DIALIGHT: Lightweight Multilingual Development and Evaluation of Task-Oriented Dialogue Systems with Large Language Models

Songbo Hu[1] **Xiaobin Wang**[2] **Zhangdie Yuan**[3] **Anna Korhonen**[1] **Ivan Vulić**[1]
[1]Language Technology Lab, University of Cambridge, UK
[2]Independent Researcher
[3]Department of Computer Science and Technology, University of Cambridge, UK
[1,3]{sh2091,zy317,alk23,iv250}@cam.ac.uk
[2]wxb9585@gmail.com

Abstract

We present DIALIGHT, a toolkit for developing and evaluating multilingual Task-Oriented Dialogue (TOD) systems which facilitates systematic evaluations and comparisons between TOD systems using fine-tuning of Pretrained Language Models (PLMs) and those utilising the zero-shot and in-context learning capabilities of Large Language Models (LLMs). In addition to automatic evaluation, this toolkit features (i) a secure, user-friendly web interface for fine-grained human evaluation at both local utterance level and global dialogue level, and (ii) a microservice-based backend, improving efficiency and scalability. Our evaluations reveal that while PLM fine-tuning leads to higher accuracy and coherence, LLM-based systems excel in producing diverse and likeable responses. However, we also identify significant challenges of LLMs in adherence to task-specific instructions and generating outputs in multiple languages, highlighting areas for future research. We hope this open-sourced toolkit will serve as a valuable resource for researchers aiming to develop and properly evaluate multilingual TOD systems and will lower, currently still high, entry barriers in this field.

1 Introduction

Task-oriented dialogue (TOD) systems are designed to model interactions between human users and system agents, focusing on accomplishing specific, predefined tasks such as assisting with hotel or restaurant bookings, or providing domain-specific FAQ information (Gupta et al., 2006; Tür et al., 2010; Young, 2010). These systems serve not only as access points to cutting-edge AI applications but also as drivers of technological expansion.

The prevailing approach in TOD system development has predominantly involved fine-tuning Pretrained Language Models (PLMs), like T5 (Raffel et al., 2020) and BART (Lewis et al., 2020), on task-specific dialogue datasets (Budzianowski et al., 2018a; Byrne et al., 2019). However, recent research trends indicate a paradigm shift from fine-tuning to an increased reliance on Large Language Models' (LLMs) inherent capacity for in-context learning and generalisation for natural language understanding and generation. In our work, we categorise systems as fine-tuned-based using PLMs (hereafter FT-based) or in-context-learning-based using LLMs (ICL-based), noting that LLMs can be fine-tuned and smaller PLMs can use ICL.[1]

Several pilot works (Hudeček and Dusek, 2023; Heck et al., 2023; Zhang et al., 2023; Chung et al., 2023) have explored the integration of LLMs into TOD systems. These studies indicate that FT-based approaches outperform ICL-based approaches, as evidenced by superior automatic evaluation scores. This applies even with smaller PLMs and when the number of training examples is limited. On the other hand, instruction-based training of LLMs demonstrates its potential in aligning model outputs more closely with human preferences (Ouyang et al., 2022; Wang et al., 2022). In TOD, ICL-based systems (Chung et al., 2023) have been shown to generate responses that exceed previous models in critical human evaluation dimensions such as informativeness, helpfulness, and perceived humanness. Despite these early results, a systematic, comparative human evaluation of these two approaches in TOD systems remains a gap in current research.[2]

[1]**Disclaimer:** Here, we (coarsely) differentiate between PLMs and LLMs in terms of their dependency on fine-tuning with task-specific datasets for achieving optimal performance. LLMs, such as LLaMA (Touvron et al., 2023) and GPT-4 (OpenAI, 2023), are characterised by their extensive training on a broad spectrum of data. This approach enables LLMs to adapt to diverse tasks with minimal reliance on task-specific data. Empirical evidence suggests that LLMs demonstrate remarkable capabilities, nearing or 'surpassing' human-level performance on NLP benchmarks such as SuperGLUE (Wang et al., 2019) and BIG-Bench (Srivastava et al., 2023).

[2]The study by Chung et al. (2023) lacks full details on their human evaluation protocol, leading to potential ambiguities in interpretation. Moreover, current research infrastructure falls short in facilitating extensive human evaluation of end-to-end

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations), pages 36–52

Toolkit	Human Evaluation	Multilinguality	LLM+E2E	Comparative Experiment
PyDial (Ultes et al., 2017)	✓	✗	✗	✗
ConvLab2 (Zhu et al., 2020)	✓	✗	✗	✗
ConvLab3 (Zhu et al., 2022)	✓	✓	✗	✗
to-llm-bot (Hudeček and Dusek, 2023)	✗	✗	✓	✗
other E2E baselines	✗(∗)	✗(∗)	✗	✗
DIALIGHT (this work)	✓	✓	✓	✓

Table 1: A comparative overview of TOD system toolkits supporting E2E modelling. This summary excludes system components focused solely on dialogue state tracking (DST) and response generation (RG). Key features include **Human Evaluation**, indicating support for online (crowdsourcing) human evaluation, **Multilinguality** capabilities for development and evaluation across various languages and in monolingual, multilingual, and cross-lingual setups, **LLM+E2E** for in-context learning with LLMs in E2E modelling, and **Comparative Experiment**, denoting a unified framework for evaluating both FT-based and ICL-based systems. Detailed comparisons are presented in § 2. (∗) We acknowledge that while some TOD systems support human evaluation and multilingualism, it is recognised that a significant majority lack these crucial features.

Furthermore, the development of TOD systems has historically been confined to a limited number of high-resourced languages (Razumovskaia et al., 2022). The recent release of the Multi3WOZ dataset (Hu et al., 2023a) expands the linguistic scope, introducing the same-level data support for Arabic, English, French, and Turkish TOD. Nevertheless, there are still noticeable disparities in system performance across different languages even with the fully comparable training data (Hu et al., 2023b), raising questions about system utility and user satisfaction with TOD in non-English contexts. To facilitate future research in minimising these performance disparities, a toolkit that supports developing and evaluating multilingual dialogue systems is critically needed.

Aiming to address these gaps, this paper introduces DIALIGHT, a novel toolkit for developing and evaluating multilingual end-to-end (E2E) TOD systems. DIALIGHT is specifically designed for comprehensive comparative analyses between FT-based and ICL-based systems (see §3). It supports an array of TOD datasets from the MultiWOZ family (Budzianowski et al., 2018b; Ding et al., 2022; Hu et al., 2023a, *inter alia*), and enables seamless evaluations in monolingual, multilingual, and cross-lingual setups (§4). Given often the moderate correlation between automatic evaluation metrics and human judgements (Yeh et al., 2021; Mehri et al., 2022), our toolkit places a special emphasis on human evaluation, facilitating both utterance-level and full dialogue-level assessments (§5). The toolkit is available online at: `github.com/cambridgeltl/e2e_tod_toolkit`.

2 Related Work

DIALIGHT represents a novel addition to the landscape of TOD system toolkits, complementing existing frameworks such as PyDial (Ultes et al., 2017), ConvLab-3 (Zhu et al., 2022), and their antecedents (Zhu et al., 2020; Lee et al., 2019), as shown in Table 1. DIALIGHT is unique in offering support for ICL-based implementations in E2E ToD systems, a feature not yet available in existing toolkits. Moreover, DIALIGHT diverges in its core design philosophy. Instead of incorporating extensive and intricate systems and components, it is meticulously crafted to reduce entry barrier and learning curve for researchers engaging in multilingual TOD research. Our objective is to provide a streamlined codebase that facilitates rapid development of multilingual TOD systems.

Another range of publicly available implementations exists for both traditional FT-based and ICL-based E2E TOD systems, typically accompanying research publications as supplementary code. For traditional FT-based approaches, works include those by Wen et al. (2017); Bordes et al. (2017); Lei et al. (2018); Eric and Manning (2017); Eric et al. (2017); Lin et al. (2020); Peng et al. (2021); He et al. (2022) . The ICL-based category is exemplified by the work of Hudeček and Dusek (2023).[3] However, these systems lack a unified setup for comparative experimentation. Our toolkit fills this gap, enabling fair and comprehensive comparisons between the aforementioned two types of systems. Additionally, many of these systems, primarily designed to enhance benchmark results, exhibit key limitations, such as: 1. absence of implementation for lexicalisation in utterances, 2. lack of a dialogue

TOD systems, especially for comparative analyses between FT-based and ICL-based systems.

[3] To date (December 2023), the implementations by Zhang et al. (2023); Chung et al. (2023) have not been released.

system agent for real user interaction, 3. existence of English-centric heuristics in evaluation,[4] 4. inadequate support for human evaluation. DIALIGHT specifically addresses these limitations.

While existing tools for human evaluation, such as DialCrowd (Lee et al., 2018; Huynh et al., 2022), offer simple approaches for conducting human evaluation experiments in TOD systems, they are not without limitations. DialCrowd is not open-sourced, available solely via its designated website, which imposes significant constraints. These include limited customisation flexibility, increased maintenance complexity,[5] and challenges in aligning with data protection regulations such as GDPR, thus affecting its broader applicability in research. In contrast, our human evaluation tool is open-sourced, enabling 'one-click' deployment on local or cloud servers.

3 System Architecture

This section delineates the architecture and implementation of the proposed E2E dialogue system within DIALIGHT. As illustrated in Figure 2 in the Appendix, despite the term 'end-to-end', state-of-the-art E2E TOD systems typically employ a pipelined approach in the background, incorporating three key components: **1)** a dialogue state tracking (DST) model, **2)** a database interface, and **3)** a response generation (RG) model. In the following, we describe the system pipeline and provide an in-depth examination of each constituent component.

3.1 System Pipeline

Our E2E TOD system operates by processing a dialogue history, represented as the concatenation of a list of preceding dialogue utterances $[\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_{t-1}]$ with the latest user utterance $\mathbf{u}_t$. Specifically, the DST model, denoted as $\mathrm{DST}(\cdot)$, takes this concatenated input dialogue utterances to predict the current dialogue state, formulated as $\mathbf{s}_t = \mathrm{DST}([\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_t])$. Within the context of MultiWOZ datasets, a dialogue state is defined as a set of tuples $\mathbf{s} = \{(d_1, s_1, v_1), \cdots, (d_k, s_k, v_k)\}$, where each tuple consists of domain d, slot s, and slot value v. Then, this state is sent to a database $\mathrm{DB}(\cdot)$, from which a set of entities satisfying the requirements specified by the state $\mathbf{s}_t$ are retrieved. Namely, $\{\mathcal{E}_1, \cdots \mathcal{E}_l\} = \mathrm{DB}(\mathbf{s}_t)$, where $\mathcal{E}$ is a data entry in the database. The RG model, denoted as $\mathrm{RG}(\cdot)$, is then tasked to consume the sequence of dialogue utterances with the retrieved set of entities as input and generate a dialogue response $\mathbf{u}_{t+1}$. The process can be formally expressed as $\mathbf{u}_{t+1} = \mathrm{RG}([\mathbf{u}_1, \cdots, \mathbf{u}_t], \{\mathcal{E}_1, \cdots, \mathcal{E}_l\})$, where $\mathbf{u}_{t+1}$ represents the generated dialogue response to user input $\mathbf{u}_t$, taken as the next system turn.

In the proposed toolkit, DST models and RG models can be implemented using both FT-based and ICL-based methods. In what follows, we detail the implementation for each system component, applying both methods.

3.2 Dialogue State Tracking Models

Fine-Tuning with PLMs. The dialogue state $\mathbf{s}_t$ is transformed into a flattened string representation. For example, consider a dialogue state $\{(d_1, s_1, v_1), \cdots, (d_k, s_k, v_k)\}$: it gets transformed into the string d_1#s_1=v_1;s_2=v_2| $\cdots$ |d_k#s_k=v_k. In this representation, slots and their corresponding values within the same domain are merged. For example, the dialogue state '{(taxi, departure, saint johns college), (taxi, destination, pizza hut fenditton)}' is linearised as 'taxi # departure = saint johns college ; destination = pizza hut fenditton'. At each dialogue turn t, a PLM is trained to take the input of the concatenated dialogue history $[\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_t]$, and generates the linearised dialogue state.

In-Context Learning with LLMs. We use LLMs for direct generation of dialogue states in JSON format, guided by task-specific prompts and training examples for in-context learning. The instruction prompt consists of six parts: 1. *Task instruction* for generating the dialogue state. 2. *Output format instruction* to produce results in JSON format. 3. *Ontology instruction* detailing available domains and slots. 4. *Categorical slot instruction*, guiding LLMs to choose from predefined options for categorical slots. 5. *Time slot instruction* for generating times in 24-hour format (`hh:mm`). 6. *Number slot instruction*, directing LLMs to produce non-negative integer values for numeric slots, such as the quantity of individuals in a booking. After these instructions, we incorporate a list of training examples, randomly chosen from the training dataset, to provide a baseline for future research. The final

[4] For instance, the official evaluation script for MultiWOZ (Nekvinda and Dušek, 2021) employs a string matching algorithm that normalises English slot values to their canonical forms. This approach, however, introduces a bias in evaluation, leading to potentially unfair comparisons when extending the framework to other languages.

[5] E.g., DialCrowd is currently offline (December 2023).

step involves appending the concatenated dialogue history $[\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_t]$ to the prompt, enabling the LLMs to generate the corresponding state $\mathbf{s}_t$.

Implementation and Setup. In DIALIGHT, the FT-based DST models can be instantiated with any of the PLMs available in the Huggingface repository (Wolf et al., 2020). Additionally, we provide comprehensive support for various models in ICL-based systems, including: 1. Models from the Huggingface repository, 2. Models accessible through the OpenAI API,[6] 3. LLaMA.cpp models, tailored for on-device LLM inferences.[7] For evaluation of FT-based DST in this paper, we utilise the $\text{mT5}_{\text{small}}$ and $\text{mT5}_{\text{large}}$ (Xue et al., 2021). For ICL-based DST experiments, we employ the GPT-3.5, LLaMA2 (Touvron et al., 2023), and OpenChat-3.5 (Wang et al., 2023) models.[8]

3.3 Database Interface

We adapt the implementation of our database interface from the official MultiWOZ evaluation scripts (Nekvinda and Dušek, 2021) with minor modifications. Each data entity across different domains within the database can be represented as a set of slot-value pairs. In response to each user utterance $\mathbf{u}_t$, our system executes a database query using the predicted dialogue state $\mathbf{s}_t$, to independently retrieve relevant data entries from each respective domain. For each domain, a database entry is retrieved if it meets the following criteria: 1. exact matching of categorical slot values with those prescribed in the dialogue state, 2. achieving a Levenshtein distance for non-categorical slot values that is below a predefined threshold when compared with the dialogue state.[9]

3.4 Response Generation Models

Fine-Tuning with PLMs. In the context of the MultiWOZ datasets, our approach follows the conventional two-step process: initially generating a delexicalised response and subsequently lexicalising this response with slot values from dialogue states and retrieved entities.[10] It is worth noting that the majority of prevalent automatic evaluation metrics predominantly focus on delexicalised responses. However, incorporating lexicalisation is a more realistic scenario, which is also a necessity for proper human evaluation.

The generation of a delexicalised response is modelled as a transduction problem, converting dialogue history into a natural response. To integrate database outcomes, we initially create a summary (e.g., *'attraction has one result found; hotel has no result found'*). Subsequently, a PLM is trained to process the input, which is a combination of dialogue history and the database summary, to produce the delexicalised system response. The lexicalisation process is carried out through a systematic replacement of placeholders with the relevant values from the current dialogue state $\mathbf{s}_t$ and from the retrieved entities.

In-Context Learning with LLMs. The ICL-based method also follows the conventional two-step process. However, instead of fine-tuning, this approach involves prompting LLMs with a task-specific instruction that encompasses four parts: 1. *Task instruction*, which directs the generation of the dialogue response. 2. *Ontology instruction*, providing details on the available domains and slots. 3. *Delexicalisation instruction*, informing the LLM about all available placeholders and guiding it to substitute slot values with these placeholders. 4. *Language instruction*, specifying the target language for the generated response. A set of training examples, randomly selected from the dataset, is appended to these instructions. The LLM is then tasked to generate the corresponding dialogue response based on this augmented prompt, the database summary, and the concatenated dialogue history $[\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_t]$.

Implementation and Setup. Similarly as before with DST models, we employ the $\text{mT5}_{\text{small}}$ and $\text{mT5}_{\text{large}}$ models for FT-based response generation. For ICL-based experiments, we employ the GPT-3.5, LLaMA2, and OpenChat-3.5 models.

4 Automatic Evaluation

As part of DIALIGHT, we have implemented a range of automatic evaluation metrics: 1. For DST evaluation, metrics include Joint Goal Accuracy (JGA), Slot F1, Slot Recall, and Slot Precision. 2. For response generation evaluation, we utilize the BLEU score (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Banerjee

[6] openai.com/blog/openai-api

[7] LLaMA.cpp is open-sourced and accessible at github.com/ggerganov/llama.cpp.

[8] All the other experimental details are in Appendix C.

[9] We use 90% as the ratio with the fuzzywuzzy package for matching.

[10] In the case of delexicalised dialogues, all the slot values in the context and responses are replaced a predefined placeholder (e.g. *[value_name] is an [value_price] [value_food] restaurant on the [value_area] . do you need to know more ?*).

Language	Dialogue State Tracking				Response Generation			End-to-end Modelling		
	JGA	Slot F1	Slot Precision	Slot Recall	BLEU	ROUGE	METEOR	Inform Rate	Success Rate	BLEU
FT-mT5$_{small}$										
ENG	56.4	83.4	83.7	83.1	16.3	24.5	27.2	70.7	46.0	16.3
ARA	43.8	77.5	78.7	76.3	15.0	28.8	26.4	70.2	42.6	15.0
FRA	47.0	79.6	79.5	79.8	14.3	25.5	26.3	71.5	41.5	14.1
TUR	48.9	80.7	78.8	82.6	21.0	33.7	33.3	77.7	48.0	21.4
AVG.	49.0	80.3	80.2	80.5	16.6	28.1	28.3	72.4	44.4	16.7
ICL-GPT-3.5 (∗)										
ENG	13.5	37.6	26.3	65.8	1.8	12.2	11.4	33.0	16.0	1.8
ARA	7.6	31.9	22.6	60.2	0.1	2.2	1.5	36.0	18.0	0.1
FRA	12.5	39.4	29.0	61.2	0.9	8.7	7.2	40.0	24.0	0.8
TUR	7.7	34.5	24.7	57.4	0.8	6.1	5.0	32.0	13.0	0.8
AVG.	10.3	35.9	25.7	61.2	1.2	7.3	6.3	35.3	17.8	0.9

Table 2: Performance across fully supervised variants of DST models, RG models, and E2E systems on the MULTI3WOZ dataset. This table reports the performance metrics for each language, evaluated across different models. 'AVG.' represents the mean average of the evaluation scores aggregated across all four languages. We note that for these metrics the ground truth score is set at 100, with the exception of the Inform Rate and Success Rate, which are measured as 89.3 ± 0.2 and 68.6 ± 0.2 across the four languages, respectively. $(*)$ For practical considerations, the evaluation of ChatGPT-3.5-based models and systems is limited to a randomly selected sample of 100 dialogues from the full test set, due to the significant time and resource requirements of full-scale evaluation.

and Lavie, 2005). 3. For evaluating the overall systems, we report Inform Rate, Success Rate, and BLEU. Additionally, we provide an interface to facilitate easy extension for future additional metrics.

Table 2 presents the performance of implemented systems across various languages and backbone models, evaluated using the aforementioned automatic metrics.[11] The main results indicate a performance advantage of FT-based systems over their ICL-based counterparts. Specifically, the ICL-GPT-3.5-based TOD system demonstrates inferior performance when compared to the FT-based systems.[12] Subsequently, we present a detailed analysis aimed at identifying the root causes of this performance discrepancy. Initially, we observe that a substantial portion, 42.7%, of the system predictions generated by the English ICL-based DST model do not adhere to the prescribed dialogue state format specified by the instruction and ontology. Furthermore, with the given instructions, the ICL-GPT-3.5 system generates delexicalised English system responses that recall only 3.6% of the placeholders found in the ground-truth utterances. Concerning other languages, even when explicitly instructed to generate responses in the target language, the GPT-3.5 model produces utterances thare are only 18.4% in Arabic, 78.0% in French, and 70.5% in Turkish.[13] We hypothesise that this is attributed to the fact that our prompt is provided in English, and future work should experiment with additional and more sophisticated prompt designs (Shaham et al., 2024).

5 Human Evaluation

In DIALIGHT, we offer an open-sourced human evaluation tool specifically designed for TOD systems. This tool is comprehensive, providing all essential functionalities to conducting human evaluation experiments in a production setting. In the following, we detail the features supported by our web interface, provide an overview of the high-level design of our backend servers, and present a case study of human evaluation experiments.

5.1 Web Interface

The web interface of the tool offers a range of features, including user registration, account login, consent acquisition for data collection, and the execution of human evaluation tasks. In this section, we highlight two critical features that distinguish our tool from existing work.

Fine-Grained User Feedback. As shown in Figure 7 in the Appendix, our web interface is designed to support the collection of user feedback and scores at both the (local) utterance and (global) dialogue levels. Furthermore, all evaluation questions can be fully customised with minimal programming effort.[14]

Secure Authentication and Access Management. A high priority has been placed on data security

[11] We show the performance of the FT-mT5$_{large}$, ICL-LLaMA2, and ICL-OpenChat-3.5 systems in Table 6 in the Appendix. Additionally, Table 7 in the Appendix presents the evaluation results of FT-based systems on the same selected subset of 100 dialogues for consistency in comparison.

[12] It is worth noting that our system's simplicity, as opposed to the more complex system proposed by Hudeček and Dusek (2023), leads to its lower absolute performance. Our system is primarily designed to serve as a baseline for future research.

[13] Language detection was performed using the tool developed by Nakatani (2010).

[14] The tool supports all components provided by the AntDesign toolkit: `ant.design/components/overview`.

and the incorporation of authentication measures. An authentication system employing JSON Web Tokens (JWT) and a basic role management framework has been implemented. This arrangement ensures that access to specific task groups is limited to authorised users, and only task administrators are permitted to access submitted data in the database. Moreover, the web interface is integrated with an `nginx` reverse proxy, enhancing data and communication security through SSL encryption.

5.2 Back-End Servers

Our human evaluation tool is architected using a microservice design, a choice that significantly enhances its scalability and adaptability. Another standout feature of the tool is the 'one-click' deployment option, making it more accessible for users. This tool sets itself apart from its predecessors by being easy-to-use and tailored for the prevailing trend of LLMs.

Microservice with Scalability. In a microservice architecture, each service operates independently, managing a specific task or functionality. Our tool adopts this modular framework, partitioning each task model into its own independent service on a dedicated server, as depicted in Figure 3 in the Appendix. The design of these stateless services offers several benefits: firstly, it enables a single model to be concurrently shared by multiple systems, each with different configurations; secondly, it allows for the deployment of multiple instances of the same model within the same system.

On-click Deployment. The tool is designed with an 'out-of-the-box' capability, facilitated by full containerisation using docker and docker-compose (see `docs.docker.com/`) This approach ensures a simple and efficient deployment process. The versatility of these containers supports deployment in various environments, ranging from local machines to cloud-based platforms. The entire build process is governed by a central configuration file, which users can modify according to their specific requirements, thus enhancing the tool's adaptability.

5.3 A Pilot Analysis of FT-based versus ICL-based Systems in English

Relying on DIALIGHT, we conduct a human evaluation experiment comparing the performance of the English system with FT-mT5$_{\text{small}}$ and the system utilising the ICL-GPT-3.5 model. In this study, each of the 10 participants has completed 2 dialogues for each system, resulting in a total of 40 data entries. The FT-based system attains an overall score of 3.8 ± 0.9, while the ICL-based system achieves 1.6 ± 0.9. Furthermore, Figure 1 demonstrates the number of dialogues for each system based on a set of dialogue-level evaluation dimensions (Mehri and Eskenazi, 2020). The results indicate that the FT-based system outperforms the ICL-based system in terms of maintaining conversation coherence, providing consistent information, understanding the user, and delivering informative responses. Conversely, the ICL-based system generates more diverse responses and exhibits a more favourable personality.

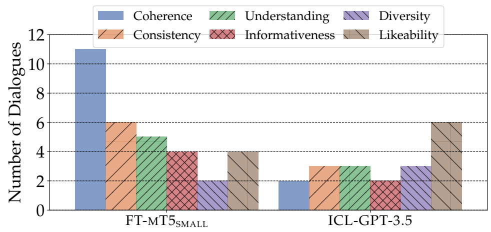


Figure 1: Number of dialogues (from the total of 20) assessed by human evaluators according to the desirable properties that align with the dialogue-level evaluation dimensions outlined in Mehri and Eskenazi (2020).

6 Conclusion and Outlook

We introduced DIALIGHT, a comprehensive toolkit for advancing multilingual TOD systems supported by language models of different families, offering the essential infrastructure for system development and evaluation. DIALIGHT supports the development of and includes baseline systems for fine-tuning and in-context learning paradigms, enabling comparative experiments within a unified setup for automatic and human evaluation. We have publicly released all our source code to facilitate future research and invite the research community to adapt and contribute to this toolkit.

Utilising this toolkit, we executed a performance analysis across several languages and models for the two modeling paradigms. The results indicate that despite their potential, LLMs, even the most powerful ones, are far from 'solving' the TOD task, especially in a multilingual context. Instead, our results open up new avenues for future research and exploration. In Appendix B, we provide some examples of future work aimed at addressing the bottlenecks identified in the baseline systems.

Ethics Statement

The experimental study obtained the full Ethics Approval from the University of Cambridge in advance of its implementation and execution. Informed consent was obtained from every individual participant involved in the study, all of whom participated voluntarily. Our models leverage two data sources: the MULTI3WOZ dataset and the pretraining data of each PLM employed in this study. Hu et al. (2023a) highlights that the creation and publication of MULTI3WOZ comply with the GDPR. Particularly, this dataset consists solely of hypothetical dialogues in which the domains and content have been restricted and predefined, minimising the risk of personal data being present. On the other hand, it is important to acknowledge that although these PLMs and LLMs are publicly available, there exists a potential risk of privacy violations (Carlini et al., 2021; Brown et al., 2022).

Acknowledgements

Songbo Hu is supported by the Cambridge International Scholarship. Anna Korhonen acknowledges the support of the UK EPSRC grant EP/T02450X/1 and the UKRI Frontier grant EP/Y031350/1. Ivan Vulić acknowledges the support of a personal Royal Society University Research Fellowship *'Inclusive and Sustainable Language Technology for a Truly Multilingual World'* (no 221137; 2022–). We are grateful to our diligent participants for their efforts and contributions in our human evaluation experiment.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. In *International Conference on Learning Representations*.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What does it mean for a language model to preserve privacy? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2280–2292, New York, NY, USA. Association for Computing Machinery.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018a. 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.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018b. 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.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4516–4525, Hong Kong, China. Association for Computational Linguistics.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *USENIX Security Symposium*, volume 6.

Willy Chung, Samuel Cahyawijaya, Bryan Wilie, Holy Lovenia, and Pascale Fung. 2023. Instructtods: Large language models for end-to-end task-oriented dialogue systems. *arXiv preprint arXiv:2310.08885*.

Bosheng Ding, Junjie Hu, Lidong Bing, Mahani Aljunied, Shafiq Joty, Luo Si, and Chunyan Miao. 2022. GlobalWoZ: Globalizing MultiWoZ to develop multilingual task-oriented dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1639–1657, Dublin, Ireland. Association for Computational Linguistics.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49, Saarbrücken, Germany. Association for Computational Linguistics.

Mihail Eric and Christopher Manning. 2017. A copy-augmented sequence-to-sequence architecture gives good performance on task-oriented dialogue. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 468–473, Valencia, Spain. Association for Computational Linguistics.

Narendra K. Gupta, Gökhan Tür, Dilek Hakkani-Tür, Srinivas Bangalore, Giuseppe Riccardi, and Mazin Gilbert. 2006. The at&t spoken language understanding system. *IEEE Transactions on Speech and Audio Processing*, 14(1):213–222.

Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, et al. 2022. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Michael Heck, Nurul Lubis, Benjamin Ruppik, Renato Vukovic, Shutong Feng, Christian Geishauser, Hsien-chin Lin, Carel van Niekerk, and Milica Gasic. 2023. ChatGPT for zero-shot dialogue state tracking: A solution or an opportunity? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 936–950, Toronto, Canada. Association for Computational Linguistics.

Songbo Hu, Han Zhou, Mete Hergul, Milan Gritta, Guchun Zhang, Ignacio Iacobacci, Ivan Vulić, and Anna Korhonen. 2023a. Multi 3 WOZ: A multilingual, multi-domain, multi-parallel dataset for training and evaluating culturally adapted task-oriented dialog systems. *Transactions of the Association for Computational Linguistics*, 11:1396–1415.

Songbo Hu, Han Zhou, Moy Yuan, Milan Gritta, Guchun Zhang, Ignacio Iacobacci, Anna Korhonen, and Ivan Vulić. 2023b. A systematic study of performance disparities in multilingual task-oriented dialogue systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6825–6851, Singapore. Association for Computational Linguistics.

Vojtěch Hudeček and Ondrej Dusek. 2023. Are large language models all you need for task-oriented dialogue? In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*, pages 216–228, Prague, Czechia. Association for Computational Linguistics.

Jessica Huynh, Ting-Rui Chiang, Jeffrey Bigham, and Maxine Eskenazi. 2022. DialCrowd 2.0: A quality-focused dialog system crowdsourcing toolkit. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1256–1263, Marseille, France. European Language Resources Association.

Kyusong Lee, Tiancheng Zhao, Alan W. Black, and Maxine Eskenazi. 2018. DialCrowd: A toolkit for easy dialog system assessment. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 245–248, Melbourne, Australia. Association for Computational Linguistics.

Sungjin Lee, Qi Zhu, Ryuichi Takanobu, Zheng Zhang, Yaoqin Zhang, Xiang Li, Jinchao Li, Baolin Peng, Xiujun Li, Minlie Huang, and Jianfeng Gao. 2019. ConvLab: Multi-domain end-to-end dialog system platform. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 64–69, Florence, Italy. Association for Computational Linguistics.

Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. 2018. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1437–1447, Melbourne, Australia. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. Bactrian-X: A multilingual replicable instruction-following model with low-rank adaptation. *CoRR*, abs/2305.15011.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 2020. MinTL: Minimalist transfer learning for task-oriented dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3391–3405, Online. Association for Computational Linguistics.

Shikib Mehri, Jinho Choi, Luis Fernando D'Haro, Jan Deriu, Maxine Eskenazi, Milica Gasic, Kallirroi Georgila, Dilek Hakkani-Tur, Zekang Li, Verena Rieser, Samira Shaikh, David Traum, Yi-Ting Yeh, Zhou Yu, Yizhe Zhang, and Chen Zhang. 2022. Report from the nsf future directions workshop on automatic evaluation of dialog: Research directions and challenges.

Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting. Association for Computational Linguistics.

Shuyo Nakatani. 2010. Language detection library for java.

Tomáš Nekvinda and Ondřej Dušek. 2021. Shades of BLEU, flavours of success: The case of MultiWOZ. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 34–46, Online. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2021. Soloist: Building task bots at scale with transfer learning and machine teaching. *Transactions of the Association for Computational Linguistics*, 9:807–824.

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. *J. Mach. Learn. Res.*, 21(1).

Evgeniia Razumovskaia, Goran Glavaš, Olga Majewska, Edoardo Maria Ponti, Anna Korhonen, and Ivan Vulić. 2022. Crossing the conversational chasm: A primer on natural language processing for multilingual task-oriented dialogue systems. *Journal of Artificial Intelligence Research*, 74:1351–1402.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. *CoRR*, abs/2401.01854.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng

He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel Stern Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, vinay uday prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Gökhan Tür, Andreas Stolcke, L. Lynn Voss, Stanley Peters, Dilek Hakkani-Tür, John Dowding, Benoît Favre, Raquel Fernández, Matthew Frampton, Michael W. Frandsen, Clint Frederickson, Martin Graciarena, Donald Kintzing, Kyle Leveque, Shane Mason, John Niekrasz, Matthew Purver, Korbinian Riedhammer, Elizabeth Shriberg, Jing Tien, Dimitra Vergyri, and Fan Yang. 2010. The CALO meeting assistant system. *IEEE Transactions on Speech Audio Processing*, 18(6):1601–1611.

Stefan Ultes, Lina M. Rojas Barahona, Pei-Hao Su, David Vandyke, Dongho Kim, Iñigo Casanueva, Paweł Budzianowski, Nikola Mrkšić, Tsung-Hsien Wen, Milica Gasic, and Steve Young. 2017. PyDial: A Multi-domain Statistical Dialogue System Toolkit. In *Proceedings of ACL 2017, System Demonstrations*, pages 73–78, Vancouver, Canada. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. *SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems*. Curran Associates Inc., Red Hook, NY, USA.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Qingyang Wu, Deema Alnuhait, Derek Chen, and Zhou Yu. 2023. Using textual interface to align external knowledge for end-to-end task-oriented dialogue systems. *arXiv preprint arXiv:2305.13710*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online. Association for Computational Linguistics.

Steve Young. 2007. Cued standard dialogue acts. *Report, Cambridge University Engineering Department, 14th October*, 2007.

Steve J. Young. 2010. Cognitive user interfaces. *IEEE Signal Processing Magazine*, 27(3):128–140.

Xiaoying Zhang, Baolin Peng, Kun Li, Jingyan Zhou, and Helen Meng. 2023. SGP-TOD: Building task bots effortlessly via schema-guided LLM prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13348–13369, Singapore. Association for Computational Linguistics.

Qi Zhu, Christian Geishauser, Hsien chin Lin, Carel van Niekerk, Baolin Peng, Zheng Zhang, Michael Heck, Nurul Lubis, Dazhen Wan, Xiaochen Zhu, Jianfeng Gao, Milica Gašić, and Minlie Huang. 2022. Convlab-3: A flexible dialogue system toolkit based on a unified data format. *arXiv preprint arXiv:2211.17148*.

Qi Zhu, Zheng Zhang, Yan Fang, Xiang Li, Ryuichi Takanobu, Jinchao Li, Baolin Peng, Jianfeng Gao, Xiaoyan Zhu, and Minlie Huang. 2020. ConvLab-2: An open-source toolkit for building, evaluating, and diagnosing dialogue systems. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 142–149, Online. Association for Computational Linguistics.

A Limitations

There are several limitations in this work, primarily stemming from the scope, design, and intended purpose of this toolkit. In this work, we only provide a set of baseline systems. As shown in Figure 9, these systems are intentionally kept as simple as possible while being fully functional, with the goal of enabling users to gain a conceptual understanding of the TOD task and implement their own systems with minimal learning effort. It is important to acknowledge that our systems may underperform compared to other more sophisticated systems, such as the one developed by Hudeček and Dusek (2023) that employs advanced techniques for retrieving positive and negative ICL examples, or the system proposed by Zhang et al. (2023) that incorporates an explicit and pre-defined task schema to guide system actions. Instead of pursuing state-of-the-art performance, we place greater emphasis on providing the essential environment and a set of tools, including automatic and human evaluation tools, to enable researchers to develop more advanced systems using our toolkit in future work.

Currently, DIALIGHT currently supports only TOD datasets that are derived from the MultiWOZ dataset (Budzianowski et al., 2018b) and its schema, that is, the ones annotated with the CUED schema (Young, 2007). We recognise the additional challenges associated with extending the toolkit to accommodate other datasets with different annotation schemata. Such extensions would typically involve the re-implementation of the data loader, the database, and some automatic evaluation metrics, such as Inform Rate and Success Rate. However, we believe that our human evaluation tool can be easily extended to evaluate systems developed on other datasets.

Furthermore, it is worth noting that a fully inclusive dialogue system should consider not only text input but also other modalities, such as spoken and sign languages. We acknowledge that DIALIGHT currently focuses on text input only, and we hope to integrate the support for speech input and output as part of future work.

B Future Work: Some Ideas

In this section, we provide examples of future work aimed at addressing the bottlenecks identified in the baseline systems.

Modernising TOD Systems for LLMs. In §4, our experimental results reveal that solely prompting LLMs resulted in system failure to comply with the instructions and predict dialogue states in the required format, leading to over 40% empty predictions. Additionally, ICL-based systems employing LLMs encountered challenges in recalling placeholders and generating delexicalised utterances in a similar fashion. This issue is arguably attributable to the misalignment between the task requirements for TOD and the inherent pretraining of LLMs. We propose that future research should critically reevaluate the current design choices of TOD systems to better tailor LLM-based TOD systems.

Multilingual Generation with LLMs. Our analysis demonstrates that when prompts are solely provided in English and LLMs are instructed to generate outputs in other languages, they often encounter difficulties in complying, resulting in the generation of outputs in English. Developing complex NLP applications like TOD systems requires a significant number of instructions to specify task requirements. However, the dominance of English instructions tends to bias model outputs towards English. Conversely, tailoring task instructions for each individual language, especially for resource-lean languages, poses a challenge. While some work has been recently conducted for other NLP tasks (Li et al., 2023), arguably less complex than TOD, the question of how to effectively control LLMs to generate target language outputs remains an open question for future research.

C Experimental Details

In this section, we describe the experimental setups for the systems developed in this paper. For specific implementation details, including the prompts used in the ICL-based systems, we direct readers to the actual implementation and documentation of DIALIGHT.

Table 3 presents the selected hyper-parameters for the conducted experimental study. All the FT-based experiments were run on a single A100 80 GiB GPU and a 32-core vCPU. Notably, the ICL-based systems deployed in our experimental study excluded training examples from the prompts. This decision was based on empirical evidence indicating that these examples adversely affect system performance. When tested with 10 ICL examples, the ICL-GPT-3.5 systems recorded a JGA of 4.3 (↓9.2), an Inform Rate of 31.0 (↓2.0), and a Success Rate of 14.0 (↓2.0). We did not conduct a

Hyper-parameter	Value
FT-mT5$_{small}$	
batch size	32
learning rate	1e-3
weight decay	0.01
evaluation per steps	5000
max training steps	50000
context window	10
early stopping patience	2
maximum generation length	512
FT-mT5$_{large}$	
batch size	8
learning rate	1e-3
weight decay	0.01
evaluation per steps	5000
max training steps	50000
context window	10
early stopping patience	2
maximum generation length	512
ICL-GPT-3.5, **ICL-LLaMA2**, and **ICL-OpenChat-3.5**	
context window	10
number of ICL examples	0 ($*$)

Table 3: The hyperparameters for E2E systems and their constituent models. Both the DST and RG models, which are based on the same PLM, utilised identical hyper-parameter setups. To select the optimal model checkpoint, we employ early stopping and select the one with the best validation performance, measured by JGA for DST and BLEU score for RG. Unless explicitly specified, all other hyper-parameters are set to their default values as defined in the HuggingFace Transformers. ($*$) Notably, our observations suggest that the introduction of training examples actually adversely affects model performance.

hyperparameter search for the number of ICL examples due to the high costs associated with such an experiment.

Table 4 lists all the language models we used in this work, along with their respective checkpoints in the Huggingface repository and the OpenAI API. Both the LLaMA2 and OpenChat-3.5 models employed in this study have 7 billion parameters.

Table 5 shows the time consumption of the models for the E2E task in the experimental study.

D Additional Results on MULTI3WOZ

In this section, we show supplementary experimental results to solidify the empirical findings in §4. Firstly, the evaluation metrics for E2E tasks, such as Inform Rate and Success Rate, are influenced by two key outputs of the system: the dialogue state and the generated response. To assess the impact of each component on overall system performance, we conduct an extra experiment where the predictions of each part were individually replaced with the ground-truth. When substituting the predicted utterances with ground-truth utterances, the FT-mT5$_{small}$ systems exhibited a notable improvement, achieving an average Inform Rate of 85.1 (↑13.3) and a Success Rate of 66.1 (↑21.5) across four languages. In contrast, substituting the predicted dialogue states resulted in a marginal increase, with the systems attaining an Inform Rate of 72.1 (↑0.3) and a Success Rate of 43.3 (↑0.7) across the languages. These findings highlight the critical role of RG model performance in determining overall

Model	Checkpoint
mT5$_{small}$	google/mt5-small
mT5$_{large}$	google/mt5-large
GPT-3.5	gpt-3.5-turbo-1106
LLaMA2	TheBloke/Llama-2-7B-GGUF
OpenChat-3.5	openchat/openchat_3.5

Table 4: The employed langauge models in our experimental study and their Huggingface or OpenAI Checkpoints. We use 7B variants of LLaMA2 and OpenChat-3.5.

Setup	Time Consumption
FT-mT5$_{small}$	
DST training per 500 steps	3:02
RG training per 500 steps	2:27
Inference on *full test*	9:30
FT-mT5$_{large}$	
DST training per 500 steps	5:09
RG training per 500 steps	5:02
Inference on *full test*	1:16:00
ICL-GPT-3.5	
Inference on *10 dialogues*	9:05
ICL-LLaMA2	
Inference on *10 dialogues*	36:00
ICL-OpenChat-3.5	
Inference on *10 dialogues*	5:35

Table 5: The average time consumption for the E2E task. For all FT-based systems, the computation was performed on a machine equipped with a single A100 80 GiB GPU and a 32-core vCPU. In the case of the ICL-GPT-3.5 systems, calculations were conducted using the OpenAI API. Meanwhile, the ICL-LLaMA2 systems were executed on an Intel 13900k CPU and ICL-OpenChat-3.5 systems were executed on a machine with an Intel 13900k CPU and a single NVIDIA RTX 4090 GPU.

system performance and the relative insensitivity of these metrics in evaluating the performance of DST models.

Table 6 shows the fully supervised performance of $\text{mT5}_{\text{large}}$, LLaMA2, and OpenChat-3.5 models across DST models, RG models, and E2E systems on the MULTI3WOZ dataset. It is noteworthy that the OpenChat-3.5 based systems exhibited a failure in generating coherent dialogue responses. These systems consistently repeated the prompts and, in each utterance, indiscriminately included all the placeholders. This simplistic method led to inflated Inform Rate and Success Rate scores, highlighting the potential vulnerability of these metrics to adversarial strategies, a concern also highlighted by Wu et al. (2023).
Table 7 presents an analysis of the performance under full supervision for both $\text{mT5}_{\text{small}}$ and $\text{mT5}_{\text{large}}$ models. This evaluation is conducted on a specifically selected subset of 100 dialogues from the entire test set, consistent with the evaluation setup applied to all other ICL-based models and systems in this study. This approach ensures a rigorous and direct comparability across the discussed ICL-based models and systems.

E Diagrams and Screenshots

In this section, we present a series of diagrams, web interface screenshots, and code snippets that illustrate the architectural design and functionalities of our toolkit.

Figure 2 illustrates the architecture of an E2E dialogue system. State-of-the-art E2E TOD systems typically employ a pipelined approach in the background, incorporating three key components: a dialogue state tracking (DST) model, a database interface, and a response generation (RG) model.

Figure 3 shows the architectural framework of our human evaluation tool. The underlying infrastructure of the back-end servers is constructed on the principles of a microservice architecture.

Figure 4 presents a screenshot capturing the login page of our human evaluation web interface. This interface serves as the entry point for evaluators to access the system.

Figure 5 presents a screenshot capturing the registration page of our human evaluation web interface.

Figure 6 displays a screenshot of the assignment page within the human evaluation web interface.

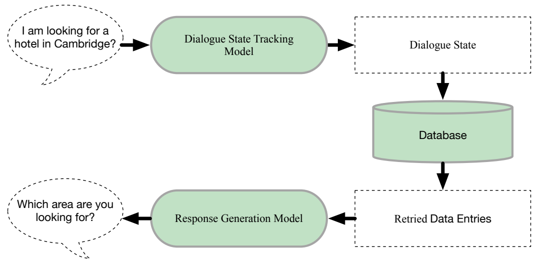


Figure 2: An E2E dialogue system contains three key components: a dialogue state tracking (DST) model, a database, and a response generation (RG) model. The DST model processes user utterances with the accumulated dialogue history to predict a dialogue state. This state is then translated into a database query to extract data entries relevant to the current dialogue context from the database. The RG model uses these entries and the dialogue history to produce the final response.

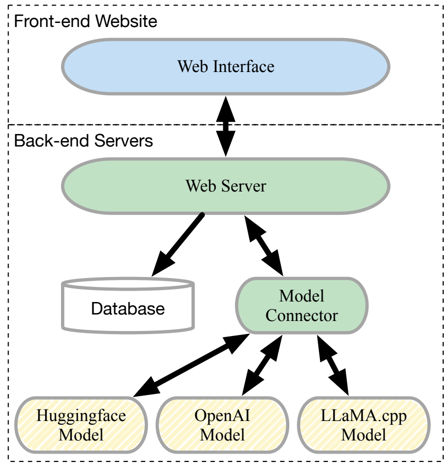


Figure 3: Architectural design of our human evaluation tool, containing two primary components: a web-based interface and a cluster of back-end servers. The server infrastructure is based on a microservice architecture, segregating each task model into its own independent service, as highlighted by the dashed yellow background in the figure. Central to this architecture is the Model Connector, functioning as an API gateway to manage and route requests to the appropriate task models. For example, each fine-tuned DST model is hosted independently on a dedicated server. These servers are designed to be stateless, enabling their shared use across various systems and dialogue sessions, thereby enhancing efficiency and scalability.

This page is specifically designed for users to carry out the task of evaluating the dialogue system.

Figure 7 presents detailed screenshots capturing both the utterance-level and dialogue-level feedback forms facilitated by our tool. These feedback forms are integrated into the assignment page.

Figure 8 presents screenshots of the code to pro-

Language	Dialogue State Tracking				Response Generation			End-to-end Modelling		
	JGA	Slot F1	Slot Precision	Slot Recall	BLEU	ROUGE	METEOR	Inform Rate	Success Rate	BLEU
					FT-mT5$_{large}$					
ENG	18.6	52.5	53.0	52.0	15.8	24.1	27.1	70.1	47.3	15.4
ARA	44.0	78.5	78.3	78.6	7.0	17.4	14.8	67.3	31.4	6.7
FRA	46.8	79.5	79.6	79.5	14.1	25.2	25.8	74.3	44.2	13.6
TUR	48.5	80.2	78.8	81.7	10.6	19.7	19.0	77.7	48.0	10.8
AVG.	39.5	72.7	72.4	73.0	11.9	21.6	21.7	70.1	42.4	11.6
					ICL-LLaMA2 $(*)$					
ENG	1.6	0.0	0.0	0.0	0.1	0.1	0.1	22.0	8.0	0.2
ARA	1.6	0.0	0.0	0.0	0.1	0.0	0.0	23.0	9.0	0.1
FRA	1.6	0.0	0.0	0.0	0.0	0.0	0.0	19.0	9.0	0.0
TUR	1.6	0.0	0.0	0.0	0.0	0.0	0.0	18.0	7.0	0.0
AVG.	1.6	0.0	0.0	0.0	0.1	0.0	0.0	20.5	8.3	0.0
					ICL-OpenChat-3.5 $(*)$ $(**)$					
ENG	1.6	0.0	0.0	0.0	0.0	0.0	0.0	67.0	61.0	0.0
ARA	1.6	0.0	0.0	0.0	0.0	0.0	0.0	67.0	60.0	0.0
FRA	1.6	0.0	0.0	0.0	0.0	0.0	0.0	67.0	60.0	0.0
TUR	1.6	0.0	0.0	0.0	0.0	0.0	0.0	67.0	60.0	0.0
AVG.	1.6	0.0	0.0	0.0	0.0	0.0	0.0	67.0	60.3	0.0

Table 6: Evaluation of fully supervised performance across DST models, RG models, and E2E systems on the MULTI3WOZ dataset. This table reports the performance metrics for each language, evaluated across different models. It should be noted that for these metrics, the ground truth score is set at 100, with the exceptions of the Inform Rate and Success Rate, which are measured as 89.3 ± 0.2 and 68.6 ± 0.2 across the four languages, respectively. $(*)$ For practical considerations, the evaluation ICL-based models and systems is limited to a randomly selected sample of 100 dialogues from the full test set, due to the significant time and resource requirements of a full-scale evaluation. $(**)$ Additionally, it is noteworthy that the OpenChat-3.5 based systems exhibited a failure in generating coherent dialogue responses. These systems repetitively echoed the prompts and for each utterance, generated all the placeholders. This simplistic approach resulted in artificially high Inform Rate and Success Rate scores, revealing the vulnerability of these metrics to adversarial strategies.

Language	Dialogue State Tracking				Response Generation			End-to-end Modelling		
	JGA	Slot F1	Slot Precision	Slot Recall	BLEU	ROUGE	METEOR	Inform Rate	Success Rate	BLEU
					FT-mT5$_{small}$$(*)$					
ENG	54.4	83.0	84.6	81.5	18.7	26.5	29.7	66.0	47.0	18.5
ARA	41.9	77.5	79.9	75.3	16.1	30.0	27.4	67.0	43.0	15.8
FRA	42.9	80.3	81.3	79.3	12.9	25.0	25.6	66.0	34.0	13.4
TUR	48.0	81.5	81.1	81.9	22.8	34.6	34.1	75.0	48.0	23.1
AVG.	46.8	80.6	81.7	79.5	17.6	29.0	29.2	68.5	43.0	17.7
					FT-mT5$_{large}$$(*)$					
ENG	19.1	53.3	55.0	51.6	18.0	25.7	29.0	69.0	49.0	17.8
ARA	41.9	79.1	80.2	78.1	9.2	19.4	17.0	62.0	30.0	9.0
FRA	44.6	80.3	82.1	78.5	13.1	24.4	24.6	77.0	50.0	13.4
TUR	45.2	80.8	80.8	80.8	11.2	20.5	19.7	68.0	28.0	11.4
AVG.	30.2	73.4	74.5	72.3	12.9	22.5	22.6	69.0	39.3	12.9

Table 7: Evaluation of fully supervised performance across DST models, RG models, and E2E systems on the MULTI3WOZ dataset. This table reports the performance metrics for each language, using both mT5$_{small}$ and mT5$_{large}$ models. 'AVG.' represents the mean average of the evaluation scores aggregated across all four languages. We note that for these metrics the ground truth score is set at 100, with the exception of the Inform Rate and Success Rate, which are measured as 89.3 ± 0.2 and 68.6 ± 0.2 across the four languages, respectively. $(*)$ In this table, the evaluation is limited to a randomly selected sample of 100 dialogues from the full test set, ensuring direct comparability with other ICL-based models and systems discussed herein.

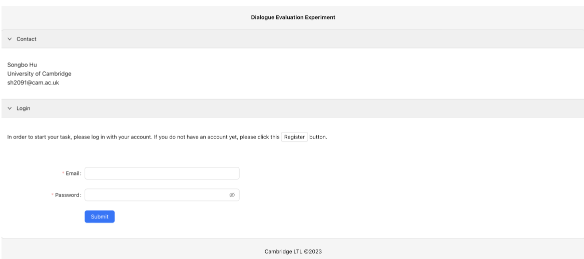


Figure 4: A screenshot capturing the login page of the human evaluation web interface.

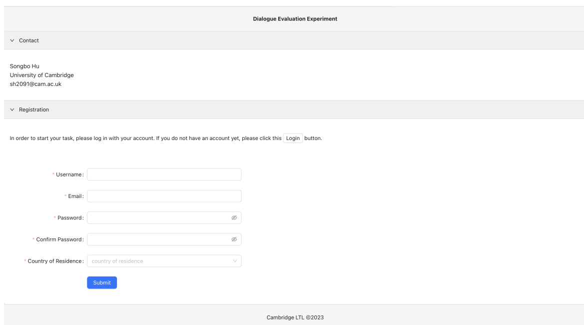


Figure 5: A screenshot of the registration page.

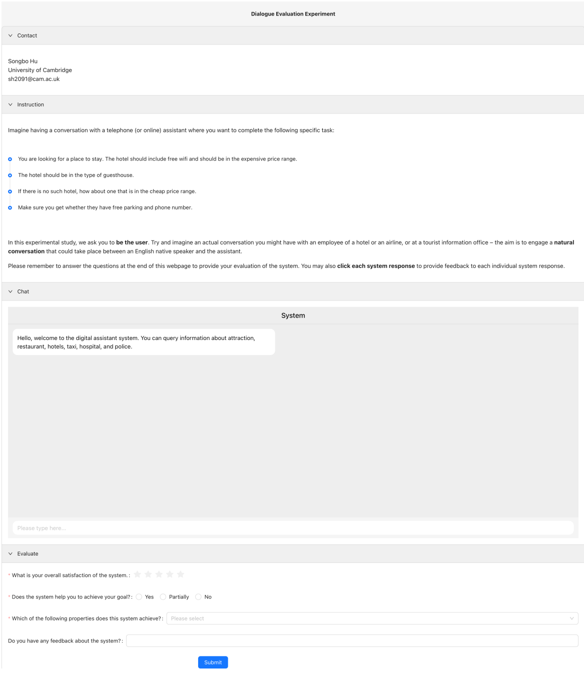
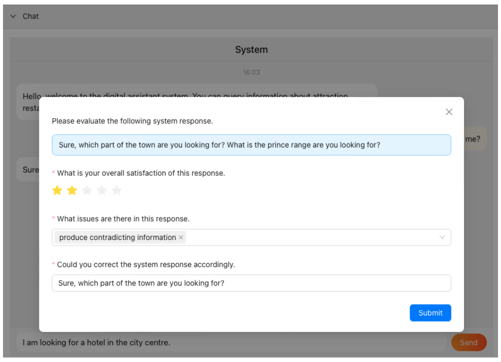
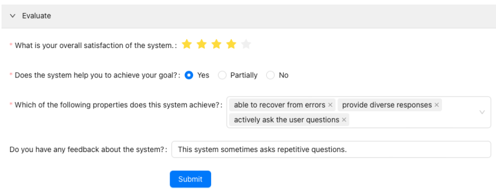


Figure 6: A screenshot of the task assignment page.



(a) Utterance Level



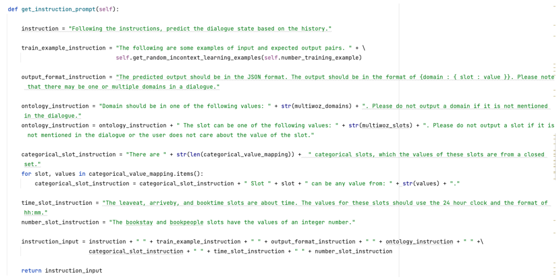
(b) Dialogue Level

Figure 7: Our human evaluation tool is designed to collect user feedback at both the (a) utterance and (b) dialogue levels. This tool allows for full customisation of evaluation questions with minimal programming effort.

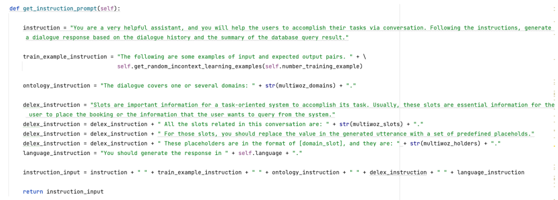
duce the prompts for both our ICL-based DST and RG models.

Figure 9 presents a screenshot of the inference code for our FT-based E2E systems. The code is modularised and intentionally designed to be both simple and fully functional. Our goal is to facilitate users in acquiring a clear understanding of the TOD task, as well as to provide insights into the implementation of our system.

Figure 10 presents a screenshot of the backend web server code for our RESTful API for the storage of human evaluation results. This setup incorporates a JWT-based authentication system to secure access. Additionally, it is structured to permit only authorised users with specific permissions to record evaluation results in the database.



(a) DST Prompt



(b) RG Prompt

Figure 8: Screenshots of the code to produce the prompts for our ICL-based (a) DST and (b) RG models are provided above.

```
# Step 1): DST.
dst_prediction = get_dst_prediction(dst_config, ["test"], current_language)

# Step 2): DB query.
db_result = get_db_summery(dst_config, dst_prediction, current_language)

# Step 3): RG.
response_prediction = get_response_prediction(rg_config, db_result, dst_prediction, current_language)
```

Figure 9: A screenshot of the inference code for our FT-based E2E systems is provided above. The code is intentionally designed to be both simple and fully functional, aiming to assist users in gaining a conceptual understanding of the TOD task and the implementation of our system.

```python
@auth.route("/save_result", methods=['POST', "OPTIONS"])
@jwt_required()
@require_permission('submit_task')
@require_additional_permission_with_access_control("submit_task_with_access_control")
def save_result():
    if request.method == 'OPTIONS':
        return _handle_preflight()
    current_user_id = get_jwt_identity()

    try:
        _currentUser = LoginUser().check_user_with_id(current_user_id)
        save_data = request.get_json()
        logging.info(save_data)
        if save_data:
            save_data['create_time'] = datetime.now()
            save_data['feedback_user'] = _currentUser.email
            _feedback = FeedBack(**save_data)
            insert_id = _currentUser.save_result(_feedback)

            return jsonify({'success': True, 'msg': 'Save success.', "insert_id": insert_id}), 200

    except Exception as e:
        logging.error('Error while saving task data', exc_info=True)
        return jsonify({'success': False, 'msg': 'Failed to save task data'}), 500
```

Figure 10: A screenshot of the backend web server code for our RESTful API, designed for storing system evaluation results. It features a JWT-based authentication mechanism to ensure secure access. Furthermore, the system is configured to allow only users with specific permissions to save evaluation results to the database, thereby enhancing data integrity and security.