Knowledge-driven Augmentation and Retrieval for Integrative Temporal Adaptation

Weisi Liu
University of Memphis
wliu9@memphis.edu

Guangzeng Han
University of Memphis
ghan@memphis.edu

Xiaolei Huang
University of Memphis
xiaolei.huang@memphis.edu

Abstract

Time introduces fundamental challenges in model development and deployment: models are usually trained on historical data while deployed on future data where semantic distributions and domain knowledge may evolve. Unfortunately, existing studies either overlook temporal shifts or hardly capture rich shifting patterns of both semantic and knowledge. We develop Knowledge-driven Augmentation and Retrieval for Integrative Temporal Adaptation (KARITA) to capture diverse temporal shifts (e.g., uncertainty and feature shift), construct and integrate rich knowledge sources (e.g., medical ontology like MeSH), and leverage shifting insights for selecting-retrieval augmented learning. We evaluate KARITA on classification tasks across multiple domains, clinical, legal, and scientific corpora, demonstrating consistent improvements across multiple domains with temporal adaptation. Our results show that knowledge integration can be more critical and effective in temporal augmentation and learning.

1 Introduction

Standard training and validation settings commonly overlook varying patterns between training and real-world deployment by randomly splitting data into multiple sets with the same distribution. Time plays a critical role in evolving data and its distributions when data corpora span over years and seasons, where we can call the varying patterns between training and test as *temporal shifts*. Studies have shown such temporal shifts can be complex and its cause can be heterogeneous (Liu et al., 2024; Agarwal and Nenkova, 2022). For example, (Liu et al., 2024) found that biomedical classifiers perform much worse when time spans are closer while other tasks may perform worse when time spans are farther (Liu et al., 2025); and such patterns may vary across other domains. Therefore, a question to be answered is: *How can we capture the complexity and heterogeneity of temporal shift and further mitigate its effects on models?*

Temporal learning, a framework to understand and integrate temporal shifts, aims to promote consistent model performance over periods of time. In contrast to test-time adaptation focusing on discrete data splits (Shi et al., 2024), the framework will learn evolving patterns over multiple continuous time splits. Existing studies in this track primarily follow changing semantic meanings or representations, such as word sense shifting (Su et al., 2022) or feature embedding distances (Huang and Paul, 2019; Röttger and Pierrehumbert, 2021). For example, Huang and Paul (2019) model temporal shifts by learning diachronic word embeddings and adapting classifiers across time intervals, treating changes in word usage and representation spaces as the primary signal of temporal variation. However, the unified feature representations may not be adequate to fully capture the diversity and heterogeneity of temporal shifts and therefore mislead model learning steps, particularly when corpus domains are versatile and high-stakes. For example, in clinical and legal document classification, temporal shifts may stem from multiple co-occurring factors, such as evolving practices, changing conventions, and shifts in disease prevalence, with different types of shifts dominating at different time periods. Treating these heterogeneous shifts as a single unified change can obscure their individual impacts and lead to unstable model performance over time. Thus, a second question to be answered by our study is: *can temporal learning better inform and keep model robust over time?*

To answer the questions, we present *K*nowledge-driven *A*ugmentation and *R*etrieval method for *I*ntegrative *T*emporal *A*daptation (*KARITA*) that discover shifting patterns from diverse aspects, enable time-aware learning process, and promote data quality by knowledge-driven augmentations over

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 36130–36143

Dataset	Time Span	Task	Labels	Data Size
MIMIC	2008–2019	Phenotype inference	Top 50 frequent ICD-10 codes	331,794 clinical notes
EurLex	1958–2016	Legal topic classification	21 Top-level EuroVoc categories	65,000 legal documents
arXiv-CS	1991–2025	Scientific paper categorization	Top 30 CS subject categories	622,419 paper abstracts

Table 1: Overview of three time-varying datasets with time span, task, labels, and data size.

time. We introduce three major modules to achieve our goals, multi-aspect shift detector, source backtracking retrieval, and data augmentation. We construct domain knowledge and validate our approach with multiple state-of-the-art methods in temporal learning (T.y.s.s et al., 2024), closer work (Agarwal and Nenkova, 2022), and without adapting time (Niu et al., 2022, 2023) over health, legal, and scientific domains. Our results highlight temporal shifts come diverse sources beyond regular embedding features, inform model with temporal shifts can be critical, and domain knowledge is efficient to augment model performance in temporal learning.[1]

2 Related Work

Data Drift Data drift is common in real-world NLP applications, and prior studies have repeatedly observed that temporal mismatches between training and test data can impair model performance (Agarwal and Nenkova, 2022; Liu et al., 2025) Such drift can stem from multiple sources: e.g., word senses shifting over time (Kutuzov et al., 2018; Tang et al., 2023), changes in data distributions (Guo et al., 2023), or factual knowledge updates as real-world entities and relationships evolve (Jang et al., 2022). Correspondingly, many studies have focused on specific types of shifts to improve model generalization, including topic shift (Huang et al., 2018), lexical semantic adaptation (Su et al., 2022), temporal concept shift (Chalkidis and Søgaard, 2022; Margatina et al., 2023), and overall data distribution drift (Guo et al., 2023). However, these different types of shifts do not typically occur in isolation; rather, they often co-occur across time periods. Methods that focus on a single shift type may overlook other concurrent changes, thereby limiting adaptation effectiveness. In contrast, our method explicitly detects multiple shift patterns and addresses them through targeted data augmentation that combines the generative capabilities of large language models (Han et al., 2025; Rao et al., 2026) and external domain Knowledge.

[1]Our code is available at https://github.com/trust-nlp/TemporalLearning-KARITA.

Temporal Learning Existing efforts to address temporal drift can be broadly categorized into three directions. *Model-centric* approaches modify network architectures to explicitly encode time or capture time-varying patterns (Liu et al., 2025). *Feature-centric* methods focus on learning temporally robust or aligned representations to mitigate distributional mismatch across time (Dhingra et al., 2022). *Training-centric* solutions redesign learning strategies, including time-aware training curricula (T.y.s.s et al., 2024) and continual learning frameworks (Röttger and Pierrehumbert, 2021; Agarwal and Nenkova, 2022; Shang et al., 2022). In contrast, *data-centric* approaches that explicitly manipulate training data to cope with temporal drift remain relatively underexplored as a primary lever for temporal learning.

A common limitation of most existing approaches is that temporal adaptation is often treated as a one-time or stage-wise adjustment process, implicitly assuming relatively clear boundaries between temporal regimes (Liu et al., 2025; Röttger and Pierrehumbert, 2021). However, in realistic temporal scenarios, distributional drift is typically gradual, overlapping, and heterogeneous, which challenges such assumptions and limits the effectiveness of one-shot adaptation strategies. Our approach departs from prior work in two key aspects. First, instead of performing one-off adaptation, we model temporal learning as an iterative and selective process that continuously identifies and reacts to emerging shifts. Second, we adopt a data-centric perspective that explicitly retrieves and augments data conditioned on detected temporal shifts.

3 Data

Biomedical, legal, and scientific domains are characterized by continuous evolution of terminology and knowledge, which poses substantial challenges for text classification models deployed over time. We select three corpora from these domains to systematically study temporal shifts and adaptation in multi-label classification tasks: phenotype inference from clinical notes (MIMIC-IV-Notes), legal topic classification (EurLex), and scientific docu-

ment classification (arXiv-CS). Table 1 reports the main statistics of the datasets.

MIMIC-IV-Notes (MIMIC) (Johnson et al., 2023; Goldberger et al., 2000) is a collection of de-identified clinical notes including discharge summaries and radiology reports. We choose the discharge summaries and focus on the phenotype inference task. Each discharge summary is annotated with International Classification of Diseases (ICD) codes, which indicate the presence of diseases, symptoms, injuries, and other health conditions. The earlier notes are annotated with ICD-9 codes, and we convert them to ICD-10 using a standard ICD mapping toolkit [2] for consistency. We choose the 50 most frequent ICD-10 codes as labels. The de-identification procedure in MIMIC-IV assigns each note to a three-year time interval, spanning from 2008 to 2022, naturally forming five temporal domains. Due to data sparsity in the final interval, We only use the first four intervals (2008–2019) in our experiments. Detailed data partitioning is provided in Appendix A.1.

EurLex (Chalkidis et al., 2021) is a large-scale legal corpus consisting of European Union laws published between 1958 and 2016. The documents are annotated with concepts from the EuroVoc (Walhain et al., 2025) taxonomy by the Publications Office of the European Union. Each EuroVoc concept is associated with a semantic descriptor, and documents are originally labeled with one or more concepts drawn from different levels of the hierarchy. EuroVoc is organized into eight hierarchical levels. In this work, we focus on the top level of the EuroVoc concept, which contains 21 legal domain categories (e.g. Economics), and focus the English portion of the corpus.

arXiv-CS is a scientific document corpus derived from arXiv abstracts in the computer science category, covering publications from 1991 to 2025. Each paper is associated with one or more subject categories that reflect its research area (e.g., *cs.LG*, *cs.CV*, *cs.AI*). We choose the 30 most frequent computer science subdomains as the labels.

4 Method

In this section, we present our KARITA framework in Figure 1. The key idea is to automatically detect shifts from target data and use the detected shifted samples to retrieve similar source samples and use LLM augmented knowledge and external ontology thesaurus to augment the retrieved source data, and use the augmented data to adapt the model. It consists of three major modules: 1) Shift Detection, 2) Source Backtracking Retrieval, and 3) Knowledge-driven Data Augmentation. We include detailed workflow in Algorithm 1 under the Appendix.

4.1 Shift Detection

For a target instance x, we compute three normalized shift scores capturing uncertainty, feature-level deviation, and ontology-based terminology drift. These scores are then combined to determine whether x requires retrieval.

Uncertainty-Based Shift Model uncertainty is often associated with distribution shift, since out-of-distribution samples are typically harder to classify. Maximum predicted probability reflects the model's confidence, while predictive entropy measures prediction dispersion. Combining them provides a reliable indicator of inputs likely to deviate from the training distribution. Let $p_l(x) = \sigma(z_l(x))$ be the sigmoid probability for label l, where $z_l(x)$ is the model's logit output. We define the maximum predicted probability as $p(x) = \max_l p_l(x)$ and the average binary entropy as $H(x) = \frac{1}{L}\sum_{l=1}^{L} -\big[p_l(x)\log p_l(x) + (1-p_l(x))\log(1-p_l(x))\big]$, where L is the total number of labels. We define a binary uncertainty indicator: $U(x) = \mathbf{1}\big[p(x) < \tau_p \ \wedge \ H(x) > \tau_H\big]$. We set $\tau_p = 0.5$ (i.e. empty prediction threshold) and $\tau_H = 0.25$ in all experiments.

Feature-Based Shift Data shift can manifests in the representation space, where changes in input distributions lead to deviations in learned feature embeddings. Measuring feature-level deviation therefore provides a meaningful proxy for detecting distribution mismatch. Mahalanobis distance (MAHALANOBIS, 1936) has been shown effective for out-of-distribution detection (Lee et al., 2018), and we adopt it here to measure feature-level deviation from the source embedding statistics. Let $E(x)$ denote the embedding of x and μ, Σ be the mean and covariance of source-domain sample embeddings. The Mahalanobis distance is $d(x) = \sqrt{(E(x)-\mu)^\top \Sigma^{-1}(E(x)-\mu)}$. We normalize it into $[0,1]$ via $F(x) = \mathrm{clip}\Big(\frac{d(x)-d_{\min}}{d_{\max}-d_{\min}+\varepsilon},\ 0,\ 1\Big)$, where $d_{\min}, d_{\max}$ are computed from source-domain sample embeddings.

Ontology-Based Terminology Shift Ontology shift reflects how domain-specific terminology

[2] https://github.com/snovaisg/ICD-Mappings

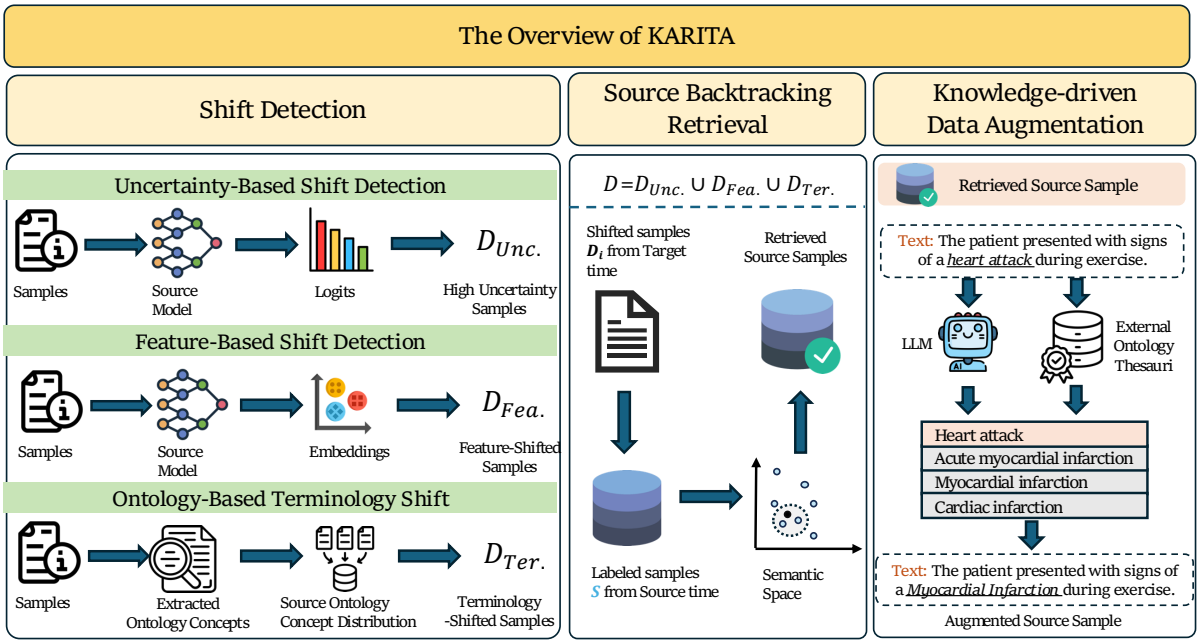

Figure 1: The KARITA method overview. For each incoming target batch, the model performs shift detection, retrieves semantically related source instances, and applies knowledge-driven data augmentation. The augmented data are then used to update the model parameters, and the same process is repeated for the next target batch.

changes over time, where newly emerging concepts or shifts in usage frequency push test samples away from the ontology distribution observed during training, thereby increasing the risk of model degradation when such terms were unseen or underrepresented in early data. To quantify this shift, we treat the source ontology as a probability distribution over concepts and measure how strongly a target document deviates from it. Let $\mathcal{C}(x)$ denote the set of ontology concepts detected in document x, and define $p_{t_1}(c) = f_{t_1}(c)/N_{t_1}$ as the source-period document-frequency probability of concept c, where $f_{t_1}(c)$ is its count and N_{t_1} is the total number of source documents. Inspired by information theoretic surprisal, we compute $I(c) = -\log(p_{t_1}(c) + \varepsilon)$, where a larger value indicates rarer or unseen concepts relative to the training distribution, hence more likely to cause temporal shift. We then aggregate across all concepts in a target document to obtain the ontology-tail shift score

$$O_{\text{tail}}(x) = \frac{1}{|\mathcal{C}(x)|} \sum_{c \in \mathcal{C}(x)} -\log(p_{t_1}(c) + \varepsilon),$$

where a greater $O_{\text{tail}}(x)$ indicates that the document relies on long-tail terminology under the source ontology distribution. This formulation gives a continuous notion of terminology drift rather than relying on the maximum over concepts, and naturally highlights documents whose vocabularies fall in the distribution tail, making them strong candidates for targeted augmentation and test-time adaptation.

Detection Trigger We use the uncertainty trigger to directly collect target samples satisfying $U(x) = 1$, while feature- and ontology-based metrics produce continuous scores and are used to rank samples. For $F(x)$ and $O(x)$, we select the top ρ fraction (e.g., $\rho = 10\%$) of high-scoring target samples: $\mathcal{D}_F = \text{Top}_\rho(F(x)), \mathcal{D}_O = \text{Top}_\rho(O(x))$. Let $\mathcal{D}_U = \{x \mid U(x) = 1\}$ be the set activated by uncertainty. The final detected shift set is $\mathcal{D}_{shift} = \mathcal{D}_U \cup \mathcal{D}_F \cup \mathcal{D}_O$, which is then passed to the retrieval module for augmentation.

We set $\rho = 0.1$ (top 10%) in all experiments (see Appendix B.1 for sensitivity analysis).

4.2 Source Backtracking Retrieval

Temporal shifts do not necessarily imply the complete absence of relevant information in historical data. Even when domain-specific terminology or feature distributions evolve over time, target instances may still share underlying semantic patterns with source data. As a result, shifted target samples often admit semantically similar counterparts from the source period. This motivates us to propose source backtracking retrieval: this module identifies semantically similar source samples for each detected shifted target instance. The

retrieved samples provide semantically relevant source-domain information used for model adaptation under temporal shift.

Specifically, given a target sample x_t, we first obtain its semantic representation (e.g. the [CLS] token) $\mathbf{z}_t = f^{enc}_{\Theta_s}(x_t)$, Θ_s is the source-trained model. We compute source sample embedding $\mathbf{z}_s$ using the same encoder. We then measure the semantic similarity between target and source samples using cosine similarity in the embedding space: $\text{sim}(x_t, x_s) = \cos(\mathbf{z}_t, \mathbf{z}_s)$. For each target instance, we retrieve the top-k source samples with the highest similarity scores. In this paper, k is set to three (see Appendix B.1 for sensitivity analysis). These retrieved samples serve as semantically aligned source references that remain relevant under temporal shift, and are subsequently used for knowledge-driven data augmentation in the next stage.

4.3 Knowledge-driven Data Augmentation

In many classification settings, label assignment depends on knowledge-bearing expressions whose surface forms may evolve over time. Pretrained language models, however, may have limited exposure to such evolving expressions, leading to degraded performance when terminology or phrasing shifts across time. To address this, we apply knowledge-driven data augmentation to the retrieved source samples associated with shifted target instances by aligning the terminology between the source domain and target domain. From a representation perspective, synonym-based augmentation can be viewed as a form of controlled lexical perturbation. This encourages invariance to terminology variation and improves robustness to temporal shift.

We employ two complementary sources of synonym knowledge for data augmentation: 1) LLM-based task-relevant Term identification and synonym generation and 2) external ontology thesauri. For the EurLex and arXiv-CS datasets, we prompt LLM to identify terminology relevant to label assignment and generate corresponding synonyms. For the MIMIC-IV-Notes dataset, due to privacy constraints on non-open-source models, we rely exclusively on external ontology thesauri to provide structured terminological relations.

Task-Relevant Term Identification and Synonym Mapping For EurLex and arXiv-CS documents, we employ GPT-4o-mini[3] to analyze target texts and identify task relevant terms that are informative for the classification task. Based on these terms, the LLM provides corresponding synonymous or historically used variants, and this forms a mapping between current expressions and alternative surface forms. Each target instance is provided together with the set of candidate labels to guide term identification. We provide target instance along with the set of all candidate labels to GPT-4o-mini. The prompt template used for experiments is detailed in Appendix A.2.

External Ontology Thesauri To ensure comprehensive and reliable terminological coverage, we incorporate domain-specific ontology thesauri that provide structured synonym relationships: Medical Subject Headings (MeSH), EuroVoc and Computer Science Ontology (CSO). Details on the specific versions and processing procedures for these resources are provided in Appendix A.3.

MeSH (U.S. National Library of Medicine, 2025) is the controlled and hierarchically-organized vocabulary thesaurus produced by the U.S. National Library of Medicine. It's widely used for indexing, cataloging, and searching for biomedical and health-related information.

EuroVoc (Walhain et al., 2025) is a multidisciplinary thesaurus maintained by the Publications Office of the European Union. It contains keywords organized into 21 domains and 127 sub-domains, used to describe and index the content of legal and policy documents for the European Union.

CSO (Salatino et al., 2020) is a large-scale ontology of Computer Science research areas, comprising approximately 14,000 topics and 162,000 semantic relationships, automatically generated from scholarly publications in Computer Science.

5 Experiment

To examine the effectiveness of our proposed KARITA framework under temporal shift, we conduct experiments on three specialized multi-label classification datasets across different time periods. We compare KARITA with a source model trained on historical data, several state-of-the-art baselines, and an target model fine-tuned directly on target-time data as an upper bound.

All methods are initialized from the same source model trained on the earliest time interval and are evaluated on target time-domain test sets. Performance is evaluated using both instance-level and label-level multi-label classification metrics.

[3]https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

Dataset	MIMIC-IV-Notes					EurLex					arXiv-CS				
Method	P	R	sa-F1	mi-F1	ma-F1	P	R	sa-F1	mi-F1	ma-F1	P	R	sa-F1	mi-F1	ma-F1
I. Source Pretrained Performance on Source															
Source Model	66.83	51.92	55.40	59.31	47.52	86.09	80.39	80.92	81.11	66.97	66.69	58.46	60.56	62.36	43.65
I. Target Pretrained Performance on Target															
Target Model	73.63	57.91	62.12	76.66	65.78	89.27	85.28	85.73	86.10	71.74	81.60	76.05	74.98	72.18	65.51
II. Source Pretrained Performance on Target															
Source Model	58.78	44.07	47.41	52.86	40.65	79.65	56.86	63.73	64.48	46.75	52.12	40.21	43.36	45.85	34.86
III. Source Model Performance on Target after adaptation															
Ours	**64.14**	**62.13**	**60.15**	**63.95**	**52.12**	**82.79**	**69.49**	**73.09**	**73.71**	**56.15**	**68.14**	**64.83**	**62.63**	**61.55**	**49.82**
IFT	58.60	41.56	49.64	55.24	43.05	81.17	52.08	60.54	61.92	37.12	58.43	46.23	49.17	49.75	40.67
SAR	59.41	40.97	45.68	51.96	36.65	73.71	60.47	63.84	64.21	48.30	53.80	40.20	43.40	50.60	38.10
Self-Labeling	58.52	43.52	46.99	52.34	40.55	80.84	54.22	61.62	62.42	42.02	52.25	40.33	43.46	45.95	34.94
EATA	55.85	34.75	40.37	45.98	28.02	77.40	58.92	64.32	64.83	47.97	41.40	33.29	34.90	37.84	27.63

Table 2: Overall classification performance (%) on MIMIC-IV-Notes, EurLex, and arXiv-CS before and after adaptation. All results are averaged over 3 runs using different random seeds. Results compare base models fine-tuned on source data (source models) evaluated on both source and target test sets, base models fine-tuned on target data (target models) as upper bounds, and source models adapted to the target domain using KARITA and baseline adaptation methods. We **bold** the highest and underline the lowest performance of each column.

Specifically, we report sample-averaged precision, recall, and F1 score to assess prediction quality at the sample level, along with micro- and macro-averaged F1 scores to measure overall performance across labels under varying degrees of label imbalance. All results are reported on the target time test sets. This experimental design allows us to assess (i) the extent of performance degradation caused by temporal shift and (ii) how effectively different adaptation strategies mitigate this degradation.

5.1 Baselines

We compare our method with representative baselines that cover three major paradigms for temporal adaptation: training-based adaptation, data-centric pseudo-labeling, and test-time adaptation (TTA).

IFT (T.y.s.s et al., 2024) is a training-based temporal adaptation method that updates the model sequentially across time. This strategy improves upon traditional fine-tuning by exposing the model to chronologically ordered training data incrementally, while keeping the model architecture and loss function unchanged.

Self-Labeling (Agarwal and Nenkova, 2022) represents a data-centric adaptation approach. It first applies the source-trained model to generate pseudo-labels for target-domain samples. These silver-labeled target instances are then combined with the original gold-labeled source data to further fine-tune the model.

SAR (Niu et al., 2023) and **EATA** (Niu et al., 2022) are test-time adaptation (TTA) methods that update the model online during inference without access to target labels. Both methods are built upon the TENT (Wang et al., 2021) framework, which adapts models by minimizing prediction entropy at test time. SAR improves robustness by incorporating sharpness-aware optimization to stabilize model updates under distribution shift, while EATA selectively updates the model using only reliable low-entropy samples and constrains parameter drift to alleviate catastrophic forgetting.

6 Results

As shown in Table 2, our proposed method consistently achieves the state-of-the-art (SOTA) performance across all datasets and metrics in the adaptation scenarios. Specifically, compared to the source model evaluated directly on the target domain (Section II), our method yields substantial improvements, e.g., an absolute increase of **11.47%** in ma-F1 on MIMIC-IV-Notes and **14.96%** on arXiv-CS. This demonstrates the high effectiveness of our adaptation strategy in bridging the domain gap.

The consistent gains across diverse domains suggest that our strategy of shift-aware retrieval followed by synonym-based augmentation effectively addresses the limitations of purely unsupervised or pseudo-label-based adaptation. By retrieving semantically similar samples from the source domain, the model leverages reliable ground-truth labels to bridge the distributional gap, thereby avoiding the error accumulation common in self-labeling or the instability inherent in entropy-minimization-based test-time adaptation (e.g., the performance degradation of EATA on MIMIC-IV-Notes). This targeted alignment allows the model to capture target-specific features while maintaining the high-quality supervision signals preserved from the source domain, resulting in particularly robust performance

Dataset	MIMIC-IV-Notes					EurLex					arXiv-CS				
Method	P	R	sa-F1	mi-F1	ma-F1	P	R	sa-F1	mi-F1	ma-F1	P	R	sa-F1	mi-F1	ma-F1
Full Method	64.14	**62.13**	**60.15**	**63.95**	**52.12**	**82.79**	**69.49**	**73.09**	**73.71**	**56.15**	68.14	**64.83**	**62.63**	**61.55**	**49.82**
w/o detection	64.67	57.98	58.18	61.88	49.33	79.60	69.03	71.39	71.69	48.77	65.36	54.80	56.56	56.62	31.02
w/o augmentation	**66.75**	57.53	58.85	61.80	48.13	80.33	67.55	70.78	71.43	54.60	**68.79**	62.59	62.01	61.01	43.74
w/o retrieval	62.77	59.49	58.08	62.00	50.67	75.57	65.83	67.30	68.88	44.16	57.68	54.09	52.40	52.89	36.40

Table 3: Ablation study on the contribution of each module. We **bold** the highest and underline the lowest performance in each column.

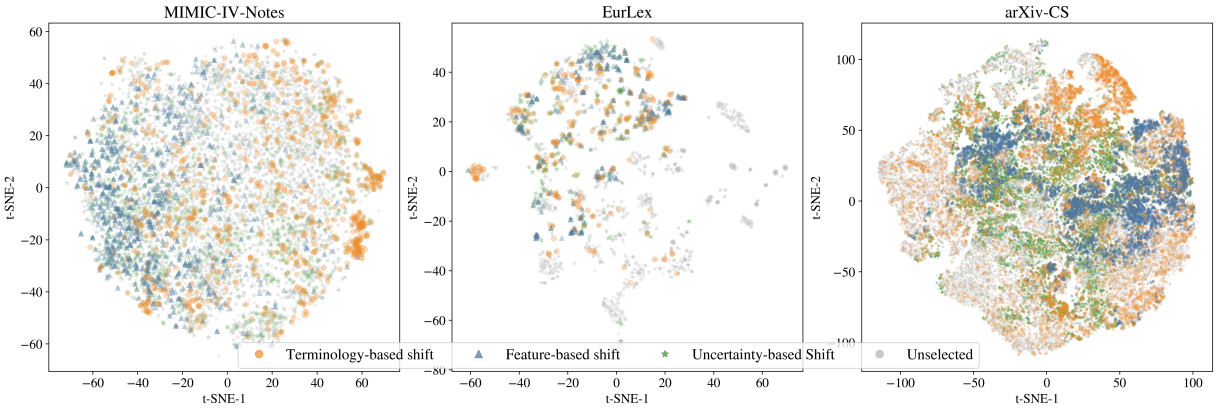

Figure 2: t-SNE visualization of target-domain representations on MIMIC-IV-Notes, EurLex, and arXiv-CS. Samples detected by ontology-based shift (orange circles), feature-based shift (blue triangles), and uncertainty-based shift (green stars) are highlighted, with all other target samples shown in grey.

in recall-sensitive tasks.

6.1 Ablation Study

We conduct an ablation study to verify the necessity of each component in our framework. As shown in Table 3, removing any single module leads to a consistent performance drop across all three datasets. Specifically, replacing our shift-aware retrieval with the purposeful selection of low-similarity samples (*w/o retrieval*) results in the **lowest sa-F1 scores** across all datasets (e.g., 58.08% on MIMIC-IV-Notes and 52.40% on arXiv-CS). This suggests that without semantically relevant "bridges" from the source domain, the model struggles to maintain prediction quality at the individual sample level. It highlights that semantic relevance is not just beneficial for overall distribution alignment, but fundamental for effective per-sample knowledge transfer.

Furthermore, replacing our detection module with random sampling (*w/o detection*) significantly impairs the model's ability to adapt, with the macro-F1 on arXiv-CS plummeting from 49.82% to 31.02%. This suggests that identifying specific shift dimensions is critical for selecting the "correct" knowledge to transfer. Removing augmentation (*w/o augmentation*) also yields inferior results compared to the full method. This is most evident on arXiv-CS, where ma-F1 drops from 49.82% to 43.74%, suggesting that terminology alignment contributes beyond what retrieval alone provides. Overall, the synergy between these three modules is what drives robust adaptation capabilities.

7 Analysis

In this section, we examine how the three shift signals in KARITA relate to each other and contribute to adaptation. We first investigate their spatial distribution in the feature space, asking whether they capture redundant or complementary information. We then quantify their pairwise overlap on target samples and track how each type of shift evolves over time. Finally, we compare adaptation performance when each shift detector is used in isolation versus in combination.

7.1 How do different types of data shift relate in the feature space?

Figure 2 shows t-SNE projections of target-domain representations, with samples detected by each shift type marked in different colors. We observe that ontology-based shift consistently highlights regions in the target-domain embedding space that are weakly captured by feature-based and uncertainty-based shift signals. Across all three datasets, ontology-shifted samples form distinct

spatial patterns, including clusters that are separated from feature-shifted regions, suggesting that terminology-level changes are not always reflected as representation drift.

This effect is most evident in MIMIC, where feature-based and ontology-based shifts concentrate along different directions of the embedding space. While feature shift follows a broad global trend, ontology shift forms several localized clusters with limited overlap. This suggests that relying solely on feature-based shift may miss samples whose difficulty arises from evolving medical terminology, as such changes do not necessarily produce detectable deviations in the feature space.

A similar separation is observed in EurLex, where ontology-shifted samples form compact clusters that are only weakly aligned with feature-based shift, despite the overall sparsity of the embedding. This suggests that legal terminology evolution can introduce temporal challenges that are not fully captured by representation-level deviation.

In arXiv-CS, feature-based and uncertainty-based shifts mainly concentrate in dense regions of the embedding space, whereas ontology-shifted samples span a much broader area and blend more closely with samples considered unshifted. This pattern indicates that terminology-driven label ambiguity can persist even when documents remain close in feature space, further motivating the need for ontology-aware signals. Overall, these observations show that ontology-based shift is not redundant with feature- or uncertainty-based criteria, but reveals terminology-driven temporal effects that would otherwise remain difficult to detect.

7.2 How do the three shift signals complement each other?

To further quantify the relationship among the three shift detectors, we examine both their overlap on detected target samples and the temporal trends of their respective metrics on the target time period.

Dataset	$\lvert\mathcal{D}_t\rvert$	U	U∩O	U∩F	O∩F	U∩O∩F
MIMIC	6,827	27.74%	3.05%	6.09%	0.37%	0.13%
EurLex	2,542	27.50%	4.48%	8.03%	1.49%	1.26%
arXiv-CS	113,495	25.77%	2.71%	4.38%	1.45%	0.62%

Table 4: Overlap among shift detectors, reported as the percentage of target-domain samples ($|\mathcal{D}_t|$) detected by each signal or their intersections. U denotes uncertainty-based shift; O and F denote ontology-based and feature-based shift, both detecting the top ρ=10% of $\mathcal{D}_t$.

Overlap across shift types Table 4 reports the sample counts and pairwise intersections among uncertainty-based (U), ontology-based (O), and feature-based (F) shift detections. Across all three datasets, the pairwise intersections are small: on MIMIC, only 3.05% of target samples fall under both U and O, and the three-way intersection (U∩O∩F) accounts for merely 0.13%. The overlap between ontology shift and the other two types is particularly limited (e.g., O∩F = 0.37% on MIMIC), confirming that terminology-level drift captures a distinct aspect of temporal shift that is missed by feature- or uncertainty-based signals.

	F score		O score		Entropy	
Year	mean	med.	mean	med.	mean	med.
EurLex (2011–2015)						
2011	.551	.585	1.536	1.493	.194	.177
2012	.579	.629	1.561	1.489	.209	.209
2013	.576	.599	1.596	1.560	.206	.199
2014	.601	.643	1.585	1.548	.213	.213
2015	.580	.637	1.526	1.491	.207	.206
arXiv-CS (2021–2025)						
2021	.655	.686	1.357	1.350	.134	.137
2022	.659	.690	1.358	1.355	.135	.138
2023	.667	.696	1.367	1.362	.136	.139
2024	.670	.696	1.391	1.388	.136	.138
2025	.673	.698	1.427	1.424	.138	.140

Table 5: Year-wise statistics of feature shift (F score), ontology shift (O score), and predictive entropy in the target period for EurLex and arXiv-CS.

Temporal trends of shift scores To examine whether the three signals capture temporal evolution consistently, we compute year-wise statistics of feature shift scores (F score), ontology shift scores (O score), and predictive entropy for EurLex and arXiv-CS target time data (Table 5). We exclude MIMIC because its de-identification procedure maps timestamps to three-year intervals and year-level breakdowns is unavailable.

Across both datasets, all three scores trend upward over time, indicating that temporal distance from the source period amplifies all three types of shift. Notably, the scores increase at different rates: on EurLex, ontology scores rise more steeply than feature scores, while on arXiv-CS the two grow at comparable rates. This suggests that while all three detectors respond to increasing temporal distance, they do so with different sensitivities depending on the domain, further motivating their joint use.

Dataset	MIMIC-IV-Notes					EurLex					arXiv-CS				
Method	P	R	sa-F1	mi-F1	ma-F1	P	R	sa-F1	mi-F1	ma-F1	P	R	sa-F1	mi-F1	ma-F1
Full Method	64.14	**62.13**	**60.15**	**63.95**	**52.12**	**82.79**	**69.49**	**73.09**	**73.71**	**56.15**	**68.14**	**64.83**	**62.63**	**61.55**	**49.82**
Feature-based Shift Only	63.73	59.23	58.61	63.84	51.58	77.33	57.54	62.96	63.4	44.57	54.28	51.43	49.59	52.01	23.00
Ontology-based Shift Only	65.42	56.24	57.26	59.99	40.94	72.45	69.32	68.32	68.98	50.48	59.92	54.43	53.78	55.51	29.69
Uncertainty-Based Shift Only	**67.96**	50.72	55.15	58.36	42.45	82.11	68.65	71.78	72.27	54.97	67.68	56.24	58.25	57.03	42.64

Table 6: Performance of using individual shift detectors versus combining all three. We **bold** the highest and underline the lowest performance of each column.

7.3 How do different types of data shift affect adaptation performance?

Table 6 compares adaptation performance when only a single type of shift is detected. Feature-based shift detection performs the worst on both EurLex and arXiv-CS, with a large drop in macro-F1. For example, on arXiv-CS, macro-F1 declines to 23.00%, compared to 49.82% achieved by the full method. This shows that relying solely on feature or embedding distance fails to identify shifts that are helpful for adaptation and can even harm performance for certain label categories.

Uncertainty-based shift detection performs the worst on MIMIC-IV-Notes, where macro-F1 drops to 42.45%, well below the full method (52.12%). This shows that outcome-driven shift detection based on predictive uncertainty (e.g., high entropy or low maximum probability) is useful but not sufficient on its own, and combining it with text-level shift detection leads to more effective adaptation.

By contrast, ontology-based shift detection leads to more stable performance across datasets. As shown in Fig. 2, ontology-aware signals capture semantically meaningful shifts that are not clearly separated in the feature space and provide more useful samples for retrieval and augmentation. Overall, these results show that combining complementary shift signals is necessary for robust adaptation.

8 Conclusion

We introduced **KARITA**, a knowledge-driven augmentation framework for integrative temporal adaptation. KARITA detects temporal shift signals from multiple complementary perspectives (uncertainty, feature deviation, and ontology-based terminology drift), backtracks to retrieve semantically aligned labeled instances from historical source data, and performs knowledge-driven synonym augmentation using external thesauri and LLM-based terminology mapping. Across clinical, legal, and scientific corpora, KARITA consistently improves target-time performance over strong temporal adaptation baselines, highlighting that temporally robust learning benefits from both shift-aware sample selection and domain knowledge integration.

Our analyses show that different shift types occupy distinct regions in the representation space and contribute differently to the adaptation process, indicating that a single drift signal is often insufficient for reliable temporal learning. By unifying multi-aspect shift detection and terminology-level augmentation, the proposed KARITA framework provides a flexible, data-centric approach that is model-agnostic and effective across domains, offering a promising direction for maintaining deployed language models under evolving language.

Limitations

While KARITA demonstrates effective temporal adaptation across classification tasks in three major domains, two important limitations should be acknowledged: First, the proposed framework relies on the availability of external terminological resources, such as large language models or curated ontology thesauri, to construct mappings between alternative expressions. In highly specialized, low-resource, or privacy-sensitive domains where well-developed ontologies or non-open-source LLMs are unavailable, this requirement may reduce direct applicability. In such cases, alternative solutions, including domain-adapted language models, expert-curated terminological mappings, or corpus-induced synonym discovery from the data itself, may be necessary to maintain the effectiveness.

Second, KARITA primarily addresses temporal shifts manifested at the lexical and terminological level. While this captures a common and impactful form of temporal variation, the framework does not explicitly model deeper forms of knowledge evolution, such as the emergence of entirely new concepts, changes in label definitions, or shifts in task formulation over time. Extending the approach to handle such structural knowledge changes remains an important direction for future work.

Acknowledgment

The authors thank anonymous reviewers for their insightful feedback. The project was partially supported by the National Science Foundation (NSF) under awards TI-2434589 (OpenAI API expenses) and IIS-2440381. We thank the computing resources provided by the iTiger GPU cluster (Sharif et al., 2025) supported by the NSF MRI program under the award CNS-2318210.

References

Oshin Agarwal and Ani Nenkova. 2022. Temporal effects on pre-trained models for language processing tasks. *Transactions of the Association for Computational Linguistics*, 10:904–921.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. Multieurlex – a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ilias Chalkidis and Anders Søgaard. 2022. Improved multi-label classification under temporal concept drift: Rethinking group-robust algorithms in a label-wise setting. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2441–2454, Dublin, Ireland. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. 2000. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220.

Yue Guo, Chenxi Hu, and Yi Yang. 2023. Predict the future from the past? on the temporal data distribution shift in financial sentiment classifications. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1029–1038, Singapore. Association for Computational Linguistics.

Guangzeng Han, Weisi Liu, and Xiaolei Huang. 2025. Attributes as textual genes: Leveraging LLMs as genetic algorithm simulators for conditional synthetic data generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19367–19389, Suzhou, China. Association for Computational Linguistics.

Xiaolei Huang, Lixing Liu, Kate Carey, Joshua Woolley, Stefan Scherer, and Brian Borsari. 2018. Modeling temporality of human intentions by domain adaptation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 696–701, Brussels, Belgium. Association for Computational Linguistics.

Xiaolei Huang and Michael J. Paul. 2019. Neural temporality adaptation for document classification: Diachronic word embeddings and domain adaptation models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4113–4123, Florence, Italy. Association for Computational Linguistics.

Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun KIM, Stanley Jungkyu Choi, and Minjoon Seo. 2022. Towards continual knowledge learning of language models. In *International Conference on Learning Representations*.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023. Mimic-iv-note: Deidentified free-text clinical notes.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Weisi Liu, Guangzeng Han, and Xiaolei Huang. 2025. Examining and adapting time for multilingual classification via mixture of temporal experts. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6151–6166, Albuquerque, New Mexico. Association for Computational Linguistics.

Weisi Liu, Zhe He, and Xiaolei Huang. 2024. Time matters: Examine temporal effects on biomedical language models. In *AMIA Annual Symposium Proceedings*, volume 2024, pages 723–732, San Francisco, CA, USA. American Medical Informatics Association.

PC MAHALANOBIS. 1936. On the generalised distance in statistics. In *Proceedings of the National Institute of Science of India*, volume 12, pages 49–55.

Katerina Margatina, Shuai Wang, Yogarshi Vyas, Neha Anna John, Yassine Benajiba, and Miguel Ballesteros. 2023. Dynamic benchmarking of masked language models on temporal concept drift with multiple views. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2881–2898, Dubrovnik, Croatia. Association for Computational Linguistics.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. 2022. Efficient test-time model adaptation without forgetting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16888–16905. PMLR.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. 2023. Towards stable test-time adaptation in dynamic wild world. In *The Eleventh International Conference on Learning Representations*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Hanshu Rao, Weisi Liu, Haohan Wang, I-Chan Huang, Zhe He, and Xiaolei Huang. 2026. A Scoping Review of Synthetic Data Generation by Language Models in Biomedical Research and Application: Data Utility and Quality Perspectives. *Journal of Healthcare Informatics Research*.

Paul Röttger and Janet Pierrehumbert. 2021. Temporal adaptation of BERT and performance on downstream document classification: Insights from social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2400–2412, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Angelo A. Salatino, Francesco Osborne, Thiviyan Thanapalasingam, and Enrico Motta. 2019. The cso classifier: Ontology-driven detection of research topics in scholarly articles. In *Digital Libraries for Open Knowledge: 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, Oslo, Norway, September 9-12, 2019, Proceedings*, page 296–311, Berlin, Heidelberg. Springer-Verlag.

Angelo A. Salatino, Thiviyan Thanapalasingam, Andrea Mannocci, Aliaksandr Birukou, Francesco Osborne, and Enrico Motta. 2020. The computer science ontology: A comprehensive automatically-generated taxonomy of research areas. *Data Intelligence*, 2(3):379–416.

Chao Shang, Guangtao Wang, Peng Qi, and Jing Huang. 2022. Improving time sensitivity for question answering over temporal knowledge graphs. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8017–8026, Dublin, Ireland. Association for Computational Linguistics.

Mayira Sharif, Guangzeng Han, Weisi Liu, and Xiaolei Huang. 2025. Cultivating multidisciplinary research and education on gpu infrastructure for mid-south institutions at the university of memphis: Practice and challenge. *Preprint*, arXiv:2504.14786.

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Haotian Sun, Hang Wu, Carl Yang, and May Dongmei Wang. 2024. MedAdapter: Efficient test-time adaptation of large language models towards medical reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22294–22314, Miami, Florida, USA. Association for Computational Linguistics.

Zhaochen Su, Zecheng Tang, Xinyan Guan, Lijun Wu, Min Zhang, and Juntao Li. 2022. Improving temporal generalization of pre-trained language models with lexical semantic change. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6380–6393, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiaohang Tang, Yi Zhou, Taichi Aida, Procheta Sen, and Danushka Bollegala. 2023. Can word sense distribution detect semantic changes of words? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3575–3590, Singapore. Association for Computational Linguistics.

Santosh T.y.s.s, Tuan-Quang Vuong, and Matthias Grabmair. 2024. ChronosLex: Time-aware incremental training for temporal generalization of legal classification tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3022–3039, Bangkok, Thailand. Association for Computational Linguistics.

U.S. National Library of Medicine. 2025. Medical subject headings (MeSH). `https://www.nlm.nih.gov/mesh/`. Accessed: 2025-01-03.

Lucy Walhain, Sébastien Albouze, Anikó Gerencsér, Mihai Paunescu, Vassilis Tzouvaras, and Cosimo

Palma. 2025. The EuroVoc thesaurus: Management, applications, and future directions. In *Proceedings of the 5th Conference on Language, Data and Knowledge*, pages 340–350, Naples, Italy. Unior Press.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. 2021. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Implementation Details

A.1 Temporal Data Partitioning

The released MIMIC-IV-Notes data employs a de-identification mechanism that assigns the time information of each clinical note to three-year-long intervals spanning from 2008 to 2022, which naturally forms five temporal domains. We selected the first four temporal intervals for our experiments due to data scarcity in the last time interval, denoted as T_1, T_2, T_3, T_4 in chronological order. To maintain consistency across datasets and enable systematic comparison, we adopt the same four-interval partitioning strategy. Table 7 presents the detailed temporal splits for each dataset.

The last interval is set as the target data, and earlier as the source data. Within each temporal interval, we randomly split the data into training (70%) and test (30%) sets. We train models on the training set of T_1 and evaluate on the test sets of both T_1 (in-domain) and T_4 (cross-domain) to test the temporal generalization of the model.

Dataset	Time Intervals
MIMIC	2008–2010, 2011–2016, 2017–2019
EurLex	1958–1985, 1986–2010, 2011–2016
arXiv-CS	1991–2010, 2011–2020, 2021–2025

Table 7: Temporal splits of datasets.

A.2 Prompt Template

```
You are helping to build a terminology
 lexicon for a multi-label
text classification task on the
dataset: {dataset_name}.
You will receive:
- One document text.
- The full list of possible labels for
 this dataset.
Your tasks:
1. Read the document text carefully.
2. Choose 3-10 words or short phrases
("terms") that are especially
informative for deciding which labels
to assign from the given label set.
These terms should:
- Be specific to the subject or
content of the document, not generic
words
like "data", "paper", "regulation", "
method", "result", etc.
- Strongly indicate one or more labels
 when they appear.

2. For each selected term, generate:
- "synonyms": close paraphrases,
especially earlier phrasings that
could appear in older literature or
legal/technical writing, preferred
terms if applicable,
You MUST return a SINGLE JSON object
with EXACTLY this structure:
{
"entities": [
{
"term": "...",
"synonyms": ["...", "..."]
}
]
}
No extra keys. No comments. No
explanations outside JSON.
--------------------
DATASET NAME: {dataset_name}
FULL LABEL SET:
{label_block}
--------------------
DOCUMENT TEXT:
{text}
```

A.3 External Knowledge Resources

We leverage domain-specific ontology thesauri to provide structured terminological relationships for data augmentation.

MeSH (Medical Subject Headings) (U.S. National Library of Medicine, 2025) We used the MeSH 2025 release [4], including both the descriptor file (`desc2025.xml`) and the supplementary concept file (`supp2025.xml`). These XML files provide hierarchically-organized medical terminology and synonym relationships used for augmentation of the MIMIC-IV-Notes data.

[4] https://www.nlm.nih.gov/mesh/

Algorithm 1 Knowledge-driven Augmentation and Retrieval for Integrative Temporal Adaptation

Require: Source-trained model Θ_s, source dataset $\mathcal{D}_s$, target dataset $\mathcal{D}_t$

Initialize model $\Theta \leftarrow \Theta_s$
for each target batch $\mathcal{B}_t \subset \mathcal{D}_t$ **do**
  Detect shifted samples $\mathcal{B}_{shift} \subset \mathcal{B}_t$ using uncertainty, feature, and ontology-based scores
  for each $x_t \in \mathcal{B}_{shift}$ **do**
    Compute embedding $\mathbf{z}_t \leftarrow f_\Theta^{enc}(x_t)$
    Retrieve top-k source samples $\mathcal{R}(x_t)$ from $\mathcal{D}_s$ using cosine similarity
  end for
  Augment retrieved samples using LLM-based and/or ontology-based synonym augmentation
  Update model Θ using the augmented source samples
end for
Prediction: Generate predictions on $\mathcal{D}_t$ using the final adapted model Θ

EuroVoc (Walhain et al., 2025) We use version 4.22 of the EuroVoc thesaurus (released 2025-07-02), specifically the English version (`eurovoc_export_en-4.22.xlsx`). EuroVoc provides a multi-level hierarchical taxonomy of concepts spanning 21 domains and 127 sub-domains. The thesaurus defines relationships between preferred terms (PT), which represent canonical concept labels, and non-preferred terms (NPT), which serve as synonyms or alternative expressions. We exploit these PT-NPT relationships to identify terminological variants for data augmentation in EUR-Lex legal documents.

CSO (Computer Science Ontology) (Salatino et al., 2020) We use CSO version 3.5[5], the latest release of the ontology. We extract terminological relationships from CSO using the CSO Classifier toolkit (v4.0.0) (Salatino et al., 2019)[6], which provides pre-processed ontology files including the topic hierarchy and term-to-topic mappings. These resources supplement the GPT-4o-mini-identified terminology and generated synonyms for the arXiv-CS dataset.

[5] https://cso.kmi.open.ac.uk/

[6] https://github.com/angelosalatino/cso-classifier

A.4 Experiment details and Baselines

Source Model We initialize all methods from a common base model trained on the earliest time interval T_1. For EurLex and arXiv-CS, we use *XLM-RoBERTa-base* (Conneau et al., 2020) as the base encoder, while for MIMIC-IV we adopt Longformer (Beltagy et al., 2020) to better handle long clinical documents. The source model is fine-tuned on gold-labeled data from T_1 for 10 epochs, using a learning rate of 3×10^{-5}. We use a batch size of 32 for EurLex and arXiv-CS, and 8 for MIMIC, with gradient accumulation steps set to 2. Unless otherwise specified, we choose the same batch size and learning rate for all baseline methods and KARITA.

KARITA KARITA adapts the source-trained model sequentially over the target data stream. For each target batch, shifted samples are first identified using the proposed shift detection module. Only the detected samples trigger source backtracking retrieval, where semantically similar source instances are selected. The retrieved samples are then augmented using knowledge-driven terminology expansion and used to update the model before processing the next target batch. For efficiency, LLM-based terminology identification and synonym generation are performed once for the detected target samples and reused throughout adaptation, rather than being invoked within each update step. After adaptation is completed, predictions are generated once on the full target split using the final adapted model.

Self-Labeling We follow the protocol in (Agarwal and Nenkova, 2022). The source model fine-tuned on gold-labeled source data is first used to generate pseudo-labels for target-domain samples. The pseudo-labeled target data is then merged with the original source training set to fine-tune a new model.

IFT For IFT (T.y.s.s et al., 2024), we train the model sequentially on chronologically ordered temporal splits. We split the first interval of the data into three incremental splits and train on every split for 3 epochs. The model parameters fine-tuned from the previous splits are used to initialize training for the next splits.

SAR (Niu et al., 2023) We set the entropy minimization objective to operate on per-label sigmoid outputs to apply SAR to multi-label classification setting. Specifically, prediction entropy is com-

puted independently for each label dimension and aggregated across labels. We adopt the sharpness-aware optimizer proposed in SAR to stabilize online updates and perform adaptation using unlabeled target samples at test time. For SAR, we use a learning rate of 1×10^{-4} with a sharpness-aware optimizer, momentum set to 0.9, and a perturbation radius of $\rho = 0.05$. The entropy threshold is defined as $E_0 = 0.4 \cdot \ln(\texttt{max\_token\_len})$.

EATA (Niu et al., 2022) Similar to SAR, we apply EATA to the multi-label setting by computing entropy over sigmoid-based label probabilities. Only target samples with low aggregated entropy are selected for model updating. In addition, we apply the regularization mechanism proposed in EATA to limit parameter drift and mitigate catastrophic forgetting during test-time adaptation. For EATA, we adopt a learning rate of 5×10^{-5} and apply Fisher regularization with $\beta = 1/2000$ to constrain parameter updates. The entropy threshold is set to $E_0 = 0.4 \cdot \ln(\texttt{max\_token\_len})$ to select reliable target samples for adaptation.

A.5 Hardware and Software

The experiments are conducted on a server equipped with 8x H100 GPUs, 2x EPYC Genoa 9334 CPUs, and 768GB of RAM. The system runs on Linux kernel 5.14. The system utilizes PyTorch 2.3.0 (CUDA 12.1) (Paszke et al., 2019) alongside HuggingFace Transformers 4.57.1 (Wolf et al., 2020).

B Additional Results

B.1 Sensitivity Analysis

We examine the sensitivity of KARITA to two key hyper-parameters: the number of retrieved source samples k and the shift detection proportion ρ. We conduct experiments on MIMIC-IV-Notes as a representative case, varying one parameter while fixing the other.

Effect of k We vary k from 1 to 5 with ρ fixed at 0.1. As shown in Table 8, performance improves from k=1 to k=3 and remains relatively stable beyond that, with k=3 achieving the best overall balance across metrics. Retrieving too few samples (k=1) limits the diversity of source supervision, while larger values (k=4, 5) do not yield further gains and increase computational cost.

Effect of ρ We vary ρ across $\{0.05, 0.1, 0.2, 0.3\}$ with k fixed at 3. As shown in Table 9, performance improves as ρ increases from 0.05 to 0.1, but does not improve further at 0.2 and 0.3. A smaller ρ limits the coverage of shifted samples, while larger values introduce non-shifted samples into retrieval without additional benefit. We therefore choose ρ=0.1 as a practical balance between effectiveness and efficiency.

Metric	k=1	k=2	k=3	k=4	k=5
P	64.04	67.34	67.00	66.07	69.34
R	55.74	61.18	63.35	63.55	60.87
sa-F1	56.05	59.86	61.80	61.03	61.07
mi-F1	59.60	64.11	65.13	64.78	64.83
ma-F1	46.57	51.78	53.12	52.60	52.15

Table 8: Sensitivity analysis on k ($\rho = 0.1$, MIMIC-IV-Notes).

Metric	ρ=0.05	ρ=0.1	ρ=0.2	ρ=0.3
P	68.14	67.00	65.43	67.75
R	57.51	63.35	63.24	61.56
sa-F1	58.76	61.80	61.10	61.41
mi-F1	62.37	65.13	64.32	64.51
ma-F1	49.05	53.12	53.68	53.33

Table 9: Sensitivity analysis on ρ ($k = 3$, MIMIC-IV-Notes).