BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Clinical-ShiftEval: a framework for simulating and evaluating model adaptation in dynamic clinical NLP tasks

Fabián Villena[1,2], Felipe Bravo-Marquez[2*] and Jocelyn Dunstan[3,4*]

Abstract

Background Clinical natural language processing (NLP) models are widely used to extract information from electronic health records (EHR) and support healthcare decision-making. However, most existing models are evaluated under the assumption of static data distributions and fixed task definitions, overlooking the dynamic and evolving nature of real clinical environments. Distributional changes arising from updated guidelines, emerging diseases, or institutional restructuring can lead to substantial model degradation in practice.

Methods We introduce Clinical-ShiftEval, a framework designed to simulate and evaluate model adaptation in dynamic clinical NLP tasks using real-world clinical text data. Our framework parameterizes two prevalent types of evolution observed in healthcare: (1) label-set incompatibility (LSI), where the set of prediction targets changes over time, and (2) task definition evolution (TDE), where the clinical meaning or labeling rule is revised. These changes are operationalized through controlled transitions between two periods in real datasets. As a case study, we applied Clinical-ShiftEval to the Chilean Waiting List Corpus, using referral specialty classification to model LSI and disease prioritization to model TDE, and systematically compared continual training, in-context learning with large language models, and a hybrid approach.

Results We validated that Clinical-ShiftEval reliably simulates realistic levels of clinical change and produces controlled, interpretable performance drops. This confirms its utility as a framework to generate reproducible scenarios for benchmarking model robustness. As a case study, we compared multiple model adaptation strategies under LSI and TDE. Conventional supervised models showed severe degradation under these shifts (up to a drop of 82% F1 in LSI and 43% in TDE). In contrast, in-context learning reduced the drop to 35% in LSI and 10% in TDE, while a hybrid method further improved robustness, limiting the drop to approximately 8% in both settings. Continual adaptation with incremental retraining gradually restored performance, but only surpassed in-context and hybrid approaches after incorporating at least 30% of data from the second period.

Conclusions Clinical-ShiftEval enables robust benchmarking of model adaptation in realistic, dynamic settings using authentic EHR text. Both in-context learning and hybrid approaches substantially mitigate the impact of evolving

*Correspondence:
Felipe Bravo-Marquez
fbravo@dcc.uchile.cl
Jocelyn Dunstan
jdunstan@uc.cl

Full list of author information is available at the end of the article

clinical data, achieving strong performance even without access to new labeled data. For optimal recovery of baseline accuracy, continual training with a moderate amount of new data is most effective, eventually surpassing the other methods as more new data become available. These findings provide practical guidance for the deployment of resilient NLP models in dynamic healthcare settings.

Keywords Artificial intelligence, Natural language processing, Concept drift, Concept evolution

Introduction

Clinical natural language processing (NLP) has become a fundamental component to transform large volumes of unstructured text in electronic health records (EHR) into actionable clinical insights. Applications such as patient prioritization, clinical entity recognition, and automated disease coding have demonstrated significant value in supporting clinical workflows and decision making [1–3].

Despite these advances, the standard development and evaluation paradigm for clinical NLP models typically assumes a static data distribution and a fixed task definition [4–7]. Most models are trained on retrospective datasets and evaluated on test samples extracted from the same temporal or institutional context, not considering the range of distributional changes that occur once models are deployed in real clinical environments [8–10]. This prevalent practice can lead to an overestimation of model robustness and generalizability, as it overlooks data drift, and the evolving nature of clinical documentation and workflows, factors recognized as risks to the reliability of machine learning in medicine [11, 12].

From this perspective, clinical NLP practitioners face three major challenges that structure the problem space:

Data scarcity The limited availability of annotated clinical text remains one of the main barriers for clinical NLP. In our previous work [13], we studied this challenge in static settings, showing how different modeling paradigms behave under restricted data availability.

Data dynamism Beyond scarcity, models deployed in healthcare face continuous changes due to evolving clinical guidelines, emerging diseases, and institutional restructuring. These shifts alter distributions and label spaces, often leading to severe model degradation if not properly addressed. This paper focuses on this second challenge by operationalizing realistic scenarios of dataset and task evolution.

Paradigm multiplicity Whereas in static settings paradigm choice can often be guided primarily by the amount of labeled data available (as explored in our prior study), in dynamic environments the search for an effective paradigm is more complex. Models must not only work with limited data, but also remain robust as label spaces and task definitions evolve.

By structuring the problem with these three challenges, we highlight why the dynamic clinical setting is particularly difficult: the combined effect of evolving data and competing paradigms makes paradigm selection harder than in static scenarios.

The consequences of ignoring these distributional changes have been demonstrated: after deployment, models often exhibit a marked deterioration in performance as clinical practice evolves, new diseases emerge, and institutional processes change [4–7]. For example, the onset of the COVID-19 pandemic led to the rapid introduction of new terminology and clinical concepts, quickly making many previously reliable models obsolete or inaccurate [14]. Similarly, regulatory updates, changes in disease prioritization, or institutional restructuring can introduce or eliminate clinical categories, and thus target labels, undermining the validity of legacy models. Organizational changes, such as the creation of new specialty services or updated patient flows, further shift document patterns and label distributions, often necessitating model adaptation.

These realities expose a limitation of current clinical NLP practice: existing models are poorly equipped to handle real-world changes in label sets and task definitions, making them fragile in the face of clinical or policy-driven evolution. Recent studies have highlighted that such model inflexibility can compromise generalizability [5]. Interest is growing in techniques such as domain adaptation and transfer learning, yet there remains a lack of systematic tools to simulate and quantitatively evaluate these adaptation challenges as they appear in dynamic clinical contexts [15, 16].

Building on our previous work [13], where we analyzed different modeling paradigms for clinical NLP under conditions of restricted data availability, we now extend this line of research to dynamic clinical environments. While that earlier study offered recommendations for static data-scarcity scenarios, it did not address the challenges of distributional change and evolving task definitions that appear in practice. Here, we construct on those findings by asking how different paradigms perform when data itself changes over time.

To bridge this gap, we present Clinical-ShiftEval, a systematic framework to simulate and evaluate model adaptation in dynamic clinical NLP tasks. In addition to introducing the framework, we present a case study based on the Chilean Waiting List Corpus, a clinical referral dataset written in Spanish. This case study plays a central role in our contribution, as it demonstrates the framework's applicability to real EHR text and enables

a comparison between different adaptation paradigms, including continual training, in-context learning with large language models (LLMs), and a hybrid approach. A central advantage of Clinical-ShiftEval is that it operates directly on real-world clinical text data, rather than relying on synthetic or artificially generated narratives. Thus, all simulations and evaluations reflect the complexity and variability of actual EHRs. More specifically, Clinical-ShiftEval is designed to simulate the effects of real-world clinical change on model behavior. In the framework, label-set incompatibility represents changes in active prediction categories that may arise from processes such as institutional restructuring or administrative updates, whereas task definition evolution represents changes in labeling criteria or clinical priorities, such as those introduced by revised guidelines or prioritization policies. Furthermore, Clinical-ShiftEval uniquely supports multiple shift levels, enabling precise, fine-grained benchmarks that could mirror both incremental and sudden changes. These properties fill a critical methodological gap and position Clinical-ShiftEval as a necessary tool for robust, actionable clinical NLP research.

Within this framework, we focus on three representative adaptation paradigms that reflect distinct approaches to handling dynamic clinical NLP tasks. Supervised models with continual training represent the conventional strategy, where models are incrementally updated with new data to recover performance across periods. In-context learning with LLMs offers an alternative that relies on task descriptions and examples provided at inference time, making it attractive when little or no new labeled data is available. Finally, the hybrid approach combines a legacy supervised model trained on earlier data with the adaptability of an LLM, leveraging the strengths of both paradigms. By comparing these three methods, our study not only evaluates Clinical-ShiftEval as a framework, but also provides practical insights into the trade-offs between retraining, prompt-based adaptation, and integrative hybrid solutions in realistic clinical settings.

Our methodology formalizes and operationalizes two common forms of data drift: (1) *label-set incompatibility (LSI)*, where the set of prediction targets evolves due to new or deprecated clinical categories; and (2) *task definition evolution (TDE)*, where changes in clinical guidelines or priorities modify the mapping from the input data to the target labels. By parameterizing these scenarios and applying them to real clinical text datasets, we enable rigorous benchmarking of model robustness and adaptation strategies, including static supervised models, in-context learning with LLMs, and hybrid approaches.

Given the current proliferation of modeling paradigms and the rapidly evolving NLP technologies in healthcare, there is a pressing need for benchmarking platforms that enable systematic, realistic, and fair comparison of model adaptability. Clinical-ShiftEval addresses this gap, offering a versatile testbed for evaluating the reliability of diverse approaches under the complex and changing conditions currently faced in clinical NLP practice.

Clinical-ShiftEval enables researchers to systematically investigate the reliability and adaptability of clinical NLP models in dynamic environments. In this work, we focus on the following key questions:

- How robust are clinical NLP models and adaptation strategies to data and task changes typical in healthcare settings?
- How does the severity and type of distributional change affect model performance and adaptation?
- What is the relationship between the amount and nature of new data and the ability of a model to recover its performance after the shift?
- Under what conditions do specific adaptation strategies provide advantages over others in dynamic clinical NLP tasks?

While these represent the central research questions we address through our case study on the Chilean Waiting List Corpus, we note that Clinical-ShiftEval is a general framework that can be used to explore many other aspects of distributional robustness and model adaptation across broader clinical NLP scenarios.

Our contribution is twofold. First, we introduce Clinical-ShiftEval, a framework that provides a systematic path to understanding the risks associated with the deployment of clinical NLP models in dynamic healthcare settings. Clinical-ShiftEval enables researchers and practitioners to rigorously assess model reliability, identify critical points of weakness, and evaluate adaptation strategies under controlled and clinically meaningful scenarios. Second, we present a case study that applies this framework to compare continual training, in-context learning, and a hybrid method at varying levels of LSI and TDE. Together, the framework and case study demonstrate both the utility of Clinical-ShiftEval for simulating realistic clinical changes and its practical value for guiding robust and safe deployment of NLP models in evolving healthcare settings.

The Clinical-ShiftEval framework is released as an open-source Python library and can be accessed at https://github.com/fvillena/clinical-shifteval.

Background

Robust clinical NLP systems must operate effectively in environments where both the data and prediction targets evolve. In real-world healthcare, this change is routine: new diseases and clinical categories arise, diagnostic guidelines are updated, and the underlying distributions of both patient characteristics and clinical

documentation practice change over time. Understanding and addressing such changes is fundamental to the advancement of reliable, adaptive NLP models for longitudinal clinical applications.

In the machine learning literature, these types of changes are commonly referred to under two related concepts: *concept drift*, which concerns alterations in the relationship between inputs and outputs, and *concept evolution*, which captures the appearance or disappearance of labels over time. Clearly defining these notions in the clinical NLP context is important, as they provide a precise vocabulary and theoretical foundation for characterizing the distributional challenges this work seeks to address.

Label-set incompatibility

Label-set incompatibility or *concept evolution* arises when entirely new output classes appear in the data, or when old classes become obsolete, a phenomenon also known as label space drift or class-incremental learning. In clinical NLP, this can take the form of new disease codes introduced during a pandemic or outdated conditions. Machine learning approaches to concept evolution must detect candidate instances of novel classes and incorporate them into future model updates or predictive ensembles. Concept evolution is an especially pressing challenge in fields that face rapid expansion of knowledge, regulatory change, or frequent updates to disease ontologies, and must be studied alongside traditional concept drift to ensure model robustness in dynamic healthcare settings [17].

Task definition evolution

Task definition evolution or *concept drift* refers to temporal changes in the data generation process that affect the relationship between predictors X and target labels Y. In machine learning, this is typically formalized as a change in the conditional distribution $P(Y|X)$, with or without concurrent changes in the marginal distribution $P(X)$. Real concept drift occurs when the rule used to assign labels to examples changes, such as when new clinical guidelines redefine disease diagnoses or when prioritization criteria are revised. On the other hand, virtual concept drift occurs when only the input distribution $P(X)$ changes, but the relationship to the outcome remains the same. Different types of drift can be observed, including sudden, gradual, partial (affecting only part of the space), or recurrent (where past distributions reappear, such as seasonal effects) [17]. In clinical practice, task definition evolution is critical because guideline updates or policy changes can quickly render existing models obsolete if not adapted.

Related work

The challenge of concept drift, concept evolution, and maintaining reliable performance in clinical machine learning has received increasing attention in recent years. However, most studies have focused on structured data such as clinical measurements, diagnostic codes, and physiological time series.

Several works have systematically documented the negative effects of dataset shift on model reliability in health contexts. Park et al. [12] demonstrated that out-of-distribution predictions in medical machine learning can be highly unreliable and that out-of-distribution detection methods effectively quantify uncertainty, potentially increasing trust in model recommendations. Similar concerns about performance deterioration have been observed as shifts in data, covariates, or clinical context changes occur, especially during critical events such as the COVID-19 pandemic [11, 14]. These findings underscore the need for systematic monitoring and adaptation when machine learning models are deployed in dynamic health environments.

Recent reviews and case studies have surveyed strategies to mitigate the effects of temporal dataset shift. Guo et al. [4] performed a systematic review of approaches to preserve machine learning performance in the presence of temporal shift, finding that although some techniques can help preserve model performance, they do not fully solve the problem. Auxiliary methods such as feature selection, domain adaptation, and retraining on new data have also been investigated to extend model robustness over time [8, 9].

Several theoretical reviews and opinion pieces highlight the centrality of data drift and dataset shift in the lifecycle of artificial intelligence in healthcare [5–7, 10]. These works emphasize that model performance is intimately linked to the characteristics of the development dataset and that generalization under real-world conditions is a key unaddressed challenge for clinical adoption.

Despite this growing attention, nearly all prior work has focused on structured data types, such as vital signs, laboratory values, and coded diagnoses, rather than the unstructured free text typical of clinical narratives. At the same time, recent benchmark efforts have begun to examine distribution shift in clinical NLP, although with a different scope from ours. CLIFT [18] introduced a testbed for analyzing natural distribution shift in clinical question answering, showing that strong performance on the original test set may not transfer to shifted test distributions. More recently, ConflictMedQA [19] proposed a benchmark to evaluate how LLMs handle medical knowledge drift and conflicting clinical guidelines, particularly in question-answering settings involving updated recommendations. These resources are highly relevant to the broader problem of robustness under clinical change.

Our work differs from these benchmarks in its objective and level of abstraction. Whereas CLIFT and ConflictMedQA are task-specific benchmark datasets, Clinical-ShiftEval is a general framework for constructing controlled dynamic evaluation scenarios on real clinical text. Rather than defining a fixed benchmark for a single task, it parameterizes two reusable forms of change and enables systematic comparison of adaptation strategies across multiple levels of shift severity. In this sense, our contribution is complementary to existing benchmarks: we provide a methodology for generating clinically motivated dynamic scenarios, while prior benchmarks provide fixed testbeds for particular tasks and forms of drift.

Methods

To evaluate the robustness of clinical NLP models in dynamic data settings, we developed Clinical-ShiftEval, a systematic framework to simulate and benchmark model adaptation in two major real-world change scenarios: label-set incompatibility (LSI) and task definition evolution (TDE). All validation experiments were carried out on real-world clinical text corpora, with controlled and parameterized simulations of the shift for each scenario.

Clinical-ShiftEval

To systematically assess the effects of real-world change on clinical NLP model robustness, we constructed controlled simulation scenarios that reflect sources of distributional change in healthcare settings. Clinical-ShiftEval focuses on two prototypical challenges: changes in the set of possible prediction targets (*label-set incompatibility*), and changes in the assignment or definition of task labels without altering the label space (*task definition evolution*). By parameterizing and simulating these challenges in real clinical datasets, we enable fine-grained analyses of model adaptability and performance in a range of plausible healthcare settings. The following subsections detail the formalization and implementation of each simulation scenario.

Label-set incompatibility (LSI)

Label-set incompatibility simulates scenarios in which the set of possible prediction targets changes between two periods, reflecting real-world shifts such as the introduction of new diseases, regulatory requirements, or updates to diagnostic vocabularies. In our framework, both period-specific datasets D_A and D_B are derived from a single underlying dataset D, which can in principle be any clinical dataset containing labeled instances. Formally, let X denote the input space and Y the complete set of labels. For period A, we define the dataset as

$$D_A = \{(x_i, y_i)\}_{i=1}^{n}, \quad \text{with } y_i \in Y_A, \quad Y_A \subseteq Y.$$

For period B, let

$$D_B = \{(x_j, y_j)\}_{j=1}^{m}, \quad \text{with } y_j \in Y_B, \quad Y_B \subseteq Y.$$

The set of labels Y_A and Y_B differs, so $Y_A \neq Y_B$. This formulation allows for both the appearance of new labels in period B (where $Y_B \supset Y_A$), as well as the removal of labels that were present in period A (where $Y_B \subset Y_A$), depending on the specific configuration.

This challenging setting is constructed by altering the label set between period A and B. It simulates real-world clinical scenarios in which models may encounter newly introduced categories or find that previously relevant labels are no longer applicable. We evaluate the model's ability to remain robust and generalizable under these conditions with a focus on how it handles both unseen labels and cases where previously seen labels are no longer part of the active classification space.

To simulate LSI in a flexible and realistic way, we define two parameters that allow us to control different sources of mismatch between periods A and B. These parameters help to evaluate the robustness of NLP models when dealing with emerging and deprecated labels.

- The proportion of deprecated (A-only) labels, denoted $\alpha \in [0, 1]$, is defined as

$$\alpha = \frac{|Y_\mathrm{A} \setminus Y_\mathrm{B}|}{|Y_\mathrm{A}|}.$$

  Here, the operator "\" denotes the set difference, so $Y_\mathrm{A} \setminus Y_\mathrm{B}$ are the labels that were present in period A but do not appear in period B. A value of $\alpha = 0$ means no labels were deprecated, while larger values simulate scenarios where more labels disappear between periods.
- The proportion of new (B-only) labels, denoted $\beta \in [0, 1]$, is defined as

$$\beta = \frac{|Y_\mathrm{B} \setminus Y_\mathrm{A}|}{|Y_\mathrm{B}|}.$$

  In this case, $Y_\mathrm{B} \setminus Y_\mathrm{A}$ are the labels that appear for the first time in period B and did not exist in period A. A value of $\beta = 0$ means no new labels were added, while larger values simulate increasingly larger expansions of the label set.

By varying α and β, we support a wide range of realistic evaluation scenarios, ranging from expanding label spaces (new clinical targets) to shrinking ones (deprecated labels). Together, these parameters provide a comprehensive framework for testing model robustness in

dynamic data environments. A special case arises when both $\alpha = 0$ and $\beta = 0$, which implies that $Y_A = Y_B$. In this setting, the label set remains unchanged between periods and no label-set incompatibility is introduced.

Task definition evolution (TDE)

Task definition evolution models changes in clinical priorities or criteria that redefine what constitutes a positive case, while keeping the input distribution and label set constant. Let X be the input space, Y the fixed label space, and $z \in Z$ a set of hidden attributes associated with each input $x \in X$ (for example, the presence of specific clinical conditions such as diabetes or hypertension). During period A, the model observes a dataset $D_{\mathrm{A}} = \{(x_i, y_i)\}_{i=1}^{n}$, where labels y_i are determined according to a task-specific decision function $f_{\mathrm{A}} : Z \to Y$ that maps hidden attributes to output labels based on current priorities. In period B, the same attributes z can lead to different labeling outcomes under a new decision function $f_{\mathrm{B}} : Z \to Y$, such that $f_{\mathrm{A}}(z) \neq f_{\mathrm{B}}(z)$. The latent space Z and the input distribution $p(X)$ remain unchanged, but the definition of a positive or relevant label under Y evolves.

For example, a clinical note that mentions both diabetes and hypertension can be labeled negative during training if only cancer-related cases are prioritized. If hypertension becomes a priority condition in period B, the same input should now be labeled positive. This change simulates a realistic shift in prediction goals over time without modifying the data, posing challenges for models that assume static task definitions.

In our formulation, a decision function assigns a binary label based on the presence of a subset of hidden attributes associated with each input. Given a set of binary hidden attributes $Z = \{z_1, z_2, \ldots, z_n\}$ extracted from input $x \in X$, both the period A and the B decision functions are defined as disjunctions over subsets of attributes:

$$f(z) = \bigvee_{z_i \in S} z_i = 1,$$

where $S \subseteq Z$ is the set of attributes considered relevant under a specific task definition. This equation indicates that the output of the decision function f is positive (1) if any of the attributes in the relevant subset S are present (that is, $z_i = 1$ for at least one z_i in S). The operation $\bigvee$ denotes the logical OR, so $f(z)$ implements a logical "any" over the attributes of interest.

In our case, if the hidden attributes represent diseases encoded using the ICD-10 vocabulary, then the size of Z corresponds to the number of distinct ICD-10 codes and $S \subseteq Z$ represents the subset of codes that are prioritized at a given point in time. A patient is labeled as positive if they are associated with at least one prioritized condition and negative otherwise. It is important to note that simulating task definition evolution requires access to such hidden attributes. Therefore, not all datasets are suitable for this type of experiment: only datasets that include auxiliary or structured labels (e.g., ICD codes linked to free-text notes) can be used to model task definition evolution in this way.

We define two task-specific decision functions:

$f_{\mathrm{A}} : Z \to Y$ uses a set $S_{\mathrm{A}} \subseteq Z$ of relevant attributes during period A.

$f_{\mathrm{B}} : Z \to Y$ uses a different set $S_{\mathrm{B}} \subseteq Z$ of relevant attributes during period B.

To systematically simulate task definition evolution across periods, we introduce two parameters, γ_1 and γ_2 , that govern the transitions in the relevance of the attributes between period A and period B. Specifically, for each attribute in Z:

- γ_1 denotes the probability that an attribute not relevant in period A becomes newly relevant in period B (i.e., is added to S_{B} but not S_{A});
- γ_2 is the probability that an attribute relevant in period A loses relevance in period B (i.e., is removed from S_{B}).

Adjusting γ_1 and γ_2, we can represent a wide range of change scenarios, from environments where clinical priorities remain highly stable to those where criteria for positive cases change substantially.

Parameter sweep for scenario parameterization

To ensure that our evaluation encompassed a realistic range of change stengths, we systematically explored the parameter spaces for both LSI and TDE using a parameter sweep. By parameter sweep, we mean a grid-based exploration in which all relevant parameters are varied across their possible ranges in a systematic manner. For each scenario, we evaluated all combinations of relevant parameters by training a supervised model on period A data and measuring its performance drop when tested on period B data, relative to its baseline performance in period A. This approach allowed us to robustly quantify how different degrees and types of clinical shifts degrade model performance.

Based on the distribution of observed performance drops across this grid, we selected representative parameter settings that resulted in a performance drop of 25, 50 and 75% in the F_1 score. These levels were used to define the simulation strengths of *low*, *medium*, and *high* for both dynamic scenarios. All comparative analyses of adaptation strategies were performed in these calibrated settings, ensuring that our results reflect a range of

practically meaningful and interpretable degrees of distributional and task change.

Case study

To demonstrate the practical utility of Clinical-ShiftEval, we conducted a case study that applies the framework to real-world clinical data. The purpose of this case study is twofold: (i) to validate that the framework reliably simulates realistic distributional and task shifts with controlled severity, and (ii) to compare representative adaptation strategies under these conditions. By grounding the simulations in authentic EHR text, we ensure that the evaluation reflects both the complexity of clinical documentation and the evolving challenges faced in deployment. The case study focuses on two clinically significant tasks: referral specialty classification and disease prioritization, chosen because they naturally lend themselves to scenarios of LSI and TDE. Through this setup, we can assess how different adaptation paradigms recover performance under dynamic clinical conditions.

Adaptation strategies

Clinical-ShiftEval is agnostic with respect to the choice of adaptation method: any learning strategy that addresses distributional shifts or task redefinitions can be evaluated within the framework. To illustrate its practical use, we designed a case study comparing three representative approaches: continual training, in-context learning, and a hybrid method. These were selected because they capture three complementary paradigms of model retraining, prompt-based adaptation of large language models (LLMs), and coordinated integration of legacy discriminative models with generative LLMs. Importantly, Clinical-ShiftEval is not limited to these strategies.

Continual training In this approach, models are initially trained on data from period A and then incrementally updated with data from period B. This allows the model to gradually adapt its parameters as new distributions or label sets emerge in the deployment environment.

For continual training experiments, we used the `PlanTL-GOB-ES/bsc-bio-ehr-es`, a RoBERTa-base model [20], as it represents the current state of the art for the Spanish biomedical and clinical text. Each task was fine-tuned using the default model hyperparameters, and training was performed for five epochs using the HuggingFace `transformers` library.

In-context learning For this strategy, LLMs are conditioned in inference time using prompt-based techniques. Task descriptions and label definitions are included directly in the prompt, allowing the model to rapidly adapt to new tasks or changes in the label space without the need for parameter updates or retraining.

In our experiments, we used the `dspy` Python library [21] together with the `MIPROv2` optimizer [22] to automatically optimize prompt instructions and demonstrations. We used only 1% of the examples for prompt optimization. For each task, the optimizer was run using the light auto configuration, with 10 trials, minibatch optimization enabled, 6 few-shot candidates, 3 instruction candidates, and a validation set size of 100 examples.

We evaluated the following LLMs: `GPT-4.1` and `GPT-5` through the OpenAI API, and `DeepSeek-V3-0324` and `DeepSeek-R1` through OpenRouter. Unless otherwise noted, inference used the following parameters: temperature = 1.0, top_p = 1.0, and max_tokens = 2048. We evaluated the models in a zero-shot setting with optimized instructions, where the prompt included the task description and the current label space or task definition, but no manually selected annotated examples from period B at inference time.

Hybrid method The hybrid adaptation strategy combines a legacy discriminative model trained on period A data with an LLM acting as an adaptive decision agent. During inference, the agent receives: (i) the clinical text, (ii) the current task definition for period B, including the valid label space for LSI or the updated prioritization criteria for TDE, and (iii) the prediction produced by the legacy model, which is treated as an auxiliary suggestion rather than a binding output.

The legacy model output is provided as an auxiliary suggestion rather than a binding prediction. The agent is explicitly instructed to evaluate whether this suggestion is consistent with the current period B rules. If the legacy prediction is compatible with the updated label set or task definition, the agent may retain it; otherwise, it is instructed to override it and produce a corrected prediction based on the clinical text and the current instructions. This design allows the agent to leverage the legacy model as a source of prior information while remaining aligned with the updated clinical requirements.

The hybrid prompts were also optimized using `dspy` with `MIPROv2`, using the same parameters as in the last subsection. The underlying LLMs and inference parameters were the same as those used in the in-context learning experiments.

An overview of this orchestrator workflow, including the agent's decision-making process given both the task context and the legacy model outputs, is presented in Fig. 1.

Evaluation protocol and metrics

For each simulation scenario, model performance was measured with respect to both standard and dynamic metrics.

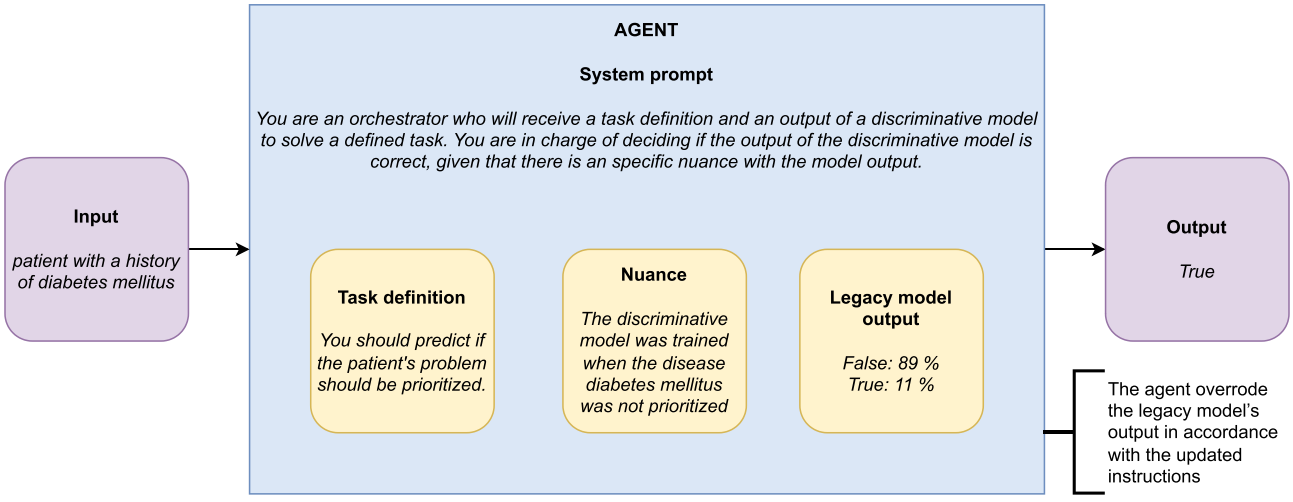


Fig. 1 Illustrative example of the agent in the hybrid method fixing a discriminative output because the task definition changed

- **Macro F_1 score**, measuring the conventional classification performance.
- **Performance drop**, defined as the relative reduction in F_1 between baseline (period A or pre-shift) and evaluation (period B or post-shift):

$$\text{Performance drop} = \frac{F_1^A - F_1^B}{F_1^A}$$

- **Adaptation speed**: For continual adaptation experiments, the proportion of period B data required for a model to recover 85% of its pre-shift performance.

All experiments were repeated with three different random seeds, and the results are reported as the mean across the runs.

Data sources and experimental setup

To demonstrate the applicability of Clinical-ShiftEval, we conducted the case study using clinical data derived from the Chilean Waiting List Corpus [2, 23, 24]. This corpus originates from Chilean public health institutions and consists of referral records written in Spanish, where the reason for patient referral is expressed in free text. The multitask nature of this dataset provides a suitable setting for simulating clinically meaningful scenarios involving both stable and evolving prediction targets.

It is important to emphasize that Clinical-ShiftEval is not tied to this dataset: any clinical corpus could be used to construct different case studies. We encourage future work to build on this framework by applying it to diverse datasets, languages, and healthcare domains. To further illustrate this portability, we include supplementary experiments on two English-language biomedical datasets in Appendix 1.

The period A and period B splits for our case study were generated using Clinical-ShiftEval, which parameterizes changes in label sets and task definitions to mimic realistic clinical scenarios.

Referral specialty classification To model LSI, we consider the task of predicting the appropriate medical specialty for each referral based solely on the free-text reason for referral. This is a multilabel classification problem with 48 possible specialty classes. Period A and period B were simulated by partitioning the set of specialty labels: for each configuration, different subsets of specialties were assigned to periods A and B according to the LSI parameters α and β. Consequently, the set of available specialties differed across periods, allowing new specialties to appear in period B and others to become obsolete (present only in period A). The dataset is available at https://doi.org/10.57967/hf/6298

Disease prioritization To model TDE, we used the ICD-10 annotations available in the dataset. Each referral is associated with a set of diseases encoded as hidden attributes. Within our framework, a subset of these diseases is designated as "prioritized" for each period: period A and period B are assigned different subsets, simulating changes in clinical guidelines, prioritization criteria, or health policies. This setup enables systematic evaluation of how models handle evolving task definitions while keeping the input data fixed. The dataset is available at https://doi.org/10.5281/zenodo.7757971

Error analysis

To complement the aggregate evaluation metrics, we performed an additional error analysis to better characterize model failure modes under both dynamic scenarios.

For the LSI scenario, we analyzed normalized confusion matrices for the referral specialty classification task. These matrices were used to compare the behavior of the continual training model in three representative settings: baseline evaluation on period A, evaluation on period B without adaptation, and evaluation on period B after retraining with period B data.

For the TDE scenario, we complemented Macro F_1 with two error-type metrics tailored to the binary prioritization setting: the false negative rate (FNR) and the false positive rate (FPR). These are defined as

$$\text{FNR} = \frac{FN}{TP + FN}, \qquad \text{FPR} = \frac{FP}{TN + FP},$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. As in the LSI analysis, we compared these quantities across baseline evaluation on period A, evaluation on period B without adaptation, and evaluation on period B after retraining.

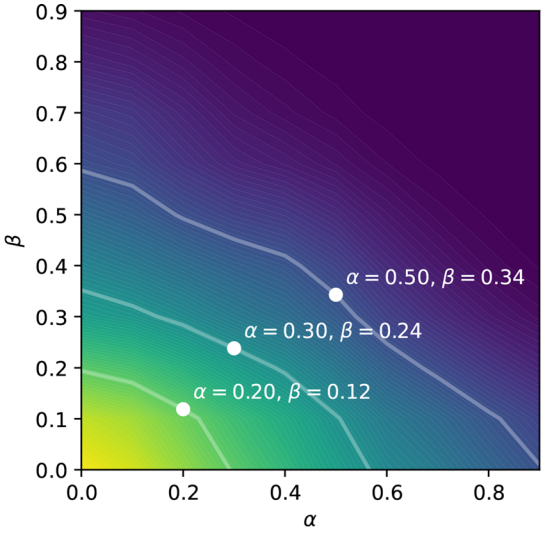
For both scenarios, this analysis was conducted using a representative Clinical-ShiftEval configuration whose parameters were selected from the parameter sweep to produce an approximate performance drop of 0.5, corresponding to the medium shift level.

Results

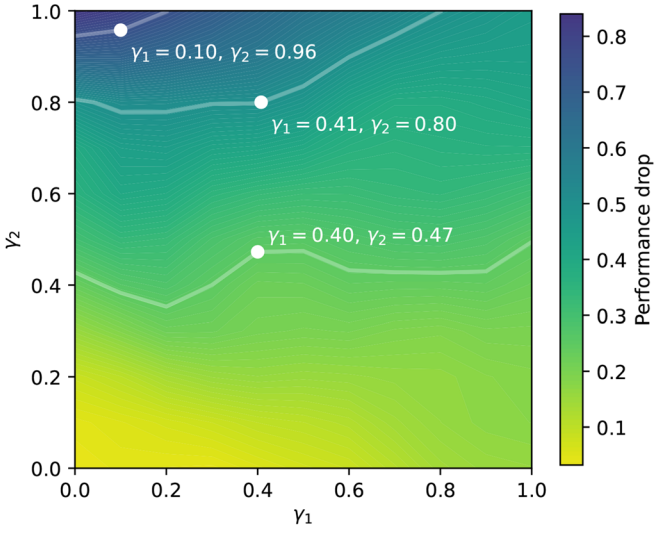
We first analyzed the relationship between simulation parameters and model performance using a comprehensive parameter sweep in Clinical-ShiftEval. For each scenario (LSI and TDE), we systematically evaluated all combinations of parameters, quantifying the performance drop experienced by the baseline models under each condition. This approach enabled us to identify and define low, medium, and high simulation strengths. The complete results of the parameter sweep, visualized in Fig. 2 for LSI and TDE, reveal the range and structure of the performance degradation in the parameter space.

Continual training

Figure 3 shows the performance drop for the continual training method as a function of the proportion of period B data used for retraining, for both the LSI and TDE scenarios. Across both scenarios and all simulation strengths (low, medium and high), the performance drop was highest when the period B data were not incorporated and decreased as more fractions of new data were used for retraining. In the LSI setting (Fig. 3a), more severe simulation strength was associated with a greater initial performance degradation, but this difference decreased as retraining included more period B data, with the performance drop converging between strengths when at least 40% of period B data was used. In the TDE scenario (Fig. 3b), a similar trend was observed: all simulation strengths began with high performance drop, but retraining with as little as 20% of period B data reduced the drop to minimal levels. Regarding adaptation speed, we observed a

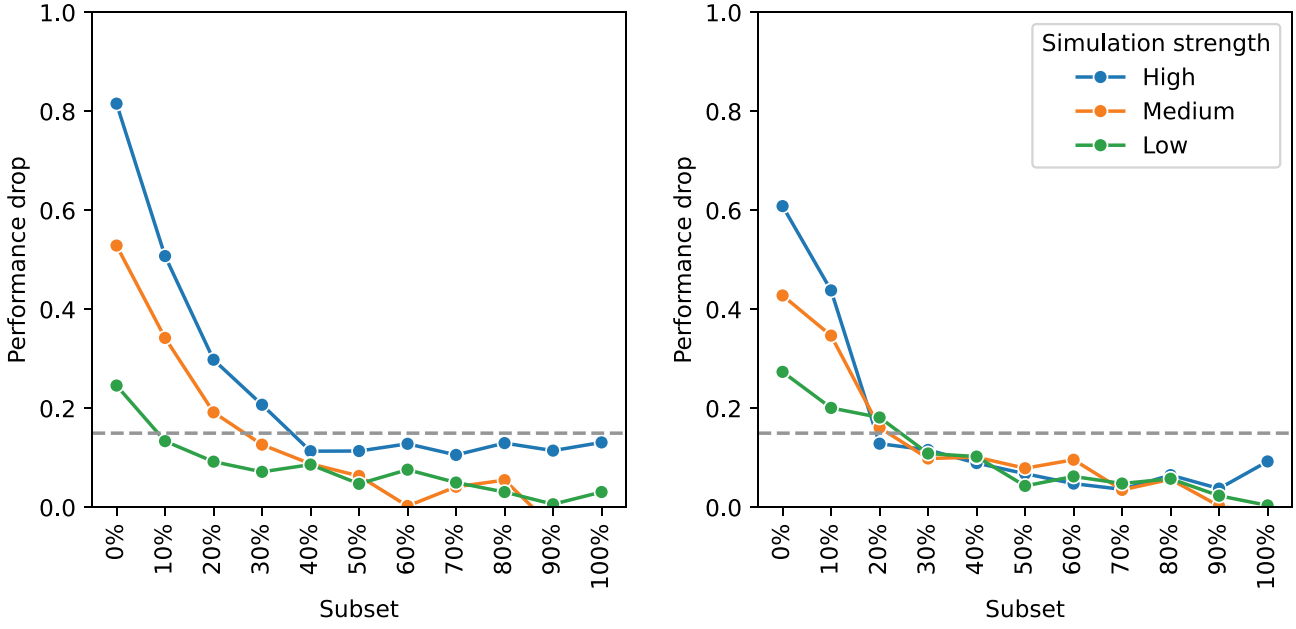


(a) Label set incompatibility (parameters α and β).



(b) Task definition evolution (parameters γ_1 and γ_2).

Fig. 2 Performance drop as a function of scenario parameters for (**a**) label set incompatibility and (**b**) task definition evolution. Contour lines correspond to performance drop levels of 0.25 (lowest controur line), 0.5, and 0.75 (highest contour line), and markers indicate representative points on each contour



(a) Label set incompatibility scenario. (b) Task definition evolution scenario.

Fig. 3 Performance drop for the continual training method as a function of the proportion of period B data used for retraining: (**a**) label set incompatibility and (**b**) task definition evolution. The dashed line represent and 15% performance drop, used to measure adaptation speed

clear correlation in the LSI setting: a higher simulation strength required a higher proportion of period B data for the model to achieve a 15% performance drop. This indicates that more severe shifts require greater data exposure for effective adaptation. In contrast, for TDE, the adaptation speed did not show such a correlation, all simulation strengths converged to a 15% performance drop after incorporating a similar amount of period B data, regardless of the initial severity of the change.

In-context learning

Figure 4 compares the performance of in-context learning and continual training paradigms in both LSI and TDE scenarios. In both scenarios, in-context learning achieves competitive performance without requiring any new period B training data. However, as the proportion of period B data used for continual training increases, the performance drop for continual training decreases and eventually falls below that of in-context learning. This crossover point indicates the regime where ongoing model retraining begins to outperform in-context learning, especially in more severe shift settings or as more new data become available. In particular, the performance advantage of in-context learning is most pronounced when only a small fraction (or none) of the new data are accessible for model update.

A detailed comparison of the LLM-based paradigms is presented in Fig. 6, which shows that among the in-context learning paradigm models, GPT-5 consistently achieved the lowest performance drop, outperforming the other LLMs in both the LSI and TDE scenarios.

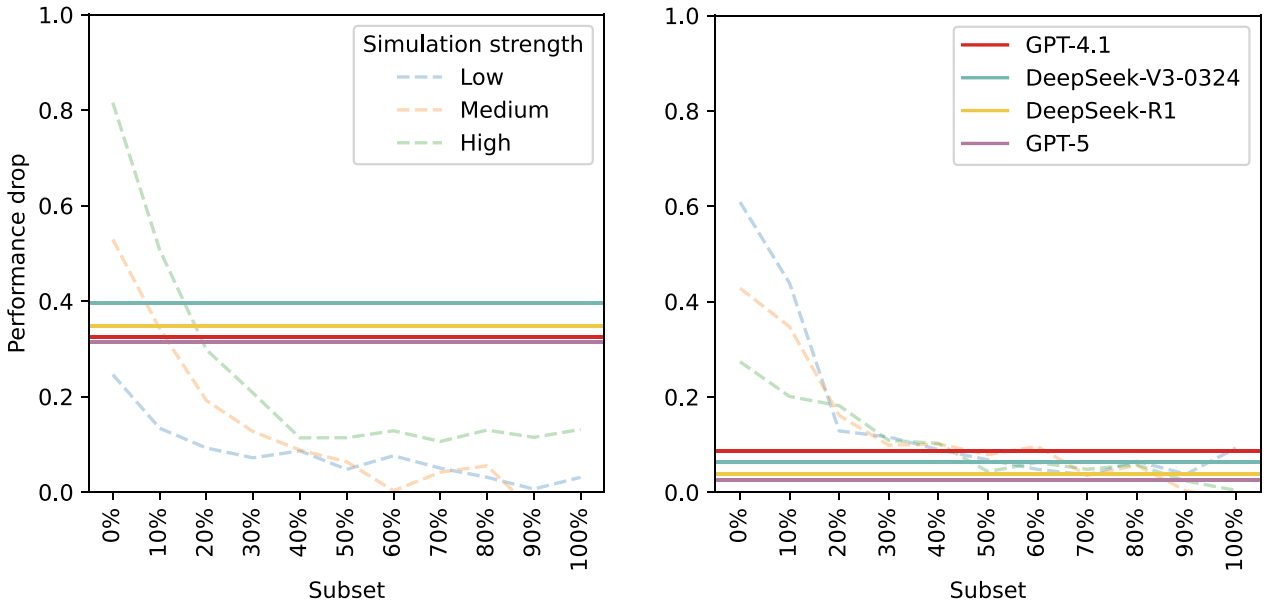
Hybrid method

Figure 5 presents the performance drop for the hybrid paradigm in the LSI and TDE scenarios. In both scenarios, the hybrid approach achieves a low performance drop. For the LSI scenario, the hybrid paradigm some models consistently outperforms continual training in low proportions of period B data and maintains its advantage until continual retraining is performed with approximately 40% of new data, at which point the performance becomes comparable. In the TDE scenario, a similar pattern is observed: the hybrid method is highly robust initially and remains competitive as more new data is used for retraining, where continual training only surpasses it after substantial integration of period B samples.

Figure 6 further illustrates that the hybrid paradigm achieves lower performance drops than in-context learning, the advantage being particularly marked under LSI, while in TDE the difference between the two approaches is smaller.

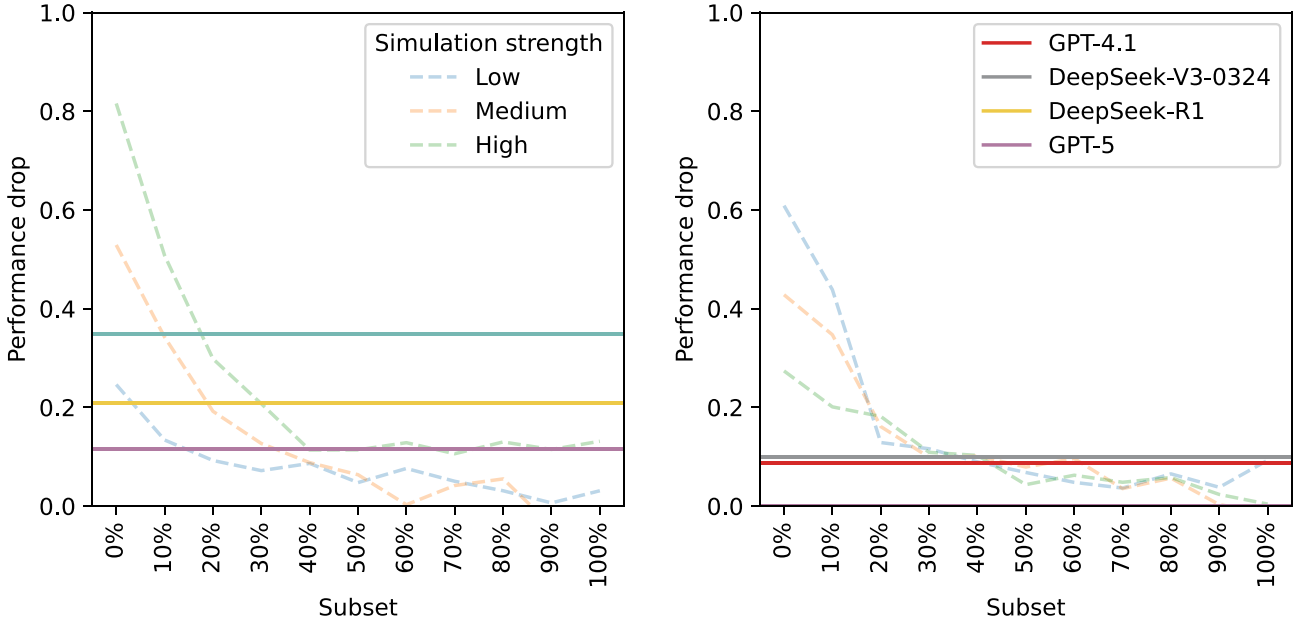
Error analysis

For the LSI scenario, Fig. 7 shows the confusion matrices for the specialty classification task under three representative conditions: evaluation on period A, evaluation on period B without retraining, and evaluation on period B after retraining with all available period B data. On period A, the model exhibits a clear diagonal structure.



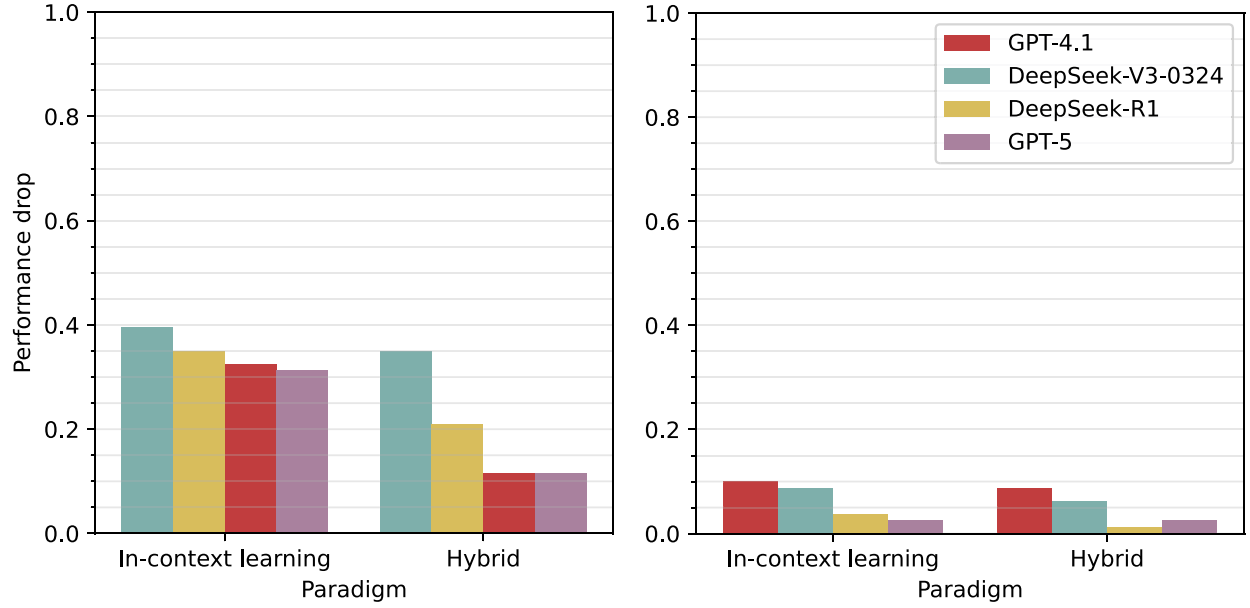
(a) Label set incompatibility scenario. (b) Task definition evolution scenario.

Fig. 4 The solid lines show the performance drop for the in-context learning paradigm in (**a**) label set incompatibility and (**b**) task definition evolution scenarios. Dashed lines in the background indicate the performance drop for the continual training paradigm as the proportion of period B data used for retraining increases, providing context for when continual training surpasses in-context learning



(a) Label set incompatibility scenario. (b) Task definition evolution scenario.

Fig. 5 The solid lines show the performance drop for the hybrid paradigm in (**a**) label set incompatibility and (**b**) task definition evolution scenarios. Dashed lines show the performance drop for the continual training paradigm as the proportion of period B data used for retraining increases, providing context for when continual training surpasses the hybrid method



(a) Label set incompatibility scenario. (b) Task definition evolution scenario.

Fig. 6 Performance drop for the LLM-based paradigms in (**a**) label set incompatibility and (**b**) task definition evolution scenarios

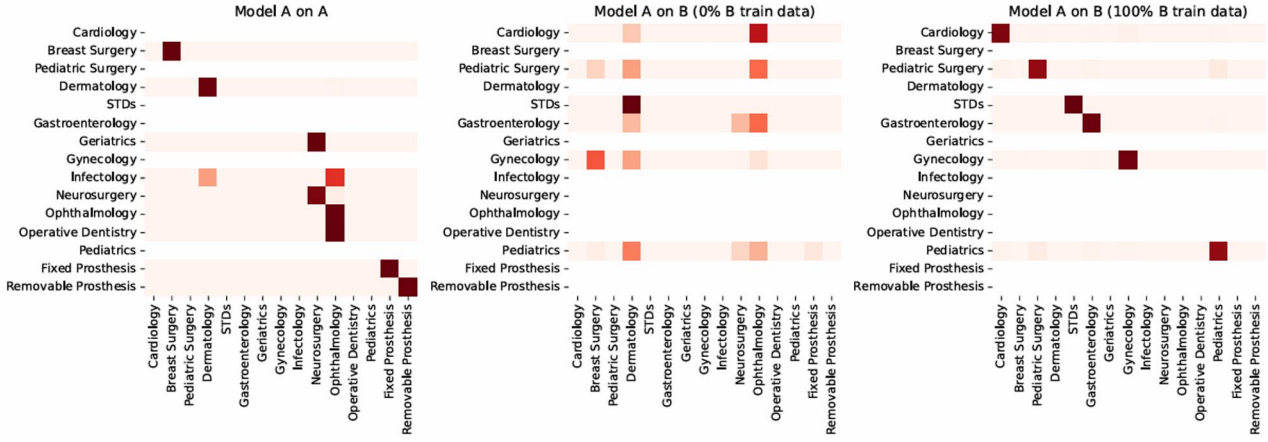


Fig. 7 Normalized confusion matrices for the LSI scenario under three representative conditions: baseline evaluation on period A, evaluation on period B without retraining, and evaluation on period B after retraining with 100% of period B data

When the same model is applied to period B without adaptation, the diagonal weakens substantially and systematic off-diagonal confusions emerge. After retraining on period B data, the diagonal structure is largely restored.

For the TDE scenario, Fig. 8 reports the FNR and FPR for the binary prioritization task. When the model trained on period A was directly evaluated on period B without retraining, both error rates increased substantially relative to the baseline evaluation on period A. After retraining on period B data, both FNR and FPR decreased markedly, approaching baseline levels.

Discussion

In this study, we introduced Clinical-ShiftEval, a benchmarking framework designed to systematically evaluate model adaptation in dynamic clinical NLP tasks using real-world EHR text. Our results demonstrate the importance of modeling both LSI and TDE, revealing the distinct ways in which each type of distributional shift degrades model performance and influences adaptation requirements.

Our experiments show that conventional supervised models trained under a static regime are highly sensitive to both LSI and TDE. In line with previous evidence from conventional machine learning studies on data set shift and temporal drift [4, 11, 14], we also observe

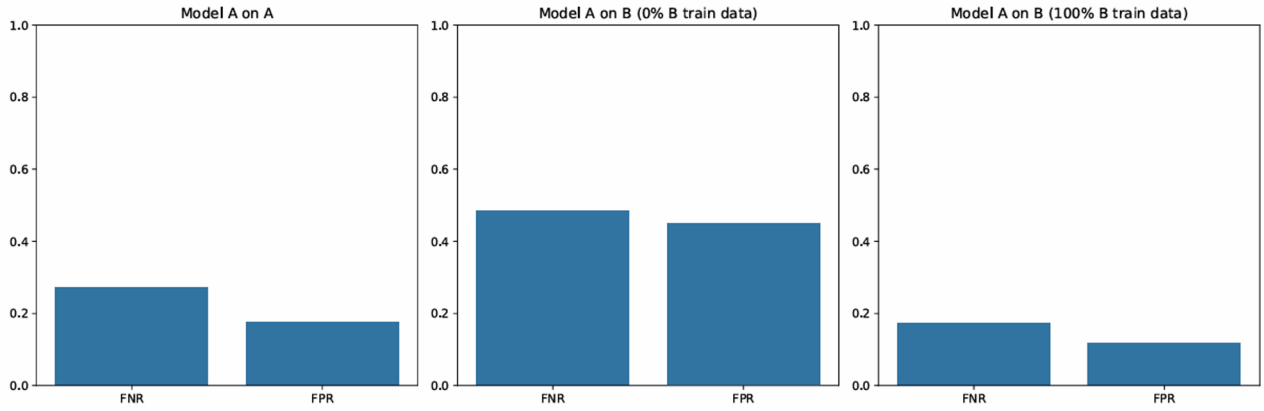


Fig. 8 FNR and FPR for the TDE scenario under three representative conditions: baseline evaluation on period A, evaluation on period B without retraining, and evaluation on period B after retraining with 100% of period B data

severe performance drops, up to 82% in F1 for LSI and 43% for TDE, which underscores the risks of deploying fixed models in evolving clinical environments. In particular, we observed that in LSI, the proportion of newly introduced labels (controlled by β) exerts a significantly stronger effect on performance degradation than the removal of deprecated labels (α). This pattern mirrors observations reported in the conventional machine learning literature on data streams [17, 25], and here we confirm that similar challenges also arise in the context of clinical NLP, where models struggle most when required to handle previously unseen categories. In the TDE setting, the performance drop increased more strongly with higher values of γ_2, which represents the probability that a relevant attribute becomes irrelevant. The gradient of the heatmap is noticeably steeper along the γ_2 axis than along the γ_1 axis, indicating that the removal of previously relevant attributes leads to a greater degradation in model performance than the introduction of new ones. This suggests that the model reliability under the evolution of the task definition is particularly sensitive to the loss of existing criteria in the definition of positive cases.

Among the adaptation strategies tested, in-context learning using LLMs achieved strong robustness to shift without access to new data, yielding a 35% drop in F1 in the LSI scenario and up to a 10% drop in F1 in the TDE setting. These results are consistent with previous studies in general NLP that demonstrate the effectiveness of in-context learning for rapid task adaptation [26, 27], and here we show that the same paradigm also provides substantial benefits in clinical NLP tasks. The hybrid method, which combines in-context learning with access to a legacy discriminative model, further reduced this drop to around 8% in both scenarios. Continual adaptation through incremental retraining outperformed the other methods only after a substantial amount (at least 30%) of period B data was incorporated, illustrating a key trade-off between data requirements and adaptation speed [28, 29]. These trends were held consistently across both types of shift and across all levels of simulation strength.

An additional observation from our experiments is the clear differentiation in performance between the LLMs evaluated. Among the in-context learning models, GPT-5 consistently achieved the lowest performance drop, with DeepSeek-R1 also performing competitively in the TDE scenario. This suggests that not all LLMs are equally suited for adaptation in clinical NLP tasks, and choosing the underlying model can influence robustness outcomes. Furthermore, when comparing paradigms, the hybrid approach consistently outperformed in-context learning, with the advantage being more pronounced for LSI than for TDE. This indicates that combining the discriminative strength of legacy models with the adaptability of LLMs is especially advantageous when models must handle entirely new labels, whereas TDE can often be effectively managed by LLMs alone.

A closer inspection of the confusion patterns in the LSI scenario also suggests that not all errors have the same practical severity. Some misclassifications remain clinically plausible because they occur between related specialties. For example, referrals labeled as STDs were misclassified only as Dermatology, which may be less concerning in practice, since dermatology services often manage sexually transmitted infections with cutaneous or mucosal manifestations. Likewise, Gynecology referrals were sometimes misclassified as Breast Surgery, a confusion that is also clinically understandable given the overlap in referral pathways and women's health services in some healthcare systems. In contrast, referrals from broader categories such as Pediatrics and Pediatric Surgery were distributed across several specialties, suggesting a more diffuse error pattern. This may reflect the more general scope of these services, whose

referrals can reasonably span multiple organ systems and subspecialties.

In the TDE scenario, the simultaneous increase in both error rates after the shift suggests that legacy models not only fail to detect newly prioritized cases, but also continue to act on outdated criteria. In practice, this may compromise both the efficiency and the equity of prioritization decisions.

Collectively, these findings suggest that prompt-based and hybrid LLM approaches provide a practical means for rapid adaptation in healthcare settings where labeled data for new periods may be scarce or delayed, especially in languages other than English [30], while continual training remains the most powerful option when sufficient new data becomes available. This complex understanding is especially relevant for clinical deployments, where timely operations and patient safety often require robust solutions even before large-scale data collection or model retraining is feasible.

The flexibility and fine-grained control of Clinical-ShiftEval enable researchers and practitioners to probe robustness and adaptation strategies under a spectrum of realistic, clinically motivated scenarios. By modeling both abrupt and gradual changes and employing real EHR narratives instead of synthetic data, our framework produces benchmarks that closely mirror the complexity of operational healthcare data. This is crucial for modern clinical NLP, where guideline updates, institutional workflow changes, and evolving disease landscapes are frequent and consequential.

In addition, Clinical-ShiftEval accommodates the evaluation of emerging adaptation paradigms, supporting fair and interpretable comparisons between continuous learning, prompt-based, and hybrid solutions. This is timely given the proliferation of new foundation models and in-context learning strategies in clinical NLP.

Conclusion

In this study, we introduced Clinical-ShiftEval, a novel framework for simulating and evaluating model adaptation in dynamic clinical NLP tasks using real-world EHR text. By parameterizing both LSI and TDE, our framework enables systematic benchmarking of model robustness under a wide range of realistic clinical change scenarios.

Our results show that static models are vulnerable to substantial performance degradation as data and labeling priorities change, while in-context learning and hybrid approaches substantially reduce this impact even without additional retraining. Continual training remains the most effective long-term strategy when sufficient new data is available.

Beyond aggregate performance, our error analysis shows that dynamic shifts give rise to clinically meaningful failure modes, including incorrect referral routing under label-set changes and missed or spurious prioritization under evolving task definitions

Clinical-ShiftEval offers a fine-grained, data-driven platform to compare and improve adaptation strategies in a rapidly evolving technological landscape. This broader applicability is further illustrated by supplementary experiments on English-language biomedical datasets presented in Appendix 1. We believe that our framework will accelerate progress toward the deployment of safe, reliable, and adaptable NLP models in real clinical settings and motivate further research into dynamic, robust clinical language understanding.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-026-03538-6.

Supplementary Material 1

Acknowledgements
Not applicable.

Author contributions
All authors led the conception and design of the work; FV and JD led the acquisition of the data; FV analyzed the data; all authors interpreted the data; FV drafted the work and all authors and revised it; all authors approved the submitted version; all authors have agreed both to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and the resolution documented in the literature.

Funding
This work was funded by ANID Chile: Millennium Science Initiative Program ICN17_002 - IMFD, ICN2021_004 - iHEALTH, Basal Funds FB210017 - CENIA and CIA250006 - AC3E, National Doctoral Scholarship 21220200 (FV), and Fondecyt grant 1241825 (JD).

Data availability
All datasets used in this publication are available for download:– Referral specialty classification: https://doi.org/10.57967/hf/6298 – Disease prioritization: https://doi.org/10.5281/zenodo.7757971

Declarations

Ethics approval and consent to participate
The ethics committee of the South East Metropolitan Health Service has waived the need for approval and the datasets have been published before this publication. The Waiting List Dataset was collected through the Chilean Transparency Law (http://www.portaltransparencia.cl).

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]School of Dentistry, Pontificia Universidad Católica de Chile, Santiago, Chile
[2]Department of Computer Science, Universidad de Chile, Santiago, Chile
[3]Department of Computer Science, Pontificia Universidad Católica de Chile, Santiago, Chile

[4]Institute for Mathematical and Computing Engineering, Pontificia Universidad Católica de Chile, Santiago, Chile

Received: 9 September 2025 / Accepted: 27 April 2026
Published online: 11 May 2026

References

1. Villena F, Pérez J, Lagos R, Dunstan J. Supporting the classification of patients in public hospitals in Chile by designing, deploying and validating a system based on natural language processing. BMC Med Inf Decis Mak. 2021;21(1). https://doi.org/10.1186/s12911-021-01565-z.
2. Báez P, Bravo-Marquez F, Dunstan J, Rojas M, Villena F. Automatic extraction of nested entities in clinical referrals in Spanish. ACM Trans Comput Healthcare. 2022;3(3):1–22. https://doi.org/10.1145/3498324.
3. Villena FNAIN, Rojas M, Arias F, Pacheco J, Vera P, Dunstan J. In: Naumann T, Ben Abacha A, Bethard S, Roberts K, Rumshisky A, editors. Automatic coding at scale: design and deployment of a nationwide system for normalizing referrals in the Chilean public healthcare system. Proceedings of the 5th Clinical Natural Language Processing Workshop. Toronto, Canada: Association for Computational Linguistics; 2023, pp. 335–43. https://aclanthology.org/2023.clinicalnlp-1.37/.
4. Guo LL, Pfohl SR, Fries J, Posada J, Fleming SL, Aftandilian C, et al. Systematic review of approaches to preserve machine learning performance in the presence of temporal dataset shift in clinical medicine. Appl Clin Inf. 2021;12(4):808–15. https://doi.org/10.1055/s-0041-1735184.
5. Zhang A, Xing L, Zou J, Wu JC. Shifting machine learning for healthcare from development to deployment and from models to data. Nat Biomed Eng. 2022;6(12):1330–45. https://doi.org/10.1038/s41551-022-00898-y.
6. Sahiner B, Chen W, Samala RK, Petrick N. Data drift in medical machine learning: implications and potential remedies. The Br J Radiol. 2023;96(1150). https://doi.org/10.1259/bjr.20220878.
7. Lea AS, Jones DS. Mind the gap — machine learning, dataset shift, and history in the age of clinical algorithms. N Engl J Med. 2024;390(4):293–95. https://doi.org/10.1056/nejmp2311015.
8. Guo LL, Pfohl SR, Fries J, Johnson AEW, Posada J, Aftandilian C, et al. Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. Sci Rep. 2022;12(1). https://doi.org/10.1038/s41598-022-06484-1.
9. Lemmon J, Guo LL, Posada J, Pfohl SR, Fries J, Fleming SL, et al. Evaluation of feature selection methods for preserving machine learning performance in the presence of temporal dataset shift in clinical medicine. Methods Inf Med. 2023;62(1/02):060–070. https://doi.org/10.1055/s-0043-1762904.
10. Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health ai. Biostatistics. 2019. https://doi.org/10.1093/biostatistics/kxz041.
11. Rahmani K, Thapa R, Tsou P, Casie Chetty S, Barnes G, Lam C, et al. Assessing the effects of data drift on the performance of machine learning models used in clinical sepsis prediction. Int J Multiling Med Inf. 2023;173:104930. https://doi.org/10.1016/j.ijmedinf.2022.104930.
12. Park C, Awadalla A, Kohno T, Patel S. Reliable and trustworthy machine learning for health using dataset shift detection. Proceedings of the 35th International Conference on Neural Information Processing Systems. NIPS '21. Red Hook, NY, USA: Curran Associates Inc; 2021.
13. Villena F, Bravo-Marquez F, Dunstan J. NLP modeling recommendations for restricted data availability in clinical settings. BMC Med Inf Decis Mak. 2025;25(1). https://doi.org/10.1186/s12911-025-02948-2.
14. Duckworth C, Chmiel FP, Burns DK, Zlatev ZD, White NM, Daniels TWV, et al. Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19. Sci Rep. 2021;11(1). https://doi.org/10.1038/s41598-021-02481-y.
15. Laparra E, Mascio A, Velupillai S, Miller, Miller T. A review of recent work in transfer learning and domain adaptation for natural language processing of electronic health records. Yearb Med Inf. 2021;30(1):239–44. https://doi.org/10.1055/s-0041-1726522.
16. Sarafraz G, Behnamnia A, Hosseinzadeh M, Balapour A, Meghrazi A, Rabiee HR. Notice of removal March 3, 2026: domain adaptation and generalization of functional medical data: a systematic survey of brain data. ACM Comput Surv. 2024;56(10):1–39. https://doi.org/10.1145/3654664.
17. Bifet A, Gavaldà R, Holmes G, Pfahringer B. Machine learning for data streams: with practical examples in MOA. Cambridge, MA: The MIT Press; 2018. https://doi.org/10.7551/mitpress/10654.001.0001.
18. Pal A. CLIFT: analysing natural distribution shift on question answering models in clinical domain. NeurIPS 2022 Workshop on Robustness in Sequence Modeling. 2022. https://openreview.net/forum?id=9PQFROOfqm.
19. Wu W, Xu X, Gao C, Diao X, Li S, Salas LA, et al. Assessing and mitigating medical knowledge drift and conflicts in large language models. In: Christodoulopoulos C, Chakraborty T, Rose C, Peng V, editors. Findings of the association for computational linguistics: eMNLP 2025. Suzhou, China: Association for Computational Linguistics; 2025. p. 707–30. https://aclanthology.org/2025.findingsemnlp.38/.
20. Carrino CP, Llop J, Pàmies M, Gutiérrez-Fandiño A, Armengol-Estapé J, Silveira-Ocampo J, et al. Pretrained biomedical language models for clinical NLP in Spanish. Proceedings of the 21st Workshop on Biomedical Language Processing. Dublin, Ireland: Association for Computational Linguistics; 2022, pp. 193–99). https://aclanthology.org/2022.bionlp-1.19.
21. Khattab O, Singhvi A, Maheshwari P, Zhang Z, Santhanam K, Vardhamanan S, et al. Dspy: compiling declarative language model calls into self-improving pipelines. 2024.
22. Opsahl-Ong K, Ryan MJ, Purtell J, Broman D, Potts C, Zaharia M, et al. Optimizing instructions and demonstrations for multi-stage language model programs. 2024. https://arxiv.org/abs/2406.11695.
23. Báez P, Villena F, Rojas M, Durán M, Dunstan J. The Chilean waiting list corpus: a new resource for clinical named entity recognition in Spanish. Proceedings of the 3rd Clinical Natural Language Processing Workshop. Online: Association for Computational Linguistics; 2020). https://doi.org/10.18653/v1/2020.clinicalnlp-1.32. 291–300. https://aclanthology.org/2020.clinicalnlp-1.32.
24. Báez P, Campillos-Llanos L, Núñez F, Dunstan J. Entity normalization in a Spanish medical corpus using a UMLS-based lexicon: findings and limitations. Lang Resour Evaluation. 2025;59(2):1013–41. https://doi.org/10.1007/s10579-024-09755-7.
25. Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia, Bouchachia A. Survey on concept drift adaptation. ACM Comput Surv. 2014;46(4):1–37. https://doi.org/10.1145/2523813.
26. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.). Adv. Neural Inf. Process. Syst. 2020;33:1877–901. Curran Associates, Inc., Vancouver, Canada. https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html
27. Dong Q, Li L, Dai D, Zheng C, Ma J, Li R, et al. In: Al-Onaizan Y, Bansal M, Chen Y-N, editors. A survey on in-context learning. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami, Florida, USA: Association for Computational Linguistics; 2024, pp. 1107–28. https://aclanthology.org/2024.emnlp-main.64/.
28. Delange M, Aljundi R, Masana M, Parisot S, Jia X, Leonardis A, et al. A continual learning survey: defying forgetting in classification tasks. IEEE Trans Pattern Anal Mach Intell. 2021;1–1. https://doi.org/10.1109/tpami.2021.3057446.
29. Parisi GI, Kemker R, Part JL, Kanan C, Wermter S. Continual lifelong learning with neural networks: a review. Neural Networks. 2019;113:54–71. https://doi.org/10.1016/j.neunet.2019.01.012.
30. Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical natural language Processing in languages other than English: opportunities and challenges. J Biomed Semant. 2018;9(1). https://doi.org/10.1186/s13326-018-0179-8.

Publisher's Note