





Evaluating Large Language Models in Semantic Parsing for Conversational Question Answering over Knowledge Graphs

Phillip Schneider¹^a, Manuel Klettner¹, Kristiina Jokinen²^b, Elena Simperl³^c
and Florian Matthes¹^d

¹Department of Computer Science, Technical University of Munich, Germany

²AI Research Center, National Institute of Advanced Industrial Science and Technology, Japan

³King's College London, Department of Informatics, U.K.

Keywords: Conversational Question Answering, Knowledge Graphs, Large Language Models, Semantic Parsing.


Abstract: Conversational question answering systems often rely on semantic parsing to enable interactive information retrieval, which involves the generation of structured database queries from a natural language input. For information-seeking conversations about facts stored within a knowledge graph, dialogue utterances are transformed into graph queries in a process that is called knowledge-based conversational question answering. This paper evaluates the performance of large language models that have not been explicitly pre-trained on this task. Through a series of experiments on an extensive benchmark dataset, we compare models of varying sizes with different prompting techniques and identify common issue types in the generated output. Our results demonstrate that large language models are capable of generating graph queries from dialogues, with significant improvements achievable through few-shot prompting and fine-tuning techniques, especially for smaller models that exhibit lower zero-shot performance.


1 INTRODUCTION


Conversational search has emerged as a growing area of interest within the information retrieval research field in recent years. This burgeoning search paradigm casts the information-seeking process into multi-turn dialogues with conversational agents. The latter are designed to facilitate exploring and gradually narrowing down the search scope to relevant information items (Aliannejadi et al., 2021). A fundamental aspect of these agents is their ability to access data stored in knowledge bases with an inherent semantic structure, such as relational databases or knowledge graphs. Hence, a key challenge lies in bridging the gap between natural language utterances, where users express their information needs, and corresponding formal representations, such as logical forms or executable queries. In the field of natural language processing (NLP), the task of semantic parsing focuses on this problem by deriving machine-readable meaning representations given linguistic in-


puts. Conversational question answering (CQA) over knowledge graphs is a specialized facet of conversational search that revolves around responding to user queries in a dialogue given an underlying knowledge graph. Through semantic parsing mechanisms, dialogue utterances are transformed into executable queries to retrieve answer triples from a graph.

With the advent of pre-trained large language models (LLMs), the field of NLP has witnessed a shift in methodologies. Unlike conventional supervised learning approaches that rely on annotated datasets, LLMs are trained in a self-supervised manner, predicting tokens within vast amounts of unlabeled data. Combined with scaling up model size and training corpora, this approach has demonstrated remarkable emergent capabilities of LLMs and their prowess in multi-task learning (Radford et al., 2019). Given a carefully defined input prompt, LLMs have the advantage of prompt-based learning, or in-context learning, which allows them to perform a range of generative tasks like, for instance, question answering, machine translation, or semantic parsing (Liu et al., 2023). There has been a growing interest in optimizing LLMs for conversational interactions using instruction fine-tuning and reinforcement learning from

^a <https://orcid.org/0000-0001-9492-2927>

^b <https://orcid.org/0000-0003-1229-239X>

^c <https://orcid.org/0000-0003-1722-947X>

^d <https://orcid.org/0000-0002-6667-5452>

human feedback (OpenAI, 2022). Although LLMs offer tremendous potential, it is crucial to acknowledge their inherent limitations, such as the risk of hallucinating or omitting information and a lack of accountability in high-risk scenarios, with limited transparency of the information sources from which they derive their outputs (Ji et al., 2023). Therefore, it becomes imperative to ground their generated outputs in verifiable facts contained in knowledge bases, which can be facilitated through semantic parsing.

The goal of our study lies in investigating how LLMs perform in semantic parsing of dialogues for CQA over knowledge graphs. To answer the stated question, we systematically compare generated outputs from LLMs of varying sizes and training objectives, with a primary focus on models optimized for conversational interaction. Based on an extensive benchmark dataset, we evaluate the models' performance in the generation of *SPARQL* queries from dialogues about knowledge graph facts and discuss insights about their individual capabilities as well as limitations. Our contributions include a benchmark study evaluating four LLMs, utilizing both automatic metrics and human evaluation to identify eight common error types in generated graph queries, and a detailed discussion of prompting and fine-tuning strategies aimed at improving model performance. To ensure full reproducibility of our experiments, we have established a GitHub repository encompassing all model scripts, datasets, and evaluation outputs.¹

2 RELATED WORK

Early semantic parsing methods were characterized by symbolic approaches rooted in production grammars and handcrafted linguistic features. With significant advancements driven by deep learning, there has been a shift towards neural approaches that cast semantic parsing as a machine translation problem by training neural networks that convert natural language input into a formal target language. The adaptability of neural networks eliminates the necessity of defining lexicons or manually crafted features, enabling models to generalize across various domains and meaning representation languages (Wang et al., 2020).

While the majority of existing work focuses on parsing independent natural language utterances without considering broader contextual information, a growing body of literature is about contextualized semantic parsing that takes surrounding information be-

yond the current utterance into account, such as interaction histories (Li et al., 2020). Therefore, context-aware parsing is particularly relevant for CQA scenarios characterized by a series of interrelated utterances, ambiguous queries, and evolving search intents. Evaluating CQA methods poses a considerable challenge, primarily due to the scarcity of available datasets constructed for this task. Since the few benchmarks for CQA are often limited in scale, domain specificity, or dialogue length, our study takes advantage of the recently published dataset SPICE (Perez-Beltrachini et al., 2023), preventing benchmark leakage, where data from evaluation sets is occasionally used for LLM pre-training. Derived from the CSQA dataset (Saha et al., 2018), an established benchmark for retrieval-based CQA, SPICE extends CSQA by pairing dialogues with executable SPARQL queries along with answer triples from the Wikidata knowledge graph. Further details about this dataset are provided in Section 3.

The creators of the SPICE benchmark tested two strong baseline models and analyzed their performance across various question types. The first approach *BertSP* adopts a standard sequence-to-sequence architecture for generating complete SPARQL queries (Gu et al., 2021). To tackle the extensive vocabulary of the Wikidata knowledge graph, dynamic vocabularies for entities and relations are derived from knowledge subgraphs corresponding to each question and its contextual information. This approach combines a BERT-based encoder (Devlin et al., 2019), fine-tuned specifically for semantic parsing, with a randomly initialized transformer network as the decoder. The second approach *LasagneSP* adapts the *LASAGNE* architecture, as proposed by Kacupaj et al. (2021). It employs an encoder-decoder transformer network to generate base logical forms (i.e., SPARQL templates), while a graph attention model is used to produce node representations by exploiting correlations between entity types and relations. An entity recognition module based on an inverted index is also part of the architecture.

To the best of our knowledge, we are the first to evaluate the emergent capabilities of LLMs that have not been explicitly trained for conversational semantic parsing. Unlike previously mentioned models, which chain multiple fine-tuned neural networks in a sequence, we apply in-context learning and post-processing. This enables the end-to-end generation of structured queries. We seek to investigate how LLMs perform in understanding dialogues, resolving vocabularies, and generating SPARQL queries with correct syntax. Consequently, our objective transcends the scope of individual models; it strives for a compre-

¹<https://github.com/sebischair/LLM-SP-CQA>

Table 1: Overview of the two applied prompts. Parts marked with “<>” denote variables that are inserted at runtime based on the current test example. The complete few-shot prompt with three examples is provided in the linked repository.

Prompt	Content
Zero-shot	<p>SYSTEM: Generate a SPARQL query that answers the given 'Input question:'. Use 'Entities:', 'Relations:' and 'Types:' specified in the prompt to generate the query. The SPARQL query should be compatible with the Wikidata knowledge graph. Prefixes like 'wdt' and 'wd' have already been defined. No language tag is required. Use '?x' as variable name in the SPARQL query. Remember to provide only a SPARQL query in the response without any notes, comments, or explanations.</p> <p>USER: <conversation_history> Input question: <utterance> Entities: <entities> Relations: <relations> Types: <types></p>
Few-shot example	<p>[...] USER: Conversation history: USER: Which administrative territory is the native country of Cirilo Villaverde? SYSTEM: {'Q241': 'Cuba'} Input question: Which is the national anthem of that administrative territory ? Entities: {'Q241': 'Cuba'} Relations: {'P85': 'anthem'} Types: {'Q484692': 'hymn'} ASSISTANT: SPARQL query: SELECT ?x WHERE { wd:Q241 wdt:P85 ?x . ?x wdt:P31 wd:Q484692 . } [...]</p>

hensive understanding of using LLMs in conversational semantic parsing over knowledge graphs.

3 EXPERIMENTAL SETUP

Benchmark Dataset. Our study aims to assess the performance of LLMs in knowledge-based conversational search, with a particular focus on semantic parsing for CQA. For the evaluation, we have chosen to use the *SPICE* dataset from Perez-Beltrachini et al. (2023). The dataset comprises conversational interactions between a user and an assistant. Each independent conversation is paired with SPARQL queries, which are executable against a knowledge graph engine to retrieve answers from a Wikidata snapshot. Furthermore, the dialogue transcripts within the SPICE dataset showcase conversational phenomena such as coreference, ellipsis, and clarifications. In some instances, clarifying questions and user responses accompany the SPARQL parse and query results. Obtaining correct SPARQL queries and corresponding answers requires handling a variety of different questions. The higher-order question types can be distinguished into logical reasoning, quantitative reasoning, comparative reasoning, verification, and simple questions.

In total, the SPICE dataset consists of 197,000 dialogues, with an average of 9.5 conversation turns. Because the dataset only contains Wikidata references for entities, types, and relations, we carried out pre-processing steps to map references to their lexical forms. This was done to avoid relying on the models' intrinsic knowledge for resolving these lexical forms. For example, if an entity reference corresponds to the Wikidata ID Q30, we include the label "United States of America" through a lookup via the Wikidata.org website. From the more than 150,000 training ex-

amples, we constructed a smaller fine-tuning dataset with 30,000 conversations. These conversations contained between one and four independent conversational turns to simulate zero- and few-shot prompting, maintaining the same system message and prompt structure as during inference. Another preparation step was to sample down the test examples to a subset with 1,500 conversations. The reason for this was the resource constraint of keeping each model's required inference run time below 24 hours. To construct this test subset, we computed the distribution of the entire test set across all question categories and then determined the required samples for each category.

Models. We compare four large language models of varying sizes with different prompting techniques. As a popular state-of-the-art LLM that is closed-source, we opted to include *GPT-3.5-Turbo (ChatGPT)* (OpenAI, 2022) in our comparison. It is optimized for dialogue interaction and has demonstrated remarkable zero-shot performance on various NLP tasks and is often used as a benchmark for comparing LLMs. We conducted our semantic parsing experiments with the model version GPT-3.5-Turbo-0613. Further, we decided to test *LLaMA*, a collection of LLMs developed and open-sourced by Meta (Touvron et al., 2023). We include three model variations with 7B parameters of the first LLaMA version. In addition to the non-conversational base model, we included a fine-tuned model we abbreviate as *LoRA*. The training was done through low-rank adaptation (LoRA) with 30,000 examples, a method that fine-tunes only a subset of the model's parameters, referred to as low-rank matrices, rather than updating the entire parameter space, improving the fine-tuning efficiency (Hu et al., 2022). Another fine-tuned LLaMA model we tested is named *Vicuna*, which was trained on a corpus of roughly 70,000 user-shared ChatGPT conversations crawled

from the ShareGPT website (Chiang et al., 2023).

We set the token limit to 128 and the temperature parameter to 0, maximizing deterministic generation by favoring tokens with the highest probability. All models are prompted in the chat completion structure of the FastChat² platform, with a structured list of system, user, and assistant messages. Table 1 displays the structure of each prompt. The main instruction is given as a system message. The user message contains the question, lexical forms of entities as well as relations, and the conversation history, which is created by including up to three last dialogue turns. The zero-shot prompt contains only a system message with a semantic parsing instruction. The few-shot prompt expands the instruction with three in-context examples of the task with different SPARQL constructs, such as ASK, SELECT, or COUNT. Furthermore, one example demonstrates using the conversation history to resolve an entity referenced from a previous conversational turn.

4 RESULTS AND DISCUSSION

Automatic Evaluation Results. The performance metrics of the semantic parsing experiments for simple questions and complex questions are presented in Table 2. To ensure consistency in the evaluation, all metrics were computed on post-processed model predictions, which involved normalizing whitespace characters and removing “SPARQL query:” from the beginning of the generated output. Based on 1,500 conversations of the SPICE test dataset, we computed the exact match (EM) ratio, comparing the predicted queries to the ground truth queries. While F1 scores were calculated for question categories yielding sets of entities, the accuracy (ACC) metric was employed to measure performance in cases where the results constituted count or boolean values. The range for each metric lies between 0 to 1, with the optimal score being 1. The provided tables report results for 7 out of the 10 question types, focusing on those for which at least one model achieved a reasonable performance score above zero. Aside from the four LLMs, we added results from a “LoRA-7B-512” model, which used a maximum token length of 512 instead of 128 tokens. This was done to assess if increasing the token limit could enhance the top-performing model.

Upon inspecting the metrics for simple questions in Table 2, it becomes evident that LLaMA and Vicuna show the worst performance, particularly with zero-shot prompting, where they fail to pro-

duce valid queries regardless of the question type. Although the provision of in-context examples significantly improves their performance on direct and coreference questions, achieving F1 scores of up to 0.352 for LLaMA and 0.189 for Vicuna, few-shot prompting does not extend to their ability to handle questions that involve ellipsis. The GPT-3.5-Turbo model demonstrates superior performance compared to LLaMA and Vicuna, being capable of parsing queries in zero-shot settings and effectively addressing questions with an ellipsis. Notably, the fine-tuned LoRA model consistently surpasses all models, producing exact ground truth SPARQL queries with an ACC ranging from 75% to 97%. It can be observed that LoRA often performs better without few-shot

Table 2: Semantic parsing performance for simple and complex questions evaluated by F1 score (F1), accuracy (ACC), and ratio of exact matches (EM). Bold values indicate the best performance for each metric.

Model	Zero-Shot		Few-Shot	
	F1	EM	F1	EM
Simple Question (Direct)				
LLaMA-7B	0.000	0.000	0.352	0.724
Vicuna-7B	0.003	0.000	0.127	0.230
GPT-3.5-Turbo	0.324	0.337	0.804	0.741
LoRA-7B	0.867	0.970	0.963	0.917
LoRA-7B-512	0.867	0.970	-	-
Simple Question (Coreference)				
LLaMA-7B	0.000	0.000	0.350	0.568
Vicuna-7B	0.000	0.000	0.189	0.321
GPT-3.5-Turbo	0.491	0.234	0.636	0.623
LoRA-7B	0.882	0.867	0.844	0.786
LoRA-7B-512	0.892	0.873	-	-
Simple Question (Ellipsis)				
LLaMA-7B	0.000	0.000	0.000	0.000
Vicuna-7B	0.000	0.000	0.000	0.000
GPT-3.5-Turbo	0.342	0.158	0.609	0.351
LoRA-7B	0.855	0.754	0.618	0.526
LoRA-7B-512	0.855	0.754	-	-
Logical Reasoning (All)				
LLaMA-7B	0.000	0.000	0.109	0.000
Vicuna-7B	0.000	0.000	0.001	0.000
GPT-3.5-Turbo	0.631	0.000	0.912	0.246
LoRA-7B	0.900	0.926	0.810	0.779
LoRA-7B-512	0.900	0.926	-	-
Comparative Reasoning (All)				
LLaMA-7B	0.000	0.000	0.001	0.000
Vicuna-7B	0.000	0.000	0.072	0.000
GPT-3.5-Turbo	0.015	0.000	0.006	0.000
LoRA-7B	0.000	0.000	0.001	0.000
LoRA-7B-512	0.315	0.114	-	-
Zero-Shot ACC				
Model	Zero-Shot		Few-Shot	
	ACC	EM	ACC	EM
Verification (Boolean) (All)				
LLaMA-7B	0.000	0.000	0.000	0.000
Vicuna-7B	0.000	0.000	0.365	0.162
GPT-3.5-Turbo	0.000	0.000	0.926	0.480
LoRA-7B	0.939	0.851	0.926	0.777
LoRA-7B-512	0.939	0.858	-	-
Quantitative Reasoning (Count) (All)				
LLaMA-7B	0.000	0.000	0.152	0.027
Vicuna-7B	0.000	0.000	0.091	0.008
GPT-3.5-Turbo	0.197	0.008	0.485	0.212
LoRA-7B	0.591	0.561	0.492	0.417
LoRA-7B-512	0.591	0.561	-	-

²FastChat: <https://github.com/lm-sys/FastChat>

Table 3: Overview of the eight identified error types with examples from model generated predictions (PRED) and ground truth (GT) queries. Errors in the predictions are highlighted in red color.

Error Type	Definition	Example
Cutoff	PRED matches GT exactly but ends abruptly.	GT: [...] WITH { SELECT DISTINCT ?x (0 AS ?upcount) WHERE { { { ?x wdt:P122 ?b . ?x wdt:P31 wd:Q7275 . } } FILTER NOT EXISTS [...] } PRED: [...] WITH { SELECT DISTINCT ?x (0 AS ?upcount) WHERE { { { ?x wdt:P122 ?b . ?x wdt:P31 w
Deviating entities	PRED uses entity reference not specified in the prompt.	GT: SELECT DISTINCT ?x WHERE { ?x wdt:P101 ?y . VALUES ?y { wd:Q1622272 wd:Q170790 . } . ?x wdt:P31 wd:Q502895 . } } PRED: SELECT ?x WHERE { ?x wdt:P101 wd:Q1622272 . ?x wdt:P101 wd:Q170790 . ?x wdt:P31 wd:Q5 . }
Alternative query	Alternative SPARQL query but correct result.	GT: SELECT ?x WHERE { wd:Q6177791 wdt:P451 ?x . ?x wdt:P31 wd:Q502895 . } PRED: SELECT ?x WHERE { ?x wdt:P451 wd:Q6177791 . ?x wdt:P31 wd:Q502895 . }
Incorrect result	Valid SPARQL query but incorrect result.	GT: SELECT ?x WHERE { wd:Q6177791 wdt:P451 ?x . ?x wdt:P31 wd:Q502895 . } PRED: SELECT ?x WHERE { ?x wdt:P451 ?p . ?p wdt:Q502895 ?type . ?type wdt:commonName ?x . }
Language filter	PRED contains language filter.	GT: SELECT ?x WHERE { wd:Q123179 wdt:P69 ?x . ?x wdt:P31 wd:Q163740 . } PRED: SELECT ?x WHERE { ?x wdt:P69 ?y . FILTER (LANG(?y)= 'en') . } LIMIT 1
Namespace definition	PRED uses name-spaces instead of wd and wdt	GT: SELECT ?x WHERE { [...] } PRED: PREFIX wdt: <http://www . wikidata . org/entity/>PREFIX wd:<http://www . wikidata . org/prop/direct/> SELECT ?x WHERE { [...] }
Off-prompt	PRED is unrelated to prompt and contradicts desired output format.	GT: SELECT ?x WHERE { wd:Q23487488 wdt:P702 ?x . ?x wdt:P31 wd:Q863908 . } PRED: Input question: What is the nucleic acid sequence that is encoded by 16S rRNA methyltransferase GidB SSA_0605 ? Entities: { 'Q23487488': '16S rRNA methyltransferase [...]
Syntax error	PRED is invalid SPARQL	GT: SELECT DISTINCT ?x WHERE { ?x wdt:P166 ?y . VALUES ?y { wd:Q918055 wd:Q133160 wd:Q920783 . } . ?x wdt:P31 wd:Q502895 . } PRED: SELECT ?x WHERE { ?x wdt:P166 ?award ?award wdt:Q918055 ?award wdt:Q133160 ?award wdt:Q920783 }

prompting. We hypothesize that the examples of the few-shot prompt might introduce a superfluous information bias since the model already learned from task-specific examples during fine-tuning. When dealing with dialogue phenomena such as coreferences (i.e., linguistic expressions like pronouns referring back to entities mentioned in a previous turn) and ellipsis (i.e., omission of one or more words for brevity because they can be inferred from the dialogue context), generating parses that precisely match the ground truth proves challenging for all LLMs. This aligns with observations from Perez-Beltrachini et al. (2023), where the SPICE baseline models also struggled with these two phenomena. Still, the fine-tuned LoRA model handles these complexities well, achieving similar F1 scores across all simple question types.

Concerning the LLMs’ performance metrics on more complex question types, semantic parsing proves to be more difficult. Complex questions require a number of logical and numerical operations over entity sets associated with longer SPARQL parses (e.g., *How many bodies of water or water-courses are situated nearby Lübeck?*). LLaMA and Vicuna exhibit inferior performance compared to simple questions, with the exception of verification questions that result in a boolean value. The latter is the only category where Vicuna outperforms LLaMA with an F1 score of 0.365. The substantially larger GPT-3.5-Turbo model excels in logical and verification questions in few-shot scenarios, even though it took few-shot examples to get the model to use the ASK instead of the SELECT operator for verification questions that should return a boolean value. The exceptional performance could be attributed to the model’s explicit training on tasks that involve logical reasoning operations. Another interesting observa-

tion about these question types is that GPT-3.5-Turbo demonstrates the ability to infer parses that yield correct results, achieving comparable F1 scores to LoRA, even though the EM ratio is considerably lower, suggesting that it has learned to convey the same question intent through an alternative SPARQL expression.

Overall, LoRA emerges again as the best-performing LLM, producing the highest number of EM queries; however, for the most complex question types, such as quantitative and comparative reasoning, its performance is limited, akin to the other models. It is worth noting that using LoRA with 512 instead of 128 maximum tokens leads only to better performance on comparative reasoning questions. This indicates that the model successfully generated correct outputs for a few queries but often terminated abruptly upon reaching the token limit, which, in turn, resulted in syntax errors. Given that the ground truth queries for comparative reasoning are relatively lengthy, extending the maximum token limit even further could yield enhancements in performance.

Human Evaluation Results. Besides our metric-based performance assessment of semantic parsing, we carried out a qualitative human evaluation to get further insights into the LLMs’ output. Two researchers manually analyzed a sample of 15 generated queries for each of the 10 question types, resulting in the examination of a total of 150 predictions, although six ground truth queries were absent from the SPICE dataset and were thus excluded from the analysis. By employing an iterative process involving the creation and consolidation of error categories, we successfully identified eight prevalent error types, as delineated in Table 3. For each error type, we provide a short def-

Table 4: Relative frequency of error categories for zero-shot and few-shot prompts in the evaluated sample of 150 predictions. The asterisk symbols denote: “*” excluding off-prompt predictions, “**” excluding off-prompt and cutoff predictions, and “***” excluding off-prompt and syntax error predictions.

Error Type	LLaMA-7B	Vicuna-7B	GPT-3.5-Turbo	LoRA-7B
	Relative error frequency: zero-shot / few-shot			
Cutoff	- / -	- / -	- / -	0.33 * / 0.24 *
Deviating entities	- / 0.05 *	0.04 * / 0.02 *	0.06 * / 0.03 *	0.02 * / 0.04 *
Alternative query	- / -	- / -	0.08 / 0.06	0.01 / 0.01
Incorrect result	- / 0.82 ***	0.97 *** / 0.86 ***	0.69 *** / 0.63 ***	0.12 *** / 0.20 ***
Language filter	- / -	0.33 * / 0.06 *	0.07 * / 0.02	- / -
Namespace definition	- / -	0.11 * / -	- / -	- / -
Off-prompt	1.00 / 0.10	0.13 / 0.10	0.10 / 0.10	0.10 / 0.10
Syntax error	- / 0.16 **	0.71 ** / 0.26 **	0.20 ** / 0.17 **	0.01 ** / 0.10 **

initiation accompanied by an example to juxtapose the ground truth query with the erroneous generated output. For instance, the LLMs sometimes ignored the instruction given in the prompt. In other cases, they included entities that were wrong or not specified previously, cut off abruptly in the middle of the query, or generated parses with syntactical errors.

To gain a more profound understanding of the error occurrence rates specific to each model and prompt combination, we present the relative frequencies of these error types in Table 4. Many of these errors manifested in the predictions generated by LLaMA and Vicuna. Outputs from GPT-3.5-Turbo and LoRA exhibited a higher degree of reliability and a diminished incidence of such errors. Vicuna, GPT-3.5-Turbo, and LoRA demonstrate the ability to generate zero-shot output that aligns with the desired output written in the prompt. This outcome is consistent with expectations for instruction-tuned and fine-tuned models, suggesting their efficacy in aligning with user instructions and prompts. However, all outputs produced by the LLaMA zero-shot model are off-prompt (1.00), meaning that they did not contain a SPARQL query as the only desired output format, suggesting that a textual description of a complex task without including in-context examples is insufficient for LLaMA. An intriguing finding is that the issue of off-prompt errors can be effectively mitigated in all models by including SPARQL examples within the prompt, thus enhancing model performance and alignment with user intent. It is worth mentioning that across all LLMs, 0.10 is the lower bound for the corresponding relative frequencies. The reason for this observation is the clarification question type. Because our study exclusively focuses on semantic parsing, we do not consider clarifying questions in the instructions leading to off-prompt behavior.

Output classified as an *incorrect result* represents the relative frequency of a syntactically valid generated query that retrieved the wrong result (e.g., boolean, entity set, or integer). Within the few-shot setting, Vicuna demonstrated the least favorable performance, with an error rate of 0.86, followed by

LLaMA (0.82) and GPT-3.5-Turbo (0.63). In contrast, LoRA produced queries with the fewest incorrect results, with a ratio of 0.20. Notably, in the case of LoRA, introducing few-shot examples resulted in nearly double the number of incorrect results compared to its zero-shot performance (0.12). This phenomenon suggests that the inclusion of few-shot examples may exert a negative bias on the already fine-tuned LoRA model.

A similar pattern of few-shot behavior becomes evident when assessing queries with *syntax errors*. This type was used to classify non-executable queries. Except for the LoRA model, few-shot prompting improves the ratio of syntactically valid queries compared to zero-shot prompting. GPT-3.5-Turbo (0.17) is significantly better than Vicuna (0.26) in both zero-shot and few-shot scenarios, while LLaMA (0.16) few-shot achieves a very similar occurrence rate as GPT-3.5-Turbo. LoRA generates the smallest number of syntax mistakes with few-shot (0.10) and even less in zero-shot prompting (0.01). We hypothesize that this is due to its exposure to 30,000 examples of correct SPARQL queries during fine-tuning.

Another common error type pertains to *deviating entities*, wherein Wikidata references are used in the predicted query without being explicitly specified as part of the prompt. This error type has a uniform relative frequency across all model-prompt combinations. Looking at these errors more closely, we see cases in which parts of the original reference are omitted, for example, using the Wikidata ID Q5 instead of the provided one Q502895. Moreover, two analyzed samples offer limited information regarding the conversation history of the prompt, leading to models hallucinating other entities or relation references when the information is unavailable. This issue could potentially be alleviated by including references for all relevant entities, types, and relations within the full conversation history of the prompt. In the system prompt, the LLMs are explicitly instructed to only use specified entities, relations, and types.

Furthermore, we instructed the models to refrain from defining namespace prefixes and to use the

Wikidata internal prefixes “wdt” and “wd” instead. Considering errors with *namespace definitions*, we measure the relative frequency where the model does not follow that instruction. Only Vicuna (zero-shot) shows this undesirable behavior with a frequency of 0.11. Similar to this error type, we also analyze how well the model follows the system instruction to refrain from using *language filters*. The SPICE knowledge graph only contains English triples, whereas no language labels are provided. Consequently, filtering in the query for a language does not yield any results, even if it is syntactically valid. Therefore, in the system prompt, we specify that the generated query should not filter for languages. This error type was only observed in generations from Vicuna and GPT-3.5-Turbo, with few-shot prompting leading to significant improvements over zero-shot. In-context examples reduced the error for Vicuna from 0.33 to 0.06 and GPT-3.5-Turbo from 0.07 to 0.02.

A prediction is considered to be *cutoff* if it matches the expected output, but the generation stops abruptly before completing the query. The maximum token length, a hyperparameter for LLMs, is the cause of this issue. Increasing it can mitigate the problem. Regarding the analyzed model-prompt combinations, such errors are only found in predictions from LoRA. All models except LoRA deviate from the ground truth query before reaching maximum tokens. To see if we can improve the generations for LoRA, we experimented with increasing the limit from 128, which we use as standard for all model-prompt configurations, to 512. Although the four times higher limit improves the performance as shown in Table 2, the token limit is only reached for very complex question types, such as comparative reasoning questions.

Lastly, the error type named *alternative query* was used to classify predictions that constitute valid SPARQL queries that yield a correct result, albeit not exactly matching with the ground truth query from the SPICE dataset. Hence, this error type reduces the EM performance while leaving the ACC and F1 scores unaffected. As presented in Table 4, the generation of alternative queries was only observed in outputs from GPT-3.5-Turbo and LoRA. The GPT-3.5-Turbo model, specifically in the zero-shot setting, produced the highest number of instances within this category (0.08), with few-shot prompting decreasing it further (0.06). We assume that this may be attributed to GPT-3.5-Turbo’s extensive pre-training, which likely equipped it with a deeper understanding of SPARQL queries, enabling it to formulate alternative queries that still yield the correct results. Conversely, LoRA generated this type of substitute query in merely 1% of the analyzed outputs, with no discernible difference

between zero-shot and few-shot prompting settings.

Discussion. Through our study’s experimental results, we gained several valuable insights into how LLMs perform in semantic parsing for CQA. Each model we evaluated demonstrated at least a degree of proficiency in generating structured SPARQL queries, even if they were not explicitly trained for this specific task. Some LLMs showed the ability to handle coreference and ellipsis within the context of simple questions. This aptitude indicates their capacity to leverage contextual information from the dialogue histories to produce correct SPARQL queries from ambiguous user questions. Nevertheless, when faced with these linguistic phenomena, especially in more complex question types, the LLMs’ performance experienced a significant decrease.

When analyzing overall performance as a weighted average score comprised of ACC and F1 scores across all question types, LLaMA base model demonstrates almost twice as good performance as the fine-tuned Vicuna model. This may suggest that fine-tuning on conversational data, as in the case of Vicuna, might have a trade-off, potentially leading to a decrease in generative capabilities concerning structured query languages. The significantly larger GPT-3.5-Turbo model outperformed LLaMA and Vicuna with zero- and few-shot prompting. We found GPT-3.5-Turbo’s ability to generate alternative queries especially interesting. Although these queries did not match the ground truth query, they still managed to return correct results, which may be attributed to extensive pre-training on documents containing structured, formal languages, equipping the model with substantial knowledge of SPARQL. Our fine-tuned 7B parameter LoRA model surpassed the performance of the considerably larger GPT-3.5-Turbo model. Our analysis of common errors also revealed that LoRA consistently generated the fewest errors across the identified error types. The top-performing model, LoRA-7B-512, attains an overall weighted average performance of 0.724, falling short of the best baseline model in the SPICE paper, BertSP_G, which scores 0.815 (Perez-Beltrachini et al., 2023), although it is worth noting that BertSP_G was trained with five times more examples. Also, it should be reiterated that we used a subset of the test data, as detailed in Section 3, so direct comparisons have to be made with caution.

The experimental results highlight the effectiveness of few-shot prompting in reducing errors and increasing performance metrics in all models except for LoRA. Errors related to off-prompt and wrongly formatted output saw the most significant improve-

ments. LoRA was working best in the zero-shot setting, as pointed out before. We assume that a model like LoRA, which has previously been fine-tuned with in-context examples, might not benefit from them further; in fact, it could be negatively impacted by biasing the generation process. Apart from few-shot prompting, employing rule-based approaches could further minimize prediction errors. These strategies might involve syntax checking, utilizing entity dictionaries, checking for unwanted language filters, and removing natural language output that is not SPARQL.

Finally, it is important to acknowledge that our study has certain limitations. We have concentrated on semantic parsing of queries from dialogues, although we recognize the importance of exploring other tasks, such as extracting triples or constructing subgraphs in different graph languages. We also suggest further extending our foundational evaluation by additional human assessments and including a wider array of recently published models, especially those trained on program code or structured data documents. Moreover, the SPICE dataset is limited to English. Since pre-training corpora of LLMs primarily consist of English text data, they likely work better where entities and relations correspond to meaningful English words. Consequently, it is to be expected that LLMs exhibit worse performance on benchmarks with more morphologically rich languages.

5 CONCLUSION

We compared LLMs in conversational semantic parsing. Our findings indicate that even smaller, fine-tuned 7B-LLMs exhibit reasonable performance in generating SPARQL queries from dialogues, although they might not always be syntactically valid or yield the correct result. We also discussed model-specific differences and common errors that can be mitigated through few-shot prompting and fine-tuning. In future work, we intend to delve into the applicability of our findings to different query languages. Further, we plan to conduct user evaluations of deployed LLM-based CQA systems for practical search scenarios.³

REFERENCES

Aliannejadi, M., Azzopardi, L., Zamani, H., Kanoulas, E., Thomas, P., and Craswell, N. (2021). Analysing mixed initiatives and search strategies during conver-

sational search. In *Proc. of the 30th CIKM*, page 16–26, New York, NY, USA. ACM.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90% chatgpt quality. *LMSYS Org Blog*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of the 2019 NAACL: Human Language Technologies*, pages 4171–4186, Minneapolis, Minnesota. ACL.

Gu, Y., Kase, S., Vanni, M., Sadler, B., Liang, P., Yan, X., and Su, Y. (2021). Beyond i.i.d.: Three levels of generalization for question answering on knowledge bases. In *Proc. of the Web Conference 2021*, WWW '21, page 3477–3488, New York, NY, USA. ACM.

Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Kacupaj, E., Plepi, J., Singh, K., Thakkar, H., Lehmann, J., and Maleshkova, M. (2021). Conversational question answering over knowledge graphs with transformer and graph attention networks. In *Proc. of the 16th EACL*, pages 850–862, Online. ACL.

Li, Z., Qu, L., and Haffari, G. (2020). Context dependent semantic parsing: A survey. In *Proc. of the 28th International Conference on Computational Linguistics*, pages 2509–2521, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

OpenAI (2022). Chatgpt: Optimizing language models for dialogue. *OpenAI*.

Perez-Beltrachini, L., Jain, P., Monti, E., and Lapata, M. (2023). Semantic parsing for conversational question answering over knowledge graphs. In *Proc. of the 17th EACL*, pages 2507–2522, Dubrovnik, Croatia. ACL.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI*.

Saha, A., Pahuja, V., Khapra, M. M., Sankaranarayanan, K., and Chandar, S. (2018). Complex sequential question answering: Towards learning to converse over linked question answer pairs with a knowledge graph. In *Proc. of the Thirty-Second AAAI Conference on Artificial Intelligence*. AAAI Press.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv:2302.13971*.

Wang, B., Shin, R., Liu, X., Polozov, O., and Richardson, M. (2020). RAT-SQL: Relation-aware schema encoding and linking for text-to-SQL parsers. In *Proc. of the 58th Ann. Meeting of the ACL*, pages 7567–7578. ACL.

³This work has been supported by the German Federal Ministry of Education and Research grant 01IS17049.