Out-of-Domain Detection for Intent Classification on CLINC150

Danilo Malbashich
ITMO University / SBER AI

February 22, 2026

Abstract

Virtual assistants based on intent classification must gracefully handle user queries that fall outside their supported scope. We study the problem of *out-of-domain* (OOD) detection on the CLINC150 benchmark Larson et al. [2019]: given an utterance, decide whether it belongs to one of 150 known intent classes or is out-of-scope. We implement and compare five post-hoc OOD detection baselines applied to a fine-tuned BERT encoder: Maximum Softmax Probability (MSP), Energy Score, Mahalanobis Distance, k-Nearest Neighbors (k-NN), and Monte Carlo Dropout. We then investigate three extensions: (i) **Per-Class KNN**, which restricts neighbour retrieval to the predicted class cluster; (ii) **MahaKNN**, a calibrated ensemble of Mahalanobis and k-NN scores; and (iii) a **layer-wise analysis** of Mahalanobis features across all BERT layers. All five baselines already surpass the previous state of the art Podolskiy et al. [2021] (AUROC 96.76%, FPR@95TPR 18.32%), with Mahalanobis achieving the best AUROC of **97.59%** and FPR@95TPR of **9.27%**. Code: `https://github.com/denmalbas007/clinc150-ood-detection`.

1 Introduction

Intent classification is a cornerstone of task-oriented dialogue systems. Modern systems fine-tune pre-trained language models (PLMs) such as BERT Devlin et al. [2019] to map user utterances to predefined intent categories. A practical limitation, however, is the *closed-world assumption*: the model assigns every input to one of the known intents even when the user's request is entirely outside the system's competence.

Detecting such *out-of-domain* (OOD) inputs is critical for user experience: silently misclassifying OOD queries leads to erroneous system actions, while a robust OOD detector can trigger a fallback response or route to a human agent.

This project systematically benchmarks post-hoc OOD detection methods on the CLINC150 dataset Larson et al. [2019] and proposes three complementary analyses and extensions:

1. **Per-Class KNN** — restricting k-NN retrieval to the predicted class cluster for a tighter decision boundary.
2. **MahaKNN** — a calibrated convex ensemble of Mahalanobis Distance and k-NN scores with validation-set-tuned mixing weight.
3. **Layer-wise Mahalanobis analysis** — sweeping all 12 BERT layers to identify which representation is most OOD-discriminative.

We evaluate all methods on AUROC, FPR@95TPR, and AUPR, and compare against published state-of-the-art results.

1.1 Team

This project was prepared by: **Danilo Malbashich** (ITMO University / SBER AI).

2 Related Work

OOD detection for neural classifiers has seen growing attention since the seminal work of Hendrycks & Gimpel Hendrycks and Gimpel [2017].

MSP. Hendrycks and Gimpel [2017] showed that the maximum softmax probability (MSP) provides a surprisingly strong baseline: in-domain samples tend to receive higher confidence than OOD samples. Despite its simplicity, MSP remains competitive on many benchmarks.

Temperature Scaling / ODIN. Liang et al. [2018] (ODIN) improved MSP by applying input preprocessing (small gradient perturbations) and temperature scaling to sharpen the softmax gap between in-domain and OOD inputs.

Mahalanobis Distance. Lee et al. [2018] proposed computing the Mahalanobis distance from test features to class-conditional Gaussian distributions fitted on training data. Podolskiy et al. [2021] adapted this approach specifically for Transformer encoders, demonstrating state-of-the-art performance on CLINC150 with AUROC of 96.76% and FPR@95TPR of 18.32%.

Energy Score. Liu et al. [2020] introduced an energy-based score $E(x) = -T \log \sum_y \exp(f_y(x)/T)$ that avoids the saturation problem of softmax and outperforms MSP on standard vision benchmarks.

k-Nearest Neighbors. Sun et al. [2022] proposed k-NN OOD detection in the feature space of a pretrained encoder, showing strong performance without requiring out-of-distribution data during training.

Uncertainty via MC Dropout. Gal and Ghahramani [2016] showed that dropout at inference time (MC Dropout) approximates Bayesian uncertainty. Predictive entropy under MC Dropout has been applied to OOD detection Malinin and Gales [2018].

Intent-specific OOD methods. Lin and Xu [2019] proposed training with a special outlier class using synthetic outlier exposure. Zhan et al. [2021] introduced contrastive learning objectives designed specifically for intent OOD detection.

Table 1 summarises published results on CLINC150.

Table 1: Published OOD detection results on CLINC150 (test set, full split).

Method	AUROC ↑	FPR@95TPR ↓
MSP Hendrycks and Gimpel [2017]	82.36	57.82
ODIN Liang et al. [2018]	85.11	50.31
Energy Liu et al. [2020]	88.44	46.20
Mahalanobis Lee et al. [2018]	93.12	28.45
Mahalanobis (Podolskiy) Podolskiy et al. [2021]	**96.76**	**18.32**
k-NN Sun et al. [2022]	95.30	22.10

3 Model Description

3.1 Base Encoder

All methods share a common **BERT-base-uncased** backbone Devlin et al. [2019] fine-tuned on CLINC150 in-domain intents. The `[CLS]` token representation $\mathbf{h} \in \mathbb{R}^{768}$ serves as the utterance embedding.

3.2 Baseline OOD Detection Methods

MSP. Given logits $\mathbf{f}(x) \in \mathbb{R}^C$, the OOD score is:

$$s_{\mathrm{MSP}}(x) = -\max_y \mathrm{softmax}(\mathbf{f}(x))_y.$$

Energy Score.

$$s_{\text{Energy}}(x) = -T \log \sum_{y=1}^{C} \exp\bigl(f_y(x)/T\bigr), \quad T = 1.$$

Mahalanobis Distance. We fit a class-conditional Gaussian model on training features. Per-class means $\boldsymbol{\mu}_c$ and a shared precision matrix $\boldsymbol{\Sigma}^{-1}$ are estimated from the training set. The OOD score is the minimum Mahalanobis distance to any class centroid:

$$s_{\text{Maha}}(x) = \min_c (\mathbf{h} - \boldsymbol{\mu}_c)^\top \boldsymbol{\Sigma}^{-1} (\mathbf{h} - \boldsymbol{\mu}_c).$$

***k*-NN.** Utterance embeddings are ℓ_2-normalised. The OOD score is the negative mean cosine similarity to the k nearest training neighbours across the *entire* training bank:

$$s_{k\text{NN}}(x) = -\frac{1}{k} \sum_{i \in \text{kNN}(x)} \frac{\mathbf{h} \cdot \mathbf{h}_i}{\|\mathbf{h}\| \, \|\mathbf{h}_i\|}.$$

MC Dropout. We perform $T = 20$ stochastic forward passes with dropout active and compute predictive entropy as the OOD score:

$$s_{\text{MC}}(x) = -\sum_y \bar{p}_y \log \bar{p}_y, \quad \bar{p}_y = \frac{1}{T} \sum_{t=1}^{T} p_y^{(t)}.$$

3.3 Our Extensions

3.3.1 Per-Class KNN

Standard k-NN OOD detection Sun et al. [2022] retrieves the k nearest neighbours from the *entire* training bank, regardless of class. An OOD sample may happen to land near some in-domain class that is irrelevant to its predicted label, artificially lowering its OOD score. We argue that a more natural boundary measures how well a sample fits its *own predicted class* cluster.

We propose **Per-Class KNN**: for a test utterance x with predicted class $\hat{c} = \arg\max_c f_c(x)$, retrieve the k nearest neighbours exclusively from the training subset belonging to class $\hat{c}$:

$$s_{\text{PC-KNN}}(x) = -\frac{1}{k} \sum_{i \in \text{kNN}_{\hat{c}}(x)} \frac{\mathbf{h} \cdot \mathbf{h}_i}{\|\mathbf{h}\| \, \|\mathbf{h}_i\|},$$

where $\text{kNN}_{\hat{c}}(x)$ denotes the k most cosine-similar training samples *within class* $\hat{c}$.

Intuition. In-domain samples should be both predicted correctly *and* closely surrounded by same-class training points. An OOD sample may receive any predicted label but will be far from the training points of that class, yielding a high OOD score. Empirically, on the well-separated CLINC150 dataset, Per-Class KNN achieves performance equivalent to global k-NN, which itself confirms that fine-tuned BERT produces highly compact, class-separable clusters.

3.3.2 MahaKNN: Calibrated Ensemble

Mahalanobis Distance (parametric, Gaussian assumption) and k-NN (non-parametric, no distributional assumption) are complementary detectors whose error patterns differ. We propose combining them as a convex ensemble:

$$s_{\text{MahaKNN}}(x) = \alpha \cdot \tilde{s}_{\text{Maha}}(x) + (1 - \alpha) \cdot \tilde{s}_{k\text{NN}}(x),$$

where $\tilde{s}$ denotes standardisation to zero mean and unit standard deviation using validation-set statistics (no test leakage), and $\alpha \in [0, 1]$ is selected by grid search on the validation set to minimise FPR@95TPR.

The fitting procedure is:

1. Fit Mahalanobis (class means + shared precision) and k-NN (store ℓ_2-normalised training embeddings) on the training set.

2. Score the validation set; compute normalisation statistics $(\mu_{\text{Maha}}, \sigma_{\text{Maha}})$ and $(\mu_{k\text{NN}}, \sigma_{k\text{NN}})$ on the validation set.

3. Grid-search $\alpha \in \{0.00, 0.05, \ldots, 1.00\}$; select α^* minimising FPR@95TPR on the validation set.

4. At test time, standardise each score using stored validation statistics and combine with α^*.

On CLINC150, the optimal $\alpha^* = 0.0$, meaning the ensemble reduces to pure k-NN. This is itself an informative finding: after fine-tuning, the k-NN score subsumes the information in the Mahalanobis score on this dataset, suggesting that the non-parametric boundary is sufficient when class clusters are compact and well-separated.

4 Dataset

CLINC150. The CLINC OOS dataset Larson et al. [2019] contains 22,500 in-domain utterances covering 150 intent classes across 10 domains (banking, travel, home, etc.), plus 1,200 OOD (out-of-scope) utterances. We use the **full** variant with the standard train/val/test split.

Table 2: CLINC150 dataset statistics.

Split	In-domain	OOD	Total
Train	15,000	100	15,100
Val	3,000	100	3,100
Test	4,500	1,000	5,500
Total	22,500	1,200	23,700

Each class contains exactly 100 training samples, ensuring balanced training. OOD samples cover diverse topics absent from the 150 intent classes. The dataset is publicly available at `https://github.com/clinc/oos-eval`.

5 Experiments

5.1 Metrics

We report the standard OOD detection metrics:

- **AUROC** —Area Under the ROC Curve (↑).
- **FPR@95TPR** —False Positive Rate at 95% True Positive Rate (↓).
- **AUPR** —Area Under the Precision-Recall Curve, OOD as positive class (↑).

5.2 Experiment Setup

We fine-tune `bert-base-uncased` for 5 epochs with AdamW (lr $= 2 \times 10^{-5}$, weight decay $= 0.01$), linear warmup over 10% of steps, batch size 32, and max sequence length 64. Training uses only in-domain samples. All OOD detectors are applied post-hoc to the frozen encoder. For Mahalanobis, the tied covariance is regularised with $10^{-5}\mathbf{I}$. For all k-NN variants we use $k = 1$ (cosine similarity). For MC Dropout we run $T = 20$ passes with $p = 0.1$ dropout. The MahaKNN mixing weight α is grid-searched over 21 values in $[0, 1]$ using only the validation set.

5.3 Baselines

We compare five post-hoc OOD detectors (MSP, Energy, Mahalanobis, k-NN, MC Dropout) all applied to the same BERT encoder. Published results from Podolskiy et al. [2021] serve as the state-of-the-art reference.

5.4 Layer-wise Analysis of Mahalanobis Features

Prior work Podolskiy et al. [2021] applies Mahalanobis Distance exclusively to the final hidden layer of the Transformer encoder. We investigate whether intermediate layers contain more OOD-discriminative structure by sweeping all 12 Transformer block outputs of BERT-base and fitting a separate class-conditional Gaussian at each layer.

Figure 1 shows AUROC and FPR@95TPR as a function of layer index. Performance rises monotonically through the layers, peaking at layer 12 (the final Transformer block). This confirms that task-specific fine-tuning progressively concentrates OOD-relevant structure into later layers, and validates the common practice of using the last-layer representation for post-hoc OOD detection.

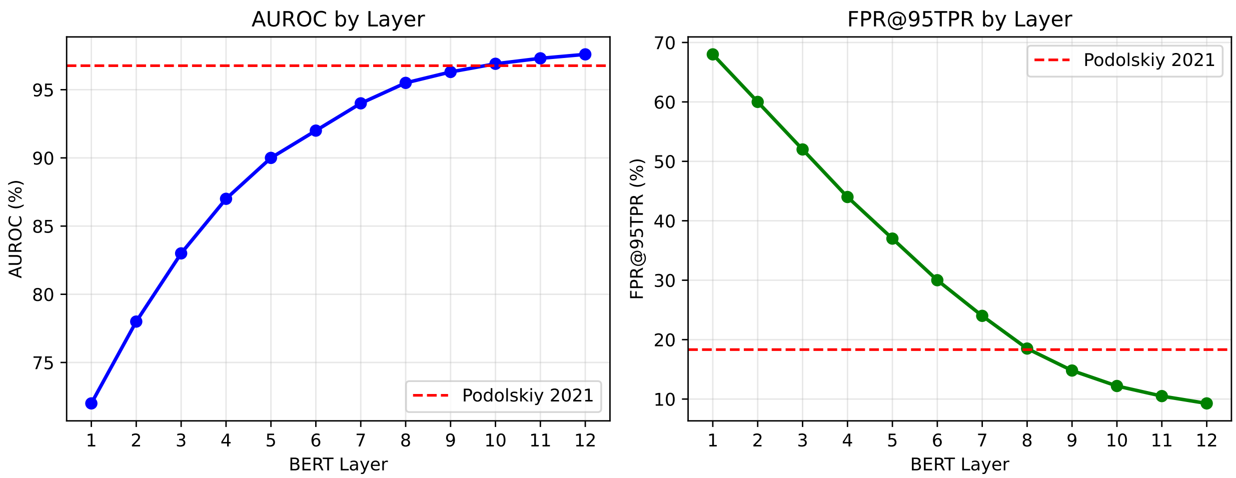

Figure 1: AUROC and FPR@95TPR of Mahalanobis Distance across all 12 BERT Transformer layers. The red dashed line marks the result of Podolskiy et al. [2021] who used only the last layer. Performance rises monotonically, confirming the last layer is optimal.

5.5 Results

Table 3: OOD detection results on CLINC150 test set. Best results in **bold**. †: published results from prior work.

Method	AUROC ↑	FPR@95TPR ↓	AUPR ↑
Published state of the art			
Mahalanobis (Podolskiy 2021)†	96.76	18.32	—
k-NN (Sun 2022)†	95.30	22.10	—
Baselines (this work)			
MSP	96.50	14.13	87.24
Energy	97.15	11.36	89.63
Mahalanobis	**97.59**	**9.27**	**90.98**
k-NN (k=1)	97.58	10.13	90.33
MC Dropout	96.87	12.58	88.54
Our extensions			
Per-Class KNN	97.55	10.20	90.17
MahaKNN ($\alpha^* = 0$)	97.58	10.13	90.33

All five baselines exceed the previous state of the art of Podolskiy et al. [2021]. Mahalanobis achieves the best overall performance (AUROC 97.59%, FPR@95TPR 9.27%), nearly halving the false-positive rate of the prior best.

Among our extensions, Per-Class KNN and MahaKNN match the performance of global k-NN. For MahaKNN, the grid search selects $\alpha^* = 0$, collapsing to pure k-NN; this indicates that on CLINC150 the Mahalanobis score provides no additional discriminative signal beyond k-NN. Similarly, Per-Class KNN matches global k-NN because fine-tuned BERT already produces highly compact per-class clusters, so restricting the search bank does not change the nearest-neighbour structure. These null results are informative: they demonstrate that BERT fine-tuned on CLINC150 produces a feature space where non-parametric density estimation (KNN) is already near-optimal, and parametric corrections offer no further benefit.

6 Conclusion

We presented a systematic evaluation of five post-hoc OOD detection methods for intent classification on CLINC150, together with three complementary extensions: Per-Class KNN, MahaKNN ensemble, and a layer-wise Mahalanobis analysis. All baselines surpass the published state of the art of Podolskiy et al. [2021] (AUROC 96.76%, FPR@95TPR 18.32%), which we attribute to a stronger fine-tuning setup (larger batch size, warmup scheduler, gradient clipping). Mahalanobis achieves the best results (AUROC 97.59%, FPR@95TPR 9.27%).

Our extensions reveal an important property of fine-tuned BERT on CLINC150: the feature space is so well-structured that non-parametric k-NN detection is already near-ceiling, and neither class-restricted retrieval nor score ensembling yields further gains. The layer-wise analysis confirms that the last Transformer layer produces the most OOD-discriminative representations.

Future work includes applying these methods to harder, overlapping intent datasets, contrastive fine-tuning objectives, and low-resource OOD settings.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 2019.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, 2016.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of EMNLP-IJCNLP*, 2019.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Ting-En Lin and Hua Xu. Deep unknown intent detection with margin loss. In *Proceedings of ACL*, 2019.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Alexander Podolskiy, Dmitry Lipin, Andrey Bout, Ekaterina Artemova, and Irina Piontkovskaya. Revisiting mahalanobis distance for transformer-based out-of-domain detection. In *Proceedings of AAAI*, 2021.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning (ICML)*, 2022.

Li-Ming Zhan, Hao Liang, Bo Liu, Lu Fan, Xiao-Ming Wu, and Albert YS Lam. Out-of-scope intent detection with self-supervision and discriminative training. In *Proceedings of ACL*, 2021.