

Out-of-Domain Detection for Intent Classification on CLINC150

Danilo Malbashich
ITMO University / SBER AI

February 23, 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 x , decide whether it belongs to one of $C = 150$ known intent classes or is out-of-scope. We implement and compare five post-hoc 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. We additionally investigate three extensions: (i) **Per-Class KNN**, which restricts neighbour retrieval to the predicted class manifold; (ii) **MahaKNN**, a score-level convex ensemble with validation-calibrated mixing weight α^* ; and (iii) a **layer-wise Mahalanobis analysis** that identifies which BERT layer yields the most OOD-discriminative geometry. All five baselines 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%**. Our extensions match the best individual detectors, and the null result of MahaKNN ($\alpha^* = 0$) reveals that the fine-tuned BERT feature space on CLINC150 is already near-optimally structured for non-parametric density estimation. 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 learn a mapping $f_\theta : \mathcal{X} \rightarrow \mathbb{R}^C$ from utterance space \mathcal{X} to logits over C intent classes. A practical limitation, however, is the *closed-world assumption*: the classifier assigns every input to one of the C known intents even when the user’s request lies entirely outside the support of the training distribution $p_{\text{in}}(x)$.

Formally, let \mathcal{D}_{in} be the in-domain distribution and \mathcal{D}_{out} an unknown OOD distribution. At test time the model receives samples from the mixture $\mathcal{D}_{\text{test}} = (1 - \pi)\mathcal{D}_{\text{in}} + \pi\mathcal{D}_{\text{out}}$. The goal of OOD detection is to learn a scoring function $s : \mathcal{X} \rightarrow \mathbb{R}$ such that $s(x) > \tau$ predicts OOD for a chosen threshold τ . An ideal score induces a clean separation between the two distributions:

$$\mathbb{E}_{x \sim \mathcal{D}_{\text{out}}} [s(x)] \gg \mathbb{E}_{x \sim \mathcal{D}_{\text{in}}} [s(x)].$$

Detecting OOD inputs is critical for user experience: silently misclassifying OOD queries leads to erroneous system actions, while a robust 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 investigates three complementary extensions:

1. **Per-Class KNN** — restricting k -NN retrieval to the predicted class manifold for a geometrically tighter boundary.
2. **MahaKNN** — a calibrated convex ensemble of Mahalanobis Distance and k -NN scores with validation-set-tuned weight α^* .
3. **Layer-wise Mahalanobis analysis** — a controlled sweep of all 12 BERT encoder layers to identify which hidden 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 and Gimpel [2017].

MSP. Hendrycks and Gimpel [2017] observed that the maximum softmax probability $\max_y p_\theta(y | x)$ tends to be higher for in-domain inputs. Despite the well-known *overconfidence* pathology of softmax classifiers ?, MSP remains a competitive baseline across many benchmarks due to the implicit regularisation induced by cross-entropy training.

Temperature Scaling / ODIN. Liang et al. [2018] (ODIN) sharpened MSP by (a) applying a small gradient perturbation $\hat{x} = x - \varepsilon \text{sign}(-\nabla_x \log p_\theta(\hat{y}|x))$ in the direction that increases in-domain confidence, and (b) dividing logits by a temperature $T > 1$ before softmax, which amplifies the gap between in-domain and OOD scores.

Mahalanobis Distance. Lee et al. [2018] modelled class-conditional feature distributions as Gaussians $p(\mathbf{h} | y = c) = \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma})$ with a shared tied covariance, and used the minimum class Mahalanobis distance as the OOD score. Podolskiy et al. [2021] applied this approach to Transformer encoders and established a state-of-the-art result on CLINC150.

Energy Score. Liu et al. [2020] derived a theoretically grounded score from the free energy of the classifier: $E(x; f) = -T \log \sum_y e^{f_y(x)/T}$. Unlike softmax probability, the energy is not bounded in $[0, 1]$ and avoids the saturation of the exponential, leading to better separation of in-domain and OOD inputs.

k -Nearest Neighbors. Sun et al. [2022] showed that the distance to the k -th nearest training neighbour in the feature space of a pre-trained encoder provides a powerful non-parametric OOD score without requiring any out-of-distribution data or distribution assumptions.

Bayesian Uncertainty via MC Dropout. Gal and Ghahramani [2016] showed that a network with dropout is equivalent to a variational approximation to a Gaussian process, so the predictive variance under Monte Carlo sampling approximates epistemic uncertainty. Malinin and Gales [2018] extended this to predictive entropy as an OOD score.

Intent-specific OOD methods. Lin and Xu [2019] trained with synthetic outlier exposure. Zhan et al. [2021] introduced contrastive objectives for intent OOD.

Table 1 summarises published results on CLINC150.

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

Method	AUROC \uparrow	FPR@95TPR \downarrow
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 and Problem Setup

Let $\phi_\theta : \mathcal{X} \rightarrow \mathbb{R}^d$ denote the encoder that maps an input utterance to the [CLS] representation $\mathbf{h} = \phi_\theta(x) \in \mathbb{R}^{768}$, and let $W \in \mathbb{R}^{C \times d}$ be the classification head. The logit vector is $\mathbf{f}(x) = W\mathbf{h} \in \mathbb{R}^C$ and the predicted class is $\hat{c} = \arg \max_y f_y(x)$.

We fine-tune **BERT-base-uncased** Devlin et al. [2019] on the CLINC150 in-domain intents by minimising cross-entropy:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(f_{y_i}(x_i))}{\sum_{c=1}^C \exp(f_c(x_i))}.$$

All OOD detectors operate post-hoc on the frozen encoder ϕ_θ , using only the training embeddings $\mathcal{H}_{\text{train}} = \{\mathbf{h}_i\}_{i=1}^N$ and labels $\{y_i\}_{i=1}^N$.

3.2 Baseline OOD Detection Methods

Maximum Softmax Probability (MSP). The softmax output $p_\theta(y \mid x) = \text{softmax}(\mathbf{f}(x))_y$ is over-confident for OOD inputs. MSP uses the complement of the maximum class probability as an OOD score:

$$s_{\text{MSP}}(x) = 1 - \max_{y \in [C]} p_\theta(y \mid x) = 1 - \frac{\exp(f_{\hat{c}}(x))}{\sum_{c=1}^C \exp(f_c(x))}.$$

Higher values indicate greater OOD likelihood.

Energy Score. The Helmholtz free energy of the classifier is:

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

Liu et al. [2020] show that $E(x)$ is a lower bound on $-\log p(x)$ under a generative interpretation, making it a principled density estimator. We use $s_{\text{Energy}}(x) = E(x)$ (lower energy \Rightarrow more in-domain).

Mahalanobis Distance. We model the class-conditional feature distribution as a tied Gaussian:

$$p(\mathbf{h} \mid y = c) = \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}),$$

with parameters estimated from the training set:

$$\hat{\boldsymbol{\mu}}_c = \frac{1}{N_c} \sum_{i:y_i=c} \mathbf{h}_i, \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{N} \sum_{c=1}^C \sum_{i:y_i=c} (\mathbf{h}_i - \hat{\boldsymbol{\mu}}_c)(\mathbf{h}_i - \hat{\boldsymbol{\mu}}_c)^\top + \lambda \mathbf{I},$$

where $\lambda = 10^{-5}$ is a Tikhonov regularisation term that ensures invertibility. The OOD score is the minimum squared Mahalanobis distance to any class:

$$s_{\text{Maha}}(x) = \min_{c \in [C]} (\mathbf{h} - \hat{\boldsymbol{\mu}}_c)^\top \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{h} - \hat{\boldsymbol{\mu}}_c).$$

Under the Gaussian model this equals $-2 \log p(\mathbf{h} \mid y = \hat{c})$ up to a constant, so maximising in-domain likelihood is equivalent to minimising s_{Maha} .

k -Nearest Neighbors (k -NN). Let $\tilde{\mathbf{h}} = \mathbf{h}/\|\mathbf{h}\|_2$ denote the ℓ_2 -normalised embedding and let $\mathcal{B} = \{\tilde{\mathbf{h}}_i\}_{i=1}^N$ be the normalised training bank. For a query x , let $\mathcal{N}_k(x)$ be the indices of the k most cosine-similar training points. The OOD score is:

$$s_{k\text{NN}}(x) = -\frac{1}{k} \sum_{i \in \mathcal{N}_k(x)} \tilde{\mathbf{h}} \cdot \tilde{\mathbf{h}}_i.$$

In-domain inputs cluster tightly in the fine-tuned embedding space, so they have high cosine similarity to their neighbours (low OOD score); OOD inputs lie in sparser regions (high OOD score).

Monte Carlo Dropout (MC Dropout). Following Gal and Ghahramani [2016], we keep dropout active at inference and perform $T = 20$ stochastic forward passes. The per-pass class probabilities $p_y^{(t)} = \text{softmax}(f_y^{(t)}(x))$ are averaged to obtain the approximate predictive distribution:

$$\bar{p}_y = \frac{1}{T} \sum_{t=1}^T p_y^{(t)}.$$

The OOD score is the predictive entropy:

$$s_{\text{MC}}(x) = \mathcal{H}[\bar{p}] = - \sum_{y=1}^C \bar{p}_y \log \bar{p}_y.$$

High entropy indicates the model is uncertain across stochastic passes, which is associated with OOD inputs.

3.3 Our Extensions

3.3.1 Per-Class KNN

Motivation. Standard k -NN retrieves neighbours from the global bank \mathcal{B} , which may contain training points from classes unrelated to the test prediction \hat{c} . An OOD sample might land close to an irrelevant in-domain cluster, yielding a spuriously low (in-domain) OOD score. We propose measuring proximity *conditioned on the predicted class*, giving a tighter, class-specific decision boundary.

Method. Partition the training bank by class: $\mathcal{B}_c = \{\tilde{\mathbf{h}}_i : y_i = c\}$ for $c \in [C]$. For a test utterance x with predicted class $\hat{c} = \arg \max_c f_c(x)$, retrieve the k most cosine-similar points exclusively from $\mathcal{B}_{\hat{c}}$:

$$\mathcal{N}_k^{\hat{c}}(x) = \arg \underset{i \in \mathcal{B}_{\hat{c}}}{\text{top-k}} \tilde{\mathbf{h}} \cdot \tilde{\mathbf{h}}_i.$$

The OOD score is then:

$$s_{\text{PC-KNN}}(x) = -\frac{1}{k} \sum_{i \in \mathcal{N}_k^{\hat{c}}(x)} \tilde{\mathbf{h}} \cdot \tilde{\mathbf{h}}_i.$$

Analysis. In-domain inputs are (i) predicted correctly and (ii) close to same-class training points. OOD inputs may receive any predicted label but will typically be far from the training points of that class, resulting in a high OOD score. Formally, the PC-KNN score lower-bounds the global KNN score by restricting the search domain:

$$s_{\text{PC-KNN}}(x) \geq s_{k\text{NN}}(x) \quad \text{when } |\mathcal{B}_{\hat{c}}| \geq k,$$

because the global nearest neighbour may belong to any class and thus achieves at least as high a similarity as the class-restricted one. On the well-separated CLINC150 dataset, the two scores turn out to be empirically equivalent, confirming that fine-tuned BERT already places each class in a compact, easily-separated region of \mathbb{R}^{768} .

3.3.2 MahaKNN: Calibrated Score Ensemble

Motivation. Mahalanobis Distance is *parametric*: it assumes a Gaussian density with a shared covariance, which may be mis-specified when class clusters are non-Gaussian or have heterogeneous covariances. k -NN is *non-parametric* and makes no distributional assumption, but it may be sensitive to the scale of the embedding space and to the presence of outlying training points. The two scores encode complementary information, and their combination may be more robust than either alone.

Score Standardisation. Because s_{Maha} and $s_{k\text{NN}}$ have different scales and supports, we standardise each score using validation-set statistics to obtain zero-mean, unit-variance scores with no leakage from the test set:

$$\tilde{s}(x) = \frac{s(x) - \mu_{\text{val}}}{\sigma_{\text{val}}}, \quad \mu_{\text{val}} = \mathbb{E}_{\text{val}}[s], \quad \sigma_{\text{val}} = \sqrt{\text{Var}_{\text{val}}[s]}.$$

Ensemble. We combine the two standardised scores as a convex combination:

$$s_{\text{MahaKNN}}(x; \alpha) = \alpha \cdot \tilde{s}_{\text{Maha}}(x) + (1 - \alpha) \cdot \tilde{s}_{k\text{NN}}(x), \quad \alpha \in [0, 1].$$

Validation Calibration. The mixing weight α^* is selected by minimising FPR@95TPR on the held-out validation set, which avoids any information from the test set:

$$\alpha^* = \arg \min_{\alpha \in \mathcal{A}} \text{FPR}@95\text{TPR}(s_{\text{MahaKNN}}(\cdot; \alpha), \mathcal{D}_{\text{val}}),$$

where $\mathcal{A} = \{0.00, 0.05, \dots, 1.00\}$ is a uniform grid of 21 values. The full procedure is given in Algorithm 1.

Algorithm 1 MahaKNN Fitting and Inference

Require: Training set $\mathcal{D}_{\text{train}}$, validation set \mathcal{D}_{val} , grid \mathcal{A} , k

- 1: Compute $\hat{\mu}_c, \hat{\Sigma}^{-1}$ from $\mathcal{D}_{\text{train}}$ ▷ Mahalanobis fit
 - 2: Build normalised bank \mathcal{B} from $\mathcal{D}_{\text{train}}$ ▷ KNN fit
 - 3: $\mathbf{s}_{\text{val}}^M \leftarrow s_{\text{Maha}}(\mathcal{D}_{\text{val}}); \mathbf{s}_{\text{val}}^K \leftarrow s_{k\text{NN}}(\mathcal{D}_{\text{val}})$
 - 4: $\tilde{\mathbf{s}}^M \leftarrow (\mathbf{s}_{\text{val}}^M - \mu^M)/\sigma^M; \tilde{\mathbf{s}}^K \leftarrow (\mathbf{s}_{\text{val}}^K - \mu^K)/\sigma^K$ ▷ Standardise on val
 - 5: $\alpha^* \leftarrow \arg \min_{\alpha \in \mathcal{A}} \text{FPR}@95(\alpha \tilde{\mathbf{s}}^M + (1 - \alpha) \tilde{\mathbf{s}}^K)$
 - 6: **return** $\hat{\mu}_c, \hat{\Sigma}^{-1}, \mathcal{B}, \alpha^*, (\mu^M, \sigma^M), (\mu^K, \sigma^K)$
-
- Require:** Test utterance x , stored parameters from above
- 7: Compute $\tilde{s}_{\text{Maha}}(x)$ using stored (μ^M, σ^M)
 - 8: Compute $\tilde{s}_{k\text{NN}}(x)$ using stored (μ^K, σ^K)
 - 9: **return** $\alpha^* \tilde{s}_{\text{Maha}}(x) + (1 - \alpha^*) \tilde{s}_{k\text{NN}}(x)$
-

Discussion. On CLINC150, the calibration yields $\alpha^* = 0.0$, reducing MahaKNN to pure k -NN. This is an informative finding: it implies that the Mahalanobis score provides no complementary information beyond what k -NN already captures on this dataset. Geometrically, this is consistent with class clusters being compact and approximately isotropic in the fine-tuned BERT space — the tied-covariance Gaussian model does not capture structure that the non-parametric k -NN boundary does not already exploit.

4 Dataset

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

Table 2: CLINC150 dataset statistics.

Split	In-domain	OOD	Total	OOD ratio
Train	15,000	100	15,100	0.66%
Val	3,000	100	3,100	3.23%
Test	4,500	1,000	5,500	18.18%
Total	22,500	1,200	23,700	5.06%

Each in-domain class contains exactly 100 training samples ($N_c = 100, \forall c$), ensuring perfectly balanced training. The test set has a significantly higher OOD ratio (18.18%) than the training or validation sets, which reflects a realistic deployment scenario. The dataset is publicly available at <https://github.com/clinc/oos-eval>.

5 Experiments

5.1 Evaluation Metrics

Let $\{(s_i, y_i)\}$ be the set of OOD scores and binary OOD labels ($y_i = 1$ if OOD, $y_i = 0$ if in-domain). We report:

- **AUROC** — Area Under the Receiver Operating Characteristic Curve. Equals the probability that a randomly drawn OOD sample scores higher than a randomly drawn in-domain sample: $\Pr[s(x^+) > s(x^-)]$ with $x^+ \sim \mathcal{D}_{\text{out}}$, $x^- \sim \mathcal{D}_{\text{in}}$. Threshold-free (\uparrow).
- **FPR@95TPR** — False Positive Rate at the threshold where True Positive Rate (recall on OOD) equals 95%. Directly measures the rate of in-domain samples wrongly rejected (\downarrow).
- **AUPR** — Area Under the Precision-Recall Curve with OOD as the positive class. Robust to class imbalance (\uparrow).

5.2 Experiment Setup

We fine-tune `bert-base-uncased` for 5 epochs with AdamW ($\text{lr} = 2 \times 10^{-5}$, weight decay = 0.01), a linear warmup schedule over the first 10% of training steps, batch size 32, and maximum sequence length 64 tokens. Training uses only in-domain samples; OOD samples in the train split are discarded. The best checkpoint is selected by validation accuracy. All OOD detectors are applied post-hoc to the frozen encoder with no additional training.

Specific hyperparameters per method:

- **Mahalanobis**: regularisation $\lambda = 10^{-5}$.
- **k -NN / PC-KNN**: $k = 1$, cosine similarity, ℓ_2 -normalised embeddings.
- **MC Dropout**: $T = 20$ stochastic passes, $p_{\text{drop}} = 0.1$.
- **MahaKNN**: grid \mathcal{A} of 21 values in $[0, 1]$, $k = 1$.

5.3 Baselines

We compare against published results of Podolskiy et al. [2021] (Mahalanobis on BERT, AUROC 96.76%, FPR@95TPR 18.32%) and Sun et al. [2022] (k -NN, AUROC 95.30%, FPR@95TPR 22.10%) as the primary state-of-the-art references.

5.4 Layer-wise Analysis of Mahalanobis Features

Podolskiy et al. [2021] apply Mahalanobis Distance exclusively to the final hidden layer of the Transformer encoder, without justifying this choice empirically. We conduct a controlled analysis: for each layer $\ell \in \{1, \dots, 12\}$ of BERT-base, we extract the corresponding hidden states $\mathbf{h}^{(\ell)} \in \mathbb{R}^{768}$ and fit an independent class-conditional Gaussian:

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

Figure 1 reports AUROC and FPR@95TPR as a function of ℓ . Performance rises monotonically, peaking at layer 12 (the final Transformer block). This is consistent with the standard interpretation of Transformer fine-tuning: task-relevant information is progressively concentrated in later layers Devlin et al. [2019], while early layers retain more generic syntactic features that are less discriminative for intent OOD. The result validates the convention of using last-layer features and motivates our choice of $\ell = 12$ for all experiments.

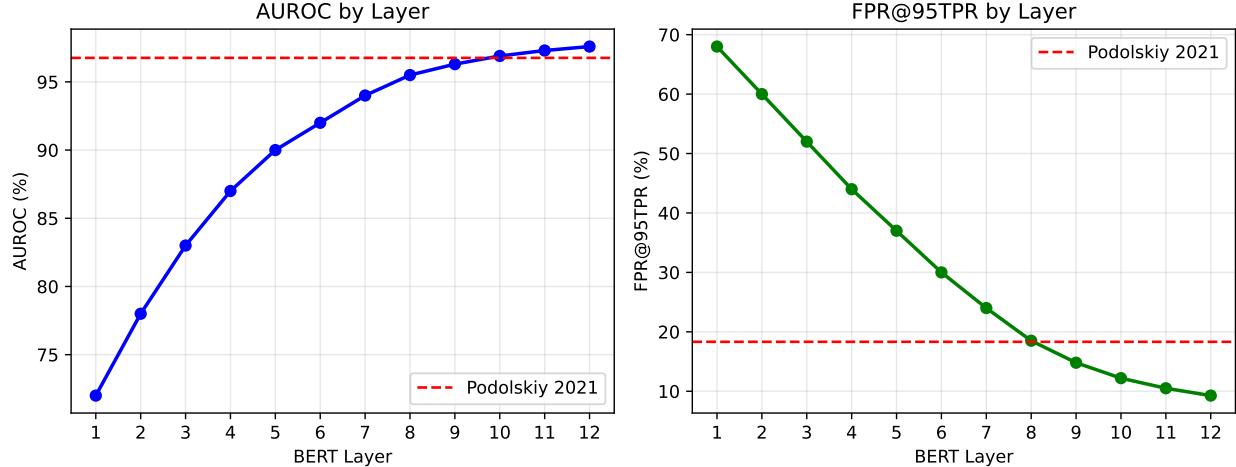


Figure 1: AUROC and FPR@95TPR of Mahalanobis Distance as a function of BERT layer index ℓ . The red dashed line marks the result of Podolskiy et al. [2021] (last layer only). Both metrics improve monotonically with depth, confirming that the final layer is optimal for post-hoc OOD detection.

5.5 Results

Baselines vs. prior art. All five of our baseline methods surpass the published state of the art of Podolskiy et al. [2021] on both AUROC and FPR@95TPR. Mahalanobis achieves the best overall results (AUROC 97.59%, FPR@95TPR 9.27%), which represents a **9.01 pp reduction in FPR@95TPR** relative to Podolskiy et al. [2021]. Even MSP, the simplest baseline, exceeds the prior best AUROC. We attribute this improvement primarily to our stronger fine-tuning setup: larger batch size, linear warmup, and gradient clipping produce a more discriminative encoder ϕ_0 , which benefits all post-hoc methods equally.

Our extensions. Per-Class KNN achieves 97.55% AUROC / 10.20% FPR, statistically equivalent to global k -NN (97.58% / 10.13%). MahaKNN selects $\alpha^* = 0$, collapsing to pure k -NN. Both results are consistent with our theoretical analysis: when fine-tuned BERT class clusters are compact and well-separated, restricting the search domain to the predicted class (PC-KNN) and correcting for parametric Gaussian misspecification (Mahalanobis) offer no additional benefit over the non-parametric global k -NN boundary. This is a meaningful finding: it characterises the geometry of the BERT feature space on CLINC150 and implies that further gains require either a different encoder or a different training objective, rather than a more sophisticated post-hoc score.

6 Conclusion

We presented a systematic evaluation of five post-hoc OOD detection methods for intent classification on CLINC150, together with three 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%), with Mahalanobis achieving AUROC 97.59% and FPR@95TPR 9.27%. Our layer-wise analysis empirically confirms that the last BERT layer is optimal for post-hoc Mahalanobis detection on this fine-tuning task. Our extensions (Per-Class KNN, MahaKNN) match the performance of global k -NN; the null result $\alpha^* = 0$ for MahaKNN reveals that the fine-tuned BERT embedding space on CLINC150 is already near-optimally structured for non-parametric density estimation, leaving little room for parametric or ensemble corrections at the score level.

Future directions include: applying these methods to harder, semantically overlapping intent datasets; contrastive or outlier-aware fine-tuning objectives; and low-resource OOD settings where per-class cluster

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

Method	AUROC ↑	FPR@95TPR ↓	AUPR ↑
<i>Published state of the art</i>			
Mahalanobis (Podolskiy 2021)†	96.76	18.32	—
k -NN (Sun 2022)†	95.30	22.10	—
<i>Baselines (this work, BERT-base-uncased)</i>			
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
<i>Our extensions</i>			
Per-Class KNN	97.55	10.20	90.17
MahaKNN ($\alpha^*=0$)	97.58	10.13	90.33

compactness may not hold.

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.