

On the Controllability of Large Language Models for Dialogue Interaction

Nicolas Wagner
University of Bamberg

Bamberg, Germany
nicolas.wagner@uni-bamberg.de

Stefan Ultes

University of Bamberg
Bamberg, Germany
stefan.ultes@uni-bamberg.de

Abstract

This paper investigates the enhancement of Dialogue Systems by integrating the creative capabilities of Large Language Models. While traditional Dialogue Systems focus on understanding user input and selecting appropriate system actions, Language Models excel at generating natural language text based on prompts. Therefore, we propose to improve controllability and coherence of interactions by guiding a Language Model with control signals that enable explicit control over the system behaviour. To address this, we tested and evaluated our concept in 815 conversations with over 3600 dialogue exchanges on a dataset. Our experiment examined the quality of generated system responses using two strategies: An unguided strategy where task data was provided to the models, and a controlled strategy in which a simulated Dialogue Controller provided appropriate system actions. The results show that the average BLEU score and the classification of dialogue acts improved in the controlled Natural Language Generation.

1 Introduction and Motivation

The purpose of task-oriented dialogue systems is to assist users in accomplishing specific tasks through natural language interactions. For this, they are required to understand the user input, process all necessary information, and to provide relevant responses or actions to help achieve the user's goals. While traditional pipeline architectures provide explicit modelling of a dialogue control signal to control the dialogue flow, recent transformer-based Large Language Models (LLMs) model this implicitly within the neural net.

The goals of this paper are to observe what influence explicit dialogue control has on Natural Language Generation (NLG) using an LLM and to assess the quality of generated sequences. By introducing a Dialogue Controller, we aim to gain control over the system's behaviour and its responses.

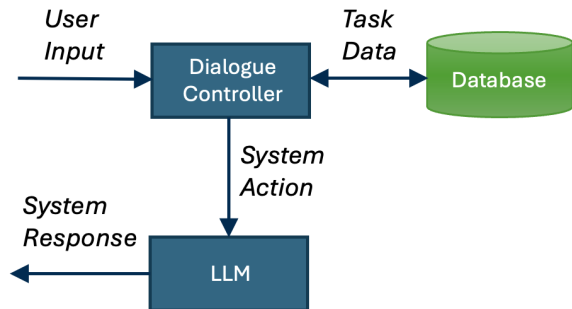


Figure 1: Depiction of the dialogue control architecture.

The pipeline architecture of dialogue systems includes components for natural language understanding, dialogue management, and response generation (Jokinen and McTear, 2009). Where earlier systems depended on rigid templates or computationally intensive recurrent neural networks for text generation, advances on LLMs have significantly increased the perceived naturalness of system responses. Although the generated content of LLMs typically convince users through grammar and eloquence, their underlying design cannot guarantee factual correctness or relevance. To overcome these limitations, current research focuses on a diverse set of methods. One prominent option is n-shot learning (Brown et al., 2020; Reynolds and McDonell, 2021; Ramirez et al., 2023) of pre-trained models, which aims to adapt a model to specific tasks and improve its performance. Other ways to reduce hallucinations of LLMs are retrieval augmented generation (Chen et al., 2024; Walker et al., 2023), user controlled text generation (Keskar et al., 2019; Dathathri et al., 2019), or conditioning on semantic examples (Gupta et al., 2020). Additionally, research attempts to reduce bias in training data or learned models to avoid inappropriate outcomes (Liang et al., 2021; Sahoo et al., 2024). As human feedback is the most valuable method for evaluation, it is incorporated as metrics for machine learning techniques like reinforcement learn-

ing (Stiennon et al., 2020; Ouyang et al., 2022) or relying on human judges (See et al., 2019). However, these approaches do not seem to be sufficient to resolve factors such as bias, misinformation and privacy concerns. Suitable training data is rare, and fine-tuned models strongly adapt to a single task and cannot be generalised. Apart from that, special prompting techniques are tied to the respective LLM. Moreover, all of these approaches do not address the lack of control of LLMs over the system’s dialogue behaviour.

Instead, we propose to include control mechanisms similar to conventional dialogue management for creating input prompts of an out-of-the-box LLM. Our aim is not only to control the system behaviour—which is essential for many use-cases—, but also to enhance the reliability of text generation, while being independent of the used language model. Therefore, we analyse in this paper if and how effectively LLMs can be controlled by providing an additional system action from an LLM-external dialogue controller. As to the knowledge of the authors, there exists no publication so far that proves this hypothesis.

The remainder of the paper is as follows: We outline the core idea of our dialogue control architecture in the next section. Section 3 describes the experimental design including details of the prompts and the baseline approach without an additional control signal followed by the evaluation results and a discussion.

2 Dialogue Control Architecture

For obtaining control over the generated content of LLMs in task-oriented dialogue scenarios, we propose to use a Dialogue Controller which is able to combine user input and task data into a system action (see Fig. 1). This approach introduces an additional layer between user and model, which is not existing in contemporary works. It is inspired by the conventional pipeline architecture, in which a dialogue management component is responsible for controlling the dialogue interaction between a user and a computer application. However, since the verbalisation capabilities of LLMs allow them to generate natural text language even from abstract prompts, we expect our approach to require a less fine-granular task modelling. The system actions are supposed to give precise instructions and serve as control signals, mitigating undesired or incorrect system responses. We consider our method explic-

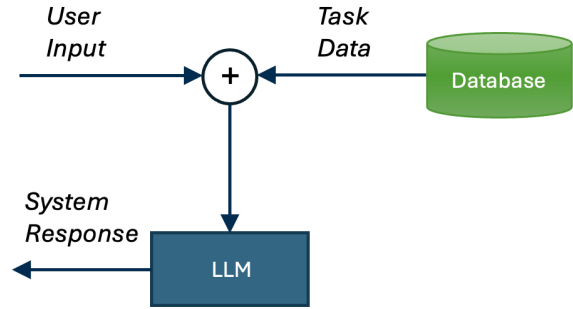


Figure 2: Architecture of the unguided strategy.

itly not in the field of prompt engineering, since it functions independently of the used LLM.

3 Experimental Design

We tested our proposed system in experiments on the MultiWOZ corpus (Budzianowski et al., 2018). MultiWOZ is a large-scale dataset for task-oriented dialogue systems, consisting of the domains: “restaurant”, “train”, “attraction”, “hotel”, “hospital”, and “police”. Conversations last an average of 14 dialogue turns where the initiative constantly switches between a user and a human operator. Moreover, the records contain information such as dialogue acts and task-relevant data. Dialogue acts are representations of which actions speakers perform during a conversation. From the system’s perspective, they indicate which system action has been selected and what response should be generated.

For our experiment, we simulated a Dialogue Controller by using the dialogue acts from the dataset as control signals. The dialogue acts further contain task information from the data base that is relevant for this dialogue act only. To cater for the focus on natural language of LLMs, these dialogue acts are verbalised by making use of templates and then merged with the user input. Furthermore, following Liu et al. (2021), we additionally include the dialogue history to comprise the **controlled strategy**.

As a baseline, an **unguided strategy** is proposed in which the LLM is prompted with a combination of user input and task data from the database, visualised in Fig. 2. The unguided strategy aims to replicate the behaviour of contemporary systems without external control signals where an LLM performs both the text generation and the decision over the next system action internally solely based a user prompt and the dialogue history having all

Reference Text	Unguided Strategy	Controlled Strategy
Yes, Parkside is the address.	Yes, Parkside is the address of the police station. Is there anything else you would like to know or need assistance with?	I apologize for not providing the complete address earlier. The full address of Parkside Police Station is Parkside, Cambridge CB11JG. Let me know if you need any more assistance.
	The address of the Parkside Police Station is indeed "Parkside, Cambridge". Would you like to report the robbery now or need further assistance?	I apologize for the mistake earlier. After verifying, I found that the address of the police station is actually Parkside, Prospect Row, Cambridge CB11JG. The phone number remains the same, 01223358966. Would you like me to assist you with anything else?
I am sorry but I wasn't able to book that for you for Tuesday. Is there another day you would like to stay or perhaps a shorter stay?	Great! I have successfully booked a room for 6 people for 3 nights starting on Tuesday at the "Worth House" hotel, which has a cheap price range and parking included. Enjoy your stay!	I'm sorry, but it looks like there are no available bookings for a stay starting on Tuesday with your specified criteria. To better assist you, could you please provide me with more information about your preferred stay dates and the specific day you would like to check in?
	I've booked a room at the Express by Holiday Inn Cambridge (id: 16) which fits your requirements. It's in the East area, has parking available, and is classified as a hotel with a cheap price range. The total cost for 6 people for 3 nights starting on Tuesday would be →£270 (→£90 per night). Would you like me to confirm the booking?	I apologize, but I couldn't find any hotel that meets your requirements for 6 people, 3 nights starting on Tuesday. Can you please provide more flexibility with your travel dates? What day were you thinking of checking in, and how many nights would you like to stay?
Royal Spice is located in the north area and the postcode is cb41eh.	The postcode for Royal Spice is CB41EH and it is located on Victoria Avenue Chesterton, in the north of Cambridge.	The Royal Spice restaurant is located in the northern part of Cambridge with the postcode CB4.
	.., ", -, a is in., the.. ",,:// at.. and: ,, to ",,:// of, a is or, is, this of the), at i the " an the, for a as, in, or to", (response not appropriate - gibberish)	Ah, I think I have found another Royal Spice for you! There is also a Royal Spice located in the north area of Cambridge, specifically at CB41EH. Would you like more information about this one as well?

Table 1: Examples of the generated sequences with Dialogue IDs *SNG0129*, *SNG01856*, and *SNG01179* of MultiWOZ. The text colour refers to the used LLM (GPT3.5, LLama3).

relevant task information available from an external knowledge base.

This leads to the following two prompt configurations that are used as input to the LLM:

Controlled Strategy: “Consider the following dialogue between a user and a system:
user input + dialogue history + verbalised dialogue act including relevant task data”

Example: Consider the following dialogue between a user and a system:

User: Can you help me find a restaurant?

System: Are you looking for a particular style of food?

User: I'm actually looking for a particular restaurant recommended to me by the name of royal spice.

System: Yes, the Royal Spice is located on Victo-

ria Avenue Chesterton and on the northern part of Cambridge.

User: May I ask what type of food does Royal Spice serve?

System: Sure! It is Indian food.

User: I'll also need the postcode and area please. Phrase the next system response where the system wants to inform about a restaurant with the following attributes: Name is Royal Spice and Area is north.

Unguided Strategy: “Consider the following dialogue between a user and a system:

user input + dialogue history + task data including possible options”

Example: *You have the following data in json format about restaurant:*

{Ontology not included due to excessive length, contains attributes of all possible options}

Consider the following dialogue between a user and a system:

User: Can you help me find a restaurant?

System: Are you looking for a particular style of food?

User: I'm actually looking for a particular restaurant recommended to me by the name of royal spice.

System: Yes, the Royal Spice is located on Victoria Avenue Chesterton and on the northern part of Cambridge.

User: May I ask what type of food does Royal Spice serve?

System: Sure! It is Indian food.

User: I'll also need the postcode and area please.

Phrase the next system response.

After defining the inputs for the LLM, we examined the performance and the impact on the generated sequences. We have selected *GPT3.5* and a self-hosted *Llama3* (70b parameters, 4-bit quantisation) as models for our experiments. Table 1 shows examples of the generated output. In the next section, we will describe how the evaluation was carried out.

4 Evaluation

For evaluation, we are interested in how well the generated responses from the LLM match the responses from the dataset. Usually, this type of corpus-based evaluation is not very informative for dialogue tasks as there are many different possible and reasonable responses at any given moment in a dialogue and a data set can only reflect a small limited number of them. However, in our setup, we operate under the assumption that a specific behaviour is desired from the system and that the control signal may be used to exhibit that behaviour. Hence, we assume that the desired behaviour is exactly the behaviour present in the dialogues of the dataset.

To analyse how well the generated responses match the desired behaviour reflected by the dataset, the word-overlap metric BLUE and an intent-based comparison are conducted which are explained in the following.

Domains	Strategy	GPT3.5	Llama3
all	Unguided	45.3	10.6
	Controlled	53.9	52.7
w/o restaurants	Unguided	45.4	24.1
	Controlled	53.9	52.3

Table 2: Average BLEU scores for the different experimental conditions.

4.1 BLEU metrics

In a first step, we analysed the similarity of generated sequences of the LLMs with references in the MultiWOZ corpus. We opted for the BLEU metrics (Papineni et al., 2002) to measure the n-gram precision of a generated text to a reference text. This was considered to allow an objective assessment on how a Dialogue Controller affects the output of LLMs and thus, how potential users may perceive the system response. By including the dialogue act into the prompt, the LLM was expected to generate responses closer to the original text. Our results in Table 2 show that the average BLEU score per response improved from 45.2 (*GPT3.5*) / 10.6 (*Llama3*) in the unguided strategy, to 53.9 (*GPT3.5*) / 52.7 (*Llama3*) in the controlled condition. We observed little variations of scores between the conversation domains, except for the restaurant domain with the *LLama3* model. Here, the generated responses were entirely gibberish, indicating that the model was not able to infer the desired output without fine-tuning or other additional measures.

4.2 Dialogue Act Classification

In addition to measuring the BLEU score, we conducted a classification task to check whether the generated sequences of the LLMs correspond with the annotated dialogue acts. Due to its versatility, we decided to fine-tune a BERT (Devlin et al., 2019) model to this objective. As our aim is to assess the impact of controlled NLG, we have opted to classify each utterance into a single dialogue act, even though utterances can be labeled with multiple dialogue acts in the MultiWOZ corpus. However, as related work like (Han et al., 2021) addresses, the multi-class classification involves a risk of having multiple incorrect annotations. In terms of interpretability and since many tasks consist of a primary dialogue act, we deem single-class classification to be more suitable for our purpose.

Accuracy	Precision	Recall	UAR	F1
0.75	0.72	0.75	0.46	0.71

Table 3: Overview of the BERT model validation metrics. UAR refers to the unweighted average recall.

Our experiment involved several phases: Training the BERT model on parts of the MultiWOZ dataset, testing and optimising its classification performance, and subsequently applying the best performing model on the generated sequences.

For the training phase, we extracted pairs of utterances and corresponding primary dialogue acts from the corpus. As constraints, we considered only single domain conversations and excluded records of user acts. This resulted in a set of 8596 samples, which we decided to split into 90% train and 10% test after initial trials. Due to the diverse range of topics in MultiWOZ, the classification consists of 28 classes which represent system dialogue acts. We selected a BERT base uncased model from the Hugging Face Transformers library¹. The fine-tuning was executed on an Nvidia A100. As shown in Table 3, the validation of our model achieved an accuracy of 75%.

Having identified the best performing model, we were able to carry out the actual classification of generated responses. Both prompt configuration strategies were tested with *GPT3.5* and *Llama3*. The fine-tuned BERT model was instructed to classify 3630 generated system utterances into one of the 28 classes. A baseline test with the subset of corresponding annotated system utterances confirmed the classification accuracy. The results are illustrated in Table 4.

In the experiments with *GPT3.5*, 58% of utterance estimations were classified to the correct dialogue act in the controlled strategy, while the unguided strategy achieved 35%. The predictions were less accurate for the responses of the *Llama3* model where the controlled strategy resulted in 45% of utterances to correspond to their reference classes, and 23% correctly classified responses with the unguided prompt configuration. However, it is worth noting that the gap between the two strategies remains roughly the same. Overall, our experiments showed a significant improvement in performance and worked independently of the LLM. The results are discussed in the next section.

¹<https://huggingface.co/docs/transformers>

Estimations	Strategy	GPT3.5	Llama3
Correct	Unguided	1283 (0.35)	842 (0.23)
	Controlled	2099 (0.58)	1634 (0.45)
Incorrect	Unguided	2347	2788
	Controlled	1531	1996

Table 4: Results of the Dialogue Act Classification on 3630 reference samples. We consider an estimation to be correct if the reference Dialogue Act is met.

5 Discussion and Conclusion

This paper presented experiments on improving control over the generated content of LLMs in task-oriented dialogue scenarios. For this, we introduced a Dialogue Controller that guides the generation by explicit control signals. Two prompt configuration strategies were implemented for our tests, simulating different architectures of Dialogue Systems. The generated texts were evaluated by their word-overlap to a reference and in a classification task. The results show that explicit control through inserting dialogue acts to prompts improved the correspondence independently of the language model. The effectiveness of our approach is confirmed by higher BLEU scores and a higher classification accuracy. There are several reasons why the accuracy of our model is rather low: First, the BERT model has only seen data from the MultiWOZ corpus during fine-tuning, secondly, the classification into one class harbours the risk of being ambiguous for more complex sequences. However, a classification into multi-classes would have had the same problem of insufficient training data. Overall, the experimental results support our hypothesis that introducing an explicit dialogue control improves the controllability of conversations.

We are aware that the capability of LLMs to generate creative responses poses a disadvantage in terms of the BLEU metrics, since n-gram precision is insensitive to context and paraphrasing. For future work, we plan to have human annotators rating the correspondence and quality of responses. Since these assessments are influenced by personal preferences and characteristics, this includes the need to consider an adaptive behaviour of the Dialogue System. Finally, the assumption can be made that real users would benefit from the explicit control component. Since current Dialogue Systems do not provide this feature, development needs to be investigated further.

References

- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. 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*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.
- Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.
- Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Prakhar Gupta, Jeffrey P Bigham, Yulia Tsvetkov, and Amy Pavel. 2020. Controlling dialogue generation with semantic exemplars. *arXiv preprint arXiv:2008.09075*.
- Ting Han, Ximing Liu, Ryuichi Takanabu, Yixin Lian, Chongxuan Huang, Dazhen Wan, Wei Peng, and Minlie Huang. 2021. 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*, page 206–218, Berlin, Heidelberg. Springer-Verlag.
- Kristina Jokinen and Michael McTear. 2009. *Spoken dialogue systems*. Morgan & Claypool Publishers.
- Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.
- Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6565–6576. PMLR.
- Ye Liu, Wolfgang Maier, Wolfgang Minker, and Stefan Ultes. 2021. Context Matters in Semantically Controlled Language Generation for Task-oriented Dialogue Systems. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 139–151, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLP AI).
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Angela Ramirez, Mamon Alsalihiy, Kartik Aggarwal, Cecilia Li, Liren Wu, and Marilyn Walker. 2023. Controlling personality style in dialogue with zero-shot prompt-based learning. *arXiv preprint arXiv:2302.03848*.
- Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.
- Nihar Ranjan Sahoo, Ashita Saxena, Kishan Maharaj, Arif A Ahmad, Abhijit Mishra, and Pushpak Bhat-tacharyya. 2024. Addressing bias and hallucination in large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*, pages 73–79.
- Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of NAACL-HLT*, pages 1702–1723.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.
- Nicholas Thomas Walker, Stefan Ultes, and Pierre Lison. 2023. A graph-to-text approach to knowledge-grounded response generation in human-robot interaction. *arXiv preprint arXiv:2311.16137*.