A Notation

Table 5 provides a summary of the key notations used throughout this paper.

Table 5: Summary of notations.

Symbol	Description
$\mathcal{Y}, \mathcal{Y}_k, \mathcal{Y}_u$	Sets of all, known, and unknown intents.
K, M	Total and known intent counts.
$\mathcal{D}_l, \mathcal{D}_u$	Sets of labeled and unlabeled utterances.
$\mathcal{D}_{train}, \mathcal{D}_{test}$	Training and testing sets.
X, x_i	Set of utterances and the i-th utterance.
$\mathcal{X}, \mathbf{x}_i$	Set of utterance embeddings and the embedding for x_i.
y_i	The i-th utterance's intent.
N, d	Number of utterances and embedding dimension.
T, t	Total iterations and current iteration index.
$\mathsf{PTE}(\cdot)$	Pre-trained Text Encoder.
$\mathsf{LLM}(\cdot)$	Large Language Model for generation and refinement.
p_{smry}	The prompt template for summary generation.
$C_k, \boldsymbol{\mu}_k$	The k-th cluster and its Euclidean centroid.
$\boldsymbol{\theta}_k, s_k$	The semantic centroid and textual summary for cluster C_k.
$\boldsymbol{\theta}_{\text{nbr}(k)}$	The nearest neighboring semantic centroid to $\boldsymbol{\theta}_k$.
$\mathcal{S}_k$	Set of representative exemplars from cluster C_k.
$f(\cdot)$	The clustering cost function.
$\cos(\cdot)$	Cosine similarity between two utterance embeddings.
$\mathcal{L}^{\text{ED}}$	The Euclidean distance cost.
$\mathcal{L}^{\text{SC}}, \alpha$	The semantic cohesion cost and its weight.
$\mathcal{L}^{\text{SS}}, \beta$	The semantic separation cost and its weight.
p_{ref}	The prompt template for utterance refinement.
$\mathcal{H}$	Set of identified hard samples for refinement.
$H(\cdot), \delta$	Shannon entropy function and the number of hard samples.
K_{nbr}	The number of neighboring clusters for HSR context.
$\tilde{x}_i, \tilde{\mathbf{x}}_i$	Refined utterance and its new embedding.
$s_h, \mathcal{S}_h$	Summary and exemplars of a sample's home cluster.
$\{\boldsymbol{\mu}_k^0\}_{k=1}^K$	The initial Euclidean centroids.
$\{\boldsymbol{\mu}_j^*\}_{j=1}^M$	Seed centroids from labeled data $\mathcal{D}_l$.
$\mathcal{L}^{\text{SP}}, \gamma$	The supervised cost and its weight.
$\mathsf{Mean}(\cdot)$	The mean of a set of embeddings.

B Baseline Repositories

Table 6 lists the public code repositories used for the baseline methods in our experiments.

Table 6: Code repositories for baselines.

Baseline	Code Repository
SAE/DEC	https://github.com/piiswrong/dec
DCN	https://github.com/boyangumn/DCN
CC	https://github.com/XLearning-SCU/2021-AAAI-CC
SCCL	https://github.com/amazon-science/sccl
KCL/MCL	https://github.com/GT-RIPL/L2C
DTC	https://github.com/k-han/DTC
CDAC+	https://github.com/thuiar/CDAC-plus
GCD	https://github.com/sgvaze/generalized-category-discovery
DeepAligned	https://github.com/HanleiZhang/DeepAligned-Clustering
MTP-CLNN	https://github.com/fanolabs/NID_ACLARR2022
LatentEM	https://github.com/zyh190507/Probabilistic-discovery-new-intents
USNID	https://github.com/thuiar/TEXTOIR
SDC	https://github.com/Lackel/SDC
LANID	https://github.com/floatSDSDS/LANID

C Hyperparameter Settings

Table 7 details the hyperparameter configurations. Across all settings, we employ USNID as $\mathsf{PTE}(\cdot)$ and GPT-4o-Mini as $\mathsf{LLM}(\cdot)$. We use fixed parameters for DCS and HSR: $|\mathcal{S}_k| = 10$, $\delta = 10$, and $K_{\text{nbr}} = 10$. For brevity, these constant values are omitted from the main hyperparameter table.

We can observe that MMR proves to be the most effective choice in the majority of cases. Notably, for the semi-supervised experiments on the BANKING and CLINC datasets, we opt for a similarity-based mapping strategy instead of one reliant on the LLM. This is because these datasets feature a large number of known intents (over a hundred). The LLM struggles to fully comprehend and accurately map such a wide range of intents while strictly adhering to the required output format. Consequently, a direct similarity-based mapping provides a more robust and effective solution in high-cardinality scenarios.

Table 7: Hyperparameter settings for NILC.

Setting	Dataset	Selection Strategy	T	α	β	γ	Mapping Strategy
Unsupervised	BANKING	MMR	3	0.5	0.5	–	–
	CLINC	MMR	3	0.5	0.3	–	–
	DBPedia	MMR	3	0.9	0.5	–	–
	M-CID	NN	2	0.3	0.3	–	–
	SNIPS	MMR	3	0.3	0.5	–	–
	StackOverflow	MMR	3	0.9	0.1	–	–
Semi-Supervised	BANKING	MMR	3	0.9	0.3	0.5	Similarity-based
	CLINC	MMR	3	0.5	0.5	0.5	Similarity-based
	DBPedia	MAD	3	0.5	0.5	0.5	LLM-based
	M-CID	NN	3	0.3	0.1	0.5	LLM-based
	SNIPS	MMR	3	0.5	0.5	0.5	LLM-based
	StackOverflow	MMR	3	0.9	0.7	0.1	LLM-based

D Selection Strategies for $\mathcal{S}_k$

K-Means++. This strategy adapts the seeding procedure of K-Means++ to select a geometrically diverse set of exemplars. The selection is iterative. Let $\mathcal{S}_k^{(i)}$ be the set of i selected exemplars. The first exemplar, $\mathbf{x}_1$, is chosen uniformly at random from C_k to form $\mathcal{S}_k^{(1)}$. For $i = 2, \ldots, |\mathcal{S}_k|$, each subsequent exemplar $\mathbf{x}_i$ is chosen from the remaining embeddings $C_k \setminus \mathcal{S}_k^{(i-1)}$ with a probability proportional $G(\mathbf{x}_i)$ to its minimum squared Euclidean distance to the set of already-selected exemplars $\mathcal{S}_k^{(i-1)}$:

$$G(\mathbf{x}_i) = \frac{\min_{\mathbf{x}_s \in \mathcal{S}_k^{(i-1)}} ||\mathbf{x}_i - \mathbf{x}_s||^2}{\sum_{\mathbf{x}_j \in C_k \setminus \mathcal{S}_k^{(i-1)}} \min_{\mathbf{x}_s \in \mathcal{S}_k^{(i-1)}} ||\mathbf{x}_j - \mathbf{x}_s||^2} \tag{14}$$

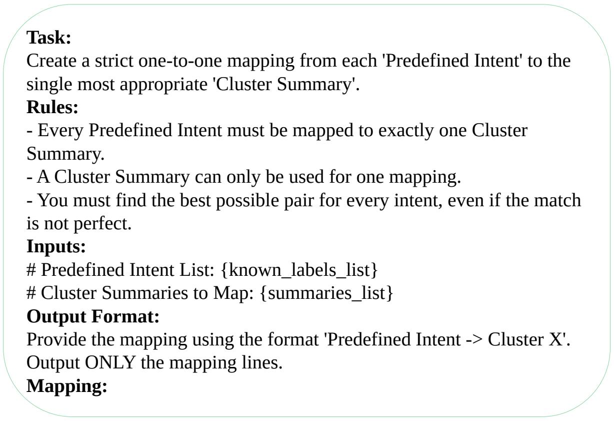
This method is designed to maximize the diversity of $\mathcal{S}_k$, ensuring broad coverage of the cluster's semantic space.

Mean Average Distance (MAD). The MAD strategy identifies exemplars from the cluster's periphery by selecting those that are, on average, most dissimilar from other members of the cluster. We select the set $\mathcal{S}_k$ by maximizing the mean distance:

$$\mathcal{S}_k = \underset{\mathcal{S} \subset C_k, |\mathcal{S}| = |\mathcal{S}_k|}{\arg\max} \sum_{\mathbf{x}_i \in \mathcal{S}} \frac{1}{|C_k| - 1} \sum_{\mathbf{x}_j \in C_k, j \neq i} ||\mathbf{x}_i - \mathbf{x}_j|| \tag{15}$$

Table 8: Evolution of s_{10} for C_{10} on M-CID.

Iteration	Summary s_{10} for Cluster C_{10}
1	What cleaning and disinfecting practices are effective in preventing the spread of COVID-19 on surfaces?
2	How long does the coronavirus survive on various surfaces and materials, and what cleaning practices are recommended?
3	What are the best cleaning practices and precautions to prevent COVID-19 transmission from surfaces and packages?



Task:
Create a strict one-to-one mapping from each 'Predefined Intent' to the single most appropriate 'Cluster Summary'.
Rules:
- Every Predefined Intent must be mapped to exactly one Cluster Summary.
- A Cluster Summary can only be used for one mapping.
- You must find the best possible pair for every intent, even if the match is not perfect.
Inputs:
# Predefined Intent List: {known_labels_list}
# Cluster Summaries to Map: {summaries_list}
Output Format:
Provide the mapping using the format 'Predefined Intent -> Cluster X'.
Output ONLY the mapping lines.
Mapping:

Figure 16: Prompt template for LLM-based mappings.

The theoretical justification is that these boundary points are crucial for defining the cluster's extent and improving its separation from neighboring clusters.

Maximal Marginal Relevance (MMR). MMR provides a formal framework for balancing relevance to the cluster's central theme with the diversity of the selected exemplars. After an initial exemplar is chosen based on maximum similarity to the geometric centroid $\boldsymbol{\mu}_k$, subsequent exemplars are selected iteratively to maximize the following objective function:

$$\underset{\mathbf{x}_j \in C_k \setminus \mathcal{S}_k}{\arg\max} \left[\cos(\mathbf{x}_j, \boldsymbol{\mu}_k) - \max_{\mathbf{x}_s \in \mathcal{S}_k} \cos(\mathbf{x}_j, \mathbf{x}_s) \right] \tag{16}$$

This ensures that $\mathcal{S}_k$ is composed of exemplars that are both highly representative and non-redundant.

Nearest Neighbors (NN). This strategy selects the most central and prototypical instances of the cluster. The centrality $C(\mathbf{x}_i)$ of an embedding is defined as its cumulative similarity to all other embeddings within the cluster:

$$C(\mathbf{x}_i) = \sum_{\mathbf{x}_j \in C_k, j \neq i} \cos(\mathbf{x}_i, \mathbf{x}_j) \tag{17}$$

The set $\mathcal{S}_k$ is formed by the utterances corresponding to the embeddings with the highest centrality scores. The premise is that the most central points are the most faithful representatives of the underlying intent.

E Mapping Strategies for π^t

Embedding-based Mapping. This approach matches the known seed centroids $\{\boldsymbol{\mu}_j^*\}_{j=1}^M$ to the current semantic centroids $\{\boldsymbol{\theta}_k^t\}_{k=1}^K$ by solving the assignment problem that minimizes cosine dissimilarity, using the Hungarian algorithm:

$$\min_{\pi^t} \sum_{m=1}^{M} \left(1 - \cos(\boldsymbol{\mu}_j^*, \boldsymbol{\theta}_{\pi^t(j)}^t)\right) \tag{18}$$

LLM-based Mapping. This strategy employs the LLM to perform a direct semantic mapping between the known intent labels $\mathcal{Y}_k$ and the generated cluster summaries $\{s_k^t\}_{k=1}^K$. As detailed in Fig. 16, the prompt p_{map} is designed to constrain the LLM to behave like an optimal assignment algorithm to create a strict one-to-one mapping by enforcing rules that require every intent to be matched with a unique cluster summary. This process compels the LLM to find the best possible pairing for each intent:

$$\pi^t = \text{LLM}\left(p_{\text{map}}, \mathcal{Y}_k, \{s_k^t\}_{k=1}^K\right) \tag{19}$$

F Case Studies

F.1 Evolution of Semantic Centroids

To understand how NILC iteratively refines the semantic understanding of each cluster, we can dive into the evolution of the LLM-generated cluster summaries, which act as semantic centroids. Table 8 presents a case study from M-CID, tracking the summary s_{10} for C_{10} over 3 iterations.

Initially, the cluster is summarized with a broad question about general cleaning practices. As the cluster assignments and embeddings are refined, the summary evolves to become more specific and detailed. By Iteration 2, it narrows its focus to the survivability of the virus on surfaces. Finally, in Iteration 3, the summary crystallizes into a precise and actionable utterance about the "best practices and precautions" for preventing surface transmission. This progressive refinement demonstrates how NILC's iterative process enhances the coherence and specificity of the discovered intents.

F.2 Successful Refinement of Hard Samples

To illustrate the utility of HSR, we present a qualitative example on StackOverflow. HSR clarifies ambiguous utterances by leveraging the LLM's contextual understanding.

For instance, the utterance x_i = "Confusion regarding laziness" is ambiguous. It was initially misclassified into a cluster about LINQ queries because "laziness" can relate to deferred execution, which is not the utterance's core intent. NILC identifies this high-uncertainty sample and provides the LLM with the context of its assigned and neighboring clusters, as detailed in Table 9. Note that the "True Cluster" is shown for illustrative purposes; the LLM only receives the "home" cluster and neighboring clusters as the context, not its ground-truth identity.

The LLM analyzes the competing intents and, recognizing that "laziness" is a core concept in Haskell, rewrites the utterance into a clear and specific question. The new embedding $\tilde{\mathbf{x}}_i$ for the refined utterance has a much lower clustering cost and is confidently reassigned to the correct Haskell-related cluster. This case study

Table 11: Analysis of Semi-Supervised Mapping Strategies.

Dataset	Mapping Strategy	NMI	ARI	ACC
DBPedia	Similarity-based	89.41	83.96	91.43
	LLM-based	89.36	84.19	91.57
M-CID	Similarity-based	81.49	70.56	83.09
	LLM-based	83.06	72.48	84.53
StackOverflow	Similarity-based	80.13	75.89	86.85
	LLM-based	80.53	76.47	87.18

Table 12: Analysis of Representative Sampling Methods.

Dataset	Selection Strategy	NMI	ARI	ACC
DBPedia	MMR	89.36	84.19	91.57
	MAD	89.99	84.88	92.00
	NN	88.90	83.81	91.43
	K-Means++	89.36	84.03	91.43
M-CID	MMR	83.06	72.48	84.53
	MAD	81.83	70.64	83.09
	NN	83.36	73.12	85.10
	K-Means++	82.62	71.66	83.67
StackOverflow	MMR	80.53	76.47	87.18
	MAD	80.19	75.86	86.77
	NN	80.33	76.15	86.93
	K-Means++	80.28	76.16	86.95

Table 9: An example of the HSR process for an ambiguous utterance on StackOverflow.

Component	Content
Hard Sample (x_i)	Confusion regarding laziness
Assigned Cluster (C_4)	**Summary** (s_4)**:** What are the various techniques and best practices for effectively using LINQ to query and manipulate data, including handling distinct values, dynamic queries, joins, and return types?
True Cluster (C_6)	**Summary** (s_6)**:** What are some common challenges and best practices when working with Haskell, including syntax, error handling, and functional constructs?
LLM Task	Given the context, analyze the best fit for the utterance and rewrite it to be an unambiguous exemplar of that theme.
Refined Utterance ($\tilde{x}_i$)	Understanding laziness in functional programming languages like Haskell

Table 10: Comparison of mapping strategies on DBPedia.

Mapping Strategy	Summary s_k for Cluster C_k	Mapped Known Intent
Similarity-based	C_0: Books and Publications	WrittenWork
	C_1: Notable Individuals	OfficeHolder
	C_2: Plant Species	Plant
	C_3: Organisms	Animal
	C_4: Historic and Cultural Institutions	Building
	C_6: Historical Vehicles and Vessels	MeanOfTransportation
	C_8: Diverse Companies and Organizations	Company
	C_9: Films	Film
	C_{10}: Geographical Features	NaturalPlace
	C_{13}: **Professional Athlete**	**Artist**
LLM-based	C_0: Literary Works	WrittenWork
	C_1: Notable Individuals	OfficeHolder
	C_2: Plant Species	Plant
	C_3: Organisms	Animal
	C_4: Historic and Cultural Institutions	Building
	C_5: **Music Albums and Compilations**	**Artist**
	C_6: Historical Vehicles and Vessels	MeanOfTransportation
	C_8: Corporations and Organizations	Company
	C_9: Films	Film
	C_{10}: Geographical Features	NaturalPlace

demonstrates how HSR actively corrects the data manifold, improving cluster cohesion and separation by resolving ambiguity.

F.3 Comparison of Two Mapping Strategies

To showcase the superiority of our LLM-based mapping approach over the traditional similarity-based method, we present a detailed comparison of the mappings generated for DBPedia, as shown in Table 10.

While both methods successfully map many clusters, the similarity-based approach, which relies on cosine distance between centroids, makes a critical error. It incorrectly maps C_{13}, summarized as “Professional Athlete”, to the known intent “Artist”. Although athletes can be metaphorically considered “artists” of sports, this is not the correct semantic relationship on DBPedia. This error underscores a fundamental limitation of relying purely on embedding similarity in a Euclidean space; such an approach is confined to geometric proximity and lacks the awareness of semantic context, possibly causing it to be misled by abstract or metaphorical connections that an LLM, with its richer world knowledge, can correctly disambiguate.

In contrast, our LLM-based method leverages its world knowledge and reasoning capabilities. It correctly discerns that “Professional Athlete” does not fit the “Artist” category and instead makes the more semantically sound decision to map C_5 (“Music Albums and Compilations”) to “Artist”. This leads to a more effective injection of semi-supervised signals, a more accurate must-links constraint relationship, and, ultimately, a more accurate final clustering.

G Empirical Studies of Mapping and Sampling Strategies

We further analyze the specific strategies used to represent clusters. Table 11 compares the LLM-based mapping strategy against a traditional similarity-based approach for semi-supervised NID. The LLM-based strategy consistently outperforms the similarity-based one, particularly on the M-CID dataset, where it yields a 1.92% improvement in ARI. This shows that LLMs can capture the semantic alignment between known intent labels and cluster summaries more effectively than simple embedding similarity. In Table 12, we analyze different representative sampling methods for generating cluster summaries. While all methods perform well, the MAD, NN, and MMR strategies show slight advantages on DBPedia, M-CID, and StackOverflow, respectively, suggesting that the optimal sampling strategy can be dataset-dependent.