ECLM: Entity Level Language Model for Spoken Language Understanding with Chain of Intent

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have shown remarkable success in language generation, demonstrating broad competence across different tasks. However, their direct application to spoken language understanding (SLU) remains challenging. This is particularly true for token-level tasks, where the autoregressive architecture of LLMs can lead to error propagation and misalignment problems. In this paper, we present the Entity-level Language Model (ECLM) framework for SLU, which addresses these challenges by transforming the traditional token-level slot-filling task into an entity recognition problem. In addition, we propose a novel concept, "Chain of Intent", which enables LLMs to effectively handle multi-intent recognition in a step-by-step manner. Our experiments demonstrate that ECLM achieves substantial improvements over state-of-the-art pre-trained models like Uni-MIS, with overall accuracy gains of 3.7% on the MixATIS dataset and 3.1% on the MixSNIPS dataset. Moreover, the ECLM framework surpasses conventional supervised fine-tuning of LLMs, delivering improvements of 8.5% and 21.2% on MixATIS and MixSNIPS, respectively.

1 Introduction

The rapid advancement of large language models (LLMs) has markedly accelerated progress in the field of natural language processing (NLP) (Google., 2023; Touvron et al., 2023). Trained on extensive datasets, these models demonstrate exceptional performance across a wide range of NLP tasks, including natural language inference, summarization, and dialog systems, often achieving impressive results through in-context learning alone (Kavumba et al., 2023; Hu et al., 2022).

Spoken language understanding (SLU) is a critical component of task-oriented dialog systems, which are designed to construct a semantic frame that accurately captures the user's request. This semantic frame is typically built through two sub-tasks: intent detection, which identifies the user's intent, and slot filling, which extracts relevant semantic elements. Given the close interdependence of these sub-tasks (Tur and Mori, 2011), state-of-the-art SLU systems often employ joint models to effectively capture the correlations between them (Goo et al., 2018; Qin et al., 2019).

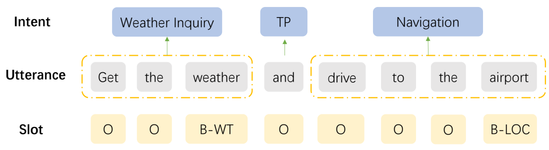

Figure 1: An example with Multi-Intent SLU, where B-WT donates B-Weather, B-LOC donates B-Location and "TP" denote "Transition Point".

In real-life scenarios, users often express multiple intents within a single utterance, and the Amazon internal dataset showed that 52% of examples are multi-intent (Gangadharaiah and Narayanaswamy, 2019). Figure 1 shows a two-intent example, which contains a classification task to classify the intent labels (i.e., predict the intents as : `Weather_Inquiry` and `Navigation`) and a sequence labeling task to predict the slot label sequence (i.e., label the utterance as {`O`, `O`, `B-WT`, `O`, `O`, `O`, `O`, `B-LOC` }). To deal with multi-intent scenarios, an increasing number of studies have begun to focus on modeling SLU in multi-intent settings. Xu and Sarikaya (2013) and Kim et al. (2017) first explored the multi-intent SLU. Then Qin et al. (2020a, 2021b) incorporated graph attention networks to model fine-grained intent-slot guiding. Recently, Huang et al. (2022) proposed a chunk-level intent detection (CLID) framework to split multi-intent into single-intent with an intent transition point. Furthermore, Yin et al. (2024) develop an united multi-view intent-slot interaction framework(Uni-MIS), achieving promising performance.

Whether LLMs can effectively handle multi-intent SLU remains an open question. While a straightforward approach might involve fine-tuning LLMs for this specific task, several challenges persist. For example, although LLMs exhibit strong capabilities in entity-level intent detection, their autoregressive architecture can lead to issues such as error propagation and misalignment, particularly in token-level slot filling tasks. This is because LLMs may generate undesirable outputs that do not align one-to-one with the original tokens from the utterance.

To address these challenges, we introduce a novel method that leverages the strengths of LLMs for multi-intent SLU by transforming the traditional token-level slot-filling task into an entity detection problem. By shifting the focus to entity-level slot detection, LLMs can concentrate on identifying relevant slot labels without the need to label every token within a sentence. This approach effectively mitigates the issues of misalignment and uncontrolled generation length. Moreover, we propose the concept of a **chain of intent**, inspired by the chain-of-thought reasoning framework (Wei et al., 2022). This strategy enhances the ability of LLMs to differentiate and separate multi-intent utterances into distinct sub-intent segments, enabling the models to handle multi-intent recognition in a systematic, step-by-step manner.

Our experimental results demonstrate that ECLM achieves substantial improvements over state-of-the-art pre-trained models, such as Uni-MIS. Specifically, ECLM achieves overall accuracy gains of 3.7% on the MixATIS dataset and 3.1% on the MixSNIPS dataset. Furthermore, the ECLM framework surpasses conventional supervised fine-tuning of LLMs, delivering improvements of 8.5% and 21.2% in overall accuracy on MixATIS and MixSNIPS, respectively. In terms of slot filling F1 score, ECLM outperforms vanilla LLM fine-tuning by 22% and 8.1%. We also conduct further experiments to evaluate the performance of ECLM across different numbers of intents within the datasets. Our model consistently outperforms Uni-MIS in overall accuracy across all settings, particularly in scenarios with a high number of intents, showing improvements of 1.1%, 4.3%, and 7.8% for intent counts ranging from 1 to 3. Additionally, we find that ECLM requires only 60% of the data to surpass Uni-MIS, with more training further enhancing its performance. In summary, the contributions of this work can be outlined as follows: (1) We design an entity-slot framework that transforms the traditional token-level slot-filling task into an entity detection problem, thereby mitigating issues of misalignment and uncontrolled generation length. (2) We introduce the chain of intent concept, which enables LLMs to effectively handle multi-intent recognition in a step-by-step manner. (3) We demonstrate that our proposed model, ECLM, outperforms strong baselines on two widely used datasets, MixATIS and MixSNIPS, across the majority of metrics.

2 Problem Definition

2.1 Multi-Intent Detection

Given an input sequence $x = (x_1, \ldots, x_n)$, multi-intent detection can be defined as a multi-label classification task that outputs a sequence of intent labels $o_I = (o_1^I, \ldots, o_m^I)$, where m is the number of intents in a given discourse and n is the length of the discourse.

2.2 Slot Filling

Slot filling can be considered as a sequence annotation task that maps the input discourse x to a slot output sequence $o_S = (o_1^S, \ldots, o_n^S)$.

3 Approach

As depicted in Figure 2, our methodology establishes a comprehensive framework for integrating large language models (LLMs) into the domain of multi-intent spoken language understanding (SLU). The left side of the figure illustrates the prompt structure used for training ECLM, alongside standard supervised fine-tuning (SFT) prompts. On the right, we present an example of the ECLM training process, highlighting the key components: the Entity Slot and the Chain of Intent. Finally, we perform supervised fine-tuning to adapt the LLM to the multi-intent SLU task.

3.1 Entity Slots Construction and Recovery

Our approach introduces a novel two-phase process: Entity Slots Construction for training, and Entity Slots Recovery for inference, designed to bridge the gap between traditional sequence labeling and the generative capabilities of large language models (LLMs).

3.1.1 Entity Slots Construction

In the Entity Slots Construction phase, we transform conventional BIO sequence labeling into a

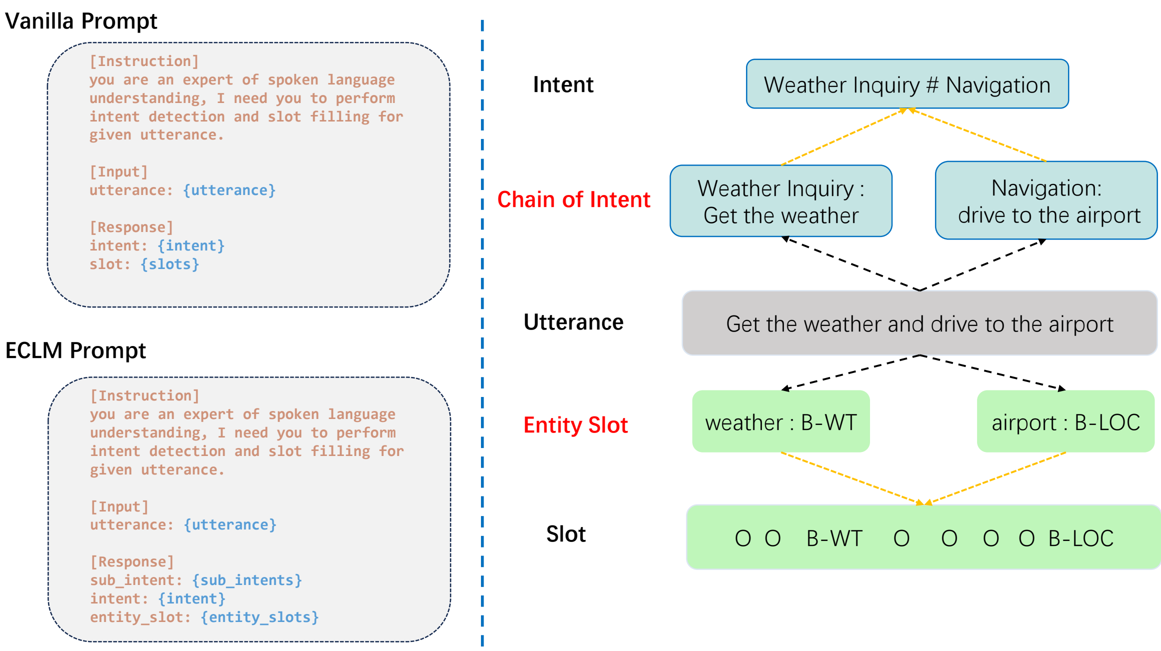

Figure 2: Brief introduction of the workflow of ECLM. The left shows the prompt structure for ECLM training and vanilla SFT prompts. The right illustrates an example training process of ECLM.

structured entity-slot representation, optimizing for generative modeling with LLMs. Given a token sequence $T = \{t_1, t_2, \ldots, t_n\}$ and its corresponding BIO-annotated slots $S = \{s_1, s_2, \ldots, s_n\}$, we map these to a set of entity slots $E = \{e_1, e_2, \ldots, e_m\}$, where m is the number of identified entities. This mapping is defined by a function c as follows:

$$c(T, S) = \left\{ \left(k_i, \bigcup_{j \in I_i} t_j \right) \right\}_{i=1}^{m}, \quad (1)$$

where k_i is the entity type derived from the 'B-' tag, and I_i is the index set of tokens corresponding to the i-th entity, identified by contiguous 'B-' and 'I-' tags in S. This function systematically extracts and maps each entity in S, ensuring all tokens related to each entity are correctly grouped and labeled.

3.1.2 Entity Slots Recovery

During the inference stage, we implement an Entity Slots Recovery process to convert the generated structured entity slots back into a BIO-tagged sequence. This recovery process, defined by a function r, can be expressed as:

$$r(T, E) = \{s_1, s_2, \ldots, s_n\}, \quad (2)$$

where s_j is determined for each token t_j based on its presence in the entity slots E. The recovery follows these rules: (1). If t_j is the first token of an entity in E, s_j is assigned a 'B-' tag with the corresponding entity type. (2). If t_j is a non-initial token of an entity in E, s_j is assigned an 'I-' tag with the corresponding entity type. (3). If t_j does not belong to any entity in E, s_j is assigned an 'O' tag.

3.2 Chain of Intent

To effectively manage the complexity of multi-intent spoken language understanding, we propose a novel framework termed the "Chain of Intent," inspired by the "Chain of Thought" reasoning process (Wei et al., 2022). This framework enhances the model's ability to discern and process multiple intents within a single utterance by segmenting it into distinct sub-intent utterances, enabling more granular understanding and response generation.

Consider an utterance U consisting of n intents. Each intent I_i (where $i = 1, 2, \ldots, n$) corresponds to a specific segment of the utterance U_i. The process of decomposing the utterance U can be formally expressed as a mapping:

$$U \mapsto \{(I_1 : U_1), (I_2 : U_2), \ldots, (I_n : U_n)\} \quad (3)$$

Here, the structured pairs $(I_i : U_i)$ represent each intent I_i paired with its associated sub-utterance U_i. During training, the model is presented with this mapping to learn the relationship between each intent and its corresponding segment of the utterance, thereby improving its ability to generate contextually accurate and intent-specific responses.

3.3 Supervised Fine-tuning

We employ supervised fine-tuning to enhance the generative capabilities of LLMs, ensuring they meet the structured requirements of multi-intent spoken language understanding (SLU). This process involves adjusting the model parameters θ to minimize a loss function $\mathcal{L}$ across a set of training examples. Given a training set $\{(U_j, T_j)\}_{j=1}^{M}$, where U_j represents the j-th input utterance and T_j denotes the corresponding target output, including segmented sub-intents and entity slots, the fine-tuning objective is defined as:

$$\theta^* = \arg\min_{\theta} \sum_{j=1}^{M} \mathcal{L}\left(\mathrm{LLM}(U_j;\theta), T_j\right) \quad (4)$$

Here, $\mathrm{LLM}(U_j;\theta)$ represents the output generated by the LLM given the input U_j with parameters θ. The supervised fine-tuning process iteratively updates θ to more accurately map input utterances U_j to their corresponding intent and entity slot outputs T_j, thereby improving the model's effectiveness in multi-intent SLU tasks.

4 Experiments

4.1 Datasets

We conducted experiments on two widely used multi-intent SLU datasets: MixATIS (Hemphill et al., 1990; Qin et al., 2020a) and MixSNIPS (Coucke et al., 2018; Qin et al., 2020a). The MixATIS dataset contains 13,162 training instances and 828 test instances, primarily focusing on airline-related queries. In contrast, the MixSNIPS dataset spans a broader range of domains, including restaurants, hotels, and movies, with 39,776 training instances and 2,199 test instances. These datasets are designed to mimic real-world scenarios, featuring utterances with 1 to 3 intents, distributed in ratios of 30%, 50%, and 20%, respectively and detail information can be found in Table 1.

4.2 Experimental Settings

We utilize Llama3.1-8B-Instruct as base model and our experiments were conducted with a carefully selected set of hyperparameters. We employed FlashAttention v2 to optimize memory usage and accelerate training. To determine the optimal settings, we performed a grid search over the learning rate $[1 \times 10^{-5}, 2 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}]$ and the number of epochs $[1, 2, 3]$. Based on the results, we settled on a learning rate of 2×10^{-5} and a batch size of 32, tuning the model for 1 epoch on both datasets. During inference, a generation temperature of 0.0 was used to ensure deterministic and consistent outputs.

Dataset	MixATIS	MixSNIPS
Vocabulary Size	722	11241
Intent categories	17	6
Slot categories	116	71
Training set size	13162	39776
Test set size	828	2199

Table 1: Dataset statistics

4.3 Baselines

In our study, we benchmark LLMs performance against a range of established baselines in the multi-intent SLU domain. These include **vanilla models** like Stack-Propagation (Qin et al., 2019): a stack-propagation framework to explicitly incorporate intent detection for guiding slot filling. AGIF (Qin et al., 2020b): an adaptive interaction network to achieve fine-grained multi-intent information integration, GL-GIN (Qin et al., 2021b): a local slot-aware and global intent-slot interaction graph framework to model the interaction between multiple intents and all slots within an utterance, SDJN (Chen et al., 2022): a multiple instance learning and self-distillation framework for weakly supervised multiple intent information capturing, CLID (Huang et al., 2022): a chunk-level intent detection framework for recognizing intent within a fragment of an utterance and SSRAN (Cheng et al., 2023): a transformative network built on the Transformer model, designed to reduce the complexity of multi-intent detection in SLU through scope recognition and bidirectional interaction between results of slot filling and intent detection. We also included **PLM-based models** such as Uni-MIS (Yin et al., 2024): a unified multi-intent slu framework via multi-view intent-slot interaction. Additionally, SDJN(Bert) and CLID(Roberta) extend their respective base models by incorporating pre-trained language model backbones.

4.4 Main Result Analysis

The evaluation metrics included slot F1 score, intent accuracy and semantic accuracy to compre-

Model	MixATIS			MixSNIPS		
	Slot(F1)	Intent(Acc)	Overall(Acc)	Slot(F1)	Intent(Acc)	Overall(Acc)
Stack-Propagation (Qin et al., 2019)	87.8	72.1	40.1	94.2	96.0	72.9
AGIF (Qin et al., 2020b)	86.9	72.2	39.2	93.8	95.1	72.7
GL-GIN (Qin et al., 2021b)	87.2	75.6	41.6	93.7	95.2	72.4
SDJN (Chen et al., 2022)	88.2	77.1	44.6	94.4	96.5	75.7
CLID (Huang et al., 2022)	88.2	77.5	49.0	94.3	96.6	75.0
SSRAN (Cheng et al., 2023)	89.4	77.9	48.9	95.8	98.4	77.5
SDJN + Bert	87.5	78.0	46.3	95.4	96.7	79.3
RoBERTa+Linear	86.0	80.3	48.4	96.0	97.4	82.1
CLID + Roberta	85.9	80.5	49.4	96.0	97.0	82.2
Uni-MIS (Yin et al., 2024)	88.3	78.5	52.5	96.4	97.2	83.4
ECLM (Ours)	**90.2**	**80.7**	**56.2***	**97.0**	97.0	**86.5***

Table 2: Multi-Intent SLU performance on MixATIS and MixSNIPS datasets. Values with * indicate that the improvement from our model is statistically significant over all baselines ($p < 0.05$ under t-test).

Model	MixATIS Dataset			MixSNIPS Dataset		
	Slot(F1)	Intent(Acc)	Overall(Acc)	Slot(F1)	Intent(Acc)	Overall(Acc)
ECLM (Ours)	90.2	80.7	**56.2**	97.0	97.0	**86.5**
-w/o Entity Slot	73.5	78.7	54.9	92.7	97.6	69.7
-w/o Chain of Intent	89.4	82.6	52.9	96.8	98.0	85.1
-w/o Both (Vanilla SFT)	68.2	74.0	47.7	88.9	97.4	65.3

Table 3: Ablation experiments on the MixATIS and MixSNIPS datasets. Interestingly, we observe that entity slots play a more significant role in the MixSNIPS dataset compared to MixATIS, while the chain of intent does not explicitly improve intent accuracy but instead enhances overall performance.

hensively assess the sentence-level semantic frame parsing capabilities. These metrics, adhering to the methodologies delineated by Qin et al. (2021b); Huang et al. (2022); Yin et al. (2024) facilitate a nuanced evaluation of SLU systems. The paramount metric, semantic overall accuracy, quantifies the system's proficiency in simultaneously and correctly predicting both intents and slots within a single sentence.

Our main experiments yield several important observations: (1) As shown in Table 2, ECLM outperforms the strong baseline in slot filling F1 scores in both datasets. This improvement indicates that the ECLM interaction effectively utilises entity slots to improve it's slot filling ability. (2) For the single-domain MixATIS dataset, ECLM outperforms Uni-MIS with a 1.9 % point improvement in slot filling F1 scores (90.2%), a 2.2 % point improvement in intent prediction accuracy (80.7%), and a 3.7 % point improvement in overall sentence-level semantic frame parsing accuracy (56.2%). For the multi-domain MixATIS dataset, ECLM outperforms Uni-MIS by 0.6 % points in slot-filling F1 score (97.0%) and 3.1 % points in overall sentence-level semantic frame parsing accuracy (86.5%). These results highlight the competitive advantage of robust language models in multi-intent SLU tasks. (3) Importantly, our framework achieves state-of-the-art performance for most evaluation metrics, highlighting a promising research direction for multi-intent SLU using LLM-based methodologies.

4.5 Ablation Study

To understand the impact of key components in ECLM, we conducted ablation experiments on the MixATIS and MixSNIPS datasets. As shown in Table 3, the results illustrate the contribution of entity slots and the chain of intent to overall performance.

4.5.1 Without Entity Slot

Removing the entity slot significantly reduces performance, with a drop of 16.7 % in slot F1 score and 1.3 % points in overall accuracy on MixATIS. Similarly, on MixSNIPS, we observe a drop of 4.3 % in slot F1 score, and the overall accuracy decreases by 16.8 %. This highlights the crucial role of entity slots in maintaining high performance. Especially in the multi-domain dataset MixSNIPS, the absence of entity slots may cause significant misalignment, as the majority of slot labels are "O". This could lead to the model incorrectly labeling

Model	intent num = 1			intent num = 2			intent num = 3		
	Slot(F1)	Intent(Acc)	Overall(Acc)	Slot(F1)	Intent(Acc)	Overall(Acc)	Slot(F1)	Intent(Acc)	Overall(Acc)
GL-GIN	88.0	91.3	72.6	87.3	76.2	39.1	86.8	63.1	23.0
CLID	88.6	94.7	76.4	88.1	77.5	48.4	87.6	64.3	28.5
CLID + Roberta	88.6	95.8	77.6	85.4	80.3	48.8	84.7	66.8	29.0
Uni-MIS	89.2	95.1	78.6	87.6	78.3	50.5	86.7	66.7	31.7
ECLM(Ours)	92.1	93.7	**79.7**	90.3	79.4	**54.8**	90.3	70.0	**39.5**

Table 4: The result comes from the dataset MixATIS. The intent num denotes the number of intents in an utterance.

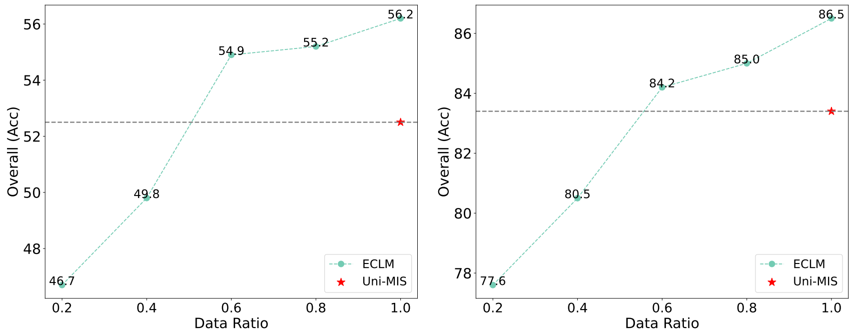

Figure 3: Performance of ECLM on the MixATIS and MixSNIPS datasets at different training data proportions

words as "O" rather than their corresponding slot tags.

4.5.2 Without Chain of Intent

Eliminating the chain of intent structure leads to a 0.8 % point drop in slot F1 score and a 3.3 % decline in overall accuracy on MixATIS. On MixSNIPS, the overall accuracy decreases by 1.4 %, emphasizing the importance of intent chaining in enhancing the model's semantic understanding. However, we observe that the improvement in intent detection accuracy is less pronounced, suggesting that the chain of intent mainly contributes to the joint effect and compromises some intent accuracy.

4.5.3 Without Both (Vanilla SFT)

When both components are removed, the performance suffers dramatically. The slot F1 score drops by 22.0 % and the overall accuracy by 8.5 % on MixATIS. The MixSNIPS dataset also shows a significant decrease, with the overall accuracy dropping by 21.2 %. This indicates that the Vanilla SFT method cannot effectively adapt LLMs to this domain.

5 Further Exploration

5.1 Influence of Different Intent Numbers

The analysis of MixATIS dataset results, categorized by the number of intents as shown in Table 4, reveals significant insights into the performance of our ECLM model compared to baseline approaches. For single-intent utterances, ECLM achieves superior performance with a slot F1 score of 92.1% and overall accuracy of 79.7%, outperforming the strong Uni-MIS over Uni-MIS (89.2% and 78.6% respectively). As the complexity increases with multi-intent scenarios, ECLM's advantages become more pronounced. In two-intent cases, ECLM maintains its lead with a slot F1 of 90.3% and overall accuracy of 54.8%, showing a substantial improvement over Uni-MIS (87.6% and 50.5% respectively). The performance gap widens further for three-intent utterances, where ECLM achieves a slot F1 of 90.3%, intent accuracy of 70.0%, and overall accuracy of 39.5%, significantly surpassing Uni-MIS (86.7%, 66.7%, and 31.7% respectively). This consistent outperformance, particularly in challenging multi-intent scenarios, underscores ECLM's robustness and efficacy in handling complex spoken language understanding tasks. The results demonstrate ECLM's capacity to maintain high performance across varying levels of intent complexity, indicating its potential as a versatile solution for advanced SLU systems.

5.2 Influence of Training Data Ratio

Figure 3 illustrates the impact of varying training data volumes on ECLM's performance, focusing on overall semantic accuracy across the MixATIS and MixSNIPS datasets. We systematically adjusted the training data ratios at 0.2, 0.4, 0.6, 0.8, and 1.0 to assess model proficiency under different

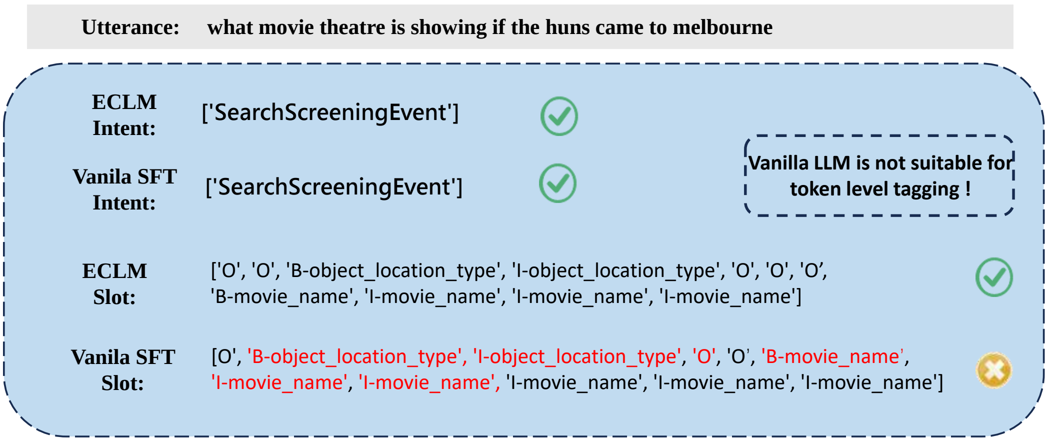

Figure 4: Comparative analysis of ECLM and vanilla SFT performance on a complex multi-intent utterance, highlighting ECLM's superior slot filling capabilities and the limitations of LLMs in token-level tagging tasks.

data availability scenarios. The results demonstrate a consistent positive correlation between the data ratio and performance improvements across both datasets. For MixATIS, ECLM's semantic accuracy rises from 46.7% at 0.2 data ratio to 56.2% at full data utilization, surpassing the Uni-MIS baseline (52.5%) with just 60% of the training data. Similarly, on MixSNIPS, ECLM's performance increases from 77.6% to 86.5%, exceeding the Uni-MIS benchmark (83.4%) also at approximately 60% data ratio. Notably, ECLM exhibits robust performance even with limited data, achieving competitive results at lower data ratios. The performance gains are more pronounced in the MixSNIPS dataset, suggesting ECLM's particular effectiveness in multi-domain scenarios. As the data ratio approaches 1.0, the performance improvement rate gradually stabilizes, indicating a potential plateau effect at higher data volumes.

5.3 Influence of Different Backbone LLMs in the ECLM Framework

Table 5 presents a comparative analysis of overall accuracy across various large language models (LLMs) when integrated into our ECLM framework, evaluated on both the MixATIS and MixSNIPS datasets. The results demonstrate a clear progression in performance as we move towards more advanced LLM architectures. Llama2-7B-Chat, while competent, shows the lowest performance with overall accuracies of 48.2% and 81.5% on MixATIS and MixSNIPS respectively. Mistral-7B-Instruct-v0.1 exhibits a notable improvement, achieving 50.1% and 83.9% on the same datasets, highlighting the rapid advancements in LLM capabilities. The Llama3.1 series showcases significant performance gains. The base Llama3.1-8B model achieves impressive results of 55.6% and 85.9% on MixATIS and MixSNIPS, respectively. However, the instruction-tuned variant, Llama3.1-8B-Instruct, emerges as the top performer, reaching 56.2% accuracy on MixATIS and 86.5% on MixSNIPS. The superior performance of Llama3.1-8B-Instruct underscores the importance of instruction tuning in enhancing model capabilities for specific tasks like multi-intent SLU. This model's consistent outperformance across both datasets justifies its selection as the default backbone for our ECLM framework.

Model	MixATIS	MixSNIPS
Llama2-7B-Chat	48.2	81.5
Mistral-7B-Instruct-v0.1	50.1	83.9
Llama3.1-8B	55.6	85.9
Llama3.1-8B-Instruct	**56.2**	**86.5**

Table 5: The impact of different backbone LLMs Integrated into the ECLM Framework.

5.4 Case Analysis

As illustrated in Figure 4, we present a comparative analysis of ECLM and vanilla LLM-based SFT approaches on a complex multi-intent utterance. The example, "what movie theatre is showing if the huns came to melbourne", demonstrates the superior performance of ECLM in handling intricate spoken language understanding tasks. Both ECLM and vanilla SFT correctly identify the primary intent as "SearchScreeningEvent". However, the critical distinction emerges in the slot filling task. ECLM accurately labels each token, precisely identifying "movie theatre" as the "object_location_type" and "if the huns came to melbourne" as the "movie_name". In contrast, the vanilla SFT model, despite its correct intent classification, exhibits significant errors in slot filling. The vanilla SFT incorrectly labels "what" as part of the "object_location_type" and mistakenly extends the "movie_name" to include "showing". This misalignment highlights a fundamental limitation of autoregressive LLMs in token-level tagging tasks. The sequential nature of their predictions can lead to error propagation and misalignment with the original utterance tokens.

6 Related Work

6.1 Intent Detection and Slot Filling

The inherent interconnected of intent detection and slot filling has spurred the development of unified models that foster mutual interaction between the two elements. Joint learning techniques, acknowledging the potent correlation between intents and slots, have proven particularly efficacious in recent years. Certain methodologies facilitating simultaneous slot filling and intent detection employ shared parameters (Liu and Lane, 2016; Wang et al., 2018; Zhang and Wang, 2016), while others model the relationship between the two via either unidirectional interaction or bidirectional-flow interaction (Qin et al., 2021c). Models adopting unidirectional interaction, such as those by (Goo et al., 2018; Li et al., 2018; Qin et al., 2019), primarily emphasize the flow from intent to slot. Gating mechanisms, functioning as specialized guiding forces for slot filling, have seen extensive use (Goo et al., 2018; Li et al., 2018). Qin et al. (2019) put forth a token-level intent detection model to curtail error propagation. Bidirectional-flow interaction models (E et al., 2019; Zhang et al., 2019; Liu et al., 2019; Qin et al., 2021a), on the other hand, examine the reciprocal influence of intent detection and slot filling. E et al. (2019) utilized iterative mechanisms to enhance intent detection and slot filling in both directions. Fine-grained intent detection and intent-slot interaction models have also seen remarkable advancements. Chen et al. (2022) developed a Self-distillation Joint SLU model exploitating multi-task learning, and treated multiple intent detection as a weakly-supervised problem solved through Multiple Instance Learning (MIL). Similarly, Huang et al. (2022) introduced a chunk-level intent detection framework that employs an auxiliary task to pinpoint intent transition points within utterances, thereby augmenting the recognition of multiple intents. Furthermore, Cheng et al. (2023) proposed a transformative network rooted in the Transformer model, designed to diminish the complexity of multi-intent detection in SLU. Recently, Yin et al. (2024) further develop an united multi-view intent-slot interaction framework(Uni-MIS), archiving promising performance.

6.2 Open Source Large Language Models

The advent of open-source Large Language Models (LLMs) such as Llama2 (Touvron et al., 2023), Vicuna (Peng et al., 2023), and Mistral (Jiang et al., 2023) has dramatically reshaped the landscape of Natural Language Processing. These models, characterized by their vast parameter spaces and diverse training corpora, have significantly expanded the capabilities and applications of NLP technologies. The rapid evolution of LLMs has accelerated progress across a broad spectrum of NLP tasks, including natural language inference, summarization, and dialogue systems (Geogle., 2023; Kavumba et al., 2023). Complementing these advancements, the "Chain of Thought" method (Wei et al., 2022) has emerged as a pivotal technique in enhancing the reasoning capabilities of LLMs. This approach enables models to break down complex problems into interpretable steps, significantly improving performance on tasks requiring multi-step reasoning or complex problem-solving.

7 Conclusion

In this paper, we introduced the Entity-level Large Language Model framework ECLM for multi-intent spoken language understanding. By transforming token-level slot-filling into an entity recognition problem and introducing the "Chain of Intent" concept, we effectively addressed the challenges of applying LLMs to SLU tasks. Our approach significantly outperformed state-of-the-art models, including Uni-MIS and conventional LLM fine-tuning, on the MixATIS and MixSNIPS datasets. ECLM demonstrated robust performance across various intent counts, particularly excelling in complex multi-intent scenarios.

8 Limitations

(1) *Scaling up Model Size of ECLM*: Due to computational resource constraints, we were unable to

experiment with ECLM models larger than 8 billion parameters. However, we believe that scaling to larger model sizes could potentially yield further improvements in performance. Recent trends in language model research suggest that larger models often demonstrate enhanced capabilities across various NLP tasks. Future work with access to more substantial computational resources could explore the impact of increased model size on ECLM's performance in multi-intent SLU tasks. (2) *Prospects for Improvement through Data Curation and Prompt Optimization*: Our current research framework does not extend to the advanced strategies of selective data curation or intricate prompt engineering. Recognizing this as a limitation, we propose that future investigations will embrace these crucial techniques.

References

Lisong Chen, Peilin Zhou, and Yuexian Zou. 2022. Joint multiple intent detection and slot filling via self-distillation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 7612–7616. IEEE.

Lizhi Cheng, Wenmian Yang, and Weijia Jia. 2023. A scope sensitive and result attentive model for multi-intent spoken language understanding. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 12691–12699. AAAI Press.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *CoRR*, abs/1805.10190.

Haihong E, Peiqing Niu, and Zhongfu Chen. 2019. A novel bi-directional interrelated model for joint intent detection and slot filling. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5467–5471.

Rashmi Gangadharaiah and Balakrishnan Narayanaswamy. 2019. Joint multiple intent detection and slot labeling for goal-oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 564–569.

Geogle. 2023. Palm 2 technical report. *CoRR*, abs/2305.10403.

Chih-Wen Goo, Guang Gao, and Yun-Kai Hsu. 2018. Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 753–757.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, USA, June 24-27, 1990*. Morgan Kaufmann.

Yushi Hu, Chia-Hsuan Lee, Tianbao Xie, Tao Yu, Noah A. Smith, and Mari Ostendorf. 2022. In-context learning for few-shot dialogue state tracking. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2627–2643. Association for Computational Linguistics.

Haojing Huang, Peijie Huang, Zhanbiao Zhu, Jia Li, and Piyuan Lin. 2022. CLID: A chunk-level intent detection framework for multiple intent spoken language understanding. *IEEE Signal Process. Lett.*, 29:2123–2127.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Pride Kavumba, Ana Brassard, Benjamin Heinzerling, and Kentaro Inui. 2023. Prompting for explanations improves adversarial NLI. is this true? yes it is true because it weakens superficial cues. In *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 2120–2135. Association for Computational Linguistics.

Byeongchang Kim, Seonghan Ryu, and Gary Geunbae Lee. 2017. Two-stage multi-intent detection for spoken language understanding. *Multim. Tools Appl.*, 76(9):11377–11390.

Changliang Li, Liang Li, and Ji Qi. 2018. A self-attentive model with gate mechanism for spoken language understanding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3824–3833.

Bing Liu and Ian Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. In *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, pages 685–689.

Yijin Liu, Fandong Meng, Jinchao Zhang, Jie Zhou, Yufeng Chen, and Jinan Xu. 2019. CM-Net: A novel collaborative memory network for spoken language understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1051–1060.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with GPT-4. *CoRR*, abs/2304.03277.

Libo Qin, Wanxiang Che, and Yangming Li. 2019. A stack-propagation framework with token-level intent detection for spoken language understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2078–2087.

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 *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, 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 Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 178–188.

Libo Qin, Tianbao Xie, Wanxiang Che, and Ting Liu. 2021c. A survey on spoken language understanding: Recent advances and new frontiers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4577–4584. ijcai.org.

Libo Qin, Xiao Xu, Wanxiang Che, and Ting Liu. 2020a. AGIF: An adaptive graph-interactive framework for joint multiple intent detection and slot filling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1807–1816, Online. Association for Computational Linguistics.

Libo Qin, Xiao Xu, Wanxiang Che, and Ting Liu. 2020b. Towards fine-grained transfer: An adaptive graph-interactive framework for joint multiple intent detection and slot filling. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1807–1816.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Gokhan Tur and Renato De Mori. 2011. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. Wiley, New York.

Yu Wang, Yilin Shen, and Hongxia Jin. 2018. A bi-model based RNN semantic frame parsing model for intent detection and slot filling. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 309–314. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Puyang Xu and Ruhi Sarikaya. 2013. Convolutional neural network based triangular CRF for joint intent detection and slot filling. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, December 8-12, 2013*, pages 78–83.

Shangjian Yin, Peijie Huang, and Yuhong Xu. 2024. Uni-mis: United multiple intent spoken language understanding via multi-view intent-slot interaction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19395–19403.

Chenwei Zhang, Yaliang Li, Nan Du, Wei Fan, and Philip S. Yu. 2019. Joint slot filling and intent detection via capsule neural networks. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5259–5267.

Xiaodong Zhang and Houfeng Wang. 2016. A joint model of intent determination and slot filling for spoken language understanding. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 2993–2999. IJCAI/AAAI Press.