Intent Detection in the Age of LLMs

Gaurav Arora
Amazon
gaurvar@amazon.com

Shreya Jain
IIT Jammu*
2020uee0135@iitjammu.ac.in

Srujana Merugu
Amazon
smerugu@amazon.com

Abstract

Intent detection is a critical component of task-oriented dialogue systems (TODS) which enables the identification of suitable actions to address user utterances at each dialog turn. Traditional approaches relied on computationally efficient supervised sentence transformer encoder models, which require substantial training data and struggle with out-of-scope (OOS) detection. The emergence of generative large language models (LLMs) with intrinsic world knowledge presents new opportunities to address these challenges. In this work, we adapt 7 SOTA LLMs using adaptive in-context learning and chain-of-thought prompting for intent detection, and compare their performance with contrastively fine-tuned sentence transformer (SetFit) models to highlight prediction quality and latency tradeoff. We propose a hybrid system using uncertainty based routing strategy to combine the two approaches that along with negative data augmentation results in achieving the best of both worlds (i.e. within 2% of native LLM accuracy with 50% less latency). To better understand LLM OOS detection capabilities, we perform controlled experiments revealing that this capability is significantly influenced by the scope of intent labels and the size of the label space. We also introduce a two-step approach utilizing internal LLM representations, demonstrating empirical gains in OOS detection accuracy and F1-score by >5% for the Mistral-7B model.

1 Introduction

Task oriented dialogue systems (TODS) have gained significant traction and investment from industry because of their efficiency, accessibility and 24x7 availability to serve customers. Automation through TODS is expected to save billions of dollars in labor costs by 2026 (Gartner, 2022).

Intent Detection is a vital part of natural language understanding (NLU) layer of TODS. Traditionally, intent detection has been used to understand and map the user query to a bot action (e.g., respond with a static answer, execute a pre-configured flow etc) (Dialogflow, 2010; LEX, 2017). With increasing use of LLMs such as ChatGPT (OpenAI, 2022), Claude (Anthropic, 2023), Mistral (Mistral, 2023), Llama (Meta, 2023) as retrieval augmented generators to generate answers to user queries in TODS, intent detection is being used to identify the right knowledge sources, APIs, and tools to call for retrieval augmented generation. This ensures efficient utilization of tools, APIs and various other knowledge sources.

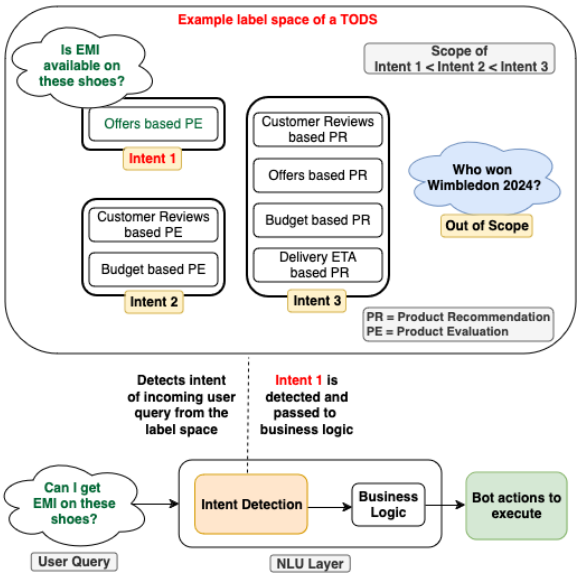

Figure 1: Example of broad/specific intent scopes and OOS queries which Intent Detection systems deal with in a typical TODS.

An intent detection system of a conversational AI service is expected to handle intents anywhere in the spectrum of very-broad to very-specific scopes[1] depending upon actionability of intents and bot usecases as shown in Fig 1. They are also

*Contributed to this work during her internship at Amazon

[1]By "scope of intent" we mean semantic space of all natural language utterances which can fall in that intent.

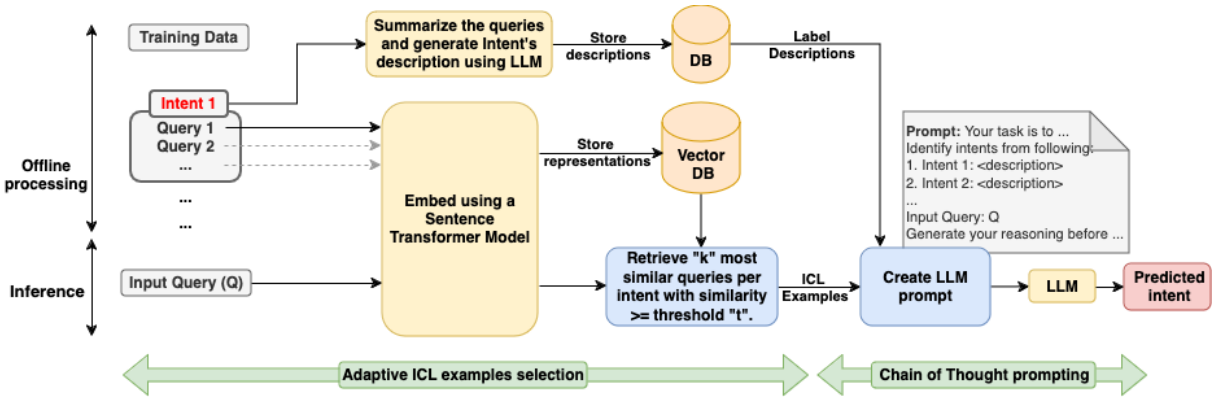

Figure 2: Methodology for adaptive ICL and CoT based intent detection using LLMs.

expected to accurately reject out-of-scope (OOS) queries[2] without having access to any training data for such queries as universe of OOS queries for any TODS is infinitely large. Since for a typical conversational AI service, data for intent detection training comes from bot developers who are not experts in ML, intent detection systems have to also deal with imbalanced training datasets. Additionally, these systems are expected to work with very few utterances per intent.

Traditionally, intent detection systems have been built using supervised classification or similarity based models (Zhang et al., 2021; Liu and Lane, 2016; Casanueva et al., 2020). LLMs, due to their few-shot learning capabilities, world knowledge and impressive performance across multiple NLP tasks (Qin et al., 2023; Zhao et al., 2023), have the potential to improve intent detection systems in TODS. In this work, we explore how LLMs can be best leveraged for the task of intent detection and assess their ability to handle OOS queries and varying scope of intents.

Contributions. 1. We employ generative LLMs using adaptive in-context learning (ICL) and chain of thought (CoT) prompting for the task of intent detection and compare them against contrastively fine-tuned sentence transformer (SetFit) models, highlighting performance/latency trade-offs. We evaluate 7 SOTA LLMs from Claude and Mistral families on 3 open-source and 3 internal real world datasets.

2. We propose a hybrid system that combines SetFit and LLM by conditionally routing queries to LLM based on SetFit's predictive uncertainty determined using Monte Carlo Dropout. We also propose a negative data augmentation technique that improves SetFit's performance by >5% across datasets. The resulting system achieves performance within ∼2% of native LLM performance with ∼50% less latency than native LLM.

3. We study the behavior of adaptive ICL based intent detection through controlled experiments and show that LLM's OOS detection capability significantly depends upon the scope of intent labels (class design) and the number of labels.

4. We also propose a novel two step methodology utilizing internal LLM representations to help improve LLM's OOS detection capabilities and show empirical gains in OOS detection accuracy and F1-score by >5% across datasets for Mistral-7B.

We intend to also share the three internal datasets after necessary approvals as a community resource and to ensure reproducibility.

2 Related Work

Evaluation of LLMs. LLMs like ChatGPT (OpenAI, 2022), GPT-4 (OpenAI et al., 2024), Claude (Anthropic, 2023), Mistral (Mistral, 2023), Llama (Meta, 2023) have shown impressive performance on multiple NLP tasks and benchmarks (Zhao et al., 2023). Supervised BERT (Devlin et al., 2018) based models have been widely used for intent detection but now with the advent of LLMs it is not clear what benefits they bring for intent detection in the real world. Hence in this work, we evaluate LLMs on the critical task of intent detection for TODS on real world intent detection datasets and highlight performance/latency tradeoffs by benchmarking LLMs with traditional sentence transformers. Recent work (Wang et al., 2024; Liu et al., 2024) majorly focused on evaluation of LLMs on datasets like CLINC150 (Larson et al., 2019),

[2]Out-of-scope (OOS) queries are the ones which do not fall into any of the system's supported intents (Larson et al., 2019).

	SOF Mattress	Curekart	Power Play11	ALC	ADP	OADP	Avg Score	Avg p50 Latency
Claude v1 Instant	0.613	0.528	0.295	0.840	0.687	0.630	0.599	2.297
Claude v2	0.763	0.773	0.665	0.891	0.703	0.630	**0.737**	11.795
Claude v3 Haiku	**0.815**	0.775	0.646	0.849	0.715	0.619	0.736	1.697
Claude v3 Sonnet	0.739	0.647	0.566	0.895	**0.765**	**0.653**	0.711	4.592
Mistral 7B	0.699	0.615	0.384	0.804	0.624	0.453	0.597	1.624
Mixtral 8x7B	0.694	0.614	0.434	0.824	0.653	0.587	0.634	1.992
Mistral Large	0.767	**0.779**	**0.668**	**0.907**	0.688	0.601	0.735	3.565
SetFit (Baseline)	0.632	0.511	0.612	0.769	0.617	0.462	0.600	0.030
SetFit + Neg Aug	0.672	0.709	0.639	0.848	0.625	0.459	0.658	0.030

Table 1: Comparison of F1 Score of various SOTA LLMs with fine tuned sentence transformer models across AID3 and HINT3 datasets

BANKING77 (Casanueva et al., 2020) which are: (i) not real world intent detection datasets (queries are not from deployed TODS), (ii) not multi-label (every query maps to single intent). Instead, our evaluation is on real world intent detection datasets wherein queries are from deployed TODS which have real world challenges like intents with very-broad to very-specific scopes, imbalanced training datasets with very few examples per intent and 3 out of 6 of our datasets are also multi-label which makes our evaluation more comprehensive.

Improving OOS detection performance of LLMs. Recent work (Liu et al., 2024) fine-tuned LLMs to improve OOS performance which is prohibitive both from development and maintenance perspective for a typical Conversational-AI platform which needs to support hundreds of different TODS (because fine-tuning and deploying a separate instance of LLM for every TODS is prohibitively expensive which makes fine-tuning LLMs impractical). Hence, we propose an alternative approach without LLM fine-tuning which improves both OOS accuracy and overall performance by >5% and allows use of the same instance of foundational LLM across TODS.

Hybrid intent detection system which uses LLMs. Unlike prior work, our focus is not just on evaluation of LLMs and/or improving OOS detection performance of LLMs, but we also focus on building a deployable intent detection system which can benefit from LLMs but does not have prohibitive cost and latency, as part of which we propose a hybrid system using uncertainty based routing strategy to combine LLMs and SetFit approaches that along with negative data augmentation results in achieving the best of both worlds (i.e. within 2% of native LLM accuracy with 50% less latency).

Better understanding of LLM's OOS detection capabilities. In this work we do controlled experiments to study the effect of scope of labels and size of label space. Recent work (Wang et al., 2024) also investigated the effect of the size of the label space on LLM's OOS performance and their findings are inline with our findings. However, our findings on how LLM OOS detection capabilities are influenced by the scope of intent labels are novel and would inform label space design during development of TODS.

3 Leveraging LLMs for Intent Detection

In this section we see how LLMs can be best leveraged for intent detection and propose a hybrid system which leverages LLMs conditionally, achieving a balance between performance and cost.

3.1 Methodology

3.1.1 Fine-Tuned Sentence Transformers

We fine tune sentence transformer (SetFit) models in two steps (Tunstall et al., 2022a) and use them as our baseline. In the first step, a sentence transformer model is fine-tuned on the training data in a contrastive, siamese manner on sentence pairs. In the second step, a text classification head is trained using the encoded training data generated by the fine-tuned sentence transformer from the first step.

Negative Data Augmentation. To help SetFit learn better decision boundaries, we augment training data by modifying keywords in sentences by (a) removing, or (b) replacing them with random strings. These modified sentences are considered OOS during training. Since these augmented OOS sentences have similar lexical pattern as in-scope