# Logistic Regression-Based Example Selection for Enhanced Few-Shot Learning in Intent Classification

Gyutae Park
*Department of Artificial Intelligence*
*Chung-Ang University*
Seoul, Republic of Korea
pkt0401@cau.ac.kr

Hwanhee Lee[†]
*Department of Artificial Intelligence*
*Chung-Ang University*
Seoul, Republic of Korea
hwanheelee@cau.ac.kr

2025 IEEE International Conference on Consumer Electronics (ICCE) | 979-8-3315-2116-5/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICCE63647.2025.10930071

***Abstract*—This study introduces a novel approach to example selection in few-shot learning scenarios for dialog intent classification, leveraging logistic regression to refine the set of examples retrieved through traditional similarity-based methods. We evaluate our method on three benchmark datasets: BANKING77, CLINC150, and HWU64, using 5-shot and 10-shot learning setups with 20 demonstrations. Our results show improvements in classification accuracy compared to baseline similarity-based retrieval methods, particularly for semantically similar intents. Notably, we observe a reduction in misclassifications within similar domains after applying our proposed approach. This work contributes to the growing body of research on efficient few-shot learning techniques for natural language understanding tasks, offering insights into enhancing performance in challenging, domain-specific scenarios.**

***Index Terms*—Few-shot learning, Multi-class classification, Intent classification, Logistic regression, Example selection**

## I. Introduction

Intent classification is the task of identifying the underlying intent behind a user's input in a dialog system. This task plays a key role particularly in building dialogue systems and customer service automation. Traditional intent classification methods require large labeled datasets and have the disadvantage of needing to retrain the model whenever new intents are added. [1] As an alternative, few-shot learning methods based on Large Language Models (LLMs), utilizing in-context learning, have gained attention. But they face limitations in including examples for all intents in the prompt when the number of intents is large. [2], [3] Thus, research on optimizing the application of LLMs for multi-class intent classification remains still insufficient. [4] To overcome these limitations, we propose a new approach that combines similarity-based example selection methods using a retriever with logistic regression. Our method enables effective handling of a large number of intents by selecting only a small number of examples most similar to the query to include in the prompt. Our experiments on three benchmark datasets (HWU64, CLINC150 and BANKING77) demonstrate consistent performance improvements over baseline methods in both 5-shot and 10-shot settings. Notably, we observe a significant reduction in misclassifications within similar domains across all datasets. These results suggest that our approach is particularly effective at distinguishing between subtle differences in semantically similar intents, even with limited examples per intent.

## II. Methods

Our proposed method for enhancing few-shot learning in intent classification combines the strengths of similarity-based retrieval and logistic regression to select the most relevant and informative examples for a given query. Our approach aims to overcome the limitations of traditional few-shot learning methods when dealing with a large number of intents.

### *A. Initial Similarity-Based Example Selection*

In multi-class classification problems, it is practically impossible to include examples for all intents in the prompt. Therefore, our work selects initial examples through the following process:

1) Sentence Embedding: Use a pre-trained SentenceTransformer model (all-mpnet-base-v2) [1] to embed all training data and queries into vector space.
2) Cosine Similarity Computation: Compute the cosine similarity between the query embedding and all training data embeddings.
3) Initial Example Selection: Select k examples (k = 60, 80, 100, 120) with the highest similarity.

### *B. Embedding Space Transformation through Logistic Regression*

Simple similarity-based selection has limitations in distinguishing examples that are semantically similar but have different intents. [5] To overcome this limitation, we perform embedding space transformation through logistic regression with the following formula:

$$P(y = c|x) = \sigma(w_c^T x + b_c), \tag{1}$$

[†] Corresponding Author.

[1] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

979-8-3315-2116-5/25/$31.00 ©2025 IEEE 979-8-3315-2116-5

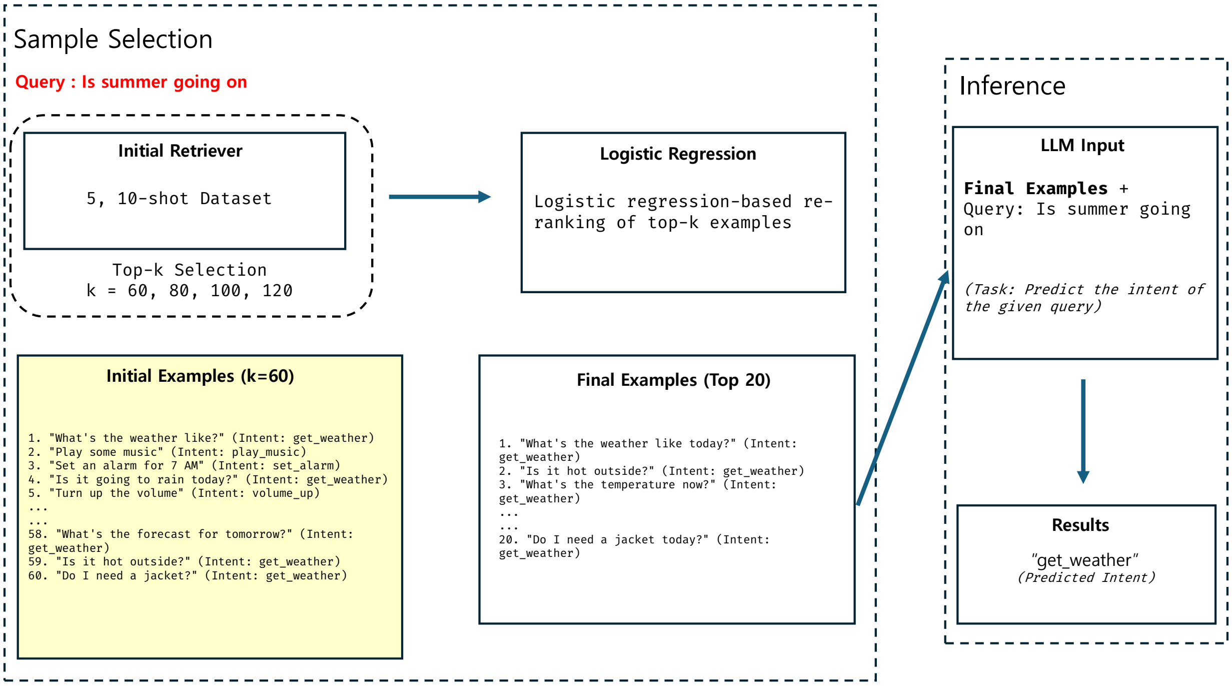
Sample Selection
Query : Is summer going on
Initial Retriever
5, 10-shot Dataset
Top-k Selection
k = 60, 80, 100, 120
Logistic Regression
Logistic regression-based re-ranking of top-k examples
Initial Examples (k=60)
1. "What's the weather like?" (Intent: get_weather)
2. "Play some music" (Intent: play_music)
3. "Set an alarm for 7 AM" (Intent: set_alarm)
4. "Is it going to rain today?" (Intent: get_weather)
5. "Turn up the volume" (Intent: volume_up)
...
...
58. "What's the forecast for tomorrow?" (Intent: get_weather)
59. "Is it hot outside?" (Intent: get_weather)
60. "Do I need a jacket?" (Intent: get_weather)
Final Examples (Top 20)
1. "What's the weather like today?" (Intent: get_weather)
2. "Is it hot outside?" (Intent: get_weather)
3. "What's the temperature now?" (Intent: get_weather)
...
...
20. "Do I need a jacket today?" (Intent: get_weather)
Inference
LLM Input
Final Examples +
Query: Is summer going on
(Task: Predict the intent of the given query)
Results
"get_weather"
(Predicted Intent)

Fig. 1. Detailed flow diagram of the proposed method

where $\sigma$ is the sigmoid function, $w_c^T$ is the weight vector for class c, and $b_c$ is the bias term. We transform the original embedding space using the weight matrix $W_x$ of the trained logistic regression model as follows:

$$x' = W_x x \tag{2}$$

To train the logistic regression model, we use the initially selected k examples as training data. We train the model using cross-entropy loss and optimize it using the L-BFGS algorithm, which is particularly effective for small to medium-sized datasets. [6], [7]

```
**Algorithm 1** Logistic Regression-based Example Selection
**Require:** Query $q$, Dataset $D$, Initial selection size $k$, Final selection size $m = 20$
**Ensure:** Final set of $m$ examples
$E \leftarrow \text{Retriever}(q, D, k)$ {Select initial $k$ examples}
$X \leftarrow \text{Embed}(E)$ {Embed selected examples}
$y \leftarrow \text{Labels}(E)$ {Get labels of examples}
$W \leftarrow \text{TrainLogisticRegression}(X, y)$
    $X_{\text{new}} \leftarrow WX$ {Transform embedding space}
    $S \leftarrow \text{CosineSimilarity}(\text{Embed}(q), X_{\text{new}})$
    $E_{\text{final}} \leftarrow \text{TopK}(E, S, m)$
**return** $E_{\text{final}}$
```

### C. Final Example Selection and Constructing LLM Input

In the transformed embedding space, we select the final examples and construct input for the LLM through the following process:

1) New Similarity Calculation: Recalculate the similarity between the query and examples in the transformed space.
2) Final Example Selection: Sort by similarity score in descending order and select the top 20 examples.
3) Constructing LLM Input: Arrange the selected examples from lowest to highest similarity and input them to the LLM. This allows the LLM to process progressively more relevant information, potentially giving more weight to the most important information in the final prediction.

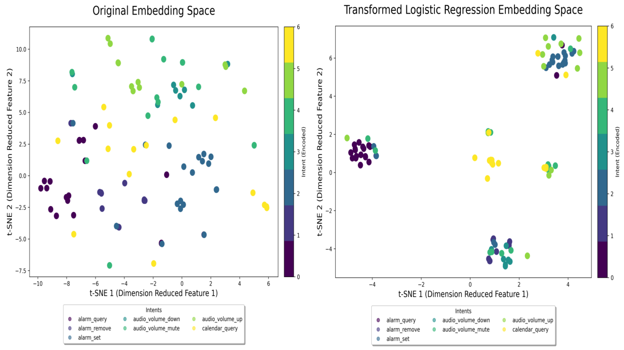

Fig. 2. Comparison of t-SNE visualizations for HWU64 dataset with 60 examples: (left) Original Embedding Space, (right) Transformed Logistic Regression Embedding Space

## III. Experimental Setup

In this section, we outline the datasets and experimental configuration for evaluating our logistic regression-based example selection method for few-shot intent classification. We describe the models used, the datasets employed, and the specific experimental settings for our study.

**Models Used:**

- Large Language Model: We use the Llama 2 7B model [8] developed by Meta AI. This model has 7 billion parameters and was trained on public web crawling data and conversational data.
- Retriever: We use the all-mpnet-base-v2 model based on Sentence-BERT. This model is effective in generating sentence embeddings and captures semantic similarities well.

**Datasets:**

- HWU64 [9]: A dataset related to smart home device control, covering 21 domains with 64 intents.
- CLINC150 [10]: A dataset covering various daily tasks and services across 10 domains, containing 150 intents.
- BANKING77 [11]: A dataset related to financial and banking services, containing 77 intents.

**Experimental Configuration:**

- Few-shot Learning Setup: This study used the existing split from the Dialoglue dataset [12]. This split provides 5-shot and 10-shot settings, including 5 or 10 examples per intent.
- Baseline: A method that directly selects the top 20 examples based on simple similarity.
- Proposed Method: Initially selects K examples (60, 80, 100, 120), then chooses the final 20 through logistic regression.
- Evaluation Metric: Accuracy

## IV. Results and Analysis

We conduct experiments on three intent classification datasets: BANKING77, CLINC150, and HWU64, using both 5-shot and 10-shot settings. The baseline method directly uses the top 20 examples based on retriever similarity. Our proposed method initially selects 60, 80, 100, or 120 examples, then chooses the final 20 using logistic regression.

### A. Performance Comparison

Table I,II shows the accuracy results for each dataset.

Upon analyzing the results, we demonstrate that the proposed method consistently outperforms the baseline across the majority of cases. Based on the results, we identify the following notable findings.

- The HWU64 dataset showed the largest performance improvement. In the 5-shot setting, accuracy increased from the baseline of 85.32% to a maximum of 87.14%, and in the 10-shot setting, from 86.62% to 88.57%.
- For the CLINC150 dataset, although the improvement was smaller, it was consistent. This may be due to the already high baseline performance (over 93%), making further improvements challenging.
- The BANKING77 dataset also showed modest improvements. In the 5-shot setting, accuracy improved from the baseline of 85.55% to a maximum of 86.40%.

TABLE I
Accuracy Results (%) for 5-shot Setting

| Dataset | Baseline | Proposed Method (K) | | | |
|---|---|---|---|---|---|
| | (K=20) | 60 | 80 | 100 | 120 |
| HWU64 | 85.32 | 85.78 | 85.87 | **87.14** | 86.71 |
| CLINC150 | 93.22 | 93.36 | **93.49** | 93.38 | 93.25 |
| BANKING77 | 85.55 | 85.62 | **86.40** | 86.17 | 86.33 |

TABLE II
Accuracy Results (%) for 10-shot Setting

| Dataset | Baseline | Proposed Method (K) | | | |
|---|---|---|---|---|---|
| | (K=20) | 60 | 80 | 100 | 120 |
| HWU64 | 86.62 | 86.90 | 88.29 | **88.57** | 88.20 |
| CLINC150 | 94.26 | **94.39** | 94.28 | 94.20 | 94.18 |
| BANKING77 | 89.22 | **89.38** | 89.18 | 89.10 | 89.14 |

These results suggest that the proposed method is particularly effective at distinguishing between subtle differences in similar intents. In other words, the logistic regression-based transformation of the embedding space appears to help in selecting more semantically relevant examples.

### B. Domain-specific Misclassification Analysis

To further understand the effectiveness of our method, we conduct a domain-specific misclassification analysis on the HWU64 dataset, in the 10-shot setting. Figure 3 shows the number of misclassifications within the same domain and across different domains for both the baseline and proposed methods.

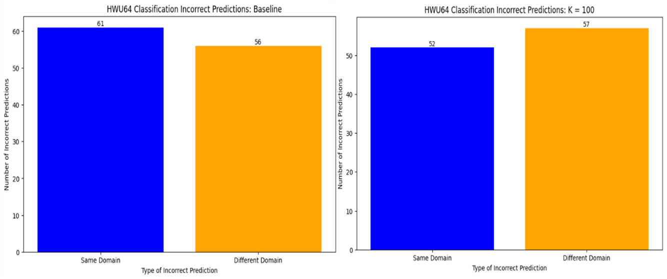

Fig. 3. Domain-specific misclassification analysis for (left) the baseline model and (right) after applying logistic regression-based example selection (K=100) on HWU64 (10-shot)

The analysis reveals the following findings:

- The baseline model (Fig. 3, left) had 61 misclassifications within the same domain and 56 misclassifications across different domains. This suggests that the baseline model struggled to distinguish between intents within similar domains.

- After applying our proposed method (Fig. 3, right), misclassifications within the same domain decreased to 52, while misclassifications across different domains slightly increased to 57.
- This reduction in within-domain misclassifications (from 61 to 52) indicates that our method significantly improved the model's ability to distinguish between similar intents within the same domain.

These findings indicate that our logistic regression-based example selection method is particularly effective at improving intent classification performance for semantically similar intents within the same domain.

### C. Impact of Initial Example Set Size

We varies the initial number of selected examples (K) from 60 to 120 and observe the following findings:

- Performance generally peaked when K was 80 or 100.
- Smaller K values (60) may not provide sufficient information for effective logistic regression.
- Larger K values (120) might introduce noise, slightly degrading performance.

## V. Conclusion and Future Work

This study proposed a novel approach to example selection in few-shot learning scenarios for intent classification, leveraging logistic regression to refine the set of examples retrieved through traditional similarity-based methods. Our experiments on three benchmark datasets (HWU64, CLINC150, and BANKING77) yielded several key findings:

- Performance Improvement: Our method consistently outperformed the baseline similarity-based retrieval across all datasets, with the most significant improvements observed in the HWU64 dataset.
- Effective Disambiguation: The logistic regression-based transformation of the embedding space proved particularly effective in distinguishing between semantically similar intents, as evidenced by the reduction in within-domain misclassifications.
- Optimal Example Set Size: We found that initial selection of 80 to 100 examples generally yielded the best performance, balancing between sufficient information and minimal noise.
- Generalizability: The proposed method showed effectiveness across various domains (smart home, daily tasks, and banking), suggesting good generalization capabilities.

These results demonstrate the potential of our approach in enhancing few-shot learning performance for intent classification, particularly in scenarios with many semantically similar intents.

## Acknowledgement

This research was supported by Institute for Information & Communications Technology Planning & Evaluation (IITP) through the Korea government (MSIT) under Grant No. 2021-0-01341 (Artificial Intelligence Graduate School Program (Chung-Ang University)).

## References

[1] C. Zhang, Y. Li, N. Du, W. Fan, and P. Yu, "Joint Slot Filling and Intent Detection via Capsule Neural Networks," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019, pp. 5259–5267.

[2] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.(2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901. Curran Associates, Inc.

[3] O. Rubin, J. Herzig, and J. Berant, "Learning To Retrieve Prompts for In-Context Learning," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States, 2022, pp. 2655–2671.

[4] D. Yu, L. He, Y. Zhang, X. Du, P. Pasupat, and Q. Li, "Few-shot Intent Classification and Slot Filling with Retrieved Examples," arXiv preprint arXiv:2104.05763, 2021.

[5] J. Zhang et al., "Discriminative Nearest Neighbor Few-Shot Intent Detection by Transferring Natural Language Inference," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 5064–5082.

[6] G. Andrew and J. Gao, "Scalable training of L1-regularized log-linear models," in *Proc. 24th Int. Conf. Mach. Learn.*, Corvalis, OR, USA, 2007, pp. 33–40.

[7] Q. V. Le, J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, and A. Y. Ng, "On optimization methods for deep learning," in *Proc. 28th Int. Conf. Mach. Learn.*, Bellevue, WA, USA, 2011, pp. 265–272.

[8] H. Touvron et al., "Llama 2: Open Foundation and Fine-Tuned Chat Models," arXiv preprint arXiv:2307.09288, 2023. [Online]. Available: https://huggingface.co/meta-llama/Llama-2-7b

[9] I. Casanueva, T. Temčinas, D. Gerz, M. Henderson, and I. Vulić, "Efficient Intent Detection with Dual Sentence Encoders," *arXiv preprint arXiv:2003.04807*, 2020. [Online]. Available: https://arxiv.org/abs/2003.04807

[10] X. Liu, A. Eshghi, P. Swietojanski, and V. Rieser, "Benchmarking Natural Language Understanding Services for Building Conversational Agents," in *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, E. Marchi, S. M. Siniscalchi, S. Cumani, V. M. Salerno, and H. Li, Eds. Singapore: Springer Singapore, 2021, pp. 165–183.

[11] S. Larson et al., "An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction," in *Proc. 2019 Conf. Empirical Methods Natural Language Process. 9th Int. Joint Conf. Natural Language Process. (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 1311–1316.

[12] S. Mehri, M. Eric, and D. Hakkani-Tur, "DialoGLUE: A Natural Language Understanding Benchmark for Task-Oriented Dialogue," arXiv preprint arXiv:2009.13570, 2020.