Beyond the Metrics: an Investigation into the Reliability of Evaluation Metrics for Domain Specific Graph-based Question Answering

Lia Draetta[1], Marco Antonio Stranisci[1], Flaviana Corallo[1], Pier Felice Balestrucci[1], Michael Oliverio[1], Rossana Damiano[1] and Alessandro Mazzei[1]

[1]Department of Computer Science, University of Turin, Italy

Abstract

Recently, knowledge graph-based approaches have gained wider adoption across domains thanks to their ability to enhance explainability and reduce hallucination in domain specific tasks. Although graph-based architectures have shown promising results, however, evaluation remains an open issue due to the complexity of the analysis and the inherent subjectivity and variability involved when it comes to practical use scenarios and stakeholders' needs. In this context, we present GRADES (Graph-based Reliability Assessment of Domain-specific Evaluation Systems) an evaluation framework for graph-based question answering. To investigate the reliability of the current state-of-the-art evaluation strategies we insert both automatic and qualitative human-based evaluation at each step (information extraction, entity linking and verbalization) of a reference graph-based QA pipeline. At the final step domain experts are engaged to asses both correctness and soundness of the verbalized output. We apply the pipeline and evaluation framework to a case study in the literary domain, showing that the punctual evaluation of each step is able to highlight the limits of off-the-shelf tools in a practical use case.

Keywords

Question Answering, Knowledge Graph, Human-in-the-Loop

1. Introduction

In recent years, Large Language Models (LLMs) have demonstrated remarkable capabilities across a wide range of tasks [1]. Despite these advancements, several scholars have begun to highlight the limitations of such models across multiple levels. For instance, LLMs outputs have been shown to suffer from issues such as hallucinations [2], outdated knowledge [3], and a lack of domain-specific expertise [4]. Additionally, studies have pointed out that LLMs often reflect societal, cultural biases [5] and under-represent marginalized groups [6]. These limitations undermine the explainability and trustworthiness of LLMs, particularly when applied to niche domains [7, 2, 8]. In this context, graph-based retrieval-augmented generation (RAG) approaches, and knowledge graph-based methods more broadly, have shown promising results across various tasks [9, 10, 11]. These approaches have emerged as effective strategies to mitigate the aforementioned limitations by incorporating external knowledge while leveraging the capabilities of LLMs [12, 13], demonstrating strong potential in various fields [14, 15], particularly in areas that require precise and up-to-date information and involve specialized knowledge such as question answering (QA). Despite the growing interest and the potential of graph-based QA, however, the lack of a standardized evaluation framework remains an open research challenge [16, 17].

In this regard, several studies have stressed the importance of integrating traditional quantitative metrics with human feedback, particularly in narrative domains such as literature or digital humanities, to enhance explainability and ensure output reliability [7, 18]. Additional challenges emerge when LLM-generated outputs are evaluated, particularly since scholars have questioned the validity of using LLMs to assess their own responses [19, 20]. These concerns underscore the need for new proposals in the context of hybrid evaluation methodologies that combine human and automatic feedback. Finally,

RAGE-KG 2025: The Second International Workshop on Retrieval-Augmented Generation Enabled by Knowledge Graphs, co-located with ISWC 2025, November 2–6, 2025, Nara, Japan

lia.draetta@unito.it (L. Draetta)

it is important to note that since graph-based QA pipelines are inherently multi-component, relying on a single evaluation strategy carried out on the final output may be overly restrictive.

Having in mind the potential of graph-based QA and the challenges that evaluation sets in this domain (e.g., lack of human involvement, poor reliability on rare entities, lack of multi-step evaluation), we developed GRADES (Graph-based Reliability Assessment of Domain-specific Evaluation Systems), a multi-step evaluation framework that combines automatic and human-generated metrics deployed on a reference graph-based pipeline[1]. Aware of the potential of a collaborative approach [21, 22], we propose a methodology in which evaluation is performed separately in every step (e.g. information extraction, entity linking, triple extraction); for each step an evaluation is provided from both quantitative and qualitative, human-based point of view.

The goal of this approach is to deliver an evaluation framework able to face the still open challenge of integrating quantitative metrics about the extracted and linked entities, and a more qualitative evaluation by domain experts on the soundness and completeness of the answer. Finally, aiming to assess the effectiveness of the methodology, we present a case study in which it is applied to a QA pipeline in the literary domain, selected due to its large size and its capability to encompass mainstream as well as rare entities. This case study highlights how the characteristics of current tools impact the various stages of the pipeline, significantly hindering the final outcome. At the same time, however, it indicates some research directions to integrate graphs and LLMs more effectively in a pipeline tailored to the needs of the reference communities.

The paper is organized as follows: in Section 2, we review the main works on graph-based question answering approaches and their evaluation, highlighting the open challenges in the field. Section 3 provides a detailed description of the pipeline and the evaluation framework. The case study is presented in Section 4. In sections 5 and 6 we respectively present the conclusions and discuss the limitations of our work.

2. Background

Leveraging information retrieved from knowledge graphs to reduce factual errors in various LLM-based tasks has become increasingly common in recent years [23, 13, 24]. Graph-based tasks are typically implemented as a multi-stage pipelines comprising steps such as entity extraction, graph retrieval, triple verbalization, and prompt tuning, depending on the specific methodology employed. In the context of QA, several graph-based approaches have been proposed recently [25, 26, 27, 28], each leveraging the structural advantages of graphs, such as nodes and their relations, in distinct ways.

In terms of evaluation, Graph-RAG approaches are commonly assessed on question answering benchmarks using metrics such as F1 score, accuracy, and recall [29, 26] or leveraging lexical similarity metrics [30] such as ROUGE [31] or BLEU [32]. While several benchmarks for evaluating question answering tasks are now available [33, 34], the evaluation of graph-based methods remains an open challenge [35].

Recent studies [36, 35] have highlighted the limitations of automated evaluation methods, highlighting their poor correlation with human judgment [37] and their fallacy in capturing factuality or faithfulness issues in text [38]. Concurrently, other works [39, 40] are beginning to emphasize the potential of hybrid approaches that combine the strengths of LLMs with human judgment, aiming to balance validity and reliability while minimizing data requirements. These studies emphasize that dimensions such as correctness, clarity, and informativeness are not objective and are often interpreted differently by human evaluators and automated systems. In this context, some recent studies positively adopted a hybrid approach to verify the output of a RAG pipeline. Yu et al. [41] successfully integrate human judgment in the evaluation process of their RAG pipeline to ensure the reliability and robustness of their results. Gienapp and colleagues [42] highlight the limitations of using LLMs to evaluate LLM-generated responses and, consequently, advocate for the validity of human evaluation. To this end, they propose a crowdsourcing-based methodology for the evaluation of RAG systems. While collaborative

[1]code available at: https://anonymous.4open.science/r/Talk-5AC5/README.md

methodologies, well-established in other fields such as human-in-the-loop [43, 44], appear promising, they remain underexplored and relatively novel within the graph-based QA domain.

Moreover, when assessing the generated answers, multiple dimensions must be considered, such as correctness, coherence, completeness, and alignment with the actual needs of stakeholders. Recent studies have proposed LLM-based evaluation frameworks. For example, Es et al. [45] introduce RAGAs, a framework for evaluating Retrieval-Augmented Generation pipelines without reference data. Their methodology assesses Faithfulness, Answer Relevance, and Context Relevance through a multi-step LLM-based evaluation that takes the question and generated answer as input. While the framework is promising, the authors acknowledge its limitations, as it relies heavily on the performance of the LLMs used for evaluation. In addition, since LLMs are known to struggle with handling rare entities [46, 47], basing the evaluation entirely on such models may yield less reliable results.

In our work, acknowledging the current challenges in evaluation, we aim, on one hand, to assess the reliability of state-of-the-art tools for the different steps of the pipeline, and on the other, to propose a framework that integrates automatic metrics with human judgment.

3. Pipeline and Evaluation Design

To test our comprehensive evaluation framework, we developed a graph-based QA pipeline leveraging open-source, state of the art tools. The pipeline comprises several components, including information extraction, entity linking, triple extraction, and triple verbalization. Given a user query, the pipeline first extracts target entities using an LLMs-based approach, and then links these entities to those present in knowledge bases. Subsequently, relevant nodes are retrieved from a knowledge graph using SPARQL queries. A pruned subgraph is constructed, and the most relevant triples are verbalized through fine-tuned LLMs. Figure 1 illustrates the pipeline workflow from user query to system-generated answer and the evaluation steps provided for each phase.

3.1. Information Extraction

The first step of the pipeline is an Information Extraction (IE) task, aimed at extracting relevant entities and entity types from a question formulated in natural language. However, with respect to a standard Named Entity Recognition task, this step includes also generic mentions of entity types to drive the search in the KG. For instance, consider the sentence "Which are the **Italian women** who won the **Nobel** Prize?" where it is possible to identify not only two named entities, but also a demonym (*Italian*) and a prize (*Nobel*) respectively, and the type *women*, which is relevant for retrieval.

Baseline. We use as a baseline for this step `NuExtract-1.5`[2] [48], an open-source, lightweight, text-to-JSON fine-tuned large language model designed to extract complex information from text and organize it into structured data. We adopted the small version of NuExtract (NuExtract-tiny), as it demonstrates remarkable performance even in zero-shot settings when compared to state-of-the-art models, despite its significantly smaller size. Furthermore, prioritizing an open-source solution was essential to ensure the reproducibility of our results and resources. The task consists in completing a predefined template containing empty entity slots with their classes based on the input question. The template was defined by selecting the top-level classes of the leveraged knowledge-base:

```
Input: "entities": {
        "work": [],
        "person": [],
        "subject":[],
        "publisher":[]
    }
```

[2] https://huggingface.co/numind/NuExtract-1.5

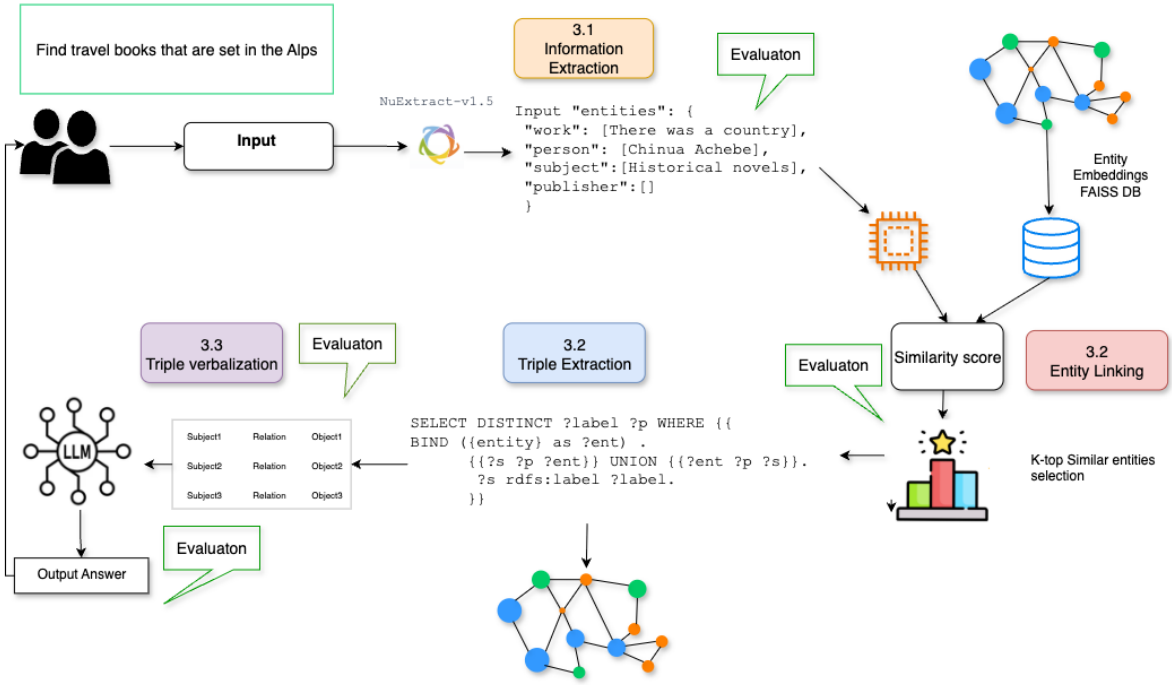

Figure 1: The graph-based QA pipeline. Given an input question, relevant entities are extracted from the question (Information Extraction, Section 3.1), then linked to the most similar ones in the KG, and the relevant information is retrieved from the graph via SPARQL (Entity Linking and Triple Extraction, Section 3.2); finally, the extracted triples (Triple Verbalization, Section 3.3) are verbalized. As shown in the figure, an evaluation strategy is provided for each step.

Evaluation Metrics. In this phase, the evaluation consists of calculating the precision and recall and F1 score of the information extracted by the baseline model against a manually annotated gold standard labeled by domain experts. The manual creation of the gold standard by two researchers enables precise assessment of the baseline model's performance and supports qualitative analysis.

3.2. Entity Linking

The second step of the pipeline is Entity Linking (EL), aimed at linking the extracted knowledge to the corresponding entities in the KG. Since our aim is to evaluate the overall performance of a system in interacting with structured knowledge, the EL phase takes as input the knowledge extracted in the previous phase. Models are fed with entities and their classes and must return the top-k candidates in the KG and all the triples where they appear as subject or object.

Baseline. Our baseline for this step is a vector-based similarity search carried out by adopting two distinct embedding models: `sentence-transformers/ all-MiniLM-L6-v2` [49] and `GIST-Embedding-v0` [50]. Each input entity, previously extracted from a text and associated with its semantic class (e.g., Work, Person), is encoded into a vector representation using both embedding models. These vectors are then used to query a pre-built FAISS [3] index corresponding to the entity type. Each index stores vectorized representations of known entities, enabling efficient nearest neighbor search. Given a query entity, the index returns the top-k most similar candidates based on vector similarity. The retrieved entities are identified via their FAISS index positions and then resolved to their corresponding names and labels using a lookup table stored in Parquet format[4], an open source data file format that enables

[3]https://github.com/facebookresearch/faiss
[4]https://parquet.apache.org/

efficient data storage and retrieval. All linked entities derived from the output template are aggregated into a unified list representing the complete set of candidate links for the input entities. Finally, the top-ranked entities are used to retrieve triples from the knowledge graph. This is accomplished via a SPARQL query that extracts all triples from the knowledge base where the target entity appears as object or subject.

```
SELECT DISTINCT ?label ?p WHERE {{
BIND ({entity} as ?ent) .
  {{?s ?p ?ent}} UNION {{?ent ?p ?s
      }} .
  ?s rdfs:label ?label .
}}
```

Evaluation. This evaluation phase involves manual validation by domain experts. Specifically, the top ten entities linked to each input question are assessed for relevance. Average precision is then computed to quantify the proportion of relevant entities, based on how many of the ten retrieved entities are pertinent. This evaluation serves a dual purpose: it enables a comparative performance analysis of the two embedding models used in the entity linking process and provides insights into the types of entities that are most challenging to link.

3.3. Triples verbalization

Finally, verbalization of the extracted triples is performed, aimed at converting the RDF data in natural language sentences. This step is crucial to assess to what extent a model is able to recognize the semantics of triples in their verbalization.

Baseline. For the triples verbalization, inspired by Oliverio and colleagues [51], three different midsize open-weight English LLMs were fine-tuned: `LLaMA 3.1 8B Instruct`[5], `Qwen 2.5 7B Instruct` [52], and `Mistral-Nemo-Instruct-2407`[6]. The fine-tuning phase was performed using the WebNLG corpus [53], a linguistic resource consisting of data units, each represented as a set of RDF triples (subject, predicate, object) extracted from 15 distinct DBpedia categories. Each data unit is accompanied by one or more human-written verbalizations produced by expert annotators. For the experiments, we adopted WebNLG 3.0, which was released during the WebNLG 2020 Challenge.[7] The dataset was split into train, dev, and test sets, with each data unit containing between 1 and 7 RDF triples. The parameters used to fine-tune the models are shown in Table 1. After fine-tuning, we used the models to generate verbalizations for all the extracted RDF triples.

Evaluation. To evaluate the generated outputs, we structured a two-fold strategy. On one hand, aiming to assess the completeness and the overall correctness of the produced sentences, we followed the taxonomy proposed by Kasner and Dušek [54] and manually assessed 50 randomly sampled triple verbalizations per model. The taxonomy classifies errors into four categories: *Incorrect*, where the text contradicts the data; *Not Checkable*, where the information cannot be verified; *Misleading*, where the text is deceptive given the context or the information is missing; and *Other*, for problematic cases not fitting the other categories.

On the other hand, to assess the relevance of the generated answers and their validity from a domain-specific perspective, a separate annotation task was conducted. Subject matter experts were asked to judge the triples, using a structured evaluation template to asses if the output was pertinent with the

[5] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[6] https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407
[7] https://synalp.gitlabpages.inria.fr/webnlg-challenge/challenge_2020/

Parameter	Value
QLoRA parameters	
LoRA attention dimension	64
Alpha parameter	16
Dropout probability	0.1
bitsandbytes parameters	
Activate 4-bit precision	True
Compute dtype for 4-bit	float16
Quantization type	nf4
Activate nested quantization	False
TrainingArguments parameters	
Number of training epochs	2
Enable fp16 training	False
Enable bf16 training	True
Batch size per GPU for training	4
Batch size per GPU for evaluation	4
Gradient accumulation steps	1
Maximum gradient norm	0.3
Initial learning rate	2e-4
Weight decay	0.001
Optimizer	p_adamw_32bit
Learning rate schedule	cosine
Warmup ratio	0.03

Table 1
Hyperparameters used in the experiments.

input question. This phase of the evaluation goes beyond traditional quantitative metrics by determining whether the extracted, linked, and verbalized information is not only technically correct but also useful in a real-world scenario, providing insight into the applicability of the system's outputs.

4. Case Study: QA for a Specialized Literary Domain

To assess the effectiveness of the proposed pipeline, we present a case study in which the evaluation framework is applied in practice. The leveraged knowledge base is the World Literature Knowledge Graph [55], a collection of writers and works (10.8 million of type human) derived from Wikidata, Goodreads, and Open Library, created to detect and contrast cultural underrepresentation through the integration of minor literary archives from selected research projects. Given the well-known limitations of LLMs in handling entities from niche or specialized domains, this case study focuses on the literary domain, and on less-represented, minority literary traditions. The rationale behind this choice is to evaluate the pipeline in a scenario where domain-specific knowledge is critical but potentially underrepresented in general-purpose language models. The case study consists of an initiative aimed to foster interdisciplinary research on cultural heritage through a networked representation of writers, literary works, and places. To ensure that the case study emphasizes underrepresented entities while maintaining relevance for domain experts, the evaluation focused on a set of specific domains from ongoing research projects (Travel and literature in the French-speaking World, Pyrenees in sounds and pictures, Transylvania and the Banat in British travel writing), and was developed in collaboration with the scholars involved in these projects.

4.1. Case study: Input Question

Recognizing the value of collaborative and participatory approaches in research design [21], we engaged domain experts and collaborators during the initial phase of input design. First, a research team proposed a set of questions in natural language; these questions were subsequently reviewed and validated by two scholars from the foreign literature department.

This process resulted in a final benchmark of 15 different questions to evaluate whether more intricate queries pose greater challenges for the pipeline. Specifically, four questions focused on identifying narratives set in specific geographic regions (e.g., "Find travel books that are set in the Alps"), five combined author's origin and book setting (e.g., "Find books written by French authors and located in Morocco"), four combined the language of the text and the location it is about (e.g., "Find books written in French that talk about the Pyrenees"), and two posed highly specific questions related to books about locations in Romania (e.g., "Find books about Casa Mureșenilor in Brașov").

4.2. Case study: Information Extraction

For each input question, the model was provided with the following template reflecting the semantics of the literary domain:

```
{"work": [], "subject": [], "publisher": [], "person": [],
        "language": [], "date": [], "place": [] }}
```

The task consists of extracting relevant entities from the input sentence and populating the corresponding fields of the predefined template. For the evaluation a gold standard dataset was manually annotated by two expert annotators using the same set of input questions provided to the models (see section 4.1). Annotators were asked to fill the entity templates based solely on the question. Finally, evaluation metrics were computed by comparing the output of NuExtract against the annotated gold standard. Precision and recall were first calculated for each input question, and then averaged to obtain the overall metrics. We report a macro-averaged precision of 0.47, a macro-averaged recall of 0.69, and a macro-averaged F1 score of 0.56. (Results are presented in Table 2.). While the model demonstrates the ability to extract the relevant entities (Recall) from the input sentence, it often struggles to correctly assign them to the appropriate category (Precision).

From a qualitative perspective, it was observed that the entity extraction task appears to be more challenging for simpler questions involving fewer entities. This may be attributed to the model's tendency to fill all available slots in the given template, even when fewer relevant entities are present. In such cases, recall tends to be high, since most relevant entities are retrieved, but precision is relatively low, as the extracted entities are often incorrectly associated with the target classes.

These findings indicate that this step requires model capabilities beyond those addressed by standard information extraction (IE) tasks. For example, extracting information from a query like "Find all books that talk about Marseille" goes beyond traditional Named Entity Recognition (NER), as "book" represents a generic mention rather than a named entity. Moreover, the models demonstrate difficulty in handling rare or less frequent entity types. While they correctly associate "Marseille" with location-related types, they often fail to do so with entity type as "Publisher".

4.3. Case Study: Entity Linking

Following the proposed methodology (3.2) the extracted entities were vectorized, and the k-most similar entities were retrieved using a pre-built FAISS index corresponding to each entity type in the knowledge graph. The evaluation was conducted on the top 10 retrieved entities, with two annotators assessing the relevance of each result. Annotators were asked to indicate how many out of the ten entities were relevant to the input question. They were presented with a list of input questions (e.g. Find travel books that are set in the Alps) each associated with a ranked list of linked entities retrieved by the model (e.g. {'entity': 'urb:urb_subject_6978', 'label': 'alps', 'distance': 4.0611733268837436e-13}), and they were asked to decide how many of the top 10 retrieved entities are relevant to the input question.

Information Extraction	
Metric	**Value**
Macro Precision	0.47
Macro Recall	0.69
Macro F1	0.56
Entity Linking	
SBERT Avg Precision	0.326
GIST Avg Precision	0.32

Table 2
Macro precision, recall and F1 for the IE task and Average precision of the Entity Linking task.

	I↓	NC↓	M↓	O↓
LLaMA 3.1 8B Instruct	0.25	0.00	0.02	0.02
Qwen 2.5 7B Instruct	0.32	0.01	0.00	0.07
Mistral-Nemo-Instruct-2407	0.30	0.00	0.02	0.00

Table 3
Average scores assigned by annotators for each label and model, along with the average agreement metrics. Symbols used: *I* stands for Incorrect, *NC* for Not Checkable, *M* for Misleading, and *O* for Other.

The results of the human evaluation were obtained by calculating the Average Precision at 10 over the entire set of input questions. The two models, SBERT and GIST, demonstrated comparable performance, achieving average precision scores of 0.326 and 0.32, respectively (see results in Table 2). Both models tended to perform better on the input that showed higher precision and recall in the entity extraction task, suggesting that entity linking performance is partially based on the quality of the extracted entities.

The findings highlight that vector similarity correlates closely with surface-level textual similarity, frequently resulting in the linking of entities with similar names or spellings. Furthermore, our results highlight a persistent gap in the field of entity linking: current state-of-the-art models, such as Relik [56], while demonstrating strong overall performance, are typically trained on a limited set of taxonomies (e.g., Wikidata). As a result, they are not easily transferable to other knowledge graphs and exhibit limitations in handling long-tail entities [57].

4.4. Case study: Triple Verbalization

The ten top ranked entities per question are then used to extract triples from the KG. After fine-tuning, a model to generate verbalizations for all RDF triples corresponding to the 15 target questions is used. This process resulted in 358 verbalizations. For each question, the corresponding verbalizations were concatenated. Two of the authors manually evaluated the generations and achieved an agreement of 0.98 for Krippendorff's alpha and 0.96 for Cohen's kappa, indicating an almost perfect agreement. Table 3 shows the results of this manual evaluation.

All three models exhibit a preference for passive constructions even when the original predicate is active. For instance, the triple `Alps, has topic, Melissa Hill` is verbalized by Llama 3.1 8B as "Melissa Hill is a topic related to the Alps''. Among the most recurrent error types are those concerning named entities, which are often modified, e.g., through alterations of book titles or geographic names. Notably, the most frequent and systematic error across all three models involves triples with the predicate *publishing language*, which are regularly erroneously verbalized due to confusion between subject and object or through incorrect paraphrasing. For example, the triple `French, publishing language, Rosa Montero` is incorrectly verbalized by Mistral-Nemo-Instruct-2407 as "Rosa Montero

Task	Metric	Value
Task_1	Cohen's K	0.315
Task_2	Cohen's K	0.154

Table 4
Inter-annotator agreement scores. Task_1: whether the raw triples or their verbalized forms are more informative; Task_2: perceived relevance of the system's output with respect to the input question.

is a French publisher.". This behavior may be attributed to the fact that, during training, the models were never exposed to verbalizations involving this predicate. Importantly, the models almost never exhibit hallucination or omission phenomena. These results suggest that, unexpectedly, the employed models lack awareness of entity types. For example, the triple `Alps has topic Messner` was verbalized as "Reinhold Messner is a topic of the Alps", indicating that the model failed to recognize that Messner is a person (i.e., an agent) and the Alps is a location, more plausibly the subject of a novel or an author's interest.

4.5. Case study: Domain Expert Validation

To validate the extracted triples and their corresponding verbalizations, we involved two researchers from the Department of Foreign Literature. Their role was to evaluate the reliability and relevance of the extracted content. For each instance, the annotators were provided with the input question, the raw triples, and their verbalized versions (some examples are presented in Table 5). They were then asked to answer two evaluation questions: *In the context of the input question, are the raw data or the verbalized triple more explanatory?* (task_1), *Is the extracted knowledge pertinent to the input question?* (task_2).

The opinions of the interviewed experts were not unanimous, and they encountered some difficulties in carrying out the evaluation. The assessment results were poorer than expected, as the majority of outputs were labeled as non-pertinent. This outcome is likely attributable to the limited performance observed in the initial steps of the pipeline.

To assess the difficulty of the annotation tasks and the consistency between annotators, we calculated the inter-annotator agreement (IAA) for both tasks (Table 4). We obtain a Cohen's K of 0, 315 for Task_1 and 0.154 for task_2, indicating slight agreement. To further examine the outputs of the verbalization model in Task_1, a third expert was consulted to resolve cases of disagreement and provide the final judgment.

Out of 48 analyzed examples in 29 cases the annotators judged the verbalized form as more informative than the raw triple representation. However, the analysis revealed that the verbalization model exhibited limited semantic understanding, particularly in interpreting the meaning of relations within the triples. For instance, the input triple `books has topic This is Not the End of the Book`, which refers to the work by Umberto Eco and Jean-Claude Carrière, was verbalized as "This is Not the End of the Book is a topic in books". In contrast, a more common input such as `books has topic Mark Twain's Burlesque Autobiography` was correctly verbalized as "Mark Twain's Burlesque Autobiography is a book." This suggests that the model performs better on more straightforward or lexically familiar inputs. Additionally, it appears that verbalization models face greater challenges when dealing with rare or underrepresented entities. To assess the diffusion and representation of these entities, we examined their presence in Wikidata. We found that the entity "Mark Twain's (Burlesque) Autobiography and First Romance" (wd:Q6770070) is associated with 47 triples, whereas "This is Not the End of the Book" is not present in Wikidata at all. This disparity may indicate that LLMs struggle with handling rare or low-resource entities across a variety of tasks.

Question	Triple	Verbalization
Find travel books that are set in the Alps	alps has topic Mont Blanc aux sept vallées alps has topic Melissa Hill	Melissa Hill is a topic related to the Alps. Melissa Hill is a topic related to the Alps.
Find books written by French and located in Morocco	French publishing language Rosa Montero morocco has topic Abdallah Laroui	Rosa Montero writes in French. Abdallah Laroui is a topic of Morocco.
Find Romanian books that talk about Banat	Romanian publishing language Panait Istrati Romanian publishing language Miodrag Bulatović	Panait Istrati is a Romanian author. AMiodrag Bulatović is a Romanian publisher.

Table 5
The table presents examples of the data analyzed by the two experts during the annotation process. The first column contains the input questions, the second column lists the most relevant triples, and the third column shows their corresponding verbalizations.

5. Conclusion and Future Works

In this work, we presented a human-centered framework for the evaluation of graph-based question answering. We developed a graph-based QA system composed of discrete, interpretable steps, and conducted both quantitative and qualitative evaluations for each phase. Recognizing the value of collaborative approaches, we involved domain experts in the initial phase to validate the soundness of the input questions, as well as in the overall evaluation phase to assess the relevance and quality of the system's outputs. Our focus was particularly centered on the evaluation phase, as the literature indicates that determining how, and through which parameters, exhaustively evaluate such systems remains an open challenge. We first observed that tools based on LLMs often struggle with handling structured knowledge, and that without human supervision, such as evaluation through gold standards or manual checks, they do not appear to be suitable for exhaustively completing knowledge graph-related tasks. The evaluation of the system's output conducted by domain experts provides additional insights on how LLMs struggle to handle rare and underrepresented entities. This issue, which is of high relevance within the AI community, becomes even more critical in the context of graph-based frameworks. Since many pipelines that incorporate external knowledge bases are specifically designed to mitigate the limitations of LLMs, relying solely on LLMs to evaluate such systems is at risk of creating a self-reinforcing loop. As recently argued by several scholars [20, 19], substituting human judgment with LLM-based assessments may be an overly simplistic and potentially misleading approach.

Our work highlights the importance of integrating intermediate evaluation steps into the pipeline and demonstrates that human evaluation remains a crucial component, particularly when assessing the final output. We argue that the outputs generated by such models must be validated by domain experts, as they are ultimately responsible for determining the utility and relevance of the information provided. In addition, our findings highlight the limitations of current models when used for handling semantic data, emphasizing its inherent complexity. Through qualitative evaluation, we provide updated insights into ongoing challenges and offer perspectives on how these models could be improved.

This work while proposing initial metrics and reflections on the role of evaluation, paves the way to several future works. First, we plan to develop a large-scale evaluation framework involving scholars and experts from diverse disciplines to ensure broader and more robust validation. Second, in the context of verbalization, we intend to enhance the model with additional semantic information, such as the top-level classes of the involved entities, and assess the impact of this information on the quality

and accuracy of the generated verbalizations. In conclusion, by proposing an analysis and evaluation framework that extends beyond traditional metrics, this study provides a deeper understanding of the current challenges associated with the various phases of graph-based question answering and its evaluation.

6. Limitations

The present study offers an in-context evaluation of tools introduced at various stages of a graph-based question answering pipeline. While the limited number of tools evaluated could be seen as a limitation, we intentionally selected state-of-the-art models to provide up-to-date metrics and linked qualitative considerations. Additionally, the use of a single case study could be not exhaustive, however, we chose this niche area because it aligns with ongoing domain-specific projects that are of interest to various scholars collaborating on larger initiatives.

In terms of generalizability, the objective of this study was to develop a infrastructure for evaluating different models at various stages, and across diverse knowledge graphs. In a research landscape, where assessing the reasoning abilities of LLMs remains an open challenge, we argue that providing a flexible and domain-independent framework is a critical step toward understanding how these models perform on specific tasks, topics, and rare entities. While a limitation of this work is its focus on a single case study, the insights gained from both our qualitative and quantitative analyses inform not only future enhancements to the evaluation pipeline, but also broader discussions on the socio-technical implications of deploying such systems in domain-specific contexts.

Finally, the involvement of a small number of field experts means that the results may not be fully representative. As part of future work, we plan to conduct a large-scale analysis involving more experts and assess a broader range of input questions.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] J. Huang, K. C.-C. Chang, Towards reasoning in large language models: A survey, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1049–1065. URL: https://aclanthology.org/2023.findings-acl.67/. doi:10.18653/v1/2023.findings-acl.67.

[2] A. Abusitta, M. Q. Li, B. C. Fung, Survey on explainable ai: Techniques, challenges and open issues, Expert Systems with Applications 255 (2024) 124710.

[3] J. Kasai, K. Sakaguchi, R. Le Bras, A. Asai, X. Yu, D. Radev, N. A. Smith, Y. Choi, K. Inui, et al., Realtime qa: What's the answer right now?, Advances in neural information processing systems 36 (2023) 49025–49043.

[4] X. Li, S. Chan, X. Zhu, Y. Pei, Z. Ma, X. Liu, S. Shah, Are ChatGPT and GPT-4 general-purpose solvers for financial text analytics? a study on several typical tasks, in: M. Wang, I. Zitouni (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, Association for Computational Linguistics, Singapore, 2023, pp. 408–422. URL: https://aclanthology.org/2023.emnlp-industry.39/. doi:10.18653/v1/2023.emnlp-industry.39.

[5] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, N. K. Ahmed, Bias and fairness in large language models: A survey, Computational Linguistics 50 (2024) 1097–1179.

[6] M. H. Lee, J. M. Montgomery, C. K. Lai, Large language models portray socially subordinate groups as more homogeneous, consistent with a bias observed in humans, in: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, 2024, pp. 1321–1340.

[7] V. Armant, A. Mouakher, F. Vargas-Rojas, D. Symeonidou, J. Guérin, I. Mougenot, J.-C. Desconnets, Can knowledge graphs and retrieval-augmented generation be combined to explain query/answer relationships truthfully?, in: DAO-XAI 2024 Data meets Ontologies in Explainable AI co-located with the 27th European Conference on Artificial Intelligence (ECAI 2024), volume 3833, 2024.
[8] R. Jia, B. Zhang, S. J. R. Méndez, P. G. Omran, Leveraging large language models for semantic query processing in a scholarly knowledge graph, arXiv preprint arXiv:2405.15374 (2024).
[9] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, J. Larson, From local to global: A graph rag approach to query-focused summarization, arXiv preprint arXiv:2404.16130 (2024).
[10] Z. Xu, M. J. Cruz, M. Guevara, T. Wang, M. Deshpande, X. Wang, Z. Li, Retrieval-augmented generation with knowledge graphs for customer service question answering, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024, pp. 2905–2909.
[11] W. Fan, Y. Ding, L. Ning, S. Wang, H. Li, D. Yin, T.-S. Chua, Q. Li, A survey on rag meeting llms: Towards retrieval-augmented large language models, in: Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining, 2024, pp. 6491–6501.
[12] T. Vu, M. Iyyer, X. Wang, N. Constant, J. Wei, J. Wei, C. Tar, Y.-H. Sung, D. Zhou, Q. Le, T. Luong, FreshLLMs: Refreshing large language models with search engine augmentation, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 13697–13720. URL: https://aclanthology.org/2024.findings-acl.813/. doi:10.18653/v1/2024.findings-acl.813.
[13] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, H. Wang, Retrieval-augmented generation for large language models: A survey, arXiv preprint arXiv:2312.10997 2 (2023).
[14] S. Aghaei, E. Raad, A. Fensel, Question answering over knowledge graphs: A case study in tourism, IEEE Access 10 (2022) 69788–69801.
[15] A. Tauqeer, I. Hammid, S. Aghaei, P. Parvin, E. M. Postma, A. Fensel, Smell and taste disorders knowledge graph: answering questions using health data, Expert Systems with Applications 234 (2023) 121049.
[16] D. Galla, S. Hoda, M. Zhang, W. Quan, T. D. Yang, J. Voyles, Courage: A framework to evaluate rag systems, in: International Conference on Applications of Natural Language to Information Systems, Springer, 2024, pp. 392–407.
[17] S. Simon, A. Mailach, J. Dorn, N. Siegmund, A methodology for evaluating rag systems: A case study on configuration dependency validation, arXiv preprint arXiv:2410.08801 (2024).
[18] E. Kamalloo, A. Jafari, X. Zhang, N. Thakur, J. Lin, Hagrid: A human-llm collaborative dataset for generative information-seeking with attribution, arXiv preprint arXiv:2307.16883 (2023).
[19] C. L. Clarke, L. Dietz, Llm-based relevance assessment still can't replace human relevance assessment, arXiv preprint arXiv:2412.17156 (2024).
[20] I. Soboroff, Don't use llms to make relevance judgments, Information retrieval research journal 1 (2025) 10–54195.
[21] F. Delgado, S. Yang, M. Madaio, Q. Yang, The participatory turn in ai design: Theoretical foundations and the current state of practice, in: Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, 2023, pp. 1–23.
[22] R. Sapkota, S. Raza, M. Karkee, Comprehensive analysis of transparency and accessibility of chatgpt, deepseek, and other sota large language models, arXiv preprint arXiv:2502.18505 (2025).
[23] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in neural information processing systems 33 (2020) 9459–9474.
[24] Y. Tang, Y. Yang, Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries, arXiv preprint arXiv:2401.15391 (2024).
[25] D. Taunk, L. Khanna, S. V. P. K. Kandru, V. Varma, C. Sharma, M. Tapaswi, Grapeqa: Graph augmentation and pruning to enhance question-answering, in: Companion Proceedings of the ACM Web Conference 2023, 2023, pp. 1138–1144.

[26] J. Zhang, X. Zhang, J. Yu, J. Tang, J. Tang, C. Li, H. Chen, Subgraph retrieval enhanced model for multi-hop knowledge base question answering, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5773–5784. URL: https://aclanthology.org/2022.acl-long.396/. doi:10.18653/v1/2022.acl-long.396.

[27] C. Mavromatis, G. Karypis, Gnn-rag: Graph neural retrieval for large language model reasoning, arXiv preprint arXiv:2405.20139 (2024).

[28] J. Kim, Y. Kwon, Y. Jo, E. Choi, KG-GPT: A general framework for reasoning on knowledge graphs using large language models, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 9410–9421. URL: https://aclanthology.org/2023.findings-emnlp.631/. doi:10.18653/v1/2023.findings-emnlp.631.

[29] J. Sun, C. Xu, L. Tang, S. Wang, C. Lin, Y. Gong, L. M. Ni, H.-Y. Shum, J. Guo, Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph, arXiv preprint arXiv:2307.07697 (2023).

[30] P. Schmidtova, S. Mahamood, S. Balloccu, O. Dusek, A. Gatt, D. Gkatzia, D. M. Howcroft, O. Platek, A. Sivaprasad, Automatic metrics in natural language generation: A survey of current evaluation practices, in: S. Mahamood, N. L. Minh, D. Ippolito (Eds.), Proceedings of the 17th International Natural Language Generation Conference, Association for Computational Linguistics, Tokyo, Japan, 2024, pp. 557–583. URL: https://aclanthology.org/2024.inlg-main.44/. doi:10.18653/v1/2024.inlg-main.44.

[31] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.

[32] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[33] J. Chen, H. Lin, X. Han, L. Sun, Benchmarking large language models in retrieval-augmented generation, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 2024, pp. 17754–17762.

[34] R. Friel, M. Belyi, A. Sanyal, Ragbench: Explainable benchmark for retrieval-augmented generation systems, arXiv preprint arXiv:2407.11005 (2024).

[35] H. Yu, A. Gan, K. Zhang, S. Tong, Q. Liu, Z. Liu, Evaluation of retrieval-augmented generation: A survey, in: CCF Conference on Big Data, Springer, 2024, pp. 102–120.

[36] B. Peng, Y. Zhu, Y. Liu, X. Bo, H. Shi, C. Hong, Y. Zhang, S. Tang, Graph retrieval-augmented generation: A survey, arXiv preprint arXiv:2408.08921 (2024).

[37] J. Novikova, O. Dušek, A. Cercas Curry, V. Rieser, Why we need new evaluation metrics for NLG, in: M. Palmer, R. Hwa, S. Riedel (Eds.), Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2241–2252. URL: https://aclanthology.org/D17-1238/. doi:10.18653/v1/D17-1238.

[38] S. Gehrmann, E. Clark, T. Sellam, Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text, Journal of Artificial Intelligence Research 77 (2023) 103–166.

[39] S. Tsaneva, D. Dessì, F. Osborne, M. Sabou, Knowledge graph validation by integrating llms and human-in-the-loop, Information Processing & Management 62 (2025) 104145.

[40] G. Faggioli, L. Dietz, C. L. Clarke, G. Demartini, M. Hagen, C. Hauff, N. Kando, E. Kanoulas, M. Potthast, B. Stein, et al., Perspectives on large language models for relevance judgment, in: Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, 2023, pp. 39–50.

[41] H. Q. Yu, F. McQuade, Rag-kg-il: A multi-agent hybrid framework for reducing hallucinations and enhancing llm reasoning through rag and incremental knowledge graph learning integration, arXiv preprint arXiv:2503.13514 (2025).

[42] L. Gienapp, T. Hagen, M. Fröbe, M. Hagen, B. Stein, M. Potthast, H. Scells, The viability of

crowdsourcing for rag evaluation, arXiv preprint arXiv:2504.15689 (2025).
[43] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, Á. Fernández-Leal, Human-in-the-loop machine learning: a state of the art, Artificial Intelligence Review 56 (2023) 3005–3054.
[44] X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, L. He, A survey of human-in-the-loop for machine learning, Future Generation Computer Systems 135 (2022) 364–381.
[45] S. Es, J. James, L. Espinosa Anke, S. Schockaert, RAGAs: Automated evaluation of retrieval augmented generation, in: N. Aletras, O. De Clercq (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, St. Julians, Malta, 2024, pp. 150–158. URL: https://aclanthology.org/2024.eacl-demo.16/. doi:10.18653/v1/2024.eacl-demo.16.
[46] C. Holtermann, P. Röttger, T. Dill, A. Lauscher, Evaluating the elementary multilingual capabilities of large language models with MultiQ, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 4476–4494. URL: https://aclanthology.org/2024.findings-acl.265/. doi:10.18653/v1/2024.findings-acl.265.
[47] H. Li, Y. Ning, Z. Liao, S. Wang, X. L. Li, X. Lu, W. Zhao, F. Brahman, Y. Choi, X. Ren, In search of the long-tail: Systematic generation of long-tail inferential knowledge via logical rule guided search, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 2348–2370. URL: https://aclanthology.org/2024.emnlp-main.140/. doi:10.18653/v1/2024.emnlp-main.140.
[48] S. Bogdanov, A. Constantin, T. Bernard, B. Crabbé, E. Bernard, Nuner: Entity recognition encoder pre-training via llm-annotated data, 2024. URL: https://arxiv.org/abs/2402.15343. arXiv:2402.15343.
[49] N. Reimers, I. Gurevych, Making monolingual sentence embeddings multilingual using knowledge distillation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2020. URL: https://arxiv.org/abs/2004.09813.
[50] A. V. Solatorio, Gistembed: Guided in-sample selection of training negatives for text embedding fine-tuning, arXiv preprint arXiv:2402.16829 (2024). URL: https://arxiv.org/abs/2402.16829. arXiv:2402.16829.
[51] M. Oliverio, P. F. Balestrucci, A. Mazzei, V. Basile, Dipinfo-unito at the gem'24 data-to-text task: Augmenting llms with the split-generate-aggregate pipeline, INLG 2024 (2024) 59.
[52] Q. Team, Qwen2.5: A party of foundation models, 2024. URL: https://qwenlm.github.io/blog/qwen2.5/.
[53] C. Gardent, A. Shimorina, S. Narayan, L. Perez-Beltrachini, Creating training corpora for NLG micro-planners, in: R. Barzilay, M. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, Association for Computational Linguistics, 2017, pp. 179–188. URL: https://doi.org/10.18653/v1/P17-1017. doi:10.18653/v1/P17-1017.
[54] Z. Kasner, O. Dusek, Beyond traditional benchmarks: Analyzing behaviors of open LLMs on data-to-text generation, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 12045–12072. URL: https://aclanthology.org/2024.acl-long.651/. doi:10.18653/v1/2024.acl-long.651.
[55] M. A. Stranisci, E. Bernasconi, V. Patti, S. Ferilli, M. Ceriani, R. Damiano, The world literature knowledge graph, in: International Semantic Web Conference, Springer, 2023, pp. 435–452.
[56] R. Orlando, P.-L. H. Cabot, E. Barba, R. Navigli, Relik: Retrieve and link, fast and accurate entity linking and relation extraction on an academic budget, 2025. URL: https://arxiv.org/abs/2408.00103. arXiv:2408.00103.
[57] M. Boscariol, L. Bulla, L. Draetta, B. Fiumanò, E. Lenzi, L. Piano, Evaluation of llms on long-tail entity linking in historical documents, arXiv preprint arXiv:2505.03473 (2025).