Are you a Good Assistant? Assessing LLM Trustability in Task-oriented Dialogues

Tiziano Labruna[1,2,*], Sofia Brenna[1,2], Giovanni Bonetta[2] and Bernardo Magnini[2]

[1] *Free University of Bozen-Bolzano, 3 Dominikanerplatz 3 - Piazza Domenicani 3, Bozen-Bolzano, 39100, Italy*
[2] *Fondazione Bruno Kessler, Via Sommarive 18, Povo, Trento, 38123, Italy*

Abstract

Despite the impressive capabilities of recent Large Language Models (LLMs) to generate human-like text, their ability to produce contextually appropriate content for specific communicative situations is still a matter of debate. This issue is particularly crucial when LLMs are employed as assistants to help solve tasks or achieve goals within a given conversational domain. In such scenarios, the assistant is expected to access specific knowledge (e.g., a database of restaurants, a calendar of appointments) that is not directly accessible to the user and must be consistently utilised to accomplish the task. In this paper, we conduct experiments to evaluate the trustworthiness of automatic assistants in task-oriented dialogues. Our findings indicate that state-of-the-art open-source LLMs still face significant challenges in maintaining logical consistency with a knowledge base of facts, highlighting the need for further advancements in this area.

Keywords

task-oriented dialogues, constraint satisfaction, knowledge base coherence, Llama3 8B

1. Introduction

Conversational assistants [1] are widely used to help human users achieve specific goals through dialogue. In a typical scenario (e.g., booking a restaurant, scheduling an appointment, selecting a song in a playlist, etc.), the assistant interprets the user's goals, searches a database for relevant options, and provides the user with responses (e.g., a restaurant reservation, a new appointment in a calendar, a song playing on a smartphone). A key ability for an assistant is to maintain consistency between user requests and domain knowledge [2]. This is crucial because, in a typical setting, the user does not know the actual content of the database (e.g., all the restaurants in a city) and, as a consequence, cannot verify whether the assistant's response is correct.

While in traditional approaches [3], this consistency was ensured by a dedicated component responsible for retrieving information from a domain database, recent end-to-end approaches [4, 5] rely on a single LLM-based model for utterance understanding, domain knowledge retrieval, and response generation. In this setting, the LLM must generate responses that are as aligned with the database as possible. However, the ability of current end-to-end assistants to maintain consistency between the generated responses and the actual content of the domain knowledge is questionable (e.g., due to LLM confabulations), and there is a clear lack of empirical evidence on this crucial issue.

```
R1: N=Grafton Hotel Restaurant A=east F=british
    P=expensive
R2: N=The River Bar Steakhouse and Grill A=centre
    F=modern european P=cheap
R3: N=Two Two A=north F=french P=expensive
```

T3 User: I would like to have british cuisine.

T4 LLama: I've got just the thing for you! I've found a few options that serve British cuisine. [...] **One option is the Grafton Hotel Restaurant** [...]. Another option is the River Bar Steakhouse and Grill, located in the **centre** of the city. They have a mix of British and **modern European** dishes [...].

T5 User: Nothing in the north?

T6 LLama: [...] I found one option that serves British cuisine in the north area of the city. Restaurant Two Two, located in **Chesterton**, has a menu that features a range of British dishes, including classic comfort food and seasonal specials. [...]

Figure 1: An inconsistent dialogue with respect to a Knowledge Base (KB). Red values indicate inconsistencies between the system-generated text and the KB, whereas the green elements in bold indicate correct information.

To be more concrete, Figure 1 shows an example of an inconsistent dialogue with respect to the conversational knowledge base. Here, although there are two Spanish restaurants in the knowledge base, the system (turn S1) informs the user that there are three Spanish restaurants, providing incorrect information. This is an example of

---

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

[*] Corresponding author.

tlabruna@fbk.eu (T. Labruna); sbrenna@fbk.eu (S. Brenna); gbonetta@fbk.eu (G. Bonetta); magnini@fbk.eu (B. Magnini)
0000-0001-7713-7679 (T. Labruna); 0009-0001-3748-1448 (S. Brenna); 0000-0003-4498-1026 (G. Bonetta); 0000-0002-0740-5778 (B. Magnini)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

inconsistency generated by an LLM, which is the focus of this research.

Our aim is to shed new light on the trustworthiness of an LLM playing the role of an assistant in a task-oriented conversational domain while interacting with a user. We aim to answer the following research questions: (i) How can we operationally define the consistency between a task-oriented dialogue and the domain database behind the dialogue? (ii) How can we quantify the degree of trustworthiness of an assistant-LLM? (iii) Can we collect empirical evidence on a sufficiently large amount of task-oriented dialogues?

To address these research questions, we set up an experimental framework allowing large-scale analysis, where task-oriented dialogues are first automatically generated by two instances of a state-of-the-art LLM, LLama-3 8B [6], and then a more powerful LLM, GPT-4o [7], is used to detect potential inconsistencies between a dialogue and a corresponding domain knowledge base. We hope that new large-scale experimental data can be used to develop more reliable and effective task-oriented dialogue systems, ultimately enhancing the capabilities of conversational agents in various applications.

2. Methodology and Experimental Setting

Our experimental setting consists of two phases. In the preliminary phase, referred to as the Human-Llama Interaction phase (cfr. Section 3), we test the capabilities of an open-source LLM (i.e. LLama-3) to generate adequate task-oriented dialogues through interactive conversations with humans.

In the second phase, referred to as the Llama-Llama Interaction phase (cfr. Section 4), we automate both the generation and evaluation of task-oriented dialogues, creating a Llama-Llama generated MultiWOZ dialogue corpus, The Dining Llamas of Oz[1]. Following in this section, the description of the MultiWOZ dataset and the metrics used to check and quantify the reliability of the generated dialogs in both phases.

2.1. The MultiWOZ 2.3 Dataset

Since the primary focus of this work is about task-oriented dialogues, we used the MultiWOZ (Multi-Domain Wizard-Of-Oz) dataset [8], one of the most prominent datasets in this area. MultiWOZ has been extensively employed to develop and test models for natural language understanding, dialogue management, and natural language generation.

MultiWOZ is a widely known task-oriented dialogue dataset collected via the Wizard of Oz approach. The dataset comprises over 10,000 dialogues between a customer and the *Cambridge InfoTown* assistant, designed to help customers navigate Cambridge's amenities. The conversations span over seven different domain concepts, including train ticket reservations, tourist attraction searches, and restaurant reservations. For our experiments, we selected data related to the restaurant domain (version 2.3 [9]).

The MultiWOZ dialogues were collected with a system that provides information to the user relying on a specific database, known as the Knowledge Base (KB), describing properties of the Cambridge domain. Each domain concept has its own KB; for our experiments, we consider only the restaurant KB. The restaurant KB holds information about 110 different instances (i.e., restaurants), where each instance comprises a series of properties (e.g., NAME, FOOD, AREA) and corresponding values (e.g., THE OLD CAMBRIDGE, BRITISH, NORTH).

All system turns in the dialogues are expected to consistently rely on the information contained in the KB to provide accurate information to the user.

2.2. Consistency Metrics

To assess the consistency of a generated turn against its Knowledge Base, we analysed each system-generated conversational turn referring to any piece of information provided in the KB. Each turn was assessed based on two separate binary metrics:

- KB-ALIGNMENT: Assesses whether the system turn is consistent with the KB, meaning that does not contradict any information provided in the KB.
- KB-GROUNDING: Assesses whether the system turn refrains from hallucinating and introducing information not present in the KB, ensuring all mentioned details are grounded in the existing KB.

For instance, the assessments for the system turns in Figure 1 would be as follows: T4 (KB-Alignment = 0, KB-Grounding = 1), T6 (KB-Alignment = 0, KB-Grounding = 0). In addition to this, we used two evaluation metrics to assess the overall quality of each turn and provide a global evaluation of the whole corpus:

- CORRECT TURNS: Indicates the percentage of turns that have both KB-Alignment and KB-Grounding annotated as 1.
- CORRECT DIALOGUES: Indicates the percentage of dialogues that have all turns with both KB-Alignment and KB-Grounding annotated as 1.

---

[1] The generated dataset is publicly available at: https://github.com/tLabruna/The-Dining-Llamas-of-Oz

These metrics offer a comprehensive understanding of the dialogue system's ability to maintain consistency and accuracy throughout the conversation.

3. Human-Llama Interaction Phase

In this phase, we simulated the dialogue collection approach of the MultiWOZ dataset through the human-Llama interactive generation of novel dialogues. Although this phase required substantial human effort, it was crucial for obtaining an initial high-quality set of dialogues.

We aimed to generate dialogues where a human interacts with a system played by Llama-3 8B in two languages: English and Italian. The model was prompted to play the role of the *Cambridge InfoTown* system. The system's goal was to guide the user towards reserving a restaurant in Cambridge. For each dialogue, we utilised 10 restaurant instances taken from the MultiWOZ KB. We selected 6 distinct sets of instances, which had the following characteristics:

1. All with the same Food;
2. All with different Food (or as different as possible);
3. All with the same Price;
4. All with different Price (or as different as possible);
5. All with the same Area;
6. All with different Area (or as different as possible).

We chose the slots Food, Price, and Area to differentiate the sets since they are the *informable* slots within the Restaurant concept.

The human users were instructed to follow a scenario that involved reserving a restaurant, providing a realistic context for the dialogues. Five distinct instructions were employed for the interactive generation of a human-LLM dialogue, each paired with the 6 sets of KB instances, resulting in a total of 30 dialogue scenarios. The process was repeated in both English and Italian, leading to the creation of 30 dialogues in each language, for a total of 60 dialogues.

3.1. Manual Evaluation

The manual evaluations were conducted by three annotators who assessed the dialogues based on the binary metrics KB-Alignment and KB-Grounding. Each of the 60 dialogues was annotated by at least two different annotators to ensure reliability. The inter-annotator agreement between human evaluators was measured using Cohen's Kappa (κ) to provide a measure of the inter-rater reliability (IRR) level. As per Table 1, we obtained an average κ in both metrics and languages that indicates *substantial agreement* on Landis and Koch's agreement scale [10].

Table 1
Cohen's κ values for inter-annotator agreement on human-LLama generated dialogues.

Annotators	Metric	ITA	ENG
human-human	KB-Alignment	0.71	0.65
human-human	KB-Grounding	0.79	0.59
human-GPT-4o	KB-Alignment	0.60	0.58
human-GPT-4o	KB-Grounding	0.58	0.39

3.2. Automated Evaluation

We instructed GPT-4o[2] to perform the same evaluations as the human annotators. This consisted in feeding the model with a given KB/dialogue pair, asking it to output two lists of turn assessments: one for the KB-Grounding and another for the KB-Alignment. Then we computed the agreement between GPT-4o's evaluations and the human evaluations. The precise prompt used to instruct GPT-4o can be found in Appendix B. Although the agreement with GPT-4o (see Table 1) was slightly lower than the *substantial* agreement observed between human annotators, it was still classified as *moderate* on Landis and Koch's agreement scale [10]. Due to these results we assumed GPT-4o to be a valuable automatic judge and deployed it the same way for the LLama-LLama evaluation phase (cfr. Section 4).

4. The Dining Llamas of Oz

After recognising the ability of Llama-3 to generate dialogues and the evaluation skills of GPT-4o (cfr. Section 3.2), we conducted further experiments by generating 1,311 dialogues using Llama-3 8B and following the MultiWOZ dataset. For each dialogue of the original dataset, we utilised the instructions provided to the human user in the Wizard-of-Oz setting to guide a Llama acting as the user, interacting with a Llama acting as the system. During the dialogue generation phase, we randomly selected 70 instances from the entire Knowledge Base for each simulated dialogue, ensuring that each dialogue was staged in a varied KB scenario. This approach, a.k.a LLama-Llama phase, allowed us to create a large set of automatically generated dialogues, each based on a different subset of the KB. We call this generated dataset "The Dining Llamas of Oz," which comprises 1,049 training instances, with 131 instances each for the validation and test sets.

---

[2] GPT-4o was used via the Microsoft Azure APIs. The API version was 2024-02-01. The cost for the API interactions was about $400.

Table 2 presents statistics for the dataset, including the average number of turns per dialogue, the average length in number of tokens for user and system turns, and the Standardized Type-Token Ratio (STTR) [11] for user and system turns. The STTR is calculated by merging all turns, segmenting them into chunks (we used a segmentation size of 1000), and computing the average TTR for all chunks.

Table 2
Statistics of the Llama-Llama dialogues dataset.

Statistic	Value
Number of Dialogues	1311
Average Dialogue Length	6.21
Average User Turns Length	25.69
Average System Turns Length	124.52
User Turns STTR	0.29
System Turns STTR	0.41

4.1. Turn-by-Turn Evaluation

To assess the quality of the Dining Llamas of Oz dataset, we employed GPT-4o, as in our previous experiments. Using the same approach as in Section 3.2, we obtained a KB-Alignment score of 49.73% and a KB-Grounding score of 38.59% for the entire dataset. To verify the annotation quality of these new dialogues, we manually annotated 30 dialogues from the evaluation split and compared these annotations with GPT-4o's evaluations on the same dialogues. This initial comparison resulted in a not ideal κ of 0.15 for KB-Alignment and 0.06 for KB-Grounding (*slight agreement*). To enhance these performance metrics and establish a reliable evaluation pipeline, we revised our approach: instead of passing the entire dialogue to GPT-4o, we evaluated one turn at a time. The detailed methodology was as follows:

1. Provide GPT-4o with a user utterance and the corresponding system response, and prompt it to determine if the system's response references the KB.
2. If GPT-4o indicates a reference to the KB:
    a) Prompt GPT-4o with the same user-system turn and the KB to determine if the system's turn shows KB-Alignment.
    b) Prompt GPT-4o with the same user-system turn and the KB to determine if the system's turn shows KB-Grounding.

The full prompt is available at Appendix B. This method allows for a more precise scoring of each turn, though it increases OpenAI API usage and associated costs. We discovered that this *turn-by-turn evaluation* approach significantly improved the agreement: we obtained a κ of 0.68 for KB-Alignment and 0.49 for KB-Grounding (*moderate/substantial agreement*). Consequently, we decided to use this technique for automated evaluation.

Using this approach, we assessed 262 dialogues (from the evaluation and test splits) using GPT-4o. This provided a broader understanding of the KB consistency of Llama-generated dialogues across a larger dataset. The KB consistency evaluation is summarised in Table 3. The turns were filtered by removing those that were judged to have no reference to the KB. In addition to evaluating the metrics for all 262 dialogues, we further analysed the dataset by dividing it based on two criteria: the success of the dialogues and the dialogue length. For the success criterion, we distinguished between dialogues with a user instruction that, in the original MultiWOZ dataset, led to a successful restaurant booking (successful dialogues) and those that did not lead to any restaurant reservation (unsuccessful dialogues). For the dialogue length criterion, we distinguished between dialogues that had three or fewer turns (a maximum of three user utterances and three system utterances) and those that had four or more turns.

5. Discussion

Our investigation into the performance of state-of-the-art Large Language Models (LLMs) like Llama-3 in task-oriented dialogue systems reveals several critical insights about their current limitations. The central finding is that while these models exhibit advanced capabilities in generating text, their quality in managing task-oriented dialogues remains unsatisfactory.

Initially, we compared human evaluations with GPT-4o's evaluations to assess its effectiveness in evaluating dialogue quality. This comparison was instrumental in determining that GPT-4o could be useful for dialogue evaluation, but it highlighted that the model's performance degrades significantly when scaled from a smaller to a larger Knowledge Base. The annotation agreement dropped notably as the number of KB instances increased from 10 to 70, indicating that GPT-4o struggles with larger, more complex datasets.

To address this, we shifted our approach to a turn-by-turn evaluation method. After extensive experimentation and prompt engineering, this method yielded improved results in terms of annotation agreement. However, this approach proved to be highly resource-intensive, pushing up costs significantly due to increased OpenAI API usage.

Our automated evaluations on 262 dialogues provided some revealing observations, as shown in Table 3. Notably, only around 40% of system turns demonstrated KB-Alignment and KB-Grounding. When considering

Table 3
Turn-by-turn GPT-4o evaluation of KB consistency in The Dining Llamas of Oz validation and test splits.

Dialogues	# Dialogues	# Turns	KB-Alignment	KB-Grounding	Correct Turns	Correct Dialogues
All	262	656	41.46%	38.26%	26.35%	8.78%
Successful Bookings	196	494	42.51%	41.50%	28.59%	11.29%
Failing Bookings	66	162	38.27%	28.40%	19.62%	0.5%
Short dialogues	187	411	42.09%	38.44%	29.02%	11.23%
Long dialogues	75	245	40.41%	37.96%	22.80%	3.17%

both metrics together for Correct Turns and Correct Dialogues, the results were even more concerning: just 26% of turns and less than 9% of dialogues met the criteria for both metrics. These numbers underscore the inadequacy of current systems, indicating that a system producing such a low percentage of correct dialogues is not practical for real-world applications.

Further analysis showed that dialogues with successful bookings performed better than those with failed bookings. Specifically, dialogues with successful bookings had 28.59% of correct turns and 11.29% of correct dialogues, compared to dialogues with failed bookings, which had 9 percentage points fewer correct turns and only 0.5% correct dialogues. This discrepancy likely arises because when no suitable restaurants are available, the Llama model tends to hallucinate, providing restaurants not present in the KB. While these restaurants may exist in Cambridge, they are absent from the provided dataset, highlighting the model's failure to adhere to the instructions given in the prompt.

We also explored the impact of dialogue length on performance. Shorter dialogues achieved nearly 30% correct turns and 11.23% correct dialogues, while longer dialogues showed a significant drop: 7 percentage points fewer correct turns and only 3.17% correct dialogues. This suggests that as the conversation progresses, the likelihood of errors increases, possibly due to the model's difficulty in managing and integrating information from previous turns.

Overall, our findings highlight that current state-of-the-art open-source LLMs, such as Llama-3, are still unable to effectively serve as task-oriented dialogue systems while maintaining consistency with a provided KB. This underscores the need for further advancements in LLM capabilities and evaluation methodologies before such systems can be reliably used in practical applications.

6. Limitations

While our study makes significant contributions to understanding the capabilities of state-of-the-art LLMs in performing task-oriented-dialogue tasks, it is important to acknowledge certain limitations that may affect the generalizability and scalability of our findings. The turn-by-turn evaluation approach, while effective in enhancing evaluation accuracy, proved to be computationally expensive. The quality of GPT-4o's evaluations was highly dependent on effective prompt engineering. Crafting the right prompts to ensure accurate evaluation results was challenging and time-consuming. Additionally, employing a diverse set of models for generating and evaluating dialogues could provide more comprehensive findings. Using multiple models might help in understanding the strengths and limitations of different approaches, potentially offering a more robust analysis of dialogue quality and consistency. This could also help in mitigating the limitations inherent in any single model or evaluation approach.

7. Conclusions and Future Work

In this study, we explored the capabilities of state-of-the-art LLMs in generating task-oriented dialogues, focusing on maintaining consistency with a provided KB and avoiding hallucinations. Our experiments demonstrated that Llama-3, despite its advancements, struggles to perform reliably in these settings. The model showed significant limitations, especially in dialogues that led to failed outcomes (where the desired restaurant was not in the KB) and longer interactions. As a side contribution, we release The Dining Llamas of Oz, a corpus of 1,311 dialogues generated through user-Llama and system-Llama interactions, to aid future research. Our findings highlight the need for further development to improve LLM reliability and accuracy in task-oriented dialogue applications.

Aknowledgments

This work has been partially supported by the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by NextGenerationEU.

References

[1] M. McTear, Conversational ai: Dialogue systems, conversational agents, and chatbots, Synthesis Lectures on Human Language Technologies 13 (2020) 1–251.

[2] T. Labruna, B. Magnini, Addressing domain changes in task-oriented conversational agents through dialogue adaptation, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, 2023, pp. 149–158.

[3] S. Young, M. Gašić, B. Thomson, J. D. Williams, Pomdp-based statistical spoken dialog systems: A review, Proceedings of the IEEE 101 (2013) 1160–1179.

[4] S. Louvan, B. Magnini, Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 480–496. URL: https://www.aclweb.org/anthology/2020.coling-main.42. doi:`10.18653/v1/2020.coling-main.42`.

[5] V. Balaraman, S. Sheikhalishahi, B. Magnini, Recent neural methods on dialogue state tracking for task-oriented dialogue systems: A survey, in: Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2021, pp. 239–251.

[6] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open foundation and fine-tuned chat models, 2023. `arXiv:2307.09288`.

[7] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, Łukasz Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, J. H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Łukasz Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. de Avila Belbute Peres, M. Petrov, H. P. de Oliveira Pinto, Michael, Pokorny, M. Pokrass, V. H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng,

J. Zhuang, W. Zhuk, B. Zoph, Gpt-4 technical report, 2024. URL: https://arxiv.org/abs/2303.08774. arXiv:2303.08774.

[8] P. Budzianowski, T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, M. Gašić, MultiWOZ - a large-scale multi-domain wizard-of-Oz dataset for task-oriented dialogue modelling, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 5016–5026. URL: https://www.aclweb.org/anthology/D18-1547. doi:10.18653/v1/D18-1547.

[9] T. Han, X. Liu, R. Takanabu, Y. Lian, C. Huang, D. Wan, W. Peng, M. Huang, Multiwoz 2.3: A multi-domain task-oriented dialogue dataset enhanced with annotation corrections and co-reference annotation, in: Natural Language Processing and Chinese Computing: 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13–17, 2021, Proceedings, Part II 10, Springer, 2021, pp. 206–218.

[10] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, biometrics (1977).

[11] B. Richards, Type/token ratios: What do they really tell us?, Journal of child language 14 (1987) 201–209.

A. Llama Prompts

The following prompt has been used to instruct a Llama to play the role of a Cambridge InfoTown system, in English:

```
"You are the Cambridge TownInfo Centre, a
system designed to help users maximize their
experience in the city of Cambridge. Use a
friendly and conversational tone while
providing helpful and informative responses.
All the information you provide must
strictly rely on the Knowledge Base that you
have been provided with. Ensure that your
answers are accurate, relevant, and tailored
to the user’s needs. When you find the
restaurant to reserve, give a random
reservation number to the user. Be brief."
```

The following prompt has been used to instruct a Llama to play the role of a Cambridge InfoTown system, in Italian:

```
"Sei l’assistente Cambridge InfoCittà, un
sistema progettato per aiutare gli utenti a
trarre il meglio dalla loro esperienza nella
città di Cambridge. Usa un tono amichevole e
onversazionale, fornendo risposte
informative e utili. Tutte le informazioni
che fornisci devono basarsi strettamente
sulla Knowledge Base che ti è stata data.
Assicurati che le tue risposte siano accurate,
pertinenti, e mirate ai bisogni dell’utente.
Sii breve."
```

The following prompt has been used to instruct a Llama to play the role of a user looking for a restaurant in Cambridge, in English:

```
"You are a turist in the city of Cambridge
and you are looking for a restaurant to dine
in. Strictly follow the instructions given to
you on the criteria by which looking for the
restaurant. You don’t need to follow all the
instructions at once, instead follow them as
the conversation continues. Be very brief,
and go straight to the point. At the end,
thank the system and say goodbye. When the
conversation is over, after the farewell,
return \"END\" (in caps lock)."
```

The following prompt has been used to instruct a Llama to play the role of a user looking for a restaurant in Cambridge, in Italian:

```
"Sei un turista nella città di Cambridge e
stai cercando un ristorante dove cenare.
Basati strettamente sulle istruzioni che ti
vengono fornite riguardo i criteri in base ai
quali cercare il ristorante. Non seguire
tutte le istruzioni subito, invece seguile
passo passo durante la conversazione. Sii
molto breve e vai subito al punto."
```

B. GPT Prompts

The following system prompt has been used has general instruction for telling GPT to behave like a dialogue evaluator:

```
"You are a dialogue evaluator. Given a
dialogue you have to return a list of symbols
separated by commas, where each symbol is an
evaluation of each turn in the dialogue. Only
system turns must be considered."
```

The following prompt has been used to instruct GPT to determine if a system turn talks about information contained in a KB:

```
"Given the following user and system turns,
return 1 if the system turn contains
information that requires verification from
```

an external source to ensure its accuracy, 0 otherwise."

The following prompt has been used to instruct GPT to determine if a system turn constitute a KB-Error:

"Given the following user turn, system turn, and Knowledge Base (KB), return 0 if the system contradicts the KB (e.g. says that a restaurant is at north, but it’s actually at south), 1 otherwise."

The following prompt has been used to instruct GPT to determine if a system turn constitute an KB-Grounding error:

"Given the following user turn, system turn, and Knowledge Base, return 1 if the system doesn’t mention properties outside of the Knowledge Base, 0 otherwise (e.g. says that the restaurant serves british and indian, but only indian is present in the KB)."