The 7th International Conference on Interdisciplinary Research
on Computer Science, Psychology, and Education (2023)

# Active Learning Enhanced by LLMs: Empirical Strategies to Rectify Intent Classifier Discrepancies

Changwoo Chun°, HeuiSeok Lim
Korea University
{chun8629,limhseok}@korea.ac.kr

**Abstract**

Intent classification plays a key role in voice assistance systems that aim to accurately identify user commands and take action accordingly. However, as user interaction environments continue to evolve, it becomes essential to detect when the performance of existing models is degrading. Misclassification of intent remains a prevalent problem, especially considering the overwhelming volume of daily user logs. This study introduces an novel methodology that harnesses the capabilities of Large Language Models (LLMs) to address these misclassifications. Through a meticulous examination of sentences yielding divergent predictions across multiple intent classifier versions, the LLM acts as an adjudicator, offering refined intent predictions. Misclassification determination using LLM can be leveraged to evaluate utterances that deviate from the learned distribution and measure performance. It can also be coupled with active learning to achieve performance enhancement of intent classifiers, providing a streamlined solution to enhance intent classification precision in a wide range of production environments.

Keywords: Large Language Models (LLMs), Intent Classification, Misclassification Detection, Automated Active Learning, Voice Assistant Systems

## 1. Introduction

With the widespread use of Bluetooth earphones and the trend towards contactless interaction, voice interfaces have become a common means of communication between humans and machines. From smartphones, vehicle infotainment systems, and even TVs and set-top boxes, voice agent systems are one of the essential elements in our daily lives. [1] Users can simply perform actions that previously used a remote controller, such as "Turn on the radio," "Turn off the TV," and "Turn down the volume," with a single word.

To handle the variety of commands, the system leverages intent classifiers trained on predefined user intent. Over time, the diversity and complexity of user utterances often forces the classifier to process input that falls outside of its training distribution, leading to an increasingly high incidence of misclassification.

Collecting training data is a key strategy for improving the performance of intent classifiers. While collecting additional data is the best way to scale and improve intent classifiers, this can be very difficult in some environments. For this reason, much of the research has been directed towards enriching datasets already collected in order to improves the robustness. However, the most fundamental improvements ultimately require manual work to curate and enrich datasets. This job is still very much a human endeavour, analysing and generating data [2]

In response to these challenges, this study propounds an automated methodology anchored in the prowess of Large Language Models (LLMs). By systematically evaluating sentences which elicit divergent predictions from distinct versions of intent classifiers, the LLM assumes the role of an arbitrator, furnishing a more refined intent classification. This strategy not only helps us find misclassifications, but also provides a feedback loop to continuously improve the classifier. The goal of our approach is to leverage the vast capabilities of LLM to enhance intent classification while automating the task of reducing misclassifications. By minimising the work of human annotators and automated labelling and evaluation from data accumulated in real-time, the system will be able to increase its reliability and accuracy through iterative improvement.

## 2. Related Works

Intent classification, an integral component of voice-assistance systems, has seen continuous evolution aimed at enhancing accuracy in understanding user intent. With advancements in this research area, two primary challenges stand out: the rectification of misclassified recognized intents and the unveiling of novel intentions. [3] While numerous studies utilizing public datasets have made strides in intent classification, many have faced constraints in integrating additional data. Consequently, the emphasis has largely been on identifying new intents and refining the accuracy of classifiers within known domains.

A multitude of studies have adopted various techniques, from traditional text encoders to advanced contrastive learning methods, to address both In-Domain (IND) and Out-of-Domain (OOD) utterances [4, 5, 6]. Even though these methods have been crucial in discovering new intents, they often necessitate either manual intervention from domain experts or semi-automatic processes to sift through expansive user logs [7]. However, a central challenge persists: the misclassification of intents within familiar domains. This issue is particularly pressing, given the substantial daily interactions voice assistant systems handle. Such misclassifications not only compromise user experience but can also hinder the systematic growth of these systems.

Recently, the spotlight has turned to the potential of Large Language Models (LLMs). Prior works have primarily utilized LLMs for new intent detection, tapping into their extensive knowledge to identify patterns overlooked by traditional classifiers [2]. Yet, there remains a noticeable gap in literature concerning the proactive use of LLMs to identify misclassifications. Considering these misclassifications frequently manifest in daily user logs, an automated approach to swiftly detect and address them becomes essential. Analyzing variances in predictions from diverse intent classifiers might offer a valuable solution in this context.

## 3. Methodology

In this study, we propose a simple but effective approach for detecting and improving misclassification in intent inference systems that process large amounts of data. First, we explore which sentences mainly cause misclassifications. Next, we prepare several intent classifiers with different versions, and feed large amounts of data to each intent classifier as input. Based on the outputs of the intent classifiers, we explore which sentences give different results. Last, select candidate cases with a high probability of misclassification, and use LLM on the selected candidate sentences to infer which intent classification result is likely to be correct and what the true intent is. Samples determined to be misclassified are fed back into the training data of the intent classifier and included in an active learning routine for retraining.

### 3.1 Causes of Misclassification

After preliminary examination has identified two main causes of intent misclassification in voice-enabled systems. One is when a sentence has a completely different intent but contains key words used in the existing intent. The other is when new words and expressions that the intent classifier did not see when learning appear. The former is called IND (In-Domain) misclassification, and the latter is called OOD (Out-of-Distribution) misclassification. We decided to focus on finding and improving misclassifications that occur in IND. And misclassifications that occurred near domain boundaries were included as an intersection of IND and OOD misclassifications.

| **Utterance** | **Predicted Intent** |
|---|---|
| Close the window | closeWindow |
| **Clean** the window | closeWindow |
| Close the window | closeWindow |
| Close the **door** | closeWindow |
| Find the traffic light | checkTrafficSignal |
| Find **Lee Moojin's** traffic light | checkTrafficSignal |
| When is my appointment? | checkCalendar |
| When is the **rainy** season? | checkCalendar |

Table 1. examples of misclassification

**Word-Dominant Misclassification** When a particular word consistently appears in a specific intent, classifiers tend to be biased towards predicting that intent whenever the word is present. We named this the Word-Dominant Misclassification (WDM) problem. In order to improve WDM, we needed to understand which tokens were focused on specific intents. However, in most cases, it was unavoidable that certain words had to be used intensively to express the purpose. For example, the word 'sunroof' only appears in intents that open and close the sunroof, but it also appears in other

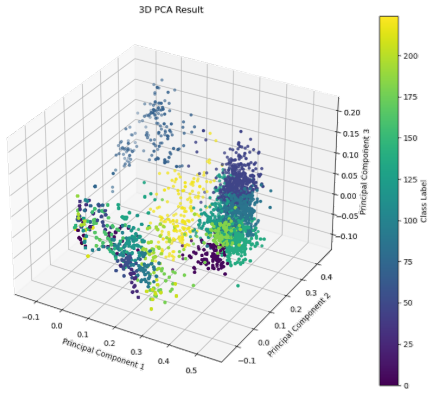


Figure 1. utterance distribution

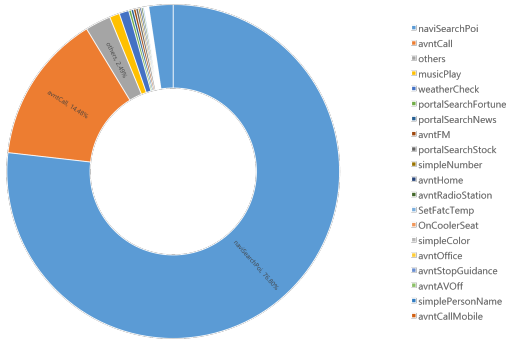


Figure 2. intent statistics

intents, or it is difficult to create sentences using alternative words in intents that control the sunroof. Therefore, a module that judges the meaning of sentences independent of biased learning is needed. Introducing these auxiliary such as LLMs means can improve the overall understanding of the input by alleviating the intent classifier's excessive dependence on single words.

**Unseen Expression Misclassification** Resolving misclassifications caused by completely new words or expressions is more complex and extremely difficult to correct without data enrichment. There are existing studies that define areas outside the existing domain as OOD and determine OOD. However, it is still difficult to accurately classify ambiguous sentences that occur at the border between IND and OOD. In our work, we exclude this area.

### 3.2 Candidate Selection for Misclassification Detection with Ensembled Intent Classifiers

The sheer volume of user logs coming into a large-scale system makes it challenging to sift through them to find only those with a high probability of misclassification. Many commercial intent classifiers use probability-based calibration methods to detect misclassification. However, this also makes it difficult to detect when the intent classifier makes an incorrect judgment and misclassifies with a high probability.

Recent study [8] has shown that LLMs can make inferences as accurate and consistent as humans. That said, feeding all tens of thousands of potentially misclassified sentences into a large language model and inferring the results is not a good approach given the computational cost and efficiency. The problem of culling sentences where intent classifiers are likely to be wrong and updating training data with LLM-based inference should be approached carefully.

We ran multiple versions of the intent classifier in parallel so that different intents could be predicted for a single utterance. We compared four ways to build intent classifiers that train or infer in different ways.

- Dropout Ratio: Applying different dropouts ratio to the same intent classifier
- Diff Epochs: Useing models trained different epochs while being trained on the same data
- Diff Sampling: Using the same intent classes set but samples the training data differently
- Diff Intent set: Using different intent classes set

Applying different dropouts ratio to the same intent classifier has been previously attempted [9] by looking at the same sentence from different perspectives. Choosing models from different epochs is a way to consider both under-optimized and overfitted models from a model optimization perspective, which can lead to different intent classification results. Similarly, using the same intent classes but sampling the training data differently is an attempt to find samples that fall on the class boundaries where subtle differences in representation can lead to different intent classification results. Finally, in our method using different intent classes, we create intent sets $N_1$, $N_2$, ... and $N_K$ that share a very important intent from the overall intent set $N$ but contain other less important intents. We then train an intent classifier on each data separately. We select a total of three intent classifiers from each method to perform large-scale data inference. By comparing the predicted intent outcomes, the misclassifiable sentences are sorted by the diversity of the distribution over which the intent predictions vary. The sentences with the most varied outcomes are considered the most ambiguous,

and these are the candidates for error by the intent classifier in the next step.

**Ensemble of k Intent Classifiers with Diversified Strategies** The effect of an ensemble of k intent classifiers is particularly amplified when each classifier has unique characteristics. In this study, we examined four variants of the same algorithm, trained on different dropouts, different training epochs, different sentence sampling, and different intent sets. Of these approaches, we found that using k intent classifiers trained on k different datasets, each sharing i intents aimed at misclassification detection and varying the remaining intents across the entire N, yielded the broadest set of misclassification candidates.

Table 2. Discrepancy Rates in Intent Classification

| **Method** | **Rate of Discrepancy** |
|---|---|
| Dropout Ratio | 1.15% |
| Diff_epochs | 1.07% |
| Diff_sampling | 1.24% |
| Diff_intent set | **3.55%** |

In a scenario with N intents, each dataset will have distinct training characteristics for each classifier by varying the composition of the remaining intents while retaining the i intents that are important for misclassification detection. As a result, an ensemble of k intent classifiers can be used to identify nuances in user utterances to obtain a candidate set of potential misclassifications for the target intent.

In our experiments, out of a total evaluation dataset of 161,843 utterances, 5,746 (3.55%) resulted in discrepancies between the results of all three intent classifiers. Among these, excluding variants based on different learned intents, 2,320 utterances (1.43%) were identified by the intent classifier as confusing data within a common intent set.

**LLM as the Adjudicator** It has been shown that LLMs can act as human-like evaluators with their vast built-in knowledge and reasoning abilities. [10] In this study, utterance data selected in order of highest ambiguity is input into LLM as the adjudicator. The LLM determines the relationship between the actual meaning of the utterance and the predicted intention. At first, the LLM analyzes the meaning of the utterance, then uses the generated result as a prompt, and finally, LLM outputs a judgment result as to whether

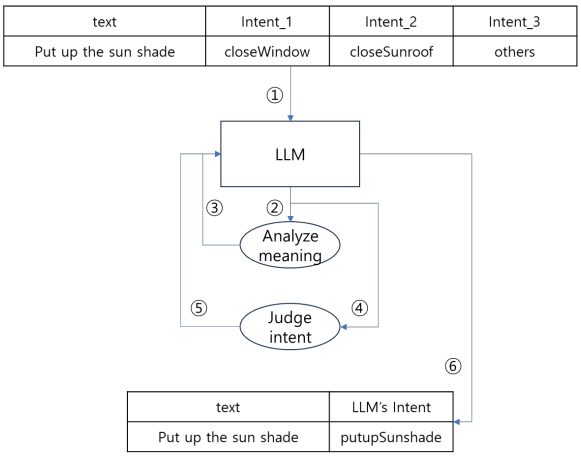


Figure 3. LLM's Adjudicating Flows

the intent prediction of the intent classifier is appropriate or incorrect.

| | **A group** | **B group** | **C group** |
|---|---|---|---|
| Agreement rate | 91% | 83% | 81% |

Table 3. Agreement rate between LLM judgment results and human expert evaluation

The 2,320 selected ambiguous utterances were divided into three groups and LLM was used to evaluate the prediction results of the intent classifier. Based on the frequency of utterance, it was divided into three groups: Group A (395 sentences) uttered more than 100 times, Group B (1,348 sentences) uttered more than 30 times but less than 100 times, and Group C (577 sentences) uttered 30 or less times. We randomly sampled 100 sentences from each group respectively, and let LLM judge the results of intent classifiers and infer the correct intent. When human experts evaluated the final results of the LLM, the average agreement rate was 85.46%. Empirical evidence from our experiments shows that the adjudication ability of the LLM can mimic that of human experts. However, it was also confirmed that the agreement rate between LLM and human experts decreased as one went to the low-frequency group.

$$\exists x_i \rightarrow I_i$$

If the intent classifier results parallelized in an ensemble are different for a sentence $x_i$ whose meaning is confusing, the LLM selects the actual intent that reflects the meaning of the sentence. (Correct answer intent label $y_i$) The intention determination process can be expressed as input to a large-

scale language model as follows.

$$LLM(P, f(x_i, y_1), ..., f(x_i, y_n)) \rightarrow \hat{y}_i$$

- $P$: Instruction prompt
- $f$: Demonstrations
- $x_i$: input sentence
- $y_k$: Intent predicted by kth intent classifier
- $\hat{y}_i$: Actual intent determined by LLM

### 3.3 Active Learning for Rectifying the Intent Classifier

Based on LLM's rich built-in knowledge and reasoning capabilities, we utilize refined data to retrain the intent classifier to improve performance. By accurately labeling confusing data near class boundaries, feeding it to the classifier, and retraining it, we observed a improvement in system performance.

The experiments showed an average performance improvement of 0.58% per retraining whenever 100 IND misclassifications candidates were revised to the correct label by the LLM adjudicator. This shows that the approach combined with active learning helps to continuously improve and strengthen the intent classifier by effectively detecting IND misclassifications and inferring the correct label based on LLM.

## 4. Conclusion

In this paper, we attempted to solve problems arising from new words and new expressions, which are one of the main causes of misclassification in intent classifiers. Based on LLM's extensive knowledge and reasoning capabilities, we inferred the meaning of new sentences and determined whether the interpreted meaning matched the predicted intended outcome. And if there was no match, we generated the correct intent label. The proposed methodology offers two major advantages. First, the LLM-based labeling task helps improve the accuracy of the intention classifier and ensures robust classification performance. Second, labeling tasks at the level of human experts can be automated and combined with active learning. This ensures that the model remains adaptable and reliable in a constantly evolving and changing user interaction environment.

## 5. Limitation

While our research has made significant strides in addressing the challenges of intent misclassification, there are inherent limitations that merit acknowledgment. Among the sentences identified by the intent classifier as having a high potential for misclassification, several were notably ambiguous even upon human evaluation. Such sentences presented inherent challenges as they were inherently equivocal, making it difficult even for human experts to discern the underlying intent with clarity.

Furthermore, in the context of these ambiguous sentences, there was observed discrepancy between human experts and the Large Language Models (LLMs). Specifically, while human experts grappled with discerning a clear intent, the LLMs, in their design to execute directives, proceeded to classify the intent. This resulted in occasional variances in intent classification outcomes when compared with human expert assessments. It underscores the broader challenge of ensuring alignment between human interpretation and machine-driven classifications, especially when the input data is inherently fraught with ambiguity.

## Acknowledgements

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2022-2018-0-01405) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation). This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425).

## Reference

[1] 김민경, 김채원, and 원숙영, "A study on the development of context framework for understanding users intention in an environment of automotive voice interface," *한국 HCI 학회 학술대회*, pp. 582–587, 2020.

[2] R. Kumar, M. Patidar, V. Varshney, L. Vig, and G. Shroff, "Intent detection and discovery from user logs via deep semi-supervised contrastive clustering," *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1836–1853, 2022. [Online]. Available: https://aclanthology.org/2022.naacl-main.134

| Utterance | Intent1 | Intent2 | Intent3 | LLM | Human |
|---|---|---|---|---|---|
| 미세먼지 | weatherCheckDust | naviSearchPoi | weatherCheck | weatherCheckDust | 1 |
| 그만 | naviSearchPoi | avntExit | avntStop | avntStop | 1 |
| 멜론 | naviSearchPoi | avntGoto | playMusic | playMusic | 1 |
| 에어컨 | naviSearchPoi | fatcAirconOn | others | fatcAirconOn | 1 |
| 하하하 | naviSearchPoi | chitchatExclamation | others | laughter | 1 |
| 운세 | naviSearchPoi | portalSearchFortune | portalSearchFortune | portalSearchFortune | 1 |
| 조수석 창문 올려줘 | closeWindowPassenger | closeWindowPosition | closeWindowPassenger | closeWindowPassenger | 1 |
| 통풍 시트 | seatCoolingOn | naviSearchPoi | seatCoolingOn | seatCoolingOn | 1 |
| 에어컨 좀 줄여줘 | fatcAirconOff | fatcTempDown | setDownWind | fatcAirconOff | 0 |
| 창문 다 내려줘 | openWindowAll | openWindow | openWindowAll | openWindow | 0 |
| tv 꺼줘 | showoffCam | avntDMB | avntOffVolume | tvOff | 1 |
| 비상등 켜줘 | wheelHeatingOn | seatHeatingOn | others | turnonHazardlights | 1 |
| 와이퍼 작동 | settingsOn | fatcAirconOn | others | wiperOn | 1 |
| 업비트 비트코인 | portalSearchStock | naviSearchPoi | others | cryptoCurrencyTrade | 1 |
| 다른 경로 | avntReroute | avntRouteOption | avntResumeGuidance | avntReroute | 1 |

Table 4. Results of LLM Adjudicator

Table 5. Iterations of Active Learning

| Step | Accuracy | Improvement Rate |
|---|---|---|
| Initial | 0.951 | - |
| 1-step | 0.958 | 0.74% |
| 2-step | 0.959 | 0.10% |
| 3-step | 0.968 | 0.94% |
| 4-step | 0.974 | 0.62% |
| 5-step | 0.979 | 0.51% |

| Utterance |
|---|
| 창문 다 열어 |
| 바람 중간으로 |
| 에어컨 좀 내려줘 |
| 길 안내 전화 |
| 응 아니야 |
| 오늘 추워 |

Table 6. examples of ambiguous utterance

[3] D. R. Changoo Chun, "Novel intent discovery utilizing large language models and active learning strategies," 2023.

[4] H.-Y. K. Hyeok-Ju Ahn, "Comparing the performances of intent classifications by encoder layer," pp. 410–413, 2021.

[5] J. Lim, S. Son, S. Lee, C. Chun, S. Park, Y. Hur, and H. Lim, "Intent classification and slot filling model for in-vehicle services in korean," *Applied Sciences*, Vol. 12, No. 23, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/23/12438

[6] Y. Zhang, H. Zhang, L.-M. Zhan, X.-M. Wu, and A. Y. S. Lam, "New intent discovery with pre-training and contrastive learning," 2022.

[7] Y. Mou, K. He, P. Wang, Y. Wu, J. Wang, W. Wu, and W. Xu, "Watch the neighbors: A unified k-nearest neighbor contrastive learning framework for ood intent discovery," 2022.

[8] C.-H. Chiang and H.-y. Lee, "Can large language models be an alternative to human evaluations?" *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., pp. 15 607–15 631, Jul. 2023. [Online]. Available: https://aclanthology.org/2023.acl-long.870

[9] T. Gao, X. Yao, and D. Chen, "Simcse: Simple contrastive learning of sentence embeddings," 2022.

[10] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang,

Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen, "A survey of large language models," 2023.