Out-of-Domain Detection for Intent Classification on CLINC150

Your Name

February 20, 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 OOD detection methods applied to a fine-tuned BERT encoder: Maximum Softmax Probability (MSP), Energy Score, Mahalanobis Distance, k-Nearest Neighbors (k-NN), and Monte Carlo Dropout. Mahalanobis Distance achieves the best AUROC of **97.59%** and FPR@95TPR of **9.27%**, surpassing the previous state of the art. 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 benchmarks a range of OOD detection methods applied to the CLINC150 dataset Larson et al. [2019]. The dataset provides 150 intent classes and a dedicated OOS (out-of-scope) partition, making it a de facto standard for intent-OOD research. We evaluate methods along the standard metrics AUROC and FPR@95TPR and compare against published state-of-the-art results.

1.1 Team

This project was prepared by: **Your Name**.

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 pre-trained 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 OOD Detection Methods

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

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

Energy Score.

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

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:

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

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

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

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

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

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 k-NN we use $k = 1$ (cosine similarity). For MC Dropout we run $T = 20$ passes with $p = 0.1$ dropout.

5.3 Baselines

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

5.4 Results

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

Method	AUROC ↑	FPR@95TPR ↓	AUPR ↑
Mahalanobis (Podolskiy 2021)†	96.76	18.32	—
k-NN (Sun 2022)†	95.30	22.10	—
MSP (ours)	96.50	14.13	87.24
Energy (ours)	97.15	11.36	89.63
Mahalanobis (ours)	**97.59**	**9.27**	**90.98**
k-NN k=1 (ours)	97.58	10.13	90.33
MC Dropout (ours)	96.87	12.58	88.54

6 Conclusion

We presented a systematic comparison of five OOD detection methods for intent classification on the CLINC150 benchmark. All methods are applied post-hoc to a fine-tuned BERT encoder, requiring no modification of the training procedure. All five of our methods surpass the previous state of the art of Podolskiy et al. [2021] (AUROC 96.76%, FPR@95TPR 18.32%). Mahalanobis Distance achieves the best AUROC of 97.59% and the lowest FPR@95TPR of 9.27%, nearly halving the false-positive rate of the prior best method. The k-NN detector (k=1) is a close second (AUROC 97.58%), confirming that cosine-similarity-based retrieval in the BERT embedding space is highly effective for OOD detection. Even the simplest baseline, MSP, exceeds the published SotA with AUROC 96.50%, which we attribute to the stronger fine-tuning setup (larger batch, warmup scheduler, gradient clipping).

Future work includes exploring contrastive pre-training objectives, score ensembling, and few-shot OOD exposure.

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.