

DuCEI: When to Introduce Metadata in Cross-Entity Interaction for Entity Disambiguation

Jinwoo Min¹, Roseline Nyange¹, Wonkee Lee², Hyun-Je Song^{1,†}, Seung-Hoon Na^{3,†}

¹Jeonbuk National University

{mabongza, roselinenyange, hyunje.song}@jbnu.ac.kr

²LG AI Research

wklee@lgresearch.ai

³UNIST

nash@unist.ac.kr

Abstract

Entity disambiguation links mentions in text to the entities they refer to in a knowledge base. Recent cross-entity interaction models improve candidate comparison by jointly encoding multiple candidates, but incorporating entity metadata into this joint encoding is not straightforward. Appending metadata to every candidate forces joint encoding to process longer inputs with detailed distractor information, which can weaken the contrastive cues that make cross-entity interaction effective. We propose DuCEI, a dual-stage framework that avoids this broad metadata interaction by reducing the candidate set before introducing entity metadata. DuCEI first compares all candidates using title-based cross-entity interaction, and then performs metadata-enriched interaction only among the remaining candidates. This design makes metadata use selective at the candidate level, applying detailed evidence only where it is likely to help. Experiments on six standard benchmarks show that DuCEI achieves the best overall performance among models trained solely on AIDA-CoNLL. Further analyses show that confidence-based selective refinement preserves nearly all of the full two-stage performance while applying Stage 2 to only 27.80% of mentions, reducing inference cost from $1.90\times$ to approximately $1.27\times$ that of Stage 1 alone. Our code is available at <https://github.com/jinwoomin/DuCEI>.

1 Introduction

Entity disambiguation (ED) is a core subtask of entity linking (Mrini et al., 2022; Zhang et al., 2022). Given a mention in context, ED aims to link it to the entity it refers to in a knowledge base (KB). This task is often formulated as selecting the most plausible entity from a candidate set. Because ED requires comparing multiple candidates for each mention, models must support fine-grained disambiguation while keeping candidate comparison efficient.

To handle this candidate selection problem, many ED models adopt independent mention–candidate matching, where each candidate is compared with the mention context separately (van Hulst et al., 2020; Logeswaran et al., 2019; Wu et al., 2020; Agarwal and Bikel, 2020). These models often enrich candidate representations with KB-derived metadata, such as entity descriptions and types, in addition to entity titles to better discriminate among candidates (Gupta et al., 2017; Chen et al., 2020; Zhu et al., 2020; Tianran et al., 2021).

Recent studies have sought more effective comparison among competing candidates through cross-entity interaction. Rather than matching each candidate independently with the mention context, these models jointly encode multiple candidates in a single sequence, enabling direct comparison among them (Barba et al., 2022; Procopio et al., 2023; Xu et al., 2023a,b). This joint encoding allows candidates to be interpreted relative to one another, making it possible to capture contrastive cues among competing entities. In particular, ExtEnD (Barba et al., 2022) achieves strong performance through cross-entity interaction even when relying only on entity titles. This result suggests that direct comparison among candidates can itself provide useful disambiguation signals.

While direct comparison among candidates provides useful disambiguation signals, entity titles alone provide limited evidence about each candidate. KB-derived metadata such as entity descriptions and types can provide additional evidence, yet incorporating such metadata into joint cross-entity encoding is not straightforward. Unlike independent matching, joint encoding places information from all candidates in a shared interaction space. Appending metadata to every candidate therefore increases the input length and introduces detailed information from many distractors into the same interaction scope. This can dilute contrastive cues and make the relevant differences among candi-

[†]Corresponding authors.

dates less salient.

We propose DuCEI (**D**ual-stage **C**ross-**E**ntity **I**nteraction), a dual-stage framework that organizes cross-entity interaction into broad candidate comparison and focused metadata-enriched disambiguation. DuCEI follows a broad-to-focused interaction strategy, using compact candidate representations for initial comparison and introducing detailed metadata only after the candidate set has been reduced. Specifically, DuCEI first identifies promising candidates through title-based cross-entity interaction and then applies metadata-enriched interaction with descriptions and types within the reduced candidate set. This design concentrates detailed semantic evidence on candidates that require fine-grained disambiguation while preserving broad relative comparison over the full candidate set.

DuCEI also uses this idea adaptively during inference, skipping metadata-enriched refinement when the first stage makes a confident prediction. Rather than using two stages as a generic reranking pipeline, DuCEI uses them to make metadata-enriched cross-entity interaction selective and efficient, introducing detailed metadata only within reduced candidate sets and only for mentions that require further refinement. We validate this design on six standard ED benchmarks under the common AIDA-CoNLL-only training setting. DuCEI achieves the best overall performance among models trained solely on AIDA-CoNLL, reaching 86.4 average and 85.6 out-of-domain average in in-KB Micro F1. Selective refinement further preserves nearly all of the full two-stage performance while applying metadata-enriched interaction to only a fraction of mentions.

Our contributions are summarized as follows:

1. We propose **DuCEI**, a dual-stage cross-entity interaction framework that organizes interaction into broad candidate comparison over the full candidate set and focused metadata-enriched disambiguation over a reduced candidate set.
2. We demonstrate that introducing metadata after reducing the candidate scope is more effective than single-step full-information encoding, especially for larger candidate sets.
3. Across six standard ED benchmarks, DuCEI achieves the best overall performance among models trained solely on AIDA-CoNLL. We further show that selective metadata-enriched refinement preserves nearly all of the full two-

stage performance while invoking Stage 2 for only 27.80% of mentions at approximately $1.27\times$ the inference cost of Stage 1 alone.

2 Related Work

2.1 Entity Disambiguation Models

Early ED models largely follow a scoring-based paradigm that computes a relevance score between the mention context and each candidate entity (Ganea and Hofmann, 2017; Le and Titov, 2018; Wu et al., 2020). Within this paradigm, bi-encoder models encode mentions and entities separately, enabling efficient retrieval through precomputed entity representations, whereas cross-encoder models jointly encode the mention context and each candidate entity to capture fine-grained interactions at a higher computational cost (Logeswaran et al., 2019; Wu et al., 2020; Agarwal and Bikel, 2020).

Generative models directly produce the entity title or identifier given the mention context, enabling end-to-end disambiguation and not necessarily requiring an explicit candidate set (De Cao et al., 2021; Mrini et al., 2022). This paradigm aligns naturally with large language models (LLMs), and recent LLM-based EL/ED systems often combine prompting with candidate retrieval (or tool-augmented pipelines) to improve accuracy and controllability by grounding generation in a KB-backed candidate space (Xiao et al., 2023; Shlyk et al., 2024; Ding et al., 2024; Ye and Mitchell, 2025). However, such approaches can still suffer from output inconsistency and may incur substantial computational cost, especially when retrieval returns many candidates or when multiple rounds of generation or reranking are required.

Extractive models (Barba et al., 2022; Procopio et al., 2023; Zhang et al., 2024) instead treat ED as a machine-reading comprehension task, concatenating the mention context and a set of candidates (e.g., titles) to predict the correct entity-title span. This formulation has provided a natural basis for recent cross-entity interaction models.

2.2 Cross-Entity Interaction

Cross-entity interaction methods jointly encode multiple candidates within a single sequence, enabling explicit comparison among them (Barba et al., 2022). Later work further incorporated entity descriptions (Procopio et al., 2023), but gains were inconsistent, suggesting a trade-off between information richness and sequence efficiency. Re-

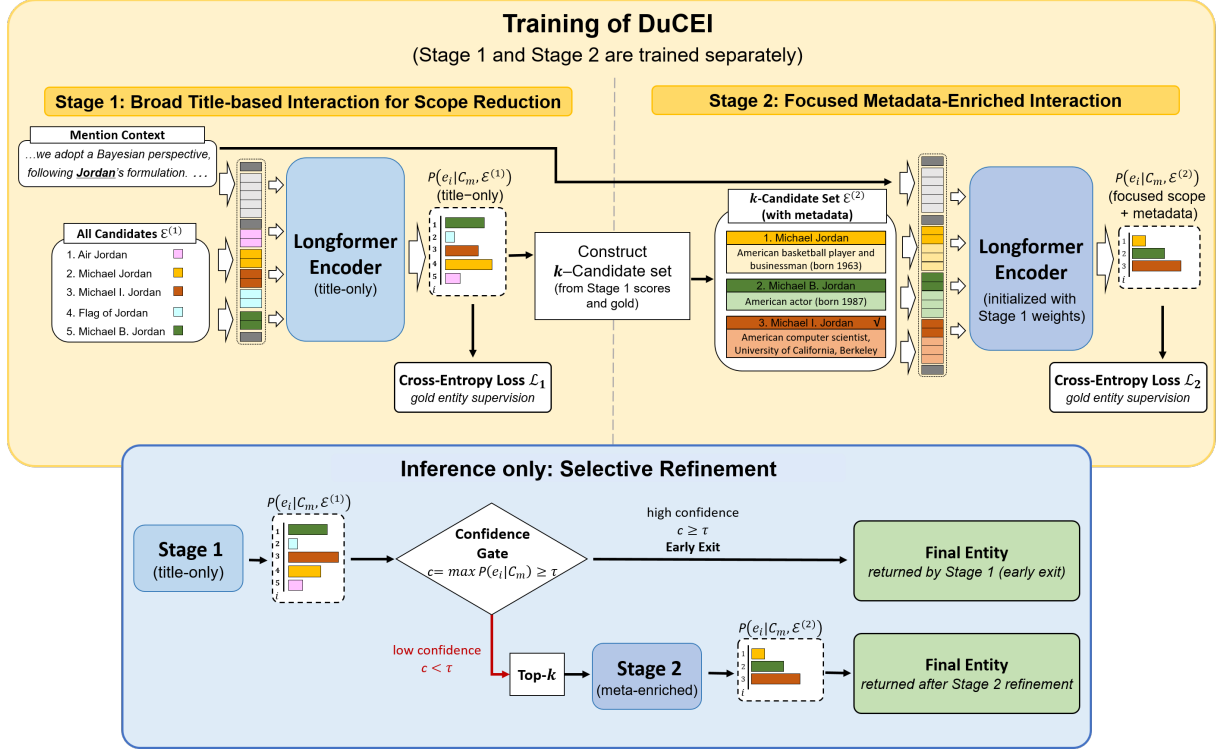


Figure 1: Overview of DuCEI. The upper panel shows the training procedure, where Stage 1 performs broad title-based interaction over all candidates and Stage 2 is trained for focused metadata-enriched interaction over up-to- k candidate sets constructed using Stage 1 scores and the gold entity. The lower panel shows inference-time selective refinement: DuCEI exits early when the Stage 1 prediction is confident, and otherwise applies Stage 2 to the top- k candidates for metadata-enriched refinement.

lated forms of candidate interaction have also appeared in retrieval and reranking (Song et al., 2024), biomedical entity linking (Xu et al., 2023a), and zero-shot entity linking (Xu et al., 2023b). Together, these studies highlight the usefulness of cross-candidate comparison, but they do not systematically study how the interaction scope should be controlled before introducing rich entity metadata under long-sequence constraints.

Conventional retrieve-then-rerank pipelines generally reduce the candidate set using a lightweight independent scorer, such as a mention-entity prior or a bi-encoder, and then disambiguate the retained top- k candidates using a pointwise cross-encoder reranker (Wu et al., 2020). In contrast, DuCEI defines the metadata-enriched interaction scope through cross-entity interaction itself. Specifically, its first stage performs broad title-based comparison over the full candidate set to identify a compact scope, and its second stage introduces descriptions and types only within this reduced scope for focused comparison. By doing so, DuCEI treats interaction scope and the timing of metadata introduction as explicit design variables rather than as byproducts of a fixed candidate-reduction pipeline.

3 Preliminaries

We formally define the ED task as follows. Given a mention m appearing in a textual context C_m , where $C_m = [\text{Ctx}_{\text{left}}; \langle M \rangle m \langle /M \rangle; \text{Ctx}_{\text{right}}]$ denotes the sequence consisting of the mention and its surrounding left and right contexts, the goal is to identify the referent entity e^* in a knowledge base $\mathcal{K}$. $\langle M \rangle$ and $\langle /M \rangle$ are special tokens to indicate the mention boundaries within the context.

For each mention m , we assume a predefined candidate set $\mathcal{E}_m = \{e_1, \dots, e_n\} \subseteq \mathcal{K}$. Each entity e_i is associated with a title t_i and entity metadata. We use descriptions d_i and type sequences y_i as metadata, and denote each candidate as $e_i = (t_i, d_i, y_i)$. Given the mention context C_m and the candidate set $\mathcal{E}_m$, an ED model predicts the entity with the highest posterior probability:

$$\hat{e} = \underset{e_i \in \mathcal{E}_m}{\operatorname{argmax}} P(e_i | C_m, \mathcal{E}_m).$$

Under a scoring-based formulation, the model first computes a relevance score for each candidate:

$$s_i = f_\theta(C_m, e_i; \mathcal{E}_m).$$

The posterior over candidates is then obtained by applying a softmax over these scores:

$$P(e_i | C_m, \mathcal{E}_m) = \frac{\exp(s_i)}{\sum_{e_j \in \mathcal{E}_m} \exp(s_j)}. \quad (1)$$

Here, f_θ denotes a trainable encoder. The dependence on $\mathcal{E}_m$ allows the score of each candidate to be computed with information from other candidates, as in cross-entity interaction models.

4 Methods

4.1 DuCEI Overview

DuCEI organizes cross-entity interaction into two consecutive stages designed to use metadata only after the candidate scope has been reduced (Figure 1): **Stage 1: Broad Title-based Interaction for Scope Reduction** and **Stage 2: Focused Metadata-Enriched Interaction**. In Stage 1, DuCEI takes the mention context and the titles of all candidate entities as input. It encodes the candidates in a single sequence through title-based cross-entity interaction and scores them jointly. The model then selects the top- k candidates as the reduced candidate set for the next stage. In Stage 2, DuCEI augments these candidates with entity metadata, such as descriptions and types, and performs cross-entity interaction again within the reduced set for final disambiguation. This two-stage design keeps broad candidate comparison lightweight in the first stage and reserves metadata-enriched interaction for the smaller set of candidates that require finer-grained evidence.

4.2 Model Architecture

DuCEI uses the same encoding–representation–scoring architecture in both stages, while varying the candidate set and input composition across stages. Let $\mathcal{E}^{(1)} = \mathcal{E}_m = \{e_1, \dots, e_n\}$ denote the candidate set in Stage 1, and let $\mathcal{E}^{(2)} \subseteq \mathcal{E}^{(1)}$ denote the up-to- k candidate set used in Stage 2.

Input Sequence Construction. Each stage constructs an input sequence by concatenating the mention context C_m with candidate representations. We use $\langle /E \rangle$ to separate candidates, and in Stage 2 we additionally use $\langle /T \rangle$ and $\langle /D \rangle$ to separate titles, descriptions, and type sequences. In Stage 1, candidates are provided in descending order of their empirical mention–entity frequency, which reflects the candidate prior $P(e | m)$.

Stage 1: Broad Title-based Interaction for Scope Reduction.

$$x^{(1)} = [\text{CLS}] C_m [\text{SEP}] t_1 \langle /E \rangle \dots t_n \langle /E \rangle [\text{SEP}]$$

Stage 1 jointly encodes the mention context and all candidate titles to score candidates and define a reduced interaction scope.

Stage 2: Focused Metadata-Enriched Interaction.

$$x^{(2)} = [\text{CLS}] C_m [\text{SEP}] t_1 \langle /T \rangle d_1 \langle /D \rangle y_1 \langle /E \rangle \dots t_k \langle /T \rangle d_k \langle /D \rangle y_k \langle /E \rangle [\text{SEP}]$$

Stage 2 jointly encodes the mention context and the retained candidates augmented with descriptions and types for focused, fine-grained disambiguation.

Encoding. Each stage takes as input a token sequence $\mathbf{X}$ (i.e., the tokenized sequence of $x^{(1)}$ or $x^{(2)}$) that concatenates the mention context and candidate representations (titles only in Stage 1; titles with metadata in Stage 2).

The sequence is encoded by a Transformer encoder (Vaswani et al., 2017) to produce contextualized hidden representations:

$$\mathbf{H} = \text{Encoder}(\mathbf{X}), \quad \mathbf{H} = \{\mathbf{h}_0, \dots, \mathbf{h}_N\},$$

where $\mathbf{h}_0$ corresponds to the [CLS] token representing the entire input sequence, and $\mathbf{h}_j$ denotes the contextual representation of the j -th token in $\mathbf{X}$.

Both stages use the same encoder architecture, but are trained stagewise: we first train Stage 1, and then fine-tune Stage 2 from the Stage 1 checkpoint. To efficiently process long concatenated sequences, we adopt the Longformer encoder (Beltagy et al., 2020), which replaces the quadratic full attention with sparse global and local attention patterns, achieving linear scalability with respect to input length.

Candidate Representation and Scoring. From the encoder output $\mathbf{H} = \{\mathbf{h}_0, \dots, \mathbf{h}_N\}$, we obtain the representation of each candidate entity based on its title span. Let (a_i, b_i) denote the start and end token indices of the i -th entity title within the input sequence $\mathbf{X}$. The candidate representation $\mathbf{r}_i$ is obtained by concatenating the corresponding hidden vectors:

$$\mathbf{r}_i = [\mathbf{h}_{a_i}; \mathbf{h}_{b_i}] \in \mathbb{R}^{2d}.$$

Each $\mathbf{r}_i$ is then projected to a scalar score using a linear transformation:

$$s_i = \mathbf{W}_s \mathbf{r}_i + b_s, \quad \mathbf{W}_s \in \mathbb{R}^{1 \times 2d}.$$

These scores are converted into a posterior distribution by a softmax over the stage-specific candidate set, $\mathcal{E}_m = \mathcal{E}^{(1)}$ for Stage 1 and $\mathcal{E}_m = \mathcal{E}^{(2)}$ for Stage 2, as in Equation (1).

For a candidate set $\mathcal{E}$ used in a given stage, the training loss is the negative log-likelihood of the gold referent entity:

$$\mathcal{L} = -\log P(e^* | C_m, \mathcal{E}).$$

This setup enables stagewise training and evaluation under a unified architecture, where the two stages differ primarily in input composition and candidate set size.

4.3 Confidence-Based Selective Refinement

DuCEI’s two-stage structure can also be used selectively at inference time. Not every mention requires rich metadata such as descriptions and types. In many cases, Stage 1 can already produce a sufficiently confident prediction using title-based cross-entity interaction alone. Therefore, DuCEI uses the Stage 1 prediction confidence to determine whether Stage 2 refinement should be applied.

Let $P^{(1)}(e_i | C_m, \mathcal{E}^{(1)})$ denote the posterior distribution computed by Stage 1. We define the confidence of the top Stage 1 prediction as

$$c = \max_{e_i \in \mathcal{E}^{(1)}} P^{(1)}(e_i | C_m, \mathcal{E}^{(1)}).$$

Given a threshold τ , if $c \geq \tau$, we use the Stage 1 prediction as the final output and skip Stage 2. Otherwise, if $c < \tau$, we regard the Stage 1 prediction as insufficiently confident and apply metadata-enriched Stage 2 interaction over the retained top- k candidates. The final prediction is defined as

$$\hat{e} = \begin{cases} \arg \max_{e_i \in \mathcal{E}^{(1)}} P^{(1)}(e_i | C_m), & c \geq \tau, \\ \arg \max_{e_i \in \mathcal{E}^{(2)}} P^{(2)}(e_i | C_m), & c < \tau. \end{cases}$$

This selective refinement avoids applying metadata-enriched interaction uniformly to all instances and instead invokes Stage 2 only when Stage 1 is uncertain. Thus, DuCEI makes metadata use selective not only at the candidate level, through candidate-scope reduction, but also at the instance level by deciding whether metadata-enriched refinement is needed.

5 Experiments

5.1 Experimental Setup

Datasets. We followed the standard evaluation setup widely used in prior work on ED (Hoffart

et al., 2011; Guo and Barbosa, 2018). Specifically, we used the AIDA-CoNLL dataset (Hoffart et al., 2011) for training, model selection, and in-domain evaluation. For out-of-domain evaluation, we used the cleaned MSNBC, AQUAINT, ACE2004, WNED-CWEB (CWEB), and WNED-WIKI (WIKI) benchmarks released by Guo and Barbosa (2018).

We use the candidate sets constructed by Le and Titov (2018), following common practice in recent ED work (De Cao et al., 2021; Barba et al., 2022). Detailed dataset statistics and top- K gold inclusion rates are provided in Tables 8 and 9.

Knowledge Base. For entity identification, both gold and candidate entities were normalized to canonical English Wikipedia titles (with redirects resolved). All titles were standardized using the English Wikipedia dump enwiki-20230120. Entity descriptions and types were extracted from Wikidata (Vrandečić and Krötzsch, 2014), using the wikidata-20200601 dump. We represent entity types as a sequence of Wikidata relation-object pairs (details in Appendix B.1).

Baselines. We compare DuCEI against representative fine-tuned ED systems trained solely on the AIDA-CoNLL dataset without large-scale Wikipedia pretraining. Specifically, we include GENRE (De Cao et al., 2021), ReFinED (Ayoola et al., 2022), cross-entity interaction baselines ExtEnD and ExtEnD_{def} (Barba et al., 2022; Procopio et al., 2023), and FusionED (Wang et al., 2024). We report DuCEI Stage 1 only as an internal ablation to isolate the contribution of focused metadata-enriched interaction in Stage 2. We additionally report ChatEL (Ding et al., 2024) and LLMAEL×ReFinED (Xin et al., 2025) as LLM-based/LLM-assisted references, with broader positioning discussed in Appendix E.

Implementation Details. We use Longformer-Large as the encoder for both stages of DuCEI. During inference, Stage 1 ranks all candidates and, when refinement is applied, passes the top- k candidates to Stage 2. We set $k = 10$ as a practical operating point that keeps Stage 2 compact while preserving a sufficient recall margin for focused metadata-enriched interaction, as analyzed in Section 6. Further implementation and training details are provided in Appendix C.

Evaluation. We used **in-KB Micro F1** as the main evaluation metric, following standard ED

Model	In-domain	Out-of-domain					Averages	
	AIDA	MSNBC	AQUAINT	ACE2004	CWEB	WIKI	Avg	Avg _{OOD}
<i>LLM-based / LLM-assisted systems</i>								
ChatEL (Ding et al., 2024)	–	87.8	86.5	89.0	71.7	78.0	–	82.6
LLMAEL×ReFinED (Xin et al., 2025)	93.2	88.6	89.9	91.2	77.3	87.6	88.0	86.9
<i>AIDA-CoNLL-only supervised systems under comparable candidate-based settings</i>								
GENRE (De Cao et al., 2021)	88.6	88.1	77.1	82.3	71.9	71.7	80.0	78.2
ReFinED (Ayoola et al., 2022)	88.2	92.3	86.8	90.6	75.1	74.5	84.6	83.9
ExtEnD (Barba et al., 2022)	90.0	94.5	87.9	88.9	76.6	76.7	85.8	84.9
ExtEnD _{def} (Procopio et al., 2023)	92.4	93.2	87.0	87.7	76.4	78.3	85.8	84.5
FusionED (Wang et al., 2024)	91.7	92.4	82.0	87.1	75.8	78.6	84.6	83.2
DuCEI (Stage 1 only)	89.8	94.6	88.5	89.3	77.1	76.6	86.0	85.2
DuCEI (Stage 1+2)	90.6	94.7	89.0	90.1	77.1	77.1	86.4	85.6

Table 1: ED results on AIDA-CoNLL and five out-of-domain datasets. We re-evaluate ChatEL and LLMAEL×ReFinED using the same Le and Titov (2018) candidates and in-KB Micro F1 protocol as DuCEI. ChatEL is candidate- and evaluation-aligned, whereas LLMAEL×ReFinED is contextual because its underlying ReFinED variant uses additional Wikipedia-scale training. For the supervised comparison, we report representative systems trained only on AIDA-CoNLL under comparable candidate-based settings. Bold scores indicate the best results among the AIDA-CoNLL-only supervised systems.

evaluation practice (Röder et al., 2018; Le and Titov, 2018; Barba et al., 2022). A mention is treated as NIL when its mention-specific candidate set becomes empty after applying our KB snapshot and title mapping; such cases are excluded from precision but counted in recall when computing Micro F1. We report scores on AIDA-CoNLL (in-domain), the average over all six datasets (denoted as Avg), and the average over the five out-of-domain datasets (denoted as Avg_{OOD}). Comprehensive results are summarized in Table 1.

5.2 Main Results

We report “DuCEI (Stage 1 only)” as the variant that runs only the title-based first stage without Stage 2 refinement, and “DuCEI (Stage 1+2)” as the full two-stage model. As shown in Table 1, DuCEI achieves the best overall performance among AIDA-CoNLL-only supervised systems under comparable candidate-based settings, obtaining the highest Avg and Avg_{OOD} scores.

Notably, DuCEI (Stage 1 only) already provides a strong title-based cross-entity interaction baseline, outperforming all prior AIDA-CoNLL-only supervised baselines in both Avg and Avg_{OOD}. This indicates that broad title-based interaction, together with the prior-based candidate order, provides effective signals for candidate comparison even without rich metadata. DuCEI (Stage 1+2) further improves over this strong Stage 1 baseline by introducing descriptions and types only within the reduced top- k candidate scope.

While ExtEnD_{def} incorporates Wikidata defini-

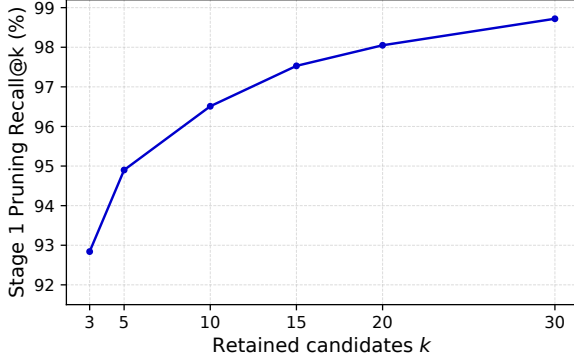
tions in a single-stage formulation, DuCEI first reduces the candidate scope through title-based cross-entity interaction and then introduces descriptions and types only within the compact top- k set. This design reduces the overhead of encoding full metadata for all candidates and allows richer metadata to be leveraged more effectively during final disambiguation.

6 Analysis

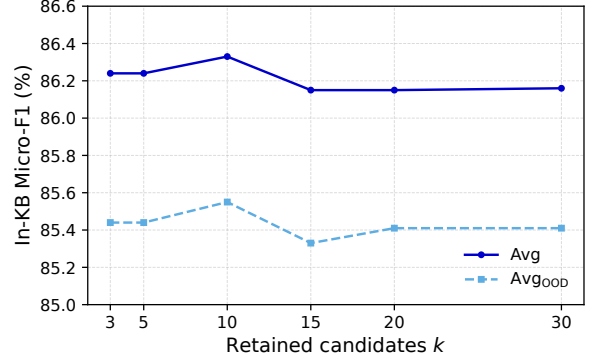
6.1 Model Ablation

Efficiency–Accuracy Trade-off w.r.t. the Retained Candidates (k). We analyze how the number of candidates retained for Stage 2 refinement affects both pruning coverage and final disambiguation accuracy. For each k , we train and evaluate a model with k fixed. For a controlled comparison across different k , we set the mention masking probability to 0 in this analysis; mention masking randomly replaces tokens within the marked mention span with [MASK] during training (see Appendix C).

Figure 2a reports Stage 1 pruning Recall@ k across the six evaluation benchmarks. We define pruning Recall@ k as the fraction of mentions whose gold entity is contained in the Stage 1 top- k retained set, conditioned on the gold entity being present in the original candidate set. As k increases, pruning Recall@ k improves consistently, indicating that larger retained sets reduce the risk of irreversible pruning errors. At our default setting of $k=10$, Stage 1 pruning Recall@10 reaches



(a) Stage 1 pruning Recall@ k vs. k .



(b) Final Avg and AvgOOD vs. k .

Figure 2: Effect of varying the retained candidate size k for Stage 2 refinement.

96.51%, corresponding to a 3.49% pruning-failure rate under the gold-in-candidate condition.

In contrast, Figure 2b shows that final performance does not monotonically improve as k increases. Although retaining more candidates consistently improves pruning Recall@ k (Figure 2a), Avg and AvgOOD peak around $k = 10$ and slightly decrease as the Stage 2 interaction scope becomes broader. This counterintuitive trend highlights a trade-off between candidate coverage and interaction interference: retaining more candidates reduces the risk of pruning the gold entity, but also introduces more distractor metadata into the shared interaction space, which does not necessarily improve final refinement. Thus, $k = 10$ provides a practical balance between preserving correct entities and keeping metadata-enriched interaction focused.

Ablation Study on Description and Type Information in Stage 2. To quantify the contribution of each semantic component used in Stage 2, we conduct an ablation study with four Stage 2 variants that differ only in their input composition over the retained top- k candidates: (i) Title only, (ii) Title + Description, (iii) Title + Type, and (iv) Title + Description + Type (Full). Note that (i) still operates on the top- k candidates retained by Stage 1. It simply omits description/type fields in Stage 2.

Input Composition	Avg	AvgOOD
Title only	85.91	85.14
Title + Description	86.28	85.44
Title + Type	86.10	85.33
Title + Description + Type (Full)	86.42	85.60

Table 2: Ablation study on Stage 2 input composition. All variants use the same Stage 1 retained candidates and differ only in the entity information provided to Stage 2.

As shown in Table 2, adding either descriptions or types improves over the title-only Stage 2 variant, and using both yields the best performance. This indicates that descriptions and types provide complementary semantic cues: descriptions offer richer contextual grounding about the entity, while types supply high-level categorical information that helps rule out semantically incompatible candidates.

6.2 Limitations of Single-step Full-information Encoding

Single-step Full-info setting. We define a single-step full-information baseline that follows the same input sequence construction described in Section 4, but incorporates full entity information (title, description, and type) for all candidates in a single pass. Concretely, instead of first reducing the interaction scope and then introducing metadata only within the reduced scope, this baseline concatenates the full semantic fields for every candidate and performs disambiguation in one step under a maximum input length constraint. This provides a strong single-pass baseline for isolating the effect of reducing the interaction scope before introducing rich metadata.

Setting	Avg	AvgOOD
DuCEI (Stage 1 only)	86.00	85.23
Single-step Full-info (max_len=1024)	85.06	84.13
Single-step Full-info (max_len=1536)	85.17	84.06
Single-step Full-info (max_len=2048)	85.18	84.36

Table 3: Single-step full-information encoding results under different maximum input lengths. DuCEI (Stage 1 only) is included as a title-only reference.

Overall comparison. Table 3 compares this single-step baseline with DuCEI (Stage 1 only). Despite the richer semantics, the single-step base-

Model	$N=1$	2–5	6–10	11–20	21–50	51–100
Single-step Full-info (max_len=2048)	100.0	88.15	78.87	75.35	71.06	73.70
DuCEI (Stage 1+2)	100.0	88.99	81.04	76.88	72.14	76.05

Table 4: Top-1 accuracy stratified by candidate pool size N on the full six-dataset evaluation suite ($N \geq 1$), conditioned on the gold entity being present in the candidate set; this directly measures ranking quality under varying candidate pool sizes.

Reduction Method	Stage 1		Recall@10	Stage 2	
	Avg	Avg _{good}		Avg	Avg _{good}
Prior	77.86	79.07	95.18	82.85	83.07
Bi-encoder	78.39	76.89	92.06	81.78	80.55
DuCEI Stage 1	86.00	85.23	96.51	86.42	85.60

Table 5: Comparison of Stage 1 candidate-reduction methods under the same fixed metadata-enriched Stage 2 model. The Stage 1 columns report each reduction method’s own top-1 performance.

line consistently underperforms DuCEI (Stage 1 only) across all maximum input lengths (1024/1536/2048), suggesting that simply adding rich metadata to the full candidate set can introduce interference in the shared interaction space rather than improve joint comparison. This result naturally motivates our stagewise design: Stage 1 first defines a reduced interaction scope through broad title-based interaction, allowing Stage 2 to introduce description/type information only where fine-grained semantic evidence is most useful.

Analysis by candidate pool size. We further analyze performance by stratifying instances from the full six-dataset evaluation suite with non-empty candidate pools ($N \geq 1$) into bins by candidate pool size (conditioned on gold-in-candidates), and compare DuCEI with the single-step full-information baseline under max_len=2048. Table 4 shows that DuCEI outperforms the single-step baseline in every non-trivial bin, by at least one point in every bin beyond five. This supports our key claim: reducing the interaction scope before introducing description/type information is beneficial whenever multiple plausible candidates must be compared.

6.3 Beyond Conventional Retrieve-then-Rerank

Although Section 2.2 conceptually distinguishes DuCEI from conventional retrieve-then-rerank pipelines, this distinction alone does not empirically isolate the source of its improvement. We therefore examine whether DuCEI benefits from the two-stage structure alone or specifically from cross-entity interaction in candidate reduction and refinement.

Controlled Comparison: Stage 1 Candidate Reduction. We replace DuCEI’s interaction-based Stage 1 with either the mention–entity prior or a bi-encoder, while keeping the same trained metadata-enriched DuCEI Stage 2 model fixed. Table 5 summarizes the results.

DuCEI Stage 1 achieves the highest standalone performance and Recall@10 among the three reduction methods. Although prior-based reduction retains a high Recall@10 of 95.18, DuCEI improves it to 96.51 and yields substantially better final performance when the retained candidates are processed by the same fixed Stage 2 model. The bi-encoder similarly produces lower candidate retention and downstream performance. These results show that interaction-based candidate reduction provides an advantage over conventional independent reduction methods.

Controlled Comparison: Stage 2 Refinement.

Using exactly the same top- k candidates produced by DuCEI Stage 1, we replace the interaction-based Stage 2 with a standard pointwise cross-encoder that scores each mention–candidate pair independently. Table 6 summarizes the results.

Refinement	Avg	Avg _{good}
DuCEI Stage 1 only	86.00	85.23
+ Standard cross-encoder	85.84	84.48
+ Interaction-based Stage 2	86.42	85.60

Table 6: Comparison of refinement methods over the same top- k candidates produced by DuCEI Stage 1.

The standard cross-encoder does not improve over the Stage 1-only model. In contrast, DuCEI’s interaction-based Stage 2 improves performance under the same retained candidate set. Thus, the

Setting	Avg	Avg _{OOD}	Stage 2 Usage	Inference Time (ms/sample)
DuCEI (Stage 1 only)	86.00	85.23	0.00%	69.50
DuCEI (Stage 1+2)	86.42	85.60	100.00%	131.88
DuCEI (Selective, $\tau = 0.991$)	86.42	85.59	27.80 %	88.60

Table 7: Selective Stage 2 refinement using Stage 1 confidence. Stage 2 is applied only when the highest Stage 1 prediction confidence is below the threshold τ . Inference time is the average model forward-pass time per sample, measured on a single NVIDIA RTX A6000 GPU with batch size 1 and dynamically padded inputs up to 1024 tokens.

gain is not explained solely by candidate reduction, metadata enrichment, or the use of a second-stage reranker; joint comparison among the remaining candidates also contributes to final refinement.

Together, these results empirically distinguish DuCEI from a conventional retrieve-then-rerank pipeline at both stages. Its contribution lies not in the two-stage structure itself, but in organizing cross-entity interaction into broad title-based comparison over the full candidate set and focused metadata-enriched comparison over a reduced candidate set.

6.4 Selective Stage 2 Refinement

We evaluate the accuracy–efficiency trade-off of the confidence-based selective refinement introduced in Section 4.3. Based on the AIDA-dev trade-off, we use $\tau = 0.991$ as a representative operating point and keep this threshold fixed across all test sets. Additional development-set analysis of the threshold is provided in Appendix D.4.

As shown in Table 7, selective DuCEI preserves nearly all of the full two-stage performance while applying Stage 2 to only 27.80% of mentions. Its inference time is 88.60 ms per sample, compared with 69.50 ms for Stage 1 alone and 131.88 ms for full two-stage inference. This corresponds to approximately $1.27\times$ the inference cost of Stage 1 alone, substantially below the $1.90\times$ cost of always applying Stage 2.

These results indicate that the benefit of metadata-enriched refinement is concentrated in cases where Stage 1 remains less confident. For many high-confidence cases, broad title-based cross-entity interaction is sufficient to make the final decision without additional metadata.

This selective behavior is enabled by the strong discriminative signal produced by Stage 1. Unlike independent candidate scoring, Stage 1 jointly encodes and compares all candidates through cross-entity interaction. As a result, it can produce confident title-based decisions for many instances, al-

lowing rich metadata to be introduced only where it is needed. This finding suggests that Stage 1 confidence provides a practical criterion for deciding whether metadata-enriched interaction is required for a given instance.

7 Conclusion

We studied how to make metadata-enriched cross-entity interaction more selective and efficient for entity disambiguation. DuCEI follows a broad-to-focused interaction strategy, separating broad candidate comparison over the full candidate set from focused metadata-enriched disambiguation over a reduced candidate set. This design introduces rich metadata only after the interaction scope has been reduced, keeping detailed semantic evidence focused on candidates that require fine-grained comparison.

Experiments on six standard ED benchmarks show that DuCEI consistently improves over a strong title-based cross-entity interaction baseline and achieves the best overall performance among models trained solely on AIDA-CoNLL. Further analyses show that single-step full-information encoding becomes less effective as candidate sets grow, supporting the importance of controlling the interaction scope before introducing rich metadata. Confidence-based selective refinement preserves nearly all of the full two-stage performance while invoking Stage 2 for only 27.80% of mentions.

Future work includes end-to-end joint training through differentiable candidate selection, dynamic mechanisms that jointly adjust the retained candidate scope and metadata usage, alternative selective-refinement criteria beyond maximum softmax confidence (e.g., score margins, predictive entropy, calibrated confidence, or learned routing), improved metadata selection or compression, and evaluation with stronger candidate generation pipelines and in long-context or cross-domain settings.

Limitations

DuCEI has several limitations.

- DuCEI relies on a predefined candidate set and cannot recover when the correct entity is absent from the candidate pool. Our experiments use the fixed candidate sets of [Le and Titov \(2018\)](#), and we do not evaluate generalization across alternative candidate generation pipelines with different coverage or candidate quality.
- Because Stage 2 operates only on the Stage 1 top- k candidates, errors made during broad title-based scope reduction can propagate to the final prediction. In particular, if Stage 1 removes the gold entity from the retained top- k set, Stage 2 cannot recover it.
- The focused metadata-enriched interaction stage depends on Wikidata descriptions and type fields, which can be missing, overly generic, or noisy. In addition, our type representation uses a selected set of Wikidata relation–object pairs, so performance may be sensitive to relation selection and serialization choices.
- Although the stagewise design alleviates the overhead of full-information encoding, Stage 1 still scores the entire candidate set once, which may limit scalability when the candidate pool is very large.
- Our experiments rely on specific snapshots of Wikipedia and Wikidata; temporal mismatch or outdated metadata may reduce coverage for time-sensitive entities.
- DuCEI may inherit coverage and popularity biases from Wikipedia/Wikidata resources and frequency-based candidate priors. These biases may disadvantage long-tail entities, underrepresented entity types, and languages or domains with lower knowledge-base coverage.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea Government (MSIT) (No. RS-2020-II201336, Artificial Intelligence Graduate School Program (UNIST); No. RS-2023-00216011, Development of Artificial Complex Intelligence for Conceptually Understanding and Inferring like Human; No. RS-2026-25615995, AI Star Fellowship Support Program (JBNU)), and by LG AI Research.

References

- Oshin Agarwal and Daniel M Bikel. 2020. [Entity linking via dual and cross-attention encoders](#). *arXiv preprint arXiv:2004.03555*.
- Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022. [ReFinED: An efficient zero-shot-capable approach to end-to-end entity linking](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 209–220, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.
- Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2022. [ExtEnD: Extractive entity disambiguation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2478–2488, Dublin, Ireland. Association for Computational Linguistics.
- Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. [Longformer: The long-document transformer](#). *arXiv preprint arXiv:2004.05150*.
- Shuang Chen, Jinpeng Wang, Feng Jiang, and Chin-Yew Lin. 2020. [Improving entity linking by modeling latent entity type information](#). In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7529–7537.
- Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. [Autoregressive entity retrieval](#). In *International Conference on Learning Representations*.
- Yifan Ding, Qingkai Zeng, and Tim Weninger. 2024. [ChatEL: Entity linking with chatbots](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3086–3097, Torino, Italia. ELRA and ICCL.
- Octavian-Eugen Ganea and Thomas Hofmann. 2017. [Deep joint entity disambiguation with local neural attention](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2619–2629, Copenhagen, Denmark. Association for Computational Linguistics.
- Zhaochen Guo and Denilson Barbosa. 2018. [Robust named entity disambiguation with random walks](#). *Semantic Web*, 9:459–479.
- Nitish Gupta, Sameer Singh, and Dan Roth. 2017. [Entity linking via joint encoding of types, descriptions, and context](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2681–2690, Copenhagen, Denmark. Association for Computational Linguistics.
- Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstena, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum.

2011. [Robust disambiguation of named entities in text](#). In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- Phong Le and Ivan Titov. 2018. [Improving entity linking by modeling latent relations between mentions](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1595–1604, Melbourne, Australia. Association for Computational Linguistics.
- Belinda Z. Li, Sewon Min, Srinivasan Iyer, Yashar Mehdad, and Wen-tau Yih. 2020. [Efficient one-pass end-to-end entity linking for questions](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6433–6441, Online. Association for Computational Linguistics.
- Yajie Li, Albert Galimov, Mitra Datta Ganapaneni, Pujitha Thejaswi, De Meng, Priyanshu Kumar, and Saloni Potdar. 2025. [Leveraging the power of large language models in entity linking via adaptive routing and targeted reasoning](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 871–882, Suzhou (China). Association for Computational Linguistics.
- Xukai Liu, Ye Liu, Kai Zhang, Kehang Wang, Qi Liu, and Enhong Chen. 2024. [OneNet: A fine-tuning free framework for few-shot entity linking via large language model prompting](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13634–13651, Miami, Florida, USA. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [RoBERTa: A robustly optimized BERT pretraining approach](#). *arXiv preprint arXiv:1907.11692*.
- Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. [Zero-shot entity linking by reading entity descriptions](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3449–3460, Florence, Italy. Association for Computational Linguistics.
- Khalil Mrini, Shaoliang Nie, Jiatao Gu, Sinong Wang, Maziar Sanjabi, and Hamed Firooz. 2022. [Detection, disambiguation, re-ranking: Autoregressive entity linking as a multi-task problem](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1972–1983, Dublin, Ireland. Association for Computational Linguistics.
- Kristian Noullet, Ayoub Ourgani, Niklas Thomas Lakner, Lukas Kinder, and Tobias Käfer. 2025. [Agnus LLM: Robust and flexible entity disambiguation with decoder-only language models](#). In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 3266–3284, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. [PyTorch: An imperative style, high-performance deep learning library](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Luigi Procopio, Simone Conia, Edoardo Barba, and Roberto Navigli. 2023. [Entity disambiguation with entity definitions](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1297–1303, Dubrovnik, Croatia. Association for Computational Linguistics.
- Michael Röder, Ricardo Usbeck, and Axel-Cyrille Ngonga Ngomo. 2018. [GERBIL – benchmarking named entity recognition and linking consistently](#). *Semantic Web*, 9(5):605–625.
- Darya Shlyk, Tudor Groza, Marco Mesiti, Stefano Montanelli, and Emanuele Cavalleri. 2024. [REAL: A retrieval-augmented entity linking approach for biomedical concept recognition](#). In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 380–389, Bangkok, Thailand. Association for Computational Linguistics.
- Jonghyun Song, Cheyon Jin, Wenlong Zhao, Andrew McCallum, and Jay-Yoon Lee. 2024. [Comparing neighbors together makes it easy: Jointly comparing multiple candidates for efficient and effective retrieval](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22255–22269, Miami, Florida, USA. Association for Computational Linguistics.
- 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 *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937, Singapore. Association for Computational Linguistics.
- Li Tianran, Yang Erguang, Zhang Yujie, Chen Yufeng, and Xu Jinan. 2021. [Improving entity linking by encoding type information into entity embeddings](#). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1087–1095, Huhhot, China. Chinese Information Processing Society of China.

- Johannes M. van Hulst, Faegheh Hasibi, Koen Dercksen, Krisztian Balog, and Arjen P. de Vries. 2020. **REL: An entity linker standing on the shoulders of giants**. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2197–2200. Association for Computing Machinery.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. **Attention is all you need**. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Daniel Vollmers, Hamada Zahera, Diego Moussallem, and Axel-Cyrille Ngonga Ngomo. 2025. **Contextual augmentation for entity linking using large language models**. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8535–8545, Abu Dhabi, UAE. Association for Computational Linguistics.
- Denny Vrandečić and Markus Krötzsch. 2014. **Wiki-data: a free collaborative knowledgebase**. *Commun. ACM*, 57(10):78–85.
- Junxiong Wang, Ali Mousavi, Omar Attia, Ronak Pradeep, Saloni Potdar, Alexander Rush, Umar Farooq Minhas, and Yunyao Li. 2024. **Entity disambiguation via fusion entity decoding**. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6524–6536, Mexico City, Mexico. Association for Computational Linguistics.
- Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. **Scalable zero-shot entity linking with dense entity retrieval**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.
- Zilin Xiao, Ming Gong, Jie Wu, Xingyao Zhang, Linjun Shou, and Daxin Jiang. 2023. **Instructed language models with retrievers are powerful entity linkers**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2267–2282, Singapore. Association for Computational Linguistics.
- Amy Xin, Yunjia Qi, Zijun Yao, Fangwei Zhu, Kaisheng Zeng, Bin Xu, Lei Hou, and Juanzi Li. 2025. **LL-MAEL: Large language models are good context augmenters for entity linking**. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM '25*, pages 3550–3559, New York, NY, USA. Association for Computing Machinery.
- Zhenran Xu, Yulin Chen, and Baotian Hu. 2023a. **Improving biomedical entity linking with cross-entity interaction**. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13869–13877.
- Zhenran Xu, Yulin Chen, Baotian Hu, and Min Zhang. 2023b. **A read-and-select framework for zero-shot entity linking**. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13657–13666, Singapore. Association for Computational Linguistics.
- Christophe Ye and Cassie S. Mitchell. 2025. **LLM as entity disambiguator for biomedical entity-linking**. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 301–312, Vienna, Austria. Association for Computational Linguistics.
- Guobiao Zhang, Xueping Peng, Tao Shen, Guodong Long, Jiasheng Si, Libo Qin, and Wenpeng Lu. 2024. **Extractive medical entity disambiguation with memory mechanism and memorized entity information**. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13811–13822, Miami, Florida, USA. Association for Computational Linguistics.
- Wenzheng Zhang, Wenyue Hua, and Karl Stratos. 2022. **EntQA: Entity linking as question answering**. In *International Conference on Learning Representations (ICLR)*.
- Ming Zhu, Busra Celikkaya, Parminder Bhatia, and Chandan K Reddy. 2020. **LATTE: Latent type modeling for biomedical entity linking**. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9757–9764.

A Dataset Statistics and Candidate Coverage

A.1 Dataset Overview

Table 8 summarizes the number of mentions and the average number of candidate entities for each dataset. For AIDA-CoNLL, all evaluation statistics in this appendix are computed on the test split.

Dataset	Num Mentions	Avg. Candidates
AIDA-CoNLL	4,485	51.5
MSNBC	656	27.2
AQUAINT	727	32.8
ACE2004	257	47.7
CWEB	11,154	41.3
WIKI	6,821	32.6

Table 8: Statistics of datasets used in our experiments.

A.2 Top- K Gold Inclusion Rates

To verify the coverage of the candidate dictionary, we measured the proportion of mentions whose gold entity appears within the top- K candidates ($K \in \{1, 3, 5, 10, 30, 100\}$). Table 9 summarizes the results.

Dataset	@1	@3	@5	@10	@30	@100
AIDA-CoNLL	71.8	89.7	93.6	96.5	98.8	99.8
MSNBC	89.3	97.3	97.9	98.6	98.6	98.6
AQUAINT	83.1	92.2	93.4	94.1	95.1	95.5
ACE2004	84.4	87.9	88.3	89.5	91.1	92.2
CWEB	70.0	85.6	88.5	91.3	94.3	95.6
WIKI	64.3	81.6	85.4	89.3	93.5	96.4

Table 9: Gold entity inclusion rates (%) within the top- K candidates. High recall values indicate that gold entities are usually covered by a small candidate set (e.g., top-30), supporting the feasibility of defining a reduced candidate scope in our framework.

B Metadata Construction and Coverage

B.1 Entity Type Selection

Previous work often restricts type-related relations to a small set (e.g., *instance of*, *occupation*, *country*, and *sport*) (Ayoola et al., 2022). To increase coverage beyond this limited relation set, we expand the relation inventory used for type construction and explicitly serialize the resulting type information into a compact sequence.

Relation set used for type construction (ours)
instance of
occupation
country
sport
country of citizenship
country for sport
country of origin
located in the administrative territorial entity
continent
genre
subclass of
part of
home venue
location

Table 10: Relation set used for type sequence construction in our model.

Type sequence construction. We construct a compact type sequence by collecting type-related facts from the expanded relation inventory listed in Table 10. This expanded set allows the model to capture more diverse semantic distinctions beyond commonly used relations. To keep the sequence concise, we retain at most the first two objects for each relation in their original order, merge multiple objects under the same relation as $rel: obj_1, obj_2$, and separate different relations using a dedicated delimiter token $|$. The resulting type sequence takes the form: $rel_1: obj_1, obj_2 | rel_2: obj_1 | rel_3: obj_1, obj_2$. This format preserves diverse categorical cues while controlling

input length.

B.2 Metadata Coverage

We report coverage statistics of the metadata fields used in Stage 2, computed only for gold entities in the evaluation benchmarks, i.e., per evaluation instance. If the same gold entity appears in multiple instances, it is counted multiple times.

Statistic	Value
Gold entities with a description	86.24%
Gold entities with type information	94.12%
Gold entities with both description and type	85.12%
Avg. description length (gold)	7.64 tokens

Table 11: Metadata coverage for gold entities in the evaluation benchmarks. Statistics are computed per evaluation instance; repeated gold entities are counted multiple times.

As shown in Table 11, metadata coverage for gold entities is high: 86.24% of gold-entity instances have descriptions, 94.12% have type information, and 85.12% have both fields available. Descriptions are concise on average, with 7.64 tokens per gold entity.

B.3 Robustness to Missing Metadata

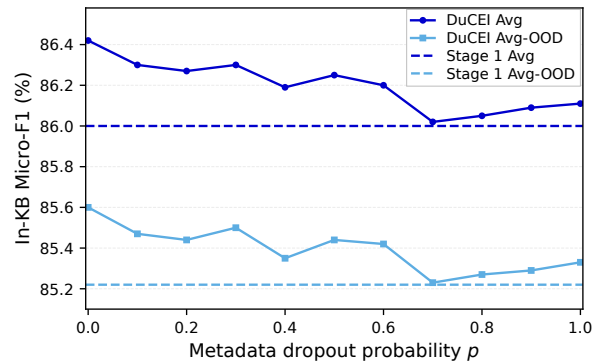


Figure 3: Inference-time metadata dropout robustness. Descriptions and type fields are randomly dropped in Stage 2 with probability p at inference time. Dashed lines indicate the performance of the Stage 1-only baseline.

We further evaluate whether DuCEI is robust to incomplete entity metadata. At inference time, we randomly drop descriptions and type fields in Stage 2 with probability p for each candidate, while keeping the model trained with full metadata. As shown in Figure 3, DuCEI shows only modest degradation as metadata becomes increasingly incomplete. The lowest performance is observed around $p = 0.7$, where Avg_{OOD} drops from 85.60 to 85.23, but the drop remains small. Both Avg and Avg_{OOD} remain

at or above the corresponding Stage 1-only baselines across all dropout probabilities. Moreover, even when all metadata fields are removed at inference time ($p = 1.0$), DuCEI still outperforms Stage 1-only on both Avg and Avg_{OOD}. This result suggests that the benefit of metadata is not limited to its direct availability at inference time; exposure to metadata during training may also help the second-stage model learn more robust candidate representations.

C Implementation Details

C.1 Training and Inference

Both stages of DuCEI use Longformer-Large with a maximum input length of 1024 tokens and global attention on the [CLS] token, the mention start token, and the first token of each entity title. When the concatenated input exceeds 1024 tokens, we retain candidates in their original order up to the last candidate that fits within the maximum input length. All DuCEI and controlled-baseline experiments were conducted on a single NVIDIA RTX A6000 GPU (48 GB) using PyTorch (Paszke et al., 2019) with mixed-precision training (fp16 autocast with loss scaling).

Since candidate order can act as an implicit prior in cross-entity interaction, we preserve the original frequency-based ordering, i.e., descending $P(e | m)$, for Stage 1 instances that use the original candidate set rather than shuffling candidates. As shown in Appendix D.2, randomizing candidate order during Stage 1 training consistently degrades performance, indicating that maintaining the prior-aligned ordering is beneficial. During inference, Stage 1 receives candidates in this prior-aligned order and ranks them. When refinement is applied, the top- k candidates are selected according to the Stage 1 scores and passed to Stage 2 in this score-based order.

During Stage 2 training, we construct inputs by sampling up to $k-1$ negatives according to Stage 1 candidate probabilities, adding the gold entity, and ordering the resulting candidates by their Stage 1 scores.

During Stage 1 training, some mentions have either an empty candidate set or only a single candidate due to dictionary coverage. To avoid degenerate training instances and improve learning efficiency, we uniformly choose the number of additional negative entities from $\{1, \dots, 99\}$ and sample them from the global training candidate pool,

after which the gold entity is inserted at a uniformly random position to form a non-trivial candidate list. This augmentation is applied only in Stage 1 (Stage 2 uses the candidate construction with up to k candidates described above).

C.2 Controlled Baseline Architectures

For the controlled comparisons in Section 6.3, we implement a bi-encoder for Stage 1 candidate reduction and a pointwise cross-encoder for Stage 2 refinement. Both baselines score candidates independently and therefore do not perform cross-entity interaction.

Bi-encoder for Stage 1 Candidate Reduction.

The bi-encoder independently encodes the mention context and each candidate entity. To match the information available to DuCEI Stage 1, the candidate-side input contains only the entity title. The bi-encoder follows the same Stage 1 candidate construction procedure described in Appendix C.1. Let X_m denote the mention-context sequence and let X_i^{BE} denote the title-only sequence for candidate e_i :

$$X_i^{\text{BE}} = [\text{CLS}] t_i [\text{SEP}].$$

The mention context and candidate title are encoded separately:

$$\mathbf{h}_m = \text{ENC}_m(X_m)_{[\text{CLS}]}, \mathbf{h}_i = \text{ENC}_e(X_i^{\text{BE}})_{[\text{CLS}]}.$$

The mention representation is projected through an MLP:

$$\mathbf{z}_m = g_m(\mathbf{h}_m).$$

Because the original candidate ordering provides an informative mention–entity prior signal, we also provide this signal explicitly to the independent bi-encoder. Let r_i^{prior} denote the rank of candidate e_i in the original mention–entity prior ordering, and let

$$\boldsymbol{\rho}_i^{\text{prior}} = \text{Emb}_r(r_i^{\text{prior}})$$

denote its learned rank embedding. The candidate representation is obtained by concatenating this rank embedding with the independently encoded title representation and applying an MLP:

$$\mathbf{z}_i = g_e([\mathbf{h}_i; \boldsymbol{\rho}_i^{\text{prior}}]).$$

The relevance score is then computed by a dot product:

$$s_i = \mathbf{z}_m^\top \mathbf{z}_i.$$

The scores are normalized with a softmax over the candidate set and optimized using cross-entropy

loss. Thus, the bi-encoder has access to the same title-level information and an explicit candidate-prior signal, but each candidate is represented and scored independently without cross-entity interaction.

Pointwise Cross-encoder for Stage 2 Refinement.

For the Stage 2 control, we retain exactly the same top- k candidates produced by DuCEI Stage 1 and replace DuCEI’s interaction-based Stage 2 with a pointwise cross-encoder. Unlike the title-only Stage 1 bi-encoder, this baseline uses the same entity metadata as DuCEI Stage 2, including the title, description, and type sequence. For each retained candidate e_i , we independently construct a mention–candidate sequence

$$X_{m,i}^{\text{CE}} = [\text{CLS}] C_m [\text{SEP}] t_i \langle /T \rangle d_i \langle /D \rangle y_i \langle /E \rangle [\text{SEP}],$$

following the same metadata composition used in DuCEI Stage 2. Each mention–candidate pair is encoded independently:

$$\mathbf{h}_i^{\text{CE}} = \text{Enc}_{\text{CE}}(X_{m,i}^{\text{CE}})_{[\text{CLS}]}.$$

We additionally encode the candidate’s rank within the Stage 1 retained set. Let $r_i^{(1)}$ denote the rank of candidate e_i according to the Stage 1 score, and let

$$\boldsymbol{\rho}_i^{(1)} = \text{Emb}_r(r_i^{(1)})$$

be the corresponding learned rank embedding. The scalar relevance score is computed as

$$s_i = g_{\text{CE}} \left([\mathbf{h}_i^{\text{CE}}; \boldsymbol{\rho}_i^{(1)}] \right),$$

where g_{CE} is an MLP. The resulting scores are normalized with a softmax over the retained top- k candidates and optimized using cross-entropy loss.

This baseline therefore controls for both the retained candidate set and the metadata supplied during refinement. A key structural difference is that the pointwise cross-encoder processes each mention–candidate pair separately, whereas DuCEI Stage 2 jointly encodes all retained candidates within a shared sequence and allows explicit cross-entity interaction.

C.3 Hyperparameter Tuning

We tuned Stage 1 and Stage 2 with separate learning-rate ranges. For Stage 1, we searched learning rates in $\{5\text{e-}6, 1\text{e-}5, 2\text{e-}5, 3\text{e-}5\}$, and

for Stage 2 in $\{1\text{e-}6, 2.5\text{e-}6, 5\text{e-}6, 7.5\text{e-}6\}$. For both stages, we tuned the number of training epochs in $\{3, 4, 5\}$. Mention masking is a training-time augmentation that replaces tokens inside the marked mention span with [MASK] with probability p , encouraging the model to rely on surrounding context and candidate evidence rather than only surface-form cues. After selecting the best learning rate and number of epochs, we selected the final model by additionally tuning the mention masking probability in $\{0.1, 0.2, 0.3, 0.4, 0.5\}$.

Tables 12 and 13 summarize the final hyperparameters used for Stage 1 and Stage 2.

Setting	Value
Encoder	Longformer-Large
Window size	256
Dropout	0.1
Train / Eval batch	8 / 64
Optimizer	AdamW
β_1/β_2	0.9 / 0.999
ϵ	1e-8
Scheduler	warmup linear
Warmup ratio	0.1
Weight decay	0.01

Table 12: Common hyperparameters used in both stages.

Setting	Stage 1	Stage 2
Learning rate	1e-5	5e-6
Epochs	3	4
Top- k	—	10
Mention masking p	0.3	0.2

Table 13: Stage-specific hyperparameters.

Controlled Baselines. For the controlled bi-encoder and pointwise cross-encoder baselines, we use RoBERTa-Large (Liu et al., 2019) instead of Longformer-Large, while keeping the remaining common optimization settings identical to those in Table 12. We fix the learning rate to 1e-5 and select the number of training epochs from $\{3, 4, 5\}$ based on AIDA-dev performance. After selecting the best epoch, we tune the mention masking probability p over $\{0.0, 0.1, 0.2, 0.3\}$. Each baseline is tuned independently using the same procedure, and we report its best-performing configuration.

D Additional Experiments

D.1 Scoring Function Comparison in Stage 1

Besides our main **classification** scoring function (defined in Section 4.2), we additionally compare two alternative designs frequently used in prior work.

Dot-product similarity. We compute the matching score as the dot-product between the mention-context representation $\mathbf{r}_m$ and entity representation $\mathbf{r}_i$:

$$s_i = \mathbf{r}_m^\top \mathbf{r}_i.$$

Extractive (span-based) scoring. Inspired by MRC-style formulations, the model predicts start and end logits for each entity title span (a_i, b_i) within the sequence:

$$s_i = \frac{1}{2} (\mathbf{W}_{\text{start}} \mathbf{h}_{a_i} + \mathbf{W}_{\text{end}} \mathbf{h}_{b_i}),$$

where $\mathbf{W}_{\text{start}}$ and $\mathbf{W}_{\text{end}}$ are learnable projection vectors.

Table 14 summarizes the comparison results. Based on these results, we use the classification scoring function by default.

Scoring Function	Avg	Avg _{good}
Dot-product	85.39	84.48
Extractive (span-based)	85.37	84.45
Classification (ours)	85.62	84.80

Table 14: Comparison of scoring functions for Stage 1 with mention masking disabled ($p = 0$).

D.2 Effect of Prior-Based Candidate Order during Stage 1 Training

In our candidate dictionary, entities are sorted in descending order of their empirical mention-entity frequency, which implicitly reflects the prior probability $P(e \mid m)$. Accordingly, Stage 1 receives not only the mention context and entity titles, but also an informative candidate ordering induced by the mention-entity prior. Since Stage 1 jointly encodes the ordered candidate list, it can exploit this ordering through positional information together with title-based semantic evidence and cross-entity interaction.

To examine the contribution of this signal, we conduct a control experiment with mention masking probability $p = 0$, where the candidate list is randomly shuffled during training while the original prior-based order is retained during evaluation. This prevents the model from learning a stable association between candidate position and mention-entity frequency. For reference, we also report a prior-only top-1 baseline that directly selects the highest-prior candidate.

As shown in Table 15, the full Stage 1 model with the original prior-based order substantially outperforms the prior-only top-1 baseline. This indicates that Stage 1 does not simply reproduce the

Stage 1 Setting ($p = 0$)	Avg	Avg _{good}
Prior-only top-1	77.86	79.07
Original prior-based order	85.62	84.80
Shuffled order during training	82.76	81.60

Table 15: Effect of prior-based candidate order during Stage 1 training under $p = 0$. The gap from the prior-only baseline shows that Stage 1 does not merely reproduce the candidate prior, while the degradation under shuffled-order training supports the usefulness of prior-induced ordering as an additional signal.

candidate-prior ranking, but combines mention context, entity-title information, and cross-entity interaction with the prior-induced ordering signal.

At the same time, shuffling candidates during training substantially degrades both Avg and Avg_{good}. This suggests that the consistent prior-based ordering provides an important additional cue that the model can exploit through positional information. Thus, the strong performance of Stage 1 appears to result from combining cross-entity interaction with effective use of candidate-prior information, rather than relying exclusively on either signal.

D.3 Negative Candidate Sampling Strategies in Stage 2

During Stage 2 training, we construct an up-to- k candidate set by inserting the gold entity and selecting up to $k - 1$ negatives using one of the following strategies:

- **Greedy:** take the top- $(k-1)$ negatives by Stage 1 scores.
- **Random:** sample $k - 1$ negatives uniformly from the candidate pool.
- **Probability-based:** sample $k - 1$ negatives from the Stage 1 softmax distribution $p_i \propto \exp(s_i)$.

Sampling Strategy	Avg	Avg _{good}
Probability (used in our main setting)	86.33	85.55
Random	86.32	85.56
Greedy	86.33	85.51

Table 16: Comparison of Stage 2 negative sampling strategies with mention masking disabled ($p = 0$).

Table 16 summarizes the results. Although all three strategies achieve comparable accuracy, we adopt probability-based sampling in our final model to introduce controlled stochasticity while largely preserving the Stage 1 ranking.

D.4 Development Threshold Analysis for Selective Stage 2 Refinement

τ	Dev F1	Stage 1 Decision	Stage 2 Usage
0.000	94.75	100.00%	0.00%
0.900	94.98	93.40%	6.60%
0.970	95.00	87.72%	12.28%
0.990	95.08	83.82%	16.18%
0.991	95.10	83.36%	16.64%
0.993	95.13	82.00%	18.00%
0.995	95.15	80.65%	19.35%
0.999	95.17	70.58%	29.42%
1.000	95.19	0.00%	100.00%

Table 17: Analysis of the confidence threshold τ for selective Stage 2 refinement. On the AIDA-CoNLL development set, Stage 2 is applied only when the highest Stage 1 prediction confidence is below τ . Boldface indicates the final selected threshold.

We analyze the accuracy–efficiency trade-off induced by the confidence threshold τ on AIDA-dev. As shown in Table 17, the sweep exhibits a gradual trade-off rather than a broad performance plateau: increasing τ invokes Stage 2 for more mentions and progressively approaches the performance of full two-stage inference.

Based on this trade-off, we identify the range $\tau = 0.99\text{--}0.995$ as a reasonable operating region, where development performance is close to full Stage 2 while Stage 2 usage remains substantially lower. Within this region, $\tau = 0.991$ is the smallest tested threshold whose development F1 is within 0.1 points of full Stage 2 (95.10 vs. 95.19). We therefore use it as a representative efficiency-oriented operating point and keep it fixed for all test sets.

D.5 Prediction Transitions and Conditional Analysis of Stage 2

To better characterize when metadata-enriched refinement helps or hurts, we analyze prediction transitions from Stage 1 to Stage 2 and examine how these transitions vary with metadata availability and the retained candidate scope.

Transition	Full Stage 2	Selective Stage 2
Correct $\rightarrow$ Correct	19,206 (80.01%)	19,216 (80.06%)
Correct $\rightarrow$ Wrong	103 (0.43%)	93 (0.39%)
Wrong $\rightarrow$ Correct	170 (0.71%)	159 (0.66%)
Wrong $\rightarrow$ Wrong	4,524 (18.85%)	4,535 (18.89%)

Table 18: Prediction transitions from Stage 1 under full and selective Stage 2 refinement over 24,003 mentions with non-empty candidate sets. Selective refinement uses $\tau = 0.991$.

Prediction transitions. Table 18 summarizes changes in prediction correctness from Stage 1 to full and selective Stage 2 refinement. For full Stage 2, 170 Stage 1 errors are corrected, while 103 initially correct predictions become incorrect, yielding a net improvement of 67 predictions. Selective refinement with $\tau = 0.991$ retains most of these beneficial transitions, correcting 159 errors while introducing 93 new errors, for a net improvement of 66 predictions.

These results show that the overall Stage 2 improvement arises from a relatively small number of prediction changes, with beneficial changes outweighing harmful ones. Moreover, selective refinement preserves nearly all of the net improvement of full Stage 2 while avoiding refinement for most mentions, consistent with the accuracy–efficiency results in Section 6.4.

For the following conditional analyses, we focus on the 22,473 mentions for which the gold entity is contained in the top-10 candidate set retained for Stage 2. Here, *Help* denotes a wrong-to-correct transition and *Hurt* denotes a correct-to-wrong transition.

Effect of gold-entity metadata availability. We first examine whether Stage 2 behavior depends on the availability of the gold entity’s description and type information. Table 19 shows that the proportion of helpful changes is similar whether both metadata fields are available or at least one is missing.

When both fields are available, 62.55% of changed predictions are helpful, compared with 60.00% when at least one field is missing. Both groups exhibit a positive net effect, and the difference between them is small. Thus, gold-entity metadata availability alone does not appear to strongly predict whether Stage 2 refinement will help or hurt.

Effect of retained candidate scope. We further analyze prediction changes according to the number of candidates retained for Stage 2. As shown in Table 19, Stage 2 changes predictions more frequently as the retained candidate set grows.

Although the net effect remains positive across all non-trivial candidate sizes, the proportion of harmful changes increases from 25.71% for 2–5 candidates to approximately 38–40% for 6–9 and 10 candidates. This suggests a trade-off between candidate coverage and interaction interference: retaining more candidates reduces the risk of exclud-

Condition	<i>N</i>	Changed	Help	Hurt	Net
<i>(a) Gold-entity metadata availability</i>					
Both available	19,325	243 (1.26%)	152 (62.55%)	91 (37.45%)	+61
At least one missing	3,148	30 (0.95%)	18 (60.00%)	12 (40.00%)	+6
Total	22,473	273 (1.21 %)	170 (62.27 %)	103 (37.73 %)	+67
<i>(b) Number of candidates retained for Stage 2</i>					
1	4,982	0 (0.00%)	—	—	0
2–5	3,678	35 (0.95%)	26 (74.29%)	9 (25.71%)	+17
6–9	1,368	21 (1.54%)	13 (61.90%)	8 (38.10%)	+5
10	12,445	217 (1.74%)	131 (60.37%)	86 (39.63%)	+45
Total	22,473	273 (1.21 %)	170 (62.27 %)	103 (37.73 %)	+67

Table 19: Conditional analysis of prediction changes under full Stage 2 refinement. Panel (a) groups instances by gold-entity metadata availability, and Panel (b) groups them by the number of candidates retained for Stage 2. Percentages in the Help and Hurt columns are computed among changed predictions.

ing the gold entity, but also introduces more competing metadata into the shared interaction scope.

This pattern is consistent with Figure 2, where increasing k improves pruning Recall@ k but does not monotonically improve final performance, and with the degradation of single-step full-information encoding for larger candidate pools. Together, these analyses support keeping the metadata-enriched interaction scope compact: Stage 2 can resolve difficult cases when multiple plausible candidates remain, but broadening its interaction scope can also increase harmful refinements.

E Broader Positioning among Recent ED/EL Systems

Our main comparison focuses on AIDA-CoNLL-only supervised systems under comparable candidate-based settings. Recent LLM-based and LLM-assisted systems are discussed here primarily for broader positioning, since their prompting protocols, candidate construction procedures, and evaluation settings often differ from our primary supervised comparison.

Table 20 positions DuCEI among recent ED/EL systems and is not intended as a strict head-to-head comparison. The listed systems differ in benchmark coverage, task formulation, candidate construction, supervision, model access, prompting protocol, metric, or evaluation setting. We also include representative retrieval-based ED/EL systems to clarify how DuCEI differs from retrieve-then-rerank pipelines.

Connections with LLM-based ED. DuCEI is complementary to recent LLM-based ED approaches in several ways:

- **LLM-based context augmentation.** LLM-augmented mention contexts, as used in LL-

MAEL, could be provided as input to DuCEI, combining richer contextual evidence with cross-entity candidate comparison.

- **LLM-based candidate-scope reduction.** DuCEI’s first-stage scope reduction could be implemented using LLM-based listwise candidate comparison, following recent LLM-based reranking approaches (Sun et al., 2023), with prompting (e.g., in-context learning) or fine-tuning, followed by metadata-enriched interaction over only the reduced candidate set.
- **Selective LLM routing.** DuCEI’s confidence-based selective mechanism could be extended to route only difficult or uncertain mentions to an LLM, as explored in ARTER (Li et al., 2025), while resolving confident cases without additional LLM inference.

These directions preserve DuCEI’s broad-to-focused principle while allowing LLMs to augment the mention context, participate in candidate-scope reduction, or handle selectively routed difficult cases.

F Case Studies

We provide four case examples that illustrate the main behaviors and limitations of DuCEI. Tables 21 and 22 show two boundary cases: an easy case where Stage 1 already ranks the gold entity first, and a gold-not-in-candidates failure where the gold entity is absent from the initial candidate pool. Tables 23 and 24 show two ambiguous cases that reveal the key behavior of the stagewise design: when Stage 1 retains the gold entity within the top- k set, Stage 2 can resolve fine-grained ambiguity through focused metadata-enriched interaction; however, when Stage 1 prunes the gold entity out of the top- k set, Stage 2 cannot recover it.

Model	Setting / Relevance
<i>Ours</i> DuCEI	Fixed-candidate in-KB ED model that introduces metadata only after reducing the interaction scope.
<i>Recent LLM-based or LLM-assisted ED/EL systems</i>	
ChatEL (Ding et al., 2024)	Three-step LLM prompting framework for ED that generates candidates, augments the mention with auxiliary context, and performs multiple-choice selection.
OneNet (Liu et al., 2024)	Fine-tuning-free few-shot ED framework with LLM-prompted entity reduction, dual-perspective linking, and consensus judgment.
AGNUS (Noullet et al., 2025)	Zero-shot decoder-only LLM ED framework with order-invariant encoding and constrained decoding.
ARTER (Li et al., 2025)	Adaptive ED framework that routes difficult mentions to targeted LLM reasoning.
INSGENEL (Xiao et al., 2023)	Instruction-tuned generative EL system with a retrieval-augmented variant.
AugmentedEL (Vollmers et al., 2025)	T5-based EL system with LLM-based mention and abbreviation expansion.
<i>Related retrieval-based ED/EL systems</i>	
BLINK (Wu et al., 2020)	Retrieval-based ED/EL system with bi-encoder candidate retrieval and cross-encoder reranking.
ELQ (Li et al., 2020)	One-pass end-to-end EL system with joint mention detection and linking.
EntQA (Zhang et al., 2022)	End-to-end EL system that retrieves entities and finds their mentions via QA.

Table 20: Broader positioning among recent ED/EL systems.

Field	Content
Context	... In 2000, King Diamond released the album House of God , which peaked at No. 60 in Sweden. After the release of 2002’s album, which peaked at No. 42 in Sweden and No. 24 in Finland, the band’s line-up has remained stable to this day. ...
Total candidates	The House of God; House of God (album) ; Church of God, House of Prayer; Place of worship; Temple in Jerusalem
Stage 1 (top-5)	House of God (album) ; The House of God; Church of God, House of Prayer; Place of worship; Temple in Jerusalem
Stage 2 (final)	House of God (album)

Table 21: Easy case: Stage 1 ranks the gold entity at the top, and Stage 2 preserves the prediction.

Easy case. Table 21 shows a straightforward instance where the surrounding context explicitly indicates that the mention refers to an *album*. Stage 1 ranks the gold entity at the top based on titles alone, and Stage 2 preserves the same prediction.

Gold-not-in-candidates failure. Table 22 shows a coverage error where the gold entity is absent from the candidate pool. Although the context suggests a cycling-related interpretation, neither stage can recover the correct entity because it is not included among the initial candidates.

Two-stage success. Table 23 illustrates a challenging case where the surface form “League” is highly ambiguous and Stage 1 retains many plausible sports-related candidates. Stage 2 resolves the ambiguity by leveraging KB metadata of *Lega Nord*: its description (“Italian political party”) and type fields (e.g., *instance of*: political party; *country*: Italy) align strongly with the political context (e.g., references to separatism, federal reform, and

Italian politics). As a result, Stage 2 corrects the decision by favoring the only candidate whose metadata is compatible with the surrounding evidence.

Stage 1 pruning failure. Table 24 shows a pruning failure where the gold entity is present in the initial candidate set but removed by Stage 1. Although the context clearly refers to Iowa counties, the ambiguous surface form “Palo Alto” is strongly biased toward California-related entities. Without metadata cues in Stage 1, the gold entity *Palo Alto County, Iowa* is ranked below the top- k cutoff. Unlike gold-not-in-candidates failures, this case is not caused by candidate coverage; rather, it exposes an unrecoverable bottleneck of the two-stage design.

Field	Content
Context	... “I gave it everything my body had,” said Wellons. Verge NECCS Series Results. Elite Men: Wins at the Gloucester and Northampton races highlighted Anthony’s series victory in his first year as an Elite. For Chris Jones NERAC, the series was an accelerated class in cyclocross. In his first full season as a crosser, Jones won at Farmington, Conn., and finished the series in second. Cross legend Mark McCormack, who has won the series seven times, was on the podium again in his first season as a part-time racer. Elite Women: Bessette dominated the series even while skipping two events. ...
Total candidates	Chris Jones (1990s outfielder); Christopher Jones; Chris Jones (rugby union); Obsidian Entertainment; Chris Jones (Access Software); Chris Jones (filmmaker); Chris Jones (footballer, born 1945); Chris Jones (footballer, born 1989); Chris Jones (American musician); Chris Jones (1980s outfielder); Chris Jones (footballer, born 1956); Chris Jones (gymnast); Adventure Game Studio; Chris T. Jones; John Chris Jones; Chris Jones (cricketer); Christopher Jones (Mayflower captain); Chris Jones (punter); Baltimore Orioles minor league players; Chris Jones (rugby, born 1982)
Stage 1 (top-10)	Christopher Jones; Chris Jones (filmmaker); Chris Jones (footballer, born 1945); Chris Jones (1980s outfielder); Chris Jones (1990s outfielder); Chris Jones (Access Software); Chris Jones (footballer, born 1956); Chris Jones (cricketer); Christopher Jones (Mayflower captain); Chris Jones (gymnast); (<i>gold not included</i>)
Stage 2 (final)	Chris Jones (footballer, born 1945) (<i>cannot recover without gold in candidates</i>)
Gold entity	Chris Jones (sprinter)

Table 22: Gold-not-in-candidates failure: the gold entity is absent from the candidate set, so Stage 2 cannot recover.

Field	Content
Context	... Italian President Oscar Luigi Scalfaro visited the symbolic heartland of the separatist Northern League on Friday and appealed to its supporters to drop their campaign for a breakaway state. Addressing a convention on Italian unity in Mantua, Scalfaro made a direct appeal to what he called “my friends from the League ,” urging them to work instead for federal reform. He later warned, “But turn back from the line you are taking now,” and was jeered by a small group of League supporters. ...
Total candidates	English Football League; League (unit); Football League First Division; Leagues of China; La Liga; Scottish Premier League; Rugby league; 1994 NFL season; Premier League; ...; Lega Nord ; ...; Serbian SuperLiga; Russian Premier League
Stage 1 (top-10)	Ligue de Football Professionnel; Primeira Liga; La Liga; Lega Nord ; Super League XIV; Liga Nacional de Fútbol Profesional de Honduras; National League North; Super League XIII; League (unit); Ekstraklasa
Stage 2 (final)	Lega Nord

Table 23: Hard case: the mention is ambiguous from titles alone (e.g., sports leagues), but political context cues lead to the correct entity. Stage 1 ranks the gold entity within top- k , and Stage 2 correctly selects it using richer entity information.

Field	Content
Context	... From 1883 to 1886, the district included the north central Iowa counties of Worth, Cerro Gordo, Franklin, Hardin, Story, Wright, Hamilton, Humboldt, Hancock, Kossuth, Winnebago, Webster, and Boone. Reapportionment in 1886 reflected the increasing population balance between eastern and western Iowa, resulting in a westward shift of the district’s boundaries. From 1886 until 1933, the district was made up of Boone, Calhoun, Carroll, Crawford, Emmet, Greene, Hamilton, Humboldt, Kossuth, Palo Alto , Pocahontas, Winnebago, and Webster counties. ...
Gold entity	Palo Alto County, Iowa
Stage 1 (top-10)	Palo Alto, California; Palo Alto (disambiguation); East Palo Alto, California; Palo Alto Airport; Palo Alto, Mississippi; Palo Alto High School; Palo Alto, Aguascalientes; Palo Alto, Pennsylvania; Palo Alto (short story collection); Palo Alto Unified School District (<i>gold not included</i>)
Stage 2 (final)	Palo Alto, California (<i>cannot recover because the gold entity is pruned</i>)
Failure reason	The surface form is highly ambiguous and strongly biased toward California-related entities. Although the local context indicates an Iowa county, the title-only Stage 1 model ranks the gold county entity below the top- k cutoff without metadata cues.

Table 24: Stage 1 pruning failure: the gold entity is present in the initial candidate set but excluded from the Stage 1 top- k set, making recovery impossible in Stage 2.