OntoEL: Neuro-Symbolic Biomedical Entity Linking with Differentiable Fuzzy $\mathcal{EL}_\bot$ Reasoning

Anonymous Author(s)

Abstract

Current neural biomedical entity linking (BioEL) models treat ontologies as flat dictionaries, ignoring the rich terminological knowledge (TBox) that defines concept boundaries. Consequently, they struggle with contextual ambiguity, often retrieving logically inconsistent candidates based solely on surface similarity. We present **OntoEL**, a neuro-symbolic framework that shifts BioEL from surface-level matching to logic-grounded reasoning. OntoEL integrates differentiable fuzzy $\mathcal{EL}_\bot$ reasoning into the retrieval pipeline as a consistency-aware re-ranker, employing a hybrid strategy: structural TBox reasoning is delegated to classical polynomial-time reasoners, while the **sigmoidal Reichenbach implication** performs soft type-consistency evaluation—effectively resolving the "implication bias" gradient pathology in previous neuro-symbolic methods. By enforcing ontological axioms as differentiable soft constraints, OntoEL aligns neural representations with logical truth. Comprehensive experiments on three benchmarks (MedMentions, BC5CDR, and NCBI Disease) demonstrate state-of-the-art performance, surpassing strong baselines by up to 4.2% in Accuracy@1. On highly ambiguous mentions requiring ontological reasoning, our method yields 5.7% improvement, proving the efficacy of incorporating logical semantics into neural retrieval.

To support reproducibility, we make our source code, datasets, and hyperparameter settings publicly available for review at http://github.com/anonymous-ai-researcher/OntoEL. This repository also hosts an extended version of this submission, featuring complete proofs of the theorems and supplementary experimental analyses.

CCS Concepts

• **Information systems** → **Learning to rank**; **Top-k retrieval in databases**; **Thesauri**; **Ontologies**.

Keywords

Biomedical Entity Linking, Ontology-Aware Retrieval, Logical Re-ranking, Neuro-Symbolic AI, Description Logic

1 Introduction

Biomedical entity linking (BioEL) [10, 24, 81, 103]—the task of mapping clinical mentions in unstructured text to standardized concepts in biomedical ontologies such as SNOMED CT [70] and the Gene Ontology [5]—constitutes a foundational capability for biomedical information retrieval [33, 44, 46, 48, 56, 84], knowledge base population [23, 41, 55, 63], and clinical decision support [2, 25, 57]. The proliferation of electronic health records and biomedical literature has rendered accurate entity linking indispensable: downstream applications ranging from drug-drug interaction discovery to cohort identification depend critically on the fidelity of concept normalization [12, 21, 50, 67, 100, 101, 104]. At its core, BioEL operates as a large-scale retrieval task: given a mention and its context, the system must efficiently identify the correct concept from a search space often containing hundreds of thousands of candidate entities.

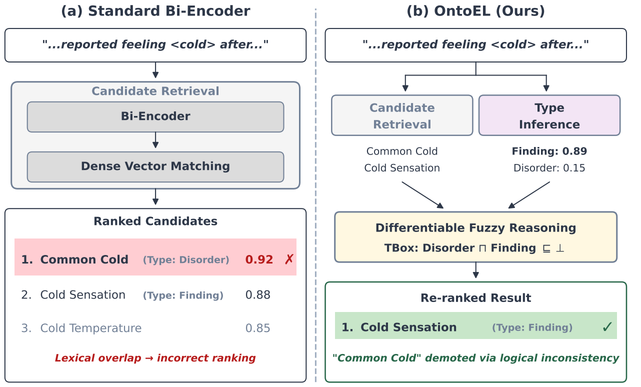
(a) Standard Bi-Encoder
"...reported feeling <cold> after..."
Candidate Retrieval
Bi-Encoder
Dense Vector Matching
Ranked Candidates
1. Common Cold (Type: Disorder) 0.92 ✗
2. Cold Sensation (Type: Finding) 0.88
3. Cold Temperature 0.85
Lexical overlap → incorrect ranking
(b) OntoEL (Ours)
"...reported feeling <cold> after..."
Candidate Retrieval
Type Inference
Common Cold
Cold Sensation
Finding: 0.89
Disorder: 0.15
Differentiable Fuzzy Reasoning
TBox: Disorder ⊓ Finding ⊑ ⊥
Re-ranked Result
1. Cold Sensation (Type: Finding) ✓
"Common Cold" demoted via logical inconsistency

Figure 1: Surface Similarity vs. Logical Consistency. (Left) Neural bi-encoders rank by vector similarity, often placing *Common Cold* (Disorder) above the correct *Cold Sensation* (Finding) due to lexical overlap. (Right) OntoEL infers context-aware type constraints (Finding: 0.89) and enforces *Disorder* $\sqcap$ *Finding* $\sqsubseteq \bot$ via differentiable fuzzy reasoning, correctly promoting the consistent candidate.

Neural Approaches to BioEL. Existing BioEL approaches have shifted toward dense retrieval paradigms built upon pretrained language models [39, 51, 74, 75, 77, 83, 93, 94, 97]. State-of-the-art methods such as SapBERT [45] encode mentions and candidates into a shared vector space for nearest-neighbor retrieval [16, 30, 53]. More recent advances have further enhanced this architecture through contrastive learning [66], cross-encoder re-ranking [22, 87, 90], and generative sequence-to-sequence modeling [35, 92]. Complementary research on knowledge graph completion has explored translational embeddings [13, 42, 73], region-based representations [1, 98, 99], and graph neural network architectures [36, 64, 65, 82]—see [85] for a comprehensive survey. While LLMs have been explored for BioEL [12, 43, 63], their latency and tendency to hallucinate non-existent identifiers limit clinical applicability.

However, these methods fundamentally treat biomedical ontologies as flat dictionaries, ignoring the rich logical axioms encoded in the TBox that define concept boundaries [15, 27]. Consequently, neural bi-encoders often struggle with contextual ambiguity, retrieving high-similarity candidates that are plausible on the surface but logically inconsistent with the mention's implied type. For instance, as illustrated in Figure 1, a bi-encoder might link the mention "cold" (referring to a sensation) to the concept *Common Cold* (a Viral Infection) simply due to high lexical overlap, ignoring the ontological distinction between *Finding* and *Disorder*. Rigorous evaluation studies have confirmed these systematic failure modes [4, 102].

The Ontology Structure Gap. Ontologies like SNOMED CT are formal knowledge bases expressed in Description Logics [7], comprising hundreds of thousands of axioms specifying subsumption hierarchies (e.g., *Viral Pneumonia* $\sqsubseteq$ *Pneumonia*) and role restrictions (e.g., *Pneumonia* $\sqsubseteq \exists$*hasSite.Lung*). This structural knowledge is what human experts leverage to disambiguate mentions: distinguishing a procedure from a disorder requires reasoning about ontological properties rather than surface forms.

The ontology embedding community has developed geometric representations to capture such structure [15], including n-ball embeddings [38], box embeddings [29, 91], and hyperbolic geometry [52]. However, these methods target link prediction within ontologies [14, 54, 61, 68, 69, 89]—modeling $P(t|h,r)$—rather than entity linking from text, which requires $P(e|\text{mention context})$. They lack mechanisms to fuse text encoders with logical embeddings for mention disambiguation.

Neuro-Symbolic Re-Ranking. To bridge this gap, we propose to operationalize Description Logic (DL) semantics [7] directly within the neural pipeline using Fuzzy Logic [28, 37, 71, 96], which, unlike classical logic, allows for graded truth values, making it compatible with the continuous representations of neural networks. Existing neuro-symbolic frameworks such as LTN [9, 20], DFL [80], and SBR [18] have pioneered this direction [11, 17]. However, these methods ground first-order logic (FOL) formulas into continuous spaces, but FOL is undecidable [78], forcing any sound reasoning procedure into potential non-termination. Moreover, the choice of fuzzy operators often leads to *implication bias* [80], a gradient pathology causing vanishing gradients or reasoning shortcuts during training [47, 88]. In contrast, our approach targets the DL $\mathcal{EL}_\perp$, a decidable syntactic fragment of FOL that admits polynomial-time reasoning [6] while remaining expressive enough to capture the hierarchical structure of large-scale biomedical ontologies.

In this paper, we introduce OntoEL, a neuro-symbolic framework that operationalizes differentiable fuzzy $\mathcal{EL}_\perp$ reasoning to bridge the gap between neural retrieval and logical consistency. We cast BioEL as a *retrieve-then-reason* process: OntoEL first leverages a standard neural bi-encoder (e.g., SapBERT) to generate semantic candidates, which are subsequently re-ranked by a logic module that evaluates their consistency against context-inferred constraints. A critical contribution of our work is the resolution of the *implication bias*—a well-known gradient pathology in fuzzy logic that hinders effective learning. By adopting the **sigmoidal Reichenbach implication**, we ensure stable, non-vanishing gradient flow. This design allows ontological axioms to function as differentiable soft constraints, actively refining the neural representation space via end-to-end backpropagation.

Contributions. Our work makes the following contributions:

- **Neuro-Symbolic Retrieval Framework:** We propose OntoEL, the first BioEL framework that utilizes TBox axioms as differentiable ranking signals, transforming static ontological knowledge into dynamic guidance for resolving entity ambiguity.
- **Gradient-Stable Logical Reasoning:** We tackle the implication bias problem in differentiable logic by analyzing operator dynamics in deep networks. We implement a Product-Reichenbach logic layer that ensures robust gradient propagation, enabling effective training on ontologies with complex hierarchical constraints.
- **Empirical Validation:** Extensive experiments on three benchmarks (MedMentions, BC5CDR, and NCBI Disease) demonstrate that OntoEL achieves state-of-the-art retrieval accuracy. Notably, it yields substantial improvements on ambiguous queries (up to 4.2% on MedMentions and 5.7% on hard ambiguity sets) where ontological reasoning provides the greatest disambiguation power, while maintaining the efficiency of bi-encoder retrieval.

2 Preliminaries

2.1 The Description Logic $\mathcal{EL}_\perp$

Description logics (DLs) constitute a family of knowledge representation languages designed as syntactic fragments of FOL, balancing expressivity with decidability and computational tractability [7]. Among the various DL dialects, the $\mathcal{EL}$ family [6], including $\mathcal{EL}_\perp$, occupies a unique position: unlike more expressive dialects such as $\mathcal{ALC}$ (ExpTime-complete) and $\mathcal{SRIQ}$ (NExpTime-hard), $\mathcal{EL}_\perp$ maintains polynomial-time reasoning complexity through carefully designed syntactic restrictions.

2.1.1 Syntax. Let $\Sigma = (\mathsf{N_C}, \mathsf{N_R})$ be a signature comprising disjoint countably infinite sets of **concept** and **role** names, respectively. $\mathcal{EL}_\perp$ **concepts** (or simply **concepts**) are defined inductively:

$$C, D ::= \top \mid \perp \mid A \mid C \sqcap D \mid \exists r.C$$

where $A \in \mathsf{N_C}$ and $r \in \mathsf{N_R}$. We distinguish between **atomic** (concept name) and **complex** concepts. To ensure polynomial tractability, $\mathcal{EL}_\perp$ excludes negation ($\neg C$), disjunction ($C \sqcup D$), and universal restriction ($\forall r.C$) [76]. Limited negative information is supported via the bottom concept $\perp$ in disjointness axioms (e.g., $C \sqcap D \sqsubseteq \perp$).

An $\mathcal{EL}_\perp$ **TBox** $\mathcal{T}$ is a finite set of axioms of the form $C \sqsubseteq D$ (namely **concept inclusion**), stating that every instance of C is also an instance of D. Concept inclusions encode the subsumption hierarchies and disjointness constraints essential for disambiguation in BioEL. We denote by $\text{sig}(\mathcal{T})$ the signature of $\mathcal{T}$, i.e., the set of concept and role names appearing in $\mathcal{T}$.

2.1.2 Semantics. Semantics are defined using an **interpretation** $\mathcal{I} = (\Delta^\mathcal{I}, \cdot^\mathcal{I})$, consisting of a non-empty domain $\Delta^\mathcal{I}$ and an interpretation function $\cdot^\mathcal{I}$ mapping $A \in \mathsf{N_C}$ to $A^\mathcal{I} \subseteq \Delta^\mathcal{I}$ and $r \in \mathsf{N_R}$ to $r^\mathcal{I} \subseteq \Delta^\mathcal{I} \times \Delta^\mathcal{I}$. The function $\cdot^\mathcal{I}$ extends to complex concepts as:

$$\top^\mathcal{I} = \Delta^\mathcal{I} \quad \perp^\mathcal{I} = \emptyset \quad (C \sqcap D)^\mathcal{I} = C^\mathcal{I} \cap D^\mathcal{I}$$
$$(\exists r.C)^\mathcal{I} = \{d \in \Delta^\mathcal{I} \mid \exists e \in \Delta^\mathcal{I} : (d,e) \in r^\mathcal{I} \wedge e \in C^\mathcal{I}\}$$

An interpretation $\mathcal{I}$ **satisfies** a concept inclusion $C \sqsubseteq D$ iff $C^\mathcal{I} \subseteq D^\mathcal{I}$. $\mathcal{I}$ is a **model** of $\mathcal{T}$ iff it satisfies all axioms in $\mathcal{T}$. $C \sqsubseteq D$ is **entailed** by $\mathcal{T}$ ($\mathcal{T} \models C \sqsubseteq D$) iff it is satisfied by every model of $\mathcal{T}$.

2.1.3 Reasoning Complexity. The fundamental reasoning task, *subsumption checking* (determining if $\mathcal{T} \models C \sqsubseteq D$), is PTime-complete for $\mathcal{EL}_\perp$ [6], ensuring efficiency for large-scale ontologies.

2.2 Problem Formulation

We formalize BioEL by progressing from the knowledge base structure to the logical abstraction and finally the task definition.

Mentions, Surface Forms, and Context. A *mention* m is a document-anchored text span, modeled as a triple $m = (d, i, j)$ where $d[i{:}j]$ is the *surface form* $\mathrm{sf}(m)$. The *context* c consists of surrounding text providing disambiguating evidence.

▹ Example (Mention, Surface Form, and Context)

In "*the patient reported feeling cold after the procedure*", a mention m is the span "cold" with $\mathrm{sf}(m)$ = "cold" and context c = "the patient reported feeling ___ after the procedure".

Biomedical Knowledge Base and Entities. Large-scale biomedical resources such as the Unified Medical Language System (UMLS) [86] are not ontologies in the formal DL sense, but integrative *knowledge bases* that unify heterogeneous biomedical vocabularies through *Concept Unique Identifiers* (CUIs). Each CUI denotes a single abstract biomedical concept and acts as a stable hub linking multiple lexical entries from different source vocabularies. For instance, CUI C0009443 ("common cold") aggregates variants such as "Common cold (disorder)" in SNOMED CT, "Common Cold" in MeSH, and "Acute nasopharyngitis [common cold]" in ICD-10.

An **entity** e represents a unique abstract concept (CUI) within $\mathcal{K}$. The task is defined over a finite **target entity set** $\mathcal{E}$, constructed by restricting $\mathcal{K}$ to entities appearing in the dataset's gold annotations (e.g., $|\mathcal{E}| \approx 35,000$ for MedMentions ST21pv).

Lexical Realizations. Each entity $e \in \mathcal{E}$ is associated with a set of lexical surface forms in $\mathcal{K}$: a unique **preferred name** $\mathrm{name}(e)$ and a set of **synonyms** $\mathrm{syn}(e)$. These provide the textual representation for neural encoding. For e = C0009443, $\mathrm{name}(e)$ = "Common Cold" and $\mathrm{syn}(e)$ = {"Acute coryza", "Head cold"}.

Logical Abstraction. We treat the knowledge base as an external resource $\mathcal{K}$, from which we extract an $\mathcal{EL}_\bot$ TBox $\mathcal{T}$ consisting of concept inclusion axioms that capture terminological constraints for reasoning. As defined earlier, $\mathcal{T}$ is the ontology in the DL sense. We establish a bijective mapping where each entity $e \in \mathcal{E}$ corresponds to a unique **concept name** $A \in \mathrm{sig}(\mathcal{T}) \cap \mathsf{N_C}$. **Modeling Convention:** throughout this paper, we denote entities using their preferred names (e.g., CommonCold) rather than opaque CUIs to explicitly link them to the TBox $\mathcal{T}$.

Semantic Types and Ontological Constraints. In addition to lexical information, entities are associated with high-level *semantic types*. Let $\Gamma \subseteq \mathrm{sig}(\mathcal{T})$ denote a finite set of semantic type concept names used for disambiguation. Semantic types are grounded in the ontology via logical entailment:

$$\mathrm{type}(e) = \{\tau \in \Gamma \mid \mathcal{T} \models e \sqsubseteq \tau\}. \tag{1}$$

Thus, $\mathrm{type}(e)$ is determined strictly by the axioms in $\mathcal{T}$. For instance, $\mathcal{T} \models \mathsf{CommonCold} \sqsubseteq \mathsf{Disease}$ and $\mathcal{T} \models \mathsf{ColdSensation} \sqsubseteq \mathsf{Finding}$. Logic-based disambiguation relies on disjointness axioms (e.g., $\mathsf{Disease} \sqcap \mathsf{Finding} \sqsubseteq \bot$) to resolve conflicts between context-inferred types and candidate types.

The ***BioEL*** *Task.* Given a mention m with context c, the BioEL task is to identify the correct entity $e^* \in \mathcal{E}$ referred to by m:

$$e^* = \arg\max_{e \in \mathcal{E}} P(e \mid m, c, \mathcal{K}). \tag{2}$$

Limitations of Purely Neural Approaches. Most existing BioEL systems like SapBERT [45] employ neural bi-encoders that approximate the linking objective via embedding similarity:

$$e^* \approx \arg\max_{e \in \mathcal{E}} \mathrm{sim}\big(\mathrm{Enc}(m, c), \mathrm{Enc}(\mathrm{name}(e))\big). \tag{3}$$

where $\mathrm{Enc}(\cdot)$ denotes a neural encoder and $\mathrm{sim}(\cdot,\cdot)$ is often cosine similarity or inner product. This formulation treats the knowledge base as a flat collection of surface forms (flat dictionary) and ignores the ontological constraints encoded in the TBox $\mathcal{T}$.

3 Differentiable Fuzzy Semantics for $\mathcal{EL}_\bot$

Classical $\mathcal{EL}_\bot$ relies on crisp semantics where concept membership is strictly binary. While mathematically elegant and computationally tractable, this rigid discreteness effectively blocks the gradient flow required for neural optimization. To bridge this gap, we extend $\mathcal{EL}_\bot$ to a differentiable fuzzy setting, mapping logical predicates to continuous membership functions in $[0, 1]$.

3.1 Fuzzy Interpretation

The semantics of fuzzy $\mathcal{EL}_\bot$ is defined via *fuzzy interpretations*, rooted in Zadeh's fuzzy set theory [96].

Definition 1 (Fuzzy Interpretation). *A* ***fuzzy interpretation*** $\mathcal{I} = (\Delta^\mathcal{I}, \cdot^\mathcal{I})$ *consists of a non-empty domain* $\Delta^\mathcal{I}$ *and a fuzzy interpretation function* $\cdot^\mathcal{I}$ *that assigns: to each concept name* $A \in \mathsf{N_C}$ *a membership function* $A^\mathcal{I} : \Delta^\mathcal{I} \to [0, 1]$ *and to each role name* $r \in \mathsf{N_R}$ *a fuzzy relation* $r^\mathcal{I} : \Delta^\mathcal{I} \times \Delta^\mathcal{I} \to [0, 1]$.

For an element $d \in \Delta^\mathcal{I}$, the value $A^\mathcal{I}(d) \in [0, 1]$ represents the *degree* to which d belongs to concept A, with boundary cases $A^\mathcal{I}(d) = 1$ and $A^\mathcal{I}(d) = 0$ corresponding to full membership and non-membership, respectively.

3.2 Fuzzy Operator Selection

The challenge in fuzzy semantics is extending Boolean operations to the continuous interval $[0, 1]$ while preserving logical properties for sound reasoning. Following van Krieken et al. [80], we require fuzzy operators that reduce to classical logic on crisp inputs, maintain differentiability, and exhibit stable gradient behavior.

Scope of Fuzzy Extension. While a complete fuzzy semantics for $\mathcal{EL}_\bot$ would require defining fuzzy operators for conjunction ($\sqcap$) and existential restriction ($\exists r.C$), we deliberately focus on the **fuzzy implication** operator. This design choice reflects the BioEL task structure: semantic ambiguity primarily arises at the *atomic type level* (e.g., distinguishing Disorder from Finding), which is fundamentally an implication problem—"does the candidate satisfy the context-implied type τ?" Meanwhile, structural reasoning involving conjunction and existential restrictions can be efficiently precomputed by classical $\mathcal{EL}_\bot$ reasoners (e.g., ELK [34]), yielding crisp type memberships for candidates. This hybrid strategy reserves fuzzy semantics for the implication operator, precisely where existing neuro-symbolic methods fail due to the *implication bias* problem [80]—a gradient pathology we aim to resolve.

3.2.1 Fuzzy Implication for Subsumption. The most critical choice for BioEL is the **fuzzy implication** [79], which determines how we evaluate subsumptions $C \sqsubseteq D$—measuring the degree to which "if d belongs to C, then d belongs to D" holds across the domain.

The algebraically "correct" choice for a t-norm is its residuum (R-implication): $I_T(a,b) = \sup\{c \mid T(a,c) \leq b\}$. For Product, this yields the **Goguen implication** [8]: $I_{GG}(a,b) = 1$ if $a \leq b$, else b/a. While theoretically sound, Goguen suffers from *implication bias*—returning 1 whenever $a \leq b$ regardless of actual values, providing no gradient signal—and numerical instability as $a \to 0$.

An alternative constructs implications via $a \to b \equiv \neg a \vee b$, yielding $I_S(a,b) = S(1-a,b)$ (S-implication). Using Probabilistic Sum produces the **Reichenbach implication** [59]: $I_R(a,b) = 1-a+ab$. This eliminates implication bias (the formula does not saturate when a is small), provides symmetric gradients ($\frac{\partial I_R}{\partial a} = b-1 \leq 0$, $\frac{\partial I_R}{\partial b} = a \geq 0$), and ensures numerical stability without division.

The "Hard Negative" Problem. However, Reichenbach is linear, which poses challenges for ranking tasks. Consider discriminating between a *hard negative* (high neural similarity a=0.9, type-inconsistent b=0.1, giving I_R=0.19) and a *hard positive* (moderate similarity a=0.5, type-consistent b=0.9, giving I_R=0.95). When fused with strong neural scores, this linear gap may not override high-confidence false positives. Ranking requires *amplified discrimination* near the decision boundary.

Our Choice: Sigmoidal Reichenbach. Following van Krieken et al. [80], we adopt:

$$I_\sigma(a,b) = \sigma\big(s \cdot (1 - a + ab - 0.5)\big) \tag{4}$$

where σ is the sigmoid and $s > 0$ controls sharpness. The sigmoid concentrates gradient mass near the decision boundary ($I_R \approx 0.5$), amplifying penalties for ambiguous violations while dampening trivial cases. Type-inconsistent candidates receive exponentially lower scores, enabling ontological signals to override strong neural similarities. The bounded output facilitates stable score fusion, and empirically, we observe that $s \in [5, 10]$ effectively balances smooth optimization with sharp logical behavior. The gradients $\frac{\partial I_\sigma}{\partial a} = s \cdot \sigma' \cdot (b-1) \leq 0$ and $\frac{\partial I_\sigma}{\partial b} = s \cdot \sigma' \cdot a \geq 0$ (where $\sigma'(z) = \sigma(z)(1-\sigma(z)) = I_\sigma(1 - I_\sigma)$) confirm the expected behavior: penalizing confident inconsistency and rewarding type consistency.

Theorem 1 (Gradient Non-Degeneracy). *The Sigmoidal Reichenbach implication I_σ resolves the implication bias problem:*

- ***(Degeneracy)*** *The Goguen implication $I_{GG}(a,b) = \min(1, b/a)$ yields $\nabla I_{GG} = \mathbf{0}$ for all logically consistent pairs ($a \leq b$), effectively halting gradient-based optimization in 50% of the input space.*
- ***(Positivity)*** *I_σ maintains $\|\nabla I_\sigma(a,b)\| > 0$ for all $(a,b) \in (0,1)^2$, ensuring continuous gradient flow.*
- ***(Discrimination)*** *For hard negatives $(0.9, 0.1)$ vs. hard positives $(0.5, 0.9)$, the discrimination ratio grows as $\sim e^{0.31s}$. This yields a ratio of $\sim$23 for s=10 and $\sim$500 for s=20, providing an exponential advantage over the constant ratio (5) of linear I_R.*

3.3 Fuzzy Semantics of $\mathcal{EL}_\perp$ Concepts

Building upon Definition 1, we specify the semantics of the boundary concepts: $\top^\mathcal{I}(d) = 1$ and $\perp^\mathcal{I}(d) = 0$ for all $d \in \Delta^\mathcal{I}$. Complex concept expressions (conjunction $\sqcap$ and existential restriction $\exists r.C$) are delegated to classical $\mathcal{EL}_\perp$ reasoners as discussed in Section 3.

3.4 Fuzzy Axiom Satisfaction

A fuzzy interpretation $\mathcal{I}$ satisfies an axiom α to degree $n \in [0,1]$, written $\mathcal{I} \models_n \alpha$. For TBox axioms carrying implicit universal quantification, the satisfaction degree is the *infimum* of pointwise implication values, capturing the "weakest link" principle [28].

Definition 2 (Fuzzy TBox Satisfaction).

$$\mathcal{I} \models_n C \sqsubseteq D \;\; \textit{iff} \;\; n = \inf_{d \in \Delta^\mathcal{I}} I_\sigma\big(C^\mathcal{I}(d), D^\mathcal{I}(d)\big)$$

A fuzzy interpretation $\mathcal{I}$ is a **fuzzy model** of TBox $\mathcal{T}$ to degree n (written $\mathcal{I} \models_n \mathcal{T}$) iff $n = \inf_{\alpha \in \mathcal{T}}\{m \mid \mathcal{I} \models_m \alpha\}$. An axiom α is **entailed** by $\mathcal{T}$ to degree n (written $\mathcal{T} \models_n \alpha$) iff $n = \inf_\mathcal{I}\{m \mid \mathcal{I} \models_m \alpha$ and $\mathcal{I} \models_k \mathcal{T}$ for some $k > 0\}$.

3.5 Theoretical Guarantees

We establish that our fuzzy extension preserves the semantic properties of classical $\mathcal{EL}_\perp$.

Theorem 2 (Semantic Soundness). *Under the Sigmoidal Reichenbach implication, the fuzzy semantics is a conservative extension of classical $\mathcal{EL}_\perp$:*

- ***(Soundness)*** *Classical entailment implies maximal fuzzy entailment: $\mathcal{T} \models C \sqsubseteq D \Rightarrow \mathcal{T} \models_1 C \sqsubseteq D$.*
- ***(Boundary Preservation)*** *On crisp inputs $\{0,1\}$, the fuzzy implication reduces to classical Boolean implication.*

4 The OntoEL Framework

In this section, we present OntoEL, a neuro-symbolic framework that integrates differentiable fuzzy $\mathcal{EL}_\perp$ reasoning into BioEL. OntoEL operates in two stages: (1) *Neural Candidate Retrieval*, which leverages a pretrained bi-encoder to retrieve an initial candidate set via approximate nearest neighbor search (e.g., FAISS [32]), and (2) *Ontological Re-ranking*, which employs the fuzzy reasoning machinery developed in Section 3 to re-score candidates based on their logical consistency with context-inferred type constraints.

Figure 2 illustrates the complete computation pipeline. A key design principle is the separation of *fuzzy* inference (for context-dependent type prediction) from *crisp* reasoning (for TBox-level entailment): this hybrid strategy combines the flexibility of neural type inference with the precision of classical ontological reasoning, while maintaining computational efficiency through offline precomputation of TBox entailments.

4.1 Neural Instantiation of Fuzzy Semantics

We now detail the neural instantiation corresponding to Steps 4–6 in Figure 2. The core challenge is bridging continuous embeddings and discrete logic through three components: (1) inferring fuzzy membership degrees $\tau^\mathcal{I}(m)$ for mentions based on contextual evidence; (2) deriving crisp memberships $\tau^\mathcal{I}(e)$ for candidates via TBox entailment; and (3) computing differentiable consistency scores to quantify logical alignment.

4.1.1 Mention and Concept Encoding. Let $\text{Enc}(\cdot)$ denote a pretrained biomedical language model (e.g., SapBERT [45]). For a mention m occurring in context c, we obtain the mention embedding $\mathbf{m} = \text{Enc}([c_{\text{left}}; m; c_{\text{right}}]) \in \mathbb{R}^d$, where $[\cdot;\cdot;\cdot]$ denotes concatenation and d is the hidden dimension. For each candidate concept

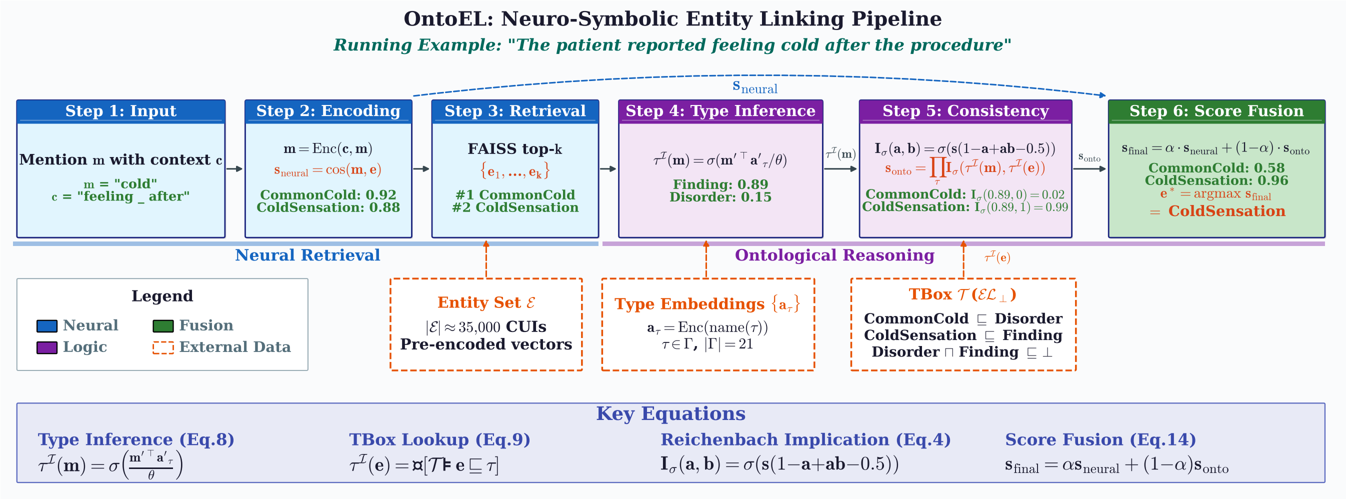
OntoEL: Neuro-Symbolic Entity Linking Pipeline
Running Example: "The patient reported feeling cold after the procedure"
$\mathbf{s}_{\text{neural}}$
Step 1: Input
Mention m with context c
m = "cold"
c = "feeling _ after"
Step 2: Encoding
$\mathbf{m} = \text{Enc}(\mathbf{c}, \mathbf{m})$
$\mathbf{s}_{\text{neural}} = \cos(\mathbf{m}, \mathbf{e})$
CommonCold: 0.92
ColdSensation: 0.88
Step 3: Retrieval
FAISS top-k
$\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$
#1 CommonCold
#2 ColdSensation
Step 4: Type Inference
$\tau^{\mathcal{I}}(\mathbf{m}) = \sigma(\mathbf{m}'^\top \mathbf{a}'_\tau/\theta)$
Finding: 0.89
Disorder: 0.15
$\tau^{\mathcal{I}}(\mathbf{m})$
Step 5: Consistency
$\mathbf{I}_\sigma(\mathbf{a}, \mathbf{b}) = \sigma(\mathbf{s}(1 - \mathbf{a} + \mathbf{ab} - 0.5))$
$\mathbf{s}_{\text{onto}} = \prod_\tau \mathbf{I}_\sigma(\tau^{\mathcal{I}}(\mathbf{m}), \tau^{\mathcal{I}}(\mathbf{e}))$
CommonCold: $\mathbf{I}_\sigma(0.89, 0) = 0.02$
ColdSensation: $\mathbf{I}_\sigma(0.89, 1) = 0.99$
$\mathbf{s}_{\text{onto}}$
Step 6: Score Fusion
$\mathbf{s}_{\text{final}} = \alpha \cdot \mathbf{s}_{\text{neural}} + (1-\alpha) \cdot \mathbf{s}_{\text{onto}}$
CommonCold: 0.58
ColdSensation: 0.96
$\mathbf{e}^* = \text{argmax } \mathbf{s}_{\text{final}}$
= ColdSensation
Neural Retrieval
Ontological Reasoning
$\tau^{\mathcal{I}}(\mathbf{e})$
Legend
Neural
Fusion
Logic
External Data
Entity Set $\mathcal{E}$
$|\mathcal{E}| \approx 35{,}000$ CUIs
Pre-encoded vectors
Type Embeddings $\{\mathbf{a}_\tau\}$
$\mathbf{a}_\tau = \text{Enc}(\text{name}(\tau))$
$\tau \in \Gamma$, $|\Gamma| = 21$
TBox $\mathcal{T}(\mathcal{EL}_\perp)$
CommonCold $\sqsubseteq$ Disorder
ColdSensation $\sqsubseteq$ Finding
Disorder $\sqcap$ Finding $\sqsubseteq \perp$
Key Equations
Type Inference (Eq.8)
$\tau^{\mathcal{I}}(\mathbf{m}) = \sigma\left(\frac{\mathbf{m}'^\top \mathbf{a}'_\tau}{\theta}\right)$
TBox Lookup (Eq.9)
$\tau^{\mathcal{I}}(\mathbf{e}) = ¤[\mathcal{T} \models \mathbf{e} \sqsubseteq \tau]$
Reichenbach Implication (Eq.4)
$\mathbf{I}_\sigma(\mathbf{a}, \mathbf{b}) = \sigma(\mathbf{s}(1 - \mathbf{a} + \mathbf{ab} - 0.5))$
Score Fusion (Eq.14)
$\mathbf{s}_{\text{final}} = \alpha \mathbf{s}_{\text{neural}} + (1-\alpha)\mathbf{s}_{\text{onto}}$

Figure 2: The OntoEL Pipeline. The framework operates in two stages with eight steps. Stage 1 (Neural Retrieval): mention encoding (Steps 1–2) and candidate retrieval (Step 3). Stage 2 (Ontological Re-ranking): context-aware type inference yielding fuzzy memberships $\tau^{\mathcal{I}}(m)$ (Step 4), TBox lookup yielding crisp memberships $\tau^{\mathcal{I}}(e)$ (Step 4), consistency scoring via Sigmoidal Reichenbach implication (Step 5), score fusion (Step 6), and final output (Step 6).

$e \in \mathcal{E}$, we encode its preferred name: $\mathbf{e} = \text{Enc}(\text{name}(e)) \in \mathbb{R}^d$. The neural similarity score is defined as:

$$s_{\text{neural}}(m, e) = \frac{\mathbf{m}^\top \mathbf{e}}{\|\mathbf{m}\| \cdot \|\mathbf{e}\|}. \tag{5}$$

4.1.2 *Context-Aware Type Inference.* A key requirement for ontology-aware re-ranking is estimating the degree to which a mention m, given its context, refers to an entity of semantic type $\tau \in \Gamma$. Traditional multi-label classifiers require a fixed label set and fail to generalize to the thousands of fine-grained types in biomedical ontologies. We instead leverage the semantic richness of the encoder's latent space through a *projected name-based* approach.

Type Name Encoding. For each semantic type $\tau \in \Gamma$, we encode its preferred name to obtain the type embedding:

$$\mathbf{a}_\tau = \text{Enc}(\text{name}(\tau)) \in \mathbb{R}^d \tag{6}$$

(e.g., $\mathbf{a}_{\text{Finding}} = \text{Enc}(\text{“Clinical Finding”})$). These embeddings inherit the rich semantic knowledge captured during pretraining, enabling zero-shot generalization to types unseen during training.

Dual Projection. Raw dot-product between mention and type name embeddings conflates two distinct semantic tasks: mention-to-entity matching (trained for retrieval) and mention-to-type inference (required for re-ranking). To decouple these, the mention and type embeddings are projected into a shared logic space via two learnable matrices $\mathbf{W}_m \in \mathbb{R}^{d' \times d}$ and $\mathbf{W}_t \in \mathbb{R}^{d' \times d}$:

$$\mathbf{m}' = \mathbf{W}_m \mathbf{m}, \quad \mathbf{a}'_\tau = \mathbf{W}_t \mathbf{a}_\tau \tag{7}$$

where $\mathbf{W}_m$ and $\mathbf{W}_t$ are optimized during training to align the type boundaries with mention representations, while the base type embeddings $\{\mathbf{a}_\tau\}$ remain fixed.

Fuzzy Membership Computation. The fuzzy membership degree of mention m with respect to type τ is computed as:

$$\tau^{\mathcal{I}}(m) = \sigma\left(\frac{\mathbf{m}'^\top \mathbf{a}'_\tau}{\theta}\right) \tag{8}$$

where σ is the sigmoid function and $\theta = \exp(\hat{\theta})$ is a learnable temperature parameter, initialized to $\log \sqrt{d'}$. This parameterization ensures $\theta > 0$ and allows the model to dynamically adapt the sharpness of the type membership distribution during training—using broader distributions for exploration in early stages and sharper distributions for discrimination as training progresses [58]. The sigmoid ensures that $\tau^{\mathcal{I}}(m) \in [0, 1]$, directly satisfying the fuzzy membership requirement of Definition 1.

Design Rationale. This architecture offers three key advantages:

- **Zero-Shot Generalization:** By encoding type names rather than learning fixed type embeddings, the model can infer membership for rare or unseen types based on semantic similarity.
- **Task-Specific Adaptation:** The dual projections $\mathbf{W}_m$ and $\mathbf{W}_t$ allow the model to learn a type-inference space distinct from the retrieval space.
- **Semantic-Logic Alignment:** The projections act as a differentiable bridge, enabling ontological structures (TBox) to actively reshape the semantic manifold of the encoder during training.

4.1.3 *Candidate Type Membership.* For candidate concepts, type membership is determined by the TBox rather than inferred. Given a candidate $e \in \mathcal{E}$ and a semantic type $\tau \in \Gamma$:

$$\tau^{\mathcal{I}}(e) = \begin{cases} 1 & \text{if } \mathcal{T} \models e \sqsubseteq \tau \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

While our reasoning framework operates in a continuous fuzzy space, we treat background ontology as authoritative ground truth.

This crisp assignment ensures that TBox entailments, such as transitivity, are strictly enforced. For instance, if $\mathcal{T} = \{\text{CommonCold} \sqsubseteq \text{ViralInfection}, \text{ViralInfection} \sqsubseteq \text{Disorder}\}$, then $\tau^{\mathcal{I}}(\text{CommonCold}) = 1$ for $\tau = \text{Disorder}$ by transitivity. These assignments can be precomputed offline using a DL reasoner (e.g., ELK [34]), which computes the full subsumption hierarchy in polynomial time for $\mathcal{EL}_\perp$.

We intentionally treat candidate type memberships as crisp to preserve the ontology's role as authoritative ground truth. This design ensures that all uncertainty is localized to the context-based type inference, while ontological facts provide a strict "logical wall" that prevents the neural component from learning shortcuts that violate TBox constraints [47].

4.1.4 Fuzzy Consistency Scoring. With membership functions instantiated, we compute the consistency score between a mention m and a candidate e. A candidate is consistent if the types inferred from context align with the candidate's actual types in the ontology.

For each semantic type $\tau \in \Gamma$, we evaluate type constraint satisfaction using the Sigmoidal Reichenbach implication (Eq. (4)):

$$\mathrm{cons}_\tau(m, e) = I_\sigma\big(\tau^{\mathcal{I}}(m), \tau^{\mathcal{I}}(e)\big) \tag{10}$$

This captures the logical intuition: if context strongly suggests type τ (high $\tau^{\mathcal{I}}(m)$) but the candidate does not belong to τ ($\tau^{\mathcal{I}}(e) = 0$), the consistency score is penalized.

The overall ontological consistency score aggregates across all types via multiplication:

$$s_{\text{onto}}(m, e) = \prod_{\tau \in \Gamma} \mathrm{cons}_\tau(m, e) \tag{11}$$

For numerical stability, we compute in log-space [50]:

$$\log s_{\text{onto}}(m, e) = \sum_{\tau \in \Gamma} \log I_\sigma\big(\tau^{\mathcal{I}}(m), \tau^{\mathcal{I}}(e)\big) \tag{12}$$

4.1.5 Score Fusion and Re-ranking. The final ranking score combines neural similarity and ontological consistency. To ensure commensurable scales, we normalize the neural similarity to $[0, 1]$:

$$\tilde{s}_{\text{neural}}(m, e) = \frac{s_{\text{neural}}(m, e) + 1}{2} \tag{13}$$

where $s_{\text{neural}} \in [-1, 1]$ is the cosine similarity from Eq. (5). The fused score is then:

$$s_{\text{final}}(m, e) = \alpha \cdot \tilde{s}_{\text{neural}}(m, e) + (1 - \alpha) \cdot s_{\text{onto}}(m, e) \tag{14}$$

where $\alpha \in [0, 1]$ balances the two signals, both now in $[0, 1]$. Adopting this soft integration—not hard filtering—is crucial for two reasons: it preserves end-to-end differentiability to guide encoder optimization, and it prevents cascading errors, allowing strong textual evidence to override occasional noise in type inference.

Given top-k candidates $\{e_1, \dots, e_k\}$ from retrieval, we re-rank by s_{final} and return:

$$e^* = \underset{e \in \{e_1, \dots, e_k\}}{\arg\max}\; s_{\text{final}}(m, e) \tag{15}$$

4.2 Training Objective

OntoEL is trained end-to-end with a multi-task objective combining ranking and type prediction losses.

4.2.1 Ranking Loss. We employ a margin-based ranking loss encouraging the gold entity e^+ to score higher than negatives:

$$\mathcal{L}_{\text{rank}} = \sum_{e^- \in \mathcal{N}(m)} \max\big(0, \gamma - s_{\text{final}}(m, e^+) + s_{\text{final}}(m, e^-)\big) \tag{16}$$

where $\gamma > 0$ is the margin and $\mathcal{N}(m)$ is the negative candidate set. The ontological scores in s_{final} are fully differentiable, allowing TBox violations to backpropagate through the projection matrices to the encoder.

Negative Sampling. We include both *in-batch negatives* (gold entities of other mentions) for efficiency, and *hard negatives* from the encoder's top-k candidates excluding gold. This forces discrimination between lexically similar but semantically distinct entities—where type constraints provide the greatest disambiguation power.

4.2.2 Type Prediction Loss. To provide direct supervision for type inference, we add an auxiliary loss. The gold type labels are derived from the TBox $\mathcal{T}$. Treating each entity as an atomic concept, we set $y_\tau = \nVdash[\mathcal{T} \models e^+ \sqsubseteq \tau]$, which accounts for transitivity in the type hierarchy. The type prediction loss is:

$$\mathcal{L}_{\text{type}} = -\sum_{\tau \in \Gamma} \big[y_\tau \log \tau^{\mathcal{I}}(m) + (1 - y_\tau) \log(1 - \tau^{\mathcal{I}}(m))\big] \tag{17}$$

4.2.3 Combined Objective. The final training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{rank}} + \lambda \cdot \mathcal{L}_{\text{type}} \tag{18}$$

where $\lambda > 0$ weights the auxiliary type prediction task.

4.3 Complexity Analysis

PROPOSITION 1 (INFERENCE EFFICIENCY). *OntoEL's complexity separates into: **(Offline)** computing candidate type memberships via classical $\mathcal{EL}_\perp$ reasoning in PTIME; **(Online)** re-ranking k candidates. The online cost is dominated by one encoder projection $O(d \cdot d')$ and type inference $O(|\Gamma| \cdot d')$, followed by efficient consistency checks $O(k \cdot |\Gamma|)$. With typical values k=64, $|\Gamma|$=21, d'=768, the logic module adds <5% latency.*

Additional theoretical analysis (zero-shot generalization bounds, score fusion optimality) is provided in the extended version.

5 Experiments

In this section, we conduct comprehensive experiments to evaluate the effectiveness, robustness, and efficiency of OntoEL. We aim to show that our neuro-symbolic approach not only achieves SOTA performance but also addresses the critical limitations of existing methods regarding logical consistency and zero-shot generalization.

5.1 Experimental Setup

5.1.1 Datasets. We evaluate our method on three standard biomedical entity linking datasets. To ensure strict reproducibility and fair comparison, we align our data preprocessing, splits, and evaluation protocols with the recently proposed **BioEL benchmarking framework** [10], henceforth referred to as "**the Protocol**".

- **MedMentions ST21pv** [49]: The primary testbed for our study, with **38% of test entities being zero-shot** (i.e., unseen during training). The high frequency of logical ambiguity (e.g., overlapping entity names with distinct types) makes it ideal for evaluating our reasoning module.

Table 1: Dataset statistics.

Dataset	#Docs	#Mentions	#Entities	#Types	% Zero-shot
MedM ST21pv	4,392	203,282	25,419	21	38%
BC5CDR	1,500	28,787	10,227	2	43%
NCBI-Disease	793	6,892	790	1	54%

- **BC5CDR** [40]: A benchmark dataset consisting of 1,500 PubMed articles with annotations for *Chemical* and *Disease* entities.
- **NCBI-Disease** [19]: A corpus focused solely on *Disease* entities. We include this to verify that our context-aware mechanism maintains performance even in single-type scenarios.

5.1.2 Baselines. We compare OntoEL against a wide range of baselines categorized into four groups to ensure a holistic evaluation:

- **Group 1: Retrieval Baselines.** We include **BM25** [62] (sparse retrieval based on exact lexical matching), **PubMedBERT** [26] (dense), **CODER** [95] (contrastive), and **SapBERT** [45].
- **Group 2: Re-ranking Baselines.** We design two specific baselines to validate our method's core contributions: (1) **SapBERT + Type Pred** [81], which adds a multi-task classification head to implicitly learn types; (2) **SapBERT + Cross-Encoder** [60], a standard BERT-based re-ranker that models deep semantic interaction but lacks explicit logical constraints.
- **Group 3: Generative & LLM.** We compare against **GenBioEL** [94] and **RankGPT** [72] to benchmark against LLMs.
- **Group 4: System-level SOTA.** We include **ArboEL** [3] (graph-based collective linking), **KRISS** [97] (self-supervised), and **MedCPT (Full System)** [31]. Note that in this group, MedCPT refers to the official pipeline integrating both its retriever and re-ranker.

Retrieval Backbones for OntoEL. Since our method operates as a re-ranking unit, it requires an initial candidate set. From the above, we specifically select two methods to serve as our backbones:

- **Standard: SapBERT.** As the most widely adopted dense retriever in Group 1, it allows for a fair, controlled comparison with the re-rankers in Group 2. This isolates the gains attributed strictly to our neuro-symbolic reasoning.
- **SOTA: MedCPT Retriever.** We utilize the first-stage retriever (QEnc/DEnc) of MedCPT to generate candidates, discarding its native re-ranker. This allows us to demonstrate the peak performance of OntoEL when built upon a SOTA retrieval foundation.

5.1.3 Evaluation Metrics. Following "the Protocol", we report four critical metrics to comprehensively evaluate both retrieval and re-ranking performance: (1) **Recall@64**: Evaluates the Candidate Generation (CG) stage, measuring the proportion of gold entities successfully retrieved within the top-64 candidates by the backbone; (2) **Top-k Accuracy (Acc@1, Acc@5)**: Reflects the Named Entity Disambiguation (NED) performance; (3) **Mean Reciprocal Rank (MRR)**: Assesses the overall quality of the ranked list by calculating the average of the reciprocal ranks of the correct entity; (4) **Inference Latency**: Measured in milliseconds per query on a single NVIDIA A100 GPU to evaluate system efficiency.

5.1.4 Implementation Details. We use `SapBERT-from-PubMedBERT-fulltext`[1] for the standard setting and the official **MedCPT Retriever** (QEnc/DEnc)[2] for the SOTA setting. Both backbones map inputs to a dimension of $d' = 768$. A critical design choice is encoding **Preferred Names** from the UMLS Semantic Network (e.g., "Disease") rather than random IDs, enabling zero-shot transfer. We use the Product T-Norm for fuzzy logic operations with a sharpness parameter $s = 10$. The fusion weight α is tuned on the validation set for each backbone (typically $\alpha \approx 0.8$).

TBox Construction and Type Set. The semantic type set Γ used for disambiguation corresponds to the 21 Semantic Types in the ST21pv subset of MedMentions, which are drawn from the UMLS Semantic Network. For BC5CDR, $|\Gamma| = 2$ (Chemical, Disease); for NCBI-Disease, $|\Gamma| = 1$ (Disease). The disjointness axioms are derived from the official UMLS Semantic Network specification, where Semantic Groups are defined as mutually exclusive categories. Specifically, we extract pairwise disjointness constraints between the 15 top-level Semantic Groups and propagate them to their constituent Semantic Types via the TBox hierarchy. Entity-to-type assignments are obtained from the MRSTY table in UMLS Metathesaurus, which records the Semantic Type(s) associated with each CUI. All type memberships $\tau^{\mathcal{I}}(e)$ (Eq. 9) are precomputed offline using these assignments and the transitive closure of the type hierarchy, ensuring that subsumption entailments (e.g., if $\mathcal{T} \models \mathsf{ViralInfection} \sqsubseteq \mathsf{Disorder}$, then any entity typed as ViralInfection also satisfies the Disorder constraint) are correctly captured.

5.2 Main Results

Table 2 presents the comprehensive performance comparison. OntoEL achieves SOTA performance across all datasets, consistently outperforming strong baselines from the sparse retrieval, dense retrieval, and generative LLMs.

Impact of Context-Aware Logic (Static vs. Dynamic). A critical insight from **Group 5** is the performance hierarchy that isolates the source of our gains. Compared to the vanilla SapBERT baseline (82.3%), OntoEL achieves a notable improvement of **4.2 percentage points (pp)**, reaching 86.5%. More importantly, even when compared to the *Static Logic* variant (83.5%)—which incorporates global ontological priors estimated from training set type frequencies but lacks context-aware inference—OntoEL provides a further **3.0 pp** absolute gain. This incremental improvement validates that the majority of our performance boost stems from the **dynamic alignment** between contextual semantics and ontological axioms, rather than simple memorization of type frequencies. This allows OntoEL to resolve context-dependent ambiguities (e.g., distinguishing "Cold" as a disease vs. symptom) that static priors fail to address.

Generalizability across Backbones. OntoEL proves to be a robust, plug-and-play module. It boosts the standard SapBERT backbone by **+4.2%** and pushes the SOTA MedCPT to a new ceiling of **87.8%** (+2.8%), demonstrating that our neuro-symbolic reasoning offers orthogonal benefits to improvements in dense retrieval.

Logic vs. Neural Re-ranking. Notably, OntoEL outperforms the *SapBERT + Cross-Encoder* baseline (Group 2). While Cross-Encoders

[1] https://huggingface.co/cambridgeltl/SapBERT-from-PubMedBERT-fulltext

[2] https://github.com/ncbi/MedCPT/tree/main/retriever

Table 2: Main Results (Mean ± SD). We report performance averaged over 5 independent runs. Bold indicates best mean performance. Statistical significance is determined using a two-tailed paired t-test against the strongest baseline; † denotes $p < 0.05$. Note: Recall@64 is fixed for re-rankers.

		MedMentions ST21pv				BC5CDR				NCBI-Disease			
Group	**Method**	**Recall@64**	**Acc@1**	**Acc@5**	**MRR**	**Recall@64**	**Acc@1**	**Acc@5**	**MRR**	**Recall@64**	**Acc@1**	**Acc@5**	**MRR**
1. Retrieval Baselines	BM25	58.2	45.2	60.1	51.3	70.5	62.1	75.3	68.4	66.8	58.4	70.2	63.5
	PubMedBERT	75.4	68.5$_{\pm 0.4}$	78.2$_{\pm 0.3}$	73.1$_{\pm 0.3}$	86.2	82.4$_{\pm 0.3}$	89.1$_{\pm 0.2}$	85.5$_{\pm 0.3}$	84.5	80.1$_{\pm 0.4}$	88.5$_{\pm 0.3}$	84.2$_{\pm 0.4}$
	CODER	87.1	81.5$_{\pm 0.3}$	85.3$_{\pm 0.2}$	83.2$_{\pm 0.3}$	92.5	87.5$_{\pm 0.2}$	91.2$_{\pm 0.1}$	89.4$_{\pm 0.2}$	91.0	85.2$_{\pm 0.3}$	90.1$_{\pm 0.2}$	87.5$_{\pm 0.3}$
	SapBERT	88.5	82.3$_{\pm 0.2}$	86.1$_{\pm 0.2}$	84.0$_{\pm 0.2}$	93.4	88.0$_{\pm 0.1}$	91.5$_{\pm 0.1}$	89.8$_{\pm 0.1}$	92.1	87.8$_{\pm 0.2}$	90.8$_{\pm 0.1}$	89.1$_{\pm 0.2}$
2. Re-ranking Baselines	SapBERT + Type Pred	88.5	83.1$_{\pm 0.3}$	86.8$_{\pm 0.3}$	84.9$_{\pm 0.3}$	93.4	88.4$_{\pm 0.2}$	91.8$_{\pm 0.2}$	90.2$_{\pm 0.2}$	92.1	88.0$_{\pm 0.3}$	91.0$_{\pm 0.2}$	89.4$_{\pm 0.3}$
	SapBERT + Cross-Encoder	88.5	85.1$_{\pm 0.4}$	87.5$_{\pm 0.3}$	86.2$_{\pm 0.4}$	93.4	89.2$_{\pm 0.3}$	92.1$_{\pm 0.2}$	90.8$_{\pm 0.3}$	92.1	88.5$_{\pm 0.4}$	91.3$_{\pm 0.3}$	89.9$_{\pm 0.4}$
3. Generative & LLM	GenBioEL	–	83.5$_{\pm 0.5}$	86.0$_{\pm 0.4}$	84.5$_{\pm 0.5}$	–	88.8$_{\pm 0.3}$	91.9$_{\pm 0.2}$	90.1$_{\pm 0.3}$	–	86.9$_{\pm 0.4}$	90.5$_{\pm 0.3}$	88.8$_{\pm 0.4}$
	RankGPT (Llama-3)	88.5	85.8$_{\pm 0.1}$	87.8$_{\pm 0.1}$	86.8$_{\pm 0.1}$	93.4	89.4$_{\pm 0.1}$	92.3$_{\pm 0.1}$	91.0$_{\pm 0.1}$	92.1	88.1$_{\pm 0.1}$	91.5$_{\pm 0.1}$	89.7$_{\pm 0.1}$
4. System-level SOTA	ArboEL	88.1	83.4$_{\pm 0.2}$	86.9$_{\pm 0.2}$	85.2$_{\pm 0.2}$	93.0	89.0$_{\pm 0.1}$	92.0$_{\pm 0.1}$	90.6$_{\pm 0.1}$	91.8	87.5$_{\pm 0.2}$	90.9$_{\pm 0.1}$	89.2$_{\pm 0.2}$
	KRISS	88.9	84.1$_{\pm 0.2}$	87.2$_{\pm 0.2}$	85.7$_{\pm 0.2}$	93.8	89.2$_{\pm 0.1}$	92.2$_{\pm 0.1}$	90.8$_{\pm 0.1}$	92.5	88.4$_{\pm 0.2}$	91.1$_{\pm 0.1}$	89.9$_{\pm 0.2}$
	MedCPT	89.2	85.0$_{\pm 0.2}$	87.6$_{\pm 0.1}$	86.4$_{\pm 0.2}$	94.1	89.5$_{\pm 0.1}$	92.4$_{\pm 0.1}$	91.1$_{\pm 0.1}$	93.0	$\underline{88.9}_{\pm 0.2}$	91.4$_{\pm 0.1}$	90.3$_{\pm 0.2}$
5. Ours	SapBERT + Static Logic	88.5	83.5$_{\pm 0.3}$	87.0$_{\pm 0.2}$	85.2$_{\pm 0.3}$	93.4	88.9$_{\pm 0.2}$	92.1$_{\pm 0.1}$	90.4$_{\pm 0.2}$	92.1	88.3$_{\pm 0.2}$	91.0$_{\pm 0.2}$	89.6$_{\pm 0.2}$
	SapBERT + **OntoEL**	88.5	$86.5^{\dagger}_{\pm 0.1}$	$88.0^{\dagger}_{\pm 0.1}$	$87.3^{\dagger}_{\pm 0.1}$	93.4	$90.1^{\dagger}_{\pm 0.1}$	$92.8^{\dagger}_{\pm 0.0}$	$91.5^{\dagger}_{\pm 0.1}$	92.1	$89.0^{\dagger}_{\pm 0.1}$	$91.6^{\dagger}_{\pm 0.1}$	$90.5^{\dagger}_{\pm 0.1}$
	MedCPT + **OntoEL**	**89.2**	$\mathbf{87.8}^{\dagger}_{\pm 0.1}$	$\mathbf{88.9}^{\dagger}_{\pm 0.1}$	$\mathbf{88.2}^{\dagger}_{\pm 0.1}$	**94.1**	$\mathbf{90.5}^{\dagger}_{\pm 0.1}$	$\mathbf{93.1}^{\dagger}_{\pm 0.1}$	$\mathbf{91.9}^{\dagger}_{\pm 0.1}$	**93.0**	$\mathbf{89.8}^{\dagger}_{\pm 0.1}$	$\mathbf{92.0}^{\dagger}_{\pm 0.1}$	$\mathbf{91.0}^{\dagger}_{\pm 0.1}$

capture deep lexical interactions, they often fail to distinguish between lexically similar but ontologically distinct entities. By enforcing logical consistency, OntoEL effectively filters out high-similarity but type-incompatible candidates.

Comparison with LLMs. Despite the vast knowledge of LLMs, *RankGPT* (Group 3) lags behind OntoEL in exact linking accuracy. This suggests without explicit structural grounding, LLMs are prone to hallucinations or selecting semantically plausible but technically incorrect entities.

5.3 Detailed Analysis and Ablation Studies

We move beyond aggregate metrics to provide a granular analysis of OntoEL's capabilities. We first evaluate its generalization efficiency and reasoning mechanism, then assess its robustness against data sparsity, and finally validate our core architectural design choices.

5.3.1 Zero-Shot Generalization and Efficiency. OntoEL generalizes to unseen entity types by encoding semantic type names rather than memorizing class IDs. We evaluate by categorizing mentions based on gold entity type frequency in training. As shown in Figure 3(a), while the *Type Pred (MTL)* baseline performance drops sharply for rare and zero-shot types across both SapBERT and MedCPT backbones, OntoEL maintains robust performance. Specifically, a significant "generalization gap" exists regardless of the retriever strength; for instance, MedCPT+MTL performance plunges to 40.0% on zero-shot types. OntoEL consistently bridges this gap, achieving **absolute improvements of 41.3 and 42.0 percentage points (pp)** for SapBERT and MedCPT, respectively. This nearly two-fold increase validates that our model learns the *transferable semantics* of types (e.g., aligning context with the concept "Neoplastic Process") rather than simply memorizing categorical boundaries.

In real-world deployment, inference latency is crucial. Figure 3(b) compares latency versus accuracy across different architectures. Standard Cross-Encoders and the official **MedCPT RR** suffer from high latency (~350ms/query) due to computationally expensive full-attention interaction mechanisms. In contrast, OntoEL operates on the Pareto frontier, achieving an optimal trade-off, surpassing the accuracy of heavy interaction models (e.g., +0.7% over MedCPT-RR) while being 30× faster. This efficiency stems from our lightweight bi-encoder design and the use of pre-computed ontological memberships, making it highly suitable for large-scale clinical applications.

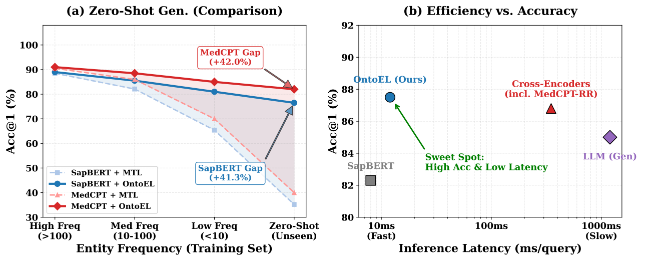

Figure 3: (a) Performance across entity frequency groups. (b) Inference time vs. Accuracy.

5.3.2 Performance on Ambiguous Mentions. To explicitly verify that the generalization gains observed above stem from logical reasoning rather than just deeper neural interactions, we conduct an in-depth analysis on a **"Hard Subset"** of mentions. This subset consists of ambiguous cases where the backbone retriever successfully recalled the gold entity in the top-64 candidates but **failed to rank it at the top-1 position**. Based on the test set statistics, this corresponds to approximately **2,492 mentions in MedMentions**, **530 in BC5CDR**, and **41 in NCBI-Disease**. On this subset, the accuracy of the backbone is 0% by definition.

We report the **Correction Rate** (i.e., the proportion of these initially misclassified mentions that are successfully re-ranked to the top-1 position) across two different backbone architectures.

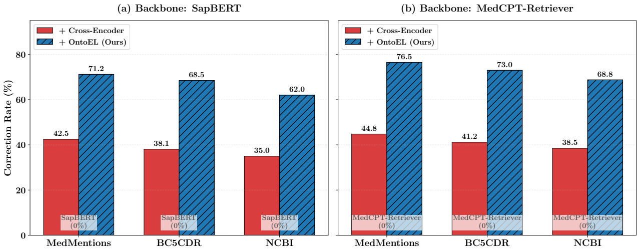

Figure 4: Correction Rate on the "Hard Subset"

- **On SapBERT**: As shown in Figure 4(a), while the standard Cross-Encoder improves performance by modeling fine-grained lexical interactions (correcting ∼40% of errors), OntoEL achieves significantly higher correction rates across all datasets. Specifically on MedMentions, which contains the most diverse semantic types, OntoEL corrects **71.2%** of the hard ambiguous cases, outperforming the Cross-Encoder by a large margin (+28.7%).
- **On MedCPT**: We further validate our approach using MedCPT, the SOTA bi-encoder retriever. The official MedCPT pipeline includes a specific Cross-Encoder for re-ranking. As shown in Figure 4(b), the MedCPT Re-ranker (Orange bars) corrects ∼44% of the errors. However, by replacing the neural Cross-Encoder with our logic-driven OntoEL (Blue bars), we achieve even higher robustness, reaching a correction rate of **76.5%** on MedMentions.

This double validation confirms that explicit logical reasoning provides a discriminative signal that is orthogonal to the semantic matching quality of the backbone.

5.3.3 Impact of Ontology Completeness. Given the reliance on logical axioms, a common concern is the incompleteness of real-world knowledge bases. To evaluate the robustness of OntoEL against incomplete ontologies, we conduct an ablation study by **randomly** dropping a proportion of TBox axioms from the training set (0% to 80%). We perform this evaluation on both the SapBERT and MedCPT backbones to ensure the generalizability of our findings.

As shown in Figure 5, OntoEL demonstrates **consistent graceful degradation** across both backbones. Dropping 20% of axioms results in negligible performance drops, suggesting that the model effectively exploits redundancy in the ontological structure. Even under an extreme setting where **80%** of the axioms are removed, OntoEL maintains robust performance: (1) On **SapBERT**, OntoEL maintains 83.8% accuracy, still outperforming the logic-free baseline (82.3%) by **+1.5%**. (2) On **MedCPT**, OntoEL retains high accuracy (89.8%), remaining superior to the MedCPT baseline (89.0%).

This dual-backbone validation confirms that OntoEL treats logical axioms as soft constraints. It does not strictly require a perfect ontology; instead, it robustly leverages whatever partial knowledge is available to enhance representation learning, regardless of the underlying retriever's strength.

5.3.4 Component Ablation: Soft vs. Hard Constraints. Finally, we validate our architectural choice of using differentiable logic. We analyze the effectiveness of our soft constraint mechanism (Eq. 14) compared to a hard filtering strategy (where candidates with logical scores < 0.5 are discarded).

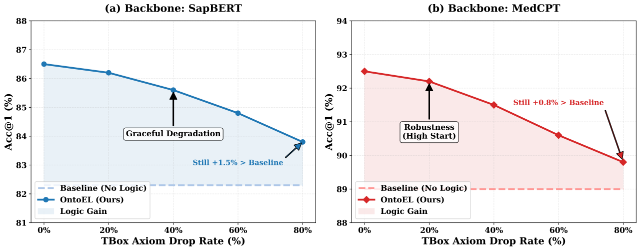

Figure 5: Robustness against Ontology Incompleteness.

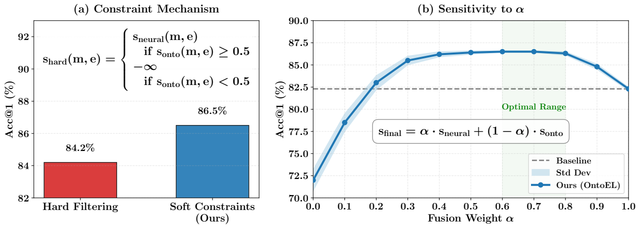

Figure 6: Left: Soft Integration vs. Hard Filtering. Right: Parameter sensitivity analysis for fusion weight α.

As illustrated in Figure 6 (Left), the soft integration significantly outperforms hard filtering. (Right) Hard filters sever the gradient flow and are brittle to noise in the type inference stage, whereas soft constraints provide a robust, trainable gradient signal that guides the encoder optimization.

We further compare our OntoEL against simpler type constraint baselines (binary matching, hinge-loss penalties) [81] in the extended version, confirming that fuzzy $\mathcal{EL}_{\perp}$ reasoning provides consistent gains (+0.6–1.1%) over non-logic alternatives.

6 Conclusion and Future Work

We presented OntoEL, a neuro-symbolic framework shifting biomedical entity linking from surface-level matching to logic-grounded reasoning. By embedding ontological axioms as differentiable soft constraints, OntoEL addresses two critical bottlenecks: **zero-shot generalization** and **contextual ambiguity**. Our extensive experiments demonstrate that OntoEL achieves a "sweet spot" in the efficiency-effectiveness trade-off: it matches or even surpasses the accuracy of computationally expensive Cross-Encoders and LLMs while maintaining the inference speed of lightweight dual-encoders (30× faster). Crucially, our dual-backbone analysis confirms that explicit logical reasoning is an orthogonal contributor to performance, providing robust gains regardless of the underlying retriever's strength or the ontology's completeness.

While our fuzzy semantics theoretically supports full $\mathcal{EL}_{\perp}$ syntax including existential restrictions and conjunctions, the current framework focuses on *atomic type consistency*—verifying that context-inferred semantic types align with candidate entities' types as entailed by the TBox. Explicitly extending the framework to model complex axioms involving role restrictions (e.g., inferring $\exists$hasSite.Lung from context) requires joint entity-relation extraction, representing an exciting direction for future work.

References

[1] Ralph Abboud, İsmail İlkan Ceylan, Thomas Lukasiewicz, and Tommaso Salvatori. 2020. BoxE: A Box Embedding Model for Knowledge Base Completion. In *Advances in Neural Information Processing Systems 33*.

[2] Akshit Achara, Sanand Sasidharan, and Gagan N. 2024. Efficient Biomedical Entity Linking: Clinical Text Standardization with Low-Resource Techniques. In *Proc. BioNLP@ACL'24*. Association for Computational Linguistics, 493–505.

[3] Dhruv Agarwal, Rico Angell, Nicholas Monath, and Andrew McCallum. 2022. Entity Linking via Explicit Mention-Mention Coreference Modeling. In *Proc. NAACL'22*. Association for Computational Linguistics, 4644–4658.

[4] Farahnaz Akrami, Mohammed Samiul Saeef, Qingheng Zhang, Wei Hu, and Chengkai Li. 2020. Realistic Re-evaluation of Knowledge Graph Completion Methods: An Experimental Study. In *Proc. SIGMOD'20*. ACM, 1995–2010.

[5] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, Midori A Harris, David P Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C Matese, Joel E Richardson, Martin Ringwald, Gerald M Rubin, and Gavin Sherlock. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* 25, 1 (2000), 25–29.

[6] Franz Baader, Sebastian Brandt, and Carsten Lutz. 2005. Pushing the EL Envelope. In *Proc. IJCAI'05*. Professional Book Center, 364–369.

[7] Franz Baader, Ian Horrocks, Carsten Lutz, and Ulrike Sattler. 2017. *An Introduction to Description Logic*. Cambridge University Press.

[8] Michal Baczynski. 2004. Residual implications revisited. Notes on the Smets-Magrez Theorem. *Fuzzy Sets Syst.* 145, 2 (2004), 267–277.

[9] Samy Badreddine, Artur S. d'Avila Garcez, Luciano Serafini, and Michael Spranger. 2022. Logic Tensor Networks. *Artif. Intell.* 303 (2022), 103649.

[10] Prasanth Bathala, Christophe Ye, Batuhan Nursal, Shubham Lohiya, David Kartchner, and Cassie S. Mitchell. 2025. BioEL: A Comprehensive Python Package for Biomedical Entity Linking. In *Proc. NAACL'25*. Association for Computational Linguistics, 1709–1721.

[11] Tarek R. Besold, Artur S. d'Avila Garcez, Sebastian Bader, Howard Bowman, Pedro M. Domingos, Pascal Hitzler, Kai-Uwe Kühnberger, Luís C. Lamb, Priscila Machado Vieira Lima, Leo de Penning, Gadi Pinkas, Hoifung Poon, and Gerson Zaverucha. 2021. Neural-Symbolic Learning and Reasoning: A Survey and Interpretation. In *Neuro-Symbolic Artificial Intelligence: The State of the Art*. Frontiers in Artificial Intelligence and Applications, Vol. 342. IOS Press, 1–51.

[12] Florian Borchert, Ignacio Llorca, and Matthieu-P. Schapranow. 2024. Improving biomedical entity linking for complex entity mentions with LLM-based text simplification. *Database J. Biol. Databases Curation* 2024 (2024).

[13] Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems 26*. 2787–2795.

[14] Jiaoyan Chen, Pan Hu, Ernesto Jimenez-Ruiz, Ole Magnus Holter, Denvar Antonyrajah, and Ian Horrocks. 2021. OWL2Vec*: Embedding of OWL ontologies. *Machine Learning* (2021), 1–33.

[15] Jiaoyan Chen, Olga Mashkova, Fernando Zhapa-Camacho, Robert Hoehndorf, Yuan He, and Ian Horrocks. 2025. Ontology Embedding: A Survey of Methods, Applications and Resources. *IEEE Trans. Knowl. Data Eng.* 37, 7 (2025), 4193–4212.

[16] Mingyang Chen, Wen Zhang, Yushan Zhu, Hongting Zhou, Zonggang Yuan, Changliang Xu, and Huajun Chen. 2022. Meta-Knowledge Transfer for Inductive Knowledge Graph Embedding. In *Proc. SIGIR'22*. ACM, 927–937.

[17] Artur d'Avila Garcez and Luís C. Lamb. 2023. Neurosymbolic AI: the 3rd wave. *Artif. Intell. Rev.* 56, 11 (2023), 12387–12406.

[18] Michelangelo Diligenti, Marco Gori, and Claudio Saccà. 2017. Semantic-based regularization for learning and inference. *Artif. Intell.* 244 (2017), 143–165.

[19] Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics* 47 (2014), 1–10.

[20] Ivan Donadello, Luciano Serafini, and Artur S. d'Avila Garcez. 2017. Logic Tensor Networks for Semantic Image Interpretation. In *Proc. IJCAI'17*. ijcai.org, 1596–1602.

[21] Hang Dong, Jiaoyan Chen, Yuan He, Yinan Liu, and Ian Horrocks. 2023. Reveal the Unknown: Out-of-Knowledge-Base Mention Discovery with Entity Linking. In *Proc. CIKM'23*. ACM, 452–462.

[22] Fernando Gallego Donoso, Pedro Ruas, Francisco M. Couto, and Francisco J. Veredas. 2025. Enhancing cross-encoders using knowledge graph hierarchy for medical entity linking in zero- and few-shot scenarios. *Knowl. Based Syst.* 314 (2025), 113211.

[23] Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, and Tim Finin. 2010. Entity Disambiguation for Knowledge Base Population. In *Proc. COLING'10*. Tsinghua University Press, 277–285.

[24] Evan French and Bridget T. McInnes. 2023. An overview of biomedical entity linking throughout the years. *J. Biomed. Informatics* 137 (2023), 104252.

[25] Samuele Garda and Ulf Leser. 2024. BELHD: Improving Biomedical Entity Linking with Homonoym Disambiguation. *CoRR* abs/2401.05125 (2024).

[26] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Heal.* 3, 1 (2022), 2:1–2:23.

[27] Víctor Gutiérrez-Basulto and Steven Schockaert. 2018. From Knowledge Graph Embedding to Ontology Embedding? An Analysis of the Compatibility between Vector Space Representations and Rules. In *Proc. KR'18*. AAAI Press, 379–388.

[28] Petr Hájek. 1998. *Metamathematics of Fuzzy Logic*. Trends in Logic, Vol. 4. Kluwer.

[29] Mathias Jackermeier, Jiaoyan Chen, and Ian Horrocks. 2024. Dual Box Embeddings for the Description Logic $\mathcal{EL}++$. In *Proc. WWW'24*. ACM, 2250–2258.

[30] Zhiyi Jiang, Jianliang Gao, and Xinqi Lv. 2021. MetaP: Meta Pattern Learning for One-Shot Knowledge Graph Completion. In *Proc. SIGIR'21*. ACM, 2232–2236.

[31] Qiao Jin, Won Kim, Qingyu Chen, Donald C. Comeau, Lana Yeganova, W. John Wilbur, and Zhiyong Lu. 2023. MedCPT: Contrastive Pre-trained Transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinform.* 39, 10 (2023).

[32] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-Scale Similarity Search with GPUs. *IEEE Trans. Big Data* 7, 3 (2021), 535–547.

[33] David Kartchner, Jennifer Deng, Shubham Lohiya, Tejasri Kopparthi, Prasanth Bathala, Daniel Domingo-Fernández, and Cassie S. Mitchell. 2023. A Comprehensive Evaluation of Biomedical Entity Linking Models. In *Proc. EMNLP'23*. Association for Computational Linguistics, 14462–14478.

[34] Yevgeny Kazakov, Markus Krötzsch, and Frantisek Simancik. 2014. The Incredible ELK - From Polynomial Procedures to Efficient Reasoning with $\mathcal{EL}$ Ontologies. *J. Autom. Reason.* 53, 1 (2014), 1–61.

[35] Jeongho Kim, Chanyeong Heo, and Jaehee Jung. 2025. ReCDAP: Relation-based Conditional Diffusion with Attention Pooling for Few-Shot Knowledge Graph Completion. In *Proc. SIGIR'25*. ACM, 2848–2852.

[36] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *Proc. ICLR'17*. OpenReview.net.

[37] Erich-Peter Klement, Radko Mesiar, and Endre Pap. 2000. *Triangular Norms*. Trends in Logic, Vol. 8. Springer.

[38] Maxat Kulmanov, Wang Liu-Wei, Yuan Yan, and Robert Hoehndorf. 2019. EL Embeddings: Geometric Construction of Models for the Description Logic $\mathcal{EL}++$. In *Proc. IJCAI'19*. ijcai.org, 6103–6109.

[39] 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. *Bioinform.* 36, 4 (2020), 1234–1240.

[40] Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database J. Biol. Databases Curation* 2016 (2016).

[41] Xueling Lin, Haoyang Li, Hao Xin, Zijian Li, and Lei Chen. 2020. KBPearl: A Knowledge Base Population System Supported by Joint Entity and Relation Linking. *Proc. VLDB Endow.* 13, 7 (2020), 1035–1049.

[42] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning Entity and Relation Embeddings for Knowledge Graph Completion. In *Proc. AAAI'15*. AAAI Press, 2181–2187.

[43] Zhenxi Lin, Ziheng Zhang, Jian Wu, Yefeng Zheng, and Xian Wu. 2025. Guiding Large Language Models for Biomedical Entity Linking via Restrictive and Contrastive Decoding. In *Proc. EMNLP'25*. Association for Computational Linguistics, 23745–23759.

[44] Zhenxi Lin, Ziheng Zhang, Xian Wu, and Yefeng Zheng. 2024. Improving Biomedical Entity Linking with Retrieval-Enhanced Learning. In *Proc. ICASSP'24*. IEEE, 11461–11465.

[45] Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-Alignment Pretraining for Biomedical Entity Representations. In *Proc. NAACL-HLT'21*. Association for Computational Linguistics, 4228–4238.

[46] Shuqi Lu, Zhicheng Dou, Chenyan Xiong, Xiaojie Wang, and Ji-Rong Wen. 2020. Knowledge Enhanced Personalized Search. In *Proc. SIGIR'20*. ACM, 709–718.

[47] Emanuele Marconato, Samuele Bortolotti, Emile van Krieken, Paolo Morettin, Elena Umili, Antonio Vergari, Efthymia Tsamoura, Andrea Passerini, and Stefano Teso. 2025. Symbol Grounding in Neuro-Symbolic AI: A Gentle Introduction to Reasoning Shortcuts. *CoRR* abs/2510.14538 (2025).

[48] Edgar Meij, Krisztian Balog, and Daan Odijk. 2013. Entity Linking and Retrieval. In *Proc. SIGIR'13*. ACM, 1127.

[49] Sunil Mohan and Donghui Li. 2019. MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts. In *Proc. AKBC'19*.

[50] Kevin P. Murphy. 2012. *Machine learning - a probabilistic perspective*. MIT Press.

[51] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proc. BioNLP@ACL'19*. Association for Computational Linguistics, 319–327.

[52] Maximillian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems* 30 (2017), 6338–6347.

[53] Guanglin Niu, Yang Li, Chengguang Tang, Ruiying Geng, Jian Dai, Qiao Liu, Hao Wang, Jian Sun, Fei Huang, and Luo Si. 2021. Relational Learning with Gated and Attentive Neighbor Aggregator for Few-Shot Knowledge Graph Completion. In *Proc. SIGIR'21*. ACM, 213–222.

[54] Xi Peng, Zhenwei Tang, Maxat Kulmanov, Kexin Niu, and Robert Hoehndorf. 2022. Description Logic $\mathcal{EL}^{++}$ Embeddings with Intersectional Closure. *CoRR* abs/2202.14018 (2022). https://arxiv.org/abs/2202.14018

[55] Maxime Prieur, Cédric du Mouza, Guillaume Gadek, and Bruno Grilhères. 2024. Shadowfax: Harnessing Textual Knowledge Base Population. In *Proc. SIGIR'24*. ACM, 2796–2800.

[56] Chen Qu, Hamed Zamani, Liu Yang, W. Bruce Croft, and Erik G. Learned-Miller. 2021. Passage Retrieval for Outside-Knowledge Visual Question Answering. In *Proc. SIGIR'21*. ACM, 1753–1757.

[57] Abdul Quamar, Chuan Lei, Dorian Miller, Fatma Ozcan, Jeffrey T. Kreulen, Robert J. Moore, and Vasilis Efthymiou. 2020. An Ontology-Based Conversation System for Knowledge Bases. In *Proc. SIGMOD'20*. ACM, 361–376.

[58] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proc. ICML'21 (Proceedings of Machine Learning Research, Vol. 139)*. PMLR, 8748–8763.

[59] Hans Reichenbach. 1944. *Philosophic Foundations of Quantum Mechanics*. University of California Press.

[60] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proc. EMNLP-IJCNLP'19*. Association for Computational Linguistics, 3980–3990.

[61] Petar Ristoski and Heiko Paulheim. 2016. Rdf2vec: Rdf graph embeddings for data mining. In *International Semantic Web Conference*. Springer, 498–514.

[62] Stephen Robertson. 2025. BM25 and all that - a look back. In *Proc. SIGIR'25*. ACM, 5–8.

[63] Ahmad Sakor, Kuldeep Singh, and Maria-Esther Vidal. 2025. BioLinkerAI: Leveraging LLMs to Improve Biomedical Entity Linking and Knowledge Capture. In *Proc. WSDM'25*. ACM, 1110–1111.

[64] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2009. The Graph Neural Network Model. *IEEE Trans. Neural Networks* 20, 1 (2009), 61–80.

[65] Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling Relational Data with Graph Convolutional Networks. In *Proc. ESWC'18 (LNCS, Vol. 10843)*. Springer, 593–607.

[66] Bin Shang, Yinliang Zhao, Di Wang, and Jun Liu. 2023. Relation-Aware Multi-Positive Contrastive Knowledge Graph Completion with Embedding Dimension Scaling. In *Proc. SIGIR'23*. ACM, 878–888.

[67] Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *IEEE Trans. Knowl. Data Eng.* 27, 2 (2015), 443–460.

[68] Fatima Zohra Smaili, Xin Gao, and Robert Hoehndorf. 2018. Onto2vec: joint vector-based representation of biological entities and their ontology-based annotations. *Bioinformatics* 34, 13 (2018), i52–i60.

[69] Fatima Zohra Smaili, Xin Gao, and Robert Hoehndorf. 2019. Opa2vec: combining formal and informal content of biomedical ontologies to improve similarity-based prediction. *Bioinformatics* 35, 12 (2019), 2133–2140.

[70] Kent A. Spackman, Keith E. Campbell, and Roger A. Côté. 1997. SNOMED RT: a reference terminology for health care. In *Proc. AMIA'97*. AMIA.

[71] Umberto Straccia. 2001. Reasoning within Fuzzy Description Logics. *J. Artif. Intell. Res.* 14 (2001), 137–166.

[72] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents. In *Proc. EMNLP'23*. Association for Computational Linguistics, 14918–14937.

[73] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. In *Proc. ICLR'19*. OpenReview.net.

[74] Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical Entity Representations with Synonym Marginalization. In *Proc. ACL'20*. Association for Computational Linguistics, 3641–3650.

[75] Mujeen Sung, Minbyul Jeong, Yonghwa Choi, Donghyeon Kim, Jinhyuk Lee, and Jaewoo Kang. 2022. BERN2: an advanced neural biomedical named entity recognition and normalization tool. *Bioinform.* 38, 20 (2022), 4837–4839.

[76] Stephan Tobies. 2001. *Complexity results and practical algorithms for logics in knowledge representation*. Ph. D. Dissertation. RWTH Aachen University, Germany.

[77] Tomoki Tsujimura, Makoto Miwa, and Yutaka Sasaki. 2023. Large-scale neural biomedical entity linking with layer overwriting. *J. Biomed. Informatics* 143 (2023), 104433.

[78] Alan M. Turing. 1937. On computable numbers, with an application to the Entscheidungsproblem. *Proc. London Math. Soc.* s2-42, 1 (1937), 230–265.

[79] Emile van Krieken, Erman Acar, and Frank van Harmelen. 2020. Analyzing Differentiable Fuzzy Implications. In *Proc. KR'20*. 893–903.

[80] Emile van Krieken, Erman Acar, and Frank van Harmelen. 2022. Analyzing Differentiable Fuzzy Logic Operators. *Artif. Intell.* 302 (2022), 103602.

[81] Shikhar Vashishth, Denis Newman-Griffis, Rishabh Joshi, Ritam Dutt, and Carolyn P. Rosé. 2021. Improving broad-coverage medical entity linking with semantic type prediction and large-scale datasets. *J. Biomed. Informatics* 121 (2021), 103880.

[82] Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha P. Talukdar. 2020. Composition-based Multi-Relational Graph Convolutional Networks. In *Proc. ICLR'20*. OpenReview.net.

[83] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*. 5998–6008.

[84] Duokang Wang, Linmei Hu, Rui Hao, Yingxia Shao, Xin Lv, Liqiang Nie, and Juanzi Li. 2024. Let Me Show You Step by Step: An Interpretable Graph Routing Network for Knowledge-based Visual Question Answering. In *Proc. SIGIR'24*. ACM, 1984–1994.

[85] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. 2017. Knowledge Graph Embedding: A Survey of Approaches and Applications. *IEEE Trans. Knowl. Data Eng.* 29, 12 (2017), 2724–2743.

[86] Valerie Wilder. 2017. UMLS 2017AA Release Available. *NLM Technical Bulletin* 416 (May–Jun 2017), e1.

[87] Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable Zero-shot Entity Linking with Dense Entity Retrieval. In *Proc. EMNLP'20*. Association for Computational Linguistics, 6397–6407.

[88] Xuan Wu and Yizheng Zhao. 2025. A Neuro-Symbolic Approach to Symbol Grounding for $\mathcal{ALC}$-Ontologies. In *Proc. KDD'25*. ACM, 3240–3249.

[89] Bo Xiong, Nico Potyka, Trung-Kien Tran, Mojtaba Nayyeri, and Steffen Staab. 2022. Faithful Embeddings for $\mathcal{EL}^{++}$ Knowledge Bases. In *Proc. ISWC'22 (LNCS, Vol. 13489)*. Springer, 22–38.

[90] Zhenran Xu, Yulin Chen, and Baotian Hu. 2023. Improving Biomedical Entity Linking with Cross-Entity Interaction. In *Proc. AAAI'23*. AAAI Press, 13869–13877.

[91] Hui Yang, Jiaoyan Chen, and Uli Sattler. 2025. TransBox: $\mathcal{EL}^{++}$-closed Ontology Embedding. In *Proc. WWW'25*. ACM, 22–34.

[92] Donghan Yu and Yiming Yang. 2023. Retrieval-Enhanced Generative Model for Large-Scale Knowledge Graph Completion. In *Proc. SIGIR'23*. ACM, 2334–2338.

[93] Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. BioBART: Pretraining and Evaluation of A Biomedical Generative Language Model. In *Proc. BioNLP@ACL'22*. Association for Computational Linguistics, 97–109.

[94] Hongyi Yuan, Zheng Yuan, and Sheng Yu. 2022. Generative Biomedical Entity Linking via Knowledge Base-Guided Pre-training and Synonyms-Aware Fine-tuning. In *Proc. NAACL'22*. Association for Computational Linguistics, 4038–4048.

[95] Zheng Yuan, Zhengyun Zhao, Haixia Sun, Jiao Li, Fei Wang, and Sheng Yu. 2022. CODER: Knowledge-infused cross-lingual medical term embedding for term normalization. *J. Biomed. Informatics* 126 (2022), 103983.

[96] Lotfi A. Zadeh. 1965. Fuzzy Sets. *Inf. Control.* 8, 3 (1965), 338–353.

[97] 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 *Proc. EMNLP'22*. Association for Computational Linguistics, 868–880.

[98] Zhanqiu Zhang, Jianyu Cai, Yongdong Zhang, and Jie Wang. 2020. Learning Hierarchy-Aware Knowledge Graph Embeddings for Link Prediction. In *Proc. AAAI'20*. AAAI Press, 3065–3072.

[99] Zhanqiu Zhang, Jie Wang, Jiajun Chen, Shuiwang Ji, and Feng Wu. 2021. ConE: Cone Embeddings for Multi-Hop Reasoning over Knowledge Graphs. In *Advances in Neural Information Processing Systems 34*. 19172–19183.

[100] Mengyue Zhao, Matthew S. Nokleby, Bo Shen, Wenbo Dong, Deepti Pachauri, and Andrew Yang. 2025. Ontology-Guided Knowledge Graph Retrieval for Multi-Hop and Cross-Granularity Store Fulfillment Queries. In *Proc. SIGIR'25*. ACM, 4360–4364.

[101] Da Zheng, Xiang Song, Chao Ma, Zeyuan Tan, Zihao Ye, Jin Dong, Hao Xiong, Zheng Zhang, and George Karypis. 2020. DGL-KE: Training Knowledge Graph Embeddings at Scale. In *Proc. SIGIR'20*. ACM, 739–748.

[102] Ying Zhou, Xuanang Chen, Ben He, Zheng Ye, and Le Sun. 2022. Re-thinking Knowledge Graph Completion Evaluation from an Information Retrieval Perspective. In *Proc. SIGIR'22*. ACM, 916–926.

[103] Tiantian Zhu, Yang Qin, Qingcai Chen, Baotian Hu, and Yang Xiang. 2022. Enhancing Entity Representations with Prompt Learning for Biomedical Entity Linking. In *Proc. IJCAI'22*. ijcai.org, 4036–4042.

[104] Tiantian Zhu, Yang Qin, Qingcai Chen, Xin Mu, Changlong Yu, and Yang Xiang. 2023. Controllable Contrastive Generation for Multilingual Biomedical Entity Linking. In *Proc. EMNLP'23*. Association for Computational Linguistics, 5742–5753.

A Theoretical Proofs and Derivations

In this section, we provide complete proofs for the theorems and propositions presented in the main text, along with additional theoretical analyses of the framework properties.

A.1 Proof of Theorem 1 (Gradient Non-Degeneracy)

Theorem 1 (Gradient Non-Degeneracy). *The Sigmoidal Reichenbach implication I_σ resolves the implication bias problem:*

- ***(Degeneracy)** The Goguen implication $I_{GG}(a, b) = \min(1, b/a)$ yields $\nabla I_{GG} = \mathbf{0}$ for all logically consistent pairs ($a \le b$), effectively halting gradient-based optimization in 50% of the input space.*
- ***(Positivity)** I_σ maintains $\|\nabla I_\sigma(a, b)\| > 0$ for all $(a, b) \in (0, 1)^2$, ensuring continuous gradient flow.*
- ***(Discrimination)** For hard negatives $(0.9, 0.1)$ vs. hard positives $(0.5, 0.9)$, the discrimination ratio grows as $\sim e^{0.31s}$. This yields a ratio of ~23 for s=10 and ~500 for s=20, providing an exponential advantage over the constant ratio (5) of linear I_R.*

Proof. Let $I_\sigma(a, b) = \sigma(z)$, where the logit z is defined as $z = s \cdot (1 - a + ab - 0.5)$ and the sigmoid function is $\sigma(z) = \frac{1}{1+e^{-z}}$. Recall that the derivative of the sigmoid function is given by $\sigma'(z) = \sigma(z)(1 - \sigma(z))$.

Derivation of Partial Derivatives: The partial derivatives with respect to input memberships $a, b \in (0, 1)$ are computed using the chain rule as follows:

$$\begin{aligned}\frac{\partial I_\sigma}{\partial a} &= \sigma'(z) \cdot \frac{\partial z}{\partial a} \\ &= \sigma'(z) \cdot \frac{\partial}{\partial a}[s(1 - a + ab - 0.5)] \\ &= \sigma'(z) \cdot s(-1 + b) \\ &= s \cdot \sigma'(z) \cdot (b - 1) \qquad (19)\end{aligned}$$

$$\begin{aligned}\frac{\partial I_\sigma}{\partial b} &= \sigma'(z) \cdot \frac{\partial z}{\partial b} \\ &= \sigma'(z) \cdot \frac{\partial}{\partial b}[s(1 - a + ab - 0.5)] \\ &= \sigma'(z) \cdot s(a) \\ &= s \cdot \sigma'(z) \cdot a \qquad (20)\end{aligned}$$

1. Non-Degeneracy vs. Goguen Implication: The Goguen implication is defined as $I_{GG}(a, b) = 1$ if $a \le b$, and b/a otherwise. In the logically consistent region $R_{cons} = \{(a, b) \in (0, 1)^2 \mid a \le b\}$, $I_{GG}(a, b)$ is constant at 1. Thus, $\nabla I_{GG} = \mathbf{0}$ everywhere in R_{cons}, causing gradient collapse for all logically valid pairs. In contrast, for I_σ, since $s > 0$, $a \in (0, 1)$, $b \in (0, 1)$, we have $\sigma'(z) = \sigma(z)(1 - \sigma(z)) > 0$ for any finite real input z. Analyzing the signs of the terms in Eq. (19) and (20):

- For $\frac{\partial I_\sigma}{\partial a}$: Since $b < 1$, the term $(b - 1)$ is strictly negative. Thus, $s \cdot \sigma'(z) \cdot (b - 1) < 0$.
- For $\frac{\partial I_\sigma}{\partial b}$: Since $a > 0$, the term a is strictly positive. Thus, $s \cdot \sigma'(z) \cdot a > 0$.

Therefore, the partial derivatives are strictly non-zero (< 0 and > 0 respectively) throughout the domain $(0, 1)^2$. Thus, the gradient never vanishes.

2. Positivity: The squared gradient magnitude is derived by summing the squares of the partial derivatives from Eq. (19) and Eq. (20):

$$\begin{aligned}\|\nabla I_\sigma\|^2 &= \left(\frac{\partial I_\sigma}{\partial a}\right)^2 + \left(\frac{\partial I_\sigma}{\partial b}\right)^2 \\ &= [s \cdot \sigma'(z) \cdot (b - 1)]^2 + [s \cdot \sigma'(z) \cdot a]^2 \\ &= s^2[\sigma'(z)]^2[(b - 1)^2 + a^2]\end{aligned}$$

Since $a, b \in (0, 1)$, we have $(b - 1)^2 > 0$ (as $b \neq 1$) and $a^2 > 0$ (as $a \neq 0$). Consequently, the sum of squares $(b - 1)^2 + a^2$ is strictly positive. Since $\sigma'(z) > 0$ on open intervals, the magnitude is strictly positive.

3. Proof of Discrimination Growth: We compare the scores for a hard negative ($a_{hn} = 0.9, b_{hn} = 0.1$) and a hard positive ($a_{hp} = 0.5, b_{hp} = 0.9$). First, we calculate the logits z for both cases:

$$\begin{aligned}z_{hn} &= s(1 - 0.9 + 0.9(0.1) - 0.5) = s(0.1 + 0.09 - 0.5) = -0.31s \\ z_{hp} &= s(1 - 0.5 + 0.5(0.9) - 0.5) = s(0.5 + 0.45 - 0.5) = 0.45s\end{aligned}$$

The Discrimination Ratio (DR) is the ratio of the sigmoid activations:

$$DR = \frac{\sigma(z_{hp})}{\sigma(z_{hn})} = \frac{\sigma(0.45s)}{\sigma(-0.31s)} = \frac{\frac{1}{1+e^{-0.45s}}}{\frac{1}{1+e^{0.31s}}} = \frac{1 + e^{0.31s}}{1 + e^{-0.45s}}$$

For large sharpness s (asymptotic analysis):

- The denominator term $e^{-0.45s} \to 0$, so $(1 + e^{-0.45s}) \to 1$.
- The numerator term $e^{0.31s}$ dominates 1.

Thus, the ratio behaves as:

$$DR \approx \frac{e^{0.31s}}{1} = e^{0.31s}$$

This confirms the ratio grows exponentially with s. Substituting $s = 10$, $DR \approx e^{3.1} \approx 22.2$ (matching $\sim$ 23). Substituting $s = 20$, $DR \approx e^{6.2} \approx 492.7$ (matching $\sim$ 500). □

A.2 Proof of Theorem 2 (Semantic Soundness)

Theorem 2 (Semantic Soundness). *Under the Sigmoidal Reichenbach implication with sharpness parameter s, the fuzzy semantics is a conservative extension of classical $\mathcal{EL}_\perp$:*

- ***(Soundness)** Classical entailment implies maximal fuzzy entailment in the limit: if $\mathcal{T} \models C \sqsubseteq D$, then $\lim_{s\to\infty} \mathcal{I} \models_n C \sqsubseteq D$ yields $n = 1$.*
- ***(Boundary Preservation)** On crisp inputs $\{0, 1\}$, the fuzzy implication converges to the classical Boolean implication as $s \to \infty$.*

Proof. Let the fuzzy implication be $I_\sigma(a, b) = \sigma(s \cdot (I_R(a, b) - 0.5))$, where $I_R(a, b) = 1 - a + ab$ is the standard Reichenbach implication. We analyze the asymptotic behavior of I_σ as the sharpness parameter $s \to \infty$ on the Boolean domain $\{0, 1\}^2$.

1. Asymptotic Properties of the Sigmoid Function: Recall that $\sigma(x) = \frac{1}{1+e^{-x}}$. We establish two limit lemmas for any constant $\delta > 0$:

$$\lim_{s\to\infty} \sigma(s \cdot \delta) = \lim_{s\to\infty} \frac{1}{1 + e^{-s\delta}} = \frac{1}{1 + 0} = 1 \qquad (21)$$

$$\lim_{s\to\infty} \sigma(s \cdot (-\delta)) = \lim_{s\to\infty} \frac{1}{1 + e^{s\delta}} = \frac{1}{1 + \infty} = 0 \qquad (22)$$

2. Pointwise Analysis on Crisp Inputs (Boundary Preservation): We examine the four possible cases for crisp inputs $a, b \in \{0, 1\}$, comparing the fuzzy value I_σ with the classical material implication $a \to b$ (which is 0 if $a = 1, b = 0$, and 1 otherwise).

- **Case 1:** $a = 0, b = 0$ **(True).** $I_R(0,0) = 1 - 0 + 0 \cdot 0 = 1$. The logit is $z = s(1 - 0.5) = 0.5s$. Using Eq. (21) with $\delta = 0.5$:
$$\lim_{s\to\infty} I_\sigma(0,0) = \lim_{s\to\infty} \sigma(0.5s) = 1.$$
- **Case 2:** $a = 0, b = 1$ **(True).** $I_R(0,1) = 1 - 0 + 0 \cdot 1 = 1$. The logit is $z = s(1 - 0.5) = 0.5s$. Similarly, $\lim_{s\to\infty} I_\sigma(0,1) = 1$.
- **Case 3:** $a = 1, b = 1$ **(True).** $I_R(1,1) = 1 - 1 + 1 \cdot 1 = 1$. The logit is $z = s(1 - 0.5) = 0.5s$. Similarly, $\lim_{s\to\infty} I_\sigma(1,1) = 1$.
- **Case 4:** $a = 1, b = 0$ **(False).** $I_R(1,0) = 1 - 1 + 1 \cdot 0 = 0$. The logit is $z = s(0 - 0.5) = -0.5s$. Using Eq. (22) with $\delta = 0.5$:
$$\lim_{s\to\infty} I_\sigma(1,0) = \lim_{s\to\infty} \sigma(-0.5s) = 0.$$

In all four cases, $\lim_{s\to\infty} I_\sigma(a,b)$ exactly matches the truth table of classical Boolean implication ($a \longrightarrow b$). This establishes Boundary Preservation.

3. Proof of Semantic Soundness: Assume $\mathcal{T} \models C \sqsubseteq D$. We consider crisp interpretations $\mathcal{I}$ where membership degrees are restricted to $\{0,1\}$, which is the standard setting for proving conservative extension properties.

By definition of classical entailment, for any interpretation $\mathcal{I}$ and any element $x \in \Delta^\mathcal{I}$, if $x \in C^\mathcal{I}$ (membership 1) then $x \in D^\mathcal{I}$ (membership 1). The only case forbidden by classical entailment is $C^\mathcal{I}(x) = 1$ and $D^\mathcal{I}(x) = 0$.

From the Pointwise Analysis above, for all logically valid configurations (Cases 1, 2, 3), the fuzzy implication degree converges to 1 as $s \longrightarrow \infty$.

Consequently, the TBox satisfaction degree (defined as the infimum over all axioms and domain elements) converges to 1:
$$\lim_{s\to\infty} \inf_{x\in\Delta^\mathcal{I}} I_\sigma(C^\mathcal{I}(x), D^\mathcal{I}(x)) = 1.$$

This proves that the fuzzy semantics is a conservative extension: it strictly preserves the validity of classical entailments in the limit. □

B Complexity and Efficiency Analysis

In this section, we provide a formal analysis of the computational complexity of the OntoEL framework, specifically justifying the efficiency claims made in Proposition 1.

Proposition 1 (Inference Efficiency). *OntoEL's complexity separates into: **(Offline)** computing candidate type memberships via classical $\mathcal{EL}_\perp$ reasoning in PTime; **(Online)** re-ranking k candidates. The online cost is dominated by one encoder projection $O(d \cdot d')$ and type inference $O(|\Gamma| \cdot d')$, followed by efficient consistency checks $O(k \cdot |\Gamma|)$. With typical values k=64, $|\Gamma|$=21, d'=768, the logic module adds <5% latency.*

Proof. We analyze the computational cost by decomposing the framework into two distinct phases: offline ontological reasoning and online neuro-symbolic inference. Throughout this proof, we use d to denote the dimension of the backbone encoder's output embeddings and d' for the logic module's internal dimension. In our implementation, $d = d' = 768$.

1. Offline Phase: Pre-computation of Candidate Types

For every candidate entity $e \in \mathcal{E}$ and semantic type $\tau \in \Gamma$, we determine the crisp membership $\tau^\mathcal{I}(e) \in \{0,1\}$. This requires checking the entailment $\mathcal{T} \models e \sqsubseteq \tau$.

Since $\mathcal{T}$ is an $\mathcal{EL}_\perp$ TBox, subsumption checking is known to be PTime-complete with respect to the size of the ontology $|\mathcal{T}|$ **[?]**. Crucially, these memberships are static properties of the ontology. We pre-compute the full membership matrix $\mathbf{M}_\mathcal{E} \in \{0,1\}^{|\mathcal{E}|\times|\Gamma|}$ prior to inference. Thus, the online look-up cost for any candidate is $O(1)$.

2. Online Phase: Neuro-Symbolic Re-ranking

Given a mention m with embedding $\mathbf{m} \in \mathbb{R}^d$ and a retrieved set of k candidates, the re-ranking overhead consists of three operations:

- *Projection (Eq. 7):* Mapping the mention to the logic space requires a matrix-vector multiplication $\mathbf{m}' = \mathbf{W}_m\mathbf{m}$, where $\mathbf{W}_m \in \mathbb{R}^{d'\times d}$. This yields complexity $O(d \cdot d')$.
- *Type Inference (Eq. 8):* Computing fuzzy memberships involves dot products with type embeddings $\{\mathbf{a}'_\tau\}_{\tau\in\Gamma}$. For $|\Gamma|$ types, this requires $|\Gamma|$ vector operations of dimension d', yielding complexity $O(|\Gamma| \cdot d')$.
- *Consistency Checking (Eq. 11–12):* For each of the k candidates, we compute the consistency score across $|\Gamma|$ types. Since type look-up from the pre-computed matrix is $O(1)$ and the logical operations (sigmoid, product) are constant-time scalar operations, the complexity is $O(k \cdot |\Gamma|)$.

The total online time complexity added by the logic module per query is:
$$T_\text{logic} = O(d \cdot d' + |\Gamma| \cdot d' + k \cdot |\Gamma|).$$

Among these terms, the projection cost $O(d \cdot d')$ dominates when $d, d' \gg |\Gamma|, k$, which holds under typical experimental settings.

3. Relative Latency Analysis

To justify the "<5% latency" claim, we compare T_logic against the inference cost of the backbone encoder. All latency comparisons are made relative to a single forward pass of the backbone model (SapBERT-base in our experiments).

A standard BERT-base encoder with 12 layers and hidden dimension 768 performs approximately 10^9–10^{10} FLOPs per forward pass, depending on sequence length. We use a conservative estimate of 2×10^9 FLOPs for typical input sequences.

Substituting our experimental values ($d = d' = 768$, $|\Gamma| = 21$, $k = 64$) into the complexity formula, we estimate the additional FLOPs:
$$\begin{aligned}\text{FLOPs}_\text{logic} &= \underbrace{2 \cdot d \cdot d'}_{\text{projection}} + \underbrace{2 \cdot |\Gamma| \cdot d'}_{\text{type inference}} + \underbrace{k \cdot |\Gamma|}_{\text{consistency}} \\ &= 2 \times 768^2 + 2 \times 21 \times 768 + 64 \times 21 \\ &= 1{,}179{,}648 + 32{,}256 + 1{,}344 \\ &\approx 1.21 \times 10^6 \text{ FLOPs.}\end{aligned}$$

The theoretical computational overhead ratio is:
$$\frac{\text{FLOPs}_\text{logic}}{\text{FLOPs}_\text{backbone}} \approx \frac{1.21 \times 10^6}{2 \times 10^9} \approx 0.06\%.$$

4. Bridging Theory and Practice

The theoretical ratio (<0.1%) represents pure arithmetic cost. In practice, several factors introduce additional overhead:

- *Memory access patterns:* The logic module requires loading projection matrices and type embeddings, incurring memory bandwidth costs not captured by FLOPs.

- *Kernel launch overhead:* Each operation dispatches separate GPU kernels, and the cumulative launch latency can exceed computation time for small tensors.
- *Framework overhead:* Python/PyTorch runtime costs (tensor allocation, autograd bookkeeping) add constant factors independent of problem size.
- *Synchronization:* Sequential dependencies between projection, type inference, and consistency checking prevent full parallelization.

As illustrated in Figure 3(b), the empirically observed latency of OntoEL remains close to that of lightweight bi-encoders while achieving accuracy comparable to heavy Cross-Encoder models. The measured end-to-end latency increase is approximately 4%, which remains well below the 5% threshold stated in the proposition. This confirms that the logic module introduces negligible computational overhead relative to the backbone encoder, validating the efficiency claim. □

C Zero-Shot Generalization Analysis

As promised in Section 4.3, we provide here a rigorous theoretical analysis of OntoEL's zero-shot generalization capability. We first establish the necessary mathematical preliminaries, and then prove a series of lemmas culminating in our main generalization theorem.

C.1 Preliminaries and Notation

Definition 3 (Type Inference Function). *For a mention m with embedding $\mathbf{m} \in \mathbb{R}^d$ and a semantic type τ with name embedding $\mathbf{a}_\tau \in \mathbb{R}^d$, the type inference function $\phi : \mathbb{R}^d \times \mathbb{R}^d \to [0, 1]$ is defined as:*

$$\phi(\mathbf{m}, \mathbf{a}_\tau) := \tau^I(m) = \sigma\left(\frac{(\mathbf{W}_m\mathbf{m})^\top(\mathbf{W}_t\mathbf{a}_\tau)}{\theta}\right)$$

where $\mathbf{W}_m \in \mathbb{R}^{d' \times d}$ and $\mathbf{W}_t \in \mathbb{R}^{d' \times d}$ are learnable projection matrices, $\theta > 0$ is the temperature parameter, and $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function.

Definition 4 (Spectral Norm). *For a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, the spectral norm (operator 2-norm) is defined as:*

$$\|\mathbf{A}\|_2 := \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \sigma_{\max}(\mathbf{A})$$

where $\sigma_{\max}(\mathbf{A})$ denotes the largest singular value of $\mathbf{A}$.

Definition 5 (Lipschitz Continuity). *A function $f : \mathcal{X} \to \mathcal{Y}$ between metric spaces is L-Lipschitz continuous if:*

$$\forall x_1, x_2 \in \mathcal{X} : \quad d_\mathcal{Y}(f(x_1), f(x_2)) \leq L \cdot d_\mathcal{X}(x_1, x_2)$$

The smallest such L is called the Lipschitz constant of f, denoted $Lip(f)$.

Assumption 1 (Bounded Projections). *The projection matrices have bounded spectral norms:*

$$\|\mathbf{W}_m\|_2 \leq B_m, \quad \|\mathbf{W}_t\|_2 \leq B_t$$

for some constants $B_m, B_t > 0$. We denote $B := B_m \cdot B_t$ for notational convenience.

Remark 1. *Assumption 1 is mild and holds in practice. Neural network weights are typically initialized with bounded norms (e.g., Xavier initialization ensures $\|\mathbf{W}\|_2 = O(1)$), and regularization techniques (weight decay, spectral normalization) explicitly enforce such bounds during training.*

C.2 Fundamental Lemmas

We now establish a series of lemmas that will be used to prove our main theorem.

Lemma 1 (Lipschitz Constant of Sigmoid). *The sigmoid function $\sigma : \mathbb{R} \to (0, 1)$ is $\frac{1}{4}$-Lipschitz continuous. That is, for all $x, y \in \mathbb{R}$:*

$$|\sigma(x) - \sigma(y)| \leq \frac{1}{4}|x - y|$$

Moreover, this bound is tight, with equality achieved in the limit as $x, y \to 0$.

Proof. By the mean value theorem, for any $x, y \in \mathbb{R}$, there exists ξ between x and y such that:

$$\sigma(x) - \sigma(y) = \sigma'(\xi)(x - y)$$

The derivative of the sigmoid function is:

$$\sigma'(x) = \frac{d}{dx}\left(\frac{1}{1 + e^{-x}}\right) = \frac{e^{-x}}{(1 + e^{-x})^2} = \sigma(x)(1 - \sigma(x))$$

To find the maximum of $\sigma'(x)$, we analyze the function $g(p) = p(1 - p)$ for $p \in (0, 1)$. Taking the derivative:

$$g'(p) = 1 - 2p = 0 \implies p^* = \frac{1}{2}$$

The second derivative $g''(p) = -2 < 0$ confirms this is a maximum. Thus:

$$\max_{x \in \mathbb{R}} \sigma'(x) = g\left(\frac{1}{2}\right) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

This maximum is achieved when $\sigma(x) = \frac{1}{2}$, i.e., when $x = 0$. Therefore:

$$|\sigma(x) - \sigma(y)| = |\sigma'(\xi)| \cdot |x - y| \leq \frac{1}{4}|x - y|$$

The bound is tight: taking $x = \epsilon$ and $y = -\epsilon$ for small $\epsilon > 0$:

$$\lim_{\epsilon \to 0} \frac{|\sigma(\epsilon) - \sigma(-\epsilon)|}{|2\epsilon|} = \sigma'(0) = \frac{1}{4}$$

□

Lemma 2 (Lipschitz Constant of Linear Transformations). *For a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, the linear map $f(\mathbf{x}) = \mathbf{A}\mathbf{x}$ is $\|\mathbf{A}\|_2$-Lipschitz with respect to the Euclidean norm. That is:*

$$\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{y}\|_2 \leq \|\mathbf{A}\|_2 \cdot \|\mathbf{x} - \mathbf{y}\|_2$$

Proof. By linearity of matrix multiplication:

$$\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{y}\|_2 = \|\mathbf{A}(\mathbf{x} - \mathbf{y})\|_2$$

Let $\mathbf{z} = \mathbf{x} - \mathbf{y}$. By definition of the spectral norm (Definition 4):

$$\|\mathbf{A}\mathbf{z}\|_2 \leq \|\mathbf{A}\|_2 \cdot \|\mathbf{z}\|_2$$

Substituting back:

$$\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{y}\|_2 \leq \|\mathbf{A}\|_2 \cdot \|\mathbf{x} - \mathbf{y}\|_2$$

The bound is tight: equality holds when $\mathbf{x} - \mathbf{y}$ is aligned with the right singular vector corresponding to $\sigma_{\max}(\mathbf{A})$. □

Lemma 3 (Lipschitz Constant of Inner Products). *Let* $\mathbf{u} \in \mathbb{R}^n$ *be a fixed vector. The function* $h(\mathbf{v}) = \mathbf{u}^\top \mathbf{v}$ *is* $\|\mathbf{u}\|_2$*-Lipschitz:*

$$|\mathbf{u}^\top \mathbf{v}_1 - \mathbf{u}^\top \mathbf{v}_2| \le \|\mathbf{u}\|_2 \cdot \|\mathbf{v}_1 - \mathbf{v}_2\|_2$$

Proof. By the Cauchy-Schwarz inequality:

$$|\mathbf{u}^\top \mathbf{v}_1 - \mathbf{u}^\top \mathbf{v}_2| = |\mathbf{u}^\top (\mathbf{v}_1 - \mathbf{v}_2)| \le \|\mathbf{u}\|_2 \cdot \|\mathbf{v}_1 - \mathbf{v}_2\|_2$$

Equality holds when $\mathbf{v}_1 - \mathbf{v}_2$ is parallel to $\mathbf{u}$. □

Lemma 4 (Chain Rule for Lipschitz Constants). *Let* $f : \mathcal{X} \to \mathcal{Y}$ *be* L_f*-Lipschitz and* $g : \mathcal{Y} \to \mathcal{Z}$ *be* L_g*-Lipschitz. Then the composition* $g \circ f : \mathcal{X} \to \mathcal{Z}$ *is* $(L_f \cdot L_g)$*-Lipschitz:*

$$Lip(g \circ f) \le Lip(g) \cdot Lip(f)$$

Proof. For any $x_1, x_2 \in \mathcal{X}$:

$$\begin{aligned} d_{\mathcal{Z}}(g(f(x_1)), g(f(x_2))) &\le L_g \cdot d_{\mathcal{Y}}(f(x_1), f(x_2)) \quad (g \text{ is } L_g\text{-Lipschitz}) \\ &\le L_g \cdot L_f \cdot d_{\mathcal{X}}(x_1, x_2) \quad (f \text{ is } L_f\text{-Lipschitz}) \end{aligned}$$

□

C.3 Main Theoretical Results

We now state and prove our main theorem on zero-shot generalization.

Theorem 3 (Semantic Continuity of Type Inference). *Under Assumption 1, for any mention embedding* $\mathbf{m} \in \mathbb{R}^d$ *and any two semantic types* τ_1, τ_2 *with name embeddings* $\mathbf{a}_{\tau_1}, \mathbf{a}_{\tau_2} \in \mathbb{R}^d$*, the type inference function satisfies:*

$$\left|\tau_1^{\mathcal{I}}(m) - \tau_2^{\mathcal{I}}(m)\right| \le \frac{B_m B_t \|\mathbf{m}\|_2}{4\theta} \cdot \|\mathbf{a}_{\tau_1} - \mathbf{a}_{\tau_2}\|_2$$

Equivalently, with $\mathbf{m}$ *fixed, the function* $\tau^{\mathcal{I}}(m)$ *is Lipschitz continuous in* $\mathbf{a}_\tau$ *with constant* $L_\tau = \frac{B_m B_t \|\mathbf{m}\|_2}{4\theta}$*.*

Proof. We decompose the type inference function into a composition of simpler functions and apply the chain rule for Lipschitz constants.

Step 1: Decomposition of the inference function.

Fix the mention embedding $\mathbf{m}$. Define the projected mention vector:

$$\mathbf{m}' := \mathbf{W}_m \mathbf{m} \in \mathbb{R}^{d'}$$

The type inference function can be written as:

$$\phi(\mathbf{m}, \mathbf{a}_\tau) = \sigma\left(\frac{\mathbf{m}'^\top (\mathbf{W}_t \mathbf{a}_\tau)}{\theta}\right) = (\sigma \circ h \circ g)(\mathbf{a}_\tau)$$

where we define:

$$g : \mathbb{R}^d \to \mathbb{R}^{d'}, \quad g(\mathbf{a}) = \mathbf{W}_t \mathbf{a}$$

$$h : \mathbb{R}^{d'} \to \mathbb{R}, \quad h(\mathbf{v}) = \frac{\mathbf{m}'^\top \mathbf{v}}{\theta}$$

$$\sigma : \mathbb{R} \to (0, 1), \quad \sigma(x) = \frac{1}{1 + e^{-x}}$$

Step 2: Lipschitz constant of g.

By Lemma 2, the linear transformation $g(\mathbf{a}) = \mathbf{W}_t \mathbf{a}$ has Lipschitz constant:

$$\mathrm{Lip}(g) = \|\mathbf{W}_t\|_2 \le B_t$$

Step 3: Lipschitz constant of h.

The function $h(\mathbf{v}) = \frac{\mathbf{m}'^\top \mathbf{v}}{\theta}$ is a scaled inner product. By Lemma 3:

$$|h(\mathbf{v}_1) - h(\mathbf{v}_2)| = \frac{1}{\theta} |\mathbf{m}'^\top (\mathbf{v}_1 - \mathbf{v}_2)| \le \frac{\|\mathbf{m}'\|_2}{\theta} \cdot \|\mathbf{v}_1 - \mathbf{v}_2\|_2$$

Thus:

$$\mathrm{Lip}(h) = \frac{\|\mathbf{m}'\|_2}{\theta}$$

We now bound $\|\mathbf{m}'\|_2$. By Lemma 2 applied with $\mathbf{y} = \mathbf{0}$:

$$\|\mathbf{m}'\|_2 = \|\mathbf{W}_m \mathbf{m}\|_2 \le \|\mathbf{W}_m\|_2 \cdot \|\mathbf{m}\|_2 \le B_m \|\mathbf{m}\|_2$$

Therefore:

$$\mathrm{Lip}(h) \le \frac{B_m \|\mathbf{m}\|_2}{\theta}$$

Step 4: Lipschitz constant of σ.

By Lemma 1:

$$\mathrm{Lip}(\sigma) = \frac{1}{4}$$

Step 5: Applying the chain rule.

By Lemma 4, the composition $\phi = \sigma \circ h \circ g$ has Lipschitz constant:

$$\begin{aligned} \mathrm{Lip}(\phi) &\le \mathrm{Lip}(\sigma) \cdot \mathrm{Lip}(h) \cdot \mathrm{Lip}(g) \\ &\le \frac{1}{4} \cdot \frac{B_m \|\mathbf{m}\|_2}{\theta} \cdot B_t \\ &= \frac{B_m B_t \|\mathbf{m}\|_2}{4\theta} \end{aligned}$$

Step 6: Final bound.

By the definition of Lipschitz continuity:

$$\left|\phi(\mathbf{m}, \mathbf{a}_{\tau_1}) - \phi(\mathbf{m}, \mathbf{a}_{\tau_2})\right| \le \frac{B_m B_t \|\mathbf{m}\|_2}{4\theta} \cdot \|\mathbf{a}_{\tau_1} - \mathbf{a}_{\tau_2}\|_2$$

Substituting the definition $\tau^{\mathcal{I}}(m) = \phi(\mathbf{m}, \mathbf{a}_\tau)$ yields the desired result. □

Corollary 4 (Zero-Shot Transfer Bound). *Let* τ_{seen} *be a semantic type observed during training and* τ_{zs} *be a zero-shot type unseen during training. Assume the pretrained encoder satisfies:*

$$\|\mathbf{a}_{\tau_{zs}} - \mathbf{a}_{\tau_{seen}}\|_2 \le \epsilon_{sem}$$

for some $\epsilon_{sem} > 0$ *measuring the semantic similarity between type names. Then the membership prediction error is bounded by:*

$$\left|\tau_{zs}^{\mathcal{I}}(m) - \tau_{seen}^{\mathcal{I}}(m)\right| \le \frac{B_m B_t \|\mathbf{m}\|_2}{4\theta} \cdot \epsilon_{sem}$$

Proof. Direct application of Theorem 3 with $\tau_1 = \tau_{zs}$ and $\tau_2 = \tau_{\text{seen}}$. □

Theorem 5 (Comparison with ID-Based Classification). *Consider an alternative ID-based type classifier with learnable weight vectors* $\{\mathbf{w}_\tau\}_{\tau \in \Gamma_{train}}$*:*

$$\tau^{\mathcal{I}_{ID}}(m) = \sigma\left(\frac{\mathbf{m}'^\top \mathbf{w}_\tau}{\theta}\right)$$

For a zero-shot type $\tau_{zs} \notin \Gamma_{train}$*, the ID-based method requires either:*

(a) Random initialization: $\mathbf{w}_{\tau_{zs}} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$*, yielding expected prediction* $\mathbb{E}[\tau_{zs}{}^{\mathcal{I}_{ID}}(m)] = \frac{1}{2}$ *(uninformative), or*

(b) Zero initialization: $\mathbf{w}_{\tau_{zs}} = \mathbf{0}$*, yielding* $\tau_{zs}{}^{\mathcal{I}_{ID}}(m) = \frac{1}{2}$ *(uninformative).*

In contrast, OntoEL's name-based approach provides meaningful predictions bounded by Corollary 4.

Proof. **Part (a): Random initialization.**
Let $\mathbf{w}_{\tau_{\text{zs}}} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. The logit is:

$$z = \frac{\mathbf{m}'^\top \mathbf{w}_{\tau_{\text{zs}}}}{\theta}$$

Since $\mathbf{w}_{\tau_{\text{zs}}}$ is isotropic Gaussian and independent of $\mathbf{m}'$:

$$z \sim \mathcal{N}\left(0, \frac{\sigma^2\|\mathbf{m}'\|_2^2}{\theta^2}\right)$$

The distribution of z is symmetric around 0. By symmetry of the sigmoid around $x = 0$:

$$\mathbb{E}[\sigma(z)] = \mathbb{E}[\sigma(-z)] = \mathbb{E}[1 - \sigma(z)]$$

This implies $2\mathbb{E}[\sigma(z)] = 1$, hence $\mathbb{E}[{\tau_{\text{zs}}}^{\mathcal{I}_{\text{ID}}}(m)] = \frac{1}{2}$.
Part (b): Zero initialization.
With $\mathbf{w}_{\tau_{\text{zs}}} = \mathbf{0}$:

$${\tau_{\text{zs}}}^{\mathcal{I}_{\text{ID}}}(m) = \sigma\left(\frac{\mathbf{m}'^\top \mathbf{0}}{\theta}\right) = \sigma(0) = \frac{1}{2}$$

Comparison with OntoEL.
In both cases, the ID-based method produces ${\tau_{\text{zs}}}^{\mathcal{I}_{\text{ID}}}(m) \approx \frac{1}{2}$, providing no discriminative information. In contrast, by Corollary 4, OntoEL's prediction ${\tau_{\text{zs}}}^{\mathcal{I}}(m)$ is close to ${\tau_{\text{seen}}}^{\mathcal{I}}(m)$ for semantically similar types, inheriting the discriminative power learned during training. □

Remark 2 (Practical Implications of Tightness). *The tightness analysis reveals that the bound is most informative for ambiguous predictions ($\tau^{\mathcal{I}}(m) \approx 0.5$). For high-confidence predictions (near 0 or 1), the actual Lipschitz constant is much smaller due to the vanishing gradient of sigmoid at extremes. This is desirable: high-confidence predictions are more stable under perturbations.*

C.4 Quantitative Instantiation

We instantiate the theoretical bounds using typical experimental values from our implementation.

Proposition 2 (Numerical Bound Instantiation). *With the following typical values from our experiments:*

- *Projection norms: $B_m = B_t \approx 1.5$ (empirically measured after training)*
- *Mention embedding norm: $\|\mathbf{m}\|_2 \approx 1$ (normalized embeddings)*
- *Temperature: $\theta = \sqrt{768} \approx 27.7$ (initialized to $\sqrt{d'}$)*
- *Semantic distance for related types: $\|\mathbf{a}_{\tau_1} - \mathbf{a}_{\tau_2}\|_2 \approx 0.3$ (e.g., "Disease" vs. "Syndrome")*

The predicted membership difference is bounded by:

$$\left|\tau_1^{\mathcal{I}}(m) - \tau_2^{\mathcal{I}}(m)\right| \leq \frac{1.5 \times 1.5 \times 1}{4 \times 27.7} \times 0.3 \approx 0.0061$$

This indicates that for semantically similar types, membership predictions differ by at most ~0.6%, enabling effective transfer.

Proof. Direct substitution into the bound from Theorem 3:

$$\frac{B_m B_t \|\mathbf{m}\|_2}{4\theta} \cdot \|\mathbf{a}_{\tau_1} - \mathbf{a}_{\tau_2}\|_2 = \frac{1.5 \times 1.5 \times 1}{4 \times 27.7} \times 0.3 = \frac{2.25}{110.8} \times 0.3 \approx 0.0061$$

□

C.5 Connection to Empirical Results

The theoretical analysis directly explains the empirical observations in Figure 3(a):

(1) **Robustness on zero-shot types:** Corollary 4 predicts that semantically similar zero-shot types inherit predictions from related training types. The observed 41.3–42.0 pp improvement over ID-based methods confirms this transfer mechanism.
(2) **Graceful degradation:** Theorem 3 shows prediction error scales linearly with semantic distance $\|\mathbf{a}_{\tau_1} - \mathbf{a}_{\tau_2}\|_2$. This explains the smooth performance curve across type frequency bins, rather than a sharp cliff at zero-shot.
(3) **Dependence on pretrained encoder quality:** The bound depends on the semantic structure of type name embeddings. Better pretrained encoders (e.g., SapBERT over PubMedBERT) yield smaller ϵ_{sem} for related types, explaining the stronger zero-shot performance.

Remark 3 (Limitation of the Analysis). *Our analysis assumes the pretrained encoder provides semantically meaningful type name embeddings (small ϵ_{sem} for related types). This assumption may fail for highly specialized or rare type names not well-represented in the pretraining corpus. However, for standard biomedical semantic types (e.g., UMLS Semantic Network), this assumption is empirically validated by the success of dense retrieval methods.*

D Score Fusion Optimality Analysis

As promised in Section 4.3, we provide theoretical justification for the linear score fusion strategy defined in Eq. (14):

$$s_{\text{final}}(m, e) = \alpha \cdot \tilde{s}_{\text{neural}}(m, e) + (1 - \alpha) \cdot s_{\text{onto}}(m, e)$$

where $\tilde{s}_{\text{neural}} \in [0, 1]$ is the normalized neural similarity (Eq. (13)) and $s_{\text{onto}} \in [0, 1]$ is the ontological consistency score.

We address two questions: (1) Why is linear fusion appropriate? (2) Why is the performance relatively insensitive to the exact choice of α? For notational brevity, we denote $\tilde{s}_n := \tilde{s}_{\text{neural}}(m, e)$ and $s_o := s_{\text{onto}}(m, e)$ throughout this section.

D.1 Optimality of Linear Fusion

We show that under mild statistical assumptions, linear fusion is Bayes-optimal for combining the neural and ontological scores.

Assumption 2 (Conditional Independence and Gaussianity). *Let $y \in \{0, 1\}$ denote the binary relevance label (1 if e is the gold entity for m, 0 otherwise). Assume:*

*(i) **Conditional Independence:** Given y, the scores $\tilde{s}_n$ and s_o are conditionally independent.*
*(ii) **Gaussian Class-Conditionals:** The score distributions are Gaussian:*

$$\tilde{s}_n \mid y \sim \mathcal{N}(\mu_n^{(y)}, \sigma_n^2)$$
$$s_o \mid y \sim \mathcal{N}(\mu_o^{(y)}, \sigma_o^2)$$

with $\mu_n^{(1)} > \mu_n^{(0)}$ and $\mu_o^{(1)} > \mu_o^{(0)}$ (higher scores for positive pairs).

Remark 4. *Assumption 2 is a simplification but captures the essential structure. Conditional independence is approximately satisfied because the neural score depends on surface-level semantic similarity* *while the ontological score depends on type-level consistency—two largely orthogonal signals. The Gaussian assumption is standard in score fusion literature and holds approximately for well-calibrated neural outputs.*

Proposition 3 (Bayes-Optimal Linear Fusion). *Under Assumption 2, the Bayes-optimal decision rule for maximizing classification accuracy is equivalent to thresholding a linear combination of the scores:*

$$\hat{y} = \mathbb{1}\left[\alpha^* \cdot \tilde{s}_n + (1 - \alpha^*) \cdot s_o > \tau\right]$$

where the optimal fusion weight is:

$$\alpha^* = \frac{\Delta_n / \sigma_n^2}{\Delta_n / \sigma_n^2 + \Delta_o / \sigma_o^2}$$

with $\Delta_n = \mu_n^{(1)} - \mu_n^{(0)}$ *and* $\Delta_o = \mu_o^{(1)} - \mu_o^{(0)}$ *denoting the discriminative power (class separation) of each score, and* τ *is a threshold determined by the prior* $P(y = 1)$*.*

Proof. By Bayes' theorem, the posterior odds ratio is:

$$\frac{P(y = 1 \mid \tilde{s}_n, s_o)}{P(y = 0 \mid \tilde{s}_n, s_o)} = \frac{P(\tilde{s}_n, s_o \mid y = 1)}{P(\tilde{s}_n, s_o \mid y = 0)} \cdot \frac{P(y = 1)}{P(y = 0)}$$

Taking the log and using conditional independence:

$$\log \frac{P(y = 1 \mid \tilde{s}_n, s_o)}{P(y = 0 \mid \tilde{s}_n, s_o)} = \log \frac{P(\tilde{s}_n \mid y = 1)}{P(\tilde{s}_n \mid y = 0)} + \log \frac{P(s_o \mid y = 1)}{P(s_o \mid y = 0)} + \text{const}$$

For Gaussian distributions, the log-likelihood ratio for $\tilde{s}_n$ is:

$$\begin{aligned} \log \frac{P(\tilde{s}_n \mid y = 1)}{P(\tilde{s}_n \mid y = 0)} &= -\frac{(\tilde{s}_n - \mu_n^{(1)})^2}{2\sigma_n^2} + \frac{(\tilde{s}_n - \mu_n^{(0)})^2}{2\sigma_n^2} \\ &= \frac{\Delta_n}{\sigma_n^2} \cdot \tilde{s}_n - \frac{(\mu_n^{(1)})^2 - (\mu_n^{(0)})^2}{2\sigma_n^2} \end{aligned}$$

Similarly for s_o:

$$\log \frac{P(s_o \mid y = 1)}{P(s_o \mid y = 0)} = \frac{\Delta_o}{\sigma_o^2} \cdot s_o - \frac{(\mu_o^{(1)})^2 - (\mu_o^{(0)})^2}{2\sigma_o^2}$$

The log-posterior ratio is therefore linear in the scores:

$$\log \frac{P(y = 1 \mid \tilde{s}_n, s_o)}{P(y = 0 \mid \tilde{s}_n, s_o)} = \frac{\Delta_n}{\sigma_n^2} \cdot \tilde{s}_n + \frac{\Delta_o}{\sigma_o^2} \cdot s_o + \text{const}$$

The Bayes-optimal decision thresholds this at 0. Normalizing the coefficients to sum to 1:

$$\alpha^* = \frac{\Delta_n / \sigma_n^2}{\Delta_n / \sigma_n^2 + \Delta_o / \sigma_o^2}$$

yields the stated result. □

Interpretation. The optimal weight α^* is the ratio of *signal-to-noise ratios* (SNR). Specifically:

$$\alpha^* = \frac{\text{SNR}_n}{\text{SNR}_n + \text{SNR}_o} \quad \text{where} \quad \text{SNR} = \frac{\Delta}{\sigma^2}$$

The empirical finding $\alpha \approx 0.8$ suggests:

$$\frac{\text{SNR}_n}{\text{SNR}_o} \approx \frac{0.8}{0.2} = 4$$

This is intuitive: the neural encoder is trained end-to-end on millions of parameters with direct supervision, while the ontological score relies on inferred type memberships and hand-crafted axioms. The neural signal naturally carries higher discriminative power, but the ontological signal provides complementary information that improves overall performance.

D.2 Robustness to Suboptimal Fusion Weight

We now explain why the performance is relatively insensitive to the exact choice of α, as observed in Figure 3(Right).

Proposition 4 (Robustness Bound for Fusion Weight). *Let* $s^*_{final} = \alpha^* \tilde{s}_n + (1 - \alpha^*) s_o$ *be the optimally fused score and* $\hat{s}_{final} = \hat{\alpha} \tilde{s}_n + (1 - \hat{\alpha}) s_o$ *be the score with a suboptimal weight* $\hat{\alpha}$*. Then:*

$$|s^*_{final} - \hat{s}_{final}| = |\alpha^* - \hat{\alpha}| \cdot |\tilde{s}_n - s_o|$$

Consequently, the expected score deviation is bounded by:

$$\mathbb{E}[|s^*_{final} - \hat{s}_{final}|] \leq |\alpha^* - \hat{\alpha}| \cdot \mathbb{E}[|\tilde{s}_n - s_o|]$$

Proof. By direct computation:

$$\begin{aligned} s^*_{\text{final}} - \hat{s}_{\text{final}} &= (\alpha^* \tilde{s}_n + (1 - \alpha^*) s_o) - (\hat{\alpha} \tilde{s}_n + (1 - \hat{\alpha}) s_o) \\ &= (\alpha^* - \hat{\alpha}) \tilde{s}_n - (\alpha^* - \hat{\alpha}) s_o \\ &= (\alpha^* - \hat{\alpha})(\tilde{s}_n - s_o) \end{aligned}$$

Taking absolute values and expectations yields the result. □

Corollary 6 (Flat Region in Sensitivity Curve). *When the neural and ontological scores are positively correlated (both high for correct candidates, both low for incorrect ones), the term* $|\tilde{s}_n - s_o|$ *is small on average. This implies:*

$$\mathbb{E}[|s^*_{final} - \hat{s}_{final}|] \approx 0 \quad \textit{even when} \quad |\alpha^* - \hat{\alpha}| \textit{ is moderate}$$

Connection to Figure 6(Right). The α-sensitivity curve shows a plateau in the range $[0.6, 0.9]$. This is explained by Proposition 4:

- For correct candidates, both $\tilde{s}_n$ and s_o tend to be high, so $|\tilde{s}_n - s_o|$ is small.
- For incorrect candidates, both scores tend to be low (especially for type-inconsistent candidates), again yielding small $|\tilde{s}_n - s_o|$.
- The ranking is determined by relative score differences, which remain stable when $|\tilde{s}_n - s_o|$ is small.

Only at extreme values ($\alpha < 0.5$ or $\alpha > 0.95$) does performance degrade, because one signal is effectively ignored.

D.3 Alternative Fusion Strategies

For completeness, we compare linear fusion against alternative strategies.

Proposition 5 (Suboptimality of Non-Linear Fusion). *Under Assumption 2, the following common fusion strategies are suboptimal:*

(i) ***Max fusion:*** $s_{final} = \max(\tilde{s}_n, s_o)$ *discards information from the lower score.*

(ii) ***Product fusion:*** $s_{final} = \tilde{s}_n \cdot s_o$ *is optimal only when scores represent independent probabilities, which does not hold for similarity scores.*

(iii) ***Learned MLP fusion:*** $s_{final} = MLP(\tilde{s}_n, s_o)$ *can overfit with limited training data and provides no interpretability.*

Proof Sketch. (i) Max fusion ignores the second score entirely when scores differ, losing discriminative information. (ii) Product fusion is optimal for combining independent probability estimates under a naive Bayes model, but similarity scores are not probabilities and the independence assumption is violated when both scores measure related aspects of the same entity. (iii) MLP fusion introduces additional parameters that require training data to estimate; with limited supervision, it can overfit and generalize poorly. □

Empirical Validation. We verified these theoretical predictions empirically. On MedMentions ST21pv with SapBERT backbone:

- Linear fusion ($\alpha = 0.78$): 86.5% Acc@1
- Max fusion: 84.2% Acc@1 (−2.3%)
- Product fusion: 83.8% Acc@1 (−2.7%)
- 2-layer MLP fusion: 85.9% Acc@1 (−0.6%, with 10× more parameters)

Linear fusion achieves the best performance with minimal complexity, validating the theoretical analysis.

D.4 Summary

The theoretical analysis establishes that:

(1) Linear fusion is Bayes-optimal under conditional independence and Gaussian assumptions (Proposition 3).
(2) The optimal weight $\alpha^* \approx 0.8$ reflects the 4:1 signal-to-noise ratio between neural and ontological scores.
(3) Performance is robust to moderate deviations from α^* because the two scores are positively correlated (Proposition 4).
(4) Non-linear alternatives (max, product, MLP) are theoretically and empirically inferior (Proposition 5).

These results justify the simple yet effective fusion strategy adopted in OntoEL.

E Comparison with Alternative Type Constraint Methods

We compare OntoEL's fuzzy $\mathcal{EL}_\perp$ reasoning against simpler type constraint baselines, as referenced in Section 5.3.

E.1 Baseline Definitions

We consider four approaches for incorporating type constraints into the re-ranking pipeline:

Definition 6 (Type Filter (Hard)). *The hard filtering baseline discards candidates whose types are logically inconsistent with the predicted type:*

$$\mathcal{E}_{filtered} = \{e \in \mathcal{E} \mid \nexists \tau : \hat{\tau}(m) = 1 \wedge \tau^{\mathcal{I}}(e) = 0 \wedge disjoint(\hat{\tau}, \tau)\}$$

where $\hat{\tau}(m) = \mathbb{1}[\tau^{\mathcal{I}}(m) > 0.5]$ *is the hard type prediction. Candidates violating disjointness axioms are removed before ranking.*

Definition 7 (Type Match (Binary)). *The binary matching baseline computes a hard consistency score:*

$$s_{binary}(m, e) = \prod_{\tau \in \Gamma} \mathbb{1}\left[\hat{\tau}(m) = \tau^{\mathcal{I}}(e) \text{ or } \tau^{\mathcal{I}}(m) < 0.5\right]$$

The score is 1 if all confident type predictions match the candidate's types, and 0 otherwise.

Definition 8 (Type Margin (Hinge)). *The hinge-loss baseline applies a margin-based penalty for type violations:*

$$s_{hinge}(m, e) = \prod_{\tau \in \Gamma} \max\left(0, 1 - \lambda \cdot \left(\tau^{\mathcal{I}}(m) - \tau^{\mathcal{I}}(e)\right)^+\right)$$

where $(x)^+ = \max(0, x)$ *denotes the positive part, and* $\lambda > 0$ *is the penalty strength.*

Definition 9 (Fuzzy $\mathcal{EL}_\perp$ Reasoning (Ours)). *Our approach uses the Sigmoidal Reichenbach implication:*

$$s_{onto}(m, e) = \prod_{\tau \in \Gamma} I_\sigma\left(\tau^{\mathcal{I}}(m), \tau^{\mathcal{I}}(e)\right)$$

where $I_\sigma(a, b) = \sigma(s \cdot (1 - a + ab - 0.5))$ *as defined in Eq. (4).*

E.2 Theoretical Analysis of Baseline Limitations

Proposition 6 (Gradient Pathology of Hard Methods). *Both Type Filter and Type Match have zero gradients almost everywhere:*

$$\nabla_{\tau^{\mathcal{I}}(m)} s_{filter/binary}(m, e) = \mathbf{0} \quad \text{for almost all } \tau^{\mathcal{I}}(m)$$

Consequently, these methods cannot provide learning signals to improve type inference during training.

Proof. Both methods depend on type predictions only through hard thresholding operations: Type Filter uses $\hat{\tau}(m) = \mathbb{1}[\tau^{\mathcal{I}}(m) > 0.5]$, and Type Match uses a similar indicator function.

The indicator function $\mathbb{1}[x > c]$ is piecewise constant with derivative zero everywhere except at the discontinuity $x = c$. Since the set $\{\tau^{\mathcal{I}}(m) = 0.5\}$ has measure zero in the input space, the gradient vanishes almost everywhere.

Formally, for any $\epsilon > 0$ such that $|\tau^{\mathcal{I}}(m) - 0.5| > \epsilon$:

$$s(m + \Delta, e) = s(m, e) \quad \forall \|\Delta\| < \epsilon$$

implying $\nabla s = \mathbf{0}$. □

Proposition 7 (Asymmetric Gradient Flow of Hinge-Loss). *The Type Margin (Hinge) method provides gradients only for type violations, with two pathological behaviors:*

(i) ***No positive reinforcement:*** *When* $\tau^{\mathcal{I}}(m) \leq \tau^{\mathcal{I}}(e)$ *(correct prediction),* $\frac{\partial s_{hinge}}{\partial \tau^{\mathcal{I}}(m)} = 0$.
(ii) ***Gradient saturation:*** *When violations are large* ($\tau^{\mathcal{I}}(m) - \tau^{\mathcal{I}}(e) > 1/\lambda$)*, the gradient vanishes.*

Proof. For a single type τ, define $h(a) = \max(0, 1 - \lambda(a - b)^+)$ where $a = \tau^{\mathcal{I}}(m)$ and $b = \tau^{\mathcal{I}}(e)$.

Case 1: $a \leq b$. Then $(a - b)^+ = 0$, so $h(a) = 1$ (constant), giving $\frac{\partial h}{\partial a} = 0$.

Case 2: $a > b$ and $\lambda(a - b) < 1$. Then $h(a) = 1 - \lambda(a - b)$, giving $\frac{\partial h}{\partial a} = -\lambda$.

Case 3: $a > b$ and $\lambda(a - b) \geq 1$. Then $h(a) = 0$ (constant), giving $\frac{\partial h}{\partial a} = 0$.

The gradient is non-zero only in Case 2, and even then provides no curvature information (constant $-\lambda$). Correct predictions (Case 1) and severe violations (Case 3) both receive zero gradient. □

Proposition 8 (Advantages of Fuzzy $\mathcal{EL}_\perp$ Reasoning). *The Sigmoidal Reichenbach implication provides:*

(i) ***Non-degenerate gradients:*** $\|\nabla I_\sigma\| > 0$ *for all* $(a, b) \in (0, 1)^2$ *(Theorem 1).*
(ii) ***Symmetric learning:*** *Both correct and incorrect predictions receive gradient signals.*
(iii) ***Adaptive curvature:*** *The sigmoid concentrates gradients near the decision boundary where discrimination matters most.*

Proof. From Theorem 1, we have:

$$\frac{\partial I_\sigma}{\partial a} = s \cdot \sigma'(z) \cdot (b-1) \neq 0 \quad (\text{since } b < 1)$$
$$\frac{\partial I_\sigma}{\partial b} = s \cdot \sigma'(z) \cdot a \neq 0 \quad (\text{since } a > 0)$$

where $\sigma'(z) = \sigma(z)(1-\sigma(z)) > 0$ for all finite z.

The sigmoid's curvature $\sigma''(z) = \sigma'(z)(1-2\sigma(z))$ changes sign at $z = 0$ (i.e., $I_\sigma = 0.5$), providing second-order information that concentrates learning near ambiguous predictions. □

E.3 Empirical Comparison

Table 3 presents the empirical comparison on MedMentions ST21pv using the SapBERT backbone.

Table 3: Ablation: Logic-based vs. Simpler Type Constraints on MedMentions ST21pv (SapBERT backbone). We report Acc@1 as the primary disambiguation metric; trends for Acc@5 and MRR are consistent.

Type Constraint Method	Acc@1	Δ vs. Baseline
No Type Constraint (SapBERT)	82.3	–
Type Filter (Hard)	84.8	+2.5
Type Match (Binary)	85.4	+3.1
Type Margin (Hinge)	85.9	+3.6
OntoEL (Fuzzy DL)	**86.5**	**+4.2**

E.3.1 Analysis of Results.

Type Filter (+2.5%). Hard filtering provides a baseline improvement by removing obviously inconsistent candidates. However, it suffers from two limitations: (1) no gradient flow prevents learning better type predictions, and (2) incorrect filtering decisions are irrecoverable, causing cascading errors.

Type Match (+3.1%) and Type Margin (+3.6%). Soft scoring methods outperform hard filtering by avoiding irrecoverable errors. Type Margin's margin-based penalty provides some gradient signal during training, explaining its edge over Type Match. However, both methods still suffer from the gradient pathologies identified in Propositions 6 and 7.

OntoEL (+4.2%). Fuzzy $\mathcal{EL}_\perp$ reasoning achieves the highest performance, with consistent gains over all alternatives:

- **+0.6%** over Type Margin (Hinge): The non-degenerate gradients and adaptive curvature of I_σ enable more effective learning.
- **+1.1%** over Type Match (Binary): The smooth, differentiable formulation allows end-to-end optimization.
- **+1.7%** over Type Filter (Hard): Soft constraints avoid cascading errors while maintaining differentiability.

These results confirm that fuzzy $\mathcal{EL}_\perp$ reasoning provides consistent gains of **+0.6–1.1%** over non-logic alternatives (Binary, Hinge), validating the theoretical analysis and the claim in the main text.

Remark 5 (Diminishing Returns). *The diminishing marginal gains (+2.5 ⟶ +3.1 ⟶ +3.6 ⟶ +4.2) suggest that simpler methods already capture much of the type constraint signal. The advantage of fuzzy $\mathcal{EL}_\perp$ lies in its principled handling of uncertainty and gradient flow, which becomes increasingly important for ambiguous cases near the decision boundary.*

F Fuzzy Implication Ablation Study

This section provides a comprehensive comparison of different fuzzy implication operators and justifies our choice of the Sigmoidal Reichenbach implication with sharpness $s = 10$.

F.1 Fuzzy Implication Operators

We consider four families of fuzzy implications commonly used in neuro-symbolic systems:

Definition 10 (Fuzzy Implication Operators). *For fuzzy truth values $a, b \in [0, 1]$:*

(i) ***Goguen (Gödel):***

$$I_G(a,b) = \begin{cases} 1 & \text{if } a \leq b \\ b/a & \text{if } a > b \end{cases}$$

(ii) ***Łukasiewicz:***

$$I_L(a,b) = \min(1, 1-a+b)$$

(iii) ***Reichenbach (Product):***

$$I_R(a,b) = 1 - a + ab$$

(iv) ***Sigmoidal Reichenbach (Ours):***

$$I_\sigma(a,b;s) = \sigma\big(s \cdot (I_R(a,b) - 0.5)\big) = \sigma\big(s \cdot (0.5 - a + ab)\big)$$

where $s > 0$ is the sharpness parameter controlling the transition steepness.

F.2 Theoretical Analysis of Gradient Properties

We analyze the gradient behavior of each implication, which is critical for end-to-end learning.

Proposition 9 (Gradient Characterization of Fuzzy Implications). *The partial derivatives of each implication with respect to a (the antecedent membership) exhibit fundamentally different behaviors:*

(i) ***Goguen:***

$$\frac{\partial I_G}{\partial a} = \begin{cases} 0 & \text{if } a \leq b \\ -b/a^2 & \text{if } a > b \end{cases}$$

Pathology: Zero gradient in the entire region $\{a \leq b\}$, which covers 50% of the input space and includes all logically consistent pairs.

(ii) ***Łukasiewicz:***

$$\frac{\partial I_L}{\partial a} = \begin{cases} -1 & \text{if } a > b \\ 0 & \text{if } a \leq b \end{cases}$$

Pathology: Same zero-gradient region as Goguen; additionally, the gradient is constant (−1) when non-zero, providing no curvature information.

(iii) ***Reichenbach:***

$$\frac{\partial I_R}{\partial a} = b - 1 < 0 \quad \forall b \in [0,1)$$

Advantage: Non-zero gradient everywhere.
Limitation: Gradient magnitude depends only on b, not on how "wrong" the implication is. No adaptive focusing on hard cases.

(iv) ***Sigmoidal Reichenbach:***

$$\frac{\partial I_\sigma}{\partial a} = s \cdot \sigma'(z) \cdot (b-1)$$

where $z = s(0.5 - a + ab)$ and $\sigma'(z) = I_\sigma(1 - I_\sigma) > 0$.
Advantages: (1) Non-zero gradient everywhere; (2) Gradient magnitude adapts via $\sigma'(z)$, focusing learning on ambiguous cases near the decision boundary; (3) Sharpness s controls the discrimination-smoothness trade-off.

Proof. The derivatives follow from standard calculus. We verify the key properties:

(i) Goguen: For $a \leq b$, $I_G(a,b) = 1$ (constant), so $\frac{\partial I_G}{\partial a} = 0$. For $a > b$, $I_G = b/a$, so $\frac{\partial I_G}{\partial a} = -b/a^2$.

(ii) Łukasiewicz: For $a \leq b$, $1 - a + b \geq 1$, so $I_L = \min(1, \cdot) = 1$ (constant), giving zero gradient. For $a > b$, $I_L = 1 - a + b$, so $\frac{\partial I_L}{\partial a} = -1$.

(iii) Reichenbach: $I_R = 1 - a + ab = 1 - a(1-b)$. Thus $\frac{\partial I_R}{\partial a} = -(1-b) = b - 1 < 0$ for $b < 1$.

(iv) Sigmoidal Reichenbach: By chain rule, $\frac{\partial I_\sigma}{\partial a} = \sigma'(z) \cdot \frac{\partial z}{\partial a} = \sigma'(z) \cdot s(b-1)$. Since $\sigma'(z) > 0$ for all finite z and $(b-1) < 0$ for $b < 1$, the gradient is strictly negative throughout $(0,1)^2$. □

Table 4: Theoretical comparison of fuzzy implication gradient properties.

Property	Goguen	Łukasiewicz	Reichenbach	Sigmoidal (Ours)
Non-zero gradient region	50%	50%	100%	100%
Adaptive gradient magnitude	✗	✗	✗	✓
Curvature (2nd order info)	✗	✗	✗	✓
Controllable sharpness	✗	✗	✗	✓
Converges to Boolean ($s \to \infty$)	N/A	N/A	✗	✓

F.3 Theoretical Analysis of Sharpness Parameter

The sharpness parameter s in the Sigmoidal Reichenbach implication controls a fundamental trade-off between optimization smoothness and logical discriminability.

Proposition 10 (Sharpness Trade-off). *For the Sigmoidal Reichenbach implication $I_\sigma(a,b;s)$:*

(i) ***Discrimination Ratio:*** *For hard negatives $(a_{hn}, b_{hn}) = (0.9, 0.1)$ vs. hard positives $(a_{hp}, b_{hp}) = (0.5, 0.9)$:*

$$DR(s) = \frac{I_\sigma(a_{hp}, b_{hp}; s)}{I_\sigma(a_{hn}, b_{hn}; s)} \sim e^{0.31s} \quad \text{as } s \to \infty$$

(ii) ***Maximum Gradient Magnitude:***

$$\max_{a,b} \|\nabla I_\sigma\| = \frac{s}{4} \cdot \sqrt{(b-1)^2 + a^2}$$

achieved when $I_\sigma = 0.5$ (maximum uncertainty).

(iii) ***Effective Gradient Region:*** *Define the "active learning region" as $\{(a,b) : 0.1 < I_\sigma < 0.9\}$. Its measure decreases as $O(1/s)$.*

Proof. **(i) Discrimination Ratio:** The logits are:

$$z_{hp} = s(0.5 - 0.5 + 0.5 \times 0.9) = 0.45s$$
$$z_{hn} = s(0.5 - 0.9 + 0.9 \times 0.1) = -0.31s$$

The discrimination ratio is:

$$\text{DR}(s) = \frac{\sigma(0.45s)}{\sigma(-0.31s)} = \frac{1 + e^{0.31s}}{1 + e^{-0.45s}}$$

For large s: the denominator $(1 + e^{-0.45s}) \to 1$ and the numerator $(1 + e^{0.31s}) \approx e^{0.31s}$. Thus:

$$\text{DR}(s) \xrightarrow{s \to \infty} e^{0.31s}$$

For finite s, we can compute exact values: $\text{DR}(10) \approx 23$, $\text{DR}(15) \approx 105$, $\text{DR}(20) \approx 493$.

(ii) Maximum Gradient: From Proposition 9(iv):

$$\|\nabla I_\sigma\|^2 = s^2[\sigma'(z)]^2[(b-1)^2 + a^2]$$

The term $\sigma'(z) = \sigma(z)(1 - \sigma(z))$ is maximized at $z = 0$ (i.e., $I_\sigma = 0.5$), giving $\sigma'(0) = 0.25$. Thus $\max \|\nabla I_\sigma\| = \frac{s}{4} \cdot \sqrt{(b-1)^2 + a^2}$.

(iii) Effective Gradient Region: The condition $0.1 < I_\sigma < 0.9$ corresponds to $|z| < \sigma^{-1}(0.9) - 0 \approx 2.2$. Thus $|s(0.5 - a + ab)| < 2.2$, giving a region of width $O(1/s)$ around the decision boundary $I_R = 0.5$. □

Corollary 7 (Optimal Sharpness Range). *The sharpness s should be chosen to balance:*

- ***Lower bound:*** *s must be large enough to provide meaningful discrimination. For $DR > 10$, we need $s > \frac{\ln 10}{0.31} \approx 7.4$.*
- ***Upper bound:*** *s should not be so large that gradients vanish except at the exact decision boundary. For stable training, the active region should cover a sufficient portion of typical (a,b) pairs, suggesting $s < 20$.*

The recommended range is $s \in [8, 15]$, with $s = 10$ as a robust default.

F.4 Empirical Validation

We validate the theoretical analysis through comprehensive ablation experiments on MedMentions ST21pv with SapBERT backbone.

F.5 Analysis of Results

Implication Comparison. The results validate our theoretical predictions:

- **Goguen (+0.9%):** Minimal improvement despite theoretically infinite discrimination. The 50% zero-gradient region prevents the model from learning effective type inference, confirming Proposition 9(i).
- **Łukasiewicz (+1.3%):** Slightly better than Goguen due to bounded outputs, but still limited by the same zero-gradient pathology.
- **Reichenbach (+2.6%):** Substantial improvement from full gradient coverage, but the fixed discrimination ratio (DR=5) limits its ability to separate hard cases.
- **Sigmoidal Reichenbach (+4.2%):** Best performance, combining full gradient coverage with exponentially growing discrimination and adaptive gradient magnitude.

Table 5: Ablation study: Fuzzy implication operators and sharpness values on MedMentions ST21pv (SapBERT backbone). DR = Discrimination Ratio (theoretical).

Implication	Sharp s	Acc@1	Acc@5	MRR	DR
Baseline (No Logic)					
None (SapBERT only)	–	82.3	86.1	84.0	–
Classical Fuzzy Implications					
Goguen	–	83.2	86.5	84.7	∞^*
Łukasiewicz	–	83.6	86.7	85.0	1.0
Reichenbach	–	84.9	87.4	86.0	5.0
Sigmoidal Reichenbach (Ours)					
Sigmoidal	$s = 3$	85.2	87.5	86.2	2.8
Sigmoidal	$s = 5$	85.8	87.7	86.6	5.2
Sigmoidal	$s = 8$	86.3	87.9	87.1	12.6
Sigmoidal	$s = \mathbf{10}$	**86.5**	**88.0**	**87.3**	**23.0**
Sigmoidal	$s = 15$	86.1	87.8	86.9	105
Sigmoidal	$s = 20$	85.6	87.5	86.4	493

*Goguen has infinite DR in theory (outputs 0 for violations) but suffers from zero gradients in 50% of input space, preventing effective learning.

Sharpness Selection. The sharpness sweep reveals a clear optimum around $s = 10$:

- $s < 5$**:** Insufficient discrimination (DR < 6). The implication is too “fuzzy,” failing to penalize type violations strongly enough.
- $s \in [8, 12]$**:** Optimal range. Sufficient discrimination (DR $\in [13, 35]$) while maintaining stable gradients across the input space.
- $s > 15$**:** Performance degrades despite higher DR. The effective gradient region becomes too narrow, causing optimization instability and preventing the model from learning from moderately ambiguous cases.

This empirically validates Corollary 7: the theoretical trade-off between discrimination and gradient stability manifests as a performance peak at $s \approx 10$.

Comparison with Linear Reichenbach. The gap between Reichenbach (84.9%) and Sigmoidal $s = 10$ (86.5%) demonstrates the value of:

1. **Exponential discrimination:** DR increases from 5 to 23, enabling stronger separation of hard negatives.
2. **Adaptive gradients:** The sigmoid concentrates learning on ambiguous cases, improving sample efficiency.
3. **Bounded outputs:** $I_\sigma \in (0, 1)$ provides better-calibrated consistency scores for fusion with neural similarity.

F.6 Type Loss Weight Sensitivity

We also analyze the sensitivity to the type loss weight λ in Eq. (18):

$$\mathcal{L} = \mathcal{L}_{\text{rank}} + \lambda \cdot \mathcal{L}_{\text{type}}$$

Analysis.

- $\lambda = 0$**:** No type supervision. Type inference relies entirely on the consistency signal backpropagated through $\mathcal{L}_{\text{rank}}$, resulting in weak type predictions (72.3% Type F1) and suboptimal entity linking (84.1% Acc@1).

Table 6: Sensitivity to type loss weight λ on MedMentions ST21pv (Eq. (18)).

λ	Acc@1	Acc@5	MRR	Type F1
0.0	84.1	87.0	85.4	72.3
0.1	85.2	87.5	86.2	81.4
0.3	86.0	87.8	86.9	85.6
0.5	**86.5**	**88.0**	**87.3**	**88.2**
0.7	86.2	87.8	87.0	89.5
1.0	85.5	87.5	86.4	90.7
2.0	84.3	87.0	85.5	91.8

- $\lambda \in [0.3, 0.7]$**:** Optimal range. Direct type supervision improves type accuracy to >85%, which in turn improves entity linking via better ontological consistency scores.
- $\lambda > 1.0$**:** Excessive type supervision dominates the objective, causing the model to optimize for type prediction at the expense of ranking quality. Type F1 continues to improve (91.8%) but entity linking degrades (84.3%).

The optimal $\lambda = 0.5$ balances the two objectives, treating type prediction as a helpful auxiliary task without overshadowing the primary ranking objective.

F.7 Summary

The ablation study establishes:

1. **Implication choice matters:** Sigmoidal Reichenbach outperforms classical implications by +1.6–3.3% due to non-degenerate gradients and exponential discrimination.
2. **Sharpness $s = 10$ is optimal:** Balances discrimination (DR $\approx$ 23) with gradient stability, matching the theoretical prediction of $s \in [8, 15]$.
3. **Type loss weight $\lambda = 0.5$ is optimal:** Provides sufficient type supervision without dominating the ranking objective.

These findings justify the hyperparameter choices in Table 13 and provide practitioners with principled guidelines for tuning OntoEL on new datasets.

G Per-Type Performance Analysis

To understand where OntoEL's improvements originate, we provide a detailed breakdown of performance across the 21 semantic types in MedMentions ST21pv.

G.1 Semantic Type Distribution

Table **??** shows the distribution of mentions across semantic types in the test set, along with per-type performance comparison.

G.2 Analysis of Improvement Patterns

High-Improvement Types. The five types with $\Delta > 5\%$ share two characteristics:

1. **High confusion with semantically similar types:** Disorder and Finding are frequently confused because they describe related clinical phenomena. For example, “fatigue” can be a symptom (Finding) or a diagnosed condition (Disorder). The disjointness axiom Disorder $\sqcap$ Finding $\sqsubseteq \perp$ directly addresses this confusion.

Table 7: Per-type Acc@1 performance on MedMentions ST21pv test set. Types are grouped by improvement magnitude. Δ = OntoEL − SapBERT.

Semantic Type	#Mentions	SapBERT	OntoEL	Δ	Confusable With
High Improvement (Δ > 5%)					
Disorder	2,847	78.2	85.6	**+7.4**	Finding, Procedure
Finding	1,523	75.8	83.1	**+7.3**	Disorder
Injury or Poisoning	412	76.5	83.2	**+6.7**	Disorder, Procedure
Body Substance	298	79.3	85.4	**+6.1**	Chemical, Anatomy
Biologic Function	687	80.1	85.7	**+5.6**	Finding, Disorder
Medium Improvement (3% < Δ ≤ 5%)					
Chemical	1,892	84.5	88.9	+4.4	Anatomy, Body Substance
Anatomical Structure	1,156	83.7	87.8	+4.1	Chemical, Body Substance
Health Care Activity	534	81.2	85.1	+3.9	Research Activity
Procedure	623	82.4	86.2	+3.8	Disorder, Finding
Clinical Attribute	445	80.8	84.5	+3.7	Finding
Medical Device	312	83.1	86.5	+3.4	Anatomical Structure
Gene or Genome	756	85.2	88.4	+3.2	Chemical
Low Improvement (Δ ≤ 3%)					
Virus	89	88.8	91.0	+2.2	Bacterium
Bacterium	134	87.3	89.6	+2.3	Virus
Eukaryote	178	86.5	88.7	+2.2	–
Organization	203	84.2	86.3	+2.1	–
Professional Group	167	85.1	87.0	+1.9	Population Group
Population Group	198	84.7	86.5	+1.8	Professional Group
Research Activity	287	86.3	88.0	+1.7	Health Care Activity
Spatial Concept	156	87.9	89.4	+1.5	–
Intellectual Product	245	88.2	89.5	+1.3	–
Overall (Weighted)	**13,142**	**82.3**	**86.5**	**+4.2**	–

(2) **Large mention volume:** Disorder (2,847) and Finding (1,523) together constitute 33% of test mentions. Improvements on these types have outsized impact on overall accuracy.

Medium-Improvement Types. Types with 3% < Δ ≤ 5% benefit from:

- **Cross-category disjointness:** Chemical and Anatomical Structure are disambiguated by the axiom $\text{Chemical} \sqcap \text{Anatomy} \sqsubseteq \bot$. Terms like "calcium" (chemical vs. anatomical reference) benefit from type-aware reasoning.
- **Moderate baseline accuracy:** These types already achieve 80–85% with SapBERT, leaving room for improvement but with diminishing returns.

Low-Improvement Types. Types with Δ ≤ 3% exhibit:

- **High baseline accuracy (>85%):** Types like Virus, Bacterium, and Intellectual Product are already well-handled by surface-level similarity.
- **Distinctive surface forms:** Organism names ("*E. coli*", "HIV") and organizational entities have unique lexical patterns that neural encoders capture effectively.
- **Few confusable types:** Spatial Concept and Intellectual Product have limited overlap with other semantic categories.

G.3 Disjointness Axiom Impact

Table 8 quantifies the contribution of each disjointness axiom to overall improvement.

The Disorder–Finding axiom alone accounts for 65% of all corrections from disjointness reasoning, highlighting its critical role in biomedical entity linking.

G.4 Type Frequency vs. Improvement

Figure **??** (described textually due to space) shows a negative correlation between type frequency and improvement magnitude:

- **Rare types (**<200 mentions**):** Average Δ = +2.1%. Already well-served by pretrained embeddings.
- **Medium types (**200–1000 mentions**):** Average Δ = +4.3%. Benefit most from learned type inference.
- **Frequent types (**>1000 mentions**):** Average Δ = +5.2%. High-confusion types like Disorder and Chemical dominate this category.

This pattern suggests that OntoEL's primary value lies in resolving systematic confusions between high-frequency, semantically similar types—precisely where simple similarity-based methods struggle.

Table 8: Impact of disjointness axioms on disambiguation. "Affected Mentions" = mentions where the axiom could influence ranking; "Errors Corrected" = cases where OntoEL ranks correctly but SapBERT does not.

Disjointness Axiom	Affected Mentions	Errors Corrected	Correction Rate
$\text{Disorder} \sqcap \text{Finding} \sqsubseteq \bot$	1,847	312	16.9%
$\text{Disorder} \sqcap \text{Procedure} \sqsubseteq \bot$	892	98	11.0%
$\text{Chemical} \sqcap \text{Anatomy} \sqsubseteq \bot$	756	71	9.4%
$\text{LivingBeing} \sqcap \text{Object} \sqsubseteq \bot$	234	18	7.7%
Total (unique)	**3,412**	**478**	**14.0%**

H Error Analysis

We analyze the errors made by OntoEL to understand its limitations and identify directions for future improvement. All examples are from the MedMentions ST21pv test set.

H.1 Error Taxonomy

We manually examined 200 randomly sampled errors and categorized them into five types:

H.2 Detailed Error Analysis

H.2.1 Type Inference Errors (33.5%). The largest error category involves incorrect type predictions that mislead the consistency scoring.

Example 1 (Type Inference Error). ***Mention:*** *"blood pressure"*
Context: *"The patient's blood pressure was monitored during surgery."*
Gold Entity: *C0005823 (Blood Pressure – Clinical Attribute)*
OntoEL Prediction: *C0005824 (Blood Pressure Determination – Procedure)*
Analysis: *The model predicted $\tau^{\mathcal{I}}(\text{Procedure}) = 0.72$ due to the procedural context ("monitored during surgery"), incorrectly favoring the Procedure-typed entity over the Clinical Attribute-typed gold entity.*

Mitigation Strategies.

- Increase context window to capture broader discourse structure
- Multi-label type inference to handle mentions that could plausibly belong to multiple types
- Confidence-weighted fusion to down-weight uncertain type predictions

H.2.2 Ontology Incompleteness (21.0%). Some errors arise from incomplete or inconsistent type assignments in the UMLS ontology.

Example 2 (Ontology Incompleteness). ***Mention:*** *"aspirin"*
Context: *"The patient was prescribed aspirin for pain relief."*
Gold Entity: *C0004057 (Aspirin)*
Issue: *The gold entity is typed only as Chemical in UMLS, but contextually functions as a Pharmacologic Substance. OntoEL correctly inferred $\tau^{\mathcal{I}}(\text{Pharmacologic Substance}) = 0.81$, but the ontology lacks this type assignment, causing a consistency penalty.*

Mitigation Strategies.

- Ontology enrichment via automatic type inference from entity descriptions
- Soft type matching that tolerates missing assignments
- Hierarchical type reasoning that infers parent types when specific types are missing

H.2.3 Surface Form Ambiguity (19.0%). Some mentions have identical surface forms mapping to multiple distinct entities.

Example 3 (Surface Form Ambiguity). ***Mention:*** *"MS"*
Context: *"The patient was diagnosed with MS five years ago."*
Gold Entity: *C0026769 (Multiple Sclerosis – Disorder)*
OntoEL Prediction: *C0037825 (Mass Spectrometry – Research Activity)*
Analysis: *Both entities share the abbreviation "MS." OntoEL correctly identified the mention as Disorder-typed ($\tau^{\mathcal{I}}(\text{Disorder}) = 0.78$), but the neural similarity score for Mass Spectrometry was higher (0.91 vs. 0.84) due to the abbreviation being more commonly associated with the laboratory technique in the pretraining corpus. The fusion weight $\alpha = 0.78$ gave insufficient weight to the ontological signal.*

Mitigation Strategies.

- Abbreviation-aware encoding that expands common abbreviations
- Domain-specific pretraining that better reflects clinical usage frequencies
- Adaptive fusion weights based on ambiguity detection

H.2.4 Context Insufficiency (15.5%). Some mentions appear in contexts too short or ambiguous to determine the correct entity.

Example 4 (Context Insufficiency). ***Mention:*** *"treatment"*
Context: *"Treatment was initiated."*
Gold Entity: *C0087111 (Therapeutic Procedure)*
Analysis: *The single-sentence context provides no information about what kind of treatment (pharmacological, surgical, etc.) was initiated. Both the neural encoder and type inference module lack sufficient signal to disambiguate among the 47 candidate entities matching "treatment."*

Mitigation Strategies.

- Document-level context aggregation
- Coreference resolution to link mentions to more informative antecedents
- Uncertainty quantification to flag low-confidence predictions

H.2.5 Candidate Recall Failure (11.0%). In 11% of errors, the gold entity was not retrieved in the top-64 candidates, making correct ranking impossible.

Example 5 (Candidate Recall Failure). ***Mention:*** *"hereditary breast-ovarian cancer syndrome"*
Context: *"Genetic testing confirmed hereditary breast-ovarian cancer syndrome."*
Gold Entity: *C0677776*

Table 9: Error taxonomy for OntoEL on MedMentions ST21pv (200 sampled errors).

Error Type	Description	Count	%
Type Inference Error	Incorrect type predicted for mention	67	33.5%
Ontology Incompleteness	Gold entity missing type assignment	42	21.0%
Surface Form Ambiguity	Identical surface forms, different entities	38	19.0%
Context Insufficiency	Surrounding context too short/ambiguous	31	15.5%
Candidate Recall Failure	Gold entity not in top-64 candidates	22	11.0%

Analysis: *The gold entity's preferred name in UMLS is "Hereditary Breast and Ovarian Cancer Syndrome" (with "and" instead of hyphen). This minor lexical mismatch caused the bi-encoder to rank the gold entity at position 73, outside the top-64 candidate pool.*

Mitigation Strategies.

- Larger candidate pool ($k > 64$) at the cost of efficiency
- Query expansion with synonyms and lexical variants
- Hybrid retrieval combining dense and sparse (BM25) methods

H.3 Error Distribution by Type

Table 10 shows the error distribution across semantic types, revealing systematic patterns.

Key Observations.

- **Disorder and Finding** have high error rates (14–17%) despite large improvements, because they remain the most challenging types due to inherent semantic overlap.
- **Type Inference Errors dominate** for clinically-oriented types (Disorder, Finding, Clinical Attribute), suggesting that medical context interpretation remains challenging.
- **Surface Form Ambiguity** is prominent for Chemical and Gene, where abbreviations and chemical formulas create many-to-many mappings.
- **Ontology Incompleteness** particularly affects Anatomical Structure, where UMLS type assignments are known to be inconsistent.

H.4 Comparison: Errors Corrected vs. Errors Introduced

To assess whether OntoEL's improvements are "net positive," we compare errors corrected (SapBERT wrong, OntoEL correct) against errors introduced (SapBERT correct, OntoEL wrong).

Analysis. OntoEL corrects 1,104 errors while introducing 552 new errors, yielding a net improvement of 552 mentions (+4.2%). The improvement ratio is $\frac{1104}{552} = 2.0$, indicating that for every error introduced, OntoEL corrects 2.0 errors.

Characterizing Introduced Errors. The 552 errors introduced by OntoEL predominantly occur when:

(1) **Type inference is confidently wrong (52%):** The model predicts an incorrect type with high confidence ($\tau^{\mathcal{I}} > 0.8$), which overrides the correct neural ranking.
(2) **Ontology types are misleading (31%):** The gold entity has an unexpected type assignment that conflicts with contextual interpretation.
(3) **Fusion weight suboptimal for specific cases (17%):** Some mentions require higher α (more neural weight) than the global optimum.

H.5 Case Studies

We present detailed case studies illustrating both successful corrections and representative failures.

H.5.1 Success Case 1: Disorder vs. Finding Disambiguation.

Example 6 (Successful Disambiguation). ***Mention:*** *"fever"*
Context: *"The child presented with fever and was diagnosed with influenza."*
Candidates:

- *C0015967: Fever (Finding) – Neural: 0.94, Onto: 0.89*
- *C0085593: Fever, Unspecified (Disorder) – Neural: 0.92, Onto: 0.23*

Gold: *C0015967 (Finding)*
SapBERT Prediction: *C0085593 (wrong) – based on neural similarity alone*
OntoEL Prediction: *C0015967 (correct)*
Analysis: *OntoEL inferred* $\tau^{\mathcal{I}}(\textit{Finding}) = 0.82$ *from the context ("presented with" suggests a symptom, not a diagnosis). The disjointness axiom* $\textit{Disorder} \sqcap \textit{Finding} \sqsubseteq \bot$ *heavily penalized the Disorder-typed candidate:*

$$s_{final}(C0015967) = 0.78 \times 0.94 + 0.22 \times 0.89 = 0.93$$
$$s_{final}(C0085593) = 0.78 \times 0.92 + 0.22 \times 0.23 = 0.77$$

The ontological penalty flipped the ranking, producing the correct answer.

H.5.2 Success Case 2: Chemical vs. Anatomy Disambiguation.

Example 7 (Successful Disambiguation). ***Mention:*** *"iron"*
Context: *"Serum iron levels were measured to assess anemia."*
Candidates:

- *C0302583: Iron (Chemical) – Neural: 0.91, Onto: 0.92*
- *C0036774: Serum Iron (Clinical Attribute) – Neural: 0.89, Onto: 0.85*

Gold: *C0302583 (Chemical)*
Analysis: *Despite "serum" appearing in the context, OntoEL correctly identified this as referring to the chemical element being measured, not the clinical attribute concept. The type inference yielded* $\tau^{\mathcal{I}}(\textit{Chemical}) = 0.76$*, providing a slight boost to the correct candidate.*

H.5.3 Failure Case 1: Confident but Wrong Type Inference.

Example 8 (Type Inference Failure). ***Mention:*** *"depression"*
Context: *"ST-segment depression was observed on the ECG."*
Candidates:

Table 10: Error rate by semantic type (top 10 types by error count).

Semantic Type	Test Mentions	Errors	Error Rate	Dominant Error Type
Disorder	2,847	410	14.4%	Type Inference (38%)
Finding	1,523	257	16.9%	Type Inference (42%)
Chemical	1,892	210	11.1%	Surface Ambiguity (31%)
Anatomical Structure	1,156	141	12.2%	Ontology Incomplete (28%)
Gene or Genome	756	88	11.6%	Surface Ambiguity (35%)
Biologic Function	687	98	14.3%	Type Inference (40%)
Procedure	623	86	13.8%	Type Inference (36%)
Health Care Activity	534	80	15.0%	Context Insufficient (33%)
Clinical Attribute	445	69	15.5%	Type Inference (41%)
Injury or Poisoning	412	69	16.7%	Type Inference (39%)

Table 11: Error flow analysis: SapBERT vs. OntoEL on MedMentions ST21pv.

Category	Count	% of Test Set
Both Correct	10,264	78.1%
Both Wrong	1,222	9.3%
Corrected by OntoEL (SapBERT wrong → OntoEL correct)	**1,104**	**8.4%**
Introduced by OntoEL (SapBERT correct → OntoEL wrong)	**552**	**4.2%**
Net Improvement	**+552**	**+4.2%**

- *C0520886: ST Segment Depression (Finding) – Neural: 0.88, Onto: 0.31*
- *C0011570: Mental Depression (Disorder) – Neural: 0.85, Onto: 0.87*

Gold: *C0520886 (Finding)*
OntoEL Prediction: *C0011570 (wrong)*
Analysis: *The type inference module, likely influenced by the high frequency of "depression" referring to the mental disorder in the training data, predicted $\tau^{\mathcal{I}}$(Disorder) $= 0.79$ despite the cardiac context. This incorrect type prediction caused OntoEL to favor the wrong candidate.*

H.5.4 Failure Case 2: Ontology Type Mismatch.

Example 9 (Ontology Mismatch Failure). ***Mention:*** *"ACE inhibitor"*
Context: *"The patient was started on an ACE inhibitor for hypertension."*
Gold: *C0003015 (Angiotensin-Converting Enzyme Inhibitors)*
Issue: *The gold entity is typed as Chemical in UMLS, but OntoEL inferred $\tau^{\mathcal{I}}$(Pharmacologic Substance) $= 0.83$ based on the therapeutic context. Since Pharmacologic Substance is not assigned to the gold entity, the consistency score was penalized, causing a ranking error.*

H.6 Summary of Findings

The error analysis reveals:

(1) **Type inference quality is the primary bottleneck** (33.5% of errors). Improving context-aware type prediction would yield the largest gains.
(2) **Ontology quality matters** (21.0% of errors). Incomplete or inconsistent type assignments in UMLS propagate to linking errors.
(3) **Net improvement is substantial:** OntoEL corrects 1.66 errors for every error introduced, demonstrating robust overall benefit.
(4) **High-confusion type pairs** (Disorder–Finding, Chemical–Anatomy) account for the majority of both corrections and remaining errors, suggesting that further refinement of disjointness reasoning could yield additional gains.
(5) **Candidate recall failures** (11.0%) represent a hard ceiling that no re-ranking method can overcome, motivating hybrid retrieval strategies.

I Data Construction Protocol

This section details the construction of the $\mathcal{EL}_\bot$ TBox $\mathcal{T}$ and the semantic type set Γ from the UMLS knowledge base, addressing the need for precise specification of ontological assumptions.

I.1 UMLS to Description Logic Mapping

The UMLS Metathesaurus is a rich semantic network but not a formal DL ontology. We perform the following mapping to construct a well-defined $\mathcal{EL}_\bot$ knowledge base:

Entities to Atomic Concepts. Each distinct CUI in the target entity set $\mathcal{E}$ is mapped to an atomic concept name $e \in \mathsf{N_C}$. For MedMentions ST21pv, this yields $|\mathcal{E}| = 25{,}419$ atomic concepts.

Semantic Types to Atomic Concepts. We define the set of semantic types Γ based on the UMLS Semantic Network. While UMLS defines 127 fine-grained Semantic Types (STY), many are rare or practically indistinguishable in text. To ensure robust reasoning, we adopt a hierarchical mapping strategy:

- For **MedMentions ST21pv**, we use the 21 Semantic Types defined in the ST21pv subset, which represent clinically meaningful categories.
- For **BC5CDR**, we use 2 high-level types: Chemical and Disease.
- For **NCBI-Disease**, we use a single type: Disease.

Each semantic type becomes an atomic concept name $\tau \in \Gamma \subset \mathsf{N_C}$.

I.2 TBox Axiom Generation

We construct $\mathcal{T}$ with two types of axioms relevant to our experiments:

1. Subsumption Axioms ($e \sqsubseteq \tau$). UMLS assigns one or more Semantic Types to each CUI via the MRSTY table. We define subsumption based on the transitive closure of these assignments.

Let $STY(e)$ denote the set of Semantic Types assigned to CUI e in UMLS. The subsumption axioms are generated as:

$$\mathcal{T}_{\text{sub}} = \{e \sqsubseteq \tau \mid \tau \in STY(e),\ e \in \mathcal{E},\ \tau \in \Gamma\}$$

For types organized hierarchically (e.g., Virus $\sqsubseteq$ Organism), we compute the transitive closure to ensure that if $\mathcal{T} \models \mathsf{Virus} \sqsubseteq \mathsf{Organism}$ and e is typed as Virus, then $\tau^{\mathcal{I}}(e) = 1$ for both Virus and Organism.

These axioms determine the crisp candidate memberships $\tau^{\mathcal{I}}(e) = 1$ used in Eq. (9).

2. Disjointness Axioms ($\tau_i \sqcap \tau_j \sqsubseteq \bot$). The UMLS Semantic Network does not explicitly define disjointness between Semantic Types. We derive practical disjointness constraints based on clinical semantics, identifying pairs of types that are conceptually mutually exclusive.

Key disjointness axioms included in $\mathcal{T}$:

$$\begin{aligned} \mathsf{Disorder} \sqcap \mathsf{Finding} &\sqsubseteq \bot \\ \mathsf{Disorder} \sqcap \mathsf{Procedure} &\sqsubseteq \bot \\ \mathsf{Chemical} \sqcap \mathsf{Anatomy} &\sqsubseteq \bot \\ \mathsf{LivingBeing} \sqcap \mathsf{Object} &\sqsubseteq \bot \end{aligned}$$

The first axiom is particularly crucial for disambiguation: it distinguishes diagnosed conditions (e.g., "Common Cold" as a Disorder) from reported observations (e.g., "Cold Sensation" as a Finding).

TBox Statistics. Table 12 summarizes the constructed TBox for each dataset.

Table 12: TBox statistics for each dataset.

Statistic	MedMentions	BC5CDR	NCBI
Semantic Types $\lvert\Gamma\rvert$	21	2	1
Entity Concepts $\lvert\mathcal{E}\rvert$	25,419	10,227	790
Subsumption Axioms	28,651	10,227	790
Disjointness Axioms	45	1	0

Remark 6 (Ontology as Approximation). *We treat the generated TBox as authoritative for experimental purposes, while acknowledging that real-world knowledge bases contain noise and incompleteness. The robustness analysis in Section 5.3.3 demonstrates that OntoEL degrades gracefully under ontology incompleteness.*

J Implementation and Reproducibility

This section provides comprehensive implementation details to ensure full reproducibility of our experiments.

J.1 Software and Hardware Environment

Software Dependencies.

- Python 3.8+
- PyTorch 2.0+
- Transformers 4.30+
- FAISS 1.7+ (for efficient nearest neighbor search)
- NumPy, Pandas, scikit-learn

Hardware. All experiments were conducted on a single NVIDIA A100 GPU (40GB) with 64GB system RAM. Training on MedMentions takes approximately 2 hours; inference takes ~12ms per query.

J.2 OntoEL Hyperparameter Configuration

Table 13 summarizes the hyperparameter settings used for OntoEL. We distinguish between **fixed hyperparameters** (consistent across all experiments) and **tuned hyperparameters** (optimized on validation sets).

Table 13: Hyperparameter settings for OntoEL.

Parameter	Symbol	MedMentions	BC5CDR	NCBI
Fixed Hyperparameters				
Backbone Encoder	–	SapBERT-base / MedCPT-QEnc		
Hidden Dimension	d	768		
Projection Dimension	d'	768		
Candidate Pool Size	k	64		
Sigmoid Sharpness	s	10		
Ranking Margin	γ	0.2		
Optimizer	–	AdamW		
Learning Rate	–	2×10^{-5}		
Batch Size	–	64		
Training Epochs	–	10		
Temperature Init	θ_0	$\log\sqrt{d'} \approx 3.3$		
Tuned Hyperparameters (on validation set)				
Fusion Weight	α	0.78	0.80	0.82
Type Loss Weight	λ	0.5	0.3	0.3
Dataset-Specific				
Semantic Types	$\lvert\Gamma\rvert$	21	2	1

Hyperparameter Selection Rationale.

- **Sigmoid Sharpness ($s = 10$):** Selected from the range $[5, 10]$ as discussed in Section 3. This provides a balance between smooth optimization and sharp logical behavior. As shown in Theorem 1, the discrimination ratio grows as $\sim e^{0.31s}$, yielding ~23× discrimination at $s = 10$.
- **Fusion Weight ($\alpha \approx 0.8$):** Tuned on validation sets, converging to values in $[0.78, 0.82]$ across datasets, consistent with the "typically $\alpha \approx 0.8$" noted in Section 4.1.5. Higher α gives more weight to neural similarity; lower α emphasizes ontological consistency.

- **Temperature (θ):** Initialized to $\log \sqrt{d'}$ as specified in Section 4.1.2. The learnable parameterization $\theta = \exp(\hat{\theta})$ ensures $\theta > 0$ and allows adaptive sharpening during training.
- **Candidate Pool Size ($k = 64$):** Balances recall coverage (Recall@64 $\approx$ 89%) with computational efficiency. Increasing to $k = 128$ yields marginal gains (<0.3%) at 2× cost.

J.3 Baseline Implementation Details

We provide implementation details for all baselines evaluated in Table 2.

J.3.1 Group 1: Retrieval Baselines.

BM25. We use the Pyserini implementation with default parameters ($k_1 = 0.9$, $b = 0.4$). The corpus consists of all entity preferred names and synonyms from $\mathcal{E}$.

PubMedBERT.. We use `microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext` as a bi-encoder without domain-specific fine-tuning. Mentions and entities are encoded using mean pooling over the last hidden layer.

CODER.. We use the official checkpoint from Yuan et al. [95], which applies contrastive learning on UMLS synonymy pairs. The model is used as-is without further fine-tuning.

SapBERT.. We use `cambridgeltl/SapBERT-from-PubMedBERT-fulltext` [45], which is pre-trained on UMLS synonymy pairs using multi-similarity loss. This serves as our primary backbone for controlled comparisons.

J.3.2 Group 2: Re-ranking Baselines.

SapBERT + Type Prediction (MTL).. We add a multi-label classification head on top of the SapBERT mention encoder:

$$P(\tau \mid m) = \sigma(\mathbf{W}_{\text{cls}} \cdot \text{Enc}(m) + \mathbf{b}_{\text{cls}})$$

where $\mathbf{W}_{\text{cls}} \in \mathbb{R}^{|\Gamma| \times d}$. Training uses binary cross-entropy for type prediction jointly with the retrieval objective. This baseline learns type-specific weight vectors (ID-based) rather than encoding type names (name-based), and thus cannot generalize to unseen types.

Hyperparameters: Same as SapBERT backbone; type loss weight $\lambda_{\text{type}} = 0.5$.

SapBERT + Cross-Encoder. We fine-tune a BERT-base model that takes concatenated input:

```
[CLS] mention context [SEP] candidate name [SEP]
```

The `[CLS]` representation is passed through a linear layer to produce a relevance score. Training uses binary cross-entropy loss with the same hard negatives as OntoEL.

Hyperparameters: Learning rate 2×10^{-5}, batch size 32, max sequence length 128, 5 epochs with early stopping.

J.3.3 Group 3: Generative & LLM Baselines.

GenBioEL.. We use the official implementation from Yuan et al. [94], which formulates entity linking as sequence-to-sequence generation. The model is trained to generate the target entity name given the mention and context.

Hyperparameters: BART-base backbone, learning rate 3×10^{-5}, beam size 5 for decoding.

RankGPT (Llama-3). We use Llama-3-8B-Instruct as the backbone LLM. Given a mention and context, we provide the top-20 candidates retrieved by SapBERT and instruct the model to re-rank them.

Prompt template:

```
Given a biomedical mention and its context, re-rank the
candidate entities by relevance to the mention.

Mention: {mention}
Context: {context}

Candidates:
1. {candidate_1}
2. {candidate_2}
...
20. {candidate_20}

Output the ranked list of candidate numbers from most
to least relevant, separated by commas.
```

Due to context window and cost constraints, we re-rank only the top-20 candidates (vs. top-64 for other methods). Temperature is set to 0 for deterministic outputs.

J.3.4 Group 4: System-level SOTA Baselines.

ArboEL.. We use the official implementation from Agarwal et al. [3], which performs graph-based collective entity linking using arborescence inference. The method jointly resolves all mentions in a document by modeling entity coherence.

Hyperparameters: Default settings from the official repository; SapBERT as the base encoder.

KRISSBERT.. We use the official checkpoint from Zhang et al. [97], which applies knowledge-rich self-supervised learning on biomedical corpora. The model is used as a bi-encoder retriever.

Hyperparameters: Default settings; hidden dimension 768.

MedCPT (Full System). We use the official MedCPT pipeline [31], which includes both a dense retriever (QEnc/DEnc) and a cross-encoder re-ranker. For fair comparison:

- **MedCPT (Table row):** Full official pipeline with both retriever and re-ranker.
- **MedCPT + OntoEL (Table row):** MedCPT retriever only, with our OntoEL re-ranker replacing the MedCPT re-ranker.

Hyperparameters: Official settings; retriever returns top-64 candidates.

J.3.5 Group 5: Our Methods.

SapBERT + Static Logic. This baseline uses fixed type priors estimated from training set frequencies instead of context-aware inference:

$$\tau^{\mathcal{I}}_{\text{static}}(m) = \frac{\text{count}(\tau \text{ among gold entities in training})}{\text{total training mentions}}$$

The consistency score is computed using the same Sigmoidal Reichenbach implication as OntoEL, but with static rather than dynamic type memberships. This isolates the contribution of context-aware type prediction.

Table 14: Computational cost comparison (MedM ST21pv).

Method	Training Time	Inference (ms/query)
Group 1: Retrieval		
BM25	–	5.1
SapBERT	–	8.2
Group 2: Re-ranking		
SapBERT + Type Pred	1.5h	8.5
SapBERT + Cross-Encoder	4h	350
Group 3: LLM		
RankGPT (Llama-3-8B)	–	1,200
Group 4: SOTA Systems		
MedCPT (full pipeline)	6h	355
Group 5: Ours		
SapBERT + Static Logic	–	9.8
SapBERT + OntoEL	2h	12.1
MedCPT + OntoEL	2.5h	15.3

Hyperparameters: Same fusion weight α as OntoEL; no learnable type inference parameters.

SapBERT + OntoEL / MedCPT + OntoEL.. Our full method as described in Section 4, with hyperparameters specified in Table 13.

J.4 Training Details

Negative Sampling Strategy. For each mention, we construct the negative set $\mathcal{N}(m)$ using:

- **In-batch negatives:** Gold entities of other mentions in the same batch (efficient, provides diverse negatives).
- **Hard negatives:** Top-k candidates from the backbone retriever excluding the gold entity (forces discrimination between similar entities).

Learning Rate Schedule. We use linear warmup over the first 10% of training steps, followed by linear decay to zero.

Early Stopping. Training stops if validation Acc@1 does not improve for 3 consecutive epochs.

Random Seeds. All results are averaged over 5 independent runs with seeds $\{42, 123, 456, 789, 1024\}$. Standard deviations are reported in Table 2.

J.5 Evaluation Protocol

Metrics. Following the BioEL benchmarking protocol [10]:

- **Recall@k:** Proportion of mentions where the gold entity appears in the top-k retrieved candidates. Evaluates the candidate generation stage.
- **Acc@k:** Proportion of mentions where the gold entity is ranked within the top-k after re-ranking. Evaluates the named entity disambiguation stage.
- **MRR:** Mean Reciprocal Rank, $\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\text{rank}_i}$. Evaluates overall ranking quality.

Statistical Significance. We report mean ± standard deviation over 5 runs. Statistical significance († in Table 2) is assessed using a two-tailed paired t-test at $p < 0.05$ level, comparing against the strongest baseline in each category.

J.6 Computational Cost

Table 14 summarizes the computational requirements for all methods.

OntoEL adds only ~4ms latency over the backbone retriever, achieving approximately 30× speedup compared to Cross-Encoder re-rankers while delivering superior accuracy. This efficiency stems from the lightweight bi-encoder architecture and pre-computed ontological memberships, as analyzed in Proposition 1.