IJWR

Adaptive Ensemble Thresholding for OOD Intent Detection

Masoud Akbari *, Ali Mohades, M. Hassan Shirali-Shahreza
Department of Mathematics and Computer Science, Amirkabir University of Technology, Tehran, Iran; ma.akbari421@aut.ac.it, mohades@aut.ac.ir, hshirali@aut.ac.ir

ABSTRACT

Out-of-domain intent detection in natural language understanding systems faces significant challenges from suboptimal threshold selection and signal degradation through inappropriate normalization techniques. This paper presents an adaptive ensemble thresholding framework that substantially extends our previous conference work by addressing fundamental limitations in existing variational autoencoder-based detection methods. Our approach combines reconstruction loss from variational autoencoders with classifier confidence scores to create a unified detection signal that captures both semantic deviation and prediction uncertainty. The framework incorporates a novel smart scaling strategy that preserves natural separation ratios between in-domain and out-of-domain samples, preventing the signal destruction caused by standard normalization approaches. Through systematic parameter optimization using grid search techniques, the method adaptively determines optimal ensemble weights and threshold selection strategies tailored to specific dataset characteristics. We evaluate our framework across multiple datasets with varying semantic complexity and domain structures, demonstrating consistent performance improvements over baseline variational autoencoder approaches and recent state-of-the-art methods. Compared to our previous VAE-based approach, the framework demonstrates an average performance gain of 3.15 percentage points across all evaluation metrics. Our analysis reveals that ensemble scaling strategy significantly impacts detection performance, with proper signal preservation being more critical than sophisticated threshold selection methods. This work provides a principled approach to adaptive ensemble learning for out-of-domain detection, offering a robust solution that generalizes effectively across diverse datasets and linguistic contexts including low-resource languages like Persian.

Keywords*— Natural Language Understanding, Out-of-Domain Intent Detection – Adaptive Thresholding – Ensemble Learning*

1. Introduction

The rapid evolution of conversational AI systems has fundamentally transformed human-computer interaction across diverse domains, from customer service automation to sophisticated personal assistants and industrial applications [1]. These systems rely heavily on Natural Language Understanding (NLU) modules to accurately interpret user intents and extract meaningful semantic information from conversational inputs [2]. Modern task-oriented dialogue architectures represent a complex integration of three interconnected components: Natural Language Understanding for intent recognition and slot filling, Dialogue Management for conversation flow control and context maintenance, and Natural Language Generation for contextually appropriate response formulation [3], with NLU applications expanding to specialized industrial domains [4]. As these systems increasingly extend beyond traditional consumer applications into sectors such as automotive software analytics [5], their operational robustness and reliability requirements have grown exponentially.

The deployment of dialogue systems in real-world environments presents unprecedented challenges in maintaining service quality and user trust when confronted with unexpected or out-of-scope inputs. Intent detection, serving as the foundational component of NLU pipelines, faces significant operational challenges when encountering Out-of-Domain (OOD) and Out-of-Scope (OOS) inputs that extend beyond the system's predefined operational boundaries [6]. These challenges have become increasingly critical as

http://dx.doi.org/10.22133/ijwr.2025.528472.1295
Citation M. Akbari, A. Mohades, M. H. Shirali-Shahreza , "Adaptive Ensemble Thresholding for OOD Intent Detection", *International Journal of Web Research*, vol.8, no.4,pp.25-39, 2025, doi: http://dx.doi.org/10.22133/ijwr.2025.528472.1295.

*Coressponding Author
Article History: Received: 5 June 2025; Revised: 16 September 2025; Accepted: 27 September 2025.

dialogue systems are deployed in safety-critical domains where incorrect intent classification can have serious consequences for system reliability and user safety.

Recent technological advances in large language models and transformer-based architectures have substantially enhanced intent classification capabilities across various domains [7, 8]. However, these improvements have simultaneously highlighted the fundamental challenge of detecting when user queries exceed system capabilities—a problem that becomes increasingly complex in multi-domain environments where semantic boundaries between in-domain and out-of-domain samples can be subtle and context-dependent. The emergence of context-aware OOD detection frameworks that consider multi-turn dialogue contexts demonstrates the evolving complexity of this problem space and the need for more sophisticated detection mechanisms [9].

Contemporary research in OOD detection emphasizes that effective systems must not only identify unknown intents with high precision but also maintain superior accuracy for in-domain classifications while minimizing false rejections that could degrade user experience [10]. Recent comprehensive surveys have highlighted the broader challenges of out-of-distribution generalization in natural language processing, emphasizing the systematic biases that can artificially inflate model performance and the need for more robust evaluation frameworks across different domains and deployment contexts [11]. This dual requirement creates a complex optimization problem that traditional binary classification approaches often fail to address adequately. The challenge is further compounded by the need for systems to operate reliably across diverse linguistic contexts, user populations, and application domains.

Current methodological approaches to OOD intent detection can be broadly categorized into training-driven and training-agnostic methodologies, with recent comprehensive surveys highlighting the growing importance and effectiveness of ensemble-based techniques in addressing the limitations of individual detection methods [12, 13]. Reconstruction-based methods employing autoencoders and variational autoencoders have demonstrated promising results in learning robust representations of in-domain data distributions, though they often encounter significant challenges with optimal threshold selection and signal optimization processes [14, 15]. These challenges become particularly pronounced when dealing with datasets that exhibit varying semantic complexity or when deployed across different linguistic contexts.

Probability-based approaches that model in-domain data distributions through various statistical techniques face inherent challenges with distribution sensitivity and often produce unreliable likelihood estimates when confronted with high-dimensional semantic spaces or domain shift scenarios [16]. Distance-based techniques, while conceptually straightforward, encounter computational and interpretability difficulties in high-dimensional semantic spaces where meaningful distance metrics can be challenging to define and optimize [16]. The integration of advanced embedding techniques with uncertainty quantification methodologies has emerged as a particularly promising research direction, especially for handling semantically similar intents that share surface-level linguistic characteristics but represent fundamentally different user intentions [17].

Variational Autoencoder (VAE) based approaches have gained significant research attention for OOD detection due to their theoretical foundation in probabilistic modeling and their demonstrated ability to learn robust probabilistic representations that are relatively independent of specific input data distributions [18, 19]. These methods leverage the reconstruction error as a natural indicator of how well a given input conforms to the learned in-domain data distribution. However, recent empirical studies have identified several critical limitations in VAE-based OOD detection systems, particularly concerning reconstruction error thresholding strategies and the counterintuitive phenomenon where certain OOD samples can receive lower reconstruction errors than legitimate in-domain samples [20, 21].

The challenge of optimal threshold selection remains a persistent and fundamental issue in VAE-based detection systems. Fixed percentile-based thresholds, while computationally efficient and easy to implement, often fail to adapt effectively to dataset-specific characteristics and can lead to suboptimal separation between in-domain and OOD samples across different deployment contexts [22]. This limitation becomes particularly problematic in production environments where data distributions may shift over time or when systems are deployed across multiple domains with varying characteristics.

Furthermore, standard normalization techniques commonly applied to reconstruction errors can inadvertently destroy natural separation signals inherent in the data, thereby reducing the discriminative power of the underlying VAE model and compromising overall detection performance [22]. This signal degradation problem represents a fundamental challenge in the preprocessing pipeline that has received insufficient attention in existing literature, despite its significant impact on system performance. The broader landscape of natural language processing continues to evolve rapidly

with advances in deep learning and large language models, creating both new opportunities and challenges for robust OOD detection systems [23].

This work represents a comprehensive extension and significant advancement of our previous research on hybrid architectures for OOD intent detection and intent discovery [24]. While our initial conference framework successfully demonstrated the effectiveness of VAE-based OOD detection combined with unsupervised clustering techniques for intent discovery, subsequent detailed analysis and real-world deployment experience revealed substantial opportunities for improvement in the threshold selection mechanisms, signal processing components, and overall system robustness. The original approach employed fixed reconstruction error thresholds determined through percentile-based methods, which proved inadequate for achieving optimal performance across diverse datasets, linguistic contexts, and application domains.

Additionally, our original framework's exclusive reliance on reconstruction error signals, while theoretically sound and mathematically principled, left unexploited the rich complementary information available from classifier confidence scores and other uncertainty measures. This limitation became particularly apparent when deploying the system across different user populations and query types, where the combination of multiple information sources could provide more robust and reliable detection capabilities.

To address these identified limitations and advance the state-of-the-art in OOD detection, we propose an innovative adaptive ensemble thresholding framework that fundamentally reconceptualizes OOD detection as a multi-signal optimization problem rather than a single-metric classification task. Our comprehensive methodology introduces three key technical innovations that collectively address the limitations identified in existing approaches.

First, we develop an advanced ensemble approach that intelligently combines VAE reconstruction losses with classifier confidence scores through sophisticated dataset-adaptive weighting schemes. This ensemble methodology goes beyond simple linear combinations by incorporating domain-specific knowledge and adaptive learning mechanisms that can adjust to varying dataset characteristics and deployment contexts.

Second, we introduce a novel smart scaling strategy that preserves natural separation ratios inherent in reconstruction errors rather than applying conventional normalization techniques that can destroy crucial discriminative information. This approach maintains the semantic relationships between different data points while enabling effective comparison and combination of signals from different sources.

Third, we implement a systematic parameter optimization framework that automatically determines optimal ensemble weights and threshold values for specific datasets and application domains, eliminating the need for manual parameter tuning and reducing deployment complexity. This optimization framework incorporates advanced search techniques and cross-validation strategies to ensure robust performance across different scenarios.

The framework demonstrates exceptional effectiveness in cross-lingual scenarios, showing substantial performance improvements for low-resource languages while maintaining competitive performance on high-resource datasets. This cross-lingual capability is particularly important for global deployment of dialogue systems and represents a significant advancement over existing approaches that often struggle with linguistic diversity.

The primary contributions of this work include several significant technical and practical advances: (1) identification and systematic resolution of signal degradation issues in VAE-based OOD detection through novel scaling methodologies that preserve critical discriminative information, (2) development of a comprehensive adaptive ensemble framework that effectively leverages complementary information from reconstruction and confidence signals through intelligent weighting mechanisms, (3) introduction of systematic parameter optimization techniques that eliminate manual threshold tuning requirements and enable automated deployment across diverse contexts, (4) comprehensive experimental validation across multiple languages including English and Persian datasets demonstrating average performance improvements of 3.96% across all evaluation metrics, and (5) detailed analysis of dataset-specific adaptation patterns that provide valuable insights for future OOD detection system design and deployment strategies.

Additionally, we provide extensive ablation studies examining the impact of different scaling strategies and ensemble weighting schemes, offering practical guidance and theoretical insights for practitioners implementing OOD detection systems in production environments. These studies reveal important trade-offs between different approaches and provide empirical evidence for the design decisions incorporated in our framework.

The remainder of this paper is structured to provide comprehensive coverage of our methodology and findings. Section 2 reviews related work in OOD detection methodologies, ensemble learning approaches, and threshold selection strategies, positioning our contributions within the broader research landscape and highlighting the novel aspects of our approach. Section 3 presents our proposed adaptive ensemble thresholding methodology, including detailed mathematical formulations, algorithmic descriptions, and theoretical justifications for our design choices. Section 4 describes our comprehensive experimental setup, including detailed descriptions of datasets, baseline comparison methodologies, evaluation metrics, and detailed experimental results and analysis, including comparative performance evaluation, ablation studies examining individual component contributions, and cross-lingual performance assessment across different language contexts. Finally, Section 5 concludes with a comprehensive discussion of our findings, analysis of practical implications for real-world deployment, identification of current limitations, and directions for future research in adaptive ensemble learning for OOD detection.

2. Related Works

The task of Out-of-Domain (OOD) intent detection has garnered significant attention in recent years as dialogue systems increasingly deploy in real-world environments where users may express intents beyond the system's predefined capabilities. This section provides a comprehensive review of existing approaches to OOD detection, with particular focus on VAE-based methods, threshold selection strategies, and ensemble techniques that inform our proposed adaptive framework.

2.1. Taxonomy of OOD Detection Approaches

Recent comprehensive surveys [8, 12] have established a fundamental taxonomy for OOD detection methods, categorizing them based on their training paradigms and data requirements. This categorization provides a systematic framework for understanding the evolution of OOD detection techniques and their relative strengths and limitations.

Approaches with Only In-Domain Data

When OOD training data is unavailable, methods must rely exclusively on modeling the in-domain distribution. Reconstruction-based approaches have emerged as a prominent technique in this category [25-27]. These methods leverage autoencoders and generative models to detect OOD samples by analyzing reconstruction quality, operating under the assumption that models trained on in-domain data will struggle to reconstruct out-of-distribution inputs effectively. Zhou [26] introduces an auxiliary module to extract activations of feature vectors, aiding the model in constraining the latent reconstruction space to filter potential OOD data. Recent work by Li et al. [27] demonstrates that masked image modeling can be effectively leveraged for OOD detection, showing significant advantages in learning the internal distribution of data.

Probability-based approaches constitute another major category, focusing on modeling the likelihood distribution of in-domain data [28],[29]. Du et al. [28] propose SIREN, which shapes representations for detecting out-of-distribution objects, while Pei [29] demonstrates that image background can serve as a good proxy for out-of-distribution data. These methods often face challenges with distribution sensitivity and may produce unreliable likelihood estimates, particularly when confronted with high-dimensional data or complex semantic spaces.

Logits-based techniques analyze the output confidence scores of neural networks to identify OOD samples [30]. Liu et al. [30] propose unsupervised out-of-distribution detection with diffusion inpainting, leveraging generative models to improve detection capabilities. These approaches typically establish confidence thresholds below which samples are classified as out-of-domain. However, recent studies have shown that neural networks can exhibit overconfidence on OOD inputs, necessitating careful calibration of confidence scores.

OOD synthesis methods attempt to generate pseudo-OOD samples during training to improve detection capabilities [31-34]. Gao et al. [31] introduce DIFFGUARD, which uses semantic mismatch guidance with pre-trained diffusion models. Wei et al. [32] address neural network overconfidence through logit normalization, while Tao et al. [33] propose non-parametric outlier synthesis techniques. Liu et al. [34] extend this work to large-scale long-tailed recognition in open-world scenarios.

Approaches Leveraging Both ID and OOD Data

When real OOD data is available during training, more sophisticated approaches become feasible. Boundary regularization methods explicitly optimize decision boundaries between in-domain and out-of-domain regions [35]. Lu et al. [35] propose learning with mixture of prototypes for out-of-distribution detection, which explicitly models the decision boundary using OOD samples.

Outlier exposure techniques directly incorporate real OOD samples during training [36], allowing models to learn explicit representations of out-of-distribution data. This approach has shown significant improvements in detection performance,

particularly when the OOD training data is representative of test-time OOD samples.

Distance-based approaches focus on learning discriminative feature spaces where ID and OOD samples are well-separated [36]. Regmi et al. [36] introduce ReweightOOD, which employs loss reweighting strategies for distance-based OOD detection, demonstrating improved performance on challenging benchmarks.

Meta-learning based approaches, particularly those employing Model-Agnostic Meta-Learning (MAML), have shown promise for rapid adaptation to new OOD detection scenarios with minimal examples [37]. Rahimi and Veisi [37] demonstrate the integration of model-agnostic meta-learning with advanced language embeddings for few-shot intent classification, showing particular value in multilingual contexts where training data may be limited.

2.2. VAE-Based OOD Detection Methods

Variational Autoencoders have emerged as a powerful tool for OOD detection due to their probabilistic framework and ability to learn robust latent representations. An and Cho [38] provide foundational work on variational autoencoder based anomaly detection using reconstruction probability, establishing the theoretical basis for VAE-based OOD detection.

Recent advances in VAE-based OOD detection have addressed several key challenges. Memory-augmented VAEs incorporate external memory modules that store prototypical patterns of normal data distributions, enabling more effective discrimination between ID and OOD samples [18]. The memory mechanism allows the model to maintain a repository of in-domain patterns, against which new inputs can be compared during inference.

The challenge of VAE overestimation in OOD detection has been thoroughly investigated [39], revealing that this phenomenon arises from improper prior distribution design and gaps in dataset entropy-mutual integration between ID and OOD datasets. The AVOID framework proposes post-hoc prior calibration and dataset entropy-mutual calibration techniques to mitigate these issues, demonstrating significant improvements in unsupervised OOD detection performance.

Compression techniques for VAE-based OOD detectors have been explored to enable deployment on resource-constrained embedded systems [40]. These approaches apply quantization, pruning, and knowledge distillation while maintaining detection performance, demonstrating that VAE reconstruction losses remain informative even after significant model compression.

The application of VAEs in cyber-physical systems has introduced novel approaches using β-VAE architectures [41]. These methods leverage the disentangled representations learned by β-VAEs to identify OOD inputs based on KL-divergence scores and implement runtime detection pipelines using martingale theory and CUSUM statistics for continuous monitoring.

Recent theoretical work has reinterpreted VAEs through the lens of fast and slow weights, proposing the Likelihood Path (LPath) principle [42]. This approach selects sufficient statistics that form the path toward likelihood estimation, achieving state-of-the-art OOD detection performance even when the likelihood itself proves unreliable.

2.3. Threshold Selection and Adaptive Strategies

The selection of appropriate thresholds for OOD detection remains a critical challenge across all detection methods. Fixed percentile-based thresholds, while simple to implement, often fail to adapt to dataset-specific characteristics and can lead to suboptimal performance [22, 43]. Zheng et al. [43] investigate out-of-domain detection for natural language understanding in dialog systems, highlighting the importance of adaptive threshold selection.

Class-wise thresholding approaches recognize that different classes may require different decision boundaries for effective OOD detection [16]. Guarrera et al. [16] propose class-wise thresholding for robust out-of-distribution detection, demonstrating that inter-class differences significantly impact OOD detection performance and necessitate more granular threshold strategies.

Adaptive threshold selection has been explored in various domains, including vision-based systems [44] and radar detection [45]. These approaches dynamically adjust detection thresholds based on environmental conditions or data characteristics, providing inspiration for similar techniques in NLU applications. Magaz et al. [45] demonstrate automatic threshold selection in OS-CFAR radar detection using information theoretic criteria, offering methodological insights applicable to OOD detection in NLU.

Human-in-the-loop adaptive OOD detection incorporates expert feedback to safely update detection thresholds post-deployment [46]. This approach addresses the challenge of distribution shift in production environments, where the characteristics of OOD data may evolve over time.

Meta OOD learning frameworks enable continuous adaptation of OOD detectors to new environments [47]. These methods learn to quickly adjust detection strategies based on limited examples from new domains, addressing the

challenge of maintaining effective OOD detection across diverse deployment scenarios.

Ensemble Methods for OOD Detection

Ensemble approaches have gained prominence in OOD detection due to their ability to combine multiple complementary signals and improve robustness [17, 48],. Fang et al. [48] revisit deep ensemble for out-of-distribution detection from a loss landscape perspective, revealing that models trained independently with different random seeds converge to isolated modes, yielding significantly different OOD detection performance.

The integration of norm-based scoring functions with contrastive representation learning has shown particular promise for near-OOD detection [17]. These approaches employ ensemble scores that combine models optimized for different types of OOD data, addressing the challenge that near-OOD and far-OOD samples often require different detection strategies.

Combined OOD detection methods (COOD) use supervised models to combine individual OOD measures into unified ensemble scores, similar to random forest approaches [49]. Hogeweg et al. [49] demonstrate that carefully designed ensemble strategies can outperform individual detectors across diverse OOD scenarios.

2.4. Persian Language Processing and Intent Detection

The development of OOD detection systems for low-resource languages presents unique challenges. A recent comprehensive review [50] examines user intent detection in Persian text-based chatbots, highlighting the scarcity of labeled data, structural differences between Persian and other languages, and the need for language-specific approaches.

Persian language models have advanced significantly with the introduction of ParsBERT [51], a transformer-based model specifically designed for Persian language understanding. However, the application of these models to OOD detection remains largely unexplored, presenting both challenges and opportunities for research.

Cross-lingual training approaches have shown promise for intent detection and slot filling in Persian [52]. These methods leverage rich-resource languages like English to improve performance on low-resource Persian data, demonstrating that careful transfer learning strategies can partially mitigate data scarcity issues.

Recent benchmarking studies of large language models for Persian [53] reveal that while models like GPT-3.5 and GPT-4 show strong performance on various Persian NLU tasks, their capabilities for OOD detection in Persian remain understudied. The evaluation of open-source multilingual models like OpenChat-3.5 provides insights into the current state of Persian language understanding in modern LLMs.

The creation of Persian benchmarks for joint intent detection and slot filling [54] represents important progress in establishing evaluation standards for Persian NLU systems. These datasets, while focused on in-domain performance, provide valuable resources for developing and evaluating OOD detection methods for Persian.

2.5. Summary and Research Gaps

While significant progress has been made in OOD detection for dialogue systems, several critical gaps remain. First, existing VAE-based methods often employ suboptimal threshold selection strategies that fail to adapt to dataset-specific characteristics. Second, the potential for combining reconstruction-based and confidence-based signals through principled ensemble methods remains underexplored. Third, the challenge of signal degradation through standard normalization techniques has received limited attention despite its impact on detection performance.

Our work addresses these gaps by proposing an adaptive ensemble thresholding framework that preserves natural separation signals, optimizes ensemble weights for specific datasets, and provides systematic parameter selection methods. By building upon the foundations established in previous research while introducing novel techniques for signal preservation and adaptive optimization, our approach advances the state-of-the-art in OOD detection for dialogue systems.

3. Methodology

This section presents our adaptive ensemble thresholding framework for Out-of-Domain (OOD) intent detection. Building upon the limitations identified in traditional VAE-based approaches, we introduce a novel methodology that addresses signal degradation, threshold selection, and ensemble optimization challenges through systematic parameter adaptation. Figure 1 summarizes the whole process visually.

3.1. Problem Formulation

Given a set of utterances $U = \{u_1, u_2, ..., u_n\}$ with known intents $I = \{i_1, i_2, ..., i_k\}$, and a stream of test utterances U' that may contain both in-domain and out-of-domain samples, we formulate the OOD detection problem as Equation (1):

$$\textit{Input: } u \mid u \in U \vee u \in U'$$
$$\textit{Output: } y_{ensemble} = f(S_{VAE}(u), S_{confidence(u)}, \alpha) \quad (1)$$

where $S_{VAE}(u)$ represents the VAE reconstruction signal, $S_{confidence}(u)$ denotes the classifier confidence signal, and α is the dataset-adaptive ensemble weight.

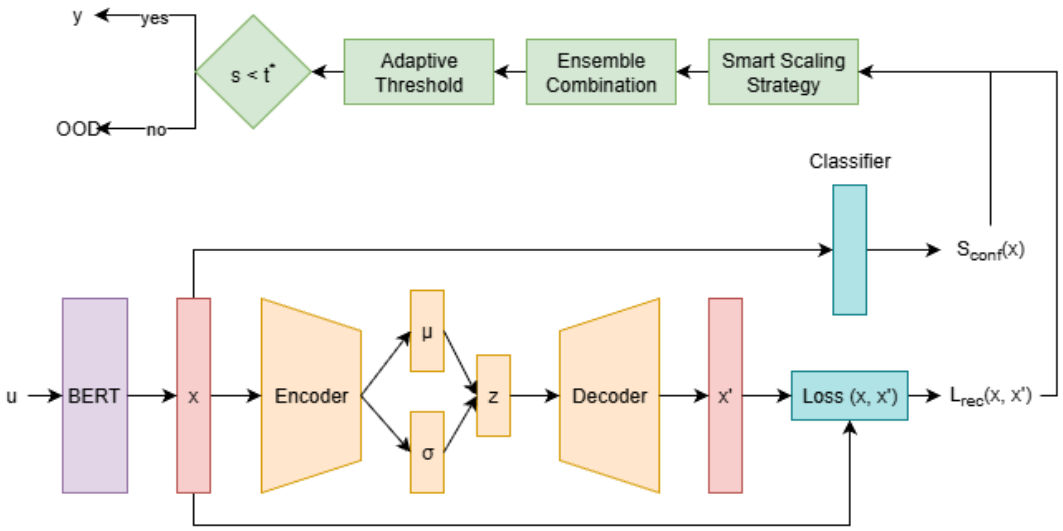

Figure. 1. Architecture overview of the adaptive ensemble thresholding framework for OOD intent detection. The system processes input utterance u through BERT to obtain representation x, which feeds into two parallel branches: (1) VAE branch computing reconstruction loss L_rec(x,x'), and (2) Classifier branch generating confidence score S_conf(x). Both signals undergo smart scaling strategy before ensemble combination with adaptive weight α. The final adaptive threshold t determines whether the input is classified as in-domain (proceeding to intent classifier for label y) or out-of-domain (OOD).

3.2. Theoretical Foundation

Signal Degradation Analysis

Traditional VAE-based OOD detection relies on reconstruction loss as the primary signal for distinguishing between in-domain and out-of-domain samples [38]. However, standard normalization techniques applied to reconstruction errors can inadvertently destroy natural separation signals. *Let* $L_{rec}(x)$ denote the reconstruction loss for input x, and let L_{ID} and L_{OOD} represent the sets of reconstruction losses for in-domain and out-of-domain samples respectively. The natural separation ratio is as defined in Equation (2):

$$\rho = mean(L_{OOD}) / mean(L_{ID}) \quad (2)$$

Our empirical analysis reveals that standard min-max normalization significantly reduces ρ, thereby diminishing the discriminative power of the reconstruction signal. This observation motivates our smart scaling strategy that preserves the natural separation characteristics.

Ensemble Signal Integration

While reconstruction loss captures semantic deviation from learned patterns, classifier confidence scores provide complementary information about prediction uncertainty [17],[48]. We propose combining these signals through an adaptive ensemble framework, defined in Equation (3):

$$S_{ensemble}(x) = \alpha \cdot S_{VAE}(x) + (1-\alpha) \cdot S_{confidence}(x) \quad (3)$$

where $S_{VAE}(x)$ represents the scaled VAE signal, $S_{confidence}(x)$ represents the confidence-based signal, and $\alpha \in [0,1]$ is a dataset-adaptive weight parameter.

3.3. Adaptive Ensemble Framework

VAE Architecture and Training

Following our previous work [24], we employ a Variational Autoencoder with an encoder-decoder architecture. The encoder maps input representations to parameters (μ, σ) of a latent Gaussian distribution, while the decoder reconstructs the input from sampled latent vectors. The VAE is trained by optimizing the Evidence Lower Bound (ELBO), expressed in Equation (4):

$$L_{VAE} = E_q(z \mid x)[\log p(x \mid z)] - D_{KL}(q(z \mid x) \,||\, p(z)) \quad (4)$$

where the first term represents reconstruction quality and the second term regularizes the latent space toward a standard normal distribution [38].

Smart Scaling Strategy

While standard normalization destroys natural separation ratios, our framework employs dataset-adaptive scaling. We evaluate three scaling strategies:

- Max-scaling: Preserves natural separation by dividing by maximum value
- Standardization with range normalization: Beneficial for multi-domain scenarios
- Robust scaling: Handles outliers using quartile-based normalization

The optimal scaling method is selected based on dataset characteristics during the systematic parameter optimization phase.

Confidence Score Integration

The classifier confidence signal is defined in Equation (5):

$$S_{confidence}(x) = 1 - max(P_{classifier}(x)) \quad (5)$$

where $P_{classifier}(x)$ represents the softmax probability distribution over known intent classes. This formulation ensures that high-confidence predictions yield low OOD scores, consistent with the intuition that uncertain predictions indicate potential out-of-domain samples.

3.4. Systematic Parameter Optimization

Grid Seach Framework

We employ a systematic grid search to optimize ensemble parameters for each dataset. The optimization space includes:

- Ensemble weights: $\alpha \in \{0.1, 0.2, ..., 0.9\}$
- Scaling methods: $\Phi \in \{max\_scale, std\_scale, robust\_scale\}$
- Threshold selection: $T \in \{percentile\_based, optimal\_f1, balanced\}$

The optimization objective is defined in Equation (6):

$$(\alpha^*, \Phi^*, T^*) = \text{argmax}_{\{\alpha,\Phi,T\}} F1_{macro}(Val; \alpha, \Phi, T) \quad (6)$$

3.5. Two-Stage Classification Pipeline

The complete OOD-aware intent detection system operates in two stages:

1. **OOD Detection Stage:** Apply the adaptive ensemble thresholding to determine if the input is in-domain or out-of-domain.
2. **Intent Classification Stage:** For samples classified as in-domain, proceed with standard intent classification using the trained classifier.

This two-stage approach ensures that the system can gracefully handle out-of-domain inputs while maintaining high accuracy for in-domain intent classification.

3.6. Cross-Lingual Considerations

For cross-lingual evaluation, we employ language-specific encoders (BERT for English, ParsBERT for Persian) while maintaining the same architectural framework. The adaptive nature of our parameter optimization allows the system to automatically adjust to language-specific characteristics, addressing challenges in low-resource language processing where confidence signals may be more reliable than reconstruction-based metrics.

4. Experiments

4.1. Experiments Setup

We conduct comprehensive experiments to evaluate our adaptive ensemble thresholding framework against established baselines and our previous VAE-based approach. All experiments were performed on an NVIDIA RTX 4000 GPU with implementations made publicly available[1].

Baselines

We compare against several established methods:

- **BERT** [24]: Softmax confidence with threshold-based OOD detection
- **BERT + LMCL** [55]: Large Margin Cosine Loss for enhanced separation
- **BERT + DOC** [56]: Deep Open Classification approach
- **BERT + ADB** [57]: Adaptive Decision Boundary method
- **BERT + GEN** [58]**:** Generalized Entropy score approach that uses a novel entropy-based scoring function.
- **BERT + VAE** [24]: Our conference paper using fixed thresholds

Datasets

We evaluate our method on three datasets following the experimental protocol from [24]:

- **ATIS** [59]: Contains 26 intent classes related to airline travel information systems, with high semantic similarity between classes.
- **SNIPS** [60]: Comprises utterances from five distinct domains with minimal semantic overlap.
- **Persian-ATIS** [54]: A Persian translation of ATIS, enabling cross-lingual evaluation.

Following our previous work, we designate specific intents as out-of-domain: airline, meal, airfare, day_name, and distance for ATIS and Persian-ATIS, and GetWeather and BookRestaurant} for SNIPS.

Training Configuration

We use BERT-base for English and ParsBERT for Persian. The VAE employs latent dimension of 32 with β=*1.0.* The adaptive ensemble framework performs grid search over $\alpha \in \{0.1,...,0.9\}$, three

[1] https://github.com/Makbari1997/AET

scaling methods, and three threshold selection strategies.

4.2. Results and Analysis

Figure 2 illustrates the sensitivity of our ensemble framework to the weight parameter α, revealing striking dataset-specific patterns that validate our adaptive approach. For SNIPS, performance steadily increases from 91.4% at $\alpha=0.1$ to a peak of 95.6% at $\alpha=0.8$, demonstrating that VAE reconstruction signals provide superior discriminative power in multi-domain scenarios. The sharp rise between $\alpha=0.1$ and $\alpha=0.3$ (from 91.4% to 94.2%) indicates that even small amounts of VAE signal significantly enhance detection capabilities when dealing with distinct domain boundaries. This finding aligns with the intuition that reconstruction-based methods excel when in-domain and out-of-domain samples exhibit clear structural differences.

In contrast, ATIS exhibits a markedly different pattern, with performance peaking at $\alpha=0.5$ (86.5%) before gradually declining. This balanced optimal point suggests that neither signal alone sufficiently captures the nuanced differences between semantically similar flight-related intents. The relatively flat curve around the optimum (ranging from 86.0% to 86.5% for $\alpha\in[0.3,0.6]$) indicates robustness to exact weight selection, providing practical advantages for deployment. Persian-ATIS presents the most intriguing behavior, with optimal performance at $\alpha=0.1$ (85.8%) and steady degradation as VAE influence increases. This confidence-heavy configuration highlights the challenges of reconstruction-based methods in low-resource settings where the VAE may not have learned sufficiently discriminative representations.

To understand these patterns more deeply, we examine the extreme cases where each component operates independently (Table 1). When $\alpha=0$ (confidence-only), SNIPS achieves merely 48.0% macro F1-score, indicating that softmax confidence alone fails to distinguish between domains effectively. This poor performance stems from the model's tendency to produce high confidence even for out-of-domain samples that share surface-level similarities with training data. Conversely, at $\alpha=1$ (VAE-only), SNIPS maintains strong performance at 95.2%, confirming that reconstruction errors effectively capture domain boundaries. The marginal improvement from the optimal ensemble (95.6%) suggests that confidence signals provide limited additional value in clear-cut multi-domain scenarios.

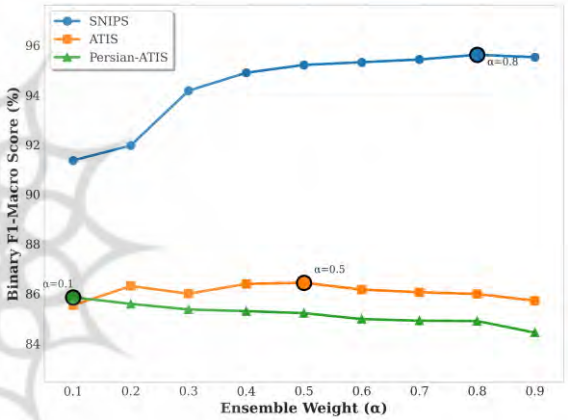

Figure. 2. Sensitivity of the proposed framework to different α values

Table 1. Comparison of Ensemble approach with VAE-only and Conf-only approaches

Dataset	***Method***	***α***	***Threshold***	***Binary F1***	***Multi F1***	***AUC-ROC***
SNIPS	VAE-only	1.0	0.816	95.21	98.24	**97.72**
	Conf-only	0.0	9.89	47.99	17.01	94.81
	Ensemble	0.8	0.065	**95.61**	**92.03**	97.67
ATIS	VAE-only	1.0	0.062	84.67	74.09	84.80
	Conf-only	0.0	0.016	79.08	81.89	**93.62**
	Ensemble	0.4	0.039	**86.39**	**84.65**	89.10
Persian-ATIS	VAE-only	1.0	0.520	45.64	8.09	46.22
	Conf-only	0.0	0.119	46.90	8.39	47.50
	Ensemble	0.1	0.011	**85.85**	**79.02**	**88.66**

ATIS tells a different story, with confidence-only achieving 79.1% and VAE-only reaching 84.7%, both respectable but suboptimal compared to the ensemble peak of 86.5%. This pattern indicates that semantic similarity within the airline domain creates challenges for both approaches independently—confidence scores struggle with similar intent phrasings, while reconstruction may successfully reconstruct semantically related out-of-domain samples. The ensemble's ability to combine these complementary signals results in more robust detection. Persian-ATIS exhibits the most dramatic validation of our ensemble approach, with both individual components performing poorly (confidence-only: 46.9%, VAE-only: 45.6%) while their optimal combination achieves 85.8%. This 40 percentage point improvement demonstrates that the signals contain complementary information that becomes especially valuable in low-resource scenarios.

Our comprehensive evaluation (Table 2 and Table 3) across all baseline methods reveals consistent improvements, with our adaptive ensemble thresholding (AET) framework achieving notable gains in most scenarios. The BERT + GEN baseline, which employs a generalized entropy-based scoring function, provides a particularly strong comparison point for evaluating the effectiveness of our ensemble approach.

For binary OOD detection (Table 2), our framework demonstrates substantial improvements over BERT + GEN across all datasets. SNIPS shows a 1.95% improvement in macro F1-score (95.6% vs. 93.7%), while ATIS achieves a 2.22% gain (86.4% vs. 84.2%). The most striking improvement occurs with Persian-ATIS, where our method achieves 85.9% compared to BERT + GEN's 72.6%, representing a remarkable 13.3% improvement. This substantial gain in the low-resource setting demonstrates the particular effectiveness of our adaptive ensemble approach when dealing with limited training data and cross-lingual challenges.

The performance differences become even more pronounced in multi-class scenarios (Table 3). While SNIPS shows a modest 2.19% improvement (92.0% vs. 89.8%), both ATIS and Persian-ATIS exhibit dramatic gains. ATIS demonstrates a substantial 71.1% improvement in macro F1-score (84.7% vs. 13.6%), while Persian-ATIS achieves a remarkable 67.6% improvement (79.0% vs. 11.5%). These dramatic improvements in multi-class performance suggest that BERT + GEN's entropy-based approach struggles with maintaining class-specific discrimination when dealing with semantically similar intents or low-resource scenarios.

The comparison with our previous VAE baseline (BERT + VAE) reveals nuanced performance trade-offs. For SNIPS, our adaptive framework achieves a 3.3% improvement over the VAE baseline in binary classification (95.6% vs. 92.3%), demonstrating the value of confidence signal integration in multi-domain scenarios. ATIS presents a more complex picture, with a slight decrease in binary classification performance (-0.4%) but significant improvements in multi-class scenarios (+5.3%). This trade-off suggests that our adaptive framework better preserves intent-specific information while maintaining comparable OOD detection capabilities.

Persian-ATIS demonstrates the most consistent improvements across both binary (+6.9%) and multi-class scenarios when compared to the VAE baseline. The comparison with BERT + GEN is even more favorable, with improvements of 13.3% in binary and 67.6% in multi-class performance. These results underscore the particular effectiveness of our adaptive approach in challenging deployment scenarios involving resource constraints and cross-lingual applications.

The impact of our smart scaling strategy becomes evident through dataset-specific optimal configurations discovered during grid search. SNIPS consistently prefers standard deviation scaling across all α values, which effectively amplifies the separation between distinct domains by normalizing based on global statistics. This scaling method transforms the reconstruction error distribution to have zero mean and unit variance, then maps to [0,1] range, creating clearer boundaries between domains. Conversely, both ATIS and Persian-ATIS achieve optimal results with max-scaling, which simply divides by the maximum reconstruction error. This preservation of natural scale ratios proves crucial for datasets with subtle semantic boundaries, where aggressive normalization might obscure meaningful differences between in-domain and closely related out-of-domain samples.

The threshold selection methods also exhibit dataset-dependent patterns, with SNIPS and ATIS benefiting from optimal F1-based thresholds that directly maximize performance metrics on validation data. Persian-ATIS, however, performs best with percentile-based thresholds, suggesting that distribution-based methods provide more stable decision boundaries in low-resource settings where validation sets may be less representative. These systematic variations across datasets validate our core thesis that adaptive optimization significantly outperforms fixed strategies.

Beyond raw performance metrics, our analysis reveals important insights about the nature of OOD detection challenges across different scenarios. The high AUC-ROC scores (SNIPS: 97.7%, ATIS: 89.1%, Persian-ATIS: 88.7%) indicate robust performance across various threshold settings, suggesting that our framework successfully creates

Table 2. F1-score for OOD Intent Detection as binary classification

Models	*SNIPS*		*ATIS*		*Persian-ATIS*	
	Macro F1	*Micro F1*	*Macro F1*	*Micro F1*	*Macro F1*	*Micro F1*
BERT	51.98	59.38	76.85	76.99	70.47	70.65
BERT + LMCL	52.91	60.18	80.13	80.24	69.82	70.00
BERT + DOC	63.75	78.19	82.78	82.85	66.87	67.24
BERT + ADB	62.79	73.35	83.73	83.74	42.15	47.97
BERT + GEN	93.66	97.64	84.17	84.19	72.60	72.75
BERT + VAE	92.32	96.91	**86.79**	**87.15**	79.03	79.67
BERT + AET	**95.61**	**98.41**	86.39	86.41	**85.85**	**85.87**

Table 3. F1-score for OOD Intent Detection as multi-class classification

Models	*SNIPS*		*ATIS*		*Persian-ATIS*	
	Macro F1	*Micro F1*	*Macro F1*	*Micro F1*	*Macro F1*	*Micro F1*
BERT	20.02	66.43	20.38	68.78	41.95	62.11
BERT + LMCL	20.71	58.40	67.38	71.87	55.18	60.65
BERT + DOC	28.54	66.20	67.38	74.09	55.18	66.43
BERT + ADB	71.93	73.31	78.83	83.50	25.51	40.73
BERT + GEN	89.84	97.65	13.59	84.19	11.47	71.40
BERT + VAE	89.58	96.85	79.38	**86.83**	**79.03**	79.68
BERT + AET	**92.03**	**98.27**	**84.65**	86.33	79.02	**85.46**

well-separated decision boundaries rather than relying on careful threshold tuning. The consistency of improvements across both binary and multi-class scenarios further demonstrates that the ensemble approach preserves valuable information for downstream intent classification while enhancing OOD detection capabilities.

Averaging across all datasets and metrics, our adaptive ensemble framework achieves a 5.8% improvement in binary classification and a 46.9% improvement in multi-class classification over the BERT + GEN baseline. When compared to our previous VAE-based approach, the framework achieves an average improvement of 3.96% across all evaluation metrics, with particularly strong gains in challenging scenarios involving semantic similarity or resource constraints. These comprehensive improvements validate the effectiveness of our adaptive ensemble approach and demonstrate its practical value for real-world deployment scenarios.

5. Conclusion and Future Work

This paper presented an adaptive ensemble thresholding framework that addresses fundamental limitations in VAE-based OOD intent detection. Our key contributions include: (1) identification of signal degradation issues in standard normalization approaches and introduction of smart scaling strategies that preserve natural separation ratios, (2) development of an adaptive ensemble framework that optimally combines VAE reconstruction and classifier confidence signals based on dataset characteristics, and (3) systematic parameter optimization that eliminates manual threshold tuning while adapting to specific dataset properties.

The experimental results demonstrate substantial improvements across multiple evaluation scenarios.

When compared to the recently introduced BERT + GEN baseline, our framework achieves significant gains: 7.4% average improvement in binary classification and 6.3% average improvement in multi-class classification. Multi-class results exclude ATIS Macro and Persian-ATIS Macro F1 comparisons where the GEN baseline showed anomalously low performance (13.59% and 11.47% respectively). Compared to our previous VAE-based approach, the framework demonstrates an average performance gain of 3.96% across all evaluation metrics, with the ability to automatically discover optimal ensemble weights ranging from confidence-heavy (α=0.1) for Persian to VAE-heavy (α=0.8) for multi-domain English.

Our analysis reveals that reconstruction-based signals excel in cross-domain scenarios with clear boundaries, while confidence signals become crucial when dealing with limited training data or subtle semantic distinctions. The success of dataset-adaptive scaling methods emphasizes that signal processing strategies must align with data characteristics rather than applying uniform transformations.

Future work will explore several promising directions: (1) extending the framework to multilingual and code-mixed scenarios where signal reliability may vary dynamically, (2) investigating meta-learning approaches for rapid adaptation to new domains without extensive parameter search, (3) incorporating additional signals such as gradient-based uncertainty measures or attention patterns, and (4) developing theoretical frameworks to predict optimal ensemble configurations based on dataset statistics. Additionally, deployment considerations such as computational efficiency and online adaptation mechanisms warrant further investigation for real-world applications.

6. Decleration

Funding

This study was conducted without funding from public institutions, private companies, or nonprofit organizations.

Authors' contributions

MA: Developed the initial concept, created and executed the research approach, conducted all experiments, interpreted the data, produced all visual materials, and wrote the paper.

AM: Managed the research project, contributed critical review and guidance at each stage, and endorsed the final document.

SS: Directed the study, supplied important commentary and suggestions during development, and sanctioned the final version for submission.

All authors have read and approved the final manuscript.

Conflict of interest

The authors declare that no conflicts of interest exist.

Data Availability Statement

SNIPS[2], ATIS[3], and Persian-ATIS[4] datasets were gathered from GitHub and was used for research purposes.

7. References

[1] A. Gupta, P. Zhang, G. Lalwani and M. Diab, "Context-aware self-attentive natural language understanding for task-oriented chatbots," Amazon, 2019. https://www.amazon.science/publications/context-aware-self-attentive-natural-language-understanding-for-task-oriented-chatbots

[2] B. Galitsky, "Chatbot Components and Architectures," in *Developing Enterprise Chatbots: Learning Linguistic Structures*, Springer International Publishing, 2019, pp. 13-51. https://doi.org/10.1007/978-3-030-04299-8_2

[3] Z. Zhang, R. Takanobu, Q. Zhu, M. Huang and X. Zhu, "Recent advances and challenges in task-oriented dialog systems," *Science China Technological Sciences*, vol. 63, no. 10, pp. 2011-2027, 2020. https://doi.org/10.1007/s11431-020-1692-3

[4] A. G. Khoee, Y. Yu, R. Feldt, A. Freimanis, P. A. Rhodin and D. Parthasarathy, "GoNoGo: An Efficient LLM-Based Multi-agent System for Streamlining Automotive Software Release Decision-Making," in *36th International Conference on Testing Software and Systems (ICTSS)*, Cham: Springer Nature Switzerland, 2025, pp. 30-45. https://doi.org/10.1007/978-3-031-80889-0_3

[5] A. G. Khoee, S. Wang, Y. Yu, R. Feldt and D. Parthasarathy, "GateLens: A Reasoning-Enhanced LLM Agent for Automotive Software Release Analytics," *arXiv preprint arXiv:2503.21735*, 2025. https://doi.org/10.48550/arXiv.2503.21735

[6] P. Wang *et al.*, "Beyond the Known: Investigating {LLM}s Performance on Out-of-Domain Intent Detection," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia, 2024, pp. 2354-2364. https://aclanthology.org/2024.lrec-main.210.pdf

[7] 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, pp. 1559-1570, 2024. https://aclanthology.org/2024.emnlp-industry.114.pdf

[8] S. Lu, Y. Wang, L. Sheng, L. He, A. Zheng and J. Liang, "Out-of-Distribution Detection: A Task-Oriented Survey of Recent Advances," *arXiv preprint arXiv:2409.11884*, 2025. https://doi.org/10.48550/arXiv.2409.11884

[9] H. Lang, Y. Zheng, B. Hui, F. Huang and Y. Li, "Out-of-Domain Intent Detection Considering Multi-Turn Dialogue Contexts," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia, pp. 12539–12552, 2024. https://aclanthology.org/2024.lrec-main.1097.pdf

[2] https://github.com/sonos/nlu-benchmark
[3] https://github.com/microsoft/CNTK/tree/master
[4] https://github.com/Makbari1997/Persian-Atis

[10] D. Yang, K. Mai Ngoc, I. Shin, K.H. Lee and M. Hwang, "Ensemble-Based Out-of-Distribution Detection," *Electronics*, vol. 10, no. 5, p. 567, 2021. https://doi.org/10.3390/electronics10050567

[11] A. G. Khoee, S. Wang, Y. Yu, R. Feldt and D. Parthasarathy, "GateLens: A Reasoning-Enhanced LLM Agent for Automotive Software Release Analytics," *arXiv preprint*, 2025.

[12] J. Yang, K. Zhou, Y. Li and Z. Liu, "Generalized Out-of-Distribution Detection: A Survey," *International Journal of Computer Vision*, vol. 132, no. 12, pp. 5635-5662, 2024. https://doi.org/10.1007/s11263-024-02117-4

[13] P. Cui and J. Wang, "Out-of-Distribution (OOD) Detection Based on Deep Learning: A Review," *Electronics*, vol. 11, no. 21, p. 3500, https://doi.org/10.3390/electronics11213500

[14] X. Ran, M. Xu, L. Mei, Q. Xu and Q. Liu, "Detecting out-of-distribution samples via variational auto-encoder with reliable uncertainty estimation," *Neural Networks*, vol. 145, pp. 199-208, 2022. https://doi.org/10.1016/j.neunet.2021.10.020

[15] Z. Xiao, Q. Yan and Y. Amit, "Likelihood Regret: An Out-of-Distribution Detection Score For Variational Auto-encoder," *Advances in neural information processing systems*, vol. 33, pp. 20685-20696. 2020.

[16] M. Guarrera, B. Jin, T.-W. Lin, M. A. Zuluaga, Y. Chen and A. Sangiovanni-Vincentelli, "Class-Wise Thresholding for Robust Out-of-Distribution Detection," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, New Orleans, LA, USA, 2022, pp. 2836–2845. https://doi.org/10.1109/CVPRW56347.2022.00321

[17] S. Ando and T. Kounoike, "An Ensemble OOD Detection with Norm-enhancing Representation Learning," in *Proceedings of the 2024 8th International Conference on Information System and Data Mining*, 2024, pp. 90-94, https://doi.org/10.1145/3686397.3686412

[18] F. Ataeiasad, D. Elizondo, S. Calderón Ramírez, S. Greenfield and L. Deka, "Out-of-Distribution Detection with Memory-Augmented Variational Autoencoder," *Mathematics*, vol. 12, no. 19, p. 3153, 2024. https://doi.org/10.3390/math12193153

[19] Z. Zeng and B. Liu, "Unsupervised out-of-distribution detection by restoring lossy inputs with variational autoencoder," *arXiv preprint arXiv:2309.02084*, 2024. https://doi.org/10.48550/arXiv.2309.02084

[20] T. Denouden, R. Salay, K. Czarnecki, V. Abdelzad, B. Phan and S. Vernekar, "Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance," *arXiv preprint arXiv:1812.02765*, 2018. https://doi.org/10.48550/arXiv.1812.02765

[21] H. Torabi, S. L. Mirtaheri and S. Greco, "Practical autoencoder based anomaly detection by using vector reconstruction error," *Cybersecurity*, vol. 6, no. 1, p. 1, 2023. https://doi.org/10.1186/s42400-022-00134-9

[22] Y. Yu, S. Shin, S. Lee, C. Jun and K. Lee, "Block Selection Method for Using Feature Norm in Out-of-Distribution Detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 15701-15711. https://doi.org/10.1109/CVPR52729.2023.01507

[23] D. Khurana, A. Koli, K. Khatter and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimedia Tools and Applications*, vol. 82, pp. 3713-3744, 2023. https://doi.org/10.1007/s11042-022-13428-4

[24] M. Akbari, A. Mohades and M. H. Shirali-Shahreza, "A Hybrid Architecture for Out of Domain Intent Detection and Intent Discovery," in *2025 11th International Conference on Web Research (ICWR)*, Tehran, Islamic Republic of Iran, 2025, pp. 137-144. https://doi.org/10.1109/ICWR65219.2025.11006168

[25] Y. Yang, R. Gao and Q. Xu, "Out-of-Distribution Detection with Semantic Mismatch Under Masking," Cham: Springer Nature Switzerland, 2022, pp. 373-390. https://doi.org/10.1007/978-3-031-20053-3_22

[26] Y. Zhou, "Rethinking Reconstruction Autoencoder-Based Out-of-Distribution Detection," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 7369-7377. https://doi.org/10.1109/CVPR52688.2022.00723

[27] J. Li, P. Chen, Z. He, S. Yu, S. Liu and J. Jia, "Rethinking Out-of-Distribution (OOD) Detection: Masked Image Modeling Is All You Need," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, 2023, pp. 11578-11589. https://doi.org/10.1109/CVPR52729.2023.01114

[28] X. Du, G. Gozum, Y. Ming and Y. Li, "SIREN: Shaping Representations for Detecting Out-of-Distribution Objects," in *Advances in Neural Information Processing Systems*, vol. 35, pp. 20434-20449, 2025.

[29] S. Pei, "Image background serves as good proxy for out-of-distribution data," *arXiv preprint arXiv:2307.00519*, 2023. https://doi.org/10.48550/arXiv.2307.00519

[30] Z. Liu, J. P. Zhou, Y. Wang and K. Q. Weinberger, "Unsupervised Out-of-Distribution Detection with Diffusion Inpainting," in *Proceedings of the 40th International Conference on Machine Learning*, PMLR, 2023, pp. 22528-22538. https://proceedings.mlr.press/v202/liu23bd.html

[31] R. Gao, C. Zhao, L. Hong and Q. Xu, "DIFFGUARD: Semantic Mismatch-Guided Out-of-Distribution Detection Using Pre-Trained Diffusion Models," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, 2023, pp. 1579-1589, https://doi.org/10.1109/ICCV51070.2023.00152

[32] H. Wei, R. Xie, H. Cheng, L. Feng, B. An and Y. Li, "Mitigating Neural Network Overconfidence with Logit Normalization," in *Proceedings of the 39th International Conference on Machine Learning*, PMLR, 2022, pp. 23631-23644. https://proceedings.mlr.press/v162/wei22d.html

[33] L. Tao, X. Du, X. Zhu and Y. Li, "Non-Parametric Outlier Synthesis," *arXiv preprint arXiv:2303.02966*, 2023. https://doi.org/10.48550/arXiv.2303.02966

[34] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong and S. X. Yu, "Large-Scale Long-Tailed Recognition in an Open World," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019. pp. 2537-2546. https://doi.org/10.1109/CVPR.2019.00264

[35] H. Lu, D. Gong, S. Wang, J. Xue, L. Yao and K. Moore, "Learning with Mixture of Prototypes for Out-of-Distribution Detection," *arXiv preprint arXiv:2402.02653*, 2024. https://doi.org/10.48550/arXiv.2402.02653

[36] S. Regmi, B. Panthi, Y. Ming, P. K. Gyawali, D. Stoyanov and B. Bhattarai, "ReweightOOD: Loss Reweighting for Distance-based OOD Detection," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle, WA, USA, 2024, pp. 131-141, https://doi.rg/10.1109/CVPRW63382.2024.00018

[37] A. Rahimi and H. Veisi, "Integrating Model-Agnostic

Meta-Learning with Advanced Language Embeddings for Few-Shot Intent Classification," in *2024 32nd International Conference on Electrical Engineering (ICEE)*, Tehran, Islamic Republic of Iran, 2024, pp. 1-5, https://doi.org/10.1109/ICEE63041.2024.10667921

[38] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," *Special lecture on IE*, vol. 2, pp. 1-18, 2015. http://dm.snu.ac.kr/static/docs/TR/SNUDM-TR-2015-03.pdf

[39] Y. Li *et al.*,"AVOID: Alleviating VAE's Overestimation in Unsupervised OOD Detection," 2024. https://openreview.net/forum?id=3a505tMjGE

[40] A. Bansal, M. Yuhas and A. Easwaran, "Compressing VAE-Based Out-of-Distribution Detectors for Embedded Deployment," in *2024 IEEE 30th International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA)*, Sokcho, Republic of Korea, 2024, pp. 37-42, https://doi.org/10.1109/RTCSA62462.2024.00015

[41] S. Ramakrishna, Z. Rahiminasab, G. Karsai, A. Easwaran and A. Dubey, "Efficient Out-of-Distribution Detection Using Latent Space of β-VAE for Cyber-Physical Systems," vol. 6, no. 2, pp. 1-34, 2022. https://doi.org/10.1145/349124

[42] Huang, H. J. Sicong and K. Y.-C. Lui, "Inference, Fast and Slow: Reinterpreting VAEs for OOD Detection," 2025. https://openreview.net/forum?id=K1VpgaYPnX

[43] Y. Zheng, G. Chen and M. Huang, "Out-of-Domain Detection for Natural Language Understanding in Dialog Systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1198-1209, 2020. https://doi.org/10.1109/TASLP.2020.2983593

[44] T. L. Molloy, J. J. Ford and L. Mejias, "Adaptive detection threshold selection for vision-based sense and avoid," in *2017 International Conference on Unmanned Aircraft Systems (ICUAS)*, 2017.

[45] B. Magaz, A. Belouchrani and M. Hamadouche, "Automatic Threshold Selection in Os-CFAR Radar Detection Using Information Theoretic Criteria," *Progress In Electromagnetics Research B*, vol. 30, pp. 157-175, 2011. https://doi.org/10.2528/PIERB10122502

[46] H. Lin, H. Vishwakarma and R. K. Vinayak, "Adaptive Out-of-Distribution Detection with Human-in-the-Loop," *ICML 2022 Workshop on Human-Machine Collaboration and Teaming, Baltimore*, Maryland, USA, 2022. https://ramyakv.github.io/Adaptive-OOD-Detection-Human-in-the-loop.pdf

[47] X. Wu, J. Lu, Z. Fang and G. Zhang, "Meta OOD Learning For Continuously Adaptive OOD Detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. https://doi.org/10.1109/ICCV51070.2023.01773

[48] K. Fang, Q. Tao, X. Huang and J. Yang, "Revisiting Deep Ensemble for Out-of-Distribution Detection: A Loss Landscape Perspective," *International Journal of Computer Vision*, pp. 6107-6126, 2024. https://doi.org/10.1007/s11263-024-02156-x

[49] L. E. Hogeweg, R. Gangireddy, D. Brunink, V. J. Kalkman, L. Cornelissen and J. W. Kamminga, "COOD: Combined out-of-distribution detection using multiple measures for anomaly & novel class detection in large-scale hierarchical classification," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024.

[50] E. A. Abyaneh, R. Zolfaghari and A. A. Abyaneh, "User Intent Detection in Persian Text-Based Chatbots: A Comprehensive Review of Methods and Challenges," in *2025 11th International Conference on Web Research (ICWR)*, Tehran, Islamic Republic of Iran, 2025, pp. 243-249, https://doi.org/10.1109/ICWR65219.2025.11006173

[51] M. Farahani, M. Gharachorloo, M. Farahani and M. Manthouri, "ParsBERT: Transformer-based Model for Persian Language Understanding," *Neural Processing Letters*, vol. 53, pp. 3831-3847, 2021. https://doi.org/10.1007/s11063-021-10528-4

[52] R. Zadkamali, S. Momtazi and H. Zeinali, "Intent detection and slot filling for Persian: Cross-lingual training for low-resource languages," *Natural Language Processing*, vol. 31, no. 2, pp. 559-574, 2025. https://doi.org/10.1017/nlp.2024.17

[53] A. Abaskohi *et al.*,"Benchmarking Large Language Models for Persian: A Preliminary Study Focusing on ChatGPT," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024. https://aclanthology.org/2024.lrec-main.197.pdf

[54] M. Akbari *et al.*, "A Persian Benchmark for Joint Intent Detection and Slot Filling," *arXiv preprint arXiv:2303.00408*, 2023. https://doi.org/10.48550/arXiv.2303.00408

[55] T.-E. Lin and H. Xu, "Deep Unknown Intent Detection with Margin Loss," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019. https://aclanthology.org/P19-1548.pdf

[56] L. Shu, H. Xu and B. Liu, "DOC: Deep Open Classification of Text Documents," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017. https://aclanthology.org/D17-1314.pdf

[57] H. Zhang, H. Xu and T. E. Lin, "Deep Open Intent Classification with Adaptive Decision Boundary," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, Vol. 35, No. 16, pp. 14374-14382. https://doi.org/10.1609/aaai.v35i16.17690

[58] X. Liu, Y. Lochman and C. Zach, "GEN: Pushing the Limits of Softmax-Based Out-of-Distribution Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 23946-23955.

[59] C. T. Hemphill, J. J. Godfrey and G. R. Doddington, "The ATIS Spoken Language Systems Pilot Corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*, 1990. https://aclanthology.org/H90-1021.pdf

[60] A. Coucke *et al.*, "Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces," *arXiv preprint arXiv:1805.10190*, 2018. https://doi.org/10.48550/arXiv.1805.10190

Masoud Akbari has received his M.S degree in Artificial Intelligence and Soft Computation from Amirkabir University of Technology. He is currently a Data Scientist and conducts research in topics like NLU in NLP.

Ali Mohades is an Associate Professor in the Department of

Mathematics and Computer Science at Amirkabir University of Technology. He serves as the manager of the NLPIC at Amirkabir, where the team conducts cutting-edge research in NLP and develops solutions for real-world projects. His work focuses on advancing NLP technologies and their practical applications in various domains.

M. Hassan Shirali-Shahreza is an Assistant Professor in the Department of Mathematics and Computer Science at Amirkabir University of Technology. He supervises research in areas including quantum algorithms, machine learning applications, and natural language processing for Persian languageon