POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Uncertainty-Aware Continual Learning for Open-World Intent Discovery Under an Evolving Label Space

Tesi di Laurea Magistrale in
Automation and Control Engineering - Ingegneria dell'automazione

Author: **Aida Pisante**

Student ID: 252402
Advisor: Prof. Simone Formentin
Co-advisors: Feng Tian
Academic Year: 2025-2026

Abstract

Intelligent systems deployed in real-world settings increasingly operate under *open-world* conditions, in which the user intent space is neither fixed nor exhaustively known a priori and may expand as novel interaction patterns emerge. Prior work has addressed intent detection, discovery and knowledge retention largely in isolation; their joint incremental treatment across multiple learning phases remains insufficiently explored.

This thesis introduces a unified uncertainty-aware, probabilistic and adaptive framework for continual new intent discovery under a dynamically evolving label space. Each utterance is encoded by an adaptive β-VAE into a latent mean, supporting classification and density modelling and a posterior uncertainty estimate, which acts as a global reliability signal. A multi-signal decision layer fuses classifier confidence, posterior uncertainty and DP-GMM likelihood to distinguish known intents from novel candidates. Reliable clusters identified by a density-based module are promoted to new intent labels, while Elastic Weight Consolidation and experience replay consolidate the model after each expansion, mitigating catastrophic forgetting.

The framework advances the field on three axes: (i) continual intent discovery is formalised as a structured multi-phase open-world problem; (ii) a label-space expansion mechanism is adaptive and proposed under *stability–plasticity* constraints; (iii) posterior uncertainty serves as a unified control signal regulating sample selection, pseudo-labelling, novelty admission and replay.

Experiments show controlled novelty detection, stable adaptation across sequential phases and limited forgetting through replay–EWC consolidation. Near-zero NMI and ARI values indicate limited recovery of the full fine-grained taxonomy, consistent with the framework's conservative promotion policy. Nevertheless, qualitative analyses show that several promoted clusters form dense and locally coherent semantic regions, supporting reliable local discovery under an evolving label space.

Keywords: continual intent discovery; open-world learning; uncertainty-aware modelling; label space expansion; adaptive thresholding.

Abstract in lingua italiana

I sistemi intelligenti reali operano in ambienti *open-world* dove lo spazio degli intenti non è fisso né noto a priori e può espandersi con l'emergere di nuovi pattern di interazione. I lavori esistenti tendono a trattare rilevamento, scoperta e conservazione della conoscenza come problemi separati, la loro integrazione congiunta in un processo incrementale multi-fase rimane ancora poco esplorata.

Questa tesi introduce un *framework probabilistico unificato, consapevole dell'incertezza e adattivo* per la *scoperta continua di nuovi intenti* in uno spazio delle etichette in evoluzione. Ciascun enunciato è codificato da una β-VAE adattiva in una media latente, a supporto di classificazione e modellazione della densità e in una stima dell'incertezza posteriore, usata come segnale di affidabilità globale. Un livello decisionale multi-segnale combina confidenza del classificatore, incertezza posteriore e verosimiglianza del DP-GMM per riconoscere nuovi candidati. I cluster affidabili identificati da un modulo basato sulla densità vengono promossi a nuove etichette, mentre Elastic Weight Consolidation e replay consolidano il modello dopo ogni espansione, mitigando la dimenticanza catastrofica.

Il framework apporta tre avanzamenti: (i) la scoperta continua è formalizzata come problema strutturato di apprendimento open-world multi-fase; (ii) un meccanismo adattivo di espansione dello spazio delle etichette è proposto sotto vincoli di *stabilità–plasticità*; (iii) l'incertezza posteriore funge da segnale di controllo unificato che regola selezione dei campioni, pseudo-etichettatura, ammissione alla novità e replay.

I risultati mostrano rilevamento controllato della novità, adattamento stabile su fasi sequenziali e dimenticanza catastrofica limitata. Valori prossimi allo zero di NMI e ARI indicano recupero limitato dell'intera tassonomia fine-grained, coerente con la politica conservativa di promozione. Le analisi qualitative mostrano tuttavia che diversi cluster promossi formano regioni semantiche dense e localmente coerenti, a supporto di una scoperta locale affidabile in uno spazio delle etichette in evoluzione.

Parole chiave: scoperta continua degli intenti; apprendimento open-world; modellazione consapevole dell'incertezza; espansione dello spazio delle etichette; soglia adattiva.

Contents

1 | Introduction

1.1. Background

1.1.1. Intelligent Dialogue Systems and Intent Understanding

Intelligent dialogue systems enable natural language interaction between humans and machines and are widely deployed in domains such as customer service, virtual assistants, healthcare and e-commerce. They are increasingly relevant to intelligent automation contexts, where natural-language interfaces can support human-machine interaction, robotic supervision, operator assistance and decision support in complex systems. Their architecture typically comprises automatic speech recognition, semantic understanding, dialogue management and response generation, with **Natural Language Understanding (NLU)** acting as the core component for extracting actionable meaning from user utterances[1].

Within NLU, **intent classification** is fundamental: it identifies the user's communicative goal and drives downstream decision-making in task-oriented dialogue systems. Traditional closed-set formulations often model this task as utterance-level classification over a predefined set of intent categories. Recent approaches increasingly rely on deep neural architectures and pre-trained language models to capture richer semantic and contextual information[2]. However, intent understanding is not inherently static, since user goals may evolve across dialogue turns and require continuous reinterpretation based on context[2].

This challenge becomes particularly critical in complex settings such as medical conversations, where intent understanding must remain sensitive to evolving user needs and contextual dependencies over time.

1.1.2. Limitations of Closed-Set Intent Classification

Traditional dialogue systems operate under a **closed-set assumption**, wherein all possible intent categories are predefined during training and each user utterance is assumed to belong to one of these known classes. This assumption renders such models inadequate for

real-world deployment. A further limitation is that the manual construction of intent taxonomies is itself costly, labor-intensive and must be repeated over time as new user needs emerge. Consequently, closed-set intent classification cannot adequately capture the dynamic, incomplete and open-ended nature of real-world conversational environments[1][2].

These limitations motivate the need for dialogue systems that can move beyond fixed label spaces and operate in more realistic open-world conditions.

1.1.3. Novel Intent Detection

To overcome the limitations of closed-set intent classification, recent research has focused on **novel intent detection**, which identifies utterances that do not belong to any known intent class seen during training. This problem is closely related to **out-of-distribution (OOD) detection**, as the system must recognize when an input lies outside the semantic distribution of in-domain data[3].

In **confidence-based approaches**[4], classifier confidence is used to determine whether a prediction should be trusted or assigned to an unknown category. **Density-based methods**[4] instead model the distribution of known intent representations in feature space, identifying novel inputs as points in low-density regions or far from learned class centres.

Together, these approaches form the main methodological basis for novel intent detection in open environments, assuming that known intents occupy compact semantic regions, while unseen intents tend to lie outside them.

1.1.4. New Intent Discovery

Novel intent detection serves as an intermediate step: unknown utterances are first detected and then passed to subsequent modules for **new intent discovery (NID)**, which uncovers and organises previously unseen intents into meaningful semantic categories (Figure 1-1).

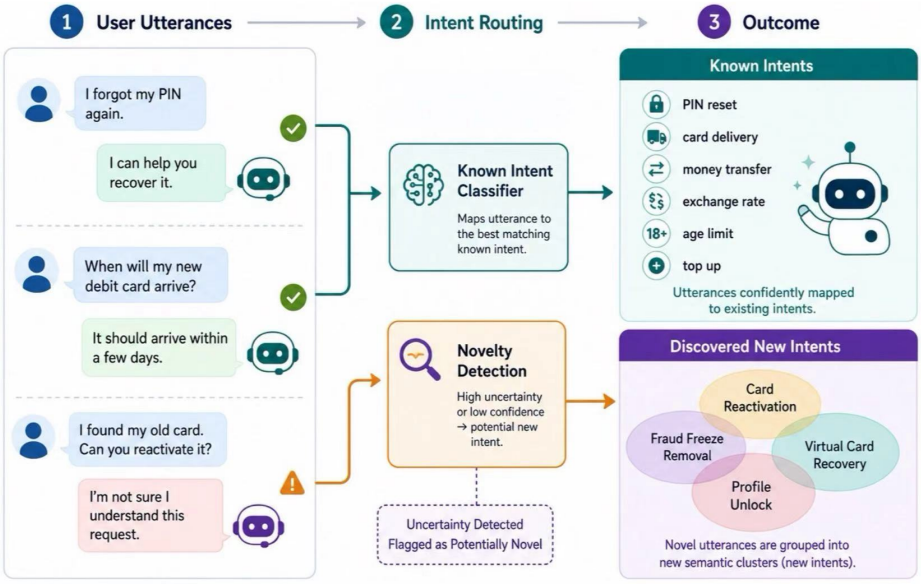

Figure 1.1: Intent discovery process based on clustering and representation learning[9].

A widely adopted paradigm in this area is **clustering-based discovery**[5][6], in which unlabelled or partially labelled utterances are first mapped into a semantic feature space and then grouped according to their similarity in order to reveal latent intent structure. The effectiveness of this process depends strongly on **representation learning**[7], since the quality of the learned utterance embeddings directly affects the separability and coherence of the resulting clusters. For this reason, recent works[8][9][17] have extensively explored contextualized representations derived from pre-trained language models such as BERT, MPNet and other sentence embedding models [40][43][44], which better capture semantic relations than traditional feature engineering methods. Building on these representations, **deep clustering approaches**[5] integrate feature learning and clustering more tightly, often through semi-supervised, contrastive, or alignment-based strategies that use known intents to facilitate the emergence of unknown ones.

Intent discovery should not be viewed solely as a clustering problem: in more complete open frameworks, it is formulated as a two-stage process that includes both semantic clustering and the generation of human-readable labels for the discovered clusters, thereby making the resulting intents directly usable in downstream dialogue systems. In addition, the performance of intent discovery methods can vary substantially depending on dataset properties such as the number of intents, class imbalance, utterance length and vocabulary diversity, which further highlights the complexity of discovering intents in realistic open-world settings.

1.2. Related Works

1.2.1. Category and Intent Discovery Methods

Category and intent discovery methods aim to uncover latent semantic structures in unlabelled data. In this context, **new intent discovery** can be regarded as a task-specific subfield of the broader **new category discovery** paradigm. While new category discovery addresses the general problem of identifying previously unseen classes in unlabelled or partially labelled data. New intent discovery applies this principle to dialogue and language-understanding systems, where the emerging categories correspond to previously unseen user intents. Therefore, methods developed for category discovery provide a relevant methodological foundation for intent discovery, as both settings require learning separable representations, detecting unknown semantic groups and assigning meaningful structure to unlabelled data. The main difference lies in the application domain: intent discovery[41] focuses on fine-grained semantic categories expressed through natural language utterances, often under stronger ambiguity and contextual variability.

Early approaches rely on classical clustering techniques such as K-means, DBSCAN and Gaussian mixture models (GMMs), which partition data based on similarity in a predefined feature space[9]. However, purely unsupervised methods have limited ability to incorporate prior knowledge and often suffer from instability and low semantic consistency in high-dimensional spaces. To address these issues, recent research has shifted toward **deep category discovery**, where neural networks learn discriminative feature spaces more suitable for clustering, while pseudo-labelling and contrastive learning iteratively refine assignments. These approaches include deep aligned clustering and prototype-based methods[6], with more advanced frameworks incorporating structural information, such as graph-based propagation or diffusion mechanisms, to improve intra-cluster compactness and inter-cluster separability[7].

A further advancement is represented by **semi-supervised discovery methods**, which exploit a small amount of labelled data to guide discovery in large unlabelled datasets. These methods bridge supervised and unsupervised learning by transferring knowledge from known to unknown categories, enabling more robust modelling of open-world data. For example, **Generalized Category Discovery (GCD)**[12] extends traditional settings by allowing unlabelled data to contain both known and novel classes and combines probabilistic models with representation learning and dynamic class estimation [10][11][13]. Semi-supervised approaches further leverage contrastive learning[42], prototypical representations and pseudo-label refinement to produce cluster-friendly embeddings and im-

prove scalability[14]. However, most existing approaches operate under a **static and offline assumption**, where the entire dataset is available during training, limiting their applicability in real-world scenarios with sequential data and evolving label spaces.

Recent works have explored **generalized and continual intent discovery settings**, where models must both classify known intents and discover new ones under evolving distributions. **Generalized Intent Discovery (GID)**[14][15][16] extends traditional pipelines by jointly modelling in-domain and out-of-domain intents, while continual variants incrementally integrate newly discovered intents over time. Additionally, controllable and human-in-the-loop approaches incorporate prior knowledge and iterative feedback to improve interpretability and adaptability[17]. The emergence of large pre-trained language models has further enabled *few-shot and prompt-based methods*, which leverage implicit knowledge to identify and expand intent categories with minimal supervision[18].

Despite significant progress, **current methods remain limited in dynamic environments** where data arrives sequentially and categories evolve over time. While recent approaches move toward more realistic settings, they do not fully address incremental adaptation without compromising previously learned knowledge. This motivates the adoption of **continual learning paradigms**, which explicitly handle non-stationary data and long-term knowledge retention, as discussed in the following section.

1.2.2. Continual Learning for Emerging Categories

In real-world dialogue systems, the distribution of user intents is not static: new intents may emerge over time, creating **dynamic environments** where models must continuously adapt to evolving semantic categories[14][36]. Addressing these scenarios requires systems capable of incorporating new knowledge without degrading previously acquired information. This problem is studied within **continual learning**[45][46][48][49][50], which focuses on sequential learning from evolving data streams. A central challenge in continual learning is **catastrophic forgetting**[36], where neural networks overwrite previously acquired knowledge when trained on new tasks or distributions (Figure 1-2).

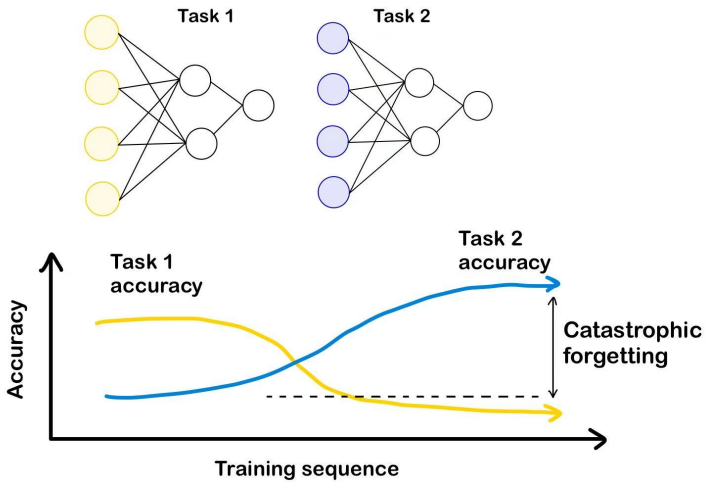

Figure 1.2: Catastrophic forgetting in a two-task incremental learning setting, showing how the acquisition of a new task can reduce the model's ability to retain previously learned knowledge. Adapted from [36].

This issue is particularly critical in category discovery settings, where newly discovered classes must be integrated while preserving existing structures, as highlighted in the *Variational Bayes Continual Generalized Category Discovery* **(VB-CGCD) framework**[18], which studies stability–plasticity trade-offs through covariance-aware representations and variational inference mechanisms to align old and new category distributions.

To address catastrophic forgetting, several strategies have been proposed. A widely used approach is **replay**[36], where a subset of past samples is retained and reused during training. In continual category discovery, replay is often combined with **prototype-based learning** to stabilize representations across stages. For example, the *Prototype-guided Learning with Replay and Distillation* **(PLRD) framework** for *Continual Generalized Intent Discovery* **(CGID)**[15] integrates replay and prototype learning to mitigate forgetting while incrementally discovering new categories. Revisiting past samples helps maintain balanced decision boundaries and prevent drift in the embedding space.

In addition to replay-based methods, **regularization-based approaches** constrain parameter updates to preserve prior knowledge. While classical techniques such as **Elastic Weight Consolidation (EWC)** operate at the parameter level, more recent methods focus on representation-level alignment to ensure consistency across learning stages. However, **these approaches mainly mitigate forgetting and do not explicitly model the evolving category structure**. This limitation motivates frameworks that jointly address category discovery in dynamic environments, as discussed in the following section.

1.2.3. Continual Category Discovery

Continual category discovery focuses on learning frameworks that *jointly perform category identification and model adaptation over sequential data*, integrating newly emerging categories into an evolving representation space without disrupting previously learned structures.

A key direction relies on **distribution-based discovery mechanisms**, where latent categories are modelled as evolving components in the embedding space, with assignments and parameters iteratively refined through procedures such as the **Expectation–Maximization (EM) algorithm**. In these approaches, mixture-based representations capture underlying semantic groups and are progressively updated via EM-style optimization to better fit evolving data distributions.

Another important direction is represented by **prompt-based continual category discovery (CCD) frameworks**, where adaptive prompts evolve with the data. The **PromptCCD framework**[19] for CCD in computer vision (Figure 1-3), exemplifies this approach by maintaining a dynamic pool of learnable prompts structured as a mixture, where each prompt acts as a category prototype. This enables incremental discovery of new categories while preserving previously acquired knowledge, effectively bridging representation learning and category discovery in sequential settings.

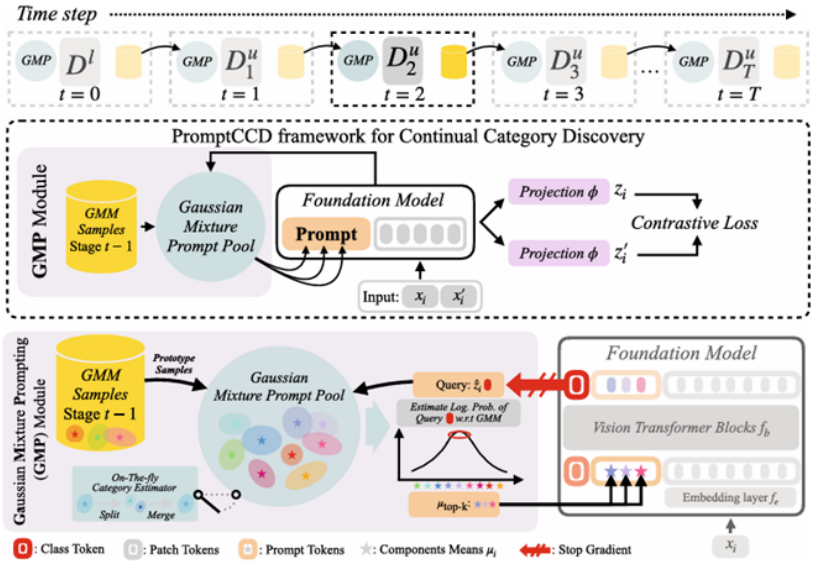

Figure 1.3: Overview of the PromptCCD framework, illustrating continual category discovery through Gaussian mixture prompting and contrastive learning across time steps. Adapted from [19].

These methods operate in **dynamic environments**, where both data distribution and category space evolve over time. Within the CGCD paradigm, frameworks such as PLRD combine prototype learning, replay mechanisms and representation alignment to handle both known and emerging categories during training. However, existing methods are typically **evaluated under simplified conditions and remain limited in addressing complex multi-phase category evolution**, where structures change continuously and unpredictably. Moreover, category distributions are often only implicitly modelled, motivating the need for more explicit probabilistic formulations.

1.2.4. Probabilistic Approaches for Category Modelling

Probabilistic approaches to category modelling aim to **explicitly represent the distributional structure underlying category discovery**, providing flexible and uncertainty-aware representations that complement learning-based frameworks.

While mixture-based formulations offer a general perspective for modelling latent category structures, recent work increasingly emphasizes **learning probabilistic representations** that capture both semantic structure and uncertainty. In particular, the *Proxy-Driven Robust Multimodal Fusion* **(P-RMF) framework**[20] introduces a probabilistic latent space to model incomplete and noisy multimodal data, enabling robust category representations through uncertainty-aware fusion. This is especially relevant in real-world scenarios with partially observed or heterogeneous data. Beyond multimodal settings, probabilistic modelling has also been applied to *capture structured latent representations and domain variability.* The *Gaussian Mixture Domain-Indexing* **(GMDI) framework**[21] models domain-specific variations through adaptive latent components, allowing representations to evolve with new data and handle distributional shifts. In parallel, **Bayesian formulations of intent modelling** further extend this perspective by learning compositional and transferable latent structures. The *Neural-Bayesian Program Learning model Dialogue-Intent Parser* (DI-Parser)[22], leverages probabilistic inference to construct structured intent representations, improving generalization in low-resource and cross-domain settings.

1.2.5. Limitations of Existing Studies in New Intent Discovery and Research Gap

Recent research has proposed several representative frameworks for new intent discovery and open-world intent modelling. **NID with Pre-training and Contrastive Learning (MTP-CLNN)**[9], by Zhang et al. revisited in 2025, combines multi-task pre-training

with contrastive learning using nearest neighbours to improve utterance representations and cluster compactness. However, it mainly treats NID as a static unsupervised or semi-supervised problem, rather than as a continual multi-phase process. **Generalized Intent Discovery (GID)**[14], by Mou et al. in 2022, extends in-domain intent classification to an open-world setting by jointly classifying labelled IND intents and discovering unlabelled OOD types, but it still relies on a specified or estimated number of OOD classes and focuses on joint IND/OOD modelling rather than continual integration. **Continual Generalized Intent Discovery (CGID)**[15], introduced by Song et al. in 2023, extends GID to a dynamic open-world setting by continuously discovering OOD intents from sequential data streams and incrementally expanding the intent classifier. The proposed PLRD method combines prototype-guided pseudo-labelling, replay memory and feature distillation to balance new intent learning with the preservation of previously acquired knowledge. However, although CGID explicitly addresses continual intent discovery and catastrophic forgetting, it does not model posterior latent uncertainty as an operational control signal, nor does it use density-based probabilistic novelty detection or uncertainty-aware cluster promotion to regulate label-space expansion. **Controllable Discovery of Intents (CDI)**[16], proposed by Rawat et al. in 2023, introduces controllable intent discovery with MPNet, unsupervised contrastive learning, pseudo-label refinement, Learning without Forgetting and human-in-the-loop feedback. Although more incremental and controllable, it remains stage-wise and feedback-driven, without explicit latent uncertainty modelling or density-based novelty detection. **Robust and Adaptive Prototypical learning (RAP)**[13], introduced by Zhang et al. in 2024, improves semi-supervised NID through robust prototypical attracting and adaptive prototypical dispersing, encouraging within-cluster compactness and between-cluster separation. Nevertheless, it is still evaluated mainly in a static clustering-oriented setting. **IntentGPT: Few-shot Intent Discovery with Large Language Models**[17], proposed by Rodriguez et al. in 2024, uses large language models for few-shot intent discovery through in-context prompting, semantic few-shot sampling and known-intent feedback, removing task-specific training but depending on prompt design, context length and LLM inference constraints, without explicit continual stabilisation across phases.

These existing studies provide strong foundations for representation learning, open-world intent modelling, continual discovery, controllable discovery, prototype-based clustering and prompt-based few-shot prediction. However, they mostly address these aspects separately and usually remain tied to static or stage-wise discovery settings. They do not jointly model uncertainty-aware latent representations, density-based novelty detection, explicit label-space expansion, replay-based memory and parameter-level regularisation

in a unified multi-phase continual framework.

1.3. Main Research Content

1.3.1. Problem Definition

This research addresses the problem of *continual new intent discovery in dynamic conversational environments*, where the set of possible user intents is not fixed but evolves over time as new interaction patterns emerge.

This thesis is formulated within an **open-world learning scenario** with an **evolving label space**, where the system must operate under incomplete knowledge of the label space. In such a setting, the model must not only classify inputs belonging to known intents but must also detect previously unseen intents, discover their underlying structure and integrate them into the existing taxonomy without degrading previously learned knowledge.

A key motivation of this work is that, to the best of current knowledge, no prior research has explicitly investigated the problem of **uncertainty-aware continual NID in an open-world environment with an evolving label space**. While several works address open-set intent detection, clustering-based intent discovery or generalized category discovery, they typically assume either static datasets or single-stage discovery procedures.

The research presented fundamentally transforms the learning task from a static classification problem into a new formulation of discovery and adaptation. The system must simultaneously solve several **interconnected challenges**: preserving previously learned intents without catastrophic forgetting, detecting previously unseen inputs, discovering coherent clusters corresponding to emerging intents, expanding the label space by promoting reliable clusters to new intent categories and stabilizing the model across successive discovery phases. In this way, discovery is no longer treated as a one-time clustering problem, but as a repeated process that occurs multiple times during the lifetime of the system.

This capability is crucial in real-world conversational systems, where user behaviour changes over time and new intents must be incorporated to preserve long-term scalability and reliability. The research therefore supports adaptive intelligent systems operating under unknown, evolving and partially observable label spaces.

1.3.2. Overview of the Proposed Framework

To achieve the research objectives outlined in the previous section, the proposed architecture integrates **three subsystems**: a *latent representation module*, a *probabilistic discovery module* and a *continual learning stabilisation module*. These components interact continuously throughout a **multi-phase learning process** that alternates between supervised learning of known intents and unsupervised discovery of emerging intents.

The **latent representation module** is implemented as an **uncertainty-aware latent encoder** based on a *β-Variational Autoencoder* (**β-VAE**). For each input utterance, the encoder produces a compact latent vector together with its posterior standard deviation, providing an explicit measure of encoding reliability. This **posterior standard deviation** functions as a **global uncertainty-based control signal** that regulates decisions across the entire framework: it determines which samples are reliable candidates for novelty analysis, how pseudo-labels are propagated and how replay samples are selected. By embedding uncertainty directly into the latent space, the system gains a unified mechanism for managing trust, limiting unreliable updates and preventing error propagation when new intents are integrated.

The **probabilistic discovery module** operates on these latent representations to detect unknown samples and identify coherent clusters corresponding to potential new intents. This is achieved through **density-based mixture modelling**, which approximates the evolving latent distribution of both known and unknown samples and serves as a probabilistic memory of intent structure. Candidate unknown inputs are first filtered using classifier confidence combined with the posterior standard deviation criterion; only samples meeting these reliability conditions proceed to clustering. Dense regions identified by the **clustering algorithm** that satisfy reliability conditions are then promoted to newly discovered intent classes, allowing the label space to expand dynamically as new intents emerge.

The **continual-learning stabilisation module** ensures that this expansion does not degrade previously acquired knowledge. Since integrating new classes can cause *catastrophic forgetting*, the deterioration of earlier representations as the model adapts to new data, the module combines *experience replay*, which reinforces previously learned decision boundaries and preserves consistency in the latent space, with *parameter regularisation*, which constrains changes to parameters identified as important for earlier tasks. The posterior standard deviation again plays a role here, guiding replay selection by distinguishing reliable low-uncertainty samples from informative high-uncertainty ones.

This integration enables the framework to balance the two core requirements of open-world

learning: **stability**, which preserves previously acquired intent representations and **plasticity**, which supports adaptation to emerging intents and evolving data distributions. Such a balance is essential in dynamic environments where the label space is unknown, non-stationary and progressively expanding.

1.3.3. Multi-Phase Open-World Continual Learning Pipeline

These three subsystems operate within an **open-world continual learning architecture**, in which the system must function under *incomplete knowledge of the label space* while adapting to emerging semantic categories. Each phase corresponds to a stage in the system's knowledge evolution and reflects the recurrent discovery cycle described above.

In the *initial phase*, the system performs supervised learning on known intents, establishing a stable latent representation and an initial classifier.

In *subsequent phases*, new unlabelled data arrive that may contain unseen intents: the framework performs **novelty detection** to identify unknown candidates, applies **density-based clustering** to uncover coherent intent structures and **promotes reliable clusters to new intent categories**, thereby expanding the label space. Once new intents are incorporated, the system enters a **consolidation stage**, in which the classifier and latent representations are updated while replay and regularisation mechanisms preserve prior knowledge.

This iterative cycle produces a **progressively expanding label space**. The setting is further characterised by the *continual* nature of learning: data arrive sequentially as a stream containing both known and unseen intents, rather than in a single batch and new categories may appear at any time and must be recognised as lying outside previously learned classes. This requires ongoing updates to internal representations while preventing catastrophic forgetting.

1.3.4. Contributions

This thesis provides a conceptual and methodological advancement in the NID problem by addressing a largely unexplored gap in the literature: multi-phase continual discovery in open-world environments with an evolving label space. Although some limitations remain, mainly due to the novelty and complexity of the proposed setting, they also define clear directions for future work. Overall, the proposed contributions move beyond the current state of the art and, with further refinement, may provide a solid foundation for adaptive real-world dialogue systems.

The contributions of this research can be summarised as three interconnected advances.

The **first methodological advancement** is the formalisation of multi-phase continual NID in an open-world setting. Existing literature generally focuses on static or single-stage intent discovery, or treats open-set detection, clustering and incremental learning as separate problems. In contrast, this work explicitly models a setting in which the **label space is dynamic**, with each phase assigned a precise role within a recurrent discovery cycle. Discovery thus becomes a **recurrent and structured process**, rather than a one-time clustering operation, better reflecting realistic dialogue systems where new user needs must be incorporated through controlled discovery and consolidation.

A **second key aspect** is the development of a unified mechanism for label-space expansion under stability–plasticity constraints. Incremental learning methods typically assume a fixed label space, while open-set approaches can detect unknown samples without necessarily converting them into stable knowledge. In this thesis, discovered clusters are promoted to new labels through a multi-signal decision process explicitly coupled with replay and EWC, allowing the system to expand its classification space progressively while mitigating catastrophic forgetting.

A **third relevant feature** concerns the explicit treatment of uncertainty as a first-class control signal for continual NID. While uncertainty is present in variational and confidence-based methods, it is typically used only as an auxiliary diagnostic or static thresholding criterion. Here, posterior uncertainty derived from the β-VAE latent space is elevated to a system-level mechanism guiding the stability and reliability of the continual learning process, reducing error propagation under non-stationary conditions while supporting the controlled discovery and integration of new intents over time.

Together, these three contributions establish a perspective on intent discovery in which adaptive intelligent systems are not limited to recognising known categories, but are designed to repeatedly detect, discover, integrate and retain emerging intents over time. This capability is particularly valuable for adaptive conversational systems, where user behaviour evolves continuously and where stable knowledge expansion directly determines the long-term usability, robustness and scalability of the system.

1.4. Thesis Organization

This thesis is structured into five chapters, progressing from foundational concepts to the design, implementation and evaluation of the proposed framework. The organization reflects a transition from theory to methodology and from controlled assumptions to real-

world complexity, ensuring clarity and methodological rigor.

Chapter 1 introduces continual intent discovery within the broader context of open-world learning. It reviews the most relevant literature, highlighting the limitations of existing approaches and positions this work at the intersection of probabilistic modelling and continual learning, establishing its motivation and direction.

Chapter 2 presents the theoretical foundations of the framework, revisiting key concepts in probabilistic latent variable models, variational inference, Gaussian mixture modelling and continual learning. Emphasis is placed on uncertainty as a structured signal and on non-parametric Bayesian methods, which form the backbone of the approach.

Chapter 3 introduces a multi-phase framework for uncertainty-aware continual intent discovery. It presents the β-VAE latent representation, the uncertainty-aware modelling strategy and the density-based discovery mechanism based on Dirichlet Process Gaussian Mixture Models, showing how representation learning, uncertainty estimation and clustering interact to enable robust discovery.

Chapter 4 extends this architecture by focusing on learning dynamics over time. It details mechanisms ensuring stability and adaptability, including replay strategies, Elastic Weight Consolidation and uncertainty-guided decision rules and formalizes novelty detection, cluster promotion and label space expansion as a balance between plasticity and preservation.

Chapter 5 reports an extensive experimental evaluation, analysing performance across configurations and ablations, with focus on novelty detection, cluster quality and stability. In addition to quantitative metrics, it includes qualitative analysis of discovered clusters, providing interpretability and insight into model behaviour under uncertainty.

Chapter 5 concludes the thesis by summarising the main findings, discussing limitations and outlining directions for future research.

2 | Theoretical Background and Core Technologies

2.1. Sentence Embedding Models

2.1.1. Text Representation in Natural Language Processing

The goal of **Natural Language Processing (NLP)** is the representation of textual data in a form that can be effectively processed by computational models. As highlighted in the literature[23][24], the NLP pipeline begins with the *transformation of raw text into numerical representations*, which serve as the basis for subsequent modelling tasks. Given a sentence $s = (w_1, w_2, \ldots, w_n)$, the goal of text representation is to learn a mapping function: $f : S \to \mathbb{R}^d$ where S denotes the space of all possible sentences and $\mathbb{R}^d$ is a continuous d-dimensional embedding space. The resulting vector $z = f(s) \in \mathbb{R}^d$ encodes the semantic content of the sentence. Modern approaches rely on **dense embeddings**, where semantic similarity is encoded geometrically (Appendix A-1). By projecting sentences into a continuous vector space, embeddings enable models to capture both syntactic and semantic properties, supporting operations such as similarity search, clustering and semantic reasoning.

2.1.2. Sentence Transformer Architectures

Recent advances in NLP have led to the development of **Sentence Transformers**, a class of models that generate semantically meaningful sentence embeddings optimizing the embedding space using similarity-based objectives. These models are typically built upon pre-trained transformer architectures, such as BERT, which leverage self-attention mechanisms to model contextual dependencies across tokens (Appendix A-2). A common formulation is the *contrastive loss*, which enforces that semantically similar sentences have closer embeddings than dissimilar one (Appendix A-3). In this framework, the embedding function $f_\theta(s) = z$ parameterized by θ, directly maps sentences into a continuous

semantic space where distances reflect meaning relationships. This property makes Sentence Transformers particularly effective for tasks such as semantic search, clustering and intent discovery, where the geometry of the embedding space encodes latent semantic structure, enabling scalable and robust modelling of utterances in open-world settings.

2.2. Dimensionality Reduction

2.2.1. High-Dimensional Embedding Spaces

Modern representation learning models map input data into high-dimensional embedding spaces, where each sample is represented as a dense numerical vector. While these embeddings enable the capture of rich semantic relationships, they introduce several well-known challenges associated with high-dimensional data[25]:

- **Curse of dimensionality**: as dimensionality increases, distances become less discriminative, reducing the effectiveness of similarity-based methods and clustering algorithms.
- **Informational redundancy**: increases computational complexity and obscures the intrinsic low-dimensional structure underlying the data, leading to inefficient representations.
- **Instability in probabilistic models**: covariance estimates can become ill-conditioned or singular, leading to unreliable inference and degraded performance in models such as GMMs or Variational Autoencoder (VAE) frameworks.

These challenges motivate the use of dimensionality reduction techniques, which aim to extract the most informative structure from high-dimensional embeddings while mitigating noise and redundancy.

2.2.2. Principal Component Analysis

Principal Component Analysis (PCA) is one of the most widely used techniques for dimensionality reduction in multivariate statistics. In the context of representation learning, PCA is often used to *remove noise, reduce redundancy and improve the stability* of downstream models such as clustering algorithms or probabilistic latent-variable models, in order to obtain more compact and semantically meaningful representations, facilitating tasks such as category discovery.

According to Johnson and Wichern[25], PCA relies on the covariance matrix Σ, which, for

a dataset $X \in \mathbb{R}^{n \times d}$, captures how variables vary together and thus provides a compact description of the structure of multivariate data. PCA identifies the most informative directions in the data through the eigenvalue decomposition of the covariance matrix. The eigenvectors define the principal axes, corresponding to directions of maximum variance, while the eigenvalues measure the variance explained along each axis. After ordering the eigenvectors by decreasing eigenvalue, the data are projected onto the first k principal components, producing a lower-dimensional representation that preserves most of the original variance and retains the essential structure of the data.

2.2.3. PCA Whitening

PCA whitening[51][52] extends standard PCA by transforming the data into a representation in which features are both decorrelated and normalized, resulting in a more isotropic distribution. After projection onto the principal components, each component is divided by the square root of its corresponding eigenvalue, so that all transformed components have unit variance and the covariance matrix of the whitened data becomes the identity matrix. From a statistical perspective, PCA whitening standardizes the covariance structure of the data by removing scale differences across dimensions and ensuring that each component contributes equally to the representation (Figure 2-1). This preprocessing step improves the numerical stability of covariance-based methods, supports more reliable parameter estimation and is particularly useful for clustering and Gaussian-based models, where well-conditioned covariance estimates are essential.

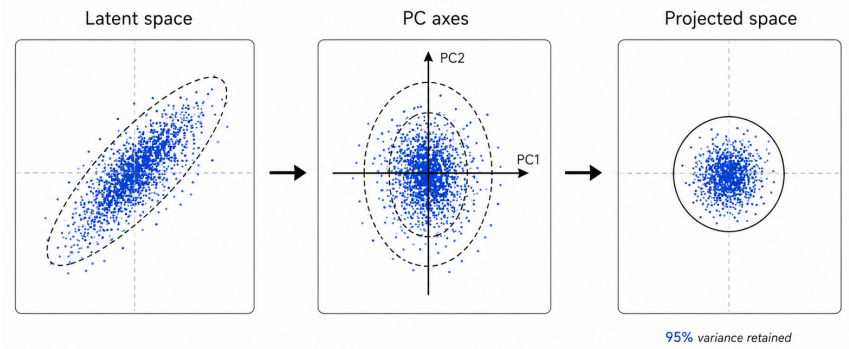

Figure 2.1: PCA projection of the latent space, retaining 95% of the variance in a compact representation. Adapted from [38].

2.3. Variational Autoencoders

2.3.1. Latent Variable Models

Latent variable models constitute a fundamental class of probabilistic models in which the observed data are assumed to be generated from underlying, unobserved variables that capture the intrinsic structure of the data[26] (Appendix A-4). These models are particularly valuable for **representation learning**, as they enable the construction of structured latent spaces in which semantically similar data points are mapped to nearby regions. As highlighted in recent multimodal learning approaches (e.g., Zhang et al.[9]), mapping data into a Gaussian latent space promotes compact, regularized and semantically consistent representations, improving robustness to noise and missing information.

2.3.2. Variational Autoencoder Architecture

The Variational Autoencoder (VAE) is a deep generative model that combines latent variable modelling and neural networks (NNs) to learn probabilistic representations and generate data. Following the formulation of Kingma and Welling[26], it assumes a **generative process** in which a latent variable is sampled from a prior $z \sim p(z)$ and the data are generated through $x \sim p_\theta(x \mid z)$. The VAE architecture consists of three main components (Figure 2-2):

1. Encoder Network

The encoder maps an input x to an approximate posterior distribution over the latent variable, denoted as $q_\phi(z \mid x)$. Unlike a deterministic encoder, it does not produce a single latent vector, but predicts the parameters of a probability distribution, typically a Gaussian[27]:

- **Mean vector**: $\mu = f_\mu(x)$;
- **Variance**: $\sigma^2 = f_\sigma(x)$.

These parameters define the approximate posterior distribution $q_\phi(z \mid x) = \mathcal{N}(\mu, \sigma^2)$. The mean provides the central latent representation of the input, while the variance quantifies the uncertainty associated with that representation. As shown in recent works[9], this probabilistic formulation allows the model to explicitly represent uncertainty and can improve robustness in the presence of noisy, ambiguous or incomplete data.

2. Latent Variable

The latent variable z is a compressed probabilistic representation of the input data, typically modelled as $z \sim \mathcal{N}(\mu, \sigma^2)$. This continuous latent formulation enables meaningful interpolation and sampling while preserving the main structure of the data distribution[26].

3. Decoder Network

The decoder *reconstructs the input from the latent representation by modelling the conditional distribution* $p_\theta(x \mid z)$. It acts as a generative model that maps latent variables back to the data space, allowing the generation of new samples by sampling from the prior distribution $p(z)$.

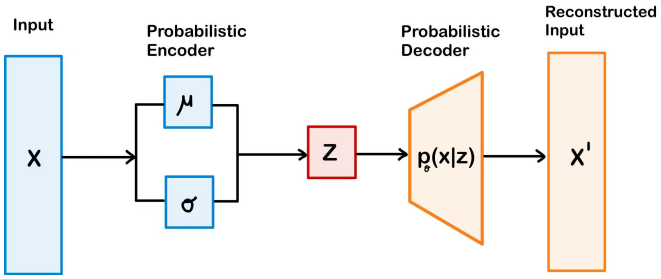

Figure 2.2: VAE framework illustrating probabilistic encoding, latent sampling with controlled variance and reconstruction through a regularized latent representation. Adapted from [39].

2.3.3. Variational Inference and Reparameterization

Variational inference addresses the intractability of the true posterior distribution $p(z \mid x)$ in latent variable models. In VAEs, this problem is handled by introducing an approximate posterior $q_\phi(z \mid x)$, parameterized by the encoder and typically modelled as a Gaussian distribution whose mean and variance are predicted by a neural network. A simple prior, usually a standard normal distribution, is imposed on the latent space to regularize the representation and promote a continuous and structured latent geometry.

The model is trained by maximizing the **Evidence Lower Bound (ELBO)**, which combines two objectives: a **reconstruction term**, encouraging the decoder to preserve the information needed to represent the input accurately and a **regularization term**, which constrains $q_\phi(z \mid x)$ to remain close to the prior. In the β-VAE formulation, the regularization term is weighted by a scaling factor β, allowing explicit control over the trade-off between reconstruction accuracy and the degree of latent-space regularization.

Since sampling from the latent distribution is stochastic, propagating gradients directly through z would be difficult. The reparameterization trick solves this issue by expressing the latent variable as a differentiable transformation of the encoder outputs and an external noise term, enabling end-to-end training through backpropagation[26]. A detailed formulation of the objective function and its probabilistic components is provided in Appendix A-5.

2.4. β-Variational Autoencoders

2.4.1. KL Regularization

The VAE is trained by maximizing the ELBO, which can be decomposed into two main components:

$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)} \left[\log p_\theta(x \mid z)\right] - D_{\mathrm{KL}} \left(q_\phi(z \mid x) \parallel p(z)\right) \tag{2.1}$$

Reconstruction Loss

The first term encourages accurate reconstruction of the input: $\mathbb{E}_{q(z|x)}[\log p(x \mid z)]$. It ensures that the latent variable retains sufficient information about the input data.

KL Divergence

The second term regularizes the latent space: $D_{\mathrm{KL}}(q(z \mid x) \parallel p(z))$. It enforces similarity between the approximate posterior and the prior, promoting smoothness and preventing overfitting[26].

2.4.2. β-VAE Objective Function

The β-VAE extends the standard VAE by introducing a **weighting parameter** β that controls the strength of the KL regularization:

$$\mathcal{L}_\beta(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)} \left[\log p_\theta(x \mid z)\right] - \beta\, D_{\mathrm{KL}} \left(q_\phi(z \mid x) \parallel p(z)\right) \tag{2.2}$$

When $\beta > 1$, the model emphasizes latent space regularization, encouraging disentangled representations at the cost of reconstruction accuracy. This trade-off allows the model to learn independent and interpretable latent factors, which is particularly useful in representation learning tasks[28].

2.5. Probabilistic Clustering Models

2.5.1. Mixture Models

Probabilistic clustering models assume that complex data distributions can be represented as mixtures of simpler latent components. In mixture models, each data point is assumed to originate from one of these hidden components (Appendix A-6). This makes

them suitable for modelling *multi-modal distributions*, which often arise in real-world data due to heterogeneous subpopulations and cannot be adequately captured by a single parametric distribution.

2.5.2. Gaussian Mixture Models (GMMs)

A **Gaussian Mixture Model (GMM)** is a specific instance of a mixture model where each component distribution is Gaussian (Appendix A-7). GMMs provide a *soft clustering mechanism* (Figure 2-3): instead of assigning each point to a single cluster, they compute posterior probabilities[19]: $\gamma_{ik} = p(z_i = k \mid x_i)$. These probabilities represent the degree to which each data point belongs to each cluster. Recent works[21] further extend GMMs by modeling latent spaces as mixtures of Gaussians to capture complex semantic structures and improve generalization across domains.

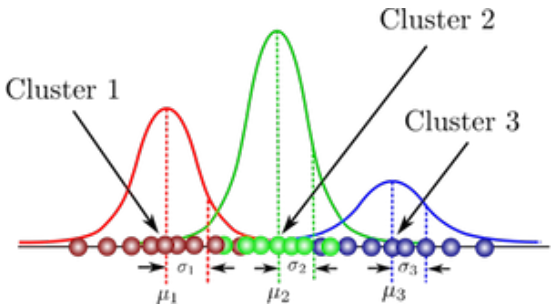

Figure 2.3: Illustration of GMM clustering, where data is modelled as a weighted combination of multiple Gaussian distributions. Adapted from [40].

2.5.3. Expectation-Maximization Algorithm

The standard approach for parameter estimation of a GMM is the **Expectation-Maximization (EM) algorithm**, which is an iterative procedure that alternates between estimating latent variables and optimizing parameters[29].

1. E-step (Expectation Step)

In the E-step, the algorithm computes the posterior probability (also called responsibility) that a data point x_i belongs to component k:

$$\gamma_{ik} = p(z_i = k \mid x_i) = \frac{\pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i \mid \mu_j, \Sigma_j)} \tag{2.3}$$

These responsibilities act as soft assignments of points to clusters.

2. M-step (Maximization Step)

In the M-step, the parameters are updated by maximizing the expected log-likelihood using the responsibilities computed in the E-step (Appendix A-8).

Iterative Optimization

The EM algorithm alternates between E-step and M-step until convergence, measured via the log-likelihood:

$$\log p(X) = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k) \right) \tag{2.4}$$

Due to the presence of the logarithm of a sum, direct maximization is intractable, which motivates the EM procedure[29].

2.6. Density-Based Category Modelling

2.6.1. Latent Space Clustering

The latent space provides a structured representation in which geometric and density-based relationships can reflect underlying semantic category boundaries[30]. For this reason, clustering methods are commonly applied in the latent space, where each mixture component can represent a potential category. This strategy is widely used in category discovery and representation learning, as it supports the identification of both known and novel classes. Recent studies[30][57] further show its relevance in open-world and continual discovery settings, where categories are not predefined but must be inferred from the data distribution.

2.6.2. Cluster Structure in Representation Spaces

Clusters arise when the learned representation captures the statistical structure of the data. Semantically similar samples are mapped to nearby regions of the latent space, forming dense groups, whereas dissimilar samples tend to be separated by lower-density regions. The quality of this structure depends strongly on the representation itself: well-regularized latent spaces generally promote more separable and semantically coherent clusters, supporting downstream tasks such as classification and novelty detection.

From a probabilistic perspective, clusters can be interpreted as modes of the underlying data distribution. Gaussian-based models approximate each cluster through a mean

and a covariance matrix, thereby capturing both its central tendency and variability. This view is fundamental to mixture-based approaches, where each component represents a potential latent category and the overall model can describe complex, multimodal distributions[19][53][55][56].

Recent Bayesian approaches to category discovery[52] further show that explicitly modelling class distributions in the latent space can improve both robustness and interpretability. In particular, Gaussian mixture-based frameworks can adapt the number and shape of clusters over time, making them suitable for dynamic environments where category structures evolve[21][54].

2.7. Uncertainty Modelling in Latent Spaces

2.7.1. Posterior Variance as Uncertainty

Within the VAE framework, the posterior distribution is parameterized by a mean vector μ that represents the central latent embedding and a variance vector σ^2 that provides a direct measure of uncertainty associated with the input sample.

Posterior variance serves as a *principled and interpretable measure of uncertainty* directly derived from the probabilistic formulation of the model: high posterior variance indicates that the model is uncertain about the latent representation of a given input, which may occur in the presence of ambiguous, noisy, or out-of-distribution (OOD) data. Conversely, low variance suggests that the model has high confidence in the encoded representation.

Recent works explicitly leverage this property by interpreting the variance as an uncertainty signal. For example, in multimodal representation learning, the variance is used to quantify modality-specific uncertainty and guide fusion strategies, improving robustness under missing or corrupted data conditions[20].

2.7.2. Reliability of Latent Representations

The uncertainty encoded in latent representations enhances interpretability and robustness by providing a principled criterion to evaluate the reliability of both learned representations and model predictions. In particular, high uncertainty may indicate:

- **Semantic ambiguity**: inputs may belong to multiple categories or lie near decision boundaries;
- **Outliers**: samples that do not conform to the learned data distribution;

- **Noisy or corrupted samples**: input features are unreliable or incomplete.

By incorporating uncertainty into the representation, models can distinguish between confident and uncertain predictions, enabling more robust decision-making. This is especially important in open-world and continual learning scenarios, where the model must detect and adapt to novel or evolving categories. Bayesian approaches further emphasize the importance of uncertainty-aware representations, showing that explicitly modelling distributional properties improves both category discovery and knowledge retention over time[18].

2.8. Continual Learning Fundamentals

2.8.1. Catastrophic Forgetting

A central challenge in continual learning is **catastrophic forgetting**: the tendency of the NNs to lose previously acquired knowledge when they are trained sequentially on new tasks or data distributions. In contrast to static learning settings, continual learning assumes that data arrive sequentially over time and that the model has limited or no direct access to past training samples. Consequently, the losses associated with previous data distributions become difficult to evaluate, since past data are no longer fully available. This limitation biases the optimization process toward the current task, leading to a degradation in performance on earlier tasks[31][32].

Theoretically, catastrophic forgetting can be understood as a manifestation of the stability–plasticity dilemma[33]: a continual learner must remain plastic enough to absorb new information while also being stable enough to preserve previous knowledge. Excessive plasticity causes the parameters to drift too far from earlier solutions, whereas excessive stability prevents effective adaptation to new tasks.

Recent surveys[31][32][33][56] emphasize that successful continual learning therefore requires a principled trade-off between memory retention and adaptation, rather than the maximization of either property alone. This issue is particularly relevant in modern language models, where sequential adaptation across domains, time periods, or instructions can easily disrupt previously learned capabilities.

2.8.2. Replay-Based Learning

Replay-based methods represent one of the most widely adopted strategies to **mitigate catastrophic forgetting**. The core idea consists in maintaining a memory buffer of

samples from previously observed tasks and interleaving them with data from the current task during training. This allows the model to approximate joint training over past and present data distributions, which would otherwise be inaccessible in sequential learning settings. Formally, let D_t denote the current task distribution and M_{t-1} a memory buffer containing samples from previous tasks. The training objective can be written as:

$$\mathcal{L}_{\text{replay}}(\theta) = \mathbb{E}_{(x,y)\sim D_t}\left[l(f_\theta(x), y)\right] + \lambda\, \mathbb{E}_{(x,y)\sim M_{t-1}}\left[l(f_\theta(x), y)\right] \tag{2.5}$$

where f_θ is the model parameterized by θ, l is the task loss and λ balances the influence of current and replayed data, where λ controls the balance between learning new information and preserving past knowledge.

However, since access to all past data is infeasible, replay-based methods rely on a limited buffer, resulting in a trade-off between memory capacity and knowledge retention. Classical replay strategies construct this buffer by selecting representative samples from previous tasks. While effective, this approach becomes increasingly limited as the number of tasks grows, since the buffer cannot fully capture the diversity of past data, making sample selection a critical factor for performance[33].

Recent works have revisited this paradigm by considering *more flexible memory assumptions*: when storage is not the primary constraint, the challenge shifts from deciding how much data to store to determining which data should be replayed. In this context, adaptive replay strategies propose prioritizing samples that are more susceptible to forgetting, rather than simply relying on static or uniformly sampled memory buffers[34].

2.8.3. Elastic Weight Consolidation

A second major family of continual learning methods is based on parameter regularization with **Elastic Weight Consolidation (EWC) method**, commonly derived from a Bayesian or Laplace-approximation perspective. EWC preserves previously acquired knowledge by penalizing deviations in parameter space from solutions learned in earlier tasks, constraining the parameter update trajectory to respect the importance of what the model has already learned.

After learning task $t-1$, the parameters θ^*_{t-1} are treated as the center of an approximate posterior distribution. When learning task t, the objective becomes:

$$\mathcal{L}_{\text{EWC}}(\theta) = \mathcal{L}_{D_t}(\theta) + \frac{\lambda}{2}\sum_i F_i\left(\theta_i - \theta^*_{t-1,i}\right)^2 \tag{2.6}$$

where F_i is the i-th diagonal element of the Fisher Information Matrix, θ_i is the current value of the i-th parameter, $\theta^*_{t-1,i}$ is its value after the previous task and λ is a regularization coefficient. The quadratic penalty discourages large deviations from previously important parameter values, with stronger penalties applied to parameters associated with higher Fisher scores.

The intuition behind EWC is that not all model parameters contribute equally to previously learned tasks: if some parameters are particularly important for past knowledge, leading to a substantial deterioration in the likelihood of previously observed data, then changes to those parameters should be penalized more strongly during the learning of new tasks.

The **Fisher Information Matrix** plays a central role by serving as a proxy for parameter importance. In practice, a diagonal approximation is typically adopted to reduce computational complexity, thereby making EWC scalable to a wide range of neural architectures[35] (Figure 2-4).

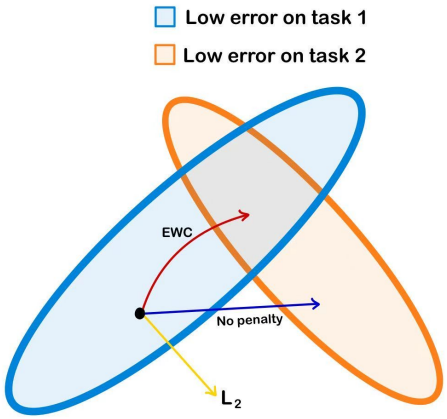

Figure 2.4: EWC regularization, penalizing changes to important parameters based on Fisher information to prevent catastrophic forgetting. Adapted from [36].

Compared with replay-based learning, EWC is more memory-efficient because it does not require storing raw historical samples. However, this advantage comes with a trade-off: the old task distributions are represented only indirectly through parameter importance estimates, which may be insufficient when task shifts are large or highly non-stationary. For this reason, recent continual learning systems often combine regularization with replay or other mechanisms[34][35].

2.9. Chapter Summary

The chapter establishes a coherent theoretical foundation focusing on representation learning, probabilistic modelling and continual learning within a unified framework for open-world intent discovery.

It begins by positioning sentence embeddings as the core of modern NLP, highlighting the transition from sparse representations to dense semantic spaces where meaning is encoded geometrically.

Limitations of high-dimensional embeddings is addressed through PCA and whitening, introduced as principled techniques to improve numerical stability while preserving the most informative structure of the data.

A probabilistic perspective is developed through latent variable models, with a focus on VAE that learns structured latent spaces, while the β-VAE formulation balances reconstruction and regularization, promoting disentangled representations and posterior variance emerges as an explicit and interpretable measure of uncertainty.

The GMMs provide a flexible approach to modelling multi-modal data and performing soft clustering. Through the EM algorithm, latent structure and parameters are iteratively refined and clustering in latent space enables the identification of categories as density modes. Uncertainty modelling plays a central role in open-world settings by enabling the system to distinguish between confident and ambiguous representations, improving robustness to noise, outliers and unseen inputs. Finally, these components are framed within continual learning, where catastrophic forgetting is addressed through replay-based methods and EWC.

The chapter outlines a unified perspective in which semantic representation, probabilistic latent modelling, uncertainty estimation and continual adaptation jointly support dynamic and reliable learning in evolving environments.

3 | Framework Architecture

3.1. System Overview

The proposed framework addresses **new intent discovery in open-world settings**, where the label space evolves dynamically and previously unseen intents may emerge after deployment. To tackle this problem, the system combines *probabilistic representation learning, uncertainty-aware decision mechanisms, density-based clustering and continual learning strategies* within a unified multi-phase pipeline designed to support both the discovery of novel intents and the retention of previously acquired knowledge.

Table 3-1 summarizes the main research objectives of the proposed framework and links each objective to the corresponding methodological approach, implemented components and expected benefits. The table provides a compact overview of how the framework addresses the key challenges of open-world continual new intent discovery, including the separation between known and unknown samples, the discovery of emerging intents, the preservation of previously acquired knowledge and the progressive expansion of the intent space. In this way, it clarifies the role of each module within the overall architecture and shows how probabilistic representation learning, uncertainty-aware filtering, density-based discovery and continual learning mechanisms jointly contribute to the proposed solution.

The system operates as a tightly coupled self-regulating architecture and is organized into four main components that continuously interact and influence each other:

1. **Semantic Representation Layer**: raw utterances are encoded into high-dimensional semantic embeddings using a pretrained sentence encoder.

2. **Latent Representation Learning Module**: a β-VAE maps embeddings into a compact probabilistic latent space with an adaptive β mechanism.

3. **Uncertainty-Aware Decision Layer**: posterior variance is used to regulate downstream operations through adaptive σ-thresholding.

4. **Density-Based Discovery Model**: a Dirichlet Process Gaussian Mixture Model (DP-GMM) enables automatic discovery of emerging intent clusters. The model operates

on reliability-filtered latent representations and leverages a truncated variational formulation to support adaptive and scalable clustering. This design allows uncertainty to actively control the flow of information across modules, ensures robust cluster formation through hard filtering and supports stable expansion of the intent space via the separation between known and novel distributions.

Table 3.1: Summary of Research Objectives, Methods and Expected Benefits in the Proposed Framework

Research Objective	**Methodological Approach**	**Implemented Components**	**Expected Benefits**
Continual new intent discovery in open-world settings	Probabilistic latent representation learning	β-VAE, transformer-based sentence embeddings	Robust intent representations enriched with uncertainty information
Reliable separation of known and unknown samples	Uncertainty-aware filtering mechanism	Posterior variance estimation, adaptive σ-quantile thresholding, trusted-sample masking	Improved filtering of ambiguous samples and reduced propagation of noisy representations
Automatic discovery of emerging intents	Bayesian nonparametric density modelling	DP-GMM, variational inference	Flexible identification of novel intent clusters without requiring a predefined number of classes
Stability across sequential learning phases	Continual learning regularization	Replay buffer with hard/easy sampling, EWC	Reduced catastrophic forgetting and more stable retention of previously acquired knowledge
Progressive expansion of the intent space	Multi-phase continual learning framework	Phase-wise training for known-intent learning, novelty screening, consolidation and label-space expansion	Controlled integration of newly discovered intents over time
Reliable cluster formation under uncertainty	Uncertainty-based sample selection	σ-based filtering, high-confidence pseudo-label assignment	Improved cluster purity and greater robustness in novel intent discovery

3.2. Multi-Phase Continual Learning Pipeline

The proposed framework adopts a multi-phase continual learning pipeline through **three phases**, that enables the model to progressively expand its knowledge of the intent space

while preserving previously acquired information. The core objective of this framework is to **balance plasticity** (ability to incorporate new intents) **with stability** (retention of past knowledge).

In the experimental implementation, the **continual setting** is simulated by partitioning the available intent space into two subsets: **50 initially known intents** and **50 novel intents**. The framework is organized into three sequential phases. In Phase 1, the model is trained in a supervised manner using labelled samples belonging to the initial set of known intents. In Phase 2, data associated with the novel intent set are introduced within an unlabelled incoming batch, together with samples from previously known intents and the framework performs novelty detection and discovery. In Phase 3, the accepted discoveries are consolidated through retraining, replay and elastic weight consolidation (EWC), allowing the model to expand its knowledge while preserving previously acquired information.

3.2.1. Phase 1: Known Intent Learning

The first phase *establishes the initial semantic structure of the model* by learning from a labelled dataset: $D_0 = \{(x_i, y_i)\}_{i=1}^{N_0}$ containing a predefined set of 50 known intents K_0. The training data consists of supervised utterance–label pairs $\{(x_i, y_i)\}$, with $y_i \in K_0$, where each input x is first encoded into a semantic embedding and subsequently mapped into a probabilistic latent space of dimension $z_{\text{dim}} = 64$ through a variational encoder:

$$z \sim q_\phi(z \mid x_i) = \mathcal{N}(\mu(x_i), \sigma^2(x_i)) \tag{3.1}$$

This latent representation, learned via a β-VAE, provides a compact and structured embedding space in which semantically similar utterances are clustered together. On top of this representation, a classifier is trained to assign each sample to one of the known intents:

$$\hat{y} = \arg\max_{k \in K_0} p(y = k \mid z) \tag{3.2}$$

Additionally, a subset of representative samples is selected and stored in a replay buffer $M \subset D_0$, which will be reused in subsequent phases to preserve previously acquired knowledge. Beyond standard supervised learning, this phase plays a crucial preparatory role for subsequent adaptation. In particular, mechanisms are introduced to *mitigate catstrophic forgetting*, ensuring that the learned latent space remains stable and reusable when the system evolves, as formalized in Chapter 4.

3.2.2. Phase 2: Novelty Detection and Discovery

The second phase shifts **from supervised to partially unsupervised**, processing a stream of **unlabelled data** $D_t^{\text{unlabeled}} = \{x_j\}$ drawn from evolving distributions, which contain both known and new intents and the proportion of new intents is unknown and may vary over time. The primary challenge is *identifying which samples belong to the known intent space and detecting those that correspond to novel semantic patterns.* To address this, the model evaluates the confidence of its predictions over the known classes and samples for which no class achieves sufficiently high posterior probability are considered candidates for novelty:

$$\max_{k \in K_0} p(y = k \mid z) < \tau_{\text{conf}} \tag{3.3}$$

However, in ambiguous regions of the latent space, the confidence alone is insufficient. For this reason, the framework leverages the uncertainty encoded in the latent representation, quantified by the variance σ_i^2 and high-uncertainty samples are treated cautiously, reducing their influence on downstream decisions.

The framework adopts a *DP-GMM* (see Section 3.5) to model the distribution of latent representations, allowing the number of clusters to grow adaptively and enabling the discovery of new intent groups without prior specification. At this stage, the system **identifies coherent clusters** likely corresponding to emerging intents, which are not yet incorporated into the classifier. The final decision process is formulated as a **multi-criteria decision mechanism**, relying on the joint evaluation of multiple probabilistic signals (classifier confidence, uncertainty-based filtering and density-based likelihood estimation) integrated within a unified pipeline to robustly distinguish between known and novel samples, reducing ambiguity in boundary regions of the latent space.

3.2.3. Phase 3: Consolidation and Expansion

In the final phase, the **model is retrained** using a composite training set $D_t^{\text{train}} = D_0 \cup \tilde{D}_t^{\text{new}} \cup M$, that integrates multiple data sources:

- Original labelled data from known intents D_0;
- Pseudo-labelled samples from newly discovered and validated clusters $\tilde{D}_t^{\text{new}}$;
- Replay samples stored in memory M.

The latent distribution $q_\phi(z \mid x_i)$ is updated to incorporate new clusters that are promoted

to new intent classes and the uncertainty threshold τ_σ is recalibrated based on the updated variance statistics. **The intent set is expanded as**: $K_{t+1} = K_t \cup K_{\text{new}}$. This process enables the system to *continuously integrate newly discovered intents while maintaining stability on previously learned knowledge*. During this phase, cluster assignments may be noisy; therefore, their contribution to training is modulated through uncertainty estimates, reducing the influence of unreliable samples. The model is then jointly optimized on labelled and pseudo-labelled data, incorporating EWC regularization and replay-based learning, whose implementation is described in Chapter 4.

3.2.4. Three-fases Process Overview

The three phases collectively establish a *cyclic and progressive learning paradigm*, enabling the model to alternate between supervised learning, uncertainty-guided novelty detection and incremental expansion of the intent space through knowledge consolidation (Figure 3-1), (see Appendix B-1 for pseudocodes).

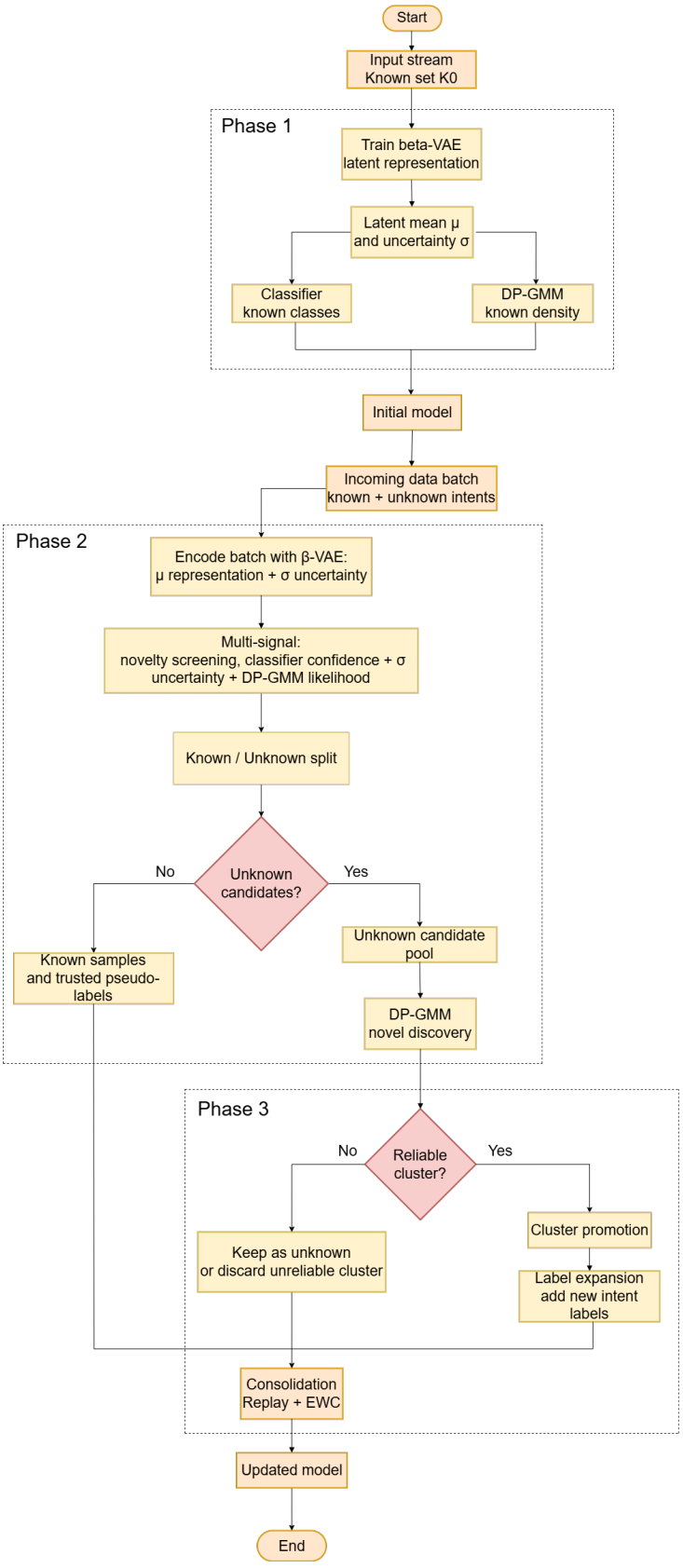

Figure 3.1: three-phase continual learning pipeline for representation learning, novelty detection and label space expansion.

3.2.5. Extended Five-Stage Continual Learning Architecture

Building upon the three-phase continual learning framework, the architecture was extended to more closely represent the progressive evolution of a **real-world dialogue system**. This configuration enables the model to encounter new intent groups progressively, thereby simulating a dialogue environment in which the semantic space evolves over time as new user requirements or interaction patterns arise.

The methodological structure remains unchanged: **Phase 1** is executed only once at the beginning of the learning process to establish the initial knowledge base, whereas **Phases 2 and 3** preserve their original functions and are repeated consecutively for each incoming batch of novel intents. Specifically, the model is initialized on a set of $|K_0| = 50$ known intents and subsequently processes two novel-intent batches, with $|N_1| = |N_2| = 50$.

The resulting five-stage architecture therefore consists of one initial supervised learning stage followed by two novelty detection–consolidation cycles, enabling controlled expansion of the intent space while preserving previously acquired knowledge (Figure 3-2), (see Appendix B-2 for pseudocodes).

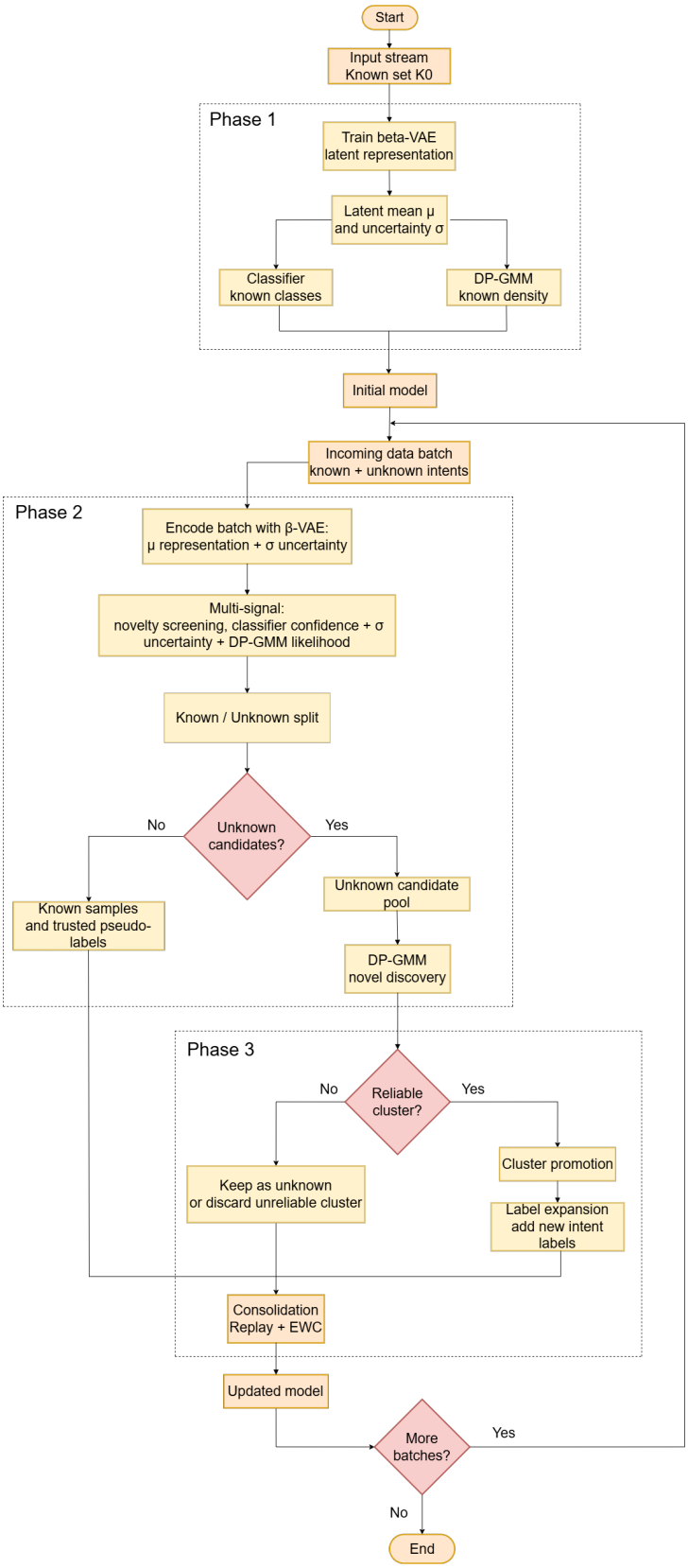

Figure 3.2: Multi-phase continual learning pipeline for representation learning, novelty detection and label space expansion.

3.3. Latent Representation Learning

A central component of the framework is a **lightweight β-Variational Autoencoder (β-VAE)**, which maps high-dimensional semantic embeddings into a compact probabilistic latent space, representing each input as a distribution parameterised by a mean and a variance rather than a deterministic point. The **latent mean** captures the stable semantic position of the sample, while the **posterior variance** provides an explicit estimate of sample-level reliability. This choice is motivated by the need to obtain a compressed representation and an uncertainty signal that, as described above, functions as a system-level control mechanism across the discovery pipeline.

Given an input embedding x, the *encoder* produces the parameters of a diagonal Gaussian posterior distribution:

$$q_\phi(z \mid x) = \mathcal{N}\left(\mu(x), \sigma^2(x)\right) \tag{3.4}$$

Latent samples are obtained using the reparameterization trick: $z = \mu + \sigma \odot \epsilon$, $\epsilon \sim \mathcal{N}(0, I)$.

The model is *trained* by maximizing the **Evidence Lower Bound (ELBO)**:

$$\mathcal{L}_{\text{VAE}} = \frac{1}{N_0} \sum_{i=1}^{N_0} \left[\mathbb{E}_{q_\phi(z|x_i)} \log p_\theta(x_i \mid z) - \beta_{\text{eff}} D_{\text{KL}}\left(q_\phi(z \mid x_i) \parallel p(z)\right)\right], \tag{3.5}$$

where the prior is defined as $p(z) = \mathcal{N}(0, I)$.

The coefficient β controls the *trade-off between reconstruction fidelity and latent regularization*. The reconstruction term encourages the model to preserve the semantic information contained in the original embedding, while the KL divergence term regularizes the posterior distribution by pushing it toward a structured standard Gaussian prior. If this regularization is too weak, the latent space may become noisy or poorly structured; if it is too strong, the representation may lose semantic expressiveness. For this reason, the framework adopts an **adaptive regularization strategy** based on an effective coefficient β_{eff}, (Figure 3-3).

More specifically, the adaptive regularization is controlled through a *KL reference level* γ. At each training step, the framework computes the average KL divergence over the current mini-batch:

$$D_{\text{KL}} = \frac{1}{B} \sum_{i=1}^{B} D_{\text{KL}}\left(q_\phi(z \mid x_i) \parallel p(z)\right) \tag{3.6}$$

The coefficient γ defines a *target reference level* for the KL divergence. If the current KL divergence moves away from this target, β_{eff} increases, encouraging a more compact and regularized latent posterior, intead when the KL divergence is close to the target, the regularization remains bounded, preventing the KL term from dominating reconstruction. Compared with a fixed-β formulation, this adaptive strategy reduces the risk of both posterior collapse and excessive latent dispersion, while preserving the regularity required for downstream density estimation and clustering (see Section 3.5). This balance is especially important in a continual learning setting, where the representation must remain stable as new intents are progressively introduced.

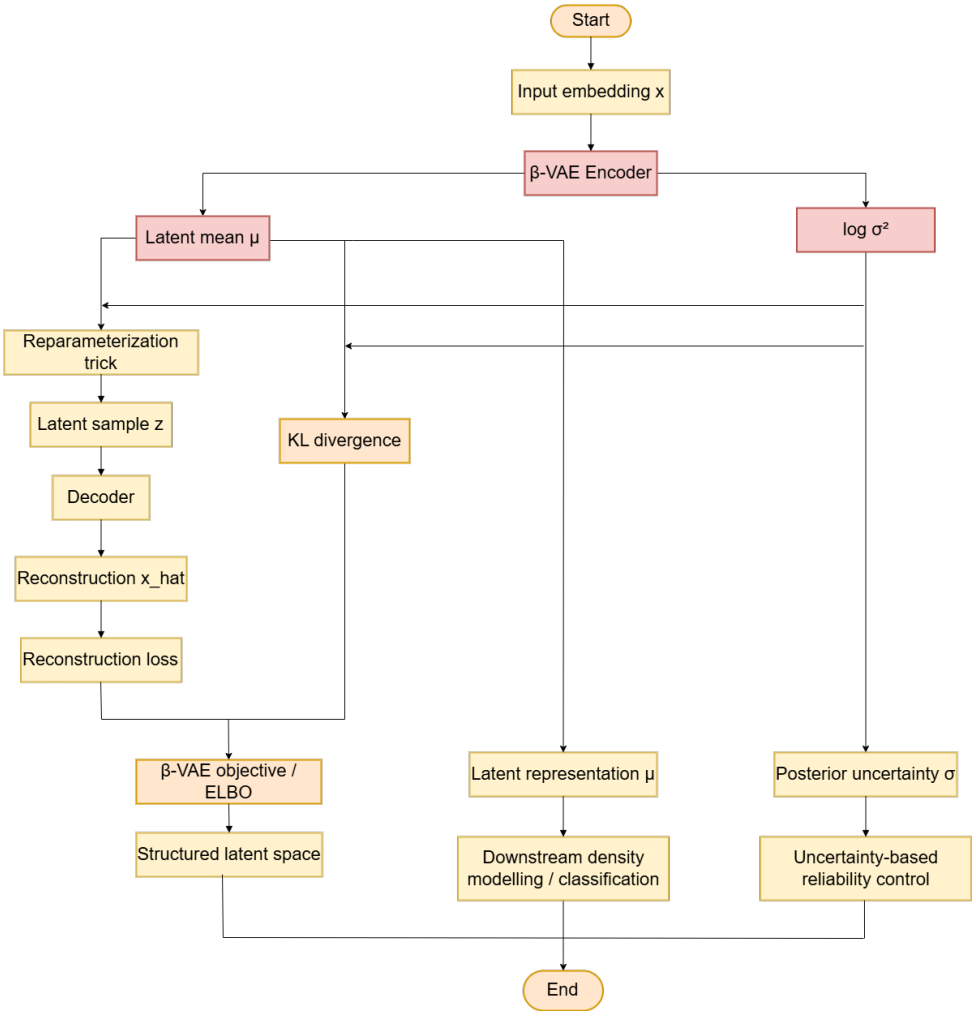

Figure 3.3: Uncertainty-aware β-VAE architecture for probabilistic latent representation and uncertainty-driven filtering.

In the proposed multi-phase pipeline, the β-VAE is first *initialized during Phase 1* using labelled samples from the initial known intents. This phase defines the reference latent

manifold of the known-intent space. In subsequent phases, the model is continually updated together with the classifier using the current data stream and the replay buffer, while EWC constrains changes to parameters that are important for previously learned intents. This allows the latent space to *adapt to emerging intents* without drifting excessively from the structure learned in Phase 1.

3.4. Uncertainty-Aware Modelling

Building upon the probabilistic latent representation introduced in Section 3.3, the framework associates each encoded sample with an uncertainty value derived from the posterior standard deviation of the β-VAE.

For each sample i, a scalar uncertainty score is computed as the average posterior standard deviation across latent dimensions. Since the encoder outputs the logarithm of the variance, the standard deviation is recovered as $\sigma_{ij} = \exp\left(\frac{1}{2}\log\sigma_{ij}^2\right)$, leading to:

$$\sigma_i = \frac{1}{d_z}\sum_{j=1}^{d_z}\exp\left(\frac{1}{2}\log\sigma_{ij}^2\right) \tag{3.7}$$

In the proposed framework σ_i is treated as an **operational control signal** that regulates how each sample is allowed to influence the continual learning process, indicating how stable and trustworthy that position is. Low values of σ_i suggest that the encoder maps the input to a compact and reliable latent region, whereas high values indicate a more dispersed, ambiguous or less dependable representation.

This reliability criterion is implemented through an **adaptive quantile-based threshold** τ_σ, computed from the empirical distribution of posterior uncertainties. The quantile is updated across phases using an adaptation rate η, which controls how quickly the threshold reacts to deviations from the target trusted-sample ratio:

$$q_t = q_{t-1} + \eta\,(r^\star - r_t) \tag{3.8}$$

where q_t is the uncertainty quantile at phase t, $r^\star$ is the target trusted-sample ratio and r_t is the observed trusted ratio.

A sample is considered trusted when: $\sigma_i \leq \tau_\sigma$. Since τ_σ is updated from the current uncertainty distribution, reliability is not fixed a priori, but remains a **dynamic criterion** that is recalibrated across continual phases as the latent space evolves. This mechanism is distinct from the Fisher smoothing coefficient α, which is introduced later in the EWC

module to regulate the retention of parameter-importance information across phases.

Uncertainty is *initialized* during Phase 1, when the β-VAE is first trained on the labelled samples of the initial known intents. At this stage, the posterior standard deviation provides the first reliability scale over the known latent space and is used to calibrate the initial trusted-sample threshold. In the following continual phases, uncertainty values are recomputed whenever new data and replay samples are encoded by the updated model.

The influence of σ_i extends across multiple components of the framework, where it also helps to regulate the **stability–plasticity trade-off**:

- **Density model calibration**: the DP-GMM is fitted primarily on trusted samples, ensuring that the estimated known latent structure is not dominated by noisy or unstable embeddings;
- **Pseudo-labelling**: low-uncertainty samples are preferred candidates for receiving and propagating pseudo-labels, while high-uncertainty samples are prevented from reinforcing potentially incorrect decisions;
- **Novelty analysis**: uncertainty provides one of the reliability signals combined with classifier confidence and density-based evidence to identify samples whose latent position is weakly supported by the current model;
- **Replay buffer**: σ_i structures sample retention by balancing reliable low-uncertainty examples with harder high-uncertainty cases, preserving stable knowledge while exposing the model to ambiguous regions of the latent space;
- **Consolidation**: limiting the influence of unstable representations on classifier updates, density estimation and label-space expansion reduces the risk of error propagation.

3.5. Density-Based Discovery Model

Building upon the uncertainty-aware latent representation introduced in Section 3.4, novel intent discovery is performed through probabilistic density modelling of the latent space by means of a **Dirichlet Process Gaussian Mixture Model (DP-GMM)**. Its role is to estimate where reliable known-intent representations lie, how strongly a new sample is compatible with this known structure and whether groups of uncertain samples form sufficiently dense regions to justify the creation of new intent labels. In other words, the mixture model provides both soft assignments and density scores, which are later used

for known-intent modelling, novelty detection and cluster promotion.

Each utterance is encoded by the β-VAE into a latent Gaussian representation, then to improve numerical conditioning and reduce redundancy before mixture fitting, the latent vectors are further projected through a **PCA transformation** that retains *95% of the variance*. This step is important because Gaussian mixture estimation can become unstable in redundant or high-dimensional latent spaces, especially when covariance matrices must be estimated from a limited number of trusted samples.

The resulting reduced representation is modelled through a **Bayesian Gaussian Mixture**, which implements a truncated approximation of a non-parametric Dirichlet Process mixture:

$$p(z) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(z \mid \mu_k, \Sigma_k\right), \tag{3.9}$$

where π_k, μ_k and Σ_k denote, respectively, the mixture weights, component means and covariance matrices. Each Gaussian component can be interpreted as a local density region in the latent space: the mean μ_k represents the centre of that region, the covariance Σ_k describes its spread and orientation and the weight π_k reflects its relative importance in the mixture. Therefore, the model can evaluate which component is most likely to explain a sample and how likely the sample is under the learned distribution.

Parameter estimation in Gaussian mixture models is classically performed through the **Expectation-Maximization (EM) algorithm**, which alternates between estimating latent component responsibilities and updating the mixture parameters. In the classical GMM setting, EM monotonically increases the observed-data log-likelihood at each iteration, although convergence is guaranteed only to a local optimum. In this pipeline, these responsibilities are useful because they provide a soft association between latent samples and mixture components rather than forcing a rigid assignment. This is particularly important in open-world intent discovery, where ambiguous samples may lie between known regions and emerging novel structures.

The use of a truncated approximation of a non-parametric Dirichlet Process mixture is motivated by the need to balance modelling flexibility and computational tractability. By imposing an upper bound on the number of components, the model remains efficient to train while still retaining the ability to adaptively allocate probability mass across components, while avoiding the need to predefine the effective number of active clusters. This property is particularly important in continual new intent discovery, where the underlying latent structure may evolve as new intents emerge over time. In the implementation, the upper bound is set to 200 components for the known-intent latent density model and to

100 components for the novelty discovery module (see Appendix B-3 for pseudocode).

The role of the DP-GMM evolves across the learning phases. During **Phase 1**, the framework observes only labelled known intents and the DP-GMM is fitted on their trusted latent representations, establishing the reference density against which future samples are compared. Once fitted, the known-intent mixture provides both component assignments and per-sample log-likelihood scores:

- component assignments allow the framework to relate local density regions to known intents;
- log-likelihood scores quantify how well each sample is explained by the learned known-intent distribution.

Each component is associated with an intent label by majority assignment over the trusted labelled samples belonging to that component, yielding a probabilistic cluster-to-intent correspondence. The framework then derives **adaptive decision thresholds** from the empirical log-likelihood distribution of trusted samples using robust statistics, namely the **median** and the **median absolute deviation (MAD)**.

In particular, let:

$$m_\ell = \text{median}\left(\{\ell_i\}_{i \in \mathcal{D}_{\text{trusted}}}\right) \tag{3.10}$$

be the median trusted-sample log-likelihood and let:

$$\text{MAD}_\ell = \text{median}\left(|\ell_i - m_\ell|\right) \tag{3.11}$$

denote the corresponding median absolute deviation. The soft and strict likelihood cutoffs are then defined as:

$$\tau_\ell^{\text{soft}} = m_\ell - \alpha_{\text{soft}}\,\text{MAD}_\ell, \qquad \tau_\ell^{\text{strict}} = m_\ell - \alpha_{\text{strict}}\,\text{MAD}_\ell, \tag{3.12}$$

where $\alpha_{\text{strict}} > \alpha_{\text{soft}} > 0$. This produces soft and strict cutoffs that characterize compatibility with the learned known distribution. These thresholds are reused in later phases to determine whether incoming samples remain within the known latent structure or should be considered candidates for novelty analysis.

In the **subsequent continual phases**, the DP-GMM is used to model the density structure of the evolving latent space and to distinguish regions that remain compatible with known intents from regions that may correspond to emerging intents. Samples assigned high likelihood are interpreted as belonging to dense areas already captured by the mix-

ture, whereas low-likelihood samples are treated as candidates for novelty analysis. However, likelihood alone is not sufficient to justify the creation of a new intent: it is combined with classifier confidence and posterior uncertainty so that novelty decisions rely on multiple complementary signals. In this setting, the DP-GMM provides a probabilistic indication of where the current latent model is no longer sufficient to explain the incoming data. For samples identified as candidate unknowns, a separate Bayesian mixture is fitted to their latent representations in order to assess whether these low-density regions exhibit coherent internal structure. The purpose of this second mixture is to identify compact and repeatable clusters within the candidate unknown space. Candidate clusters are then evaluated according to their support and average log-likelihood, ensuring that only dense and statistically reliable structures are eligible for promotion. In this way, the DP-GMM supports controlled discovery by separating meaningful latent aggregations from isolated outliers or unstable fragments.

This dual-mixture design reinforces the reliability-oriented nature of the framework, allowing the label space to expand only when candidate unknown regions form coherent, sufficiently supported and statistically reliable latent structures.

3.6. Chapter Summary

This chapter presents a unified framework for continual intent discovery in open-world environments, where the label space evolves over time.

The framework is first operationalized through a three-phase continual learning pipeline. **Phase 1** initializes the latent representation, classifier and known-intent density model using 50 labelled base intents. **Phase 2** then detects and discovers candidate novel intents from a mixed known–unknown batch containing 50 novel intents, while **Phase 3** consolidates the accepted discoveries through label-space expansion, replay and regularization. To better approximate a real-world scenario in which novel intents emerge over time, this architecture is subsequently extended to a five-stage setting: Phase 1 is performed only once, whereas Phases 2 and 3 are repeated for each of two sequential batches of 50 novel intents.

The model is built on a probabilistic latent space learned through a β-VAE, where each input is encoded by both its semantic representation and an associated uncertainty measure. This design enables the system to distinguish between reliable and ambiguous regions of the latent space. Uncertainty acts as a control variable that directly influences data selection and downstream decisions.

The uncertainty-aware decision layer regulates information flow through adaptive σ-based thresholding, ensuring that only high-confidence samples contribute to modelling the known intent space, reducing noise propagation and stabilizing learning. In parallel, a density-based discovery model based on a truncated DP-GMM enables the identification of emerging intent clusters without predefined assumptions on their number.

The framework introduces a principled approach to open-world intent discovery by combining probabilistic representations, uncertainty-driven control and nonparametric density modelling within a continual learning setting. This chapter therefore establishes the methodological foundation upon which the implementation details and experimental evaluation presented in the following chapters are built.

4 | Decision and Continual Learning in Dynamic Intent Spaces

4.1. Novelty Detection Strategy

In an open-world setting, the system is continuously exposed to incoming utterances whose semantic origin is inherently uncertain. Novelty detection therefore constitutes a central component of the framework, requiring a decision mechanism capable of operating under evolving data distributions while preserving the structure of previously learned intents. The proposed approach adopts a **multi-signal strategy** in which classifier confidence, latent uncertainty and density-based evidence are sequentially integrated.

Classifier confidence, derived from the softmax posterior of the classification head, provides the first discriminative signal in the novelty detection strategy. It measures how strongly the current classifier assigns a sample to one of the intents already present in the current label space. High confidence suggests that the sample is compatible with an existing decision region, while low confidence indicates that the classifier cannot reliably associate the sample with any known or previously discovered intent. However, classifier confidence alone is not sufficient for novelty detection, since ambiguous known samples and truly novel samples can both produce low-confidence predictions.

The second signal is the **latent uncertainty score**, defined in Section 3.4. In this stage, uncertainty is not reintroduced as a separate modelling component, but is used operationally to assess whether the latent representation of a sample is reliable enough to support downstream decisions. In particular, the adaptive trusted-sample filtering mechanism selects low-uncertainty labelled samples to estimate the known-intent latent distribution. At phase t, the trusted subset is defined as:

$$T_t = \{i \in L_t \ : \ \sigma_i < Q_{q_t}(\sigma)\}\,, \tag{4.1}$$

where $Q_{q_t}(\sigma)$ denotes the empirical quantile of the current uncertainty distribution. The quantile is initialized at $q_0 = 0.90$ and then adjusted according to the discrepancy between the observed trusted ratio and the target trusted ratio. This makes the reliability threshold adaptive across continual phases, rather than fixed a priori.

The third signal is the **DP-GMM log-likelihood**, defined in Section 3.5. After the trusted subset has been selected, the known-intent latent structure is modelled by fitting a Dirichlet Process Gaussian Mixture Model on the corresponding latent mean representations. Therefore, the log-likelihood used in the decision pipeline is not obtained directly from the classifier or from the β-VAE loss, but from the fitted DP-GMM density model. For a latent representation z_i, the score is computed as:

$$\log p(z_i \mid \Theta_t) = \log \sum_{k=1}^{K_t} \pi_k \mathcal{N}\left(z_i \mid \mu_k, \Sigma_k\right), \tag{4.2}$$

where $\Theta_t = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K_t}$ are the mixture parameters learned from trusted known samples. A high DP-GMM log-likelihood indicates that the sample lies in a dense region compatible with the known latent distribution, whereas a low log-likelihood suggests that the sample falls in a low-density or poorly explained region and may therefore correspond to a novelty candidate (Appendix B-4 for pseudocode).

The overall decision mechanism follows a structured multi-stage pipeline (Figure 4-1):

1. **Classifier confidence**: low confidence provides an initial indication of potential novelty;
2. **Latent uncertainty**: unreliable representations are filtered out to prevent degradation of the density estimation;
3. **Density-based likelihood**: remaining samples are evaluated to determine consistency with the known latent distribution.

By combining these signals, the framework avoids interpreting every low-confidence prediction as novelty. Novelty detection is triggered only when three conditions are simultaneously satisfied: low classifier confidence, high latent uncertainty and low DP-GMM likelihood under the known-intent density model. This strict agreement requirement gives the decision mechanism a conservative behaviour, since novelty is admitted only when discriminative, uncertainty-based and density-based evidence jointly provide weak support for the known-intent hypothesis.

A related design choice is the adoption of a **non-forced assignment strategy**, imple-

mented by setting the fallback mechanism to “OFF”. As a result, samples are not forced toward the closest known component when the available evidence is insufficient. They are assigned to known intents only when the active decision signals provide adequate support; otherwise, they remain unassigned and are passed as novelty candidates to the discovery module described in Section 3.5.

In that stage, clustering and promotion mechanisms evaluate whether these candidates form sufficiently dense and coherent structures in the latent space. Cluster promotion is driven by minimum-support and likelihood-based quality criteria, while uncertainty contributes indirectly by ensuring that the known-density model is estimated from reliable representations. This design supports robust novelty identification while preserving the stability of previously learned representations, providing the operational foundation for the continual learning mechanisms introduced in the subsequent sections.

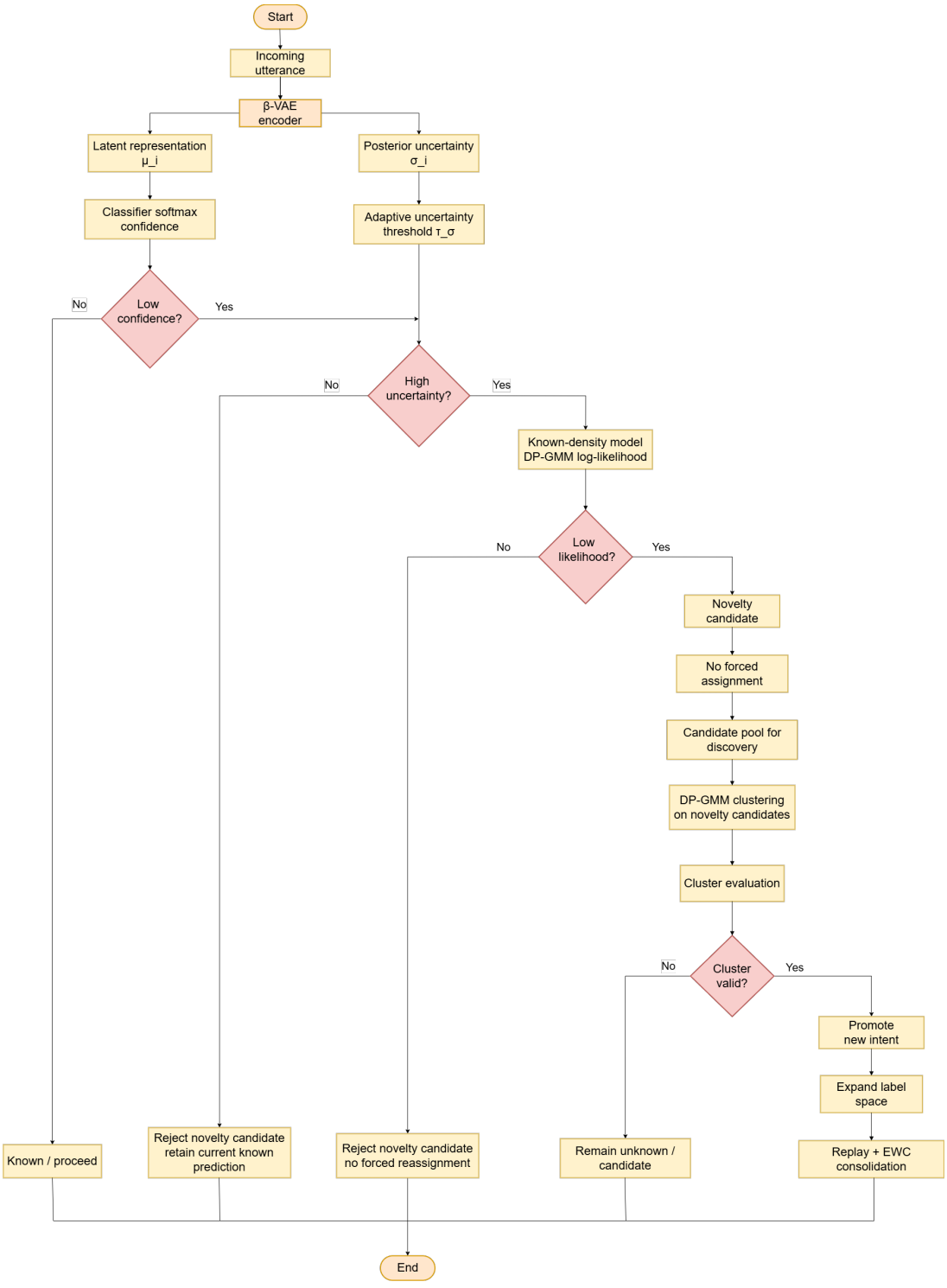

Figure 4.1: Novelty detection strategy combining uncertainty estimation and likelihood scoring to identify and validate new intent candidates.

4.2. Continual Stability Mechanisms

Although the experiments are built from a single benchmark dataset, the continual setting is simulated by reorganizing the data into sequential learning phases. The first phase contains labelled samples from the initial known intents, while the following phases receive new batches containing previously seen intents together with novel intents whose labels are masked during training. This simulates a temporal scenario in which new user intents progressively appear over time.

Under this setting, the model must learn from newly arriving batches and expand its label space without overwriting previously acquired representations. This problem, known as catastrophic forgetting, is addressed through two complementary stability mechanisms: a structured replay buffer acting at the data-distribution level and Elastic Weight Consolidation (EWC) acting at the parameter level.

4.2.1. Replay Buffer

The replay buffer serves as an **explicit memory of past observations**, ensuring that previously learned intents remain represented during subsequent training phases. At each phase, samples from the current data stream are first added to the memory together with previously stored instances, including labelled samples, pseudo-labelled samples and newly discovered intents. Replay selection is therefore not applied before new discoveries enter the memory; rather, it is applied only when the buffer exceeds its fixed capacity. This means that newly discovered samples do not need to be selected as hard samples in order to be retained. If they exhibit low posterior uncertainty after discovery and consolidation, they may be preserved through the easy subset of the replay buffer. If, in later phases, their latent representations become less stable due to distributional drift or interference with newly arriving intents, their uncertainty may increase and they may instead be selected as hard replay samples.

In the experimental configuration, the replay buffer is constrained to a fixed size of 5000 samples and updated through a hybrid uncertainty-aware selection strategy. Replay explicitly leverages the latent uncertainty score σ produced by the β-VAE to retain both informative and representative past samples. When the buffer exceeds its maximum capacity, stored samples are ranked according to their latent uncertainty and reduced to 5000 elements by keeping 4750 high-uncertainty samples and 250 low-uncertainty samples, corresponding to a hard/easy replay ratio of 0.95 / 0.05. Let D denote the available

pool of past samples, the replay buffer R is constructed as:

$$|R_{\text{hard}}| = \rho M, \qquad |R_{\text{easy}}| = (1-\rho)M, \tag{4.3}$$

where M is the buffer size. Easy samples correspond to low-uncertainty observations, acting as anchors that preserve the core structure of previously acquired knowledge: $R_{\text{easy}} = \{i \in D : \sigma_i \leq \tau_{\text{low}}\}$, and represent stable, high-confidence regions of the learned intent space. Hard samples are selected from higher-uncertainty regions: $R_{\text{hard}} = \{i \in D : \sigma_i > \tau_{\text{low}}\}$, and correspond to boundary or ambiguous cases. They are highly informative for preserving decision boundaries and preventing over-simplification of the model. The replay ratio is therefore skewed toward hard samples, while still retaining a smaller subset of easy examples to preserve a stable representation of previously learned intents. By combining informative hard samples with representative easy samples, the buffer supports both local decision sensitivity and global consistency, helping the model retain previous intents while adapting to new ones.

However, this mechanism should be interpreted as a mitigation strategy rather than as a formal guarantee of per-intent retention. Since replay selection is performed globally according to posterior uncertainty, older discovered clusters that become very stable may be underrepresented when the buffer is full, especially under a high hard-sample ratio. In the implemented framework, this risk is partially mitigated by the inclusion of easy samples, by the repeated consolidation of discovered intents and by the complementary use of EWC. Nevertheless, an explicit stratified replay policy, reserving a minimum quota for each known or discovered intent before applying uncertainty-based hard/easy selection within each group, would provide a stronger guarantee against forgetting in longer continual streams.

4.2.2. Elastic Weight Consolidation (EWC)

EWC is a **regularization-based continual learning method** that constrains updates of important parameters of previously learned knowledge. In the proposed framework, after each learning phase, the model estimates the importance of the parameters of the β-VAE encoder and classifier through an approximation of the Fisher Information Matrix. Parameters associated with high Fisher values are interpreted as more relevant for retaining previously learned intent representations. During the following phases, EWC adds a quadratic penalty to the training objective, discouraging these important parameters

from moving too far from their previous optimal values:

$$\mathcal{L}_{\text{EWC}}(\theta) = \frac{\lambda_{\text{EWC}}}{2} \sum_i F_i \left(\theta_i - \theta_i^*\right)^2, \tag{4.4}$$

where θ_i denotes the current value of parameter i, θ_i^* is the value stored after the previous phase, F_i is the corresponding diagonal Fisher Information estimate and $\lambda_{\text{EWC}} = 7500$ controls the strength of the regularization. In this way, EWC does not freeze the model entirely, but introduces a parameter-level resistance against changes that would damage previously consolidated intents. This is particularly important in the multi-phase continual NID setting, where the model must adapt to newly arriving and potentially novel intents while preserving the latent structure and classification boundaries learned from earlier phases.

4.2.3. Joint Effect: Data-Level and Parameter-Level Stability

The combination of replay and EWC provides a twofold **stabilization mechanism**:

- Replay operates at the data level by reintroducing past samples and preserving the empirical distribution of known intents.
- EWC operates at the parameter level by constraining the optimization trajectory in regions of the parameter space that are critical for past performance.

This dual strategy is essential in the context of continual intent discovery and its integration ensures a balanced trade-off between stability and plasticity, allowing the framework to expand the intent space while maintaining consistency with previously learned knowledge (see Appendix B-5 for pseudocode).

4.3. Label Expansion and Cluster Promotion

A key property of the framework is that novelty detection serves as an intermediate mechanism rather than an end goal. Detected patterns are evaluated and, when reliable, incorporated as new semantic classes. Through cluster validation, label creation and classifier expansion, the model converts previously unseen patterns into stable components. This **dynamic extension of the label space** enables genuine open-world intent discovery, allowing the set of semantic categories to evolve with incoming data (see Appendix B-6 for pseudocode).

4.3.1. When a Cluster Becomes a New Intent

Clusters obtained from the discovery DP-GMM are not automatically accepted as new intents: promotion is governed by a **set of stability criteria** designed to ensure that only statistically meaningful and semantically coherent structures are retained.

First, a **minimum support constraint** is imposed to avoid promoting spurious clusters formed by noise or isolated samples. Let C_k denote a candidate cluster; it is considered only if $|C_k| \geq n_{\min}$, where $n_{\min}$ is implicitly enforced in the implementation through qualitative inspection thresholds.

Second, **internal consistency** is assessed through cluster purity, defined as the proportion of samples sharing the dominant underlying intent and clusters with high purity (purity ≥ 0.75–0.95) are preferentially selected, as observed in the qualitative inspection logs.

Third, **separation from known intents** is enforced through the likelihood-based filtering described in the previous chapter. A candidate cluster must lie outside high-density regions of the known DP-GMM, ensuring that it does not simply correspond to a poorly classified instance of an existing intent.

Finally, **uncertainty** plays a critical role in *validating cluster reliability*: only samples passing the adaptive uncertainty filter are considered during clustering. Consequently, promoted clusters are composed of low-uncertainty embeddings, ensuring that their structure is not driven by ambiguous or noisy representations.

$$\text{Promote}(C_k) = \mathbb{I}\big[|C_k| \geq n_{\min} \ \wedge \ \bar{\ell}_k \geq \tau_\ell^{\text{disc}} \ \wedge \ \bar{\sigma}_k \leq \tau_\sigma^{\text{disc}}\big] \tag{4.5}$$

where $\bar{\ell}_k$ is the average log-likelihood of the samples in C_k under the discovery mixture model, τ_ℓ^{disc} is the corresponding minimum likelihood threshold, $\bar{\sigma}_k$ is the mean posterior uncertainty of the cluster and $\tau_\sigma^{\text{disc}}$ is the maximum uncertainty threshold allowed for promotion.

Together, these conditions define a **strict promotion policy**: only candidate clusters with sufficient support, high internal density and low average uncertainty are promoted to new intents. Separation from known-intent regions is enforced upstream by the novelty-detection gate. This prevents uncontrolled label-space expansion and ensures that newly introduced labels are supported by robust statistical evidence.

4.3.2. Creation of New Labels

Once a cluster satisfies the promotion criteria, it is assigned a new label representing a previously unseen intent, reflecting that each cluster corresponds to a semantic hypothesis rather than a predefined category. At this stage, the system does not assume prior knowledge about the meaning of the new intent; instead, it treats the cluster as an emergent structure whose semantic interpretation can be refined over time or through external validation.

4.3.3. Classifier Expansion

Following label creation, the classifier is explicitly expanded to accommodate the newly discovered intents, preserving previously learned parameters. Let the classifier output dimension at phase t be $|K_t|$. After incorporating K_{new} discovered labels, the output layer is extended to size $|K_t| + |K_{\text{new}}|$. The weights associated with existing classes remain unchanged, ensuring that prior knowledge is retained. New neurons corresponding to the discovered intents are initialized separately with standard initialization schemes and are subsequently trained during the consolidation phase. This **incremental expansion mechanism** allows the classifier to grow in capacity without retraining from scratch, maintaining continuity across phases while enabling adaptation to new semantic structures.

4.3.4. Label Space Update

The integration of new intents results in an explicit update of the label space: K_t denotes the set of known intents at phase t and the updated set at the next phase is given by $K_{t+1} = K_t \cup K_{\text{new}}$. Once incorporated, newly discovered intents are treated as standard known classes in subsequent phases. They contribute to the replay buffer, participate in density modelling and are subject to the same classification and uncertainty mechanisms as the original intents. This **recursive update** defines a continual expansion cycle: detection leads to discovery, discovery leads to promotion and promotion leads to integration. The process then repeats as new data becomes available (Figure 4-3).

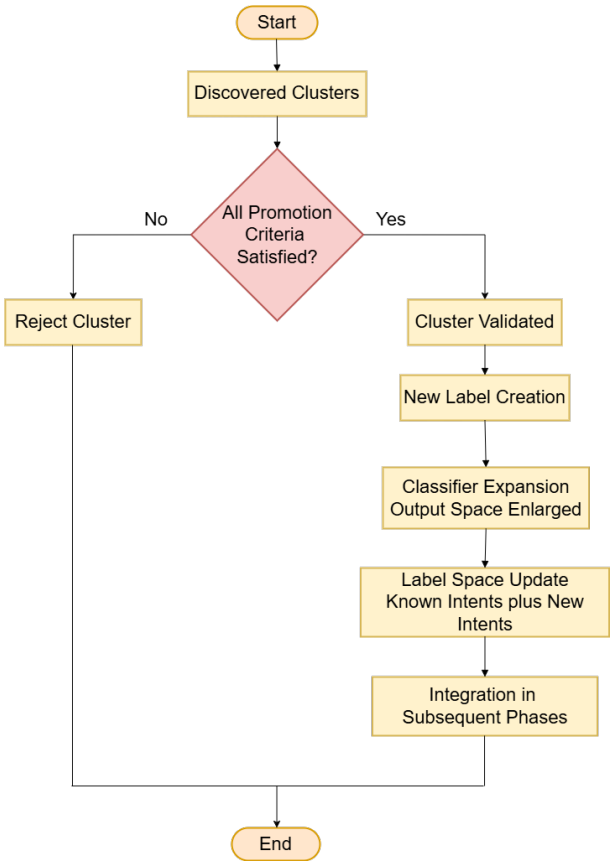

Figure 4.2: Label expansion process where validated clusters are promoted to new intents, updating the classifier and enlarging the label space.

4.4. Qualitative Cluster Inspection Module

Global clustering metrics are insufficient to fully characterize the quality of the learned structure in an open-world setting: **Normalized Mutual Information (NMI)** and **Adjusted Rand Index (ARI)**, although useful for benchmarking, fail to capture local semantic coherence and may remain close to zero even when clusters exhibit meaningful internal organization. To address this gap, the framework incorporates a **qualitative cluster inspection module**, designed to evaluate discovered intents at a finer granularity. Rather than relying exclusively on aggregate statistics, this module analyses each cluster individually through a set of interpretable descriptors.

For a given discovered cluster C_k, three key quantities are computed:

First, **support**,

$$\text{support}(C_k) = |C_k|, \tag{4.6}$$

measures the number of samples assigned to the cluster and provides an indication of

its statistical relevance. Clusters with very low support are typically unstable and less reliable, whereas larger clusters reflect recurring patterns in the data.

Second, **internal consistency** is quantified through *purity*,

$$\text{purity}(C_k) = \max_y \frac{|\{i \in C_k : y_i = y\}|}{|C_k|} \tag{4.7}$$

which evaluates the proportion of samples sharing the dominant underlying label. High purity indicates strong semantic alignment, even when global clustering metrics are low due to label permutation effects or class imbalance.

Third, the **distribution of underlying labels** within each cluster is examined, providing insight into how different semantic categories are grouped in latent space. This distributional view allows the identification of mixed clusters, emerging sub-intents, or potential over-segmentation.

Beyond numerical summaries, the module includes *representative examples from each cluster*, reporting the most frequent label along with its associated samples to enable *direct human interpretation*. These examples link latent representations to their semantic meaning, helping practitioners assess whether a cluster reflects a coherent intent or a spurious grouping. This qualitative analysis complements quantitative evaluation, supporting debugging, model refinement and interpretability, that is an aspect often overlooked in automated intent discovery (Figure 4-4), (see Appendix B-7 for pseudocode).

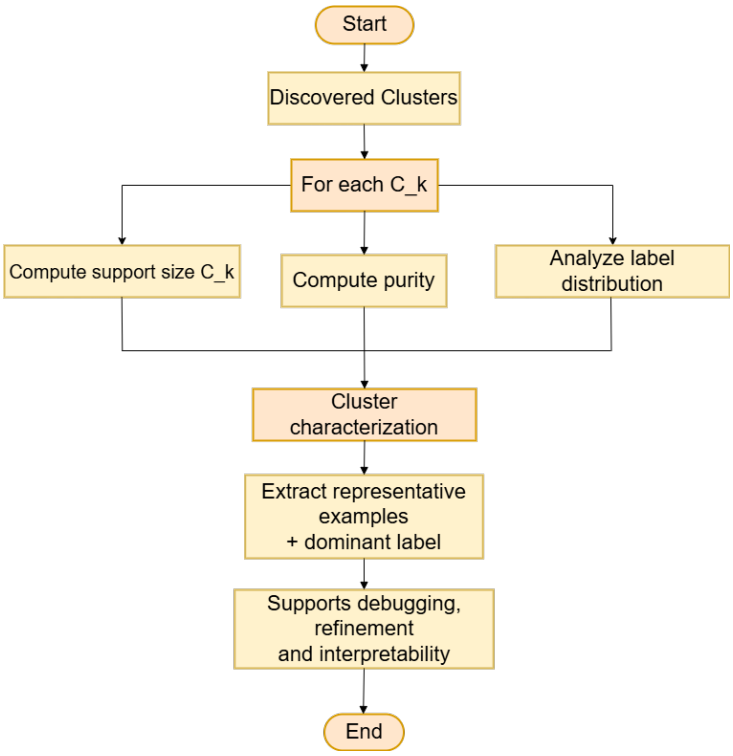

Figure 4.3: Qualitative cluster inspection module analyzing support, purity and label distribution to characterize clusters and extract representative examples.

4.5. Full Continual Learning Cycle

This section presents a comprehensive overview of the complete framework architecture and its operational cycle, showing how representation learning, uncertainty-aware novelty detection, cluster promotion and knowledge consolidation are integrated into a unified continual open-world learning process.

The framework starts with an initial supervised phase and then proceeds through repeated discovery–consolidation steps as new data batches arrive. The learning process is thus organized as a progressive cycle in which the model updates its latent space, evaluates incoming samples and selectively expands the label space while preserving previously acquired knowledge. The formal definition of the multi-phase process is provided in Section 3.2.

In **Phase 1**, the model is trained on the initial set of labelled known intents. This phase is executed once and initializes the main components of the framework: the uncertainty-aware latent representation, the classifier boundaries and the density model of known intents. These elements provide the reference structure used in the following open-world stages.

In **Phase 2** the model receives the mixed batch containing both known intents and novel samples whose labels are masked. Incoming samples are evaluated through the multi-signal decision mechanism, which combines classifier confidence, latent uncertainty and density likelihood. Samples that deviate from the known-intent structure are selected as novelty candidates and passed to the discovery module, where only reliable and coherent clusters can be promoted to new intent labels.

In **Phase 3** the same detection, discovery and consolidation procedure is repeated over subsequent mixed batches. At this point the label space may already include previously discovered intents, which are treated as part of the evolving known set. The model therefore continues to identify remaining novel structures, integrate validated clusters and retrain on current data together with replayed samples, maintaining alignment between old and newly acquired representations.

Throughout this cycle, two stabilization mechanisms support continual learning:

- the replay buffer preserves representative samples from previous phases, reducing distributional drift and maintaining exposure to past knowledge;
- EWC penalizes changes to parameters that are important for earlier phases, limiting abrupt forgetting during adaptation.

In the implemented experimental protocol, the continual process is finite and follows a predefined sequence of batches. Once all scheduled discovery and consolidation batches have been processed, the global cycle terminates. However, within this finite setup, the framework still operates according to a continual learning principle: knowledge is updated progressively, the intent space can expand over time and cluster promotion remains selective rather than automatic. Only candidate clusters satisfying the required reliability, density and support criteria are assigned new intent labels, while uncertain or weakly supported structures remain unassigned. This design allows the model to balance plasticity and stability through the joint action of probabilistic decision mechanisms, replay and regularization strategies (see Appendix B-8 for pseudocode).

4.6. Pseudocode

The following pseudocodes provide a unified view of the proposed continual learning framework, summarizing its full multi-phase operation. They integrate uncertainty-aware representation learning, adaptive filtering, density-based modelling and continual adaptation mechanisms, providing a compact operational view of the implemented system and illustrating how uncertainty, clustering and decision rules interact to support reliable nov-

elty detection and incremental label expansion (Algorithm 1, Algorithm 2, Algorithm 3).

Algorithm 1: Three-Phase Continual Intent Discovery Framework

Data: Initial labelled data $\mathcal{D}_0$ and unlabelled batch $\mathcal{N}$

Result: Updated model and expanded label space $\mathcal{K}$

Initialize encoder, classifier, DP-GMM and label set $\mathcal{K}_0$;

Phase 1: Representation Learning;

Train encoder and classifier on known intents $\mathcal{K}_0$;

Encode samples to obtain latent means μ_i and uncertainties σ_i;

Fit the known-intent DP-GMM on trusted latent samples;

Introduction of new unalabelled data:

Phase 2: Novelty Detection and Discovery;

Identify candidate novel samples using confidence, uncertainty and likelihood;

Discover latent clusters from filtered samples;

Phase 3: Consolidation;

Promote reliable clusters to new intents $\mathcal{K}_{new}^{(b)}$;

Expand classifier and update label space $\mathcal{K}_b = \mathcal{K}_{b-1} \cup \mathcal{K}_{new}^{(b)}$;

Update model using replay buffer and EWC regularization;

Algorithm 2: Five-Phase Continual Intent Discovery Framework

Data: Initial labelled data $\mathcal{D}_0$ and unlabelled batches $\mathcal{N}_1, \mathcal{N}_2$

Result: Updated model and expanded label space $\mathcal{K}$

Initialize encoder, classifier, DP-GMM and label set $\mathcal{K}_0$;

Phase 1: Representation Learning;

Train encoder and classifier on known intents $\mathcal{K}_0$;

Encode samples to obtain latent means μ_i and uncertainties σ_i;

Fit the known-intent DP-GMM on trusted latent samples;

for *each unlabelled batch* $\mathcal{N}_b \in \{\mathcal{N}_1, \mathcal{N}_2\}$ **do**

    Phase 2: Novelty Detection and Discovery;

    Identify candidate novel samples using confidence, uncertainty and likelihood;

    Discover latent clusters from filtered samples;

    Phase 3: Consolidation;

    Promote reliable clusters to new intents $\mathcal{K}_{new}^{(b)}$;

    Expand classifier and update label space $\mathcal{K}_b = \mathcal{K}_{b-1} \cup \mathcal{K}_{new}^{(b)}$;

    Update model using replay buffer and EWC regularization;

Algorithm 3: Uncertainty-Aware Novelty Detection and Discovery

Data: Latent means μ_i, uncertainties σ_i, classifier outputs

Result: Discovered clusters $\mathcal{C}_{new}$

Compute uncertainty scores σ_i;

Compute adaptive uncertainty threshold τ_σ;

Select trusted samples:

$$\mathcal{D}_{trusted} = \{i \mid \sigma_i \leq \tau_\sigma\}$$

Apply PCA on trusted latent means μ_i (retain 95% variance);

Fit known-intent DP-GMM on reduced trusted representations;

Compute robust likelihood threshold τ_ℓ from trusted samples;

for *each sample* i **do**

    Compute classifier confidence c_i;

    Compute uncertainty σ_i;

    Compute log-likelihood ℓ_i under the known-intent DP-GMM;

    if c_i *low* ***and*** σ_i *high* ***and*** $\ell_i < \tau_\ell$ **then**

        Assign i to novel candidate set;

    else

        Keep i within the known-intent decision space;

Fit discovery DP-GMM on novel candidate samples;

for *each candidate cluster* C_k **do**

    Compute support, mean likelihood and mean uncertainty;

    if *support, density, consistency and reliability criteria are satisfied* **then**

        Add C_k to $\mathcal{C}_{new}$;

4.7. Chapter Summary

This chapter establishes a operational framework in which uncertainty-aware modelling, density-based discovery and continual learning mechanisms are tightly integrated, enabling robust intent discovery under evolving and open-world conditions.

The novelty detection strategy is structured as a multi-stage process in which classifier confidence, latent uncertainty and density-based likelihood are integrated sequentially and each signal refines the reliability of the decision. A non-forced assignment policy is adopted where uncertain samples are retained as candidates for discovery, prioritizing reliable predictions.

Catastrophic forgetting is mitigated through an uncertainty-aware replay buffer that preserves past knowledge by retaining both low-uncertainty samples, stabilizing the training data distribution. In parallel EWC constrains parameter updates to protect information that is critical for previously learned intents.

Label expansion is performed conservatively: only reliable clusters with sufficient support, consistency, separation from known intents and low uncertainty are promoted to new intents and integrated into the classifier for subsequent phases.

A qualitative cluster inspection module complements quantitative evaluation by analysing cluster support, purity, label distribution and representative examples. This provides interpretability and captures semantic structure that global metrics may fail to reflect.

All components are integrated within a continual learning cycle: after the initial supervised phase, Phase 2 and Phase 3 are repeated for each new data batch, enabling the model to detect novelty, consolidate reliable discoveries and progressively update the label space over time.

This chapter formalizes the decision-making and continual learning mechanisms that enable the framework to operate in dynamic and open-world intent spaces.

5 | Experimental Evaluation

5.1. Experimental Setup

5.1.1. Dataset

The experimental evaluation is conducted on the `CLINC150` intent classification dataset, a benchmark composed of short user utterances annotated with fine-grained intent labels. In this thesis, the original out-of-scope (OOS) labels are not used as novel intents. Instead, the experiment focuses on the in-domain intent set: OOS-like labels are removed before constructing the continual stream and novelty is simulated by withholding a subset of in-domain intents from the initial supervised phase.

After filtering, 100 in-domain intents are selected and divided into two disjoint groups: $K_0 = 50$ base known intents and $N = 50$ novel intents. The known intents (5000 labelled samples) are used in Phase 1 to initialize the system under supervised conditione.In phases 2 and 3, the novel intents, containing mixed streams of known samples and samples from the current novel macro-batch, are introduced.

The dataset is used in a controlled and approximately balanced configuration, without explicitly introducing class imbalance or long-tail distributions. This allows the experiments to focus on the behaviour of the proposed continual discovery framework, while the study of naturally imbalanced intent streams is left for future work.

5.1.2. Data Pre-processing

The data pre-processing stage transforms the selected `CLINC150` intent data into the input format required by the continual discovery framework. The implementation operates on pre-computed semantic embeddings, together with their corresponding intent labels. When available, the original utterance text is also retained only for qualitative inspection, while all learning and inference steps are performed on the embedding representations.

After the in-domain intent set has been defined, the data are reorganized into a phase-

wise continual stream. A fixed global evaluation set is first constructed from all selected known and novel intents and is kept unchanged throughout the experiment. This allows performance to be assessed after each phase on the same evaluation space, including both previously known categories and truly novel categories.

The training stream is then generated according to the continual protocol. The first phase contains only labelled samples from the initial known intent set K_0. In the following two phases, each stream contains a controlled mixture of known samples and samples from the same novel intent set N. Known-intent labels are preserved, whereas all labels belonging to novel intents are replaced with the placeholder `_NO_LABEL`. This masking step prevents direct supervision on emerging categories and forces the framework to rely on novelty detection, density modelling and clustering for their identification (Table 5-1).

Sampling is performed with a controlled known–novel ratio, so that the model is exposed to both old and new categories during each continual phase. These phase-specific streams are then used by the continual learning pipeline, where replay, uncertainty-based filtering, pseudo-labelling and cluster promotion are applied. For evaluation purposes, novelty detection is converted into a binary problem. Samples whose ground-truth label belongs to the novel intent set are treated as positive instances, while samples from K_0 are treated as negative instances. This binary formulation does not replace the multi-signal decision mechanism of the framework; rather, it is used only to evaluate the final prediction produced by the model. A prediction is counted as novel whenever the final output is either the generic `unknown_intent` label or a promoted discovered label of the form `__NOVEL_DISCOVERED__k`:

$$\hat{y}_{\text{novel}} = \begin{cases} 1 & \text{if output} \in \{\texttt{unknown\_intent}, \texttt{\_\_NOVEL\_DISCOVERED\_\_k}\} \\ 0 & \text{if output} \in K_0 \end{cases}$$

Table 5.1: Summary of the dataset and the three-phases protocol

Component	Setting
Dataset	CLINC150 in-domain intent dataset
Total selected intents	100
Base known intents K_0	50
Novel intents N	50
Total learning steps	3: one supervised phase and two continual updates
Phase 1 data	Labelled samples from K_0 only
Continual phase data	Mixed known and current novel samples
Novel-label handling	Novel labels masked as `UNKNOWN_NO_LABEL`
Evaluation set	Fixed global set with known and novel intents
Input representation	Pre-computed semantic embeddings
Main latent dimension	$z = 64$
Replay buffer	5000 samples
Replay hard ratio	0.95
EWC weight	$\lambda = 7500$

The extension of the framework to five phases is characterized by the use of all 150 in-domain intents, divided into $K_0 = 50$ known intents and $N = 100$ novel intents. After the initial supervised phase on K_0, the novel intents are split into two macro-batches of 50 intents each. For each macro-batch, novelty detection and discovery are followed by consolidation and label-space expansion, resulting in four continual updates after Phase 1 (Table 5-2).

Table 5.2: Summary of the dataset and the five-phases protocol

Component	Setting
Dataset	CLINC150 in-domain intent dataset
Total selected intents	150
Base known intents K_0	50
Novel intents N	100
Novel macro-batches	2 macro-batches of 50 novel intents each
Total learning steps	5: one supervised phase and four continual updates
Phase 1 data	Labelled samples from K_0 only
Continual phase data	Mixed known and current novel samples
Novel-label handling	Novel labels masked as `UNKNOWN_NO_LABEL`
Evaluation set	Fixed global set with known and novel intents
Input representation	Pre-computed semantic embeddings
Main latent dimension	$z = 64$
Replay buffer	5000 samples
Replay hard ratio	0.95
EWC weight	$\lambda = 7500$

5.1.3. Evaluation Metrics

The evaluation protocol assesses the proposed framework as a reliability-oriented continual intent discovery system operating under an evolving label space. Rather than assuming that all unseen intents must be exhaustively recovered as separate categories, the evaluation focuses on three core behaviours: stable decision-making over previously learned intents, detection of deviations from the known intent distribution and selective promotion of statistically reliable novel structures. Accordingly, the experimental analysis combines complementary metrics that capture different levels of the continual discovery process:

- **Global accuracy** measures the overall correctness of the final open-world prediction

on the global evaluation set. It reflects whether the model can correctly classify known intents while preventing novel samples from being forced into existing known classes. This metric is influenced by known-intent classification quality, novelty thresholds, classifier confidence, VAE uncertainty, GMM likelihood scores, replay and EWC-based stability. However, since it compresses known-intent recognition and novelty handling into a single score, it is used mainly as a global stability indicator rather than as the sole measure of discovery quality.

- **Novelty detection precision, recall and F1-score** evaluate the binary known–novel decision produced by the framework. A sample is predicted as novel when the final output is either the generic `unknown_intent` label or a promoted discovered label of the form `__NOVEL_DISCOVERED__k`. Precision measures how reliable the novel predictions are, recall measures how many truly novel samples are detected and F1-score summarizes their balance. These metrics are influenced by classifier confidence, VAE uncertainty, GMM likelihood thresholds and the conservative gating strategy. For this reason, they represent the main indicators of novelty detection quality, especially because the framework prioritizes reliable known–novel separation over assigning every novel sample to a specific discovered class.
- **Normalized Mutual Information (NMI)** and **Adjusted Rand Index (ARI)** inspect the fine-grained structure of the discovered novel clusters. They are computed only on truly novel samples and compare the promoted cluster labels with the ground-truth novel intents. These metrics are influenced by latent-space separability, DP-GMM density estimation, uncertainty filtering, minimum cluster size and the strictness of the promotion criteria. However, they are not the primary objective of the thesis, since the framework is designed to perform reliable novelty detection and selective label-space expansion rather than to exhaustively reconstruct the complete fine-grained taxonomy of all novel intents.

Taken together, these metrics separate the main evaluation targets from the auxiliary diagnostic analysis. The primary focus is on verifying whether the framework can remain stable across continual phases, detect novel samples under uncertainty and expand the label space only when sufficient density-based evidence is available. The clustering metrics provide an additional view of how much fine-grained semantic structure is recovered among novel intents, but they are interpreted in light of the framework's conservative design rather than as the central criterion of success.

5.1.4. Experimental Design and Implementation Details

The experimental setup uses a controlled **five-run protocol** to evaluate the robustness and reproducibility of the framework under different stochastic continual-learning trajectories. The five runs are not different model configurations: the architecture, thresholds, β-VAE, DP-GMM modules, replay, EWC and decision logic are kept fixed. Only the random seed changes, using the set $\{42, 7, 13, 21, 123\}$.

The seed controls NumPy and PyTorch randomness, the deterministic shuffling of intents, the construction of the continual stream and the initialization of trainable parameters. Therefore, the differences across runs reflect stochastic factors such as intent ordering, data sampling and model initialization, rather than hyperparameter sensitivity. This is important in continual new intent discovery because the order in which novel intents appear can affect representation learning, replay selection, clustering and label-space expansion.

The implementation is kept consistent across all runs:

- input representation: pre-computed semantic embeddings;
- original utterances: loaded only for qualitative inspection, not used as direct training inputs;
- device: CPU;
- latent dimension: $z_{\text{dim}} = 64$;
- batch size: 64;
- number of epochs: 60;
- optimizer: Adam;

Repeating the same configuration over five seeds allows the results to be summarized with mean and standard deviation, giving a more reliable estimate of framework stability than a single execution.

5.1.5. Hyperparameter Selection Strategy

The experimental study adopts a controlled configuration strategy rather than an exhaustive grid search. In the final implementation, the main hyperparameters are kept fixed across all runs in order to evaluate the behaviour of the proposed framework under a stable and reproducible setting.

The continual learning setup uses an EWC regularization strength of $\lambda = 7500$ and a

smoothing factor of $\alpha = 0.7$, a replay buffer size of 5000 and a replay hard-sample ratio of 0.95. The classifier loss multiplier is set to 20.0, while the confidence threshold for prediction and novelty-related decisions is set to 0.75. The density-based modules use a maximum of 200 components for the known-intent DP-GMM and 100 components for the novelty discovery DP-GMM.

This design enables:

- evaluation of robustness to stream ordering;
- estimation of variability across independent runs;
- assessment of the stability of novelty detection and continual learning behaviour.

5.2. Experimental Results

5.2.1. Results of the Three-Phase Framework

Overall Performance

The experimental results reveal a **moderate but stable performance** of the proposed continual intent discovery framework across the five runs executed (Figure 5-1).

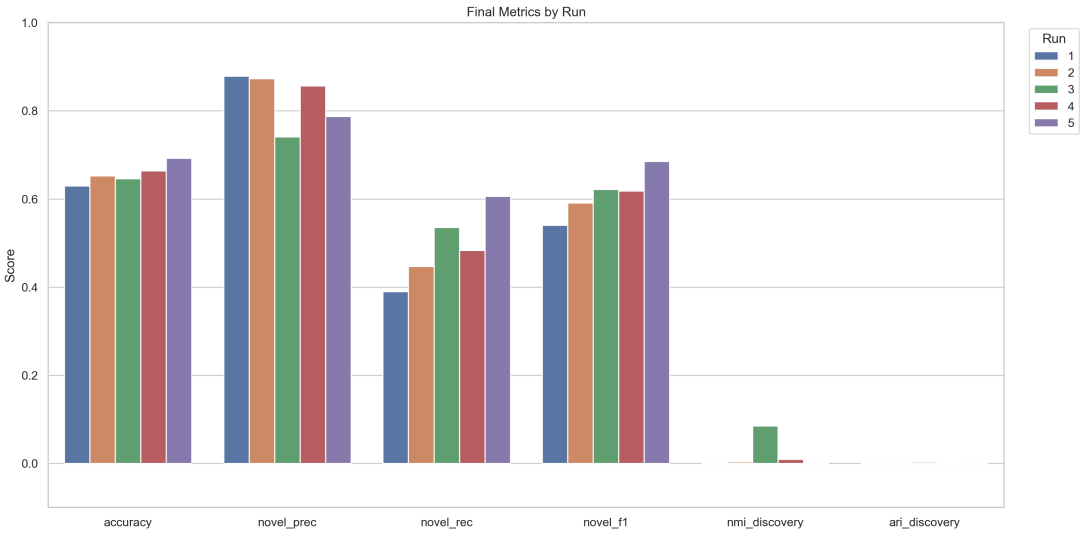

Figure 5.1: Final Metrics by Run bar plot: overall performance evaluation.

Global accuracy remains relatively stable, ranging from 0.6296 to 0.6924, with Run 5 achieving the highest value. This limited variation suggests that the classification component of the framework is reasonably robust across the tested configurations.

However, a markedly different behaviour is observed in the *novelty detection metrics*: **novel precision** remains consistently high across all runs, ranging from 0.7413 to 0.8791, indicating that the framework adopts a conservative and reliability-oriented detection strategy. At the same time, **novel recall** remains in a narrower but still variable range, from 0.3898 to 0.6064, showing that the model detects a substantial portion of novel samples but does not attempt to assign every uncertain sample to novelty.

This behaviour directly determines the **Novel F1 score**, which ranges from 0.5410 to 0.6852. Run 5 achieves the best overall balance between precision and recall, while Runs 2 and 5 show very close performance, confirming that the framework is relatively stable across seeds rather than being dependent on a single favourable run.

The two scatter plots (Figure 5-2) provide a compact view of this behaviour. The first plot compares *global accuracy and Novel F1*, showing that better classification stability is generally associated with stronger novelty-detection performance. This indicates that the best runs are not only better at recognizing known intents, but also maintain a more effective known–novel separation.

The second plot focuses specifically on the *novelty-detection trade-off*. Precision remains high in all runs, while recall varies more visibly. This pattern is consistent with the decision logic implemented in the framework: novelty is assigned only when classifier confidence, posterior uncertainty and density-based evidence jointly support the decision. Therefore, the plots show a conservative but stable operating regime, where the model prioritizes reliable novelty assignments while still preserving a useful level of recall.

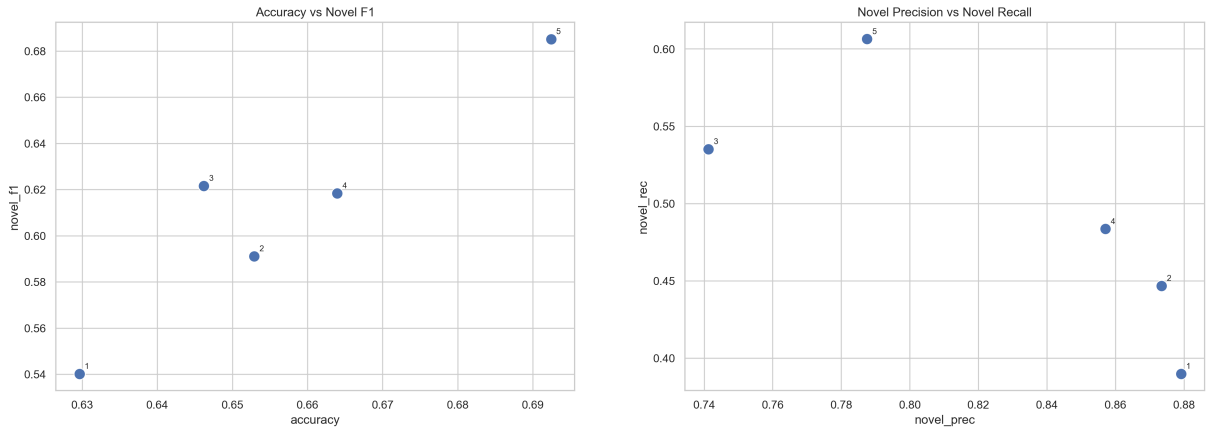

Figure 5.2: Novelty detection analysis highlighting the trade-off between precision, recall and overall performance across different runs.

The results confirm a stable and reliability-oriented behaviour across different seeds, par-

ticularly in terms of novelty detection precision and controlled performance variability. At the same time, the performance is not perfectly uniform across runs, recall remains lower than precision due to the conservative decision strategy and the final outcome is still sensitive to threshold calibration, replay configuration and the stochastic structure of the continual stream.

A clear limitation is that NMI and ARI remain very low overall, despite a modest increase in the reduced-complexity Run 3, with NMI ranging from 0.0000 to 0.0847 and ARI from 0.0000 to 0.0027. However, this result is consistent with the conservative design of the framework: NMI and ARI can remain close to zero even when some promoted clusters satisfy the framework's local promotion criteria, because these metrics evaluate the agreement between the discovered clusters and the complete fine-grained ground-truth partition of the novel intents. In this work, the objective is controlled and trustworthy discovery under continual-learning constraints rather than exhaustive global clustering reconstruction. Only clusters satisfying strict density, support and consistency constraints are promoted, while ambiguous or overlapping samples are kept unassigned or conservatively treated as unknown. Therefore, low NMI and ARI indicate that the framework does not recover the full taxonomy of novel intents, but do not imply that it degenerates into pure continual new intent detection, since it still performs discovery through latent-space clustering, selective cluster promotion and label-space expansion. For this reason, global clustering metrics alone are not sufficient to fully characterize the learned structure in an open-world setting: they may fail to capture local semantic coherence within individual promoted clusters, motivating the qualitative cluster inspection module introduced in this work, which analyses discovered intents at a finer granularity through interpretable cluster-level descriptors.

The framework does not implement a classical explicit **forgetting metric**, such as the average accuracy drop on old classes between their best previous performance and their final performance. Instead, forgetting is estimated through a phase-wise stability analysis. Specifically, after each continual phase, the model is evaluated on a fixed global held-out set and the evolution of global accuracy is used as the main retention proxy. A strong decrease in this metric across phases would indicate that the model is losing previously acquired classification ability, whereas stable values suggest that previous knowledge is largely preserved. Novel F1 is monitored in parallel to ensure that stability on known intents is not achieved at the expense of novelty detection, while NMI and ARI describe the alignment between discovered clusters and true novel intents. Thus, forgetting is assessed indirectly through the temporal degradation, or stability, of the model's evaluation behaviour, rather than through a dedicated old-class forgetting formula. Replay and Elastic

Weight Consolidation act as the main anti-forgetting mechanisms, mitigating respectively representation-level drift and parameter-level drift during continual updates.

Overall, the results confirm that the framework achieves stable and high-precision novelty detection under a challenging multi-batch continual setting, while also highlighting acknowledged limitations related to its conservative design, including limited reconstruction of the underlying novel-intent taxonomy, sensitivity to threshold calibration and the trade-off between reliable discovery and broader novelty coverage.

Continual Learning Behaviour

Across phases, the framework shows a progressive degradation in **classification-oriented performance**, accompanied by a more variable but still controlled evolution of novelty detection (Figure 5-3).

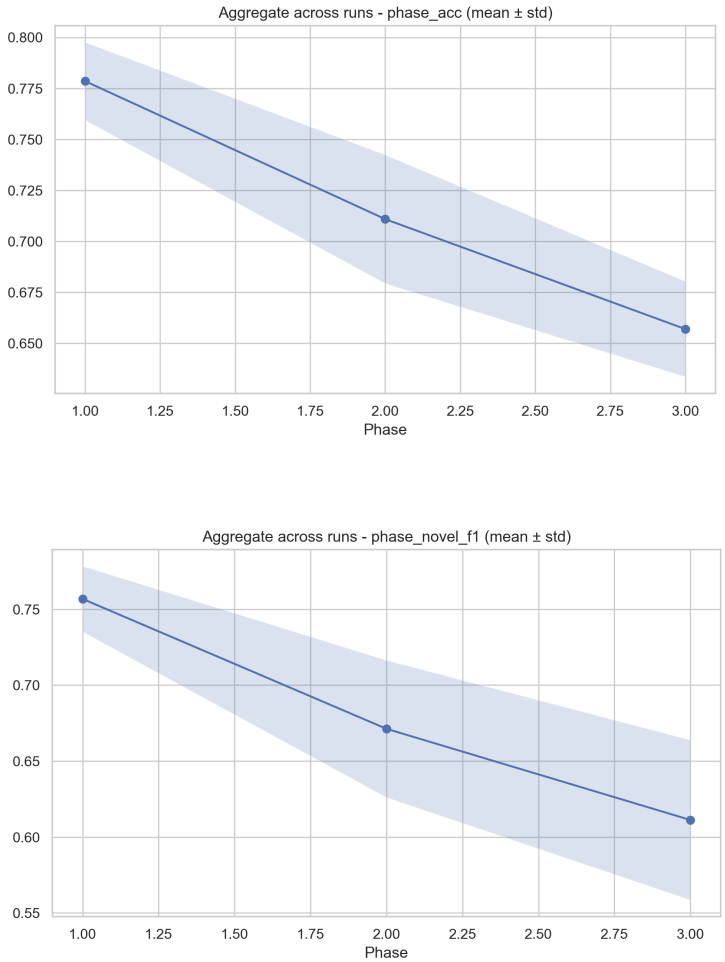

Figure 5.3: Continual learning behaviour across phases, showing gradual performance degradation and variability as new tasks are introduced.

Phase 1 is the only purely supervised stage, where the classifier and β-VAE are optimized on labelled base intents only.

By contrast, **Phase 2 and 3** are intrinsically harder: the system must preserve the old decision structure while simultaneously identifying, rejecting and eventually integrating new semantic classes into an enlarged label space. This growing difficulty is clearly visible in the phase-wise accuracy curve averaged across runs.

The **first graph** shows a monotonic decrease of mean phase accuracy from approximately 0.778 in Phase 1 to 0.657 in Phase 3, with the uncertainty band also widening slightly over time. The model gradually faces increasing difficulty in maintaining clear separation between known and newly introduced classes as the latent space becomes more complex. This reflects structured interference rather than catastrophic forgetting. While replay and EWC effectively stabilize training and prevent abrupt degradation, the expansion of the label space naturally makes it more challenging to preserve existing knowledge while integrating new classes.

The **phase-wise Novel F1 curve** shows that mean novelty-detection F1 also decreases over time, from approximately 0.757 in Phase 1 0.611 in Phase 3, again with increasing variability across runs. This confirms that novelty detection becomes harder as the continual process progresses. As training progresses, this mechanism operates on an increasingly crowded and heterogeneous latent space, due to the addition of new classes, pseudo-labels and replayed samples. This leads to greater decision ambiguity, making the distinction between known and novel inputs less stable. The expansion of the label space further increases this complexity, resulting in a trade-off between classification stability and novelty sensitivity, reflected in the gradual decline of both accuracy and Novel F1.

The **uncertainty plots** provide a more reassuring perspective. Both `sigma_quantile` and `actual_trusted_ratio` remain very close to 0.90 across all phases and runs, reflecting the behaviour of the adaptive quantile mechanism implemented (Figure 5-4).

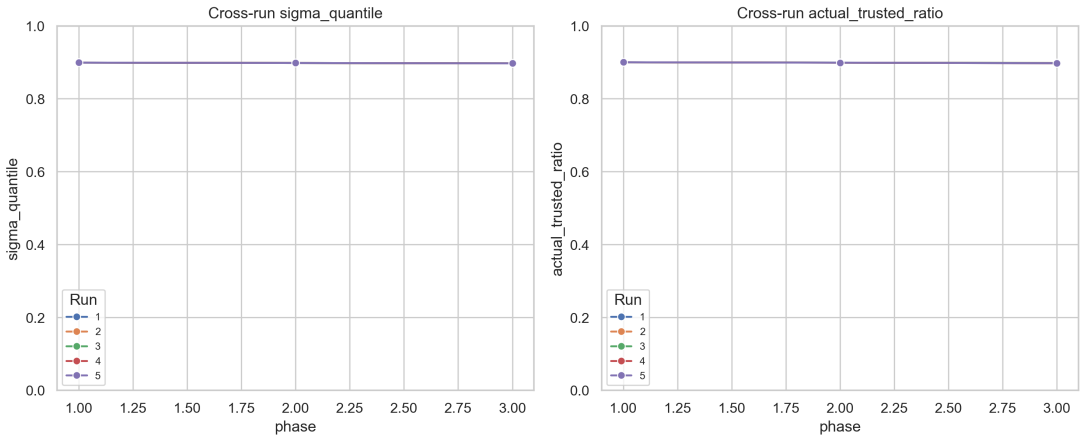

Figure 5.4: Stability of uncertainty-based filtering across phases, with consistent sigma quantile and trusted sample ratio over multiple runs.

These values should not be interpreted as fixed thresholds manually kept at 0.90: in the implementation, the uncertainty quantile is initialized at 0.90 and then updated phase by phase according to the discrepancy between the observed trusted-sample ratio and the configured target ratio. In the representative experimental logs, the updated quantile evolves slightly from 0.8990 in Phase 1 to 0.8980, 0.8971, 0.8961 and 0.8952 in later phases, while the corresponding sigma cutoff moves from 1.0495 to 0.9360, 0.9189, 0.9258 and 0.9088. Therefore, the nearly horizontal curves in Figure 5-5 indicate that the adaptive controller performs only small corrections because the observed trusted ratio remains close to the desired filtering regime.

This behaviour indicates a stable uncertainty mechanism, with no signs of threshold instability or filtering collapse and with a consistent proportion of trusted samples throughout the continual process. The apparent flatness of the plot is therefore not evidence of a fixed threshold, but rather of an adaptive mechanism whose updates remain small under stable uncertainty calibration. This stability is crucial for interpreting the observed performance degradation: since uncertainty remains well calibrated, the decline in phase-wise accuracy and Novel F1 cannot be attributed to instability in the β-VAE gating mechanism. Instead, it reflects the increasing difficulty of the task as new classes are introduced and the label space expands. In this framework, uncertainty acts as a reliable regulator of decision quality, but cannot fully offset the growing interference caused by continual learning dynamics.

Figure 5-5 provides a phase-wise view of the **decision mechanism** for the Run 2. The σ cutoff remains almost stable across phases, indicating that uncertainty-based filtering keeps a consistent trusted-sample regime. Meanwhile, the GMM likelihood thresholds are

progressively recalibrated as new batches are introduced, reflecting the adaptation of the density model to an expanding latent space. Accuracy and Novel F1 remain within a controlled range, while the decision thresholds evolve across the continual process.

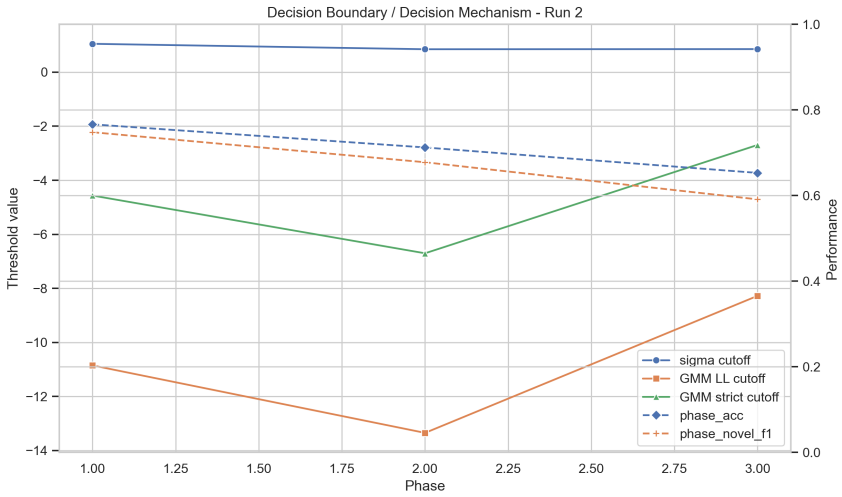

Figure 5.5: Evolution of the multi-signal decision mechanism across continual phases, showing uncertainty and GMM likelihood thresholds together with phase accuracy and Novel F1.

The results show that the framework maintains a stable control signal, while the joint objectives of classification and discovery become progressively more challenging as the process evolves.

Clustering Analysis

Although the framework is effective at the novelty-detection level, its ability to reconstruct the fine-grained taxonomy of novel intents remains limited. The **NMI and ARI metrics** evaluate whether the discovered clusters align with the ground-truth partition of the novel classes (Figure 5-6).

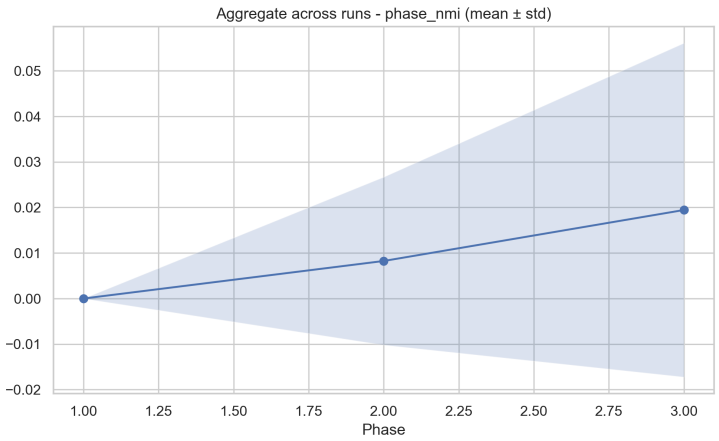

Figure 5.6: Clustering analysis showing consistently low alignment (NMI, ARI) across runs and phases, with only marginal improvements in later stages.

The **qualitative inspection** provides a more nuanced interpretation. Some discovered clusters show high local purity, with values around 0.91–1.00, indicating well-separated and internally consistent structures. Other clusters are more mixed, with purity ranging from approximately 0.54 to 0.88, while the remaining clusters exhibit lower purity values reflecting stronger semantic overlap or incomplete separation. This explains the discrepancy between local observations and aggregate metrics: the framework can identify dense and semantically coherent regions, but these regions are not consistently aligned with the complete set of true novel labels.

This outcome is consistent with the implemented discovery strategy: candidate samples are filtered through uncertainty-based trusted masks, DP-GMM density estimation, likelihood-based selection, minimum support constraints and conservative promotion rules. These steps reduce unreliable discoveries, but also restrict the diversity of samples available for clustering and favour only dense, statistically reliable regions.

The learned latent space appears more suitable for robust known–unknown discrimination than for exhaustive fine-grained clustering of unseen intents. This limitation is coherent with the conservative design of the framework: the goal is not full reconstruction of the novel-intent taxonomy, but selective and reliable discovery under continual open-world conditions.

5.2.2. Results of the Five-Phase Framework

Overall Performance

Compared with the three-phase framework, the five-phase extension maintains a **moderate and relatively stable overall performance**, although under a more demanding continual setting in which novel intents are introduced through additional sequential updates (Figure 5-7).

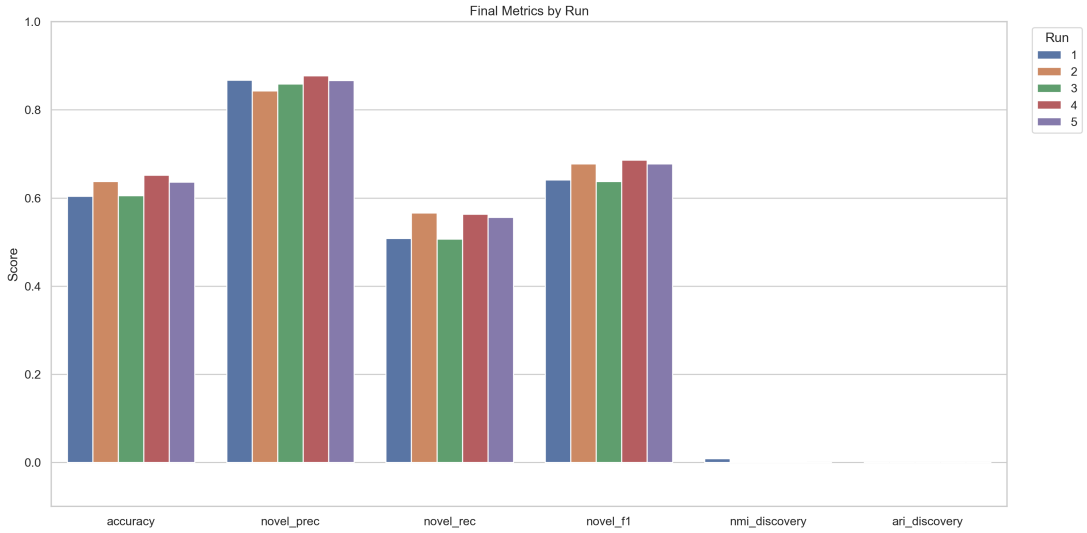

Figure 5.7: Final Metrics by Run bar plot: overall performance evaluation.

Global accuracy ranges from 0.6043 to 0.6521, showing a moderate decrease with respect to the three-phase configuration. This reduction indicates that extending the continual stream slightly increases the difficulty of preserving globally correct predictions as the model is updated over a longer sequence of learning phases. A different pattern emerges for *novelty detection*.

Novel precision remains consistently high, ranging from 0.8431 to 0.8776 and is less variable than in the three-phase framework, indicating that the model continues to identify novel samples with limited false positive behaviour. **Novel recall** ranges from 0.5072 to 0.5664, remaining above the lowest values observed in the three-phase configuration. This suggests that distributing novel-intent arrival across multiple updates improves the consistency with which novel samples are detected, even though it does not increase overall classification accuracy.

Accordingly, the **Novel F1 score** ranges from 0.6378 to 0.6860, with Run 4 achieving the best balance between precision and recall. Compared with the three-phase framework, the five-phase extension produces a narrower and generally higher Novel F1 interval, indicating a more stable novelty-detection behaviour across runs. Therefore, while the

longer continual process introduces a moderate cost in terms of global accuracy, it provides a more balanced and consistent identification of novel samples.

The two scatter plots (Figure 5-8) provide a compact view of the behaviour of the five-phase framework. The first plot compares *global accuracy and Novel F1*, showing a clear positive association between the two metrics: runs with stronger overall classification performance also achieve better novelty-detection results. In particular, Run 4 obtains the best joint performance, indicating that the extended continual process can preserve an effective balance between global prediction quality and known–novel separation despite the additional sequential updates.

The second plot focuses on the *novelty-detection trade-off*. Precision remains consistently high across all runs, whereas recall is distributed within a relatively narrow interval. Compared with the three-phase framework, where recall exhibited greater variability and reached substantially lower values in some runs, the five-phase extension produces a more regular detection behaviour. This suggests that introducing novel intents progressively across multiple updates allows the model to manage the novelty stream more consistently, rather than confronting the entire expansion within a shorter continual process.

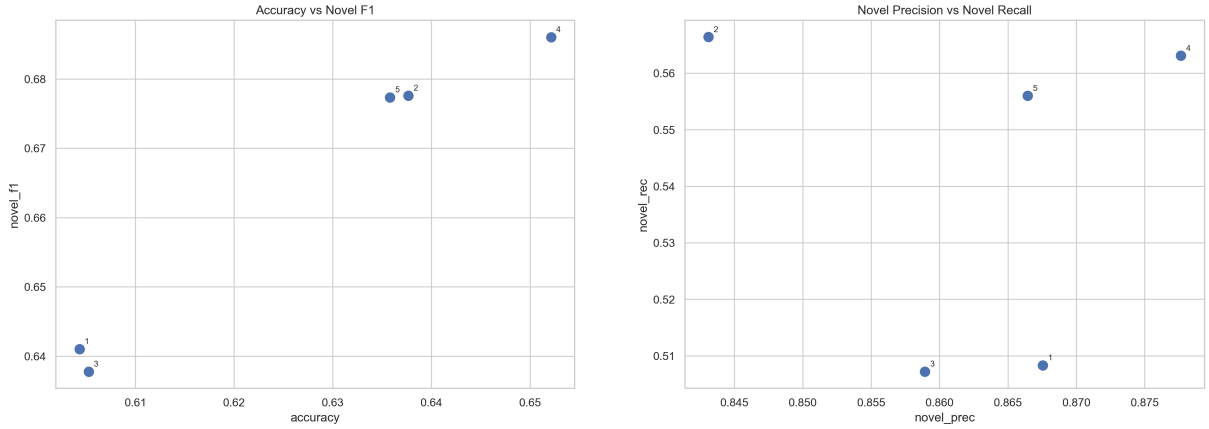

Figure 5.8: Novelty detection analysis highlighting the trade-off between precision, recall and overall performance across different runs.

NMI and **ARI** remain essentially zero also in the five-phase framework indicating that the additional sequential updates do not improve the global reconstruction of the fine-grained novel-intent partition. The model remains still oriented toward selective and reliable cluster promotion rather than exhaustive recovery of the complete novel-label taxonomy.

The results therefore confirm that the five-phase framework preserves the reliability-

oriented behaviour of the original model while reducing variability in novelty detection across random seeds. Precision remains higher than recall, consistently with the conservative decision logic based on classifier confidence, posterior uncertainty and density-based evidence. However, the more concentrated performance pattern indicates that the extended temporal structure supports a more stable known–novel separation, although with a moderate reduction in global accuracy compared with the three-phase setting.

Continual Learning Behaviour

Across phases, the five-phase framework exhibits an initial reduction in **classification-oriented performance** followed by a substantially more stable evolution over the subsequent continual (Figure 5-9).

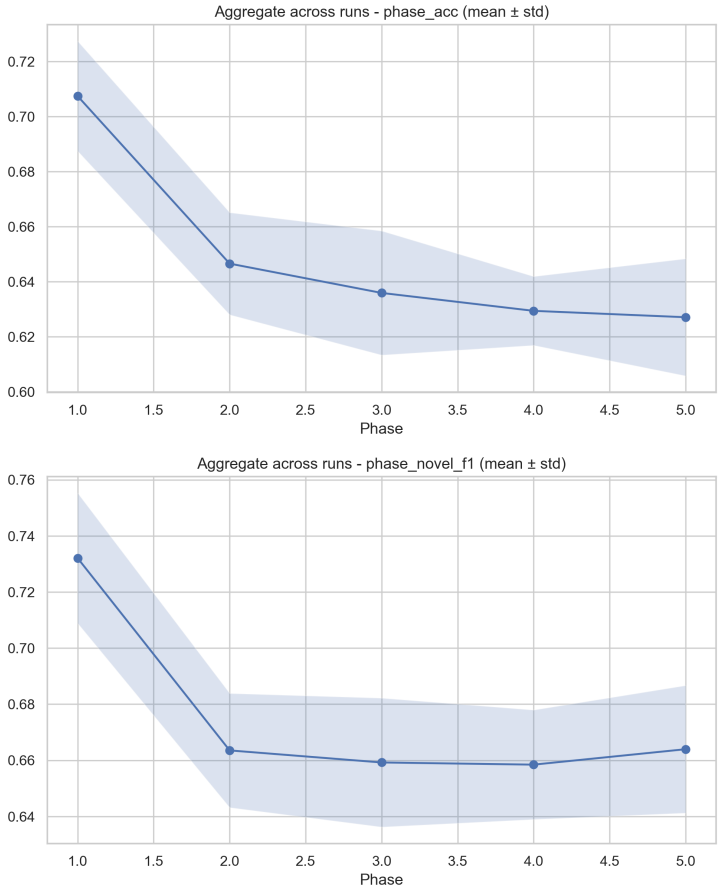

Figure 5.9: Continual learning behaviour across phases, showing gradual performance degradation and variability as new tasks are introduced.

Phase 1 remains the supervised initialization stage, whereas the following phases progressively introduce novel intents and require the model to update an increasingly expanded

label space. Mean phase accuracy decreases from approximately 0.707 in Phase 1 to 0.652 in Phase 5.

Compared with the three-phase framework, where accuracy continues to decrease more markedly up to the final stage, the extended configuration shows that most of the performance loss occurs at the first transition from supervised learning to open-world adaptation. After this initial adjustment, replay and EWC appear to support a relatively stable consolidation process, despite the repeated integration of new intent batches.

A similar pattern is observed for the **phase-wise Novel F1**. The mean score decreases from approximately 0.751 in Phase 1 to 0.595 in Phases 3–5, with a slight recovery in the final phase. In contrast to the continuing decrease observed in the three-phase setting, this plateau indicates that the additional sequential updates do not progressively destabilize novelty detection. The five-phase framework introduces an early adaptation cost, but distributes later novelty integration more gradually, resulting in a more stable balance between retention of previous knowledge and detection of newly arriving intents.

The uncertainty-related results (Figure 5-10) confirm that the five-phase extension preserves the stable filtering behaviour already observed in the three-phase framework. Both `sigma_quantile` and `actual_trusted_ratio` remain close to 0.90 across all runs and phases. The updated quantile evolves slightly from 0.8990 in Phase 1 to 0.8980, 0.8971, 0.8961 and 0.8952 in later phases, while the corresponding sigma cutoff moves from 1.0495 to 0.9360, 0.9189, 0.9258 and 0.9088. However, in the extended setting this stability is maintained over a longer sequence of updates, indicating that the uncertainty-based selection mechanism does not become unstable as additional novel-intent batches are introduced.

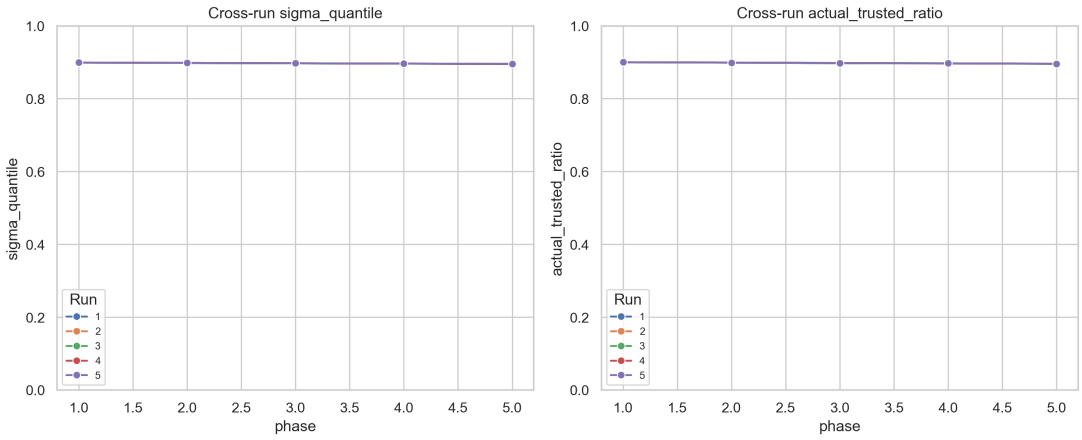

Figure 5.10: Stability of uncertainty-based filtering across phases, with consistent sigma quantile and trusted sample ratio over multiple runs.

The decision-mechanism analysis (Figure 5-11) further clarifies this behaviour. In the best-performing five-phase run, the σ cutoff remains approximately stable, while the GMM likelihood thresholds undergo a more pronounced phase-wise recalibration as the continual stream progresses. This indicates that the uncertainty signal continues to define a consistent trusted-sample regime, whereas the density model adapts to the progressively evolving latent distribution. Compared with the three-phase setting, the five-phase framework therefore provides additional evidence that uncertainty-based gating remains reliable under repeated label-space updates, while the main adaptation burden is absorbed by the density-based decision boundaries.

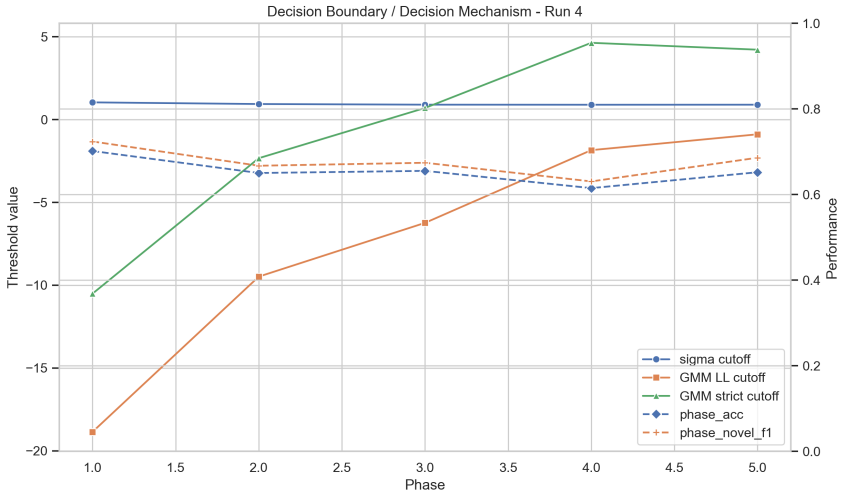

Figure 5.11: Evolution of the multi-signal decision mechanism across continual phases, showing uncertainty and GMM likelihood thresholds together with phase accuracy and Novel F1.

Clustering Analysis

Compared with the three-phase framework, the extension to five phases does not improve global clustering alignment. Therefore, introducing novel intents through additional sequential updates supports selective discovery but does not enable the model to reconstruct the fine-grained organization of the 100 novel intents. The qualitative inspection provides a more nuanced interpretation. Some discovered clusters show high local purity, with values around 0.91–0.97, indicating well-separated and internally consistent structures. Other clusters are more mixed, with purity ranging from ap proximately 0.60 to 0.88, while the remaining clusters exhibit lower purity values, between 0.19 and 0.50, reflecting stronger semantic overlap or incomplete separation. This explains the discrepancy between local observations and aggregate metrics: the framework can identify dense and

semantically coherent regions, but these regions are not consistently aligned with the complete set of true novel labels (Figure 5-12).

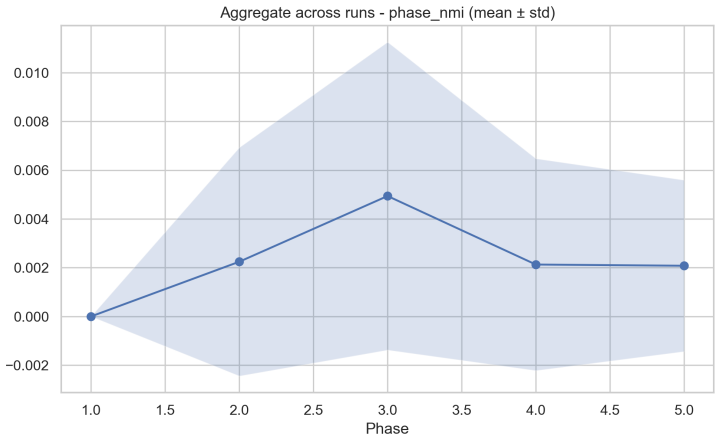

Figure 5.12: Clustering analysis showing consistently low alignment (NMI, ARI) across runs and phases, with only marginal improvements in later stages.

5.2.3. Comparison between the Three and Five-Phase Protocols

Table 5-3 provides a direct comparison of the final results obtained under the three-phase and five-phase continual settings. Extending the learning stream to five phases leads to a moderate reduction in global accuracy, while novelty detection becomes more consistent across runs, particularly in terms of precision, recall and Novel F1. However, the very low NMI and ARI values observed in both settings indicate that increasing the number of phases does not improve the recovery of the fine-grained novel-intent structure.

Table 5.3: Comparison of final results between the three-phase and five-phase frameworks.

Metric	Three-Phase Framework	Five-Phase Framework	Comparison
Global Accuracy	0.6296–0.6924	0.6043–0.6521	Slight decrease in the longer continual setting.
Novel Precision	0.7413–0.8791	0.8431–0.8776	Higher minimum value and lower variability.
Novel Recall	0.3898–0.6064	0.5072–0.5664	More stable detection of novel samples.
Novel F1	0.5410–0.6852	0.6378–0.6860	Higher and narrower performance range.
NMI	0.0000–0.0847	0.0000–0.0083	Very low in both settings.
ARI	0.0000–0.0027	≈ 0.0000	Negligible in both settings.

5.3. Ablation Studies

5.3.1. Clustering Strategy Comparison

To evaluate the sensitivity of the proposed framework to the choice of clustering backend, a controlled ablation was conducted in which the density-based discovery module (DP-GMM) was replaced with a centroid-based alternative: K-Means with the number of clusters selected via a silhouette-based heuristic ($K = 16$ in the reported experiment). All other components of the pipeline remain unchanged and cluster promotion follows the same conservative criteria in both cases, ensuring a consistent expansion of the intent space.

The comparison is conducted at Phase 2 of the continual learning cycle, where the clustering module is explicitly activated to perform novel intent discovery. From a computational standpoint, K-Means provides a moderate efficiency gain, reducing Phase-2 runtime from 157.94 seconds to 126.86 seconds (-19.7%). From a detection perspective, the two methods exhibit comparable short-term performance: K-Means achieves a novelty-detection F1-score of 0.6233, driven by a recall of 0.4731 and a precision of 0.9130. These results indicate that, under the same uncertainty-filtered input distribution, a centroid-based approach can approximate the decision boundary between known and unknown samples in the early discovery stage.

However, the key differences emerge in the discovery process itself. The proposed DP-

GMM-based approach promotes a larger number of clusters (25) compared to K-Means (8), suggesting a higher sensitivity to heterogeneous and irregular density structures in the latent space. This difference should not be interpreted as direct evidence of better semantic clustering quality, but rather as a different expansion behaviour: DP-GMM tends to identify and promote more candidate structures, whereas K-Means produces a more compact and computationally cheaper partition.

Unlike K-Means, which enforces rigid, spherical partitions around fixed centroids, the DP-GMM models the data as a flexible mixture of Gaussian components with adaptive covariance and component usage. This property is particularly relevant in open-world intent discovery, where emerging intents are not uniformly distributed and often occupy complex, uneven regions of the latent space. As a result, the DP-GMM is more consistent with the probabilistic and uncertainty-aware design of the proposed framework, although this does not imply that it reconstructs the ground-truth novel-intent taxonomy more accurately.

It is important to note that the final clustering alignment metrics remain near zero regardless of whether K-means or DP-GMM is used (NMI = 0.0009, ARI = 0.0000). This indicates that the low NMI and ARI values are not mainly determined by the specific clustering algorithm, but rather by the conservative discovery strategy adopted by the framework.

Overall, while K-Means represents a lightweight and computationally efficient alternative with competitive early detection performance, it lacks the probabilistic flexibility required for robust open-world discovery. The DP-GMM, as employed in the proposed framework, is therefore more consistent with the probabilistic and uncertainty-aware design of the system and better suited for modelling complex latent density patterns. Thus, this ablation supports the use of DP-GMM mainly from the perspective of modelling assumptions, discovery dynamics and label-space expansion behaviour, rather than as evidence of superior global clustering alignment (Table 5-4).

Table 5.4: Comparison of Clustering Strategies under the Continual Discovery Pipeline

Metric / Property	DP-GMM	K-Means	Interpretation
Clustering paradigm	Density-based (Bayesian mixture)	Centroid-based (partitioning)	DP-GMM is more flexible
Cluster selection	Adaptive (non-parametric truncation)	Fixed K=16 (silhouette)	K-Means requires explicit cluster selection
Promoted clusters	25	8	DP-GMM promotes more structures in the final Run 4 discovery step, while K-Means is evaluated only in Phase 2
Runtime	157.94 s	126.86 s	K-Means is approximately 20% faster in the Phase-2 ablation
Novelty Precision	0.8776	0.9130	K-Means is more conservative in the Phase-2 ablation
Novelty Recall	0.5631	0.4731	DP-GMM Run 4 achieves higher final novelty coverage
Novelty F1	0.6860	0.6233	DP-GMM Run 4 achieves a stronger final precision–recall balance
NMI	0.0009	0.0000	Fine-grained alignment remains very limited in both settings
ARI	0.0000	0.0000	Same limitation

5.3.2. Fallback Mechanism (ON vs OFF)

A targeted ablation was conducted to evaluate the impact of introducing the GMM fallback mechanism into the decision process in the open-world continual learning setting.

In the fallback-enabled variant, the fallback acts as a density-based recovery mechanism integrated into both pseudo-labelling and inference. The model learns a mapping between latent clusters and semantic labels (`cluster2intent`) using trusted samples, allowing uncertain inputs to be reassigned to known classes if they fall within high-likelihood regions of the known-intent GMM. In the baseline configuration, this reassignment is disabled and uncertain samples are not forcibly mapped back to known intents. The comparison therefore contrasts a strict rejection policy with a density-assisted assignment strategy.

The ablation was performed under a controlled configuration ($z_{dim} = 64$, $K_0 = 50$, Novel = 100, $\lambda_{EWC} = 7500$, Replay Hard = 0.95), with all other components unchanged. Introducing the fallback leads to a degradation in both global performance and novelty detection: accuracy decreases from 0.6521 to 0.6280 and Novel F1 drops from 0.6860 to 0.5592. This decline is due to reductions in both Novel Precision (0.8776 to 0.8295) and Novel Recall (0.5631 to 0.4218).

These results highlight the effect of the fallback mechanism: enabling it makes the system more permissive, allowing uncertain samples to be absorbed into the known-intent space, but at the cost of increased misassignments. In contrast, the baseline configuration enforces a more conservative decision rule, avoiding forced assignments and preserving higher precision and recall. The performance gap is therefore not due to improved detection of novel samples, but to the prevention of incorrect reassignment of ambiguous inputs.

Removing the fallback does not eliminate the underlying probabilistic modelling: the GMM remains fully active for uncertainty filtering, trusted-sample selection and DP-GMM-based discovery. Only the final density-based reassignment step is discarded. From a novelty detection perspective, enabling the fallback blurs the open-world decision boundary by overriding uncertainty signals, resulting in a less reliable separation between known and novel regions.

The ablation demonstrates that the fallback mechanism weakens decision reliability by introducing confident misassignments. For this reason, it is excluded from the final architecture and the configuration without fallback is adopted throughout this work (Table 5-5).

Table 5.5: Ablation of the GMM Fallback Mechanism

Method	GMM Fallback	Accuracy	Novel Precision	Novel Recall	Novel F1	Interpretation
Fallback variant	Enabled	0.6280	0.8295	0.4218	0.5592	More permissive; density-based reassignment to known intents
Main framework	Disabled	0.6521	0.8776	0.5631	0.6860	More conservative; fewer misassignments, improved reliability
Difference (OFF – ON)	–	+0.0241	+0.0481	+0.1413	+0.1268	Gain driven by improved novelty separation

5.4. Discussion of Results

5.4.1. Analysis of Experimental Findings

The experimental evaluation presented in this chapter investigates the proposed framework across multiple complementary dimensions. The results show that the framework is able to separate known from unknown samples and to maintain a stable behaviour across runs and phases.

In particular, the combination of replay, EWC regularization and uncertainty-aware filtering contributes to preventing abrupt performance degradation, leading to a gradual and controlled evolution over time. The persistently low NMI and ARI values can be explained by the conservative nature of the discovery mechanism: the framework does not aim to reconstruct the full fine-grained partition of all novel intents, but rather to promote only those latent regions that satisfy strict reliability, density and uncertainty-based criteria.

A key observation is the consistency of qualitative patterns across runs, despite variations in hyperparameters. While different configurations shift the balance between precision and recall, the global behaviour remains stable: novelty detection tends to be reliable but conservative, whereas clustering appears to be the most challenging component.

The ablation studies reinforce this interpretation, showing that modifications such as alternative clustering strategies or fallback mechanisms affect specific aspects of the system without fundamentally changing its overall behaviour.

These findings are consistent with known challenges in the literature, where detecting novelty is generally easier than reconstructing a fine-grained semantic structure, especially in high-dimensional latent spaces under continual learning constraints.

5.4.2. Interpretation of Model Behaviour

From a behavioural perspective, the framework follows a **conservative and reliability-oriented decision strategy**. Phase 1 initializes the known-intent structure, while the later continual phases process separated batches of novel intents. In these phases, predictions are not determined by the classifier alone, but by the joint effect of classifier confidence, latent uncertainty and density-based evidence.

Samples with low confidence, high uncertainty or weak compatibility with the known latent density are treated cautiously: they may be rejected, marked as unknown candidates or excluded from promotion. This reduces large misclassification errors and limits the propagation of incorrect pseudo-labels as the label space expands.

The model implements a controlled stability–plasticity compromise. Replay and EWC

preserve consolidated knowledge, while the uncertainty-aware decision pipeline regulates how new information enters the system. The goal is therefore not exhaustive assignment of all novel samples, but selective label-space expansion based on sufficiently reliable structures.

5.4.3. Implications of the Open-World Setting

The results should be interpreted within an open-world learning scenario, where the **set of classes is not fixed** and **new categories may emerge over time**.

In this context, the ability to reject uncertain samples is not a limitation, but a key requirement for maintaining system reliability and the absence of fallback assignment in the experimental setup reinforces this principle.

By avoiding forced assignment of uncertain inputs, the model reduces systematic errors and improves robustness, even at the cost of reduced coverage. This reflects a shift from traditional classification, where every sample must be assigned, to a more flexible paradigm where abstention is a valid outcome.

The framework therefore operates as a selective continual learner, prioritizing reliable decisions over exhaustive labelling. This behaviour is particularly relevant in realistic applications, where incorrect assignments can be more harmful than unassigned samples.

5.4.4. Comparative Analysis with Existing Intent Discovery Frameworks

Before comparing the proposed framework with existing methods, it is necessary to distinguish the learning scenario addressed in this thesis from the settings commonly considered in prior work. Most intent discovery approaches assume a static, semi-supervised or few-shot formulation, where the label space is fixed, the data are available in a single training stage, or the emergence of new classes is only partially modelled.

This thesis addresses a more demanding setting in which new intents appear across multiple phases, prior knowledge must be retained and the label space must expand without full retraining. Unlike most existing benchmarks, which treat new intent discovery, generalized category discovery, open-set recognition and continual learning as separate tasks, the proposed framework integrates these requirements into a single pipeline.

For this reason, a direct numerical comparison with state-of-the-art methods would be misleading. The following comparison is primarily methodological rather than purely

performance-based. It focuses on differences in assumptions, learning conditions, system capabilities and design principles (Table 5-6, Table 5-7).

Table 5.6: Methodological Comparison Across Different Intent Discovery Settings

Method	**Year**	**Task & Setting**	**Core Representation**	**Discovery Mechanism**
MTP-CLNN	2022-25	Static NID (unsupervised / semi-supervised)	BERT + multi-task pre-training	Contrastive learning with nearest-neighbor clustering
GID	2022	Static generalized discovery (IND/OOD, non-continual)	BERT encoder	End-to-end pseudo-OOD classification
CGID	2023	Continual generalized intent discovery (dynamic open-world)	BERT encoder + joint classifier	PLRD with prototype-guided pseudo-labeling, replay and feature distillation
CDI	2023	Incremental intent discovery (stage-wise, non open-world)	MPNet + contrastive learning	Iterative clustering with pseudo-label refinement
RAP	2024	Static semi-supervised NID	BERT + prototype learning	Prototype attraction/dispersion with clustering
IntentGPT	2024	Few-shot NID via in-context learning (ICL)	LLM + Sentence-BERT embeddings	Prompt-based grouping with clustering
Proposed Framework	2026	Multi-phase continual NID (open-world)	β-VAE latent space + classifier	DP-GMM density modelling with uncertainty-aware novelty detection

Table 5.7: Comparison of Uncertainty Modeling and Continual Learning Capabilities

Method	Year	Uncertainty Modeling	Pseudo-label Strategy	Continual Mechanisms	Label-Space Evolution	Key Strength
MTP-CLNN	2022-25	×	Neighborhood-based positives (contrastive)	×	×	Strong semantic representation learning
GID	2022	×	Alignment-based pseudo-labeling (IND/OOD split)	×	Implicit (static partition)	Joint IND/OOD modelling
CGID	2023	×	Prototype-guided pseudo-labeling	Replay memory + feature distillation	Explicit continual expansion	Dynamic open-world intent recognition
CDI	2023	×	Iterative pseudo-label refinement	LwF (knowledge distillation)	Partial (stage-wise expansion)	Controlled incremental discovery
RAP	2024	×	Robust pseudo-labeling (ITS strategy)	×	×	High clustering accuracy in static settings
IntentGPT	2024	×	Prompt-based labeling with feedback	×	Dynamic (prompt-dependent)	Training-free flexible discovery
Proposed Framework	2026	✓ (σ)	Multi-signal gating (confidence + uncertainty + likelihood)	Replay buffer + EWC regularization	Explicit dynamic expansion	Stable continual discovery under non-stationarity

As shown from the comparison tables, the proposed framework addresses a methodological gap that remains only partially covered by existing approaches. Previous methods achieve strong results in specific settings: MTP-CLNN and RAP focus on static new intent discovery, GID targets generalized discovery under a non-continual protocol, CDI considers stage-wise incremental discovery, IntentGPT exploits prompt-based few-shot grouping and CGID explicitly introduces continual expansion through replay and distillation. However, these methods generally optimize one specific aspect of the problem rather than jointly modelling uncertainty, density-based novelty evidence, continual retention and controlled label-space evolution.

The main advancement of the proposed framework lies in this joint formulation: it operates under a multi-phase open-world setting in which novel intents may appear repeatedly over time. Unlike few-shot or prompt-based methods, it does not require predefined support examples or labelled descriptions of emerging intents during discovery. Compared

with continual generalized discovery approaches such as CGID, the framework adds an explicit reliability layer: cluster promotion is not driven only by prototypes, pseudo-labels or classifier updates, but by the agreement between classifier confidence, posterior uncertainty and DP-GMM likelihood.

This makes uncertainty and density modelling central elements of the discovery process. Posterior uncertainty is used to select trusted samples, regulate pseudo-label acceptance and prevent ambiguous instances from being prematurely consolidated. DP-GMM likelihoods provide a density-based criterion for detecting deviations from known intent regions and for deciding whether candidate clusters have sufficient statistical support to become new labels. Replay and EWC then preserve previously learned knowledge while the classifier is expanded, allowing the system to adapt without full retraining.

A further distinction concerns the goal of discovery. Many clustering-oriented methods aim to recover the complete partition of unknown classes. The proposed framework instead follows a selective and reliability-oriented strategy: only clusters satisfying uncertainty, density and support constraints are promoted, while ambiguous samples are conservatively left unresolved. This design favours robust known–unknown discrimination and controlled label-space expansion over exhaustive taxonomy reconstruction.

The comparison highlights that the proposed framework extends the state of the art by treating continual new intent discovery as an open-world reliability problem. Its contribution is not limited to improving a single module, but consists in coordinating probabilistic latent representations, uncertainty-aware decision making, density-based discovery and stability-preserving continual learning within one unified multi-phase pipeline. A direct quantitative comparison with recent state-of-the-art models under the same multi-batch continual protocol remains an important direction for future work.

5.5. Chapter Summary

This chapter has critically examined the proposed framework through a detailed discussion of the empirical results, ablation studies and methodological limitations of the three-phase pipeline, followed by its extension to a five-phase setting. In doing so, it has provided a grounded assessment of both the framework's strengths and its current constraints. The results show that the integration of probabilistic latent representations, uncertainty-aware filtering and density-based clustering enables the model to operate under open-world conditions, in which both the data distribution and the label space evolve over time.

Across the five experimental runs, the framework maintains controlled performance con-

firming a conservative but reliable novelty detection behaviour. At the same time, the very low NMI and ARI values show that the framework is not able to reconstruct the full fine-grained taxonomy of novel intents. This is consistent with the design of the method: the discovery module prioritizes dense, reliable and low-uncertainty regions rather than exhaustive clustering of all novel classes. The qualitative inspection supports this interpretation, since some discovered clusters reach high purity, while others remain mixed or fragmented.

The ablation studies demonstrate that each core component contributes non-trivially to the overall performance. Their interaction appears to support a more stable and interpretable discovery process than simplified or partially disabled configurations. The results also show that the framework follows a conservative reliability-oriented strategy: novelty precision remains high, while recall and clustering metrics reflect the selective nature of the discovery mechanism rather than an exhaustive reconstruction of the full novel-intent taxonomy.

The chapter consolidates the experimental and conceptual contributions of the thesis, situating the proposed framework within the broader landscape of continual learning and intent discovery.

6 | Conclusion and Future Works

6.1. Conclusion

This thesis proposed a multi-phase framework for continual new intent discovery in open-world environments, where the set of possible intents is not fixed but evolves over time as new user needs emerge. The framework was designed as an adaptive and uncertainty-aware system that progressively updates its internal knowledge, expands its class space and consolidates previously acquired information. Rather than treating novelty detection, clustering and knowledge retention as independent tasks, the proposed approach integrates them into a unified continual process of learning, discovery, adaptation and consolidation.

Architecturally, the pipeline combines probabilistic representation learning, density-based modelling and continual-learning mechanisms. Input utterances are mapped into semantic embeddings and encoded through a lightweight β-VAE, which provides both latent representations and posterior uncertainty estimates. These representations support classification, DP-GMM-based modelling of known intents and the identification of uncertain or low-density regions as candidates for novelty. A multi-signal decision strategy combines classifier confidence, posterior uncertainty and GMM likelihood to detect potential novel intents. Candidate samples are then clustered and only reliable, dense and sufficiently supported clusters are promoted to new labels. These labels are integrated into the classifier and replay memory, while replay and Elastic Weight Consolidation help preserve prior knowledge and mitigate catastrophic forgetting across phases.

The framework introduces three main methodological contributions. First, it formalises continual new intent discovery as a multi-phase open-world problem in which the label space evolves through a recurrent cycle of novelty detection, reliability assessment, cluster promotion and integration. Second, it defines a controlled label-space expansion mechanism, where new intents are added only when supported by classifier confidence, posterior uncertainty and density-based evidence, while replay and Elastic Weight Consolidation preserve previously acquired knowledge. Third, it uses posterior uncertainty from the β-VAE as a central control signal, regulating trusted-sample selection, novelty admission,

pseudo-labelling, replay selection and information flow across phases.

Experimental results show that the framework achieves high novelty precision and maintains relatively stable behaviour across continual phases. The ablation studies further clarify the role of the main components: uncertainty gating encourages a more exploratory discovery process, while the GMM fallback reveals a trade-off between recovering difficult known samples and avoiding ambiguous misassignments. Although the low NMI and ARI values reveal a limitation of the current framework, they are consistent with its selective and reliability-oriented design. The objective is not to exhaustively reconstruct the full fine-grained taxonomy of all novel intents, but to promote reliable and well-supported local structures under uncertainty and continual-learning constraints. Qualitative cluster analyses provide additional insight beyond aggregate clustering metrics, showing that some promoted clusters correspond to dense and locally coherent semantic regions. Therefore, although global cluster-label alignment remains limited, the model demonstrates the ability to recover reliable local structures and integrate them into the evolving label space.

This thesis shows that continual intent discovery can be understood as a process of probabilistic adaptation. The proposed framework evolves its class space, latent structure and decision rules as new data arrive, combining uncertainty-aware representations, density-based modelling, selective cluster promotion and continual consolidation. In this way, open-world intent discovery is framed not only as a clustering task, but as an adaptive learning problem in which uncertainty and probabilistic evidence guide the integration of new intents over time.

6.2. Limitations

6.2.1. Implementation Challenges

The proposed framework is based on a tightly integrated architecture combining β-VAE latent modelling, uncertainty-driven filtering, dual DP-GMM modules, replay mechanisms, EWC regularization and dynamic classifier expansion.

This design introduces structural complexity, but enables a unified treatment of representation learning, uncertainty estimation and continual discovery within a single pipeline.

The overall behaviour emerges from the interaction among components, making the system configuration-sensitive. In particular, the effectiveness of the DP-GMM modules depends on the quality of the latent space, while uncertainty-driven filtering improves robustness by focusing the discovery process on reliable samples, even if this leads to

uneven data distributions across phases.

More generally, the framework is governed by **multiple interacting hyperparameters** which jointly define its operating regime. The observed variability across runs indicates that performance arises from their coordinated balance rather than from any single dominant factor.

These aspects reflect both the implementation challenges and the flexibility of the framework, highlighting its potential for adaptation and further optimization.

6.2.2. Threshold Sensitivity

A central limitation of the framework lies in its **strong sensitivity to decision thresholds**, which directly regulate novelty detection and cluster promotion. The novelty assignment is based on a multi-stage gating mechanism that improves robustness against naive overconfidence, it also introduces a complex decision surface that depends on multiple interacting thresholds.

Empirically, the results clearly show that different configurations lead to distinct operating regimes: precision-oriented settings enforce stricter thresholds, yielding high novelty precision but low recall, whereas recall-oriented settings relax these constraints, increasing coverage at the cost of more false positives.

This confirms that the framework inherently operates along a precision–recall trade-off curve, where small changes in threshold values can significantly alter final performance. Although the adaptive quantile mechanism stabilizes the uncertainty filtering process, this does not eliminate global sensitivity.

The remaining thresholds, particularly those governing GMM likelihoods and confidence-based filtering, remain effectively fixed or weakly adaptive. As a result, the system can be described as locally stabilized but globally threshold-dependent, with overall performance strongly influenced by manually tuned decision boundaries rather than purely learned representations.

6.2.3. Clustering Limitations

A relevant limitation of the framework concerns the **alignment between discovered clusters and ground-truth novel intent**, as reflected by consistently low NMI and ARI values across runs. While this representation effectively separates known from unknown samples, it provides limited semantic resolution for distinguishing among multiple unseen

intents. As a result, different novel classes may remain partially overlapping, making fine-grained clustering inherently challenging.

In addition, the framework prioritizes stability and precision through conservative mechanisms preventing catastrophic forgetting and reducing noisy cluster formation, but also limiting the plasticity required to separate sparsely represented novel intents.

Consequently, the discovery module tends to form dense and statistically robust clusters, while low-support intents are more likely to be merged into broader groups or remain underrepresented.

This behaviour induces a many-to-one mapping between true novel labels and discovered clusters, directly limiting cluster-label correspondence as measured by NMI and ARI.

Importantly, this does **not indicate clustering collapse**: qualitative inspection confirms that clusters are locally coherent. Instead, it reflects a fundamental trade-off between reliable novelty detection and the reconstruction of a fine-grained intent taxonomy.

Clustering should be interpreted as locally effective and adaptive across phases, contributing to selective discovery while providing a partial yet meaningful approximation of the underlying intent structure.

6.2.4. Dataset Bias

The interpretation of the results is influenced by both dataset characteristics and evaluation design.

Although the CLINC dataset offers a standardized benchmark, its intent space includes several semantically similar classes, which can make separation more challenging, particularly in open-world settings.

At the same time, the evaluation protocol reflects different dimensions of performance that are not fully aligned. While classification and novelty detection metrics capture the model's ability to distinguish known from unknown inputs, clustering metrics such as NMI and ARI focus on the internal organization of discovered samples.

These perspectives are complementary but do not necessarily correlate, leading to situations where effective novelty detection is not matched by equally structured clustering. Taken together, these aspects highlight the complexity of the task, while the proposed framework shows stable behaviour and supports controlled adaptation as new intents emerge.

6.3. Future Directions

The limitations identified in Section 6.2 highlight clear opportunities to extend the proposed framework toward a more expressive and scalable continual intent discovery system.

A first direction is the **explicit shaping of the latent space for semantic separability**. The current representation supports reliable known–unknown discrimination but does not organize novel intents at a sufficiently fine-grained level. Future work should therefore introduce structure-inducing objectives, such as contrastive or metric learning, to enforce inter-intent margins directly in the latent space. This would allow the density model to operate on a representation that is inherently more clusterable, reducing many-to-one mappings between true intents and discovered clusters and improving the alignment between discovered clusters and true intent semantics.

A second key direction is the **replacement of heuristic decision rules with a learned and unified decision mechanism for novelty detection**. Instead of combining confidence, uncertainty and likelihood through thresholds, these signals can be integrated into a single probabilistic model. This would enable data-driven decision boundaries improving robustness and reducing sensitivity to configuration choices.

A promising direction for future work is the **integration of large language models (LLMs)** with the proposed probabilistic continual intent discovery framework. In the current implementation, novel intents are detected and promoted exclusively through statistical criteria derived from semantic embeddings, β-VAE posterior uncertainty, classifier confidence and DP-GMM likelihood estimation. This design ensures that label-space expansion remains controlled by reliability-oriented mechanisms rather than by unconstrained generative predictions.

Future extensions could investigate the use of an LLM as a semantic interpretation layer operating after the discovery stage. Once a candidate cluster has satisfied the density, support, coherence and uncertainty requirements imposed by the probabilistic pipeline, an LLM could analyse representative utterances from that cluster and generate a concise human-readable intent description or propose an appropriate label. This would address an important practical limitation of the current framework, in which promoted clusters are identified structurally but are not automatically assigned meaningful semantic names.

An additional direction would be to employ LLMs in a human-in-the-loop validation process. Instead of directly determining whether a sample or cluster is novel, the language model could summarize ambiguous cases, compare discovered clusters with existing intent descriptions and support manual review before final deployment. In this configuration,

the statistical discovery module would preserve its role as the decision-making component, while the LLM would provide semantic assistance and improve interpretability.

This integration could therefore combine the reliability of uncertainty-aware density-based discovery with the semantic expressiveness of large language models. Future research should evaluate whether such a hybrid architecture improves cluster interpretability and validation efficiency without increasing false promotions or weakening the conservative behaviour of the continual discovery process.

Another crucial direction concerns the **adaptive regulation of stability–plasticity dynamics**. The observed sensitivity to replay and EWC parameters indicates that fixed policies are insufficient. Future work should move toward dynamic, data-dependent control mechanisms, where replay composition and regularization strength evolve with the learning phase and the level of novelty. This would allow the model to preserve past knowledge while remaining responsive to emerging intents, improving long-term continual performance.

From a computational perspective, scalable density modelling is essential because DP-GMM is the main bottleneck as the number of intents grows, limiting applicability to larger scenarios. Future work should therefore focus on **incremental or approximate inference strategies**, enabling efficient updates without compromising the probabilistic structure of the model.

These directions would allow the framework to operate effectively in realistic, heterogeneous and non-stationary environments. Advancing the framework along these axes would transform it from a conservative novelty detector into a fully structured, adaptive and semantically consistent continual learning system.

Bibliography

[1] C. Chandrakala, R. Bhardwaj, and C. Pujari, "An intent recognition pipeline for conversational ai," *International Journal of Information Technology*, vol. 16, no. 1, pp. 421–431, 2024. DOI: 10.1007/s41870-023-01584-6.

[2] Z. Ran, Y. Chen, Q. Jiang, and K. E, "Intent recognition in dialogue systems," in *Proceedings of the 2025 4th Asia-Pacific Artificial Intelligence and Big Data Forum (AIBD 2025)*, New York, NY, USA: Association for Computing Machinery, 2025, pp. 165–173. DOI: 10.1145/3718491.3718520.

[3] G. Arora, S. Jain, and S. Merugu, "Intent detection in the age of LLMs," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 1558–1571. DOI: 10.18653/v1/2024.emnlp-industry.114. arXiv: 2410.01627 [cs.CL]. [Online]. Available: https://aclanthology.org/2024.emnlp-industry.114.

[4] H. Weld, X. Huang, S. Long, J. Poon, and S. C. Han, "A survey of joint intent detection and slot-filling models in natural language understanding," *ACM Computing Surveys*, vol. 55, no. 3, pp. 1–38, Jul. 2022. DOI: 10.1145/3514227. [Online]. Available: https://doi.org/10.1145/3514227.

[5] G. Anderson, E. Hart, D. Gkatzia, and I. Beaver, "An open intent discovery evaluation framework," in *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Kyoto, Japan: Association for Computational Linguistics, Sep. 2024, pp. 760–769. DOI: 10.18653/v1/2024.sigdial-1.64. [Online]. Available: https://aclanthology.org/2024.sigdial-1.64.

[6] H. Zhang, H. Xu, T.-E. Lin, and R. Lyu, "Discovering new intents with deep aligned clustering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, May 2021, pp. 14 357–14 364. DOI: 10.1609/aaai.v35i16.17689. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17689.

[7] W. Shi, W. An, F. Tian, and Q. Zheng, "A diffusion weighted graph framework for new intent discovery," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore: Association for Computational Linguis-

tics, Dec. 2023, pp. 8126–8137. DOI: 10.18653/v1/2023.emnlp-main.499. [Online]. Available: https://aclanthology.org/2023.emnlp-main.499.

[8] C. Cheng, S. Cao, G. Tang, F. Ma, and D. Cui, "Dynamic BERT-reinforcement learning model for intent recognition in medical dialogue systems," *Informatica*, vol. 50, no. 5, 2026. DOI: 10.31449/inf.v50i5.10324. [Online]. Available: https://www.informatica.si/index.php/informatica/article/view/10324.

[9] Y. Zhang, H. Zhang, L.-M. Zhan, X.-M. Wu, and A. Lam, "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)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds., Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 256–269. DOI: 10.18653/v1/2022.acl-long.21. [Online]. Available: https://aclanthology.org/2022.acl-long.21/.

[10] B. Zhao, X. Wen, and K. Han, "Learning semi-supervised gaussian mixture models for generalized category discovery," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 16 577–16 587. DOI: 10.1109/ICCV51070.2023.01524. [Online]. Available: https://doi.org/10.1109/ICCV51070.2023.01524.

[11] Z. He, Y. Liu, and K. Han, "Category discovery: An open-world perspective," *arXiv preprint arXiv:2509.22542*, Sep. 2025. DOI: 10.48550/arXiv.2509.22542. arXiv: 2509.22542 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2509.22542.

[12] S. Vaze, K. Han, A. Vedaldi, and A. Zisserman, "Generalized category discovery," *arXiv preprint arXiv:2201.02609*, 2022. DOI: 10.48550/arXiv.2201.02609. arXiv: 2201.02609 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2201.02609.

[13] S. Zhang et al., "New intent discovery with attracting and dispersing prototype," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia: ELRA and ICCL, May 2024, pp. 12 193–12 206. [Online]. Available: https://aclanthology.org/2024.lrec-main.1067.

[14] Y. Mou et al., "Generalized intent discovery: Learning from open world dialogue system," in *Proceedings of the 29th International Conference on Computational Linguistics*, Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 707–720. [Online]. Available: https://aclanthology.org/2022.coling-1.59/.

[15] X. Song et al., "Continual generalized intent discovery: Marching towards dynamic and open-world intent recognition," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 4370–4382. DOI: 10.

18653/v1/2023.findings-emnlp.289. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.289/.

[16] M. Rawat, H. Sankararaman, and V. Barres, "Controllable discovery of intents: Incremental deep clustering using semi-supervised contrastive learning," in *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Nusa Dua, Bali: Association for Computational Linguistics, Nov. 2023, pp. 793–803. [Online]. Available: https://aclanthology.org/2023.ijcnlp-main.51.

[17] J. A. Rodriguez, N. Botzer, D. Vazquez, C. Pal, M. Pedersoli, and I. Laradji, "Intentgpt: Few-shot intent discovery with large language models," *arXiv preprint arXiv:2411.10670*, 2024. DOI: 10.48550/arXiv.2411.10670. arXiv: 2411.10670 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2411.10670.

[18] H. Dai, Y. Wu, K. He, and W. Xu, "Continual category discovery from a bayesian perspective," *arXiv preprint arXiv:2507.17382*, Jul. 2025. DOI: 10.48550/arXiv.2507.17382. arXiv: 2507.17382 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2507.17382.

[19] F. J. Cendra, X. Li, and K. Han, "Effective prompt pool learning for continual category discovery," *arXiv preprint arXiv:2407.19001*, 2024. DOI: 10.48550/arXiv.2407.19001. arXiv: 2407.19001 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2407.19001.

[20] A. Zhu, M. Hu, X. Wang, J. Yang, Y. Tang, and N. An, "Proxy-driven robust multimodal sentiment analysis with incomplete data," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 22 123–22 138. DOI: 10.18653/v1/2025.acl-long.1075. [Online]. Available: https://aclanthology.org/2025.acl-long.1075.

[21] Y. Ling, S.-J. Huang, and Z.-H. Zhou, "Bayesian domain adaptation with gaussian mixture domain-indexing," in *Advances in Neural Information Processing Systems*, vol. 37, Curran Associates, Inc., 2024, pp. 33 728–33 754. [Online]. Available: https://papers.nips.cc/paper_files/paper/2024/hash/9ebc79569f5e356b1ecfd1892d1b0a2e-Abstract-Conference.html.

[22] M. Hong, D. Jiang, Y. Song, and C. J. Zhang, "Neural-Bayesian program learning for few-shot dialogue intent parsing," *arXiv preprint arXiv:2410.06190*, Oct. 2024. DOI: 10.48550/arXiv.2410.06190. arXiv: 2410.06190 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2410.06190.

[23] R. Patil and P. K. Sharma, "Survey of text representation and embeddings: Evolution, challenges, and future directions," *IEEE Access*, vol. 11, pp. 33 734–33 764, 2023. DOI: 10.1109/ACCESS.2023.3262657. [Online]. Available: https://ieeexplore.ieee.org/document/10098736.

[24] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 6th. Upper Saddle River, NJ: Prentice Hall, 2007, ISBN: 978-0131877153.

[25] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations (ICLR)*, 2014. [Online]. Available: https://arxiv.org/abs/1312.6114.

[26] A. Asesh, "Variational autoencoder frameworks in generative AI model," in *2023 24th International Arab Conference on Information Technology (ACIT)*, IEEE, 2023, pp. 1–6. DOI: 10.1109/ACIT58814.2023.10444342. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10453782.

[27] M. Ebrahimabadi and S. M. J. Mashhadban, "Disentangled representation learning using β-VAE: A survey," *arXiv preprint arXiv:2208.04549*, Aug. 2022. DOI: 10.48550/arXiv.2208.04549. arXiv: 2208.04549 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2208.04549.

[28] R. Sridharan, *Gaussian mixture models and the EM algorithm*, Lecture Notes for 6.438 Estimation and Detection, Massachusetts Institute of Technology, 2014. [Online]. Available: https://www.semanticscholar.org/paper/Gaussian-mixture-models-and-the-EM-algorithm-Sridharan/efa8185bae6608669e26b2c5ade4f00fe96becb6.

[29] P. Bhattacharjee and P. Mitra, "A survey of density-based clustering algorithms," *Frontiers of Computer Science*, vol. 15, no. 1, p. 151 308, 2021. DOI: 10.1007/s11704-019-9059-3. [Online]. Available: https://link.springer.com/article/10.1007/s11704-019-9059-3.

[30] H. Shi et al., "Continual learning of large language models: A comprehensive survey," *arXiv preprint arXiv:2404.16789*, Apr. 2024. DOI: 10.48550/arXiv.2404.16789. arXiv: 2404.16789 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2404.16789.

[31] Y. Yang et al., "Recent advances in continual learning for foundation models: A survey," *arXiv preprint arXiv:2405.18653*, May 2024. DOI: 10.48550/arXiv.2405.18653. arXiv: 2405.18653 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2405.18653.

[32] L. Wang, X. Zhang, H. Su, J. Zhu, and Y. Zhong, "A comprehensive survey of continual learning: Theory, method and application," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5362–5383, 2024. DOI: 10.

1109/TPAMI.2024.3367344. [Online]. Available: https://ieeexplore.ieee.org/document/10444954.

[33] J. S. Smith et al., "Adaptive memory replay for continual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, IEEE, 2024, pp. 3605–3615. DOI: 10.1109/CVPRW63382.2024.00364. [Online]. Available: https://ieeexplore.ieee.org/document/10678401.

[34] J. Kim et al., "An efficient forgetting-aware fine-tuning framework for pretrained universal machine-learning interatomic potentials," *npj Computational Materials*, vol. 11, no. 1, p. 26, 2025. DOI: 10.1038/s41524-025-01895-w. [Online]. Available: https://www.nature.com/articles/s41524-025-01895-w.

[35] N. Thanh Tung and M. Park, "EL-GNN: A continual-learning-based graph neural network for task-incremental intrusion detection systems," *Electronics*, vol. 14, no. 14, p. 2756, Jul. 2025. DOI: 10.3390/electronics14142756. [Online]. Available: https://www.researchgate.net/publication/393599554_EL-GNN_A_Continual-Learning-Based_Graph_Neural_Network_for_Task-Incremental_Intrusion_Detection_Systems.

[36] C. Cheng, *Principal component analysis (PCA) explained visually with zero math*, Towards Data Science, Feb. 2022. [Online]. Available: https://towardsdatascience.com/principal-component-analysis-pca-explained-visually-with-zero-math-1cbf392b9e7d/.

[37] E. M. Dodds, J. A. Livezey, and M. R. DeWeese, "Spatial whitening in the retina may be necessary for V1 to learn a sparse representation of natural scenes," *bioRxiv*, Sep. 2019. DOI: 10.1101/776799. bioRxiv: 776799. [Online]. Available: https://www.biorxiv.org/content/10.1101/776799v1.

[38] P. Singh, "From autoencoders to β-VAE: A survey," Computational Science Research Center, San Diego State University, Technical Report, Dec. 2018. [Online]. Available: https://pdeep.xyz/documents/Autoencoders_Report.pdf.

[39] S. Markkandan, K. Aggarwal, K. Ashok, K. Selvakumarasamy, R. Kaushal, and M. Jadhav, "RETRACTED ARTICLE: Design of precoder for a MIMO–NOMA system using Gaussian mixture modelling," *Optical and Quantum Electronics*, vol. 56, no. 1, p. 128, Dec. 2023. DOI: 10.1007/s11082-023-05655-2. [Online]. Available: https://link.springer.com/article/10.1007/s11082-023-05655-2.

[40] J. Kim, D.-K. Kim, L. Logeswaran, S. Sohn, and H. Lee, "Auto-intent: Automated intent discovery and self-exploration for large language model web agents," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds., Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 16 531–16 541. DOI: 10.18653/v1/2024.

findings-emnlp.964. [Online]. Available: https://aclanthology.org/2024.findings-emnlp.964.

[41] M. De Raedt, F. Godin, T. Demeester, and C. Develder, "IDAS: Intent discovery with abstractive summarization," in *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, Y.-N. Chen and A. Rastogi, Eds., Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 71–88. DOI: 10.18653/v1/2023.nlp4convai-1.7. [Online]. Available: https://aclanthology.org/2023.nlp4convai-1.7.

[42] X. Shen, Y. Sun, Y. Zhang, and M. Najmabadi, "Semi-supervised intent discovery with contrastive learning," in *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, A. Papangelis, P. Budzianowski, B. Liu, E. Nouri, A. Rastogi, and Y.-N. Chen, Eds., Online: Association for Computational Linguistics, Nov. 2021, pp. 120–129. DOI: 10.18653/v1/2021.nlp4convai-1.12. [Online]. Available: https://aclanthology.org/2021.nlp4convai-1.12.

[43] J. Liang, L. Liao, H. Fei, and J. Jiang, "Synergizing large language models and pre-trained smaller models for conversational intent discovery," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 14 133–14 147. DOI: 10.18653/v1/2024.findings-acl.840. [Online]. Available: https://aclanthology.org/2024.findings-acl.840.

[44] X. Song et al., "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*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 10 291–10 304. DOI: 10.18653/v1/2023.emnlp-main.636. [Online]. Available: https://aclanthology.org/2023.emnlp-main.636.

[45] Y. Wu, Z. Chi, Y. Wang, and S. Feng, "MetaGCD: Learning to continually learn in generalized category discovery," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE/CVF, Oct. 2023, pp. 1655–1665. DOI: 10.1109/ICCV51070.2023.00159. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2023/html/Wu_MetaGCD_Learning_to_Continually_Learn_in_Generalized_Category_Discovery_ICCV_2023_paper.html.

[46] S. Ma, F. Zhu, Z. Zhong, W. Liu, X.-Y. Zhang, and C.-L. Liu, "HAPPY: A debiased learning framework for continual generalized category discovery," in *Advances in Neural Information Processing Systems (NeurIPS)*, A. Globerson et al., Eds., vol. 37, Curran Associates, Inc., 2024, pp. 50 850–50 875. DOI: 10.52202/079017-1609.

[Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/hash/5ae0f7cfd65d8e2b39da4177fef82015-Abstract-Conference.html.

[47] J. Han et al., "GOAL: Geometrically optimal alignment for continual generalized category discovery," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 6, pp. 4565–4573, Mar. 2026. DOI: 10.1609/aaai.v40i6.42456. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/42456.

[48] L. Wang et al., "Federated continuous category discovery and learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE/CVF, Oct. 2025, pp. 2429–2439. DOI: 10.1109/ICCV55074.2025.00227. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2025/html/Wang_Federated_Continuous_Category_Discovery_and_Learning_ICCV_2025_paper.html.

[49] C. Li, S. Wang, and H. Zhang, "Learning a fix and explore framework for continuous generalized category discovery," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 8, pp. 6064–6072, Mar. 2026. DOI: 10.1609/aaai.v40i8.37530. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/37530.

[50] F. Gao, W. Zhong, Z. Cao, X. Peng, and Z. Li, *OpenGCD: Assisting open world recognition with generalized category discovery*, Aug. 2023. DOI: 10.48550/arXiv.2308.06926. arXiv: 2308.06926 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2308.06926.

[51] X. Shi, Z. Guo, F. Nie, L. Yang, J. You, and D. Tao, "Two-dimensional whitening reconstruction for enhancing robustness of principal component analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 2130–2136, Oct. 2016. DOI: 10.1109/TPAMI.2015.2501810. [Online]. Available: https://ieeexplore.ieee.org/document/7331304.

[52] B.-J. Zhang, G.-H. Liu, Z. Li, and S.-X. Song, "Multistage PCA whitening: A robust method to dimensionality reduction in image retrieval," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2026. DOI: 10.1109/TNNLS.2026.3669538. [Online]. Available: https://ieeexplore.ieee.org/document/11478387.

[53] W. Chen, X. Wang, Z. Cai, C. Liu, Y. Zhu, and W. Lin, "DP-GMM clustering-based ensemble learning prediction methodology for dam deformation considering spatiotemporal differentiation," *Knowledge-Based Systems*, vol. 222, p. 106 964, Jun. 2021, ISSN: 0950-7051. DOI: 10.1016/j.knosys.2021.106964. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705121002276.

[54] X. Yan, Y. Gan, Y. Mao, Y. Ye, and H. Yu, "Live and learn: Continual action clustering with incremental views," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 15, pp. 16 264–16 271, Mar. 2024. DOI: 10.1609/aaai.v38i15.29561. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/29561.

[55] H. Ruan et al., "Unsupervised continual learning: A review of challenges, techniques, and evaluation metrics," *IEEE Access*, vol. 14, pp. 56 919–56 937, 2026. DOI: 10.1109/ACCESS.2026.3679348. [Online]. Available: https://ieeexplore.ieee.org/document/11458775.

[56] A. Aghasanli, Y. Li, and P. Angelov, "Prototype-based continual learning with label-free replay buffer and cluster preservation loss," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, IEEE/CVF, Jun. 2025, pp. 4064–4073. DOI: 10.1109/CVPRW64194.2025.00407. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2025W/DG-EBF/papers/Aghasanli_Prototype-Based_Continual_Learning_with_Label-free_Replay_Buffer_and_Cluster_Preservation_CVPRW_2025_paper.pdf.

[57] N. Masuyama, T. Takebayashi, Y. Nojima, C. K. Loo, H. Ishibuchi, and S. Wermter, *A parameter-free adaptive resonance theory-based topological clustering algorithm capable of continual learning*, 2023. arXiv: 2305.01507 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2305.01507.

Appendix A

A-1 Given two sentences s_i and s_j, their similarity is typically computed using cosine similarity:

$$\text{sim}(s_i, s_j) = \cos(z_i, z_j) = \frac{z_i \cdot z_j}{\|z_i\| \, \|z_j\|}$$

In this embedding space, semantically similar sentences are expected to satisfy:

$$\|z_i - z_j\| \ll \|z_i - z_k\|$$

where s_i and s_j are semantically related, while s_k is unrelated.

A-2 Given an input sentence $s = (w_1, \ldots, w_n)$, a transformer encoder produces contextualized token representations:

$$h_i = \text{Transformer}(w_1, \ldots, w_n)_i$$

A fixed-length sentence embedding is then obtained through a pooling operation, such as mean pooling:

$$z = \frac{1}{n} \sum_{i=1}^{n} h_i$$

or by selecting a special token representation (e.g., [CLS]):

$$z = h_{[\text{CLS}]}$$

A-3 Contrastive loss, which enforces that semantically similar sentences have closer embeddings than dissimilar ones:

$$L = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_k \exp(\text{sim}(z_i, z_k)/\tau)}$$

where τ is a temperature parameter and $s_i s_j$ is a positive pair.
Alternatively, in siamese architectures, the model learns a function $f(\cdot)$ such that:

$$\mathrm{sim}(s_i, s_j) \approx y_{ij}$$

where $y_{ij} \in [0, 1]$ encodes semantic similarity.

A-4 Given an observed variable x, a latent variable model introduces a hidden variable z and defines a joint distribution $p(x, z) = p(z)p(x \mid z)$. The latent variable z serves as a compact representation encoding the essential factors of variation that explain the observed data. The learning objective is to maximize the marginal likelihood $p(x)$, obtained by:

$$p(x) = \int p(x \mid z)p(z)\, dz$$

However, this integral is often intractable for complex models, especially when the likelihood $p(x \mid z)$ is parameterized by deep neural networks.

A-5

- Variational Objective (ELBO):

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)] - D_{KL}(q_\phi(z \mid x) \| p(z))$$

- β-VAE Objective:

$$\mathcal{L}_{\beta\text{-VAE}} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)] - \beta\, D_{KL}(q_\phi(z \mid x) \| p(z))$$

- Kullback–Leibler Divergence:

$$D_{KL}(q(z) \| p(z)) = \mathbb{E}_{q(z)} \log \frac{q(z)}{p(z)}$$

- Gaussian Assumptions:

$$q_\phi(z \mid x) = \mathcal{N}(\mu(x), \sigma^2(x)), \qquad p(z) = \mathcal{N}(0, I)$$

- Closed-form KL (Gaussian case):

$$D_{KL}(q_\phi(z \mid x) \| p(z)) = \frac{1}{2} \sum_{j=1}^{d} \left(\mu_j^2 + \sigma_j^2 - \log \sigma_j^2 - 1 \right)$$

- Reparameterization Trick:

$$z = \mu + \sigma \cdot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I)$$

A-6 Given a dataset $\{x_i\}_{i=1}^N$, a mixture model assumes that the probability density function can be expressed as:

$$p(x) = \sum_{k=1}^{K} \pi_k \, p(x \mid z = k)$$

where:

- K is the number of mixture components,
- π_k are the mixing coefficients, such that $\sum_{k=1}^K \pi_k = 1$ and $\pi_k \geq 0$,
- z is a latent variable indicating the component assignment,
- $p(x \mid z = k)$ is the component-specific distribution.

This formulation introduces a latent variable model where each observation is generated through a two-step process:

- Sample a component: $z \sim \text{Categorical}(\pi)$;
- Sample data: $x \sim p(x \mid z)$.

A-7 The probability density function of a GMM is given by:

$$p(x) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(x \mid \mu_k, \Sigma_k)$$

where:

- $\mu_k \in \mathbb{R}^d$ is the mean vector of the k-th component,
- $\Sigma_k \in \mathbb{R}^{d \times d}$ is the covariance matrix,
- π_k are the mixing coefficients.

Each Gaussian component is defined as:

$$\mathcal{N}(x \mid \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right)$$

The parameters of a GMM are therefore:

- Means μ_k: represent the center of each cluster,
- Covariances Σ_k: define the shape and orientation of clusters,
- Mixing coefficients π_k: represent the relative importance of each component.

A-8 Define:

$$N_k = \sum_{i=1}^{N} \gamma_{ik}$$

Then the parameter updates are:

- Mixing coefficients:

$$\pi_k = \frac{N_k}{N}$$

- Means:

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{ik} x_i$$

- Covariances:

$$\Sigma_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{ik}(x_i - \mu_k)(x_i - \mu_k)^T$$

Appendix B

B-1 Single-Batch Continual Learning Pipeline

Algorithm: Single-Batch Continual Learning Pipeline

Input: $D_0 = \{(x_i, y_i)\}$ (Initial labelled dataset with intents K_0)
B_1 (Single mixed batch containing known intents and masked novel intents);
M (Replay buffer, initially empty); $z_{\text{dim}} = 64$ (Latent space dimension)

Output: Updated model and expanded label space

Phase 1: Known Intent Learning;
Train β-VAE on D_0 to map $x_i \to q_\phi(z|x_i) = \mathcal{N}(\mu_i, \sigma_i^2)$ and initialize the latent structure;
Train classifier $P(y|z)$ over the initial known intent set K_0;
Fit the known-intent DP-GMM on trusted latent representations;
Update M with samples from D_0;
Construct a single mixed stream containing known samples from K_0 and masked novel samples from B_1;
Phase 2: Novelty Detection and Discovery;
for *each* x_j *in the mixed stream* **do**
    Encode x_j to obtain latent mean μ_j and uncertainty $\sigma(x_j)$;
    Evaluate novelty candidates through the multi-signal decision mechanism:;
    Classifier confidence: $\max_k P(y = k|\mu_j) < \tau_{\text{conf}}$;
    Uncertainty filtering: compare $\sigma(x_j)$ with the adaptive threshold τ_σ;
    Density estimation: compute DP-GMM log-likelihood in the latent space;
Apply the novelty discovery DP-GMM to candidate unknown samples;
Promote reliable dense clusters as discovered labels `__NOVEL_DISCOVERED__`k;
Phase 3: Consolidation and Expansion;
Update the label space: $K_1 = K_0 \cup K_{\text{new},1}$
Expand the classifier output layer to include newly discovered intents;
Construct the training set using current data, discovered pseudo-labelled samples and replay buffer M;
Update encoder and classifier using reconstruction loss, classification loss, replay and EWC regularization;
Update M using uncertainty-aware hard/easy replay selection;
Recalibrate adaptive uncertainty and density thresholds;

B-2 Multi-Batch Continual Learning Pipeline

Algorithm: Multi-Batch Continual Learning Pipeline

Input: $D_0 = \{(x_i, y_i)\}$, initial labelled dataset with intents K_0;
$\{N_b\}_{b=1}^{B}$, sequential novel macro-batches mixed with known intents;
M, replay buffer initially empty;
$z_{\text{dim}} = 64$, latent space dimension

Output: Updated model and expanded label space

Phase 1: Known intent learning;
Train the β-VAE on D_0 to map each input x_i to $q_\phi(z|x_i) = \mathcal{N}(\mu_i, \sigma_i^2)$ and initialize the latent structure;
Train the classifier $P(y|z)$ over the initial known intent set K_0;
Fit the known-intent DP-GMM on trusted latent representations;
Update the replay buffer M with samples from D_0;

for *each novel macro-batch* N_b, $b = 1, \ldots, N$ **do**
    Construct a mixed stream containing known samples from K_b and masked novel samples from B_b;
    Phase 2: Novelty detection and discovery;
    for *each sample* x_j *in the current mixed stream* **do**
        Encode x_j to obtain latent mean μ_j and uncertainty $\sigma(x_j)$;
        Evaluate novelty through the multi-signal decision mechanism;
        Compute classifier confidence $\max_k P(y = k|\mu_j)$ and compare it with τ_{conf};
        Compare $\sigma(x_j)$ with the adaptive uncertainty threshold τ_σ;
        Compute the DP-GMM log-likelihood of μ_j in the latent space;
    Apply the novelty-discovery DP-GMM to candidate unknown samples;
    Promote reliable dense clusters as discovered labels `__NOVEL_DISCOVERED__`k;
    Phase 3: Consolidation and expansion;
    Update the label space: $K_{b+1} = K_b \cup K_{\text{new},b}$;
    Expand the classifier output layer to include newly discovered intents;
    Construct the training set using current data, discovered pseudo-labelled samples and replay buffer M;
    Update the encoder and classifier using reconstruction loss, classification loss, replay and EWC regularization;
    Update M using uncertainty-aware hard/easy replay selection;
    Recalibrate adaptive uncertainty and density thresholds;

return *updated model, replay buffer* M *and expanded label space*;

B-3 Density-Based Discovery Model

Algorithm: Density-Based Discovery Model

Input: Latent representations (means μ_i, uncertainty σ_i) from β-VAE
Trusted labelled subset

Output: Known intent assignments
Promoted novel intent clusters

Step 1: Upstream Uncertainty Filtering;
for *each sample* $i = 1, \ldots, N$ **do**
    Retrieve deterministic latent mean μ_i;
    Filter μ_i through adaptive quantile-based trusted mask (using σ_i);
    if *sample i retained* **then**
        Use μ_i for density modelling of the known intent structure;
Step 2: Pre-mixture Dimensionality Reduction;
Project retained latent vectors through PCA transformation;
Retain 95% of the variance;
Step 3: Known Intent Density Module (Module 1);
Initialize truncated Bayesian Gaussian Mixture ($K_{max} = 200$);
Fit Module 1 using filtered latent means;
for *each mixture component k* **do**
    Associate k with intent label via majority assignment over trusted samples;
Compute per-sample log-likelihood scores;
Derive cutoffs using Median and MAD;
Step 4: Novelty Discovery Module (Module 2);
Identify candidate unknown samples;
Fit separate truncated Bayesian Gaussian Mixture ($K_{max} = 100$) on candidates;
Step 5: Cluster Promotion Procedure;
for *each candidate cluster C in Module 2* **do**
    if *criteria satisfied* **then**
        Promote C as a New Intent;
return known assignments, adaptive thresholds and promoted new intents;

B-4 Novelty Detection Strategy

Algorithm: Novelty Detection Strategy

Input: Incoming utterances

Classifier softmax posteriors

Latent uncertainty scores σ_i

Labelled set L_t Initial quantile $\mathrm{q}_0 = 0.90 Target trusted ratio$

Output: Known samples and novelty candidates

Compute classifier confidence;

for *each labelled sample* $i \in L_t$ **do**

  $\sigma_i = (1/d_z) \sum_j \exp(1/2 \log var_{ij})$;

Update quantile q_t;

$T_t = \{i \in L_t : \sigma_i < Q_{q_t}(\sigma)\}$ trusted subset;

Estimate density of known intents;

Identify low-density samples;

if *decision signals sufficient* **then**

  Assign as known;

else

  Treat as novelty candidate;

Process candidates through discovery module;

Promote clusters if criteria satisfied;

B-5 Continual Stability Mechanisms

Algorithm: Continual Stability Mechanisms (Replay + EWC)

Input: D_k, D, σ_i, τ_{low}, M

Output: Updated model

Partition past samples: $\mathrm{R}_{\text{easy}} = \{i \in D : \sigma_i \leq \tau_{\text{low}}\}, \quad R_{\text{hard}} = \{i \in D : \sigma_i > \tau_{\text{low}}\}$

Construct replay buffer: $\mathrm{R} = \mathrm{R}_{\text{easy}} \cup R_{\text{hard}}, \quad |R| = M$

Form training set: $_k = D_k \cup R$

Train model with replayed samples;

Apply EWC regularization;

Update model parameters;

B-6 Label Expansion and Cluster Promotion

Algorithm: Label Expansion and Cluster Promotion

Input: Clusters $\{C_k\}$, known intents K_t

Output: Updated label set $\mathrm{K}_{t+1} and expand classifier$

Initialize $K_{new} := \emptyset$
for *each* C_k **do**
    Evaluate support;
    Evaluate purity;
    Evaluate separation;
    Verify low-uncertainty;
    if *criteria satisfied* **then**
        Assign new label y^k_{new};
        Add to K_{new};
Expand classifier from $|K_t|$ to $|K_t| + |K_{new}|$;
Update label space: $K_{t+1} \leftarrow K_t \cup K_{new}$;

B-7 Qualitative Cluster Inspection Module

Algorithm: Qualitative Cluster Inspection Module

Input: Clusters $\{C_k\}$ and labels $\{y_i\}$

Output: Cluster descriptors

for *each* $\{C_k\}$ **do**
    Compute: support$(\{C_k\}) = |\{C_k\}|$;
    Compute label distribution;
    Determine dominant label;
    Compute purity: purity$(C_k) = \frac{\max_y(|\{i \in C_k : y_i = y\}|)}{|C_k|}$;
    Analyze distribution;
    Select representative examples;
Return descriptors and examples;

Appendix B-8 Full Continual Learning Cycle

Algorithm: Full Continual Learning Cycle

Input: Model, initial labelled dataset D_0, sequential macro-batches $\{B_b\}_{b=1}^{N}$, replay buffer M

Output: Updated model, expanded intent space, refined latent representation and decision boundaries

Phase 1: Known Intent Learning (Initialization);
Train the β-VAE and classifier on the initial known intents K_0;
Establish the initial latent representation, known-intent density model and decision boundaries;
Initialize the replay buffer M with samples from D_0;
for *each macro-batch* N_b, $N = 1, \ldots, N$ **do**
    Phase 2: Open-World Processing;
    Process the current mixed batch containing known samples and masked novel samples;
    Apply the multi-signal decision mechanism based on classifier confidence, latent uncertainty and DP-GMM likelihood;
    Identify candidate novel samples under the open-world assumption;
    Filter candidates using the adaptive uncertainty threshold τ_σ;
    Organize reliable candidates through the novelty discovery DP-GMM;
    Phase 3: Consolidation and Expansion;
    Validate discovered clusters using density, support and internal consistency criteria;
    Promote reliable clusters as new intent classes `__NOVEL_DISCOVERED__`k;
    Expand the classifier output space according to the updated intent set;
    Retrain the model using current samples, pseudo-labelled discovered samples and replayed samples from M;
    Apply replay and EWC regularization to preserve previously learned intents;
    Update the replay buffer through uncertainty-aware hard/easy sample selection;
    Recalibrate uncertainty and density thresholds;
    Refine latent representation and decision boundaries;
    Expand the intent space when new clusters are accepted;

List of Figures

List of Tables

Glossary

ARI	Adjusted Rand Index
Beta-VAE	Beta-Variational Autoencoder
CCD	Continual Category Discovery
CDI	Controllable Discovery of Intents
CGID	Continual Generalized Intent Discovery
DI-Parser	Dialogue-Intent Parser
DP-GMM	Dirichlet Process Gaussian Mixture Model
ELBO	Evidence Lower Bound
EM	Expectation-Maximization
EWC	Elastic Weight Consolidation
GCD	Generalized Category Discovery
GID	Generalized Intent Discovery
GMDI	Gaussian Mixture Domain Indexing
GMM	Gaussian Mixture Model
LLM	Large Language Model
MAD	Median Absolute Deviation
NID	New Intent Discovery
NLP	Natural Language Processing
NLU	Natural Language Understanding
NMI	Normalized Mutual Information
NN	Neural Networks
OOD	Out-of-Distribution Detection
OOS	Out-of-Scope
PCA	Principal Component Analysis
PLRD	Prototype Guided Learning with Replay and Distillation

P-RMF	Proxy Driven Robust Multimodal Fusion
RAP	Robust and Adaptive Prototypical learning
VAE	Variational Autoencoder
VB-CGCD	Variational Bayes Continual Generalized Category Discovery

List of Symbols

Variable	Description	Unit
$\mathcal{D}_t$	Dataset processed at phase t	–
z_i	Latent representation of sample i	–
z_{dim}	Latent-space dimensionality	–
$\boldsymbol{\mu}_i$	Posterior latent mean of sample i	–
σ_i	Scalar posterior uncertainty of sample i	–
$q_\phi(z \mid x)$	Approximate latent posterior	–
ϕ	Variational encoder parameters	–
θ	Trainable model parameters	–
$\mathcal{L}_{\text{VAE}}$	Total β-VAE loss	–
$\mathcal{L}_{\text{rec}}$	Reconstruction loss	–
$\mathcal{L}_{\text{KL}}$	KL regularisation loss	–
β	KL regularisation coefficient	–
π_k	Weight of DP-GMM component k	–
$\boldsymbol{m}_k$	Mean of DP-GMM component k	–
$\boldsymbol{\Sigma}_k$	Covariance of DP-GMM component k	–
γ_{ik}	Responsibility of component k for sample i	–
ℓ_i	DP-GMM log-likelihood of sample i	–
τ_σ	Uncertainty threshold for trusted samples	–
q_t	Adaptive uncertainty quantile	–
τ_{conf}	Classifier-confidence threshold	–
τ_ℓ	Log-likelihood threshold for novelty detection	–
C_k	Candidate novel cluster k	–

Variable	Description	Unit
$\bar{\ell}_k$	Mean log-likelihood of cluster k	–
$\bar{\sigma}_k$	Mean uncertainty of cluster k	–
$n_{\min}$	Minimum support for cluster promotion	–
$\mathcal{R}$	Replay buffer	–
M	Replay-buffer capacity	–
$\mathcal{L}_{\text{EWC}}$	EWC regularisation loss	–
λ_{EWC}	EWC regularisation strength	–
F_i	Fisher importance of parameter θ_i	–
$\theta_i^\star$	Consolidated value of parameter θ_i	–
$\text{Precision}_{\text{novel}}$	Novelty-detection precision	–
$\text{Recall}_{\text{novel}}$	Novelty-detection recall	–
$F1_{\text{novel}}$	Novelty-detection F1-score	–