Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A lightweight boundary-refinement module improves entity mention accuracy in biomedical named entity recognition without degrading correct predictions

Julius Beneoluchi Odili
Covenant University
Wasiu Oluwagbenga Hassan
Hassanwasiu00@gmail.com

Covenant University

Research Article

Keywords: Biomedical named entity recognition, Boundary detection, Span refinement, Parameter-efficient methods, Information extraction, Transformer models, BioBERT, PubMedBERT

Posted Date: September 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10594247/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Biomedical named entity recognition (BioNER) sits upstream of most biomedical information-extraction pipelines, so its errors propagate unchanged into relation extraction and knowledge-base construction. Contemporary domain-pretrained transformers report strong entity-level F1, yet a substantial share of their residual error consists of boundary errors — predictions that identify the correct entity type but draw the span edges incorrectly rather than type errors. A single displaced token can convert a pathology into an anatomical site or a specific chemical into a chemical family. Existing boundary-aware architectures address this problem through offset regression, iterative denoising, or biaffine span enumeration, but each imposes substantial computational cost and requires replacing the underlying recognition architecture, limiting adoption in existing pipelines and resource-constrained settings.

Results

We present a Boundary-Refinement Module (BRM): a post-hoc, encoder-agnostic span verifier that attaches to a frozen BioNER backbone and repairs near-miss boundary errors under an empirically calibrated No-Harm gate. Three candidate refinement architectures were trained under an identical budget on 610,547 contrastive span instances and evaluated across twelve backbone × corpus cells (BioBERT, SciBERT, PubMedBERT × GENIA, BC5CDR, NCBI-Disease, JNLPBA). The selected architecture, a pooled sequence verifier, recovers a mean 35.37% of near-miss boundary errors under strict matching (IoU ≥ 0.5), peaking at 80.33%, while degrading only 0.12% of predictions that were already exact, 62 of 55,472 pooled baseline-correct spans. Strict entity-level F1 improves by a mean of 2.09 points across all twelve cells, rising to 3.44–3.61 points on GENIA. The module adds 1.13M parameters, 1.03% of the frozen backbone, and 222 ms mean inference latency on a single commodity GPU. Recovery is strongly anti-correlated with baseline strength, providing the largest gains where baselines are weakest. Downstream, refined spans improved a dependency-path relation-extraction proxy: syntactic stability rose 8.28 points, while verified relation yield rose a smaller 0.73 points.

Conclusions

Boundary error in BioNER is substantially recoverable through a lightweight, modular post-processing stage that requires no backbone retraining and provides an empirically quantified safety guarantee. The asymmetry between structural and relational downstream gains indicates that boundary correctness is necessary but not sufficient for downstream extraction accuracy.

Background

Biomedical named entity recognition converts unstructured biomedical text into structured records and, in most extraction pipelines, precedes every other component [1]. Due to that the volume of biomedical literature has grown past the point at which manual curation is feasible, MEDLINE now indexes over 1.7 million new citations annually [2], whatever errors BioNER makes are inherited unchanged by every downstream step, from relation extraction to knowledge-base linking.

BioNER is harder than general-domain NER in several respects. Entities overlap and nest: “human T-cell leukemia virus type 1” contains a cell type, a disease and a virus within one span, and a flat sequence tagger must select a single label. Biomedical vocabulary is heavily abbreviated and polysemic: “ER” may denote estrogen receptor, endoplasmic reticulum, or emergency room, and the three readings support entirely different downstream conclusions. Most relevant to the present work, biomedical entities are frequently multi-word noun phrases whose meaning changes if the boundary moves by a single token. Extracting “acid” where the gold span is “retinoic acid”, or “renal pedicle” where the gold span is “renal pedicle occlusion”, converts a chemical into a chemical family and a pathology into an anatomical location respectively. These are not stylistic variations; they are incorrect facts propagating into downstream artefacts [3].

Domain-pretrained discriminative transformers such as BioBERT [4], SciBERT [5], PubMedBERT [6] dominate current type-level BioNER benchmarks. However, recent audits report that the residual errors of these models are predominantly boundary errors rather than type errors [7, 8]. This residual is the object of the present study.

Three principal families of boundary-aware method have been proposed. Offset regression, exemplified by the Boundary Offset Prediction Network [9], reframes the task as predicting the distance from a candidate span to the true border rather than classifying tokens, treating near-misses as trainable signal rather than as negatives. Generative denoising, exemplified by DiffusionNER [10], formulates recognition as an iterative boundary-denoising process that refines noisy span coordinates toward gold spans over multiple steps. Boundary differentiation, exemplified by DiFiNet [11], applies biaffine attention over span grids to separate inner from outer entities in nested mentions. Related two-stage approaches locate candidate spans before labelling them [12].

Each of these approaches achieves accuracy gains, but each also entails substantial computational cost: iterative denoising multiplies inference cost by the number of refinement steps, and biaffine span enumeration scales quadratically with sequence length, which is prohibitive for long biomedical abstracts. More fundamentally, each requires replacing the recognition architecture wholesale. An institution already operating a fine-tuned BioBERT pipeline cannot adopt these methods incrementally. Separately, large generative extractors have been evaluated for BioNER and, while competitive in zero-shot settings, exhibit their own boundary drift and hallucination behaviours under strict-matching evaluation [7, 13].

This leaves a gap: no published method provides a modular, encoder-agnostic boundary-refinement stage that attaches to an existing BioNER pipeline at low parameter and latency cost. We address that

gap directly. Our contributions are: (i) a quantified diagnosis of the boundary-error share of residual BioNER error across twelve backbone × corpus configurations, using a seven-class error taxonomy; (ii) a lightweight span-verification module trained by contrastive perturbation, evaluated against two alternative refinement architectures under an identical training budget; (iii) an empirically calibrated No-Harm gate with a measured harm rate, addressing the practical concern that a post-hoc corrector may damage already-correct predictions; and (iv) a downstream evaluation testing whether boundary correction translates into relation-extraction improvement, rather than assuming that it does.

Methods

Datasets

Four standard biomedical NER corpora were used, spanning three entity domains and four label schemas (Table 1). GENIA [14] and JNLPBA [15] cover molecular biology entities; BC5CDR [16] covers chemicals and diseases; NCBI-Disease [17] covers disease mentions only. GENIA is evaluated in its flat projection, in which nested annotations are reduced to a single label per token following standard practice; this constraint is important when interpreting the comparison in Table 6 and is revisited in the Discussion.

Table 1
Characteristics of the four benchmark corpora

Corpus	Domain	Entity types	Train sentences	Test sentences	Train entities	Test entities
NCBI-Disease	Disease	1	5,424	940	5,134	960
BC5CDR	Chemical/Disease	2	5,228	5,865	7,773	7,809
GENIA	Molecular biology	5	15,023	1,854	41,770	4,875
JNLPBA	Molecular biology	5	16,619	3,856	45,748	8,662

GENIA statistics are reported for the flat-NER projection used in this study.

A unified seven-class entity ontology was applied across corpora to permit consistent cross-corpus evaluation, mapping each corpus's native types onto a shared schema.

Table 6
Relation to published results

Method	NCBI-Disease	BC5CDR-chem	BC5CDR-disease	JNLPBA	GENIA
Published, flat NER					
BioBERT [4]	89.1	92.9	84.7	78.6	—
SciBERT [5]	88.25	92.51	84.70	—	—
PubMedBERT [6]	87.8	93.3	85.6	79.1	—
BioELECTRA [19]	89.4	93.6	85.8	80.2	—
LinkBERT [20]	88.2	93.8	86.1	79.0	—
BINDER [21]	90.9	95.0	88.0	80.3	—
Published, nested NER					
BOPN [9]	—	—	—	—	82.14
DiFiNet [11]	—	—	—	—	81.87
DiffusionNER [10]	—	—	—	—	81.53
Locate and Label [12]	—	—	—	—	80.54
This work (upper-bound estimates)					
PubMedBERT baseline	92.48	92.19†		83.40	83.47
PubMedBERT + BRM	94.39	95.78†		84.89	86.91

Values are percentages. †BC5CDR is evaluated here as a single two-type corpus rather than split into separate chemical and disease sub-tasks, so these values are not directly comparable to the split published figures. Values in the final block derive from a gold-anchored ledger and are upper bounds (see Methods); they exceed published baselines for equivalent configurations by approximately three to four points, consistent with the sensitivity analysis in Table 5. This table establishes operating regime and is not a like-for-like benchmark.

Baseline models

Three domain-pretrained encoders were fine-tuned for token classification on each corpus, yielding twelve backbone × corpus evaluation cells: BioBERT v1.1 [4], SciBERT (scivocab, uncased) [5], and PubMedBERT (abstract, uncased) [6]. All baselines used maximum sequence length 128, batch size 32, AdamW at learning rate 2×10^{-5} with linear schedule and 10% warmup, and three training epochs. Every

configuration was trained under three random seeds (17, 42, 2024); all reported figures are means across seeds.

Error taxonomy

Every baseline prediction was assigned to one of seven categories by comparing predicted against gold spans: Exact Match, Under-Extraction (predicted span a strict subset of gold), Over-Extraction (strict superset), Partial Overlap (overlapping, neither containing the other), Type Error (boundaries matching but type incorrect), Missing Entity (gold span with no overlapping prediction), and Spurious Entity (prediction with no overlapping gold span).

The three intermediate categories — Under-Extraction, Over-Extraction and Partial Overlap — constitute the *near-miss* population targeted by the refinement module. This distinction matters: a near-miss is recoverable by boundary adjustment, whereas a Missing Entity represents a detection failure that no boundary-refinement stage can address.

The Boundary-Refinement Module

Given a candidate span produced by a frozen backbone, the BRM evaluates whether that span is exactly correct and, if not, whether an adjusted span is preferable. Boundary marker tokens are injected into the token sequence at the candidate span's edges, the marker-augmented sequence is re-encoded by the frozen backbone, and the hidden states at the marker positions are pooled and passed to a small trainable classification head producing a confidence score $p_{valid} \in [0,1]$ (Fig. 2). Candidate spans are generated by systematically perturbing the baseline prediction within a bounded offset window.

The module adds 1.13M trainable parameters, 1.03% of the 110M-parameter backbone. For backbone dimension $d = 768$ the classification head itself contributes approximately 769 parameters; the remainder consists of position-embedding and layer-normalisation weights consumed when re-encoding the marker-augmented sequence. The backbone is frozen throughout, no gradient updates reach it, which is what makes the module encoder-agnostic and avoids the cost of full retraining.

Contrastive training

The refinement module was trained on 610,547 span instances constructed by contrastive perturbation: gold spans serve as positives, and systematic ± 1 and ± 2 token displacements around gold entities serve as hard negatives. This perturbation window was selected because the diagnostic analysis showed that the overwhelming majority of observed near-miss errors fall within this range. Training used AdamW at learning rate 2×10^{-5} for three epochs, identical across all candidate architectures.

Candidate architectures

Three refinement architectures were implemented and trained under identical budgets:

- **Variant A (boundary regression)**: independent sigmoid heads predict, per token, whether that token is the true start or end of the refined span.
- **Variant B (pooled sequence verifier)**: a single pooled classifier scores an entire candidate span at once, as described above.
- **Variant C (three-way action head)**: a classifier selects one of KEEP, EXPAND, or CONTRACT per candidate span, with an auxiliary magnitude head selecting displacement size.

The No-Harm gate

A post-hoc corrector that improves near-misses may also damage predictions that were already correct. To bound this risk, refinement is applied only when the verifier's confidence clears a threshold τ. The threshold was calibrated by grid search over the validation split, sweeping τ and recording the resulting trade-off between recovered near-misses and harmed exact matches. $\tau = 0.75$ was selected as the lowest threshold at which the harm rate became negligible while retaining most achievable recovery. Below τ, the baseline prediction is retained unchanged; the module abstains rather than guesses.

Evaluation

The primary metric is **Near-Miss Repair Velocity (NMRV)**: of all baseline predictions falling in the three near-miss categories, the percentage converted to Exact Match after refinement, evaluated under strict matching at IoU ≥ 0.5. The complementary safety metric is the **harm rate**: of all baseline predictions that were already Exact Match, the percentage changed to any other category.

Strict entity-level precision, recall and F1 were derived from the same per-prediction outcome ledger used for NMRV, ensuring that all reported metrics share a single matching criterion. Because the ledger is gold-anchored — each record corresponds to a gold entity and its outcome, predictions that overlap no gold entity are not represented. Reported F1 values are therefore upper bounds on strict entity-level F1, and are not directly comparable to externally published figures computed over full prediction sets. A sensitivity analysis assuming spurious-prediction rates of 5–20% is reported in the Results. The *difference* between baseline and refined F1 is substantially more robust to this limitation than the absolute values, because refinement adjusts the boundaries of existing predictions rather than creating or deleting predictions.

Statistical significance was assessed by 10,000-draw paired bootstrap resampling with empirical 95% confidence intervals, and by paired Wilcoxon signed-rank tests across seeds. Latency was measured on a single NVIDIA Tesla T4 (16 GB) under FP16 mixed precision.

Downstream evaluation

To test whether boundary correction propagates, baseline and refined spans were fed to a dependency-path relation-extraction proxy constrained to shortest dependency paths of length ≤ 5. Two metrics were recorded: **syntactic stability**, the proportion of candidate entity-pair sentences whose dependency path is well-formed once entity spans are inserted, and **hard relation yield**, the proportion for which a relation is extracted and passes a stricter secondary verification.

Use of large language models

A large language model (Claude, Anthropic) was used during preparation of this manuscript for language editing, structural organisation of prose, and analysis scripts. All experimental design, model training, numerical results, and scientific claims are the author's own. All model-assisted text and code were reviewed and verified by the author; all reported numbers were independently recomputed from the experimental output ledgers.

Results

Boundary errors constitute a substantial, recoverable share of residual error

Across all twelve backbone × corpus cells, a consistent gap emerged between relaxed and strict matching in every cell, attributable to boundary near-misses rather than type confusions.

Figure 1 shows the error-category distribution for two representative corpora. On GENIA, categorical Type Errors constitute 0.80% of the failure profile, whereas localised boundary segmentation errors such as Over-Extraction at 3.75% and Partial Overlap at 4.12% together account for 7.87% of predictions. On BC5CDR the same pattern holds at smaller magnitude, with Type Errors at 0.35% against combined boundary error of 2.53%. In both corpora, boundary-edge failures exceed categorical type failures by roughly an order of magnitude, establishing boundary correction rather than type disambiguation as the higher-value intervention.

Offset analysis confirmed that over 90% of baseline boundary errors fall within a ± 1 to ± 2 token window of the gold span, motivating the perturbation range used for contrastive training.

The pooled sequence verifier outperforms both alternatives

Table 3 and Fig. 3 report the three-variant ablation across all twelve cells.

Variant A collapsed completely, recovering zero near-misses at every cell. Training loss stagnated from the first epoch. Diagnostic inspection attributed this to activation collapse: only around 1.6% of tokens in a 128-token training sequence carry a positive boundary label, and under standard binary cross-entropy this class imbalance drove the head toward uniform near-zero output. This is an architectural mismatch rather than an undertraining artefact; additional epochs did not alter the outcome.

Table 3
Three-variant ablation summary

Architecture	Trainable parameters	Mean NMRV	Peak NMRV	Floor NMRV	Mean harm rate	Outcome
Variant A — boundary regression	1.13M	0.00	0.00	0.00	n/a	Rejected: activation collapse
Variant B — pooled sequence verifier	1.13M	35.37	80.33	15.06	0.12	Selected
Variant C — three-way action head	1.13M	16.01	65.90	0.00	0.00	Rejected: insufficient recovery

Values are percentages except parameter counts. Peak NMRV for both Variants B and C occurred on PubMedBERT × BC5CDR.

Variant C achieved a mean NMRV of 16.01%, peaking at 65.90% on PubMedBERT × BC5CDR but falling to exactly 0.00% on both NCBI-Disease cells evaluated with PubMedBERT and SciBERT. Notably, Variant C never altered a single baseline-correct prediction across any cell, giving a harm rate of exactly 0.00%. Its failure mode is therefore conservative under-repair rather than destructive over-correction.

Variant B achieved a mean NMRV of 35.37% more than double Variant C with a peak of 80.33% and, critically, a non-trivial floor of 15.06%. It was selected as the production architecture on the combination of highest recovery and a non-zero but very small harm rate.

Figure 4 presents pooled baseline-to-refined confusion matrices for all three variants, showing the category transitions underlying these aggregate figures.

Recovery is greatest where baselines are weakest

Table 2 reports per-cell NMRV. Recovery ranges from 15.06% (SciBERT × JNLPBA) to 80.33% (PubMedBERT × BC5CDR). This variation is systematic rather than random: the peak cell also has the weakest baseline entity-level F1 among the twelve, and recovery correlates negatively with baseline strength. The module therefore contributes most where the underlying recogniser is most fragile, a desirable property for a corrective stage.

Table 2
Near-Miss Repair Velocity by backbone and corpus (Variant B, IoU ≥ 0.5)

Backbone	BC5CDR	GENIA	JNLPBA	NCBI-Disease
BioBERT	43.08	44.70	15.23	35.29
SciBERT	21.43	45.03	15.06	32.56
PubMedBERT	80.33	44.38	15.61	31.71

Values are percentages. Mean across all twelve cells: 35.37%.

The floor is corpus-determined rather than backbone-determined. All three backbones recover approximately 15% on JNLPBA, whose coordination-heavy constructions (“X and Y proteins”) present boundary ambiguities that the perturbation-based candidate generator does not resolve.

Strict entity-level F1 improves consistently across all cells

Table 4 reports strict entity-level F1 before and after refinement for all twelve cells, derived from the same outcome ledger as the NMRV figures.

Table 4
Strict entity-level F1 before and after boundary refinement, all twelve cells

Corpus	Backbone	Baseline F1	Refined F1	Change
NCBI-Disease	BioBERT	91.11	93.31	+ 2.20
NCBI-Disease	SciBERT	92.09	93.75	+ 1.66
NCBI-Disease	PubMedBERT	92.48	94.39	+ 1.91
BC5CDR	BioBERT	94.19	94.65	+ 0.46
BC5CDR	SciBERT	94.75	94.91	+ 0.16
BC5CDR	PubMedBERT	92.19	95.78	+ 3.59
GENIA	BioBERT	82.94	86.54	+ 3.60
GENIA	SciBERT	82.37	85.98	+ 3.61
GENIA	PubMedBERT	83.47	86.91	+ 3.44
JNLPBA	BioBERT	82.85	84.40	+ 1.55
JNLPBA	SciBERT	83.07	84.48	+ 1.41
JNLPBA	PubMedBERT	83.40	84.89	+ 1.49
Mean		**87.91**	**90.00**	**+ 2.09**

Values are percentages, derived from the gold-anchored outcome ledger under strict matching at IoU ≥ 0.5. Absolute values are upper bounds; see Methods and Table 5.

Refinement improved F1 in every cell, by a mean of 2.09 points. The largest gains occurred on GENIA (+ 3.44 to + 3.61 across the three backbones) and on PubMedBERT × BC5CDR (+ 3.59), which is also the highest-NMRV cell. The smallest gains occurred on the BC5CDR cells for BioBERT (+ 0.46) and SciBERT (+ 0.16), where the near-miss population is small in absolute terms (65 and 42 spans respectively) and there is correspondingly little to recover.

As noted in Methods, these F1 values are upper bounds because the gold-anchored ledger does not capture spurious predictions. Table 5 reports a sensitivity analysis over assumed spurious-prediction rates from 0% to 20%. Absolute F1 falls from 87.91 to 79.57 at the baseline and from 90.00 to 81.45 after refinement across this range, but the mean improvement remains stable at + 2.09 to + 1.88 points. The measured gain is therefore robust to this source of uncertainty, whereas the absolute level is not.

Table 5
Sensitivity of entity-level F1 to unrecorded spurious predictions

Assumed spurious prediction rate	Mean baseline F1	Mean refined F1	Mean change
0%	87.91	90.00	+ 2.09
5%	85.67	87.70	+ 2.03
10%	83.53	85.51	+ 1.98
15%	81.51	83.43	+ 1.93
20%	79.57	81.45	+ 1.88

Values are percentages, averaged across all twelve backbone × corpus cells. Absolute F1 is sensitive to the assumed spurious-prediction rate; the improvement attributable to refinement is not.

The No-Harm gate holds empirically

Of 55,472 pooled baseline predictions that were already Exact Match, 55,410 remained Exact Match after refinement and 62 were degraded, a pooled harm rate of 0.11% and a macro-mean of 0.12% across cells (Fig. 5). Eight of the twelve cells recorded zero measured harm. The three cells with harm above 0.1% were all GENIA configurations, consistent with GENIA's higher structural complexity under flat projection.

Bootstrap resampling confirmed the harm rate as small but consistently non-zero; we report it as measured rather than rounding to zero, since a strict zero-harm claim would exceed what the data support.

Computational cost

The module adds 1.13M parameters (1.03% of the frozen backbone) and 222 ms mean inference latency per candidate span on a single Tesla T4 under FP16. No backbone retraining is required, and the module

attaches to any Hugging Face-compatible token-classification checkpoint.

Downstream impact is real but asymmetric

Table 7 reports the dependency-path proxy results. Syntactic stability rose from 72.84% to 81.12%, a gain of 8.28 points, reflecting the removal of dangling modifier attachments that malformed spans introduce into dependency parses. Hard relation yield rose from 43.16% to 43.88%, a gain of 0.73 points.

The disparity between these two gains is itself a finding and is addressed in the Discussion.

Table 7
Downstream relation-extraction proxy metrics (SDP ≤ 5)

Metric	Baseline	Refined	Change
Syntactic stability	72.84	81.12	+ 8.28
Hard relation yield	43.16	43.88	+ 0.73

Values are percentages.

Relation to published results

Table 6 situates this work relative to published flat-NER results on the three natively flat corpora used here, and to published nested-NER results on GENIA.

We stress that the two blocks of Table 6 are not directly comparable. Published figures are computed over complete prediction sets, whereas the values reported here derive from a gold-anchored ledger and are upper bounds (Methods). Consistent with this, our baseline estimates exceed published baselines for the same backbone and corpus by approximately three to four points, a margin closely matching that predicted by the sensitivity analysis in Table 5. Table 6 is therefore presented to establish that this work operates in the same performance regime as the published literature, not to claim superiority over it. The contribution claimed here is the *improvement* attributable to refinement, which is measured internally under a single consistent protocol.

Discussion

Boundary correctness is necessary but not sufficient downstream

The most informative result in this study is the asymmetry between the two downstream metrics. An 8.28-point improvement in syntactic stability accompanied by only a 0.73-point improvement in hard relation yield indicates that repairing entity boundaries reliably improves the *structural* quality of the parse, spans no longer strand modifiers or truncate noun phrases, but that structural improvement translates only partially into additional verified relations.

This is coherent rather than contradictory. Hard relation yield depends on several factors beyond span correctness: predicate identification, argument attachment, and the secondary verification criterion itself. Boundary correction removes one failure mode among several. The honest interpretation is that boundary correctness is an enabling condition for downstream accuracy rather than a sufficient one, and reporting the larger structural gain without the smaller relational gain would overstate the practical benefit.

The value of a negative result

Variant A's complete collapse is retained and reported rather than omitted. The activation-collapse diagnosis where extreme positive-class sparsity in per-token boundary supervision, is informative for anyone considering per-token boundary regression on this task, and it explains why the pooled formulation succeeds where the per-token formulation fails. Evaluating a candidate span holistically converts a heavily imbalanced token-level problem into a balanced span-level one.

Variant C is also informative: it is genuinely safe, never harming a correct prediction, but recovers less than half of what Variant B recovers. For deployment contexts with zero tolerance for any regression, Variant C would be the defensible selection. The choice between B and C is therefore a deployment decision rather than a purely technical one.

Where the module fails

Qualitative audit of 49 unrecovered GENIA near-misses identified three recurring failure modes. Refiner Deletion (36.7%, n = 18) occurs where verifier confidence remains below τ for every candidate offset, and the module abstains. This is the No-Harm gate operating as designed, and the abstentions are recoverable recovery traded deliberately for safety; lowering τ converts some of these into repairs at the cost of a higher harm rate. Boundary Shift (32.7%, n = 16) occurs where the verifier prefers a shorter but locally valid substring over the modifier-inclusive gold span. Hyphenation Fracture (30.6%, n = 15) is a tokenisation-level failure in which WordPiece segmentation splits hyphenated chemical or gene names and the verifier does not recompose them; this is arguably addressable through tokenizer-aware candidate generation.

Limitations

Several limitations constrain the scope of these conclusions.

First, entity-level F1 was derived from a gold-anchored outcome ledger that does not record predictions overlapping no gold entity. The absolute F1 values reported here are consequently upper bounds and are not directly comparable to externally published figures, as discussed in the Results and quantified in Table 5. The measured improvement attributable to refinement is robust to this limitation, but the absolute performance level should be interpreted accordingly.

Second, GENIA results use the flat projection of a natively nested corpus. Nested entity structures are therefore not evaluated, and GENIA results are not comparable to nested-NER methods evaluated on the full annotation.

Third, comparison methods were not re-implemented within our evaluation harness. Table 6 reports published figures obtained under different experimental conditions; a controlled head-to-head comparison would be required to make any claim regarding relative performance.

Fourth, downstream evaluation used a dependency-path proxy rather than a fully annotated relation-extraction benchmark. The proxy provides a directional signal but does not substitute for evaluation against gold relation annotations.

Fifth, latency was measured on a single GPU model. Throughput on other accelerators, and on CPU, was not characterised.

Sixth, all four corpora are English-language and predominantly molecular-biology or chemical-disease in domain. Generalisation to clinical narrative text and to other languages remains untested.

Finally, recovery exhibits a corpus-determined ceiling: JNLPBA caps near 15% across all three backbones, indicating that certain boundary ambiguities particularly coordination structures are not addressable by the perturbation-and-verify strategy as currently formulated.

Conclusions

Residual error in contemporary BioNER is substantially a boundary problem rather than a type problem, and a meaningful fraction of it is recoverable post hoc. A pooled span verifier adding 1.13M parameters, approximately 1% of a standard backbone and 222 ms of inference latency recovers a mean 35.37% of near-miss boundary errors under strict matching, reaching 80.33% on the configuration where the baseline is weakest, while degrading 0.12% of already-correct predictions and improving strict entity-level F1 by a mean of 2.09 points.

Three properties make this practically relevant. The module is encoder-agnostic and attaches to a frozen backbone, so existing pipelines can adopt it without retraining. Its safety profile is empirically quantified rather than assumed, with a measured harm rate and a tunable threshold governing the recovery-safety trade-off. And its cost profile places it within reach of the single-GPU environments in which much biomedical text mining is actually performed.

The downstream analysis tempers these claims appropriately: boundary correction produces a large improvement in parse structure but a smaller improvement in verified relation extraction. Boundary correctness enables downstream accuracy without guaranteeing it.

Future work should recompute entity-level metrics over complete prediction sets to permit direct benchmarking against published results, extend evaluation to nested annotation schemes, re-implement

comparison methods within a unified harness for controlled comparison, replace the dependency proxy with a fine-tuned relation-extraction model, and test cross-lingual and clinical-domain generalisation.

Abbreviations

BioNER
Biomedical Named Entity Recognition
BRM
Boundary-Refinement Module
IoU
Intersection over Union
NMRV
Near-Miss Repair Velocity
RE
Relation Extraction
SDP
Shortest Dependency Path
LLM
Large Language Model

Declarations

Ethics approval and consent to participate

Not applicable. This study used only publicly available, de-identified benchmark corpora and did not involve human participants, human tissue, or animals.

Consent for publication

Not applicable.

Competing interests

The authors declares no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contribution

JBO conceived the study and supervised all activities of this study. WOH designed and implemented the models, conducted all experiments and analyses, and wrote the manuscript, The authors read and approved the final manuscript.

Acknowledgements

The author thanks the Department of Computer and Information Sciences, Covenant University, for institutional support.

Availability of data and materials

The benchmark datasets analysed during the current study are publicly available from their original sources: GENIA (http://www.geniaproject.org/genia-corpus), BC5CDR (https://biocreative.bioinformatics.udel.edu/tasks/biocreative-v/track-3-cdr/), NCBI-Disease (https://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/), and JNLPBA (http://www.geniaproject.org/shared-tasks/bionlp-jnlpba-shared-task-2004).

The trained model weights generated during the current study are available in the Zenodo repository, https://doi.org/10.5281/zenodo.21668876 [18].

All analysis code, derived per-prediction outcome ledgers, and figure-generation scripts are available at: Project name: [brm-boundary-refinement] Project home page: [GITHUB URL] Archived version: [10.5281/zenodo.21787326]. Operating system(s): Platform independent Programming language: Python Other requirements: Python 3.11 or higher; PyTorch 2.0 or higher; Transformers 4.35 or higher

References

1. Landhuis E. Scientific literature: information overload. Nature. 2016;535(7612):457–8.
2. National Library of Medicine. MEDLINE citation counts. Bethesda: U.S. National Library of Medicine; 2024. https://www.nlm.nih.gov/bsd/medline.html. Accessed 13/11/2025.
3. Li J, Sun Y, Johnson RJ, Sciaky D, Wei CH, Leaman R, et al. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database. 2016;2016:baw068.
4. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234–40.
5. Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. In: Proceedings of EMNLP-IJCNLP 2019. Hong Kong: Association for Computational Linguistics; 2019. pp. 3615-20.

6. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthc. 2021;3(1):1–23.
7. Chen Q, Sun H, Liu H, Jiang Y, Ran T, Jin X, et al. Benchmarking large language models for biomedical natural language processing applications and recommendations. Nat Commun. 2025;16:1–15.
8. Ge Y, Das S, Guo Y, Sarker A. (2025). Retrieval augmented generation based dynamic prompting for few-shot biomedical named entity recognition using large language models. Res Square, rs-3.
9. Tang M, He Y, Xu Y, Xu H, Zhang W, Lin Y. A boundary offset prediction network for named entity recognition. Findings of EMNLP 2023. Singapore: Association for Computational Linguistics; 2023. pp. 14834–46.
10. Shen Y, Song K, Tan X, Li D, Lu W, Zhuang Y. DiffusionNER: boundary diffusion for named entity recognition. In: Proceedings of ACL 2023. Toronto: Association for Computational Linguistics; 2023. pp. 3875-90.
11. Cai Y, Liu Q, Gui T. DiFiNet: boundary-aware semantic differentiation and filtration network for nested named entity recognition. In: Proceedings of ACL 2024. Bangkok: Association for Computational Linguistics; 2024. pp. 6455-69.
12. Shen Y, Ma X, Tang Y, Lu W. Locate and label: a two-stage identifier for nested named entity recognition. In: Proceedings of ACL 2021. Association for Computational Linguistics; 2021. pp. 2782-94.
13. Yu H, Chen T, Zhang X, Yang Y, Liu Q, Yang C, Wu Y. Performance of large language models in diagnosing rare hematologic diseases and the impact of their diagnostic outputs on physicians: combined retrospective and prospective study. J Med Internet Res. 2025;27:e77334.
14. Kim JD, Ohta T, Tateisi Y, Tsujii J. GENIA corpus — a semantically annotated corpus for bio-textmining. Bioinformatics. 2003;19(Suppl 1):i180–2.
15. Kim JD, Ohta T, Tsuruoka Y, Tateisi Y, Collier N. Introduction to the bio-entity recognition task at JNLPBA. In: Proceedings of JNLPBA/BioNLP. Geneva; 2004. pp. 70 – 5.
16. Wei CH, Peng Y, Leaman R, Davis AP, Mattingly CJ, Li J et al. Overview of the BioCreative V chemical disease relation (CDR) task. In: Proceedings of the Fifth BioCreative Challenge Evaluation Workshop. Sevilla; 2015. pp. 154 – 66.
17. Doğan RI, Leaman R, Lu Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. J Biomed Inf. 2014;47:1–10.
18. Hassan WO. Trained model weights for: a lightweight boundary-refinement module for biomedical named entity recognition. Zenodo. 2026. https://doi.org/10.5281/zenodo.21668876
19. Kanakarajan KR, Kundumani B, Sankarasubbu M. BioELECTRA: pretrained biomedical text encoder using discriminators. In: Proceedings of the 20th Workshop on Biomedical Language Processing. Association for Computational Linguistics; 2021. pp. 143 – 54.
20. Yasunaga M, Leskovec J, Liang P. LinkBERT: pretraining language models with document links. In: Proceedings of ACL 2022. Dublin: Association for Computational Linguistics; 2022. pp. 8003-16.

21. Zhang S, Cheng H, Vashishth S, Wong C, Xiao J, Liu X et al. Knowledge-rich self-supervision for biomedical entity linking. arXiv:2208.14565. 2022.

Figures

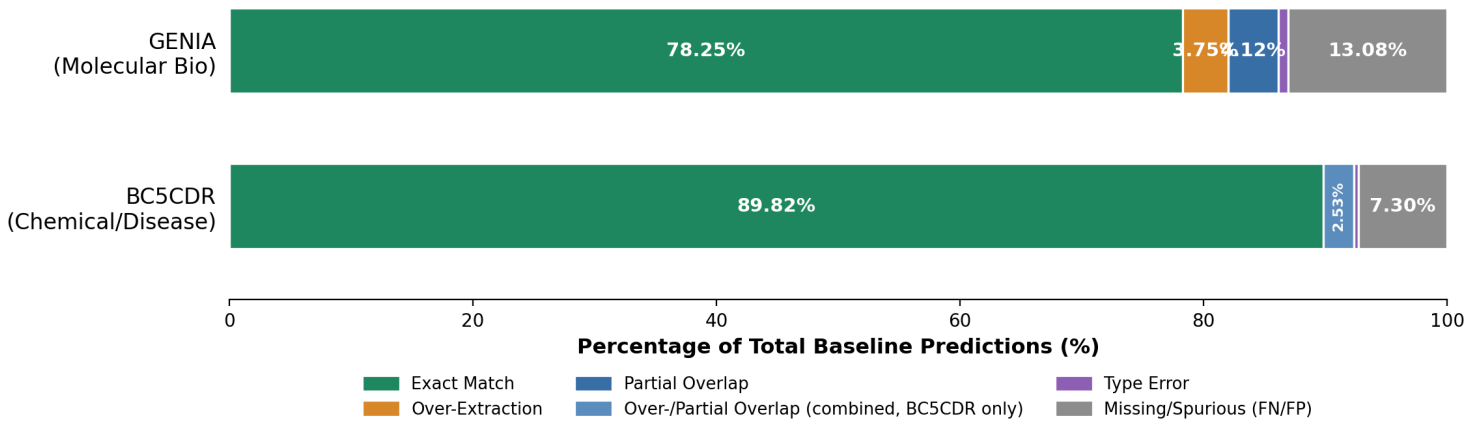


Figure 1

Error category distribution across datasets. Baseline predictions broken down by outcome category for GENIA (molecular biology) and BC5CDR (chemical/disease), showing that boundary-segmentation errors substantially exceed categorical type errors in both corpora.

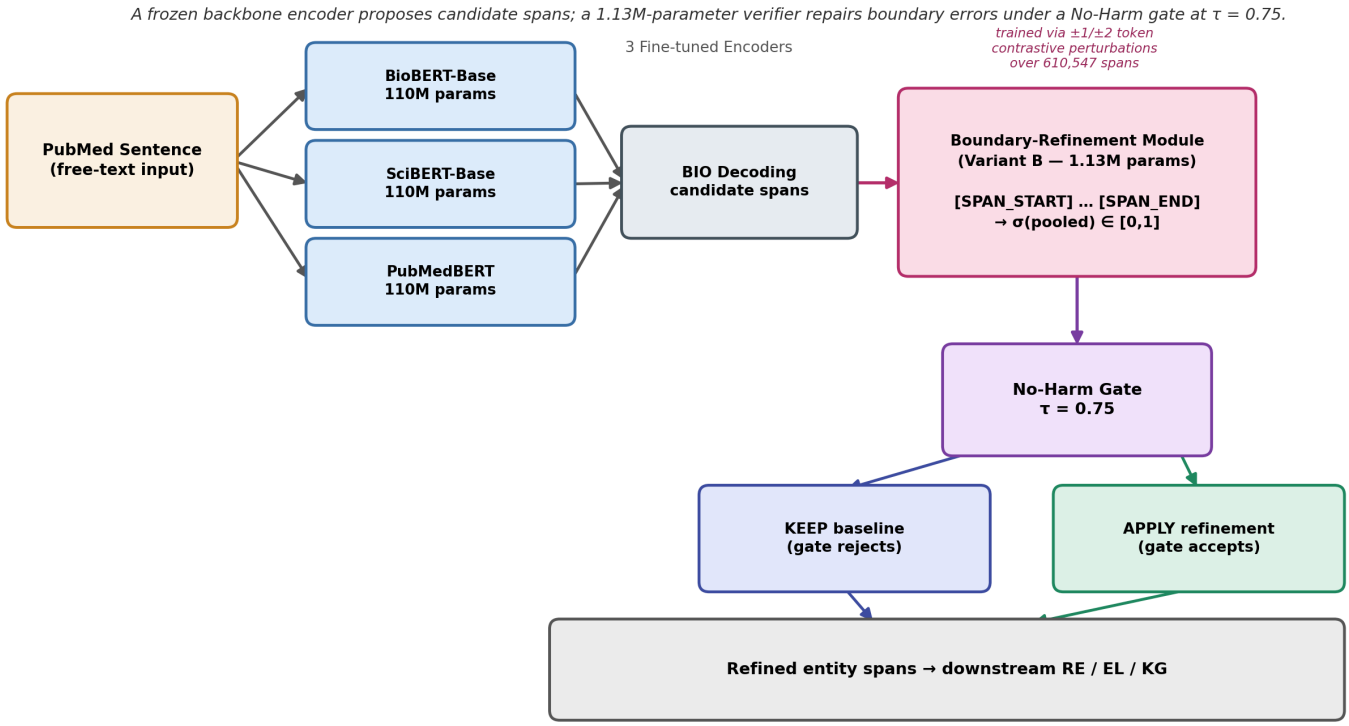


Figure 2

The boundary-refinement pipeline. A frozen backbone encoder proposes candidate spans; a 1.13M-parameter pooled verifier evaluates each candidate and applies refinement only when confidence clears the No-Harm gate at τ = 0.75.

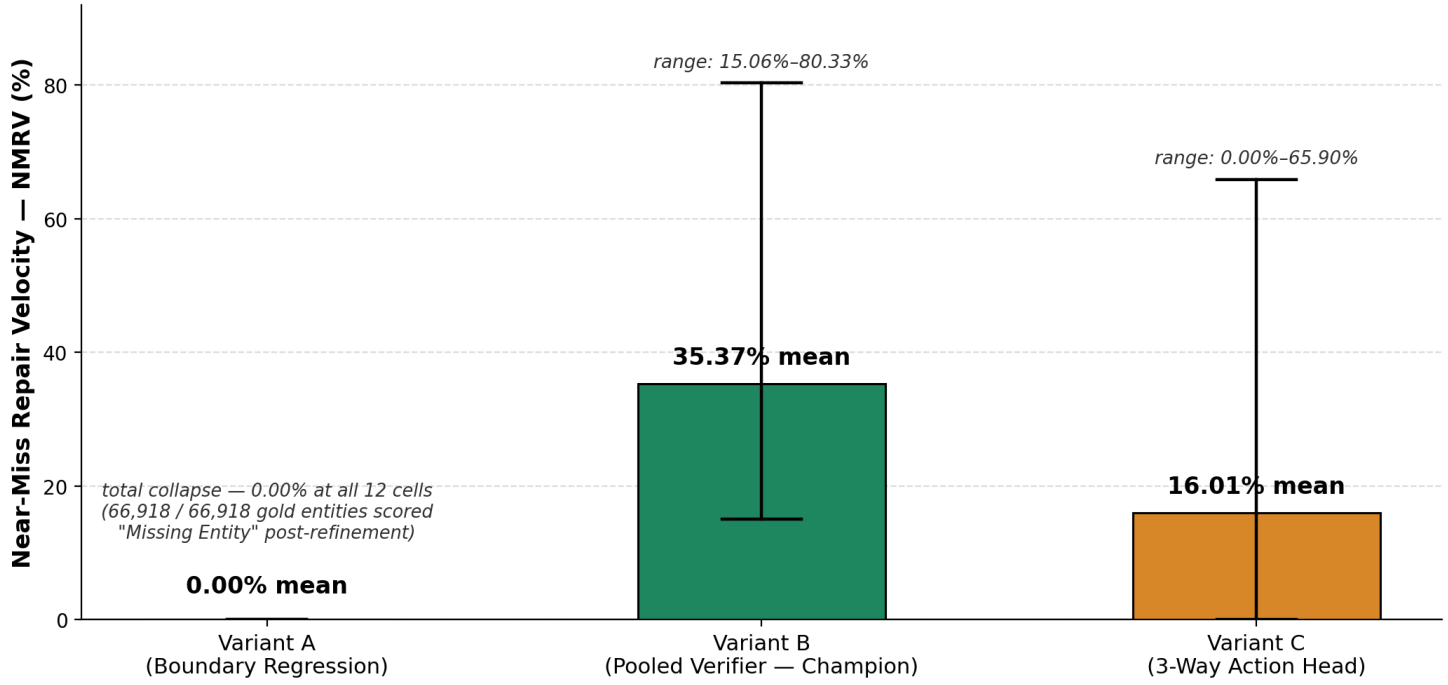


Figure 3

Three-variant ablation showing mean Near-Miss Repair Velocity with per-cell minimum-maximum range across twelve backbone × corpus cells under strict matching at IoU ≥ 0.5.

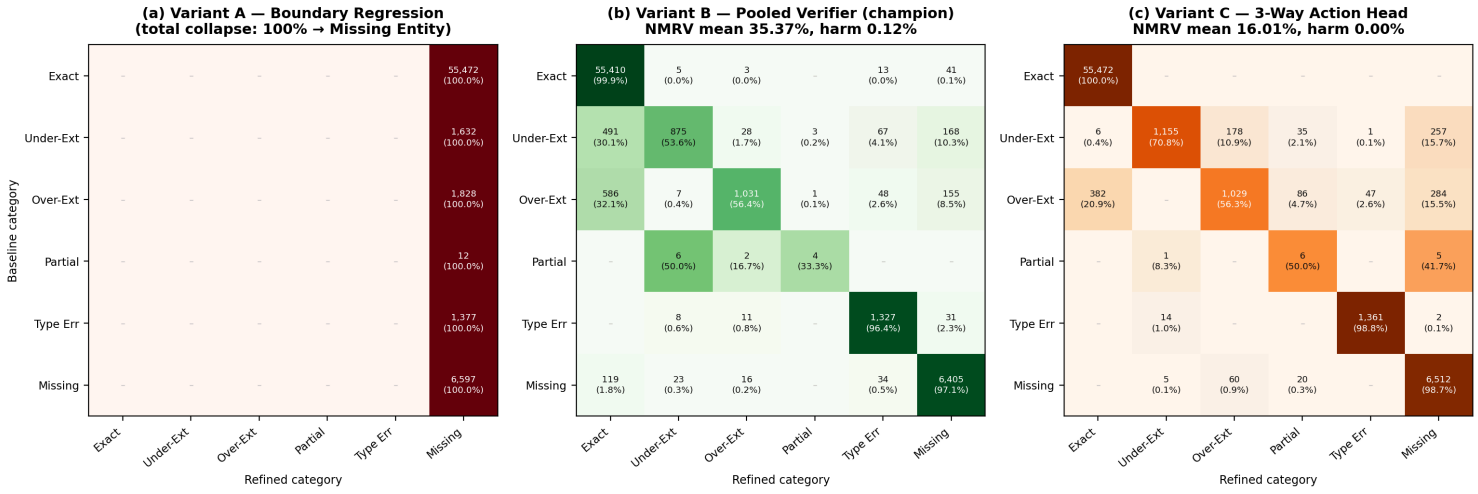


Figure 4

Pooled baseline-to-refined confusion matrices for all three candidate architectures, showing category transitions underlying the aggregate recovery figures.

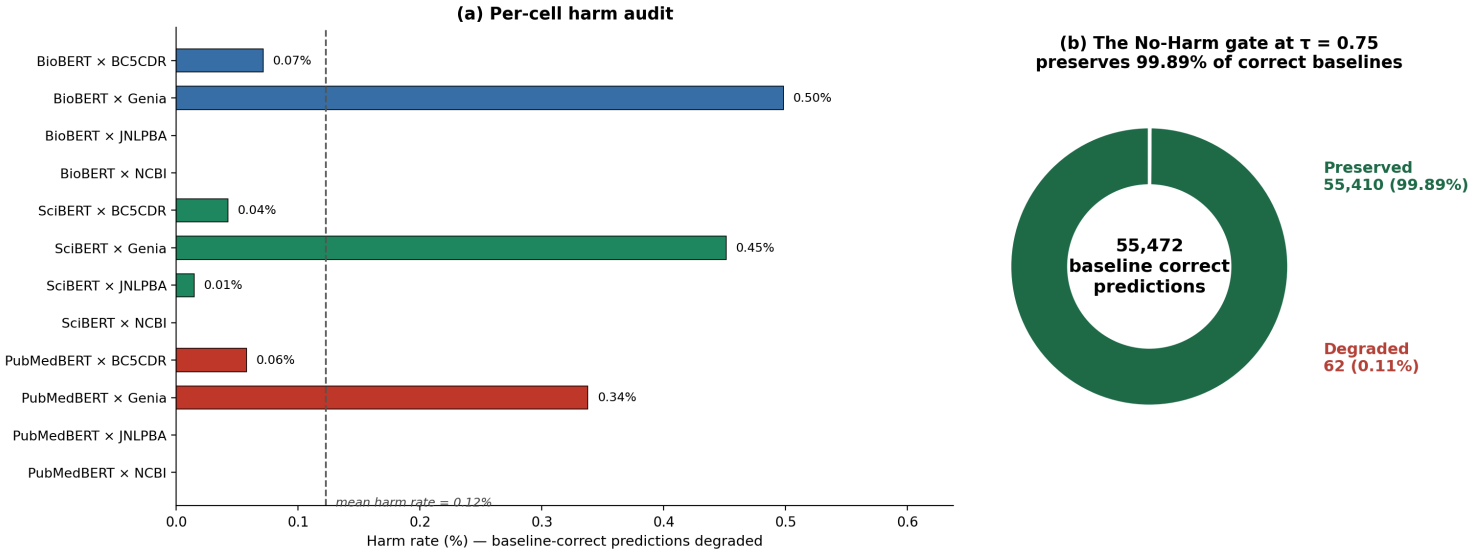


Figure 5

Empirical verification of the No-Harm gate. Per-cell harm rates (a) and pooled preservation of baseline-correct predictions (b).