

Bridging Information Gaps in Dialogues With Grounded Exchanges Using Knowledge Graphs

Phillip Schneider¹, Nektarios Machner¹, Kristiina Jokinen², and Florian Matthes¹

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

²National Institute of Advanced Industrial Science and Technology, AI Research Center, Japan
{phillip.schneider, nektarios.machner, matthes}@tum.de
kristiina.jokinen@aist.go.jp

Abstract

Knowledge models are fundamental to dialogue systems for enabling conversational interactions, which require handling domain-specific knowledge. Ensuring effective communication in information-providing conversations entails aligning user understanding with the knowledge available to the system. However, dialogue systems often face challenges arising from semantic inconsistencies in how information is expressed in natural language compared to how it is represented within the system’s internal knowledge. To address this problem, we study the potential of large language models for conversational grounding, a mechanism to bridge information gaps by establishing shared knowledge between dialogue participants. Our approach involves annotating human conversations across five knowledge domains to create a new dialogue corpus called *BridgeKG*. Through a series of experiments on this dataset, we empirically evaluate the capabilities of large language models in classifying grounding acts and identifying grounded information items within a knowledge graph structure. Our findings offer insights into how these models use in-context learning for conversational grounding tasks and common prediction errors, which we illustrate with examples from challenging dialogues. We discuss how the models handle knowledge graphs as a semantic layer between unstructured dialogue utterances and structured information items.

1 Introduction

Conversational grounding is an integral aspect of dialogues where interlocutors share information and build up a common understanding. This mutually established knowledge serves as context for subsequent interactions. For building effective dialogue systems, the natural language processing (NLP) community has long focused on conversational grounding, which involves inferential reasoning, dynamic feedback, and repair strategies (Udagawa

and Aizawa, 2021). Despite extensive research, challenges remain in adapting to different conversation domains, addressing semantic vocabulary mismatches, overcoming information gaps between user knowledge and the system’s internal knowledge model, as well as the lack of appropriate training data (Lemon, 2022). Owing to rapid technical advances regarding large language models (LLMs), novel opportunities arise to comprehend contextual intricacies within dialogues and reconcile information expressed in natural language with that stored in machine-readable data structures.

Recognizing the limited research on LLM-based conversational grounding, we investigated the capabilities of LLMs on knowledge grounding tasks. This involved annotating an existing corpus containing dialogues about different domain-specific tabular datasets. In addition to labeling grounding acts, we annotated grounded knowledge items in a knowledge graph structure, a powerful representation of complex relationships between entities and their attributes. Knowledge graphs have proven valuable in various NLP tasks, such as disambiguating ambiguous utterances by providing contextual information (Hogan et al., 2021; Schneider et al., 2022). For example, in dialogue systems, knowledge graphs can help identify the correct meaning of a word with multiple senses or resolve references to specific entities, enhancing the overall understanding and coherence of conversations. We opted for the JSON-LD format due to its simplicity and acceptance as a web standard, allowing interoperability by reusing existing namespaces with shared vocabularies to model knowledge from different sources and domains.

While JSON-LD primarily uses a tree-like structure, it can represent more complex graph structures by linking nodes using identifiers like *@id* and *@type*. As a serialization format for Resource Description Framework (RDF) data, JSON-LD can be transformed into other formats, such as

N-Triples, RDF/XML, or Turtle. This flexibility allows JSON-LD to be integrated with graph databases and other RDF tools, enhancing its utility in various applications. Table 1 shows an example annotation of grounded knowledge in JSON-LD format from a conversation about nature parks.

Our contributions include (1) creating a novel dialogue corpus called *BridgeKG* with over 250 conversational grounding annotations across five knowledge domains, (2) conducting a range of zero- and few-shot experiments by evaluating four LLMs on two grounding tasks, and (3) summarizing common prediction errors and prompting techniques for improving model performance. To ensure the reproducibility of our experiments, we provide the *BridgeKG* dataset, source code, and evaluation outputs in a public GitHub repository.¹

2 Related Work

In regard to the literature on grounding in NLP, it is essential to first define the broadly used term. Grounding can be categorized into three main types. Conversational grounding ensures a common understanding of shared knowledge within a conversation (Traum, 1994). Perceptual grounding links language to sensory experiences of the real world like visual information (Cangelosi, 2010). Knowledge grounding incorporates external information sources to support NLP systems, such as providing factual knowledge to generative language models (Lewis et al., 2020).

Our study focuses solely on conversational grounding by employing LLMs, a topic addressed in only a few recent studies. One related work by Shaikh et al. (2024) examines whether LLM generations contain grounding acts, simulating turn-taking from various conversation datasets. They found that LLMs generate language with less conversational grounding than humans, often producing text that appears to assume common ground. Both their study and ours focus on the three grounding acts: explicit grounding, implicit grounding, and clarification, as proposed by Clark and Schaefer (1989). Two other closely related studies, conducted by Jokinen et al. (2024) and Mohapatra et al. (2024), involve annotating dialogue corpora and employing language models to classify grounding acts and extract grounded knowledge items. While the former conducts preliminary experiments on two conversations with GPT-3.5-Turbo, the lat-

ter presents two annotated dialogue corpora with grounding acts, grounding units, a measure of their degree of grounding, and a baseline evaluation with the open-source T5 model (Raffel et al., 2020).

Unlike the mentioned related work, we are the first to conduct a series of LLM experiments aimed at knowledge identification in information-seeking conversations utilizing an in-context knowledge graph structure for identifying referenced and grounded knowledge items in dialogues.

3 Method

Dataset Annotation The source dialogue corpus we reuse was collected in a study on exploratory information-seeking conversations from Schneider et al. (2023). It comprises 26 conversations about tabular datasets on real-world knowledge spanning the domains of geography, history, media, nutrition, and sports. Every conversation involved a pair where one person was the information seeker and the other was the information provider, using a text-based chatroom for communication. The information seekers were instructed to discover and gather new information about their partner’s previously unknown dataset. Two researchers annotated each written dialogue with labels for grounding acts (explicit, implicit, and clarification). Explicit grounding involves a response that clearly confirms understanding or acceptance of received information (e.g., “okay, thanks”), whereas implicit grounding moves the conversation forward without explicitly acknowledging or questioning the recently shared information (implicit acceptance). Clarification occurs when a conversation partner seeks more information about thus far presented knowledge, which does not result in grounded knowledge since mutual acceptance has not yet been reached.

Example Annotation of Grounded Knowledge

```

[{"@context": ["http://www.w3.org/ns/csvw", {"schema": "http://schema.org"}], "@id": "http://example.org/nature-parks", "url": "nature-parks.csv", "schema:description": "The table contains information about nature parks in Germany", "tableSchema": {"columns": [{"name": "name", "datatype": "string"}, {"name": "state", "datatype": "string"}, {"name": "year", "datatype": "integer"}, {"name": "area_in_km2", "datatype": "integer"}, {"name": "summary", "datatype": "string"}], "primaryKey": "name"}, {"@type": "schema:Place", "name": "Barnim", "state": "Brandenburg Berlin", "year": 1999, "area_in_km2": 749, "summary": "The park includes the Barnim heath habitats dating back to the ice age. It lies between the glacial valleys of Eberswalde in the north and Berlin in the south, and is more than half forested. The region is shaped by many individual lakes and meltwater gullies."}]

```

Table 1: Example JSON-LD annotation of grounded knowledge from the *BridgeKG* dataset, representing the system’s knowledge concerning a dialogue about nature parks. Properties are displayed in blue color.

¹github.com/philotron/Bridge-KG

Model	Zero-Shot Prompt				Few-Shot Prompt			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
GPT-3.5-Turbo (n=1)	0.64	0.50	0.46	0.43	0.55	0.50	0.51	0.50
GPT-3.5-Turbo (n=3)	0.66	0.81	0.50	0.50	0.69	0.59	0.54	0.54
GPT-3.5-Turbo (n=all)	0.59	0.39	0.44	0.41	0.57	0.51	0.45	0.45
GPT-4o (n=1)	0.39	0.55	0.54	0.42	0.64	0.66	0.64	0.61
GPT-4o (n=3)	0.59	0.66	0.67	0.59	0.73	0.74	0.69	0.70
GPT-4o (n=all)	0.64	0.68	0.66	0.62	0.71	0.73	0.67	0.67
Llama-3-8B (n=1)	0.61	0.54	0.53	0.54	0.59	0.65	0.69	0.59
Llama-3-8B (n=3)	0.65	0.60	0.60	0.60	0.57	0.60	0.61	0.55
Llama-3-8B (n=all)	0.44	0.55	0.39	0.38	0.55	0.54	0.51	0.51
Llama-3-70B (n=1)	0.41	0.54	0.56	0.43	0.51	0.61	0.63	0.53
Llama-3-70B (n=3)	0.59	0.66	0.67	0.59	0.65	0.68	0.69	0.64
Llama-3-70B (n=all)	0.71	0.66	0.64	0.64	0.76	0.70	0.70	0.70

Table 2: Zero-shot and few-shot performance metrics for grounding act classification evaluated by macro-averaged accuracy, precision, recall, and F1-score. The variable n denotes the number of preceding input utterances. Bold values highlight the best value for each metric.

For explicit and implicit labels, the grounded knowledge items that have been shared until this point in the dialogue were annotated as a knowledge graph structure in JSON-LD format (Sporny et al., 2020). Annotation disagreements were collaboratively resolved to reach a consensus. Knowledge is incorporated into the grounding annotation only if it is a subset of the underlying tabular dataset and can be represented within the modeled internal system knowledge, which we defined using vocabulary from the namespaces *Schema.org* and *CSVW* (W3C, 2017, 2024). An example conversation illustrating labeled grounding acts and grounded knowledge items for individual dialogue utterances is provided in Table 4 in Appendix A.

Experimental Setup Based on the annotated dataset with conversational grounding labels, we conducted several experiments using four state-of-the-art LLMs: the open-source Llama-3-8B-Instruct as well as Llama-3-70B-Instruct (Meta AI, 2024) from the Llama 3 model family, and the closed-source models GPT-3.5-Turbo (version: 0125) and GPT-4o (version: 2024-05-13) (OpenAI, 2022, 2024). We defined two model prompts: one for classifying grounding acts and another for identifying grounded knowledge. For the knowledge identification prompt, which tasked the LLM to predict the grounded knowledge subset in the conversation thus far, we provided both the input dialogue and the complete system knowledge (i.e., the annotated grounded knowledge for the entire conversation). All models were prompted using a chat completion format, which included a system instruction and, in the few-shot setting, three in-context examples presented as user and assistant turns. Both model prompts are provided in

the Appendix in full length (Tables 5 and 6). To promote deterministic generation, we set the generation seed to 1 and the temperature parameter to 0. The maximum token limit was set to 128 for classification and 4096 for grounded knowledge identification. All generated outputs with extra text were preprocessed using a regular expression to match and extract the first occurrence of either the grounding act or JSON-LD array.

4 Results and Discussion

Classification of Grounding Acts Table 2 shows the performance for classifying grounding acts, using macro-averages to ensure equal class importance. Nearly all tested LLMs benefited from the added context of few-shot examples, with F1-scores generally improving; however, this improvement diminishes as the number of input dialogue turns (n) increases, suggesting potential redundancy when in-context examples are already provided. The results indicate that n=3 often optimizes

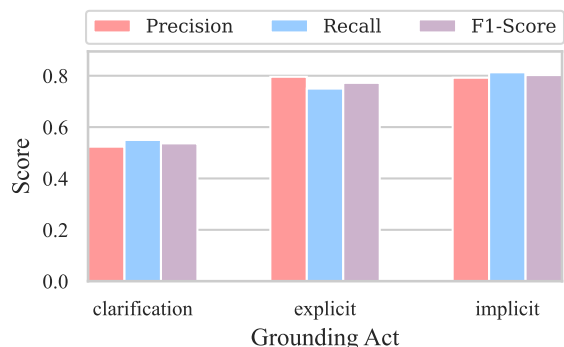


Figure 1: Performance comparison of precision, recall, and F1-score by grounding act for the Llama-3-70B model with all input utterances (n=all).

Issue Type	GPT-3.5-Turbo	GPT-4o	Llama-3-8B	Llama-3-70B
	Relative Frequency: Zero-Shot / Few-Shot			
Invalid JSON-LD	0.00 / 0.01	0.00 / 0.00	0.02 / 0.09	0.20 / 0.00
Property Hallucination	0.01 / 0.00	0.00 / 0.02	0.08 / 0.22	0.38 / 0.26
Value Hallucination	0.02 / 0.00	0.01 / 0.03	0.22 / 0.05	0.46 / 0.07
Property Excess	0.49 / 0.48	0.29 / 0.24	0.50 / 0.38	0.61 / 0.51
Property Deficit	0.37 / 0.22	0.31 / 0.09	0.50 / 0.36	0.39 / 0.20
Value Excess	0.68 / 0.63	0.40 / 0.31	0.66 / 0.32	0.76 / 0.47
Value Deficit	0.22 / 0.22	0.29 / 0.28	0.34 / 0.62	0.24 / 0.34

Table 3: Relative frequency of issues in zero- and few-shot predictions for grounded knowledge identification.

performance in both zero- and few-shot settings by balancing context retention, noise reduction, and efficient usage of tokens. While Llama-8B’s performance drops from 0.54 F1-score at n=1 to 0.38 at n=all, larger LLMs like Llama-70B and GPT-4o handle longer input better, probably due to a higher parameter count and superior noise handling.

Another significant finding is the competitive performance of open-source LLMs against proprietary ones: Llama-8B surpasses GPT-3.5 in the zero-shot run, and Llama-70B matches GPT-4o in the few-shot run. The breakdown of Llama-70B’s performance by grounding act, illustrated in Figure 1, reveals clarification as the most challenging act to classify, consistent with our observation of the other LLMs. For instance, the models often struggled when users tried to clarify a previously introduced concept. Instead of recognizing the clarification (e.g., “And category describes whether it is a movie, tv show, or work of literature?”), the models often misinterpreted it as introducing a new topic, falsely assuming that the previous concept is already implicitly grounded. Contrary to clarification acts, the F1-scores for explicit and implicit classification are comparable. Despite achieving the same overall F1-score, GPT-4o tends to overpredict implicit labels in contrast to the more balanced Llama-70B, as revealed by the confusion matrices in Figure 3 in Appendix A. The latter shows that GPT-4o excels at predicting explicit grounding accurately, avoiding false positives altogether, but it tends to overpredict the implicit class, particularly in cases where participants acknowledge information explicitly before asking a new question (e.g., “Ok very interesting! What is the highest level of protein in the chart?”).

Identification of Grounded Knowledge The second series of experiments aimed at identifying grounded knowledge for a suitable dialogue context, which is a significantly more complex task than classifying grounding acts (Wu et al., 2021;

Oh et al., 2023). Knowledge identification required the LLMs to uniquely pinpoint specific knowledge items from a set of possibilities within the system knowledge model, bridging between vague conversation utterances and structured JSON-LD arrays.

Figure 2 depicts the count of JSON-LD generations accurately matching our 127 annotations with valid properties, values, or completely identical content. The open-source models notably struggle more compared to the proprietary LLMs. While both open-source Llama models produce multiple valid outputs for properties and values with few-shot prompting, they fail to generate any valid predictions in the zero-shot setting. Therefore, these model runs are not displayed in the chart. Remarkably, GPT-4o outperforms GPT-3.5 by almost double, even in the zero-shot experiment, surpassing all other models by a great margin. In the few-shot cases, every third prediction from GPT-4o is identical to our annotated groundings, totaling 42 out of 127 instances. In some cases, The GPT-4o model even succeeded in precisely matching the annotated JSON-LD in a given conversation across a number of subsequent turns.

Table 3 provides a detailed analysis of the most common prediction issues and their relative fre-

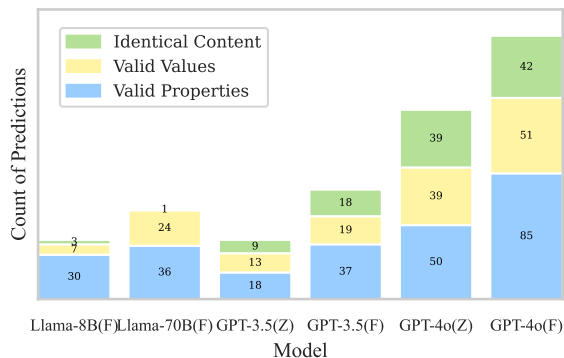


Figure 2: Count of predictions in JSON-LD format with valid properties, valid values, or identical content for evaluated models in zero- (Z) and few-shot (F) settings.

quencies for each model-prompt experiment. Examples for each issue type are listed in Table 7 in Appendix A. Open-source models generally produce more invalid JSON-LD arrays and hallucinate properties and values that are not part of the system knowledge. All tested LLMs tend to overpredict properties and values in zero-shot settings, even though these are grounded later in the conversation. Few-shot prompting can reduce excess properties and values, as well as counteract property deficits. However, in few-shot prompting, open-source models, particularly Llama-3-8B, tend to increase value deficits, becoming too hesitant to identify knowledge. This often results in empty JSON-LD arrays with generated statements such as “The conversation does not mention any specific knowledge items from the system knowledge.”

Our findings corroborate existing benchmarks, highlighting the sophisticated reasoning abilities of state-of-the-art proprietary LLMs such as GPT-4o in highly complex tasks. A similar task complexity-based LLM performance gap is also observable in the direct comparison of the MMLU and HumanEval benchmark scores between GPT-4o and Llama-3 (Hendrycks et al., 2020; Chen et al., 2021; OpenAI, 2024). While Llama-70B performs competitively in the language-focused grounding act classification task, the superiority of GPT-4o becomes apparent in identifying knowledge when handling structured JSON-LD data and fragmented information from dialogue utterances.

In short, when designing dialogue systems augmented with LLMs to handle conversational grounding, smaller open-source models like Llama-3-8B, especially fine-tuned versions, seem to be generally sufficient for basic NLP tasks such as detecting and classifying grounding-related dialogue acts. However, more complex tasks, such as identifying and integrating grounded knowledge from dialogue utterances with structured knowledge representations, require the use of more advanced and larger models like GPT-4o, which possess superior reasoning capabilities and proficiency in processing structured data formats.

5 Conclusion and Future Work

Our study examined LLMs for handling grounding-related knowledge in information-sharing dialogues. We found that classifying grounding acts was feasible for both open- and closed-source LLMs, with open-source LLMs performing on par

compared with leading proprietary ones. However, identifying grounded knowledge proved to be a distinctly more complex task. For the latter, the proprietary LLMs had a competitive edge, and the open-source models underperformed due to their higher predisposition to generate erroneous output. The experiment results from our newly created dataset highlight common prediction issues and demonstrate how few-shot prompting can enhance model outputs, offering valuable insights to advance research on conversational grounding.

Future work should concentrate on developing LLM-based dialogue systems that handle conversational grounding through a multi-component pipeline approach for recognizing grounding-specific dialogue acts as well as grounded knowledge (Jokinen et al., 2024). In previous studies, we have shown that LLMs can augment dialogue systems by performing semantic parsing for conversational question answering over knowledge graphs (Schneider et al., 2024a) and by verbalizing retrieved semantic triples into text responses (Schneider et al., 2024b). We believe conversational grounding is essential as it links the processes of semantic parsing of dialogue utterances, knowledge identification, and response generation, aligning the user’s prior knowledge with the system’s available knowledge base while maintaining the relevance and coherence of conversations.

6 Limitations

Our study has certain limitations that should be acknowledged. First, the experiments are based on a relatively small dataset, consisting of only 26 information-seeking conversations and 669 dialogue turns collected in a controlled laboratory setting. While these conversations span five distinct domains, the findings should be interpreted with caution, as they may not generalize to larger or more diverse dialogue corpora.

Additionally, the grounded knowledge annotations in our study are represented using the JSON-LD syntax. We chose the JSON-LD format because it is widely used, and many LLMs are trained to process JSON sequences effectively. However, it is important to recognize that other encoding formats, such as Turtle, RDF/XML, and N-Triples, may produce different performance results. Further, our experiments were restricted to the open-source Llama (Meta AI, 2024) and closed-source GPT (OpenAI, 2022, 2024) model families. It is

advisable for future work to explore an even bigger variety of LLMs, particularly those that are specifically trained on code and structured data like Codestral or Code Llama.

Lastly, conversational grounding in dialogue systems entails both the classification of grounding acts and the identification of grounded knowledge. While we have introduced and evaluated these tasks separately, incorporating our approach into an end-to-end evaluation could offer a more holistic understanding of end-to-end performance in more realistic dialogue scenarios.

7 Ethical Considerations

In our experiments, we used a publicly available dialogue dataset from [Schneider et al. \(2023\)](#) while ensuring that no personal identifying information of the participants was processed or disclosed. The information-seeking conversations from the dataset discuss only domain-specific knowledge from publicly accessible websites, such as Wikipedia. Moreover, to ensure optimal computing efficiency, evaluations of the Llama and GPT models were conducted on cloud computing platforms, with each inference run taking less than an hour.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful suggestions. Kristiina Jokinen acknowledges the support of Project JPNP20006 commissioned by the New Energy and Industrial Technology Development Organization (NEDO), Japan.

References

Angelo Cangelosi. 2010. Grounding language in action and perception: From cognitive agents to humanoid robots. *Physics of life reviews*, 7(2):139–151.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. [Evaluating large language models trained on code](#). *arXiv preprint arXiv:2107.03374*.

Herbert H Clark and Edward F Schaefer. 1989. Contributing to discourse. *Cognitive science*, 13(2):259–294.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. [Measuring massive multitask language understanding](#). In *International Conference on Learning Representations*.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D’amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. [Knowledge graphs](#). *ACM Comput. Surv.*, 54(4).

Kristiina Jokinen, Phillip Schneider, and Taiga Mori. 2024. [Towards harnessing large language models for comprehension of conversational grounding](#). In *In 14th International Workshop on Spoken Dialogue Systems Technology (IWSDS 2024)*, Sapporo, Japan.

Oliver Lemon. 2022. [Conversational grounding in emergent communication—data and divergence](#). In *Emergent Communication Workshop at ICLR 2022*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Meta AI. 2024. [Introducing Meta Llama 3: The most capable openly available LLM to date](#). *Meta AI Blog*.

Biswesh Mohapatra, Seemab Hassan, Laurent Romary, and Justine Cassell. 2024. [Conversational grounding: Annotation and analysis of grounding acts and grounding units](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3967–3977, Torino, Italia. ELRA and ICCL.

Minsik Oh, Joosung Lee, Jiwei Li, and Guoyin Wang. 2023. [PK-ICR: Persona-knowledge interactive multi-context retrieval for grounded dialogue](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16383–16395, Singapore. Association for Computational Linguistics.

OpenAI. 2022. [ChatGPT: Optimizing language models for dialogue](#). *OpenAI Blog*.

OpenAI. 2024. [Hello GPT-4o](#). *OpenAI Blog*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Phillip Schneider, Anum Afzal, Juraj Vladika, Daniel Braun, and Florian Matthes. 2023. [Investigating conversational search behavior for domain exploration](#). In *European Conference on Information Retrieval*, pages 608–616. Springer.

- Phillip Schneider, Manuel Klettner, Kristiina Jokinen, Elena Simperl, and Florian Matthes. 2024a. [Evaluating large language models in semantic parsing for conversational question answering over knowledge graphs](#). In *International Conference on Agents and Artificial Intelligence*.
- Phillip Schneider, Manuel Klettner, Elena Simperl, and Florian Matthes. 2024b. [A comparative analysis of conversational large language models in knowledge-based text generation](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 358–367, St. Julian’s, Malta. Association for Computational Linguistics.
- Phillip Schneider, Tim Schopf, Juraj Vladika, Mikhail Galkin, Elena Simperl, and Florian Matthes. 2022. [A decade of knowledge graphs in natural language processing: A survey](#). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 601–614, Online only. Association for Computational Linguistics.
- Omar Shaikh, Kristina Gligorić, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. 2024. [Grounding gaps in language model generations](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.
- Manu Sporny, Dave Longley, Gregg Kellogg, Markus Lanthaler, Pierre-Antoine Champin, and Niklas Lindström. 2020. [JSON-LD 1.1. W3C Recommendation](#).
- David Traum. 1994. [A computational theory of grounding in natural language conversation](#). *PhD thesis, Univ. Rochester*.
- Takuma Udagawa and Akiko Aizawa. 2021. [Maintaining common ground in dynamic environments](#). *Transactions of the Association for Computational Linguistics*, 9:995–1011.
- W3C. 2017. [CSVW Namespace Vocabulary Terms. W3C Document](#).
- W3C. 2024. [Schema.org. Schema.org](#).
- Zeqiu Wu, Bo-Ru Lu, Hannaneh Hajishirzi, and Mari Ostendorf. 2021. [DIALKI: Knowledge identification in conversational systems through dialogue-document contextualization](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1852–1863, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Appendix

The Appendix provides one annotated conversation example (Table 4), the model prompts in full length (Tables 5 and 6), an overview of common issue types identified in the predictions (Table 7), and two confusion matrices of the classification results of the two best-performing model inference runs (Figure 3).

Dialogue Utterances	Dialogue Grounded Knowledge Act	
S: What is your dataset about?	-	-
P: it contains information about 11341 historical figures, including their full name, sex, birth year, city, country, continent, occupation, historical popularity index (HPI). The HPI represents the degree of this person’s online popularity	-	-
S: Who is the most popular?	implicit	<pre>[{"@context": ["http://www.w3.org/ns/csvw", {"schema": "http://schema.org"}], "@id": "http://example.org/historical-figures", "url": "historical-figures.csv", "schema:description": "The table contains information about historical figures", "tableSchema": {"columns": [{"name": "full_name", "datatype": "string"}, {"name": "sex", "datatype": "string"}, {"name": "birth_year", "datatype": "integer"}, {"name": "city", "datatype": "string"}, {"name": "country", "datatype": "string"}, {"name": "continent", "datatype": "string"}, {"name": "occupation", "datatype": "string"}, {"name": "historical_popularity_index", "datatype": "float"}], "primaryKey": "full_name"}]}</pre>
P: Aristotle, who is from Greece and has a largest HPI value: 31.9938.	-	-
S: I see, is there Socrate in the dataset?	explicit	<pre>[{"@context": ["http://www.w3.org/ns/csvw", {"schema": "http://schema.org"}], "@id": "http://example.org/historical-figures", "url": "historical-figures.csv", "schema:description": "The table contains information about historical figures", "tableSchema": {"columns": [{"name": "full_name", "datatype": "string"}, {"name": "sex", "datatype": "string"}, {"name": "birth_year", "datatype": "integer"}, {"name": "city", "datatype": "string"}, {"name": "country", "datatype": "string"}, {"name": "continent", "datatype": "string"}, {"name": "occupation", "datatype": "string"}, {"name": "historical_popularity_index", "datatype": "float", "maximum": 31.9938}], "primaryKey": "full_name"}, {"@type": "schema:Person", "full_name": "Aristotle", "country": "Greece", "historical_popularity_index": 31.9938}]</pre>
P: Yes, Socrate is in the dataset.	-	-
S: What is is popularity index?	implicit	<pre>[{"@context": ["http://www.w3.org/ns/csvw", {"schema": "http://schema.org"}], "@id": "http://example.org/historical-figures", "url": "historical-figures.csv", "schema:description": "The table contains information about historical figures", "tableSchema": {"columns": [{"name": "full_name", "datatype": "string"}, {"name": "sex", "datatype": "string"}, {"name": "birth_year", "datatype": "integer"}, {"name": "city", "datatype": "string"}, {"name": "country", "datatype": "string"}, {"name": "continent", "datatype": "string"}, {"name": "occupation", "datatype": "string"}, {"name": "historical_popularity_index", "datatype": "float", "maximum": 31.9938}], "primaryKey": "full_name"}, {"@type": "schema:Person", "full_name": "Aristotle", "country": "Greece", "historical_popularity_index": 31.9938}, {"@type": "schema:Person", "full_name": "Socrates"}]}</pre>
P: Historical popularity index (HPI) is metric that aggregates information on a biography’s on-line popularity. It aggregates information on the age and attention received by biographies in multiple language editions of Wikipedia to provide a summary statistic of their global popularity.	-	-

Table 4: Example of dialogue excerpt from the history domain with annotated grounding dialogue acts and grounded knowledge in JSON-LD format. Seeker (S) and provider (P) roles are abbreviated for each turn. Utterances are taken from the dialogue logs and may contain spelling errors. Newly grounded knowledge is displayed in blue color.

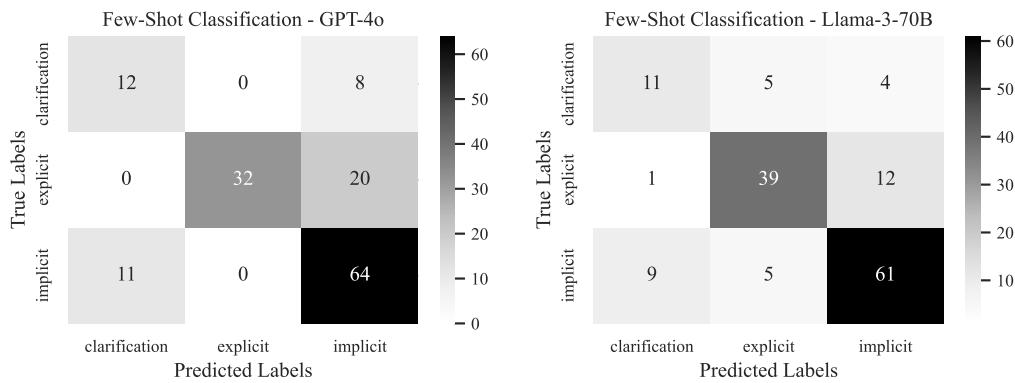


Figure 3: Confusion matrices for few-shot classification results of GPT-4o with three input utterances and Llama-3-70B with all input utterances.

Grounding Act Classification Prompt

Zero-Shot

SYSTEM: Predict the grounding label for the last response in the 'Input Dialogue:'. The label indicates whether the knowledge in the dialogue was accepted. Choose one of the following labels:

explicit: The response confirms understanding or acceptance (e.g., 'okay', 'thanks', 'alright', 'nice') without seeking clarification.

clarification: The response seeks clarification about a previous dialogue snippet.

implicit: The response moves the conversation forward without explicitly confirming or seeking clarification.

Few-Shot

SYSTEM: Predict the grounding label for the last response in the 'Input Dialogue:'. The label indicates whether the knowledge in the dialogue was accepted. Choose one of the following labels:

explicit: The response confirms understanding or acceptance (e.g., 'okay', 'thanks', 'alright', 'nice') without seeking clarification.

clarification: The response seeks clarification about a previous dialogue snippet.

implicit: The response moves the conversation forward without explicitly confirming or seeking clarification.

USER: Input Dialogue:

seeker: Can you give me some information about your dataset?

provider: My dataset includes information on buildings of Gothic architecture.

seeker: How tall is the Cologne Cathedral?

ASSISTANT: Output Label: implicit

USER: Input Dialogue:

provider: Monitors have different attributes like size or panel technology.

provider: There are some with an aspect ratio of 21:9.

seeker: What is aspect ratio?

ASSISTANT: Output Label: clarification

USER: Input Dialogue:

provider: An elephant's average lifespan is around 65 years.

seeker: I see, good to know.

ASSISTANT: Output Label: explicit

Table 5: Overview of applied zero-shot and few-shot prompts for classification.

Grounded Knowledge Identification Prompt

Zero-Shot

SYSTEM: Your task is to identify the knowledge items that have been grounded by the conversation partners in the 'Input Dialogue'. The items of mutually grounded knowledge must be explicitly mentioned in the dialogue. Based on the complete set of 'System Knowledge', your task is to generate the subset of knowledge items that have been grounded so far. Ensure that the output is a valid JSON-LD structure (an array of JSON objects) and only include knowledge items from the formatted 'System Knowledge'.

Few-Shot

SYSTEM: Your task is to identify the knowledge items that have been grounded by the conversation partners in the 'Input Dialogue'. The items of mutually grounded knowledge must be explicitly mentioned in the dialogue. Based on the complete set of 'System Knowledge', your task is to generate the subset of knowledge items that have been grounded so far. Ensure that the output is a valid JSON-LD structure (an array of JSON objects) and only include knowledge items from the formatted 'System Knowledge'.

USER: System Knowledge: [{"@context": ["http://www.w3.org/ns/csvw", {"schema": "http://schema.org"}], "@id": "http://example.org/american-presidents", "url": "american-presidents.csv", "schema:description": "The table contains information about American presidents", "tableSchema": {"columns": [{"name": "name", "datatype": "string"}, {"name": "term", "datatype": "string"}, {"name": "party", "datatype": "string"}, {"name": "election_year", "datatype": "integer"}]}, "primaryKey": "name"}, {"@type": "schema:Person", "name": "Barack Obama", "party": "Democratic"}]
Input Dialogue:

seeker: Can you give me an example entry from your dataset?

provider: One of the presidents in the list is Barack Obama.

seeker: Thanks. What party does he belong to?

ASSISTANT: Output JSON-LD: [{"@context": ["http://www.w3.org/ns/csvw", {"schema": "http://schema.org"}], "@id": "http://example.org/american-presidents", "url": "american-presidents.csv", "schema:description": "The table contains information about American presidents", "tableSchema": {"columns": [{"name": "name", "datatype": "string"}]}, "primaryKey": "name"}, {"@type": "schema:Person", "name": "Barack Obama"}]

USER: System Knowledge: [{"@context": ["http://www.w3.org/ns/csvw", {"schema": "http://schema.org"}], "@id": "http://example.org/greek-islands", "url": "greek-islands.csv", "schema:description": "The table contains information about islands in Greece", "tableSchema": {"columns": [{"name": "island", "datatype": "string"}, {"name": "area_in_km2", "datatype": "integer", "minimum": 64, "maximum": 8336}, {"name": "cluster", "datatype": "string"}]}, "primaryKey": "island"}, {"@type": "schema:Place", "island": "Crete", "area_in_km2": 8336, "cluster": "Cretan"}, {"@type": "schema:Place", "island": "Alonnisos", "area_in_km2": 64, "cluster": "Sporades"}, {"@type": "schema:Place", "island": "Lesbos", "area_in_km2": 1633, "cluster": "North Aegean Islands"}]
Input Dialogue:

provider: My dataset contains information on Greek islands. For example, there is Crete with an area of 8336 square kilometers.

provider: That makes it the largest island in Greece.

seeker: Which one is the smallest and what is its area?

ASSISTANT: Output JSON-LD: [{"@context": ["http://www.w3.org/ns/csvw", {"schema": "http://schema.org"}], "@id": "http://example.org/greek-islands", "url": "greek-islands.csv", "schema:description": "The table contains information about islands in Greece", "tableSchema": {"columns": [{"name": "island", "datatype": "string"}, {"name": "area_in_km2", "datatype": "integer", "maximum": 8336}], "primaryKey": "island"}, {"@type": "schema:Place", "island": "Crete", "area_in_km2": 8336}]

USER: System Knowledge: [{"@context": ["http://www.w3.org/ns/csvw", {"schema": "http://schema.org"}], "@id": "http://example.org/android-smartphones", "url": "android-smartphones.csv", "schema:description": "The table contains information about Android smartphones", "tableSchema": {"columns": [{"name": "model", "datatype": "string"}, {"name": "developer", "datatype": "string"}, {"name": "release_year", "datatype": "integer", "minimum": 2008, "maximum": 2024}, {"name": "android_version", "datatype": "string"}]}, "primaryKey": "model"}, {"@type": "schema:Product", "model": "HTC Dream", "developer": "HTC", "release_year": "2008"}, {"@type": "schema:Product", "model": "LG Wing", "developer": "LG", "release_year": "2020", "android_version": "Android 10"}, {"@type": "schema:Product", "release_year": "2024"}]
Input Dialogue:

provider: I can provide technical information about Android smartphones.

provider: One column contains data about the model and another specifies its release year.

seeker: I see, good to know.

ASSISTANT: Output JSON-LD: [{"@context": ["http://www.w3.org/ns/csvw", {"schema": "http://schema.org"}], "@id": "http://example.org/android-smartphones", "url": "android-smartphones.csv", "schema:description": "The table contains information about Android smartphones", "tableSchema": {"columns": [{"name": "model", "datatype": "string"}, {"name": "release_year", "datatype": "integer"}]}, "primaryKey": "model"}]

Table 6: Overview of applied zero-shot and few-shot prompts for grounding knowledge.

Issue Type	Definition	Example
Invalid JSON-LD	Prediction is overall not a valid JSON-LD structure.	<p>PRED: [{"@type": "schema:CreativeWork", "category": ["films", "series", "literature"]}]]</p> <p>→ JSON syntax elements are missing.</p>
Property Hallucination	Predicted property is not part of system knowledge.	<p>PRED: [{"@type": "schema:Place", "name": "English Garden", "state": "Bavaria", "city": "Munich", "summary": "The English Garden is a large public park in Munich, Bavaria. It is one of the world's largest urban public parks, larger than New York's Central Park. The park is known for its beer gardens, streams, lakes, and Japanese teahouse."}]</p> <p>→ English Garden was mentioned in the dialogue but is not part of the system knowledge. The "city" property does not exist.</p>
Value Hallucination	Predicted value is not part of system knowledge.	<p>PRED: [{"@context": ["http://www.w3.org/ns/csvw", "http://schema.org"], "@id": "http://example.org/time-travel-media", "url": "time-travel-media.csv", "schema:description": "The table contains information about time travel media", "tableSchema": {"columns": [{"name": "year", "datatype": "string"}, {"name": "title", "datatype": "string"}, {"name": "author", "datatype": "string"}, {"name": "short text description", "datatype": "string"}, {"name": "category", "datatype": "string"}]}, "primaryKey": ["year", "title", "category"]}]]</p> <p>→ System knowledge contains the columns "creators" and "description" but "author" and "short text description" are hallucinated.</p>
Property Excess	Predicted property is part of system knowledge but grounded too early.	<p>PRED: [{"@context": ["http://www.w3.org/ns/csvw", "http://schema.org"], "@id": "http://example.org/football-records", "url": "football-records.csv", "schema:description": "The table contains information about football records", "tableSchema": {"columns": [{"name": "record", "datatype": "string"}]}, "primaryKey": "record"}]]</p> <p>→ Properties should have been grounded at a later stage.</p>
Property Deficit	Property from system knowledge was not identified.	<p>PRED: [{"@context": ["http://www.w3.org/ns/csvw", "http://schema.org"], "@id": "http://example.org/nature-parks", "url": "nature-parks.csv", "schema:description": "The table contains information about nature parks in Germany", "tableSchema": {"columns": [{"name": "name", "datatype": "string"}, {"name": "state", "datatype": "string"}, {"name": "year", "datatype": "integer"}, {"name": "area_in_km2", "datatype": "integer"}, {"name": "summary", "datatype": "string"}]}, "primaryKey": "name"}], [{"@type": "schema:Place", "name": "Altmühl Valley Nature Park"}]]</p> <p>→ Highlighted properties were not grounded.</p>
Value Excess	Predicted value is part of system knowledge but grounded too early.	<p>PRED: [{"@context": ["http://www.w3.org/ns/csvw", "http://schema.org"], "@id": "http://example.org/nature-parks", "url": "nature-parks.csv", "schema:description": "The table contains information about nature parks in Germany", "tableSchema": {"columns": [{"name": "name", "datatype": "string"}, {"name": "state", "datatype": "string"}, {"name": "year", "datatype": "integer"}, {"name": "area_in_km2", "datatype": "integer"}, {"name": "summary", "datatype": "string"}]}, "primaryKey": "name"}]]</p> <p>→ Values should have been grounded at a later stage.</p>
Value Deficit	Value from system knowledge was not identified.	<p>PRED: [{"@context": ["http://www.w3.org/ns/csvw", "http://schema.org"], "@id": "http://example.org/historical-figures", "url": "historical-figures.csv", "schema:description": "The table contains information about historical figures", "tableSchema": {"columns": [{"name": "full_name", "datatype": "string"}, {"name": "birth_year", "datatype": "integer", "minimum": -3500, "maximum": 2005}], "primaryKey": "full_name"}], [{"@type": "schema:Person", "birth_year": -3500}, {"@type": "schema:Person", "birth_year": 2005}]]</p> <p>→ Highlighted values were not grounded.</p>

Table 7: Overview of six identified issue types with examples from generated model predictions (PRED). The manifestation of issues are highlighted in red color.