Balancing Accuracy and Efficiency in Multi-Turn Intent Classification for LLM-Powered Dialog Systems in Production

Junhua Liu[1], **Yong Keat Tan**[2], **Bin Fu**[2], **Kwan Hui Lim**[3,1]
[1]Forth AI, Singapore
[2]Shopee, Singapore
[3]Singapore University of Technology and Design, Singapore
j@forth.ai, yongkeat.tan@shopee.com, bin.fu@shopee.com, kwanhui@acm.org

Abstract

Accurate multi-turn intent classification is critical for advancing conversational AI systems but remains challenging due to limited datasets and complex contextual dependencies across dialogue turns. This paper presents two novel approaches leveraging Large Language Models (LLMs) to enhance scalability and reduce latency in production dialogue systems. First, we introduce Symbol Tuning, which simplifies intent labels to reduce task complexity and improve performance in multi-turn dialogues. Second, we propose Consistency-aware, Linguistics Adaptive Retrieval Augmentation (CLARA), a framework that employs LLMs for data augmentation and pseudo-labeling to generate synthetic multi-turn dialogues. These enriched datasets are used to fine-tune a small, efficient model suitable for deployment. Experiments on multilingual dialogue datasets show that our methods result in notable gains in both accuracy and resource efficiency, with improvements of 5.09% in classification accuracy, a 40% reduction in annotation costs, and effective deployment in low-resource multilingual industrial settings.

Introduction

Dialogue systems are critical for automating interactions between customers and agents, enabling efficient communication and improved user experience. They play a pivotal role in global e-commerce platforms by meeting the growing demand for instantaneous customer service. Intent classification, a core component of natural language understanding, identifies users' goals from their inputs, thereby reducing waiting times and operational costs (Weld et al. 2021). User interactions often involve multi-turn dialogues, particularly for complex requests, which complicates the development of multi-turn intent classification (MTIC) models despite their similarity to standard text classification tasks. Moreover, real-world multilingual systems require scalable and inclusive solutions, especially in low-resource settings. This complexity arises from the need to model contextual factors such as historical utterances and prior intents; without proper session-level understanding, systems risk misinterpreting user intentions and producing incorrect or irrelevant responses (Xu and Sarikaya 2014). Consequently, MTIC remains a challenging problem in real-world and industrial dialogue systems.

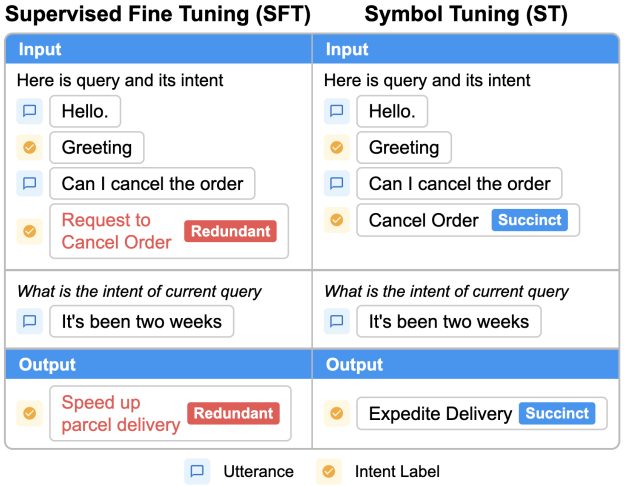

Figure 1: Comparison of instruction tuning and symbol tuning. Simplifying verbose intent labels (e.g., "Request to Cancel Order" → "Cancel Order") reduces redundancy, enhancing LLM classification performance by 5.09%, addressing key challenges in production intent classification.

There are two main challenges in multi-turn intent classification. First, intents in industrial dialogue systems are typically longer than those used in general text classification tasks. Figure 1 shows that real-world intents often comprise multiple words in the knowledge base, as operators (Ops) assign clear and descriptive intent names for knowledge management, resulting in redundancy. While recent advances in large language models (LLMs) offer opportunities to simplify and optimize text classification (Wang, Pang, and Lin 2024), and prior work shows that LLMs perform well in sentiment analysis (Přibáň et al. 2024) with short labels (e.g., positive, negative), LLMs still struggle with context dependency in multi-turn conversations and long intent labels common in industrial systems. Second, collecting multi-turn datasets remains challenging. Although several studies on MTIC exist (Qu et al. 2019; Wu, Su, and Juang 2021), they typically assume access to comprehensive multi-turn train-

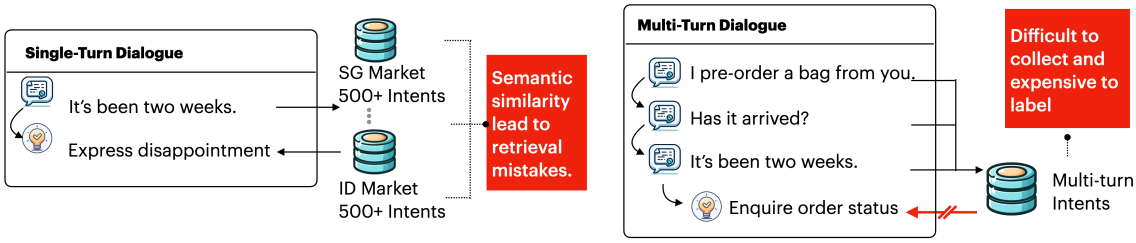

Figure 2: Annotation pipeline of multi-turn intent classification datasets, and two major challenges: (1) managing a large number of intents (500+) across markets with redundant labels, and (2) the high cost of collecting multi-turn training data.

ing data, which is rarely available in real-world applications.

Figure 2 shows the annotation pipeline for MTIC tasks. Even when redundant information within intent labels is ignored, dialogue systems typically maintain hundreds of intents operated by local Ops in the knowledge base to cover users' diverse and specific intents across markets, which substantially increases the complexity of multi-turn classification and annotation. This complexity is significantly greater than that of dialogue act classification, which typically involves fewer than 10 classes in dialogue state tracking (DST) (Qin et al. 2020). As a result, the annotation process becomes costly and time-consuming, making large-scale manual annotation of multi-turn datasets impractical. Moreover, insufficient training data can significantly hinder model performance, even when using LLMs. Together, these challenges highlight the need for more efficient methods to address data scarcity and classification complexity.

To address these challenges, we first examine the feasibility of using LLMs for supervised fine-tuning (SFT) to perform MTIC via a generative approach. The large number of diverse intents increases task complexity, as generating longer token sequences degrades LLM performance (Rust et al. 2021). To mitigate this issue, we compress redundant information in intent labels into concise forms using GPT-4 and adopt these compressed labels for SFT, a process we term symbol tuning, which reduces the difficulty of multi-turn intent classification for generative LLMs.

Secondly, to address the shortage of multi-turn data, we propose a novel pseudo-labeling and data generation framework termed Consistency-aware Linguistics Adaptive Retrieval Augmentation (CLARA). Extending beyond existing synthetic data generation approaches (Liu et al. 2024), CLARA serves as an effective pseudo-labeling method for generating multi-turn data from users' unlabeled utterances via self-consistency. Specifically, CLARA orders retrieved examples in multiple ways to construct adaptive prompts, capturing diverse reasoning paths and filtering noise during in-context learning to improve label quality. The resulting data are then used to train a compact model for efficient online inference. Designed for multi-turn intent classification, CLARA addresses limitations of prior methods by leveraging adaptive retrieval and self-consistency to improve pseudo-labeling accuracy, and directly optimizes for zero-shot multi-turn classification and scalable deployment.

In summary, the contributions of this paper are as follows:

1. We introduce symbol-tuning, leveraging compressed intents to enhance LLM performance for MTIC, demonstrating a 5.09% improvement in SFT results.
2. We propose CLARA, a novel framework for multi-turn data generation that effectively improves MTIC results.
3. We show that fine-tuning smaller models on CLARA-generated data enables scalable and accurate deployment of MTIC systems in low-resource industrial settings.

Problem Formulation

Multi-Turn Intent Classification (MTIC) involves identifying the intent I of the final query q_n from a predefined set $\mathcal{I}$, based on a sequence of user queries $\mathcal{Q} = \{q_i\}_{i=1}^{n}$ in a chatbot session. This task relies on the conversational context $\mathcal{C} = \{q_i\}_{i=1}^{n-1}$, which includes prior queries. Context-dependency adds complexity, requiring models to interpret nuanced conversational dynamics and evolving user intentions. Each intent I has a local-language title y and a hierarchical English category z (e.g., Indonesia: y = 'Cara membatalkan pesanan', z = 'Logistics > Order > Cancellation').

Supervised Fine-tuning (SFT) adapts pre-trained LLMs for specific tasks using labeled datasets. This process achieves high benchmark accuracy through task-specific supervision.

Problem Definition Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i is an input query and y_i is the corresponding label, the objective is to optimize model parameters θ to maximize the conditional likelihood $p(y_i|x_i;\theta)$:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\log p(y_i|x_i;\theta).$$

Conditional Probability Modeling For structured outputs, y_i is a sequence of tokens $\{t_1, t_2, \ldots, t_m\}$, with probability factorized autoregressively:

$$p(y|x;\theta) = \prod_{j=1}^{m} p(t_j|t_{<j}, x;\theta).$$

The training objective becomes:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{m}\log p(t_j|t_{<j}, x_i;\theta).$$

Symbol Tuning, instead of replacing task labels with unrelated symbols (Wei et al. 2023b), focuses on intent classification. Verbose labels in industrial systems disperse semantic information, hindering model performance. To address this, we compress labels into concise phrases using GPT-4. For example, "Request to Cancel Order" becomes "Cancel Order," serving as compact semantic anchors that enhance shallow and deep layer representations.

Let the original intent label be $L = \{t_1, t_2, \ldots, t_m\}$. The compressed label L', with $n \ll m$, is generated by optimizing:

$$L' = \text{argmin}_{L'} \ \mathcal{C}(L') + \mathcal{E}(L', L),$$

where: - $\mathcal{C}(L')$: Compactness of L' (e.g., token count). - $\mathcal{E}(L', L)$: Semantic divergence, computed as:

$$\mathcal{E}(L', L) = 1 - \text{cosine\_sim}(\phi(L'), \phi(L)),$$

with $\phi(\cdot)$ as an embedding function.

Objective Function Given $\mathcal{D} = \{(x_i, L_i)\}_{i=1}^{N}$, where L_i is the original label, the supervised fine-tuning loss becomes:

$$\mathcal{L}_{\text{ST}}(\theta) = -\mathbb{E}_{(x,L')\sim\mathcal{D}} \sum_{j=1}^{n} \log p(t_j | t_{<j}, x; \theta),$$

where $t_{<j}$ denotes preceding tokens in L'.

Solutions

Symbol Tuning on LLM

Our Symbol Tuning (ST) method involves supervised fine-tuning (SFT) of an LLM with compressed intent labels. Given a complete chat session $\mathcal{S} = \{q_1, I_1, ..., q_{n-1}, I_{n-1}, q_n\}$, the model is trained to generate the representative question r_n corresponding to the correct intent I_n of the final query q_n. Queries and intents are structured in a natural question-answering flow, as shown below:

> **SYSTEM**: "A chat between a curious user and an artificial intelligence assistant. The assistant provides helpful, detailed, and polite responses to the user's questions. ...
> **USER**: "{q˙1}"
> **ASSISTANT**: "The intent title is {r˙1}."
> ...
> **USER**: "{q˙n}"
> **ASSISTANT**: "The intent title is {r˙n}."

The generated r_n is compared with intents in $\mathcal{I}$ using cosine similarity in embedding space to ensure semantic alignment between the model output and predefined intent titles.

Compressed Generation Intent representative queries r often comprise approximately 12 tokens, making them inefficient generation targets. To address this, we use an LLM to compress r into concise phrases, typically two words, while preserving semantic meaning. This process ensures each compressed intent label r_c is unique. The compression reduces the average length of r_c to four tokens, optimizing it for generation tasks and improving classification accuracy.

Cross-Lingual Labels In non-English markets, intent labels r are compressed into English while retaining the original language for input queries $\mathcal{Q}$. Leveraging English, the main language in LLM pretraining corpora, simplifies label generation and enhances model performance in multilingual settings. This cross-lingual strategy reduces complexity and improves alignment with pretraining distributions.

Consistency-aware Linguistics Adaptive Retrieval Augmentation (CLARA)

To enhance in-context learning, we propose the Consistency-aware Linguistics Adaptive Retrieval Augmentation (CLARA) framework. CLARA incorporates a fine-tuned single-turn model $\mathcal{M}_c$ in a retrieval-augmented pipeline. This framework enables zero-shot Multi-Turn Intent Classification (MTIC) using only single-turn demonstrations, and operates offline as a pseudo-labeling tool.

Since the pipeline operates offline, response time is not a critical consideration. Self-consistency checking was performed on the LLM outputs to ensure the quality of pseudo-labels. As shown in Figure 3, the in-context learning phase is run three times per sample, with the in-context demonstrations sorted in three orders according to their similarity scores to the session queries: ascending, descending, and random. This self-consistency checking approach can also be implemented when using a black-box LLM. Online chat logs are sampled for pseudo-labeling, and only those having consistent labels for all 3 runs will be kept for training.

Hierarchical Text Classification (HTC) $\mathcal{M}_c$ is an ensemble of label-attention encoder and a hierarchical-aware tree-based encoder with 3-layered global and local intent classifiers.

The label-attention encoder has one classifier head for each intent layer. Each classifier head has one hidden linear layer to obtain the layer intermediate output L_l, which encodes the layer information. This layer information will be utilised in the input of the next layer classifier head.

$$L_l = \begin{cases} HW_l^1 + b_l^1, & \text{if } l = 1, \\ (H \oplus L_{l-1})W_l^1 + b_l^1, & \text{if } l > 1, \end{cases}$$

where $W_l^1 \in \mathbb{R}^{d\times d}$ for $l = 1$ and $W_l^1 \in \mathbb{R}^{2d\times d}$ for $l > 1$. $b_l^1 \in \mathbb{R}^d$, l is the layer number, $\oplus$ denotes tensor concatenation. Finally, we obtain the local logits H_{local}^l for each layer classes by using another linear layer

$$H_{local}^l = L_l \cdot W_l^2 + b_l^2, W_l^2 \in \mathbb{R}^{d\times|\mathcal{I}_l|}, b_l^2 \in \mathbb{R}^{|\mathcal{I}_l|}$$

where $|\mathcal{I}_l|$ is the number of classes in the layer.

To inject awareness of the overall hierarchical structure, we adopt a state-of-the-art HTC global approach introduced in HiTIN (Zhu et al. 2023). Specifically, the original taxonomy structure is simplified and constructed a tree network. The messages are propagated bottom-up in an isomorphism manner, which complements the label-attention model used. The embedding for leaf nodes are obtained by broadcasting the text representation H. After propagation, all embedding from all layers are aggregated to form single embedding and passed to a classification layer to obtain the logits H_{global}

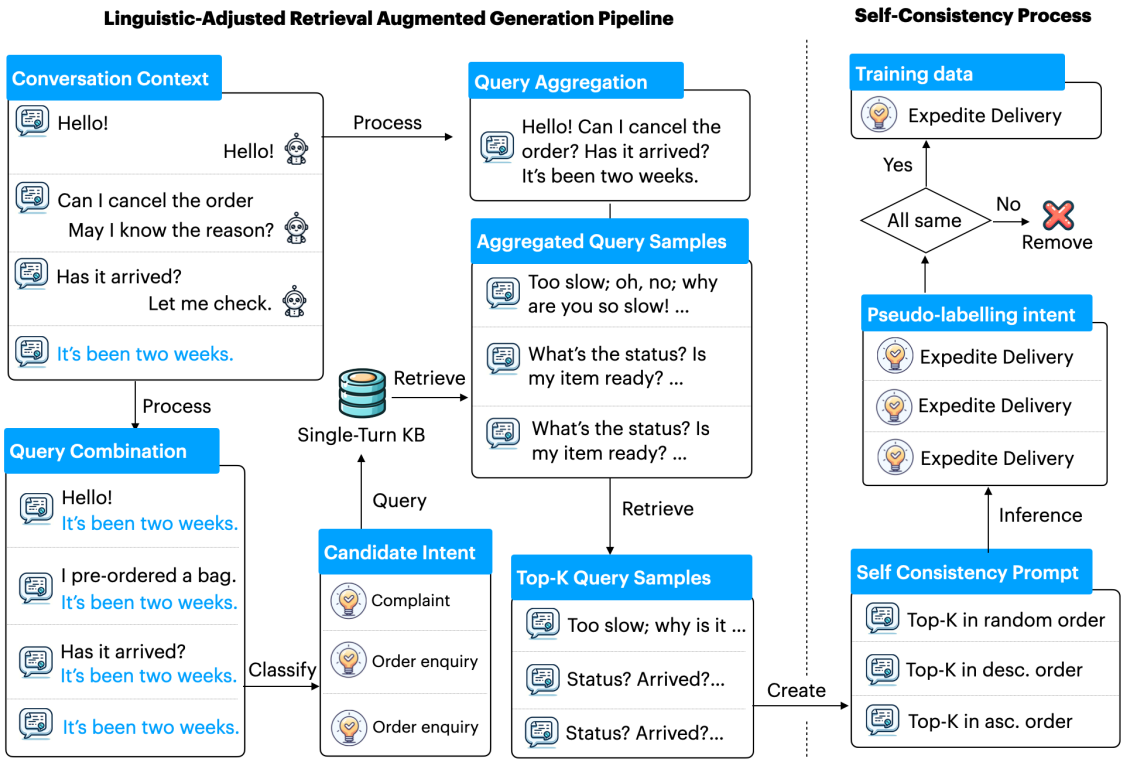

Figure 3: Illustration of CLARA: Merging LARA with Self-Consistency effectively combines query aggregation, knowledge base retrieval, and self-consistency mechanism to generate high-quality pseudo-labels for multi-turn dialogues. The self-consistency process improves labeling accuracy by validating intent predictions across different prompt orderings.

Market	Lang.	Intents	Train(ST)	Test(MT)
BR	pt	316	66k	372
ID	id	481	161k	1145
MY	en,ms	473	74k	1417
PH	en,fil	237	33k	189
SG	en	360	76k	737
TH	th	359	60k	502
TW	zh-tw	373	31k	353
VN	vi	389	178k	525

Table 1: Multilingual dataset statistics for Single Turn (ST) and Multi-Turn (MT).

MKT	Model	r_c	CL-Label	Accuracy
SG	Naive Concat.	-	-	60.52%
SG	Selective Concat.	-	-	56.99%
SG	Llama2-7B	✘	✘	56.24%
SG	Llama2-7B	✔	✘	61.33%
SG	Domain-Llama2-7B	✔	✘	**63.23%**
ID	Naive Concat.	-	-	60.61%
ID	Selective Concat.	-	-	**63.23%**
ID	Llama2-7B	✔	✘	49.96%
ID	SeaLLM-7B-chat	✔	✘	52.49%
ID	SeaLLM-7B-chat	✔	✔	55.02%

Table 2: Performance of LLM with symbol tuning.

of all tree nodes. The logits are then split by the number of classes in each layer to obtain H^l_{global}.

The final class probabilities for each layer P_l is then obtained by:

$$P_l = softmax(H^l_{local} + H^l_{global})$$

Experiments

Dataset

The dataset used in our experiments is derived from the conversation history of a large e-commerce platform. It includes user queries in the local languages of eight markets: Brazil (BR), Indonesia (ID), Malaysia (MY), Philippines (PH), Singapore (SG), Thailand (TH), Taiwan (TW), and Vietnam (VN), as detailed in Table 1. Labeled data were manually annotated by local customer service teams, with only samples achieving label consistency across three independent taggers being selected to ensure quality.

Single-turn training data collected over years of business operations form the basis for supervised fine-tuning and in-context learning. For multi-turn evaluation, real online sessions are annotated by local customer service teams, with

only the last query q_n labeled in each session $\mathcal{Q}$. For preprocessing, we remove noisy annotations, standardize intents, and augment multi-turn sessions using dialogue state transition probabilities derived from chat logs.

Symbol Tuning. We perform symbol tuning on LLMs for the SG and ID datasets, where SG mainly uses English while ID uses Bahasa Indonesia. The training data comprises a mix of existing single-turn samples and about 60k semi-automatically crafted multi-turn samples added to each market. Some are obtained by cleaning online chat logs to identify more accurate intents using an LLM with few-shot chain-of-thought prompts. The rest are constructed by combining several dialogues sampled from the existing single-turn training dataset to form one session. The transition of intents in a session is calculated from the online chat logs.

HTC with CLARA. 70k of online chat logs are sampled for pseudo-labelling. After self-consistency checking, around 12% of the data yield inconsistent results and are discarded from training. 1.5k samples are split from the pseudo-labeled data to serve as the validation set for early stopping.

Metrics

The primary evaluation metric is the accuracy of predicted labels for the final query q_n in each conversation session $\mathcal{Q}$. Metrics accounting for class imbalance were not used, as the sampled sessions reflect the distribution of online traffic across intents, providing a realistic approximation of live performance.

Implementation Details

Symbol Tuning on LLM FastChat framework is used to fine-tune 7B LLMs using LoRA method on their q_proj, v_proj, o_proj, and k_proj modules with a learning rate of 2e-5 over 10 epochs. The 7B models used are Llama-2-7B (for SG) and SeaLLM-7B-chat (for ID) on Hugging Face. Before the models are fine-tuned on the multi-turn intent recognition task, they are further pre-trained on ShareGPT dataset with the same setting above, and the weights are then merged. For the sake of simplicity, we will refer to the LLMs further pre-trained on ShareGPT dataset as base models. During training for intent classification task, loss is calculated on all the model output including those after history queries. During inference, greedy decoding strategy is used to generate the target r part, the prefix "The intent title is " is not generated but instead appended at the end of the prompt. When the generated label has no exact match with any r in $\mathcal{I}$, gestalt string matching is used to find the closest one.

HTC with CLARA The in-house Hierarchical Text Classification (HTC) model is a BERT-based model fine-tuned using the combination of the pseudo-labeled multi-turn data and existing single-turn data, as shown in Section 4.1. We use AdamW to finetune the HTC with a learning rate of 5e-6. All tests are run on a single Nvidia V100 GPU card with 32GB of GPU memory.

Baseline settings

For a fair comparison, we adopt three methods fine-tuned on HTC ($\mathcal{M}_c$) as our global baselines across two methods:

1. **Single-turn method**: where only the last query of a session is considered by $\mathcal{M}_c$;
2. **Naive concatenation**: all queries are concatenated together before being fed into $\mathcal{M}_c$;
3. **Selective concatenation**: where a concatenation selection model is trained to select the most suitable historical query with the last query to serve as the input to $\mathcal{M}_c$.

ST on LLM In SG, except Llama2-7B, we also tried to continue pre-training the base models on in-domain corpus to strengthen the language understanding of local languages and the corresponding slang used, as humans usually converse with the chatbot in a non-formal way. We term the domain specific base model as **Domain-Llama2-7B**. In ID, we switched Llama2-7B model to **SeaLLM-7B-chat** (Nguyen et al. 2024) which was introduced specifically for languages in South East Asia.

The ST approach was adapted for supervised intent recognition using compressed generation targets (r_c) and cross-lingual labels (CL_label). These adjustments optimized performance by simplifying the generative task while maintaining semantic integrity. Comparisons with baseline methods in Table 2 show that ST achieves competitive results in English markets but faces challenges in non-English ones due to limitations in pre-training for low-resource languages.

HTC with CLARA This experiment uses Vicuna-13B as our base model for pseudo-labeling within LARA and CLARA. We designed three pipelines with four prompt templates in (Liu et al. 2024) to demonstrate that using CLARA for pseudo-labeling can effectively improve the HTC model's performance in multi-turn classification tasks. The detailed introduction is listed as follows:

- **LARA**: Using LARA directly as a classifier.
- **LARA-PL**: Using LARA as a naive pseudo-labeling tool and fine-tune HTC model with generated data.
- **CLARA**: Using CLARA to filter out the noise and generate high-quality data to fine-tune the HTC model.

Offline Experiments

Symbol Tuning on LLM Table 2 illustrates the effectiveness of Symbol Tuning (ST) on LLMs. Compressing the generation target r reduces task complexity and improves accuracy by 5.09% in the SG market. This compression also mitigates hallucination, reducing instances of unmatched generated labels from 2.5% to 0%.

Interestingly, this technique also stopped LLM hallucination, i.e. generating label with no match in the $\mathcal{I}$. The hallucination rate without using compressed r is about 2.5%. In ID, which is a non-English market, we find that **cross-lingual label** which changes the generation target to English rather than in the local language also improved the performance by 2.53%. Using **different base models** which were trained specifically on the in-domain corpus or for the local language also proves to be useful. Domain-Llama2-7B improves the performance by 1.90% in SG while SeaLLM-7B-chat improves the performance by 2.53% in ID compared to Llama2-7B. While the ST approach outperforms the baselines in English market, it still leaves a lot to be desired in

Pipeline	Model	Prompt	SC	BR	ID	MY	PH	SG	TH	TW	VN	avg
Fine-tuning	Single-turn	-	-	30.98%	52.14%	56.81%	40.21%	51.13%	52.99%	58.07%	65.90%	53.76%
Fine-tuning	Naive Concat.	-	-	50.81%	60.61%	57.02%	47.62%	60.52%	56.97%	65.44%	76.95%	60.08%
Fine-tuning	Select. Concat.	-	-	52.69%	63.23%	60.20%	51.32%	56.99%	57.77%	64.02%	74.10%	60.97%
LARA	Vicuna-13B	$\mathcal{P}$	✘	52.69%	61.48%	65.42%	54.50%	65.26%	60.96%	67.14%	77.90%	64.18%
CLARA	Vicuna-13B	$\mathcal{P}$	✔	55.38%	63.58%	65.00%	54.50%	66.21%	63.75%	71.10%	79.24%	65.52%
LARA	Vicuna-13B	$\mathcal{P}_{symbolic}$	✘	51.88%	60.00%	64.57%	53.97%	65.26%	58.96%	65.44%	74.67%	62.92%
CLARA	Vicuna-13B	$\mathcal{P}_{symbolic}$	✔	54.57%	62.62%	65.56%	50.79%	66.76%	62.95%	69.97%	76.76%	64.94%
LARA	Vicuna-13B	$\mathcal{P}_{prepend}$	✘	54.03%	61.75%	64.50%	53.44%	65.94%	61.55%	66.86%	75.81%	63.97%
CLARA	Vicuna-13B	$\mathcal{P}_{prepend}$	✔	53.76%	63.84%	65.70%	52.91%	**68.11%**	63.15%	69.97%	78.48%	65.65%
LARA	Vicuna-13B	$\mathcal{P}_{format.}$	✘	55.65%	62.88%	64.71%	**55.03%**	65.40%	61.95%	66.86%	78.10%	64.64%
LARA-PL	Vicuna-13B	$\mathcal{P}_{format.}$	✘	**55.91%**	64.19%	64.43%	49.21%	66.49%	61.95%	69.41%	**81.14%**	65.29%
CLARA	Vicuna-13B	$\mathcal{P}_{format.}$	✔	**55.91%**	**65.33%**	**66.27%**	51.85%	67.16%	**63.35%**	**72.80%**	78.86%	**66.35%**

Table 3: Performance of CLARA compared to baselines. CLARA with formatted prompts and Self Consistency (SC) achieves the best average accuracy (66.35%) across all markets, validating our approach's effectiveness in both English and non-English markets. The average scores are weighted on the number of test samples in each market.

non-English market. This phenomenon may arise as a result of the ST approach employed for non-dominant languages during pre-training, which necessitates a greater quantity or higher quality of data to achieve satisfactory performance in a task that was not included in the pre-training phase. This is particularly true when the model lacks knowledge pertaining to domain intents.

Pseudo-labeling using CLARA As shown in Table 3, CLARA improves pseudo-label quality through self-consistency validation, resulting in a 1.06% performance gain over LARA. This validation process identifies and removes approximately 12% of inconsistent samples, ensuring high-quality synthetic labels. While this approach requires additional offline training resources, it significantly lowers deployment costs by relying on a single, lightweight classification model.

This performance gain can likely be attributed to the strengths of the discriminative method in classification tasks, as training process also exposed the model to the comprehensive high quality single-turn dataset. Besides, the pre-trained model used for $\mathcal{M}_c$ was also pre-trained specifically on the in-domain multi-lingual corpus, making it a strong suit for our multilingual e-commerce setting. CLARA's integration of self-consistency within the pseudo-labeling pipeline significantly enhances the quality of synthetic labels, resulting in a 1.06% improvement in performance, as indicated in the last row of Table 3. When the LLM lacks confidence in its ICL responses, minor changes in the input prompt can significantly alter the output. This method effectively identifies potential inaccuracies in ICL outputs for black-box models where direct output scores are unavailable. While the offline training may require more time, the resulting deployment is more efficient and scalable, requiring only one compact classification model.

Online Deployment Evaluation

ST on LLM Using the LMDeploy framework, LARA weights were merged with the 7B base model, enabling faster inference times. Deployed on a single 32GB V100 GPU, the Symbol Tuning (ST) approach achieved an average latency of 170ms at 0.5 QPS in the SG market. In contrast, CLARA models converted to ONNX format (1.1GB per model) achieved an average latency of 80ms at 1 QPS on an 8-core CPU machine with 16GB memory, demonstrating superior scalability and cost-efficiency.

CLARA We deploy CLARA across all eight markets. The models were first converted to ONNX format, reducing their size to 1.1GB. Deployed on an 8-core CPU machine with 16GB memory, CLARA achieved an average latency of 80ms at 1 QPS, which is less than half the latency of the ST on LLM method. This deployment significantly reduced both costs and complexity, making it more scalable for industrial applications. Due to its versatility, an Auto-Training Portal (ATP) ecosystem is built around the LARA-PL method (Figure 4). ATP enables seamless and continuous improvements for multi-turn intent recognition system. Using online chat logs, local operations teams can update the Knowledge Base (KB) by adding new intents and crafting example queries. Subsequently, they can trigger CLARA for pseudo-labeling multi-turn chat logs, generating data to train lightweight models. Once training is complete, the models are deployed through the portal for online A/B testing, creating an iterative cycle of improvement. For fair comparisons, the intents and single-turn training data was kept consistent across control and test groups.

Online Performance

We leverage the following two metrics:

1. **Resolution rate (RR)** which is measured by the rate of user completing the answer flow, not transferring to live agent, and not giving bad rating to the answer.
2. **Customer Service Satisfaction (SCSAT)** where users will be asked about their satisfaction towards our chatbot for chatbot only sessions (no intervention from live agents). The score is calculated by *# good rated sessions*/(*# good rated sessions* + *# bad rated sessions*).

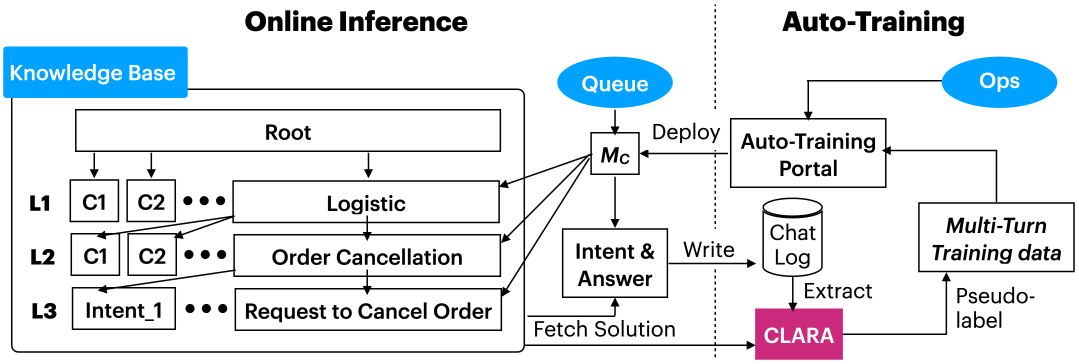

Figure 4: Online Deployment of Multi-turn Intent Classification model demonstrates our production architecture integrating CLARA for automated training data generation. The system handles real-time inference while continuously improving through automated training.

We use the **selective concatenation** method as the baseline for all experiments, with paired t-test to evaluate statistical significance.

ST on LLM In the SG market, ST on LLMs was deployed to 50% of online traffic for three weeks, yielding ∼14k chatbot sessions per group. The test group exhibited a +2.19% improvement in Customer Service Satisfaction (SCSAT), but Resolution Rate (RR) declined by -0.11%. Neither result was statistically significant, indicating limited benefits from ST given its resource-intensive nature.

CLARA The multi-turn dialogue model was replaced, while single-turn models remained unchanged. Aggregated results from over 108k chatbot sessions per group showed statistically significant improvements: Resolution Rate (RR) increased by +0.78% and Customer Service Satisfaction (SCSAT) by +1.39% (p-value < 0.05). These gains translate to overall session improvements of RR +0.47% and SCSAT +0.84%, as multi-turn dialogues comprise 60.60% of total sessions.

Furthermore, adding pseudo-labeled multi-turn data enhanced single-turn intent recognition. Substituting single-turn dialogue models with CLARA models yielded an RR improvement of +0.06% and a statistically significant SCSAT increase of +0.27%.

Ablation Study

Effect of Target Length

We investigate how the amount of information in ST generation target affects the intent recognition performance using two rather extreme approaches and their conversation semantic fluidity.

Longer Target Length To achieve this, the model is trained to summarize all queries in $\mathcal{Q}$ before outputting the target r. For instance, the new output format of model will be "**You are asking about {**$summary$**}. So,** the intent title is $\{r_n\}$". The rationale is to utilize the summarization ability of LLMs to better understand the context. For our training data, the summaries are obtained by prompting the original LLM backbones in a zero-shot manner. We chose this over increasing the length of r statically to impose more information on the model's generation target. Empirically, extending the generation target to include query summaries decreases performance by 3.82%. While this approach enhances semantic coherence, excessive information overloads the model, reducing its ability to focus on the core intent classification task.

Shorter Target Length Inspired by (Wei et al. 2023b), we compress rs by replacing with meaningless symbols. Compressing target labels to purely symbolic representations results in a significant 8.91% performance drop. This highlights the importance of preserving semantic richness in target labels for generative fine-tuning. Effective compression methods must retain key information from the original labels to avoid loss in classification accuracy.

Impact of Self-Consistency in MTIC

Using our multi-turn test sets, we evaluate the performance of MTIC with and without self-consistency checking. We remove the samples with inconsistent outputs and calculate the precision of the remaining samples. On average, 12% of test samples will be removed in each market. Incorporating self-consistency checking into MTIC evaluations improves accuracy across all prompt variations, as shown in Table 4. By removing approximately 12% of test samples with inconsistent outputs, this method effectively filters out erroneous predictions, ensuring higher-quality pseudo-labels and more reliable results. This ensures the quality of pseudo-labels.

Effect of Model Size

For fair comparison between LLM ST and CLARA, we use vicuna-7b-v1.5 as the base model with prompt $\mathcal{P}$ and $\mathcal{P}_{formatted}$, without self-consistency. The results of LLM ST method are taken from the best of each market, including base models pre-trained on in-domain corpus. Table 5 compares CLARA and LLM ST using models of the same size (Vicuna-7B-v1.5) without self-consistency. Despite the simpler pipeline, CLARA consistently outperforms LLM ST, avoiding the complexity of multi-turn sample crafting. However, smaller models exhibit reduced instruction-following

Prompt	SC	BR	ID	MY	PH	SG	TH	TW	VN	avg
$\mathcal{P}$	✘	52.69%	61.48%	65.42%	54.50%	65.26%	60.96%	67.14%	77.90%	64.18%
$\mathcal{P}$	✔	58.59%	68.13%	69.93%	56.44%	69.58%	66.75%	71.30%	81.14%	69.11%
$\mathcal{P}_{symbolic}$	✘	51.88%	60.00%	64.57%	53.97%	65.26%	58.96%	65.44%	74.67%	62.92%
$\mathcal{P}_{symbolic}$	✔	56.63%	64.71%	68.48%	55.19%	68.77%	65.59%	71.61%	78.02%	67.27%
$\mathcal{P}_{prepend}$	✘	54.03%	61.75%	64.50%	53.44%	65.94%	61.55%	66.86%	75.81%	63.97%
$\mathcal{P}_{prepend}$	✔	59.49%	66.08%	68.60%	55.90%	68.85%	68.19%	71.79%	81.36%	68.43%
$\mathcal{P}_{formatted}$	✘	55.65%	62.88%	64.71%	55.03%	65.40%	61.95%	66.86%	78.10%	64.64%
$\mathcal{P}_{formatted}$	✔	59.24%	67.01%	69.47%	56.79%	68.81%	68.69%	72.35%	82.93%	69.12%

Table 4: Precision of CLARA variants after filtering inconsistent predictions demonstrates the effectiveness of Self-Consistency (SC) across different prompt types. Self-Consistency improves accuracies by 4-5%, with $\mathcal{P}_{formatted}$ achieving the highest precision (69.12%).

Method	Prompt	ID	SG	avg
LLM ST	-	58.17%	63.23%	60.15%
CLARA	$\mathcal{P}$	60.44%	64.31%	61.96%
CLARA	$\mathcal{P}_{formatted}$	59.83%	64.04%	61.48%

Table 5: Results of LARA using 7B LLM.

capabilities, as demonstrated by the lower performance of $\mathcal{P}_{formatted}$ compared to $\mathcal{P}$. We observe that the performance of CLARA while using $\mathcal{P}_{formatted}$ is lower than $\mathcal{P}$. As LLMs of smaller size could be weaker in instruction following, it implies that semantic meaning of the labels in demonstrations are critical.

Related Work

Synthetic Data Generation

The scarcity of annotated dialogue data, particularly in low-resource languages, has driven research into synthetic data generation. Borisov et al. (2022) proposed a method leveraging auto-regressive generative models to create realistic tabular datasets, highlighting their utility in data augmentation. Similarly, Li et al. (2023) demonstrated that synthetic data generated by LLMs can significantly enhance model performance in classification tasks. Tang, Laban, and Durrett (2024) utilized synthetic data to craft challenging examples for fact-checking, improving the factual accuracy of LLM outputs.

Modeling Multi-turn Dialogue Context

Multi-turn dialogue modeling is essential for dialogue understanding tasks. Early methods used bidirectional contextual LSTMs (Ghosal et al. 2021) to capture context-aware utterance representations for tasks such as MultiWOZ intent classification (Budzianowski et al. 2018). Other approaches, such as multi-channel graph convolutional networks, were applied to query classification in E-commerce (Yuan et al. 2024). Recent advancements leverage pre-trained language models (PLMs) as sentence encoders (Shen et al. 2021), particularly for emotion recognition in conversations (ERC). For instance, Lee and Lee (2022) encoded both context and speaker memory using PLMs, while Qin et al. (2023) incorporated multi-turn information from utterances and dialogue structure through fine-tuning. Despite their effectiveness, these methods depend heavily on multi-turn training datasets, which are difficult to acquire in real-world e-commerce settings (Liu and Fu 2024). In contrast, our approach employs LLMs within an augmentation-based pipeline to generate multi-turn data, enabling zero-shot intent classification using smaller models.

LLM on Text Classification

Recent studies investigated the performance of LLMs as domain-specific text classifiers. Wei et al. (2023a) highlighted the benefits of fine-tuning LLMs on domain-specific datasets, improving performance in legal document review. Wei et al. (2023b) introduced symbol tuning, where natural language labels were replaced with unrelated symbols during fine-tuning to enhance classification. Loukas et al. (2023) analyzed the trade-offs between performance and cost when using LLMs for text classification. Liu, Lee, and Lim (2025) employed GPT-4o to perform zero-shot classification on multi-level semi-structured text with retrieval augmentation. In social science research, Chae and Davidson (2023) investigated LLMs for sociological text classification, demonstrating their potential. Our work differs by compressing longer intent labels into semantically meaningful phrases, enabling easier generation and improving accuracy for tasks with a large number of classes.

Conclusion

Multi-turn intent classification in production faces unique challenges, including highly variable and often lengthy intents. To address these issues, we introduce Symbol Tuning, which fine-tunes LLMs with compressed intents and improves classification accuracy by 5.09% compared to original intents. We also propose the CLARA pipeline for generating high-quality multi-turn datasets, significantly reducing annotation costs by automating pseudo-labeling based on users' latest utterances in dialogue history, thereby improving model iteration efficiency and enabling scalable deployment and online inference. Looking ahead, we plan to incorporate additional signals such as user profiles and order history into CLARA to support more diverse dialogue tasks, and to explore cross-lingual transfer and advanced tokenization techniques for low-resource languages.

Acknowledgements

This work is led and supported in part by Forth AI for the use of computational resources, software licenses, in-kind R&D contributions and domain expertise.

This research is supported in part by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Award No. MOE-T2EP20123-0015). Any opinions, findings and conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of the Ministry of Education, Singapore.

References

Borisov, V.; Seßler, K.; Leemann, T.; Pawelczyk, M.; and Kasneci, G. 2022. Language Models Are Realistic Tabular Data Generators. *ArXiv*, abs/2210.06280.

Budzianowski, P.; Wen, T.-H.; Tseng, B.-H.; Casanueva, I.; Ultes, S.; Ramadan, O.; and Gašić, M. 2018. MultiWOZ–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

Chae, Y.; and Davidson, T. 2023. Large language models for text classification: From zero-shot learning to fine-tuning. *Open Science Foundation*.

Ghosal, D.; Majumder, N.; Mihalcea, R.; and Poria, S. 2021. Exploring the role of context in utterance-level emotion, act and intent classification in conversations: An empirical study. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 1435–1449.

Lee, J.; and Lee, W. 2022. CoMPM: Context Modeling with Speaker's Pre-trained Memory Tracking for Emotion Recognition in Conversation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5669–5679.

Li, Z.; Zhu, H.; Lu, Z.; and Yin, M. 2023. Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Liu, J.; and Fu, B. 2024. Responsible Multilingual Large Language Models: A Survey of Development, Applications, and Societal Impact. *ArXiv*.

Liu, J.; Lee, R. K.-W.; and Lim, K. H. 2025. BGM-HAN: A Hierarchical Attention Network for Accurate and Fair Decision Assessment on Semi-Structured Profiles. In *Proceedings of the 17th International Conference on Advances in Social Networks Analysis and Mining (ASONAM'25)*.

Liu, J.; Tan, Y. K.; Fu, B.; and Lim, K. H. 2024. LARA: Linguistic-Adaptive Retrieval-Augmentation for Multi-Turn Intent Classification. *Proceedings of the Empirical Methods in Natural Language Processing*.

Loukas, L.; Stogiannidis, I.; Diamantopoulos, O.; Malakasiotis, P.; and Vassos, S. 2023. Making llms worth every penny: Resource-limited text classification in banking. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, 392–400.

Nguyen, X.-P.; Zhang, W.; Li, X.; Aljunied, M.; Hu, Z.; Shen, C.; Chia, Y. K.; Li, X.; Wang, J.; Tan, Q.; Cheng, L.; Chen, G.; Deng, Y.; Yang, S.; Liu, C.; Zhang, H.; and Bing, L. 2024. SeaLLMs – Large Language Models for Southeast Asia. arXiv:2312.00738.

Přibáň, P.; Šmíd, J.; Steinberger, J.; and Mištera, A. 2024. A comparative study of cross-lingual sentiment analysis. *Expert Systems with Applications*, 247: 123247.

Qin, L.; Che, W.; Li, Y.; Ni, M.; and Liu, T. 2020. Dcr-net: A deep co-interactive relation network for joint dialog act recognition and sentiment classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 8665–8672.

Qin, X.; Wu, Z.; Zhang, T.; Li, Y.; Luan, J.; Wang, B.; Wang, L.; and Cui, J. 2023. Bert-erc: Fine-tuning bert is enough for emotion recognition in conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 13492–13500.

Qu, C.; Yang, L.; Croft, W. B.; Zhang, Y.; Trippas, J. R.; and Qiu, M. 2019. User Intent Prediction in Information-seeking Conversations. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, CHIIR '19. ACM.

Rust, P.; Pfeiffer, J.; Vulić, I.; Ruder, S.; and Gurevych, I. 2021. How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3118–3135.

Shen, W.; Wu, S.; Yang, Y.; and Quan, X. 2021. Directed Acyclic Graph Network for Conversational Emotion Recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1551–1560.

Tang, L.; Laban, P.; and Durrett, G. 2024. MiniCheck: Efficient Fact-Checking of LLMs on Grounding Documents. *ArXiv*, abs/2404.10774.

Wang, Z.; Pang, Y.; and Lin, Y. 2024. Smart Expert System: Large Language Models as Text Classifiers. *arXiv preprint arXiv:2405.10523*.

Wei, F.; Keeling, R.; Huber-Fliflet, N.; Zhang, J.; Dabrowski, A.; Yang, J.; Mao, Q.; and Qin, H. 2023a. Empirical study of LLM fine-tuning for text classification in legal document review. In *2023 IEEE International Conference on Big Data (BigData)*, 2786–2792. IEEE.

Wei, J.; Hou, L.; Lampinen, A.; Chen, X.; Huang, D.; Tay, Y.; Chen, X.; Lu, Y.; Zhou, D.; Ma, T.; and Le, Q. V. 2023b. Symbol tuning improves in-context learning in language models. arXiv:2305.08298.

Weld, H.; Huang, X.; Long, S.; Poon, J.; and Han, S. C. 2021. A survey of joint intent detection and slot-filling models in natural language understanding. arXiv:2101.08091.

Wu, T.-W.; Su, R.; and Juang, B.-H. 2021. A Context-Aware Hierarchical BERT Fusion Network for Multi-turn Dialog Act Detection. arXiv:2109.01267.

Xu, P.; and Sarikaya, R. 2014. Contextual domain classification in spoken language understanding systems using recurrent neural network. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 136–140.

Yuan, C.; Pang, M.; Fang, Z.; Jiang, X.; Peng, C.; and Lin, Z. 2024. A Semi-supervised Multi-channel Graph Convolutional Network for Query Classification in E-commerce. In *Companion Proceedings of the ACM on Web Conference 2024*, 56–64.

Zhu, H.; Zhang, C.; Huang, J.; Wu, J.; and Xu, K. 2023. HiTIN: Hierarchy-aware Tree Isomorphism Network for Hierarchical Text Classification. In *Annual Meeting of the Association for Computational Linguistics*.