Stale by Design: Temporal Entity Drift in Biomedical NER Across Encoder and Generative Models

Anonymous ACL submission

Abstract

Biomedical Named Entity Recognition (NER) models are routinely evaluated on static benchmarks compiled before 2020, leaving behaviour under post-International Classification of Diseases, 11th Revision (ICD-11) terminology largely unexplored. A corpus of 17,610 PubMed abstracts was stratified into pre-2020 and post-2022 cohorts surrounding the January 2022 ICD-11 activation. Four biomedical language models, comprising two frozen encoders and two generative models under 3-shot prompting, were evaluated using silver labels derived from ICD-11 and Medical Subject Headings (MeSH) ontologies. Novel entity density increased by 58% across the temporal boundary (0.26% to 0.41%). Frozen encoder models exhibited consistent era-level F1 degradation after ICD-11 activation, whereas the generative models showed no comparable trend. Novel-entity Expected Calibration Error (ECE) exceeded non-novel ECE for both encoders despite near-unity confidence. A lightweight ontology-aware adapter trained only on pre-2020 silver labels recovered 20.1 absolute F1 points over the frozen baseline while keeping the base encoder fully frozen.

1 Introduction

Biomedical language models frequently assign high confidence to entity spans that they incorrectly classify or semantically misunderstand. This failure becomes particularly pronounced when biomedical terminology emerges after the model's effective training horizon. The global activation of ICD-11 in January 2022 created a natural ontological boundary for studying this phenomenon. Disease names and synonyms standardized in ICD-11 were largely absent from pre-2020 biomedical NER corpora and benchmarks, yet post-2022 PubMed literature adopts these terms at a measurable rate.

As a result, modern biomedical NER systems can encounter clinically valid terminology that is ontologically standardized but temporally novel relative to their pretraining and supervision data. Existing biomedical benchmarks do not explicitly evaluate this setting. Prior clinical NER work based on the OMOP Common Data Model compares encoder and decoder architectures across entity types, but does not perform temporal stratification or examine ontology-driven distribution shifts across the ICD-11 boundary (Abdul et al., 2024). Broader studies of temporal robustness in biomedical language models analyze classification and extraction tasks without anchoring the analysis to ICD-11 or MeSH ontology evolution, and propose no model-level mitigation strategy (Liu et al., 2024). Work on temporal NER drift has also focused primarily on social media domains rather than biomedical ontologies (Rijhwani and Preotiuc-Pietro, 2020).

To address this gap, this work introduces an ontology-aware temporal evaluation framework for biomedical NER. (i) A novel entity density is operationalized at the ICD-11/MeSH boundary using dual PhraseMatcher silver labeling over 17,610 PubMed abstracts. (ii) Encoder-based and generative biomedical language models are compared under frozen zero-shot evaluation on the same temporally stratified benchmark. (iii) A lightweight 4.6 MB ontology-aware adapter is proposed that improves robustness to temporally novel terminology, recovering 20.1 absolute F1 points over a frozen PubMedBERT baseline without updating base encoder weights.

2 Related Work

Research relevant to ontology-driven temporal drift in biomedical NER spans three largely separate areas: static biomedical benchmarking, temporal robustness analysis, and parameter-efficient adaptation.

2.1 Static Biomedical NER Benchmarks

Biomedical NER benchmarks have historically prioritised domain adaptation and entity coverage rather than temporal robustness. BioBERT(Lee et al., 2020) demonstrated that large-scale pre-training on PubMed and PMC corpora substantially improves downstream biomedical NER performance (Lee et al., 2020), while PubMedBERT further strengthened this paradigm through fully in-domain pre-training from scratch (Gu et al., 2021). However, both the underlying corpora and standard evaluation benchmarks largely predate the ICD-11 transition and therefore contain limited exposure to post-2022 clinical terminology. Abdul *et al.* later introduced a comprehensive clinical NER benchmark across OMOP-modelled entity types (Abdul et al., 2024), but the evaluation remains temporally static and does not examine ontology-driven vocabulary shift across the ICD-11 boundary.

2.2 Temporal Effects in Biomedical NLP

Recent work has begun to examine temporal degradation in biomedical language models (Liu et al., 2024). Liu *et al.* analysed temporal effects across multiple biomedical NLP tasks and reported measurable performance variation over time (Liu et al., 2024). However, their study does not ground novelty in a formal ontology boundary such as ICD-11 or MeSH, does not evaluate generative biomedical models, and proposes no mechanism for mitigating temporal drift. Beyond the biomedical domain, Rijhwani *et al.* investigated temporal NER drift on Twitter and showed that entity distributions evolve substantially across years (Rijhwani and Preotiuc-Pietro, 2020). Their work establishes the broader importance of temporal robustness but does not address ontology-aligned biomedical terminology evolution.

2.3 Adapter-Based Remedies

Parameter-efficient fine-tuning methods have shown that frozen language models can be adapted to new domains and tasks by training only lightweight inserted modules (Li and Liang, 2021). Houlsby *et al.* introduced adapter layers as a scalable alternative to full-model fine-tuning while preserving the underlying pre-trained representations (Houlsby et al., 2019). Building on this principle, the present work evaluates an ontology-aware adapter designed specifically for temporally novel biomedical terminology. The proposed module ingests live ICD-11 term embeddings while keeping the underlying PubMedBERT encoder fully frozen during both training and inference.

3 Data and Methods

This section describes the temporally stratified PubMed corpus, the ontology-driven silver labelling pipeline, the frozen-model evaluation setup, and the proposed ontology-aware adapter architecture used to study temporal drift in biomedical NER.

3.1 Corpus Construction

PubMed abstracts were collected using National Centre for Biotechnology Information (NCBI) E-utilities (ESearch and EFetch). A pre-ICD-11 cohort spans 2015–2019 with approximately 1,900–2,000 abstracts per year, while a post-ICD-11 cohort spans 2022–2025 at a comparable density. The transition years 2020–2021 were excluded to minimise contamination between pre- and post-adoption terminology distributions. The final corpus contains 17,610 abstracts.

No manual annotation was performed; all labels are silver-generated. Consequently, the reported metrics should be interpreted as conservative lower bounds on true biomedical entity performance rather than absolute clinical accuracy estimates.

3.2 Ontologies and Silver Labelling

To operationalise ontology-driven temporal novelty, two complementary biomedical vocabularies were used. MeSH 2024 XML descriptors were parsed to represent pre-transition biomedical terminology (Lipscomb, 2000), while ICD-11 disease chapter titles and synonyms were retrieved through the World Health Organisation (WHO) public API (World Health Organization, 2022).

Each abstract was processed using two independent spaCy PhraseMatcher passes (Honnibal et al., 2020). MeSH matching produced `mesh_spans`, while ICD-11 matching produced `icd11_spans`. Token-level BIO labels were then generated using longest-span-first overlap resolution with ICD-11 priority under ties. A span was classified as novel when matched by ICD-11, and no overlapping MeSH span shared any token. No fuzzy matching or embedding-based retrieval was used. The resulting corpus contains 293 unique novel entity spans.

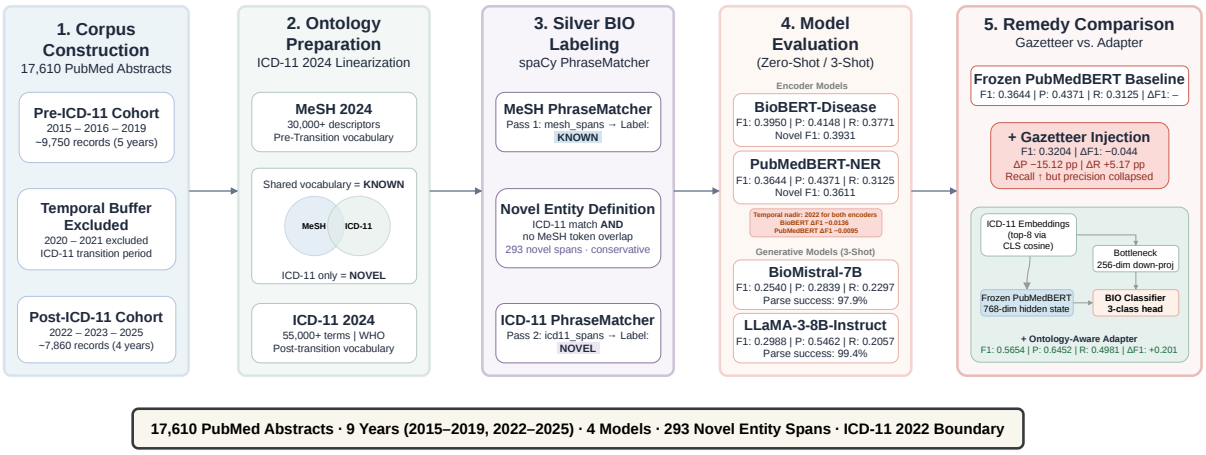

Figure 1: Proposed research workflow illustrating temporally stratified corpus construction, ontology-driven silver labelling, frozen-model evaluation, and ontology-aware adaptation across the ICD-11 boundary.

3.3 Models

Four frozen pre-trained language models were evaluated to determine whether temporal drift is architecture-specific or a broader property of biomedical language models. BioBERT Disease (Lee et al., 2020) and PubMedBERT-NER (Gu et al., 2021) represent encoder-based biomedical NER systems, while BioMistral-7B (Labrak et al., 2024) and LLaMA-3-8B-Instruct (Grattafiori et al., 2024) extend the comparison to generative instruction-tuned models.

BioBERT and PubMedBERT were executed as HuggingFace NER pipelines with batch size 64. BioMistral-7B and LLaMA-3-8B-Instruct were evaluated under 3-shot prompting with JSON list outputs, maximum generation length 128 tokens, batch size 8, and character-to-token BIO alignment.

The selected models also capture different pre-training regimes. BioBERT adapts a general-domain BERT checkpoint to biomedical text, whereas PubMedBERT is trained entirely from scratch on PubMed corpora (Lee et al., 2020; Gu et al., 2021). Similarly, BioMistral-7B represents a biomedical-specialised generative model, while LLaMA-3-8B-Instruct serves as a general-purpose instruction-tuned baseline. Evaluating both encoder and generative families on the same temporally stratified benchmark enables architecture-level temporal sensitivity to be isolated from dataset-level confounds.

3.4 Ontology-Aware Adapter

The ontology-aware adapter uses a fully frozen PubMedBERT base (Gu et al., 2021). For each token, the contextual hidden state (768 dimensions) was concatenated with the mean embedding of the top-8 retrieved ICD-11 concepts obtained through cosine similarity retrieval using the document CLS embedding, yielding a 1,536-dimensional fused representation (Lewis et al., 2021).

Let $h_i \in \mathbb{R}^{768}$ denote the contextual representation for token i, and let $o \in \mathbb{R}^{768}$ denote the retrieved ontology embedding. The fused representation is constructed via vector concatenation as

$$x_i = [h_i; o] \tag{1}$$

where $[\cdot;\cdot]$ denotes the concatenation operator yielding $x_i \in \mathbb{R}^{1536}$.

A bottleneck adapter then down-projects and subsequently up-projects this concatenated vector into the original hidden dimension space:

$$z_i = W_{\text{up}}\,\text{GELU}(W_{\text{down}} x_i) \tag{2}$$

where $W_{\text{down}} \in \mathbb{R}^{256\times1536}$ and $W_{\text{up}} \in \mathbb{R}^{768\times256}$ represent the trainable bottleneck weights.

To modulate the integration of external taxonomic knowledge and reduce the risk of false-positive over-triggering, ontology influence is explicitly regulated through a gated residual mechanism:

$$g_i = \sigma(W_g z_i) \tag{3}$$

where $W_g \in \mathbb{R}^{768\times768}$ is a trainable gating weight matrix, σ is the element-wise sigmoid activation function, and $g_i \in [0,1]^{768}$ acts as a learned, token-specific gating vector. Values near 0 suppress ontology influence when the retrieved concepts mismatch the local sequence context, while values near 1 amplify ontology-conditioned features.

The final adapted representation $\tilde{h}_i \in \mathbb{R}^{768}$ is subsequently computed via a gated residual shortcut:

$$\tilde{h}_i = h_i + g_i \odot z_i \tag{4}$$

where $\odot$ denotes the Hadamard (element-wise) product. This final representation $\tilde{h}_i$ is directly passed to the linear BIO prediction head for sequence labeling.

Training used Adam with Weight Decay (AdamW) (Loshchilov and Hutter, 2019) and a one-cycle learning rate schedule over 5 epochs on pre-2020 silver labels. Only adapter parameters were updated while all base encoder weights remained frozen (Houlsby et al., 2019). The resulting checkpoint size is 4.6 MB. Training was conducted on an NVIDIA A100-SXM4-80GB MIG instance with 39.2 GB allocated VRAM.

3.5 Adapter Design Rationale

Several architectural choices were designed to balance ontology injection against over-triggering risk. CLS-based cosine retrieval was selected because the CLS representation provides a document-level semantic summary suitable for identifying ICD-11 concepts relevant to the current abstract before token-level prediction. The top-8 retrieval cutoff balances ontology coverage against the introduction of semantically irrelevant concepts.

The gated residual connection is the central architectural component. It allows the adapter to dynamically control how much ontology information is injected into each token representation. When retrieved ICD-11 concepts are weakly related to the local token context, the gate suppresses ontology influence, reducing the false-positive behaviour observed in the gazetteer injection baseline.

3.6 Evaluation Protocol

Span-level Precision, Recall, and F1 score were computed using seqeval exact-match evaluation (Nakayama, 2018). A separate Novel-Entity F1 metric was additionally computed over spans marked as novel=True in order to isolate performance on temporally introduced disease terminology. Era-level mean F1 was computed as an unweighted average across yearly scores rather than as a pooled metric over all abstracts. Because yearly abstract counts are approximately balanced across cohorts, weighting would not materially affect aggregate performance. More importantly, the unweighted formulation ensures direct interpretability of temporal comparisons by assigning equal contribution to each year and preventing density differences from dominating cohort-level results.

4 Results

This section analyses ontology-driven vocabulary change across the ICD-11 boundary, evaluates temporal drift in frozen biomedical language models, and examines whether ontology-aware adaptation mitigates the resulting degradation.

4.1 Novel Entity Analysis

A novel entity span is defined as a contiguous token sequence matched by the ICD-11 2024 PhraseMatcher that shares no token overlap with any span matched by the MeSH 2024 PhraseMatcher. No fuzzy or embedding-based matching was used; the definition is intentionally conservative and exact. Across the full corpus, 293 unique novel entity spans were identified.

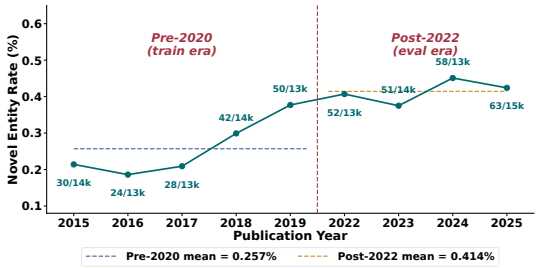

Figure 2: Per-year novel entity rate (%) for pre-2020 and post-2022 cohorts with bootstrap 95% confidence intervals.

The central cross-era finding is clear. Novel entity density increased from 0.26% in the pre-2020 cohort to 0.41% in the post-2022 cohort, representing a 58% relative increase across the ICD-11 boundary (Figure 2). The difference was statistically significant under a two-proportion z-test. Because the PhraseMatcher vocabularies remained fixed, corpus sampling density was matched across years, and the 2020–2021 transition window was excluded, the increase is unlikely to reflect a measurement artefact. Instead, it indicates a measurable shift in published biomedical vocabulary composition following ICD-11 adoption.

A secondary temporal pattern further supports this interpretation. Within the pre-2020 cohort alone, the novel entity rate increased from 0.19% in 2016 to 0.38% in 2019, nearly doubling before official ICD-11 activation. This trajectory is consistent with ICD-11 draft terminology entering biomedi-

cal literature during the review and pre-adoption period. The World Health Organisation (WHO) released ICD-11 for member review in 2018 and formally endorsed it in May 2019 prior to its January 2022 activation (World Health Organization, 2022). Authors therefore appear to have adopted portions of the emerging terminology several years before the formal boundary (Chan et al., 2020). Consequently, the pre-2020 cohort should not be interpreted as entirely ICD-11-free, but rather as containing a gradually increasing pre-adoption signal.

4.2 Model Performance

Overall model performance across all 17,610 abstracts is presented in Table 1. Among encoder models, BioBERT-Disease achieved the highest overall F1 at 0.3950 (P 0.4148, R 0.3771), followed by PubMedBERT-NER at 0.3644 (P 0.4371, R 0.3125). Among generative models, LLaMA-3-8B-Instruct achieved F1 0.2988 (P 0.5462, R 0.2057), while BioMistral-7B achieved F1 0.2540 (P 0.2839, R 0.2297).

Table 1: Main NER Results (all 17,610 records)

Model	Type	F1	P	R	Nov. F1
BioBERT-Dis.	Enc.	0.3950	0.4148	0.3771	0.3931
PubMedBERT-NER	Enc.	0.3644	0.4371	0.3125	0.3611
BioMistral-7B	Gen.	0.2540	0.2839	0.2297	0.2689
LLaMA-3-8B-Ins.	Gen.	0.2988	0.5462	0.2057	0.3078

The encoder family outperformed the generative family by approximately 10 F1 points at the top of each family. This gap likely reflects the structural mismatch between exact-match BIO evaluation and free-form generative span extraction under prompting rather than a fundamental difference in biomedical language understanding capability.

Novel-entity F1 follows the same rank ordering as overall F1: BioBERT 0.3931, PubMedBERT 0.3611, LLaMA-3 0.3078, and BioMistral 0.2689. The relatively small difference between overall and novel-entity F1 within each model suggests that novel ICD-11 spans are not categorically harder in isolation. Instead, the degradation observed across eras appears to emerge from broader distributional shifts affecting prediction precision.

LLaMA-3-8B-Instruct exhibits a strong precision-recall asymmetry (P 0.5462, R 0.2057). Under 3-shot JSON-list prompting with a 128-token generation limit, the model predicts conservatively and emits only spans associated with high surface-form confidence. Parse coverage reached 99.4%, indicating that the low recall is semantic rather than a consequence of formatting failure.

4.3 Temporal Drift

The era-level comparison reveals a consistent degradation pattern across both frozen encoder models (Table 2; Figure 3). BioBERT F1 decreased from 0.4010 in the pre-2020 cohort to 0.3874 post-2022 ($\Delta = -0.0136$), corresponding to a 3.4% relative decline from its own baseline. PubMedBERT exhibited a similar reduction, falling from 0.3687 to 0.3592 ($\Delta = -0.0095$). Both models reached their lowest yearly F1 in 2022, the first year of official ICD-11 activation, before partially recovering during 2023–2025 as the post-transition vocabulary stabilised in the literature.

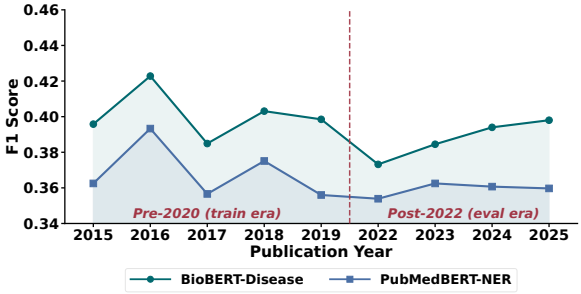

Figure 3: Per-year F1 for BioBERT-Disease and PubMedBERT-NER across 2015–2019 and 2022–2025. Both models reached their lowest F1 in 2022.

Table 2: Era-Level Encoder F1 Means

Model	Pre-2020 F1	Post-2022 F1	Δ F1
BioBERT	0.4010	0.3874	−0.0136
PubMedBERT	0.3687	0.3592	−0.0095

The degradation pattern is driven primarily by precision rather than recall. BioBERT showed $\Delta P = -0.0208$ against $\Delta R = -0.0077$, while PubMedBERT showed $\Delta P = -0.0204$ against $\Delta R = -0.0035$. In both cases, precision deteriorated substantially faster than recall. A model that simply failed to recognise novel terminology would be expected to exhibit primarily recall degradation. Instead, the dominant effect is increased false-positive prediction on post-2022 abstracts. Temporally unfamiliar spans appear to activate entity-class features within the frozen encoder, producing confident but incorrect predictions rather than abstention. The generative models exhibit a qualitatively different temporal behaviour. BioMistral-

7B changed only marginally between eras, from 0.2525 to 0.2555 ($\Delta = +0.003$), while LLaMA-3-8B-Instruct showed no measurable degradation. The absence of drift under prompting, paired with its consistent presence in frozen encoders, suggests an architecture-level explanation rather than a purely dataset-level one.

Frozen encoder representations commit to a fixed partition of the token embedding space during training. Token sequences that fall near, but outside, learned entity boundaries therefore produce systematic prediction errors that cannot be corrected at inference time. Instruction-tuned generative models operate differently: span extraction under prompting depends more heavily on contextual reasoning at generation time rather than on static hidden-state decision boundaries, making the output less sensitive to post-activation vocabulary shifts. Year-level t-tests were underpowered given the limited number of years per cohort and should therefore not be interpreted as statistically confirmatory. The era-level deltas are better treated as consistent effect-size estimates.

4.4 Error Analysis

Error decomposition reveals distinct behaviour between novel and non-novel spans. Non-novel spans accumulate 63.76% Type 1 errors and 57.42% Type 3 errors. Novel spans, by contrast, exhibit a lower Type 1 rate of 43.21% but a substantially higher Type 2 boundary error rate of 39.78% compared to 25.03% for non-novel spans.

This pattern suggests that encoder models partially detect novel ICD-11 terminology but fail to localise span boundaries accurately. The dominant failure mode is therefore imprecise delimitation rather than complete entity blindness.

The elevated boundary-error rate on novel spans is consistent with the behaviour of frozen encoder representations on out-of-distribution token sequences. For in-distribution biomedical entities, repeated exposure during pre-training and fine-tuning allows the model to learn stable boundary-sensitive hidden-state patterns, particularly around B-tag transitions. Novel ICD-11 terminology is largely absent from those corpora. As a result, the encoder detects semantic irregularity at the token position but fails to activate sufficiently discriminative boundary features, leading to partial detection without clean span resolution.

Table 3 presents representative post-2022 failure cases illustrating the dominant error patterns observed in frozen PubMedBERT predictions, including boundary over-extension, missed spans, acronym-triggered false positives, and morphological confusion. In each case, the ontology-aware adapter produces a corrected span prediction.

Table 3: Representative post-2022 encoder failure cases illustrating boundary over-extension, missed spans, acronym-triggered false positives, and morphological confusion.

Text Snippet	Target Span	PubMedBERT Prediction	Error Type	Adapter Prediction
...especially **Bipolar II disorder**, is a mental disorder...	*Bipolar disorder*	*Bipolar II disorder*	Boundary over-extension	*Bipolar disorder*
...not considering relevant factors such as **substance use**...	*substance use*	*(not tagged)*	False negative (missed span)	*substance use*
...first episode of **psychosis** (FEP) have been widely...	*psychosis*	*psychosis + FEP*	Spurious prediction (acronym)	*psychosis*
...caregivers of first episode **psychosis**...**psychotic** features...	*psychosis*	*psychotic*	Morphological false positive	*psychosis*

4.5 Calibration

Both encoder models remain highly overconfident on temporally novel spans despite measurable degradation in prediction accuracy (Table 4). BioBERT exhibits overall Expected Calibration Error (ECE) of 0.0413, increasing to 0.0653 on novel spans. PubMedBERT shows overall ECE of 0.0377 and novel-span ECE of 0.0611. Mean confidence remains extremely high in both models (0.9975 and 0.9874 respectively).

Table 4: Expected Calibration Error (10 bins, token-level)

Model	Overall	Novel	Non-nov.	Mean Conf.
BioBERT	0.0413	0.0653	0.0409	0.9975
PubMedBERT	0.0377	0.0611	0.0373	0.9874

This combination of high confidence and elevated novel-span ECE constitutes a silent failure mode. The pattern is consistent with prior observations on discriminative classifiers under out-of-distribution inputs (Hendrycks and Gimpel, 2018). Softmax prediction heads partition the learned representation space into fixed class regions without explicitly modelling epistemic uncertainty (Guo et al., 2017). Consequently, temporally novel entity spans are mapped to the nearest learned entity region with high confidence even when the underlying representation differs substantially from the training distribution.

ECE analysis exposes this mismatch directly because it measures the alignment between predicted confidence and empirical correctness. Standard F1 evaluation captures the existence of prediction

errors but does not quantify the confidence with which those errors are made.

4.6 Remedy Comparison

Remedy comparison results are shown in Table 5 and Figure 4. Gazetteer injection reduced BioBERT F1 from 0.3950 to 0.3467 ($\Delta = -0.048$) and reduced PubMedBERT F1 from 0.3644 to 0.3204 ($\Delta = -0.044$). Precision collapsed sharply despite modest recall gains, indicating that surface-form over-matching dominates the retrieval process. Novel-specific gains were negligible (BioBERT ΔF1 -0.0038; PubMedBERT ΔF1 -0.0004).

Table 5: Remedy Comparison

System	F1	P	R	Δ F1
Frozen PubMedBERT	0.3644	0.4371	0.3125	–
+ Gazetteer Injection	0.3204	0.2859	0.3642	−0.044
+ Ontology-Aware Adapter	0.5654	0.6452	0.4981	+0.201

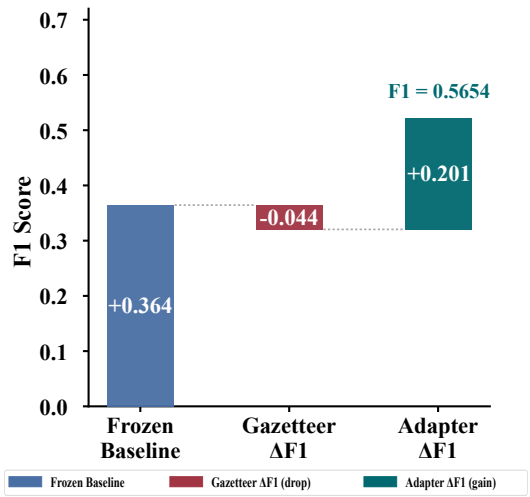

Figure 4: F1 progression from frozen PubMedBERT (0.3644), through gazetteer injection ($\Delta = -0.044$), to the ontology-aware adapter ($\Delta = +0.201$, reaching 0.5654).

The ontology-aware adapter produces a markedly different outcome. PubMedBERT F1 increased from 0.3644 to 0.5654, while precision increased from 0.4371 to 0.6452 and recall increased from 0.3125 to 0.4981. This corresponds to a gain of 0.201 absolute F1 points (20.1 percentage points; 55.2% relative improvement) while keeping the underlying PubMedBERT encoder fully frozen. The resulting checkpoint size is only 4.6 MB.

The contrast between the two remedies reflects their treatment of contextual information. Gazetteer injection functions as a context-independent lookup mechanism: any token sequence matching an ICD-11 surface form receives an entity prediction regardless of sentence context. Because biomedical surface forms can appear in negated, descriptive, or modifier contexts, the method generates substantial false positives and collapses precision.

The ontology-aware adapter instead learns when ontology information should influence prediction. The gated residual suppresses ontology contributions when retrieved ICD-11 concepts are semantically distant from the local token context and amplifies them when alignment is strong. Consequently, ontology information shifts the encoder representation only when sentence-level semantics and ontology-level signal agree, allowing both precision and recall to improve simultaneously.

5 Discussion

The results collectively reveal a coherent failure pattern in frozen biomedical encoder models exposed to ontology-driven vocabulary shift. The observed era-level degradation is primarily precision-driven rather than recall-driven. Instead of simply failing to recognise post-2022 biomedical entities, the encoders produce increasing numbers of false-positive predictions on temporally unfamiliar spans. The error analysis and calibration findings help explain this behaviour mechanistically. Novel ICD-11 spans are often partially detected but poorly localised, producing elevated boundary-error rates rather than complete entity omission. At the same time, both BioBERT and PubMedBERT remain highly confident on these incorrect predictions despite measurable degradation in empirical accuracy. Together, these findings suggest that frozen encoder representations map temporally novel spans toward the nearest learned entity-class region, generating confident but spatially imprecise predictions. Under exact-match BIO evaluation, these localisation failures manifest primarily as precision degradation.

The ontology-aware adapter mitigates this behaviour by conditioning token-level predictions on ICD-11 semantic context before classification. Unlike context-independent gazetteer injection, the gated residual selectively amplifies ontology information only when semantic alignment with the local sequence context is strong, improving both precision and recall simultaneously.

Several limitations of the silver-labelling frame-

work require explicit acknowledgement. PhraseMatcher supervision operates through exact surface-form matching; any true biomedical entity absent from the MeSH or ICD-11 vocabularies remains unlabeled. Consequently, the silver annotations represent conservative lower bounds on true entity density, and absolute F1 values should be interpreted accordingly. The primary interpretive value therefore lies in relative comparisons across temporal cohorts and remediation strategies rather than in the absolute metric magnitudes themselves. Statistical interpretation also requires caution. Year-level t-tests were underpowered given the limited number of yearly observations (five pre-2020 years and four post-2022 years). BioBERT yielded $p = 0.156$, while PubMedBERT yielded $p = 0.282$. These results should therefore not be interpreted as statistically confirmatory evidence of causal temporal degradation. Instead, the era-level deltas are more appropriately treated as consistent effect-size estimates.

The present analysis is additionally restricted to the biomedical publication domain. PubMed abstracts differ substantially from clinical EHR notes in sentence structure, abbreviation density, and documentation style. The conclusions therefore apply specifically to biomedical literature and may not generalise directly to clinical documentation settings without separate evaluation. The recall limitations observed in generative models appear to be semantic rather than formatting-related. BioMistral achieved parse coverage of 97.9%, while LLaMA-3 achieved 99.4%. Output parsing failures were therefore negligible. The low recall instead reflects a conservative span-prediction tendency under prompted generation.

6 Limitations

Although the observed temporal trends and remediation effects are consistent across multiple analyses, several limitations constrain the interpretation and generalisability of the findings.

1. Silver labels provide conservative lower bounds on entity coverage because PhraseMatcher supervision is limited to exact surface-form matches. Biomedical entities absent from the ICD-11 or MeSH vocabularies remain unlabeled, causing absolute F1 values to underestimate true model capability. Quantitative comparisons should therefore be interpreted primarily as relative measures.

2. Year-level statistical testing is underpowered due to the limited number of yearly observations (n=5 pre-2020 years; n=4 post-2022 years). BioBERT produced t=1.5894, p=0.156, while PubMedBERT produced t=1.1654, p=0.282. Temporal degradation should therefore be interpreted as a consistent empirical trend rather than as statistically definitive evidence.

3. The corpus consists entirely of PubMed biomedical abstracts. The findings may not directly generalise to clinical EHR notes, discharge summaries, or other forms of medical documentation that exhibit substantially different linguistic structure.

4. Novel-entity precision and recall values for BioMistral-7B were not recoverable from the stored raw predictions; consequently, only novel-entity F1 is reported for that model.

7 Conclusion

This study demonstrates that ontology-driven temporal drift at the ICD-11/MeSH boundary is measurable in biomedical literature. Across 17,610 PubMed abstracts, novel entity density increased by 58% between the pre-2020 and post-2022 cohorts following ICD-11 activation. Frozen biomedical encoder models exhibited consistent era-level degradation that was primarily precision-driven and accompanied by elevated calibration error on temporally novel spans. The evidence suggests that frozen encoder representations remain vulnerable to ontology-level vocabulary evolution even when overall benchmark degradation appears modest.

To address this limitation, an ontology-aware adapter was introduced that conditions token-level predictions on ICD-11 semantic context while keeping the underlying PubMedBERT encoder fully frozen. The adapter improved F1 from 0.3644 to 0.5654, corresponding to an absolute improvement of 0.201 F1 (20.1 percentage points; 55.2% relative improvement) with only a 4.6 MB checkpoint. More broadly, the findings suggest that ontology activation boundaries provide natural temporal evaluation checkpoints for biomedical language models deployed in domains where controlled vocabularies continue to evolve. Lightweight ontology-aware adaptation may therefore represent a practical path toward temporally resilient biomedical NER systems.

References

Wadood M Abdul, Marco AF Pimentel, Muhammad Umar Salman, Tathagata Raha, Clément Christophe, Praveen K Kanithi, Nasir Hayat, Ronnie Rajan, and Shadab Khan. 2024. Named clinical entity recognition benchmark. *Preprint*, arXiv:2410.05046.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. 3(1).

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Dan Hendrycks and Kevin Gimpel. 2018. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Preprint*, arXiv:1610.02136.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. *Preprint*, arXiv:1902.00751.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *Preprint*, arXiv:2101.00190.

Carolyn Lipscomb. 2000. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88:265–6.

Weisi Liu, Zhe He, and Xiaolei Huang. 2024. Time matters: Examine temporal effects on biomedical language models. *Preprint*, arXiv:2407.17638.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

H. Nakayama. 2018. seqeval: A python framework for sequence label evaluation. `https://github.com/chakki-works/seqeval`.

Shruti Rijhwani and Daniel Preotiuc-Pietro. 2020. Temporally-informed analysis of named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7605–7617, Online. Association for Computational Linguistics.

World Health Organization. 2022. *ICD-11: International Classification of Diseases, 11th Revision*. WHO Press, Geneva.