

New Intent Discovery: From Local Relations to Global Structure

Peng Chen¹, Bang Wang^{2,*},

¹School of Software Engineering,

Huazhong University of Science and Technology, Wuhan, China

²School of Electronic Information and Communications,

Huazhong University of Science and Technology, Wuhan, China

{hustchenpeng, wangbang}@hust.edu.cn

Abstract

New Intent Discovery (NID) aims to uncover unseen intent categories from unlabeled user queries, with the core objective of learning an intent representation space with a clear clustering structure. However, in the absence of explicit supervision, how to reliably form a global clustering structure remains challenging. To address this issue, we start from the constituent units of the global structure, namely local relations, and introduce more accessible and more stable forms of local supervision. Based on this idea, we propose LR2GS, which directly trains the intent encoder by leveraging multiple complementary local supervision signals, allowing local constraints to continuously act on the training process and thereby progressively shaping a well-separated global intent structure. Extensive experimental results demonstrate that LR2GS achieves state-of-the-art performance in both unsupervised and semi-supervised settings. Our code is available at: <https://github.com/hustchenpeng/LR2GS>

1 Introduction

Traditional intent detection systems have achieved remarkable success under supervised learning paradigms in closed-set scenarios, where all intent categories are pre-defined and each *query* q in a *query set* Q is annotated with a corresponding intent label (Singhal et al., 2023; Zhang et al., 2023). However, real-world user queries are inherently dynamic, with new intents continuously emerging beyond pre-defined categories (Chen and Wang, 2025). With the goal of uncovering previously unseen intent categories from user queries, Zhang et al. (2021); Tang et al. (2024) proposed the task of *New Intent Discovery* (NID).

NID aims to identify latent intent clusters from Q under an unsupervised setting. Its objective can be formulated as training an *intent encoder* f_θ to

encode Q into an *intent representation space* $\mathbb{R}$, such that queries with the same latent intent are grouped into the same *intent cluster* in this space, while the boundaries between different intent clusters are clearly separable. We refer to the intent clusters obtained in the intent representation space as the *global structure*. Obviously, an ideal global structure should have both *intra-cluster compactness* and *inter-cluster separability*.

A common baseline for NID is to apply the K-Means clustering to queries' representations encoded by a pretrained language encoder, such as the BERT (Devlin et al., 2019). However, in the absence of explicit supervision, the encoder often fails to encode the Q into a well-structured $\mathbb{R}$ for an ideal global structure. Previous NID method (Zhang et al., 2021) mitigates this issue by adopting iterative pseudo-label alignment; while recent works leverage *large language models* (LLMs) to generate auxiliary supervision signals for training an intent encoder (Viswanathan et al., 2024; Liang et al., 2024a,b; Rodriguez et al., 2024). Despite their effectiveness, local relations in between queries have not been sufficiently explored for contributing a better global structure.

In this paper, we argue that an ideal global structure essentially emerges from the accumulation of precise local relations: Queries with the consistent latent intents are continuously pulled closer; while those with inconsistent intents are stably pushed apart. These local relations explicitly specify which queries should be only regarded as variants of the same intent, thereby providing the most fundamental structural constraints for intra-cluster compactness and inter-cluster separability. Compared with directly inferring a global structure, some local semantic consistency is easier to reliably evaluate, and thus can provide supervision signals with lower noise and finer granularity.

We focus on two complementary aspects of local relations: One is inter-query semantic neighbor

*Corresponding author.

relations, which constrain two semantically similar queries to be pulled closer in the intent representation space; the other is intra-query intent consistency, which requires the variants of the same query to be assigned to the same cluster. We argue that the correctness of these two types of local relations helps determine credible same-cluster judgments, and through their continual accumulation over all queries, shapes a better global structure.

For exploiting inter-query relations, we leverage the retrieval capability of a semantic retrieval model and the semantic understanding of an LLM to obtain precise neighbor relations for each query. However, these neighbor relations do not naturally emerge in the representation space of the intent encoder. Therefore, we inject these neighbor pairs as explicit supervision to train the encoder. Beyond inter-query neighbors, we further exploit intra-query consistency to provide complementary local supervision. Real-world user queries vary widely in form. Some are overly brief, while others are verbose and noisy. Consequently, a single raw query may not reliably expose the core intent to the encoder. By using an LLM to generate a descriptive variant and a prototype variant for each query, we enforce consistency among their representations.

After the intent encoder has learned the above local relations, the global structure becomes clearer, while a small number of queries near cluster boundaries remain ambiguous. For these ambiguous queries, we use the collaboration between an LLM and a semantic retrieval model to identify and correct unreliable local relations. This process yields calibrated local relations that serve as explicit supervision for further refining the intent representation space and disambiguating the global structure.

Based on the above analysis, we propose **LR2GS**, a framework from local relations to global structure. As illustrated in Figure 1, LR2GS consists of three core modules. First, it constructs *local relations from semantic neighbors*, where a semantic retrieval model is used to retrieve neighbors for each query, and an LLM-based neighbor refinement step is further applied to improve the precision of the retrieved relations. These neighbor relations are formulated as alignment objectives that encourage representations of neighboring queries to be pulled closer in the intent representation space. Second, LR2GS derives *local relations from intra-instance variants* by using an LLM to generate descriptive and prototype variants for each query. Consistency between a query and its

variants is enforced through contrastive-style objectives. Third, it introduces a *global structure disambiguation* mechanism to handle ambiguous queries near cluster boundaries. For these queries, calibrated positives are used as contrastive-style correction supervision, pulling ambiguous queries toward their corrected positives. By integrating the above training objectives into a unified loss function, LR2GS trains an intent encoder with an ideal global structure.

Comprehensive experiments on three benchmark datasets demonstrate that LR2GS achieves state-of-the-art performance in both unsupervised and semi-supervised settings.

A comprehensive discussion of related work is deferred to Appendix A.

2 Method

2.1 Task Definition and Approach Overview

The NID task aims to discover the complete intent space $\mathcal{I} = \mathcal{I}_k \cup \mathcal{I}_{uk}$, where $\mathcal{I}_k$ is the set of known intents and $\mathcal{I}_{uk}$ is the mutually exclusive set of unknown intents. The training data contain a small labeled part $\mathcal{D}^l = \{(x_i, y_i) \mid y_i \in \mathcal{I}_k\}$, and an unlabeled part $\mathcal{D}^u = \{x_j \mid \text{intent}(x_j) \in \mathcal{I}\}$, covering both known and unknown intents. We denote the entire training set as $\mathcal{D} = \mathcal{D}^l \cup \mathcal{D}^u$, which contains N queries in total. Based on $\mathcal{D}$, the objective is to train an intent encoder f_θ . Note that if $\mathcal{I}_k = \emptyset$, the task degenerates into a fully unsupervised setting (Viswanathan et al., 2024; Liang et al., 2024a,b).

We propose the LR2GS framework to train an intent encoder f_θ , which explicitly constructs multiple forms of supervision signals for training the encoder, thereby continuously shaping the local relations of the intent representation space, such that they can gradually evolve to a global structure with favorable properties of intra-cluster compactness and inter-cluster separability. Figure 1 illustrates our LR2GS framework, which consists of three core modules.

2.2 Local Relation from Semantic Neighbors

We use a semantic retrieval model T (such as the all-mpnet-base-v2 (Reimers and Gurevych, 2019; Song et al., 2020)) to construct neighbor pairs. Specifically, for each query x_i in $\mathcal{D}$, we encode its representation with T and search for the most

LR2GS: From Local Relations to Global Structure

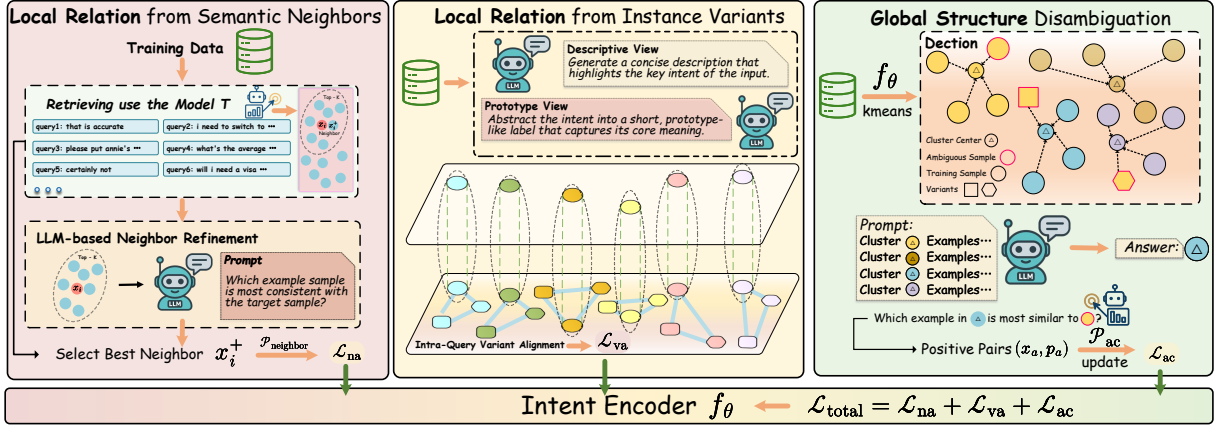


Figure 1: Overview of the proposed framework under the unsupervised setting. In the semi-supervised case, the overall objective additionally incorporates the cross-entropy loss from labeled queries.

semantically similar query in $\mathcal{D}$:

$$x_i^+ = \arg \max_{x_j, j \neq i} \text{sim}(T(x_i), T(x_j)).$$

The resulting local neighbor pair set is:

$$\mathcal{P}_{\text{neighbor}} = \{(x_i, x_i^+)\}_{i=1}^N.$$

This set remains unchanged during training, providing stable local neighbor supervision.

LLM-based Neighbor Refinement. Although the model T provides strong semantic retrieval signals, similarity-based retrieval alone may be insufficient for fine-grained queries. To further refine neighbor pairs, we introduce an optional *LLM-based Neighbor Refinement* (LNR) module that collaborates with T to provide complementary semantic judgments. For each query x_i , we first retrieve its Top- K_{nn} most semantically similar candidates based on the retrieval model T , forming a candidate set $C_i = \{x_{ik}\}_k$. We design a prompt using these candidates together with x_i . Then, the LLM compares intent consistency within the candidate set and selects the query most consistent with x_i . The prompt details are provided in Appendix F.1. Only the refined neighbor pair is kept in $\mathcal{P}_{\text{neighbor}}$ and used as a more reliable supervision signal for encoder training. It is worth noting that although this module increases computational cost, it is fully plug-and-play and is designed to provide a more precise supervision signal for achieving better clustering performance.

Inter-Query Neighbor Alignment. We construct local neighbor supervision by pairing each

query x_i with its semantic neighbor x_i^+ . Let $\mathbf{z}_i = f_\theta(x_i)$ and $\mathbf{z}_{p_i} = f_\theta(x_i^+)$ denote their representations, respectively. We define the neighbor pair similarity as $s_i = \text{sim}(\mathbf{z}_i, \mathbf{z}_{p_i})$ using cosine similarity. For all the N queries, we apply a soft-max normalization over their similarity scores to obtain a calibrated weight distribution, and define the *inter-query neighbor alignment loss* as

$$\mathcal{L}_{na} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(s_i/\tau)}{\sum_{j=1}^N \exp(s_j/\tau)}.$$

By continuously increasing the similarity between each query and its neighbor, $\mathcal{L}_{na}$ encourages semantically similar neighboring queries to be pulled closer in the intent representation space, thereby enhancing intra-cluster compactness for semantic neighbors with the same intent and ultimately facilitating the formation of a clear global structure.

2.3 Local Relation from Query Variants

We argue that relying solely on inter-query semantic neighborhood relations provides insufficient local relations for training an intent encoder. So we propose to further capture a kind of intra-query intent consistency also as local relations for encoder training. To this end, we design two prompts to let an LLM generate two complementary variants for each query x_i : a **Descriptive variant** (x_i^{v1}), which produces a semantically complete description for the query x_i ; and a **Prototype variant** (x_i^{v2}), which summarizes the query into a compact prototype-level label. The prompt details are provided in Appendix F.2.

The two variants reflect the descriptive and abstract perspectives for a query: The descriptive variant retains fine-grained semantic cues; while the prototype variant provides highly condensed conceptual expressions. Although the two variants differ in their expressions, the semantic cores of both variants must remain consistent with the raw query, which can be exploited to train the intent encoder to encode them into the same cluster in the intent representation space.

Intra-Query Variant Alignment. We use contrastive learning to enforce consistency between the raw query representation and its variants' representations in the intent representation space. We compute cosine similarities between $\mathbf{z}_i$ and the k -th variant representations $\mathbf{z}_j^{v_k}$ for all queries. Given N queries, the *intra-query variant alignment loss* is defined as:

$$\mathcal{L}_{va} = -\frac{1}{2N} \sum_{i=1}^N \sum_{k=1}^2 \log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_i^{v_k})/\tau)}{\sum_{j=1}^N \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j^{v_k})/\tau)}.$$

It is worth noting that this loss helps the intent encoder f_θ better capture the intent-level semantics of queries by treating variants as alternative semantic views of the same query, while the training set size remains N rather than increasing data volume.

2.4 Global Structure Disambiguation

With supervision signals imposed from local relations including both intra-query variant consistency and inter-query semantic neighborhood relations, the intent encoder f_θ learns a representation space where most queries exhibit stable clustering results. However, it is still possible that a few queries have unstable clustering results and tend to lie near cluster boundaries. To explicitly address such cases, we propose the *ambiguity detection and correction* (ADC) mechanism to identify and correct such *ambiguous queries*. The ADC module is activated starting from the t -th training epoch and is subsequently re-executed every t epochs thereafter.

Detection Let $\mathbf{z}_i = f_\theta(x_i)$ denote the current representation of query x_i , $\mathbf{z}_i^{v_1}$ and $\mathbf{z}_i^{v_2}$ denote the representations of its two variants, respectively. We perform K-Means clustering on the $\{\mathbf{z}_i\}_{i=1}^N$, obtaining cluster centers $\{\mathbf{c}_k\}_{k=1}^K$ and the corresponding cluster assignments $\{\hat{k}_i\}$ for each query. For each variant representation $\mathbf{z}_i^{v_j}$, we identify its nearest cluster center by

$$\hat{k}_i^{v_j} = \arg \min_k \|\mathbf{z}_i^{v_j} - \mathbf{c}_k\|_2, \quad j \in \{1, 2\}.$$

If there exists a variant $x_i^{v_j}$ with $\hat{k}_i^{v_j} \neq \hat{k}_i$, we regard x_i as an *ambiguous query*, as its clustering result is not consistent with that of its variants. All such ambiguous queries are collected into the ambiguous set

$$\mathcal{D}_{ac} = \{x_a \mid \hat{k}_a^{v_j} \neq \hat{k}_a, \exists j \in \{1, 2\}\}.$$

Correction For each ambiguous query $x_a \in \mathcal{D}_{ac}$, we compute the cosine similarity between $\mathbf{z}_a$ and each cluster center $\mathbf{c}_k$ as

$$s_{ak} = \text{sim}(\mathbf{z}_a, \mathbf{c}_k).$$

Let $\mathcal{C}_{\text{cand}}$ denote the set of the clusters whose centers have the top- K_{cand} highest s_{ak} scores. For each candidate cluster $c_k \in \mathcal{C}_{\text{cand}}$, we use the retrieval model T to retrieve a small number of representative queries from the training set $\mathcal{D}$ that are closest to the cluster center of $\mathbf{c}_k$, denoted by $\{q_m^{(k)}\}$. Note that each $q_m^{(k)}$ has been assigned to this cluster c_k . We construct a prompt using the ambiguous query x_a , the candidate cluster set $\mathcal{C}_{\text{cand}}$ and their representative queries $\{q_m^{(k)}\}$. The prompt details are provided in Appendix F.3. Given the prompt, the LLM evaluates the semantic consistency between x_a and the candidate clusters and outputs the best-matching cluster k^* .

Then, we further use the retrieval model T to identify the calibrated positive query by selecting from $\{q_m^{(k^*)}\}$ (i.e., the representative queries associated with the selected cluster k^*), the one most similar to x_a (i.e., with the highest $\text{sim}(T(x_a), T(q_m^{(k^*)}))$). Let p_a denote such a retrieved query as the positive for x_a . We maintain a calibration pool $\mathcal{P}_{ac} = \{(x_a, p_a)\}$ that stores pairs of ambiguous queries and their corresponding positives. Note that $\mathcal{P}_{ac}$ is subject to being updated at every t -th epoch.

Ambiguity Correction Loss. For each pair $(x_a, p_a) \in \mathcal{P}_{ac}$, we compute $\mathbf{z}_a = f_\theta(x_a)$ and $\mathbf{z}_{p_a} = f_\theta(p_a)$, and $s_a = \text{sim}(\mathbf{z}_a, \mathbf{z}_{p_a})$. In addition, we compute the similarities between x_a and the positives associated with other ambiguous queries, denoted as $s_j = \text{sim}(\mathbf{z}_a, \mathbf{z}_{p_j})$. Given N_{amb} ambiguous queries, the *ambiguity correction loss* is defined as:

$$\mathcal{L}_{ac} = -\frac{1}{N_{\text{amb}}} \sum_{a=1}^{N_{\text{amb}}} \log \frac{\exp(s_a/\tau)}{\sum_{j=1}^{N_{\text{amb}}} \exp(s_j/\tau)}.$$

2.5 Overall Training Objective

Our final training objective consists of three core losses: the inter-query neighbor alignment loss $\mathcal{L}_{na}$, the intra-query variant alignment loss $\mathcal{L}_{va}$, and the ambiguity correction loss $\mathcal{L}_{ac}$.

In a fully unsupervised setting, the training objective is:

$$\mathcal{L}_{total} = \mathcal{L}_{na} + \mathcal{L}_{va} + \mathcal{L}_{ac}.$$

In semi-supervised settings, where a small set of labeled queries with known intent labels is available, we further incorporate a standard cross-entropy loss to supervise the classification of known intents. Specifically, for labeled queries $(x_i, y_i) \in \mathcal{D}^l$ with $y_i \in \mathcal{I}_k$, the *cross-entropy loss* is defined as:

$$\mathcal{L}_{ce} = -\frac{1}{|\mathcal{D}^l|} \sum_{(x_i, y_i) \in \mathcal{D}^l} \log p_\theta(y_i | x_i),$$

where $p_\theta(y_i | x_i)$ denotes the predicted probability of the known intent label y_i .

To examine the effectiveness of each component, we set all loss weights to 1.0 in experiments without additional tuning.

3 Experiment Settings

Datasets: We conduct experiments on three widely used datasets: BANKING (Casanueva et al., 2020), CLINC150 (Larson et al., 2019), and StackOverflow20 (Xu et al., 2015). They cover dialogue system scenarios, such as banking services and voice assistants. Following prior work (Tang et al., 2024; Liang et al., 2024a,b), we adopt the NID task setting based on the Known Intent Ratio (KIR), which controls the proportion of intent classes treated as known during training. We evaluate our method under both unsupervised (KIR = 0%) and semi-supervised settings, with a focus on challenging low-KIR regimes (KIR = 25%).

Competitors: The first group of competitors contains BERT, BERT-CE, and Retriever. BERT denotes directly applying the pretrained encoder BERT (Devlin et al., 2019) without any additional training. BERT-CE denotes fine-tuning BERT on the labeled known-intent data using the $\mathcal{L}_{ce}$. Retriever denotes a semantic retriever based on the Sentence-Transformers model all-mpnet-base-v2 (Reimers and Gurevych, 2019; Song et al., 2020), whose representations are

directly used for clustering. The second group contains state-of-the-art NID methods, including DeepCluster (Caron et al., 2018), DeepAligned (Zhang et al., 2021), MTP-CLNN (Zhang et al., 2022), IDAS (De Raedt et al., 2023), IntentGPT (Rodriguez et al., 2024), LLM Clust (Viswanathan et al., 2024), CsePL (Liang and Liao, 2023), LANID (Fan et al., 2024), USNID (Zhang et al., 2024), INTENDD (Singhal et al., 2023), GeoID (Tang et al., 2024), SynCID (Liang et al., 2024a), ALUP (Liang et al., 2024b) and SPILL (Lin et al., 2025).

Metrics: Following (Tang et al., 2024; Liang et al., 2024a,b), we adopt the following metrics: Accuracy (ACC), Adjusted Rand Index (ARI), and Normalized Mutual Information (NMI).

In our framework, all components that involve large language models consistently use Qwen3-32B (8-bit quantized). Appendix B provides all the details.

4 Experiment Results

4.1 Main Results

We evaluate our method on three widely used intent discovery benchmarks. Table 1 presents our model consistently outperforms all competitive baselines in all settings. Results for all competing methods are obtained from their reported results.

Unsupervised Setting. In the most challenging case without any known intent labels, our framework consistently achieves state-of-the-art performance across all datasets, where higher ACC indicates that the majority of queries are correctly assigned to their intent categories, and improvements in ARI and NMI reflect more compact intra-cluster organization and clearer inter-cluster boundaries in the learned representation space. Compared with the Retriever baseline, which provides the semantic retrieval signals used in our framework, LR2GS demonstrates that training f_θ to integrate such signals together with other local relational cues can reorganize the intent representation space and ultimately surpass the Retriever itself.

Semi-Supervised Setting. With a small proportion of labeled intents, our method achieves further improvements and continues to outperform all competitors, indicating that limited known-intent supervision can effectively complement local relation signals in shaping the representation space.

Methods	BANKING			CLINC			StackOverflow		
	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI
Unsupervised Setting									
BERT	16.79	5.77	36.75	27.11	10.33	58.24	15.20	2.02	12.36
Retriever	63.02	51.79	79.95	72.76	64.14	88.88	73.70	59.00	72.87
DeepCluster (ECCV 2018)	20.69	8.95	41.77	35.70	19.11	65.58	-	-	-
MTP-CLNN (ACL 2022)	65.90	55.75	81.80	-	-	-	81.43	67.63	78.71
IDAS (NLP4ConvAI 2023)	67.43	57.56	82.84	85.48	79.02	93.82	83.82	72.20	81.26
INTENDD (EMNLP 2023)	58.74	47.88	75.91	63.87	51.76	83.12	79.32	62.49	73.88
IntentGPT (ICLR 2024 Workshop)	64.22	55.18	81.42	83.20	78.69	<u>94.35</u>	-	-	-
LLM Clust (TACL 2024)	66.40	-	84.10	81.30	-	<u>93.10</u>	-	-	-
LANID (LREC-COLING 2024)	70.58	60.40	84.12	-	-	-	86.60	72.96	81.25
USNID (IEEE TKDE 2024)	54.83	43.33	75.30	75.87	68.54	91.00	69.28	52.25	72.00
SynCID (ACL 2024)	<u>72.89</u>	<u>62.42</u>	84.20	<u>86.80</u>	<u>80.85</u>	94.23	<u>86.90</u>	<u>74.42</u>	<u>81.95</u>
SPILL (ACL 2025)	71.48	-	<u>85.12</u>	84.43	-	93.63	-	-	-
Ours	76.69 (+3.80)	67.48 (+5.06)	86.84 (+1.72)	89.07 (+2.27)	84.35 (+3.50)	95.38 (+1.03)	87.30 (+0.40)	75.51 (+1.09)	82.08 (+0.13)
Semi-Supervised Setting									
BERT-CE	33.51	23.35	62.45	-	-	-	29.70	14.22	31.17
DeepAligned (AAAI 2021)	49.08	37.62	70.50	74.07	64.63	88.97	54.50	37.96	50.86
MTP-CLNN (ACL 2022)	71.43	61.29	84.11	-	-	-	83.77	70.17	79.68
CsePL (EMNLP 2023)	71.06	60.36	83.32	86.16	79.65	94.07	79.47	64.92	74.88
INTENDD (EMNLP 2023)	67.50	57.85	76.79	83.11	76.31	92.32	84.82	71.64	78.93
LANID (LREC-COLING 2024)	71.40	64.23	<u>85.51</u>	-	-	-	81.23	63.23	79.55
USNID (IEEE TKDE 2024)	65.85	56.53	81.94	83.12	77.95	94.17	75.76	65.45	74.91
GeoID (ACL 2024)	68.75	57.92	82.54	88.13	81.24	94.37	81.80	66.35	77.62
SynCID (ACL 2024)	<u>75.41</u>	<u>65.40</u>	85.39	87.85	82.39	<u>94.85</u>	<u>87.86</u>	<u>76.11</u>	<u>82.46</u>
ALUP (NAACL 2024)	74.61	62.64	84.06	<u>88.40</u>	<u>82.44</u>	94.84	82.20	64.54	76.58
Ours	78.64 (+3.23)	68.53 (+3.13)	86.83 (+1.32)	92.36 (+3.96)	89.23 (+6.79)	96.95 (+2.10)	88.70 (+0.84)	77.80 (+1.69)	83.52 (+1.06)

Table 1: Comparison of our method and representative baselines across three benchmark datasets under the unsupervised setting (upper part) and the semi-supervised setting with KIR (Known Intent Ratio) = 25% (lower part). The best results are highlighted in **bold**, the second-best results are underlined, and the relative improvements of **Ours** over the second-best scores are shown in parentheses. Results are reported as the average over three runs.

Variant	BANKING			CLINC			StackOverflow		
	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI
BERT	16.79	5.77	36.75	27.11	10.33	58.24	15.20	2.02	12.36
further + $\mathcal{L}_{va}$	64.22	52.15	78.89	72.53	63.63	88.14	81.90	65.19	75.86
further + $\mathcal{L}_{na}$	70.32	61.21	84.45	85.87	80.58	94.53	82.63	69.90	80.89
further + $\mathcal{L}_{ac}$	76.69	67.48	86.84	89.07	84.35	95.38	87.30	75.51	82.08

Table 2: Incremental effects of progressively introducing components under the unsupervised setting.

Variant	BANKING			CLINC			StackOverflow		
	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI
Ours (full)	78.64	68.53	86.83	92.36	89.23	96.95	88.70	77.80	83.52
w/o $\mathcal{L}_{na}$	72.37	63.45	84.69	90.04	85.43	95.63	85.90	73.34	80.44
w/o $\mathcal{L}_{va}$	76.17	66.80	87.06	88.53	85.91	96.23	87.90	76.69	82.76
w/o $\mathcal{L}_{ac}$	72.18	61.57	84.30	88.98	85.20	95.92	85.90	73.16	80.44
w/o $\mathcal{L}_{ce}$	75.13	65.73	86.10	91.60	88.69	96.83	87.80	76.22	82.41

Table 3: Semi-supervised ablation study.

4.2 Ablation Study

We analyze the contribution of different components from two complementary perspectives. First, we examine the incremental effects of progressively adding training components. Table 2 presents introducing the intra-query variant alignment loss $\mathcal{L}_{va}$ leads to a substantial improvement over the baseline, indicating that enforcing intra-instance consistency already provides strong intent-level supervision. Further incorporating the inter-query neighbor alignment loss $\mathcal{L}_{na}$ yields additional gains,

suggesting that inter-query neighbor relations offer complementary structural guidance. Finally, adding the ambiguity correction loss $\mathcal{L}_{ac}$ consistently brings further improvements, highlighting the importance of refining local relations near intent boundaries. These trends are also clearly reflected in the t-SNE visualizations (Figure 2), where clusters become progressively more compact and better separated as each component is introduced.

Second, we study the effect of removing individual components. Table 3 presents removing any of the core losses generally leads to performance

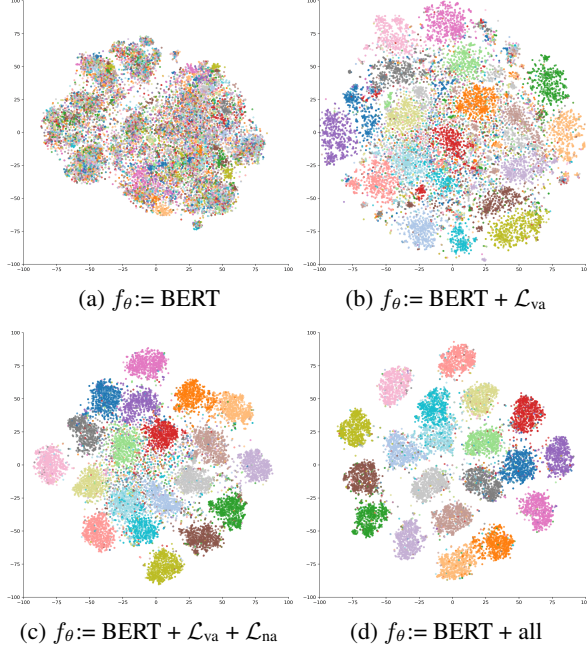


Figure 2: t-SNE visualizations on StackOverflow on $\mathcal{D}$.

degradation, demonstrating that each component plays a complementary role. Notably, even after removing a single module, most ablation variants still outperform the strongest previous competitors, indicating that the overall framework remains robust and effective.

In Section 4.5, we analyze whether ADC (from Section 2.4) can adaptively increase the number of ambiguous queries to compensate for performance degradation when key modules are ablated.

4.3 Local Neighbor Quality Analysis

Method	BANKING		CLINC		StackOverflow	
	RCA	ACC	RCA	ACC	RCA	ACC
BERT	58.84	48.54	78.43	82.29	37.58	40.40
Retriever	91.57	71.69	94.59	87.64	81.93	78.40
LNR	92.55	72.18	96.43	89.47	87.81	85.90
Full	92.02	78.64	97.42	92.36	88.44	88.70

Table 4: Effect of different retrieval strategies on clustering performance under the semi-supervised setting. The best results are highlighted in **bold**.

To examine the role of local neighbor relation supervision quality in shaping the intent representation space, we conduct comparative experiments on three benchmark datasets, as shown in Table 4. We remove the ambiguity correction loss $\mathcal{L}_{ac}$ to avoid its interference with the analysis.

We first compare multiple neighbor construction

approaches and evaluate the quality of the induced local relations using a unified metric termed *Relation Consistency Accuracy* (RCA). We use the term “relation” to denote any pair of queries that is treated as positive under a specific construction rule. Specifically, given a relation set $\mathcal{R} = \{(x_i, x_j)\}$ constructed under a certain relation definition rule, we measure how often the paired queries share the same intent label. Formally, the RCA over a relation set $\mathcal{R}$ is defined as

$$\text{RCA}(\mathcal{R}) = \frac{1}{|\mathcal{R}|} \sum_{(x_i, x_j) \in \mathcal{R}} \mathbf{1}[y_i = y_j],$$

where y_i and y_j denote the ground-truth intent labels of queries x_i and x_j , respectively, and $\mathbf{1}[\cdot]$ is the indicator function. This metric directly reflects the purity of the constructed local relations and serves as an important indicator of the supervision quality used for local relation.

Table 4 presents the quality of neighbor relations $\mathcal{P}_{\text{neighbor}}$ constructed using different strategies and the corresponding downstream clustering performance. Constructing neighbors based on raw BERT representations yields limited consistency, whereas using a semantic retrieval model substantially improves RCA and the resulting clustering accuracy. Further refining retrieved neighbors with LNR leads to consistent gains, demonstrating strong complementarity between retrieval models and LLMs in semantic understanding. For these baselines, ACC reflects the clustering performance obtained by training the intent encoder with the corresponding $\mathcal{P}_{\text{neighbor}}$. Overall, the results reveal a clear correlation between RCA and clustering accuracy: higher-quality neighbor relations consistently lead to better intent clustering.

Full reports the performance of the intent encoder trained with the complete LR2GS framework. Notably, the trained intent encoder further improves retrieval quality, achieving RCA comparable to LNR, indicating that relational knowledge provided by retrieval models and LLMs is effectively internalized into the representation space.

4.4 Variant Alignment Analysis

Table 5 presents the effect of variants, where the model is trained for three epochs using only Variant Alignment to isolate its contribution. Using either the descriptive variant or the prototype variant alone already yields reasonable clustering performance, indicating that different types of variants

Variant Configuration	BANKING			CLINC			StackOverflow		
	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI
Base Encoder	16.79	5.77	36.75	27.11	10.33	58.24	15.20	2.02	12.36
only Descriptive	58.47	46.41	76.40	56.84	43.61	79.28	79.70	66.08	75.69
only Prototype	60.49	49.23	77.86	70.80	61.20	87.46	74.10	58.60	70.90
both variants	64.22	52.15	78.89	72.53	63.63	88.14	81.90	65.19	75.86

Table 5: Variant quality analysis under the unsupervised setting, comparing different LLM-generated variant configurations in contrastive learning.

can provide effective semantic supervision. The descriptive variant preserves richer semantic details and supplements intent characterization from the perspective of concrete description, while the prototype variant captures intent semantics at a more abstract level. When both variants are introduced simultaneously, the model is able to establish a more stable correspondence between abstract intent prototypes and concrete semantic descriptions, leading to further performance gains. We provide concrete examples in Appendix C.2 and further analyze the semantic differences captured by different variants.

4.5 Global Structure Disambiguation Analysis

Epoch t	#Ambiguous	$RCA(\mathcal{R}_{center}^{(t)})$	$RCA(\mathcal{R}_{pos}^{(t)})$	$\overline{ACC}$
none	none	none	none	85.08
5	1722	68.14	74.49	89.40
10	1350	68.16	76.61	89.80
15	1197	63.02	72.62	90.76
20	1117	61.84	73.45	91.67

Table 6: Iteration-wise analysis on **CLINC** under the semi-supervised setting, where ambiguous queries are identified by the **ADC**.

ADC is performed only after epochs $t \in \{5, 10, 15, 20\}$. Let $\mathcal{D}_{ac}^{(t)}$ denote the ambiguous set identified by the ADC detection process conducted after epoch t . For each ambiguous query $x_i \in \mathcal{D}_{ac}^{(t)}$, we consider two local relations. Specifically, $c(i)$ denotes the representative center query of the cluster assigned to x_i before correction, while $p(i)$ denotes the calibrated positive selected by the dynamic ambiguity correction module. Accordingly, we define $\mathcal{R}_{center}^{(t)} = \{(x_i, x_{c(i)})\}$ and $\mathcal{R}_{pos}^{(t)} = \{(x_i, x_{p(i)})\}$ for $x_i \in \mathcal{D}_{ac}^{(t)}$, and measure their quality by $RCA(\mathcal{R}_{center}^{(t)})$ and $RCA(\mathcal{R}_{pos}^{(t)})$.

Table 6 presents although the **CLINC** training set contains 18,000 queries, only a relatively small subset is identified as ambiguous at each ambiguity detection round, indicating that the proposed mechanism effectively limits correction to a small

number of ambiguous queries and substantially reduces the overall cost. Here, $\overline{ACC}$ denotes the average clustering accuracy over the five training epochs following each ambiguity detection round. Comparing the stage without ambiguity detection to the first detection round, we observe a significant improvement in $\overline{ACC}$, demonstrating the effectiveness of introducing ADC. At each round, $RCA(\mathcal{R}_{pos}^{(t)})$ consistently exceeds $RCA(\mathcal{R}_{center}^{(t)})$, highlighting the effectiveness of large-small model collaboration in correcting unreliable local relations. Moreover, most of the performance gain is obtained at the first detection round, where the model already surpasses the previous strongest baseline SynCID (Liang et al., 2024a), suggesting a favorable trade-off between performance improvement and correction cost. As training proceeds, $\overline{ACC}$ continues to increase while the number of ambiguous queries gradually decreases, indicating that the model becomes increasingly stable and that ambiguity correction progressively reduces structural uncertainty in the representation space.

Setting (Round 1)	#Queries	$RCA(\mathcal{R}_{center}^{(t)})$	$RCA(\mathcal{R}_{pos}^{(t)})$	ACC	ARI	NMI
none	0	\	\	87.29	82.62	95.30
ADC-ambiguous	1722	68.14	74.49	89.91	86.18	96.07
Random-ambiguous	1722	88.81	90.63	87.91	82.19	94.55
All-ambiguous	18000	91.08	90.30	89.02	85.64	95.93

Table 7: Analysis of ADC on **CLINC** under the semi-supervised setting, comparing ADC-ambiguous with Random-ambiguous (randomly selected ambiguous queries of the same size) and All-ambiguous (using all queries).

Setting (Round 1)	#Ambiguous	$RCA(\mathcal{R}_{center})$	$RCA(\mathcal{R}_{pos})$	ACC	ARI	NMI
Full	1722	68.14	74.49	89.91	86.18	96.07
w/o $\mathcal{L}_{na}$	1939	58.60	77.23	87.60	82.23	94.64
w/o $\mathcal{L}_{va}$	11771	64.07	82.98	83.38	79.62	94.29

Table 8: Robustness analysis on **CLINC** under the semi-supervised setting.

The experimental results further validate the effectiveness of the proposed ADC mechanism in identifying and leveraging ambiguous queries. As shown in Table 7, the ambiguous queries identified by ADC exhibit significantly lower RCA than those selected by *Random-ambiguous* and *All-ambiguous* settings, indicating that ADC can more precisely focus on genuinely difficult queries with higher intrinsic ambiguity and more blurred semantic boundaries. Nevertheless, ADC-ambiguous achieves the best downstream clustering performance, demonstrating that ADC selects key informative queries that are most valuable for model learning. Under

a lower computational budget, correcting and optimizing only these critical ambiguous queries can more effectively refine local relations near decision boundaries, thereby improving overall clustering performance.

As shown in Table 8, we simulate more challenging training environments by removing modules that significantly contribute to performance. The multi-variant alignment module provides a stable and reliable basis for ADC, as it enforces consistency among representations of the same query across different semantic variants and effectively compresses ambiguous regions in the representation space. At the same time, removing either $\mathcal{L}_{va}$ or $\mathcal{L}_{na}$ leads to a noticeable increase in the number of ambiguous queries, further demonstrating that the ADC mechanism can adaptively identify varying degrees of ambiguity as the training environment changes.

4.6 LLM Variants

Model	ACC	ARI	NMI	Cost (\$)	Total Cost (\$)
Qwen3-14B	76.66	68.35	87.08	0.14 / 0.56	≈ 0.57
Qwen3-32B	76.40	67.40	86.60	0.28 / 1.12	≈ 1.14
Qwen2.5-72B	76.27	68.28	87.46	0.56 / 1.68	≈ 2.27
DeepSeek-V3	78.21	68.30	87.39	0.28 / 1.12	≈ 1.14

Table 9: Evaluation of open-source LLM variants under the unsupervised setting on BANKING. Cost reports the inference price per 1M tokens in USD, formatted as input/output. Total Cost denotes the total cost incurred during a single training run of our method.

We evaluate the proposed LR2GS framework using open-source LLMs with different parameter scales on BANKING. All LLMs are accessed via APIs provided by the Alibaba Bailian platform¹, and the corresponding pricing follows the official specifications released by the platform.

As shown in Table 9, our approach consistently achieves competitive performance across LLMs of varying sizes, including Qwen2.5 (Yang et al., 2024), Qwen3 (Yang et al., 2025), and DeepSeek-V3 (DeepSeek-AI et al., 2025), indicating that LR2GS does not rely on extremely large LLMs’ capacities to function effectively. Notably, Qwen3-14B delivers performance comparable to that of Qwen3-32B, suggesting that increasing model size does not necessarily translate into substantial gains in our framework. This can be attributed to the fact that LR2GS emphasizes fine-grained local relation

modeling, which is generally easier for LLMs to reason about and does not impose stringent requirements on model scale.

By accurately capturing local relations at the sample level and aggregating them progressively, LR2GS constructs an effective global structure that leads to robust overall performance. Furthermore, as discussed in Section 4.5, the proposed ADC mechanism significantly reduces token consumption in the most expensive module by selectively invoking LLM-based correction only for ambiguous samples. As a result, the total cost remains at a relatively low level while preserving strong performance.

5 Conclusion

We propose **LR2GS**, a new framework for New Intent Discovery that directly trains the intent encoder by leveraging multiple forms of local relational supervision. By jointly modeling semantic neighbor relations, intra-instance variant consistency, and ambiguity-aware correction, LR2GS enables the representation space to evolve toward a well-structured global intent clustering.

Experimental results across multiple benchmarks show that LR2GS consistently outperforms existing methods in both unsupervised and semi-supervised settings, demonstrating its effectiveness and robustness. These results indicate that explicitly incorporating local relational constraints provides a practical and reliable way to improve intent discovery under limited supervision.

Since the true number of intents is typically unknown in real-world scenarios, we additionally study the setting where K is predicted and analyze its impact on clustering performance, as reported in Appendix C.4.

Limitations

Although we have entered the era of LLMs, leveraging LLMs to provide semantic understanding and discrimination for the NID task has become an effective trend, LLMs still exhibit certain limitations. Their semantic understanding and fine-grained discriminative capabilities have not yet reached a fully reliable level in complex scenarios or in cases with ambiguous intent boundaries.

As shown in the annotation experiments in Section 4.5, the ambiguity correction mechanism based on LLMs can significantly improve the consistency of local relations (RCA) and further enhance over-

¹<https://bailian.console.aliyun.com>

all clustering performance, yet there remains considerable room for improvement in these metrics. This observation indicates that current LLMs have not fully covered the fine-grained semantic distinctions required for NID tasks, especially when dealing with highly similar intents or blurred semantic boundaries.

Nevertheless, this limitation does not diminish the value of LLMs within the NID framework. On the contrary, it provides important insights into the future development of LLMs in terms of semantic discrimination ability, task alignment capability, and collaborative learning with smaller models.

References

- Qian Cao, Deyi Xiong, Qinlong Wang, and Xia Peng. 2022. [CoCoID: Learning contrastive representations and compact clusters for semi-supervised intent discovery](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 226–236, Abu Dhabi, UAE. Association for Computational Linguistics.
- Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. 2018. [Deep clustering for unsupervised learning of visual features](#). In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part XIV*, page 139–156, Berlin, Heidelberg. Springer-Verlag.
- Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. [Efficient intent detection with dual sentence encoders](#). In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.
- Peng Chen and Bang Wang. 2025. [On collaborating small and large models for few-shot intent detection](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13937–13953, Suzhou, China. Association for Computational Linguistics.
- Maarten De Raedt, Frédéric Godin, Thomas Demeester, and Chris Develder. 2023. [IDAS: Intent discovery with abstractive summarization](#). In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 71–88, Toronto, Canada. Association for Computational Linguistics.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. [Deepseek-v3 technical report](#). Preprint, arXiv:2412.19437.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#). Preprint, arXiv:1810.04805.
- Lu Fan, Jiashu Pu, Rongsheng Zhang, and Xiao-Ming Wu. 2024. [LANID: LLM-assisted new intent discovery](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10110–10116, Torino, Italia. ELRA and ICCL.
- Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of classification*, 2(1):193–218.
- Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. [An evaluation dataset for intent classification and out-of-scope prediction](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316, Hong Kong, China. Association for Computational Linguistics.
- Jinggui Liang and Lizi Liao. 2023. [ClusterPrompt: Cluster semantic enhanced prompt learning for new intent discovery](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10468–10481, Singapore. Association for Computational Linguistics.
- Jinggui Liang, Lizi Liao, Hao Fei, and Jing Jiang. 2024a. [Synergizing large language models and pre-trained smaller models for conversational intent discovery](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14133–14147, Bangkok, Thailand. Association for Computational Linguistics.
- Jinggui Liang, Lizi Liao, Hao Fei, Bobo Li, and Jing Jiang. 2024b. [Actively learn from LLMs with uncertainty propagation for generalized category discovery](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7845–7858, Mexico City, Mexico. Association for Computational Linguistics.
- I-Fan Lin, Faegheh Hasibi, and Suzan Verberne. 2025. [SPILL: Domain-adaptive intent clustering based on selection and pooling with large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15723–15737, Vienna, Austria. Association for Computational Linguistics.
- Yutao Mou, Keqing He, Pei Wang, Yanan Wu, Jingang Wang, Wei Wu, and Weiran Xu. 2022. [Watch the neighbors: A unified k-nearest neighbor contrastive learning framework for OOD intent discovery](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1517–1529, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Juan A. Rodriguez, Nicholas Botzer, David Vazquez, Christopher Pal, Marco Pedersoli, and Issam Laradji. 2024. [Intentgpt: Few-shot intent discovery with large language models](#). *Preprint*, arXiv:2411.10670.
- Bhavuk Singhal, Ashim Gupta, V P Shivasankaran, and Amrith Krishna. 2023. [IntenDD: A unified contrastive learning approach for intent detection and discovery](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14204–14216, Singapore. Association for Computational Linguistics.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. [Mpnnet: Masked and permuted pre-training for language understanding](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc.
- Kai Tang, Junbo Zhao, Xiao Ding, Runze Wu, Lei Feng, Gang Chen, and Haobo Wang. 2024. [Learning geometry-aware representations for new intent discovery](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5641–5654, Bangkok, Thailand. Association for Computational Linguistics.
- Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2010. [Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance](#). *Journal of Machine Learning Research*, 11(95):2837–2854.
- Vijay Viswanathan, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. 2024. [Large language models enable few-shot clustering](#). *Transactions of the Association for Computational Linguistics*, 12:321–333.
- Hongtao Wang, Renchi Yang, and Wenqing Lin. 2025. [Nilc: Discovering new intents with llm-assisted clustering](#). *Preprint*, arXiv:2511.05913.
- Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. [Unsupervised deep embedding for clustering analysis](#). *Preprint*, arXiv:1511.06335.
- Jiaming Xu, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fanyuan Wang, and Hongwei Hao. 2015. [Short text clustering via convolutional neural networks](#). In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 62–69, Denver, Colorado. Association for Computational Linguistics.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.
- Hanlei Zhang, Hua Xu, Ting-En Lin, and Rui Lyu. 2021. [Discovering new intents with deep aligned clustering](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14365–14373.
- Hanlei Zhang, Hua Xu, Xin Wang, Fei Long, and Kai Gao. 2024. [A clustering framework for unsupervised and semi-supervised new intent discovery](#). *IEEE Transactions on Knowledge and Data Engineering*, 36(11):5468–5481.
- Haode Zhang, Haowen Liang, Liming Zhan, Albert Y.S. Lam, and Xiao-Ming Wu. 2023. [Revisit few-shot intent classification with PLMs: Direct fine-tuning vs. continual pre-training](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11105–11121, Toronto, Canada. Association for Computational Linguistics.
- Yuwei Zhang, Haode Zhang, Li-Ming Zhan, Xiao-Ming Wu, and Albert Lam. 2022. [New intent discovery with pre-training and contrastive learning](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 256–269, Dublin, Ireland. Association for Computational Linguistics.

A Related Work

A.1 New Intent Discovery

Intent Detection has traditionally been formulated as a closed-set supervised classification task, relying on a fixed set of intents and labeled data, and has achieved significant success (Singhal et al., 2023; Zhang et al., 2023; Chen and Wang, 2025). However, in real-world applications, systems often can only access a small number of labeled queries for known intents, while facing a large volume of unlabeled user queries that may even contain unknown intents. This practical demand has drawn increasing attention to the problem of New Intent Discovery (NID), whose goal is to discover potential new intent categories from unlabeled data while retaining the capability of recognizing known intents.

Early representative work such as Deep Aligned Clustering (Zhang et al., 2021) combines deep representation learning with the clustering process and

alleviates pseudo-label inconsistency during iterative clustering through an alignment mechanism, achieving promising results on the NID task. Subsequent studies further focus on learning semantic representation spaces that are friendly to intent clustering. MTP-CLNN (Zhang et al., 2022) jointly optimizes representation learning and clustering objectives through multi-task pre-training and contrastive learning, achieving significant improvements under both semi-supervised and unsupervised settings. GeoID (Tang et al., 2024) further investigates the geometric structure of the representation space and explicitly enhances the separability among different intent centers, alleviating the issue of new intents being entangled in the embedding space.

There are also methods that first leverage labeled data of known intents to optimize the encoder, and then utilize semantic distance relationships derived from the encoder to assist the clustering process. For example, Watch the Neighbors (Mou et al., 2022) constructs neighborhood relations based on the current encoder representations and exploits neighborhood information for contrastive learning, while CoCoID (Cao et al., 2022) introduces intra-cluster predictive distribution distillation within a contrastive learning framework to enhance cluster compactness. However, the relational signals in these methods are mainly induced by the model itself through similarity measures or clustering results, and their effectiveness heavily depends on the quality of the current representation space. When the model has not yet formed a stable intent structure, such self-induced relational constraints are prone to introducing noise, which can be further amplified during training and thus limit the model’s ability to continuously refine the intent structure.

In the absence of any labeled data, new intent discovery degenerates into a purely unsupervised clustering problem, whose difficulty is substantially increased. In this scenario, models cannot leverage any explicit intent labels and must rely solely on self-supervised objectives or global structural assumptions to learn separable intent representations. INTENDD (Singhal et al., 2023) constructs contrastive learning signals under the unsupervised setting to enhance representation discriminability and improves clustering performance through post-processing strategies. USNID (Zhang et al., 2024) proposes a unified clustering framework that mitigates cluster allocation inconsistency across different settings via multi-stage self-supervised and

semi-supervised representation learning together with centroid-guided clustering. GeoID (Tang et al., 2024) can also operate under the unsupervised setting, enhancing intent separability through global geometric constraints.

Although these methods improve the performance of new intent discovery to a certain extent, they mainly rely on global objectives or overall structural assumptions, and the modeling of query-level discriminative signals remains relatively indirect. In the absence of supervision, especially under unsupervised settings, how to obtain more fine-grained and reliable constraint signals to continuously guide the representation space toward a clear and stable intent structure remains a core challenge in new intent discovery.

A.2 LLM-assisted New Intent Discovery

With the rapid advancement of large language models (LLMs) in semantic understanding and reasoning, an increasing number of studies have explored leveraging LLMs to provide auxiliary supervision for new intent discovery. IDAS (De Raedt et al., 2023) prompts LLMs to generate abstractive summary-style labels for utterances and performs clustering over the summarized representations, demonstrating that semantic compression can improve the consistency of unsupervised intent clustering. SynCID (Liang et al., 2024a) proposes to synergize LLMs with smaller models by enriching the semantic representations of utterances and intent labels via LLMs and aligning the enhanced descriptions back into the representation space of the smaller model to correct clustering distortion. LANID (Fan et al., 2024) utilizes LLMs to judge semantic relations between pairs of samples and constructs training signals to fine-tune a smaller encoder, thereby improving performance under both unsupervised and semi-supervised settings. NILC (Wang et al., 2025) further explores deeply embedding LLMs into the clustering process to assist sample refinement and intent discovery.

Although these approaches validate the advantages of LLMs in semantic discrimination and relation judgment, most existing works treat LLMs as stage-wise or external auxiliary tools. The judgment signals provided by LLMs have not yet been systematically modeled as structural constraints that persist throughout the training process. How to continuously incorporate high-confidence and precisely judged semantic relations into representation learning under constrained computational

budgets, so as to progressively shape a clear and stable intent structure, remains an open research problem.

B Detailed Experiment Settings

B.1 Competitors Details

BERT: This baseline uses BERT (Devlin et al., 2019) as the encoder without any additional training, and directly applies its representations for downstream clustering.

BERT-CE: This model fine-tunes BERT on the labeled known-intent data using the cross-entropy loss $\mathcal{L}_{ce}$, serving as a supervised baseline.

Retriever: This baseline employs a semantic retriever based on the Sentence-Transformers model all-mpnet-base-v2 (Reimers and Gurevych, 2019; Song et al., 2020), whose sentence representations are directly used for clustering without further training.

DeepCluster (Caron et al., 2018): A general deep clustering method that iteratively assigns pseudo-labels via K-Means and uses them as supervision to train the feature extractor; it is adopted here as an intent discovery baseline.

DeepAligned (Zhang et al., 2021): A new intent discovery method that performs iterative deep clustering and introduces an alignment strategy to resolve the inconsistency of cluster assignments across training iterations.

MTP-CLNN (Zhang et al., 2022): A framework that first conducts multi-task pre-training on external labeled and in-domain unlabeled data, and then performs contrastive learning with nearest neighbors to learn clustering-friendly intent representations.

IDAS (De Raedt et al., 2023): An intent discovery approach that prompts a large language model to generate abstractive summary labels for utterances and then clusters over the encoded summaries.

IntentGPT (Rodriguez et al., 2024): A large language model-based method that performs few-shot intent discovery via prompt-driven intent induction.

LLM Clust (Viswanathan et al., 2024): A few-shot clustering framework that incorporates large language models at multiple stages, i.e., enriching input features, imposing pairwise constraints during clustering, and post-correcting cluster assignments.

CsePL (Liang and Liao, 2023): A prompt-based new intent discovery method that learns cluster-semantic-enhanced soft prompts through two-level contrastive learning with label-semantic alignment.

LANID (Fan et al., 2024): An LLM-assisted new intent discovery framework that leverages a large language model to judge the semantic relations of utterance pairs selected by KNN and DBSCAN, and fine-tunes a smaller encoder with a triplet loss.

USNID (Zhang et al., 2024): A unified clustering framework for both unsupervised and semi-supervised new intent discovery that adopts centroid-guided clustering together with multi-stage self-supervised and semi-supervised representation learning.

INTENDD (Singhal et al., 2023): A graph-based intent discovery method that combines unsupervised contrastive learning with post-processing techniques such as residual propagation and label smoothing.

GeoID (Tang et al., 2024): A geometry-aware intent discovery approach that learns discriminative representations by modeling the geometric structure of intent clusters.

SynCID (Liang et al., 2024a): A synergistic intent discovery framework that combines large language models and small language models to enhance cluster quality.

ALUP (Liang et al., 2024b): An active learning-based new intent discovery method that leverages large language model feedback to improve clustering efficiency.

SPILL (Lin et al., 2025): A training-free, domain-adaptive intent clustering method that casts clustering as a selection problem, using embeddings to select candidate neighbors for each seed utterance, an LLM to retain those sharing the same intent, and pooling to derive a refined representation.

B.2 Performance Metrics

Clustering Accuracy (ACC) (Xie et al., 2016): This metric evaluates clustering performance by measuring the best one-to-one alignment between predicted clusters and ground-truth intent labels, reflecting the overall correctness of cluster assignments.

Adjusted Rand Index (ARI) (Hubert and Arabie, 1985): This metric measures the similarity between the predicted clustering and the ground-truth partition, adjusted for chance agreement, and assesses the consistency of pairwise query assignments.

Normalized Mutual Information (NMI) (Vinh et al., 2010): This metric quantifies the mutual dependence between predicted clusters and ground-truth labels, normalized to account for cluster size variations, and reflects the overall quality of clus-

tering structure.

B.3 Experiment Details

Dataset	# Intent	# Train	Domain
CLINC	150	18000	Multi-domain
BANKING	77	9003	Banking
StackOverflow	20	18000	Programming

Table 10: Statistics of datasets used in NID.

Table 10 presents the datasets’ statistics. We follow the experimental protocol of [Liang et al. \(2024b\)](#) and use the original dataset splits without re-partitioning. The main intent encoder f_θ is based on **BERT** ([Devlin et al., 2019](#)), and we keep the encoder configuration consistent with prior baselines. All models are trained for 25 epochs with a batch size of 50.

In our framework, all components that involve large language models consistently use **Qwen3-32B (8-bit quantized)**, which is deployed on a single NVIDIA A800 GPU. In the *LLM Variants* experiments, we use open-source large language model APIs provided by the Alibaba Bailian platform.

For constructing local semantic neighbor relations, we adopt **all-mpnet-base-v2** ([Reimers and Gurevych, 2019](#); [Song et al., 2020](#)) as the semantic retrieval model T . On all benchmark datasets, we retrieve $\text{Top-}K_{\text{nn}} = 10$ nearest candidates for each query. The ambiguity detection and correction mechanism is activated and updated at a fixed interval of $t = 5$ epochs; for each identified ambiguous query, we select $K_{\text{cand}} = 10$ candidate clusters and sample $M = 5$ representative queries closest to the corresponding cluster centers.

We do not perform hyperparameter tuning for $\{\text{Top-}K_{\text{nn}}, K_{\text{cand}}, M, t\}$, and instead fix these values across all datasets to validate the proposed methodology. These hyperparameters are not tightly coupled to specific datasets and can be dynamically adjusted according to computational budget and data scale in practice. In addition, we provide a hyperparameter study and analysis of the number of prompt candidates in Appendix D.

All intent encoder training experiments are conducted on two NVIDIA RTX 5080 GPUs.

C Supplementary Experiments

C.1 Other t-SNE Visualizations

[Click here to return to Figure 2.](#)

These trends are also consistently reflected in the additional t-SNE visualizations (Figure 3 and Figure 4). Regardless of whether the scenario involves a smaller number of intents or a larger and more diverse intent space, clusters become progressively more compact and better separated as each component is introduced, demonstrating the robustness of our framework across different intent cardinalities.

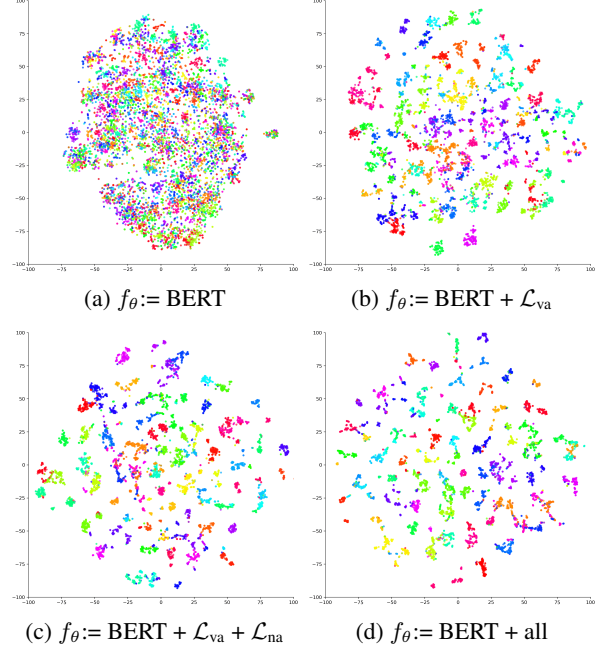


Figure 3: t-SNE visualizations on BANKING on $\mathcal{D}$.

C.2 Multi-Variant Cases Analysis

Table 11 presents several concrete examples from the BANKING dataset, which are used to compare the prototype variants (Prototype Variant) and the descriptive variants (Descriptive Variant) corresponding to the same raw query. Prototype variants usually summarize the core user intent in a highly compressed and label-like manner, emphasizing abstraction at the intent level; in contrast, descriptive variants adopt natural language sentences to provide a more complete explanation of user needs with richer pragmatic information. These two types of variants exhibit clear differences in semantic abstraction levels and expression styles, while remaining consistent at the intent level, thereby providing complementary supervision signals for multi-variant contrastive learning.

C.3 Evaluation under Different KIR Settings

To examine the effect of varying supervision levels, we evaluate our method under multiple Known Intent Ratio (KIR) settings: 0%, 25%, 50%, 75%, and

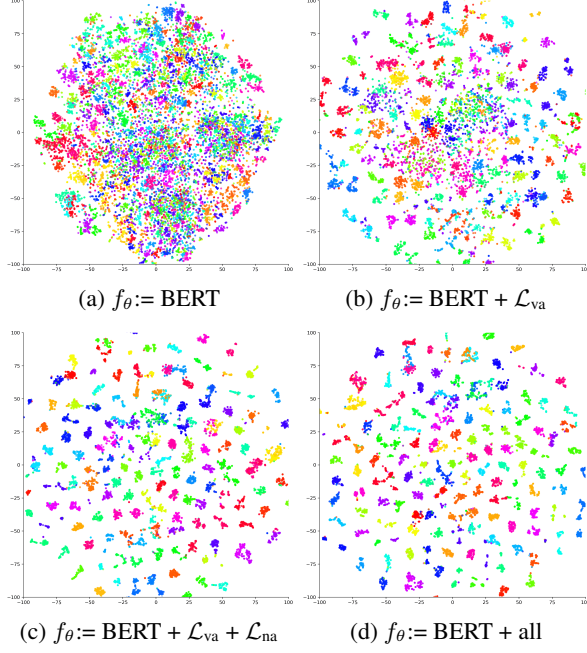


Figure 4: t-SNE visualizations on CLINC on $\mathcal{D}$.

100%. Figure 5 plots curves for ACC, ARI, and NMI to illustrate how model performance changes as the proportion of labeled known-intent data increases.

As KIR increases, the clustering performance consistently improves. Notably, LR2GS already achieves performance comparable to higher-KIR settings under the unsupervised and low-supervision regime (KIR = 25%), demonstrating that exploiting local relations among all samples can provide strong and effective supervision signals.

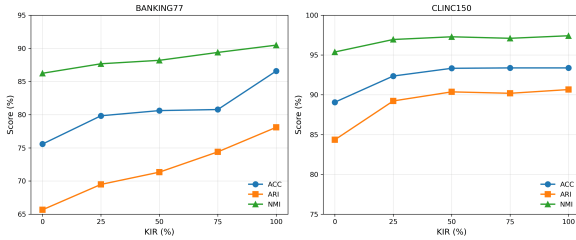


Figure 5: Evaluation under different Known Intent Ratio (KIR) settings. Curves show ACC/ARI/NMI as KIR increases.

C.4 Predicting the Total Number of Intents K

In practical intent discovery scenarios, the true number of intent categories is typically unknown. Following prior studies (Zhang et al., 2021; Liang et al., 2024b), we estimate the number of intent categories under a predefined overestimated cluster

Example 1

Raw Query: Could you help me figure out the exchange fee?

Intent Label: exchange_charge

Prototype Variant: Exchange fee inquiry

Descriptive Variant: The user wants to know the fee for currency exchange.

Example 2

Raw Query: Hello - I'm on the app and trying to purchase crypto. It's not going through. What am I doing wrong?

Intent Label: beneficiary_not_allowed

Prototype Variant: Purchase crypto issue

Descriptive Variant: The user is experiencing issues purchasing cryptocurrency through the banking app.

Example 3

Raw Query: Why is it saying I have a pending payment?

Intent Label: pending_card_payment

Prototype Variant: Pending payment inquiry

Descriptive Variant: The user is inquiring about the reason for a pending payment on their account.

Example 4

Raw Query: Do top-up limits exist?

Intent Label: top_up_limits

Prototype Variant: Top-up limits inquiry

Descriptive Variant: The user is asking about the existence of limits on top-up transactions.

Table 11: Concrete cases comparing prototype variants and descriptive variants for the same raw query.

number.

Specifically, we initialize the clustering process with an intentionally large number of clusters, denoted as K' . A trained intent representation encoder is then used to encode queries in the test set, producing semantic representations on which K-Means clustering is performed to obtain K' initial clusters. Based on the resulting clusters, we filter clusters according to their sizes and regard clusters with sufficient query support as valid intent categories. Formally, the final number of intents K is estimated as

$$K = \sum_{i=1}^{K'} \mathbf{1}[|S_i| \geq t], \quad (1)$$

where $|S_i|$ denotes the number of queries assigned to the i -th cluster, t is a predefined threshold, and $\mathbf{1}[\cdot]$ is the indicator function.

To evaluate the effect of estimating the number of intent categories, we conduct experiments under both the ground-truth K and the predicted $\hat{K}$. We compare our approach with representative semi-supervised intent discovery methods, including CsePL (Liang and Liao, 2023) and ALUP (Liang et al., 2024b), under identical experimental settings.

Table 12 presents the clustering performance of different methods under both the ground-truth number of categories K and the predicted $\hat{K}$. It can be observed that, compared with existing strong competitors CsePL and ALUP, the proposed LR2GS framework achieves the best performance under both settings, demonstrating its superior capability in modeling the intent representation space for intent discovery.

More importantly, when the number of intent categories is unknown and needs to be predicted in a data-driven manner, although the estimated $\hat{K}$ deviates slightly from the true value, LR2GS is still able to maintain performance close to that obtained under the ground-truth K . This indicates that the intent representation space learned by LR2GS exhibits good structural properties and discriminability, enabling the model to form reasonable clustering results even when the category cardinality is uncertain.

Cluster Num K	Methods	ACC	ARI	NMI
77 (gold)	CsePL	71.06	60.36	83.32
	ALUP	74.61	62.64	84.06
	Ours	78.64	68.53	86.83
74 (predict)	CsePL	69.75	56.70	81.30
74 (predict)	ALUP	72.55	61.04	82.78
75 (predict)	Ours	77.53	67.00	86.39

Table 12: Effect of estimating cluster number K on the BANKING dataset under the semi-supervised setting.

D Impact of the Number of Candidates in Prompt

We conduct a hyperparameter study on the $Top-K_{nn}$ parameter used in the LLM-based neighbor refinement stage of the first module in LR2GS. This parameter controls the number of candidate semantic neighbors retrieved for each query before being refined by the LLM, and thus directly affects the quality of constructed neighbor relations. We evaluate the impact of different $Top-K_{nn}$ values on the **CLINC** and **BANKING** datasets, while keeping all other settings fixed. Following Section 2.2, we adopt *Relation Consistency Accuracy* (RCA) (see Section 4.3) as the evaluation metric, which measures the proportion of constructed neighbor pairs that share the same ground-truth intent label.

Figure 6 shows that incorporating LNR effectively improves RCA, thereby enhancing the quality of constructed neighbor relations. Moreover, the

performance remains stable across different $Top-K_{nn}$ settings. The results demonstrate that RCA is not sensitive to the choice of $Top-K_{nn}$, as a moderate number of candidates is sufficient to provide reliable semantic context for LLM-based neighbor refinement. We consistently set $K_{nn} = 10$ for all datasets in the main experiments, following a unified design choice of LR2GS.

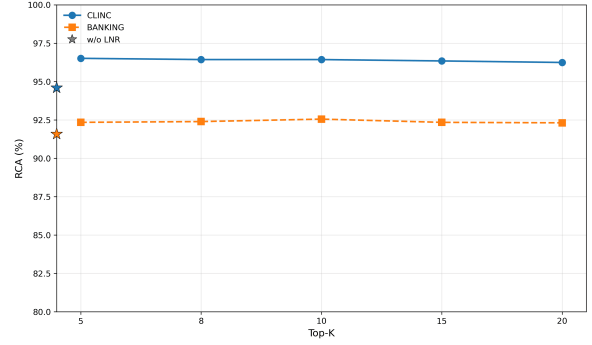


Figure 6: Impact of the Number of $Top-K_{nn}$ in LLM-based neighbor refinement (LNR).

E Use of AI Assistants

AI assistants were used to support parts of the code implementation and to improve the clarity and fluency of the manuscript writing. All generated code, experimental settings, and textual revisions were carefully reviewed and verified by the authors. The authors take full responsibility for the correctness of the implementation, experimental results, and the content of this paper.

F Prompt Design

F.1 Prompt for LLM-based Neighbor Refinement

Example prompt on the CLINC dataset
Click here to return to Section 2.2.

Select the candidate query that is most semantically similar to the target query.

Target Query: "will i need a visa to go to moldova"

Candidate Queries:

1. "do i need a visa to russia"
2. "do i need to get a visa for russia"
3. "do i need a travel visa to visit russia"
- ...
9. "must i have an international visa when travelling to russia"
10. "will i need an international visa to go to madagascar"

Answer with only the number (e.g., 3):

F.2 Prompt for Variants

Descriptive Variant

Example prompt on the CLINC dataset
Click here to return to Section 2.3.

You are an expert in semantic analysis, specializing in understanding conversational user intents.

Your task is to read the 'Raw Query' and generate a short, generic description of the user's underlying goal.

CRITICAL RULES:

1. The output **MUST** be a complete sentence describing the user's goal.
2. Be as concise as possible (maximum 10 words).
3. Use clear, natural language appropriate for conversational assistance.
4. Respond **ONLY** with the description sentence and nothing else.

Raw Query:

"find my current location"

Description:

Prototype Variant

Example prompt on the CLINC dataset
Click here to return to Section 2.3.

You are an expert at identifying the core user intent from conversational queries. Your task is to read the 'Raw Query' and distill it into a concise intent phrase.

CRITICAL RULES:

1. Be as concise as possible (typically 2-4 words).
2. Focus on the action or inquiry the user wants to perform.
3. Respond **ONLY** with the intent phrase and nothing else.

Raw Query:

"find my current location"

Intent:

F.3 Correction

Example prompts on the CLINC dataset
Click here to return to Section 2.4.

I will provide you with a 'Target Query' and a list of 'Candidate Queries'.

Your task is to identify which candidate query is **MOST LIKELY** to share the same conversational service intent as the target query.

Target Query: "i'm looking to call a group"

Cluster Examples:

Cluster 127 Examples:

1. "would you be able to help me with a meeting room on wednesday at 3pm"
2. "can you hook me up with a meeting room on saturday morning at 10am"
- ...

Cluster 128 Examples:

1. "would you be able to help me with a meeting room on wednesday at 3pm"
- ...

Cluster 84 Examples:

1. "can you call robby"
2. "i need to call mark"
3. "call my mother"
4. "call mom"
5. "can you please call nata"
- ...

Answer (cluster number only):