HOME-KGQA: A Benchmark Dataset for Multimodal Knowledge Graph Question Answering on Household Daily Activities

Shusaku Egami, Aoi Ohta, Tomoki Tsujimura, Masaki Asada, Tatsuya Ishigaki, Ken Fukuda, Masahiro Hamasaki, Hiroya Takamura
National Institute of Advanced Industrial Science and Technology (AIST)
Koto-ku, Tokyo, Japan
{s-egami, oota.aoi, tsujimura.res, masaki.asada, ishigaki.tatsuya, ken.fukuda, masahiro.hamasaki, takamura.hiroya}@aist.go.jp

Abstract

Large Language Models (LLMs) provide flexible natural language processing capabilities, while knowledge graphs (KGs) offer explicit and structured knowledge. Integrating these two in a complementary manner enables the development of reliable and verifiable AI systems. In particular, knowledge graph question answering (KGQA) has attracted attention as a means to reduce LLM hallucinations and to leverage knowledge beyond the training data. However, existing KGQA benchmark datasets are biased toward encyclopedic knowledge, limited to a single modality, and lack fine-grained spatiotemporal data, which limits their applicability to real-world scenarios targeted by Embodied AI. We introduce HOME-KGQA, a novel KGQA benchmark dataset built on a multimodal KG of daily household activities. HOME-KGQA consists of complex, multi-hop natural language questions paired with graph database query languages. Compared to existing benchmarks, it includes more challenging questions that involve multi-level spatiotemporal reasoning, multimodal grounding, and aggregate functions. Experimental results show that the LLM-based KGQA methods fail to achieve performance comparable to that on existing datasets when evaluated on HOME-KGQA. This highlights significant challenges that should be addressed for the real-world deployment of KGQA systems. Our dataset is available at https://github.com/aistairc/home-kgqa.

Keywords: Knowledge Graph Question Answering, Large Language Models, Text-to-SPARQL, Spatiotemporal Reasoning, Embodied AI

1. Introduction

Large language models (LLMs) and knowledge graphs (KGs) are mutually complementary. By integrating the flexible natural language processing capabilities of LLMs with the structured and explicit knowledge provided by KGs, it is possible to build AI systems that are more reliable and verifiable. Knowledge Graph Question Answering (KGQA) (Steinmetz and Sattler, 2021; Jiang and Usbeck, 2022; Lehmann et al., 2023), also referred to as Knowledge Base Question Answering (KBQA) (Tan et al., 2023; Li et al., 2024; Xiong et al., 2024), is a challenging task that takes diverse natural language questions as input and outputs specific entities or aggregated results from a KG. This task plays a key role in bridging large language models (LLMs) and KGs. Recently, the KGQA approaches that generate SPARQL queries from natural language using LLMs have become an active research topic (Gashkov et al., 2025).

With the growing interest in KGQA, many benchmark datasets consisting of natural language questions, SPARQL queries, and corresponding answers have been released. Several studies have analyzed existing KGQA benchmark datasets from various perspectives, such as bias (Steinmetz and Sattler, 2021) and generalization capability (Jiang and Usbeck, 2022).

We further point out the lack of diversity in the source KGs targeted by existing KGQA datasets. Almost all of the commonly used KGQA datasets are built on DBpedia (Auer et al., 2007), Freebase (Bollacker et al., 2008), or Wikidata (Vrandečić, 2012). As a result, current KGQA studies primarily evaluate question answering over textual, encyclopedic facts at a single point in time (e.g., "Who are the parents of Barack Obama?"). However, as many LLMs have already acquired such general factual knowledge, the importance of evaluating KGQA performance on these datasets has diminished. Moreover, there is a gap between these QA tasks and the QA encountered in real life, limiting the practical applicability of KGQA systems. For example, in real-world environments like households, service industries, and caregiving settings, various questions can arise for purposes such as navigation or activity log analysis (Egami et al., 2023b,a; Vizcarra et al., 2021). To answer such questions, it is necessary to recognize human–robot–object interactions from videos and sensor data and ground them in natural language representations. As an intermediate representation that bridges natural language and observation data, multi-modal KGs (MMKGs) (Zhu et al., 2024) are required to capture fine-grained knowl-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7845–7856
11–16 May 2026.

edge, including 3D spatial knowledge, 2D visual knowledge, and temporal knowledge of human activities.

In this paper, we go beyond conventional KGQA systems that target textual encyclopedic facts and propose a novel benchmark dataset, HOME-KGQA, to facilitate the development of event-centric KGQA systems designed for real-world home environments. In contrast to existing QA over encyclopedic KGs, our benchmark focuses on episodic KGs, where question answering is grounded in fine-grained multimodal spatiotemporal events. First, we probabilistically generate a 100-day episodic KG containing over 150M triples based on the event-centric MMKG of daily life simulation videos (Egami et al., 2024). By performing question answering on this episodic KG, users of our dataset can evaluate the ability to answer factual questions about daily household activities – when, where, who, and what – grounded in structured knowledge. Next, we generate natural language questions and corresponding SPARQL queries by combining various qualifiers for time, space, actions, and objects. The query execution results are then used as answers to construct our QA dataset. Experimental results show that our dataset presents a higher level of difficulty for state-of-the-art KGQA models than existing benchmarks such as KQA Pro (Cao et al., 2022), ComplexWebQuestions (Talmor and Berant, 2018), WebQuestionsSP (Yih et al., 2016), and MetaQA (Zhang et al., 2018).

The main contributions of this study are summarized as follows:

(1) we introduce HOME-KGQA, the first KGQA benchmark dataset for episodic KGs in household environments, built on multimodal event-centric KGs with fine-grained spatiotemporal structure;
(2) we present a data generation process (episodic KG population, question text and SPARQL generation, and paraphrasing questions) with available code, enabling flexible augmentation of the dataset; and
(3) we provide a comprehensive experimental analysis showing that both Text-to-SPARQL and interactive agent-based methods still face considerable challenges on HOME-KGQA compared to conventional KGQA benchmarks.

Our dataset and code are available at GitHub[1].

2. Related Work

Many KGQA benchmark datasets have been released to date, and these datasets have been analyzed from various perspectives. Steinmetz and Sattler (2021) analyzed existing KGQA datasets such as LC-QuAD 1.0 (Trivedi et al., 2017), QALD (Usbeck et al., 2017), and SimpleDBpediaQA (Azmy et al., 2018) in terms of ambiguity, lexical gaps, complex query structures, template usage, ontology types, and answer types. Their analysis revealed that existing datasets contain highly ambiguous expressions and exhibit biases in SPARQL query operators, query graph patterns, and answer types. Jiang and Usbeck (2022) systematically investigated 25 KGQA benchmark datasets, including LC-QuAD 2.0 (Dubey et al., 2019), WebQuestionsSP (Yih et al., 2016), and GrailQA (Gu et al., 2021), from the perspective of generalization. They demonstrated that most KGQA datasets have no capability to evaluate compositional generalization. In contrast, our dataset is designed to evaluate compositional generalization. We further emphasize that 23 out of the 25 datasets analyzed in their study use only Freebase, Wikidata, or DBpedia as the targeting KGs. The remaining two datasets, Event-QA (Souza Costa et al., 2020) and MetaQA (Zhang et al., 2018), are based on EventKG (Kejriwal et al., 2019) and WikiMovies (Miller et al., 2016), respectively, both of which are extracted from Wikipedia. Consequently, existing KGQA benchmarks remain limited to question answering over general, encyclopedic facts.

Motivated by the potential to deploy KGQA systems in real-world home environments, we construct a novel benchmark dataset to facilitate the development of models capable of handling questions grounded in daily life scenes.

3. Task Definition

The input is a natural language question $q \in Q$, and the output is a corresponding SPARQL query $s \in S$. The task is to translate q into $s = f_\theta(q)$, where f_θ denotes a KGQA model. The following shows an example (q, s) pair.

The natural language question q: *How many times did the agent put a water glass in the kitchen between 7:56 p.m. on April 3, 2024, and 6:38 a.m. on May 27, 2024?*

The corresponding SPARQL query s:

```
PREFIX ho: <http://www.owl-
    ontologies.com/VirtualHome.
    owl#>
# ...
SELECT DISTINCT ?object (COUNT(
    DISTINCT ?event) AS ?answer)
WHERE { { SELECT * WHERE {
?event vh2kg:action ac:put ;
    vh2kg:mainObject ?object .
# ...
} GROUP BY ?object
```

[1] https://github.com/aistairc/home-kgqa

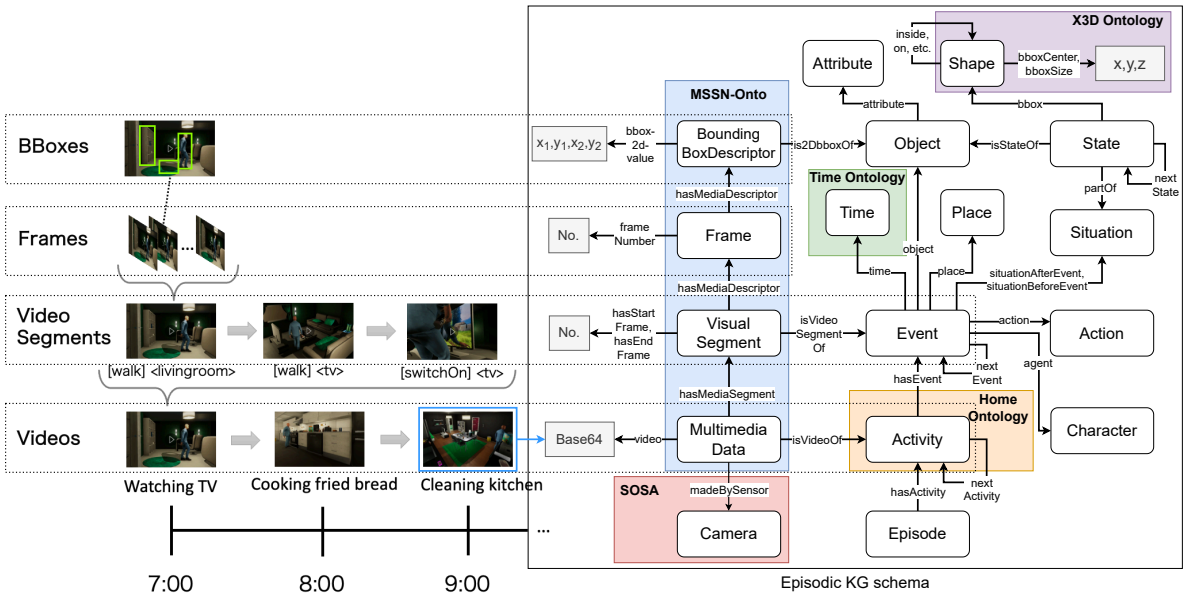

Figure 1: Daily activity videos and episodic KG

The target MMKG is formally represented as $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{L}, \mathcal{T}\}$, where $\mathcal{E}, \mathcal{R}, \mathcal{L}$ are sets of entities, relations, and literal values, respectively, and $\mathcal{T} = \mathcal{E} \times \mathcal{R} \times (\mathcal{E} \bigcup \mathcal{L})$ are sets of triples. The set $\mathcal{L}$ includes both symbolic literal values and multimodal data, defined as $\mathcal{L} = \mathcal{L}_{\mathcal{K}} \cup \mathcal{L}_{\mathcal{M}}$, where $\mathcal{L}_{\mathcal{K}}$ denotes the set of textual or numerical literals in the KG, and $\mathcal{L}_{\mathcal{M}}$ denotes the set of multimodal data such as images and videos. Figure 1 illustrates the correspondence between the videos of daily activities and the schema of our MMKG.

4. HOME-KGQA Construction

In this section, we describe the dataset construction process for HOME-KGQA. We first explain how the target MMKG is constructed, then detail the question generation process, and finally present an analysis of the constructed dataset.

4.1. Episodic KG Construction

We create an episodic KG of daily life using VHAKG (Egami et al., 2024), an MMKG of daily activities, as the source data. The episodic KG serves as the target knowledge base for the KGQA task. VHAKG is an MMKG constructed from synthetic data generated by the virtual environment simulator VirtualHome (Puig et al., 2018).

4.1.1. Episodic KG schema

The episodic KG follows the MMKG schema shown in Figure 1. Due to space limitations, only a portion of the schema is shown. The episodic KG reuses five different ontologies. Daily activities are modeled using an event-centric KG structure. Multimodal data are modeled based on the Multimedia Semantic Sensor Network Ontology (MSSN) (Angsuchotmetee et al., 2020) and SOSA (Janowicz et al., 2019), representing data captured from cameras installed in the household environment. Temporal information, including time points, intervals, and durations, is described based on the Time Ontology[2]. Activity concepts are extended from HomeOntology (Vassiliades et al., 2020), which defines activity categories (e.g., HouseCleaning) and their subclasses (e.g., Cleaning_kitchen). To represent 3D bounding boxes and spatial coordinates, the X3D Ontology (Brutzman and Flotyński, 2020) is reused.

In environments where heterogeneous data integration is required, such as in daily life, different ontologies are often reused based on the domain and modality of each data source. HOME-KGQA is also designed for question answering over KGs that integrate heterogeneous data.

4.1.2. Episode generation

In VHAKG, 700 independent activities exist in various scenarios, but no long-form episodes composed of multiple activities are included. In this study, we probabilistically generate plausible daily-life episodes by combining multiple activities using a Markov chain. The episode generation process follows the same method as in the existing study and reuses the crowdsourced data (Egami et al.,

[2] https://www.w3.org/TR/owl-time/

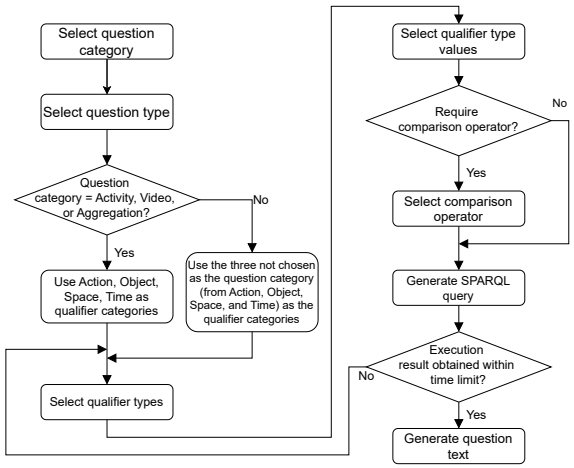

Figure 2: Flow of question generation process

2021).

Using first-order Markov chains calculated from 600 crowdsourced activity sequences, we probabilistically generated 100 daily-life episodes, each containing 18 activities representing plausible household routines.

4.1.3. Episodic KG population

We create entities corresponding to the generated episodes and represent them as instances of the *Episode* class. Each *Episode* instance is linked to 18 *Activity* instances through the *hasActivity* relation, with temporal order relationships added between consecutive activities. Since the same activities may appear multiple times across the 100-day episodes, each entity is assigned a sequential ID to distinguish individual instances within the KG.

Since VHAKG consists of a collection of independent activities, only the duration of each event was originally provided as temporal information. Therefore, we set the start time of the first event in the first day's episode to 05:00:00 on April 1, 2024, and assigned absolute start and end times to all events accordingly.

4.2. Question Dataset Generation

4.2.1. Procedure

Daily activities can be described as combinations of five elements: *Agent*, *Action*, *Object*, *Space*, and *Time*. Therefore, questions can be generated by combining these conditions. Since the episodic KG represents the daily activities of a single-person household, the *Agent* is always the same individual and is thus excluded from the question conditions. In this study, to handle questions that are temporally multi-granular, we target not only *Action* but also *Activity*. Moreover, leveraging the characteristics of the MMKG, we also include questions related to video data (*Video*). The questions are designed to cover not only those whose answers are entities or literal values but also those that require the execution of aggregate functions (*Aggregation*).

Figure 2 shows the procedure for generating SPARQL queries and corresponding natural language questions. Table 1 shows the question categories, question types, and example questions. Table 2 shows the types and values of the qualifiers used as query conditions.

4.2.2. SPARQL query and question text generation

SPARQL queries are generated based on templates defined by the question types and qualifier types. The values of randomly selected qualifier types are stored in a JSON structure as shown below, which is then used by an LLM to generate the corresponding natural language question. The values for each key are filled with the textual examples provided in Table 1 and Table 2.

```
{ "subject": "agent",
  "time": "Qualifier Text",
  "space": "Qualifier Text",
  "object": "Qualifier Text",
  "action": "Qualifier Text",
  "question": "Question Text" }
```

If one of the qualifiers is the target of the question, its value is left as an empty string. The question sentences are generated using few-shot prompting, where examples of JSON data and their corresponding expected questions are provided. We use OpenAI gpt-4o-mini as the LLM for question generation.

4.2.3. Paraphrasing question

Some expressions are unnatural from a conversational perspective because the generated questions directly contain entity URI suffixes and literal values. To address this, we apply a paraphrasing approach inspired by retrieval-augmented generation (RAG) to create more natural questions.

First, we define a set of rules for paraphrasing question sentences as follows. These rules are used as a system prompt when performing question paraphrasing with the LLM.

1. Correct grammatical errors.
2. Paraphrase time expressions in a more natural way.
3. Paraphrase attribute expressions in a more natural way.
4. Paraphrase state expressions in a more natural way.
5. Paraphrase object names in a more natural way.
6. Paraphrase type expressions in a more natural way.
7. Paraphrase class expressions in a more natural way.

Question Category	Question Type	Question Text Example
Object	None	What is the object ···
	Type	What is the type of the object ···
	Superclass	What is the superclass of the object ···
	State	What is the state of the object ···
	Attribute	What is the attribute of the object ···
	Size	What are the width, height, and depth of the object ···
Action	None	What did the agent do ···
Space	Place	What is the place ···
	3D Coordinates	What are the 3D coordinates of the object ···
Time	Temporal Instant	What is the time ···
	Temporal Interval	From when to when ···
	Duration	How long ···
Activity	None	What is the activity ···
	Previous/Next	What is the previous/next activity ···
Video	Video	What is the video ···
	Video Frame	What are the start and end frames ···
	2D Coordinates	What is the 2D coordinates of the object ···
Aggregation	Count	How many times ···
	Minimum	What is the minimum height of the object ···
	Maximum	What is the maximum width of the object ···
	Average	What is the average duration ···
	Summation	What is the total duration ···

Table 1: Question categories, types, and text examples

Qualifier Category	Qualifier Type	Qualifier Type Value	Comparison Operator	Qualifier Text Example
Object	Type	e.g., Computer, Sofa, and Towel		an object whose type is Towel
	Superclass	e.g., Electronics, Furniture, and Food		an object which is a subclass of Food
	State	e.g., ON, CLOSED, and CLEAN		an object whose state is CLEAN
	Attribute	e.g., containers, has_switch, and cream		an object whose attribute is has_switch
	Size	X, Y, Z	<, >, <=, >=	an object whose height is less than 0.6m
Action	None	e.g., walk, grab, and sit		walk
Space	Place	e.g., Livingroom, Bathroom, and Kitchen		livingroom
	3D Coordinates	X, Y, Z	<, >, <=, >=	at a position where the X coordinate is less than 2.47
Time	Temporal Instant	YYYY-MM-DD'T' HH:MM:SS		at 2024-04-25T12:01:00
	Temporal Interval	(YYYY-MM-DD'T' HH:MM:SS, YYYY-MM-DD'T' HH:MM:SS)	"< <", "<= <", "< <=", "<= <="	after 2024-05-29T03:25:00 and before or at 2024-06-05T20:08:00
	Duration	seconds	<, >, <=, >=	less than 9 seconds
	Before / After			after putting the remotecontrol460
	Just Before / Just After			just after putting the salmon332

Table 2: Qualifier categories, types, comparison operators, and text examples

8. Paraphrase activity expressions in a more natural way.
9. Paraphrase expressions describing what is shown in the video frame in a more natural way.
10. If the question is not about something that happened in the past, use the past tense in the question.
11. Don't change the original meaning.

Next, we manually create a gold dataset of paraphrased question sentences for each question type defined in Table 1. As a result, 22 pairs of raw and paraphrased question examples are pre-

Class	Relation	Instance	Triple
882	76	13,191,977	154,860,255
(882)	(86)	(13,192,053)	(162,609,309)

Table 3: Statistics of our episode KG. The lower row shows values when RDFS-Plus (Allemang and Hendler, 2011) reasoning is enabled.

pared.

Finally, for a given question, we retrieve the top-k most similar questions from the gold dataset and use the retrieved question – paraphrase pairs as multi-turn few-shot examples to prompt the LLM. In this study, we set $k = 5$. Based on the system prompt and the few-shot examples, the LLM then generates the paraphrased question sentences.

In total, we generate 150 questions for each question category, resulting in a QA dataset of 1,050 examples. Examples of the raw question and the paraphrased question are shown below.

(Raw) *How many times did an agent open an object whose type is Fridge in the kitchen from 2024-04-09T18:55:00 to 2024-06-23T02:02:00?*
(Paraphrased) *How many times did the agent open the fridge in the kitchen between 6:55 p.m. on April 9, 2024, and 2:02 a.m. on June 23, 2024?*

4.2.4. Data splitting for generalization

To enable the evaluation of KGQA methods in terms of generalization performance, we divide the dataset into train and test splits with respect to *i.i.d.* and compositional generalization. In the *i.i.d.* generalization setting, all relations $\mathcal{R}$, classes $\mathcal{C}$, and logical form constructs $\odot$, such as SPARQL modifiers, `FILTER` expressions, and operators, have been seen while training, but not the actual entities $\mathcal{E}$ and literals $\mathcal{L}$. Therefore, the dataset is split such that the qualifier values in test questions include unseen entities.

In the compositional generalization setting, all relations $\mathcal{R}$ and classes $\mathcal{C}$ are known, but specific logical form constructs and their operators $\odot$ appearing in the test set must be unseen. We set the operators `COUNT`, `MIN`, `AVG`, `SUM`, `<`, and `>` as unseen operators and split the dataset such that every SPARQL query in the test set includes at least one of these unseen operators.

4.3. Dataset Analysis

4.3.1. Statistics of episode KG

Table 3 shows the statistics of the constructed episodic KG. Compared with KQA Pro, which is based on FB15K-237 (Toutanova et al., 2015) and Wikidata, our KG is larger in both the number of classes and instances. Specifically, it contains 882 classes (concepts), exceeding KQA Pro's 794,

Class	# of instances
mssn:BoundingBoxDescriptor	6,364,212
x3do:SFVec3f	2,234,720
mssn:VisualSegment	2,173,685
vh2kg:State	1,117,360
vh2kg:Shape	1,117,360
mssn:MediaTimePointDescriptor	106,120
vh2kg:Situation	23,190
time:Duration	21,378
vh2kg:Event	17,784
mssn:MultimediaData	9,000

Table 4: Number of Instances (Top 10)

and 13,191,977 instances (entities), far surpassing KQA Pro's 16,960. This significant increase results from the more fine-grained temporal and spatial granularity of our KG compared with Freebase and Wikidata. In contrast, the number of relations (predicates) is smaller, 76 compared to KQA Pro's 363, because our KG is specifically limited to the household domain. Table 4 shows the top 10 classes with the largest number of instances. Since 2D bounding boxes are created for every 5 frames, the class *mssn:BoundingBoxDescriptor* has the highest number of instances.

4.3.2. Statistics of QA dataset

Figure 3 shows the distribution of query hop counts. All queries in our dataset are multi-hop. Figure 4 shows the distribution of question lengths. Compared with other datasets, HOME-KGQA has a wide range of question lengths, and both its mean and median lengths are longer than those of all other datasets.

Out of 100 randomly sampled questions evaluated manually, the generation of raw questions from JSON templates was accurate in 99 cases, whereas 94 were successfully paraphrased while preserving the original meaning, and 6 were incorrectly paraphrased.

Although we generated a relatively small number of QA pairs in this study to reduce the computational cost of experimental evaluation, all scripts for episodic KG construction, question–SPARQL generation, and paraphrasing are publicly available. Therefore, researchers can freely augment the dataset by reusing our scripts.

5. Experiments

The purpose of this experiment is to demonstrate the difficulty of HOME-KGQA compared to existing KGQA datasets and to clarify the challenges of KGQA in real-world daily life applications.

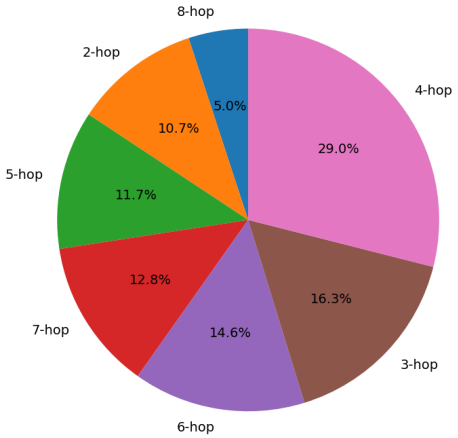

Figure 3: Distribution of query hops

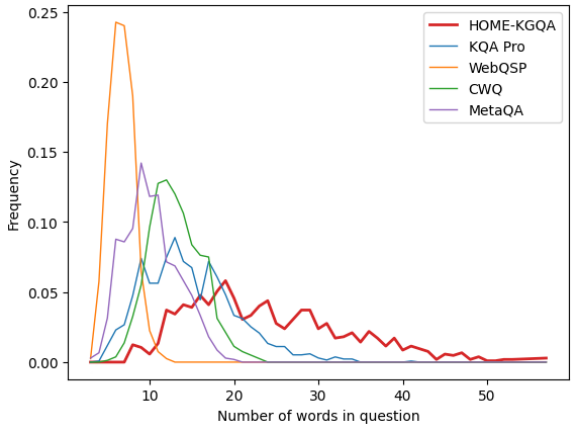

Figure 4: Question length distribution

5.1. Experimental Settings

5.1.1. Benchmark settings

Experiments are conducted using two datasets: one for *i.i.d.* generalization and the other for compositional generalization, with both having a train/test split of 350/700. As comparison datasets, we use KQA Pro (Cao et al., 2022), WebQuestionsSP (WebQSP) (tau Yih et al., 2016), ComplexWebQuestions (CWQ) (Talmor and Berant, 2018), and MetaQA (Zhang et al., 2018), which were employed in the previous study (Xiong et al., 2024). The train/test split sizes for each dataset follow the same samples as in the previous study and are as follows: KQAPro=450/900, WebQSP=100/300, CWQ=200/600, and MetaQA=150/900. We adopt exact match as the evaluation metric, which measures whether the SPARQL execution results exactly match the ground-truth answers.

5.1.2. Approaches

We conduct evaluation experiments using two approaches: a direct Text-to-SPARQL method based on LLMs, and an interactive agent-based method.

Text-to-SPARQL: The model receives a fixed-format prompt consisting of a system message and a user input containing a natural language question, and it outputs only the corresponding SPARQL query. Experiments are conducted under three settings: zero-shot prompting, multi-turn few-shot prompting, and fine-tuning. We use gpt-4o-mini-2024-07-18 and Llama3.1-8B-Instruct as the models. The system message and user input are defined as follows.

```
[System Message]
You are a SPARQL query generator.
    Generate a SPARQL query based
    on the given question. Do not
    output anything other than
   the SPARQL query.

[User Input]
Question:
{{question_text}}

SPARQL:
```

In the multi-turn few-shot prompting setting, multiple pairs of user inputs and model outputs are provided in a dialogue format. In our experiments, we adopt 5-shot prompting, allowing the model to reference these examples to generate an appropriate query for an input question. In the fine-tuning setting, supervised fine-tuning (SFT) is performed using pairs of prompts, consisting of the system message and user input, and their corresponding SPARQL queries.

In existing KGQA methods, models are often provided with a list of entity names and IDs required to answer a question. However, in real-world daily life scenarios, users do not formulate questions while being aware of entity IDs within the KG. Therefore, in this experiment, we do not provide the LLM with a list of entity names and ID pairs. Note that entity names are included in the question text itself for the raw questions.

Interactive-KBQA: For the interactive agent-based approach, we adopt Interactive-KBQA (Xiong et al., 2024) (w/o SFT), a state-of-the-art method that can be reliably reproduced. Interactive-KBQA performs semantic parsing and SPARQL generation through interactive reasoning. It provides several tools accessible to the LLM: `SearchNodes(name)` for entity linking, `SearchGraphPatterns(sparql, semantic)` for subgraph extraction, and `ExecuteSPARQL(sparql)` for query execution.

The LLM executes actions based on its initial thought. In each subsequent turn, the LLM

Approach	Model	Strategy	HOME-KGQA (ours)		KQAPro	WebQSP	CWQ	MetaQA
			Raw	Paraphrased				
Text-to-SPARQL	GPT-4o-mini	Zero-shot	0.000	0.000	0.026	0.000	0.000	0.000
		5-shot	0.117	0.056	0.115	0.050	0.095	0.059
		Fine-tuning	0.462	0.148	0.628	0.283	0.200	0.244
	Llama-3.1-8B-Instruct	Zero-shot	0.000	0.000	0.021	0.000	0.000	0.000
		5-shot	0.000	0.000	0.050	0.070	0.003	0.064
		Fine-tuning	0.148	0.047	0.590	0.200	0.245	0.217
Interactive-KBQA	GPT-4o-mini	all+1-shot	0.137	0.126	0.637	0.480	0.140	0.857

Table 5: Experimental results

Approach	Model	Strategy	I.I.D. generalization		Compositional generalization	
			Raw	Paraphrased	Raw	Paraphrased
Text-to-SPARQL	GPT-4o-mini	Zero-shot	0.000	0.000	0.003	0.003
		5-shot	0.117	0.056	0.066	0.043
		Fine-tuning	0.462	0.148	0.521	0.444
	Llama-3.1-8B-Instruct	Zero-shot	0.000	0.000	0.001	0.001
		5-shot	0.000	0.000	0.000	0.000
		Fine-tuning	0.148	0.047	0.267	0.162
Interactive-KBQA	GPT-4o-mini	all+1-shot	0.137	0.126	0.053	0.046
		same+2-shot	0.087	0.077	0.076	0.069

Table 6: Experimental results on the compositional generalization dataset

Approach	Generalization	Object	Action	Space	Time	Activity	Video	Aggregation
Text-to-SPARQL	I.I.D.	0.166	0.238	0.336	0.253	0.636	0.315	0.171
	Compositional	0.280	0.278	0.868	0.256	0.470	0.226	0.088
Interactive-KBQA	I.I.D.	0.053	0.234	0.194	0.036	0.000	0.000	0.053
	Compositional	0.112	0.299	0.132	0.037	0.008	0.000	0.034

Table 7: Evaluation results by question type (raw question, model: GPT-4o-mini)

generates a new thought based on the observations, constructs, and executes an action, utilizing Python syntax for tool execution. This cycle of thought generation and tool execution continues until the action completes with the final result (i.e., when the action is `Done`).

In our experiments, we use gpt-4o-mini-2024-07-18 as the LLM. The maximum turn number is set to the default value of 20. Following the original paper (Xiong et al., 2024), we adopt two configurations for the number of few-shot examples: (1) one example from each question category (all+1-shot), and (2) two examples from the same question category as the input question (same+2-shot).

5.2. Experimental Results

Table 5 shows the experimental results on the *i.i.d.* generalization dataset. In the Text-to-SPARQL zero-shot prompting setting, the results show that answering questions in HOME-KGQA is as difficult as in other datasets.

In the Text-to-SPARQL fine-tuning setting, the results show that HOME-KGQA (Raw) questions are more difficult to answer correctly than those in KQA Pro, but easier than those in WebQSP, CWQ, and MetaQA. In contrast, HOME-KGQA (Paraphrased) questions were found to be the most difficult to answer among all the datasets.

In the Interactive-KBQA experiments, HOME-KGQA was found to be significantly more challenging than the others.

Table 6 shows the experimental results on the compositional generalization dataset. These results indicate that the Text-to-SPARQL approach is effective for acquiring compositional generalization capabilities in HOME-KGQA. In contrast, Interactive-KBQA is less suitable for achieving compositional generalization in this specific environment.

Table 7 shows the accuracy for each question category in HOME-KGQA. The Text-to-SPARQL results are from the fine-tuning setting, and the Interactive-KBQA results are from the cls+2-shot setting. The main reason why Interactive-KBQA failed to answer any *Video* questions correctly is that it often reached the maximum number of allowed turns before finding the correct answer.

In HOME-KGQA, 19.1% of the SPARQL queries generated by the fine-tuned Text-to-SPARQL model contained syntax errors. In contrast, the final SPARQL queries generated by Interactive-KBQA at the end of the dialogue contained no syntax errors. However, 73.8% of Interactive-KBQA cases failed to output `Done` because the final answer could not be observed within the maximum number of turns.

This difficulty is caused by HOME-KGQA integrating five different ontologies into a single, complex schema and by containing many nodes without explicit entity labels. As a result, more interaction turns are required to retrieve the temporal and spatial conditions specified in the question, increasing the risk of action failures at each step. These findings suggest that KGQA, which involves temporally and spatially fine-grained KGs integrating heterogeneous data, such as HOME-KGQA, remains a significant challenge for LLM agent-based approaches.

6. Conclusion

In this paper, we introduced HOME-KGQA, a benchmark dataset for evaluating KGQA models in home environments. By integrating multiple ontologies into a multimodal episodic KG and generating complex question–SPARQL pairs, HOME-KGQA provides a challenging benchmark for real-world reasoning beyond textual encyclopedic facts. Through comparative experiments with existing datasets, HOME-KGQA demonstrated the challenges of introducing current KGQA models into daily activity environments. We expect that HOME-KGQA serves as a foundation for advancing KGQA in real-world contexts.

7. Ethical Considerations

The dataset used in this study, HOME-KGQA, is constructed entirely from synthetic data generated by the VirtualHome simulator (Puig et al., 2018) and contains no personal, biometric, or privacy-related information. We used crowdsourcing solely to collect abstract representations of activity sequences describing typical daily routines. All collected data are non-personal, non-sensitive, and do not include any demographic data.

8. Limitations

The dataset is constructed from synthetic simulations of single-person households and therefore does not capture the full diversity of real-world daily activities, such as multi-person interactions.

From a language resource perspective, the generated questions may reflect the stylistic and lexical tendencies of the underlying LLMs and may lack the linguistic diversity observed in human-authored text. Several paraphrased questions may contain minor errors or unintended shifts in meaning since paraphrasing relies on an LLM. Although manual evaluation showed that the overall paraphrasing accuracy was high (94 out of 100 samples were correctly paraphrased), the possibility of such errors remains a limitation of the current dataset. This issue will be addressed in future work through improved validation and human-in-the-loop refinement. In addition, the questions are not collected from real-world interactions but initially generated using the template-based process. The questions are designed to evaluate whether KGQA models can comprehend the complex structure of knowledge graphs in daily-life environments. The questions are also intended to support future applications in Embodied AI and robotic navigation, where agents must identify spatial coordinates and temporal instants in real-world settings. As a result, the questions tend to be more complex than the questions that naturally occur in everyday communication. We plan to include simpler, more natural questions in future work to broaden the coverage and practical applicability of the dataset.

The target KG represents household daily activities and reuses multiple ontologies to integrate heterogeneous data. However, further verification is needed to evaluate whether the KGQA approaches can generalize to KGs with different schema designs in the domain of daily activities.

9. Acknowledgements

This paper is based on results obtained from JSPS KAKENHI Grant Numbers JP23H03688 and JP25K03232, and AIST policy-based budget project “R&D on Generative AI Foundation Models for the Physical Domain.”

10. Bibliographical References

Dean Allemang and James Hendler. 2011. *Semantic web for the working ontologist: effective modeling in RDFS and OWL*. Elsevier.

Chinnapong Angsuchotmetee, Richard Chbeir, and Yudith Cardinale. 2020. MSSN-Onto: An ontology-based approach for flexible event processing in Multimedia Sensor Networks. *Future Generation Computer Systems*, 108:1140–1158.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary

Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web*, pages 722–735, Berlin, Heidelberg. Springer.

Michael Azmy, Peng Shi, Jimmy Lin, and Ihab Ilyas. 2018. Farewell Freebase: Migrating the SimpleQuestions Dataset to DBpedia. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2093–2103, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jinheon Baek, Alham Fikri Aji, and Amir Saffari. 2023. Knowledge-Augmented Language Model Prompting for Zero-Shot Knowledge Graph Question Answering. In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pages 78–106, Toronto, Canada. Association for Computational Linguistics.

Debayan Banerjee, Pranav Ajit Nair, Jivat Neet Kaur, Ricardo Usbeck, and Chris Biemann. 2022. Modern Baselines for SPARQL Semantic Parsing. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2260–2265. ArXiv:2204.12793 [cs].

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, SIGMOD '08, pages 1247–1250, New York, NY, USA. Association for Computing Machinery.

Don Brutzman and Jakub Flotyński. 2020. X3D Ontology for Querying 3D Models on the Semantic Web. In *Proceedings of the 25th International Conference on 3D Web Technology*, Web3D '20, pages 1–6, New York, NY, USA. Association for Computing Machinery.

Shulin Cao, Jiaxin Shi, Liangming Pan, Lunyiu Nie, Yutong Xiang, Lei Hou, Juanzi Li, Bin He, and Hanwang Zhang. 2022. 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)*, pages 6101–6119, Dublin, Ireland. Association for Computational Linguistics.

Mohnish Dubey, Debayan Banerjee, Abdelrahman Abdelkawi, and Jens Lehmann. 2019. 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*, pages 69–78, Berlin, Heidelberg. Springer-Verlag.

Shusaku Egami, Satoshi Nishimura, and Ken Fukuda. 2021. A Framework for Constructing and Augmenting Knowledge Graphs using Virtual Space: Towards Analysis of Daily Activities. In *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1226–1230. ISSN: 2375-0197.

Shusaku Egami, Mikiko Oono, Mai Otsuki, Takanori Ugai, and Ken Fukuda. 2023a. Analysis of Annotation Quality of Human Activities Using Knowledge Graphs. In *HCI International 2023 Posters*, pages 483–489, Cham. Springer Nature Switzerland.

Shusaku Egami, Takanori Ugai, Swe Nwe Nwe Htun, and Ken Fukuda. 2024. VHAKG: A Multimodal Knowledge Graph Based on Synchronized Multi-view Videos of Daily Activities. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, pages 5360–5364, New York, NY, USA. Association for Computing Machinery.

Shusaku Egami, Takanori Ugai, Mikiko Oono, Koji Kitamura, and Ken Fukuda. 2023b. Synthesizing Event-Centric Knowledge Graphs of Daily Activities Using Virtual Space. *IEEE Access*, 11:23857–23873.

Aleksandr Gashkov, Aleksandr Perevalov, Maria Eltsova, and Andreas Both. 2025. SPARQL Query Generation with LLMs: Measuring the Impact of Training Data Memorization and Knowledge Injection. ArXiv:2507.13859 [cs] version: 1.

Yu Gu, Sue Kase, Michelle Vanni, Brian Sadler, Percy Liang, Xifeng Yan, and Yu Su. 2021. Beyond I.I.D.: Three Levels of Generalization for Question Answering on Knowledge Bases. In *Proceedings of the Web Conference 2021*, WWW '21, pages 3477–3488, New York, NY, USA. Association for Computing Machinery.

Krzysztof Janowicz, Armin Haller, Simon J.D. Cox, Danh Le Phuoc, and Maxime Lefrançois. 2019. SOSA: A lightweight ontology for sensors, observations, samples, and actuators. *Web Semant.*, 56(C):1–10.

Longquan Jiang and Ricardo Usbeck. 2022. Knowledge Graph Question Answering Datasets and Their Generalizability: Are They Enough for Future Research? In *Proceedings of the 45th International ACM SIGIR*

Conference on Research and Development in Information Retrieval, SIGIR '22, pages 3209–3218, New York, NY, USA. Association for Computing Machinery.

Mayank Kejriwal, Vanessa Lopez, Juan F. Sequeda, Simon Gottschalk, Elena Demidova, Mayank Kejriwal, Vanessa Lopez, and Juan F. Sequeda. 2019. EventKG – the hub of event knowledge on the web – and biographical timeline generation. *Semant. web*, 10(6):1039–1070.

Jens Lehmann, Preetam Gattogi, Dhananjay Bhandiwad, Sébastien Ferré, and Sahar Vahdati. 2023. Language Models as Controlled Natural Language Semantic Parsers for Knowledge Graph Question Answering. In Kobi Gal, Ann Nowé, Grzegorz J. Nalepa, Roy Fairstein, and Roxana Rădulescu, editors, *Frontiers in Artificial Intelligence and Applications*. IOS Press.

Zhenyu Li, Sunqi Fan, Yu Gu, Xiuxing Li, Zhichao Duan, Bowen Dong, Ning Liu, and Jianyong Wang. 2024. FlexKBQA: A Flexible LLM-Powered Framework for Few-Shot Knowledge Base Question Answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18608–18616. Number: 17.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-Value Memory Networks for Directly Reading Documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1409, Austin, Texas. Association for Computational Linguistics.

Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. 2018. VirtualHome: Simulating Household Activities Via Programs. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8494–8502, Salt Lake City, UT. IEEE.

Tarcísio Souza Costa, Simon Gottschalk, and Elena Demidova. 2020. Event-QA: A Dataset for Event-Centric Question Answering over Knowledge Graphs. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, pages 3157–3164, New York, NY, USA. Association for Computing Machinery.

Nadine Steinmetz and Kai-Uwe Sattler. 2021. What is in the KGQA Benchmark Datasets? Survey on Challenges in Datasets for Question Answering on Knowledge Graphs. *Journal on Data Semantics*, 10(3):241–265.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M. Ni, Heung-Yeung Shum, and Jian Guo. 2024. Think-on-Graph: Deep and Responsible Reasoning of Large Language Model on Knowledge Graph. ArXiv:2307.07697 [cs].

Alon Talmor and Jonathan Berant. 2018. The Web as a Knowledge-Base for Answering Complex Questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 641–651, New Orleans, Louisiana. Association for Computational Linguistics.

Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023. Can ChatGPT Replace Traditional KBQA Models? An In-Depth Analysis of the Question Answering Performance of the GPT LLM Family. In *The Semantic Web – ISWC 2023*, pages 348–367, Cham. Springer Nature Switzerland.

Kerry Taylor, Armin Haller, Maxime Lefrancois, Simon Cox, Raul Garcıa-Castro, Danh Le-Phuoc, Joshua Lieberman, and Claus Stadler. 2019. The Semantic Sensor Network Ontology, Revamped. *Proceedings of the Journal Track co-located with the 18th International Semantic Web Conference (ISWC 2019)*.

Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. 2015. Representing Text for Joint Embedding of Text and Knowledge Bases. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Lisbon, Portugal. Association for Computational Linguistics.

Priyansh Trivedi, Gaurav Maheshwari, Mohnish Dubey, and Jens Lehmann. 2017. 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*, pages 210–218, Berlin, Heidelberg. Springer-Verlag.

Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Bastian Haarmann, Anastasia Krithara, Michael Röder, and Giulio Napolitano. 2017. 7th Open Challenge on Question Answering over Linked Data (QALD-7). In *Semantic Web Challenges*, pages 59–69, Cham. Springer International Publishing.

Alexandros Vassiliades, Nick Bassiliades, Filippos Gouidis, and Theodore Patkos. 2020. A Knowledge Retrieval Framework for Household Objects and Actions with External Knowledge. In

Eva Blomqvist, Paul Groth, Victor De Boer, Tassilo Pellegrini, Mehwish Alam, Tobias Käfer, Peter Kieseberg, Sabrina Kirrane, Albert Meroño-Peñuela, and Harshvardhan J. Pandit, editors, *Semantic Systems. In the Era of Knowledge Graphs*, volume 12378, pages 36–52. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Julio Vizcarra, Satoshi Nishimura, and Ken Fukuda. 2021. Ontology-based human behavior indexing with multimodal video data. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, pages 262–267. ISSN: 2325-6516.

Denny Vrandečić. 2012. Wikidata: a new platform for collaborative data collection. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, pages 1063–1064, New York, NY, USA. Association for Computing Machinery.

Guanming Xiong, Junwei Bao, and Wen Zhao. 2024. Interactive-KBQA: Multi-Turn Interactions for Knowledge Base Question Answering with Large Language Models. ArXiv:2402.15131 [cs].

Wen-tau Yih, Matthew Richardson, Chris Meek, Ming-Wei Chang, and Jina Suh. 2016. The Value of Semantic Parse Labeling for Knowledge Base Question Answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–206, Berlin, Germany. Association for Computational Linguistics.

Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alexander J. Smola, and Le Song. 2018. Variational reasoning for question answering with knowledge graph. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18, pages 6069–6076, New Orleans, Louisiana, USA. AAAI Press.

Xiangru Zhu, Zhixu Li, Xiaodan Wang, Xueyao Jiang, Penglei Sun, Xuwu Wang, Yanghua Xiao, and Nicholas Jing Yuan. 2024. Multi-Modal Knowledge Graph Construction and Application: A Survey. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 36(2).

11. Language Resource References

Shulin Cao and Jiaxin Shi and Liangming Pan and Lunyiu Nie and Yutong Xiang and Lei Hou and Juanzi Li and Bin He and Hanwang Zhang. 2022. *KQA Pro*. 1.0.

Alon Talmor and Jonathan Berant. 2018. *ComplexWebQuestions*.

Wen-tau Yih and Matthew Richardson and Christopher Meek and Ming-Wei Chang and Jina Suh. 2016. *WebQuestions Semantic Parses Dataset*. Microsoft, 1.0.

Yuyu Zhang and Hanjun Dai and Zornitsa Kozareva and Alexander J. Smola and Le Song. 2018. *MetaQA*.