QAWiki: A Knowledge Graph Question Answering & SPARQL Query Generation Dataset for Wikidata

Alberto Moya Loustaunau[1,2], Aidan Hogan[1,2]

[1]IMFD, Santiago, Chile
[2]DCC, Universidad de Chile, Santiago, Chile

Abstract

In this resource paper, we present QAWiki: a multilingual, handcrafted, knowledge graph question answering and SPARQL query generation dataset for Wikidata. QAWiki consists of 526 questions over Wikidata, of which 518 are associated with SPARQL queries, and 8 are disambiguation questions. Each question is presented in both English and Spanish, and includes paraphrased versions of the question, as well as annotations of entity and relation mentions for Wikidata. The dataset is hosted in a Wikibase instance, which allows for collaborative editing and refinement of the dataset by the community, among other features. Further metadata include tagging questions with issues (e.g., incompleteness, imprecision, ambiguity) as well as defining relations between questions (e.g., a question whose answers are contained in another question, etc.). QAWiki can thus be used as an evaluation (and training) dataset for knowledge graph question answering & query generation systems. We provide illustrative experiments over QAWiki using GPT-4o to generate SPARQL queries over Wikidata, comparing performance with and without passing entity mentions to the model via the prompt.

1. Introduction

Knowledge Graphs (KGs) are powerful abstractions of structured knowledge, representing entities and the relationships between them as nodes and edges in a graph. They have found wide application across domains such as search engines, recommendation systems, digital assistants, and scientific knowledge management [1]. Prominent examples include Wikidata [2] and DBpedia [3]. These graphs are typically queried using formal languages such as SPARQL, the W3C-standard query language for RDF data. While SPARQL is expressive and precise, its usage requires technical expertise that is out of reach for most end users.

To bridge this accessibility gap, the field of Knowledge Graph Question Answering (KGQA) has emerged, aiming to enable users to access rich, factual, and structured knowledge via natural language questions, without requiring expertise in formal query languages [4]. Closely related benchmark tasks are Question Answering over Linked Data (QALD) [5], and SPARQL query generation, which addresses KGQA by translating natural language questions into SPARQL queries [6, 7]. Neural models for these tasks have achieved good results in handling relatively simple questions, such as those answerable by a single triple (or a single fact/claim).

Wikidata'25: Wikidata workshop at ISWC 2025
amoya@dcc.uchile.cl (Alberto Moya Loustaunau); ahogan@dcc.uchile.cl (Aidan Hogan)
https://aidanhogan.com/ (Aidan Hogan)
0000-0002-7003-5087 (Alberto Moya Loustaunau); 0000-0001-9482-1982 (Aidan Hogan)

Recent research has highlighted the growing importance of addressing complex natural language questions that require reasoning over multiple relations, constraints, and inference conditions [8]. However, most existing QA datasets primarily consist of simple, single-hop questions [9], whereas real-world user queries often involve multi-hop reasoning, aggregations, temporal or comparative logic [10, 11]; indeed, the true power of knowledge graphs lies in being able to address more complex questions. In addition, the performance of KGQA systems heavily depends on accurate entity and relation linking: a step that remains challenging due to ambiguity, discontinuity, and linguistic variability, particularly in multilingual or low-resource settings [12]. Recent evaluations have also shown that existing KGQA datasets often lack the diversity and complexity required to support compositional or zero-shot generalization, limiting their usefulness for training systems that can handle real-world queries [13].

To support the development of robust KGQA systems, datasets should ideally provide: (i) gold-standard SPARQL queries for execution supervision [14]; (ii) explicit annotations of entity and property mentions for robust linking [15]; (iii) diverse and expressive question forms, including paraphrases and natural ambiguity [16]; and (iv) multilingual coverage [2].

In this resource paper, we introduce QAWiki: a novel, multilingual, and collaboratively-curated dataset for benchmarking or training question answering and query generation systems over Wikidata. QAWiki complements existing datasets by offering a rich, diverse, and extensible collection of questions, annotated with entity and property mentions, SPARQL queries, quality tags, and semantic relations between questions. QAWiki includes complex multi-hop questions that require diverse SPARQL operators to represent, and explicitly tags questions with ambiguity or incompleteness, enabling fine-grained error analysis. It further includes annotations of entity mentions and relation mentions to help evaluate systems in more detail, and to facilitate creating larger synthetic datasets (e.g., by replacing the entities). Its multilingual design—with parallel questions and annotations in English and Spanish—supports comparison across languages. A key difference to other larger datasets is that QAWiki has been almost entirely handcrafted (periodically, over a span of more than two years) to ensure diverse, high-quality questions and queries with human-like phrasing. Moreover, by being hosted in a publicly accessible Wikibase instance, QAWiki supports community-driven editing, extensibility, and provenance tracking.

2. Related Work

KGQA has seen significant advancements over the past decade. Nevertheless, the datasets used for research exhibit limitations, particularly regarding complexity, diversity and quality.

Early KGQA datasets include the following:

- WebQuestionsSP [17] and SimpleQuestions [18] provide large sets of relatively simple question–answer pairs over Freebase.
- LC-QuAD 1.0 [19] provides DBpedia-based multi-hop questions generated using templates, and thus exhibiting limited linguistic variety.
- LC-QuAD 2.0 [7] incorporates crowdsourced natural language paraphrases over Wikidata.

Initially, most KGQA datasets were English-only, limiting accessibility. The need for broader linguistic coverage led to the inclusion of multiple languages:

- The QALD series (e.g., QALD-7 [20], QALD-9 [5], QALD-10 [21]) introduced manual, crowdsourced translations across various languages (e.g., English, German, Chinese, Russian, Spanish), though often at a limited scale.
- Other datasets, such as RuBQ [22, 23] and MCWQ [24], extend language support through automatic translation and human validation.

Several recent KGQA datasets have been developed natively over Wikidata:

- WikiWebQuestions [25]: An adaptation of WebQuestions to Wikidata, pairing natural-language questions with SPARQL annotations. Despite its real-world question style, it remains English-only and contains relatively simple questions.
- QALD Series (QALD-7, QALD-9, QALD-10): These benchmarks are manually curated and multilingual. In particular, QALD-10 was built from scratch: English-language questions were authored by proficient speakers and translated into multiple languages via native translators. Final SPARQL queries were crafted manually by domain-aware experts to suit Wikidata's schema and handle its labeling inconsistencies. (These are the most closely related datasets to what we provide; a comparison is provided later.)
- WikidataQA [26] is a small, handcrafted dataset with 100 questions and their corresponding queries over Wikidata.
- SPINACH [27]: This dataset begins with real SPARQL queries harvested from Wikidata's "Request a Query" forum. Experts then crafted corresponding natural-language questions that faithfully reflect each query's intent, albeit in a more formal style than typical forum phrasing. SPINACH exhibits high structural complexity, but is monolingual (English-only), static, and does not include mention-level annotations.

Construction methodologies for KGQA datasets vary widely, each with trade-offs:

- **Synthetic Generation**: Datasets like LC-QuAD 1.0 [19], KQA Pro [28], MCWQ [24], ComplexWebQuestions [16], and CFQ [29] are built using grammar rules or templates to generate SPARQL queries and corresponding natural language questions. While effective for generating large-scale data, they often suffer from lack of realism and diversity.
- **Crowdsourcing**: LC-QuAD 2.0 [7] and QALD-10 [21] leverage human contributors for paraphrasing existing questions or directly translating them, ensuring more natural language but potentially leading to inconsistent quality or scale limitations.
- **Automatic Translation**: RuBQ [22, 23] and MCWQ [24] use machine translation to cover more natural languages. MCWQ mitigates translation errors via human review.
- **Expert Curation**: Datasets such as SPINACH [27] and parts of QALD-10 [21] involve expert linguists or domain specialists to ensure high quality, correctness, and executable SPARQL queries. This method typically results in smaller, but higher-quality datasets.

A critical evaluation by Jiang and Usbeck [13] of 25 KGQA datasets revealed that most resources lack sufficient support for zero-shot or compositional generalization, often due to template-based construction and lack of diversity. Furthermore, datasets like WebQuestionsSP and CWQ were found to have factual correctness rates below 60%, reflecting annotation errors

and outdated knowledge graph links. A quick review of many such datasets confirms such quality issues, where for example, in LC-QuAD 2.0, we can find many questions of a similar form to "*Where is {disciples} of {Nadia Boulanger}, which has {location of death} is {Azores} ?*" [*sic.*] where it seems the crowdsourced paraphrasing was not done as intended.

In contrast, QAWiki is designed to be a high-quality, multilingual, and extensible resource. It combines expert curation with community collaboration through a Wikibase instance, includes explicit mention-level annotations, and features both simple and complex questions with human-like phrasing in diverse domains and of diverse forms. These design choices aim to meet the evolving needs of KGQA research and respond directly to critiques in the literature [13]. To the best of our knowledge, all questions and queries are handcrafted by design (the vast majority by the authors, with some community contributions). In comparison to other datasets, we argue that QAWiki is of higher quality and diversity, though smaller than many synthetic datasets. However, depending on the need, QAWiki can serve as input for synthetic generation, crowdsourcing, automatic translation or paraphrasing, etc., in order to generate larger datasets.

3. QAWiki: Resource Description

We now provide an overview of QAWiki.

3.1. Question set

QAWiki currently includes 526 hand-crafted questions, all written or paraphrased in fluent English and Spanish (*not* automatically generated or translated). Of these, 8 questions are disambiguation questions (i.e., not specific enough to have a clear intent). While the questions are not equally distributed across categories, the set is composed of the following broad categories based on what the question returns:

- **Entity questions** ("*Which has the most … ?*", "*Which is the latest … ?*"): 77
- **Entity-set questions** (lists or tables of entities satisfying a condition): 284
- **Numeric questions** (e.g., "*How many … ?*", "*What is the average … ?*"): 45
- **Temporal questions** (e.g., "*In which year … ?*", "*Since when … ?*"): 54
- **Boolean questions** ("*Is … deceased?*", "*Does … exist?*"): 35
- **Miscellaneous attribute questions** (measurements, identifiers, locations, etc.): 31

3.2. Multilinguality and Paraphrasing

All QAWiki questions are presented in both English and Spanish, and all questions also have one or more paraphrased versions. Translations are intended to be idiomatic, and paraphrased versions to be distinctive but still natural; for example:

- **Label (EN)**: *"How old was Daniel Day-Lewis when he won his first Academy Award?"*
  Alias (EN): *"What age was Daniel Day-Lewis when he was first awarded an Oscar?"*
- **Label (ES)**: *"¿Cuántos años tenía Daniel Day-Lewis cuando ganó su primer premio Óscar?"*
  Alias (ES): *"¿A qué edad recibió Daniel Day-Lewis su primer premio Óscar?"*

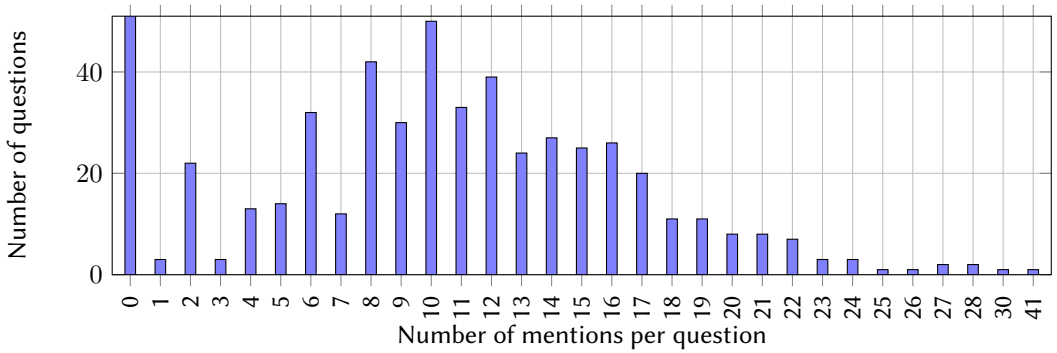

Figure 1: Distribution of mentions per question

As observed here, the translation intends to be natural, not direct, where "*Premio de la Academia*" in Spanish is much less idiomatic than its direct translation "*Academy Award*" in English, and thus is not used in the Spanish question. Some questions are further expressed in specific dialects where different; for example: "*In which countries is Elon Musk a naturalized citizen?*" is tagged @en, while "*In which countries is Elon Musk a naturalised citizen?*" is also provided, tagged @en-GB. This occurs in a handful (7) of cases. Finally, the community has provided 22 questions in Italian, and 7 in Danish. QAWiki currently contains 3,000 question forms.

3.3. Entity and Property Mentions

QAWiki includes fine-grained annotations of entity and relation mentions. These mentions are substrings of the English or Spanish questions and paraphrases, and are manually linked to the corresponding Wikidata identifiers: Q-IDs for entities and P-IDs for properties.

Importantly, the annotation process is independent of the SPARQL query. That is, mentions are identified based on the natural language surface form, regardless of whether or not they are used directly in the query. Indeed, in many cases, there may be more than one way to formulate a query using different mentions; for example, if a question refers to the *U.S. President*, a query might use the Wikidata entity *President of the United States* (Q11696), or might use *United States* (Q30) and from there traverse the property *head of state* (P35) or indeed the property *office held by head of state* (P1906). Mentions are thus intended to be exhaustive: overlapping entities are included, and some entities may be annotated with several alternatives.

Property mentions are considerably more complex than entity mentions, where we thus include a number of different types of mentions in our dataset:

- **Direct**: Indicates that the property is directly mentioned, e.g., "*What is the population of Qatar?*" contains the direct mention "*population*" of *population* (P1082) on Wikidata.
- **Inverse**: Indicates that the property is inversely mentioned, e.g., "*Who played Eleven in Stranger Things?*" contains the inverse mention "*played*" of *performer* (P175) on Wikidata.
- **Existence**: Indicates that the property has some value, e.g., "*Which popes were married?*" contains the existence mention "*married*" of *spouse* (P20).

- **Non-existence**: Indicates that the property has no value, e.g., "*Which living people have an element named after them?*" contains the non-existence mention "*living*" of *place of death* (P20), *date of death* (P570), etc.
- **Specific value**: Indicates that the property is implicitly mentioned via a specific value, e.g., "*Who is the drummer of the band Battles?*" contains the specific-value mention "*drummer*" of *occupation* (P106); we also indicate the value, in this case, *drummer* (Q386854); other cases include gender, nationality, etc., but we exclude references to *instance of* (P31) and *subclass of* (P279), which we assume to be so common as to be understood.
- **Superlative**: Indicates that the property has a superlative value, e.g., "*What is the tallest mountain in the world outside of Asia?*" contains the superlative mention "*tallest*" of *elevation above sea level* (P2044); we indicate if the superlative is a maximum or minimum.

We further capture discontinuous mentions, where the mention is split across the sentence; for example, "*Where was tellurium discovered?*" involves the discontinuous mention "*Where [...] discovered*" of `location of discovery` (P189), whereby "`Where`" differentiates the property from others that indicate who discovered something, when something was discovered, etc.

In total, 5,475 mentions are annotated by hand, pointing to 1,258 distinct Wikidata identifiers: 1,038 distinct items (Q) and 219 distinct properties (P). All QAWiki questions currently have mentions for English and Spanish (if applicable). Figure 1 lists the distribution of mentions per question. We allow questions to import mentions from another similar question to save manual effort, which results in some questions having zero mentions.

3.4. SPARQL Queries

Each non-ambiguous QAWiki question—such that the intent is clear (518 questions in total)—is annotated with one or more SPARQL queries that retrieve its correct answers from Wikidata. The queries are executable, enabling both supervised training of query generation and validation of question answering systems via their results. For example, for the question "*What was the last novel published by Harper Lee?*", we include the following query (with comments and indentation added here for readability purposes):

```
SELECT DISTINCT ?sbj
WHERE { ?sbj wdt:P7937 wd:Q8261 . # form of creative work: novel
        ?sbj wdt:P50 wd:Q182658 . # author:                 Harper Lee
        ?sbj wdt:P577 ?obj . }    # publication date
ORDER BY DESC(?obj) LIMIT 1
```

These queries are handcrafted. An important criterion is to formulate the query as generally as possible for the question type. For example, for the question "*Which university in Pakistan has the most students?*", the phrase "*in Pakistan*" is captured with the SPARQL property path:

```
?sbj wdt:P131*/wdt:P17?/wdt:P30? wd:Q843
```

which captures a wide variety of cases, including being transitively located in the place (`wdt:P131*`), with the place being a region (`wdt:P131`), a country (`wdt:P17`), or continent (`wdt:P30`), etc. Thus the query will function if "*Pakistan*" is replaced by "*Islamabad*", *Asia*", etc.

Table 1
Number of queries using a particular SPARQL feature

Feature	qald_7	qald_9_plus	qald_10	wikiwebq	spinach	qawiki
SELECT	42	468	333	1885	319	483
ASK	8	39	61	0	0	35
DISTINCT	41	340	220	1885	94	205
ORDER BY	2	58	17	48	98	73
LIMIT	2	58	17	60	15	73
OFFSET	0	3	3	1	0	3
FILTER	6	45	75	70	120	56
UNION	6	16	5	4	30	3
OPTIONAL	0	0	1	0	114	7
MINUS / NOT EXISTS	1	5	9	11	60	38
BIND	0	12	34	0	43	20
VALUES	0	6	0	0	41	2
SUBQUERY	1	2	11	0	35	17
GROUP BY	3	36	95	0	62	39
HAVING	2	3	1	0	5	3
Property Paths	12	91	88	217	134	291
Recursive Paths (*,+)	4	67	36	198	105	246
Total Queries	50	507	394	1885	319	518

Compared to many other KGQA benchmarks, QAWiki exhibits a rich set of query operators, as shown in Table 1. The average number of predicates per query is 2.95 (std. 1.84), higher than in QALD-10 (1.84) or WikiWebQuestions (1.58), but below SPINACH (4.67). Indeed, SPINACH has more complex queries overall than QAWiki, due to how it was generated: by taking existing SPARQL queries for Wikidata and generating questions from them. However, this leads to rather unnatural questions, such as "*Which properties are used in claims related to items that are public elections? Include a count of the number of times it was used.*", which are more akin to an explanation or verbalization of a query than a question a person would naturally ask.

3.5. Quality Tags

QAWiki incorporates optional metadata in the form of **quality tags**, which identify issues that may complicate question answering systems. They intend to annotate issues in questions a user is likely to ask. For example:

- `ambiguous`: The question has more than one plausible interpretation or scope; for example, the question "*What is the largest country in Africa?*" does not clarify if this is by area, or by population.
- `subjective criteria`: The question contains non-crisp criteria; for example, the question *Which conferences focus on Machine Learning?* requires a subjective interpretation of "*focus on*".

Tags are also used on queries to indicate potential issues relating to the results returned:

- `incomplete`: The query should be expected to return incomplete answers.
- `no ties`: The query uses `LIMIT 1` on a superlative question, and thus will not return ties (if any).
- `controversial or unconfirmed data`: The query returns answers based on unconfirmed information.

SPARQL queries provided in these cases aim to provide best-effort answers to the question. The inclusion of such questions is a deliberate design choice considering that Wikidata will not always be able to provide sound and complete results for all users' questions in practice. Hence we foresee that such tags can be used to flag to the user that the answers provided are best-effort, and may exhibit the aforementioned issues.

3.6. Question relations

QAWiki further captures *semantic relations* between questions. Originally, this began as a way to disambiguate questions, but evolved over time to capture more complex relations, including:

- `disambiguates`: A question may represent one way to disambiguate another more ambiguous question; for example: "*What is the largest country in Africa by area?*" disambiguates "*What is the largest country in Africa?*".
- `broader/narrower`: The results for a narrow question may be contained in those of a broader question; for example: "*Which official languages of the European Union are not Indo-European?*" is narrower than "*What are the official languages of the European Union?*".
- `contingent`: One question may assert an assumption underlying another; for example, "*Did Ian Curtis commit suicide?*" verifies the assumption of "*When did Ian Curtis commit suicide?*" (the latter is contingent on the former).
- `count of`/`boolean of`, etc.: One question might count the answers of another question, or ask if another question has any answers; for example: "*How many cities are twinned with Port-au-Prince?*" counts "*What cities are twinned with Port-au-Prince?*".

We believe that these relations could be useful for evaluating the consistency of answers generated by a system (e.g., to see if the results of a broader question are indeed contained in the narrower question, or to see if the count question actually returns the number of results in the base question), to train systems to detect and resolve ambiguity interactively with the user (e.g., to suggest questions that disambiguate an ambiguous one), and to also avoid issues relating to false premises in questions.

3.7. Implementation, Availability & Quality Control

QAWiki is hosted on a public Wikibase instance available at https://qawiki.org/, enabling users to collaboratively extend and refine the dataset, and thus for the dataset to evolve over time.

Wikibase offers various desirable features for such a dataset, including multilingual support, flexible schema, identifier schemes with autocompletion for editing, etc. It is also accompanied by a SPARQL endpoint for querying and validating annotations, which proved to be very useful

for checking and resolving errors. In preparing this version of QAWiki, we use 14 quality-control SPARQL queries[1] to find, for example, mentions not contained in a question, Wikidata item links that do not start with 'Q', Wikidata property links that do not start with 'P', etc. We also employed LLMs to verify and suggest corrections for question phrasing, which were manually reviewed and applied. We foresee that similar processes can be employed in future to provide high-quality snapshots of QAWiki incorporating community contributions.

The dataset is already integrated into systems like Templet [30], which enables template-based question answering over Wikidata powered by QAWiki (and its explicit entity mentions).

The entire dataset is licensed under CC0, ensuring free and unrestricted use. We provide a snapshot (v1) of QAWiki on Zenodo corresponding to this paper [31].

3.8. Usage

QAWiki is primarily intended to support the evaluation of knowledge graph question answering & SPARQL query generation approaches. The queries provided can be used to generate answer sets from Wikidata for evaluating question answering. QAWiki may also be useful for training or fine-tuning models, potentially in combination with methodologies that use QAWiki as a seed dataset from which to synthetically generate more instances (e.g., by replacing the entity mentions provided with parameters to create question–query templates for generating more instances, or by using machine translation to test further natural languages, or by using LLMs to paraphrase questions, etc.). It may also support few-shot learning or prompting.

4. Evaluation

To illustrate the utility of the QAWiki resource, we evaluate the ability of a large language model (GPT-4o) to generate accurate SPARQL queries from natural language questions in the QAWiki and SPINACH datasets. We further evaluate how performance improves when relevant entity mentions from QAWiki are passed to the model.

4.1. Evaluation Measures

To quantify the quality of the SPARQL queries generated by a large language model (GPT-4o), we might first consider comparing the predicted query against the gold standard query. However, there is no canonical way to write a query, and the problem of deciding if two SPARQL queries are *equivalent*—i.e., if they give the same results over any dataset—is undecidable [32]. In fact, even if we had an oracle for SPARQL query equivalence, it would still not suffice, as Wikidata contains redundancy, meaning there might be several ways to achieve valid answers via non-equivalent queries (per the *U.S. President* example presented in Section 3.3).

We thus rather follow the same approach proposed for SPINACH [27]: comparing the table of results for the predicted query with that of the gold standard query. This is perhaps more challenging than it first appears: result rows may appear in any order, queries may project a different number of variables in a different order, etc. Furthermore, in QAWiki, unless otherwise

[1] See https://qawiki.org/wiki/QAWiki:Curation/Quality_Control_Queries

required by the question, queries return Wikidata identifiers without additional information (which is trivial to retrieve in a later step). If the predicted query returns additional information via projected variables about the entities (e.g., their labels), this should not affect the measure.

These issues are addressed by computing a matrix of recall values between all row pairs. Each query yields a table of results, where each row is treated as a set of values (representing entity IDs, literals, or labels). To compare two such tables—one for the predicted query and one for the gold standard query—we compute a matrix of recall values between all row pairs. For a given pair of rows (g_i, p_j), where g_i is a gold row and p_j is a predicted row, we define recall as [27]:

$$\text{Recall}(g_i, p_j) = \frac{|g_i \cap p_j|}{|g_i|}.$$

Thus, extra variables with auxiliary information in predicted rows do not affect the measure.

We then compute a bipartite matching between gold and predicted rows that maximizes the total recall using the Hungarian algorithm. The overall precision and recall are aggregated over all matched pairs, and the final F1 score is computed as [27]:

$$\text{F1} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}}$$

where TP is the sum of per-row recalls, FP is the number of unmatched predicted rows, and FN accounts for unmatched gold rows and partial mismatches. We also report the Exact Match (EM) metric, defined as 1 if the F1 score is exactly 1.0, and 0 otherwise.

A limitation of this approach is that we consider rows as sets, not tuples, which leaves open the possibility that we consider a predicted row as "correct" if it contains the correct values but in incorrect columns. We believe such a situation to be rare in practice (the vast majority of QAWiki queries project one column, for example). And as aforementioned, this formulation allows us to compare result sets even when they differ in column projections, column order or row order, or where outputs may contain additional attributes. The evaluation is based on the execution outputs of the queries, rather than their surface form, and is thus agnostic to syntactic or structural differences in the queries themselves.

4.2. Experimental Setup

We evaluate the quality of SPARQL queries produced by GPT-4o on two datasets:

- SPINACH: A collection of expert-annotated complex queries.
- QAWiki: The handcrafted set of queries presented herein.

We evaluate GPT-4o under two configurations. In the **Base** configuration, the model is provided only with the natural language question. In the **+Linked Entities** configuration (only applicable to QAWiki), the model is also given a list of entity mentions and their corresponding Wikidata IRIs to help formulate the query. The predicted and gold standard queries are later evaluated over the Wikidata Query Service. To minimize potential differences stemming from changes on Wikidata, we evaluate all queries together in the same time frame.

Table 2
Evaluation results for SPARQL query generation

Dataset	Setup	F1	EM	TP	FP	FN
SPINACH	Base	0.1887	0.0836	38,302.6	240,648	164,843
QAWiki	Base	0.1640	0.1283	7,333.5	264,551	107,016
QAWiki	+Linked Entities	0.4064	0.3327	32,714.5	34,204	81,635

4.3. Results

Table 2 presents the evaluation scores of GPT-4o across both datasets (SPINACH and QAWiki) under two different configurations. Note that TP, FP, and FN are aggregated over all evaluated queries and rows, rather than per question.

The results highlight several trends. First, the model performs quite poorly overall. Second, the model achieves moderately better performance on SPINACH than on QAWiki in the Base setting, suggesting that the latter is a more challenging dataset. Exact matches (EM) are higher for QAWiki likely because it contains simple questions and queries alongside more complex ones; indeed QAWiki also contains 35 ASK queries that are easier to achieve exact matches on, while SPINACH has none. While exact matches remain low across the board, the gains in F1 confirm that partial correctness is common: many predicted queries retrieve subsets or supersets of the correct answers. This nuance is captured well by our measure based on row-level evaluation. Supplying the model with linked entity information (+Linked Entities) leads to a substantial improvement on QAWiki: F1 rises from 0.1640 to 0.4064 and EM from 0.1283 to 0.3327. This demonstrates again that entity linking/disambiguation plays a critical role in this task, as has also been widely observed in related works on knowledge graph question answering and query generation.

4.4. Limitations

A number of limitations of this evaluation have already been discussed, namely that (1) our measure may count as correct a row with the correct values in incorrect columns; and (2) evaluating queries on the live Wikidata instance may lead to results changing between query executions. We add to this two other important limitations: (3) basing the evaluation only on results does not consider a query getting the correct results with an incorrect query, which can be particularly problematic in the case of ASK queries, and (4) both datasets are available on the Web, and thus may have formed part of the training data for the model (though GPT-4o only has knowledge up to October 2023). Addressing these limitations is left for future work.

5. Conclusion

QAWiki is an evolving, handcrafted dataset for knowledge graph question answering & SPARQL query generation over Wikidata. It currently contains 526 questions, of which 518 questions have associated SPARQL queries and 8 are disambiguation questions. The dataset contains rich supporting metadata for these questions and queries, including questions in both Spanish and

English, 3,000 paraphrased question forms in multiple languages, and 5,443 annotated mentions of entities and relations. The dataset is hosted on a Wikibase instance at https://qawiki.org/, with a snapshot also available on Zenodo [31].

Our preliminary evaluation shows that QAWiki presents a challenging dataset, especially when no additional context is provided. We further demonstrate that providing linked entity information significantly improves performance, highlighting the importance of entity disambiguation in query generation tasks. In future work, it would be of interest to evaluate over more models, and also to explore the trade-offs (in terms of precision, runtime, etc.) between using a given LLM for direct question answering vs. generating a query to derive answers.

We welcome contributions from the community—by adding questions, defining queries, refining the dataset, adding more natural languages, etc.—and hope that QAWiki can become a collaborative project wherein the community develops a high-quality, consensus-driven dataset for knowledge graph question answering & SPARQL query generation. This in turn will benefit research on these topics by enabling the training of better models, and more robust evaluation of such models. Eventually, we hope that this will result in better natural language interfaces that unlock the true power of Wikidata for a much broader class of users.

Acknowledgments

This work was funded in part by ANID – Millennium Science Initiative Program – Code ICN17_002. We thank Daniel Diomedi and all who provided questions/queries for QAWiki.

References

[1] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. F. Sequeda, S. Staab, A. Zimmermann, Knowledge Graphs, ACM Comput. Surv. 54 (2022) 71:1–71:37. doi:10.1145/3447772.

[2] D. Vrandecic, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Commun. ACM 57 (2014) 78–85. doi:10.1145/2629489.

[3] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. G. Ives, DBpedia: A Nucleus for a Web of Open Data, in: The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007, volume 4825 of *Lecture Notes in Computer Science*, Springer, 2007, pp. 722–735. doi:10.1007/978-3-540-76298-0_52.

[4] S. Ji, S. Pan, E. Cambria, P. Marttinen, P. S. Yu, A Survey on Knowledge Graphs: Representation, Acquisition, and Applications, IEEE Trans. Neural Networks Learn. Syst. 33 (2022) 494–514. doi:10.1109/TNNLS.2021.3070843.

[5] R. Usbeck, R. H. Gusmita, A. N. Ngomo, M. Saleem, 9th Challenge on Question Answering over Linked Data (QALD-9) (invited paper), in: Joint proceedings of the 4th Workshop on Semantic Deep Learning (SemDeep-4) and NLIWoD4: Natural Language Interfaces for the Web of Data (NLIWOD-4) and 9th Question Answering over Linked Data challenge (QALD-9) co-located with 17th International Semantic Web Conference (ISWC 2018),

Monterey, California, United States of America, October 8th - 9th, 2018, volume 2241 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018, pp. 58–64.

[6] T. Soru, E. Marx, D. Moussallem, G. Publio, A. Valdestilhas, D. Esteves, C. B. Neto, SPARQL as a Foreign Language, in: Proceedings of the Posters and Demos Track of the 13th International Conference on Semantic Systems - SEMANTiCS2017 co-located with the 13th International Conference on Semantic Systems (SEMANTiCS 2017), Amsterdam, The Netherlands, September 11-14, 2017, volume 2044 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2017.

[7] M. Dubey, D. Banerjee, A. Abdelkawi, J. Lehmann, LC-QuAD 2.0: A Large Dataset for Complex Question Answering over Wikidata and DBpedia, in: The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part II, volume 11779 of *Lecture Notes in Computer Science*, Springer, 2019, pp. 69–78. doi:10.1007/978-3-030-30796-7_5.

[8] B. Fu, Y. Qiu, C. Tang, Y. Li, H. Yu, J. Sun, A Survey on Complex Question Answering over Knowledge Base: Recent Advances and Challenges, CoRR abs/2007.13069 (2020). arXiv:2007.13069.

[9] J. Berant, A. Chou, R. Frostig, P. Liang, Semantic Parsing on Freebase from Question-Answer Pairs, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL, ACL, 2013, pp. 1533–1544. doi:10.18653/V1/D13-1160.

[10] J. Bao, N. Duan, Z. Yan, M. Zhou, T. Zhao, Constraint-Based Question Answering with Knowledge Graph, in: COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan, ACL, 2016, pp. 2503–2514.

[11] S. Mitra, R. R. Ramnani, S. Sengupta, Constraint-based Multi-hop Question Answering with Knowledge Graph, in: Proceedings of the 2022 Conference of the North American Chapter of the ACL: Human Language Technologies: Industry Track, NAACL 2022, Hybrid: Seattle, Washington, USA + Online, July 10-15, 2022, ACL, 2022, pp. 280–288. doi:10.18653/V1/2022.NAACL-INDUSTRY.31.

[12] L. Logeswaran, M. Chang, K. Lee, K. Toutanova, J. Devlin, H. Lee, Zero-Shot Entity Linking by Reading Entity Descriptions, in: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, ACL, 2019, pp. 3449–3460. doi:10.18653/V1/P19-1335.

[13] L. Jiang, R. Usbeck, Knowledge Graph Question Answering Datasets and Their Generalizability: Are They Enough for Future Research?, in: SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022, ACM, 2022, pp. 3209–3218. doi:10.1145/3477495.3531751.

[14] L. Dong, M. Lapata, Language to Logical Form with Neural Attention, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers, The Association for Computer Linguistics, 2016. doi:10.18653/V1/P16-1004.

[15] M. Yu, W. Yin, K. S. Hasan, C. N. dos Santos, B. Xiang, B. Zhou, Improved Neural Relation Detection for Knowledge Base Question Answering, in: Proceedings of the 55th Annual

Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, ACL, 2017, pp. 571–581. doi:10.18653/V1/P17-1053.

[16] A. Talmor, J. Berant, The Web as a Knowledge-Base for Answering Complex Questions, in: Proceedings of the 2018 Conference of the North American Chapter of the ACL: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), ACL, 2018, pp. 641–651. doi:10.18653/V1/N18-1059.

[17] W. Yih, M. Richardson, C. Meek, M. Chang, J. Suh, The Value of Semantic Parse Labeling for Knowledge Base Question Answering, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers, The Association for Computer Linguistics, 2016. doi:10.18653/V1/P16-2033.

[18] A. Bordes, N. Usunier, S. Chopra, J. Weston, Large-scale Simple Question Answering with Memory Networks, CoRR abs/1506.02075 (2015). arXiv:1506.02075.

[19] P. Trivedi, G. Maheshwari, M. Dubey, J. Lehmann, LC-QuAD: A Corpus for Complex Question Answering over Knowledge Graphs, in: The Semantic Web - ISWC 2017 - 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part II, volume 10588 of *Lecture Notes in Computer Science*, Springer, 2017, pp. 210–218. doi:10.1007/978-3-319-68204-4_22.

[20] R. Usbeck, A. N. Ngomo, B. Haarmann, A. Krithara, M. Röder, G. Napolitano, 7th Open Challenge on Question Answering over Linked Data (QALD-7), in: Semantic Web Challenges - 4th SemWebEval Challenge at ESWC 2017, Portoroz, Slovenia, May 28 - June 1, 2017, Revised Selected Papers, volume 769 of *Communications in Computer and Information Science*, Springer, 2017, pp. 59–69. doi:10.1007/978-3-319-69146-6_6.

[21] R. Usbeck, X. Yan, A. Perevalov, L. Jiang, J. Schulz, A. Kraft, C. Möller, J. Huang, J. Reineke, A.-C. N. Ngomo, M. Saleem, A. Both, QALD-10 – The 10th challenge on question answering over linked data: Shifting from DBpedia to Wikidata as a KG for KGQA, Semantic Web 15 (2024) 2193–2207. doi:10.3233/SW-233471. arXiv:https://journals.sagepub.com/doi/pdf/10.3233/SW-233471.

[22] V. Korablinov, P. Braslavski, RuBQ: A Russian Dataset for Question Answering over Wikidata, in: The Semantic Web - ISWC 2020 - 19th International Semantic Web Conference, Athens, Greece, November 2-6, 2020, Proceedings, Part II, volume 12507 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 97–110. doi:10.1007/978-3-030-62466-8_7.

[23] I. Rybin, V. Korablinov, P. Efimov, P. Braslavski, RuBQ 2.0: An Innovated Russian Question Answering Dataset, in: The Semantic Web - 18th International Conference, ESWC 2021, Virtual Event, June 6-10, 2021, Proceedings, volume 12731 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 532–547. doi:10.1007/978-3-030-77385-4_32.

[24] R. Cui, R. Aralikatte, H. C. Lent, D. Hershcovich, Compositional Generalization in Multilingual Semantic Parsing over Wikidata, Trans. Assoc. Comput. Linguistics 10 (2022) 937–955. doi:10.1162/TACL_A_00499.

[25] S. Xu, S. Liu, T. Culhane, E. Pertseva, M. Wu, S. J. Semnani, M. S. Lam, Fine-tuned LLMs Know More, Hallucinate Less with Few-Shot Sequence-to-Sequence Semantic Parsing over Wikidata, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023, ACL, 2023, pp. 5778–

5791. doi:10.18653/V1/2023.EMNLP-MAIN.353.
[26] D. Diomedi, A. Hogan, Entity Linking and Filling for Question Answering over Knowledge Graphs, in: Proceedings of the 7th Natural Language Interfaces for the Web of Data (NLIWoD) co-located with the 19th European Semantic Web Conference (ESWC 2022), Hersonissos, Greece, May 29th, 2022, volume 3196 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 9–24.
[27] S. Liu, S. J. Semnani, H. Triedman, J. Xu, I. D. Zhao, M. S. Lam, SPINACH: SPARQL-Based Information Navigation for Challenging Real-World Questions, in: Findings of the ACL: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024, ACL, 2024, pp. 15977–16001. doi:10.18653/V1/2024.FINDINGS-EMNLP.938.
[28] S. Cao, J. Shi, L. Pan, L. Nie, Y. Xiang, L. Hou, J. Li, B. He, H. Zhang, KQA Pro: A Dataset with Explicit Compositional Programs for Complex Question Answering over Knowledge Base, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, ACL, 2022, pp. 6101–6119. doi:10.18653/V1/2022.ACL-LONG.422.
[29] D. Keysers, N. Schärli, N. Scales, H. Buisman, D. Furrer, S. Kashubin, N. Momchev, D. Sinopalnikov, L. Stafiniak, T. Tihon, D. Tsarkov, X. Wang, M. van Zee, O. Bousquet, Measuring Compositional Generalization: A Comprehensive Method on Realistic Data, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020.
[30] F. Suárez, A. Hogan, Templet: A Collaborative System for Knowledge Graph Question Answering over Wikidata, in: Companion Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023, ACM, 2023, pp. 152–155. doi:10.1145/3543873.3587335.
[31] A. Moya Loustaunau, A. Hogan, QAWiki v1: Knowledge Graph Question Answering (KGQA) / SPARQL Query Generation Dataset for Wikidata (Version v1), Zenodo, 2025. doi:10.5281/zenodo.16787599.
[32] J. Salas, A. Hogan, Semantics and canonicalisation of SPARQL 1.1, Semantic Web 13 (2022) 829–893. URL: https://doi.org/10.3233/SW-212871. doi:10.3233/SW-212871.