Creating an LLM-based conversation assistant for hotel accommodation booking

Anna Marshalova

July 2024

Abstract

This study explores the adaptation of an open-source Large Language Model for hotel booking assistance. We developed a synthetic dataset using templates generated by larger models, encompassing diverse aspects of the hotel booking process. Our approach implements a dual system where the model simultaneously functions as a conversational assistant and a slot extraction system. Experimental results demonstrate significant improvements in the fine-tuned model compared to the baseline, particularly in extracting booking details. While some dialogue quality metrics show mixed results, human evaluation confirms enhanced contextual adherence and reduced hallucinations. Despite remaining challenges, such as occasional question repetition, this work offers valuable insights into developing specialized dialogue systems for the travel industry.

1 Introduction

Artificial Intelligence (AI) is rapidly evolving and finding applications in various sectors, including travel and tourism. One practical application is the development of hotel booking assistants, which can help users book accommodations and provide hotel recommendations. There are several approaches to creating such assistants using Large Language Models (LLMs). These include simple prompting to existing models, implementing Retrieval-Augmented Generation (RAG) pipelines, or fine-tuning models for specific tasks. In this study, we explore the use of Parameter-Efficient Fine-Tuning techniques to adapt an open-source LLM for hotel booking assistance.

Our approach focuses on training a model to perform two key tasks: slot-filling (extracting relevant information from user inputs) and maintaining a dialogue with the user. To achieve this, we created a dataset using templates generated by larger, proprietary models. This dataset was designed to cover various aspects of the hotel booking process and to expose the model to a range of potential user interactions. We compared our fine-tuned model with an untrained baseline to assess the effectiveness of our approach. Our study aimed to evaluate improvements in slot-filling accuracy, the model's ability to

use contextual hotel information, and its adherence to a predefined dialogue scenario.

This paper presents an exploration of techniques for creating a specialized hotel booking assistant. Our work contributes to the ongoing efforts in adapting open-source LLMs for specific tasks in the travel industry. By sharing our methodology and results, we aim to provide practical insights that may be valuable for researchers and practitioners working on similar applications or exploring efficient ways to leverage LLMs in domain-specific contexts.

2 Related Work

AI is increasingly being applied across various sectors, and the travel industry is no exception. Several studies have explored the implementation of conversational AI in travel-related services. For instance, [Ukpabi et al., 2019] reviewed the use of chatbots in tourism, highlighting their potential for enhancing customer service and optimizing booking processes. Similarly, [Melián-González et al., 2021] found that chatbot usage in tourism is gaining popularity due to users' appreciation of their predictable yet friendly and human-like interactions.

For the last several years, there have been significant advancements in developing of LLMs. Having been trained on extensive corpora of data, fine-tuned on instructions, and aligned to human preferences, these models become capable of in-context learning [Brown et al., 2020], which allows them to tackle a wide variety of tasks without task-specific training. Consequently, they can serve as text summarizers, language translators, and code generators, as well as conversational agents, that may maintain conversations on general topics or act like task-oriented dialogue systems [Zhao et al., 2022], e.g., an assistant that helps users book hotel accommodation.

To enhance the factual accuracy of LLM responses, such models are often aided with Retrieval-Augmented Generation (RAG) [Lewis et al., 2020]. This approach demonstrates improved performance on knowledge-intensive tasks, which is particularly relevant for providing accurate hotel information, such as amenities, rates, and availability.

However, RAG and in-context learning for difficult tasks such as topic-specific multi-turn conversations work best with large and proprietary models such as GPT-4 [Achiam et al., 2023], Claude [ant,], and Gemini [Reid et al., 2024]. Smaller open-source models: Llama [AI@Meta, 2024] or Mistral [Jiang et al., 2023] may struggle to leverage context correctly, and keep factual groundedness of responses, or effectively manipulate information from previous dialogue turns. This limitation necessitates training such models on task- and domain-specific data to improve their performance in specialized applications like hotel booking systems.

There have been a vast amount of studies focusing on LLMs fine-tuning to enhance their performance in RAG scenarios and downstream tasks [Xiong et al., 2024, Padró and Saurí, 2024, Mirza et al., 2024, VM et al., 2024]. However, fine-tuning the whole model requires computational costs and memory. To address these

limitations, researchers have developed various Parameter-Efficient Fine-Tuning (PEFT) methods.

One of the most prominent PEFT methods is Low-Rank Adaptation (LoRA), introduced by [Hu et al., 2021]. LoRA adds trainable rank decomposition matrices to the weights of the pre-trained model, allowing for task-specific fine-tuning with a fraction of the parameters. Other notable PEFT techniques include Prompt Tuning [**?**] and Prefix Tuning [Li and Liang, 2021], which modify the input or internal representations of the model by adding trainable continuous vectors, allowing for task-specific adaptation without altering the majority of the pre-trained model's parameters.

3 Approach

Our approach includes prompting one model for two different tasks so that it could process each user's message from two perspectives. Firstly, it must act as an assistant that engages in conversation with the user, asking questions and providing necessary information in a friendly tone. The model is constrained to a predefined list of available hotels, ensuring accuracy and preventing misinformation about non-existent hotels. It also incorporates contextual information from previous user interactions, including the city of interest, reservation dates, number of guests, and selected hotel, and has to adhere strictly to this information throughout the conversation.

Concurrently, the same model, guided by a separate prompt, functions as a slot extraction system. This secondary task involves extracting key entities such as city, dates, number of guests, and hotel name from user messages and presenting them in a JSON dictionary, which guarantees structured, verifiable, and easily parsable outputs. These extracted slots are then integrated into the context for the primary conversational task, creating a cohesive interaction loop.

4 Dataset

The initial dataset for this study was sourced from a HuggingFace repository[1]. After deduplication, this dataset contains 644 signle-tutn dialogues, covering 374 hotels in 12 cities. However, most of them are almost identical - they are constructed using the same template, differing only in hotel and city data. This leads to a lack of diversity in the data, making the dataset unsuitable for training a model.

To solve this problem, we augmented the dataset with synthetic message templates corresponding to every step of a potential dialogue for both user and assistant, including greetings, asking for and giving information, hotel suggestion and choice, and booking confirmation. The templates we generated by Claude 3.5 Sonnet and GPT-4o. Some of them contain placeholders, which can be filled in with information about the hotel, travel dates, etc.

[1] https://huggingface.co/datasets/KvrParaskevi/hotel$_data$

Moreover, we created a basic dialogue template and applied transformations to it to get new dialogue templates. These transformations include merging, removing, and adding some dialogue turns and intents to users' messages.

Having combined the obtained dialogue and message templates and hotel data from the original dataset, we created a set of 1200 synthetic dialogues between a user and a hotel booking assistant (an example of such a dialog is shown in Figure 1). They were then divided into 5281 single-turn dialogues. This number doubles in the dataset due to the fact that for each dialogue, we contemplate two scenarios: the first implying that the model should act like a hotel booking assistant and the second requiring the model to serve as a slot-filling system.

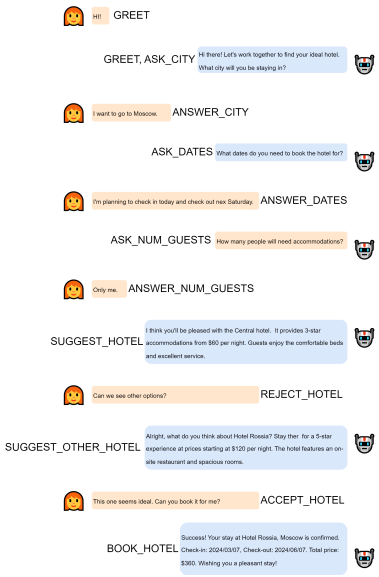

Figure 1: Example of a dialogue built upon the basic dialogue template. Each action of the dialogue is presented in capital letters, while messages themselves are built using message templates.

Each short dialogue has its own separate system prompt, containing some of the current states: the current date, slots that are already filled in, and information about hotels in the city of interest. For each sample, we randomly choose system prompt templates from a set generated by Claude and GPT-4o. This approach aims to expose the model to a wider range of task formulations during training, potentially enhancing its ability to generalize and perform consistently across different prompts during inference.

Aiming to reach more controlled and targeted generation for slot extraction, we added the SLOT_EXTRACTION token to the beginning of every system prompt for this task. This approach was hypothesized to serve as a task-specific trigger, reinforcing the model's awareness of its objective. This technique draws

inspiration from a method widely employed in text-to-image model fine-tuning [Ruiz et al., 2023], although its application in LLM fine-tuning for inference is less common. In image generation, this method, typically, involves using rare tokens to minimize the likelihood of the identifier having strong prior associations. However, we diverged from this approach, opting instead for a token semantically aligned with the task being solved. This decision resembles the embedding initialization strategy used in Prefix Tuning, where prefix embeddings are initialized with task-relevant tokens.

To preserve the model's general purpose applicability and instruction following capabilities, we enriched the dataset with 500 samples from OpenAssistant dataset [Köpf et al., 2024].

5 Experiments

5.1 Experiment Setup

In our experiments, we used Llama-2-chat 7B [Touvron et al., 2023] as a baseline and applied LoRA fine-tuning on the dataset described in Section 4. For training we used the train split of this dataset, containing 9254 samples.

The fine-tuning and inference processes were conducted using the Unsloth framework[2], which effectively reduced memory usage and accelerated the training process.

We used the following training hyperparameters:

- Batch Size: 16
- Optimizer: AdamW [Loshchilov and Hutter, 2017] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$
- Train Steps: 600
- Max sequence length: 2048
- Learning rate: 5×10^{-5}
- Warmup Steps: 100
- LoRA rank: 16
- LoRA alpha: 16

5.2 Results

5.2.1 Automatic Evaluation

We evaluated the model's general ability to maintain a dialogue on hotel booking as well as its slot filling capabilities.

[2]https://github.com/unslothai/unsloth

For dialogue evaluations, we used UniEval [Zhong et al., 2022] library, which is able to evaluate dialogues by five parameters, namely naturalness, coherence, engagingness[3], groundedness and understandability.

As shown in Table 1 the baseline model outperforms the dataset and model fine-tuned on it in engagingness. This is probably because the model is not strictly constrained to adhere to dialog scenarios and is able to output more interesting and unusual responses. The higher score of understandability of our model can be explained by its fine-tuning on a templated dataset, which implicitly controls the length and complexity of generated outputs.

While coherence remains reasonably high for our model, it falls short of the baseline. This might be due to the way the dataset is constructed: the random sampling of prompt-response pairs occasionally results in mismatches, as evidenced by the lower coherence score for the dataset.

Interestingly, other UniEval results appear to contradict human evaluation findings, according to which the fine-tuned model responses are more grounded, while probably being less natural than those of the baseline. Apparently, we should explore the evaluation framework and its applicability for our task in more detail. In particular, we should consider whether the JSON format is an appropriate context for assessing groundedness.

Metric	**Reference**	**Baseline**	**Ours**
Naturalness	0.927	0.814	**0.963**
Coherence	0.848	**0.997**	0.921
Engagingness	1.942	**9.706**	1.817
Groundedness	0.714	**0.997**	0.756
Understandability	0.925	0.863	**0.960**

Table 1: Dialogue evaluation metrics, measured using UniEval, for Reference dialogues from the test dataset, the Baseline model, and its version, fine-tuned in the proposed dataset (Ours)

For slot-filling evaluation, we compared the models' outputs with ground truth slots from the dataset and computed the accuracy score. As shown in Table 2, our model significantly outperforms the baseline across all of the slots.

5.2.2 Human Evaluation

According to human evaluation, the fine-tuned model demonstrates significant improvements over the baseline in several key areas. Firstly, it exhibits a marked reduction in hallucinations and adheres more strictly to the provided context, avoiding the insertion of information about the hotels not present in the prompt. Secondly, the model shows enhanced compliance with the required output format for slot extraction, consistently producing JSON structures enclosed in triple backticks (“‘). Furthermore, compared to the baseline, the fine-tuned

[3]The scoring range for engagingness is [0, +), while all others are [0, 1].

Slot	Accuracy	
	Baseline	**Ours**
Check-in date	0.367	**0.955**
Check-out date	0.500	**0.934**
City	0.181	**0.994**
Hotel name	0.815	**0.908**
Number of guests	0.520	**0.992**

Table 2: Slot filling performance metrics for the Baseline model and its version, fine-tuned in the proposed dataset (Ours)

model displays a decreased tendency to ask users irrelevant questions (e.g., about the gender distribution of guests).

However, the fine-tuned model still has some limitations. Particularly, it sometimes repeats questions that users have already answered, occasionally resulting in conversational deadlocks. This shortcoming suggests that there is room for improvement in the model's ability to maintain and utilize conversation history effectively.

6 Conclusion

In this study, we investigated the application of various techniques to customize an open-source Large Language Model (LLM) specifically for hotel booking assistance. Our approach aimed to enhance the model's performance in two areas: accurate slot-filling and maintaining coherent, context-aware dialogues with users.

The experimental results demonstrated some enhancements in the fine-tuned model compared to the baseline. Specifically, the model exhibited improved accuracy in extracting crucial details such as check-in/out dates, city and hotel names, and guest numbers. It also showed better utilization of contextual hotel information, leading to more reliable user interactions. Additionally, the model adhered more closely to predefined dialogue structures, resulting in more purposeful conversations.

However, we also identified areas for further improvement. Potential directions include:

- Enhancing and expanding the dataset to make the dialogues more realistic and coherent. This might involve considering another method of splitting a multi-turn dialog into samples, less random choosing of templates for each message, creating training examples that emphasize the importance of referencing previously provided information, and adding new intents to dataset dialogues (e.g., user asking about hotel facilities, breakfast availability or asking to modify some information during booking confirmation).

- Conducting more experiments, e.g., trying different models and hyperparameters. Probably, similar results could be obtained by leveraging models with 1-3 billion parameters.
- Using the model with a RAG pipeline. For instance, a list of hotels to put into the model's context could be retrieved from a vector database, based on the user's preferences, described in a message.

In conclusion, our work contributes to the growing body of work on LLM adaptation for specialized tasks, offering insights into the challenges and potential solutions for creating robust, open-source hotel booking assistants. Our findings have implications not only for the travel industry but also for the broader field of task-oriented dialogue systems based on LLM usage.

References

[ant,] Claude 3.5 sonnet, author=Aanthropic, year=2024, url = https://www.anthropic.com/news/claude-3-5-sonnet.

[Achiam et al., 2023] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

[AI@Meta, 2024] AI@Meta (2024). Llama 3 model card.

[Brown et al., 2020] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

[Hu et al., 2021] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

[Jiang et al., 2023] Jiang, A., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. (2023). Mistral 7b (2023). *arXiv preprint arXiv:2310.06825*.

[Köpf et al., 2024] Köpf, A., Kilcher, Y., von Rütte, D., Anagnostidis, S., Tam, Z. R., Stevens, K., Barhoum, A., Nguyen, D., Stanley, O., Nagyfi, R., et al. (2024). Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36.

[Lewis et al., 2020] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

[Li and Liang, 2021] Li, X. L. and Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

[Loshchilov and Hutter, 2017] Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

[Melián-González et al., 2021] Melián-González, S., Gutiérrez-Taño, D., and Bulchand-Gidumal, J. (2021). Predicting the intentions to use chatbots for travel and tourism. *Current Issues in Tourism*, 24(2):192–210.

[Mirza et al., 2024] Mirza, P., Sudhi, V., Sahoo, S. R., and Bhat, S. R. (2024). ILLUMINER: Instruction-tuned large language models as few-shot intent classifier and slot filler. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8639–8651, Torino, Italia. ELRA and ICCL.

[Padró and Saurí, 2024] Padró, L. and Saurí, R. (2024). Fine-tuning open access LLMs for high-precision NLU in goal-driven dialog systems. In Gaspari, F., Moorkens, J., Aldabe, I., Farwell, A., Altuna, B., Piperidis, S., Rehm, G., and Rigau, G., editors, *Proceedings of the Second International Workshop Towards Digital Language Equality (TDLE): Focusing on Sustainability @ LREC-COLING 2024*, pages 33–42, Torino, Italia. ELRA and ICCL.

[Reid et al., 2024] Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T., Alayrac, J.-b., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., et al. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

[Ruiz et al., 2023] Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. (2023). Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510.

[Touvron et al., 2023] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

[Ukpabi et al., 2019] Ukpabi, D. C., Aslam, B., and Karjaluoto, H. (2019). Chatbot adoption in tourism services: A conceptual exploration. In *Robots, artificial intelligence, and service automation in travel, tourism and hospitality*, pages 105–121. Emerald Publishing Limited.

[VM et al., 2024] VM, K., Warrier, H., Gupta, Y., et al. (2024). Fine tuning llm for enterprise: Practical guidelines and recommendations. *arXiv preprint arXiv:2404.10779*.

[Xiong et al., 2024] Xiong, Z., Papageorgiou, V., Lee, K., and Papailiopoulos, D. (2024). From artificial needles to real haystacks: Improving retrieval capabilities in llms by finetuning on synthetic data. *arXiv preprint arXiv:2406.19292*.

[Zhao et al., 2022] Zhao, X., He, B., Wang, Y., Li, Y., Mi, F., Liu, Y., Jiang, X., Liu, Q., and Chen, H. (2022). UniDS: A unified dialogue system for chit-chat and task-oriented dialogues. In Feng, S., Wan, H., Yuan, C., and Yu, H., editors, *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 13–22, Dublin, Ireland. Association for Computational Linguistics.

[Zhong et al., 2022] Zhong, M., Liu, Y., Yin, D., Mao, Y., Jiao, Y., Liu, P., Zhu, C., Ji, H., and Han, J. (2022). Towards a unified multi-dimensional evaluator for text generation. In *2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*.