MIDAS: Multi-level Intent, Domain, And Slot Knowledge Distillation for Multi-turn NLU

Yan Li[1,3]**, So-Eon Kim**[2]**, Seong-Bae Park**[2]**, Soyeon Caren Han**[1,3,*]**,**
[1]The University of Sydney, Sydney, [2]Kyung Hee University,
[3]The University of Melbourne
[1]yali3816@uni.sydney.edu.au, [2]{sekim0211, sbpark71}@khu.ac.kr,
[3]caren.han@unimelb.edu.au

Abstract

Although Large Language Models (LLMs) can generate coherent text, they often struggle to recognise user intent behind queries. In contrast, Natural Language Understanding (NLU) models interpret the purpose and key information of user input for responsive interactions. Existing NLU models typically map utterances to a dual-level semantic frame, involving sentence-level intent (SI) and word-level slot (WS) labels. However, real-life conversations primarily consist of multi-turn dialogues, requiring the interpretation of complex and extended exchanges. Researchers encounter challenges in addressing all facets of multi-turn dialogue using a unified NLU model. This paper introduces MIDAS, a novel approach leveraging multi-level intent, domain, and slot knowledge distillation for multi-turn NLU. We construct distinct teachers for SI detection, WS filling, and conversation-level domain (CD) classification, each fine-tuned for specific knowledge. A multi-teacher loss is proposed to facilitate the integration of these teachers, guiding a student model in multi-turn dialogue tasks. Results demonstrate the efficacy of our model in improving multi-turn conversation understanding, showcasing the potential for advancements in NLU through multi-level dialogue knowledge distillation. Our implementation is open-sourced on GitHub[1].

1 Introduction

Natural Language Understanding (NLU) within the realm of Natural Language Processing (NLP) explores the mechanisms through which computers comprehend human language. Developing a hierarchical semantic framework encompassing domain, intent, and slot has become pivotal in representing the meaning embedded in natural language (Weld et al., 2022). We present a conversation example

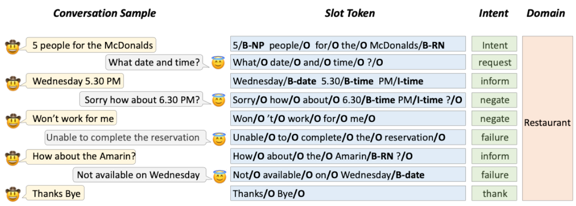

Figure 1: An example of conversations with WSs, SIs, and CD annotation from M2M. B-NP (B-Number of People), B-RN (B-Restaurant Name), O (Others).

that shows the way of annotation for WSs, SI, and CD from the M2M dataset in Figure 1. The dialogue consists of a total of 9 turns, and each turn includes WS tokens and SI information, and the dialogue corresponds to one domain, 'restaurant'.

Large Language Models (LLM) have received much attention in generating human-like text based on user prompts. However, they are still limited when it comes to deeper communication and diverse key information[2]. Hence, we investigate how to improve the state-of-the-art existing NLU techniques. While existing NLU literature predominantly concentrates on single-turn utterances within a single domain, recent advancements in multi-turn datasets have paved the way for annotations at the dialogue level, spanning across diverse domains. Interpreting more extended and intricate conversations with multiple turns necessitates understanding the ongoing context and retaining previously gathered information. Traditional NLU involves mapping single utterances to a dual-level semantic structure, encompassing SI and WS labels. With real-life conversations extending across multiple turns, there is an evident demand for research incorporating dialogue history, as demonstrated by improved performance through dialogue context. The challenge extends beyond dual-level understanding to encompass a three-level comprehen-

*Corresponding Author

[1]https://github.com/adlnlp/Midas

[2]We tested NLU benchmarks with several LLMs, including LLaMa2, LLama3.1, Gemma, QWen2, GPT3.5, and GPT4o visualised in Appendix D.2.1 and D.2.2.

Findings of the Association for Computational Linguistics: NAACL 2025, pages 8004–8027
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

Model	Year	Word (Slot)	Sentence (Intent)	Document (Domain)	Dialogue Type	Joint Integration
SeqSeq Liu and Lane (2016)	2016	○	○	×	Single-Turn	BiRNN + Attention
SDEN Bapna et al. (2017)	2017	○	○	○	Multi-Turn	BiRNN + Memory Network
Slot-Gated Goo et al. (2018)	2018	○	○	×	Single-Turn	BiLSTM + Slot Gate
BLSTM+attention Tingting et al. (2019)	2019	○	○	×	Single-Turn	BiLSTM + Attention
STD Jiang et al. (2021)	2021	○	○	×	Single-Turn	Transformer + One-teacher KD
SDJN Chen et al. (2022b)	2022	○	○	×	Single-Turn	BiLSTM + self KD
XAI Attention Gunaratna et al. (2022)	2022	○	○	×	Multi-Turn	eXplainable AI
Tri-level JNLU Weld et al. (2023)	2023	○	○	○	Multi-Turn	Cross Transformer
PACL Chen et al. (2024)	2024	○	○	×	Multi-Turn	Contrastive Learning + Attention
BiJM Luo and Feng (2024)	2024	○	○	×	Single-Turn	Transformer + Enhance Layer
Ours	2024	○	○	○	Multi-Turn	Multi-teacher KD

Table 1: Summary of existing joint NLU models and ours. Word, Sentence, and Document columns indicate whether the relevant information is used for joint integration. KD refers to knowledge distillation. The complete set of summary tables is detailed in Appendix A.

sion: SI, WS, and CD classification. However, researchers encounter challenges in handling all aspects of multi-turn dialogue conversations through a single unified NLU model, due to computational complexity and a lack of distillability of multi-level knowledge.

This paper introduces a novel multi-level multi-teacher knowledge distillation model to enhance NLU understanding in multi-turn dialogues, leveraging diverse levels of knowledge embedded in these datasets. Notably, our model is the pioneering approach in multi-teacher knowledge distillation, catering to distinct facets of knowledge within a dialogue. To achieve this, our approach involves the construction of teachers at different levels, specifically focusing on SI detection, WS filling, and CD classification. We fine-tune these multi-level teachers to acquire the relevant knowledge and combine these to educate the student model in dialogue tasks facilitated by novel multi-level teacher loss functions. There are two major contributions:
1) We introduce a novel multi-level, multi-teacher knowledge distillation model to enhance multi-turn NLU. It outperforms widely-used multi-NLU datasets, producing superior performance in all intent detection, slot filling, and domain classification, even compared with the LLMs.
2) We introduce multi-level teacher loss functions, shedding light on their impact within the multi-teacher knowledge distillation and guiding a student model.

2 Related works

There is a large body of NLU modelling literature, and we briefly introduce the joint NLU models and knowledge distillation models. A summary of these models and our model are in Table 1.

Natural Language Understanding Early works addressed slot filling and intent detection separately. Current research commonly employs joint models with transfer learning (Rongali et al., 2021; M'hamdi et al., 2021), where fine-tuning language models (Dao et al., 2021; Abro et al., 2022; Mei et al., 2023) enhances generalisation by leveraging high-quality representations. Typically, intent is classified through the [CLS] token and slots through individual token embeddings (Chen et al., 2019; Han et al., 2021a; Heo et al., 2022; Luo and Feng, 2024). Another transfer learning strategy is knowledge distillation, where a smaller student model learns from a larger teacher model, often using self-distillation (Chen et al., 2022b; Cheng et al., 2023). However, these methods primarily address single-turn dialogues or use only one teacher model. Multi-turn dialogues benefit from encoding dialogue history, leading to performance gains (Bapna et al., 2017; Weld et al., 2023; Wu and Juang, 2023; Tu et al., 2023; Chen et al., 2024). Our model is the first to employ multi-teacher knowledge distillation for multi-turn NLU. Distinct teachers specialise in intent classification, slot filling, and domain classification, thus effectively distilling multi-level knowledge. To the best of our knowledge, no previous work has explored multi-teacher distillation in multi-turn dialogue NLU.

Knowledge Distillation Knowledge Distillation (KD) defines a framework where a well-trained teacher network guides the training of a student network for various tasks. Traditional KD employs a single teacher to train one student model (Hinton et al., 2014). Multi-teacher KD, inspired by ensemble learning, integrates knowledge from multiple teachers to enhance student model performance (Wu et al., 2021a; Yuan et al., 2021; Jung et al., 2023; Wang et al., 2021b; Huang et al., 2023; Amirkhani et al., 2021; Mirzadeh et al., 2020; Son et al., 2021). Typically, multiple teachers focus on the same domain, regardless of architecture. Recent approaches involve each teacher specialising in different domains and imparting domain-

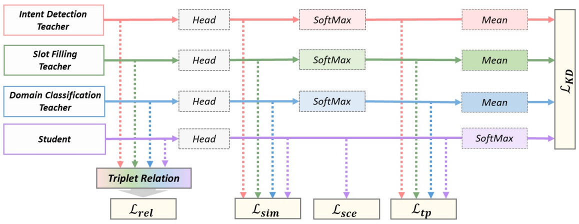

Figure 2: The proposed multi-level teacher knowledge distillation framework for the multi-turn NLU task. Note that we applied three multi-level teachers: Intent Detection, Slot Filling, and Domain Classification. In this framework, we conduct diverse Loss objectives, including $\mathcal{L}_{rel}$, $\mathcal{L}_{sim}$, $\mathcal{L}_{sce}$, $\mathcal{L}_{tp}$ and $\mathcal{L}_{KD}$, which represent relation loss, similarity loss, student cross-entropy loss, teacher prediction supervise loss, and Kullback-Leibler Divergence loss, respectively.

specific knowledge to the student (Pan et al., 2021; Ji et al., 2023). Cross-modality methods have been explored: either the teacher and student learn different modalities (Kong et al., 2019; Ni et al., 2022), or the teacher learns multiple modalities, and the student receives all modalities (Jin et al., 2021).

3 MIDAS

We propose a new multi-level dialogue teacher knowledge distillation framework, MIDAS, that trains the student model S with multi-level teachers to enhance the NLU capabilities. We have three multi-level dialogue knowledge teachers, including intent detection, slot filling and domain topic classification. To achieve this, we initially construct teachers with distinct levels of dialogue knowledge, denoted as T=$\{T_{ID}, T_{SF}, T_{DC}\}$, where T is the set of teacher models, and ID, SF, and DC correspond to Intent Detection, Slot Filling, Domain Classification. Then, we fine-tune the teacher models T to acquire knowledge from each task. On top of simple KD losses, we also introduce two novel loss functions, relation loss $\mathcal{L}_{rel}$ and teacher prediction supervised loss $\mathcal{L}_{tp}$, specifically designed for this domain. These facilitate knowledge transfer from multi-level teachers to the student model. The overall architecture is illustrated in Figure 2.

3.1 Multi-Level teacher construction

We first construct the teachers of different dialogue document component understanding levels, including WS, SI, and CD knowledge. The inputs for all teachers consist of utterances from each turn in dialogue datasets, denoted by X^i=$x_1^i, x_2^i, ..., x_l^i$, where X^i represents the i_{th} utterance in the entire dataset, l is the length of the utterance, and x_l^i signifies a word in the utterance.

1) Word-level teacher T_{SF} predicts the slot type for each word, providing knowledge to the student model about key slots in the dialogue. The output of T_{SF} is $\hat{Y}_{SF}^i$=$\hat{Y}_{SF,1}^i, \hat{Y}_{SF,2}^i, ..., \hat{Y}_{SF,l}^i$, representing the predicted slot types for each word, where $\hat{Y}_{SF,l}^i$∈$0, 1, ..., k_{SF}-1$, and k_{SF} is the number of slot types.

2) Sentence-level teacher T_{ID} predicts the intent of the utterance, aiding the student model in comprehending the overall intent of each turn. The prediction of T_{ID} is symbolised as $\hat{Y}_{ID}^i$, where $\hat{Y}_{ID}^i$∈ $0, 1, ..., k_{ID}-1$, and k_{ID} represents the number of intents in the dataset.

3) Conversation Document-level teacher T_{DC} forecasts the dialogue's domain, providing knowledge to classify it and understand its background knowledge. The prediction of T_{DC} is indicated as $\hat{Y}_{DC}^i$, where $\hat{Y}_{DC}^i$∈$0, 1, ..., k_{DC}-1$, and k_{DC} denotes the number of domains in the dataset.

Using these three levels of teachers, our objective is to instruct the student model to comprehend dialogues from multiple perspectives, incorporating WS, SI, and CD background knowledge. By doing so, we enhance the student model's grasp of dialogues across various levels. There are two primary reasons for utilising multi-level dialogue knowledge teachers to train a student. First, individually deploying a pre-trained model for each task consumes more computational resources, and some machines may not support running multiple pre-trained models. Instead, the knowledge distillation process leads to more robust models and is resistant to adversarial attacks. Incorporating soft targets from the teacher model can help the student model learn smoother decision boundaries. Secondly, we posit that diverse levels of knowledge derived from multi-turn conversation understanding datasets can enhance the comprehension of each specific natural language understanding task, surpassing the benefits of learning from single-level dialogue knowledge. Note that we use pre-trained models as the foundational structure for our teachers. After experimenting with various backbones, we determined that BERT yields one of the best results overall, as detailed in Section 5.2. These pre-trained models undergo fine-tuning using specific data for each level, resulting in distinct teachers with expertise in intent detection, slot filling, and document classification. Pre-trained models, having been trained on extensive text data, exhibit the capacity to transfer knowledge effectively. Ultimately, we leverage the collective knowledge of these refined teachers to train the student model comprehensively.

3.2 Multi-level Teacher Fine-tuning

We perform separate fine-tuning of pre-trained models on ID, SF, and DC tasks. This yields multi-level teachers, T_{ID}, T_{SF}, and T_{DC} respectively, corresponding to sentence-level, word-level and sentence-level knowledge, respectively. Each pre-trained model specialises in learning knowledge at one specific level from the dialogue datasets, resulting in teachers possessing different levels of dialogue document component understanding. It's important to note that each teacher focuses on one level of dialogue knowledge. This approach is motivated by two factors. First, learning knowledge from a single task is less complex than incorporating knowledge from all tasks, simplifying the fine-tuning of pre-trained models. Secondly, instead of burdening a single model with the challenge of mastering knowledge from all aspects of dialogues, each teacher focuses on a specific level of understanding, such as WS filling, SI detection, or CD classification. For each task, we consolidate data from two datasets (MultiWOZ and M2M) by merging split and corresponding label sets. For example, the training set for fine-tuning includes data from both datasets. We apply cross-entropy loss and fine-tune the pre-trained models for a fixed number of epochs, utilising the checkpoint from the last epoch as the teacher model. The process is described as follows:

$$\begin{aligned}X_{j,tr}{=}&X_{j,tr}^{1},\ldots,X_{j,tr}^{N_{M2M}},X_{j,tr}^{1},\ldots,\\&X_{j,tr}^{N_{MWOZ}},\\\mathcal{L}_{tce}{=}&cross\_entropy(T_j(X_{j,tr}),Y_j),\\&j{\in}\{DC,ID,SF\}\end{aligned}$$

where N_{M2M} and $N_{MultiWoz}$ are the number of training samples, and Y_j is the ground truth.

3.3 Multi teachers knowledge distillation

Following the acquisition of multi-level teachers T, we employ a blend of these teachers to instruct the student model S through multi-teacher knowledge distillation. The combination of teachers comprises different levels, such as {BERT-Base ID, BERT-Base SF, BERT-Base DC}, {BERT-Base ID, RoBERTa-Base DC, and LLaMa2-7b SF}. The student model undergoes separate training for each task, enabling it to grasp the intricacies of individual tasks with the assistance of diverse levels of dialogue knowledge.

We delve into the exploration and introduction of five distinct loss functions to assess their efficacy within the MIDAS. We propose relation loss and teacher prediction supervised loss, specifically designed for multi-level knowledge distillation. Furthermore, with MIDAS, we explore three previously established losses tailored for multi-level teacher integration. These encompass Kullback-Leibler Divergence loss, Similarity loss, and Student Cross Entropy loss, each designed to enhance the learning dynamics in the context of multi-level knowledge distillation.

The two newly proposed losses, relation loss and teacher prediction supervised loss, are designed to better handle the issue of knowledge conflicts between different teachers in multi-level teacher knowledge distillation. Relation loss uses a voting mechanism, guiding the student to learn inter-sample relationships from the majority of teachers. Teacher prediction supervised loss leverages the class probability distributions from each teacher as pseudo-labels, aligning the students' predictions with the dominant teacher perspectives. The details of the two losses are shown in the following:

Relation loss $\mathcal{L}_{rel}$: During training, for each batch of data, triplets are randomly generated, and the internal relations of the triplets are determined by aggregating the votes from the combination of teacher models. Employing TripletMarginLoss (Balntas et al., 2016), the student model learns internal relations among the batch data, aligning its understanding with that of the teacher models and ensuring a consistent perspective on the dataset.

$$\mathcal{L}_{rel}{=}\frac{1}{N}\sum_{i}^{N}TripletMarginLoss(\mathcal{T}_i)$$

where N is the batch size, and triplet $\mathcal{T}$ is generated by and articulated in Algorithm 1, see the details in Section 4.2.

Teacher prediction supervised loss $\mathcal{L}_{tp}$: In addition to utilising the ground truth for each task, we incorporate the predictions made by the teacher models as pseudo-labels to facilitate the training of the student model. We employ the probabilities assigned by the teacher models for each class, ensuring that the student comprehensively acquires the knowledge embedded in the teacher models.

$$\mathcal{L}_{tp}=\sum_{j}^{n_T} cross\_entropy(v_s, P_j)$$

In addition to the two newly proposed losses, we also introduce the previously established losses, briefly described as follows:

Kullback–Leibler divergence(KD) loss $\mathcal{L}_{KD}$: We compute the KD loss (Hinton et al., 2015) by comparing the mean probabilities generated by the combination of teacher models with the probabilities derived from the student model, i.e., $\mathcal{L}_{KD}=KLDivLoss(\frac{1}{n_T}\sum_j^{n_T} P_j, P_s)$, where $P_j=softmax(T_j(X)), P_s=softmax(S(X))$, and n_T is the number of teachers. It helps us to align the prediction probability distributions between the student model and the teacher models.

Similarity loss $\mathcal{L}_{sim}$: The similarity loss is computed by maximising the logit similarities between the student model and teacher models. With this, the student model can learn the knowledge from the teacher models in the feature space, not only the prediction probabilities. The loss equation is:

$$\begin{aligned}\mathcal{L}_{sim}&=-\sum_{j}^{n_T}\mathcal{F}_{sim}(v_j, v_s),\\ v_j&=T_j(X),\ v_s=S(X)\end{aligned} \tag{1}$$

Here, $\mathcal{F}_{sim}$ denotes the similarity function, and v_j represents the teacher logit.

Student cross entropy(SCE) loss $\mathcal{L}_{sce}$: This loss function is computed by comparing the student model's predictions with the ground truth of each task, i.e., $\mathcal{L}_{sce}=cross\_entropy(v_s, Y_{true})$, $v_s=S(X)$, where v_s represents the student logits. It's the basis loss in our supervised learning task.

We experiment with diverse combinations of the aforementioned loss functions to assess their impact on the student's performance across various datasets and NLU tasks. The summary of the loss function is described in Section 4.2 and the detailed result analysis can be found in Section 5.4.

4 Experimental setup

4.1 Datasets and Baselines

We focus on multi-turn dialogue analysis in the dialogue state tracking (DST) domain, which consists of all three NLU tasks, including intent detection ID, slot filling SF, and domain(topic) classification DC. Following by (Weld et al., 2023), we utilise two widely-used benchmark datasets in multi-turn dialogue NLU: Multi-Domain Wizard-of-Oz 2.2 (MWOZ) and Machines Talking To Machines (M2M) in the DST field. Details for datasets are shown below:

- **MWOZ** (Zang et al., 2020) is specifically designed for Dialogue State Tracking (DST) and adopts the conventional human-vs-human Wizard of Oz approach across diverse domains, including attraction, bus, hospital, hotel, police, restaurant, taxi, and train. It incorporates 30 slot types and 11 intent types. The dataset comprises 8,437 dialogues, with an average of 5.68 turns per dialogue and 14.07 tokens per turn. Following by (Liu and Lane, 2016; Goo et al., 2018; Weld et al., 2023), we do not consider any multi-label samples but utilise the data with a single domain and intent.

- **M2M** (Liu et al., 2018) is introduced with virtual agents and user-generated interactions to emulate goal-directed conversations through paraphrasing with templated utterances. M2M has movies and restaurant domains. The slots and intents are categorical, with 21 slot types and 15 intent types. The dataset comprises 1,500 dialogues, with an average of 9.86 turns per dialogue and 8.25 tokens per turn.

We adopted the three published results as baselines, **SeqSeq**(Liu and Lane, 2016), **Slot-Gated** (Goo et al., 2018) and **Tri-level JNLU** (Weld et al., 2023). Additionally, we fine-tuned the pre-trained language models commonly used in the NLU, **BERT-Base**[3], **RoBERTa-Base**[4], **ALBERT-Base**[5]. The details are shown in Appendix B.

4.2 The role of each loss function

The 5 types of losses are utilized for training the student model, each playing a distinct role:

1) $\boldsymbol{L_{kd}}$: This loss facilitates the transfer of knowledge from the teacher models to the student model, enabling the student to mimic the general behaviour of multiple teachers. 2) $\boldsymbol{L_{rel}}$: This loss is designed to capture the relationships between different samples in the input data. It helps to align the student's understanding with that of the teacher models and ensures a consistent perspective on the dataset. 3) $\boldsymbol{L_{sim}}$: This loss encourages the student

[3] https://huggingface.co/bert-base-uncased
[4] https://huggingface.co/roberta-base
[5] https://huggingface.co/albert-base-v2

model to generate outputs similar to those of the teacher models in terms of their overall structure and distribution. It helps to maintain consistency between the student and teacher predictions. 4) $\boldsymbol{L_{sce}}$: This loss function is the fundamental mechanism for training the student model. It entails the student learning to predict the correct labels associated with the input data. 5) $\boldsymbol{L_{tp}}$: This loss leverages the predictions of the teacher models to provide additional supervision signals to the student model. It helps to guide the student towards making predictions that align with those of the teachers.

4.3 Metrics and implementations

This paper evaluates the performance of baseline models and MIDAS in all three multi-turn dialogue tasks, including ID, SF, and DC for each dataset. Following by (Liu and Lane, 2016; Goo et al., 2018; Weld et al., 2023), the metrics for each task are shown as follows: *Accuracy* for ID and DC and *F1 score* for SF. Accuracy is the most commonly used metric for ID as determining the intent of an utterance is typically framed as a classification task. Accuracy is calculated as the ratio of correct predictions to the total number of tests. DC also employs accuracy as it is a classification task. On the other hand, SF employs F1 score. F1 score is directed towards assessing the prediction effectiveness for slot tokens. It computes an F1 score for each class and determines the token-based micro-averaged F1 score across all classes.

We introduce some implementation details in this section and the complete details in Appendix C. For **Multi-teacher fine-tuning**, we use BERT-Base, RoBERTa-Base and LLaMa2-7b[6] as the teacher backbones and fine-tune them on each task. For fine-tuning LLaMa2-7b, we adopt an unmask strategy used in (Li et al., 2023). We use AdamW (Loshchilov and Hutter, 2018) and CrossEntropy loss to fine-tune the pre-trained models for 3 epochs. The learning rate is 5e-5 and is warm-uped linearly from 0 to 5e-5 during the first 10% training steps. The batch size is 32. For **Multi-level Distillation**, we use AdamW and the aforementioned losses to train the student with multi-level teachers. We use Squared Euclidean distance in algorithm 1 and cosine similarity in equation 1. For the combination of these losses, we sum them without any weight. We use the same optimiser, learning rate, warm-up strategy, and batch size as the one used in teacher fine-tuning, and use a vanilla Transformer encoder as a student.

[6]https://huggingface.co/meta-LLaMA/LLaMA-2-7b

	ID		SF		DC	
	MWOZ (ACC)	M2M (ACC)	MWOZ (F1)	M2M (F1)	MWOZ (ACC)	M2M (ACC)
BERT-Base	0.6534	0.8675	0.9218	0.8543	0.8667	0.8923
RoBERTa-Base	0.8424	0.9252	0.9748	0.9132	0.8675	0.8909
ALBERT-Base	0.6531	0.8654	0.9187	0.8542	0.8694	0.8919
SeqSeq	0.6641	0.9250	0.8543	0.9172	-	-
Slot-Gated	0.6883	0.9327	0.8776	0.9279	-	-
Tri-level JNLU	0.7849	0.9419	0.9798	0.9302	0.2572	0.8938
MIDAS (BERT)	0.8464	**0.9427**	**0.9928**	**0.9856**	0.8793	**0.8952**
MIDAS (RoBERTa)	**0.8502**	0.9377	**0.9928**	0.9813	**0.8816**	0.8945

Table 2: The comparison of the MIDAS with baselines. ID, SF and DC indicate intent detection, slot filling and domain classification, respectively, as mentioned in Section 4.3. ACC and F1 stand for accuracy and micro F1, respectively, and scores in bold indicate leadership among the metrics.

5 Results

5.1 Overall performance

We compare MIDAS with fine-tuned PLM baselines and published pioneering model results for two mainstream multi-turn natural language understanding tasks, Intent Detection and Slot Filling, with the same evaluation setup. Table 2 shows that MIDAS remarkably outperforms other baselines. To demonstrate the improvement achieved through MIDAS, we conduct experiments utilising two widely recognised multi-turn dialogue understanding datasets, MWOZ and M2M. Note that all baselines and MIDAS are individually fine-tuned for each task. As detailed in Section 3.1, our approach involves the utilisation of pre-trained models, BERT or RoBERTa, for the fine-tuning of our three multi-level teacher models. These teachers encompass ID, SF, and DC. It is important to highlight that MIDAS undergoes knowledge distillation from three distinct multi-level teachers, each specialising in SI, word token-level slot, and CD topic. Thus, Table 2 shows the results *MIDAS (BERT)* and *MIDAS (RoBERTa)* that all teachers are constructed using either the BERT or RoBERTa architecture.

Two versions of MIDAS exhibit superior performance across both datasets, presenting outstanding outcomes with a slot-filling error rate below 2%. While the RoBERTa-Base model demonstrates superiority in MWOZ, the BERT-Base model excels in M2M. What should be noted is the performance difference between these models is not substantial, with both consistently outperforming other baseline models. In ID and SF tasks, MIDAS showcases notably higher performance compared to baselines. We also conduct experiments on the DC task with the same datasets to better compare the differences between MIDAS and other PLM baselines. However, while surpassing BERT-Base and ALBERT-Base, the performance difference is marginal. We assume that this discrepancy is attributed to the small number of the domain class. In contrast to other baseline models, Seq2Seq and Slot-Gated lack a structure incorporating domain information, making them unable to assess domain classification performance.

Overall, the observation highlights that bolstering multi-level conversation knowledge substantially improves the comprehension of each Natural Language Understanding (NLU) task. Specifically, enhancing results in ID is achievable by refining a student model through the distillation of multi-level knowledge, encompassing SI, WSs, and CD knowledge. The following two sections (Sections 5.2 and 5.3) delve into a more comprehensive exploration of multi-level teacher models and the combination of multi-level teachers.

5.2 Effect of pretrained model for teachers

We then evaluate the efficacy of different pre-trained models for our multi-level teachers. As detailed in Section 5.1 and illustrated in Table 2, we employed all three multi-level teachers (ID, SF, and DC) based on BERT and/or RoBERTa, resulting in a superb performance. In this section, we investigate how various pre-trained language models can impact the knowledge distillation ability of our multi-level teachers in instructing the student model. In addition to using BERT or RoBERTa, we also incorporate LLaMa2-7b, a decoder-only based pre-trained model, into our analysis.

Table 3 shows the results of the effectiveness of using various pre-trained models as base models for all three multi-level teachers[7]. Compared to the high-achieving two encoder-based models, BERT and RoBERTa, the MIDAS (LLaMa) multi-level teachers produce lower performance[8]. We assume a decoder-only model like LLaMa is primarily used for generating coherent and contextually relevant text. In contrast, BERT and RoBERTa are encoder-based models that have a deep understanding of context and relationships between words and excel in classification tasks.

	ID		SF		DC	
	MWOZ (ACC)	M2M (ACC)	MWOZ (F1)	M2M (F1)	MWOZ (ACC)	M2M (ACC)
MIDAS (BERT)	0.8464	**0.9427**	**0.9928**	**0.9856**	0.8793	**0.8952**
MIDAS (RoBERTa)	**0.8502**	0.9377	**0.9928**	0.9813	**0.8816**	0.8945
MIDAS (LLaMa)	0.8403	0.9392	0.9912	0.9833	0.8702	0.8804
MIDAS (Mixed 1)	0.8472	0.9411	0.9839	0.9745	0.8808	0.8929
MIDAS (Mixed 2)	0.8473	0.9401	**0.9928**	0.9764	0.8769	0.8925

Table 3: The performance based on the type of teacher models. The *MIDAS (BERT)* and *(RoBERTa)* are identical to those presented in the table 2 whose all teachers are either BERT or RoBERTa. *MIDAS (LLaMa)* refers to the outcome of utilising the LLaMa2-7b as teacher models of all tasks. The *MIDAS (Mixed 1 and 2)* represents the mixed type teacher combination; Mixed 1: BERT (ID), LLaMa (SF) and RoBERTa (DC); Mixed 2: BERT (ID), RoBERTa (SF) and RoBERTa (DC).

In addition to having multi-level teachers using a single pre-trained model, we adopt a mixed type of pre-trained model for preparing multi-level teachers. For instance, we can apply BERT as a pre-trained model for teaching SI knowledge, utilise RoBERTa as a teacher model for WSs, and adopt LLaMa as a CD topic teacher model. Table 3 shows that using mixed types of pre-trained teacher models is less effective than employing a consistent single pre-trained model as the teacher. This implies that knowledge distillation from teachers with inconsistencies in their feature spaces may impede the learning process for a single student model.

5.3 Effect of multi-level combinations

We explore the impact of incorporating each multi-level teacher (ID, SF, DC) in all three multi-turn dialogue understanding tasks. MIDAS is evaluated with individual teachers (ID, SF, DC), all possible pairs from $\{ID, SF, DC\}$[9], and then with all three teachers. Table 3 presents the results for each combination of teacher models for three different dialogue understanding tasks. Note that the table demonstrates the outcome of *MIDAS (BERT)* teachers, and we produce that of *MIDAS (RoBERTa)* in Table 6. The experimental findings highlight that the $ID + SF + DC$ combination attains the highest performance, underscoring the advantage of the student model integrating knowledge from all teachers for each task.

[7] Note that the *MIDAS (BERT)* and *MIDAS (RoBERTa)* models are identical to those presented in the Table 2.

[8] Any decoder-only LM produces a similar low performance.

[9] Note that we do not adopt $\mathcal{L}_{rel}$ since it is not possible to adopt when there are two teachers.

$\mathcal{L}_{KD}$	$\mathcal{L}_{sce}$	$\mathcal{L}_{sim}$	$\mathcal{L}_{rel}$	$\mathcal{L}_{tp}$	ID		SF		DC	
					MWOZ (ACC)	M2M (ACC)	MWOZ (F1)	M2M (F1)	MWOZ (ACC)	M2M (ACC)
○	○	○	×	×	0.8429	0.9411	**0.9928**	**0.9856**	0.8750	0.8928
○	○	×	○	×	0.8427	0.9411	**0.9928**	0.9791	0.8750	0.8927
○	○	○	○	×	**0.8464**	**0.9427**	**0.9928**	0.9850	0.8780	**0.8952**
○	○	○	○	○	0.8462	0.9373	0.9927	0.9761	**0.8793**	0.8903

Table 4: The comparison of the diverse loss function combinations. Only BERT is utilised as the teacher model, and the results of RoBERTa are presented in Table 7. The full names of each loss can be found in Section 3.3. We adopt two $\mathcal{L}_{KD}$ and $\mathcal{L}_{sce}$ as compulsory knowledge distillation loss and also explore three $\mathcal{L}_{rel}$, $\mathcal{L}_{sim}$, and $\mathcal{L}_{tp}$ for MIDAS. Scores in bold indicate leadership among the metrics, and underlined scores indicate the second-best.

	ID		SF		DC	
	MWOZ (ACC)	M2M (ACC)	MWOZ (F1)	M2M (F1)	MWOZ (ACC)	M2M (ACC)
ID-only	0.8406	0.9366	0.8590	0.9684	0.7977	0.7159
SF-only	0.8310	0.9377	0.9619	0.9718	0.2425	0.8930
DC-only	0.8408	0.9321	0.8888	0.9534	0.6330	0.8915
ID+SF	0.8422	0.9399	0.9924	0.9835	0.8760	0.8939
ID+DC	0.8400	0.9292	0.9923	0.9848	0.8756	0.8929
SF+DC	0.8376	0.9416	0.9923	0.9825	0.8760	0.8940
ID+SF+DC	**0.8464**	**0.9427**	**0.9928**	**0.9850**	**0.8780**	**0.8952**

Table 5: The performance based on the type of teacher models. The first column indicates the type of teacher used. For example, *ID+SF+DC* uses all intent classification, slot filling, and domain classification teachers, while *ID-only* uses only the intent classification teacher. Only BERT is utilised as the teacher model.

5.4 Effect of knowledge distillation loss

As mentioned in Section 3.3, we conducted the loss function ablation study for MIDAS. This comprehensive evaluation aims to identify the most effective combinations that enhance the student model's proficiency in handling different aspects of dialogue understanding across multiple NLU tasks. Note that we use $\mathcal{L}_{KD}$ and $\mathcal{L}_{sce}$ as compulsory knowledge distillation losses, and conduct an ablation study of three other multi-level teacher losses: $\mathcal{L}_{sim}$, $\mathcal{L}_{rel}$, and $\mathcal{L}_{tp}$. As shown in Table 4, the results indicate that incorporating $\mathcal{L}_{rel}$ with $\mathcal{L}_{sim}$ achieves the best or the second best performance across all tasks and datasets. Although $\mathcal{L}_{rel}$ and $\mathcal{L}_{sim}$ share a similar trend, their impact on model learning may be somewhat superior when employed independently, particularly when utilising $\mathcal{L}_{sim}$. While incorporating $\mathcal{L}_{tp}$ with the others led to a slight performance increase, it did not match the effectiveness observed with the sole application of the earlier losses.[10]

[10]We conducted testing with $\mathcal{L}_{tp}$ only, it produces much lower performance than any others. See the details in section 5.6

	ID		SF		DC	
	MWOZ (ACC)	M2M (ACC)	MWOZ (F1)	M2M (F1)	MWOZ (ACC)	M2M (ACC)
ID-only	0.8339	0.8097	0.9079	0.9326	0.6183	0.7147
SF-only	0.8403	0.8945	0.9620	0.9434	0.2471	0.8917
DC-only	0.8469	0.8929	0.9547	0.9251	0.7521	0.8913
ID+SF	0.8451	0.9083	0.9928	0.9802	0.8734	0.8923
ID+DC	0.8373	0.9114	0.9921	0.9797	0.8763	0.8888
SF+DC	0.8453	0.9147	0.9927	0.9805	0.8707	0.8912
ID+SF+DC	**0.8502**	**0.9377**	**0.9928**	**0.9813**	**0.8816**	**0.8945**

Table 6: The performance based on the type of teacher models. The first column indicates the type of teacher used. For example, *ID+SF+DC* uses all intent classification, slot filling, and domain classification teachers, while *ID-only* uses only the intent classification teacher. Only RoBERTa is utilised as the teacher model.

5.5 Combination-based ablation study

We explore the impact of incorporating each multi-level teacher (ID, SF, DC) in all three multi-turn dialogue understanding tasks. Table 6 presents the results for each combination of teacher models for three different dialogue understanding tasks. The experimental results are when only RoBERTa is adopted as the teacher model. MIDAS is evaluated with individual teachers (ID, SF, DC), all possible pairs from {ID, SF, DC}, and then with all three teachers. For example, ID+SF+DC uses all intent classification, slot filling, and domain classification teachers, while ID-only uses only the intent classification teacher. Note that we do not adopt $\mathcal{L}_{rel}$ while two models are used since it is not possible to adopt when there are two teachers. The experimental findings highlight that the ID+SF+DC combination attains the highest performance, underscoring the advantage of the student model integrating knowledge from all teachers for each natural language understanding task.

5.6 Loss function ablation study

We conducted the loss function ablation study for MIDAS with RoBERTa Teacher. This comprehensive evaluation aims to identify the most effective combinations that enhance the student model's proficiency in handling different aspects of dialogue

$\mathcal{L}_{KD}$	$\mathcal{L}_{sce}$	$\mathcal{L}_{sim}$	$\mathcal{L}_{rel}$	$\mathcal{L}_{tp}$	**ID**		**SF**		**DC**	
					MWOZ (ACC)	**M2M (ACC)**	**MWOZ (F1)**	**M2M (F1)**	**MWOZ (ACC)**	**M2M (ACC)**
○	○	○	×	×	0.8441	0.9362	**0.9928**	**0.9842**	0.8744	**0.8945**
○	○	×	○	×	0.8459	**0.9377**	0.9610	0.8415	**0.8816**	0.8914
○	○	○	○	×	**0.8502**	<u>0.9376</u>	**0.9928**	<u>0.9813</u>	0.8803	**0.8945**
○	○	○	○	○	<u>0.8488</u>	0.9264	<u>0.9912</u>	0.9704	<u>0.8811</u>	<u>0.8922</u>

Table 7: The comparison of the diverse loss function combinations. Only RoBERTa is utilised as the teacher model. We adopt two $\mathcal{L}_{KD}$ and $\mathcal{L}_{sce}$ as compulsory knowledge distillation loss and also explore three $\mathcal{L}_{rel}$, $\mathcal{L}_{sim}$, and $\mathcal{L}_{tp}$ for MIDAS. Scores in bold indicate leadership among the metrics, and underlined scores indicate the second-best.

understanding across multiple NLU tasks. Note that we use $\mathcal{L}_{KD}$ and $\mathcal{L}_{sce}$ as compulsory knowledge distillation losses, and conduct an ablation study of three multi-level teacher losses: $\mathcal{L}_{sim}$, $\mathcal{L}_{rel}$, and $\mathcal{L}_{tp}$. Among them, $\mathcal{L}_{rel}$, and $\mathcal{L}_{tp}$ are newly proposed losses in our work. As shown in Table 7, the results indicate that incorporating $\mathcal{L}_{sim}$ with $\mathcal{L}_{rel}$ achieves the best or the second best performance across all tasks and datasets. Although $\mathcal{L}_{rel}$ and $\mathcal{L}_{sim}$ share a similar trend, their impact on model learning may be somewhat superior when employed independently, particularly when utilising $\mathcal{L}_{sim}$. While incorporating $\mathcal{L}_{tp}$ with the others led to a slight performance increase, it did not match the effectiveness observed with the sole application of the earlier losses. We assume the reason is that it does not mainly focus on the teacher prediction supervised loss, not really consider the relations with student models

5.7 Qualitative analysis: Case study

We further evaluate MIDAS using a qualitative assessment of three NLU tasks with the M2M dataset. As shown in Table 8, we assume a three-utterance dialogue: "*how many tickets would you like to buy?*," "*1*," and "*what date do you want to go?*." Based on this conversation, we test ID, SF, and DC. We compare MIDAS with BERT, LLaMa3.1, and GPT4o. **MIDAS** is trained with three teacher models, $BERT_{ID}$, $BERT_{SF}$, and $BERT_{DC}$. BERT is a fine-tuned model (*BERT-Base*) focused on one task per prediction, while LLaMa3.1 and GPT4o are evaluated using few-shot learning, with three examples provided for each slot type, intent, and domain. The results reveal the limitations of LLMs in handling multi-level knowledge. For instance, LLaMa3.1 correctly identifies the slot **B-num_tickets** but incorrectly classifies the domain as **restaurant**.

Furthermore, both GPT4o and LLaMa3.1 struggle to follow the prompt instructions, failing to predict enough slot types even when words are explicitly separated. While fine-tuned BERT also fails to capture the conversation domain and slot types, **MIDAS**, aided by multi-level teacher models, consistently predicts slot types, intents, and domains correctly, matching the ground truth.

Model	T.	SF	ID	DC
Utterance	1	how, many, tickets, would, you, like, to, buy, ?	-	-
	2	1	-	-
	3	what, date, do, you, want, to, go, ?	-	-
GT	1	O, O, O, O, O, O, O, O, O	request	movie
	2	B-num_tickets	inform	movie
	3	O, O, O, O, O, O, O, O	request	movie
GPT4o	1	O, O, O, O, O, O, O, NaN, NaN	request	movie
	2	(O)	(O)	restau
	3	O, B-date, O, O, O, O, NaN, NaN	request	restau
LLaMa3.1	1	O, O, O, O, O, O, B-num_tickets, O, NaN	request	restau
	2	B-category	request	movie
	3	O, O, O, O, O, O, O, O	request	movie
BERT	1	O, O, O, O, O, O, O, O, O	request	restau
	2	O	inform	restau
	3	O, O, O, O, O, O, O, O	request	restau
MIDAS	1	O, O, O, O, O, O, O, O, O	request	movie
	2	B-num_tickets	inform	movie
	3	O, O, O, O, O, O, O, O	request	movie

Table 8: A Prediction example with a three-turn conversation on slot filling, intent detection, and domain classification. Green: the result that perfectly matches the Ground Truth (GT), Red: Entirely Incorrect, and Yellow: Partially Correct Results. 'NaN' means the value at this position is empty, and '(O)' means the outputs of LLMs are out of the domain defined in the prompt. Additional prediction case study examples are articulated in Appendix D.

6 Conclusion

This paper introduces a novel multi-level teacher knowledge distillation framework to enhance multi-turn natural language understanding (NLU). By fine-tuning pre-trained models at word, sentence, and document levels, we construct multi-level teachers, imparting their knowledge to a student model. Various loss functions are introduced and explored, and the experiment results demonstrate the framework's effectiveness in improving the student model's understanding across diverse NLU tasks. It shows better than the LLM result.

7 Limitation

There are some spaces for future work, including a more fine-grained analysis of the impact of each loss, covering multilingual multi-turn dialogue. The quality of the multilingual pre-trained model would be the potential risk to achieve enough multi-turn NLU performance. We believe this work will provide valuable insights into various aspects of dialogue knowledge for NLU and multi-level knowledge distillation.

8 Acknowledgement

This work was supported by the Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2022-00155911, Artificial Intelligence Convergence Innovation Human Resources Development (Kyung Hee University))

References

Waheed Ahmed Abro, Guilin Qi, Muhammad Aamir, and Zafar Ali. 2022. Joint Intent Detection and Slot Filling using Weighted Finite State Transducer and BERT. *Applied Intelligence*, 52(15):17356–17370.

Abdollah Amirkhani, Amir Khosravian, Masoud Masih-Tehrani, and Hossein Kashiani. 2021. Robust Semantic Segmentation with Multi-teacher Knowledge Distillation. *IEEE Access*, 9:119049–119066.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, and Fei Huang et al. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. 2016. Learning local feature descriptors with triplets and shallow convolutional neural networks. pages 119.1–119.11.

Ankur Bapna, Gokhan Tur, Dilek Hakkani-Tur, and Larry Heck. 2017. Sequential Dialogue Context Modeling for Spoken Language Understanding. In *Proceedings of the 18th SIGDIAL*, pages 103–114.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell et al. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Dongsheng Chen, Zhiqi Huang, Xian Wu, Shen Ge, and Yuexian Zou. 2022a. Towards Joint Intent Detection and Slot Filling via Higher-order Attention. In *Proceedings of the 35th IJCAI*, pages 4072–4078.

Guanhua Chen, Yutong Yao, Derek F Wong, and Lidia S Chao. 2024. A two-stage prediction-aware contrastive learning framework for multi-intent nlu. *arXiv preprint arXiv:2405.02925*.

Lisong Chen, Peilin Zhou, and Yuexian Zou. 2022b. Joint Multiple Intent Detection and Slot Filling via Self-Distillation. In *Proceedings of the 2022 ICASSP*, pages 7612–7616.

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. BERT for Joint Intent Classification and Slot Filling. *arXiv preprint arXiv:1902.10909*.

Xuxin Cheng, Zhihong Zhu, Wanshi Xu, Yaowei Li, Hongxiang Li, and Yuexian Zou. 2023. Accelerating Multiple Intent Detection and Slot Filling via Targeted Knowledge Distillation. In *Findings of the EMNLP 2023*, pages 8900–8910.

Mai Hoang Dao, Thinh Hung Truong, and Dat Quoc Nguyen. 2021. Intent Detection and Slot Filling for Vietnamese. In *Proceedings of the INTERSPEECH 2021*, pages 4698–4702.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, and Angela Fan et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. 2018. Slot-Gated Modeling for Joint Slot Filling and Intent Prediction. In *Proceedings of the 2018 NAACL*, pages 753–757.

Kalpa Gunaratna, Vijay Srinivasan, Akhila Yerukola, and Hongxia Jin. 2022. Explainable Slot Type Attentions to Improve Joint Intent Detection and Slot Filling. In *Findings of the EMNLP 2022*, pages 3367–3378.

Soyeon Caren Han, Siqu Long, Huichun Li, Henry Weld, and Josiah Poon. 2021a. Bi-directional joint neural networks for intent classification and slot filling. In *22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021*, pages 3931–3935. International Speech Communication Association.

Soyeon Caren Han, Siqu Long, Huichun Li, Henry Weld, and Josiah Poon. 2021b. Bi-directional joint neural networks for intent classification and slot filling. In *"Proceedings of the INTERSPEECH*, Brno, Czech Republic. ISCA.

Seong-Hwan Heo, WonKee Lee, and Jong-Hyeok Lee. 2022. mcBERT: Momentum Contrastive Learning with BERT for Zero-Shot Slot Filling. In *Proceedings of the INTERSPEECH 2022*, pages 1243–1247.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2014. Distilling the Knowledge in a Neural Network. In *Advances in NeurIPS*, pages 1–9.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Liang Huang, Senjie Liang, Feiyang Ye, and Nan Gao. 2023. A Fast Attention Network for Joint Intent Detection and Slot Filling on Edge Devices. *IEEE Transactions on Artificial Intelligence*, pages 1–11.

Zhong Ji, Jingwei Ni, Xiyao Liu, and Yanwei Pang. 2023. Teachers Cooperation: Team-Knowledge Distillation for Multiple Cross-domain Few-shot Learning. *Frontiers of Computer Science*, 17(2):172312.

Yidi Jiang, Bidisha Sharma, Maulik Madhavi, and Haizhou Li. 2021. Knowledge Distillation from BERT Transformer to Speech Transformer for Intent Classification. In *Proceedings of the INTERSPEECH 2021*, pages 4713–4717.

Woojeong Jin, Maziar Sanjabi, Shaoliang Nie, Liang Tan, Xiang Ren, and Hamed Firooz. 2021. Modality-specific Distillation. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 42–53. Association for Computational Linguistics.

Yeongseo Jung, Eunseo Jung, and Lei Chen. 2023. Towards a Unified Conversational Recommendation System: Multi-task Learning via Contextualized Knowledge Distillation. In *Proceedings of the 2023 EMNLP*, pages 13625–13637.

Quan Kong, Ziming Wu, Ziwei Deng, Martin Klinkigt, Bin Tong, and Tomokazu Murakami. 2019. MMAct: A Large-Scale Dataset for Cross Modal Human Action Understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8658–8667.

Zongxi Li, Xianming Li, Yuzhang Liu, Haoran Xie, Jing Li, Fu lee Wang, Qing Li, and Xiaoqin Zhong. 2023. Label supervised llama finetuning. *Preprint*, arXiv:2310.01208.

Bing Liu and Ian Lane. 2016. Attention-based Recurrent Neural Network Models for Joint Intent Detection and Slot Filling. In *Proceedings of the INTERSPEECH 2016*, pages 685–689.

Bing Liu, Gokhan Tür, Dilek Hakkani-Tür, Pararth Shah, and Larry Heck. 2018. Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. In *Proceedings of the 2018 NAACL*, pages 2060–2069.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *"Proceedings of the ICLR 2018*.

Bin Luo and Baiming Feng. 2024. Bi-directional joint model for intent detection and slot filling. In *2024 7th International Conference on Advanced Algorithms and Control Engineering (ICAACE)*, pages 1311–1314. IEEE.

Jie Mei, Yufan Wang, Xinhui Tu, Ming Dong, and Tingting He. 2023. Incorporating BERT With Probability-Aware Gate for Spoken Language Understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:826–834.

Meryem M'hamdi, Doo Soon Kim, Franck Dernoncourt, Trung Bui, Xiang Ren, and Jonathan May. 2021. X-METRA-ADA: Cross-lingual Meta-Transfer learning Adaptation to Natural Language Understanding and Question Answering. In *Proceedings of the 2021 NAACL*, pages 3617–3632.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2020. Improved Knowledge Distillation via Teacher Assistant. In *Proceedings of the 34th AAAI*, pages 5191–5198.

Jianyuan Ni, Raunak Sarbajna, Yang Liu, Anne H.H. Ngu, and Yan Yan. 2022. Cross-Modal Knowledge Distillation For Vision-To-Sensor Action Recognition. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4448–4452.

Haojie Pan, Chengyu Wang, Minghui Qiu, Yichang Zhang, Yaliang Li, and Jun Huang. 2021. Meta-KD: A Meta Knowledge Distillation Framework for Language Model Compression across Domains. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 3026–3036. Association for Computational Linguistics.

Thinh Pham, Chi Tran, and Dat Quoc Nguyen. 2023. MISCA: A Joint Model for Multiple Intent Detection and Slot Filling with Intent-Slot Co-Attention. In *Findings of the EMNLP 2023*, pages 12641–12650.

Libo Qin, Tailu Liu, Wanxiang Che, Bingbing Kang, Sendong Zhao, and Ting Liu. 2021a. A Co-Interactive Transformer for Joint Slot Filling and Intent Detection. In *Proceedings of the 2021 ICASSP*, pages 8193–8197.

Libo Qin, Fuxuan Wei, Tianbao Xie, Xiao Xu, Wanxiang Che, and Ting Liu. 2021b. GL-GIN: Fast and Accurate Non-Autoregressive Model for Joint Multiple Intent Detection and Slot Filling. In *Proceedings of the 59th ACL*, pages 178–188.

Subendhu Rongali, Beiye Liu, Liwei Cai, Konstantine Arkoudas, Chengwei Su, and Wael Hamza. 2021. Exploring Transfer Learning For End-to-End Spoken Language Understanding. In *Proceedings of the 35th AAAI*, volume 35, pages 13754–13761.

Wonchul Son, Jaemin Na, Junyong Choi, and Wonjun Hwang. 2021. Densely Guided Knowledge Distillation Using Multiple Teacher Assistants. In *Proceedings of the 2021 ICCV*, pages 9395–9404.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and Pouya Tafti et al. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Chen Tingting, Lin Min, and Li Yanling. 2019. Joint Intention Detection and Semantic Slot Filling Based on BLSTM and Attention. In *Proceedings of the 2019 ICCCBDA*, pages 690–694.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, and Shruti Bhosale et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Thanh Tran, Kai Wei, Weitong Ruan, Ross McGowan, Nathan Susanj, and Grant P. Strimel. 2022. Adaptive Global-Local Context Fusion for Multi-Turn Spoken Language Understanding. In *Proceedings of the 36th AAAI*, volume 36, pages 12622–12628.

Nguyen Anh Tu, Hoang Thi Thu Uyen, Tu Minh Phuong, and Ngo Xuan Bach. 2023. Joint Multiple Intent Detection and Slot Filling with Supervised Contrastive Learning and Self-Distillation. In *Proceedings of the 26th ECAI*, pages 333–343.

Jixuan Wang, Kai Wei, Martin Radfar, Weiwei Zhang, and Clement Chung. 2021a. Encoding Syntactic Knowledge in Transformer Encoder for Intent Detection and Slot Filling. In *Proceedings of the 35th AAAI*, volume 35, pages 13943–13951.

Kai Wang, Yu Liu, Qian Ma, and Quan Z Sheng. 2021b. Mulde: Multi-Teacher Knowledge Distillation for Low-Dimensional Knowledge Graph Embeddings. In *Proceedings of the Web Conference 2021*, pages 1716–1726.

Henry Weld, Sijia Hu, Siqu Long, Josiah Poon, and Soyeon Han. 2023. Tri-level Joint Natural Language Understanding for Multi-turn Conversational Datasets. In *Proceedings of the INTERSPEECH 2023*, pages 700–704.

Henry Weld, Xiaoqi Huang, Siqu Long, Josiah Poon, and Soyeon Caren Han. 2022. A survey of joint intent detection and slot filling models in natural language understanding. *ACM Computing Surveys*, 55(8):1–38.

Chuhan Wu, Fangzhao Wu, and Yongfeng Huang. 2021a. One Teacher is Enough? Pre-trained Language Model Distillation from Multiple Teachers. In *Findings of the ACL 2021*, pages 4408–4413.

Ting-Wei Wu and Biing Juang. 2023. Infusing Context and Knowledge Awareness in Multi-turn Dialog Understanding. In *Findings of the EACL 2023*, pages 254–264.

Ting-Wei Wu, Ruolin Su, and Biing-Hwang Juang. 2021b. A Context-Aware Hierarchical BERT Fusion Network for Multi-Turn Dialog Act Detection. In *Proceedings of the INTERSPEECH 2021*, pages 1239–1243.

Bowen Xing and Ivor Tsang. 2022. Group is better than individual: Exploiting Label Topologies and Label Relations for Joint Multiple Intent Detection and Slot Filling. In *Proceedings of the 2022 EMNLP*, pages 3964–3975.

Fei Yuan, Linjun Shou, Jian Pei, Wutao Lin, Ming Gong, Yan Fu, and Daxin Jiang. 2021. Reinforced Multi-Teacher Selection for Knowledge Distillation. In *Proceedings of the 35th AAAI*, pages 14284–14291.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117.

Algorithm 1 Triplet Relations

Input: The hidden states of the batch data from the teachers $H_t=\{h_1^1, h_1^2, ...h_1^n, ..., h_j^n\}$, the hidden states of the batch data from the student $H_s=\{h_s^1, h_s^2, ...h_s^n\}$, the teacher model set $T=\{T_1, T_2, ..., T_j\}$
Parameter: Distance function $\mathcal{F}_D$
Output: The batch size of triplet relations $\mathcal{T}$

Let $i=0$, $\mathcal{T}=\emptyset$.
for $i<n$ **do**
  Randomly select three samples from the batch and label their indexes in the batch as $r1, r2, r3$.
  Treat the sample indexed $r1$ as the anchor, $r2$ as the positive sample, $r3$ as the negative sample.
  Let $l=0$, $flag=0$
  for $l<j$ **do**
    $d_{1,2}=\mathcal{F}_D(h_l^{r1}, h_l^{r2})$, $d_{1,3}=\mathcal{F}_D(h_l^{r1}, h_l^{r3})$
    if $d_{1,2}>d_{1,3}$ **then**
      $flag+=1$
    else
      $flag-=1$
    end if
    $l+=1$
  end for
  if $flag>0$ **then**
    Swap the labels of h_l^{r2} and h_l^{r3}.
  end if
  $i+=1$
  $\mathcal{T}+=[h_s^{r1}, h_s^{r2}, h_s^{r3}]$
end for
return $\mathcal{T}$

A Related works

Table 9 presents a comparison of MIDAS with 23 previous joint NLU models. Recently, most NLU studies have embraced a joint learning model capable of handling all NLU tasks to mitigate error propagation inherent in pipelined approaches (Wang et al., 2021a; Han et al., 2021b; Gunaratna et al., 2022; Huang et al., 2023). The initial joint models employed traditional neural networks like RNN (Liu and Lane, 2016) and LSTM (Tingting et al., 2019; Qin et al., 2021b; Chen et al., 2022a; Xing and Tsang, 2022; Tran et al., 2022; Pham et al., 2023) with attention mechanisms.

All models leverage slot-level knowledge and intent-level knowledge, but only five previous works incorporate domain-level knowledge. This implies that only five prior studies utilised a multi-turn dialogue dataset.

Only one previous study (Weld et al., 2023) conducted tests on domain classification. Hence, we chose (Weld et al., 2023) as a representative baseline. What sets the proposed model apart is its utilisation of multi-teacher knowledge distillation. While two previous works employed self-knowledge distillation and another two adopted one-teacher knowledge distillation, MIDAS represents the first attempt at employing multi-teacher knowledge distillation for joint learning in natural language understanding.

B Details of baselines

Given the limited number of baselines available for Multi-turn Dialogue Understanding, we selected the following models as baselines. **BERT-Base**[11] is a transformer-based language model that serves as a standard benchmark. **RoBERTa-Base**[12] improves upon BERT by removing the next-sentence prediction task and optimizing hyperparameters, such as using larger mini-batches and higher learning rates. **ALBERT-Base**[13] further enhances efficiency through factorized embedding parameterization and cross-layer parameter sharing, achieving better performance with fewer parameters. **SeqSeq** (Liu and Lane, 2016) is an RNN model with attention mechanisms, developed for joint intent detection (ID) and slot filling (SF) tasks. **Slot-Gated** (Goo et al., 2018) introduces a slot-gating mechanism to capture the relationship between intent and slot labels, improving semantic understanding through global optimization. Lastly, **Tri-level JNLU** (Weld et al., 2023) incorporates domain information for enhanced joint modeling of ID and SF.

C Implementation details

C.1 Experiment hyperparameters

Table 10 presents the hyperparameters, used in our proposed Multi-level Teacher Fine-tuning, as well as Multi-Teacher Knowledge Distillation. The Implementation details can be found in Section 4.3. of the main submission.

Hyper-parameter	Fine-tuning	Knowledge Distillation
Learning Rate	5e-5	5e-5
Batch Size	32	32
Warm-up Steps	10% of Max epoch	10% of Max epoch
Mex epoch	3	100
Stop Strategy	Max Epoch	Early Stopping on loss
Stop Patience	-	10
Optimizer	AdamW	AdamW
Optimizer Weight Decay	1e-2	1e-2
Optimizer Betas	0.9, 0.999	0.9, 0.999
Margin in $\mathcal{L}_{rel}$	-	0.2
Norm in $\mathcal{L}_{rel}$	-	2
$\mathcal{F}_D$ in $\mathcal{L}_{rel}$	-	L2-Norm
Similarity in $\mathcal{L}_{sim}$	-	Cosine Similarity
Max Token Length	512	512

Table 10: The hyper-parameters used in our experiments.

We further present the results of various experiments conducted to select hyperparameters, particularly the learning rate, in Table 11. In all tests, the temperature is fixed at 20, and only the learning rate is changed to 0.0005, 0.00005, and 0.000005. In the experiments on the M2M dataset, the performance of Gemma-7b alongside LLaMa2-7b is

[11] https://huggingface.co/bert-base-uncased
[12] https://huggingface.co/roberta-base
[13] https://huggingface.co/albert-base-v2

Model	Year	Word (Slot)	Sentence (Intent)	Document (Domain)	Dialogue Type	Joint Integration
SeqSeq Liu and Lane (2016)	2016	○	○	×	Single-Turn	BiRNN + Attention
SDEN Bapna et al. (2017)	2017	○	○	○	Multi-Turn	BiRNN + Memory Network
Slot-Gated Goo et al. (2018)	2018	○	○	×	Single-Turn	BiLSTM + Slot Gate
BLSTM+attention Tingting et al. (2019)	2019	○	○	×	Single-Turn	BiLSTM + Attention
Co-Interactive Transformer Qin et al. (2021a)	2021	○	○	×	Single-Turn	BiLSTM + Attention
GL-GIN Qin et al. (2021b)	2021	○	○	×	Single-Turn	BiLSTM + GAT
SyntacticTF Wang et al. (2021a)	2021	○	○	×	Single-Turn	Transformer
STD Jiang et al. (2021)	2021	○	○	×	Single-Turn	Transformer + One-teacher KD
JointIDSF Dao et al. (2021)	2021	○	○	×	Single-Turn	CRF + Attention
CaBERT-SLU Wu et al. (2021b)	2021	○	○	○	Multi-Turn	Attention
SDJN Chen et al. (2022b)	2021	○	○	×	Single-Turn	BiLSTM + self KD
HAN Chen et al. (2022a)	2022	○	○	×	Single-Turn	BiLSTM + Attention
ReLA-NET Xing and Tsang (2022)	2022	○	○	×	Single-Turn	BiLSTM + GAT
XAI Attention Gunaratna et al. (2022)	2022	○	○	×	Multi-Turn	XAI
WFST-BERT Abro et al. (2022)	2022	○	○	×	Single-Turn	WFST
Contextual SLU Tran et al. (2022)	2022	○	○	○	Multi-Turn	BiLSTM + Attention
TKDF Cheng et al. (2023)	2023	○	○	×	Single-Turn	SSRAN + One-teacher KD
MISCA Pham et al. (2023)	2023	○	○	×	Single-Turn	BiLSTM + Attention
PAGM Mei et al. (2023)	2023	○	○	×	Single-Turn	Gate
FAN Huang et al. (2023)	2023	○	○	×	Single-Turn	Attention
Tri-level JNLU Weld et al. (2023)	2023	○	○	○	Multi-Turn	Transformer
CKA-NLU Wu and Juang (2023)	2023	○	○	○	Multi-Turn	Attention
BiSLU Tu et al. (2023)	2023	○	○	×	Single-Turn	self KD
PACL Chen et al. (2024)	2024	○	○	×	Multi-Turn	Contrastive Learning + Attention
BiJM Luo and Feng (2024)	2024	○	○	×	Single-Turn	Transformer + Enhance Layer
MIDAS (Ours)	**2024**	○	○	○	**Multi-Turn**	**Multi-teacher KD**

Table 9: Summary of previous joint NLU models and MIDAS. Word, Sentence, and Document columns indicate whether the relevant information is used for joint integration. GAT in the Joint Integration column refers to the graph attention network, KD refers to knowledge distillation, and WFST refers to Weighted Finite-State Transducers.

also measured to compare performance with the generative model. The highest accuracy is shown when the learning rate was 0.00005, and Gemma-7b shows similar performance to LLaMa2-7b, but LLaMa2-7b is slightly superior. The best performance is observed when the learning rate is 0.00005, which is also the case in experiments on the MWOZ dataset.

C.2 Model details

We display the visualisation of teacher models and our student model Vanilla Transformer Encoder together. Those two summarises can be found in Table 12. Note that we use LoRA to fine-tune LLaMa 2-7b.

C.3 Hardware information

Our experiments are run on the Linux platform with an A6000 Nvidia graphic card and an AMD Ryzen Threadripper PRO 5955WX 16-core CPU, and the RAM is 128G.

Model	Task	Learning Rate	Accuracy
		M2M	
LLaMa2-7b	ID	0.0005	0.9121
		0.00005	0.9392
		0.000005	0.9093
	SF	0.0005	0.9696
		0.00005	0.9833
		0.000005	0.9349
	DC	0.0005	0.8895
		0.00005	0.8804
		0.000005	0.8375
Gemma-7b	ID	0.0005	0.9204
		0.00005	0.9357
		0.000005	0.9102
	SF	0.0005	0.9693
		0.00005	0.9816
		0.000005	0.9429
	DC	0.0005	0.8799
		0.00005	0.8840
		0.000005	0.7890
		MWOZ	
LLaMa2-7b	ID	0.0005	0.8021
		0.00005	0.8403
		0.000005	0.7952
	SF	0.0005	0.9776
		0.00005	0.9912
		0.000005	0.9740
	DC	0.0005	0.8411
		0.00005	0.8702
		0.000005	0.7026

Table 11: Summary of performance changes according to learning rate changes.

D In-depth PLM/LLM analysis

In this section, we provide an in-depth quantitative and qualitative analysis, incorporating a detailed comparison between PLMs and LLMs.

D.1 Compared with PLMs

We evaluate MIDAS with a qualitative assessment of the three NLU tasks on MWOZ and M2M,

Conversation Sample	MIDAS		Gemma-7b		GPT3.5	
	Slot Token	Intent	Slot Token	Intent	Slot Token	Intent
Find a restaurant for breakfast	Find/O a/O restaurant/O for/O breakfast/B-Meal	Greeting	Find/B-Cat a/B-Date restaurant/B-Loc for/B-Meal breakfast/B-Movie	Inform	Find/O a/B-Meal restaurant/O for/O breakfast/O	Request
In what area and what type of rating for the place?	In/O what/O area/O and/O what/O type/O of/O rating/O for/O the/O place/O	Request	In/B-Loc what/B-Cat area/I-Rating and/N what/N type/N of/N rating/N for/N the/N place/N	Inform	In/O what/O area/O and/O what/O type/O of/O rating/ for/O the/O place/O	Request
One in Redmond and is Michelin rated	One/O in/O Redmond/B-Loc and/O is/O Michelin/B-Rating rated/I-Rating	Inform	One/B-RN in/B-Loc Redmond/B-Rating and/N is/N Michelin/N rated/N	Inform	One/I-Loc in/O Redmond/O and/B-Rating is/O Michelin/B-RN rated/I-Rating	Inform
I found the following place: the ivy, acorn or deep blue	I/O found/O the/O following/O place/O :/O the/B-RN ivy/I-RN , /O acorn/B-RN or/O deep/B-RN blue/I-RN	Select	I/B-RN found/B-RN the/B-RN following/O place/O :/O the/O, ivy/O , /O acorn/O or/O deep/O blue/O	Inform	I/O found/O the/O following/B-RN place/O :/B-RN the/I-RN ivy/I-RN, /O acorn/N or/N deep/N blue/N	Select
Show me info for deep blue	Show/O me/O info/O for/O deep/B-RN blue/I-RN	Affirm	Show/B-RN me/B-Date info/B-Time for/I-Date deep/I-Time blue/O	Inform	Show/O me/O info/O for/O deep/B-Movie blue/I-Movie	Request
Domain	Restaurant		Movie		Movie	

Slots type: O (Other), Meal, Loc (Location), RN (Restaurant Name), Cat (Category), Date, Movie, Time, N (None, LLM doesn't answer)

Intent type: Greeting, Request, Inform, Select, Affirm

Domain type: Movie, Restaurant

(a)

Conversation Sample	MIDAS		Gemma-7b		GPT3.5	
	Slot Token	Intent	Slot Token	Intent	Slot Token	Intent
Looking for a train	Looking/O for/O a/O train/O	Find Train	Looking/O for/O a/O train/O	Find Train	Looking/O for/O a/O train/O	Find Train
What's the time?						
I want to depart after 19.45	I/O want/O to/O depart/O after/O 19:45/B-TL	Find Train	I/O want/O to/O depart/O after/O 19:45/O	Find Train	I/O want/O to/O depart/O after/O 19:45/O	Find Train
How many tickets?						
I would like to book 4	I/O would/O like/O to/O book/O 4/O	Find Train	I/O would/O like/O to/O book/O 4/O	Find Train	I/O would/O like/O to/O book/O 4/O	Book Train
Do you need any other services?						
I would like to know the address of La Tasac Restaurant	I/O would/O like/O to/O know/O the/O address/O of/O La/B-RN Tasac/I-RN Restaurant/O	Find Res.	I/O would/O like/O to/O know/O the/O address/O of/O La/O Tasac/O Restaurant/O	Find Res.	I/O would/O like/O to/O know/O the/O address/O of/O La/B-RN Tasac/O Restaurant/O	Find Res.
It is located at 14 -16 Bridge Street. Would you like a reservation?						
I would like to book a table for 5 people	I/O would/O like/O to/O book/O a/O table/O for/O 5/O people/O	Book Res.	I/O would/O like/O to/O book/O a/O table/O for/B-RB 5/B-RF people/B-RN	Book Res.	I/O would/O like/O to/O book/B-RB a/O table/O for/O 5/O people/O	Book Res.
What is the date?						
At 15:30 on the same day	At/O 15:30/B-RB on/O the/O same/O day/O	Book Res.	At/O 15:30/O on/O the/O same/O day/O	Find Res.	At/O 15:30/O on/O the/O same/O day/O	Book Train
Domain	Train, Restaurant		Train		Train	

Slots type: O (Other), TL (Train Leaveat), RN (Restaurant Name), RB (Restaurant Booktime), RF (Restaurant Food)

Intent type: Find Train, Book Train, Find Res. (Find Restaurant), Book Res. (Book Restaurant)

Domain type: Train, Restaurant

(b)

Figure 3: Two examples for qualitative analysis: (a) shows the results on the M2M dataset, and (b) shows the results on the MWOZ dataset. Each example shows the results when MIDAS matches the ground truth. The three cells below each example display the type lists for slot, intent, and domain, and red text indicates errors in each column of the results table.

	BERT	RoBERTa	LLaMa	Student
Architecture	Encoder	Encoder	Decoder	Encoder
Parameters	110M	125M	7B	58M
Layers	12	12	32	6
Heads	12	12	32	8
Hidden Dim.	768	768	4096	768
Feed Forward Dim.	3072	3072	11008	2048
Dropout Rate	0.1	0.1	0.0	0.3
Rank of LoRA	-	-	64	
Alpha of LoRA	-	-	16	
Dropout of LoRA	-	-	0.1	

Table 12: The details of the models used in our work.

compared with two representative PLMs, BERT and RoBERTa. In Table 13, we test all three NLU tasks, including intent classification, slot filling, and domain classification. The first two utterances are from M2M, while the rest are from MultiWOZ 2.2 (MWOZ). The first eight results come from **MIDAS (BERT)**, trained with three teachers $BERT_{ID}$, $BERT_{SF}$, and $BERT_{DC}$, and **BERT-only** refers a single fine-tuned BERT (*BERT-Base*), whereas the remaining five results pertain to **MIDAS (RoBERTa)**, trained with three teachers $RoBERTa_{ID}$, $RoBERTa_{SF}$, and $RoBERTa_{DC}$, and **RoBERTa-only** refers a single fine-tuned RoBERTa (*RoBERTa-Base*). Although the single fine-tuned BERT *(BERT-only)* or RoBERTa *(RoBERTa-only)* can sometimes predict the slots correctly, it does not communicate/integrate with the word level and domain level classification. Instances such as these validate our hypothesis that leveraging diverse knowledge levels from multi-turn conversation datasets can improve the understanding of individual natural language understanding tasks, outperforming the advantages of learning with single-level dialogue knowledge.

D.2 Compared with LLMs

D.2.1 Quantitative analysis

We measured the performance using the zero-shot prompt method to compare performance with Large Language Model (LLM). The LLM LLaMa2(Touvron et al., 2023), LLaMa3.1(Dubey et al., 2024), Gemma(Team et al., 2024), QWen2(Bai et al., 2023), GPT3.5 and GPT4o(Brown et al., 2020), were utilized. Notably, we also tested 3-shot learning on GPT4o, QWen2 and LLaMa3.1. The prompt for each task are shown in Table 15.

Table 14 presents the experimental results of each baseline, compared with the performance of our best model. Notably, GPT4o 3-shot learning achieves the best results in all tasks, except for the DC task on the M2M dataset, though it still falls significantly short of our model's performance. In the ID and SF tasks, LLaMa's performance is markedly lower than that of Gemma and GPT, highlighting that factors such as architecture, training data, and training methods, beyond just the number of parameters, also influence LLM performance.

Even within the LLaMa series, the number of model parameters doesn't always determine performance; the 7b model sometimes outperforms the 13b and 70b models. Note that only the 70b model was used with 4-bit quantisation.

Across all three tasks, LLMs occasionally generate out-of-scope class names, despite having all class names provided. Additionally, in the SF task, LLMs don't always output answers corresponding to the length of the original text. Despite our prompt stating that no explanation is needed for efficiency, LLMs sometimes still generate explanations. These observations indicate that LLMs don't fully grasp the input.

D.2.2 Qualitative analysis

In the qualitative analysis, we first focus on two representative LLMs, Gemma-7b and GPT3.5, as shown in Figure 3. From the M2M conversation shown in Figure 3-(a), we found that both LLMs can not predict slot types based on context. For example, GPT3.5 predicts "Michelin/B-RN rated/I-Rating" instead of "Michelin/B-Rating rated/I-Rating". Except for the wrong understanding of the conversation, we found that both LLMs can not follow the prompt all the time. For example, both LLMs do not predict the slot type for each token, where the missing predictions are represented by "N". From the Multi-Domain Wizard-of-Oz 2.2 (MWOZ) conversation as shown in Figure 3-(b), we can see that both LLMs can not make predictions in terms of the whole conversation, resulting the conflicts of the predictions of the domains and intents. For example, GPT3.5 predicts "Book Train" after "Book Restaurant" and Gemma-7b predicts "Find Restaurant" after "Book Restaurant". Another example is that both LLMs failed to predict the domain "Restaurant" of the last turn dialogue, even the Gemma-7b already predicted the intent as "Find Restaurant".

We further analyse the outputs from cutting-edge large language models, including QWen2, LLaMa3.1, and GPT-4o. As outlined in Section 5.7, these models also faced challenges in making use of multi-level knowledge consistently and strictly

No.	Model	Tokens (Slot)	Intent	Domain
1	**Utterance**	near, kirkland, and, i, don, ', t, care, about, the ratings	-	-
	Ground Truth	O, B-location, O, O, O, O, O, O, O, O	inform	restaurant
	MIDAS (BERT)	O, B-location, O, O, O, O, O, O, O, O	inform	restaurant
	BERT-Only	O, B-restaurant_name, O, O, O, O, O, O, O, O	infrom	movie
2	**Utterance**	let, ', s, go, with, the, view	-	-
	Ground Truth	O, O, O, O, O, B-restaurant_name, I-restaurant_name	affirm	restaurant
	MIDAS (BERT)	O, O, O, O, O, B-restaurant_name, I-restaurant_name	affirm	restaurant
	BERT-Only	O, O, O, O, O, O, O	affirm	movie
3	**Utterance**	then, find, me, one, in, the, expensive, price, range.	-	-
	Ground Truth	O, O, O, O, O, O, O, O, O	find_hotel	hotel
	MIDAS (BERT)	O, O, O, O, O, O, O, O, O	find_hotel	hotel
	BERT-Only	O, O, O, O, O, O, O, O, O	find_restaurant	restaurant
4	**Utterance**	which, ever, is, nice., i, will, need, some, info, on, it, too.	-	-
	Ground Truth	O, O, O, O, O, O, O, O, O, O, O, O	find_attraction	attraction
	MIDAS (BERT)	O, O, O, O, O, O, O, O, O, O, O, O	find_attraction	attraction
	BERT-Only	O, O, O, O, O, O, O, O, O, O, O, O	find_hotel	restaurant
5	**Utterance**	great, we, are, meeting, friends, at, wandlebury, country, park, before, we, eat,, can, you, tell, me, about, that, place, and, where, it, is?	-	-
	Ground Truth	O, O, O, O, O, O, B-attraction-name, I-attraction-name, I-attraction-name, O, O, O, O, O, O, O, O, O, O, O, O, O, O	find_attraction	attraction
	MIDAS (BERT)	O, O, O, O, O, O, B-attraction_name, I-attraction_name, I-attraction_name, O, O, O, O, O, O, O, O, O, O, O, O, O, O	find_attraction	attraction
	BERT-Only	O, O, O, O, O, O, B-attraction-name, I-hotel-name, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O	find_restaurant	restaurant
6	**Utterance**	yes, may, i, have, the, address, post,code, and, phone, number, for, golden, house?, i'll, book, it, myself.	-	-
	Ground Truth	O, O, O, O, O, O, O, O, O, O, O, B-restaurant_name, I-restaurant_name, O, O, O, O	find_restaurant	restaurant
	MIDAS (BERT)	O, O, O, O, O, O, O, O, O, O, O, B-restaurant_name, I-restaurant_name, O, O, O, O	find_restaurant	restaurant
	BERT-Only	O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O	find_hotel	hotel
7	**Utterance**	can, you, book, for, arrival, closer, to, 17:30, for, one, person, and, give, me, the, reference, number., also, i, would, like, to, see, a, college, in, centre.	-	-
	Ground Truth	O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O	find_attraction	attraction
	MIDAS (BERT)	O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O	find_attraction	attraction
	BERT-Only	O, O, O, O, O, O, O, B-train_leaveat, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O	book_train	train
8	**Utterance**	yes, please., i, need, an, address, and, phone, number, too.	-	-
	Ground Truth	O, O, O, O, O, O, O, O, O, O	find_attraction	attraction
	MIDAS (BERT)	O, O, O, O, O, O, O, O, O, O	find_attraction	attraction
	BERT-Only	O, O, O, O, O, O, O, O, O, O	find_restaurant	restaurant
9	**Utterance**	just, need, to, know, what, area, "its", in.	-	-
	Ground Truth	O, O, O, O, O, O, O, O	find_attraction	attraction
	MIDAS (RoBERTa)	O, O, O, O, O, O, O, O	find_attraction	attraction
	RoBERTa-Only	O, O, O, O, O, O, O, O	find_hotel	hotel
10	**Utterance**	i, would, actually, like, to, book, 5, people, and, would, like, to, know, the, reference, number, for, the, tickets, and, the, address, of, la, tasca, restaurant.	-	-
	Ground Truth	O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, B-restaurant-name, I-restaurant-name, O	find_restaurant	restaurant
	MIDAS (RoBERTa)	O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, B-restaurant-name, I-restaurant-name, O	find_restaurant	restaurant
	RoBERTa-Only	O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, B-restaurant-name, B-hotel-name, O	book_train	restaurant
11	**Utterance**	"its", not, a, restaurant,, "its", an, attraction., nusha.	-	-
	Ground Truth	O, O, O, O, O, O, O, B-attraction-name	find_attraction	attraction
	MIDAS (RoBERTa)	O, O, O, O, O, O, O, B-attraction-name	find_attraction	attraction
	RoBERTa-Only	O, O, O, O, O, O, O, O	find_attraction	restaurant
12	**Utterance**	will, you, give, me, the, phone, number,, address,, and, postcode, for, graffiti,, please?	-	-
	Ground Truth	O, O, O, O, O, O, O, O, O, O, O, B-restaurant-name, O	find_restaurant	restaurant
	MIDAS (RoBERTa)	O, O, O, O, O, O, O, O, O, O, O, B-restaurant-name, O	find_restaurant	restaurant
	RoBERTa-Only	O, O, O, O, O, O, O, O, O, O, O, O, O	find_attraction	restaurant
13	**Utterance**	i, am, looking, to, get, to, the, rajmahal, restaurant, please,, how, do, i, get, there?	-	-
	Ground Truth	O, O, O, O, O, O, O, B-restaurant_name, O, O, O, O, O, O, O	find_restaurant	restaurant
	MIDAS (RoBERTa)	O, O, O, O, O, O, O, B-restaurant_name, O, O, O, O, O, O, O	find_restaurant	restaurant
	RoBERTa-Only	O, O, O, O, O, O, O, B-restaurant_name, O, O, O, O, O, O, O	find_taxi	taxi

Table 13: 13 Prediction examples with both datasets on slot filling, intent detection, and domain classification results of each model. The first two utterances are from M2M, while the rest are from MultiWOZ 2.2 (MWOZ). The first eight results come from *MIDAS (BERT)* and BERT-Only, whereas the remaining five results pertain to *MIDAS (RoBERTa)* and RoBERTa-Only. The green cell represents a result that matches the ground truth, the red cell indicates incorrect results, and the yellow cell indicates partially correct results.

	ID		SF		DC	
	MWOZ (ACC)	M2M (ACC)	MWOZ (F1)	M2M (F1)	MWOZ (ACC)	M2M (ACC)
LLaMa2-7b-chat	0.4751	0.3363	0.0217	0.0751	0.6528	0.5231
LLaMa2-13b-chat	0.1679	0.2013	0.0891	0.1092	0.5602	0.4468
LLaMa2-70b-chat	0.3896	0.3275	0.0619	0.0883	0.6987	0.6012
Gemma-7b	0.6515	0.4588	0.6653	0.4357	0.7227	0.5426
GPT3.5	0.6971	0.5100	0.8175	0.5516	0.7739	0.7740
GPT4o	0.6789	0.6410	0.8418	0.6616	0.7877	0.8503
GPT4o†	0.7614	0.7510	0.8525	0.7132	0.7941	0.7051
QWen2-7B-Ins.†	0.5459	0.278	0.1532	0.1192	0.6416	0.6541
LLaMa3.1-8b-Ins.†	0.6422	0.2715	0.6276	0.5412	0.5973	0.5076
Our best model	**0.8502**	**0.9427**	**0.9928**	**0.9856**	**0.8816**	**0.8952**

Table 14: The comparison of the proposed models with prompt tuning methods using Large Language Models. ID, SF and DC indicate intent detection, slot filling and domain classification, respectively, as mentioned in Section 4.3. ACC and F1 stand for accuracy and micro F1, respectively, and scores in bold indicate leadership among the metrics. † refers to the application of 3-shot learning in the prompt with LLMs.

adhering to prompt instructions. Those examples from both M2M dataset and MultiWOZ dataset are presented from Table 16 to Table 21.

D.3 Complete procedure cases

In this section, we provide a complete question-answer demonstration for each task and dataset, offering a clearer explanation of how we utilize LLMs for testing, as shown from Figure 4 to Figure 9.

Case 4, Intent Detection, MultiWOZ

```
Definition: In this task, you are given a dialogue. Your job is to
classify the following dialogue into one of the eleven different
intents. The intents and examples are:
Intent: find_restaurant
Dialogues:
1. i am looking for a restaurant called the gandhi.
2. i want crossover food at an expensive price. can you help?
3. is there something else in the west?
Intent: book_restaurant
Dialogues:
...
Input: [yes.,book,for,7,people,on,saturday,at,15:15.]. Output(only
output the intent):
```

Case 1 – Our Model

Book Restaurant

Case 1 - GPT-4 With Demonstration

Book Train

Figure 4: Prompt and output for a sample dialogue in **MultiWOZ** dataset, where the correct prediction is highlighted in green and wrong predictions are highlighted in red. Demonstration means few-shot (3-shot) learning. Compared to GPT4o, our model can **correctly** classify the intent of the given dialogue as **Book Restaurant**.

Case 3, Intent Detection, M2M

```
Definition: In this task, you are given a dialogue. Your job is to
classify the following dialogue into one of the fifteen different
intents. The intents and examples are:
Intent: greeting
Dialogues:
1. hi , i want to make a restaurant reservation .
2. hi , i want to make a reservation for 6 pm .
3. hi ! can you book me a restaurant reservation ?
Intent: request
Dialogues:
...
Input:
[i,found,the,following,restaurants,:,the,hudson,room,,,boats,and,casca
l]. Output(only output the intent):
```

Case 1 – Our Model

Select

Case 1 - GPT-4 With Demonstration

Inform

Figure 5: Prompt and output for a sample dialogue in **M2M** dataset, where the correct prediction is highlighted in green and wrong predictions are highlighted in red. Demonstration means few-shot (3-shot) learning. Compared to GPT4o, our model can **correctly** classify the intent of the given dialogue as **Select**.

Case 6, Slot Filling, MultiWOZ

```
"In the task of slot filling, the B-, I-, and O- prefixes are
commonly used to annotate slot types, indicating the boundaries and
types of slots. These labels typically represent:
B- (Begin): Signifies the beginning of a slot, marking the start of a
new slot. I- (Inside): Represents the interior of a slot, indicating
a continuation of the slot. O (Outside): Denotes parts of the input
that are not part of any slot.
For instance, in a sentence where we want to label a ""date"" slot,
words containing date information might be tagged as ""B-date""
(indicating the beginning of a date slot), followed by consecutive
words carrying date information tagged as ""I-date"" (indicating the
continuation of the date slot), while words not containing date
information would be tagged as ""O"" (indicating they are outside any
slot). Here are some examples:
Dialogue: ""i am looking for a restaurant called the gandhi."", slot
types: ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-restaurant-name', 'I-
restaurant-name']
...
Definition: In this task, you are given a dialogue. Your job is to
classify the words in the following dialogue into one of the thirty
different slots. The slots are: "B-attraction-name", ..., "O". Input:
[yes,i,would,like,it,to,serve,asian,oriental]. Output(Only output
slot types. And the slot types should be output as a list without any
explanation):"
```

Case 1 – Our Model

['O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-restaurant-food', 'I-restaurant-food']

Case 1 - GPT-4 With Demonstration

['O', 'O', 'O', 'O', 'O', 'O', 'O', 'I-restaurant-food, 'O']

Figure 6: Prompt and output for a sample dialogue in **MultiWOZ** dataset, where the correct prediction is highlighted in green and wrong predictions are highlighted in red. Demonstration means few-shot (3-shot) learning. Compared to GPT4o, our model can **correctly** classify the slot types of the given dialogue.

Task	Prompt
ID	Definition: In this task, you are given a dialogue. Your job is to classify the following dialogue into one of the fifteen different intents. The intents and examples are: Intent: greeting Dialogues: 1. hi , i want to make a restaurant reservation . 2. hi , i want to make a reservation for 6 pm . 3. hi ! can you book me a restaurant reservation ? Intent: request Dialogues: 1. okay , where do you want to go , and how many people will there be? ... Input: [{input}]. Output(only output the intent):
SF	In the task of slot filling, the B-, I-, and O- prefixes are commonly used to annotate slot types, indicating the boundaries and types of slots. These labels typically represent: B- (Begin): Signifies the beginning of a slot, marking the start of a new slot. I- (Inside): Represents the interior of a slot, indicating a continuation of the slot. O (Outside): Denotes parts of the input that are not part of any slot. For instance, in a sentence where we want to label a "date" slot, words containing date information might be tagged as "B-date" (indicating the beginning of a date slot), followed by consecutive words carrying date information tagged as "I-date" (indicating the continuation of the date slot), while words not containing date information would be tagged as "O" (indicating they are outside any slot). Here are some examples: Dialogue: "the sushi boat for 6 .", slot types: ['O', 'B-restaurant_name', 'I-restaurant_name', 'O', 'B-num_people', 'O'] ... Definition: In this task, you are given a dialogue. Your job is to classify the words in the following dialogue into one of the twenty-one different slots. The slots are: "B-category", "B-date", "B-location", ..., "O". Input: [{input}]. Output(Only output slot types. And the slot types should be output as a list without any explanation):
DC	In this task, you are given a dialogue. Your job is to classify the following dialogue into one of the two different intents. The domains and examples are: Domain: restaurant Dialogues: 1. hi , i want to make a restaurant reservation. 2. a reservation for cheese cake factory for 3 people on next monday. 3. ok , please choose between amarin and sakoon restaurants. Domain: movie Dialogues: 1. i would like to buy movie tickets for 6:00 pm 2. which movie , and how many tickets do you need ? 3. i need 3 tickets for the movie called a man called love Input: [{input}]. Output(only output the domain):

Table 15: The prompt we used for each dataset in our experiments.

Case 5, Slot Filling, M2M

```
In the task of slot filling, the B-, I-, and O- prefixes are commonly
used to annotate slot types, indicating the boundaries and types of
slots. These labels typically represent:
B- (Begin): Signifies the beginning of a slot, marking the start of a
new slot. I- (Inside): Represents the interior of a slot, indicating
a continuation of the slot. O (Outside): Denotes parts of the input
that are not part of any slot.
For instance, in a sentence where we want to label a "date" slot,
words containing date information might be tagged as "B-date"
(indicating the beginning of a date slot), followed by consecutive
words carrying date information tagged as "I-date" (indicating the
continuation of the date slot), while words not containing date
information would be tagged as "O" (indicating they are outside any
slot). Here are some examples:
...
Dialogue: "the sushi boat for 6 .", slot types: ['O', 'B-
restaurant_name', 'I-restaurant_name', 'O', 'B-num_people', 'O']
...
Definition: In this task, you are given a dialogue. Your job is to
classify the words in the following dialogue into one of the twenty-
one different slots. The slots are: "B-category", ..., "O". Input:
[you,can,choose,from,the,hudson,room,,,deep,blue,,,or,sakoon].
Output(Only output slot types. And the slot types should be output as
a list without any explanation):
```

Case 1 – Our Model

['O', 'O', 'O', 'O', 'B-restaurant_name', 'I-restaurant_name', 'I-restaurant_name', 'O', 'B-restaurant_name', 'I-restaurant_name', 'O', 'O', 'B-restaurant_name']

Case 1 - GPT-4 With Demonstration

['O', 'O', 'O', 'O', 'B-restaurant_name', 'I-restaurant_name', 'O', 'B-restaurant_name', 'I-restaurant_name', 'O', 'O', 'B-restaurant_name', 'O']

Figure 7: Prompt and output for a sample dialogue in **M2M** dataset, where the correct prediction is highlighted in green and wrong predictions are highlighted in red. Demonstration means few-shot (3-shot) learning. Compared to GPT4o, our model can **correctly** classify the slot types of the given dialogue.

Case 1, Domain Classification, M2M

```
Definition: In this task, you are given a dialogue. Your job is to
classify the following dialogue into one of the two different
intents. The domains and examples are:
Domain: restaurant
Dialogues:
1. hi , i want to make a restaurant reservation .
2. okay , where do you want to go , and how many people will there be
?
3. the sushi boat for 6 .
Domain: movie
Dialogues:
1. i would like to buy movie tickets for 6:00 pm
2. which movie , and how many tickets do you need ?
3. i need 3 tickets for the movie called a man called ove
Input: [what,type,of,food,and,price,range,should,i,look,for,?].
Output(only output the domain):
```

Case 1 – Our Model

Restaurant

Case 1 - GPT-4 With Demonstration

Restaurant

Figure 8: Prompt and output for a sample dialogue in **M2M** dataset, where the correct prediction is highlighted in green and wrong predictions are highlighted in red. Demonstration means few-shot (3-shot) learning. Both models can **correctly** classify the domain of the given dialogue as **Restrurant**.

Case 2, Domain Classification, MultiWOZ

```
Definition: In this task, you are given a dialogue. Your job is to
classify the following dialogue into one of the eight different
intents. The domains and examples are:
Domain: restaurant
Dialogues:
1. i am looking for a restaurant called the gandhi.
2. i'd like a reservation for 7 people monday at 15:30 please.
3. i want crossover food at an expensive price. can you help?
Domain: police
Dialogues:
...
Input:
[i,would,like,to,book,a,table,for,3,on,saturday,at,13:30.,are,there,a
ny,trains,that,go,to,broxbourne,on,sunday?]. Output(only output the
domain):
```

Case 1 – Our Model

Restaurant

Case 1 - GPT-4 With Demonstration

Train

Figure 9: Prompt and output for a sample dialogue in **MultiWOZ** dataset, where the correct prediction is highlighted in green and wrong predictions are highlighted in red. Demonstration means few-shot (3-shot) learning. Compared to GPT4o, our model can **correctly** classify the domain of the given dialogue as **Restaurant**.

Model	Turn	Tokens (Slot)	Intent	Domain
Utterance	**1**	what, are, you, in, the, mood, for, ?	-	-
	2	it, doesn, , t, matter, .	-	-
	3	how, about, the, hudson, room, or, los, altos, grill, ?	-	-
	4	sounds, good, .	-	-
GT	**1**	O, O, O, O, O, O, O, O	request	restaurant
	2	O, O, O, O, O, O	inform	restaurant
	3	O, O, B-restaurant_name, I-restaurant_name, I-restaurant_name, O, B-restaurant_name, I-restaurant_name, I-restaurant_name, O	select	restaurant
	4	O, O, O	affirm	restaurant
MIDAS (Ours)	**1**	O, O, O, O, O, O, O, O	request	restaurant
	2	O, O, O, O, O, O	inform	restaurant
	3	O, O, B-restaurant_name, I-restaurant_name, I-restaurant_name, O, B-restaurant_name, I-restaurant_name, I-restaurant_name, O	select	restaurant
	4	O, O, O	affirm	restaurant
GPT4o	**1**	O, O, O, O, O, O, NaN, NaN	request	restaurant
	2	O, O, O, O, O, NaN	negate	(O)
	3	O, O, O, B-restaurant_name, I-restaurant_name, O, B-restaurant_name, I-restaurant_name, O, NaN	select	restaurant
	4	O, O, O	affirm	movie
QWen2	**1**	B-category, NaN, NaN, NaN, NaN, NaN, NaN, NaN	request_alts	restaurant
	2	O, O, NaN, NaN, NaN, NaN	negate	restaurant
	3	B-restaurant_name, I-restaurant_name, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN	request_alts	restaurant
	4	O, O, NaN	inform	movie
LLaMa3.1	**1**	O, O, O, O, O, NaN, NaN, NaN	request	movie
	2	O, O, O, O, O, O	other	movie
	3	O, O, O, O, O, O, O, O, O, O	affirm	movie
	4	O, O, O	other	restaurant

Table 16: A four turns conversation from M2M dataset. GPT4o exhibited issues with generating out-of-domain outputs '(O)' in Turn 2, while all LLMs showed problems in the SF task by failing to follow instructions, resulting in NaN outputs.

Model	Turn	Tokens (Slot)	Intent	Domain
Utterance	**1**	what, area, is, your, preference, ?	-	-
	2	orlando	-	-
	3	i, have, the, hudson, room, ,, boats, or, the, nest, ,, which, sounds, better, ?	-	-
	4	agree, on, the, hudson, room	-	-
GT	**1**	O, O, O, O, O, O	request	restaurant
	2	B-location	inform	restaurant
	3	O, O, B-restaurant_name, I-restaurant_name, I-restaurant_name, O, B-restaurant_name, O, B-restaurant_name, I-restaurant_name, O, O, O, O, O	select	restaurant
	4	O, O, B-restaurant_name, I-restaurant_name, I-restaurant_name	affirm	restaurant
MIDAS (Ours)	**1**	O, O, O, O, O, O	request	restaurant
	2	B-location	inform	restaurant
	3	O, O, B-restaurant_name, I-restaurant_name, I-restaurant_name, O, B-restaurant_name, O, B-restaurant_name, I-restaurant_name, O, O, O, O, O	select	restaurant
	4	O, O, B-restaurant_name, I-restaurant_name, I-restaurant_name	affirm	restaurant
GPT4o	**1**	O, O, O, O, O, NaN	request	restaurant
	2	B-location	(O)	(O)
	3	O, O, O, B-restaurant_name, I-restaurant_name, O, O, O, O, O, B-restaurant_name, O, O, O, O	select	restaurant
	4	O, O, B-restaurant_name, I-restaurant_name, NaN	inform	restaurant
QWen2	**1**	B-location, NaN, NaN, NaN, NaN, NaN	request_alts	movie
	2	B-location	inform	movie
	3	B-location, B-location, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN	request_alts	movie
	4	B-restaurant_name, I-location, NaN, NaN, NaN	select	restaurant
LLaMa3.1	**1**	O, O, O, O, O, NaN	greeting	movie
	2	O	confirm	movie
	3	O, O, O, B-location, O, B-movie, O, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN	request_alts	movie
	4	O, O, O, B-location, I-location	thank_you	movie

Table 17: A four turns conversation from M2M dataset. GPT4o exhibited issues with generating out-of-domain outputs '(O)' in Turn 2, while all LLMs showed problems in the SF task by failing to follow instructions, resulting in NaN outputs.

Model	Turn	Tokens (Slot)	Intent	Domain
Utterance	1	how, many, tickets, would, you, like, to, buy, ?	-	-
	2	1	-	-
	3	what, date, do, you, want, to, go, ?	-	-
GT	1	O, O, O, O, O, O, O, O, O	request	movie
	2	B-num_tickets	inform	movie
	3	O, O, O, O, O, O, O, O	request	movie
MIDAS (Ours)	1	O, O, O, O, O, O, O, O, O	request	movie
	2	B-num_tickets	inform	movie
	3	O, O, O, O, O, O, O, O	request	movie
GPT4o	1	O, O, O, O, O, O, O, NaN, NaN	request	movie
	2	(O)	(O)	restaurant
	3	O, B-date, O, O, O, O, NaN, NaN	request	restaurant
QWen2	1	B-num_tickets, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN	request_alts	movie
	2	B-category	inform	movie
	3	B-date, NaN, NaN, NaN, NaN, NaN, NaN, NaN	request_alts	restaurant
LLaMa3.1	1	O, O, O, O, O, O, B-num_tickets, O, NaN	request	restaurant
	2	B-category	request	movie
	3	O, O, O, O, O, O, O, O	request	movie

Table 18: A three turns conversation from M2M dataset. GPT4o exhibited issues with generating out-of-domain outputs '(O)' in Turn 2, while all LLMs showed problems in the SF task by failing to follow instructions, resulting in NaN outputs. LLaMa3.1 failed to unify multi-level knowledge, resulting in predicting the domain as 'restaurant' even after identifying 'B-num_tickets'

Model	Turn	Tokens (Slot)	Intent	Domain
Utterance	1	1, pm, is, unavailable, again, ,, please, choose, from, 6, pm, or, 6.30, pm, .	-	-
	2	no, ,, thank, you	-	-
	3	please, provide, the, theatre, name, ,, date, ,, and, number, of, tickets	-	-
	4	the, theatre, name, is, aquarius, and, i, need, 1, ticket, for, friday, .	-	-
	5	what, movie, would, you, like, to, see, ?	-	-
GT	1	B-time, I-time, O, O, O, O, O, O, O, B-time, I-time, O, B-time, I-time, O	negate	movie
	2	O, O, O, O	negate	movie
	3	O, O, O, O, O, O, O, O, O, O, O, O	request	movie
	4	O, O, O, O, B-theatre_name, O, O, O, B-num_tickets, O, O, B-date, O	inform	movie
	5	O, O, O, O, O, O, O, O	request	movie
MIDAS (Ours)	1	B-time, I-time, O, O, O, O, O, O, O, B-time, I-time, O, B-time, I-time, O	negate	movie
	2	O, O, O, O	negate	movie
	3	O, O, O, O, O, O, O, O, O, O, O, O	request	movie
	4	O, O, O, O, B-theatre_name, O, O, O, B-num_tickets, O, O, B-date, O	inform	movie
	5	O, O, O, O, O, O, O, O	request	movie
GPT4o	1	B-time, I-time, O, O, O, O, O, O, O, B-time, I-time, O, O, B-time, I-time	negate	movie
	2	O, O, O, NaN	negate	(O)
	3	O, O, O, B-theatre_name, I-theatre_name, O, B-date, O, O, B-num_tickets, O, NaN	request	movie
	4	O, B-theatre_name, I-theatre_name, O, O, O, O, B-num_tickets, O, O, B-date, O, NaN	inform	movie
	5	O, B-movie, O, O, O, O, O, NaN	request	movie
QWen2	1	B-time, I-time, I-time, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN	request_alts	(O)
	2	O, O, NaN, NaN	negate	restaurant
	3	B-theatre_name, B-date, B-num_tickets, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN	request_alts	movie
	4	B-theatre_name, B-date, B-num_tickets, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN	request	movie
	5	B-movie, NaN, NaN, NaN, NaN, NaN, NaN, NaN	request_alts	movie
LLaMa3.1	1	O, O, O, B-time, I-time, O, O, O, O, O, O, O, O, O, O	request	restaurant
	2	O, O, NaN, NaN	thank_you	restaurant
	3	B-theatre_name, I-date, B-num_tickets, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN	other	movie
	4	B-theatre_name, I-theatre_name, O, O, O, B-num_tickets, (O)I-num_tickets, O, B-date, I-date, O, NaN, NaN	request	movie
	5	O, O, O, O, O, O, O, O	request	movie

Table 19: A five turns conversation from M2M dataset. GPT4o, QWen2 and LLaMa3.1 exhibited issues with generating out-of-domain outputs '(O)' in Turn 2, Turn 1 and Turn 4 respectively, while all LLMs showed problems in the SF task by failing to follow instructions, resulting in NaN outputs.

Model	Turn	Tokens (Slot)	Intent	Domain
Utterance	1	i, choose, the, ashley, hotel., what, is, their, address,, please?	-	-
	2	no,, i, just, need, the, address.	-	-
	3	hey, i, am, looking, for, a, train, from, cambridge, to, bishops, stortford., mind, helping, out?	-	-
	4	i, want, to, leave, on, monday, and, arrive, by, 18:45.	-	-
	5	could, you, give, me, the, travel, time, and, price, of, that, train, please?	-	-
	6	i, also, want, a, cheap, chinese, restaurant.	-	-
	7	id, like, to, be, in, the, centre, please.	-	-
	8	yes., monday,, 8, people,, 10:30.	-	-
	9	i, am, planning, a, trip, in, cambridge	-	-
GT	1	O, O, O, B-hotel-name, I-hotel-name, O, O, O, O, O	find_hotel	hotel
	2	O, O, O, O, O, O	find_hotel	hotel
	3	O, O, O, O, O, O, O, O, O, O, O, O, O, O, O	find_train	train
	4	O, O, O, O, O, O, O, O, O, B-train-arriveby	find_train	train
	5	O, O, O, O, O, O, O, O, O, O, O, O, O	find_train	train
	6	O, O, O, O, O, B-restaurant-food, O	find_restaurant	restaurant
	7	O, O, O, O, O, O, O, O	find_restaurant	restaurant
	8	O, O, O, O, B-restaurant-booktime	book_restaurant	restaurant
	9	O, O, O, O, O, O, O	find_train	train
MIDAS (Ours)	1	O, O, O, B-hotel-name, I-hotel-name, O, O, O, O, O	find_hotel	hotel
	2	O, O, O, O, O, O	find_hotel	hotel
	3	O, O, O, O, O, O, O, O, O, O, O, O, O, O, O	find_train	train
	4	O, O, O, O, O, O, O, O, O, B-train-arriveby	find_train	train
	5	O, O, O, O, O, O, O, O, O, O, O, O, O	find_train	train
	6	O, O, O, O, O, B-restaurant-food, O	find_restaurant	restaurant
	7	O, O, O, O, O, O, O, O	find_restaurant	restaurant
	8	O, O, O, O, B-restaurant-booktime	book_restaurant	restaurant
	9	O, O, O, O, O, O, O	find_train	train
GPT4o	1	O, O, O, B-hotel-name, I-hotel-name, O, O, O, O, O	find_hotel	hotel
	2	O, O, O, O, O, NaN	find_police	police
	3	O, O, O, O, O, O, O, (O)B-train-departure, O, (O)B-train-destination, (O)I-train-destination, O, O, O, O	find_train	train
	4	O, O, O, O, O, O, B-train-arriveby, I-train-arriveby, NaN, NaN	find_train	train
	5	O, O, O, O, O, O, O, O, O, O, O, O, O	find_train	train
	6	O, O, O, O, B-restaurant-food, I-restaurant-food, O	find_restaurant	restaurant
	7	O, O, O, O, O, O, O, NaN	find_hotel	hotel
	8	O, O, O, B-restaurant-booktime, O	book_restaurant	train
	9	O, O, O, O, O, B-bus-destination, NaN	find_attraction	attraction
QWen2	1	B-hotel-name, I-hotel-name, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN	find_hotel	hotel
	2	O, O, NaN, NaN, NaN, NaN	find_hotel	hotel
	3	(O)B-train-destination, (O)B-train-origin, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN	book_train	train
	4	B-train-leaveat, (O)I-train-arrivevi, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN	book_train	hotel
	5	O, O, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN	find_train	taxi
	6	I-restaurant-food, NaN, NaN, NaN, NaN, NaN, NaN	find_restaurant	restaurant
	7	I-hotel-stars, NaN, NaN, NaN, NaN, NaN, NaN, NaN	find_hotel	hotel
	8	B-train-leaveat, I-restaurant-booktime, NaN, NaN, NaN	book_train	hotel
	9	(O)B-city, B-hotel-name, NaN, NaN, NaN, NaN, NaN	find_attraction	hotel
LLaMa3.1	1	O, O, O, O, B-hotel-name, O, O, O, O, O	find_hotel	hotel
	2	B-restaurant-name, I-restaurant-name, O, O, NaN, NaN	find_hotel	restaurant
	3	B-train-leaveat, O, (O)B-train-destination, O, O, O, O, O, O, O, O, O, O, O, O	find_train	train
	4	B-train-leaveat, O, O, B-train-arriveby, O, O, O, O, O, O	book_train	train
	5	B-train-arriveby, B-train-leaveat, O, O, O, O, O, O, O, O, O, O, O	find_train	train
	6	O, O, O, B-restaurant-food, O, O, O	find_restaurant	restaurant
	7	O, O, O, O, O, O, O, O	find_taxi	attraction
	8	B-attraction-name, I-attraction-name, O, O, O	find_restaurant	bus
	9	O, O, O, O, O, O, O	find_attraction	restaurant

Table 20: A nine turns conversation from MultiWOZ dataset. All LLMs showed problems generating out-of-domain outputs '(O)' and problems in the SF task by failing to follow instructions, resulting in NaN outputs.

Model	Turn	Tokens (Slot)	Intent	Domain
Utterance	1	yes, i, am, looking, for, a, place, to, stay, in, cambridge, that, is, 3, stars, and, expensive., can, you, help, me?	-	-
	2	do, those, both, have, 3, star, ratings, and, are, expensive?	-	-
	3	do, they, have, free, parking?	-	-
	4	not, today, thanks., im, also, want, to, find, a, cinema, in, the, west, part, of, town.	-	-
	5	are, there, any, colleges, in, the, west, that, i, could, visit, instead?	-	-
	6	what, is, the, one, that, is, free?, can, i, get, the, phone, number, and, postcode?	-	-
GT	1	O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O	find_hotel	hotel
	2	O, O, O, O, O, O, O, O, O, O	find_hotel	hotel
	3	O, O, O, O, O	find_hotel	hotel
	4	O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O	find_attraction	attraction
	5	O, O, O, O, O, O, O, O, O, O, O, O	find_attraction	attraction
	6	O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O	find_attraction	attraction
MIDAS (Ours)	1	O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O	find_hotel	hotel
	2	O, O, O, O, O, O, O, O, O, O	find_hotel	hotel
	3	O, O, O, O, O	find_hotel	hotel
	4	O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O	find_attraction	attraction
	5	O, O, O, O, O, O, O, O, O, O, O, O	find_attraction	attraction
	6	O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O	find_attraction	attraction
GPT4o	1	O, O, O, O, O, O, O, O, O, O, O, O, O, B-hotel-stars, I-hotel-stars, O, O, O, O, O, NaN	find_hotel	hotel
	2	O, O, O, O, B-hotel-stars, I-hotel-stars, O, O, O, NaN	find_restaurant	restaurant
	3	O, O, O, O, O	find_hotel	hotel
	4	O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O	find_attraction	attraction
	5	O, O, O, O, O, O, O, O, O, O, NaN, NaN	find_attraction	attraction
	6	O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O	find_attraction	attraction
QWen2	1	B-hotel-name, B-hotel-stars, (O)B-price-range, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN	find_hotel	hotel
	2	B-hotel-stars, B-hotel-type, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN	find_hotel	restaurant
	3	O, O, NaN, NaN, NaN	find_hotel	restaurant
	4	B-attraction-name, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN	find_attraction	attraction
	5	O, O, O, O, O, O, O, O, O, O, O, O	find_attraction	attraction
	6	O, O, O, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN	find_attraction	restaurant
LLaMa3.1	1	B-hotel-stars, I-hotel-stars, O, B-hotel-type, O, B-hotel-stars, I-hotel-stars, O, B-hotel-type, O, B-hotel-stars, I-hotel-stars, O, B-hotel-type, O, NaN, NaN, NaN, NaN, NaN, NaN	book_hotel	hotel
	2	O, O, O, O, O, O, O, O, O, O	find_hotel	restaurant
	3	O, O, NaN, NaN, NaN	find_hotel	hotel
	4	O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O	find_attraction	restaurant
	5	O, O, O, O, O, O, O, O, O, O, O, O	find_attraction	attraction
	6	O, O, O, O, O, O, O, O, O, O, O, O, O, O, O, O	find_hotel	hotel

Table 21: A six turns conversation from MultiWOZ dataset. QWen2 exhibited issues with generating out-of-domain outputs ‘(O)’ in Turn 1, while all LLMs showed problems in the SF task by failing to follow instructions, resulting in NaN outputs.