# BOLT: Benchmarking Open-World Learning for Text Classification

**Chuan Qin**[1,2,♣], **Xi Chen**[3,1,♣], **Jinpeng Li**[1,2,♣], **Hengshu Zhu**[1,2,♠]
[1] Computer Network Information Center, Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences
[3] School of Computer Science and Technology, University of Science and Technology of China
{chuanqin0426, zhuhengshu}@gmail.com
chenxi0401@mail.ustc.edu.cn, lijinpeng25@mails.ucas.ac.cn
♣ Equal contribution ♠ Corresponding author

## Abstract

Text classification has long been a cornerstone of NLP, yet most prior work and benchmarks have been limited to closed-world settings, where all classes are assumed to be known in advance. In contrast, open-world learning has recently emerged as a critical paradigm for building more robust and realistic systems. However, existing benchmarks largely focus on out-of-distribution (OOD) detection, while overlooking broader challenges such as the discovery of novel categories. To address this gap, we introduce **BOLT**, a unified **B**enchmark and evaluation toolkit supporting **O**pen-world **L**earning for **T**ext classification. BOLTencompasses two representative tasks: Open-set Text Classification (OSTC), which requires models to classify in-distribution (ID) samples while rejecting OOD inputs, and Generalized Category Discovery (GCD), which aims to identify both known and novel categories from partially labeled corpora. We carefully curate 12 publicly available datasets spanning diverse domains and benchmark 22 methods, including 15 for OSTC and 7 for GCD, under a standardized protocol that explicitly accounts for varying labeled ratios and known class ratios. Our results reveal key challenges: most current methods tend to overfit training distributions and struggle to generalize to unseen classes. Moreover, by comparing our lightweight LLM-based variants with prior open-set baselines, we demonstrate the promise of leveraging LLMs for open-world text classification. BOLT provides standardized evaluation protocols that enable fair comparison and support future research in this emerging area. All datasets, baselines, and tools are available at https://github.com/CNIC-DSL/BOLT.

## 1 Introduction

Text classification has long been a central task in natural language processing (NLP), supporting diverse applications such as sentiment analysis (Wankhade et al., 2022; Qin et al., 2025c), topic labeling (Chen et al., 2020; Wang et al., 2023; Chen et al., 2024; Qin et al., 2025b), intent recognition (Zhang et al., 2024a), and document management (Shu et al., 2017). Traditional research in this area has largely relied on closed-world assumptions, where the set of target classes is predefined and assumed to be fully observable during training (Minaee et al., 2021). Such an assumption limits applicability in real-world scenarios, where label spaces are inherently open and novel categories continuously emerge (Fei and Liu, 2016).

To address these limitations, recent research has shifted toward open-world learning, a paradigm originally studied in computer vision (Li and Wechsler, 2005; Zhang et al., 2023) and later extended to NLP (Prakhya et al., 2017). Building on this foundation, open-world text classification (OTC) requires models to classify ID data while remaining robust to emerging categories. Current research on OTC primarily investigates two representative tasks. The first, OSTC, addresses the challenge of classifying ID inputs while reliably rejecting OOD samples, under the restrictive setting where only labeled ID data are available for training (Chen et al., 2026). This task requires learning robust decision boundaries that generalize beyond the observed training distribution (Shu et al., 2017; Zhang et al., 2021a). The second, GCD, extends this objective by discovering novel categories from unlabeled corpora through clustering-based pseudo-labeling, which requires not only precise ID classification but also effective structure induction to accommodate emerging classes (Zhang et al., 2021b, 2024b). By integrating these two perspectives, OTC bridges the gap between closed-world classification and the demands of dynamic real-world environments.

However, existing OSTC and GCD methods primarily rely on a few intent recognition datasets for evaluation, such as Banking and CLINC. This line of evaluation suffers from inherent limitations, as it typically focuses on a single task type and is



*Findings of the Association for Computational Linguistics: ACL 2026*, pages 13624–13658
July 2-7, 2026 

| Task | KCR | LAR | Evaluation Metrics | Paper List |
|---|---|---|---|---|
| OSTC | 0.250, 0.500, 0.750 | 1.000 | F1 | DOC [(Shu et al., 2017)], DeepUnk [(Lin and Xu, 2019)] |
| | 0.250, 0.500, 0.750 | 1.000 | ACC, F1, K-F1, N-F1 | ADB [(Zhang et al., 2021a)], AB [(Lorenc et al., 2022)],KNNCon [(Zhou et al., 2022)] |
| | 0.667 | 1.000 | ACC, F1, Recall | SCL [(Zeng et al., 2021a)] |
| | 0.250, 0.750 | 1.000 | ACC, F1, K-F1, N-F1 | DyEn [(Zhou et al., 2023)] |
| | 0.250, 0.500, 0.750 | 0.100, 0.500, 1.000 | ACC, F1, K-F1, N-F1 | **BOLT** |
| GCD | 0.750 | 0.100 | NMI, ARI, ACC | DAL [(Zhang et al., 2021b)] |
| | 0.250 | 0.100 | NMI, ARI, ACC | GeoID [(Tang et al., 2024)] |
| | 0.750 | 0.075 | ACC | DPN [(An et al., 2023)] |
| | 0.750 | 0.100 | H-score, K-ACC, N-ACC | SDC [(An et al., 2025)], TAN [(An et al., 2024b)], LOOP [(An et al., 2024a)] |
| | 0.250, 0.500 | 0.100 | ACC, ARI, NMI | ALUP [(Liang et al., 2024)] |
| | 0.250, 0.500, 0.750 | 0.100 | ACC, ARI, NMI | USNID [(Zhang et al., 2023)] |
| | 0.250, 0.500, 0.750 | 0.100, 0.500, 1.000 | ACC, ARI, NMI, H-score, K-ACC, N-ACC | **BOLT** |

Table 1: Comparison of evaluation settings for OSTC and GCD.

confined to homogeneous text domains and data characteristics; for instance, these datasets exhibit narrowly concentrated length distributions and are uniformly in English. This narrow coverage hinders systematic investigation of model generalization across diverse scenarios. In addition, existing studies often adopt inconsistent experimental protocols, such as using different proportions of known classes, varying amounts of labeled data, or divergent evaluation metrics, which further complicates fair comparison across methods as shown in Table 1. In particular, OSTC is seldom evaluated under low-resource settings, while GCD has seen only limited exploration across varying levels of supervisory signals. Consequently, **the field still lacks a comprehensive benchmark and robust evaluation toolkit** to support systematic and reliable exploration of OTC. Moreover, although LLMs have been widely adopted in NLP, their use in OTC remains limited. Current attempts mainly exploit data augmentation to improve GCD (Liang et al., 2024), while other directions, including employing LLMs as backbones for representation learning, are largely unexplored.

To this end, we present **BOLT**, a novel and comprehensive **B**enchmark and evaluation toolkit designed to support **O**pen-world **L**earning for **T**ext classification. Specifically, we first unify two representative OTC tasks, OSTC and GCD, within a single evaluation framework. To enhance the diversity of our benchmark, we carefully curate 12 publicly available datasets that span a wide range of text classification scenarios, including intent detection, topic categorization, and ontology classification. These datasets cover diverse text domains (e.g., news, conversational queries, and question-answer), exhibit varying text lengths (with an average document length ranging from 8 to 1,856 words), differ in label granularity (with class counts between 10 and 219), and vary in multiple languages. Along this line, we conduct a systematic and fine-grained evaluation of 22 representative methods, including 15 designed for OSTC and 7 for GCD, under a standardized protocol that explicitly accounts for different labeled and known-class ratios. The experimental results not only reveal the limitations of existing methods but also demonstrate the promise of **leveraging LLMs for advancing different OTC tasks**.

## 2 Related Work

In this section, we review prior work on the two representative OTC tasks, i.e., OSTC and GCD.

### 2.1 Open-Set Text Classification

OSTC is a challenging task that requires models to classify ID textual inputs into known categories while detecting OOD inputs (Prakhya et al., 2017). This dual objective contrasts with conventional OOD detection, which is typically framed as a binary discrimination problem between ID and OOD samples (Liu et al., 2020). Early studies on this problem setting mainly originated in computer vision, where maximum softmax probability (MSP) (Hendrycks and Gimpel, 2016) was introduced for OOD detection. However, this approach merely rejects uncertain predictions without explicitly modeling unknown classes (Bendale and Boult, 2016). OpenMax attempted to improve MSP by fitting a Weibull distribution to classifier logits to estimate OOD probabilities (Bendale and Boult, 2016). However, its reliance on the assumption that equiprobable logits indicate unknown classes often leads to misclassification of hard ID examples. Extending to the text domain, DOC introduced a one-vs-rest sigmoid layer to mitigate open-space risk (Shu et al., 2017). Nevertheless, threshold-based approaches as a whole remain limited in capturing the intrinsic differences between known and unknown classes.

Subsequent work shifted from thresholding to enhancing representation learning. DeepUnk (Lin

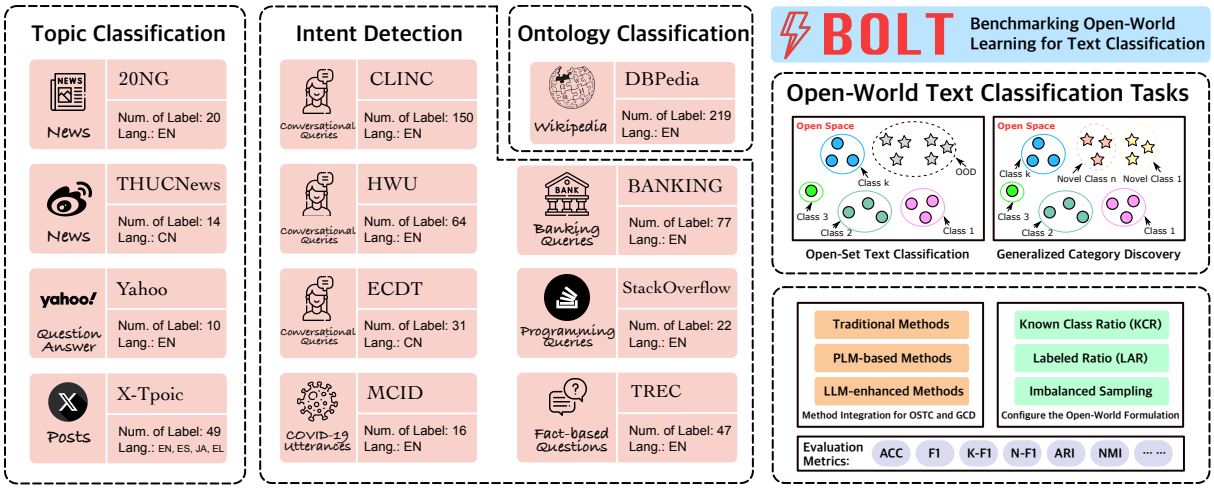


Figure 1: An overview of our BOLT benchmark for open-world text classification, covering two OTC tasks (OSTC and GCD) over 12 datasets from three scenarios (topic, intent, ontology) across eight domains. BOLT supports configurable open-world settings via KCR, LAR, and imbalanced sampling, and provides a standardized pipeline to evaluate traditional, PLM-based, and LLM-enhanced methods with consistent metrics for both tasks.

and Xu, 2019) adopted a large-margin cosine loss to better separate features, while ADB (Zhang et al., 2021a) refined decision boundaries to improve OOD discrimination. KNNCon (Zhou et al., 2022) integrated KNN-based contrastive learning with LOF-based detection, and DyEn (Zhou et al., 2023) introduced dynamic ensembling with early exits to mitigate overfitting. Recently, the progress of LLMs has significantly advanced the field of NLP, enabling notable improvements in semantic understanding and text generation (Zhao et al., 2023; Jiang et al., 2024; Tong et al., 2025; Qin et al., 2025a; Zhou et al., 2025; Huang et al., 2026; Song et al., 2026). Although a few initial studies have begun to explore the use of LLMs for OOD detection (Liu et al., 2023), their potential for OSTC has not yet been investigated. In addition, most existing OSTC evaluations focus on a narrow range of supervision settings, leaving low-resource scenarios largely underexplored.

### 2.2 Generalized Category Discovery

In contrast to OSTC, GCD aims to identify both known and novel categories from partially labeled data. Most existing approaches first learn representations from labeled known classes and then apply clustering-based pseudo-labeling to discover novel categories. Early methods such as DeepAligned (Zhang et al., 2021b) employed pairwise-similarity supervision or alignment strategies to guide clustering. However, the accuracy of pseudo-labels remains a critical bottleneck, as noise propagation can severely degrade overall performance. Subsequent studies sought to mitigate pseudo-label noise through two main strategies. Pseudo-label calibration approaches, including GeoID (Tang et al., 2024) and SDC (An et al., 2025), alleviate the dominance of known categories during clustering. Prototype-based transfer methods, including TAN (An et al., 2024b) and DPN (An et al., 2023), facilitate knowledge transfer from known to novel classes.

Recently, LLMs have been incorporated into GCD. LOOP (An et al., 2024a) and ALUP (Liang et al., 2024) leverage LLMs to refine neighborhoods or enable exemplar-based comparison.

## 3 The BOLT Benchmark

### 3.1 Problem Definition

We consider the OTC problem, where models are required to handle both ID and OOD samples under varying supervision settings. Let the textual input space be $\mathcal{X}$ and the label space be $\mathcal{Y} = \mathcal{Y}^k \cup \mathcal{Y}^n$, where $\mathcal{Y}^k = \{1, \ldots, K\}$ denotes the set of known (predefined) classes, and $\mathcal{Y}^n = \{K+1, \ldots, K+N\}$ represents the set of novel classes that may appear during testing.

Given the dataset setup: a labeled set $\mathcal{D}^l = \{(x_i, y_i)\}_{i=1}^{N_l}$, where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}^k$; an unlabeled set $\mathcal{D}^u = \{x_j\}_{j=1}^{N_u}$, which may contain samples drawn from both $\mathcal{Y}^k$ and $\mathcal{Y}^n$; and a test set $\mathcal{D}^t = \{(x_i, y_i)\}_{i=1}^{N_t}$, where $y_i \in \mathcal{Y}^k \cup \mathcal{Y}^n$. Based on the availability of labeled and unlabeled data, as well as the target prediction objectives, we define two representative OTC tasks.

| Scenario | Text Domain | Dataset | Lang. | # Inst. | # Lab. | Avg. Len./ Inst. | Avg. Len./ Lab. |
|---|---|---|---|---|---|---|---|
| Topic Classification | Usenet News | 20NG | EN | 10000 | 20 | 290.41 | 2.80 |
| Topic Classification | Sina News | THUCNews | CN | 9421 | 14 | 749.07 | 2.00 |
| Topic Classification | Yahoo pages | Yahoo | EN | 10000 | 10 | 95.19 | 1.80 |
| Topic Classification | X pages | X-Topic | Multi | 8587 | 49 | 26.86 | 1.00 |
| Intent Detection | Online banking queries | BANKING | EN | 13083 | 77 | 11.74 | 3.35 |
| Intent Detection | Conversational queries | CLINC | EN | 22500 | 150 | 8.23 | 1.94 |
| Intent Detection | Programming queries | StackOverflow | EN | 19985 | 20 | 8.36 | 1.05 |
| Intent Detection | Conversational queries | HWU | EN | 9677 | 64 | 6.60 | 2.12 |
| Intent Detection | Fact-based questions | TREC | EN | 5871 | 47 | 8.72 | 2.00 |
| Intent Detection | Conversational queries | ECDT | CN | 3069 | 31 | 7.87 | 1.00 |
| Intent Detection | COVID-19 utterance | MCID | EN | 1699 | 16 | 6.87 | 3.00 |
| Ontology Classification | Wikipedia pages | DBPedia | EN | 10000 | 219 | 98.31 | 1.00 |

Table 2: Dataset statistics in BOLT. We report the total number of instances (# Inst.), the number of labels (# Lab.), the average text length per instance (Avg. Len./ Inst.), and the average text length per label (Avg. Len./ Lab.).

**Open-Set Text Classification (OSTC)** In this setting, the model is trained only on the labeled data $\mathcal{D}^l$ containing known classes $\mathcal{Y}^k$. During testing, it encounters OOD samples belonging to the novel label space $\mathcal{Y}^n$, which is not observed during training. The objective is to learn a $K+1$ classifier:

$$f_{\text{OOD}} : \mathcal{X} \rightarrow \mathcal{Y}^k \cup \{\text{Unknown}\}, \quad (1)$$

that correctly classifies ID samples into $\mathcal{Y}^k$ and identifies OOD samples by assigning them to the "Unknown" class representing $\mathcal{Y}^n$.

**Generalized Category Discovery (GCD)** In this setting, both the labeled data $\mathcal{D}^l$ and the unlabeled data $\mathcal{D}^u$ are available during training. The unlabeled data contain a mixture of known and novel classes, i.e., $x_j \in \mathcal{X}$ with $y_j \in \mathcal{Y}^k \cup \mathcal{Y}^n$. The task requires the model to jointly recognize known categories and discover novel ones. Formally, the model learns from $\mathcal{D}^l \cup \mathcal{D}^u$ to construct a model:

$$f_{\text{GCD}} : \mathcal{X} \rightarrow \mathcal{Y}^k \cup \mathcal{Y}^n, \quad (2)$$

where the number of novel categories $|\mathcal{Y}^n|$ is assumed to be known (An et al., 2023; Shi et al., 2024). Evaluation is conducted on $\mathcal{D}^t$, assessing the model's ability to distinguish and correctly assign both seen and unseen categories.

### 3.2 Dataset and Diversity Analysis

To construct a comprehensive benchmark for OTC, BOLT aims to provide a unified evaluation framework that encompasses diverse text classification scenarios and domains. In total, BOLT includes 12 datasets spanning three major scenarios: topic classification, intent detection, and ontology classification, covering a wide range of text sources such as news articles, conversational queries, and Wikipedia entries. The datasets include 20NG (Lang, 1995), THUCNews (Sun et al., 2016), Yahoo (Zhang et al., 2015a), BANK (i.e., BANKING) (Casanueva et al., 2020), CLINC (Larson et al., 2019), S.O. (i.e., StackOverflow) (Beau and Crabbé, 2024), HWU (Liu et al., 2021), TREC (Li and Roth, 2002), ECDT (Wang et al., 2022), MCID (Arora et al., 2020a), DBPedia (Zhang et al., 2015b), and X-Topic (Antypas et al., 2024). Detailed descriptions of these datasets are provided in the Appendix B.1.

We sample 10,000 instances from each large-scale corpus, including 20 Newsgroups (20NG), THUCNews, Yahoo, and DBPedia, to ensure both fair comparison and computational efficiency. To maintain fairness and consistency, all datasets are de-duplicated and standardized into a unified format. Table 2 summarizes the statistics of the datasets provided in BOLT.

**Diverse Scenarios.** BOLT encompasses diverse text classification scenarios under open-world settings. Specifically, it includes three representative scenarios: (1) *Topic Classification*, which focuses on identifying thematic categories from longer, content-rich documents (e.g., news articles); (2) *Intent Detection*, which emphasizes fine-grained recognition of user intentions from short, query-like utterances; and (3) Ontology Classification, which involves assigning entities or documents to high-cardinality, knowledge-based categories (e.g., DBPedia), thereby evaluating the model's scalability and ability for semantic abstraction.

**Diverse Text Domains.** The benchmark encompasses multiple text domains, such as news articles, question–answer datasets, and knowledge

bases, thereby ensuring coverage of both formal and conversational contexts. It consequently exhibits substantial diversity across both linguistic and structural dimensions, covering datasets with varying text lengths (ranging from short commands to long articles, with an average of 8–1856 words) and label granularities (from 10 to 219 classes), across multiple languages such as English, Chinese, Japanese and Spanish. This diversity enables comprehensive evaluation and more faithfully reflects the complexity of real-world open-world settings.

### 3.3 BOLT Framework

The BOLT framework provides a Python-based, practical, and extensible evaluation toolkit for OTC. By standardizing dataset formats, experimental configurations, and evaluation metrics, it enables reproducible comparisons of a wide range of existing methods and facilitates the integration of new datasets and algorithms, thereby advancing research toward scalable open-world applications.

**Open-World Formulation** To emulate open-world conditions, the BOLT framework introduces a unified formulation strategy grounded in diverse real-world text classification datasets. Specifically, each dataset is restructured into controlled experimental settings by adjusting two key factors: the Known Class Ratio (KCR), which partitions existing categories into known and unknown subsets to simulate varying degrees of openness, and the Labeled Ratio (LAR), which specifies the proportion of labeled samples within the known classes to reflect different levels of supervision. By specifying appropriate KCR and LAR values, users can flexibly generate customized experimental configurations, ensuring reproducibility, comparability, and scalability across studies. This design enables consistent evaluation under diverse openness and supervision conditions. Furthermore, BOLT offers a unified dataset configuration mechanism, where users only need to provide training, validation, and test sets in a standardized CSV format. By specifying the desired KCR and LAR, BOLT automatically generates the corresponding label partitions, enabling flexible experimental settings.

**Evaluation Metrics** To ensure consistent and comparable evaluation, BOLT adopts a unified evaluation protocol across both OSTC and GCD tasks, while employing task-specific metric sets. For OSTC (Shu et al., 2017; Zhang et al., 2021a), commonly used metrics include ACC, F1, K-F1, and N-F1, whereas GCD (An et al., 2023, 2025) is evaluated with ACC, K-ACC, N-ACC, H-Score, ARI, and NMI. These metrics jointly measure model performance in classification, OOD detection, and novel category discovery. Detailed descriptions of the evaluation metrics are provided in Appendix B.4. BOLT standardizes the evaluation pipeline across datasets, ensuring fair and consistent benchmarking of models under open-set conditions.

**Method Integration** BOLT provides a unified and extensible implementation framework that supports the integration, reproduction, and evaluation of diverse algorithmic paradigms across two representative OTC tasks, OSTC and GCD. The framework prioritizes compatibility and scalability over isolated algorithmic comparisons, facilitating the seamless incorporation of both established and newly developed LLM-based methods.

For OSTC, the framework incorporates four representative families of approaches: (i) Traditional non-PLM methods, including DOC (Shu et al., 2017) and DeepUnk (Lin and Xu, 2019), which rely on shallow neural architectures or activation-based confidence estimation; (ii) PLM-based methods, including ADB (Zhang et al., 2021a), SCL (Zeng et al., 2021a), AB (Lorenc et al., 2022), KnnCon (Zhou et al., 2022), and DyEn (Zhou et al., 2023), which enhance discriminative representation learning through PLMs and adaptive decision boundaries; (iii) our newly implemented LLM-based training framework, which fine-tunes an LLM with an MLP head as a discriminative backbone and incorporates post-hoc methods for OOD detection, including LLM-MSP (Hendrycks and Gimpel, 2016), LLM-OpenMax (Bendale and Boult, 2016), LLM-TempScale (Guo et al., 2017), LLM-Energy (Liu et al., 2020), LLM-LogitNorm (Wei et al., 2022), LLM-Entropy (Chan et al., 2021), LLM-KLMatching (Basart et al., 2022), and LLM-MaxLogit (Basart et al., 2022), supporting open-world classification.

For GCD, BOLT integrates two major families of approaches: (i) PLM-based clustering methods, including DeepAligned (Zhang et al., 2021b), GeoID (Tang et al., 2024), SDC (An et al., 2025), and DPN (An et al., 2023), TAN (An et al., 2024b), which learn discriminative embeddings from pre-trained encoders and discover novel categories through clustering; and (ii) LLM-enhanced methods, including LOOP (An et al., 2024a) and

| Metric | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| K-F1 | DOC | 60.06 | 42.71 | 30.42 | 35.57 | 45.66 | 68.85 | 58.95 | 47.89 | 0.09 | 42.98 | 22.58 | 6.72 | 38.54 |
| | DeepUNK | 73.65 | 27.15 | 36.90 | 25.83 | 53.96 | 52.42 | 64.70 | 63.18 | 1.59 | 54.40 | 24.78 | 6.95 | 40.46 |
| | ADB | 78.67 | **79.94** | **55.77** | 48.30 | 68.58 | **81.72** | 82.60 | 71.49 | 66.93 | 60.69 | **73.70** | 18.88 | 65.61 |
| | SCL | 79.30 | 23.57 | 21.41 | 29.88 | 35.14 | 51.58 | 57.55 | 41.22 | 1.81 | 46.47 | 30.18 | 8.81 | 35.58 |
| | AB | 57.59 | 27.23 | 46.17 | 25.49 | 62.57 | 63.57 | 72.93 | 67.19 | 40.31 | 54.72 | 42.39 | 12.29 | 47.70 |
| | KNNCon | 92.45 | 73.72 | 55.27 | 57.19 | 76.05 | 75.78 | **85.63** | **79.91** | 77.08 | 73.90 | 65.60 | 19.49 | 69.34 |
| | DyEn | **94.73** | 73.54 | 53.21 | **57.72** | **76.07** | 71.22 | 85.41 | 76.63 | **80.06** | **75.75** | 73.12 | **19.72** | **69.77** |
| | LLM-OpenMax | 75.32 | 64.13 | 47.91 | 30.38 | 46.23 | 74.09 | 60.52 | 50.78 | 33.10 | 42.69 | 25.35 | 10.90 | 46.79 |
| | LLM-Entropy | 72.95 | 64.40 | 53.61 | 47.42 | 62.89 | 67.92 | 72.53 | 66.42 | 53.43 | 50.89 | 50.73 | 13.26 | 56.37 |
| | LLM-MaxLogit | 84.18 | 69.85 | 43.28 | 37.46 | 65.89 | 80.18 | 81.16 | 63.39 | 42.84 | 46.39 | 50.58 | 8.02 | 56.10 |
| N-F1 | DOC | 77.93 | 66.77 | 58.53 | 59.45 | 69.29 | 72.25 | 76.41 | 70.33 | 61.79 | **68.20** | 64.08 | **54.72** | 66.64 |
| | DeepUNK | 85.29 | 45.11 | 59.97 | 51.62 | 58.30 | 72.69 | 62.71 | 56.68 | 43.13 | 53.57 | 43.96 | 32.57 | 55.47 |
| | ADB | 67.16 | 78.06 | 39.91 | 51.49 | 67.75 | 83.03 | 80.48 | 70.20 | 68.43 | 55.25 | 75.39 | 31.90 | 64.09 |
| | SCL | 90.64 | 52.88 | 45.63 | 58.04 | 53.21 | 62.59 | 62.69 | 51.78 | 3.93 | 58.35 | 49.87 | 37.23 | 52.24 |
| | AB | 63.74 | 45.89 | 55.25 | 10.94 | 67.73 | 69.30 | 74.36 | 75.41 | 59.01 | 67.10 | 62.23 | 50.31 | 58.44 |
| | KNNCon | 87.68 | 54.26 | 28.44 | 53.40 | 59.21 | 67.21 | 78.32 | 59.57 | 65.25 | 56.35 | 61.45 | 28.15 | 58.27 |
| | DyEn | **93.71** | 61.59 | 21.58 | 51.19 | 64.33 | 53.72 | 80.09 | 57.42 | 66.45 | 67.06 | 64.25 | 21.47 | 58.57 |
| | LLM-OpenMax | 77.26 | 65.26 | 21.85 | 62.52 | 71.79 | 70.79 | 77.68 | 72.63 | 67.87 | 56.69 | 68.74 | 52.20 | 63.77 |
| | LLM-Entropy | 48.67 | 38.05 | 46.58 | 43.94 | 37.91 | 39.47 | 45.90 | 45.30 | 56.13 | 50.09 | 52.64 | 31.69 | 44.70 |
| | LLM-MaxLogit | 88.85 | **78.37** | **64.28** | **68.92** | **76.21** | **83.44** | **84.67** | **76.34** | **72.95** | 66.86 | **77.47** | 50.27 | **74.05** |

Table 3: Average performance on BOLT benchmark for OSTC under KCR $\in \{0.25, 0.50, 0.75\}$ and LAR = 1.0. Metrics include K-F1 and N-F1. The best results are highlighted in **bold**, while the second-best results are underlined.

ALUP (Liang et al., 2024), which leverage LLM-generated semantic priors or pseudo-labels to improve clustering alignment.

Overall, BOLT enables consistent evaluation and flexible extension from traditional classifiers to next-generation LLM-based open-world learners.

## 4 Experimental Results and Analysis

### 4.1 Experimental Setup

To simulate open-world conditions, we adopt a unified experimental protocol. Specifically, we vary the KCR at $\{25\%, 50\%, 75\%\}$ to control the proportion of known classes, and adjust the LAR at $\{10\%, 50\%, 100\%\}$. To ensure stable and comparable evaluation when varying KCR, we uniformly partition the entire label space into five predefined subsets with minimal overlap. The final results are reported as the average performance over these subsets. More dataset details are provided in Appendix B.2.

### 4.2 Overall Performance on the OSTC Task

Table 3 presents the overall performance of various methods on the OSTC task, which is the setting adopted by most previous studies. The results show that existing PLM-based methods achieve better K-F1 than N-F1, failing to generalize to unseen classes. LLM-based approaches alleviate this issue and yield more balanced performance, though their generalization to novel classes remains limited.

**Known-class performance (K-F1).** Across all KCRs, PLM-based methods such as ADB and KNNCon exhibit remarkable robustness in known-class recognition. ADB achieves the top K-F1 on structured topic such as DBPedia (73.70). KNNCon also performs consistently, outperforming all others on intent detection datasets (CLINC, HWU). LLM-based variants do not consistently surpass strong PLM-based baselines on OSTC, suggesting that the potential of directly leveraging LLMs for OSTC to be further explored.

**Novel-class performance (N-F1).** For novel-class detection, a different pattern emerges. LLM-based scoring methods (LLM-Energy, LLM-MaxLogit) dominate across nearly all datasets and KCRs, achieving the highest N-F1. These improvements demonstrate that our proposed LLM-based methods significantly enhance open-set robustness and uncertainty estimation, showcasing promising potential for future open-world understanding tasks. Besides, the non-PLM methods **DOC** and **DeepUnk** exhibit poor K-F1 performance on the ECDT dataset due to limited supervision and the absence of prior semantic knowledge for Chinese text. As a result, their predicted probabilities are nearly uniform, causing most samples to be classified as OOD. Consequently, the K-F1 is low, accompanied by an artificially inflated N-F1.

### 4.3 Overall Performance on the GCD Task

Table 4 presents the overall performance of various methods on the GCD task. PLM-based clustering methods exhibit strong known-class discrimination on long-text datasets, whereas LLM-enhanced methods show superior generalization and novel categories, particularly in short-text or low-resource settings. However, they still tend to overfit the known classes, leading to suboptimal performance on novel categories.

| Metric | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| K-ACC | DAL | 92.51 | 83.81 | 57.86 | 49.84 | 70.35 | 75.72 | 87.95 | 72.77 | 63.18 | 74.62 | 80.76 | 21.16 | 69.21 |
| | GeoID | 98.34 | 89.72 | 64.02 | 55.57 | 83.01 | 84.18 | 92.53 | 84.17 | 71.06 | 75.57 | 86.79 | 26.56 | 75.96 |
| | SDC | 80.79 | 82.78 | 57.28 | 65.58 | 62.37 | 80.24 | 72.49 | 72.00 | 73.37 | 70.04 | 72.77 | **44.66** | 69.53 |
| | DPN | 95.86 | 91.72 | 63.32 | 80.80 | 82.16 | 88.24 | 89.46 | 80.71 | **86.44** | 81.83 | 87.31 | 35.92 | 80.32 |
| | TAN | 98.39 | 93.05 | 66.23 | **82.54** | 84.31 | 86.07 | 92.98 | 81.55 | 85.04 | 83.21 | **88.83** | 34.47 | **81.39** |
| | LOOP | 96.97 | **94.47** | 68.78 | 63.31 | **87.41** | 88.97 | **95.87** | 86.35 | 77.12 | 77.70 | 85.41 | 27.06 | 79.12 |
| | ALUP | 96.95 | 92.42 | **70.64** | 69.35 | 86.44 | 88.59 | 94.53 | 85.38 | 79.97 | **84.64** | 86.61 | 29.86 | 80.45 |
| N-ACC | DAL | 81.41 | 42.32 | 25.79 | 38.67 | 51.03 | 56.85 | 71.74 | 55.22 | 51.88 | 51.40 | **76.22** | 14.82 | 51.44 |
| | GeoID | 92.73 | 66.98 | 45.69 | 43.23 | 57.14 | 75.55 | 74.92 | 63.84 | 53.67 | 33.37 | 67.86 | 9.41 | 57.03 |
| | SDC | 84.69 | **78.71** | **56.00** | **56.54** | 58.97 | **79.94** | 68.53 | **72.57** | **72.34** | 58.49 | 69.26 | **19.57** | **64.63** |
| | DPN | 83.55 | 39.33 | 24.00 | 28.60 | 28.73 | 14.71 | 49.49 | 29.55 | 47.01 | 24.53 | 68.47 | 17.78 | 37.98 |
| | TAN | 93.12 | 57.83 | 31.09 | 39.92 | 36.38 | 31.31 | 53.39 | 34.57 | 53.03 | 30.31 | 72.20 | 19.05 | 46.02 |
| | LOOP | 87.18 | 66.62 | 51.52 | 44.84 | 65.41 | 78.44 | **83.69** | 72.01 | 62.98 | 58.09 | 74.24 | 17.61 | 63.55 |
| | ALUP | **95.70** | 68.90 | 49.88 | 43.53 | 64.87 | 75.41 | 81.61 | 68.65 | 65.64 | 53.76 | 75.21 | 15.61 | 63.23 |

Table 4: Average performance on BOLT benchmark for GCD across different KCRs and LARs. Metrics include K-ACC and N-ACC. The best results are highlighted in **bold**, while the second-best results are underlined.

| Method | KCR=0.25 | | | | | | KCR=0.50 | | | | | | KCR=0.75 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | LAR=0.1 | | LAR=0.5 | | LAR=1.0 | | LAR=0.1 | | LAR=0.5 | | LAR=1.0 | | LAR=0.1 | | LAR=0.5 | | LAR=1.0 | |
| OSTC | K-F1 | N-F1 | K-F1 | N-F1 | K-F1 | N-F1 | K-F1 | N-F1 | K-F1 | N-F1 | K-F1 | N-F1 | K-F1 | N-F1 | K-F1 | N-F1 | K-F1 | N-F1 |
| DOC | 14.09 | **81.13** | 50.29 | 84.46 | 54.53 | 82.72 | 15.71 | 63.82 | 52.18 | 74.65 | 60.80 | 75.60 | 13.49 | 40.27 | 51.44 | 52.44 | 62.20 | 56.10 |
| DeepUnk | 28.33 | 42.02 | 46.95 | 82.71 | 53.99 | 82.48 | 31.19 | 41.05 | 49.28 | 69.99 | 54.49 | 72.56 | 33.28 | 28.91 | 47.49 | 51.63 | 51.60 | 54.13 |
| ADB | 45.11 | 58.99 | 64.15 | 73.99 | 67.20 | 77.92 | 59.26 | 57.05 | **76.23** | 71.08 | **77.98** | 73.16 | 64.92 | 47.18 | 80.65 | 59.85 | **82.38** | 61.71 |
| SCL | 23.92 | 52.54 | 27.75 | 70.91 | 31.02 | 68.21 | 26.89 | 50.91 | 37.13 | 56.12 | 40.15 | 52.45 | 30.81 | 36.94 | 46.37 | 40.79 | 51.25 | 39.34 |
| AB | 42.73 | 73.33 | 47.83 | 71.57 | 48.31 | 73.60 | 44.43 | 59.63 | 52.18 | 57.94 | 53.44 | 59.63 | 45.44 | 41.19 | 54.50 | 42.07 | 56.16 | 42.00 |
| KNNCon | **61.74** | 61.68 | 59.62 | 58.32 | 60.40 | 61.39 | **74.89** | 62.77 | 74.56 | 62.32 | 74.97 | 62.40 | **82.55** | **56.35** | **82.12** | 53.87 | 82.19 | 56.01 |
| LLM-OpenMax | 18.61 | 65.71 | 50.68 | 69.06 | 58.53 | 71.67 | 25.41 | 65.50 | 62.08 | 70.54 | 70.89 | 70.08 | 26.03 | 43.81 | 64.58 | 52.76 | 74.78 | 56.98 |
| LLM-Energy | 24.74 | 80.43 | 61.80 | **89.41** | 69.26 | **90.42** | 32.75 | 70.41 | 66.58 | 81.50 | 71.20 | 83.36 | 35.78 | 48.54 | 62.00 | 59.70 | 68.26 | 63.41 |
| LLM-MaxLogit | 26.32 | 79.71 | **65.43** | 89.11 | **70.77** | 90.08 | 37.55 | **71.58** | 70.69 | **82.31** | 74.42 | **83.87** | 41.52 | 50.41 | 67.07 | **62.05** | 72.50 | **65.50** |
| GCD | K-ACC | N-ACC | K-ACC | N-ACC | K-ACC | N-ACC | K-ACC | N-ACC | K-ACC | N-ACC | K-ACC | N-ACC | K-ACC | N-ACC | K-ACC | N-ACC | K-ACC | N-ACC |
| DAL | 64.18 | 46.62 | 68.01 | 48.10 | 67.19 | 48.16 | 69.24 | 53.13 | 74.51 | 56.10 | 73.64 | 55.00 | 71.57 | 53.98 | 76.09 | 58.70 | 76.82 | 55.76 |
| GeoID | 73.53 | 62.51 | 82.81 | 64.40 | 85.84 | 66.22 | 75.30 | **64.60** | 84.70 | 67.12 | 87.10 | 67.15 | 75.24 | 64.88 | 85.13 | 66.67 | 88.44 | 68.46 |
| SDC | 58.21 | 49.42 | 61.37 | 54.22 | 63.88 | 56.94 | 68.89 | 64.21 | 70.84 | **68.66** | 73.75 | **70.24** | 75.69 | **71.94** | 80.67 | **76.43** | 83.19 | **77.18** |
| DPN | 74.91 | 34.12 | 81.70 | 33.16 | 82.69 | 32.68 | 78.27 | 36.34 | 83.49 | 37.36 | 86.20 | 37.97 | 79.43 | 41.12 | 86.38 | 40.36 | 88.42 | 39.21 |
| TAN | **79.58** | 44.73 | **84.80** | 44.51 | **87.56** | 43.23 | **80.15** | 49.04 | **87.29** | 47.02 | 88.32 | 47.62 | **81.50** | 52.57 | **88.21** | 51.14 | **89.99** | 49.27 |
| LOOP | 74.65 | **63.24** | 83.43 | 64.86 | 86.02 | 64.67 | 76.49 | 63.23 | 84.55 | 64.44 | 87.26 | 65.85 | 78.00 | 64.31 | 84.74 | 65.11 | 87.15 | 64.24 |
| ALUP | 74.13 | 59.59 | 83.46 | **66.41** | 85.67 | **67.79** | 73.30 | 58.59 | 86.68 | 67.49 | **88.73** | 67.74 | 76.03 | 57.47 | 87.09 | 68.00 | 88.92 | 66.29 |

Table 5: Average performance on BOLT benchmark for OSTC and GCD under different KCRs and LARs. Due to page limitations, the complete experimental results across all datasets and metrics are provided in Appendix E.

**Known-class performance (K-ACC).** PLM-based methods (TAN, DPN) achieve the highest K-ACC on long and information-rich corpora such as 20NG, DBPedia, and THUCNews. TAN consistently attains top performance, confirming its stability in structured topic classification tasks. In contrast, LLM-enhanced methods (LOOP, ALUP) yield superior results on short-text intent detection benchmarks (CLINC, BANKING, StackOverflow), where semantic priors help compensate for limited lexical cues.

**Novel-class performance (N-ACC).** For novel-class detection, a complementary trend emerges. LLM-enhanced models (ALUP, LOOP) outperform others indicating their stronger ability to infer unseen categories under weak supervision. Among PLM-based methods, SDC demonstrates relatively strong performance due to its distinctive calibration mechanism, suggesting that effective logit calibration is a promising strategy for GCD, particularly evident when pseudo-label noise is non-negligible.

## 4.4 Impact of Known Class Ratio (KCR)

Table 5 summarizes the average performance under different KCRs across OSTC and GCD benchmarks. On OSTC, increasing KCR generally improves K-F1, but poses a clear challenge to open-set recognition: most methods exhibit a noticeable drop in N-F1 as KCR increases, indicating a stronger closed-set bias when more classes are treated as known. This challenge stems from the K+1 classification formulation of OSTC, where unknown samples must be explicitly rejected from an expanding known-class set without additional supervision. On GCD, however, the negative impact of higher KCR on unknown-class performance is far less pronounced, as the availability of extra unlabeled data enables K+N pseudo-label learning that jointly refines known and novel categories.

## 4.5 Impact of Labeled Ratio (LAR)

As shown in Table 5, increasing LAR consistently enhances performance on the known (K-F1/K-ACC) for PLM-based discriminative learners, as richer supervision sharpens decision boundaries and yields more separable known-class representations. In contrast, *open-set* robustness (N-F1/N-ACC) is less sensitive to labeled data volume, where LLM-based or LLM-enhanced methods (e.g., LLM-Energy, LLM-MaxLogit, ALUP)

| Metric | LLM | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| K-F1 | Qwen2.5-3B | 79.68 | 66.35 | 41.37 | 42.25 | 57.42 | 70.15 | 76.15 | 54.31 | 39.28 | 36.50 | 54.72 | 16.49 | 52.89 |
| | Qwen2.5-7B | 71.37 | 56.77 | 30.42 | 45.51 | 58.04 | 74.25 | 73.68 | 52.30 | 38.79 | 39.89 | 42.11 | 11.67 | 49.57 |
| | Qwen2.5-32B | 71.90 | 63.37 | 39.32 | 42.34 | 57.14 | 76.05 | 78.26 | 61.57 | 39.50 | 45.96 | 35.45 | 11.03 | 51.82 |
| N-F1 | Qwen2.5-3B | 88.53 | 76.46 | 62.45 | 48.87 | 73.54 | 81.01 | 82.61 | 76.47 | 56.80 | 64.46 | 64.84 | 11.11 | 65.60 |
| | Qwen2.5-7B | 82.07 | 74.76 | 58.79 | 49.88 | 74.87 | 82.30 | 79.77 | 73.76 | 54.01 | 64.64 | 59.96 | 9.80 | 63.72 |
| | Qwen2.5-32B | 83.80 | 78.40 | 63.84 | 50.14 | 73.69 | 84.10 | 82.82 | 78.12 | 55.87 | 69.18 | 59.37 | 11.07 | 65.86 |

Table 6: Performance of LLM-MaxLogit under the OSTC scenario with different Qwen2.5 model scales.

| Setting | Method | Random | | Similarity | |
|---|---|---|---|---|---|
| KCR, LAR | GCD | K-ACC | N-ACC | K-ACC | N-ACC |
| 0.25, 0.1 | DAL | 64.18 | 46.62 | 66.67 | 43.24 |
| | GeoID | 73.53 | 62.51 | 73.80 | 64.08 |
| | SDC | 58.21 | 49.42 | **79.95** | 47.64 |
| | DPN | 74.91 | 34.12 | 76.75 | 37.89 |
| | TAN | **79.58** | 44.73 | 73.63 | 10.20 |
| | LOOP | 74.65 | 63.24 | 75.70 | 61.66 |
| | ALUP | 74.13 | 59.59 | 79.94 | **68.42** |
| KCR, LAR | OSTC | K-F1 | N-F1 | K-F1 | N-F1 |
| 0.25, 1.0 | DOC | 54.53 | 82.72 | 56.54 | 83.73 |
| | DeepUnk | 53.99 | 82.48 | 53.48 | 83.42 |
| | ADB | 67.20 | 77.92 | 69.59 | 79.98 |
| | SCL | 31.02 | 68.21 | 36.94 | 70.65 |
| | AB | 48.31 | 73.60 | 50.52 | 76.37 |
| | KNNCon | 60.40 | 61.39 | 63.17 | 65.86 |
| | DyEn | 63.88 | 65.49 | 68.98 | 71.90 |
| | LLM-OpenMax | 58.53 | 71.67 | 61.12 | 75.38 |
| | LLM-Energy | 69.26 | **90.42** | 70.38 | **91.39** |
| | LLM-MaxLogit | **70.77** | 90.08 | **72.73** | 91.21 |

Table 7: Average performance on BOLT benchmark for OTC under different label sampling strategies.

maintain superior performance even in low-label regimes. These results suggest that while PLM-based training benefits from abundant labels, LLM-driven calibration and pseudo-supervision offer a complementary pathway toward generalization.

### 4.6 Impact of LLM Backbone

We evaluate LLM-Maxlogit with Qwen2.5 backbones of different scales (3B/7B/32B) under OSTC (Table 6). Overall, scaling does not yield monotonic improvements: 32B gives the best average OOD-aware performance (Avg. N-F1: 65.86; Avg. K-F1: 51.82), while 3B is highly competitive and even better on average K-F1 (52.89). The 7B model underperforms both in most cases, indicating that the mid-scale backbone does not offer a consistently more balanced trade-off between K-F1 and N-F1 in our OSTC setting. Dataset-wise, longer documents with clearer topical evidence (e.g., 20NG, THUCNews; Table 2) tend to benefit more from larger models on OOD detection, whereas extremely short and distribution-shifted data (X-Topic) remains challenging for all scales.

From a task perspective, larger models help more for OOD detection than for closed-set classification on several intent benchmarks (e.g., StackOverflow, MCID), consistent with improved semantic coverage for separating in- vs. out-of-domain. However, for fine-grained label spaces with minimal label descriptions (e.g., DBPedia with 219 classes and Avg. Len./Lab. $\approx$1), scaling can even degrade both K-F1 and N-F1, implying that the bottleneck is label-space ambiguity rather than representation capacity. These results highlight that simply increasing model size is insufficient in OSTC.

### 4.7 Impact of Imbalanced Sampling

To further analyze the impact of known label distribution, we compare two label sampling strategies: (1) **Random**, where known and unknown classes are randomly selected without considering their semantic relatedness; and (2) **Similarity**, where labels are grouped based on semantic proximity, so that known classes share closer meanings while unknown classes differ more significantly. This Similarity sampling better reflects real-world open-world scenarios, where initial known categories are often semantically coherent (e.g., within a specific domain). As shown in Table 7, semantically concentrated splits consistently outperform random ones, yielding clearer intra-class discrimination and more accurate OOD identification. These results suggest that OTC methods tend to deliver better performance in practical applications.

## 5 Conclusions

In this work, we presented **BOLT**, the first comprehensive benchmark and evaluation toolkit designed to support open-world learning for text classification, unifying the evaluation of two key tasks in OTC, namely OSTC and GCD. Compared with prior studies that focus narrowly on single tasks or limited domains, BOLT provides: (1) broader coverage, encompassing 12 diverse datasets across multiple text classification scenarios and domains, (2) a standardized evaluation protocol that explicitly accounts for varying known class ratios and

labeled ratios, and (3) systematic comparisons across traditional methods, PLM-based approaches, and LLM-enhanced models. The experimental results not only reveal the limitations of existing approaches but also demonstrate the great potential of leveraging LLMs to advance OTC.

## Limitations

While BOLT provides a comprehensive benchmark and evaluation toolkit for OTC, several promising directions for future work remain. One promising direction is to investigate the impact of different LLM backbones on performance, which could yield deeper insights into the trade-offs between efficiency and generalization. Moreover, incorporating additional multilingual datasets would further enhance the coverage and cross-lingual applicability of BOLT, making it a more versatile benchmark for global open-world learning research.

## Ethical Considerations

We present BOLT, a comprehensive benchmark and evaluation toolkit for OTC. This research was conducted in full compliance with the laws, rules, and regulations of our community, institution, workplace, and country. During the development of BOLT, we carefully considered multiple ethical aspects. All datasets used are publicly available and widely adopted in prior research; their licenses and terms of use were thoroughly reviewed to ensure compliance. No personally identifiable or sensitive information was collected, modified, or redistributed, and all preprocessing steps were performed with care to preserve the integrity and semantics of the original text. BOLT is intended solely for academic and research purposes to advance the understanding of OTC. To promote transparency and reproducibility, all benchmark configurations and experimental settings will be made publicly available, in line with the principles of open and responsible NLP research.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) (Grant No. 62506352), and in part by the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDB1350102).

## References

Wenbin An, Haonan Lin, Jiahao Nie, Feng Tian, Wenkai Shi, Yaqiang Wu, Qianying Wang, and Ping Chen. 2025. Unleashing the potential of model bias for generalized category discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Wenbin An, Wenkai Shi, Feng Tian, Haonan Lin, QianYing Wang, Yaqiang Wu, Mingxiang Cai, Luyan Wang, Yan Chen, Haiping Zhu, and Ping Chen. 2024a. Generalized category discovery with large language models in the loop. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Wenbin An, Feng Tian, Wenkai Shi, Yan Chen, Yaqiang Wu, Qianying Wang, and Ping Chen. 2024b. Transfer and alignment network for generalized category discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Wenbin An, Feng Tian, Qinghua Zheng, Wei Ding, QianYing Wang, and Ping Chen. 2023. Generalized category discovery with decoupled prototypical network. In *Proceedings of the AAAI conference on artificial intelligence*.

Dimosthenis Antypas, Asahi Ushio, Francesco Barbieri, and Jose Camacho-Collados. 2024. Multilingual topic classification in x: Dataset and analysis. *arXiv preprint arXiv:2410.03075*.

Abhinav Arora, Akshat Shrivastava, Mrinal Mohit, Lorena Sainz-Maza Lecanda, and Ahmed Aly. 2020a. Cross-lingual transfer learning for intent detection of covid-19 utterances.

Abhinav Arora, Akshat Shrivastava, Mrinal Mohit, Lorena Sainz-Maza Lecanda, and Ahmed Aly. 2020b. Cross-lingual transfer learning for intent detection of covid-19 utterances. In *The EMNLP 2020 Workshop NLP-COVID Submission*.

Steven Basart, Mazeika Mantas, Mostajabi Mohammadreza, Steinhardt Jacob, and Song Dawn. 2022. Scaling out-of-distribution detection for real-world settings. In *International Conference on Machine Learning*.

Nathanaël Beau and Benoit Crabbé. 2024. CodeInsight: A curated dataset of practical coding solutions from Stack Overflow. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5935–5947, Bangkok, Thailand. Association for Computational Linguistics.

Abhijit Bendale and Terrance E. Boult. 2016. Towards open set deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.

Robin Chan, Matthias Rottmann, and Hanno Gottschalk. 2021. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In *Proceedings of the ieee/cvf international conference on computer vision*.

Junyang Chen, Zhiguo Gong, and Weiwen Liu. 2020. A dirichlet process biterm-based mixture model for short text stream clustering. *Applied Intelligence*.

Xi Chen, Chuan Qin, Chuyu Fang, Chao Wang, Chen Zhu, Fuzhen Zhuang, Hengshu Zhu, and Hui Xiong. 2024. Job-sdf: A multi-granularity dataset for job skill demand forecasting and benchmarking. *Advances in Neural Information Processing Systems*, 37:129329–129356.

Xi Chen, Chuan Qin, Ziqi Wang, Shasha Hu, Chao Wang, Hengshu Zhu, and Hui Xiong. 2026. Beyond the known: An unknown-aware large language model for open-set text classification. In *The Fourteenth International Conference on Learning Representations*.

Geli Fei and Bing Liu. 2016. Breaking the closed world assumption in text classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*. PMLR.

Dan Hendrycks and Kevin Gimpel. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*.

Xiaohan Huang, Meng Xiao, Chuan Qin, Qingqing Long, Jinmiao Chen, Yuanchun Zhou, and Hengshu Zhu. 2026. Scihorizon-gene: Benchmarking llm for life sciences inference from gene knowledge to functional understanding. *arXiv preprint arXiv:2601.12805*.

Feihu Jiang, Chuan Qin, Kaichun Yao, Chuyu Fang, Fuzhen Zhuang, Hengshu Zhu, and Hui Xiong. 2024. Enhancing question answering for enterprise knowledge bases using large language models. In *International Conference on Database Systems for Advanced Applications*, pages 273–290. Springer.

Konstantin Kirchheim, Marco Filax, and Frank Ortmeier. 2022. Pytorch-ood: A library for out-of-distribution detection based on pytorch. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Harold W Kuhn. 2004. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316, Hong Kong, China. Association for Computational Linguistics.

Fayin Li and Harry Wechsler. 2005. Open set face recognition using transduction. *IEEE transactions on pattern analysis and machine intelligence*.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Jinggui Liang, Lizi Liao, Hao Fei, Bobo Li, and Jing Jiang. 2024. Actively learn from LLMs with uncertainty propagation for generalized category discovery. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Ting-En Lin and Hua Xu. 2019. Deep unknown intent detection with margin loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Bo Liu, Liming Zhan, Zexin Lu, Yujie Feng, Lei Xue, and Xiao-Ming Wu. 2023. How good are llms at out-of-distribution detection? *arXiv preprint arXiv:2308.10261*.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. *Advances in neural information processing systems*.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2021. Benchmarking natural language understanding services for building conversational agents. In *Increasing naturalness and flexibility in spoken dialogue interaction: 10th international workshop on spoken dialogue systems*, pages 165–183. Springer.

Petr Lorenc, Tommaso Gargiani, Jan Pichl, Jakub Konrád, Petr Marek, Ondřej Kobza, and Jan Šedivỳ. 2022. Metric learning and adaptive boundary for out-of-domain detection. In *International Conference on Applications of Natural Language to Information Systems*. Springer.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning–based text classification: a comprehensive review. *ACM computing surveys (CSUR)*.

Sridhama Prakhya, Vinodini Venkataram, and Jugal Kalita. 2017. Open set text classification using cnns. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*.

Chuan Qin, Xin Chen, Chengrui Wang, Pengmin Wu, Xi Chen, Yihang Cheng, Jingyi Zhao, Meng Xiao, Xiangchao Dong, Qingqing Long, and 1 others. 2025a. Scihorizon: Benchmarking ai-for-science readiness from scientific data to large language models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5754–5765.

Chuan Qin, Chuyu Fang, Kaichun Yao, Xi Chen, Fuzhen Zhuang, and Hengshu Zhu. 2025b. Cotr: Efficient job task recognition for occupational information systems with class-incremental learning. *ACM Transactions on Management Information Systems*, 16(2):1–30.

Chuan Qin, Le Zhang, Yihang Cheng, Rui Zha, Dazhong Shen, Qi Zhang, Xi Chen, Ying Sun, Chen Zhu, Hengshu Zhu, and 1 others. 2025c. A comprehensive survey of artificial intelligence techniques for talent analytics. *Proceedings of the IEEE*.

Wenkai Shi, Wenbin An, Feng Tian, Yan Chen, Yaqiang Wu, Qianying Wang, and Ping Chen. 2024. A unified knowledge transfer network for generalized category discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Lei Shu, Hu Xu, and Bing Liu. 2017. Doc: Deep open classification of text documents. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Zhiheng Song, Jingshuai Zhang, Chuan Qin, Chao Wang, Chao Chen, Longfei Xu, Kaikui Liu, Xiangxiang Chu, and Hengshu Zhu. 2026. Mobilitybench: A benchmark for evaluating route-planning agents in real-world mobility scenarios. *arXiv preprint arXiv:2602.22638*.

Maosong Sun, Jingyang Li, Zhipeng Guo, Zhao Yu, Yabiing Zheng, Xiance Si, and Zhiyuan Liu. 2016. Thuctc: An efficient chinese text classifier.

Kai Tang, Junbo Zhao, Xiao Ding, Runze Wu, Lei Feng, Gang Chen, and Haobo Wang. 2024. Learning geometry-aware representations for new intent discovery. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Zhenyu Tong, Chuan Qin, Chuyu Fang, Kaichun Yao, Xi Chen, Jingshuai Zhang, Chen Zhu, and Hengshu Zhu. 2025. From missteps to mastery: Enhancing low-resource dense retrieval through adaptive query generation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 1373–1384.

Fengjun Wang, Moran Beladev, Ofri Kleinfeld, Elina Frayerman, Tal Shachar, Eran Fainman, Karen Lastmann Assaraf, Sarai Mizrachi, and Benjamin Wang. 2023. Text2topic: Multi-label text classification system for efficient topic detection in user generated content with zero-shot capabilities. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*.

Lihua Wang, Huiting Yang, Feng Li, Wenzhong Yang, and Zhenwan Zou. 2022. Intent detection model based on dual-channel feature fusion. In *2022 IEEE 6th Information Technology and Mechatronics Engineering Conference (ITOEC)*, volume 6, pages 1862–1867. IEEE.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*.

Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. 2022. Mitigating neural network overconfidence with logit normalization. In *International conference on machine learning*. PMLR.

Zhiyuan Zeng, Keqing He, Yuanmeng Yan, Zijun Liu, Yanan Wu, Hong Xu, Huixing Jiang, and Weiran Xu. 2021a. Modeling discriminative representations for out-of-domain detection with supervised contrastive learning. *arXiv preprint arXiv:2105.14289*.

Zhiyuan Zeng, Keqing He, Yuanmeng Yan, Zijun Liu, Yanan Wu, Hong Xu, Huixing Jiang, and Weiran Xu. 2021b. Modeling discriminative representations for out-of-domain detection with supervised contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*.

Feng Zhang, Wei Chen, Fei Ding, Meng Gao, Tengjiao Wang, Jiahui Yao, and Jiabin Zheng. 2024a. From discrimination to generation: Low-resource intent detection with language model instruction tuning. In *Findings of the Association for Computational Linguistics ACL 2024*.

Hanlei Zhang, Hua Xu, and Ting-En Lin. 2021a. Deep open intent classification with adaptive decision boundary. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Hanlei Zhang, Hua Xu, Ting-En Lin, and Rui Lyu. 2021b. Discovering new intents with deep aligned clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Hanlei Zhang, Hua Xu, Xin Wang, Fei Long, and Kai Gao. 2023. A clustering framework for unsupervised and semi-supervised new intent discovery. *IEEE Transactions on Knowledge and Data Engineering*.

Shun Zhang, Chaoran Yan, Jian Yang, Changyu Ren, Jiaqi Bai, Tongliang Li, and Zhoujun Li. 2024b. Ronid: new intent discovery with generated-reliable labels and cluster-friendly representations. In *International Conference on Database Systems for Advanced Applications*.

Wei-Nan Zhang, Zhigang Chen, Wanxiang Che, Guoping Hu, and Ting Liu. 2017. The first evaluation of chinese human-computer dialogue technology. *arXiv preprint arXiv:1709.10217*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015a. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015b. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Qi Zhou, Xi Chen, Chuyu Fang, Jianji Wang, Chuan Qin, and Fuzhen Zhuang. 2025. Enhancing dual-target cross-domain recommendation via similar user bridging. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 4487–4497.

Yunhua Zhou, Peiju Liu, and Xipeng Qiu. 2022. Knn-contrastive learning for out-of-domain intent classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Yunhua Zhou, Jianqiang Yang, Pengyu Wang, and Xipeng Qiu. 2023. Two birds one stone: Dynamic ensemble for ood intent classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

## A LLM Usage Statement

LLMs were used exclusively to assist in refining the writing of this paper, including grammar correction, phrasing improvement, and enhancement of readability and clarity. It should be noted that LLMs were not involved in the ideation, research methodological, experimental design, or result analysis. All research ideas, findings, and conclusions originate solely from the authors. The authors affirm that the use of LLMs complies with ethical guidelines and does not contribute to plagiarism or any form of scientific misconduct.

## B Additional Details on BOLT

### B.1 Dataset Details

This section presents detailed descriptions of the 12 datasets incorporated in BOLT. These datasets span diverse text classification scenarios and domains, providing comprehensive coverage for the evaluation of open-world text classification systems.

- **20NG**[1]: The 20NG, also known as the 20 Newsgroups dataset, contains approximately 20,000 newsgroup documents, partitioned nearly evenly across 20 categories. We sampled 10,000 documents from it to construct our dataset.
- **THUCNews**[2]: The THUCNews dataset is a large-scale Chinese news classification corpus, based on historical news articles collected via Sina News RSS feeds between 2005 and 2011. We sampled 10,000 documents from it to construct our version of the dataset.
- **Yahoo**[3]: The Yahoo dataset, also known as the Yahoo! Answers topic classification dataset, is constructed from user-submitted questions and their best answers under the Yahoo! Webscope program, and it is categorized into 10 major topic classes. For our study, we sampled 10,000 documents from this dataset to build our version.
- **BANK (BANKING)**[4]: The BANK, also known as the BANKING77, consists of 13,083 English customer service queries from the BANK domain, annotated with 77 fine-grained intent classes.
- **S.O. (StackOverflow)**[5]: The S.O. (i.e., StackOverflow) dataset consists of 20,000 short text titles collected from the Stack Overflow question forum, each annotated with one of 20 technical topic labels.
- **CLINC**[6]: The CLINC dataset, also known as CLINC150, is an intent classification corpus covering 150 in-domain intent classes. These intents cover a broad spectrum of everyday tasks and informational queries, such as playing music, checking the weather, and setting alarms.
- **HWU**[7]: The HWU dataset, also known as HWU-64, contains English user utterances across multiple domains (e.g. news and IoT) for intent detection, covering 64 intents. It is widely adopted

[1] 20NG, https://people.csail.mit.edu/jrennie/20Newsgroups/

[2] THUCNews, http://thuctc.thunlp.org/

[3] Yahoo, https://www.kaggle.com/datasets/bhavikardeshna/yahoo-email-classification/data

[4] BANKING, https://huggingface.co/datasets/PolyAI/banking77

[5] StackOverflow, https://github.com/jacoxu/StackOverflow

[6] CLINC, https://archive.ics.uci.edu/dataset/570/clinc150

[7] HWU, https://service.tib.eu/ldmservice/en/dataset/hwu64

in conversational AI research, as its utterances simulate realistic task-oriented conversational queries.

- **ECDT**[8]: The ECDT dataset, also known as SMP-ECDT, is a Chinese intent classification corpus introduced in (Zhang et al., 2017).

- **TREC**[9]: The TREC is a question classification corpus of open-domain, fact-based questions.

- **MCID**[10]: The MCID dataset contains 6,871 utterances across multiple languages (English, Spanish, French, German, and Spanglish) with 16 COVID-19 related intents (Arora et al., 2020b). In our work, we select only the English utterances to form our version of the dataset.

- **DBpedia**[11]: DBpedia is a project that extracts structured content from Wikipedia. This dataset is an ontology classification corpus constructed by selecting 219 non-overlapping classes. For our study, we sampled 10,000 documents from this dataset to build our version.

- **X-Topic**[12]: X-Topic is a topic classification dataset based on X (formerly Twitter), featuring 19 topic labels. The classification task is multi-label, with tweets available in four languages: English, Japanese, Spanish, and Greek.

Moreover, to justify dataset selection, we conducted a systematic survey of 50 representative OTC-related papers and collected dataset usage statistics. The aggregated distribution is shown in Table S1. The results indicate that all high-frequency datasets adopted in prior OTC studies are included in BOLT, confirming that our benchmark covers the core datasets used by the community.

| Dataset | # OTC Papers Using It | Percentage (%) |
|---|---|---|
| CLINC | 44 | 88 |
| BANKING | 41 | 82 |
| StackOverflow | 26 | 52 |
| HWU | 12 | 24 |
| MCID | 3 | 6 |
| 20NG | 1 | 2 |

Table S1: Dataset usage statistics from 50 OTC-related papers.

### B.2 Dataset Split

To ensure consistent and reproducible evaluation across all tasks, we adopt a unified five-fold dataset partitioning strategy. Unlike prior studies that rely on random class splits, which often introduce instability and overlap between folds, our benchmark employs a **uniform label partitioning** scheme that minimizes class overlap while maintaining semantic balance. Specifically, the entire label space of each dataset is evenly divided into five disjoint subsets. Each fold serves as the known-class set once, while the remaining labels are treated as unknown classes for testing. This design ensures comparability across experiments and reduces randomness caused by arbitrary label sampling.

Within each known-class subset, samples are further divided according to the **Labeled Data Ratio (LAR)** and **Known Class Ratio (KCR)**:

- KCR $(0.25, 0.5, 0.75)$ controls the proportion of known classes relative to the total label set, simulating different levels of open-world exposure.

- LAR $(0.1, 0.5, 1.0)$ determines the proportion of labeled samples among the known-class instances, reflecting different supervision levels.

For each configuration, we maintain a fixed **train/-validation/test** split within the known-class set to ensure balanced learning and reliable evaluation. The unlabeled pool includes both known and unknown instances, allowing open-set and generalized category discovery tasks to share the same structural foundation.

As summarized in Table S2, all datasets are processed under identical sampling rules, ensuring that model performance reflects genuine open-world generalization rather than artifacts of random data selection. This systematic partitioning provides a stable and comparable basis for evaluating BOLT and all baseline methods across OSTC and GCD settings.

### B.3 Method Details

This section provides detailed descriptions of the seven methods for GCD and the fifteen methods for OSTC implemented and evaluated in the BOLT benchmark. All methods are reproduced

[8]ECDT, https://github.com/ygwpz/SMP2019-ECDT-NLU

[9]TREC, https://huggingface.co/datasets/CogComp/trec

[10]MCID, https://github.com/fanolabs/IntentBert

[11]DBpedia, https://huggingface.co/datasets/DeveloperOats/DBPedia_Classes

[12]X-Topic, https://huggingface.co/datasets/cardiffnlp/tweet_topic_multilingual

| | KCR | 0.25 | | | 0.50 | | | 0.75 | | | ALL | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LAR | Dataset | #label | \|Train\| | \|Eval\| | #label | \|Train\| | \|Eval\| | #label | \|Train\| | \|Eval\| | #label | \|Train\| | \|Eval\| | \|Test\| |
| 0.1 | 20NG | 5 | 183 | 25 | 10 | 351 | 48 | 15 | 518 | 71 | 20 | 701 | 96 | 2,000 |
| | THUCNews | 4 | 200 | 28 | 7 | 350 | 49 | 10 | 500 | 70 | 14 | 700 | 98 | 2,000 |
| | Yahoo | 2 | 140 | 20 | 5 | 350 | 50 | 8 | 560 | 80 | 10 | 700 | 100 | 2,000 |
| | TREC | 12 | 324 | 35 | 24 | 435 | 51 | 35 | 468 | 62 | 47 | 487 | 71 | 490 |
| | BANK | 19 | 229 | 28 | 38 | 472 | 55 | 58 | 704 | 81 | 77 | 900 | 105 | 3,080 |
| | S.O. | 5 | 300 | 50 | 10 | 600 | 100 | 15 | 900 | 150 | 20 | 1,200 | 200 | 5,991 |
| | CLINC | 38 | 457 | 76 | 75 | 901 | 150 | 112 | 1,345 | 224 | 150 | 1,801 | 300 | 2,250 |
| | HWU | 16 | 187 | 26 | 32 | 361 | 51 | 48 | 557 | 79 | 64 | 770 | 109 | 1,032 |
| | ECDT | 8 | 97 | 12 | 16 | 137 | 20 | 23 | 188 | 28 | 31 | 209 | 33 | 770 |
| | MCID | 4 | 31 | 4 | 8 | 62 | 8 | 12 | 94 | 12 | 16 | 122 | 16 | 331 |
| | DBPedia | 55 | 252 | 55 | 110 | 480 | 110 | 164 | 647 | 164 | 219 | 736 | 218 | 2,000 |
| | X-Topic | 12 | 531 | 75 | 24 | 582 | 87 | 37 | 598 | 100 | 49 | 610 | 112 | 1,767 |
| 0.5 | 20NG | 5 | 911 | 130 | 10 | 1,750 | 249 | 15 | 2,581 | 366 | 20 | 3,498 | 496 | 2,000 |
| | THUCNews | 4 | 1,000 | 144 | 7 | 1,750 | 252 | 10 | 2,500 | 360 | 14 | 3,500 | 504 | 2,000 |
| | Yahoo | 2 | 700 | 100 | 5 | 1,750 | 250 | 8 | 2,800 | 400 | 10 | 3,500 | 500 | 2,000 |
| | TREC | 12 | 1,610 | 175 | 24 | 2,176 | 239 | 35 | 2,343 | 261 | 47 | 2,426 | 273 | 490 |
| | BANK | 19 | 1,150 | 130 | 38 | 2,368 | 265 | 58 | 3,522 | 391 | 77 | 4,498 | 499 | 3,080 |
| | S.O. | 5 | 1,500 | 250 | 10 | 3,000 | 500 | 15 | 4,500 | 750 | 20 | 6,000 | 1,000 | 5,991 |
| | CLINC | 38 | 2,280 | 304 | 75 | 4,500 | 600 | 112 | 6,722 | 896 | 150 | 9,002 | 1,200 | 2,250 |
| | HWU | 16 | 943 | 110 | 32 | 1,808 | 214 | 48 | 2,784 | 328 | 64 | 3,853 | 451 | 1,032 |
| | ECDT | 8 | 487 | 48 | 16 | 684 | 68 | 23 | 938 | 92 | 31 | 1,046 | 103 | 770 |
| | MCID | 4 | 148 | 23 | 8 | 298 | 45 | 12 | 450 | 67 | 16 | 595 | 90 | 331 |
| | DBPedia | 55 | 1,204 | 177 | 110 | 2,315 | 341 | 164 | 3,111 | 463 | 219 | 3,517 | 537 | 2,000 |
| | X-Topic | 12 | 2,657 | 378 | 24 | 2,917 | 415 | 37 | 2,970 | 432 | 49 | 2,982 | 444 | 1,767 |
| 1.0 | 20NG | 5 | 1,825 | 261 | 10 | 3,505 | 501 | 15 | 5,166 | 737 | 20 | 7,000 | 1,000 | 2,000 |
| | THUCNews | 4 | 2,000 | 287 | 7 | 3,500 | 502 | 10 | 5,000 | 716 | 14 | 7,000 | 1,000 | 2,000 |
| | Yahoo | 2 | 1,400 | 200 | 5 | 3,500 | 500 | 8 | 5,600 | 800 | 10 | 7,000 | 1,000 | 2,000 |
| | TREC | 12 | 3,224 | 350 | 24 | 4,350 | 475 | 35 | 4,683 | 513 | 47 | 4,849 | 532 | 490 |
| | BANK | 19 | 2,303 | 258 | 38 | 4,740 | 527 | 58 | 7,048 | 783 | 77 | 9,003 | 1,000 | 3,080 |
| | S.O. | 5 | 2,999 | 499 | 10 | 5,997 | 998 | 15 | 8,997 | 1,498 | 20 | 11,996 | 1,998 | 5,991 |
| | CLINC | 38 | 4,560 | 570 | 75 | 9,000 | 1,125 | 112 | 13,440 | 1,680 | 150 | 18,000 | 2,250 | 2,250 |
| | HWU | 16 | 1,886 | 227 | 32 | 3,616 | 436 | 48 | 5,575 | 674 | 64 | 7,712 | 933 | 1,032 |
| | ECDT | 8 | 976 | 92 | 16 | 1,372 | 130 | 23 | 1,879 | 180 | 31 | 2,099 | 200 | 770 |
| | MCID | 4 | 298 | 43 | 8 | 601 | 85 | 12 | 905 | 127 | 16 | 1,198 | 170 | 331 |
| | DBPedia | 55 | 2,387 | 345 | 110 | 4,603 | 661 | 164 | 6,195 | 886 | 219 | 7,000 | 1,000 | 2,000 |
| | X-Topic | 12 | 5,318 | 756 | 24 | 5,836 | 825 | 37 | 5,937 | 849 | 49 | 5,954 | 866 | 1,767 |

Table S2: **Detailed dataset split statistics** in the BOLT benchmark. Each dataset is partitioned under three Known Class Ratios (KCR = 0.25/0.5/0.75) and three Labeled Data Ratios (LAR = 0.1/0.5/1.0). For each configuration, we report the number of labels, training, evaluation, and test samples. This unified partitioning scheme ensures minimal class overlap across folds and consistent evaluation across both OSTC and GCD tasks.

within a unified framework to ensure fair comparison, using consistent training–evaluation protocols and hyperparameter settings.

**OSTC Methods.** *Non-PLM-based Methods:*

- **DOC** (Shu et al., 2017): A classical open-intent detection model that replaces the softmax layer with one-vs-rest sigmoid classifiers to reduce open-space risk, enabling explicit rejection of unknown intents.
- **DeepUnk** (Lin and Xu, 2019): Introduces a large-margin cosine loss to enhance separation between known and unknown samples, yielding more discriminative intent representations.

**OSTC Methods.** *PLM-based Methods.*

- **ADB** (Zhang et al., 2021a): Employs an adaptive spherical margin loss to dynamically adjust class decision boundaries in the pre-trained language model (PLM) feature space.
- **SCL** (Zeng et al., 2021a): Utilizes supervised contrastive learning to optimize intra-class compactness and inter-class separability.
- **AB** (Lorenc et al., 2022): Enhances metric-based classification through angular boundary scaling, improving generalization to unseen categories.
- **KNNCon** (Zhou et al., 2022): Combines KNN-based neighborhood contrastive learning with outlier detection using the Local Outlier Factor (LOF).

- **DyEn** (Zhou et al., 2023): Mitigates PLM "overthinking" through dynamic ensembling and confidence-based early exits for efficient OOD rejection.

**OSTC Methods.** *Proposed LLM-based Methods.* Since no LLM-based methods have been explored for OSTC to date, we further propose a series of LLM fine-tuning frameworks, each incorporating a different post-hoc OOD detection approach, to investigate the potential of modern LLMs in addressing OSTC. Specifically, we employ eight different OOD detection methods, described as follows:

- **MSP** (Hendrycks and Gimpel, 2016): Uses the maximum softmax probability as a simple yet effective post-hoc confidence score for OOD detection.
- **LLM-OpenMax** (Bendale and Boult, 2016): Fits Weibull distributions to logits to estimate OOD probabilities, refining uncertainty estimation beyond MSP.
- **TempScale** (Guo et al., 2017): Applies temperature scaling to logits for post-hoc calibration and more reliable confidence estimation.
- **EnergyBased** (Liu et al., 2020): Reformulates confidence estimation as an energy-based scoring function, offering a theoretically grounded OOD detection approach.
- **LogitNorm** (Wei et al., 2022): Normalizes logits during training to mitigate overconfidence and improve model calibration.
- **Entropy** (Chan et al., 2021): Employs predictive LLM-Entropy as a task-agnostic uncertainty measure for identifying unknown samples.
- **KLMatching** (Basart et al., 2022): Compares class-wise probability distributions using KL divergence for fine-grained OOD differentiation.
- **MaxLogit** (Basart et al., 2022): Scores samples based on their maximum logit values, achieving strong empirical performance with minimal computational overhead.

**GCD Methods.** *PLM-based Methods:*

- **DAL** (Zhang et al., 2021b): A strong GCD baseline employing alignment-based pseudo-labeling and feature consistency constraints for novel class discovery.
- **GeoID** (Tang et al., 2024): Enhances GCD via geometry-aware clustering, calibrating pseudo-labels through similarity-based distribution alignment.
- **SDC** (An et al., 2025): Introduces self-debiasing calibration to suppress dominant known-category influence and strengthen novel-class separation.
- **DPN** (An et al., 2023): Utilizes dual-path representation learning to decouple shared and category-specific knowledge for improved clustering quality.
- **TAN** (An et al., 2024b): Implements transfer and alignment mechanisms to bridge known and novel feature spaces through prototype-based knowledge transfer.

**GCD Methods.** *LLM-enhanced Methods:*

- **LOOP** (An et al., 2024a): Leverages LLM querying and neighborhood contrastive learning to align cluster semantics and enhance pseudo-label reliability.
- **ALUP** (Liang et al., 2024): Adopts comparison-based prompting with LLMs for active pseudo-label refinement, improving clustering stability.

All implementations in BOLT follow official codebases or public releases when available. Specifically, for these LLM-based OOD detection methods, we first fine-tuned the LLM on the training set and classified samples with an OOD score exceeding a predefined threshold as OOD. Following prior work (Zeng et al., 2021b), we set this threshold to 0.5. The baselines were implemented using the pytorch-ood (Kirchheim et al., 2022) [13]. Pytorch-ood consists of various OOD Detection with Deep Neural Networks based on PyTorch. This library is open-source and publicly available and distributed under the Apache-2.0 license.

### B.4 Evaluation Metrics

We employ distinct yet complementary evaluation metrics designed for the characteristics of the OSTC and GCD tasks.

**OSTC.** Following prior studies (Zhang et al., 2021a; Zhou et al., 2022, 2023), we adopt metrics that jointly assess classification performance on ID classes and detection capability on OOD samples.

[13] https://github.com/kkirchheim/pytorch-ood

| Task | Method | Backbone | Training Time |
|---|---|---|---|
| OSTC | DOC | BERT | 5m 20s |
| | DeepUnk | BERT | 5m 20s |
| | ADB | BERT | 6m 40s |
| | SCL | BERT | 6m 40s |
| | AB | BERT | 5m 10s |
| | KnnCon | BERT | 9m 10s |
| | DyEn | BERT | 18m 50s |
| | LLM+OOD detection | LLaMA-3.1-8B | 19m 2s |
| GCD | DAL | BERT | 14m 0s |
| | Geoid | BERT | 47m 10s |
| | SDC | BERT | 11m 20s |
| | DPN | BERT | 28m 50s |
| | TAN | BERT | 27m 50s |
| | LOOP | BERT + LLM API | 1h 0m 20s |
| | ALUP | BERT + LLM API | 1h 22m 20s |

Table S3: Training time on the BANKING dataset (KCR=0.25, LAR=0.1).

- **ACC:** ACC measures the proportion of correctly predicted samples in the entire test set: $\text{ACC} = \frac{1}{N_t}\sum_{i=1}^{N_t}\mathbb{I}(\hat{y}_i = y_i)$, where $\mathbb{I}(\cdot)$ equals 1 if the condition holds and 0 otherwise. This metric reflects global classification performance but may be biased when ID and OOD sample distributions are imbalanced.

- **Macro-F1 (F1):** To account for class imbalance, we adopt the macro-averaged F1 score computed over the full label space $\mathcal{Y}^k \cup \{\text{unknown}\}$, where all novel classes $\mathcal{Y}^n$ are collapsed into a single "unknown" label. For each label $c$, we compute:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \tag{3}$$

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c}, \tag{4}$$

$$\text{F1}_c = \frac{2 \times \text{Precision}_c \times \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}. \tag{5}$$

  Here, $TP_c$, $FP_c$, and $FN_c$ denote the numbers of true positives, false positives, and false negatives for class $c$, respectively. The overall macro-F1 is defined as: $\text{F1} = \frac{1}{|\mathcal{Y}^k|+1}\sum_{c\in\mathcal{Y}^k\cup\{\text{unknown}\}}\text{F1}_c$.

- **Known-Class F1 (K-F1):** K-F1 isolates model performance on in-distribution (known) classes: $\text{K-F1} = \frac{1}{|\mathcal{Y}^k|}\sum_{c\in\mathcal{Y}^k}\text{F1}_c$. It quantifies how well the model preserves classification accuracy on ID samples while performing OOD detection.

- **Novel-Class F1 (N-F1):** To evaluate OOD detection capability, all novel classes $\mathcal{Y}^n$ are collapsed into a single "unknown" class, and F1 is computed over this merged category: $\text{N-F1} = \text{F1}_{\text{unknown}}$. Higher N-F1 indicates stronger ability to reject or correctly identify unseen classes as "unknown".

**GCD.** For GCD, we follow the standard evaluation protocol introduced by Zhang et al. (2021b); An et al. (2024b, 2025) and adopted in subsequent studies. This protocol simultaneously evaluates the recognition of known categories and the discovery of novel categories:

- **K/N-ACC:** Clustering accuracy is computed separately for known and novel categories to assess performance on both in-distribution and unseen data. Given the Hungarian matching between clusters and labels (Kuhn, 2004), let $\mathcal{Y}^k$ and $\mathcal{Y}^n$ denote the sets of known and novel labels, respectively. The accuracy on each subset is given by: $\text{ACC}_S = \frac{1}{|\mathcal{Y}^S|}\sum_{y\in\mathcal{Y}^S}\frac{|\hat{C}_y \cap C_y|}{|C_y|}, \quad S \in \{k, n\}$, where $C_y$ and $\hat{C}_y$ represent the ground-truth and predicted sample sets for label $y$. We report $\text{K-ACC} = \text{ACC}_k$ and $\text{N-ACC} = \text{ACC}_n$.

- **H-Score:** The harmonic mean of K-ACC and N-ACC: $\text{H-Score} = \frac{2\cdot\text{K-ACC}\cdot\text{N-ACC}}{\text{K-ACC}+\text{N-ACC}}$, which rewards balanced performance across known and novel classes.

- **ACC:** Clustering accuracy over all samples after Hungarian matching. While intuitive, this measure can be dominated by known-class performance under class imbalance.

- **Normalized Mutual Information (NMI):** Quantifies the dependency between predicted clusters $U$ and ground-truth labels $V$: $\text{NMI}(U, V) = \frac{2I(U;V)}{H(U)+H(V)}$, where $I(U;V)$ is mutual information and $H(\cdot)$ denotes LLM-Entropy. Values range from 0 (no agreement) to 1 (perfect match).

- **Adjusted Rand Index (ARI):** Measures pairwise consistency between predicted and true partitions, adjusted for chance: $\text{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \frac{\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_i\binom{a_i}{2}+\sum_j\binom{b_j}{2}\right] - \frac{\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}}{\binom{n}{2}}}$, where $n_{ij}$ is the number of samples shared by cluster $i$ and class $j$, and $a_i, b_j$ are marginal sums. ARI values range from $-1$ (mismatch) to 1 (perfect alignment).

| Metric | KCR | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| K-F1 | 0.25 | DOC | 27.49 | 34.16 | 17.18 | 30.75 | 0.06 | 57.45 | 0.00 | 0.00 | 0.36 | 0.00 | 0.00 | 1.23 | 14.06 |
| | | DeepUNK | 33.54 | 14.28 | 26.59 | 26.95 | 37.49 | 25.44 | 46.57 | 48.02 | 3.45 | 35.92 | 28.10 | 5.92 | 27.69 |
| | | ADB | 43.99 | 61.09 | 39.23 | 18.65 | 34.73 | **70.76** | 69.33 | 49.04 | 35.09 | 36.75 | 40.45 | 14.22 | 42.78 |
| | | SCL | **96.57** | 15.05 | 22.49 | 8.58 | 18.75 | 33.15 | 23.17 | 26.83 | 1.79 | 20.11 | 9.71 | 4.27 | 23.37 |
| | | AB | 52.25 | 23.71 | **42.56** | 27.43 | 57.24 | 55.20 | 67.50 | 60.87 | 25.92 | 41.22 | 25.47 | 4.96 | 40.36 |
| | | KNNCon | 88.67 | 59.43 | 41.14 | 50.69 | 63.07 | 59.39 | **76.88** | **71.59** | 62.07 | 58.44 | 61.90 | 16.34 | 59.13 |
| | | DyEn | 93.77 | **63.08** | 39.54 | **62.77** | **63.17** | 50.77 | 76.55 | 65.97 | **77.94** | **65.99** | **66.66** | **19.57** | **62.15** |
| | | LLM-MaxSoftmax | 21.94 | 38.44 | 25.52 | 5.83 | 35.17 | 43.32 | 47.47 | 33.22 | 15.90 | 10.49 | 6.19 | 3.81 | 23.94 |
| | | LLM-OpenMax | 20.16 | 33.97 | 24.10 | 4.19 | 7.24 | 52.62 | 13.67 | 7.37 | 11.17 | 13.62 | 0.53 | 3.88 | 16.04 |
| | | LLM-TempScale | 21.65 | 38.32 | 25.25 | 5.64 | 35.00 | 42.28 | 47.09 | 32.29 | 14.85 | 9.18 | 6.01 | 4.24 | 23.48 |
| | | LLM-Energy | 23.71 | 37.24 | 23.04 | 5.00 | 31.97 | 61.17 | 44.05 | 28.42 | 8.21 | 7.46 | 3.25 | 2.82 | 23.03 |
| | | LLM-LogitNorm | 22.46 | 35.37 | 25.13 | 5.50 | 39.19 | 38.93 | 58.01 | 34.34 | 15.09 | 10.61 | 4.63 | 4.16 | 24.45 |
| | | LLM-Entropy | 22.47 | 40.81 | 26.49 | 5.49 | 35.49 | 46.37 | 47.41 | 33.75 | 16.76 | 8.12 | 6.24 | 4.25 | 24.47 |
| | | LLM-KLMatching | 21.73 | 37.74 | 24.84 | 6.40 | 32.86 | 40.10 | 43.07 | 31.85 | 17.25 | 12.54 | 6.27 | 5.37 | 23.34 |
| | | LLM-MaxLogit | 24.44 | 38.56 | 23.42 | 5.20 | 37.16 | 63.51 | 50.56 | 31.81 | 10.30 | 7.88 | 4.40 | 2.84 | 25.01 |
| | 0.5 | DOC | 19.19 | 37.60 | 4.61 | 0.98 | 0.08 | 68.29 | 43.03 | 0.00 | 0.42 | 0.00 | 0.00 | 0.25 | 14.54 |
| | | DeepUNK | 30.36 | 12.00 | 11.46 | 27.72 | 48.99 | 26.42 | 59.97 | 58.03 | 2.04 | 50.95 | 27.79 | 4.21 | 30.00 |
| | | ADB | 70.45 | **81.23** | 56.07 | 23.66 | 57.86 | **82.56** | 72.86 | 59.14 | 47.55 | 41.33 | 58.15 | 15.91 | 55.56 |
| | | SCL | **98.47** | 14.51 | 22.88 | 12.96 | 22.57 | 47.49 | 32.10 | 27.25 | 1.39 | 29.21 | 8.13 | 5.09 | 26.84 |
| | | AB | 53.74 | 25.71 | 43.59 | 18.48 | 61.04 | 63.20 | 70.10 | 64.81 | 27.48 | 43.75 | 23.96 | 12.28 | 42.34 |
| | | KNNCon | 92.65 | 77.87 | **58.07** | **54.95** | **77.67** | 80.34 | **86.93** | **81.54** | 80.99 | **79.12** | 66.14 | **21.02** | **71.44** |
| | | DyEn | 93.76 | 75.91 | 55.16 | 51.40 | 77.63 | 74.38 | 85.94 | 77.92 | **81.57** | 76.21 | **72.12** | 17.59 | 69.97 |
| | | LLM-MaxSoftmax | 78.65 | 54.14 | 43.72 | 21.59 | 57.66 | 71.72 | 75.29 | 51.24 | 24.11 | 8.53 | 10.36 | 6.75 | 41.98 |
| | | LLM-OpenMax | 68.11 | 35.45 | 34.10 | 8.88 | 7.60 | 71.74 | 22.21 | 10.43 | 6.16 | 10.60 | 0.58 | 2.96 | 23.24 |
| | | LLM-TempScale | 78.63 | 54.16 | 43.73 | 21.48 | 57.45 | 71.47 | 74.95 | 51.24 | 24.62 | 8.26 | 10.38 | 7.31 | 41.97 |
| | | LLM-Energy | 80.84 | 44.30 | 31.41 | 11.16 | 41.02 | 74.87 | 62.65 | 25.88 | 10.16 | 5.59 | 3.52 | 3.54 | 32.91 |
| | | LLM-LogitNorm | 79.64 | 51.65 | 42.04 | 19.57 | 54.98 | 73.09 | 73.70 | 43.03 | 21.73 | 10.56 | 6.54 | 4.64 | 40.10 |
| | | LLM-Entropy | 79.55 | 55.18 | 44.74 | 20.64 | 57.51 | 72.61 | 74.67 | 52.58 | 23.80 | 6.98 | 10.49 | 6.83 | 42.13 |
| | | LLM-KLMatching | 76.05 | 53.82 | 43.60 | 21.11 | 56.12 | 68.69 | 71.89 | 51.28 | 23.40 | 12.18 | 10.46 | 8.41 | 41.42 |
| | | LLM-MaxLogit | 85.08 | 47.26 | 33.33 | 13.07 | 48.95 | 78.67 | 70.35 | 32.30 | 13.25 | 5.95 | 5.18 | 4.24 | 36.47 |
| | 0.75 | DOC | 31.92 | 27.21 | 5.70 | 23.15 | 0.05 | 72.87 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 13.41 |
| | | DeepUNK | 52.06 | 7.10 | 14.14 | 18.27 | 49.98 | 32.75 | 63.62 | 60.03 | 1.28 | 62.72 | 22.22 | 5.32 | 32.46 |
| | | ADB | 88.49 | 84.66 | 62.84 | 24.86 | 63.66 | 84.19 | 81.02 | 65.87 | 51.55 | 48.75 | 65.95 | 13.89 | 61.31 |
| | | SCL | **98.39** | 14.65 | 16.55 | 15.18 | 23.70 | 63.65 | 55.21 | 41.50 | 1.12 | 27.29 | 9.08 | 6.00 | 31.03 |
| | | AB | 57.20 | 24.99 | 48.33 | 15.46 | 62.71 | 66.60 | 71.79 | 64.71 | 28.40 | 43.16 | 22.42 | 11.64 | 43.12 |
| | | KNNCon | 97.71 | **85.35** | **66.76** | **72.68** | 86.60 | **87.34** | **92.85** | **86.50** | 88.22 | 84.72 | 73.74 | **20.65** | **78.59** |
| | | DyEn | 95.58 | 84.04 | 65.75 | 62.85 | **86.96** | 85.80 | 92.39 | 82.94 | **90.28** | **84.78** | **79.43** | 17.71 | 77.38 |
| | | LLM-MaxSoftmax | 91.67 | 72.72 | 55.87 | 28.49 | 65.93 | 84.11 | 86.33 | 61.50 | 32.59 | 36.74 | 12.52 | 8.95 | 53.12 |
| | | LLM-OpenMax | 51.59 | 49.35 | 36.73 | 8.13 | 11.23 | 73.33 | 30.90 | 9.65 | 4.43 | 8.86 | 1.18 | 2.34 | 23.98 |
| | | LLM-TempScale | 91.37 | 72.72 | 55.87 | 28.34 | 65.94 | 84.10 | 86.31 | 61.46 | 32.98 | 36.83 | 12.72 | 9.20 | 53.15 |
| | | LLM-Energy | 81.54 | 53.74 | 29.44 | 13.55 | 34.68 | 76.84 | 68.00 | 32.54 | 9.10 | 17.46 | 3.87 | 3.34 | 35.34 |
| | | LLM-LogitNorm | 93.36 | 72.59 | 54.45 | 21.13 | 49.56 | 86.76 | 78.47 | 46.16 | 26.44 | 30.22 | 6.07 | 6.89 | 47.68 |
| | | LLM-Entropy | 92.04 | 73.44 | 56.28 | 29.14 | 65.92 | 83.92 | 85.83 | 62.79 | 31.94 | 36.37 | 12.97 | 9.15 | 53.32 |
| | | LLM-KLMatching | 90.62 | 72.73 | 55.98 | 28.71 | 65.96 | 82.38 | 85.65 | 61.74 | 34.20 | 38.70 | 13.94 | 10.32 | 53.41 |
| | | LLM-MaxLogit | 87.51 | 58.59 | 32.79 | 16.42 | 43.63 | 80.39 | 75.51 | 40.04 | 11.67 | 21.35 | 6.19 | 4.54 | 39.89 |

Table S4: K-F1 performance of OSTC on the BOLT benchmark under different KCRs (LAR = 0.1)

### B.5 Implementation Details

All experiments were conducted on a workstation equipped with eight NVIDIA RTX 5090 GPUs (each with 32 GB of memory) and 512 GB of system RAM. For each KCR setting, the known-class label space was uniformly partitioned into five splits with minimal overlap to ensure stable and comparable evaluation. Experiments were conducted on all five splits, and the final results were reported as the average performance across these splits.

We adopted **LLaMA-3.1-8B**[14] as the default LLM backbone for all LLM-based methods. For PLM-based discriminative baselines, we used the **bert-base-uncased**[15] model for English datasets and **bert-base-chinese**[16] for Chinese datasets to ensure linguistic consistency. For LLM-enhanced variants, we further incorporated **DeepSeek-V3.1** as a generative verification and semantic calibration module, which supports pseudo-label validation and open-set consistency checking.

All implementations in BOLT followed official codebases or public releases when available. Each configuration was pre-trained for 100 epochs when applicable and fine-tuned for 50 epochs, with early stopping applied. Specifically, in our implementation of LLM-based OOD detection, the models were fine-tuned on the training set, and OOD detection was performed using the `pytorch-ood` library[17], an open-source framework.

## C Runtime Comparison Across Methods

Table S3 reports the end-to-end training time on BANKING with `KCR=0.25` and `LAR=0.1`. For OSTC, most BERT-based baselines are relatively lightweight, finishing within about 5–9 minutes (e.g., AB: 5m10s; DOC/DeepUnk: 5m20s; ADB/SCL: 6m40s; KnnCon: 9m10s). In contrast, methods involving heavier ensembling or addi-

[14] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[15] https://huggingface.co/google-bert/bert-base-uncased

[16] https://huggingface.co/google-bert/bert-base-chinese

[17] https://github.com/kkirchheim/pytorch-ood

| Metric | KCR | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| K-F1 | 0.25 | DOC | 74.47 | 43.80 | 38.71 | 29.59 | 60.96 | 58.47 | 77.20 | 65.11 | 0.00 | 56.39 | 38.09 | 9.41 | 46.02 |
| | | DeepUNK | 82.13 | 16.60 | 43.12 | 24.10 | 63.86 | 55.91 | 70.89 | 64.56 | 1.41 | 51.86 | 30.01 | 9.07 | 42.79 |
| | | ADB | 69.34 | 73.67 | 47.47 | 45.74 | 67.06 | **76.68** | 80.15 | 71.86 | 60.74 | 50.48 | **74.20** | 19.41 | 61.40 |
| | | SCL | 68.20 | 20.17 | 17.71 | 28.90 | 20.43 | 40.75 | 39.42 | 29.90 | 2.35 | 38.61 | 22.47 | 9.00 | 28.16 |
| | | AB | 53.40 | 25.90 | 42.02 | 26.00 | 58.20 | 59.32 | 70.82 | 67.40 | 40.12 | 58.26 | 45.38 | 8.38 | 46.27 |
| | | KNNCon | 85.52 | 54.62 | 42.36 | 38.75 | 62.27 | 60.40 | 76.40 | 72.35 | 60.35 | 60.02 | 56.78 | 16.37 | 57.18 |
| | | DyEn | **92.41** | 63.88 | 39.20 | 42.45 | 64.46 | 55.67 | 76.54 | 68.79 | **62.67** | **64.35** | 63.42 | **23.21** | 59.75 |
| | | LLM-MaxSoftmax | 53.72 | 48.62 | 40.49 | 46.97 | 49.36 | 47.32 | 56.74 | 60.88 | 46.57 | 47.34 | 50.15 | 12.47 | 46.72 |
| | | LLM-OpenMax | 79.85 | 58.68 | 40.31 | 33.57 | 56.44 | 57.02 | 70.50 | 61.48 | 37.19 | 42.25 | 18.17 | 11.47 | 47.24 |
| | | LLM-TempScale | 53.72 | 48.62 | 40.53 | 46.39 | 49.07 | 47.78 | 56.74 | 60.88 | 45.57 | 46.86 | 50.14 | 12.60 | 46.58 |
| | | LLM-Energy | 88.07 | 74.88 | 49.66 | 39.71 | 69.68 | 74.20 | 85.36 | 72.83 | 49.89 | 55.28 | 52.78 | 5.63 | 59.83 |
| | | LLM-LogitNorm | 53.46 | 47.70 | 41.28 | 46.52 | 65.27 | 50.06 | 82.68 | 69.16 | 45.97 | 46.57 | 59.31 | 11.10 | 51.59 |
| | | LLM-Entropy | 55.50 | 49.86 | 43.22 | **48.51** | 48.65 | 48.83 | 55.64 | 61.45 | 48.22 | 50.82 | 49.84 | 12.41 | 47.75 |
| | | LLM-KLMatching | 50.79 | 46.76 | 40.06 | 42.26 | 46.24 | 44.52 | 53.30 | 59.09 | 43.59 | 46.14 | 46.30 | 13.51 | 44.38 |
| | | LLM-MaxLogit | 87.58 | **76.17** | **50.37** | 40.68 | **71.77** | 76.55 | **86.28** | **74.48** | 52.93 | 57.46 | 56.47 | 7.32 | **61.50** |
| | 0.5 | DOC | 71.82 | 48.20 | 39.74 | 41.15 | 64.91 | 72.75 | 79.48 | 69.43 | 0.00 | 62.97 | 16.32 | 8.43 | 47.93 |
| | | DeepUNK | 86.98 | 38.56 | 44.72 | 24.50 | 58.77 | 61.07 | 67.23 | 65.28 | 1.22 | 51.31 | 22.24 | 7.56 | 44.12 |
| | | ADB | 86.02 | **84.80** | 56.99 | 62.24 | 77.57 | 85.26 | 86.42 | 77.66 | 81.33 | 72.21 | **84.26** | 19.43 | **72.85** |
| | | SCL | 79.46 | 26.56 | 15.72 | 35.58 | 34.75 | 50.43 | 74.35 | 38.73 | 1.71 | 55.16 | 36.39 | 9.82 | 38.22 |
| | | AB | 59.80 | 29.00 | 46.80 | 26.72 | 64.22 | 65.12 | 74.80 | 68.94 | 47.20 | 60.56 | 47.70 | 15.88 | 50.56 |
| | | KNNCon | 92.69 | 75.17 | 57.59 | 55.10 | 79.40 | 81.16 | 86.60 | **81.44** | **81.65** | 77.14 | 64.31 | **20.38** | 71.05 |
| | | DyEn | 95.30 | 76.83 | 56.18 | 56.19 | 77.64 | 74.84 | 86.63 | 78.14 | 77.85 | **77.27** | 74.40 | 19.34 | 70.88 |
| | | LLM-MaxSoftmax | 79.12 | 72.10 | 61.22 | 62.47 | 71.07 | 71.56 | 78.57 | 76.17 | 69.99 | 67.18 | 74.17 | 14.52 | 66.51 |
| | | LLM-OpenMax | 92.34 | 81.48 | 58.97 | 37.10 | 55.02 | 85.18 | 73.90 | 62.76 | 43.16 | 52.12 | 28.87 | 13.36 | 57.02 |
| | | LLM-TempScale | 79.12 | 72.22 | 61.04 | 62.19 | 70.98 | 71.56 | 78.73 | 76.04 | 69.65 | 66.91 | 74.17 | 14.92 | 66.46 |
| | | LLM-Energy | 94.70 | 76.88 | 45.83 | 39.89 | 74.82 | 84.96 | 86.67 | 73.42 | 46.61 | 63.27 | 62.64 | 6.73 | 63.03 |
| | | LLM-LogitNorm | 83.76 | 72.80 | 58.38 | **63.58** | **80.99** | 76.73 | **91.61** | 79.73 | 72.64 | 67.63 | 77.13 | 10.50 | 69.62 |
| | | LLM-Entropy | 79.93 | 72.73 | **62.05** | 62.45 | 70.10 | 71.50 | 77.29 | 76.54 | 70.05 | 67.98 | 73.25 | 14.93 | 66.57 |
| | | LLM-KLMatching | 76.49 | 69.91 | 58.05 | 59.88 | 68.99 | 70.56 | 75.25 | 74.55 | 67.88 | 65.91 | 71.19 | 18.87 | 64.79 |
| | | LLM-MaxLogit | **95.40** | 79.10 | 49.51 | 47.45 | 77.51 | **86.21** | 89.63 | 77.10 | 54.15 | 66.35 | 71.67 | 9.23 | 66.94 |
| | 0.75 | DOC | 77.65 | 43.92 | 36.91 | 41.66 | 66.98 | 78.98 | 80.49 | 71.88 | 0.00 | 62.12 | 0.00 | 8.03 | 47.39 |
| | | DeepUNK | 92.53 | 38.12 | 46.17 | 21.07 | 48.39 | 64.67 | 63.01 | 60.84 | 0.17 | 50.93 | 19.44 | 5.11 | 42.54 |
| | | ADB | 94.41 | **87.63** | 65.47 | 66.35 | 82.81 | 86.33 | 90.40 | 82.14 | 84.86 | 77.09 | **88.22** | 19.70 | 77.12 |
| | | SCL | 72.52 | 28.76 | 26.02 | 40.09 | 56.76 | 69.12 | 80.32 | 47.91 | 1.31 | 65.67 | 49.66 | 10.44 | 45.72 |
| | | AB | 63.74 | 29.72 | 51.04 | 27.36 | 67.82 | 68.68 | 77.80 | 71.06 | 50.04 | 60.88 | 49.16 | 16.00 | 52.78 |
| | | KNNCon | **97.45** | 85.00 | 66.16 | 69.33 | **86.99** | 86.65 | **92.79** | **86.30** | **87.65** | **84.95** | 73.32 | **20.89** | **78.12** |
| | | DyEn | 95.37 | 84.93 | 64.57 | 64.01 | 86.21 | 84.69 | 92.05 | 84.18 | 84.26 | 84.59 | 80.95 | 17.98 | 76.98 |
| | | LLM-MaxSoftmax | 92.40 | 82.28 | 70.09 | 70.45 | 84.07 | 85.68 | 90.17 | 84.59 | 80.91 | 77.16 | 84.66 | 17.05 | 76.63 |
| | | LLM-OpenMax | 93.57 | 85.37 | 68.85 | 38.94 | 60.49 | **89.22** | 77.66 | 67.69 | 31.28 | 54.42 | 35.42 | 13.25 | 59.68 |
| | | LLM-TempScale | 92.40 | 82.37 | **70.17** | 70.84 | 83.94 | 85.90 | 90.34 | 84.50 | 80.84 | 76.93 | 84.80 | 17.42 | 76.70 |
| | | LLM-Energy | 90.25 | 81.04 | 33.87 | 40.28 | 74.35 | 82.53 | 82.62 | 73.25 | 42.66 | 47.43 | 67.14 | 7.39 | 60.23 |
| | | LLM-LogitNorm | 96.48 | 84.56 | 67.75 | 62.48 | 81.37 | 88.99 | 91.98 | 81.91 | 79.34 | 72.07 | 75.58 | 10.41 | 74.41 |
| | | LLM-Entropy | 92.70 | 82.46 | 69.11 | **70.88** | 83.27 | 85.70 | 89.64 | 84.36 | 81.41 | 77.64 | 84.72 | 17.62 | 76.63 |
| | | LLM-KLMatching | 91.71 | 81.49 | 69.37 | 70.52 | 82.96 | 83.59 | 89.22 | 83.86 | 79.75 | 76.48 | 83.14 | 20.31 | 76.03 |
| | | LLM-MaxLogit | 92.73 | 83.03 | 44.15 | 44.64 | 77.96 | 85.02 | 87.90 | 77.20 | 49.36 | 53.32 | 74.93 | 9.92 | 65.01 |

Table S5: K-F1 performance of OSTC on the BOLT benchmark under different KCRs (LAR = 0.5)

tional components incur noticeably higher costs: DyEn requires 18m50s.

For GCD, training is generally more time-consuming than OSTC. This is partly because GCD typically leverages additional unlabeled data and performs pseudo-label learning, which introduces extra computation beyond supervised training. Under this setting, SDC and DAL finish in 11m20s and 14m00s, while DPN and TAN take around 28 minutes and Geoid reaches 47m10s. Finally, LLM-enhanced methods are the most expensive, as they additionally depend on external LLMs during training: LOOP and ALUP require 1h00m20s and 1h22m20s, respectively, underscoring the substantial overhead of LLM-augmented training.

# D Unified Experiment Runner and Configuration Example

This work provides a unified experiment entry point that supports multiple task types (e.g., GCD and OSTC) and multiple datasets within a single framework. It enables grid-style experiments over the known class ratio (`kcr`) and labeled ratio (`lar`), while allowing different methods.

## D.1 Command-line Entry

```
bolt-grid \
  --config ./config.yaml \
  --output-dir ./exp \
  --model-dir ./pretrained_models
```

where:

- `-config`: the YAML configuration file specifying tasks, methods, datasets, and grid settings.
- `-output-dir`: the output directory. Logs, aggregated results, and model artifacts are written to this directory for reproducibility and archiving.
- `-model-dir`: the local directory of pretrained models. Users should download and place required models in advance (e.g., `bert-base-chinese`, `bert-base-uncased`, `Meta-Llama-3.1-8B-Instruct`). The configuration file only references model names/identifiers.

### D.2 Configuration Example (`config.yaml`)

**Minimal example.** The following configuration illustrates the main fields with a small subset of methods and datasets for readability.

```
maps:
  gcd: [alup, deepaligned, dpn, geoid, loop,
  ↪  plm_gcd, sdc, tan]
  openset: [ab, adb, clap, deepunk, doc, dyen,
  ↪  knncon, plm_ood, scl]

methods: [alup, deepaligned]
datasets: [banking, clinc]
result_file: summary

grid:
  known_cls_ratio: [0.25, 0.5, 0.75]
  labeled_ratio:   [0.1, 0.5, 1.0]
  fold_idxs: [0, 1, 2, 3, 4]
  fold_num: 5
  seed: 2025
  cluster_num_factor: [1.0]

run:
  gpus: [0, 1]
  max_workers: 2
  num_pretrain_epochs: 100
  num_train_epochs: 50

paths:
  results_dir: results
  logs_dir: logs
```

### D.3 Field Descriptions

- `maps`: assigns each method to a task type (e.g., GCD vs. OpenSet), enabling a unified runner across tasks.
- `methods`: the methods to execute in the current run. Experiments are instantiated by combining `methods`, `datasets`, and `grid`.
- `datasets`: the datasets to evaluate. The runner is dataset-agnostic and supports multiple datasets in one configuration.
- `grid`: grid variables for controlled evaluation. `known_cls_ratio` corresponds to `kcr` and `labeled_ratio` corresponds to `lar`. `fold_*` and `seed` control reproducibility, and `cluster_num_factor` is used by some clustering/class-discovery methods.

## E Additional Details of Experimental Results

### E.1 Additional Results on OSTC

Table S4-S15 provide all experimental results across all datasets for the OSTC task, covering four evaluation metrics: K-F1, N-F1, ACC, and F1, under different settings of LAR = {0.1, 0.5, 1.0} and KCR = {0.25, 0.5, 0.75}.

### E.2 Additional Results on GCD

Table S16-S33 provide all experimental results across all datasets for the GCD task, including six evaluation metrics ACC, ARI, NMI, H-Score, K-ACC, and N-ACC evaluated under the same configurations of LAR = {0.1, 0.5, 1.0} and KCR = {0.25, 0.5, 0.75}.

| Metric | KCR | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| K-F1 | 0.25 | DOC | 76.90 | 48.18 | 39.98 | 42.81 | 66.74 | 57.17 | 80.21 | 70.46 | 0.02 | 62.61 | 52.03 | 12.53 | 50.80 |
| | | DeepUNK | 92.69 | 34.13 | 43.76 | 35.37 | 70.48 | 62.65 | 75.07 | 73.46 | 1.70 | 58.44 | 33.07 | 11.21 | 49.34 |
| | | ADB | 73.70 | 73.76 | 48.93 | 54.51 | 66.44 | 76.90 | 82.33 | 73.27 | 71.86 | 62.47 | 78.94 | 21.76 | 65.41 |
| | | SCL | 69.02 | 25.11 | 17.04 | 34.56 | 27.80 | 41.11 | 50.33 | 39.31 | 2.74 | 43.48 | 38.18 | 10.30 | 33.25 |
| | | AB | 53.52 | 26.07 | 42.27 | 28.18 | 58.56 | 59.34 | 70.90 | 67.30 | 42.04 | 60.05 | 51.71 | 10.32 | 47.52 |
| | | KNNCon | 86.81 | 62.57 | 40.13 | 44.55 | 64.16 | 60.07 | 78.33 | 71.88 | 61.22 | 58.93 | 58.21 | 17.30 | 58.68 |
| | | DyEn | **96.23** | 63.06 | 39.40 | 52.56 | 64.42 | 56.51 | 80.46 | 66.86 | **74.14** | 64.30 | 62.86 | **23.21** | 62.00 |
| | | LLM-MaxSoftmax | 59.20 | 51.25 | 39.62 | 47.16 | 50.92 | 46.55 | 58.12 | 61.96 | 50.23 | 51.86 | 59.18 | 13.73 | 49.15 |
| | | LLM-OpenMax | 81.27 | 64.07 | 39.02 | 42.39 | 68.78 | 63.70 | 81.59 | **78.36** | 51.03 | 53.90 | 41.91 | 14.56 | 56.71 |
| | | LLM-TempScale | 58.81 | 51.37 | 39.70 | 46.97 | 50.89 | 46.63 | 58.13 | 61.81 | 48.49 | 51.23 | 59.43 | 13.57 | 48.92 |
| | | LLM-Energy | 88.48 | 77.16 | 53.97 | 48.59 | 72.77 | 82.38 | 88.76 | 75.94 | 56.89 | 64.40 | 71.05 | 9.23 | 65.80 |
| | | LLM-LogitNorm | 59.32 | 50.01 | 40.63 | **55.93** | 64.35 | 45.17 | 82.62 | 71.73 | 46.43 | 49.05 | **79.38** | 11.20 | 54.65 |
| | | LLM-Entropy | 60.66 | 52.35 | 42.78 | 48.05 | 49.58 | 47.22 | 56.85 | 61.79 | 52.80 | 54.47 | 58.55 | 13.74 | 49.90 |
| | | LLM-KLMatching | 56.65 | 49.79 | 38.43 | 44.43 | 47.51 | 44.97 | 54.14 | 60.02 | 44.25 | 49.22 | 54.40 | 14.86 | 46.56 |
| | | LLM-MaxLogit | 88.64 | **77.74** | **54.32** | 53.44 | **74.92** | **82.92** | **89.70** | 77.61 | 60.38 | **66.60** | 75.89 | 10.57 | **67.73** |
| | 0.5 | DOC | 78.55 | 51.70 | 46.65 | 53.68 | 74.69 | 73.21 | 84.12 | 75.64 | 0.00 | 70.02 | 49.03 | 10.01 | 55.61 |
| | | DeepUNK | 95.61 | 41.68 | 49.91 | 30.33 | 62.73 | 69.18 | 71.20 | 71.46 | 1.81 | 63.45 | 22.51 | 8.41 | 49.02 |
| | | ADB | 87.12 | 84.64 | 58.13 | 65.76 | 81.32 | **85.92** | 88.60 | 80.46 | 82.35 | 77.11 | **84.63** | **22.60** | **74.89** |
| | | SCL | 55.15 | 36.41 | 19.70 | 36.44 | 48.01 | 55.43 | 78.09 | 48.95 | 2.10 | 64.48 | 41.61 | 10.70 | 41.42 |
| | | AB | 60.44 | 29.60 | 47.36 | 28.70 | 64.64 | 65.42 | 74.82 | 68.54 | 49.18 | 61.88 | 56.60 | 14.00 | 51.77 |
| | | KNNCon | 92.92 | 77.29 | 58.84 | 58.78 | 77.68 | 79.83 | 86.79 | 80.95 | **83.85** | 76.38 | 64.48 | 20.80 | 71.55 |
| | | DyEn | 95.45 | 76.71 | 56.44 | 57.70 | 78.42 | 74.07 | 86.01 | 79.95 | 81.57 | **79.24** | 75.29 | 19.64 | 71.71 |
| | | LLM-MaxSoftmax | 79.39 | 70.11 | 64.27 | 64.60 | 72.16 | 70.65 | 77.32 | 78.91 | 70.87 | 71.05 | 75.95 | 18.77 | 67.84 |
| | | LLM-OpenMax | 93.72 | 82.98 | 58.97 | 47.83 | 73.14 | 83.66 | 85.35 | 78.72 | 59.48 | 73.85 | 48.03 | 18.15 | 66.99 |
| | | LLM-TempScale | 79.39 | 70.11 | 64.05 | 64.60 | 72.16 | 70.76 | 77.32 | 78.91 | 70.34 | 70.97 | 76.08 | 19.05 | 67.81 |
| | | LLM-Energy | **98.44** | 84.45 | 50.02 | 55.17 | 79.66 | 83.71 | 89.64 | 79.05 | 58.23 | 66.34 | 71.00 | 9.65 | 68.78 |
| | | LLM-LogitNorm | 85.19 | 73.12 | 61.57 | **71.66** | **83.75** | 75.08 | **92.40** | **85.19** | 80.28 | 73.66 | 83.31 | 10.57 | 72.98 |
| | | LLM-Entropy | 79.97 | 70.24 | **66.11** | 63.14 | 70.85 | 70.29 | 76.22 | 78.35 | 71.16 | 72.32 | 74.83 | 19.57 | 67.75 |
| | | LLM-KLMatching | 78.21 | 69.17 | 60.00 | 63.58 | 70.88 | 69.86 | 75.19 | 78.18 | 66.96 | 69.89 | 72.42 | 20.60 | 66.24 |
| | | LLM-MaxLogit | 97.04 | **85.06** | 56.29 | 59.56 | 81.84 | 84.62 | 91.40 | 81.49 | 62.95 | 69.30 | 78.66 | 11.80 | 71.67 |
| | 0.75 | DOC | 82.52 | 49.60 | 44.31 | 56.33 | 76.51 | 80.46 | 86.04 | 78.52 | 0.00 | 72.68 | 47.71 | 10.61 | 57.11 |
| | | DeepUNK | 96.92 | 41.87 | 52.22 | 24.18 | 44.94 | 73.66 | 64.78 | 66.97 | 1.19 | 64.05 | 17.64 | 5.77 | 46.18 |
| | | ADB | 94.48 | **88.02** | 66.79 | 72.90 | 85.79 | 86.91 | 92.33 | 83.99 | 87.08 | 80.06 | **88.54** | 22.96 | **79.15** |
| | | SCL | 75.88 | 30.89 | 34.55 | 56.60 | 63.45 | 63.11 | 84.96 | 70.62 | 1.74 | 74.25 | 56.36 | 13.67 | 52.17 |
| | | AB | 64.18 | 30.36 | 51.58 | 31.06 | 68.66 | 69.26 | 77.86 | 71.12 | 52.40 | 62.74 | 59.08 | 17.14 | 54.62 |
| | | KNNCon | 97.60 | 86.14 | 66.39 | 69.91 | **86.59** | 86.82 | 93.09 | **86.66** | 87.68 | **85.44** | 71.49 | 21.68 | 78.29 |
| | | DyEn | 94.69 | 84.32 | 62.68 | 69.56 | 85.71 | 84.29 | 92.12 | 84.94 | **90.28** | 85.00 | 82.96 | 19.26 | 77.98 |
| | | LLM-MaxSoftmax | 93.14 | 82.32 | **71.79** | **78.67** | 85.12 | 85.14 | 89.90 | 86.35 | 84.53 | 83.26 | 86.35 | 20.87 | 78.95 |
| | | LLM-OpenMax | 97.30 | 85.78 | 70.14 | 52.43 | 76.10 | **90.37** | 88.93 | 80.60 | 54.00 | 74.61 | 53.49 | 18.17 | 70.16 |
| | | LLM-TempScale | 93.14 | 82.39 | 71.73 | 78.67 | 85.13 | 85.14 | 90.04 | 86.36 | 84.45 | 83.18 | 86.38 | 21.06 | 78.97 |
| | | LLM-Energy | **99.45** | 81.18 | 37.06 | 51.53 | 75.55 | 81.81 | 86.38 | 74.51 | 63.30 | 65.91 | 74.66 | 9.03 | 66.70 |
| | | LLM-LogitNorm | 96.90 | 85.84 | 70.20 | 74.18 | 83.99 | 88.40 | **93.55** | 84.08 | 86.85 | 82.26 | 84.66 | 10.79 | 78.48 |
| | | LLM-Entropy | 93.69 | 82.53 | 71.72 | 78.52 | 84.61 | 84.87 | 89.22 | 86.20 | 84.74 | 83.32 | 85.65 | 20.87 | 78.83 |
| | | LLM-KLMatching | 92.25 | 81.51 | 70.35 | 77.64 | 84.10 | 84.37 | 88.79 | 85.84 | 83.42 | 82.37 | 84.77 | **23.44** | 78.24 |
| | | LLM-MaxLogit | 99.19 | 83.15 | 45.37 | 56.65 | 79.30 | 83.73 | 89.12 | 78.44 | 70.57 | 69.28 | 81.83 | 11.76 | 70.70 |

Table S6: K-F1 performance of OSTC on the BOLT benchmark under different KCRs (LAR = 1.0)

| Metric | KCR | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| N-F1 | 0.25 | DOC | 87.39 | **84.38** | 76.33 | 67.42 | 81.92 | 84.59 | 85.50 | 85.83 | 69.09 | **86.02** | 80.16 | **82.25** | **80.91** |
| | | DeepUNK | 88.01 | 62.66 | **83.35** | 49.85 | 4.16 | 86.06 | 9.96 | 0.15 | 0.00 | 24.35 | 12.02 | 7.48 | 35.67 |
| | | ADB | 46.79 | 74.82 | 43.06 | 34.81 | 44.46 | 85.89 | 85.41 | 72.11 | 63.51 | 47.99 | 70.66 | 50.46 | 60.00 |
| | | SCL | **99.03** | 73.80 | 47.05 | 51.66 | 34.18 | 86.67 | 57.84 | 42.71 | 2.14 | 35.08 | 42.19 | 41.08 | 51.12 |
| | | AB | 84.72 | 52.99 | 80.08 | 0.00 | 84.70 | 83.81 | 88.91 | **91.61** | **82.62** | 83.60 | 83.58 | 78.14 | 74.56 |
| | | KNNCon | 90.36 | 55.74 | 34.78 | 54.45 | 67.07 | 61.43 | 84.02 | 71.83 | 53.06 | 44.92 | 69.14 | 28.85 | 59.64 |
| | | DyEn | 96.06 | 65.10 | 24.23 | 34.54 | 70.80 | 47.93 | 86.57 | 74.58 | 72.99 | 74.46 | 69.25 | 29.36 | 62.16 |
| | | LLM-MaxSoftmax | 46.36 | 42.27 | 39.16 | 56.75 | 61.68 | 50.71 | 69.30 | 61.26 | 64.26 | 52.54 | 72.98 | 49.67 | 55.58 |
| | | LLM-OpenMax | 42.37 | 66.18 | 17.64 | 60.94 | 86.15 | 67.90 | 86.22 | 85.12 | 80.22 | 47.10 | **85.31** | 75.82 | 66.75 |
| | | LLM-TempScale | 50.51 | 41.65 | 36.12 | 54.29 | 61.77 | 48.71 | 69.13 | 59.09 | 63.68 | 59.29 | 74.06 | 54.34 | 56.05 |
| | | LLM-Energy | 76.40 | 81.60 | 73.32 | **75.72** | **87.65** | **87.92** | 89.14 | 84.56 | 80.21 | 73.49 | 84.61 | 69.72 | 80.36 |
| | | LLM-LogitNorm | 31.32 | 27.06 | 47.32 | 63.18 | 76.67 | 21.36 | 88.05 | 70.23 | 38.64 | 43.12 | 81.60 | 46.89 | 52.95 |
| | | LLM-Entropy | 54.31 | 53.52 | 60.38 | 62.42 | 61.64 | 61.42 | 68.64 | 64.28 | 72.80 | 64.10 | 73.02 | 51.71 | 62.35 |
| | | LLM-KLMatching | 28.37 | 37.04 | 29.24 | 13.26 | 39.75 | 30.83 | 42.46 | 35.62 | 14.74 | 22.50 | 41.17 | 27.03 | 30.17 |
| | | LLM-MaxLogit | 74.74 | 80.59 | 74.00 | 72.60 | 87.54 | 87.49 | **89.60** | 83.26 | 78.26 | 72.74 | 83.30 | 66.33 | 79.20 |
| | 0.5 | DOC | 69.73 | 69.71 | 65.41 | 33.83 | 64.68 | 78.17 | 73.14 | 68.69 | 54.33 | 67.21 | 53.07 | 52.85 | 62.57 |
| | | DeepUNK | 71.12 | 61.03 | 66.89 | 32.62 | 31.21 | 70.13 | 29.93 | 32.60 | 0.38 | 19.40 | 10.11 | 28.31 | 37.81 |
| | | ADB | 56.37 | **79.39** | 42.02 | 26.73 | 60.38 | **85.94** | 71.25 | 67.82 | 44.19 | 43.98 | **68.71** | 29.96 | 56.39 |
| | | SCL | **98.62** | 63.88 | 48.79 | 48.22 | 44.74 | 75.48 | 45.37 | 45.47 | 2.98 | 28.27 | 44.55 | 39.06 | 48.79 |
| | | AB | 65.25 | 47.69 | 56.91 | 0.00 | 72.48 | 72.21 | 76.81 | **80.66** | 63.28 | 68.60 | 64.08 | **55.92** | 60.32 |
| | | KNNCon | 85.82 | 57.97 | 29.83 | 54.24 | 58.30 | 73.55 | 78.45 | 62.84 | **73.14** | **69.79** | 63.04 | 37.24 | 62.02 |
| | | DyEn | 89.13 | 64.49 | 23.46 | **73.75** | 63.22 | 58.45 | 78.35 | 70.60 | 69.66 | 69.69 | 65.14 | 16.55 | 61.87 |
| | | LLM-MaxSoftmax | 48.51 | 43.57 | 38.10 | 56.61 | 56.40 | 49.55 | 63.44 | 63.48 | 63.35 | 53.38 | 60.50 | 39.43 | 53.03 |
| | | LLM-OpenMax | 73.30 | 64.97 | 59.43 | 67.33 | 67.97 | 71.44 | 69.32 | 67.80 | 67.13 | 55.78 | 65.25 | 52.89 | 65.22 |
| | | LLM-TempScale | 48.44 | 43.58 | 38.09 | 51.80 | 55.47 | 48.64 | 62.38 | 60.75 | 61.07 | 55.01 | 60.05 | 40.41 | 52.14 |
| | | LLM-Energy | 87.08 | 72.65 | **67.47** | 66.38 | 72.89 | 84.11 | 79.18 | 70.39 | 66.63 | 53.58 | 65.83 | 55.46 | 70.14 |
| | | LLM-LogitNorm | 55.72 | 30.08 | 19.46 | 63.07 | 72.40 | 57.62 | **82.14** | 71.28 | 63.65 | 37.53 | 65.11 | 48.53 | 55.55 |
| | | LLM-Entropy | 52.18 | 48.13 | 43.47 | 57.66 | 55.22 | 52.88 | 60.51 | 64.64 | 65.85 | 57.62 | 59.99 | 39.38 | 54.79 |
| | | LLM-KLMatching | 36.09 | 40.37 | 34.44 | 31.68 | 45.42 | 33.33 | 47.61 | 46.93 | 12.80 | 14.96 | 35.35 | 16.78 | 32.98 |
| | | LLM-MaxLogit | 88.91 | 72.95 | 67.21 | 65.88 | **73.87** | 85.44 | 81.54 | 71.44 | 66.74 | 53.90 | 65.62 | 53.56 | **70.59** |
| | 0.75 | DOC | 48.08 | 48.12 | 34.10 | 29.62 | 34.79 | 61.14 | 40.43 | 42.60 | 42.27 | 40.74 | 23.47 | 31.21 | 39.71 |
| | | DeepUNK | 53.29 | 44.76 | 34.92 | 21.72 | 36.25 | 45.42 | 37.15 | 8.28 | 12.25 | 1.42 | 11.85 | 19.49 | 27.23 |
| | | ADB | 57.34 | **69.60** | 33.76 | 22.20 | 49.22 | **71.32** | 62.86 | 55.49 | 38.73 | 26.80 | 59.96 | **40.97** | 49.02 |
| | | SCL | **95.89** | 43.06 | 30.27 | 35.52 | 28.85 | 55.99 | 37.20 | 30.93 | 3.54 | 27.31 | 28.86 | 23.56 | 36.75 |
| | | AB | 45.86 | 34.17 | 34.56 | 0.00 | 51.08 | 53.12 | 58.06 | **60.10** | 43.60 | 44.85 | 36.59 | 30.46 | 41.04 |
| | | KNNCon | 87.78 | 54.71 | 30.78 | **52.48** | 51.64 | 67.16 | 71.44 | 53.67 | **69.90** | 56.74 | 55.24 | 18.21 | **55.81** |
| | | DyEn | 94.81 | 50.18 | 20.38 | 47.97 | **60.39** | 56.02 | **75.52** | 24.08 | 55.61 | **65.44** | **60.76** | 12.59 | 51.98 |
| | | LLM-MaxSoftmax | 50.18 | 38.81 | 30.88 | 37.76 | 43.94 | 45.22 | 56.29 | 47.77 | 45.84 | 36.25 | 38.92 | 25.93 | 41.48 |
| | | LLM-OpenMax | 50.67 | 49.54 | 35.60 | 36.99 | 40.87 | 61.04 | 44.49 | 40.49 | 43.85 | 39.15 | 39.01 | 31.23 | 42.74 |
| | | LLM-TempScale | 47.79 | 38.81 | 30.79 | 37.87 | 43.19 | 45.20 | 56.24 | 47.75 | 44.83 | 35.92 | 38.70 | 26.98 | 41.17 |
| | | LLM-Energy | 69.91 | 55.46 | 37.56 | 39.08 | 45.28 | 65.58 | 59.60 | 46.14 | 44.86 | 41.71 | 40.09 | 31.62 | 48.07 |
| | | LLM-LogitNorm | 67.64 | 44.16 | 27.55 | 43.03 | 48.34 | 66.88 | 66.20 | 48.79 | 47.77 | 38.15 | 40.31 | 29.97 | 47.40 |
| | | LLM-Entropy | 52.50 | 43.03 | 32.73 | 39.55 | 42.48 | 43.38 | 52.96 | 50.67 | 47.13 | 38.76 | 38.44 | 24.31 | 42.16 |
| | | LLM-KLMatching | 39.71 | 35.38 | 28.97 | 31.15 | 37.50 | 28.14 | 50.03 | 42.00 | 21.41 | 29.49 | 30.25 | 14.94 | 32.41 |
| | | LLM-MaxLogit | 75.11 | 57.06 | **38.04** | 40.39 | 47.12 | 68.55 | 64.23 | 48.01 | 45.25 | 41.69 | 40.38 | 31.34 | 49.76 |

Table S7: N-F1 performance of OSTC on the BOLT benchmark under different KCRs (LAR = 0.1)

| Metric | KCR | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| N-F1 | 0.25 | DOC | 94.02 | 81.51 | 70.17 | 84.29 | 86.66 | 78.54 | 92.07 | 80.71 | 86.43 | 80.20 | 89.10 | **80.77** | 83.71 |
| | | DeepUNK | 95.68 | 67.52 | 76.38 | 64.38 | 89.36 | 89.29 | 92.83 | 91.20 | 86.47 | **89.37** | 66.42 | 36.75 | 78.80 |
| | | ADB | 73.34 | 82.65 | 55.79 | 59.46 | 82.93 | 89.54 | 90.90 | 79.11 | 76.61 | 53.84 | 84.76 | 40.78 | 72.48 |
| | | SCL | 89.34 | 66.06 | 61.80 | 83.43 | 82.00 | 84.57 | 82.11 | 81.38 | 8.68 | 86.05 | 69.84 | 31.53 | 68.90 |
| | | AB | 78.72 | 54.78 | 73.12 | 14.72 | 81.22 | 83.46 | 87.58 | 88.62 | 73.06 | 83.92 | 79.34 | 68.34 | 72.24 |
| | | KNNCon | 86.76 | 35.37 | 31.58 | 50.99 | 62.03 | 63.39 | 83.92 | 60.64 | 55.41 | 46.14 | 65.99 | 26.05 | 55.69 |
| | | DyEn | **96.15** | 68.63 | 22.70 | 62.72 | 71.49 | 55.36 | 86.43 | 60.69 | 69.05 | 70.44 | 71.29 | 36.17 | 64.26 |
| | | LLM-MaxSoftmax | 38.96 | 28.00 | 19.02 | 45.31 | 39.03 | 33.96 | 49.96 | 43.33 | 56.54 | 37.17 | 62.65 | 41.66 | 41.30 |
| | | LLM-OpenMax | 83.31 | 59.20 | 25.23 | 73.42 | 89.29 | 53.46 | 92.85 | 87.96 | 67.50 | 60.64 | 86.97 | 68.01 | 70.65 |
| | | LLM-TempScale | 38.94 | 28.00 | 19.36 | 43.13 | 38.02 | 35.45 | 49.96 | 43.30 | 53.30 | 34.95 | 62.61 | 44.03 | 40.92 |
| | | LLM-Energy | 94.50 | **90.32** | **82.24** | **84.65** | **90.82** | 93.55 | **95.09** | **92.59** | **90.36** | 81.51 | **92.51** | 71.59 | **88.31** |
| | | LLM-LogitNorm | 33.46 | 24.96 | 30.62 | 64.73 | 77.54 | 36.05 | 90.95 | 75.43 | 56.90 | 36.05 | 90.96 | 53.65 | 55.94 |
| | | LLM-Entropy | 43.77 | 33.87 | 32.36 | 48.39 | 35.67 | 38.40 | 46.11 | 43.12 | 61.42 | 49.15 | 59.32 | 40.12 | 44.31 |
| | | LLM-KLMatching | 24.88 | 17.55 | 15.48 | 29.05 | 25.16 | 19.23 | 37.97 | 32.80 | 40.96 | 29.80 | 47.83 | 28.35 | 29.09 |
| | | LLM-MaxLogit | 93.85 | 90.19 | 81.72 | 83.44 | 90.37 | **93.83** | 94.69 | 92.51 | 90.21 | 80.95 | 92.38 | 68.92 | 87.75 |
| | 0.5 | DOC | 83.28 | 71.70 | 68.52 | 74.14 | 76.75 | 74.55 | 83.36 | 76.94 | 66.33 | 74.57 | 69.60 | 55.06 | 72.90 |
| | | DeepUNK | 90.17 | 39.30 | 67.66 | 67.12 | 77.85 | 77.38 | 82.74 | 80.36 | 51.21 | 74.71 | 57.32 | 51.00 | 68.07 |
| | | ADB | 76.23 | 83.88 | 28.08 | 66.42 | 76.40 | 87.77 | 83.97 | 74.95 | **82.28** | 68.07 | 80.55 | 20.40 | 69.08 |
| | | SCL | 88.83 | 50.58 | 54.98 | 73.20 | 57.09 | 56.15 | 73.54 | 57.40 | 2.97 | 74.90 | 54.48 | 42.53 | 57.22 |
| | | AB | 63.58 | 48.64 | 55.76 | 0.00 | 68.74 | 71.26 | 75.96 | 77.42 | 56.44 | 68.52 | 65.04 | 49.76 | 58.43 |
| | | KNNCon | 87.17 | 48.60 | 26.39 | 51.05 | 65.09 | 74.97 | 77.61 | 62.90 | 74.70 | 63.75 | 58.28 | 30.98 | 60.12 |
| | | DyEn | **93.87** | 65.15 | 21.31 | 58.22 | 63.87 | 56.10 | 80.26 | 58.36 | 65.74 | 66.54 | 63.00 | 18.55 | 59.25 |
| | | LLM-MaxSoftmax | 44.00 | 36.96 | 43.04 | 48.17 | 37.94 | 34.12 | 48.99 | 49.01 | 51.03 | 46.76 | 57.48 | 28.37 | 43.82 |
| | | LLM-OpenMax | 89.94 | 75.50 | 19.85 | 75.53 | 75.46 | 82.10 | 82.78 | 79.24 | 76.74 | 62.79 | 70.76 | **56.10** | 70.57 |
| | | LLM-TempScale | 44.00 | 37.47 | 41.57 | 46.93 | 37.50 | 34.12 | 49.62 | 47.67 | 50.41 | 44.91 | 57.49 | 26.60 | 43.19 |
| | | LLM-Energy | 93.86 | 83.49 | 72.79 | 78.13 | 81.61 | 88.40 | 89.23 | 82.62 | 78.43 | 77.46 | 83.54 | 54.37 | 80.33 |
| | | LLM-LogitNorm | 65.67 | 42.44 | 24.66 | 76.59 | 80.41 | 58.49 | 90.10 | 80.56 | 72.93 | 59.55 | **87.73** | 43.53 | 65.22 |
| | | LLM-Entropy | 48.54 | 39.44 | 48.35 | 47.86 | 30.94 | 32.71 | 42.25 | 47.77 | 50.25 | 53.04 | 52.89 | 23.65 | 43.14 |
| | | LLM-KLMatching | 30.36 | 24.18 | 8.35 | 33.12 | 24.06 | 27.86 | 30.99 | 32.84 | 43.72 | 37.34 | 43.77 | 10.04 | 28.89 |
| | | LLM-MaxLogit | 93.53 | **84.00** | **73.24** | **78.53** | **82.23** | **88.93** | **90.65** | **83.77** | 79.72 | **78.45** | 86.21 | 53.27 | **81.04** |
| | 0.75 | DOC | 67.11 | 49.59 | 38.25 | 47.66 | 54.49 | 62.71 | 65.32 | 57.74 | 42.27 | 53.04 | 39.29 | 30.01 | 50.62 |
| | | DeepUNK | 83.46 | 31.06 | 38.49 | 36.45 | 56.60 | 56.02 | 64.94 | 59.36 | 42.28 | 50.33 | 47.86 | 19.15 | 48.83 |
| | | ADB | 70.77 | **72.24** | 31.05 | 51.24 | 65.16 | 73.65 | 73.82 | 62.70 | **68.72** | 56.32 | **72.68** | 30.12 | 60.71 |
| | | SCL | 85.13 | 34.04 | 26.71 | 40.89 | 44.73 | 41.11 | 58.92 | 41.89 | 3.37 | 54.14 | 40.01 | 31.21 | 41.85 |
| | | AB | 46.38 | 36.10 | 32.72 | 0.00 | 51.28 | 52.88 | 59.46 | 59.82 | 41.38 | 50.52 | 45.70 | 27.08 | 41.94 |
| | | KNNCon | 85.92 | 50.00 | 9.10 | 50.99 | 54.32 | 63.74 | 71.90 | 50.84 | 63.35 | 58.53 | 55.48 | 23.17 | 53.11 |
| | | DyEn | **89.89** | 55.23 | 21.41 | 45.75 | 56.65 | 49.96 | 72.09 | 45.93 | 60.56 | **59.80** | 58.40 | 12.81 | 52.37 |
| | | LLM-MaxSoftmax | 47.42 | 33.50 | 44.90 | 38.39 | 38.61 | 43.57 | 43.10 | 42.55 | 47.92 | 45.65 | 53.36 | 23.35 | 41.86 |
| | | LLM-OpenMax | 84.93 | 67.44 | 0.20 | 47.88 | 52.91 | **75.49** | 64.20 | 58.81 | 52.39 | 50.05 | 47.06 | **32.72** | 52.84 |
| | | LLM-TempScale | 47.42 | 34.08 | 45.22 | 38.40 | 37.72 | 45.23 | 44.88 | 41.91 | 47.67 | 42.83 | 55.04 | 20.59 | 41.75 |
| | | LLM-Energy | 87.04 | 69.23 | 39.21 | 51.44 | 61.48 | 69.95 | 70.63 | 61.68 | 54.67 | 48.86 | 66.69 | 31.31 | 59.35 |
| | | LLM-LogitNorm | 79.96 | 54.59 | 40.19 | **59.82** | **66.16** | 68.46 | **79.92** | **66.01** | 65.07 | 56.11 | 70.47 | 28.87 | 61.30 |
| | | LLM-Entropy | 50.24 | 35.70 | **49.11** | 35.04 | 30.11 | 43.12 | 36.43 | 39.20 | 49.47 | 50.22 | 50.69 | 21.89 | 40.94 |
| | | LLM-KLMatching | 39.40 | 25.25 | 7.91 | 30.50 | 26.31 | 20.80 | 31.87 | 30.71 | 38.66 | 37.93 | 39.65 | 6.09 | 27.92 |
| | | LLM-MaxLogit | 87.85 | 71.01 | 41.76 | 53.25 | 63.90 | 72.28 | 75.63 | 64.28 | 57.01 | 50.26 | 71.70 | 31.32 | **61.69** |

Table S8: N-F1 performance of OSTC on the BOLT benchmark under different KCRs (LAR = 0.5)

| Metric | KCR | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| N-F1 | 0.25 | DOC | 94.62 | 74.28 | 67.86 | 80.49 | 85.58 | 74.72 | 92.33 | 82.63 | 86.80 | 78.04 | 91.23 | **78.43** | 82.25 |
| | | DeepUNK | 97.97 | 39.64 | 72.71 | 86.41 | 89.85 | 89.51 | 93.51 | **91.98** | 86.85 | **89.29** | 74.53 | 60.68 | 81.08 |
| | | ADB | 73.88 | 82.42 | 61.55 | 74.29 | 83.77 | 90.36 | 91.91 | 82.39 | 86.49 | 67.72 | 87.20 | 49.48 | 77.62 |
| | | SCL | 91.19 | 60.92 | 70.96 | 84.06 | 79.51 | 69.14 | 78.95 | 68.43 | 5.82 | 85.82 | 71.75 | 43.47 | 67.50 |
| | | AB | 79.68 | 55.50 | 75.30 | 48.78 | 81.83 | 83.69 | 87.66 | 87.46 | 71.63 | 84.51 | 79.25 | 67.72 | 75.25 |
| | | KNNCon | 89.65 | 59.71 | 31.13 | 59.80 | 67.29 | 64.66 | 85.76 | 64.15 | 55.56 | 46.22 | 72.56 | 33.56 | 60.84 |
| | | DyEn | **98.27** | 68.01 | 19.53 | 48.99 | 72.27 | 56.98 | 89.28 | 74.71 | 79.13 | 68.28 | 71.87 | 36.17 | 65.29 |
| | | LLM-MaxSoftmax | 35.04 | 21.80 | 23.40 | 39.80 | 37.50 | 28.33 | 48.57 | 38.29 | 52.27 | 36.37 | 56.22 | 43.69 | 38.44 |
| | | LLM-OpenMax | 85.88 | 67.29 | 25.70 | 77.46 | 90.90 | 71.30 | 94.74 | 88.45 | 82.84 | 60.98 | 90.71 | 66.78 | 75.25 |
| | | LLM-TempScale | 33.99 | 22.20 | 23.87 | 39.32 | 37.33 | 28.64 | 48.60 | 37.71 | 48.06 | 34.36 | 56.75 | 45.13 | 38.00 |
| | | LLM-Energy | 93.16 | **89.45** | **84.95** | **88.18** | **91.54** | **94.73** | **96.16** | 91.98 | 92.56 | 86.68 | **94.63** | 71.34 | **89.61** |
| | | LLM-LogitNorm | 37.02 | 14.36 | 43.05 | 66.98 | 73.93 | 21.51 | 90.22 | 69.97 | 45.00 | 31.33 | 92.26 | 57.43 | 53.59 |
| | | LLM-Entropy | 39.34 | 27.30 | 38.74 | 42.30 | 30.70 | 30.99 | 43.64 | 36.81 | 56.24 | 45.66 | 53.13 | 41.74 | 40.55 |
| | | LLM-KLMatching | 21.22 | 12.38 | 16.77 | 29.81 | 21.35 | 20.79 | 32.93 | 28.38 | 29.62 | 24.75 | 40.09 | 14.81 | 24.41 |
| | | LLM-MaxLogit | 92.68 | 89.02 | 84.89 | 87.62 | 91.21 | 94.64 | 96.07 | 91.60 | **92.69** | 86.25 | 94.59 | 68.51 | 89.15 |
| | 0.5 | DOC | 86.06 | 71.17 | 67.26 | 70.04 | 78.82 | 72.34 | 85.25 | 76.74 | 66.33 | 77.22 | 76.89 | 53.24 | 73.45 |
| | | DeepUNK | 96.09 | 31.88 | 62.84 | 68.70 | 78.49 | 79.72 | 84.83 | 82.31 | 66.35 | 78.10 | 67.07 | 46.29 | 70.22 |
| | | ADB | 78.76 | 83.76 | 33.12 | 73.47 | 79.72 | **88.26** | 86.42 | 74.93 | **82.44** | 73.53 | 81.15 | 18.23 | 71.15 |
| | | SCL | 81.16 | 49.80 | 44.12 | 65.43 | 61.80 | 53.43 | 70.24 | 53.78 | 2.32 | 75.32 | 52.83 | **57.47** | 55.64 |
| | | AB | 63.34 | 47.90 | 55.80 | 29.88 | 67.64 | 70.70 | 75.80 | 75.02 | 57.34 | 68.76 | 62.14 | 48.70 | 60.25 |
| | | KNNCon | 88.16 | 65.60 | 35.32 | 55.04 | 54.20 | 71.50 | 77.67 | 57.00 | 76.29 | 59.16 | 63.42 | 32.61 | 61.33 |
| | | DyEn | 97.92 | 66.27 | 18.21 | 53.70 | 64.22 | 55.15 | 79.54 | 55.02 | 69.66 | 70.57 | 63.83 | 18.16 | 59.35 |
| | | LLM-MaxSoftmax | 32.88 | 29.28 | 53.29 | 39.70 | 34.44 | 27.32 | 40.32 | 35.37 | 53.28 | 40.84 | 50.85 | 31.19 | 39.06 |
| | | LLM-OpenMax | 92.10 | 79.56 | 5.87 | 74.24 | 81.18 | 78.59 | 88.04 | 80.19 | 80.89 | 74.44 | 77.60 | 48.13 | 71.74 |
| | | LLM-TempScale | 32.88 | 29.28 | 52.00 | 39.70 | 34.44 | 27.97 | 40.34 | 35.37 | 51.87 | 40.21 | 51.48 | 29.56 | 38.76 |
| | | LLM-Energy | **98.06** | 86.82 | 74.53 | 80.09 | 83.96 | 87.56 | 90.95 | 85.44 | 81.12 | 78.59 | 86.77 | 50.55 | 82.04 |
| | | LLM-LogitNorm | 66.10 | 45.27 | 29.71 | 76.87 | 80.74 | 51.27 | 90.62 | 80.69 | 79.74 | 63.66 | **88.55** | 46.88 | 66.68 |
| | | LLM-Entropy | 35.86 | 29.13 | 63.38 | 34.60 | 25.02 | 24.84 | 34.17 | 30.05 | 54.00 | 46.44 | 44.73 | 24.60 | 37.24 |
| | | LLM-KLMatching | 24.40 | 23.48 | 11.45 | 33.75 | 25.60 | 22.29 | 27.84 | 26.80 | 36.77 | 33.03 | 32.42 | 7.66 | 25.46 |
| | | LLM-MaxLogit | 96.00 | **86.88** | **75.47** | **80.87** | **84.52** | 87.85 | **91.72** | **86.25** | 82.32 | **79.50** | 88.26 | 47.39 | **82.25** |
| | 0.75 | DOC | 71.05 | 50.47 | 38.87 | 47.52 | 59.92 | 63.45 | 70.25 | 61.10 | 42.27 | 56.72 | 53.92 | 28.62 | 53.68 |
| | | DeepUNK | 91.81 | 28.15 | 36.45 | 37.29 | 60.91 | 60.71 | 68.49 | 63.89 | 42.35 | 55.18 | 48.50 | 24.00 | 51.48 |
| | | ADB | 70.98 | **73.78** | 30.79 | 54.80 | 67.71 | 74.56 | 77.76 | 62.32 | **72.91** | 58.99 | 72.87 | 6.71 | 60.35 |
| | | SCL | 86.53 | 33.79 | 26.03 | 39.99 | 45.97 | 40.77 | 60.02 | 44.01 | 3.51 | 58.25 | 44.34 | 25.19 | 42.37 |
| | | AB | 46.14 | 35.28 | 32.96 | 5.12 | 50.62 | 52.54 | 58.98 | 57.96 | 41.78 | 50.62 | 44.32 | 26.70 | 41.92 |
| | | KNNCon | 87.47 | 60.68 | 27.07 | 51.54 | 52.96 | 64.47 | 74.13 | 52.22 | 65.82 | **61.92** | 49.87 | 22.64 | 55.90 |
| | | DyEn | 87.27 | 51.22 | 23.03 | 35.09 | 56.05 | 47.53 | 72.76 | 52.82 | 55.61 | 58.31 | 54.69 | 12.85 | 50.60 |
| | | LLM-MaxSoftmax | 57.75 | 30.79 | 45.66 | 31.06 | 35.89 | 31.19 | 36.72 | 34.61 | 48.11 | 44.17 | 49.10 | 18.51 | 38.63 |
| | | LLM-OpenMax | 92.82 | 57.69 | 7.17 | 48.92 | 61.38 | **75.77** | 76.45 | 65.60 | 59.30 | 59.30 | 56.03 | **38.09** | 58.21 |
| | | LLM-TempScale | 57.75 | 31.51 | 45.02 | 31.06 | 35.95 | 31.21 | 38.49 | 34.63 | 47.64 | 43.42 | 49.53 | 17.75 | 38.66 |
| | | LLM-Energy | **98.03** | 72.16 | 39.77 | 55.30 | 62.49 | 70.24 | 74.89 | 63.37 | 61.31 | 56.29 | 70.46 | 31.90 | 63.02 |
| | | LLM-LogitNorm | 81.65 | 59.99 | 38.70 | **63.66** | **69.23** | 63.40 | **82.64** | **67.14** | 71.18 | 60.75 | **75.62** | 27.94 | 63.49 |
| | | LLM-Entropy | 61.27 | 32.37 | **50.71** | 27.62 | 29.39 | 27.53 | 28.40 | 31.13 | 48.04 | 45.81 | 41.52 | 17.80 | 36.80 |
| | | LLM-KLMatching | 50.32 | 20.32 | 9.82 | 23.98 | 21.65 | 21.35 | 22.66 | 21.57 | 37.74 | 32.15 | 34.55 | 9.29 | 25.45 |
| | | LLM-MaxLogit | 96.98 | 73.59 | 42.20 | 57.73 | 65.17 | 71.98 | 77.90 | 65.96 | 64.38 | 57.99 | 74.79 | 31.79 | **65.04** |

Table S9: N-F1 performance of OSTC on the BOLT benchmark under different KCRs (LAR = 1.0)

| Metric | KCR | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| F1 | 0.25 | DOC | 37.47 | 44.20 | 36.90 | 34.15 | 4.15 | 61.98 | 2.19 | 5.05 | 8.00 | 17.20 | 1.43 | 7.46 | 21.68 |
| | | DeepUNK | 42.62 | 23.96 | 45.51 | 28.93 | 35.82 | 35.54 | 45.63 | 45.20 | 3.06 | 33.61 | 27.81 | 6.04 | 31.14 |
| | | ADB | 44.46 | **63.83** | 40.50 | 20.04 | 35.22 | **73.28** | 69.74 | 50.39 | 38.25 | 39.00 | 40.99 | 17.07 | 44.40 |
| | | SCL | **96.98** | 26.80 | 30.68 | 12.77 | 19.52 | 42.07 | 24.06 | 27.76 | 1.83 | 23.10 | 10.29 | 7.11 | 26.91 |
| | | AB | 57.65 | 29.56 | **55.04** | 27.43 | 58.62 | 59.96 | 68.07 | 62.64 | 32.23 | 49.72 | 26.50 | 11.50 | 44.91 |
| | | KNNCon | 88.95 | 58.70 | 39.02 | 50.91 | 63.27 | 59.73 | **77.07** | **71.60** | 61.06 | 55.74 | 62.03 | 17.37 | 58.79 |
| | | DyEn | 94.15 | 63.48 | 34.44 | **60.40** | **63.55** | 50.29 | 76.81 | 66.48 | **77.39** | **67.69** | **66.70** | **20.32** | **61.81** |
| | | LLM-MaxSoftmax | 26.01 | 39.21 | 30.07 | 10.19 | 36.50 | 44.55 | 48.03 | 34.87 | 21.27 | 18.90 | 7.38 | 7.36 | 27.03 |
| | | LLM-OpenMax | 23.87 | 40.41 | 21.95 | 8.77 | 11.18 | 55.16 | 15.53 | 11.95 | 18.84 | 20.32 | 2.05 | 9.44 | 19.96 |
| | | LLM-TempScale | 26.46 | 38.99 | 28.87 | 9.78 | 36.34 | 43.36 | 47.66 | 33.86 | 20.28 | 19.20 | 7.23 | 8.12 | 26.68 |
| | | LLM-Energy | 32.49 | 46.11 | 39.80 | 11.51 | 34.75 | 65.63 | 45.20 | 31.72 | 16.21 | 20.67 | 4.70 | 7.98 | 29.73 |
| | | LLM-LogitNorm | 23.94 | 33.71 | 32.53 | 10.47 | 41.07 | 36.01 | 58.78 | 36.45 | 17.71 | 17.11 | 6.01 | 7.46 | 26.77 |
| | | LLM-Entropy | 27.78 | 43.36 | 37.79 | 10.37 | 36.80 | 48.87 | 47.95 | 35.54 | 22.98 | 19.32 | 7.44 | 7.92 | 28.84 |
| | | LLM-KLMatching | 22.83 | 37.60 | 26.31 | 6.96 | 33.20 | 38.56 | 43.05 | 32.08 | 16.97 | 14.53 | 6.89 | 7.08 | 23.84 |
| | | LLM-MaxLogit | 32.82 | 46.97 | 40.28 | 11.48 | 39.68 | 67.51 | 51.56 | 34.84 | 17.85 | 20.85 | 5.81 | 7.74 | 31.45 |
| | 0.5 | DOC | 23.78 | 41.61 | 14.74 | 2.42 | 1.74 | 69.19 | 43.43 | 2.08 | 3.59 | 7.47 | 0.48 | 2.35 | 17.74 |
| | | DeepUNK | 34.07 | 18.13 | 20.70 | 27.93 | 48.54 | 30.40 | 59.58 | 57.26 | 1.95 | 47.44 | 27.63 | 5.18 | 31.57 |
| | | ADB | 69.17 | **81.00** | **53.73** | 23.79 | 57.92 | **82.87** | 72.83 | 59.40 | 47.35 | 41.63 | 58.24 | 16.47 | 55.37 |
| | | SCL | **98.49** | 20.68 | 27.20 | 14.67 | 23.14 | 50.04 | 32.27 | 27.81 | 1.48 | 29.11 | 8.45 | 6.45 | 28.32 |
| | | AB | 54.82 | 28.46 | 45.80 | 18.48 | 61.32 | 64.05 | 70.20 | 65.31 | 29.58 | 46.54 | 24.35 | 16.62 | 43.79 |
| | | KNNCon | 92.03 | 75.38 | 53.36 | **54.92** | 77.18 | 79.73 | **86.82** | **80.97** | 80.53 | **78.08** | 66.11 | **21.40** | **70.54** |
| | | DyEn | 93.34 | 74.48 | 49.88 | 52.34 | **77.26** | 72.93 | 85.84 | 77.70 | **80.87** | 75.48 | **72.05** | 17.54 | 69.14 |
| | | LLM-MaxSoftmax | 75.91 | 52.82 | 42.78 | 23.16 | 57.63 | 69.71 | 75.14 | 51.61 | 26.42 | 13.51 | 10.81 | 8.06 | 42.30 |
| | | LLM-OpenMax | 68.58 | 39.14 | 38.33 | 11.71 | 9.15 | 71.71 | 22.83 | 12.16 | 9.75 | 15.62 | 1.16 | 4.96 | 25.43 |
| | | LLM-TempScale | 75.89 | 52.84 | 42.79 | 22.85 | 57.40 | 69.40 | 74.79 | 51.52 | 26.77 | 13.46 | 10.83 | 8.63 | 42.26 |
| | | LLM-Energy | 81.41 | 47.84 | 37.42 | 13.80 | 41.84 | 75.71 | 62.87 | 27.22 | 13.48 | 10.92 | 4.08 | 5.61 | 35.18 |
| | | LLM-LogitNorm | 77.46 | 48.95 | 38.28 | 21.56 | 55.43 | 71.68 | 73.81 | 43.88 | 24.19 | 13.56 | 7.07 | 6.40 | 40.19 |
| | | LLM-Entropy | 77.06 | 54.30 | 44.53 | 22.30 | 57.45 | 70.82 | 74.49 | 52.94 | 26.27 | 12.60 | 10.93 | 8.13 | 42.65 |
| | | LLM-KLMatching | 72.41 | 52.14 | 42.07 | 21.59 | 55.85 | 65.47 | 71.57 | 51.15 | 22.78 | 12.48 | 10.69 | 8.75 | 40.58 |
| | | LLM-MaxLogit | 85.43 | 50.47 | 38.98 | 15.57 | 49.59 | 79.29 | 70.50 | 33.48 | 16.40 | 11.28 | 5.72 | 6.22 | 38.58 |
| | 0.75 | DOC | 32.93 | 29.11 | 8.86 | 23.36 | 0.64 | 72.14 | 0.36 | 0.87 | 1.76 | 3.13 | 0.14 | 0.85 | 14.51 |
| | | DeepUNK | 52.14 | 10.52 | 16.45 | 18.38 | 49.74 | 33.54 | 63.39 | 58.98 | 1.74 | 58.00 | 22.15 | 5.69 | 32.56 |
| | | ADB | 86.54 | **83.29** | 59.61 | 24.78 | 63.42 | 83.38 | 80.86 | 65.66 | 51.02 | 47.07 | 65.91 | 14.10 | 60.47 |
| | | SCL | **98.24** | 17.23 | 18.08 | 15.86 | 23.79 | 63.17 | 55.05 | 41.28 | 1.22 | 27.29 | 9.20 | 6.46 | 31.41 |
| | | AB | 56.52 | 25.80 | 46.80 | 15.46 | 62.51 | 65.75 | 71.69 | 64.65 | 29.04 | 43.29 | 22.52 | 14.10 | 43.18 |
| | | KNNCon | 97.09 | 82.56 | **62.76** | **72.03** | 86.01 | **86.08** | **92.66** | **85.83** | 87.45 | 82.57 | 73.62 | **20.59** | **77.44** |
| | | DyEn | 95.53 | 80.96 | 60.71 | 62.38 | **86.51** | 83.94 | 92.24 | 81.74 | **88.83** | **83.29** | **79.31** | 17.57 | 76.08 |
| | | LLM-MaxSoftmax | 89.07 | 69.64 | 53.10 | 28.78 | 65.56 | 81.68 | 86.06 | 61.22 | 33.14 | 36.70 | 12.68 | 9.40 | 52.25 |
| | | LLM-OpenMax | 51.53 | 49.37 | 36.60 | 9.10 | 11.74 | 72.56 | 31.02 | 10.27 | 6.08 | 11.19 | 1.41 | 3.10 | 24.50 |
| | | LLM-TempScale | 88.65 | 69.64 | 53.08 | 28.63 | 65.55 | 81.67 | 86.05 | 61.18 | 33.48 | 36.76 | 12.88 | 9.66 | 52.27 |
| | | LLM-Energy | 80.81 | 53.90 | 30.34 | 14.42 | 34.86 | 76.13 | 67.93 | 32.82 | 10.59 | 19.32 | 4.09 | 4.08 | 35.77 |
| | | LLM-LogitNorm | 91.75 | 70.00 | 51.46 | 21.84 | 49.53 | 85.51 | 78.36 | 46.21 | 27.33 | 30.83 | 6.28 | 7.50 | 47.22 |
| | | LLM-Entropy | 89.57 | 70.67 | 53.66 | 29.48 | 65.52 | 81.38 | 85.54 | 62.54 | 32.57 | 36.55 | 13.12 | 9.55 | 52.51 |
| | | LLM-KLMatching | 87.44 | 69.33 | 52.98 | 28.78 | 65.47 | 78.99 | 85.34 | 61.34 | 33.67 | 37.99 | 14.04 | 10.44 | 52.15 |
| | | LLM-MaxLogit | 86.74 | 58.45 | 33.37 | 17.23 | 43.69 | 79.65 | 75.41 | 40.20 | 13.07 | 22.92 | 6.40 | 5.25 | 40.20 |

Table S10: F1 performance of OSTC on the BOLT benchmark under different KCRs (LAR = 0.1)

| Metric | KCR | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| F1 | 0.25 | DOC | 77.73 | 51.34 | 49.20 | 33.15 | 62.24 | 61.82 | 77.58 | 66.03 | 9.60 | 61.16 | 39.00 | 14.90 | 50.31 |
| | | DeepUNK | 84.39 | 26.79 | 54.21 | 27.60 | 65.13 | 61.47 | 71.45 | 66.13 | 10.86 | 59.37 | 30.66 | 11.20 | 47.44 |
| | | ADB | 70.01 | 75.47 | 50.24 | 46.79 | 67.86 | 78.83 | 80.43 | 72.29 | 62.50 | 51.15 | **74.39** | 21.05 | 62.58 |
| | | SCL | 71.72 | 29.35 | 32.41 | 33.97 | 23.51 | 48.05 | 40.52 | 32.93 | 3.05 | 48.10 | 23.32 | 10.78 | 33.14 |
| | | AB | 57.66 | 31.70 | 52.38 | 26.78 | 59.34 | 63.32 | 71.26 | 68.62 | 43.76 | 63.40 | 45.98 | 13.88 | 49.84 |
| | | KNNCon | 85.73 | 50.77 | 38.76 | 39.71 | 62.26 | 60.90 | 76.59 | 71.66 | 59.80 | 57.24 | 56.94 | 17.20 | 56.46 |
| | | DyEn | **94.32** | 64.65 | 33.70 | 44.19 | 63.87 | 55.62 | 76.79 | 68.31 | **65.43** | **65.57** | 63.56 | **24.21** | 60.02 |
| | | LLM-MaxSoftmax | 51.26 | 44.50 | 33.34 | 46.69 | 48.84 | 45.09 | 56.56 | 59.84 | 47.68 | 45.31 | 50.37 | 14.80 | 45.36 |
| | | LLM-OpenMax | 80.43 | 58.79 | 35.28 | 36.82 | 58.08 | 56.42 | 71.07 | 63.04 | 40.56 | 45.93 | 19.40 | 15.95 | 48.48 |
| | | LLM-TempScale | 51.26 | 44.50 | 33.47 | 45.97 | 48.52 | 45.72 | 56.56 | 59.84 | 46.43 | 44.48 | 50.36 | 15.11 | 45.19 |
| | | LLM-Energy | 89.14 | 77.97 | 60.52 | 43.66 | 70.74 | 77.43 | 85.61 | 73.99 | 54.38 | 60.52 | 53.49 | 10.72 | 63.18 |
| | | LLM-LogitNorm | 50.13 | 43.15 | 37.73 | 47.94 | 65.89 | 47.73 | 82.90 | 69.53 | 47.18 | 44.46 | 59.87 | 14.45 | 50.91 |
| | | LLM-Entropy | 53.55 | 46.66 | 39.60 | **48.33** | 48.00 | 47.09 | 55.40 | 60.37 | 49.69 | 50.49 | 50.01 | 14.64 | 46.99 |
| | | LLM-KLMatching | 46.47 | 40.92 | 31.87 | 41.14 | 45.18 | 40.30 | 52.91 | 57.54 | 43.29 | 42.87 | 46.33 | 14.76 | 41.96 |
| | | LLM-MaxLogit | 88.62 | **78.97** | **60.82** | 44.45 | **72.70** | **79.43** | **86.49** | **75.54** | 57.07 | 62.16 | 57.11 | 12.09 | **64.62** |
| | 0.5 | DOC | 72.86 | 51.14 | 44.53 | 41.34 | 65.21 | 72.92 | 79.53 | 69.66 | 3.90 | 64.26 | 16.80 | 10.29 | 49.37 |
| | | DeepUNK | 87.27 | 38.65 | 48.54 | 26.46 | 59.26 | 62.55 | 67.44 | 65.73 | 4.16 | 53.91 | 22.56 | 9.30 | 45.49 |
| | | ADB | 85.13 | **84.69** | 52.17 | 62.38 | 77.54 | 85.49 | 86.39 | 77.58 | **81.39** | 71.75 | **84.23** | 19.46 | **72.35** |
| | | SCL | 80.31 | 29.56 | 22.26 | 37.38 | 35.32 | 50.95 | 74.34 | 39.30 | 1.78 | 57.35 | 36.56 | 11.13 | 39.69 |
| | | AB | 60.12 | 31.46 | 48.32 | 26.72 | 64.32 | 65.68 | 74.78 | 69.18 | 47.74 | 61.44 | 47.86 | 19.74 | 51.45 |
| | | KNNCon | 92.19 | 71.85 | 52.39 | 54.90 | 79.04 | 80.60 | 86.48 | **80.88** | 81.24 | 75.65 | 64.26 | **20.80** | 70.02 |
| | | DyEn | 95.17 | 75.65 | 50.37 | 56.31 | 77.28 | 72.38 | 86.54 | 77.54 | 77.69 | **76.07** | 74.29 | 19.31 | 69.88 |
| | | LLM-MaxSoftmax | 75.93 | 67.71 | 58.19 | 61.78 | 70.22 | 68.16 | 78.18 | 75.35 | 68.88 | 64.91 | 74.02 | 15.07 | 64.87 |
| | | LLM-OpenMax | 92.12 | 80.73 | 52.45 | 38.93 | 55.54 | 84.90 | 74.02 | 63.26 | 45.13 | 53.30 | 29.25 | 15.07 | 57.06 |
| | | LLM-TempScale | 75.93 | 67.88 | 57.79 | 61.46 | 70.12 | 68.16 | 78.35 | 75.18 | 68.51 | 64.47 | 74.02 | 15.39 | 64.77 |
| | | LLM-Energy | 94.63 | 77.70 | 50.33 | 41.75 | 74.99 | 85.27 | 86.71 | 73.70 | 48.48 | 64.85 | 62.83 | 8.63 | 64.16 |
| | | LLM-LogitNorm | 82.11 | 69.00 | 52.76 | **64.18** | **80.97** | 75.08 | **91.59** | 79.76 | 72.66 | 66.73 | 77.22 | 11.82 | 68.66 |
| | | LLM-Entropy | 77.08 | 68.57 | **59.77** | 61.75 | 69.10 | 67.98 | 76.83 | 75.67 | 68.89 | 66.32 | 73.07 | 15.28 | 65.03 |
| | | LLM-KLMatching | 72.29 | 64.20 | 49.76 | 58.67 | 67.84 | 66.68 | 74.67 | 73.28 | 66.46 | 62.73 | 70.95 | 18.52 | 62.17 |
| | | LLM-MaxLogit | **95.23** | 79.72 | 53.47 | 48.96 | 77.64 | **86.46** | 89.65 | 77.30 | 55.66 | 67.70 | 71.80 | 10.99 | 67.88 |
| | 0.75 | DOC | 76.99 | 44.43 | 37.06 | 41.62 | 66.77 | 77.97 | 80.36 | 71.59 | 1.76 | 61.42 | 0.24 | 8.61 | 47.40 |
| | | DeepUNK | 91.97 | 37.48 | 45.32 | 21.56 | 48.53 | 64.13 | 63.03 | 60.81 | 1.93 | 50.89 | 19.62 | 5.48 | 42.56 |
| | | ADB | 92.94 | **86.23** | 61.65 | 65.88 | 82.51 | 85.53 | 90.25 | 81.74 | 84.19 | 75.49 | **88.13** | 19.46 | 76.17 |
| | | SCL | 73.31 | 29.24 | 26.10 | 40.13 | 56.55 | 67.37 | 80.13 | 47.79 | 1.39 | 64.79 | 49.60 | 10.99 | 45.62 |
| | | AB | 62.62 | 30.28 | 49.02 | 27.36 | 67.52 | 67.70 | 77.66 | 70.82 | 49.68 | 60.10 | 49.14 | 18.06 | 52.50 |
| | | KNNCon | **96.73** | 81.82 | 59.82 | 68.72 | **86.44** | 85.22 | **92.60** | **85.57** | **86.64** | **82.92** | 73.21 | **20.78** | **76.71** |
| | | DyEn | 95.02 | 82.14 | 59.77 | 63.43 | 85.71 | 82.52 | 91.87 | 83.40 | 84.29 | 82.68 | 80.82 | 17.85 | 75.79 |
| | | LLM-MaxSoftmax | 89.59 | 77.84 | 67.29 | 69.44 | 83.30 | 83.05 | 89.76 | 83.73 | 79.53 | 74.73 | 84.48 | 17.21 | 75.00 |
| | | LLM-OpenMax | 93.03 | 83.74 | 61.22 | 39.25 | 60.36 | **88.36** | 77.54 | 67.51 | 32.16 | 54.08 | 35.49 | 13.76 | 58.87 |
| | | LLM-TempScale | 89.59 | 77.98 | **67.40** | **69.82** | 83.16 | 83.35 | 89.94 | 83.63 | 79.45 | 74.31 | 84.62 | 17.50 | 75.06 |
| | | LLM-Energy | 90.05 | 79.97 | 34.46 | 40.67 | 74.13 | 81.74 | 82.51 | 73.01 | 43.16 | 47.54 | 67.13 | 8.02 | 60.20 |
| | | LLM-LogitNorm | 95.45 | 81.84 | 64.69 | 62.41 | 81.12 | 87.70 | 91.87 | 81.58 | 78.75 | 70.84 | 75.55 | 10.90 | 73.56 |
| | | LLM-Entropy | 90.05 | 78.21 | 66.89 | 69.75 | 82.37 | 83.04 | 89.17 | 83.43 | 80.08 | 75.53 | 84.51 | 17.73 | 75.06 |
| | | LLM-KLMatching | 88.44 | 76.38 | 62.54 | 69.26 | 82.00 | 79.67 | 88.71 | 82.77 | 78.04 | 73.51 | 82.88 | 19.93 | 73.68 |
| | | LLM-MaxLogit | 92.42 | 81.93 | 43.88 | 44.94 | 77.73 | 84.23 | 87.79 | 76.94 | 49.68 | 53.08 | 74.91 | 10.49 | 64.83 |

Table S11: F1 performance of OSTC on the BOLT benchmark under different KCRs (LAR = 0.5)

| Metric | KCR | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| F1 | 0.25 | DOC | 79.85 | 53.40 | 49.28 | 43.67 | 67.68 | 60.09 | 80.52 | 71.18 | 9.66 | 65.70 | 52.73 | 17.60 | 54.28 |
| | | DeepUNK | 93.57 | 35.23 | 53.41 | 39.86 | 71.45 | 67.13 | 75.54 | 74.55 | 11.17 | 64.61 | 33.81 | 15.01 | 52.95 |
| | | ADB | 73.73 | 75.49 | 53.13 | 56.04 | 67.30 | 79.14 | 82.58 | 73.81 | 73.49 | 63.52 | 79.09 | 23.89 | 66.77 |
| | | SCL | 72.71 | 32.27 | 35.01 | 39.13 | 30.38 | 45.79 | 51.07 | 41.02 | 3.08 | 51.94 | 38.77 | 12.93 | 37.84 |
| | | AB | 57.86 | 31.95 | 53.28 | 30.21 | 59.71 | 63.41 | 71.34 | 68.51 | 45.31 | 64.95 | 52.21 | 15.60 | 51.20 |
| | | KNNCon | 87.29 | 62.00 | 37.13 | 45.74 | 64.32 | 60.83 | 78.52 | 71.43 | 60.59 | 56.39 | 58.46 | 18.64 | 58.44 |
| | | DyEn | **96.57** | 64.05 | 32.78 | 52.21 | 64.81 | 56.59 | 80.69 | 67.32 | **74.70** | 65.10 | 63.03 | **24.21** | 61.84 |
| | | LLM-MaxSoftmax | 55.17 | 45.36 | 34.21 | 46.55 | 50.25 | 43.51 | 57.87 | 60.56 | 50.46 | 48.76 | 59.13 | 16.12 | 47.33 |
| | | LLM-OpenMax | 82.04 | 64.71 | 34.58 | 45.30 | 69.88 | 64.97 | 81.93 | **78.96** | 54.56 | 55.32 | 42.78 | 18.69 | 57.81 |
| | | LLM-TempScale | 54.67 | 45.53 | 34.42 | 46.34 | 50.21 | 43.63 | 57.89 | 60.39 | 48.44 | 47.86 | 59.38 | 16.08 | 47.07 |
| | | LLM-Energy | 89.26 | 79.62 | 64.30 | 52.32 | 73.71 | 84.44 | 88.95 | 76.89 | 60.85 | 68.86 | 71.47 | 14.03 | 68.73 |
| | | LLM-LogitNorm | 55.60 | 42.88 | 41.44 | **56.80** | 64.83 | 41.22 | 82.82 | 71.63 | 46.27 | 45.51 | **79.61** | 14.80 | 53.62 |
| | | LLM-Entropy | 57.10 | 47.34 | 41.43 | 47.55 | 48.63 | 44.51 | 56.51 | 60.32 | 53.18 | 52.71 | 58.46 | 15.98 | 48.64 |
| | | LLM-KLMatching | 50.75 | 42.31 | 31.21 | 43.25 | 46.21 | 40.94 | 53.60 | 58.16 | 42.62 | 44.33 | 54.15 | 14.98 | 43.54 |
| | | LLM-MaxLogit | 89.31 | **79.99** | **64.51** | 56.64 | **75.73** | **84.87** | **89.87** | 78.44 | 63.97 | **70.53** | 76.22 | 15.06 | **70.43** |
| | 0.5 | DOC | 79.23 | 54.14 | 50.08 | 51.31 | 74.80 | 73.13 | 84.13 | 75.68 | 3.90 | 70.82 | 49.28 | 11.74 | 56.52 |
| | | DeepUNK | 95.65 | 40.45 | 52.07 | 32.14 | 63.13 | 70.14 | 71.38 | 71.79 | 5.60 | 65.07 | 22.92 | 9.92 | 50.02 |
| | | ADB | 86.36 | 84.53 | 53.96 | 66.04 | 81.28 | **86.13** | 88.57 | 80.29 | 82.36 | 76.71 | **84.59** | **22.42** | **74.44** |
| | | SCL | 57.51 | 38.08 | 23.77 | 37.87 | 48.37 | 55.24 | 77.98 | 49.09 | 2.12 | 65.68 | 41.71 | 12.57 | 42.50 |
| | | AB | 60.70 | 31.88 | 48.76 | 29.26 | 64.72 | 65.92 | 74.84 | 68.72 | 49.64 | 62.62 | 56.64 | 16.82 | 52.54 |
| | | KNNCon | 92.48 | 75.83 | 54.92 | 58.55 | 77.08 | 79.07 | 86.68 | 80.22 | **83.41** | 74.47 | 64.47 | 21.28 | 70.70 |
| | | DyEn | 95.67 | 75.40 | 50.07 | 57.48 | 78.06 | 72.35 | 85.92 | 79.20 | 80.87 | **78.28** | 75.19 | 19.58 | 70.67 |
| | | LLM-MaxSoftmax | 75.16 | 65.01 | 62.44 | 63.45 | 71.20 | 66.71 | 76.83 | 77.59 | 69.84 | 67.69 | 75.73 | 19.26 | 65.91 |
| | | LLM-OpenMax | 93.57 | 82.56 | 50.12 | 49.08 | 73.35 | 83.20 | 85.39 | 78.76 | 60.74 | 73.91 | 48.30 | 19.35 | 66.53 |
| | | LLM-TempScale | 75.16 | 65.01 | 62.04 | 63.45 | 71.20 | 66.87 | 76.84 | 77.59 | 69.26 | 67.55 | 75.86 | 19.47 | 65.86 |
| | | LLM-Energy | **98.41** | 84.75 | 54.10 | 56.37 | 79.77 | 84.06 | 89.66 | 79.24 | 59.57 | 67.70 | 71.14 | 11.28 | 69.67 |
| | | LLM-LogitNorm | 83.45 | 69.64 | 56.26 | **71.87** | **83.67** | 72.92 | **92.37** | **85.06** | 80.25 | 72.55 | 83.36 | 12.02 | 71.95 |
| | | LLM-Entropy | 75.96 | 65.10 | **65.66** | 61.87 | 69.67 | 66.16 | 75.67 | 76.89 | 70.15 | 69.44 | 74.56 | 19.77 | 65.91 |
| | | LLM-KLMatching | 73.32 | 63.46 | 51.90 | 62.26 | 69.72 | 65.54 | 74.57 | 76.62 | 65.18 | 65.79 | 72.06 | 20.08 | 63.37 |
| | | LLM-MaxLogit | 96.95 | **85.29** | 59.49 | 60.59 | 81.91 | 84.92 | 91.40 | 81.63 | 64.09 | 70.43 | 78.75 | 13.23 | 72.39 |
| | 0.75 | DOC | 81.80 | 49.68 | 43.70 | 54.64 | 76.22 | 79.40 | 85.90 | 78.16 | 1.76 | 71.45 | 47.75 | 11.09 | 56.80 |
| | | DeepUNK | 96.60 | 40.62 | 50.47 | 24.61 | 45.21 | 72.85 | 64.81 | 66.91 | 2.90 | 63.37 | 17.83 | 6.25 | 46.04 |
| | | ADB | 93.01 | **86.73** | 62.79 | 72.33 | 85.48 | 86.14 | 92.20 | 83.55 | 86.49 | 78.44 | **88.45** | 22.53 | **78.18** |
| | | SCL | 76.54 | 31.15 | 33.60 | 56.08 | 63.15 | 61.71 | 84.74 | 70.08 | 1.82 | 73.02 | 56.29 | 13.98 | 51.85 |
| | | AB | 63.00 | 30.80 | 49.52 | 31.00 | 68.34 | 68.20 | 77.66 | 70.86 | 51.96 | 61.80 | 59.02 | 19.20 | 54.28 |
| | | KNNCon | 96.97 | 83.82 | 62.03 | 69.33 | **86.01** | 85.42 | 92.92 | **85.96** | 86.78 | **83.63** | 71.36 | 21.62 | 77.15 |
| | | DyEn | 94.22 | 81.31 | 58.27 | 68.48 | 85.21 | 81.99 | 91.95 | 84.28 | **88.83** | 82.95 | 82.79 | 19.10 | 76.61 |
| | | LLM-MaxSoftmax | 90.93 | 77.63 | 68.89 | **77.14** | 84.29 | 81.76 | 89.42 | 85.30 | 83.01 | 80.25 | 86.12 | 20.80 | 77.13 |
| | | LLM-OpenMax | 97.02 | 83.23 | 63.14 | 52.32 | 75.85 | **89.46** | 88.82 | 80.29 | 54.22 | 73.43 | 53.51 | 18.69 | 69.17 |
| | | LLM-TempScale | 90.93 | 77.77 | 68.76 | 77.14 | 84.29 | 81.77 | 89.58 | 85.31 | 82.92 | 80.12 | 86.16 | 20.97 | 77.14 |
| | | LLM-Energy | **99.36** | 80.36 | 37.37 | 51.68 | 75.33 | 81.09 | 86.28 | 74.28 | 63.22 | 65.17 | 74.63 | 9.63 | 66.53 |
| | | LLM-LogitNorm | 95.94 | 83.49 | 66.70 | 73.83 | 83.74 | 86.84 | **93.46** | 83.73 | 86.20 | 80.61 | 84.61 | 11.24 | 77.53 |
| | | LLM-Entropy | 91.66 | 77.97 | **69.39** | 76.89 | 83.67 | 81.29 | 88.68 | 85.07 | 83.21 | 80.44 | 85.39 | 20.79 | 77.04 |
| | | LLM-KLMatching | 89.63 | 75.94 | 63.63 | 75.92 | 83.04 | 80.43 | 88.20 | 84.53 | 81.52 | 78.51 | 84.47 | **23.07** | 75.74 |
| | | LLM-MaxLogit | 99.05 | 82.28 | 45.02 | 56.71 | 79.06 | 82.99 | 89.02 | 78.19 | 70.32 | 68.41 | 81.79 | 12.28 | 70.43 |

Table S12: F1 performance of OSTC on the BOLT benchmark under different KCRs (LAR = 1.0)

| Metric | KCR | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ACC | 0.25 | DOC | 78.63 | **74.88** | 62.74 | **70.41** | 68.13 | 76.00 | 74.67 | 75.19 | 63.35 | **75.46** | 67.00 | **78.63** | 72.09 |
| | | DeepUNK | 79.80 | 49.02 | **73.43** | 47.70 | 19.39 | 76.51 | 26.10 | 17.74 | 7.45 | 27.42 | 21.34 | 6.55 | 37.70 |
| | | ADB | 43.36 | 69.19 | 39.75 | 31.84 | 37.25 | 80.12 | 79.16 | 61.48 | 55.37 | 40.99 | 61.60 | 52.18 | 54.36 |
| | | SCL | **98.51** | 65.50 | 46.15 | 48.13 | 38.69 | 78.13 | 57.54 | 45.80 | 7.10 | 38.47 | 42.46 | 42.00 | 50.71 |
| | | AB | 75.92 | 40.39 | 69.85 | 63.72 | 76.37 | 74.94 | **83.02** | **84.95** | 72.92 | 73.99 | 73.83 | 77.26 | **72.26** |
| | | KNNCon | 88.50 | 55.96 | 36.31 | 61.18 | 63.23 | 57.43 | 79.82 | 67.60 | 55.08 | 49.35 | 66.95 | 36.72 | 59.84 |
| | | DyEn | 94.97 | 61.85 | 29.57 | 68.53 | 64.94 | 46.85 | 81.99 | 68.17 | **76.71** | 69.20 | 66.19 | 31.60 | 63.38 |
| | | LLM-MaxSoftmax | 36.24 | 39.77 | 32.65 | 44.22 | 51.77 | 46.51 | 61.47 | 49.19 | 50.52 | 35.50 | 55.15 | 38.32 | 45.11 |
| | | LLM-OpenMax | 31.27 | 54.33 | 21.21 | 51.16 | 75.88 | 63.01 | 76.20 | 74.27 | 67.71 | 31.87 | **74.69** | 75.71 | 58.11 |
| | | LLM-TempScale | 39.94 | 39.39 | 30.98 | 41.05 | 52.47 | 45.53 | 61.70 | 47.74 | 51.01 | 41.09 | 57.00 | 44.60 | 46.04 |
| | | LLM-Energy | 62.89 | 71.30 | 60.90 | 62.76 | **78.92** | **81.20** | 81.81 | 74.22 | 67.43 | 57.25 | 73.34 | 62.16 | 69.51 |
| | | LLM-LogitNorm | 26.55 | 31.29 | 41.12 | 47.82 | 65.15 | 30.94 | 80.90 | 57.09 | 29.51 | 29.25 | 67.96 | 35.57 | 45.26 |
| | | LLM-Entropy | 43.08 | 47.04 | 47.96 | 49.42 | 52.33 | 54.74 | 61.07 | 52.15 | 60.13 | 45.87 | 55.12 | 40.45 | 50.78 |
| | | LLM-KLMatching | 24.71 | 36.88 | 27.59 | 14.42 | 35.09 | 34.55 | 40.43 | 30.77 | 18.49 | 17.62 | 24.82 | 21.73 | 27.26 |
| | | LLM-MaxLogit | 60.91 | 70.32 | 61.51 | 59.25 | 78.92 | 81.06 | 82.66 | 72.48 | 64.86 | 56.19 | 70.87 | 57.63 | 68.05 |
| | 0.5 | DOC | 56.58 | 59.66 | 48.95 | 23.01 | 45.40 | 73.79 | 64.25 | 52.33 | 45.77 | 50.62 | 36.43 | 45.27 | 50.17 |
| | | DeepUNK | 59.38 | 45.57 | 51.78 | 43.07 | 41.46 | 57.59 | 46.84 | 43.83 | 12.57 | 39.11 | 29.36 | 23.78 | 41.20 |
| | | ADB | 63.65 | **79.73** | 48.84 | 37.51 | 58.06 | **84.25** | 71.12 | 62.71 | 51.26 | 43.14 | 68.56 | 42.83 | 59.30 |
| | | SCL | **98.59** | 51.65 | 44.55 | 50.61 | 46.09 | 67.27 | 47.15 | 46.45 | 11.87 | 37.97 | 40.42 | 39.65 | 48.52 |
| | | AB | 59.24 | 36.37 | 49.56 | 52.56 | 66.62 | 67.33 | 73.56 | **74.15** | 52.68 | 58.79 | 54.71 | **54.86** | 58.37 |
| | | KNNCon | 88.92 | 70.18 | 48.23 | **67.02** | 68.97 | 76.41 | **82.08** | 70.71 | 77.28 | **74.20** | **71.20** | 43.88 | **69.92** |
| | | DyEn | 90.25 | 70.78 | 43.33 | 64.99 | **69.41** | 66.84 | 81.08 | 73.74 | **81.72** | 72.32 | 69.68 | 33.58 | 68.14 |
| | | LLM-MaxSoftmax | 65.44 | 50.08 | 41.20 | 61.18 | 56.43 | 62.67 | 69.13 | 57.40 | 53.14 | 33.60 | 42.59 | 38.17 | 52.59 |
| | | LLM-OpenMax | 71.50 | 54.36 | 48.98 | 63.92 | 52.42 | 71.24 | 56.19 | 52.85 | 53.53 | 34.80 | 49.02 | 48.12 | 54.74 |
| | | LLM-TempScale | 65.41 | 50.09 | 41.20 | 57.06 | 55.87 | 62.18 | 68.52 | 55.62 | 52.60 | 35.29 | 42.03 | 41.11 | 52.25 |
| | | LLM-Energy | 85.19 | 63.29 | 55.44 | 62.49 | 63.03 | 80.86 | 74.04 | 58.45 | 52.57 | 33.84 | 50.24 | 47.79 | 60.60 |
| | | LLM-LogitNorm | 69.04 | 44.55 | 35.05 | 62.78 | 65.71 | 66.19 | 79.24 | 62.54 | 51.95 | 23.44 | 49.11 | 41.39 | 54.25 |
| | | LLM-Entropy | 67.13 | 52.34 | 43.86 | 62.24 | 55.76 | 64.37 | 67.61 | 58.41 | 55.84 | 37.04 | 41.74 | 38.96 | 53.77 |
| | | LLM-KLMatching | 60.57 | 48.80 | 39.82 | 43.55 | 51.05 | 56.10 | 61.59 | 48.33 | 28.21 | 14.08 | 22.45 | 23.85 | 41.53 |
| | | LLM-MaxLogit | 87.70 | 64.21 | **55.52** | 62.94 | 65.55 | 82.88 | 77.88 | 60.64 | 53.64 | 33.90 | 49.77 | 45.99 | 61.72 |
| | 0.75 | DOC | 43.82 | 41.04 | 22.54 | 55.08 | 16.05 | 68.98 | 25.33 | 27.09 | 28.10 | 25.58 | 13.57 | 24.12 | 32.61 |
| | | DeepUNK | 54.28 | 30.71 | 26.24 | 34.07 | 47.17 | 40.27 | 58.31 | 48.55 | 13.69 | 51.83 | 31.67 | 19.42 | 38.02 |
| | | ADB | 81.69 | **80.01** | 56.03 | 48.16 | 60.33 | 80.06 | 76.09 | 63.93 | 57.76 | 43.08 | 74.47 | 43.73 | 63.78 |
| | | SCL | **97.76** | 32.28 | 26.08 | 46.39 | 32.99 | 62.41 | 53.34 | 44.28 | 13.79 | 35.24 | 26.61 | 30.73 | 41.83 |
| | | AB | 53.22 | 28.83 | 44.13 | 40.96 | 59.28 | 61.90 | 67.70 | 63.71 | 42.54 | 43.69 | 38.20 | 38.06 | 48.52 |
| | | KNNCon | 95.06 | 78.13 | **60.34** | **81.16** | 79.54 | **82.29** | 87.56 | **77.80** | 85.93 | 77.80 | **79.16** | **48.49** | **77.77** |
| | | DyEn | 97.15 | 76.54 | 59.03 | 75.41 | **80.83** | 79.00 | **87.96** | 70.26 | **88.46** | **79.56** | 78.03 | 45.05 | 76.44 |
| | | LLM-MaxSoftmax | 83.16 | 65.08 | 50.96 | 63.80 | 60.50 | 76.17 | 79.35 | 59.17 | 51.19 | 38.07 | 29.20 | 37.87 | 57.88 |
| | | LLM-OpenMax | 52.49 | 50.51 | 36.83 | 44.94 | 29.12 | 69.07 | 39.01 | 28.70 | 33.61 | 27.13 | 25.05 | 29.51 | 38.83 |
| | | LLM-TempScale | 82.61 | 65.08 | 50.94 | 63.27 | 60.38 | 76.17 | 79.33 | 59.09 | 51.17 | 38.43 | 29.00 | 40.36 | 57.99 |
| | | LLM-Energy | 78.65 | 55.57 | 34.62 | 47.47 | 41.27 | 73.34 | 66.18 | 42.38 | 37.69 | 31.66 | 27.45 | 30.84 | 47.26 |
| | | LLM-LogitNorm | 87.19 | 65.64 | 49.27 | 60.24 | 50.37 | 81.57 | 75.32 | 50.35 | 47.66 | 34.56 | 28.34 | 31.76 | 55.19 |
| | | LLM-Entropy | 83.75 | 66.47 | 51.59 | 64.08 | 60.46 | 75.81 | 78.56 | 60.78 | 51.87 | 38.91 | 28.75 | 38.62 | 58.30 |
| | | LLM-KLMatching | 81.00 | 64.64 | 50.91 | 62.12 | 59.84 | 72.87 | 77.96 | 58.08 | 41.58 | 36.98 | 23.62 | 33.23 | 55.24 |
| | | LLM-MaxLogit | 84.25 | 58.76 | 36.57 | 51.71 | 46.48 | 76.70 | 72.57 | 46.78 | 39.35 | 32.93 | 28.65 | 32.27 | 50.58 |

Table S13: ACC performance of OSTC on the BOLT benchmark under different KCRs (LAR = 0.1)

| Metric | KCR | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ACC | 0.25 | DOC | 90.43 | 71.63 | 59.21 | 78.98 | 79.74 | 70.78 | 88.10 | 73.04 | 77.09 | 72.63 | 82.73 | **78.12** | 76.87 |
| | | DeepUNK | 93.14 | 52.91 | 66.20 | 56.24 | 83.31 | 82.77 | 88.22 | 85.97 | 77.17 | **82.66** | 54.34 | 30.82 | 71.15 |
| | | ADB | 68.65 | 78.15 | 49.49 | 56.86 | 76.45 | 84.92 | 87.06 | 72.69 | 69.75 | 49.13 | 79.52 | 39.63 | 67.69 |
| | | SCL | 85.20 | 59.27 | 57.08 | 79.51 | 71.55 | 76.57 | 74.36 | 71.70 | 12.94 | 77.72 | 62.53 | 36.61 | 63.75 |
| | | AB | 69.72 | 42.76 | 62.36 | 59.28 | 72.72 | 75.54 | 81.82 | 81.50 | 62.66 | 76.18 | 70.70 | 62.64 | 68.16 |
| | | KNNCon | 84.02 | 45.05 | 34.85 | 51.70 | 60.18 | 59.60 | 79.56 | 59.58 | 56.39 | 49.48 | 62.57 | 37.92 | 56.74 |
| | | DyEn | **95.13** | 64.93 | 28.84 | 62.83 | 65.63 | 52.53 | 81.83 | 57.41 | 65.17 | 65.06 | 66.64 | 38.92 | 62.08 |
| | | LLM-MaxSoftmax | 43.36 | 39.07 | 26.99 | 49.42 | 41.74 | 39.48 | 49.97 | 43.93 | 52.42 | 39.88 | 58.28 | 37.22 | 43.48 |
| | | LLM-OpenMax | 82.20 | 58.63 | 30.04 | 74.83 | 82.77 | 56.04 | 88.59 | 81.47 | 61.19 | 53.47 | 78.33 | 69.33 | 68.07 |
| | | LLM-TempScale | 43.35 | 39.07 | 27.15 | 48.54 | 41.12 | 40.40 | 49.97 | 43.91 | 49.56 | 38.73 | 58.25 | 41.04 | 43.42 |
| | | LLM-Energy | 92.14 | 86.09 | **73.21** | **81.84** | **85.78** | 89.85 | **92.60** | **88.35** | **85.06** | 73.60 | **88.63** | 65.61 | **83.56** |
| | | LLM-LogitNorm | 41.76 | 37.65 | 33.81 | 60.58 | 71.16 | 43.18 | 87.40 | 69.34 | 52.88 | 39.82 | 86.91 | 45.97 | 55.87 |
| | | LLM-Entropy | 46.50 | 42.03 | 34.02 | 51.12 | 39.90 | 42.28 | 47.49 | 44.01 | 57.14 | 47.37 | 55.66 | 36.84 | 45.36 |
| | | LLM-KLMatching | 35.83 | 34.29 | 25.33 | 43.10 | 34.40 | 32.00 | 42.64 | 37.89 | 41.53 | 36.01 | 47.60 | 30.29 | 36.74 |
| | | LLM-MaxLogit | 91.33 | **86.11** | 72.66 | 80.03 | 85.39 | **90.36** | 92.16 | 88.31 | 84.96 | 73.23 | 88.60 | 62.61 | 82.98 |
| | 0.5 | DOC | 79.75 | 63.46 | 57.99 | 77.14 | 72.14 | 73.15 | 81.81 | 73.62 | 50.34 | 69.43 | 58.16 | 54.01 | 67.58 |
| | | DeepUNK | 89.04 | 37.22 | 58.59 | 62.37 | 70.32 | 71.61 | 76.76 | 75.56 | 38.55 | 67.13 | 45.42 | 50.04 | 61.88 |
| | | ADB | 80.64 | **84.00** | 45.20 | 73.06 | 76.13 | 86.50 | 84.58 | 74.52 | **82.47** | 69.00 | 81.87 | 35.09 | 72.75 |
| | | SCL | 86.60 | 48.23 | 43.54 | 74.88 | 50.42 | 59.31 | 73.98 | 52.33 | 13.30 | 68.28 | 56.66 | 46.58 | 56.18 |
| | | AB | 60.76 | 38.42 | 50.30 | 47.54 | 65.56 | 67.56 | 74.62 | 72.40 | 51.58 | 63.68 | 61.64 | 47.44 | 58.46 |
| | | KNNCon | 89.28 | 65.93 | 46.80 | 66.96 | 72.39 | 77.36 | 81.51 | 69.93 | 78.23 | 70.60 | 69.22 | 44.41 | 69.39 |
| | | DyEn | **94.37** | 71.42 | 43.88 | 71.52 | 70.11 | 65.89 | 82.48 | 66.50 | 70.67 | 71.55 | 71.01 | 33.83 | 67.77 |
| | | LLM-MaxSoftmax | 64.45 | 58.66 | 53.61 | 64.57 | 57.81 | 57.91 | 65.67 | 61.61 | 63.04 | 58.61 | 68.07 | 36.64 | 59.22 |
| | | LLM-OpenMax | 90.66 | 78.61 | 45.66 | 80.08 | 68.56 | 83.43 | 79.59 | 74.81 | 71.71 | 58.31 | 61.72 | **58.51** | 70.97 |
| | | LLM-TempScale | 64.45 | 58.87 | 53.07 | 63.76 | 57.64 | 57.91 | 65.98 | 61.09 | 62.68 | 57.82 | 68.08 | 35.94 | 58.94 |
| | | LLM-Energy | 94.02 | 81.06 | 64.40 | 81.22 | 78.79 | 87.04 | 88.33 | 79.77 | 73.87 | 72.15 | 81.91 | 50.23 | 77.73 |
| | | LLM-LogitNorm | 74.63 | 60.51 | 45.96 | 81.18 | **80.05** | 68.51 | **90.30** | 79.59 | 73.87 | 63.87 | **87.60** | 43.37 | 70.79 |
| | | LLM-Entropy | 66.27 | 59.83 | 56.64 | 64.00 | 55.36 | 57.52 | 62.89 | 61.26 | 62.81 | 61.15 | 65.75 | 33.47 | 58.91 |
| | | LLM-KLMatching | 59.25 | 54.29 | 41.87 | 57.18 | 53.05 | 55.63 | 58.68 | 55.39 | 59.71 | 54.98 | 61.91 | 30.97 | 53.58 |
| | | LLM-MaxLogit | 93.97 | 82.04 | **65.58** | **81.84** | 80.05 | **87.77** | 90.18 | **81.47** | 75.92 | **73.78** | 85.55 | 51.79 | **79.16** |
| | 0.75 | DOC | 75.05 | 48.73 | 38.03 | 69.92 | 62.75 | 74.08 | 75.90 | 68.31 | 28.10 | 59.27 | 24.95 | 35.74 | 55.07 |
| | | DeepUNK | 89.94 | 35.53 | 43.26 | 39.35 | 53.77 | 61.57 | 65.71 | 63.70 | 28.13 | 51.54 | 38.81 | 24.79 | 49.67 |
| | | ADB | 89.03 | **82.93** | 58.09 | 77.79 | 78.04 | 82.33 | 85.92 | 76.41 | **82.55** | 70.88 | **85.49** | 41.48 | 75.91 |
| | | SCL | 77.81 | 38.57 | 31.07 | 66.54 | 55.33 | 64.31 | 74.73 | 48.32 | 11.84 | 62.30 | 61.48 | 44.72 | 53.08 |
| | | AB | 58.14 | 32.34 | 45.98 | 43.48 | 63.04 | 63.46 | 72.38 | 67.48 | 48.74 | 56.98 | 56.94 | 37.78 | 53.90 |
| | | KNNCon | 94.40 | 77.41 | 58.60 | **81.70** | **80.14** | 80.94 | 87.67 | 77.32 | 81.68 | 78.46 | 78.94 | 47.01 | **77.02** |
| | | DyEn | **95.23** | 77.68 | 58.29 | 76.45 | 79.45 | 77.10 | 86.83 | 75.22 | 77.51 | **78.69** | 80.10 | 45.01 | 75.63 |
| | | LLM-MaxSoftmax | 83.13 | 72.64 | 64.07 | 77.51 | 75.27 | 77.52 | 80.62 | 75.06 | 74.65 | 69.85 | 79.91 | 46.75 | 73.08 |
| | | LLM-OpenMax | 91.25 | 80.07 | 60.39 | 72.86 | 57.92 | **85.31** | 73.06 | 67.42 | 54.99 | 53.35 | 48.61 | **49.42** | 66.22 |
| | | LLM-TempScale | 83.13 | 72.81 | **64.22** | 77.59 | 75.08 | 77.94 | 80.99 | 74.92 | 74.65 | 69.37 | 80.19 | 46.08 | 73.08 |
| | | LLM-Energy | 90.32 | 77.30 | 38.77 | 72.29 | 70.23 | 79.10 | 79.35 | 70.99 | 58.57 | 49.85 | 77.50 | 35.18 | 66.62 |
| | | LLM-LogitNorm | 92.01 | 77.35 | 61.16 | 81.51 | 76.84 | 84.00 | **88.61** | **77.73** | 78.52 | 67.49 | 81.60 | 34.15 | 75.08 |
| | | LLM-Entropy | 83.75 | 73.26 | 64.05 | 76.29 | 73.84 | 77.61 | 79.40 | 74.53 | 75.30 | 71.24 | 79.40 | 46.23 | 72.91 |
| | | LLM-KLMatching | 81.43 | 71.01 | 60.99 | 76.29 | 73.22 | 73.31 | 78.51 | 72.87 | 72.65 | 68.70 | 76.72 | 44.99 | 70.89 |
| | | LLM-MaxLogit | 91.92 | 79.23 | 45.36 | 74.86 | 73.55 | 81.42 | 84.50 | 74.22 | 62.99 | 53.60 | 82.47 | 39.63 | 70.31 |

Table S14: ACC performance of OSTC on the BOLT benchmark under different KCRs (LAR = 0.5)

| Metric | KCR | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ACC | 0.25 | DOC | 91.41 | 67.67 | 57.33 | 75.55 | 79.18 | 67.42 | 88.68 | 75.93 | 77.32 | 70.94 | 86.25 | **75.12** | 76.07 |
| | | DeepUNK | 96.85 | 34.31 | 62.46 | 79.63 | 84.52 | 83.54 | 89.29 | 87.56 | 77.43 | **82.99** | 62.12 | 58.21 | 74.91 |
| | | ADB | 69.48 | 77.96 | 55.02 | 71.41 | 77.34 | 85.91 | 88.46 | 76.51 | 81.78 | 61.75 | 82.89 | 45.01 | 72.79 |
| | | SCL | 87.29 | 57.54 | 63.47 | 79.82 | 69.79 | 65.68 | 72.61 | 61.88 | 10.81 | 78.01 | 66.06 | 47.52 | 63.37 |
| | | AB | 70.54 | 43.18 | 64.79 | 54.54 | 73.43 | 75.91 | 81.80 | 80.30 | 61.10 | 77.04 | 70.76 | 61.42 | 67.90 |
| | | KNNCon | 87.22 | 59.92 | 33.78 | 57.85 | 63.69 | 60.26 | 81.63 | 62.33 | 55.55 | 50.53 | 67.90 | 40.00 | 60.06 |
| | | DyEn | **97.53** | 64.26 | 26.85 | 66.88 | 66.36 | 53.80 | 85.38 | 68.15 | 78.02 | 63.21 | 67.16 | 38.92 | 64.71 |
| | | LLM-MaxSoftmax | 41.22 | 36.60 | 29.49 | 42.88 | 41.13 | 36.88 | 49.16 | 42.06 | 50.27 | 40.60 | 53.94 | 40.71 | 42.08 |
| | | LLM-OpenMax | 83.93 | 64.94 | 30.89 | 75.47 | 85.86 | 68.29 | 91.86 | 83.58 | 75.96 | 56.25 | 85.07 | 71.54 | 72.80 |
| | | LLM-TempScale | 40.44 | 36.82 | 29.70 | 42.49 | 41.03 | 37.07 | 49.18 | 41.67 | 46.88 | 39.40 | 54.32 | 43.25 | 41.85 |
| | | LLM-Energy | 91.00 | **85.38** | **77.23** | **85.27** | **87.00** | **91.95** | **94.22** | **87.82** | 88.39 | 80.36 | 91.88 | 65.00 | **85.46** |
| | | LLM-LogitNorm | 43.18 | 33.57 | 41.35 | 62.90 | 68.24 | 33.59 | 86.60 | 65.12 | 43.74 | 37.79 | 88.86 | 48.61 | 54.46 |
| | | LLM-Entropy | 44.30 | 39.25 | 38.40 | 44.24 | 37.45 | 38.41 | 46.05 | 41.41 | 53.94 | 46.22 | 51.73 | 39.14 | 43.38 |
| | | LLM-KLMatching | 33.75 | 32.57 | 26.06 | 36.94 | 32.84 | 33.16 | 39.99 | 36.59 | 35.45 | 34.35 | 43.32 | 21.74 | 33.90 |
| | | LLM-MaxLogit | 90.50 | 85.02 | 77.16 | 84.76 | 86.74 | 91.88 | 94.16 | 87.43 | **88.66** | 80.09 | **91.93** | 61.97 | 85.02 |
| | 0.5 | DOC | 83.65 | 63.67 | 58.56 | 73.88 | 76.63 | 72.28 | 84.58 | 75.21 | 50.34 | 74.14 | 73.06 | 53.40 | 69.95 |
| | | DeepUNK | 95.93 | 35.82 | 56.46 | 65.35 | 71.85 | 75.71 | 79.34 | 78.31 | 50.49 | 72.69 | 50.92 | 47.48 | 65.03 |
| | | ADB | 82.34 | 83.90 | 47.55 | 78.82 | 79.68 | **87.07** | 86.93 | 75.29 | **83.10** | 74.44 | 82.58 | 32.73 | 74.54 |
| | | SCL | 69.66 | 50.77 | 40.29 | 70.60 | 57.85 | 61.12 | 73.85 | 54.28 | 9.64 | 71.36 | 54.76 | **58.21** | 56.03 |
| | | AB | 60.96 | 38.38 | 50.62 | 46.02 | 65.10 | 67.30 | 74.50 | 70.54 | 52.64 | 64.58 | 61.56 | 47.20 | 58.28 |
| | | KNNCon | 89.75 | 72.20 | 49.26 | 69.51 | 67.84 | 75.35 | 81.59 | 67.40 | 80.66 | 68.85 | 71.43 | 45.48 | 69.94 |
| | | DyEn | 97.78 | 71.86 | 42.84 | 70.23 | 70.33 | 65.49 | 82.04 | 63.15 | 81.72 | 74.66 | 72.17 | 34.38 | 68.89 |
| | | LLM-MaxSoftmax | 59.97 | 56.13 | 58.79 | 60.16 | 57.71 | 55.73 | 62.23 | 57.75 | 65.97 | 59.52 | 65.59 | 41.41 | 58.41 |
| | | LLM-OpenMax | 92.63 | 81.24 | 42.48 | 82.12 | 78.04 | 80.48 | 86.95 | 79.01 | 78.03 | 73.84 | 73.62 | 52.84 | 75.11 |
| | | LLM-TempScale | 59.97 | 56.13 | 58.18 | 60.16 | 57.71 | 55.97 | 62.24 | 57.75 | 65.06 | 59.27 | 65.93 | 40.36 | 58.23 |
| | | LLM-Energy | **98.08** | 85.78 | 66.63 | 84.08 | 82.03 | 86.34 | 90.45 | 83.45 | 77.71 | 74.26 | 86.12 | 45.93 | 80.07 |
| | | LLM-LogitNorm | 74.74 | 62.49 | 50.23 | 81.80 | 81.38 | 64.99 | 90.93 | 80.99 | 80.62 | 69.00 | **88.57** | 40.80 | 72.21 |
| | | LLM-Entropy | 61.12 | 56.28 | 64.31 | 58.45 | 54.46 | 54.95 | 59.89 | 55.99 | 66.42 | 62.05 | 62.92 | 38.20 | 57.92 |
| | | LLM-KLMatching | 56.97 | 54.24 | 43.88 | 58.20 | 54.66 | 54.10 | 57.69 | 54.86 | 58.47 | 56.68 | 58.32 | 32.97 | 53.42 |
| | | LLM-MaxLogit | 96.13 | **86.04** | **68.83** | **84.94** | **83.10** | 86.79 | **91.48** | **84.55** | 79.53 | **75.71** | 88.07 | 44.62 | **80.82** |
| | 0.75 | DOC | 79.47 | 51.25 | 42.55 | 71.84 | 70.99 | 75.56 | 81.40 | 73.55 | 28.10 | 67.67 | 61.33 | 36.85 | 61.71 |
| | | DeepUNK | 95.55 | 37.38 | 47.11 | 45.39 | 53.99 | 68.99 | 68.12 | 68.84 | 28.31 | 61.03 | 36.75 | 33.98 | 53.79 |
| | | ADB | 89.09 | **83.59** | 59.09 | 80.86 | **81.01** | 83.05 | 88.36 | 77.21 | 84.61 | 73.78 | 85.69 | 42.47 | 77.40 |
| | | SCL | 80.20 | 39.54 | 37.82 | 72.27 | 60.47 | 61.65 | 78.63 | 64.37 | 13.30 | 69.39 | 64.69 | 47.10 | 57.45 |
| | | AB | 58.40 | 32.54 | 46.48 | 43.06 | 63.50 | 63.72 | 72.28 | 66.94 | 49.52 | 58.60 | 60.26 | 37.50 | 54.40 |
| | | KNNCon | 94.86 | 79.70 | 59.82 | 81.65 | 79.67 | 81.31 | 88.26 | 78.00 | 83.78 | **79.42** | 77.89 | 49.07 | **77.79** |
| | | DyEn | 94.06 | 76.59 | 57.03 | 81.88 | 78.91 | 76.31 | 87.11 | 76.75 | **88.46** | 78.79 | 82.44 | 44.65 | 76.91 |
| | | LLM-MaxSoftmax | 85.28 | 72.29 | 65.99 | 77.27 | 75.95 | 75.30 | 79.59 | 74.55 | 77.12 | 75.29 | 79.52 | 46.23 | 73.70 |
| | | LLM-OpenMax | 96.16 | 79.55 | 62.62 | 77.31 | 71.37 | **86.31** | 85.24 | 77.13 | 68.21 | 69.49 | 65.80 | **57.14** | 74.69 |
| | | LLM-TempScale | 85.28 | 72.43 | 65.86 | 77.27 | 75.96 | 75.31 | 79.95 | 74.57 | 76.99 | 75.11 | 79.56 | 46.29 | 73.72 |
| | | LLM-Energy | **99.02** | 78.75 | 40.92 | 76.69 | 71.47 | 78.84 | 83.50 | 72.40 | 69.87 | 63.26 | 81.04 | 37.23 | 71.08 |
| | | LLM-LogitNorm | 92.83 | 79.44 | 64.23 | **84.57** | 79.84 | 82.41 | **90.64** | **79.21** | 84.00 | 76.37 | **86.25** | 29.13 | 77.41 |
| | | LLM-Entropy | 86.27 | 72.78 | **66.76** | 76.57 | 74.94 | 74.78 | 78.13 | 74.19 | 77.53 | 75.77 | 78.05 | 46.53 | 73.53 |
| | | LLM-KLMatching | 83.47 | 70.54 | 62.40 | 76.12 | 73.72 | 73.74 | 77.14 | 72.64 | 74.99 | 73.17 | 76.36 | 48.31 | 71.88 |
| | | LLM-MaxLogit | 98.53 | 80.48 | 46.57 | 78.86 | 74.89 | 80.64 | 86.26 | 75.64 | 74.08 | 65.92 | 85.30 | 41.23 | 74.03 |

Table S15: ACC performance of OSTC on the BOLT benchmark under different KCRs (LAR = 1.0)

| Metric | KCR | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| K-ACC | 0.25 | DAL | 87.09 | 81.43 | 52.09 | 42.75 | 60.46 | 70.18 | 83.80 | 60.54 | 63.29 | 57.48 | 79.57 | 21.04 | 63.31 |
| | | GeoID | **98.82** | 73.77 | 62.98 | 42.90 | 71.84 | 71.87 | 88.86 | 70.20 | 52.37 | 60.04 | 81.48 | 21.10 | 66.35 |
| | | SDC | 73.88 | 81.72 | 54.25 | 57.80 | 53.92 | 74.42 | 75.20 | 64.94 | 69.60 | 40.46 | 73.44 | **33.33** | 62.75 |
| | | DPN | 94.11 | 90.39 | 61.05 | 66.57 | 72.18 | **86.64** | 84.29 | 69.59 | 75.82 | 70.82 | **82.85** | 22.88 | 73.10 |
| | | TAN | 95.13 | 91.67 | 67.87 | **78.70** | 71.54 | 82.29 | 89.17 | 69.39 | 69.16 | 70.45 | 81.91 | 26.00 | 74.44 |
| | | LOOP | 95.01 | **92.30** | 60.38 | 49.94 | **78.35** | 83.09 | **93.53** | 79.31 | 71.93 | 56.06 | 77.93 | 24.32 | 71.85 |
| | | ALUP | 98.10 | 90.22 | **71.52** | 52.37 | 73.82 | 82.40 | 88.60 | 75.69 | **76.55** | **74.84** | 82.42 | 22.99 | 74.13 |
| | 0.5 | DAL | 93.52 | 85.31 | 50.87 | 52.27 | 65.31 | 81.18 | 85.67 | 69.09 | 52.32 | 70.02 | 79.51 | 18.23 | 66.94 |
| | | GeoID | 97.87 | 87.76 | 54.80 | 41.02 | 77.11 | 81.24 | 91.73 | 79.67 | 60.35 | 49.69 | 81.34 | 25.72 | 69.02 |
| | | SDC | 85.96 | 81.08 | 54.22 | 61.43 | 57.67 | 78.08 | 70.36 | 70.53 | 67.22 | 67.35 | 70.41 | **45.30** | 67.47 |
| | | DPN | 95.61 | 90.29 | 58.03 | 78.13 | 75.80 | 82.63 | 85.12 | 74.97 | 78.81 | **77.03** | 83.51 | 36.15 | 76.34 |
| | | TAN | **99.38** | **93.87** | 59.48 | **78.24** | 78.57 | 80.91 | 89.11 | 74.25 | **79.24** | 76.81 | **87.52** | 35.79 | **77.76** |
| | | LOOP | 96.57 | 93.59 | 64.53 | 56.34 | **80.29** | **87.25** | **93.88** | **80.90** | 66.27 | 51.31 | 80.39 | 21.89 | 72.77 |
| | | ALUP | 87.48 | 90.24 | **68.02** | 53.16 | 79.69 | 83.44 | 92.86 | 78.23 | 64.81 | 76.87 | 82.13 | 22.70 | 73.30 |
| | 0.75 | DAL | 95.48 | 87.27 | 64.35 | 45.91 | 66.67 | 76.28 | 87.14 | 71.10 | 59.90 | 75.42 | 80.13 | 18.17 | 68.99 |
| | | GeoID | 97.00 | 85.43 | 62.85 | 53.44 | 79.83 | 81.32 | 90.77 | 79.52 | 66.67 | 52.20 | 83.31 | 22.52 | 71.24 |
| | | SDC | 91.67 | 85.86 | 63.34 | 67.11 | 72.25 | 83.41 | 80.09 | 77.83 | 76.30 | 74.79 | 78.06 | **50.98** | 75.14 |
| | | DPN | **99.96** | 88.79 | 62.95 | 76.24 | 76.85 | 81.64 | 90.06 | 75.56 | **84.58** | 76.90 | 84.65 | 37.93 | 78.01 |
| | | TAN | 99.94 | 84.93 | 63.51 | **78.22** | 80.08 | 82.81 | 90.40 | 76.75 | 81.59 | 79.01 | **87.47** | 36.36 | **78.42** |
| | | LOOP | 99.52 | **92.39** | 64.83 | 49.26 | **83.13** | **87.83** | 93.58 | 80.42 | 66.68 | 71.15 | 81.73 | 21.45 | 74.33 |
| | | ALUP | 94.57 | 91.27 | **68.06** | 57.65 | 80.30 | 85.00 | 91.61 | 80.32 | 73.04 | **79.45** | 84.05 | 27.05 | 76.03 |

Table S16: K-ACC performance of GCD on the BOLT benchmark under different KCRs (LAR = 0.1)

| Metric | KCR | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| K-ACC | 0.25 | DAL | 87.96 | 81.19 | 52.44 | 50.61 | 65.93 | 74.21 | 86.28 | 69.55 | 60.41 | 67.73 | 79.36 | 22.56 | 66.52 |
| | | GeoID | **100.00** | 94.01 | 51.50 | 48.82 | 80.79 | 86.40 | 85.44 | 87.11 | 67.69 | 82.72 | 87.20 | 24.69 | 74.70 |
| | | SDC | 54.69 | 72.65 | 51.92 | 51.42 | 45.04 | 72.92 | 61.87 | 57.30 | 64.37 | 62.30 | 68.84 | 30.58 | 57.83 |
| | | DPN | 88.58 | 91.04 | 65.83 | **79.56** | 79.17 | **91.03** | 84.84 | 80.43 | **87.67** | 81.63 | 85.98 | 27.59 | 78.61 |
| | | TAN | 95.86 | 93.22 | 70.40 | 79.21 | 84.82 | 86.07 | 91.58 | 82.68 | 84.28 | 82.96 | **87.97** | 29.23 | 80.69 |
| | | LOOP | 92.59 | **96.62** | 72.75 | 67.75 | **87.78** | 89.65 | **96.53** | **88.80** | 79.20 | 84.30 | 86.84 | 29.15 | **81.00** |
| | | ALUP | 97.84 | 89.51 | **74.40** | 74.35 | 86.62 | 89.16 | 95.06 | 87.70 | 73.66 | **86.30** | 84.82 | **30.67** | 80.84 |
| | 0.5 | DAL | **98.77** | 83.84 | 60.00 | 54.53 | 74.93 | 81.47 | 88.76 | 76.93 | 65.11 | 82.53 | 81.74 | 21.66 | 72.52 |
| | | GeoID | 92.32 | 93.62 | 68.39 | 46.60 | 86.64 | 81.10 | 94.49 | 86.44 | 74.75 | 85.17 | 89.30 | 28.23 | 77.25 |
| | | SDC | 81.32 | 79.41 | 52.84 | 68.64 | 61.98 | 80.35 | 65.80 | 72.33 | 72.11 | 75.11 | 69.67 | **46.94** | 68.88 |
| | | DPN | 95.34 | 90.44 | 59.18 | 84.76 | 83.88 | 89.52 | 93.22 | 83.32 | 86.70 | 87.12 | 88.82 | 39.99 | 81.86 |
| | | TAN | 97.40 | **95.23** | 65.86 | **87.84** | 86.16 | 87.58 | 94.95 | 84.76 | **91.45** | 84.66 | **90.40** | 35.82 | **83.51** |
| | | LOOP | 98.76 | 94.78 | 67.66 | 64.72 | 89.22 | 90.22 | **96.64** | 87.72 | 76.82 | 86.23 | 88.62 | 26.48 | 80.66 |
| | | ALUP | 97.49 | 94.48 | **72.28** | 76.59 | **90.13** | **91.02** | 95.91 | **89.48** | 83.15 | **89.20** | 88.17 | 29.58 | 83.12 |
| | 0.75 | DAL | 97.10 | 86.95 | 61.38 | 48.21 | 76.99 | 80.51 | 92.01 | 77.89 | 71.61 | 84.66 | 83.53 | 20.85 | 73.47 |
| | | GeoID | 99.06 | 94.44 | 68.12 | 53.66 | 89.40 | 88.96 | 94.17 | **88.60** | 78.12 | 88.21 | 90.04 | 26.48 | 79.94 |
| | | SDC | 93.92 | 86.95 | 65.24 | 78.28 | 78.69 | 86.68 | 81.26 | 82.47 | 80.76 | 82.38 | 79.33 | **51.49** | 78.95 |
| | | DPN | **100.00** | 90.99 | 67.46 | **87.17** | 88.59 | 88.51 | 93.88 | 85.69 | **92.02** | 85.72 | 90.85 | 43.07 | 84.50 |
| | | TAN | 99.55 | 93.00 | 67.92 | 86.39 | 89.02 | 87.68 | 95.07 | 86.61 | 89.79 | 86.30 | **91.81** | 41.25 | **84.53** |
| | | LOOP | 95.99 | **94.45** | 69.04 | 61.31 | **91.23** | 89.61 | **96.69** | 88.04 | 80.91 | 88.31 | 87.86 | 28.38 | 80.98 |
| | | ALUP | 99.98 | 93.44 | **70.06** | 74.23 | 91.07 | **91.23** | 95.88 | 88.26 | 87.70 | **89.64** | 89.74 | 31.52 | 83.56 |

Table S17: K-ACC performance of GCD on the BOLT benchmark under different KCRs (LAR = 0.5)

| Metric | KCR | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| K-ACC | 0.25 | DAL | 86.64 | 76.05 | 57.55 | 49.65 | 71.08 | 61.07 | 84.18 | 73.45 | 62.51 | 63.35 | 79.65 | 22.51 | 65.64 |
| | | GeoID | **100.00** | 95.74 | 72.63 | 71.30 | 79.47 | 85.67 | 94.39 | 87.68 | 74.37 | 83.94 | 87.00 | 29.24 | 80.12 |
| | | SDC | 68.47 | 79.69 | 48.30 | 52.74 | 43.13 | 74.25 | 64.60 | 62.26 | 69.74 | 65.61 | 65.97 | 30.94 | 60.47 |
| | | DPN | 95.30 | 96.07 | 64.55 | 76.48 | 83.74 | **93.26** | 85.09 | 81.54 | **89.04** | 81.01 | 87.45 | 35.48 | 80.75 |
| | | TAN | 98.24 | 95.12 | 69.39 | **82.35** | 89.37 | 88.57 | 95.33 | 84.01 | 86.65 | **89.56** | 87.95 | 29.16 | 82.98 |
| | | LOOP | 98.00 | **96.22** | **76.50** | 75.20 | **91.16** | 91.20 | **97.75** | **91.65** | 77.39 | 85.88 | 87.24 | 27.47 | 82.97 |
| | | ALUP | 98.16 | 92.38 | 68.60 | 79.56 | 90.66 | 90.80 | 95.79 | 89.58 | 79.80 | 86.90 | **88.20** | **36.56** | **83.08** |
| | 0.5 | DAL | 92.18 | 82.99 | 55.42 | 54.96 | 72.89 | 78.09 | 91.02 | 77.26 | 63.01 | 83.00 | 81.35 | 23.05 | 71.27 |
| | | GeoID | **100.00** | **95.92** | 70.06 | 71.60 | 90.66 | 89.99 | 97.65 | 88.83 | 79.01 | 87.12 | 90.47 | 30.36 | 82.64 |
| | | SDC | 82.66 | 84.51 | 56.92 | 70.59 | 67.21 | 83.41 | 70.33 | 75.51 | 77.16 | 77.11 | 69.19 | **51.54** | 72.18 |
| | | DPN | 96.22 | 93.67 | 62.08 | **88.08** | 87.74 | 90.68 | 93.67 | 86.67 | 89.04 | 87.71 | 89.85 | 36.97 | 83.53 |
| | | TAN | **100.00** | 95.73 | 63.11 | 84.68 | 87.75 | 89.10 | 95.15 | 87.07 | **90.30** | **89.85** | **91.61** | 36.48 | 84.24 |
| | | LOOP | 98.46 | 95.24 | **73.28** | 74.31 | **92.58** | 90.01 | 97.32 | **90.13** | 87.41 | 86.72 | 89.50 | 30.48 | 83.79 |
| | | ALUP | 99.92 | 94.95 | 70.58 | 80.72 | 92.25 | **92.20** | **97.87** | 88.85 | 89.72 | 87.90 | 89.21 | 34.67 | **84.90** |
| | 0.75 | DAL | 93.84 | 89.22 | 66.68 | 49.70 | 78.92 | 78.45 | 92.70 | 79.13 | 70.44 | 87.42 | 82.04 | 22.34 | 74.24 |
| | | GeoID | **100.00** | 86.81 | 64.88 | 70.75 | 91.35 | 91.05 | 95.24 | 89.51 | 86.21 | **91.06** | 90.98 | 30.73 | 82.38 |
| | | SDC | 94.54 | 93.11 | 68.52 | 82.22 | 81.42 | 88.64 | 82.93 | 84.86 | 83.06 | 85.27 | 80.01 | **60.81** | 82.12 |
| | | DPN | 97.60 | 93.82 | 68.74 | **90.23** | 91.47 | 90.28 | 95.01 | 88.66 | **94.31** | 88.54 | 91.84 | 43.26 | **86.15** |
| | | TAN | **100.00** | 94.70 | 68.55 | 87.24 | 91.49 | 89.64 | 96.07 | 88.41 | 92.89 | 89.28 | **92.86** | 40.12 | 85.94 |
| | | LOOP | 97.87 | 94.63 | 70.05 | 70.99 | 92.91 | 91.88 | 96.94 | 90.14 | 87.51 | 89.37 | 88.57 | 33.96 | 83.74 |
| | | ALUP | 99.04 | **95.28** | **72.22** | 75.51 | **93.40** | **92.08** | **97.21** | **90.31** | 91.34 | 90.65 | 90.76 | 32.99 | 85.07 |

Table S18: K-ACC performance of GCD on the BOLT benchmark under different KCRs (LAR = 1.0)

| Metric | KCR | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| N-ACC | 0.25 | DAL | 63.87 | 35.52 | 19.23 | 29.98 | 46.34 | 48.19 | 67.60 | 46.34 | 42.33 | 43.79 | 74.83 | 13.52 | 44.29 |
| | | GeoID | 78.30 | 72.81 | 40.25 | 33.86 | 47.59 | 73.94 | 71.55 | 59.59 | 44.88 | 34.40 | 68.87 | 20.65 | 53.89 |
| | | SDC | 67.34 | 56.45 | 32.25 | **37.43** | 37.77 | 62.42 | 58.03 | 57.55 | 48.64 | 35.53 | 59.23 | **24.96** | 48.13 |
| | | DPN | 75.32 | 46.53 | 17.27 | 29.88 | 23.57 | 8.52 | 41.72 | 33.13 | 44.67 | 26.89 | 67.57 | 16.19 | 35.94 |
| | | TAN | 83.56 | 54.85 | 20.14 | 31.82 | 28.47 | 13.83 | 47.05 | 37.72 | 51.31 | 29.00 | 66.79 | 22.52 | 40.59 |
| | | LOOP | 86.29 | 65.64 | 55.75 | 36.94 | 63.72 | 75.61 | **82.88** | **69.47** | **63.64** | 40.77 | 70.26 | 19.16 | 60.84 |
| | | ALUP | 79.63 | **74.68** | 51.00 | 34.91 | 60.09 | 72.66 | 79.17 | 64.35 | 63.41 | 42.80 | 72.35 | 20.01 | 59.59 |
| | 0.5 | DAL | 89.38 | 40.91 | 25.44 | 36.50 | 50.81 | 62.51 | 72.30 | 52.68 | 51.80 | 50.08 | 75.36 | 9.06 | 51.40 |
| | | GeoID | 92.35 | 67.51 | 51.50 | 39.91 | 55.13 | 73.87 | 72.71 | 63.67 | 51.13 | 39.29 | 70.25 | 9.46 | 57.23 |
| | | SDC | 91.25 | **80.49** | 56.01 | **56.48** | 57.80 | 75.50 | 70.67 | 67.92 | **69.08** | 55.64 | 68.00 | 15.38 | 63.69 |
| | | DPN | 81.74 | 33.21 | 18.42 | 24.08 | 33.22 | 13.69 | 54.72 | 29.79 | 46.85 | 21.77 | 68.46 | 12.34 | 36.52 |
| | | TAN | **97.73** | 58.74 | 31.37 | 45.58 | 37.13 | 27.00 | 54.40 | 36.58 | 56.35 | 31.85 | 72.39 | **17.10** | 47.18 |
| | | LOOP | 85.07 | 68.37 | 54.01 | 38.64 | 64.98 | **78.99** | 82.25 | **73.10** | 62.54 | 42.86 | 74.12 | 12.42 | 61.45 |
| | | ALUP | 97.62 | 73.16 | 42.10 | 41.11 | 52.24 | 64.49 | 81.33 | 55.05 | 63.89 | 54.76 | 70.54 | 6.76 | 58.59 |
| | 0.75 | DAL | 86.18 | 49.69 | 20.61 | 42.34 | 51.78 | 59.32 | 72.17 | 59.64 | 49.21 | 53.11 | 74.11 | 16.65 | 52.90 |
| | | GeoID | 97.36 | 67.08 | 55.25 | 49.61 | 58.42 | 79.24 | 75.61 | 58.25 | 56.12 | 22.35 | 61.50 | 3.85 | 57.05 |
| | | SDC | 97.87 | **86.28** | **60.78** | **72.30** | **69.79** | **83.30** | 72.77 | **81.17** | **77.28** | 60.90 | **76.29** | 18.15 | **71.41** |
| | | DPN | 93.14 | 36.54 | 26.33 | 30.16 | 35.83 | 28.74 | 57.90 | 37.10 | 45.46 | 25.56 | 68.01 | 17.30 | 41.84 |
| | | TAN | **99.73** | 70.14 | 42.50 | 70.22 | 38.96 | 31.89 | 55.93 | 34.83 | 59.29 | 25.56 | 70.55 | 18.20 | 51.48 |
| | | LOOP | 82.55 | 66.49 | 53.25 | 53.54 | 66.27 | 77.14 | **83.96** | 69.80 | 62.91 | 62.01 | 72.66 | **20.69** | 64.27 |
| | | ALUP | 98.57 | 45.70 | 48.85 | 52.37 | 66.45 | 68.29 | 72.98 | 70.62 | 65.82 | 27.06 | 69.03 | 3.85 | 57.47 |

Table S19: N-ACC performance of GCD on the BOLT benchmark under different KCRs (LAR = 0.1)

| Metric | KCR | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| N-ACC | 0.25 | DAL | 70.21 | 37.92 | 19.37 | 30.88 | 48.09 | 48.21 | 71.89 | 47.72 | 46.98 | 50.43 | 76.32 | 17.99 | 47.17 |
| | | GeoID | 89.64 | 69.22 | 45.56 | 36.89 | 63.32 | 75.75 | 76.78 | 61.78 | 50.98 | 40.80 | 73.42 | 19.94 | 58.67 |
| | | SDC | 69.80 | 62.22 | 41.87 | **46.80** | 41.85 | 72.90 | 58.74 | 60.09 | 59.67 | 41.04 | 63.36 | **26.64** | 53.75 |
| | | DPN | 75.27 | 43.02 | 19.35 | 28.00 | 24.63 | 4.92 | 42.78 | 24.08 | 45.52 | 23.58 | 66.60 | 25.96 | 35.31 |
| | | TAN | 86.13 | 57.64 | 24.99 | 33.09 | 35.03 | 19.82 | 53.39 | 29.27 | 55.69 | 29.66 | 73.98 | 23.47 | 43.51 |
| | | LOOP | 87.77 | 67.01 | 52.95 | 38.15 | **65.86** | **77.51** | **83.71** | **70.64** | 61.28 | **59.28** | 74.73 | 19.98 | 63.24 |
| | | ALUP | **96.25** | **71.42** | **53.21** | 37.65 | 65.31 | 77.01 | 81.24 | 68.94 | **66.97** | 59.20 | **77.79** | 20.25 | **64.60** |
| | 0.5 | DAL | 92.73 | 46.22 | 29.74 | 38.83 | 53.55 | 60.42 | 73.73 | 59.42 | 53.18 | 56.36 | **77.16** | 10.18 | 54.29 |
| | | GeoID | 95.79 | 69.83 | 52.10 | 42.67 | 56.60 | 74.92 | 71.91 | 64.22 | 54.61 | 43.45 | 69.07 | 4.05 | 58.27 |
| | | SDC | 88.55 | **83.26** | **60.66** | **61.26** | 60.51 | **84.51** | 67.93 | **73.78** | **76.99** | **66.86** | 71.45 | **15.75** | **67.63** |
| | | DPN | 84.93 | 44.20 | 23.58 | 30.20 | 28.00 | 12.22 | 44.83 | 27.91 | 51.64 | 18.70 | 70.38 | 13.14 | 37.48 |
| | | TAN | 94.67 | 60.88 | 28.56 | 33.72 | 37.26 | 32.29 | 53.80 | 33.13 | 51.46 | 29.56 | 73.35 | 15.10 | 45.31 |
| | | LOOP | 83.12 | 57.30 | 55.26 | 42.69 | **66.11** | 77.10 | **85.57** | 72.55 | 65.42 | 64.25 | 75.78 | 14.58 | 63.31 |
| | | ALUP | **96.76** | 68.87 | 52.13 | 46.11 | 65.83 | 80.14 | 81.68 | 71.68 | 67.06 | 61.93 | 76.83 | 13.45 | 65.21 |
| | 0.75 | DAL | 94.74 | 52.14 | 37.85 | 45.39 | 52.89 | 59.27 | 74.74 | 62.07 | 61.09 | 56.43 | 77.19 | 17.99 | 57.65 |
| | | GeoID | **98.60** | 49.63 | 23.50 | 52.63 | 57.37 | 68.56 | 72.46 | 59.65 | 56.96 | 11.76 | 65.93 | 0.00 | 51.42 |
| | | SDC | 94.53 | **91.18** | **70.40** | **64.57** | **76.60** | **88.55** | 79.33 | **85.26** | **85.40** | **75.53** | **77.78** | 17.74 | **75.57** |
| | | DPN | 94.40 | 37.65 | 31.45 | 32.85 | 32.87 | 25.65 | 55.51 | 33.11 | 44.06 | 24.48 | 67.95 | 16.52 | 41.37 |
| | | TAN | 95.20 | 55.93 | 36.78 | 42.63 | 39.08 | 54.91 | 55.86 | 38.80 | 48.86 | 36.83 | 72.82 | 17.80 | 49.63 |
| | | LOOP | 92.50 | 73.20 | 47.40 | 55.59 | 63.48 | 78.82 | 82.74 | 72.97 | 62.78 | 64.43 | 75.58 | 18.03 | 65.63 |
| | | ALUP | 94.42 | 79.15 | 53.05 | 51.55 | 67.76 | 81.16 | **85.16** | 70.31 | 64.74 | 52.83 | 77.23 | **19.91** | 66.44 |

Table S20: N-ACC performance of GCD on the BOLT benchmark under different KCRs (LAR = 0.5)

| Metric | KCR | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| N-ACC | 0.25 | DAL | 64.23 | 34.86 | 18.85 | 33.60 | 49.62 | 52.47 | 70.66 | 50.61 | 52.24 | 48.31 | 77.08 | 16.86 | 47.45 |
| | | GeoID | 83.44 | 69.71 | 47.56 | 38.16 | 55.43 | 75.75 | 75.42 | 68.15 | 59.02 | 36.00 | 69.38 | 21.34 | 58.28 |
| | | SDC | 67.75 | **72.13** | 48.14 | **44.55** | 42.91 | 76.40 | 61.03 | 60.64 | 67.87 | 48.13 | 61.53 | **25.20** | 56.36 |
| | | DPN | 72.77 | 36.97 | 16.81 | 27.15 | 23.27 | 4.48 | 42.63 | 20.02 | 44.67 | 24.29 | 67.69 | 20.60 | 33.45 |
| | | TAN | 89.41 | 56.95 | 28.53 | 31.36 | 32.38 | 15.83 | 50.82 | 29.61 | 50.47 | 27.84 | 73.03 | 22.12 | 42.36 |
| | | LOOP | 83.89 | 62.37 | 52.45 | 38.57 | 64.83 | 79.16 | **84.23** | **70.55** | 62.28 | **65.46** | 74.67 | 19.42 | 63.16 |
| | | ALUP | **98.45** | 70.39 | **53.84** | 38.37 | **67.54** | **79.56** | 83.56 | 70.18 | **69.48** | 60.78 | **78.54** | 18.06 | **65.73** |
| | 0.5 | DAL | 88.41 | 38.98 | 26.64 | 41.26 | 54.78 | 65.10 | 72.41 | 60.16 | 52.50 | 50.92 | **78.04** | 9.89 | 53.26 |
| | | GeoID | 99.11 | 64.27 | 45.00 | 44.28 | 56.03 | 80.77 | 74.84 | 71.17 | 54.89 | 41.67 | 67.30 | 5.41 | 58.73 |
| | | SDC | 90.04 | **86.75** | **62.14** | **67.88** | 64.04 | **86.57** | 67.95 | **77.71** | **79.75** | 63.22 | 68.04 | 16.07 | **69.18** |
| | | DPN | 84.60 | 43.13 | 31.62 | 29.07 | 24.90 | 11.51 | 51.00 | 29.72 | 54.13 | 25.45 | 68.68 | 16.48 | 39.19 |
| | | TAN | 96.42 | 55.69 | 32.48 | 32.13 | 39.63 | 33.10 | 55.09 | 33.52 | 53.79 | 30.28 | 74.56 | 16.26 | 46.08 |
| | | LOOP | 87.93 | 70.42 | 50.88 | 45.69 | 68.53 | 83.04 | 82.88 | 75.16 | 62.37 | 59.84 | 75.63 | 13.50 | 64.66 |
| | | ALUP | **99.79** | 70.12 | 51.58 | 41.92 | **69.41** | 76.63 | **84.68** | 73.71 | 62.72 | **65.12** | 77.71 | **16.66** | 65.84 |
| | 0.75 | DAL | 82.91 | 44.67 | 34.35 | 49.24 | 51.39 | 56.18 | 70.14 | 58.31 | 57.63 | 53.16 | 75.88 | 21.20 | 54.59 |
| | | GeoID | **100.00** | 72.78 | 50.50 | 51.09 | 64.34 | 77.18 | 82.98 | 68.07 | 54.43 | 30.59 | 65.04 | 0.00 | 59.75 |
| | | SDC | 95.05 | **89.62** | **71.75** | **57.57** | **79.42** | **89.32** | 80.32 | **88.98** | **86.41** | **79.54** | **77.70** | 16.24 | **75.99** |
| | | DPN | 89.76 | 32.76 | 31.15 | 25.97 | 32.32 | 22.68 | 54.35 | 31.11 | 46.06 | 30.01 | 70.87 | 21.46 | 40.71 |
| | | TAN | 95.24 | 49.69 | 34.44 | 38.74 | 39.46 | 53.11 | 54.18 | 37.69 | 50.01 | 32.17 | 72.29 | 18.90 | 47.99 |
| | | LOOP | 95.54 | 68.82 | 41.75 | 53.76 | 64.95 | 78.63 | **84.98** | 73.85 | 63.56 | 63.90 | 74.74 | 20.72 | 65.43 |
| | | ALUP | 99.78 | 66.61 | 43.20 | 47.79 | 69.18 | 78.71 | 84.70 | 73.05 | 66.64 | 59.38 | 76.89 | **21.54** | 65.62 |

Table S21: N-ACC performance of GCD on the BOLT benchmark under different KCRs (LAR = 1.0)

| Metric | KCR | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| H-Score | 0.25 | DAL | 73.44 | 48.88 | 27.74 | 33.33 | 52.37 | 55.88 | 74.76 | 52.38 | 50.00 | 49.19 | 77.09 | 14.83 | 50.82 |
| | | GeoID | 87.83 | 74.47 | 51.04 | 42.04 | 57.25 | 73.29 | 80.20 | 64.46 | 48.33 | 43.86 | 74.63 | 18.00 | 59.62 |
| | | SDC | 68.73 | 64.10 | 33.36 | 40.66 | 38.87 | 66.59 | 63.38 | 59.21 | 55.95 | 36.77 | 63.90 | **27.38** | 51.57 |
| | | DPN | 83.25 | 60.57 | 25.55 | 40.71 | 35.27 | 15.09 | 55.66 | 44.19 | 55.47 | 37.12 | 74.12 | 18.07 | 45.42 |
| | | TAN | 87.76 | 66.33 | 30.40 | **45.12** | 40.18 | 23.32 | 60.87 | 47.99 | 58.41 | 40.13 | 73.40 | 18.04 | 49.33 |
| | | LOOP | 90.16 | 76.45 | 57.24 | 40.51 | 70.22 | 79.04 | **87.86** | **73.91** | 66.85 | 45.94 | 73.76 | 18.61 | 65.05 |
| | | ALUP | 88.66 | **83.75** | **59.96** | 43.55 | 66.25 | 77.22 | 83.62 | 69.56 | **68.62** | **55.83** | **77.36** | 18.45 | **66.07** |
| | 0.5 | DAL | 91.34 | 55.27 | 33.46 | 40.47 | 57.01 | 70.57 | 78.38 | 59.73 | 51.47 | 57.19 | 77.33 | 10.74 | 56.91 |
| | | GeoID | 95.03 | 77.52 | 53.10 | 44.20 | 64.29 | 77.38 | 82.52 | 70.78 | 58.80 | 43.88 | 77.10 | 13.98 | 63.22 |
| | | SDC | 88.06 | 79.93 | 52.72 | 54.84 | 53.15 | 76.34 | 67.85 | 68.65 | 67.25 | 57.69 | 66.57 | 15.38 | 62.37 |
| | | DPN | 87.90 | 47.29 | 25.48 | 36.65 | 46.03 | 23.31 | 66.31 | 42.22 | 58.02 | 33.73 | 75.18 | 16.31 | 46.54 |
| | | TAN | **98.50** | 72.11 | 40.22 | **57.12** | 50.40 | 40.44 | 67.55 | 48.98 | 65.80 | 44.92 | 79.21 | **20.69** | 57.16 |
| | | LOOP | 89.98 | 78.80 | 58.49 | 42.82 | 71.75 | **82.80** | **87.66** | **76.72** | 64.25 | 45.96 | 77.06 | 13.85 | 65.84 |
| | | ALUP | 93.21 | **80.81** | 53.19 | 48.00 | 63.60 | 72.75 | 87.00 | 64.62 | **67.27** | **64.73** | 75.90 | 11.00 | 65.17 |
| | 0.75 | DAL | 89.83 | 61.91 | 30.58 | 42.72 | 58.12 | 66.47 | 78.91 | 64.76 | 53.72 | 60.71 | 76.75 | 14.04 | 58.21 |
| | | GeoID | 97.10 | 78.30 | 59.42 | 50.95 | 67.80 | 80.15 | 82.50 | 67.24 | 61.21 | 31.56 | 71.21 | 6.57 | 62.83 |
| | | SDC | 94.56 | **85.87** | **59.75** | 67.49 | 68.75 | **83.25** | 70.72 | **78.73** | **76.42** | 61.85 | 74.72 | **23.37** | 70.46 |
| | | DPN | 95.99 | 48.05 | 32.32 | 43.05 | 48.57 | 40.25 | 70.44 | 49.57 | 57.94 | 36.95 | 75.04 | 22.61 | 51.73 |
| | | TAN | **99.84** | 76.77 | 50.50 | **72.74** | 52.35 | 45.60 | 69.04 | 47.81 | 68.28 | 38.22 | 77.99 | 22.58 | 60.14 |
| | | LOOP | 88.59 | 76.90 | 57.35 | 48.11 | **73.67** | 81.80 | **88.45** | 74.54 | 64.26 | 63.49 | 76.77 | 17.75 | 67.64 |
| | | ALUP | 96.53 | 61.47 | 56.88 | 56.55 | 72.72 | 75.73 | 81.24 | 75.99 | 69.25 | 40.95 | 76.10 | 6.73 | 64.18 |

Table S22: H-Score performance of GCD on the BOLT benchmark under different KCRs (LAR = 0.1)

| Metric | KCR | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| H-Score | 0.25 | DAL | 77.93 | 51.19 | 27.99 | 37.08 | 55.47 | 58.15 | 78.42 | 56.29 | 52.28 | 57.05 | 77.76 | 16.00 | 53.80 |
| | | GeoID | 94.54 | **80.38** | 48.35 | 46.27 | 71.00 | 80.77 | 81.93 | 73.46 | 58.15 | 54.65 | 79.70 | 19.62 | 65.73 |
| | | SDC | 57.75 | 66.29 | 41.37 | 45.12 | 38.44 | 72.39 | 58.36 | 57.28 | 61.09 | 43.15 | 63.74 | 22.50 | 52.29 |
| | | DPN | 81.14 | 57.54 | 29.08 | 40.52 | 37.39 | 9.31 | 56.28 | 36.32 | 59.31 | 34.15 | 74.91 | **25.62** | 45.13 |
| | | TAN | 90.62 | 70.71 | 35.40 | 46.30 | 49.50 | 31.73 | 67.37 | 43.11 | 66.61 | 43.30 | 80.35 | 20.12 | 53.76 |
| | | LOOP | 89.70 | 78.86 | 60.75 | 46.80 | **75.23** | **83.05** | **89.66** | **78.66** | 68.77 | 69.22 | 80.27 | 21.11 | 70.17 |
| | | ALUP | **96.93** | 79.29 | **61.65** | **49.48** | 74.41 | 82.62 | 87.52 | 76.97 | **69.37** | **69.64** | **81.06** | 21.87 | **70.90** |
| | 0.5 | DAL | 95.57 | 59.24 | 39.73 | 43.85 | 62.42 | 69.06 | 80.53 | 67.03 | 58.46 | 65.80 | 79.30 | 12.36 | 61.11 |
| | | GeoID | 94.02 | 79.54 | **61.39** | 50.81 | 68.47 | 81.64 | 81.67 | 75.15 | 64.09 | 57.98 | 78.56 | 7.09 | 66.70 |
| | | SDC | 84.02 | **80.40** | 53.89 | **58.44** | 55.46 | 82.29 | 61.93 | 71.43 | 72.75 | 68.80 | 68.19 | 16.81 | 64.53 |
| | | DPN | 89.53 | 58.95 | 32.97 | 44.07 | 41.76 | 20.96 | 60.32 | 41.70 | 64.37 | 30.22 | 78.48 | 17.45 | 48.40 |
| | | TAN | 95.95 | 73.41 | 38.66 | 48.36 | 51.78 | 46.17 | 68.54 | 47.46 | 65.60 | 43.73 | 80.98 | **18.17** | 56.57 |
| | | LOOP | 90.04 | 70.19 | 60.55 | 48.26 | 75.92 | 82.96 | **90.76** | 79.29 | 70.36 | **73.22** | 81.67 | 16.43 | 69.97 |
| | | ALUP | **97.09** | 79.34 | 59.88 | 55.52 | **76.02** | **85.19** | 88.13 | **79.47** | **74.12** | 72.40 | **82.04** | 16.83 | **72.17** |
| | 0.75 | DAL | 95.90 | 63.78 | 45.75 | 44.45 | 62.20 | 67.77 | 82.37 | 68.73 | 65.20 | 65.95 | 80.03 | 15.40 | 63.13 |
| | | GeoID | **98.83** | 65.28 | 34.95 | 54.26 | 69.89 | 77.44 | 81.90 | 71.39 | 65.88 | 20.76 | 76.24 | 0.00 | 59.74 |
| | | SDC | 93.95 | **88.97** | **67.35** | **62.73** | 74.35 | **87.56** | 75.40 | **83.07** | **82.12** | **74.76** | 75.17 | 21.43 | **73.90** |
| | | DPN | 96.75 | 51.95 | 41.80 | 46.69 | 47.51 | 39.71 | 69.67 | 47.54 | 58.59 | 37.26 | 77.56 | 21.14 | 53.01 |
| | | TAN | 97.05 | 68.20 | 47.01 | 55.87 | 54.12 | 67.11 | 70.19 | 53.32 | 62.71 | 50.44 | 81.05 | **21.47** | 60.71 |
| | | LOOP | 94.19 | 82.30 | 55.70 | 56.03 | 74.59 | 83.55 | 89.15 | 79.61 | 70.26 | 72.33 | 81.14 | 19.21 | 71.51 |
| | | ALUP | 96.76 | 85.60 | 59.92 | 59.62 | **77.43** | 85.86 | **90.16** | 78.04 | 74.37 | 64.27 | **82.96** | 21.35 | 73.03 |

Table S23: H-Score performance of GCD on the BOLT benchmark under different KCRs (LAR = 0.5)

| Metric | KCR | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| H-Score | 0.25 | DAL | 73.50 | 47.42 | 28.30 | 39.08 | 58.38 | 55.02 | 76.80 | 59.74 | 56.63 | 54.27 | 78.29 | 16.00 | 53.62 |
| | | GeoID | 90.97 | **80.48** | 59.93 | 49.51 | 65.31 | 80.40 | 83.84 | 76.63 | 65.82 | 50.83 | 79.21 | 22.14 | 67.09 |
| | | SDC | 67.02 | 75.44 | 43.54 | 44.47 | 39.49 | 74.66 | 61.49 | 60.04 | 67.19 | 51.54 | 61.42 | 22.22 | 55.71 |
| | | DPN | 82.49 | 51.92 | 25.11 | 38.59 | 35.84 | 8.52 | 56.40 | 31.66 | 58.09 | 37.02 | 76.11 | **23.70** | 43.79 |
| | | TAN | 93.50 | 70.45 | 39.10 | 45.20 | 47.49 | 26.83 | 66.16 | 43.67 | 63.59 | 42.34 | 79.74 | 18.74 | 53.07 |
| | | LOOP | 90.33 | 75.59 | **61.83** | 48.67 | 75.73 | 84.58 | **90.48** | **79.71** | 68.39 | **73.98** | 80.37 | 19.60 | 70.77 |
| | | ALUP | **98.30** | 79.56 | 59.92 | **51.56** | **77.31** | **84.78** | 89.21 | 78.59 | **74.18** | 71.11 | **83.04** | 23.38 | **72.58** |
| | 0.5 | DAL | 90.18 | 52.89 | 35.68 | 45.40 | 62.42 | 70.91 | 80.64 | 67.51 | 57.02 | 61.83 | 79.61 | 12.22 | 59.69 |
| | | GeoID | 99.55 | 77.36 | 56.08 | 53.91 | 69.25 | 85.09 | 84.75 | 78.99 | 65.94 | 56.37 | 77.76 | 9.18 | 67.85 |
| | | SDC | 85.44 | **85.06** | 56.09 | **65.24** | 61.20 | 84.84 | 64.35 | 75.46 | **76.89** | 64.75 | 64.52 | 17.38 | 66.77 |
| | | DPN | 89.70 | 56.60 | 41.57 | 43.30 | 38.67 | 19.59 | 65.78 | 43.59 | 67.08 | 38.94 | 77.82 | 20.76 | 50.28 |
| | | TAN | 98.10 | 70.01 | 41.76 | 45.92 | 54.56 | 46.79 | 69.45 | 48.15 | 67.30 | 45.12 | 82.17 | 20.30 | 57.47 |
| | | LOOP | 92.30 | 80.78 | **59.90** | 53.55 | 78.74 | **86.26** | 89.48 | **81.94** | 72.74 | 69.86 | 81.95 | 17.04 | 72.05 |
| | | ALUP | **99.85** | 80.55 | 59.11 | 53.97 | **79.20** | 83.47 | **90.79** | 80.54 | 73.75 | **74.45** | **83.02** | **21.43** | **73.34** |
| | 0.75 | DAL | 87.57 | 59.12 | 44.69 | 48.04 | 61.71 | 64.90 | 79.72 | 66.91 | 63.09 | 64.90 | 78.63 | 19.21 | 61.54 |
| | | GeoID | **100.00** | 81.83 | 56.79 | 57.93 | 75.59 | 83.42 | 88.69 | 78.23 | 66.15 | 45.79 | 75.86 | 0.00 | 67.52 |
| | | SDC | 94.50 | **90.73** | **69.79** | 53.18 | 77.94 | **88.94** | 77.96 | **86.58** | **83.95** | **79.51** | 75.46 | 23.57 | **75.18** |
| | | DPN | 93.22 | 43.99 | 40.78 | 39.88 | 47.55 | 33.90 | 69.07 | 45.78 | 60.85 | 44.31 | 79.86 | **25.21** | 52.03 |
| | | TAN | 97.23 | 64.25 | 44.99 | 52.26 | 54.84 | 66.32 | 69.09 | 52.52 | 64.27 | 46.89 | 81.22 | 22.40 | 59.69 |
| | | LOOP | 96.68 | 78.35 | 51.30 | **59.46** | 76.30 | 84.40 | **90.54** | 81.00 | 73.47 | 73.10 | 80.96 | 22.72 | 72.36 |
| | | ALUP | 99.40 | 78.08 | 53.07 | 56.98 | **79.41** | 84.72 | 90.51 | 80.63 | 76.93 | 69.88 | **83.16** | 22.28 | 72.92 |

Table S24: H-Score performance of GCD on the BOLT benchmark under different KCRs (LAR = 1.0)

| Metric | KCR | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ARI | 0.25 | DAL | 63.10 | 33.83 | 8.76 | 18.46 | 37.74 | 38.46 | 62.57 | 36.89 | 36.36 | 33.77 | 73.93 | 3.52 | 37.28 |
| | | GeoID | 77.00 | 60.78 | 26.76 | 27.40 | 41.24 | 51.27 | 69.57 | 50.80 | 38.46 | 26.88 | 68.56 | 7.00 | 45.48 |
| | | SDC | 58.80 | 44.50 | 12.32 | 36.35 | 29.07 | 41.49 | 48.76 | 45.28 | 41.73 | 17.17 | 51.16 | **17.47** | 37.01 |
| | | DPN | 76.82 | 44.23 | 8.82 | 29.83 | 23.84 | 12.74 | 39.76 | 26.63 | 40.70 | 19.43 | 64.50 | 3.46 | 32.56 |
| | | TAN | 83.39 | 51.48 | 11.35 | **61.21** | 27.45 | 16.54 | 46.06 | 29.46 | 45.93 | 22.05 | 66.10 | 15.77 | 39.73 |
| | | LOOP | 86.91 | 61.33 | 34.04 | 23.63 | **56.95** | **63.15** | **79.66** | **61.23** | 51.48 | 26.55 | 68.08 | 7.09 | 51.68 |
| | | ALUP | 82.20 | **70.67** | 32.99 | 26.99 | 53.30 | 58.86 | 75.39 | 55.79 | **57.22** | **34.32** | 72.66 | 7.74 | **52.34** |
| | 0.5 | DAL | 90.60 | 48.89 | 18.64 | 24.53 | 47.12 | 56.41 | 71.46 | 48.38 | 41.76 | 47.40 | 76.10 | 4.20 | 47.96 |
| | | GeoID | 92.38 | 69.13 | 33.61 | 32.05 | 55.16 | 60.82 | 77.10 | 59.72 | 45.65 | 25.20 | 74.93 | 9.47 | 52.94 |
| | | SDC | 84.26 | 65.86 | 30.18 | 48.42 | 45.30 | 59.37 | 59.85 | 55.90 | **59.41** | 43.76 | 60.47 | **18.80** | 52.63 |
| | | DPN | 87.81 | 47.52 | 18.04 | 39.65 | 42.07 | 26.88 | 60.90 | 36.27 | 55.98 | 31.45 | 70.93 | 12.80 | 44.19 |
| | | TAN | **98.23** | 64.96 | 23.52 | **70.14** | 44.06 | 37.46 | 62.34 | 37.14 | 58.93 | 35.64 | 78.45 | 15.64 | 52.21 |
| | | LOOP | 90.90 | 68.82 | 36.49 | 25.91 | 62.21 | **70.10** | **82.91** | **66.85** | 49.92 | 27.65 | 74.20 | 6.75 | 55.23 |
| | | ALUP | 93.71 | **74.79** | 34.35 | 39.24 | 55.20 | 57.84 | 81.28 | 55.32 | 54.16 | **49.07** | 76.67 | 7.82 | **56.62** |
| | 0.75 | DAL | 93.69 | 65.48 | 32.66 | 31.81 | 52.71 | 59.26 | 77.05 | 56.31 | 47.37 | 58.65 | 77.39 | 6.81 | 54.93 |
| | | GeoID | 97.26 | 72.94 | 39.00 | 39.49 | 65.55 | 64.44 | 82.10 | 61.77 | 54.20 | 29.75 | 78.48 | 10.68 | 57.97 |
| | | SDC | 92.05 | 74.56 | 37.89 | 65.28 | 59.96 | 69.02 | 70.46 | 66.31 | 69.40 | 55.72 | 71.48 | **29.52** | 63.47 |
| | | DPN | 98.34 | 65.05 | 33.04 | 60.66 | 54.47 | 43.04 | 74.85 | 51.59 | **69.62** | 46.30 | 76.69 | 18.29 | 57.66 |
| | | TAN | **99.82** | 72.39 | 35.58 | **72.22** | 58.45 | 54.31 | 74.36 | 51.39 | 69.34 | 47.96 | **81.52** | 20.67 | 61.50 |
| | | LOOP | 95.76 | 74.76 | 40.15 | 31.58 | **69.42** | **74.25** | **86.93** | 68.53 | 50.79 | 53.06 | 76.65 | 8.51 | 60.87 |
| | | ALUP | 94.39 | 71.03 | **40.92** | 48.14 | 67.00 | 67.09 | 80.88 | **69.33** | 58.99 | 56.40 | 80.03 | 10.92 | 62.09 |

Table S25: ARI performance of GCD on the BOLT benchmark under different KCRs (LAR = 0.1)

| Metric | KCR | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ARI | 0.25 | DAL | 70.64 | 34.50 | 9.06 | 20.73 | 41.24 | 41.04 | 67.23 | 39.55 | 38.78 | 41.09 | 75.18 | 4.80 | 40.32 |
| | | GeoID | 88.77 | 64.78 | 26.86 | 25.75 | 57.09 | 62.76 | 73.54 | 57.23 | 48.63 | 34.52 | 73.84 | 7.96 | 51.81 |
| | | SDC | 54.78 | 45.82 | 18.10 | **36.89** | 31.42 | 52.56 | 46.77 | 45.83 | 47.35 | 26.66 | 53.32 | **18.37** | 39.82 |
| | | DPN | 77.10 | 42.21 | 11.13 | 30.98 | 27.31 | 12.78 | 40.65 | 19.70 | 41.98 | 19.78 | 62.97 | 9.03 | 32.97 |
| | | TAN | 87.04 | 55.20 | 15.05 | 30.15 | 36.21 | 22.43 | 52.48 | 27.95 | 50.08 | 25.71 | 72.75 | 17.87 | 41.08 |
| | | LOOP | 87.88 | 62.72 | **35.05** | 25.81 | **61.56** | **69.68** | **81.65** | **65.41** | 49.54 | **49.93** | 74.58 | 7.86 | 55.97 |
| | | ALUP | **96.32** | **66.15** | 33.66 | 32.25 | 59.72 | 65.20 | 78.53 | 62.35 | **59.54** | 49.02 | **77.73** | 8.20 | **57.39** |
| | 0.5 | DAL | 95.32 | 50.05 | 23.39 | 29.66 | 53.23 | 57.09 | 74.66 | 56.42 | 44.86 | 59.81 | 78.42 | 5.50 | 52.37 |
| | | GeoID | 92.40 | 71.10 | **39.23** | 37.77 | 60.97 | 70.53 | 77.63 | 65.73 | 55.30 | 49.26 | 78.88 | 11.24 | 59.17 |
| | | SDC | 82.49 | 67.35 | 33.51 | **54.02** | 51.74 | 67.51 | 58.38 | 60.08 | **65.09** | 53.84 | 61.72 | **18.84** | 56.21 |
| | | DPN | 90.58 | 55.96 | 22.14 | 45.85 | 44.39 | 37.78 | 59.37 | 39.95 | 60.92 | 38.27 | 74.14 | 13.02 | 48.53 |
| | | TAN | 95.77 | 66.85 | 26.00 | 48.07 | 51.44 | 44.57 | 66.22 | 44.88 | 61.09 | 40.74 | 77.52 | 15.70 | 53.24 |
| | | LOOP | 90.87 | 63.74 | 37.53 | 32.41 | **68.80** | **73.78** | **86.26** | **71.99** | 59.87 | **62.22** | 78.80 | 8.57 | 61.24 |
| | | ALUP | **97.26** | **71.63** | 37.04 | 46.10 | 68.23 | 72.58 | 83.88 | 69.68 | 63.88 | 61.46 | **80.66** | 8.72 | **63.43** |
| | 0.75 | DAL | 96.72 | 66.86 | 33.44 | 32.74 | 61.30 | 63.17 | 82.46 | 63.59 | 53.27 | 67.52 | 80.83 | 7.42 | 59.11 |
| | | GeoID | **99.14** | 75.14 | 39.41 | 39.49 | 72.90 | 73.96 | 85.44 | 70.19 | 65.40 | 57.53 | 83.62 | 12.30 | 64.54 |
| | | SDC | 94.00 | 78.53 | 42.68 | **75.53** | 68.77 | 75.62 | 74.56 | 72.29 | 74.46 | 65.93 | 73.51 | **31.45** | 68.94 |
| | | DPN | 98.86 | 67.41 | 37.62 | 69.90 | 65.32 | 59.84 | 78.32 | 60.46 | 72.26 | 56.64 | 80.28 | 24.39 | 64.28 |
| | | TAN | 98.43 | 73.18 | 38.66 | 65.18 | 66.71 | 66.11 | 79.64 | 62.65 | 70.72 | 59.60 | 83.21 | 22.34 | 65.54 |
| | | LOOP | 95.14 | 78.56 | 42.47 | 42.79 | 76.48 | 77.61 | 89.76 | **76.33** | 65.99 | **70.22** | 81.93 | 10.30 | 67.30 |
| | | ALUP | 98.52 | **80.70** | **43.73** | 64.98 | **77.12** | **77.82** | **89.82** | 75.26 | **74.68** | 68.35 | **83.65** | 12.69 | **70.61** |

Table S26: ARI performance of GCD on the BOLT benchmark under different KCRs (LAR = 0.5)

| Metric | KCR | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ARI | 0.25 | DAL | 65.78 | 28.97 | 9.78 | 21.78 | 42.86 | 40.18 | 65.67 | 43.78 | 42.08 | 38.45 | 75.80 | 5.82 | 40.08 |
| | | GeoID | 85.53 | **64.28** | 31.55 | 30.11 | 50.77 | 63.73 | 72.75 | 62.13 | 51.10 | 32.51 | 73.56 | 8.98 | 52.25 |
| | | SDC | 58.37 | 59.58 | 21.13 | **37.23** | 31.91 | 55.58 | 49.22 | 46.77 | 56.53 | 33.85 | 49.51 | 17.16 | 43.07 |
| | | DPN | 76.21 | 39.03 | 9.70 | 26.80 | 27.59 | 13.25 | 41.19 | 19.58 | 42.50 | 20.70 | 64.23 | 6.88 | 32.30 |
| | | TAN | 89.88 | 54.99 | 17.20 | 31.70 | 35.80 | 20.92 | 51.17 | 28.70 | 46.23 | 26.44 | 71.57 | **17.68** | 41.02 |
| | | LOOP | 86.41 | 58.44 | **34.56** | 26.46 | 62.10 | **71.96** | **82.48** | **66.44** | 52.06 | **55.23** | 74.11 | 8.58 | 56.57 |
| | | ALUP | **98.27** | 63.56 | 32.71 | 32.89 | **62.14** | 68.20 | 80.97 | 64.17 | **64.00** | 50.90 | **78.78** | 9.56 | **58.85** |
| | 0.5 | DAL | 90.81 | 46.04 | 20.92 | 29.84 | 53.90 | 56.65 | 74.99 | 56.63 | 44.64 | 55.87 | 78.04 | 5.58 | 51.16 |
| | | GeoID | 99.08 | 70.63 | 35.66 | 44.84 | 64.39 | 72.97 | 80.15 | 70.50 | 55.23 | 48.84 | 78.89 | 12.02 | 61.10 |
| | | SDC | 84.70 | **74.14** | 35.55 | **58.56** | 56.30 | 71.50 | 60.96 | 64.90 | **69.52** | 51.73 | 60.20 | **20.03** | 59.01 |
| | | DPN | 90.57 | 56.41 | 24.86 | 46.58 | 44.35 | 37.42 | 64.50 | 42.45 | 63.74 | 41.37 | 72.69 | 15.41 | 50.03 |
| | | TAN | 98.26 | 62.06 | 26.87 | 38.72 | 53.15 | 45.84 | 67.49 | 46.33 | 60.00 | 44.81 | 79.38 | 14.99 | 53.16 |
| | | LOOP | 93.32 | 71.31 | **38.93** | 39.95 | 71.40 | **76.61** | 85.76 | **74.10** | 66.31 | 62.27 | 79.47 | 9.34 | 64.06 |
| | | ALUP | **99.70** | 69.47 | 36.13 | 48.42 | **71.66** | 72.88 | **86.79** | 71.98 | 68.69 | **62.29** | **81.26** | 13.15 | **65.20** |
| | 0.75 | DAL | 90.72 | 64.86 | 38.10 | 39.31 | 62.10 | 59.88 | 82.22 | 63.90 | 53.05 | 69.63 | 79.74 | 9.73 | 59.44 |
| | | GeoID | **100.00** | 75.92 | 41.95 | 56.27 | 76.65 | 78.16 | 88.60 | 75.60 | 68.54 | 61.79 | 84.54 | 17.19 | 68.77 |
| | | SDC | 94.55 | **84.66** | **45.09** | **81.30** | 73.28 | 78.05 | 76.32 | 76.48 | **77.55** | 70.48 | 74.83 | **31.06** | **71.97** |
| | | DPN | 95.66 | 69.71 | 39.71 | 69.60 | 67.77 | 60.24 | 79.30 | 61.33 | 73.33 | 60.68 | 82.68 | 27.67 | 65.64 |
| | | TAN | 98.99 | 72.16 | 40.01 | 69.86 | 68.89 | 67.08 | 79.96 | 64.00 | 72.50 | 61.36 | 84.16 | 21.05 | 66.67 |
| | | LOOP | 97.42 | 77.28 | 42.50 | 54.53 | 78.88 | **79.44** | **90.95** | **78.55** | 75.23 | **72.06** | 82.46 | 15.89 | 70.43 |
| | | ALUP | 99.25 | 78.31 | 43.08 | 61.35 | **79.64** | 78.92 | 90.76 | 77.92 | 77.32 | 70.65 | **84.92** | 15.03 | 71.43 |

Table S27: ARI performance of GCD on the BOLT benchmark under different KCRs (LAR = 1.0)

| Metric | KCR | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NMI | 0.25 | DAL | 80.41 | 49.71 | 14.10 | 56.76 | 70.47 | 51.67 | 88.24 | 71.80 | 67.14 | 58.07 | 92.33 | 26.64 | 60.61 |
| | | GeoID | 85.74 | 72.83 | 34.46 | 63.61 | 72.13 | 68.22 | 91.24 | 78.25 | 67.77 | 51.59 | 90.75 | 33.54 | 67.51 |
| | | SDC | 79.04 | 63.37 | 27.16 | 60.78 | 67.44 | 61.87 | 84.86 | 76.85 | 68.51 | 42.40 | 86.88 | **36.12** | 62.94 |
| | | DPN | 90.22 | 61.47 | 15.62 | 59.58 | 61.79 | 27.63 | 79.68 | 64.10 | 65.82 | 43.42 | 91.02 | 24.00 | 57.03 |
| | | TAN | 91.14 | 65.69 | 17.76 | **69.36** | 62.79 | 27.84 | 81.10 | 65.26 | 66.07 | 46.61 | 90.78 | 27.47 | 59.32 |
| | | LOOP | 94.17 | 73.53 | 42.32 | 65.33 | **81.89** | **74.77** | **94.11** | **83.90** | **80.19** | 50.91 | 90.71 | 34.62 | 72.20 |
| | | ALUP | 90.59 | **79.30** | 41.74 | 64.41 | 79.79 | 71.73 | 92.87 | 81.59 | 79.60 | 57.73 | 92.16 | 35.40 | **72.24** |
| | 0.5 | DAL | 96.00 | 63.23 | 25.37 | 63.10 | 76.23 | 67.20 | 91.50 | 77.55 | 72.46 | 69.28 | 92.97 | 27.31 | 68.52 |
| | | GeoID | 95.61 | 78.27 | 40.44 | 69.33 | 80.24 | 73.80 | 93.32 | 82.98 | 76.35 | 49.95 | 92.56 | **36.16** | 72.42 |
| | | SDC | 92.78 | 76.10 | 38.40 | 70.16 | 76.07 | 69.17 | 88.98 | 81.20 | 77.02 | 65.74 | 89.25 | 34.99 | 71.66 |
| | | DPN | 95.78 | 65.83 | 27.72 | 65.56 | 73.93 | 48.94 | 89.20 | 71.87 | 73.24 | 53.90 | 92.23 | 28.64 | 65.57 |
| | | TAN | **99.10** | 75.64 | 30.94 | **77.34** | 73.79 | 48.72 | 88.04 | 71.48 | 76.54 | 58.20 | 93.54 | 30.30 | 68.64 |
| | | LOOP | 96.71 | 78.09 | **43.90** | 67.84 | **84.40** | **77.56** | **95.05** | **86.11** | 79.55 | 51.93 | 92.33 | 34.13 | 73.97 |
| | | ALUP | 97.40 | **81.61** | 42.25 | 71.15 | 80.83 | 72.31 | 94.39 | 81.63 | **79.79** | 67.35 | 92.96 | 35.38 | **74.75** |
| | 0.75 | DAL | 97.91 | 75.77 | 39.46 | 68.02 | 79.38 | 72.44 | 93.24 | 81.27 | 74.38 | **77.45** | 93.16 | 34.21 | 73.89 |
| | | GeoID | 98.83 | 80.56 | 45.16 | 72.80 | 85.06 | 75.72 | 94.93 | 83.99 | 78.72 | 52.00 | 93.47 | 38.94 | 75.02 |
| | | SDC | 96.76 | 81.09 | 43.69 | 77.09 | 82.55 | 74.43 | 91.89 | 85.65 | 80.89 | 72.89 | 91.88 | **41.31** | 76.68 |
| | | DPN | 99.23 | 77.11 | 39.53 | 74.74 | 79.66 | 66.69 | 92.88 | 79.64 | 79.05 | 66.66 | 93.46 | 35.92 | 73.71 |
| | | TAN | **99.88** | 80.82 | 41.39 | **80.87** | 81.14 | 62.65 | 92.21 | 78.61 | 80.30 | 66.30 | **94.20** | 36.23 | 74.55 |
| | | LOOP | 98.54 | 81.39 | 45.68 | 72.30 | **87.10** | **79.20** | **96.18** | 86.98 | 80.14 | 71.42 | 93.13 | 37.86 | 77.49 |
| | | ALUP | 97.93 | 80.04 | **46.21** | 76.90 | 86.23 | 76.16 | 94.88 | 86.79 | **82.53** | 74.90 | 94.08 | 37.97 | **77.89** |

Table S28: NMI performance of GCD on the BOLT benchmark under different KCRs (LAR = 0.1)

| Metric | KCR | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NMI | 0.25 | DAL | 85.70 | 50.92 | 15.31 | 60.13 | 72.95 | 54.91 | 89.93 | 72.80 | 70.57 | 64.78 | 92.78 | 29.92 | 63.39 |
| | | GeoID | 93.19 | 76.31 | 35.53 | 64.02 | 80.96 | 73.90 | 92.58 | 81.70 | 76.34 | 56.98 | 92.45 | 35.91 | 71.66 |
| | | SDC | 77.36 | 64.61 | 32.59 | 63.63 | 70.50 | 67.63 | 84.97 | 77.73 | 72.92 | 53.42 | 88.21 | 36.60 | 65.85 |
| | | DPN | 90.49 | 60.21 | 17.00 | 62.78 | 66.55 | 27.62 | 82.56 | 61.17 | 68.60 | 46.71 | 91.53 | 28.53 | 58.65 |
| | | TAN | 94.30 | 68.20 | 22.56 | 63.57 | 68.80 | 33.30 | 84.23 | 63.97 | 74.20 | 50.08 | 92.47 | 30.39 | 62.17 |
| | | LOOP | 95.09 | 74.45 | **42.99** | 67.82 | **84.01** | **77.18** | **94.69** | **85.61** | 80.08 | **69.08** | 92.56 | 36.70 | 75.02 |
| | | ALUP | **98.09** | **77.73** | 41.88 | **67.97** | 83.02 | 74.61 | 93.77 | 84.14 | **80.98** | 68.36 | **93.35** | **37.13** | **75.09** |
| | 0.5 | DAL | 98.25 | 64.49 | 30.48 | 68.89 | 79.45 | 67.89 | 92.56 | 81.48 | 76.10 | **79.05** | 93.46 | 31.44 | 71.96 |
| | | GeoID | 96.85 | 79.04 | 44.31 | 72.29 | 83.07 | 77.50 | 93.70 | 85.48 | 79.50 | 68.75 | 93.77 | **38.73** | 76.08 |
| | | SDC | 92.78 | 76.86 | 40.54 | **75.66** | 80.19 | 75.14 | 88.63 | 83.63 | 81.42 | 71.57 | 89.74 | 36.36 | 74.38 |
| | | DPN | 96.81 | 71.28 | 30.93 | 70.61 | 77.59 | 49.65 | 88.68 | 74.26 | 77.27 | 61.39 | 93.42 | 32.03 | 68.66 |
| | | TAN | 98.25 | 76.46 | 32.47 | 72.23 | 78.00 | 54.56 | 89.76 | 74.92 | 78.20 | 61.67 | 93.58 | 32.63 | 70.23 |
| | | LOOP | 96.48 | 75.05 | **44.41** | 72.37 | **87.25** | **79.64** | **95.93** | **88.35** | 83.03 | 77.00 | 93.75 | 37.73 | 77.58 |
| | | ALUP | **99.02** | **79.98** | 43.88 | 74.12 | 86.70 | 77.96 | 95.39 | 87.10 | **83.37** | 76.83 | **94.15** | 37.48 | **78.00** |
| | 0.75 | DAL | 98.82 | 76.68 | 40.97 | 70.91 | 83.53 | 74.46 | 94.93 | 84.62 | 80.34 | **82.45** | 94.19 | 35.64 | 76.46 |
| | | GeoID | **99.71** | 83.52 | 45.50 | 76.85 | 88.49 | 80.41 | 96.12 | 87.38 | 84.08 | 74.93 | 94.89 | 41.31 | 79.43 |
| | | SDC | 98.10 | 83.92 | 47.75 | **83.02** | 87.40 | 79.65 | 93.19 | 88.63 | 85.97 | 79.30 | 92.37 | **42.22** | 80.13 |
| | | DPN | 99.58 | 78.79 | 44.36 | 80.49 | 84.99 | 67.77 | 94.25 | 83.49 | 83.64 | 74.15 | 94.74 | 40.25 | 77.21 |
| | | TAN | 99.38 | 81.30 | 43.99 | 79.10 | 85.23 | 72.79 | 94.18 | 84.04 | 83.36 | 75.15 | **95.07** | 38.92 | 77.71 |
| | | LOOP | 98.41 | 83.62 | 47.85 | 77.24 | 90.13 | **81.67** | 97.02 | **89.90** | 85.08 | 82.30 | 94.61 | 38.47 | 80.52 |
| | | ALUP | 99.18 | **85.14** | **48.33** | 80.56 | **90.15** | 81.46 | **97.08** | 89.49 | **86.54** | 81.02 | 94.97 | 39.71 | **81.14** |

Table S29: NMI performance of GCD on the BOLT benchmark under different KCRs (LAR = 0.5)

| Metric | KCR | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NMI | 0.25 | DAL | 83.99 | 45.43 | 16.39 | 59.96 | 73.79 | 54.03 | 89.52 | 75.13 | 72.38 | 62.58 | 92.78 | 32.01 | 63.17 |
| | | GeoID | 92.33 | 75.34 | 39.24 | 67.61 | 78.09 | 72.46 | 92.07 | 84.07 | 77.89 | 56.17 | 92.42 | 35.98 | 71.97 |
| | | SDC | 80.09 | 72.31 | 35.78 | 64.22 | 70.82 | 70.48 | 85.69 | 78.58 | 78.29 | 58.32 | 87.38 | 34.21 | 68.01 |
| | | DPN | 90.50 | 58.02 | 16.63 | 61.49 | 66.78 | 27.53 | 82.42 | 59.97 | 66.73 | 46.72 | 90.83 | 29.43 | 58.09 |
| | | TAN | 95.24 | 67.57 | 24.33 | 64.43 | 68.74 | 32.31 | 83.70 | 64.54 | 71.17 | 49.26 | 92.35 | 29.96 | 61.97 |
| | | LOOP | 94.26 | 72.02 | **43.13** | **68.41** | **84.53** | **78.48** | **94.98** | **86.31** | 80.31 | **72.31** | 92.61 | **37.54** | 75.41 |
| | | ALUP | **98.96** | **75.84** | 41.35 | 68.34 | 84.18 | 76.19 | 94.48 | 84.91 | **82.68** | 69.28 | **93.63** | 37.50 | **75.61** |
| | 0.5 | DAL | 96.78 | 60.51 | 29.84 | 68.64 | 80.34 | 69.87 | 92.73 | 81.71 | 75.90 | 76.19 | 93.50 | 32.08 | 71.51 |
| | | GeoID | 99.25 | 79.73 | 42.40 | 73.12 | 84.81 | 78.58 | 94.34 | 87.50 | 80.95 | 70.08 | 93.86 | **40.09** | 77.06 |
| | | SDC | 93.96 | **80.83** | 43.06 | **77.66** | 82.86 | 77.65 | 89.32 | 85.90 | **84.71** | 71.67 | 89.05 | 37.28 | 76.16 |
| | | DPN | 96.67 | 72.49 | 35.37 | 71.35 | 78.53 | 51.49 | 90.53 | 74.82 | 78.44 | 62.78 | 93.35 | 33.10 | 69.91 |
| | | TAN | 99.31 | 72.49 | 33.96 | 70.35 | 79.02 | 55.84 | 90.30 | 75.63 | 78.06 | 64.25 | 94.12 | 33.14 | 70.54 |
| | | LOOP | 97.30 | 79.67 | **45.87** | 74.42 | **88.13** | **80.91** | 95.93 | **89.31** | 84.39 | **78.28** | 93.96 | 37.82 | **78.83** |
| | | ALUP | **99.76** | 78.78 | 43.89 | 75.56 | 88.12 | 78.44 | **96.17** | 88.12 | 83.46 | 76.96 | **94.22** | 40.02 | 78.62 |
| | 0.75 | DAL | 96.46 | 75.64 | 43.98 | 72.64 | 84.12 | 73.56 | 95.10 | 84.74 | 80.53 | **84.29** | 94.02 | 38.01 | 76.92 |
| | | GeoID | **100.00** | 83.58 | 46.99 | 79.04 | 90.18 | 81.40 | 97.02 | 89.87 | 85.49 | 77.47 | 95.28 | 42.86 | 80.77 |
| | | SDC | 98.20 | **87.32** | **50.01** | **85.96** | 89.35 | 81.38 | 93.68 | 90.47 | **88.87** | 81.95 | 92.81 | **43.52** | **81.96** |
| | | DPN | 98.69 | 80.70 | 45.87 | 81.72 | 86.41 | 68.56 | 94.89 | 84.80 | 84.87 | 76.47 | 95.22 | 41.51 | 78.31 |
| | | TAN | 99.55 | 80.59 | 45.23 | 81.35 | 86.47 | 72.93 | 94.51 | 84.66 | 84.63 | 75.81 | 95.34 | 40.59 | 78.47 |
| | | LOOP | 99.11 | 83.17 | 47.86 | 80.72 | **91.10** | **82.85** | **97.48** | **90.99** | 87.32 | 83.68 | 94.65 | 42.98 | 81.83 |
| | | ALUP | 99.72 | 84.32 | 48.26 | 81.22 | 91.00 | 82.42 | 97.27 | 90.65 | 87.68 | 82.58 | **95.39** | 42.92 | 81.95 |

Table S30: NMI performance of GCD on the BOLT benchmark under different KCRs (LAR = 1.0)

| Metric | KCR | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ACC | 0.25 | DAL | 69.62 | 48.59 | 25.80 | 30.60 | 49.83 | 53.68 | 71.70 | 49.91 | 46.42 | 47.22 | 76.20 | 12.32 | 48.49 |
| | | GeoID | 83.52 | 74.77 | 46.15 | 38.96 | 53.57 | 74.04 | 76.53 | 62.21 | 48.37 | 40.79 | 72.37 | 17.00 | 57.36 |
| | | SDC | 69.11 | 63.65 | 36.65 | 47.50 | 41.76 | 65.42 | 62.38 | 59.54 | 55.24 | 36.64 | 63.42 | **32.07** | 52.78 |
| | | DPN | 80.03 | 59.00 | 26.02 | 41.21 | 35.56 | 28.05 | 52.50 | 42.27 | 52.47 | 37.71 | 72.05 | 15.83 | 45.23 |
| | | TAN | 86.40 | 65.43 | 29.69 | **60.71** | 39.09 | 30.96 | 57.72 | 45.62 | 55.62 | 39.16 | 71.68 | 32.02 | 51.17 |
| | | LOOP | 88.53 | 73.22 | 56.68 | 37.24 | 67.33 | 77.48 | **85.58** | **71.92** | 64.64 | 44.47 | 72.52 | 17.19 | 63.07 |
| | | ALUP | 84.53 | **80.64** | 55.35 | 39.96 | 63.47 | 75.10 | 81.56 | 67.15 | **66.63** | **51.96** | 76.05 | 18.10 | **63.38** |
| | 0.5 | DAL | 91.47 | 63.04 | 38.16 | 38.60 | 57.96 | 71.84 | 78.99 | 60.82 | 52.85 | 60.03 | 77.88 | 13.10 | 58.73 |
| | | GeoID | 94.96 | 79.14 | 53.15 | 41.02 | 65.97 | 77.55 | 84.04 | 71.22 | 58.83 | 44.41 | 77.67 | 19.57 | 63.96 |
| | | SDC | 88.72 | 80.77 | 55.12 | 59.28 | 57.74 | 76.79 | 70.52 | 69.11 | 69.28 | 61.30 | 70.36 | **34.16** | 66.10 |
| | | DPN | 88.60 | 61.66 | 38.23 | 50.27 | 54.23 | 48.16 | 69.92 | 52.26 | 64.14 | 49.45 | 76.49 | 26.87 | 56.69 |
| | | TAN | **98.56** | 76.28 | 45.42 | **68.89** | 57.58 | 53.95 | 71.75 | 54.56 | 68.14 | 54.06 | **81.96** | 29.23 | 63.37 |
| | | LOOP | 90.98 | 80.95 | 59.27 | 40.45 | 72.54 | **83.12** | **88.07** | **76.98** | 64.11 | 47.06 | 77.76 | 16.77 | 66.51 |
| | | ALUP | 93.95 | **81.70** | 57.15 | 51.18 | 66.56 | 73.96 | 87.42 | 65.99 | 66.58 | **66.77** | 78.20 | 18.46 | **67.33** |
| | 0.75 | DAL | 93.35 | 76.47 | 55.61 | 41.75 | 62.99 | 72.04 | 83.35 | 68.38 | 56.88 | 69.76 | 79.26 | 15.71 | 64.63 |
| | | GeoID | 97.10 | 82.41 | 61.69 | 50.73 | 75.26 | 80.80 | 86.93 | 73.64 | 63.98 | 45.62 | 81.19 | 21.55 | 68.41 |
| | | SDC | 93.22 | 85.97 | 62.82 | 70.16 | 71.64 | 83.38 | 78.23 | 78.49 | **77.40** | 71.23 | 79.21 | **45.88** | 74.80 |
| | | DPN | 98.41 | 73.81 | 55.62 | 64.99 | 66.73 | 68.41 | 81.91 | 66.19 | 74.95 | 63.98 | 81.24 | 32.84 | 69.09 |
| | | TAN | **99.90** | 80.67 | 59.30 | **74.91** | 69.94 | 70.08 | 81.66 | 65.42 | 75.10 | 65.33 | **85.22** | 31.41 | 71.58 |
| | | LOOP | 95.67 | 84.95 | 62.51 | 45.10 | **78.97** | 85.16 | **91.14** | 77.91 | 64.71 | 68.78 | 79.95 | 19.55 | 71.20 |
| | | ALUP | 95.70 | 80.00 | **65.15** | 57.76 | 76.88 | 80.82 | 86.89 | **79.46** | 72.30 | 69.49 | 82.97 | 23.08 | 72.54 |

Table S31: ACC performance of GCD on the BOLT benchmark under different KCRs (LAR = 0.1)

| Metric | KCR | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ACC | 0.25 | DAL | 74.67 | 50.22 | 25.98 | 32.06 | 52.50 | 54.71 | 75.54 | 53.13 | 50.27 | 54.73 | 77.39 | 14.23 | 51.29 |
| | | GeoID | 92.13 | **76.89** | 46.75 | 39.23 | 67.63 | 78.46 | 80.40 | 68.90 | 58.78 | 51.06 | 77.62 | 18.88 | 63.06 |
| | | SDC | 66.23 | 65.22 | 43.88 | **50.07** | 42.63 | 72.91 | 59.54 | 59.54 | 61.85 | 46.09 | 65.43 | **35.26** | 55.72 |
| | | DPN | 78.64 | 56.69 | 28.64 | 44.59 | 38.08 | 26.45 | 53.44 | 38.18 | 56.05 | 37.97 | 72.42 | 22.06 | 46.10 |
| | | TAN | 88.63 | 67.75 | 34.07 | 43.25 | 47.31 | 36.38 | 63.07 | 42.77 | 62.15 | 42.96 | 77.78 | 34.44 | 53.38 |
| | | LOOP | 89.07 | 75.44 | 56.91 | 40.79 | **71.27** | **80.54** | **86.96** | **75.23** | 64.55 | 65.46 | 78.28 | 18.42 | 66.91 |
| | | ALUP | **96.66** | 76.58 | **57.45** | 44.37 | 70.57 | 80.05 | 84.74 | 73.63 | **68.76** | **65.89** | **80.11** | 19.34 | **68.18** |
| | 0.5 | DAL | 95.76 | 64.98 | 44.87 | 42.00 | 64.10 | 70.94 | 81.24 | 68.21 | 58.83 | 69.51 | 79.97 | 15.03 | 62.95 |
| | | GeoID | 94.15 | **81.68** | **63.40** | 49.39 | 71.43 | 82.17 | 83.20 | 76.65 | 66.29 | 64.95 | 82.33 | 22.42 | 69.84 |
| | | SDC | 85.06 | 81.31 | 56.75 | **65.18** | 61.24 | 82.43 | 66.86 | 73.00 | **75.94** | 70.82 | 71.31 | **35.62** | 68.79 |
| | | DPN | 90.11 | 67.26 | 41.38 | 58.69 | 55.58 | 50.87 | 69.03 | 55.62 | 69.80 | 53.04 | 80.07 | 26.68 | 59.84 |
| | | TAN | 96.07 | 78.03 | 47.21 | 58.86 | 61.39 | 59.93 | 74.37 | 58.89 | 71.47 | 57.16 | 82.13 | 27.92 | 64.45 |
| | | LOOP | 91.03 | 75.94 | 61.46 | 46.69 | 77.52 | 83.66 | **91.11** | 80.26 | 71.65 | 75.27 | 82.50 | 19.97 | 71.42 |
| | | ALUP | **97.11** | 81.64 | 62.21 | 55.71 | **77.82** | **85.58** | 88.80 | **80.46** | 75.58 | **75.57** | **83.32** | 19.73 | **73.63** |
| | 0.75 | DAL | 96.52 | 76.96 | 56.67 | 43.14 | 71.04 | 75.20 | 87.64 | 74.13 | 67.57 | 77.57 | 82.51 | 17.62 | 68.88 |
| | | GeoID | **98.95** | 82.19 | 59.20 | 54.08 | 81.49 | 83.86 | 88.67 | 80.81 | 75.63 | 68.58 | 86.50 | 23.80 | 73.65 |
| | | SDC | 94.13 | 88.15 | 66.27 | **80.16** | 78.18 | 87.15 | 80.77 | 82.97 | **83.07** | 80.54 | 80.71 | **48.42** | **79.21** |
| | | DPN | 98.74 | 75.72 | 60.26 | 75.92 | 74.84 | 72.79 | 84.16 | 72.75 | 78.93 | 70.33 | 85.57 | 36.73 | 73.90 |
| | | TAN | 98.57 | 82.38 | 61.69 | 71.86 | 76.70 | 79.49 | 85.14 | 74.81 | 78.76 | 73.84 | **87.28** | 33.77 | 75.36 |
| | | LOOP | 95.14 | 88.35 | 64.71 | 55.59 | 84.38 | 86.91 | 93.15 | **84.38** | 74.95 | **82.24** | 85.21 | 22.60 | 76.47 |
| | | ALUP | 98.72 | **89.34** | **66.66** | 66.65 | **85.32** | **88.71** | **93.16** | 83.93 | 81.04 | 80.26 | 86.90 | 25.53 | 78.85 |

Table S32: ACC performance of GCD on the BOLT benchmark under different KCRs (LAR = 0.5)

| Metric | KCR | Method | 20NG | THUCNews | Yahoo | TREC | BANK | S.O. | CLINC | HWU | ECDT | MCID | DBPedia | X-Topic | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ACC | 0.25 | DAL | 69.80 | 46.59 | 26.59 | 34.49 | 54.92 | 54.62 | 74.08 | 56.36 | 54.28 | 52.06 | 78.03 | 14.51 | 51.36 |
| | | GeoID | 87.42 | **77.16** | 54.25 | 42.53 | 61.36 | 78.23 | 80.22 | 73.08 | 63.25 | 48.34 | 77.25 | 21.49 | 63.72 |
| | | SDC | 67.95 | 74.29 | 48.17 | **49.88** | 42.96 | 75.86 | 61.93 | 61.10 | 69.28 | 52.38 | 63.30 | **34.70** | 58.48 |
| | | DPN | 78.43 | 53.81 | 26.36 | 41.55 | 38.19 | 26.68 | 53.39 | 35.56 | 56.07 | 38.43 | 73.12 | 19.98 | 45.13 |
| | | TAN | 91.60 | 67.79 | 36.70 | 44.42 | 46.44 | 34.02 | 62.10 | 43.42 | 59.39 | 43.20 | 77.46 | 33.59 | 53.34 |
| | | LOOP | 87.42 | 72.01 | **57.26** | 41.51 | 71.33 | 82.17 | **87.66** | **75.91** | 65.75 | **70.51** | 78.15 | 20.81 | 67.54 |
| | | ALUP | **98.39** | 76.64 | 56.79 | 46.37 | **73.25** | **82.37** | 86.66 | 75.04 | **72.53** | 67.25 | **81.32** | 22.48 | **69.92** |
| | 0.5 | DAL | 90.33 | 60.92 | 41.03 | 43.30 | 63.72 | 71.59 | 81.71 | 68.64 | 57.74 | 66.99 | 80.10 | 15.00 | 61.76 |
| | | GeoID | 99.53 | 80.71 | 59.70 | 55.02 | 73.12 | 85.38 | 86.27 | 79.98 | 67.20 | 64.05 | 82.51 | 24.01 | 71.46 |
| | | SDC | 86.48 | **85.61** | 59.53 | **68.45** | 65.60 | 84.99 | 69.14 | 76.61 | **79.27** | 69.97 | 69.84 | **37.46** | 71.08 |
| | | DPN | 90.35 | 68.35 | 46.85 | 58.74 | 55.91 | 51.09 | 72.34 | 58.04 | 72.12 | 56.62 | 79.61 | 28.36 | 61.53 |
| | | TAN | 98.24 | 75.64 | 47.79 | 52.95 | 63.37 | 61.10 | 75.12 | 60.21 | 71.91 | 60.12 | 83.67 | 28.20 | 64.86 |
| | | LOOP | 93.38 | 82.80 | **62.08** | 53.55 | 80.40 | **86.52** | 90.10 | **82.68** | 74.93 | 73.29 | 82.90 | 20.40 | 73.59 |
| | | ALUP | **99.85** | 82.52 | 61.08 | 57.47 | **80.68** | 84.41 | **91.27** | 81.30 | 76.36 | **76.56** | **83.90** | 24.84 | **75.02** |
| | 0.75 | DAL | 91.16 | 76.46 | 60.21 | 46.53 | 72.13 | 72.89 | 86.99 | 74.03 | 66.16 | 78.77 | 81.04 | 19.67 | 68.84 |
| | | GeoID | **100.00** | 84.10 | 62.00 | 62.53 | 84.73 | 87.83 | 92.13 | 84.65 | 76.18 | 75.53 | 87.22 | 28.72 | 77.14 |
| | | SDC | 94.74 | **92.10** | **69.17** | **83.84** | 80.93 | **88.81** | 82.27 | 85.76 | **84.66** | **83.75** | 81.23 | **48.86** | **81.34** |
| | | DPN | 95.71 | 76.35 | 61.22 | 76.74 | 76.88 | 73.38 | 84.71 | 74.53 | 80.34 | 73.84 | 87.04 | 40.17 | 75.08 |
| | | TAN | 98.92 | 81.80 | 61.73 | 75.10 | 78.65 | 80.51 | 85.46 | 75.72 | 80.94 | 74.93 | **88.17** | 31.94 | 76.16 |
| | | LOOP | 97.27 | 87.21 | 64.39 | 63.76 | 86.01 | 88.56 | 93.91 | **86.24** | 80.76 | 82.90 | 85.45 | 28.01 | 78.71 |
| | | ALUP | 99.22 | 87.08 | 66.42 | 66.94 | **87.42** | 88.73 | **94.04** | 86.11 | 83.88 | 82.78 | 87.70 | 26.53 | 79.74 |

Table S33: ACC performance of GCD on the BOLT benchmark under different KCRs (LAR = 1.0)