Dial-In LLM: Human-Aligned LLM-in-the-loop Intent Clustering for Customer Service Dialogues

Mengze Hong[1]*, **Wailing Ng**[1], **Chen Jason Zhang**[1], **Yuanfeng Song**[2], **Di Jiang**[1]†
[1]Hong Kong Polytechnic University, [2]AI Group, WeBank Co., Ltd

Abstract

Discovering customer intentions is crucial for automated service agents, yet existing intent clustering methods often fall short due to their reliance on embedding distance metrics and neglect of underlying semantic structures. To address these limitations, we propose an **LLM-in-the-loop (LLM-ITL)** intent clustering framework, integrating the language understanding capabilities of LLMs into conventional clustering algorithms. Specifically, this paper (1) examines the effectiveness of fine-tuned LLMs in semantic coherence evaluation and intent cluster naming, achieving over 95% accuracy aligned with human judgments; (2) designs an LLM-ITL framework that facilitates the iterative discovery of coherent intent clusters and the optimal number of clusters; and (3) introduces context-aware techniques tailored for customer service dialogue. Since existing English benchmarks lack sufficient semantic diversity and intent coverage, we further present a comprehensive Chinese dialogue intent dataset comprising over 100k real customer service calls with 1,507 human-annotated clusters. The proposed approaches significantly outperform LLM-guided baselines, achieving notable improvements in clustering quality, cost efficiency, and downstream applications. Combined with several best practices, our findings highlight the prominence of LLM-in-the-loop techniques for scalable dialogue data mining.

1 Introduction

Intent discovery is a crucial task in NLP applications, such as dialogue system design (Hengst et al., 2024), information retrieval (Jiang et al., 2016a), and utterance pattern analysis (Ghosal et al., 2020). While intent clustering techniques are extensively studied to automatically identify thematic relationships within text corpora (Gung et al., 2023), existing research primarily focuses on developing meaningful sentence representations and relies on embedding distance metrics for optimization (Yin et al., 2021; Gao et al., 2025). This approach often overlooks the distinctive characteristics of textual information, such as linguistic patterns and semantic diversity (see Figure 1), and thus restricts human-aligned evaluation and validation of clustering performance (Vinh et al., 2009). This issue is particularly pronounced in languages with rich semantics, such as Chinese, where two seemingly similar sentences can convey entirely different meanings (Chen, 1993; Jiang et al., 2016b).

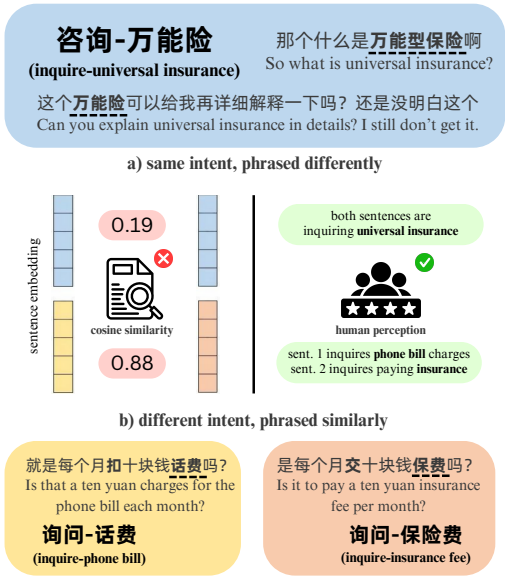

Figure 1: Comparison of embedding-distance metric and human perception: cosine similarity failed to identify the same intentions under diverse expression (top), and to distinguish distinct intentions under similar expressions (bottom).

With growing research interest in integrating large language models (LLMs) into the problem-solving pipeline (Hong et al., 2025b), LLM-guided clustering techniques have emerged, demonstrating superior performance over traditional machine learning algorithms (Zhang et al., 2023). While

*Work was partially done during internship at WeBank.
†Corresponding Author

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 5885–5900
November 4-9, 2025

these methods effectively incorporate language understanding into the clustering process, they primarily focus on data preprocessing, querying LLMs for embedding refinement or data augmentation (Viswanathan et al., 2024). This approach represents a surface-level integration, potentially missing the benefits of LLMs to contribute semantic-driven guidance within the clustering process.

In this paper, we introduce an **LLM-in-the-loop (LLM-ITL)** intent clustering framework, designed to facilitate the iterative discovery of coherent intent clusters from semantically diverse, large-scale dialogue datasets. Our approach effectively leverages intermediate clustering results by incorporating human-aligned LLM utilities, enabling computationally efficient and human-interpretable intent clustering. The key contributions of this work are:

1. We present the largest Chinese dialogue intent clustering dataset, derived from over 100,000 real-world customer service calls across the banking, telecommunication, and insurance domains. The data is annotated into 1,507 intent clusters with high semantic diversity and a substantial inclusion of noisy, out-of-domain queries, reflecting realistic and complex customer interactions.

2. We demonstrate the effectiveness of fine-tuned small LLMs in assessing the semantic coherence of intent clusters and providing accurate intent labels across various sampling strategies, offering cost-efficient utilities for designing LLM-in-the-loop solutions.

3. We propose an LLM-in-the-loop intent clustering framework that effectively combines the strengths of LLMs and conventional clustering algorithms. This approach outperforms state-of-the-art baselines and excels in downstream applications with 18.46% performance gain. Furthermore, discussions on data sampling and LLM-based crowdsourcing validate best practices for real-world deployment.

2 Related Work

LLM-guided Text Clustering. The integration of LLMs into text clustering has become increasingly prominent since 2023. Zhang et al. (2023) introduced ClusterLLM, an innovative approach that leverages instruction-tuned LLMs like ChatGPT to refine sentence embedding spaces through pairwise preference questions, aligning clustering granularity with user preferences. Viswanathan et al. (2024) highlighted the enhancement of clustering quality by LLMs through feature improvement, the imposition of constraints during clustering, and post-correction processes. Additionally, Feng et al. (2024) proposed refining edge points with LLMs, which led to notable performance gains.

Recently, Hong et al. (2025b) introduced the concept of LLM-in-the-loop machine learning, categorizing integration strategies into data-, model-, and task-level approaches. This paradigm parallels the human-in-the-loop framework (Chen et al., 2024), enabling LLMs to replicate human expertise and thereby enhance conventional problem-solving workflows in a cost-efficient and flexible manner. Within this taxonomy, existing work on LLM-guided clustering has primarily focused on data-centric aspects, with relatively limited exploration of modeling and task-solving aspects. This gap constrains a broader understanding of the potential of LLMs in tackling long-standing challenges in text clustering, such as improving cluster interpretability and moderating the clustering process (Tan et al., 1999; Jiang et al., 2021).

Intent Clustering. Intent clustering extends conventional text clustering by incorporating contextual cues and domain knowledge to uncover meaningful user intentions (Allahyari et al., 2017; Hong et al., 2024). It serves as a fundamental step in intent induction (Chandrakala et al., 2024) and in preparing data for training intent classifiers (Gung et al., 2023). While conceptually related to topic modeling (Jiang et al., 2015, 2023), which seeks to uncover latent semantic themes, intent clustering tackles the more challenging task of differentiating texts that may appear lexically similar but convey contextually distinct meanings, thereby uncovering the underlying communicative goals behind sentences (Carberry and Flowers, 1988; Qu et al., 2018). This finer granularity makes intent clustering more demanding as a data mining task and a critical component in applications such as search engines (Jiang et al., 2013, 2016a) and dialogue systems (Qin et al., 2023; Hong et al., 2025c).

While semi-supervised approaches (Kumar et al., 2022) and deep learning methods (Lin et al., 2020; Zhang et al., 2021) have substantially advanced intent clustering, recent attempts in LLM-based systems highlight their practical advantages (Liang et al., 2024). For instance, the IDAS method lever-

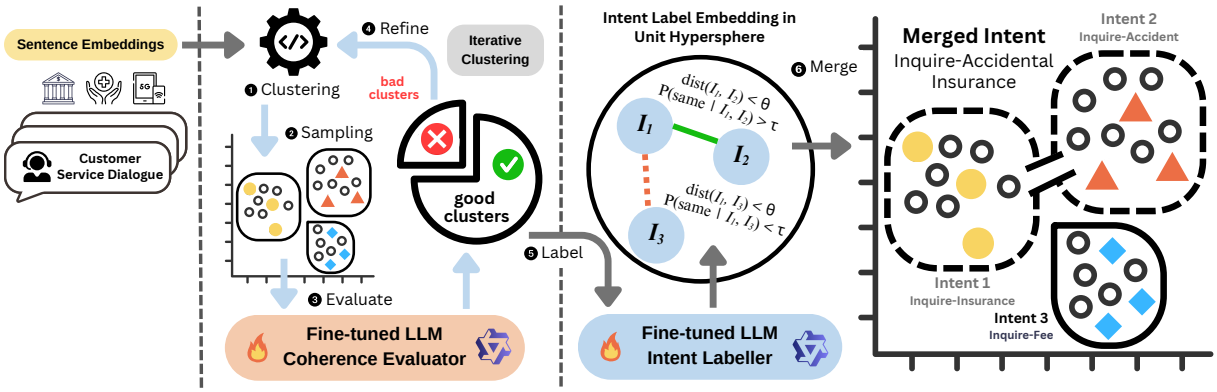

Figure 2: Overview of the proposed LLM-in-the-loop framework for dialogue intent clustering.

ages in-context learning to generate descriptive utterance labels and improve sentence embeddings (De Raedt et al., 2023). However, challenges such as high computational cost and limited model robustness remain unresolved (Song et al., 2023), and existing methods are evaluated based on small-scale datasets with limited complexity (Casanueva et al., 2020), leading to an inadequate understanding of their effectiveness in real-world applications.

Highlights. Previous studies have primarily focused on enhancing input text representations and the direct use of LLMs for cluster modification, imposing high computational cost and uncertainty. In contrast, this paper implements human-aligned LLM utilities to create an LLM-in-the-loop framework with emphasis on the intermediate clustering results. Additionally, this work releases a complex dataset from real customer service call transcriptions, enabling more practical insights and meaningful evaluations for future intent clustering.

3 Proposed Methods

In this section, we first outline the design of human-aligned LLM utilities. Then, we introduce a comprehensive LLM-in-the-loop intent clustering framework (see Figure 2) and show how each LLM utility contributes to the clustering process.

3.1 LLM Utilities

Coherence Evaluation. Inspired by human behaviors in perceiving texts and making comparisons based on semantic meaning rather than surface-level work similarity (Peter W. Foltz and Landauer, 1998), this paper proposes **semantic coherence** as a more effective metric and optimization objective, measuring the semantic consistency within a cluster. The coherence evaluation is formulated as a binary classification problem, enabling intuitive interpretation and simplifying the training of the LLM evaluator (see Section 5.6 for alternative formulations). **Good** clusters consist of sentences focused on a specific topic, whereas **Bad** clusters contain inconsistent or ambiguous intentions (see Table 11 for examples). This task is particularly challenging for traditional machine learning classifiers due to their lack of semantic understanding (Mimno et al., 2011), thus necessitating fine-tuned LLMs for robust evaluation (Gu et al., 2025).

Cluster Naming. Giving each cluster a concise and meaningful name is essential for many downstream applications (Pattnaik et al., 2024; Luo et al., 2024). In this paper, we introduce a novel naming convention, "Action-Objective," which is particularly effective for capturing dialogue intents that are typically topic-oriented (e.g., insurance, loan) and involve distinct actions (e.g., inquire, confirm). Examples of human annotations are shown in Table 12, and a comparison of different naming conventions is presented in Table 13 to demonstrate the effectiveness of the proposed approach.

3.2 LLM-in-the-loop Iterative Intent Clustering with Coherence Evaluation

At iteration t, the current set of unassigned sentences is denoted as $\mathcal{S}^{(t)}$. A special case in the first iteration, where $\mathcal{S}^{(0)} = \mathcal{S}$ represents the entire set of unique sentences derived from the dialogue corpus. For each candidate cluster number $n_i \in N$, we compute:

$$\mathcal{C}_{n_i}^{(t)} = F(\mathcal{E}^{(t)}, n_i),$$

where F is a clustering function (e.g., K-means clustering), and $\mathcal{E}^{(t)} = \{\mathbf{e}_s \mid s \in \mathcal{S}^{(t)}\}$ represents the sentence embeddings. This results in $|N|$ distinct cluster assignments as the initial outcome.

Then, the semantic coherence of each cluster is evaluated using a fine-tuned LLM $\mathcal{M}_{\text{eval}}$:

$$\mathbf{g}_{n_i}^{(t)} = \left[\mathcal{M}_{\text{eval}}(C_1^{(t)}), \ldots, \mathcal{M}_{\text{eval}}(C_{n_i}^{(t)})\right],$$

where $\mathcal{M}_{\text{eval}}(C_k) = 1$ if the cluster is coherent (i.e., "good" cluster), else 0.

While the number of clusters is often known in benchmark evaluations, determining the optimal number in a noisy text corpus is both challenging and essential. Here, we propose a local search heuristic that maximizes the "good/bad" ratio at each iteration. The optimal n_t^* is given by:

$$n_t^* = \arg\max_{n_i \in N} \frac{\sum_{j=1}^{n_i} \mathbb{I}\left(\mathbf{g}_{n_i}^{(t)}[j] = 1\right)}{\sum_{j=1}^{n_i} \mathbb{I}\left(\mathbf{g}_{n_i}^{(t)}[j] = 0\right) + 1},$$

representing the best cluster number at iteration t. This approach enables the automatic discovery of suitable cluster numbers in a step-by-step manner. The search space $\boldsymbol{N}$ should be selected carefully to balance accuracy and efficiency, and a search space pruning strategy is proposed in Section 5.6 to enhance the searching process.

Finally, the "good" clusters in the optimal $\mathcal{C}_{n^*}^{(t)}$ are retained, and the remaining sentences will be refined in the next (i.e., $t+1$) iteration, enabling the iterative discovery of high-quality clusters. The proposed method is summarized in **Algorithm 1**, and the LLM integration is highlighted.

3.3 Post-Correction with LLM-Generated Intent Labels

Preliminary results in Table 14 suggest that, in later iterations, the diminishing size of the unassigned sentence set $\mathcal{S}^{(t)}$ may lead to the formation of multiple clusters capturing similar intents, resulting in smaller and less representative clusters. The embedding distances between sentences within clusters limit the natural consolidation of clusters with similar intents but different expressions (Khan et al., 2020), necessitating a post-correction step to merge semantically aligned clusters. Previous methods typically address this by issuing direct LLM queries to validate individual cluster assignments (Viswanathan et al., 2024; Feng et al., 2024), an approach that is both computationally expensive and prone to inconsistency. In contrast, we propose a **context-aware approach**, leveraging LLMs' naming utility to robustly merge clusters based on their generated intent labels.

Algorithm 1 LLM-in-the-loop Intent Clustering

Input: Unlabeled sentence corpus $\mathcal{S}$, embedding function f_{emb}, coherence evaluator $\mathcal{M}_{\text{eval}}$, candidate cluster numbers $N = \{n_1, \ldots, n_k\}$, threshold $\epsilon > 0$, max iterations $T_{\max}$
Output: Set of clusters $\mathcal{C}$

$\mathcal{E} \leftarrow \{f_{\text{emb}}(s) \mid s \in \mathcal{S}\}$
$\mathcal{S}^{(0)} \leftarrow \mathcal{S}, \mathcal{C} \leftarrow \emptyset, t \leftarrow 0$
while $\frac{|\mathcal{S}^{(t)}|}{|\mathcal{S}|} > \epsilon$ and $t < T_{\max}$ **do**
  $\mathcal{E}^{(t)} \leftarrow \{f_{\text{emb}}(s) \mid s \in \mathcal{S}^{(t)}\}$
  for each $n_i \in N$ **do**
    $\mathcal{C}_{n_i}^{(t)} \leftarrow F(\mathcal{E}^{(t)}, n_i)$
    $\mathbf{g}_{n_i}^{(t)} \leftarrow [\mathcal{M}_{\text{eval}}(C_1^{(t)}), \ldots, \mathcal{M}_{\text{eval}}(C_{n_i}^{(t)})]$
  end for
  $n^* \leftarrow \arg\max_{n_i \in N} \frac{\sum_{j=1}^{n_i} \mathbb{I}\left(\mathbf{g}_{n_i}^{(t)}[j]=1\right)}{\sum_{j=1}^{n_i} \mathbb{I}\left(\mathbf{g}_{n_i}^{(t)}[j]=0\right)+1}$
  $\mathcal{C}_{\text{good}}^{(t)} \leftarrow \{C_j \in \mathcal{C}_{n^*}^{(t)} \mid \mathbf{g}_{n^*}^{(t)}[j] = 1\}$
  $\mathcal{C} \leftarrow \mathcal{C} \cup \mathcal{C}_{\text{good}}^{(t)}$
  $\mathcal{S}^{(t+1)} \leftarrow \mathcal{S}^{(t)} \setminus \bigcup_{C \in \mathcal{C}_{\text{good}}^{(t)}} C$
  $t \leftarrow t + 1$
end while
return $(\mathcal{C})$

At the end of the iterative clustering process, each cluster C_k receives an intent label from the fine-tuned LLM naming utility:

$$l_k = \mathcal{M}_{\text{name}}(C_k).$$

These labels concisely summarize the semantic information of each cluster and are mapped to the semantic space of sentence embeddings using the embedding function f_{emb}:

$$\mathbf{l}_k = f_{\text{emb}}(l_k) \in \mathbb{R}^d, \quad \|\mathbf{l}_k\|_2 = 1.$$

This normalization positions label embeddings $\mathbf{l}_k$ on the unit hypersphere $\mathbb{S}^{d-1} = \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_2 = 1\}$, enabling accurate measurement of semantic relationships along the sphere's surface rather than through straight-line distances (Fletcher et al., 2004), thus facilitating precise intent similarity comparisons in high-dimensional space.

Then, the clusters are structured into a semantic affinity graph $G = (V, E)$ to model relationships based on label similarity. Vertices $V = \{C_1, C_2, \ldots, C_K\}$ represent clusters, and the edges $E \subseteq V \times V$ are determined using hyperspherical geometry by computing the geodesic distance, which captures angular separation between label embeddings (Fletcher et al., 2004):

$$\text{Dist}(\mathbf{l}_i, \mathbf{l}_j) = \arccos(\langle \mathbf{l}_i, \mathbf{l}_j \rangle),$$

forming edges if $\text{Dist}(\mathbf{l}_i, \mathbf{l}_j) < \theta$, with θ is a predefined threshold ($\theta = 0.8$).

For each derived edge, a probabilistic criterion is designed to enhance the robustness of merging decisions, naturally modeling label embeddings as samples from a mixture of von Mises-Fisher distributions (Banerjee et al., 2005):

$$p(\mathbf{l}_k \mid \boldsymbol{\mu}_m, \kappa_m) = Z_d(\kappa_m) \exp(\kappa_m \langle \mathbf{l}_k, \boldsymbol{\mu}_m \rangle),$$

where $\boldsymbol{\mu}_m = \mathbf{l}_m \in \mathbb{S}^{d-1}$ is the mean direction of the m-th intent embedding, $\kappa_m > 0$ controls the distribution's tightness, and the normalization constant $Z_d(\kappa_m)$ ensures the density integrates to 1 over the hypersphere:

$$Z_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)}.$$

Thus, the probability that clusters C_i and C_j share the same intent is computed for each edge as:

$$P(\text{same} \mid \mathbf{l}_i, \mathbf{l}_j) = \sum_{m=1}^{K} \pi_m p(\mathbf{l}_i \mid \boldsymbol{\mu}_m, \kappa_m) p(\mathbf{l}_j \mid \boldsymbol{\mu}_m, \kappa_m),$$

with $\pi_m = 1/K$ as uniform mixture weights. An edge is retained only if this probability exceeds a predefined threshold τ (e.g., $\tau = 0.7$), minimizing inappropriate merges by accounting for uncertainty.

Finally, clusters linked in the affinity graph are consolidated into connected components, forming a refined cluster assignment $\mathcal{C}' = \{C'_1, C'_2, \ldots, C'_{K'}\}$ that eliminates redundancy, enhances interpretability, and maintains robust semantic alignment. Each merged cluster C'_k receives a new intent label reflecting its semantic content.

3.4 Context-Aware Role Separation with LLM-Generated Intent Labels

From a practical perspective, dialogues are often accompanied by domain-specific features or use-case scenarios (e.g., customer service calls, group discussions). This motivates the incorporation of contextual information into the clustering process (Ding et al., 2025). In particular, the customer service dialogues typically involve only the customer and the service agent (Lin et al., 2022). The classification task that assigns a sentence to its associated role is relatively simple with labeled training data. To provide an unsupervised solution, we propose that the sentence roles can be naturally determined based on the LLM-generated intent label.

By heuristics, sentences within clusters labeled as "inquire-" or similar actions are mostly sent from the customer, and vice versa. A two-step approach is proposed: initially, intermediate clustering results are obtained using previous approach, denoted as $C_{\text{inter}} = \{C_1, C_2, \ldots, C_K\}$, and these results are divided based on intent labels l_k, forming two groups with distinct roles, R_{customer} and R_{agent}, where:

$$R_{\text{customer}} = \{s \in S \mid l_k(s) \in \{\text{"inquire-"}, \ldots\}\},$$
$$R_{\text{agent}} = \{s \in S \mid l_k(s) \in \{\text{"answer-"}, \ldots\}\}.$$

In the second step, the sentences corresponding to each role are clustered again, denoted as $C'_{\text{customer}} = \text{Cluster}(R_{\text{customer}})$ and $C'_{\text{agent}} = \text{Cluster}(R_{\text{agent}})$. Finally, the resulting customer and agent clusters are merged to produce a refined clustering that maintains a clear separation of intents by role, serving either as the final output or as an improved intermediate stage for further processing.

4 Experiment

4.1 Dataset

The proposed dataset contains 1,507 high-quality intent clusters manually annotated from over 100,000 realistic customer service calls, comprising 55,085 distinct sentences with an average length of 17 Chinese characters per sentence (see Appendix A for annotation details)[1]. Among these intents, 885 are identified as domain-specific (e.g., inquire-insurance), primarily concentrated within the banking, telecommunications, and insurance industries, with a focus on the Chinese context (Hong et al., 2025a). The remaining 622 clusters are categorized as out-of-domain (e.g., provide-location), representing general queries that commonly arise in customer service interactions.

Compared with existing intent clustering benchmarks such as BANK77 (Casanueva et al., 2020) and CLINC150 (Larson et al., 2019), the proposed dataset is the first Chinese benchmark for customer service intent clustering and the largest of its kind in both the number of sentences and the number of clusters (see Table 1). It presents several new challenges, including the lexical sparsity in short-text sentences, the dependence on contextual knowledge, and the difficulties in balancing accuracy with computational efficiency due to the excessively large intent size. The semantic diversity, calculated based on the average cosine distance between individual sentences and the cluster centroid (Casanueva et al., 2022), depicts the complexity of this dataset and motivates the development of semantic-guided approaches.

[1]Data is available at GitHub repository.

Dataset	Number of sentences	Number of intents	Semantic diversity
BANK77	3080	77	0.209
NLU++	3,080	62	0.367
CLINC(I)	4,500	150	0.275
MTOP(I)	4,386	102	0.234
MASSIVE(I)	2,974	59	0.351
ours	**55,085**	**1507**	**0.538**

Table 1: Comparison of the proposed customer service intent clustering dataset with existing benchmarks.

4.2 Metrics

Normalized Mutual Information (NMI) measures the degree of similarity between ground-truth and predicted clusters, ranging from 0 (no mutual information) to 1 (perfect correlation). However, recent work has demonstrated that NMI exhibits biased behavior, particularly in favor of larger numbers of clusters (Jerdee et al., 2024), highlighting the need for complementary evaluation metrics that better reflect human-perceived clustering quality.

Goodness score measures the proportion of good clusters evaluated by the fine-tuned LLM evaluator. For intermediate steps of iteration where the number of clusters is unknown, the good/bad ratio is used as an invariant measure of clustering quality:

$$\text{goodness}_i = \frac{\text{\# good clusters}}{\text{\# bad clusters}}$$

For the final clustering results, the percentage of good clusters among all clusters is reported. Note that the evaluation uses a different LLM evaluator than the one used for intermediate evaluation to ensure the fairness of the reported metric:

$$\text{goodness}_{final} = \frac{\text{\# good clusters}}{\textbf{\# total clusters}}$$

This metric offers a comprehensive understanding of the quality of the produced intent clusters, allowing for accurate cluster-level evaluation. Additionally, it can be easily deployed in applications without ground-truth annotations, thereby eliminating the need for human involvement.

4.3 Implementations

Four open-sourced Chinese LLMs[2] are fine-tuned using LoRA (Hu et al., 2022) on 4 × Nvidia A100 GPUs. For coherence evaluation, a human-annotated training dataset of 1,772 intent clusters labeled as "good" or "bad" is used. For cluster naming, a training dataset of 2,500 clusters, each containing 20 sentences, is annotated with the "Action-Objective" labels. Semantic-rich embeddings are generated using the BAAI General Embeddings (BGE) model[3] (Xiao et al., 2024). The best-performing clustering algorithm, hierarchical clustering, and LLM utilities, *qwen14b* and *chatglm3-6b*, are selected for LLM-in-the-loop clustering. Additionally, we employ *random* and *convex* sampling to select 20 representative sentences per cluster as LLM inputs, reducing computational overhead while preserving intent coverage.

LLM	qwen7b	qwen14b	baichuan2-7b	chatglm3-6b
Accuracy	96.25%	**97.50%**	89.17%	95.83%

Table 2: Performance of fine-tuned LLMs in evaluating semantic coherence of intent clusters.

LLM	qwen7b	qwen14b	baichuan2-7b	chatglm3-6b
Accuracy	92.8%	94.3%	94.3%	**94.4%**

Table 3: Performance of fine-tuned LLMs in naming intent clusters.

5 Results and Discussions

5.1 LLM Utilities

Table 2 presents the performance of the fine-tuned LLM coherence evaluator tested on 480 unseen clusters. The results indicate that mainstream open-source LLMs can effectively serve as robust evaluators for assessing the quality of intent clusters and providing human-aligned judgments. For cluster naming, since labels are not unique, accuracy is manually evaluated by four human experts based on alignment with the true labels in the dataset. The results in Table 3 demonstrate that the fine-tuned LLMs show promising performance in generating intuitive names and adhering strictly to the predefined "Action-Objective" format.

5.2 Main Results

Table 4 reports the main results on the proposed dataset, with all evaluation metrics averaged over five random seeds. Among the LLM-guided clustering baselines, the data-centric keyphrase expansion method, which refines sentence embeddings via LLM-summarized keyphrases, achieved the most

[2]The models are available at: Qwen2.5-7B; Qwen2.5-14B; Baichuan2-7B ; ChatGLM3-6b

[3]https://huggingface.co/BAAI/bge-large-zh-v1.5

Method	NMI (Mean ± Std)	NMI Gain	#Good (Mean ± Std)	#Good Gain
Baselines				
K-Means	0.7899 ± 0.0135	-	94.8% ± 1.3%	-
GMMs	0.7903 ± 0.0140	+0.05%	91.1% ± 1.6%	-3.90%
Hierarchical	0.8001 ± 0.0128	+1.29%	94.9% ± 1.2%	+0.11%
LLM-Guided Clustering Baselines				
ClusterLLM (Zhang et al., 2023)	0.7284 ± 0.0168	-7.79%	91.2% ± 1.8%	-3.80%
IDAS (De Raedt et al., 2023)	0.8109 ± 0.0105	+2.66% *	93.6% ± 1.2%	-1.27%
Keyphrase (Viswanathan et al., 2024)	0.8371 ± 0.0098	+5.97% ***	94.5% ± 1.0%	-0.32%
LLMEdgeRefine (Feng et al., 2024)	0.7411 ± 0.0155	-6.19%	87.2% ± 2.0%	-8.02%
Proposed Method (Context-Free)				
LLM-ITL (random)	0.8202 ± 0.0095	+3.84% **	97.7% ± 0.6%	+3.06% ***
LLM-ITL (convex)	0.8208 ± 0.0090	+3.92% **	**97.8% ± 0.5%**	+3.16% ***
LLM-ITL + keyphrase	0.8378 ± 0.0085	+6.06% ***	96.4% ± 0.6%	+1.69% *
Proposed Method (Context-Aware)				
LLM-ITL + merge	0.8420 ± 0.0178	+6.59% **	97.8% ± 1.3%	+3.16% ***
LLM-ITL + role	0.8679 ± 0.0068	+9.86% ***	97.2% ± 0.5%	+2.53% **
LLM-ITL + role + merge	**0.8826 ± 0.0060**	**+11.76% *** **	97.6% ± 0.4%	+2.95% ***

Table 4: Performance of baselines and proposed methods in dialogue intent clustering. Gains are computed relative to the K-Means baseline. The best results in each column are **bolded**.

significant performance gain (Viswanathan et al., 2024). However, deriving pairwise constraints for fine-tuning the embedding model, as in ClusterLLM, was ineffective due to the excessive number of clusters in the dataset. Interestingly, goodness evaluation does not always align with the NMI score. For example, while keyphrase expansion improved NMI, the generated keyphrases could distort the original sentence's meaning, leading to less representative clusters and a slightly lower goodness score compared to the base model. This highlights the need to balance ground truth alignment with semantic coherence during evaluation.

The proposed LLM-in-the-loop intent clustering demonstrated satisfactory improvements in NMI and significant enhancements in the goodness score. In a context-free setting, iterative clustering with convex sentence sampling outperformed both ClusterLLM and IDAS baselines. Effective keyphrase data augmentation further improved NMI, highlighting the potential of integrating data-centric methods into the pipeline. Furthermore, the context-aware approaches consistently improved performance by incorporating dialogue roles and cluster merging, resulting in a 4.48% increase in NMI without compromising the goodness score. This emphasizes the importance of task-centric design in leveraging intent labels effectively for the problem of intent clustering.

Notably, the proposed method does not require prior knowledge of the number of clusters and automatically performs parameter searches during clustering, unlike baselines that depend on the true cluster count. This adaptability highlights its strong application potential for extracting intent clusters from noisy text corpora (Akama et al., 2020).

Method	Bank77	CLINC(I)	MTOP(I)	Massive(I)
SCCL	81.77	92.94	73.52	73.90
Self-supervise	80.75	93.88	72.50	72.88
ClusterLLM	**85.07**	94.00	**73.83**	77.64
IDAS	82.84	92.35	72.31	75.74
LLMEdgeRefine	-	**94.86**	72.92	76.66
ours	82.32	94.12	72.45	**78.12**

Table 5: NMI (%) performance of different clustering methods on four English intent benchmarks.

5.3 Evaluation on English Benchmarks

To validate the effectiveness of the proposed LLM-in-the-loop technique, we evaluate the context-aware (best-performing) method on four widely used English benchmarks (FitzGerald et al., 2023). The Llama-7b model is fine-tuned with 800 intent clusters (200 samples from each dataset) annotated by human experts to derive Action-Objective intent labels and bad clusters through perturbation. As shown in Table 5, our method delivers performance comparable to state-of-the-art LLM-guided methods and outperforms on Massive(I) dataset. Furthermore, the computational cost is lower than the compared methods as measured in Section 5.5.

5.4 Application Performance

One objective of intent clustering is to build a high-quality labeled dataset for training intent classifiers

Method	Accuracy	Performance Gain
K-Means (Baseline)	0.65	-
ClusterLLM	0.62	-4.62%
LLMEdgeRefine	0.63	-3.08%
IDAS	0.68	+4.62% *
Keyphrase Expansion	0.72	+10.77% ***
LLM-ITL (context-free)	0.73	+12.31% ***
LLM-ITL (context-aware)	**0.77**	**+18.46%** ***

Table 6: Accuracy of BERT classifiers trained on datasets generated by different clustering methods.

that can handle future inputs (Gung et al., 2023). To evaluate the practical effectiveness of our methods, we trained BERT classifiers (bert-base-chinese) on clustered data generated by different approaches.

As shown in Table 6, the baseline method only achieved 65% accuracy. ClusterLLM and LLMEdgeRefine performed slightly worse, likely due to their heavier reliance on LLM-driven modifications, which can introduce label noise or overly fine-grained distinctions that reduce cluster consistency. In contrast, IDAS applies a more conservative refinement strategy, leading to a modest but stable improvement. Our proposed LLM-in-the-loop methods substantially outperformed existing approaches: LLM-ITL reached 73% accuracy, and its enhanced variant achieved 77%, representing an 18.46% relative improvement. These results are consistent with the observed cluster quality and provide additional empirical evidence that our method produces high-quality, human-aligned intent clusters with significant practical advantages.

5.5 Analysis of Computational Cost

We compare the computational cost of the proposed method with existing LLM-guided clustering on the Bank77 dataset, which contains $S = 3,080$ test sentences and $N = 77$ true clusters. Our LLM-in-the-loop approach evaluates candidate cluster numbers $N = [10, 30, 50, 70]$, requiring $\sum_{n_i \in N} n_i = 160$ calls per iteration. With $T = 3$ iterations, this totals 480 calls, each processing 20 sentences for coherence evaluation. Cluster naming requires additional LLM calls based on the final cluster count, resulting in approximately 560 calls. In contrast, ClusterLLM uses a fixed triplet sampling strategy with $Q = 1,024$, resulting in a constant cost of 1,024 calls. Keyphrase expansion generates one keyphrase per sentence, totaling 3,080 calls. While our method processes more input tokens per call, it remains cost-efficient as the output is limited to a short "good" or "bad" label or an intent label.

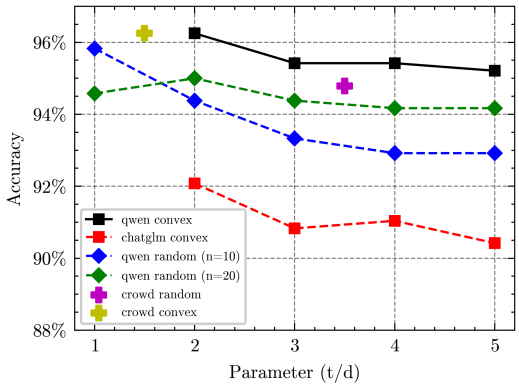

Figure 3: Performance comparison of sampling methods and hyperparameters for effective LLM evaluation.

5.6 Ablation Studies

Data Sampling and LLM Crowdsourcing for Effective Coherence Evaluation

Optimizing input to the LLM evaluator is crucial for improving performance and reducing costs (Song et al., 2025; Jiang et al., 2024). In this ablation study, we compare various sampling techniques to identify best practices for coherence evaluation. As shown in Figure 3, convex sampling outperforms random sampling in selecting representative sentences. Notably, increasing either the sample repetitions t for repeated validation or the convex hull dimension d consistently reduces performance. This suggests that simpler parameter choices yield higher accuracy, a trend also seen with the ChatGLM model (red line) and summarized in Table 15 with numerical comparison.

To better assess the consistency and alignment between LLM judgments and human perceptions, we utilize the concept of crowdsourcing for collaborative coherence evaluation (Hong et al., 2025b). The proposed LLM crowd consists of diverse entities (Zhang et al., 2024) represented by four fine-tuned LLMs and uses a majority voting mechanism to aggregate their judgments (Schoenegger et al., 2024). Based on the evaluation, we argue that **a single fine-tuned LLM can effectively serve as a robust evaluator**, as evidenced by the comparable performance to crowdsourced accuracy (cross markers in Figure 3) when using convex sampling.

Effectiveness of Cluster Merging Techniques

Table 7 evaluates the effects of distance measures and merging strategies on the proposed cluster merging method. The results show that geodesic distance in hyperspherical space outperforms cosine similarity by capturing deeper semantic re-

Method	NMI	Goodness
Geodesic distance (probabilistic)	**0.8420**	**97.8%**
Cosine similarity (deterministic)	0.8102	94.7%
Cosine similarity (probabilistic)	0.8242	95.2%
Geodesic distance (deterministic)	0.8309	95.4%

Table 7: Ablation study on cluster merging methods.

	Top1	Top2	Top3	Top4	Top5
Accuracy	26.32%	47.37%	73.68%	84.21%	89.47%

Table 8: Search space pruning by LLMs.

lationships. Furthermore, the probabilistic merging strategy consistently enhances robustness by accounting for uncertainty, leading to better clustering quality with any distance measure. These findings underscore the value of combining an appropriate distance metric with a probabilistic approach to achieve optimal merging performance.

Parameter Search Space Pruning

In practice, the search space for the number of clusters can be extensive, ranging from a few options to several hundred. To mitigate the computational costs associated with traditional model-based selection, we propose an LLM-native method for search space pruning, which predicts the optimal cluster number for subsequent iterations using logs from previous iterations (see Table 16). By pruning the search space and selecting the top 5 non-repetitive predicted solutions, we achieved an accuracy of 89.47% (see Table 8), indicating a high likelihood that the optimal cluster number is among predicted candidates. This approximation approach significantly reduces redundant model fittings and improves the efficiency of applying the proposed methods to large-scale datasets.

Binary Judgment vs. Numerical Scoring for Coherence Annotation

Finally, to identify the optimal annotation strategy for coherence evaluation, we compared a numerical 1 – 4 scoring system (1: very poor, 4: very coherent) with the proposed binary good/bad judgment. We annotated 1,000 clusters with input from five human experts, finalized scores through majority voting, and fine-tuned four Chinese LLMs using this dataset. The models were then evaluated on 200 additional clusters, assigning scores five times per cluster and consolidating results via majority voting (see results in Table 9). The 1 – 4 scale

LLM	qwen7b	qwen14b	baichuan2-7b	chatglm3-6b
Accuracy	62.50%	**65.00%**	58.50%	60.00%

Table 9: Performance of fine-tuned LLMs in assessing cluster coherence using a numerical 1 – 4 scale.

demonstrated low accuracy and inconsistency, with 19% of samples observed to have significant disagreement (e.g., three '1's and two '4's) in Qwen-2.5-14B, making it unreliable for assessing cluster quality or aligning with human judgment. In contrast, the simpler binary good/bad judgment provided more consistent, interpretable, and reliable quality assessments, demonstrating its superiority as an evaluation protocol.

To further validate the effectiveness and robustness of our annotation protocol, we conducted an inter-annotator agreement study with five experts. They annotated 100 intent clusters generated by the K-means algorithm, covering 2,000 sentences from recent customer service dialogues. Each cluster was evaluated using (i) a binary Good/Bad judgment and (ii) a finer-grained 1 – 4 rubric. The Fleiss' kappa (κ) for the binary scheme reached 0.82, indicating "almost perfect" agreement (Landis and Koch, 1977), while the 1 – 4 rubric achieved only 0.59 ("moderate" agreement) due to variability in intermediate scores (2 and 3). These results show that binary annotation provides a more reliable and interpretable measure of cluster coherence, reinforcing the robustness of our methodology.

6 Conclusion

This paper tackles dialogue intent clustering through a human-aligned LLM-in-the-loop framework. Experiments on a large-scale Chinese customer service dataset demonstrate that fine-tuned LLM utilities are highly effective for semantic coherence evaluation and cluster labeling, enabling consistent improvements over existing LLM-guided baselines in both clustering quality and computational efficiency. Beyond achieving state-of-the-art performance, our study offers strong empirical evidence for the effectiveness of LLM-in-the-loop methodologies, with ablation studies highlighting best practices. Future work should refine evaluation beyond coherence to capture interpretability and expressiveness, extend the clustering framework to multilingual settings, and explore deeper task-centric integration of LLMs to further advance intent mining in real-world applications.

Limitations

Despite the effectiveness of the proposed LLM-in-the-loop intent clustering method, this study has several limitations. First, while cluster coherence is a practical and intuitive quality indicator, its sole reliance overlooks other critical attributes, such as meaningfulness and expressiveness, which are equally important for assessing the quality of intent clusters. For example, a cluster labeled "express-feeling" may be too broad and could be refined into more specific clusters like "express-appreciation" to improve interpretability and applicability. Additionally, the binary evaluation and numerical scoring used are both deterministic metrics with fixed scales, neglecting probabilistic judgments and confidence intervals that could provide deeper assessment insights and enhance flexibility for LLM-in-the-loop integration.

Second, the proposed intent clustering dataset is limited to the Chinese language and specific domains. Although it excels in capturing large-volume, realistic customer service dialogues, its scope restricts multi-domain and multilingual generalizability, potentially limiting the applicability of findings to other languages and domains. Moreover, this paper focuses solely on intent clustering, an initial step in the broader intent discovery process. Subsequent steps, such as analyzing intent trajectories and recognizing intents at the document level rather than the sentence level, encourage further investigation with more diverse LLM utilities and varied integration strategies to better reveal the practical value of LLM-in-the-loop methodologies.

Acknowledgments

This paper is partially supported by several ongoing projects led or coordinated by Prof. Zhang Chen, including P0045948 and P0046453 (industry donations from Accel Group Holding Limited and Minshang Creative Technology Holdings Limited), P0046701 and P0046703 (PolyU internal research funding), P0048183 and P0048191 (Research Matching Grant Scheme funded by the University Grants Committee), P0048887 (Innovation and Technology Fund - ITSP, ITS/028/22FP), P0051906 (RGC Early Career Scheme, 25600624), and P0054482 (industry donations from Two Square Capital Limited).

References

Reina Akama, Sho Yokoi, Jun Suzuki, and Kentaro Inui. 2020. Filtering noisy dialogue corpora by connectivity and content relatedness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 941–958, Online. Association for Computational Linguistics.

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*.

Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, Suvrit Sra, and Greg Ridgeway. 2005. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(9).

Sandra Carberry and Margot Flowers. 1988. Modeling the user's plans and goals. *Computational Linguistics*, 14(3):23–37.

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.

Inigo Casanueva, Ivan Vulić, Georgios Spithourakis, and Paweł Budzianowski. 2022. NLU++: A multi-label, slot-rich, generalisable dataset for natural language understanding in task-oriented dialogue. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1998–2013, Seattle, United States. Association for Computational Linguistics.

CB Chandrakala, Rohit Bhardwaj, and Chetana Pujari. 2024. An intent recognition pipeline for conversational ai. *International Journal of Information Technology*, 16(2):731–743.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.

May Jane Chen. 1993. A comparison of chinese and english language processing. In *Advances in psychology*, volume 103, pages 97–117. Elsevier.

Maarten De Raedt, Fréderic 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.

Zhihao Ding, Yongkang Sun, and Jieming Shi. 2025. Retrieve-and-verify: A table context selection framework for accurate column annotations. *Preprint*, arXiv:2508.17203.

Zijin Feng, Luyang Lin, Lingzhi Wang, Hong Cheng, and Kam-Fai Wong. 2024. LLMEdgeRefine: Enhancing text clustering with LLM-based boundary point refinement. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18455–18462, Miami, Florida, USA. Association for Computational Linguistics.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.

P Thomas Fletcher, Conglin Lu, Stephen M Pizer, and Sarang Joshi. 2004. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE transactions on medical imaging*, 23(8):995–1005.

Wentao Gao, Ziqi Xu, Jiuyong Li, Lin Liu, Jixue Liu, Thuc Duy Le, Debo Cheng, Yanchang Zhao, and Yun Chen. 2025. Tsi: A multi-view representation learning approach for time series forecasting. In *AI 2024: Advances in Artificial Intelligence*, pages 291–302, Singapore. Springer Nature Singapore.

Attri Ghosal, Arunima Nandy, Amit Kumar Das, Saptarsi Goswami, and Mrityunjoy Panday. 2020. A short review on different clustering techniques and their applications. In *Emerging Technology in Modelling and Graphics*, pages 69–83, Singapore. Springer Singapore.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

James Gung, Raphael Shu, Emily Moeng, Wesley Rose, Salvatore Romeo, Arshit Gupta, Yassine Benajiba, Saab Mansour, and Yi Zhang. 2023. Intent induction from conversations for task-oriented dialogue track at DSTC 11. In *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 242–259, Prague, Czech Republic. Association for Computational Linguistics.

Floris Hengst, Ralf Wolter, Patrick Altmeyer, and Arda Kaygan. 2024. Conformal intent classification and clarification for fast and accurate intent recognition. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2412–2432, Mexico City, Mexico. Association for Computational Linguistics.

Mengze Hong, Di Jiang, Yuanfeng Song, and Chen Jason Zhang. 2024. Neural-bayesian program learning for few-shot dialogue intent parsing. *Preprint*, arXiv:2410.06190.

Mengze Hong, Wailing Ng, Chen Jason Zhang, and Di Jiang. 2025a. Qualbench: Benchmarking chinese LLMs with localized professional qualifications for vertical domain evaluation. In *The 2025 Conference on Empirical Methods in Natural Language Processing*.

Mengze Hong, Wailing Ng, Chen Jason Zhang, Yifei Wang, Yuanfeng Song, and Di Jiang. 2025b. Llm-in-the-loop: Replicating human insight with llms for better machine learning applications.

Mengze Hong, Chen Jason Zhang, Chaotao Chen, Rongzhong Lian, and Di Jiang. 2025c. Dialogue language model with large-scale persona data engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 961–970, Albuquerque, New Mexico. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Maximilian Jerdee, Alec Kirkley, and M. E. J. Newman. 2024. Normalized mutual information is a biased measure for classification and community detection. *Preprint*, arXiv:2307.01282.

Di Jiang, Kenneth Wai-Ting Leung, and Wilfred Ng. 2016a. Query intent mining with multiple dimensions of web search data. *World Wide Web*, 19(3):475–497.

Di Jiang, Kenneth Wai-Ting Leung, Lingxiao Yang, and Wilfred Ng. 2015. Teii: Topic enhanced inverted index for top-k document retrieval. *Knowledge-Based Systems*, 89:346–358.

Di Jiang, Yuanfeng Song, Rongzhong Lian, Siqi Bao, Jinhua Peng, Huang He, Hua Wu, Chen Zhang, and Lei Chen. 2021. Familia: A configurable topic modeling framework for industrial text engineering. In *International Conference on Database Systems for Advanced Applications*, pages 516–528. Springer.

Di Jiang, Yongxin Tong, and Yuanfeng Song. 2016b. Cross-lingual topic discovery from multilingual search engine query log. *ACM Transactions on Information Systems (TOIS)*, 35(2):1–28.

Di Jiang, Jan Vosecky, Kenneth Wai-Ting Leung, and Wilfred Ng. 2013. Panorama: A semantic-aware application search framework. In *Proceedings of the*

16th international conference on extending database technology, pages 371–382.

Di Jiang, Chen Zhang, and Yuanfeng Song. 2023. *Probabilistic topic models: Foundation and application*. Springer.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2024. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1658–1677, Bangkok, Thailand. Association for Computational Linguistics.

Atif Khan, Qaiser Shah, M Irfan Uddin, Fasee Ullah, Abdullah Alharbi, Hashem Alyami, and Muhammad Adnan Gul. 2020. Sentence embedding based semantic clustering approach for discussion thread summarization. *Complexity*, 2020(1):4750871.

Rajat Kumar, Mayur Patidar, Vaibhav Varshney, Lovekesh Vig, and Gautam Shroff. 2022. Intent detection and discovery from user logs via deep semi-supervised contrastive clustering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1836–1853.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

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, Lizi Liao, Hao Fei, and Jing Jiang. 2024. 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.

Haitao Lin, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2022. Other roles matter! enhancing role-oriented dialogue summarization via role interactions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2545–2558, Dublin, Ireland. Association for Computational Linguistics.

Ting-En Lin, Hua Xu, and Hanlei Zhang. 2020. Discovering new intents via constrained deep adaptive clustering with cluster refinement. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8360–8367.

Yulin Luo, Ruichuan An, Bocheng Zou, Yiming Tang, Jiaming Liu, and Shanghang Zhang. 2024. Llm as dataset analyst: Subpopulation structure discovery with large language model. In *European Conference on Computer Vision*, pages 235–252. Springer.

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, page 262–272, USA. Association for Computational Linguistics.

Anup Pattnaik, Cijo George, Rishabh Tripathi, Sasanka Vutla, and Jithendra Vepa. 2024. Improving hierarchical text clustering with llm-guided multi-view cluster representation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 719–727.

Walter Kintsch Peter W. Foltz and Thomas K Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25(2-3):285–307.

Libo Qin, Wenbo Pan, Qiguang Chen, Lizi Liao, Zhou Yu, Yue Zhang, Wanxiang Che, and Min Li. 2023. End-to-end task-oriented dialogue: A survey of tasks, methods, and future directions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5925–5941, Singapore. Association for Computational Linguistics.

Chen Qu, Liu Yang, W Bruce Croft, Johanne R Trippas, Yongfeng Zhang, and Minghui Qiu. 2018. Analyzing and characterizing user intent in information-seeking conversations. In *The 41st international acm sigir conference on research & development in information retrieval*, pages 989–992.

Philipp Schoenegger, Indre Tuminauskaite, Peter S. Park, Rafael Valdece Sousa Bastos, and Philip E. Tetlock. 2024. Wisdom of the silicon crowd: Llm ensemble prediction capabilities rival human crowd accuracy. *Science Advances*, 10(45):eadp1528.

Dingjie Song, Wenjun Wang, Shunian Chen, Xidong Wang, Michael X. Guan, and Benyou Wang. 2025. Less is more: A simple yet effective token reduction method for efficient multi-modal LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7614–7623, Abu Dhabi, UAE. Association for Computational Linguistics.

Xiaoshuai Song, Keqing He, Pei Wang, Guanting Dong, Yutao Mou, Jingang Wang, Yunsen Xian, Xunliang Cai, and Weiran Xu. 2023. Large language models meet open-world intent discovery and recognition: An evaluation of ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10291–10304, Singapore. Association for Computational Linguistics.

Ah-Hwee Tan and 1 others. 1999. Text mining: The state of the art and the challenges. In *Proceedings of*

the pakdd 1999 workshop on knowledge disocovery from advanced databases, volume 8, pages 65–70.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2009. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th annual international conference on machine learning*, pages 1073–1080.

Vijay Viswanathan, Kiril Gashteovski, 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.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 641–649, New York, NY, USA. Association for Computing Machinery.

Hui Yin, Xiangyu Song, Shuiqiao Yang, Guangyan Huang, and Jianxin Li. 2021. Representation learning for short text clustering. In *Web Information Systems Engineering–WISE 2021: 22nd International Conference on Web Information Systems Engineering, WISE 2021, Melbourne, VIC, Australia, October 26–29, 2021, Proceedings, Part II 22*, pages 321–335. Springer.

Chen Jason Zhang, Yunrui Liu, Pengcheng Zeng, Ting Wu, Lei Chen, Pan Hui, and Fei Hao. 2024. Similarity-driven and task-driven models for diversity of opinion in crowdsourcing markets. *The VLDB Journal*, 33(5):1377–1398.

Hanlei Zhang, Hua Xu, Ting-En Lin, and Rui Lyu. 2021. Discovering new intents with deep aligned clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14365–14373.

Yuwei Zhang, Zihan Wang, and Jingbo Shang. 2023. ClusterLLM: Large language models as a guide for text clustering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13903–13920, Singapore. Association for Computational Linguistics.

A Data Collection and Annotation

This section details the three-stage data collection and annotation process for the proposed Chinese customer service dialogue intent clustering dataset.

Data Collection. The raw dialogue dataset was derived from audio transcriptions of customer service calls in three major domains: banking, telecommunication, and insurance. In total, the dataset contained 11,879 calls. To ensure data integrity and protect confidentiality, strict filtering was applied to exclude sensitive content, resulting in 8,184 dialogues with 69,839 sentences. Further cleaning was conducted to remove duplicate sentences, yielding a final set of 55,085 unique sentences.

Data Annotation. To construct the intent clustering dataset, we recruited 15 human experts to conduct the annotations. Each annotator has over five years of professional experience in customer service call centers, with domain expertise aligned with the dataset domains: banking, telecommunication, and insurance. All annotators are fluent in Chinese, familiar with domain-specific terminology, and experienced in dialogue annotation or quality assurance tasks. Prior to annotation, a mandatory training session was held, during which the authors provided 50 intent clusters annotated with Good/Bad judgments and Action–Objective intent labels as demonstration and reference examples. This training ensured a consistent understanding of the guidelines across all annotators. The annotation process is outlined as follows:

1. K-means clustering was initially applied with $n = 2000$, serving as a starting point for the data annotation.

2. Each expert assessed the initial clusters by inspecting the semantic coherence of the sentences. They were instructed: "Label the cluster as 'Good' if all sentences share the same intent; otherwise, label it as 'Bad'. This resulted in 1283 good clusters and 717 bad clusters.

3. For the good clusters, annotators were asked to label the underlying intent using the naming convention "Action-Objective." After intent annotation, 1255 clusters with unique intentions were retained, and 28 clusters were merged due to replicated intent.

Model	Coherence Evaluation	Cluster Naming
LLaMA-2-7B (LoRA)	94.0% ± 1.2%	89.0% ± 1.5%
LLaMA-2-13B (LoRA)	**95.0%** ± 1.1%	**93.5%** ± 1.4%
Mistral-7B (LoRA)	91.5% ± 2.3%	82.0% ± 1.9%
GPT-3.5 (API)	91.0% ± 1.5%	86.0% ± 1.8%
GPT-4 (API)	94.5% ± 1.2%	91.0% ± 1.3%

Table 10: Performance of different LLMs on coherence evaluation and cluster naming for the English dataset.

4. For the bad clusters, annotators reassigned each sentence to the appropriate intent cluster based on the annotated labels. Sentences that did not fit into any preexisting clusters were assigned to new clusters, and the same process was repeated as in Step 2.

Data Verification. In the final stage, a separate group of 10 experts reviewed the annotated clusters for accuracy and consistency. Any discrepancies were resolved through consensus, ensuring the dataset's reliability and validity for further analysis. The final dataset consists of 1,507 high-quality intent clusters.

B Prompt Template

The prompts for coherence evaluation and cluster naming are translated into English for demonstration purposes. Each input is accompanied by a few-shot demonstration with five input-output pairs to ensure consistency in the output format and enhance understanding of the task.

Coherence Evaluation - *Your are a helpful assistant for sentence clustering. Based on the relevancy and common points of the following sentences in a cluster, classify the cluster as: "Good" or "Bad". Only provide the label without any additional content.*

Example: input:[sentences] output:[label]

input:{[sentences]} output:

Cluster Naming - *Your are a helpful assistant for sentence clustering. Based on the relevancy and common points of the following sentences in a cluster, summarize the cluster with an "Action-Objective" label. Only provide the label without any additional content.*

Example: input:[sentences] output:[label]

input:{[sentences]} output:

C More Results: Evaluation with Proprietary LLM

Given the resource-intensive nature of LLM fine-tuning, we conducted additional experiments using proprietary LLMs accessed through OpenAI APIs: GPT-3.5, GPT-4, and GPT-4o. These widely adopted models deliver strong performance across diverse tasks and do not require task-specific fine-tuning, thereby alleviating the data scarcity issue, but they also incur higher operational costs due to token consumption. On the same Chinese coherence evaluation dataset with 480 clusters, these models achieved accuracies of 89.58%, 93.54%, and 94.17%, respectively. Notably, the smaller fine-tuned Qwen-2.5-7B reached 96.25% accuracy, surpassing these advanced proprietary models while significantly reducing API-related costs. For comparison, the vanilla Qwen-2.5-7B model (without fine-tuning) obtained a much lower accuracy of 75.63%, further underscoring the importance of fine-tuning.

On the English dataset, we report results for GPT-3.5 and GPT-4, as well as two additional fine-tuned models: LLaMA-2-13B (LoRA) and Mistral-7B (LoRA). The LLaMA and Mistral models were fine-tuned on the same 800 intent clusters as LLaMA-2-7B in Section 5.3 and evaluated on 100 English intent clusters annotated by human experts with Good/Bad and Action-Objective intent labels. Each model was run five times with different random seeds to ensure robust performance metrics. Results in Table 10 show that scaling up model size (e.g., 7B vs. 13B) improves performance, while fine-tuned smaller LLMs often outperform large proprietary LLMs, consistent with the findings on the Chinese dataset. These results reinforce the value of fine-tuning and suggest that smaller, cost-efficient models can play a critical role in data mining within the LLM-in-the-loop framework.

Cluster Coherence	Original Sentences	English Translation
Good	"给企业固定资产买保险，大约能投大约得投保多少呢", "就是假如我有一百万的企业固定资金买保险大概投保多少钱啊", "请问一下我想咨>询一下企业财产保险", "您好我想了解一下这个企业财产保险", "就了解一下这个企业财产保险","你们是财险", "是哦这个属于财险了对吧","企业财产保险的", "这些都属于财产险对吗","财险人工那你这是", 就是如果我要为我这个私营企业买这个保险需要什么手续", "财产险","呃企业财产保险是以什么为保"	"If I buy insurance for a company's fixed assets, how much insurance will it cost?", "That is, if I have a million corporate fixed assets, how much will it cost to buy insurance?", "Excuse me, I would like to consult> Ask about corporate property insurance", "Hello, I want to know about this corporate property insurance", "Just want to know about this corporate property insurance", "You are a property insurance company", "Yes, this belongs to property insurance "Right?", "Enterprise property insurance", "These all belong to property insurance, right?", "What about property insurance workers?", What do I need if I want to buy this insurance for my private enterprise? Procedure","Property Insurance", "Well, what does corporate property insurance cover?"
Bad	"就是连连续，就是一直一直保","但是它连不上是怎么回事啊？", "哦就是主要是直接给公司转账对吧？","接吗","接也是吗？", "我直>接去","你是直接直接用那个" ,"就是从哪接过来再接回去","还是需要从哪儿连线这个宽带", "你直接给我说这些啊", "直接把","哦直接","嗯那个礼品是直接就发放了", "是你直接给我回复对吗","直接到那里去", "嗯，最好直飞。","请帮我连接+", "把钱打过去的话，我是直接打到那个证券公司","直接就是"	"It's continuous, it's always guaranteed", "But what's wrong with it not being able to connect?", "Oh, it's mainly to transfer money directly to the company, right?", "Yes", "Yes too?" ?", "I'll go directly", "Are you using that directly" , "Just connect it from where you are and then connect it back", "Or do you need to connect to the broadband from somewhere", "You Just tell me this directly", "Just give it directly", "Oh directly", "Well, the gift was given out directly", "You replied to me directly, right?","Go there directly" , "Well, it's best to fly directly.", "Please help me connect +", "If I call the money, I will call the securities company directly", "Directly"

Table 11: Example of "Good" and "Bad" intent clusters in coherence evaluation.

Cluster Name	Original Sentences	English Translation
询问-优惠 (Inquire-Promotion)	'那有什么优惠券什么之类的吗？', '是是怎么形式是优惠吗？', '还是不是优惠活动，是那个直接给我打我卡里吗 ，还是是什么优惠券儿啊？', '就比如新用户他有什么优惠券儿之类的吧！', '嗯你你那还有什么优惠的活动吗? 就是比较合适就是合适的动。'	'Are there any coupons or anything like that? ', 'What is the form of the discount? ', 'Still, it's not a promotion. Is it the one that directly charges my card? , or is it some kind of coupon? ', 'For example, what coupons does a new user have? ', 'Well, do you have any other discounts? It is more appropriate and appropriate to move. '
解答-金额 (Answer-Amount)	'一共是三十一块二毛', '就每个月一百三十八', '对，一个月也就是四百百四五百块钱嘛， 给您自己做个积累。', '十二月份的话是用了三十三块九毛二', '对一个月一百三十八'	'The total is thirty-one and twenty cents', 'That's one hundred and thirty-eight cents per month', 'Yes, that's four hundred, four hundred and five hundred yuan per month. Make an accumulation for yourself. ', 'In December, it cost thirty-three dollars and ninety-two cents', 'That's one hundred and thirty-eight dollars a month'

Table 12: Example of cluster naming with "Action-Objective" convention.

Dataset	Naming Convention	Example (sentence)	Example (label)
NLU	scenario-intent	Send an email to Alex and write thank you.	email sendemail
NLU++	list of keywords	How long does it usually take to get a new pin?	["how_long","pin", "arrival","new"]
OOS	objective	Please tell me why my card was declined yesterday.	card_declined
ours	action-objective	Well, do you have any other discounts?	inquire-promotion

Table 13: Comparison of cluster naming conventions in existing and proposed intent clustering datasets.

Cluster Name	Clusters with Similar Intention	English Translation
询问-意外事故 (Inquire-Accident)	"假如被车碰了或者是被楼上的砖砸了一下", "给别人儿撞的意外", "啊撞到别人然后就是", "就平时有时候开车嘛可能会遇到这个", "嗯哦这种情况，那要是就是我自己不小心撞到了那个某个地方然后", "就是把其他的东西撞到了呀什么的", "那人生意，不是就是，如果是不小心在马上被车撞了的话", "撞到人了吧", "然后不小在行驶当中被别人损害就是说拿石头砸的呀然后", "什么被车撞了之类的，是吗", "被被被撞了，还是被什么一些什么意外事故了", "我把别人的车撞了是吧"	"If I were hit by a car or hit by a brick from upstairs," "Accidentally hit by someone else," "Ah, hit someone else and then," "Just sometimes when driving, you might encounter this," "Well, if I'm not careful and hit some place myself," "Just hit something else or something," "That's a human affair, not just, if it's accidentally hit by a car on the road," "Hit someone, right?" "Then not small in the process of driving, being damaged by someone else, say, hit with a stone, and then," "What, hit by a car or something, right?" "I hit someone else's car, right?"
询问-意外死亡 (Inquire-Accident Death)	"哦猝死，那猝死算意外吗", "啊，那我想知道那个猝死的话，算是意外死亡吗", "算意外死亡吗", "那如果猝死算是意外死亡吗？", "猝死也算意外事吧", "那那猝死是意外死亡吗如果是猝死的话是", "猝死算是意外死亡吗", "猝死算意外死亡吗", "嗯，那个猝死，猝死属于意外意外险吗，意外死亡吗", "那猝死的话，算意外死亡吗？", "那那那那个就是那个猝死算是意外死亡吗"	"Oh sudden death, is sudden death considered accidental?" "Ah, I want to know if sudden death is considered accidental death?" "Is it considered accidental death?" "If sudden death is considered accidental death?" "Sudden death is also an accident, right?" "Is sudden death considered accidental death if it is sudden death?" "Is sudden death considered accidental death?" "Is sudden death considered accidental death?" "Well, that sudden death, does sudden death fall under accidental insurance, accidental death?" "Is sudden death considered accidental death?" "Is that, that, that sudden death considered accidental death?"

Table 14: Example of two high-quality intent clusters with similar intentions.

Model	Sampling Method	1	2	3	4	5
qwen14b	convex	-	**96.25**%	95.42%	95.42%	95.21%
	random (n=10)	**95.83**%	94.38%	93.33%	92.92%	92.92%
	random (n=20)	**94.58**%	95.00%	94.38%	94.17%	94.17%
chatglm3-6b	convex	-	**92.08**%	90.83%	91.04%	90.42%
	random (n=10)	**92.29**%	88.75%	85.83%	84.58%	85.21%
	random (n=20)	**90.42**%	89.58%	89.38%	89.58%	88.98%

Table 15: Comparison of sampling methods and hyperparameters for LLM coherence evaluation.

Epoch	1th				2th				3th			
	n_cluster	good	bad	rate	n_cluster	good	bad	rate	n_cluster	good	bad	rate
	20	20548	34537	0.595	20	0	2901	0.0	20	9	438	0.021
	50	24100	30985	0.778	50	372	2529	0.147	50	70	377	0.186
	100	32292	22793	1.417	100	804	2097	0.383	100	79	368	0.215
	...	...	...	...	...	...	...	...	...	...	...	...
Best	1600	52184	2901	17.988	800	2454	447	5.490	200	86	361	0.238

Table 16: Example log records from iterative intent clustering across epochs.