Context-Aware SNOMED CT Entity Linking for Clinical Text

Provia Kadusabe[1], **Demian Gholipour Ghalandari**[2], **Lauren Cassidy**[2], **Jack Boylan**[2], **Chris Hokamp**[2], **Abhishek Kaushik**[1], **Fiona Lawless**[1]
[1]Dundalk Institute of Technology, Ireland
[2]Quantexa Ltd., Dublin, Ireland
{provia.kadusabe, abhishek.kaushik, fiona.lawless}@dkit.ie
{demiangholipour, laurencassidy, jackboylan, chrishokamp}@quantexa.com

Abstract

Mapping free-text mentions in clinical notes to standardized terminologies such as SNOMED CT is essential for large-scale secondary use of electronic health records, but remains challenging due to linguistic variability, under-specified annotation guidelines, term ambiguity, and ontology scale. This work presents a two-stage entity linking pipeline that combines span detection with context-aware concept linking and evaluates it on the SNOMED CT Entity Linking Challenge dataset. Our work builds upon the SNOMED CT entity linking challenge (Davidson et al., 2025), resulting in a fully open-source system. To our knowledge, this is the first end-to-end open-source system for this task. For span detection, we compare multiple neural architectures together with dictionary-based matching. For concept linking, we adopt a context-aware bi-encoder, and construct a multi-source knowledge base enriched with context derived from the SNOMED CT ontology. Finally, we implement an agentic re-ranker and test the effectiveness of LLM-backed re-ranking with access to annotation guidelines. In contrast to findings from the original shared task submissions, we show that context is important for optimal performance, and that agentic re-ranking with a state-of-the-art LLM only marginally improves overall performance, suggesting that the current benchmark may be approaching its practical ceiling. This work provides the first fully open-source, reproducible system for SNOMED CT entity linking, offering a foundation for future research and practical deployment.

Keywords: Clinical Entity Linking, SNOMED CT, Clinical NLP, Medical Concept Disambiguation, Concept Normalization

1. Introduction

Clinical notes in Electronic Health Records (EHRs) contain rich information about patients and treatment protocols, but much of this data is unstructured, making it difficult to analyze at scale. Mapping free-text mentions in clinical notes to standardized medical terminologies such as SNOMED CT [1] is a crucial step toward enabling large-scale secondary use of clinical data, facilitating clinical decision support, pharmacovigilance, and biomedical research. This task is commonly referred to as clinical entity linking or concept normalization (French and McInnes, 2023), which requires both accurately identifying spans of text that refer to medical entities and correctly mapping them to the appropriate concept identifiers. Figure 1 illustrates the task: free-text mentions in clinical notes are mapped to their corresponding SNOMED CT concepts.

Clinical entity linking presents several challenges. Clinical narratives make pervasive use of abbreviations and shorthand that may have multiple meanings (such as RA: Rheumatoid Arthritis vs. Right Atrium), requiring disambiguation during linking (Davidson et al., 2025). Concept granularity and semantic overlap further complicate candidate selection, as closely related concepts may differ only in subtle ways (such as “chest pain” vs. “anterior chest wall pain”). Span boundary detection poses additional difficulty, as modifiers, nested expressions, and overlapping mentions can lead to multiple plausible spans, each mapping to a different clinical concept. Finally, the target concept space is large; clinical terminologies such as SNOMED CT contain hundreds of thousands of concepts, yet annotated datasets typically cover only a small fraction, producing a long-tail distribution that requires systems to generalize to rare and previously unseen concepts at inference time (Davidson et al., 2025).

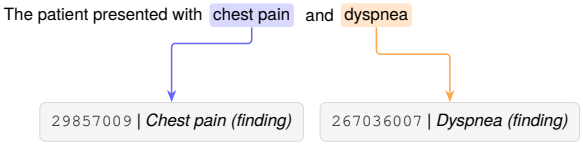

Figure 1: SNOMED CT entity linking: mentions in clinical text (highlighted) are linked to standardized concepts in the SNOMED CT ontology

The recent SNOMED CT Entity Linking Challenge (Davidson et al., 2025) provided a large, publicly available benchmark for this task, derived from de-identified discharge summaries in MIMIC-IV-Note (Johnson et al., 2023). The dataset includes about 75,000 manually annotated mentions spanning three high-level SNOMED CT subhierarchies: `Findings`, `Procedures`, and `Body Structures`. This benchmark has driven recent

[1] https://www.snomed.org

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 290–299

research on entity linking in clinical text. However, most existing top systems from the challenge largely fall into two paradigms: (i) highly engineered dictionary-based approaches that rely upon lexical matching and heuristics, or (ii) neural retrieval-based pipelines that primarily treat linking as string-to-concept matching, with limited use of broader document context.

In this work, we propose a two-stage entity linking pipeline that integrates heterogeneous span detection methods with context-aware retrieval. In the first stage, we compare three neural architectures for span detection with different inductive biases and model sizes: `GatorTron`, a large clinical language model fine-tuned with QLoRA (Yang et al., 2022); `BioClinicalBERT` (Alsentzer et al., 2019), augmented with a conditional random field decoding layer; and `GLiNER-BioMed` (Yazdani et al., 2025), a zero-shot, span-based, type-conditioned model. In the second stage, we adopt a contextual bi-encoder retrieval approach using `KRISSBERT` (Zhang et al., 2022), which encodes both mentions and knowledge base entries with surrounding document context rather than treating mentions as isolated strings. We construct a structured multi-source knowledge base that combines official SNOMED CT terminology, contextualized mention prototypes from the training data, OMOP/Athena[2] synonyms aligned to SNOMED concepts, and ontology-derived pseudo-sentences constructed from SNOMED relationship triples such as *is-a*, *finding site*, and *morphology* relations. This enables the model to leverage both local discourse cues and structured medical knowledge.

The key contributions of this work are:

- We propose and implement a fully reproducible, end-to-end framework for clinical entity linking.[3]
- We demonstrate that context-aware retrieval improves benchmark results over mention-only encoding, a finding not shown in previous SNOMED CT task submissions.
- We provide a thorough evaluation and detailed error analysis that reveal key challenges in clinical entity linking, especially with respect to the current SNOMED CT dataset.

2. Related Work

2.1. Background

Clinical entity linking aims to map text spans in clinical narratives to standardized medical concepts. As of early 2026, SNOMED CT serves as the primary knowledge resource for this task. The knowledge base contains over 350,000 clinical concepts organized in a poly-hierarchical structure, with each concept defined by a fully specified name, synonyms, and semantic relationships such as `finding site` and `causative agent`. In addition to SNOMED CT, clinical entity linking systems often rely on other complementary knowledge resources. The UMLS Metathesaurus (Unified Medical Language System[4]) (Bodenreider, 2004) integrates nearly 200 source vocabularies including SNOMED CT, ICD-10, MeSH, and RxNorm, into a shared concept space containing over 3 million concepts and more than 17 million unique concept names. UMLS provides cross-vocabulary mappings and synonym sets that are widely used for training entity linking models (Liu et al., 2021; Sung et al., 2020). The OMOP/Athena repository similarly provides standardized mappings across clinical coding systems and additional terminology.

2.2. Approaches for Clinical Entity Linking

Early approaches to clinical entity linking relied heavily on rule-based and dictionary-based systems such as MetaMap (Aronson, 2001) and cTAKES (Savova et al., 2010), which match text spans to terminology entries using curated lexicons and handcrafted heuristics. While these methods can achieve high precision for well-defined expressions, they struggle with linguistic variability, abbreviation ambiguity, and unseen terminology, and often require extensive manual engineering.

With the advent of pre-trained base models (Devlin et al., 2019), neural approaches have become dominant. Many recent systems frame entity linking as a retrieval problem, where mentions and concepts are embedded into a shared vector space and matched using cosine similarity. Encoder models such as SapBERT (Liu et al., 2021), BioSyn (Sung et al., 2020), and PubMedBERT (Gu et al., 2021) have been widely adopted for biomedical entity linking. These bi-encoder models are efficient at scale but typically encode mentions with little or no broader document context.

Prior work adopt a two-stage architecture (Kartchner et al., 2023; Gallego et al., 2024; French and McInnes, 2023; Sung et al., 2020): a bi-encoder first retrieves a small set of high-probability candidate concepts, then a context-aware re-ranker re-scores these candidates using a cross-encoder or language model that conditions on a textual window around the mention.

[2] https://athena.ohdsi.org/

[3] https://anonymous.4open.science/r/snomed-ct-entity-linking-project-E62B

[4] https://www.nlm.nih.gov/research/umls/

Recent context-aware biomedical entity linking models such as KRISSBERT (Zhang et al., 2022) are designed to capture richer contextual information than standard bi-encoders during retrieval. However, to the best of our knowledge, these context-first retrieval models have not been systematically evaluated on SNOMED CT entity linking in clinical text.

2.3. SNOMED CT Entity Linking Challenge

The SNOMED CT challenge provided a large-scale human-annotated dataset in which spans of text are labeled with their corresponding SNOMED CT concepts (Davidson et al., 2025). The top performing systems of the challenge employed different approaches. The first-place solution in the SNOMED CT Entity Linking Challenge employed a purely dictionary-based approach (Davidson et al., 2025). During inference, documents were split into sections and mentions were detected via regex matching.

The second-place solution (Kulyabin et al., 2024) employed a two-stage pipeline. First, an ensemble of six fine-tuned `BiomedBERT-large` models detected medical entities. Then `SapBERT` was used for concept linking. Additional post-processing included dictionary filtering, section removal, and confidence thresholds. However we note discrepancies between reported methods and released implementations; for example, the paper mentions a re-ranking step that is not present in the available implementation.

The third-place solution used an LLM-based approach for both entity recognition and classification. Two fine-tuned `Mistral` models detected entities on different chunk sizes, with dictionary-based post-processing. In the second stage, a Faiss index (Johnson et al., 2019) retrieved candidate concepts, and a third `Mistral` model selected the final concept. The third place uses a small window of context but only in the 3rd stage for re-ranking.

Our work takes inspiration from these solutions but also addresses some of the issues we identified, such as discrepancies between reported methods and released implementations, and unclear evaluation methodology which is not possible to fully reproduce in its presented form. In this work, we provide a fully reproducible system that additionally incorporates context at the retrieval stage, an approach not implemented by any of the challenge submissions and cleanly separate span detection evaluation from concept linking evaluation.

3. Methods

3.1. Data

We used version 1.1 of the SNOMED CT Entity Linking Challenge dataset.[5] The full dataset contains 272 annotated clinical notes with character-level entity annotations, each consisting of a `note identifier`, `start` and `end` offsets, and a SNOMED CT concept identifier. The original challenge was based on the May 2023 International Edition of SNOMED CT and restricted to three sub-hierarchies (`Clinical Findings`, `Procedures`, and `Body Structures`).[6] Table 1 reports summary statistics of the original challenge dataset. Due to availability constraints, we used the November 2025 International Edition of SNOMED CT in our experiments instead of the May 2023 version used in the challenge. This version mismatch resulted in 102 concepts being unavailable (due to deprecation or restructuring), affecting 611 annotations in total (409 in the training set and 202 in the test set).

Split	Notes	Annotations	Unique Concepts	Unseen in Train
Train	204	51,574	5,336	–
Test	68	23,234	4,082	1,288

Table 1: Summary statistics of the SNOMED CT Entity Linking Challenge dataset.

3.2. Entity Linking Pipeline

Our system follows a two-stage pipeline architecture for SNOMED CT entity linking in clinical text, as illustrated in Figure 2.

Stage 1: Span Detection We evaluate three neural architectures for entity span detection. `GatorTron` (Yang et al., 2022) a large clinical language model that we fine-tune using parameter-efficient LoRA adapters with 4-bit quantization. `BioClinicalBERT-CRF` which we build on `Bio_ClinicalBERT` (Alsentzer et al., 2019) by adding a conditional random field layer to better model dependencies between adjacent labels during decoding. Both `GatorTron` and `BioClinicalBERT-CRF` frame NER as a token classification problem using BIO tagging. `GLiNER-BioMed` (Yazdani et al., 2025) is a zero-shot span-based model that represents entity types as learnable embeddings and scores candidate text spans directly against these representations.

[5] The dataset is publicly available on PhysioNet and is derived from the MIMIC-IV-Note corpus of de-identified discharge summaries from Beth Israel Deaconess Medical Center (Hardman et al., 2025).

[6] Further details about the challenge are described in (Davidson et al., 2025).

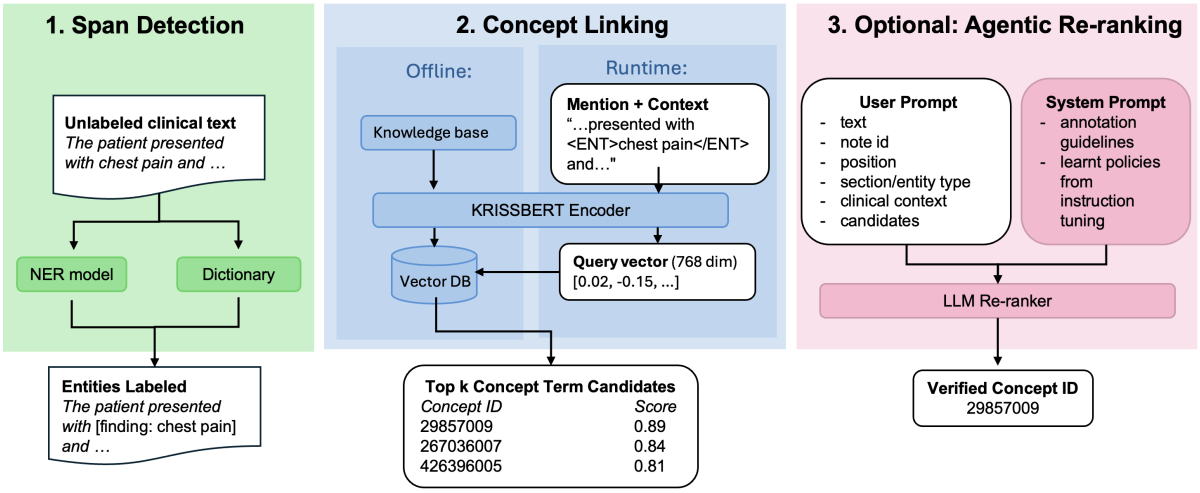

Figure 2: Our three-stage SNOMED CT entity linking pipeline: (1) span detection via neural NER (`GatorTron`, `BioClinicalBERT-CRF`, or `GLiNER`) and dictionary matching, (2) bi-encoder concept retrieval using `KRISSBERT` against a multi-source knowledge base (SNOMED terms, training prototypes, OMOP synonyms, and ontology pseudo-sentences), and (3) optional agentic re-ranking.

All models predict three SNOMED CT semantic categories: procedure, finding, and body structure. We additionally employ dictionary-based detection using an Aho-Corasick automaton (Aho and Corasick, 1975) constructed from surface forms in the training set. Terms with training precision below 0.50 are blacklisted to filter ambiguous matches. Predictions are post-processed by removing spans in medication sections or section headers, and filtering stopwords. Overlapping spans are resolved by retaining the longest match. The neural and filtered dictionary predictions are then merged and deduplicated to obtain candidate spans for linking.

Stage 2: Concept Linking Stage 2 takes the entity spans from stage 1 as input and maps them to their respective SNOMED CT concept identifiers. We adopt a bi-encoder-style retrieval approach using `KRISSBERT` (Zhang et al., 2022), a biomedical entity linking model pre-trained on PubMed and UMLS. Mentions and candidate terms are wrapped with entity markers (`<ENT>` and `</ENT>`) and encoded into L2-normalized representations. For comparison, we also implement `SapBERT` as a baseline; key architectural and training differences between `SapBERT` and KRISSBERT are summarized in Table 2. We implement the `KRISSBERT` aligned packing by tokenizing left context, mention, and right context separately as shown in Figure 3 and truncating context symmetrically to preserve the marked mention span. The resulting embeddings are stored in a Qdrant[7] vector database and retrieved using cosine similarity.

The knowledge base integrates four sources: (1) SNOMED CT terms extracted from RF2 releases and filtered to include only the three relevant subset SNOMED sub-hierarachies; (2) training mention prototypes augmented with a surrounding context window; (3) OMOP/Athena vocabulary synonyms mapped to SNOMED concepts to expand lexical coverage; and (4) ontology-derived pseudo-sentences constructed from SNOMED relationship triples by incorporating *is-a* relations, finding sites, and associated morphology.

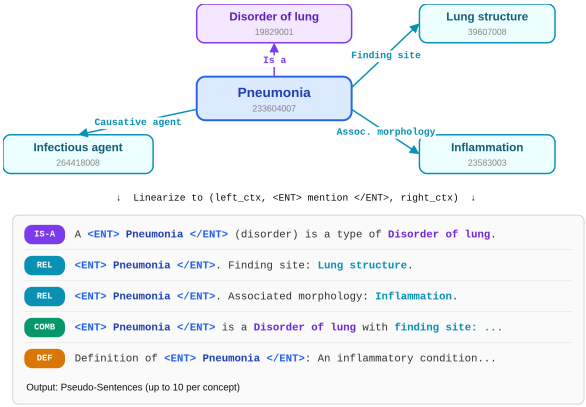

Figure 3: Pseudo-sentence construction from SNOMED CT relationship triples. Each concept's local ontology neighbourhood is linearized into up to 10 pseudo-sentences using four templates: IS-A (parent/class relationships), REL (single attribute relationships), COMB (a combined parent + key attributes in one sentence), and DEF (text definitions when available). Each pseudo-sentence is represented as (`left_ctx`, `<ENT>` mention `</ENT>`, `right_ctx`) and encoded with KRISSBERT for indexing in Qdrant.

At inference, each mention is encoded together

[7] https://github.com/qdrant/qdrant

Aspect	SapBERT	KRISSBERT
Backbone model	PubMedBERT	PubMedBERT
Training data	UMLS ontology (concept names and synonyms)	UMLS ontology + large unlabeled biomedical text (such as PubMed abstracts)
Input format	Entity name / synonym (no surrounding sentence)	Sentence with marked mention span (special entity markers)
Pooling	Mean pooling over token embeddings of the entity name	CLS representation of the contextualized mention
Training objective	Self-alignment via contrastive learning on synonyms	Knowledge-rich self-supervision + contrastive learning on contextualized mentions mined from text

Table 2: Differences between SapBERT and KRISSBERT

with its surrounding context window and entity markers. The query embedding is compared against the knowledge base using cosine similarity. Multiple matched surface forms mapping to the same concept are deduplicated by retaining only the highest-scoring match per concept ID, and the top-k ranked concepts after deduplication are returned.

Stage 3: Agentic Re-ranking (optional) We additionally implement an agentic re-ranker which has access to the top-k concepts from the concept linking step, and the annotation guidelines given to annotators [8]. The re-ranker implementation is motivated by the observation that correct candidates are often present in the top-k concepts retrieved from the knowledge base. The agent learns its re-ranking policy by sampling outputs from the training dataset. For each mention, the model receives: mention text and offsets, local note context (120 characters per side), and the top-5 linked candidates with enriched SNOMED metadata (FSN, semantic tag, embedding score, synonyms, definition, parent concepts, and defining relationships). The system prompt includes compressed annotation guidelines plus learned policies, capped at 20k characters for cost control.

The re-ranker produces one of three actions (`re-rank`, `drop`, `modify_span`). In the policy optimization step, the agent learns that the Stage 2 linker is already strong and learns conservative post-hoc guardrails to minimize harmful changes: (i) re-rank only mentions with top-1 score ≤ 0.986, (ii) keep top-1 changes only when the original top-1/top-2 score gap is ≤ 0.001, (iii) require `high` confidence for top-1 changes, and (iv) disable `drop` and non-top-1 reorders.

4. Experiments

We note that direct comparison with the top-performing systems from the SNOMED CT Entity Linking Challenge is not feasible for several reasons. First, the underlying SNOMED CT release file used in this work differs from the version used during the challenge, which affects the dataset. Second, while the challenge codebases are publicly available, the actual submission outputs are not, making it impossible to evaluate prior systems under identical conditions. Third, as discussed in Section 2, we found discrepancies between reported methods and released implementations for some solutions. For these reasons, we focus on evaluating our pipeline components in a controlled and reproducible setting rather than attempting potentially misleading comparisons with prior submissions.

For span detection, we experiment with the training dataset mapped to three entity types: *procedure*, *finding*, and *body structure*. We use a fixed 4-fold note-level split, holding out fold 0 for validation and using folds 1–3 for training. Long documents are split into overlapping chunks of 512 tokens with stride 128. We compare three architectures: (1) `UFNLP/gatortron-large` fine-tuned using QLoRA with 4-bit quantization and LoRA adapters (r=8, α=16, lr=2e-4); (2) `emilyalsentzer/Bio_ClinicalBERT` with a CRF layer for sequence labeling (lr=3e-5); and (3) `Ihor/gliner-biomed-large-v1.0`, a span-based model (lr=5e-6). All models are trained for up to 30 epochs with early stopping (patience=3)[9].

4.1. Evaluation Metrics

Span Detection Metrics We evaluate span detection using `strict`, `exact`, and `partial` matching criteria from `nervaluate`[10]. Strict matching requires exact span and entity-type agreement, exact matching ignores entity type, and partial matching awards partial credit for overlapping spans.

Concept Linking Metrics We evaluate concept linking performance using Recall@K and Mean Reciprocal Rank (MRR). Recall@K measures the pro-

[8] https://github.com/google/healthcare-text-annotation/blob/master/guidelines/defining-entity-categories.md

[9] All code is available at https://github.com/PROVIA1/snomed-ct-entity-linking-project

[10] https://github.com/MantisAI/nervaluate

portion of mentions for which the correct concept appears within the top-K predicted candidates.

Macro-Averaged Character-Level IoU The primary evaluation metric used in the in the SNOMED CT Entity Linking challenge is macro-averaged character-level Intersection-over-Union (IoU). For a given concept class c, let P^c_{char} be the set of all characters that fall within any predicted span assigned to c, and let G^c_{char} be the corresponding set of characters covered by gold spans of class c. The IoU for class c is defined as:

$$\text{IoU}_c = \frac{|P^c_{char} \cap G^c_{char}|}{|P^c_{char} \cup G^c_{char}|}.$$

Let $\mathcal{C}$ denote the set of all concept classes that appear in either the predicted annotations or the gold annotations. The overall score is obtained by averaging the per-class IoU values across this set:

$$\text{Macro-IoU} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \text{IoU}_c.$$

This metric is particularly challenging because it is highly sensitive to boundary errors and class frequency. As IoU is computed at the character level, even small span mismatches (such as missing modifiers or slight boundary shifts) can lead to substantial score reductions. Moreover, macro-averaging assigns equal weight to all concept classes, so rare concepts strongly affect the score.

5. Results

We evaluate our two-stage entity linking pipeline in two parts. First, we report end-to-end system performance. Second, we evaluate each stage performance.

5.1. End-to-End Performance

End-to-end evaluation of our system shown in Table 3 reveals that `KRISSBERT` consistently outperforms `SapBERT`. `KRISSBERT` + Dict uses the spans generated by the Stage-1 `GatorTron` + dictionary merge, whereas `KRISSBERT` and `SapBERT` use spans produced by GatorTron alone. Using spans produced by the Stage-1 dictionary merge yields further gains in end-to-end Recall and MRR.

Table 4 reports macro-averaged character IoU at different retrieval confidence score thresholds, where predictions are retained only if the model's top-1 score exceeds the given threshold. We observe that a retrieval confidence threshold of 0.98 yields the highest macro-averaged character IoU for both `KRISSBERT` and `KRISSBERT` + Dict.

Model	Recall@1	Recall@5	Recall@10	MRR
SapBERT	60.66	73.88	74.93	0.666
KrissBERT	69.56	74.93	75.43	0.720
KrissBERT + Dict	**72.65**	**78.46**	**78.99**	**0.752**

Table 3: End-to-End Concept Linking Performance

Model	None	≥0.95	≥0.98	≥0.99
SapBERT	0.3189	**0.3413**	0.3339	0.3271
KrissBERT	0.3429	0.3457	**0.3615**	0.2898
KrissBERT + Dict	0.3658	0.3686	**0.3869**	0.3154

Table 4: Macro-averaged Character IoU at Different Score Thresholds

5.2. Stage 1: Span Detection Performance

Table 5 summarizes overall model performance under strict, exact, and partial evaluation.

`GatorTron-Large` achieved the best overall F1 scores for all the metrics, while `BioClinicalBERT` obtained the highest recall under partial matching. `GLiNER-BioMed-Large` performed consistently lower overall.

Per-entity results (Table 6) show substantial variation across entity types. GLiNER-BioMed-Large performed best on *Finding* entities but showed lower performance on *Procedure* recognition. `GatorTron-Large` achieved the strongest results for *Procedure* entities, while *Body Structure* entities were the most challenging across models.

5.3. Stage 2: Concept Linking Perfomance

Table 7 reports performance of our system when evaluated on exact gold spans, `KRISSBERT` substantially improves Recall@1 over `SapBERT` (+11.6 points), confirming the benefit of contextual encoding for disambiguation.

5.4. Agentic Re-ranking Performance

Initially, we expected the agentic re-ranker, powered by a state-of-the-art LLM with access to the full annotation guidelines and rich SNOMED CT metadata, to recover a substantial portion of the near-miss errors identified in our analysis . However, the gains were surprisingly modest.

At 5,000 mentions (Table 8), the re-ranker intervenes on 12.9% of cases but retains only 49 effective top-1 changes, yielding +11 improved versus 4 degraded on gold spans. We did not scale to the full test set as the marginal gains did not justify the computational cost. The limited success of this approach may indicate that achievable performance on this dataset is approaching saturation,

Model	Strict			Exact			Partial		
	P	R	F1	P	R	F1	P	R	F1
GatorTron-Large	**0.751**	**0.750**	**0.750**	**0.770**	**0.768**	**0.769**	**0.824**	0.822	**0.823**
BioClinicalBERT	0.723	0.749	0.736	0.740	0.767	0.753	0.799	**0.828**	0.814
GLiNER-BioMed-Large	0.638	0.636	0.637	0.654	0.651	0.653	0.769	0.765	0.767

Table 5: Stage 1 NER Model Comparison across Evaluation Metrics

Model	Finding			Procedure			Body Structure		
	P	R	F1	P	R	F1	P	R	F1
GatorTron-Large	0.767	0.774	0.770	**0.774**	0.747	**0.760**	0.653	0.670	**0.661**
BioClinicalBERT	0.747	0.759	0.753	0.752	**0.765**	0.758	0.592	**0.680**	0.633
GLiNER-BioMed-Large	**0.775**	**0.777**	**0.776**	0.416	0.410	0.413	**0.663**	0.658	**0.661**

Table 6: Per-Entity-Type Performance (Strict Metric)

Model	Recall@1	Recall@5	Recall@10	MRR
SapBERT	78.99	96.20	97.56	0.867
KrissBERT	**90.58**	**97.57**	**98.21**	**0.937**
KrissBERT + Dict	90.23	97.46	98.11	0.935

Table 7: Concept Linking Performance: Linking-only evaluated on predictions with exact span matches to gold annotations.

with remaining errors driven less by insufficient reasoning and more by inherent ambiguity in both the span annotations and concept linking targets.

6. Error Analysis

We conducted a detailed error analysis on the test set predictions from our best-performing configuration (`GatorTron` NER + dictionary merge + `KRISSBERT` linking).

Of 24,555 predictions against 23,032 gold annotations, our pipeline achieved 80.5% exact span coverage of gold (18,543 matches). Among exact span matches, 90.2% (16,732) were linked to the correct SNOMED CT concept, while 9.8% (1,811) were linked to incorrect concepts. We categorize errors into three types: false negatives (missed mentions), false positives (spurious predictions), and concept linking errors.

False Negatives The pipeline missed 4,489 gold mentions (19.5% of annotations). Analysis revealed that 60.4% (2,713) of missed mentions had partial span overlap with predictions, indicating boundary detection issues rather than complete detection failures. Common boundary errors included spans differing by 1–3 characters (such as predicting “BLOOD Glucose” instead of “Glucose”). The remaining 39.6% (1,776) were completely undetected. These were predominantly generic clinical terms, such as “Evaluation procedure” (48 instances), “Recommendation to stop drug treatment” (25), and “Patient medication education ” (17).

False Positives The pipeline generated 6,012 spurious predictions not present in gold annotations. Common false positives included ambiguous short terms: “left/right” (81 instances), “BLOOD Glucose” (74), “Alert and” (59), and generic terms like “surgery,” “normal,” and “medications.”

Concept Linking Errors Among the 1,811 concept linking errors (correct span, wrong concept), we observed that 66.4% involved predictions within the same semantic category as the gold concept (such as both being procedures), indicating fine-grained disambiguation challenges rather than categorical errors. In particular, these errors occurred with high model confidence which suggests that embedding similarity alone is insufficient for distinguishing closely related concepts. Analysis of candidate rankings revealed that in 80.7% of linking errors, the gold concept appeared within the top-10 retrieved candidates. Table 9 shows the distribution of gold concept positions when the top-1 prediction was incorrect. Nearly half (49.0%) of errors had the gold concept at position 2, just below the selected prediction. An additional 25.0% had gold at positions 3–5. Only 19.3% of errors had the gold concept absent from the top-10 retrieved candidates.

7. Discussion

In this section, we discuss our findings, identify key performance bottlenecks, and outline implications for future research.

7.1. Context as the Key Differentiator in Concept Linking

On exact span matches, `KRISSBERT` achieves a Recall@1 of 90.58%, compared to 78.99% for `SapBERT`, an absolute improvement of 11.6 percentage points (Table 7). This gap is consistent across retrieval depths, with `KRISSBERT` maintaining its advantage at Recall@5 (97.57% vs. 96.20%) and

n	re-ranked	Effective top-1	Improved	Degraded	Unchanged	Non-gold	Δ Recall@1	Δ Macro-IoU	Tokens (in/out)	Cost (USD)
200	26	3	1	1	0	1	+0.0000	+0.000157	0 / 0 (cached)	0.00
500	88	9	4	1	0	4	+0.0002	+0.000456	700,462 / 13,338	2.30
5000	645	49	11	4	3	31	+0.0004	**+0.001062**	5,655,294 / 129,636	18.91

Table 8: Conservative re-ranker scaling results on the first n test mentions (offset 0). Non-gold counts indicate top-1 changes on predicted spans without a matching gold span at evaluation time.

Gold Position	Count	% of Errors
Position 2	888	49.0%
Position 3–5	452	25.0%
Position 6–10	121	6.7%
Not in candidates	350	19.3%

Table 9: Distribution of gold concept positions in candidate list when top-1 prediction is incorrect.

Recall@10 (98.21% vs. 97.56%). The performance difference can be attributed to a fundamental architectural distinction between the two models. `SapBERT` encodes mentions in isolation, while `KRISSBERT` incorporates local context, which is particularly important in clinical text where abbreviations and short expressions are common. This finding aligns with recent literature. Kartchner et al. (2023), in their systematic evaluation of nine biomedical entity linking models, identified the effective incorporation of context into linking decisions as a persistent gap across current methods. The SNOMED CT Challenge (Davidson et al., 2025) similarly emphasized the need to better leverage contextual information when training examples are scarce. The results confirm that the benefits of contextual encoding extend to the large, multi-hierarchical ontologies like SNOMED CT.

7.2. Span Detection as the Performance Bottleneck

Error propagation is a well-documented limitation of pipeline-based biomedical entity linking, where mistakes in mention detection impose an upper bound on linking performance (Noh and Kavuluru, 2021; Sarol et al., 2024; Kartchner et al., 2023). Our results confirm that span annotation is a major bottleneck in the SNOMED CT entity linking task. When correct spans are provided, KRISSBERT achieves 90.58% Recall@1 (Table 7), yet end-to-end mIoU reaches only 0.3869. This is further supported by our upper bound analysis (see Appendix A) where providing gold spans increases Recall@10 from 78.99% to 96.12%. This gap originates primarily from Stage 1, where our best NER model achieves a strict F1 of only 0.750, missing 19.5% of gold mentions. Of the missed mentions, 60.4% exhibit partial span overlap, indicating boundary errors rather than detection failures. The remaining 39.6% are dominated by implicit clinical concepts such as "Evaluation procedure" and "Patient medication education" and may reflect annotation ambiguity.

The three NER architectures we evaluated make different kinds of errors which reveals the challenging nature of this task. `GatorTron-Large` provides the most balanced performance overall (strict F1: 0.750), whereas `GLiNER-BioMed-Large` achieves the highest F1 (0.776) for Findings but performs poorly on Procedures (0.413). In contrast, `BioClinicalBERT` attains a strict F1 of 0.736 and the highest partial recall of all models (0.828) despite being approximately 80 times smaller than `GatorTron-Large`, indicating that smaller models can still be highly competitive for clinical NER. Together, these results show that no single architecture handles all entity types equally well, indicating potential for ensemble strategies that route predictions through the strongest model for each type.

The continued need to complement neural methods with dictionary-based approaches across different architectures suggests a genuine limitation of current span detection models or ambiguity in the task specification itself.

8. Conclusion

In this work, we presented a two-stage entity linking pipeline for mapping clinical text to SNOMED CT concepts, combining neural span detection with contextual bi-encoder retrieval. We also implemented an agentic re-ranker and showed that it does not substantially improve performance. Our results demonstrate that context-aware encoding via `KRISSBERT` substantially outperforms context-free alternative for concept linking, and that ontology-derived pseudo-sentences provide meaningful knowledge base enrichment, but specification, annotation quality, and evaluation metrics can likely be improved for future versions of this task. The proposed system is fully reproducible and adaptable to other clinical entity linking tasks.

9. Limitations

Several limitations should be acknowledged. First, our experiments use the November 2025 International Edition of SNOMED CT rather than the May 2023 edition used in the original challenge. This affects 102 deprecated or restructured concepts (611 annotations), potentially limiting comparability with

prior results and imposing a practical ceiling on our reported performance. This version mismatch highlights an underappreciated challenge for clinical NLP systems: ontology versioning. SNOMED CT releases new editions bi-annually, and any production system must contend with concept deprecation, merging, and restructuring over time.

Second, while our agentic re-ranker produces consistent positive gains under conservative guardrails, the improvements remain small relative to the computational cost, and the approach was not scaled to the full test set. Whether more sophisticated re-ranking strategies can yield larger gains remains an open question.

Third, our evaluation is limited to English-language clinical text from a single institution (Beth Israel Deaconess Medical Center). The generalizability to other clinical settings, documentation styles, and languages remains to be evaluated.

Finally, the macro-averaged character-level IoU metric, while providing a unified measure of system performance, conflates span detection and concept linking errors in ways that can obscure the contributions of individual pipeline components. Our separate evaluation of NER (via nervaluate) and linking (via Recall@k on exact spans) addresses this to some extent, but standardized component-level evaluation protocols would benefit the broader community.

10. Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions, which helped improve the quality of this manuscript.

The authors also acknowledge the support and collaboration of Quantexa, and in particular the Ireland-based Quantexa NLP group.

This research was partly funded through the CREATE-DkIT project, supported by the HEA TU-Rise program and co-financed by the Government of Ireland and the European Union through the Southern, Eastern & Midland Regional Program of the ERDF 2021-27 and the Northern & Western Regional Programme 2021–27.

11. References

Alfred V Aho and Margaret J Corasick. 1975. Efficient string matching: an aid to bibliographic search. *Communications of the ACM*, 18(6):333–340.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*, pages 72–78.

Alan R Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Rory Davidson, Will Hardman, Guy Amit, Yonatan Bilu, Vincenzo Della Mea, Aleksandr Galaida, Irena Girshovitz, Mikhail Kulyabin, Mihai Horia Popescu, Kevin Roitero, et al. 2025. Snomed ct entity linking challenge. *Journal of the American Medical Informatics Association*, 32(9):1397–1406.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Evan French and Bridget T McInnes. 2023. An overview of biomedical entity linking throughout the years. *Journal of biomedical informatics*, 137:104252.

Fernando Gallego, Guillermo López-García, Luis Gasco-Sánchez, Martin Krallinger, and Francisco J Veredas. 2024. Clinlinker: Medical entity linking of clinical concept mentions in spanish. In *International Conference on Computational Science*, pages 266–280. Springer.

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. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Will Hardman, Mark Banks, Rory Davidson, Donna Truran, Nindya Widita Ayuningtyas, Hoa Ngo, Alistair Johnson, and Tom Pollard. 2025. SNOMED CT Entity Linking Challenge. *PhysioNet*. Version 1.1.0.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023. MIMIC-IV-Note: Deidentified free-text clinical notes. *PhysioNet*. Version 2.2.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE transactions on big data*, 7(3):535–547.

David Kartchner, Jennifer Deng, Shubham Lohiya, Tejasri Kopparthi, Prasanth Bathala, Daniel Domingo-Fernández, and Cassie Mitchell. 2023. A comprehensive evaluation of biomedical entity linking models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14462–14478.

Mikhail Kulyabin, Gleb Sokolov, Aleksandr Galaida, Andreas Maier, and Tomas Arias-Vergara. 2024. Snobert: A benchmark for clinical notes entity linking in the snomed ct clinical terminology. In *International Conference on Pattern Recognition*, pages 154–163. Springer.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 4228–4238.

Jiho Noh and Ramakanth Kavuluru. 2021. Joint learning for biomedical ner and entity normalization: encoding schemes, counterfactual examples, and zero-shot evaluation. In *Proceedings of the 12th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 1–10.

M Janina Sarol, Gibong Hong, Evan Guerra, and Halil Kilicoglu. 2024. Integrating deep learning architectures for enhanced biomedical relation extraction: a pipeline approach. *Database*, 2024:baae079.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical entity representations with synonym marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3641–3650, Online. Association for Computational Linguistics.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, Ying Zhang, Tanja Magoc, Christopher A Harle, Gloria Lipori, Duane A Mitchell, William R Hogan, Elizabeth A Shenkman, Jiang Bian, and Yonghui Wu. 2022. A large language model for electronic health records. *npj Digital Medicine*, 5(1):194.

Anthony Yazdani, Ihor Stepanov, and Douglas Teodoro. 2025. Gliner-biomed: A suite of efficient models for open biomedical named entity recognition.

Sheng Zhang, Hao Cheng, Shikhar Vashishth, Cliff Wong, Jinfeng Xiao, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Knowledge-rich self-supervision for biomedical entity linking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 868–880.

A. Appendix A. Oracle System Upper Bound Analysis

To quantify the upper bound of the system, we performed an oracle experiment in which gold test spans were passed directly to KRISSBERT, bypassing Stage 1 entirely. Table 10 compares this setting against our best end-to-end configuration.

Setting	Recall@1	Recall@5	Recall@10	MRR	mIoU
End-to-end (best)	72.65	78.46	78.99	0.752	0.3869
Oracle	85.43	94.89	96.12	0.896	0.5384

Table 10: Linking performance with perfect (gold) span detection compared to end-to-end.

With perfect span detection, KRISSBERT achieves a macro-IoU of 0.5384 and Recall@10 of 96.1% which is a substantial improvement over the end-to-end system. Moreover, the linker retrieves the correct concept for the majority of mentions within the top candidates. However, the gap between Recall@1 (85.4%) and Recall@5 (94.9%) indicates that the primary challenge lies in disambiguation among semantically similar candidates. This observation is consistent with the candidate ranking analysis in Table 9. The remaining approximately 4% of mentions for which the gold concept is not retrieved within the top 10 candidates represents an upper bound on retrieval performance.