advancements in LLM reasoning abilities, strong interest has arisen in using generative
 1747 AI to create synthetic datasets for a variety of applications (Long et al., 2024). Two concurrent
 1748 works (Dammu et al., 2025; Zhang et al., 2025) have proposed similar pipelines to the one outlined
 1749 in Section 3 to create synthetic datasets for KGQA. Here, we discuss differences and reciprocal
 1750 advantages of these methodologies in detail.

- 1751 • Dynamic-KGQA (Dammu et al., 2025) is designed to build (question, answer subgraph)
 1752 pairs on the fly using LLMs, starting from compact seed subgraphs in YAGO 4.5 (Suchanek
 1753 et al., 2024) that group triples around a common theme. However, as observed in Zhang
 1754 et al. (2025), the generation pipeline is still prone to hallucinations, with factual correctness
 1755 of questions estimated to be not better than previous benchmarks. Moreover, no logical
 1756 query is executed on the KG: the LLM is tasked to judge whether the proposed answer
 1757 subgraph captures the correct reasoning paths from seed to answer nodes, resulting in higher
 1758 likelihood of incorrect data, compared to our validation approach based on SPARQL queries.
- 1759 • KGQAGen (Zhang et al., 2025) proposes a more cost-efficient validation pipeline, based
 1760 on SPARQL query execution and iterative revisions of incorrect queries (while we directly
 1761 discard datapoints where the SPARQL query does not execute, or returns incompatible
 1762 results with the LLM-generated ones). However, it does not keep track of all the seed entities
 1763 mentioned in the natural-language question, only providing the node from which the seed
 1764 graph is constructed (equivalent to our entity s in Algorithm 1), which limits the usefulness
 1765 for benchmarking KG-RAG models that use multiple seed entities at once. The datasets
 1766 presented in the paper, KGQAGen-10k, contains 10,787 questions with a random 80/10/10
 1767 train/validation/test split (while we carefully curate the split of GTSQA to test zero-shot
 1768 generalization abilities of models), constructed from Wikidata, with seed entities from a set
 1769 of 16,000 Wikipedia’s Vital Articles⁴.

1770 While both Dynamic-KGQA and KGQAGen are presented as multi-hop datasets, neither paper
 1771 includes statistics on the distribution of structures for their respective ground-truth answer subgraphs,
 1772 making hard to actually evaluate the degree of question complexity in these datasets. On the other
 1773 hand, our framework makes this very simple, using the metrics provided by the classification of graph
 1774 isomorphism (Appendix A.2). Similarly, while Dynamic-KGQA uses an LLM as-a-judge to assess the
 1775 presence of redundant information in the question, our approach based on decomposition of SPARQL
 1776 queries and identification of minimal subsets of seed entities sufficient to answer (Appendix A.3)
 1777 uniquely implements an exact measure of redundancy. Finally, GTSQA ensures better reliability as
 1778 a benchmark for KG-augmented LLMs by checking that the ground-truth subgraphs provided can
 1779 indeed be utilized by an LLM to answer test questions correctly.

1780 ⁴https://en.wikipedia.org/wiki/Wikipedia:Vital_articles