Entity Alignment for LLM-Generated Knowledge Graphs: Benchmark and Analysis

Anonymous ACL submission

Abstract

Recent advances in large language models (LLMs) have accelerated text-to-knowledge graph (Text2KG) construction, increasing the need to integrate LLM-generated KGs with existing human-curated KGs and with one another. Entity alignment (EA) is a fundamental step in KG integration, but existing EA benchmarks mainly focus on human-curated graph pairs and do not capture compound Text2KG-induced discrepancies in entity/triple coverage, naming, and structural variation. To address this gap, we introduce three EA datasets—ReDocRED-EA, BioREDirect-EA, and REDFM-EA—and evaluate representative EA methods under four graph-pair settings: human-curated vs. human-curated (H-H), human-curated vs. LLM-generated (H-L), same-pipeline LLM-generated (L-L), and cross-pipeline LLM-generated (L1-L2). Across datasets, performance on human-curated graph pairs does not directly transfer to generated-graph settings; same-pipeline generated KGs are consistently easier to align than cross-pipeline KGs; and upgrading the LLM inside a Text2KG pipeline does not by itself guarantee easier H-L alignment. Fine-grained analyses further show that semantic signals remain robust, while graph-derived signals become less reliable under generator-induced coverage and structural variation. Our benchmarks and analyses provide a reproducible testbed for EA in the Text2KG era.

1 Introduction

Knowledge graphs (KGs) provide structured representations of entities and their relations, supporting downstream applications such as retrieval-augmented generation (Xu et al., 2024; Min et al., 2025) and question answering (Ma et al., 2025). Because manually constructing such KGs is costly, recent work increasingly relies on large language model (LLM)-based text-to-knowledge-graph (Text2KG) pipelines to construct KGs from text (Wei et al., 2024; Zhang and Soh, 2024; Mo et al., 2025; Min et al., 2025). As these LLM-generated KGs are built from diverse documents, domains, and generation pipelines, integrating them with human-curated KGs or with other generated KGs has become an important practical problem (Lairgi et al., 2024; Yang et al., 2025).

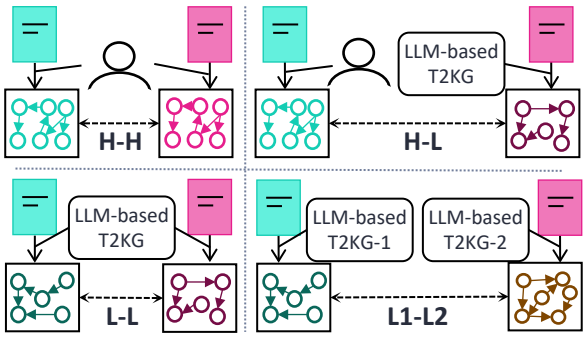

Figure 1: Overview of four practical entity alignment (EA) settings under LLM-based text-to-knowledge-graph (Text2KG) pipelines. **H-H** aligns two human-curated KGs. **H-L** aligns a human-curated KG with an LLM-generated KG. **L-L** aligns two LLM-generated KGs produced by the same Text2KG pipeline. **L1-L2** aligns two LLM-generated KGs produced by different Text2KG pipelines, including differences in the underlying LLM and/or generation procedure.

For such integration, entity alignment (EA) is a fundamental task that identifies entities that refer to the same real-world object or concept across graphs. Existing EA methods exploit graph structure, attributes, and LLMs' prior knowledge, and they have achieved strong performance on standard benchmarks (Mao et al., 2021a,b; Tang et al., 2023; Jiang et al., 2024; Chen et al., 2024; Jin et al., 2025; Long et al., 2025). However, these benchmarks mainly evaluate human-curated or synthetically perturbed KG pairs, leaving it under-explored whether the same methods remain reliable when one or both graphs are generated by Text2KG pipelines.

This setting is challenging because Text2KG pipelines can change which entities and relations are realized, how entities are named or canonicalized, and which triples are extracted. These coupled changes affect counterpart availability, lexical compatibility, and structural evidence used by EA methods. Standard EA benchmarks (Sun et al., 2017, 2018; Guo et al., 2019) and synthetic-noise protocols (Tong et al., 2022; Zhang et al., 2024; Yoon et al., 2025) are useful for studying human-curated KG differences or isolated perturbations, but they do not directly model discrepancies arising from the graph generation process itself. This motivates systematic evaluation across human-curated and LLM-generated KGs.

For this purpose, we construct three EA datasets for LLM-generated KGs: *ReDocRED-EA*, *BioREDirect-EA* and *REDFM-EA*. Each dataset starts from manually verified text-graph pairs and builds human-validated H-H alignments with non-trivial entity-name discrepancies. We then apply representative Text2KG pipelines to instantiate H-L, L-L, and L1-L2 settings (Fig. 1). This design separates semantic gold construction from configuration-specific graph realization. As a result, we can evaluate Text2KG-induced discrepancies across general and specialized domains while keeping the underlying semantic correspondences human-validated.

Our experiments show that Text2KG-induced discrepancies substantially alter the reliability of evidence used for EA. Performance on human-curated graph pairs does not directly transfer to generated-graph settings, especially in specialized-domain and cross-pipeline alignment. Across settings, same-pipeline generated graphs are consistently easier to align than cross-pipeline generated graphs, suggesting that pipeline consistency preserves more compatible entity and relation realizations. Across methods, semantic signals from entity names and LLM prior knowledge provide strong anchors, but graph-derived signals become reliability-sensitive under generator-induced coverage and structural variation. We also find that upgrading the LLM inside a Text2KG pipeline does not by itself guarantee easier H-L alignment, indicating that upstream generation capability and downstream alignability are related but distinct properties. Together, these findings indicate that Text2KG EA requires methods that assess when generated graph evidence is reliable, rather than simply combining graph and semantic signals.

Our contributions are threefold:

- We formalize EA with LLM-generated KGs through four practical settings—H-H, H-L, L-L, and L1-L2—and release three human-validated EA datasets covering both general and specialized domains.[1]
- We systematically evaluate representative EA methods across Text2KG pipelines, LLM choices, and domains, characterizing how generated-graph discrepancies affect downstream EA performance.
- We provide practical guidelines for Text2KG-era EA: use a consistent Text2KG pipeline when downstream EA is anticipated, rely on robust semantic anchors from entity names and LLM prior knowledge, and treat graph evidence as a reliability-sensitive signal under generated-graph discrepancies.

2 Related Work

2.1 LLM-based KG Construction

Text2KG research has developed a range of LLM-based pipelines for converting unstructured text into KGs. Recent approaches extract triples via prompting or constrained generation and improve output quality through decomposition, normalization, canonicalization, and post-processing (Wei et al., 2024; Bi et al., 2024; Zhang and Soh, 2024; Carta et al., 2024; Mo et al., 2025). However, LLM-generated KGs have exhibited systematic quality issues, including inconsistent naming, over- or under-coverage, relation-label drift, and hallucinations (Mihindukulasooriya et al., 2023; Zhang and Soh, 2024; Carta et al., 2024; Kamaladdini Ezzabady and Benamara, 2025). These issues matter not only for KG construction quality, but also for downstream KG integration, where generated KGs must be aligned with existing human-curated KGs or with other generated KGs produced by different pipelines. This motivates our focus on entity alignment as a downstream evaluation problem for Text2KG-generated graphs.

2.2 Entity Alignment

EA provides a natural framework for studying the KG-integration problem, as it aims to identify

[1]The datasets and reconstruction code will be publicly released upon acceptance. An anonymized supplementary package includes the reconstruction scripts and datasets.

Benchmark / Setting	H	L	Mismatch Origin	Mismatch Form	Evaluation Focus
DBP15K, SRPRS	✓	–	Natural KG mismatch	Cross-lingual names / structures	Cross-lingual EA
DWY100K	✓	–	Natural KG mismatch	Schema differences	Large-scale EA
DBP2.0	✓	–	Natural KG mismatch	Dangling entities	Dangling-aware EA
Synthetic-noise EA	✓	–	Injected perturbations	Spurious edges, noisy names	Robustness to isolated noise
Ours	✓	✓	Text2KG generation	Entity/triple coverage, naming, structural variation	EA with LLM-generated KGs

Table 1: Positioning of our benchmark relative to representative EA benchmarks and noise-aware evaluation settings. H and L denote human-curated and LLM-generated KGs, respectively. Unlike conventional EA benchmarks, our benchmark explicitly includes LLM-generated KGs and treats Text2KG-induced discrepancies as part of the evaluation condition.

equivalent entities across different KGs. Existing EA methods have addressed different assumptions and sources of difficulty in the alignment task. *Closed-set* methods assume that every source entity has a counterpart and align entities by exploiting structural, relational, or name-based compatibility (Chen et al., 2017; Wang et al., 2018; Mao et al., 2021a). *Open-set* (dangling-aware) methods relax this assumption by explicitly identifying entities that have no counterpart (Sun et al., 2021; Luo and Yu, 2022; Yin et al., 2024; Xu et al., 2025). *Noise-aware* methods further study imperfect KG conditions, such as corrupted entity names and missing or spurious edges (Tong et al., 2022; Zhang et al., 2024; Yoon et al., 2025). More recently, *LLM-based* methods have used pretrained language models to incorporate textual semantics into EA and compensate for weak graph signals or limited supervision (Jiang et al., 2024; Chen et al., 2024; Jin et al., 2025; Long et al., 2025; Cheng et al., 2025). These methods provide important foundations for EA, but they do not directly address how such difficulties interact when the KGs to be aligned are generated by Text2KG pipelines. In such LLM-generated KGs, generator-specific entity and triple coverage, naming, and structural variation jointly determine counterpart availability, lexical compatibility, and structural or relational signals. This motivates evaluating EA in settings where these discrepancies arise from the graph generation process itself, rather than only from conventional human-curated differences or synthetically injected perturbations.

2.3 Benchmarks and Noise Models for EA

EA benchmarks define the evaluation conditions under which alignment methods are tested, but existing resources primarily evaluate pairs of human-curated KGs. Representative EA benchmarks operationalize such differences as cross-lingual naming differences (Sun et al., 2017; Guo et al., 2019), schema differences (Sun et al., 2018), or curated dangling entities (Sun et al., 2021; Luo and Yu, 2022). Synthetic-noise studies broaden these evaluation settings by injecting controlled perturbations, including edge deletion (Tong et al., 2022), spurious triples (Zhang et al., 2024), and noisy entity names (Yoon et al., 2025). While these resources are valuable for isolating individual factors, they do not make LLM-generated KGs part of the evaluation condition and therefore do not capture coupled, generator-dependent discrepancies. As summarized in Table 1, our benchmark complements prior resources by explicitly including LLM-generated graphs and treating Text2KG-induced discrepancies in the evaluation setting.

3 Preliminaries

3.1 Knowledge Graphs and Entity Alignment

We denote a KG as $G = (E, R, T)$, where E, R, and $T \subseteq E \times R \times E$ are the sets of entities, relations, and triples, respectively. Given a source graph $G_{\text{src}} = (E_{\text{src}}, R_{\text{src}}, T_{\text{src}})$ and a target graph $G_{\text{tgt}} = (E_{\text{tgt}}, R_{\text{tgt}}, T_{\text{tgt}})$, EA aims to identify equivalent entity pairs $M = \{(e_{\text{src}}, e_{\text{tgt}}) \mid e_{\text{src}} \equiv e_{\text{tgt}},\ e_{\text{src}} \in E_{\text{src}},\ e_{\text{tgt}} \in E_{\text{tgt}}\}$, where $\equiv$ denotes identity with respect to the same real-world entity. Following prior work, we restrict attributes to name attributes only (Liu et al., 2023; Yin et al., 2024; Yoon et al., 2025). Throughout this paper, source and target denote the two ordered sides of an EA instance: source-side entities are used as queries and target-side entities as alignment candidates. These terms are independent of graph provenance; either side may be human-curated or LLM-generated depending on the setting.

3.2 EA Settings Considered in This Work

We distinguish human-curated KGs $\mathcal{G}^{\mathrm{H}}$ from LLM-generated KGs $\mathcal{G}^{\mathrm{L}}$ and consider the following four practical EA settings illustrated in Fig. 1:

H-H. Both graphs are human-curated KGs, i.e., $G_{\mathrm{src}}, G_{\mathrm{tgt}} \in \mathcal{G}^{\mathrm{H}}$. To the best of our knowledge, all existing EA benchmarks, including DBP15K (Sun et al., 2017), DWY100K (Sun et al., 2018), and SRPRS (Guo et al., 2019), fall into this category.

H-L. One graph is human-curated and the other is LLM-generated, i.e., $\{G_{\mathrm{src}}, G_{\mathrm{tgt}}\} \cap \mathcal{G}^{\mathrm{H}} \neq \emptyset$ and $\{G_{\mathrm{src}}, G_{\mathrm{tgt}}\} \cap \mathcal{G}^{\mathrm{L}} \neq \emptyset$. A typical use case is extending an existing KG $G_{\mathrm{src}} \in \mathcal{G}^{\mathrm{H}}$ by aligning it with an LLM-generated KG $G_{\mathrm{tgt}} \in \mathcal{G}^{\mathrm{L}}$.

L-L. Both graphs are LLM-generated KGs produced by the *same* LLM and construction pipeline. Formally, $G_{\mathrm{src}}, G_{\mathrm{tgt}} \in \mathcal{G}^{\mathrm{L}}$ share a common generation process. This setting captures scenarios where an LLM-generated KG is expanded or complemented by another KG produced with the same method.

L1-L2. Both graphs are LLM-generated KGs, but they are constructed using *different* Text2KG pipelines and/or different underlying LLMs. We represent this via two subsets $\mathcal{G}^{\mathrm{L1}}, \mathcal{G}^{\mathrm{L2}} \subseteq \mathcal{G}^{\mathrm{L}}$ corresponding to different generation pipelines, and consider $G_{\mathrm{src}} \in \mathcal{G}^{\mathrm{L1}}$ and $G_{\mathrm{tgt}} \in \mathcal{G}^{\mathrm{L2}}$. This setting also corresponds to extending or complementing an LLM-generated KG with another LLM-generated KG, but reflects the practical case where KGs are built using diverse LLMs and pipelines.

4 Dataset Construction

We construct EA datasets for evaluating the integration of human-curated and LLM-generated KGs. For this evaluation, each dataset instance must provide: (i) a paired source and target document, (ii) an H-H graph pair with a human-validated gold alignment, and (iii) LLM-generated graphs produced from the same documents for H-L, L-L, and L1-L2 settings. To satisfy these requirements, we start from an existing human-curated text-graph pair and use it as the source-side document and graph. We then create the target-side document and H-H graph by rewriting entity names into alternative surface forms while preserving referential identity, and manually verify the resulting H-H correspondences. Finally, we apply Text2KG pipelines to the paired source and target documents to obtain LLM-generated graphs, and construct configuration-specific H-L, L-L, and L1-L2 gold alignments by filtering the validated H-H alignment to entities observable in each generated-graph configuration.

4.1 Dataset Construction Pipeline

Figure 2 illustrates the five-step pipeline that implements this construction. Steps 1–3 create H-H graph-pair instances and their human-validated H-H gold alignments. Steps 4–5 generate LLM-based KGs and construct configuration-specific gold alignments for H-L, L-L, and L1-L2.

Step 1: Initialize the source-side text-graph pair. We take an existing human-curated text-graph pair $(D_{\mathrm{src}}, G^{\mathrm{H}}_{\mathrm{src}})$ as the source side of the graph-pair instance, where D_{src} is a source document and $G^{\mathrm{H}}_{\mathrm{src}}$ is its associated human-curated KG.

Step 2: Construct human-validated H-H gold. We instruct an LLM to rewrite each source-side entity name into a distinct target-side surface form. The rewritten names are intended to preserve referential identity while introducing non-trivial entity-name discrepancies. We later analyze these name discrepancies using nine categories, such as orthographic variants, acronyms, and common-vs-official names (Appendix C.3, Table 9). Date and numeric entities are left unchanged.

We verify the candidate correspondences through automatic checks and manual verification to ensure that each pair preserves referential identity without introducing generation errors or hallucinated entities. We denote the finalized mapping by $\phi : E^{\mathrm{H}}_{\mathrm{src}} \rightarrow E^{\mathrm{H}}_{\mathrm{tgt}}$, where $E^{\mathrm{H}}_{\mathrm{tgt}} := \{\phi(e) \mid e \in E^{\mathrm{H}}_{\mathrm{src}}\}$. The resulting H-H gold alignment is $M^{\text{H-H}} = \{(e, \phi(e)) \mid e \in E^{\mathrm{H}}_{\mathrm{src}}\}$. Details of the verification procedure are provided in Appendix A.1.

Step 3: Build target text and graph. Using ϕ, we obtain the H-H target graph $G^{\mathrm{H}}_{\mathrm{tgt}}$ by replacing each entity name in $G^{\mathrm{H}}_{\mathrm{src}}$. We also rewrite the source document D_{src} into D_{tgt} so that its entity mentions are consistent with the same mapping ϕ. Thus, Steps 1–3 yield a paired H-H instance $(G^{\mathrm{H}}_{\mathrm{src}}, G^{\mathrm{H}}_{\mathrm{tgt}})$ and its human-validated gold alignment $M^{\text{H-H}}$.

Step 4: Generate LLM-based graphs. We apply a Text2KG pipeline to both D_{src} and D_{tgt} to obtain the generated graphs $G^{\mathrm{L1}}_{\mathrm{src}}$ and $G^{\mathrm{L1}}_{\mathrm{tgt}}$.

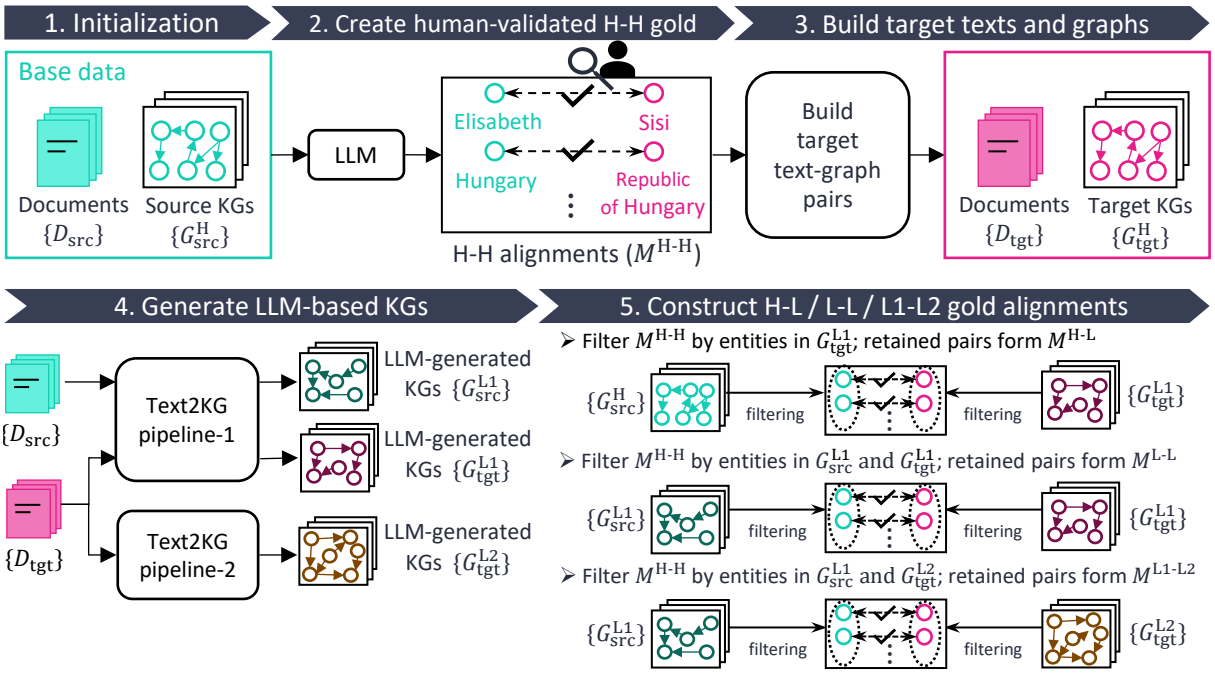

Figure 2: Overview of the dataset construction procedure. Steps 1–3 build human-curated H-H graph-pair instances, and Steps 4–5 build H-L, L-L, and L1-L2 graph-pair instances.

Together with the human-curated graphs, these graphs instantiate the H-L and L-L settings. For H-L, the graph pair is $(G^{\text{H}}_{\text{src}}, G^{\text{L1}}_{\text{tgt}})$. For L-L, the graph pair is $(G^{\text{L1}}_{\text{src}}, G^{\text{L1}}_{\text{tgt}})$, where both graphs are generated from the paired documents using the same Text2KG pipeline. For L1-L2, we apply different Text2KG pipelines and/or underlying LLMs to the source and target documents, yielding the graph pair $(G^{\text{L1}}_{\text{src}}, G^{\text{L2}}_{\text{tgt}})$.

Step 5: Construct H-L/L-L/L1-L2 gold. For each of the H-L, L-L, and L1-L2 configurations, we construct a configuration-specific gold alignment by filtering the validated H-H alignment $M^{\text{H-H}}$. We denote the resulting alignments by $M^{\text{H-L}}$, $M^{\text{L-L}}$, and $M^{\text{L1-L2}}$, respectively. Specifically, for the source and target graphs used in a given configuration, we retain a pair $(e_{\text{src}}, e_{\text{tgt}}) \in M^{\text{H-H}}$ as gold only if, after lowercasing both the gold and graph entity strings, e_{src} and e_{tgt} exactly match entities in the corresponding source and target graphs, respectively. This filtering strategy has the advantage of retaining only gold pairs that are guaranteed to be alignable in the graph pair being evaluated, since both entities must appear in the corresponding source and target graphs.

However, this strategy may miss additional alignable pairs involving entities newly introduced by Text2KG that are not present in the validated H-H alignment $M^{\text{H-H}}$. Such pairs are excluded from our derived gold alignments, and thus may not be directly reflected in the main evaluation. To estimate the impact of this omission, we manually inspected generated graphs in representative settings and found that such additional pairs were rare, accounting for only 4.62% of the expanded gold set. We also repeated the evaluation after incorporating these pairs and observed the same overall trends as in the main results. Details of this analysis are provided in Appendix C.1.

4.2 Instantiated Benchmark

We use three representative Text2KG pipelines: EDC (Zhang and Soh, 2024), KG-LLM (Carta et al., 2024), and KG-Gen (Mo et al., 2025). Following the original papers, we use `gpt-3.5-turbo-0125` for EDC and KG-LLM, and `gpt-4o-2024-08-06` for KG-Gen. To examine how the underlying LLM affects EA performance, we additionally run EDC and KG-LLM with `gpt-5.1-2025-11-13`. These choices yield five Text2KG configurations: EDC-3.5, EDC-5.1, KG-LLM-3.5, KG-LLM-5.1, and KG-Gen-4o, where numbers indicate base LLM. We use them to instantiate five H-L configurations, five L-L configurations, and nine L1-L2 cross-configuration settings. Implementation details are provided in Appendix B.2.

Using ReDocRED (Tan et al., 2022), BioREDi-

Dataset	Set.	#G.	#$E_{s/t}$	#$T_{s/t}$	#QE.
ReDoc RED-EA	H-H	127	1,493/1,493	3,174/3,174	1,151
	H-L	122	1,438/1,816	3,090/2,408	705
	L-L	116	1,783/1,768	2,480/2,363	522
	L1-L2	115	1,421/2,086	1,882/2,944	433
BioRED irect-EA	H-H	151	1,006/1,006	1,355/1,227	934
	H-L	126	868/1,983	1,186/2,320	368
	L-L	104	1,729/1,678	2,032/1,954	233
	L1-L2	94	1,175/1,871	1,404/2,192	185
REDFM -EA	H-H	191	708/708	507/507	581
	H-L	164	628/1,554	454/1,965	379
	L-L	149	1,487/1,440	1,980/1,835	303
	L1-L2	138	1,112/1,605	1,357/2,195	263

Table 2: Aggregated evaluation statistics for the constructed EA benchmarks. H-L, L-L, and L1-L2 statistics are averaged over their corresponding configurations. #G denotes retained graph-pair instances; #$E_{s/t}$ and #$T_{s/t}$ denote source/target entities and triples; and #QE. denotes evaluated query entities.

rect (Lai et al., 2025; Luo et al., 2022), and REDFM (Huguet Cabot et al., 2023) as base text-graph resources, we construct three EA benchmarks: ReDocRED-EA, BioREDirect-EA, and REDFM-EA. ReDocRED and REDFM provide general-domain text-graph pairs, while BioREDirect provides specialized text-graph pairs. For each base resource, we sample 200 text-graph pairs and apply the construction pipeline described in Section 4.1 to obtain the retained benchmark instances used in our evaluation. Table 2 reports the aggregated statistics of the constructed benchmarks, and Table 6 in Appendix A.2 reports the full configuration-level statistics.

4.3 Text2KG-Induced Graph Discrepancies

We analyze graph-level discrepancies between the source and target graphs introduced by Text2KG pipelines. To quantify these discrepancies, we use normalized Graph Edit Distance (GED) (Sanfeliu and Fu, 1983; Abu-Aisheh et al., 2015) as a diagnostic measure. Intuitively, GED measures the minimum cost of editing one graph into the other through node/edge substitutions, deletions, and insertions; after normalization, larger values indicate larger discrepancies in realized entities, triples, or labels. We use unit costs for the edit operations, with implementation details provided in Appendix B.1.

Figure 3 reports the normalized GED breakdown across H-H and generated-graph settings on ReDocRED-EA; the same analysis for

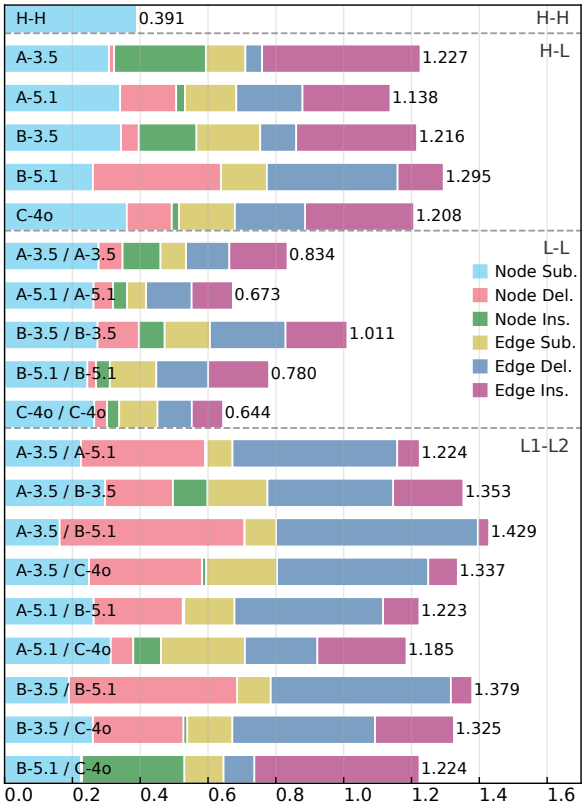

Figure 3: Normalized GED breakdown on ReDocRED-EA. Each horizontal bar corresponds to one Text2KG configuration; the bar length indicates the normalized GED, and stacked segments decompose it into node/edge substitutions, deletions, and insertions. Larger values indicate larger graph discrepancies. T2KG labels use compact notation: A=EDC, B=KG-LLM, C=KG-Gen; numbers indicate base LLM.

BioREDirect-EA and REDFM-EA is provided in Appendix C.2. Across these analyses, all settings involving LLM-generated graphs exhibit higher normalized GED than H-H, and their edit costs are distributed across multiple operation types. Node operations capture entity-label variation and entity coverage differences, while edge operations capture missing/spurious triples and relation-label differences. Thus, Text2KG-induced discrepancies are not limited to a single isolated perturbation, such as noisy names or missing edges; they jointly affect entity realization, lexical labels, graph structure, and relation labels. This motivates evaluating EA methods under the combined discrepancy profiles instantiated by our benchmarks.

5 Experiments

5.1 Experimental Settings

EA methods. Following the method families introduced in Section 2.2, we evaluate seven EA

methods covering closed-set, noise-aware, open-set, and LLM-based paradigms. NameEA is a name-only embedding baseline adapted from Tang et al. (2023). FGWEA (Tang et al., 2023), LightEA (Mao et al., 2022), SUDA (Xu et al., 2025), ERAlign (Yoon et al., 2025), and LLM4EA (Chen et al., 2024) use entity names or textual signals together with graph-structure-based representations. LLM4EA also uses LLM-generated pseudo labels. In contrast, LEA (Long et al., 2025) uses entity names and LLM prior knowledge while textualizing graph structural information as prompt context rather than learning graph-structure-based representations.

Metrics. For each EA setting $s \in \{\text{H-H, H-L, L-L, L1-L2}\}$, we report Hits@1 (H@1), aggregated by weighting each graph pair by its number of query entities: $\text{H@1}^s = \frac{1}{\sum_{i=1}^{I_s} |M_i^s|} \sum_{i=1}^{I_s} [\text{H@1}_i \times |M_i^s|]$, where I_s is the number of evaluated graph pair instances under setting s, M_i^s is the gold alignment set for pair i, and H@1_i is the instance-level Hits@1. Following prior work on name bias in EA benchmarks (Liu et al., 2020; Chakrabarti et al., 2022), we exclude gold pairs whose source and target entity names are exactly identical. All evaluation statistics, including the supplemental-gold analysis in Appendix C.1, are computed after this exact-match filtering. We also report Hits@3 and Mean Reciprocal Rank (MRR) in Appendix C.6.

5.2 Setting and Method-Level Trends

We first examine EA performance at two levels: graph-pair settings and EA method families.

Setting-level trends. Table 3 shows that involving LLM-generated KGs makes EA substantially harder. The average H@1 drops from H-H to H-L and L1-L2 on ReDocRED-EA (91.9 to 76.2/75.5), BioREDirect-EA (79.9 to 57.1/50.9), and REDFM-EA (90.3 to 74.9/72.0). These results show that strong performance on human-curated graph pairs does not directly transfer to Text2KG-based integration settings.

Among generated-graph settings, using a consistent Text2KG pipeline reduces the difficulty. L-L consistently outperforms L1-L2 on all three datasets: 82.9 vs. 75.5 on ReDocRED-EA, 61.2 vs. 50.9 on BioREDirect-EA, and 77.9 vs. 72.0 on REDFM-EA. This pattern is consistent with the GED analysis in Section 4.3, where L1-L2 exhibits larger graph discrepancies than L-L. A practical implication is that, when downstream EA is anticipated, the same Text2KG pipelines should be used whenever possible.

Method-level trends. The results also reveal a consistent difference between methods that rely primarily on semantic signals and those that additionally exploit graph structure. NameEA and FGWEA provide a useful contrast: both use the same entity-name embedding model, but FGWEA additionally incorporates graph-structural signals on top of the name-based matching used by NameEA. While this additional graph evidence is highly effective on H-H, FGWEA is less robust when LLM-generated KGs are involved. LEA remains competitive in several generated-graph settings by combining textual signals with LLM prior knowledge without learning graph-structural representations. At the same time, all methods remain below 80 H@1 on BioREDirect-EA in H-L, L-L, and L1-L2, indicating that robust semantic signals do not by themselves solve Text2KG EA. Therefore, the challenge is not simply to combine semantic and graph signals, but to decide when graph evidence is reliable under generator-induced coverage and structural variation.

5.3 Generator-level trends in H-L Alignment

We next examine whether upgrading the LLM used inside a Text2KG pipeline makes H-L alignment easier. Table 4 reports H-L results on ReDocRED-EA by Text2KG configuration. A stronger LLM does not consistently improve H-L accuracy: the average H@1 slightly decreases from 77.5 to 77.1 for EDC, and drops from 78.9 to 72.4 for KG-LLM. Figure 3 further shows that H-L configurations with stronger LLMs remain much more discrepant than H-H pairs. Thus, improving the LLM inside a Text2KG pipeline does not by itself eliminate the alignment gap between generated and human-curated KGs. These results suggest that H-L EA depends not only on the capability of the underlying LLM, but also on how the pipeline realizes entities, assigns names, and structures triples.

5.4 Entity-Name Discrepancy Trends

We further analyze whether EA performance under generated-graph settings depends on the type of entity-name discrepancy. Table 5 reports H@1 on ReDocRED-EA after grouping evalu-

Dataset	Setting	Closed-set			Open-set	Noise-aware	LLM-based		Ave.
		NameEA	FGWEA	LightEA	SUDA	ERAlign	LEA	LLM4EA	
ReDocRED-EA	H-H	93.3	99.8	98.8	77.2	99.2	85.4	89.5	91.9
	H-L	93.2	83.6	88.5	55.4	69.8	77.2	65.7	76.2
	L-L	93.5	83.1	89.3	66.0	88.1	82.3	78.2	82.9
	L1-L2	92.7	71.7	85.0	59.0	74.0	74.2	71.7	75.5
BioREDirect-EA	H-H	76.5	96.1	87.8	38.9	84.2	83.0	92.9	79.9
	H-L	72.2	47.8	61.8	31.4	50.4	71.1	64.8	57.1
	L-L	72.0	40.3	66.1	34.2	68.5	78.5	69.1	61.2
	L1-L2	71.0	31.9	54.8	26.7	44.4	69.0	58.9	50.9
REDFM-EA	H-H	92.6	98.9	97.3	78.4	85.8	89.0	90.2	90.3
	H-L	88.4	65.1	83.0	61.5	68.1	81.6	76.3	74.9
	L-L	88.4	63.6	84.2	62.1	85.8	85.5	75.7	77.9
	L1-L2	87.1	54.8	80.6	57.6	71.1	80.6	72.0	72.0

Table 3: Main Hits@1 results. Rows are aggregated by EA setting; H-L, L-L, and L1-L2 average over the corresponding Text2KG configurations. "Ave." denotes the average across EA methods.

Setting	T2KG	NameEA	FGWEA	LightEA	SUDA	ERAlign	LEA	LLM4EA	Ave.
H-L	EDC-3.5	94.0	88.5	89.7	44.8	80.4	81.3	63.9	77.5
	EDC-5.1	93.1	81.3	88.0	58.8	66.4	77.7	74.5	77.1
	KG-LLM-3.5	95.2	90.5	92.4	51.8	77.7	78.8	65.8	78.9
	KG-LLM-5.1	91.2	75.9	84.5	63.7	61.0	73.6	56.5	72.4

Table 4: H-L results on ReDocRED-EA by Text2KG method. "Ave." denotes the average Hits@1 across EA methods. Numbers in T2KG labels denote the underlying LLM (gpt-3.5/5.1).

Category	H-H	H-L	L-L	L1-L2
Orthographic	94.9	81.5	88.0	81.1
Acronym	82.4	61.3	65.4	53.4
Rephrasing/reordering	93.5	76.3	82.4	70.9
Name shortening	96.7	84.6	92.6	83.0
Common vs official	88.1	68.5	77.3	65.2
Formal designator	94.2	79.3	86.8	82.6
Parenthetical type	95.5	80.1	86.9	73.8
Cross-lingual variant	90.7	69.3	78.2	71.2
Scope qualifier	89.9	74.1	80.1	74.0

Table 5: Hits@1 by entity-name discrepancy category on ReDocRED-EA. Categories describe the variation between each source entity and its validated H-H counterpart. Generated-graph settings are averaged over the corresponding Text2KG configurations.

ated query pairs by the primary variation between aligned entity names. Appendix C.3 describes the category definitions and assignment procedure, and also reports the corresponding results on BioREDirect-EA and REDFM-EA. Across these analyses, the same qualitative setting-level trends persist: H-L is lower than H-H for almost all categories, and L-L consistently outperforms L1-L2. At the same time, degradation is not uniform across discrepancy categories. Local name differences such as orthographic changes and name shortening remain comparatively easier, whereas acronym and common-vs-official cases are consistently more difficult and show larger drops under generated-graph settings. This indicates that difficult name discrepancies are especially vulnerable in generated-graph settings, where weak lexical cues are compounded by less reliable graph evidence. Complementary analyses in Appendix C.4 and Appendix C.5 show that the generated-graph degradation is not confined to a narrow set of entity types or graph-size regimes.

6 Conclusion

We introduce three EA benchmarks for LLM-generated KGs and evaluate representative EA methods across H-H, H-L, L-L, and L1-L2 settings. The results show that Text2KG-induced discrepancies substantially change EA difficulty: H-H performance does not directly transfer to generated-graph settings, same-pipeline generated KGs are easier to align than cross-pipeline KGs, and graph-derived evidence becomes reliability-sensitive under generated-graph mismatch. Our benchmark provides a reproducible testbed for developing EA methods that combine robust semantic anchors with explicit reliability checks for generated graph evidence.

7 Limitations

Scope and external validity. Our benchmarks focus on text-grounded KGs and therefore do not directly cover large, web-scale cross-KB alignment. This scope reflects the requirements of our benchmark construction: evaluating Text2KG-induced EA discrepancies requires paired resources that provide both input text and a human-curated KG, so that semantic correspondences can be validated before generated graphs are introduced. To the best of our knowledge, no publicly available human-curated text-graph pair resource provides substantially larger graph pairs than the resources used in this work. Within this constraint, we cover three datasets across general and biomedical domains and analyze graph-size sensitivity within the available document-level range. Future progress would benefit from larger text–graph pair resources, especially multi-document or corpus-level human-curated KGs, to evaluate Text2KG EA in larger-scale integration regimes.

One-to-one alignment and granularity-aware extensions. Our benchmark follows the standard pairwise EA formulation and evaluates one-to-one correspondences in M. This design keeps the human-validated H-H alignments unambiguous, supports direct comparison with existing EA methods, and provides a controlled basis for analyzing Text2KG-induced mismatch. However, Text2KG pipelines can also induce granularity shifts and split/merge phenomena, where one entity in one graph corresponds to multiple entities in the other graph, or vice versa.

Future benchmark versions can extend our text-grounded graph-pair instances with split/merge annotations for one-to-many and many-to-one EA. Such extensions would enable direct evaluation of Text2KG-induced entity fragmentation and merging, and would test whether our findings on pipeline consistency and reliability-aware graph evidence also hold under granularity-aware alignment.

Scalability of GED. We compute GED approximately with a fixed time budget (3,000s) and uniform edit costs. In our document-level regime, we empirically verify that the best-so-far GED stabilizes as the timeout increases (Appendix B.1), suggesting that the approximation is sufficiently stable for our discrepancy analysis. However, GED remains computationally expensive and does not readily scale to much larger KGs. Therefore, our GED-based diagnostics are currently limited to document-level graph pairs; developing scalable discrepancy measures (or tighter GED approximations) for large-scale KG integration is an important direction for future work.

8 Ethical Considerations

Data licensing and redistribution. Our benchmark is derived from three publicly available resources: ReDocRED (Tan et al., 2022), REDFM (Huguet Cabot et al., 2023), and BioREDirect (Lai et al., 2025; Luo et al., 2022). We release ReDocRED-EA under the MIT License, following the license of ReDocRED and retaining the original license notice. We release REDFM-EA under CC BY-SA 4.0, following the license of REDFM and its ShareAlike condition. For BioREDirect, the public repository does not specify a formal license. We contacted an author involved in both BioRED and BioREDirect, who indicated that BioRED is released for public use and that they did not see any issue with releasing our derived BioREDirect-EA resource for public use. Accordingly, our BioREDirect-EA release does not redistribute the original BioREDirect data; it contains only benchmark-specific files created by us, such as LLM-generated graph annotations and alignment files. We provide scripts to reconstruct BioREDirect-EA after users obtain the original BioREDirect data from the official source.

Ethical review. Our work involves manual verification of entity correspondences in graphs derived from publicly available, non-sensitive texts. This verification was conducted by the paper authors and external annotators. The annotation task is minimal risk (standard data curation and quality control) and does not involve intervention with human subjects; therefore we did not seek formal ethics-board approval.

LLM output verification. We mitigate hallucination risk through manual verification because our pipeline uses an LLM to propose entity paraphrases. The paper authors and external annotators inspect the LLM outputs and correct instances that violate semantic equivalence or appropriateness requirements, ensuring that the released gold alignments are human-validated.

Potential risks and misuse. Our benchmarks and analyses can support systems that integrate

LLM-generated KGs. A key risk is incorrect integration: alignment errors can propagate to downstream applications and lead to misleading outputs or decisions. This risk is amplified when KGs are LLM-generated due to coverage shifts, inconsistent naming, and hallucinated entities or triples. Our results suggest mitigation strategies, including consistent Text2KG pipelines, semantic anchoring, and reliability-aware use of graph evidence, but human review and application-specific safeguards remain necessary in high-stakes settings.

Annotator welfare and consent. External annotation was carried out by three paid annotators recruited and managed by a third-party annotation provider. We did not use public crowdsourcing platforms. All external annotation work was conducted within legally permitted working hours and compensated at or above the applicable minimum wage. External annotators were informed that the resulting annotations would be released as part of a research dataset, and consent for public release of the annotation outputs was obtained. The paper authors also participated in the verification process as described in Appendix A.1.

References

Zeina Abu-Aisheh, Romain Raveaux, Jean-Yves Ramel, and Patrick Martineau. 2015. An exact graph edit distance algorithm for solving pattern recognition problems. In *Proceedings of the International Conference on Pattern Recognition Applications and Methods - Volume 1*, ICPRAM 2015, pages 271–278, Setubal, PRT. SCITEPRESS - Science and Technology Publications, Lda.

Zhen Bi, Jing Chen, Yinuo Jiang, Feiyu Xiong, Wei Guo, Huajun Chen, and Ningyu Zhang. 2024. CodeKGC: Code language model for generative knowledge graph construction. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(3):1–16.

Salvatore Carta, Alessandro Giuliani, Marco Manolo Manca, Leonardo Piano, Alessandro Sebastian Podda, Livio Pompianu, and Sandro Gabriele Tiddia. 2024. A zero-shot strategy for knowledge graph engineering using GPT-3.5. *Procedia Computer Science*, 246:2235–2243.

Soumen Chakrabarti, Harkanwar Singh, Shubham Lohiya, Prachi Jain, and Mausam . 2022. Joint completion and alignment of multilingual knowledge graphs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 11922–11938, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo. 2017. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 1511–1517.

Shengyuan Chen, Qinggang Zhang, Junnan Dong, Wen Hua, Qing Li, and Xiao Huang. 2024. Entity alignment with noisy annotations from large language models. *Advances in Neural Information Processing Systems*, 37:15097–15120.

Jingwei Cheng, Chenglong Lu, Linyan Yang, Guoqing Chen, and Fu Zhang. 2025. EasyEA: Large language model is all you need in entity alignment between knowledge graphs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20981–20995, Vienna, Austria. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Lingbing Guo, Zequn Sun, and Wei Hu. 2019. Learning to exploit long-term relational dependencies in knowledge graphs. In *International conference on machine learning*, pages 2505–2514. PMLR.

Pere-Lluís Huguet Cabot, Simone Tedeschi, Axel-Cyrille Ngonga Ngomo, and Roberto Navigli. 2023. RED^{fm}: a filtered and multilingual relation extraction dataset. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4326–4343, Toronto, Canada. Association for Computational Linguistics.

Xuhui Jiang, Yinghan Shen, Zhichao Shi, Chengjin Xu, Wei Li, Zixuan Li, Jian Guo, Huawei Shen, and Yuanzhuo Wang. 2024. Unlocking the power of large language models for entity alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7566–7583, Bangkok, Thailand. Association for Computational Linguistics.

Xiongnan Jin, Zhilin Wang, Jinpeng Chen, Liu Yang, Byungkook Oh, Seung-won Hwang, and Jianqiang Li. 2025. HLMEA: Unsupervised entity alignment based on hybrid language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11888–11896.

Morteza Kamaladdini Ezzabady and Farah Benamara. 2025. Entity quality enhancement in knowledge graphs through LLM-based question answering. In *Proceedings of the Workshop on Generative AI and*

Knowledge Graphs (GenAIK), pages 136–145, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Po-Ting Lai, Chih-Hsuan Wei, Shubo Tian, Robert Leaman, and Zhiyong Lu. 2025. Enhancing biomedical relation extraction with directionality. *Bioinformatics*, 41(Supplement_1):i68–i76.

Yassir Lairgi, Ludovic Moncla, Rémy Cazabet, Khalid Benabdeslem, and Pierre Cléau. 2024. iText2KG: Incremental knowledge graphs construction using large language models. In *International Conference on Web Information Systems Engineering*, pages 214–229. Springer.

Xukai Liu, Kai Zhang, Ye Liu, Enhong Chen, Zhenya Huang, Linan Yue, and Jiaxian Yan. 2023. RHGN: Relation-gated heterogeneous graph network for entity alignment in knowledge graphs. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8683–8696, Toronto, Canada. Association for Computational Linguistics.

Zhiyuan Liu, Yixin Cao, Liangming Pan, Juanzi Li, Zhiyuan Liu, and Tat-Seng Chua. 2020. Exploring and evaluating attributes, values, and structures for entity alignment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6355–6364, Online. Association for Computational Linguistics.

Meixiu Long, Jiahai Wang, Junxiao Ma, Jianpeng Zhou, and Siyuan Chen. 2025. LLM-augmented entity alignment: an unsupervised and training-free framework. *Neural Networks*, 194:108139.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. 2022. BioRED: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282.

Shengxuan Luo and Sheng Yu. 2022. An accurate unsupervised method for joint entity alignment and dangling entity detection. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2330–2339, Dublin, Ireland. Association for Computational Linguistics.

Chuangtao Ma, Yongrui Chen, Tianxing Wu, Arijit Khan, and Haofen Wang. 2025. Large language models meet knowledge graphs for question answering: Synthesis and opportunities. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24578–24597, Suzhou, China. Association for Computational Linguistics.

Xin Mao, Wenting Wang, Yuanbin Wu, and Man Lan. 2021a. Boosting the speed of entity alignment 10×: Dual attention matching network with normalized hard sample mining. In *Proceedings of the web conference 2021*, pages 821–832.

Xinnian Mao, Wenting Wang, Yuanbin Wu, and Man Lan. 2021b. From alignment to assignment: Frustratingly simple unsupervised entity alignment. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2843–2853.

Xinnian Mao, Wenting Wang, Yuanbin Wu, and Man Lan. 2022. LightEA: A scalable, robust, and interpretable entity alignment framework via three-view label propagation. In *Proceedings of the 2022 conference on Empirical Methods in Natural Language Processing*, pages 825–838.

Nandana Mihindukulasooriya, Sanju Tiwari, Carlos F Enguix, and Kusum Lata. 2023. Text2KGBench: A benchmark for ontology-driven knowledge graph generation from text. In *International semantic web conference*, pages 247–265. Springer.

Congmin Min, Rhea Mathew, Joyce Pan, Sahil Bansal, Abbas Keshavarzi, and Amar Viswanathan Kannan. 2025. Efficient knowledge graph construction and retrieval from unstructured text for large-scale RAG systems. *Preprint*, arXiv:2507.03226.

Belinda Mo, Kyssen Yu, Joshua Kazdan, Joan Cabezas, Proud Mpala, Lisa Yu, Chris Cundy, Charilaos Kanatsoulis, and Sanmi Koyejo. 2025. KGGen: Extracting knowledge graphs from plain text with language models. *Preprint*, arXiv:2502.09956.

Alberto Sanfeliu and King-Sun Fu. 1983. A distance measure between attributed relational graphs for pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(3):353–362.

Zequn Sun, Muhao Chen, and Wei Hu. 2021. Knowing the no-match: Entity alignment with dangling cases. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3582–3593, Online. Association for Computational Linguistics.

Zequn Sun, Wei Hu, and Chengkai Li. 2017. Cross-lingual entity alignment via joint attribute-preserving embedding. In *International semantic web conference*, pages 628–644. Springer.

Zequn Sun, Wei Hu, Qingheng Zhang, and Yuzhong Qu. 2018. Bootstrapping entity alignment with knowledge graph embedding. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, pages 4396–4402. AAAI Press.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022. Revisiting DocRED - addressing the false negative problem in relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8472–8487, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jianheng Tang, Kangfei Zhao, and Jia Li. 2023. A Fused Gromov-Wasserstein framework for unsupervised knowledge graph entity alignment. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3320–3334, Toronto, Canada. Association for Computational Linguistics.

Vinh Van Tong, Thanh Trung Huynh, Thanh Tam Nguyen, Hongzhi Yin, Quoc Viet Hung Nguyen, and Quyet Thang Huynh. 2022. Incomplete knowledge graph alignment. *Preprint*, arXiv:2112.09266.

Zhichun Wang, Qingsong Lv, Xiaohan Lan, and Yu Zhang. 2018. Cross-lingual knowledge graph alignment via graph convolutional networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 349–357, Brussels, Belgium. Association for Computational Linguistics.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2024. ChatIE: Zero-shot information extraction via chatting with ChatGPT. *Preprint*, arXiv:2302.10205.

Cong Xu, Mengxin Shi, Xiang Gao, Zhongkang Yin, Xiujuan Yao, Wei Li, and Jiasen Yang. 2025. A high-accuracy unsupervised statistical learning method for joint dangling entity detection and entity alignment. *Information Systems*, 133:102554.

Zhentao Xu, Mark Jerome Cruz, Matthew Guevara, Tie Wang, Manasi Deshpande, Xiaofeng Wang, and Zheng Li. 2024. Retrieval-augmented generation with knowledge graphs for customer service question answering. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, SIGIR '24, pages 2905–2909. Association for Computing Machinery.

Rui Yang, Boming Yang, Xinjie Zhao, Fan Gao, Aosong Feng, Sixun Ouyang, Moritz Blum, Tianwei She, Yuang Jiang, Freddy Lecue, and 1 others. 2025. Graphusion: A RAG framework for scientific knowledge graph construction with a global perspective. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2579–2588. Association for Computing Machinery.

Hang Yin, Liyao Xiang, Dong Ding, Yuheng He, Yihan Wu, Pengzhi Chu, Xinbing Wang, and Chenghu Zhou. 2024. Lambda: Learning matchable prior for entity alignment with unlabeled dangling cases. In *Advances in Neural Information Processing Systems*, volume 37, pages 78964–78995. Curran Associates, Inc.

Soojin Yoon, Sungho Ko, Tongyoung Kim, SeongKu Kang, Jinyoung Yeo, and Dongha Lee. 2025. Unsupervised robust cross-lingual entity alignment via neighbor triple matching with entity and relation texts. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pages 184–193. Association for Computing Machinery.

Bowen Zhang and Harold Soh. 2024. Extract, Define, Canonicalize: An LLM-based framework for knowledge graph construction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9820–9836, Miami, Florida, USA. Association for Computational Linguistics.

Yuhong Zhang, Xiaolong Zhu, and Xuegang Hu. 2024. Entity alignment in noisy knowledge graph. *Applied Intelligence*, 55(3):210.

Appendix

A Details of Dataset Construction

A.1 Gold Alignment Verification

Because Step 2 uses an LLM to propose entity rewritings, we verify the candidate H-H correspondences before finalizing them as gold. Let $\tilde{M}^{\text{H-H}} = \{(e, \tilde{\phi}(e)) \mid e \in E^{\text{H}}_{\text{src}}\}$ denote the candidate H-H alignment proposed by the LLM in Step 2. The goal of this verification is to finalize the human-validated H-H alignment $M^{\text{H-H}}$, which is then used as the gold alignment for H-H and as the basis for the Step 5 projection.

Automatic check. We first apply automatic checks to detect structurally invalid candidate mappings. The *one-to-one-alignment* check requires that each candidate target entity appear at most once, and the *consistency* check requires that the candidate mapping contain no contradictory correspondences. ReDocRED-EA contained 32 instances that violated the *one-to-one-alignment* check and 50 instances that violated the *consistency* check. BioREDirect-EA contained 24 and 19 such instances, respectively, and REDFM-EA contained 2 and 4, respectively. Detected graph-pair instances were excluded before finalizing the H-H gold alignments.

Manual verification. After the automatic checks, we manually verify the remaining candidate correspondences in two stages. First, the authors identify *paraphrase-induced LLM generation errors*, where the entity rewriting itself introduces an incorrect or unusable entity. Second, annotators, including the authors, check whether each entity pair refers to the same real-world entity or concept (*semantic validity*) and whether

the rewritten entity name is safe and appropriate (*safe entity names*).

The authors identified paraphrase-induced LLM generation errors in 6 ReDocRED-EA instances, 20 BioREDirect-EA instances, and 3 REDFM-EA instances; these instances were removed. Annotators also identified semantic-validity violations at the entity-pair level: 24 pairs in ReDocRED-EA, 42 pairs in BioREDirect-EA, and 12 pairs in REDFM-EA. These entity-pair-level violations were corrected before finalizing the gold alignments. No unsafe entity names were found in any dataset.

Finalized H-H gold alignment. After these automatic and manual verification steps, the retained and corrected alignment is denoted by $M^{\text{H-H}}$. Importantly, this verification step concerns the H-H semantic correspondences produced in Step 2; it is separate from the configuration-specific projection used later to derive H-L, L-L, and L1-L2 evaluation instances.

Annotation instructions. We provide the anonymized annotation instructions used for the second manual verification stage below. The full version will be released upon publication. The instructions focus on semantic validity and safety judgments for candidate H-H correspondences.

(Anonymized) Annotation Instructions

```
For each candidate mapping in the dataset,
     verify the following requirements:
REQ-1: The entity pair refers to the same
     real-world entity or concept.
REQ-2: The entity does not contain content
     that violates public norms (e.g., hate,
     pornography, harassment, or violent
     threats).

If REQ-1 is not satisfied, further label the
     case as either
(i) Not equivalent, or (ii) Cannot determine
     due to insufficient information.
For REQ-2, choose one of two labels: Safe or
     Unsafe.

You can use web search (including AI-
     generated summaries) and consult external
      resources (e.g., Wikipedia) to gather
     evidence, *but the final decision must be
      made by a human annotator*.

The examples below are not exhaustive; they
     are provided only to illustrate the
     intended judgments.
--------------------
EXAMPLES FOR REQ-1 (TYPICALLY EQUIVALENT)
- Abbreviation vs. full name:
  [ORG_ABBR] <-> [ORG_FULL_NAME]
- Original-language name vs. translated name:
  [INSTITUTION_LANG1] <-> [INSTITUTION_LANG2]
- Spelling variants:
  [COMPANY_NAME_VARIANT_1] <-> [
      COMPANY_NAME_VARIANT_2]
- Alternative titles of the same work:
  [WORK_TITLE_1] <-> [WORK_TITLE_2]
- Nicknames / name changes (e.g., marriage):
  [PERSON_LEGAL_NAME] <-> [
      PERSON_ALIAS_OR_NEW_NAME]

EXAMPLES FOR REQ-1 (TYPICALLY *NOT*
     EQUIVALENT)
- Hypernym / hyponym:
  [ORG_GROUP] <-> [ORG_SUBUNIT]
- Different geographic scales:
  [REGION_LEVEL_1] (e.g., state) <-> [
      REGION_LEVEL_2] (e.g., city)
- Confusion between a creator and a work:
  [DIRECTOR] <-> [WORK_NAME]
- Hallucinations:
  Non-existent people / organizations / works

EXAMPLES FOR REQ-2 (UNSAFE CONTENT)
- Content that can be interpreted as
     discrimination, pornography, abusive
     language, or violent statements.
--------------------
```

A.2 Finalized gold for H-L / L-L / L1-L2

Table 6 reports the evaluation statistics for all generated-graph configurations used in our experiments. Because Step 5 derives configuration-specific gold alignments by projecting the validated H-H alignment onto entities realized in generated graphs, the number of evaluated graph pairs and query entities can differ across Text2KG configurations.

B Detailed Experimental Settings

B.1 Graph Edit Distance Implementation

To compute graph edit distance (GED), we use NetworkX's `optimize_edit_paths` (v3.4.2). We represent each KG as a directed labeled graph: nodes correspond to entities with entity surface forms as node labels, and edges correspond to relational triples with relation names as edge labels.

We consider six edit-operation types: $\mathcal{O} = \{o_{\text{ns}}, o_{\text{nd}}, o_{\text{ni}}, o_{\text{es}}, o_{\text{ed}}, o_{\text{ei}}\}$, corresponding to node substitution, node deletion, node insertion, edge substitution, edge deletion, and edge insertion. We use unit costs for insertions and deletions. For substitutions, the cost is 0 when the corresponding labels match and 1 otherwise. Given an edit path $\pi = (o_1, \ldots, o_N)$ that transforms G_{tgt} into G_{src}, its cost is $C(\pi) = \sum_{i=1}^{N} c(o_i)$. The raw GED

Dataset	Setting	T2KG	#G.	#E$_s$.	#E$_t$.	#T$_s$.	#T$_t$.	#QEnt.
ReDocRED-EA	H-H	-	127	1,493	1,493	3,174	3,174	1,151
	H-L	**EDC**-3.5	119	1,399	865	3,045	808	446
		EDC-5.1	127	1,493	2,005	3,187	2,072	792
		KG-LLM-3.5	113	1,315	1,067	2,848	1,328	562
		KG-LLM-5.1	126	1,491	3,280	3,185	6,045	848
		KG-Gen-4o	127	1,493	1,862	3,187	1,787	876
	L-L	**EDC**-3.5 / **EDC**-3.5	110	920	823	1,004	781	309
		EDC-5.1 / **EDC**-5.1	123	1,923	1,966	2,012	2,040	576
		KG-LLM-3.5 / **KG-LLM**-3.5	98	871	942	1,116	1,198	333
		KG-LLM-5.1 / **KG-LLM**-5.1	125	3,372	3,259	6,531	6,024	681
		KG-Gen-4o / **KG-Gen**-4o	126	1,831	1,848	1,739	1,774	712
	L1-L2	**EDC**-3.5 / **EDC**-5.1	119	951	1,910	1,030	1,986	369
		EDC-3.5 / **KG-LLM**-3.5	100	820	967	841	1,220	254
		EDC-3.5 / **KG-LLM**-5.1	117	949	3,062	1,029	5,679	372
		EDC-5.1 / **KG-LLM**-5.1	122	1,921	3,203	2,011	5,956	556
		EDC-3.5 / **KG-Gen**-4o	118	951	1,740	1,031	1,679	356
		EDC-5.1 / **KG-Gen**-4o	124	1,929	1,808	2,015	1,739	564
		KG-LLM-3.5 / **KG-LLM**-5.1	105	922	2,706	1,169	4,990	385
		KG-LLM-3.5 / **KG-Gen**-4o	105	936	1,534	1,192	1,472	390
		KG-LLM-5.1 / **KG-Gen**-4o	126	3,412	1,848	6,620	1,774	651
BioREDirect-EA	H-H	-	151	1,006	1,006	1,355	1,227	934
	H-L	**EDC**-3.5	132	886	1,192	1,213	1,246	348
		EDC-5.1	145	990	1,943	1,339	1,857	462
		KG-LLM-3.5	78	561	516	782	454	158
		KG-LLM-5.1	129	910	3,313	1,254	5,437	304
		KG-Gen-4o	147	991	2,951	1,341	2,606	569
	L-L	**EDC**-3.5 / **EDC**-3.5	113	1,060	1,042	1,138	1,119	222
		EDC-5.1 / **EDC**-5.1	133	1,752	1,799	1,632	1,740	344
		KG-LLM-3.5 / **KG-LLM**-3.5	38	279	263	260	234	42
		KG-LLM-5.1 / **KG-LLM**-5.1	102	2,811	2,627	4,696	4,323	159
		KG-Gen-4o / **KG-Gen**-4o	132	2,744	2,659	2,436	2,352	396
	L1-L2	**EDC**-3.5 / **EDC**-5.1	121	1,121	1,579	1,199	1,521	232
		EDC-3.5 / **KG-LLM**-3.5	50	502	320	552	286	78
		EDC-3.5 / **KG-LLM**-5.1	99	953	2,470	1,044	3,903	173
		EDC-5.1 / **KG-LLM**-5.1	106	1,385	2,725	1,293	4,487	191
		EDC-3.5 / **KG-Gen**-4o	122	1,155	2,436	1,238	2,170	271
		EDC-5.1 / **KG-Gen**-4o	128	1,647	2,584	1,534	2,296	330
		KG-LLM-3.5 / **KG-LLM**-5.1	51	396	1,301	375	2,032	62
		KG-LLM-3.5 / **KG-Gen**-4o	58	424	1,210	394	1,061	88
		KG-LLM-5.1 / **KG-Gen**-4o	109	2,993	2,213	5,005	1,969	236
REDFM-EA	H-H	-	191	708	708	507	507	581
	H-L	**EDC**-3.5	156	603	759	440	564	316
		EDC-5.1	178	680	1,482	491	1,255	424
		KG-LLM-3.5	124	493	826	363	1,025	273
		KG-LLM-5.1	180	684	2,828	494	5,250	434
		KG-Gen-4o	182	680	1,874	481	1,729	449
	L-L	**EDC**-3.5 / **EDC**-3.5	133	686	665	515	506	228
		EDC-5.1 / **EDC**-5.1	170	1,506	1,409	1,301	1,197	379
		KG-LLM-3.5 / **KG-LLM**-3.5	99	719	671	924	862	175
		KG-LLM-5.1 / **KG-LLM**-5.1	173	2,776	2,703	5,553	4,995	356
		KG-Gen-4o / **KG-Gen**-4o	172	1,746	1,752	1,606	1,613	376
	L1-L2	**EDC**-3.5 / **EDC**-5.1	137	700	1,128	527	967	236
		EDC-3.5 / **KG-LLM**-3.5	93	464	638	345	761	168
		EDC-3.5 / **KG-LLM**-5.1	135	708	2,152	532	4,034	232
		EDC-5.1 / **KG-LLM**-5.1	165	1,511	2,645	1,306	4,958	348
		EDC-3.5 / **KG-Gen**-4o	133	686	1,378	514	1,274	240
		EDC-5.1 / **KG-Gen**-4o	164	1,453	1,680	1,260	1,551	354
		KG-LLM-3.5 / **KG-LLM**-5.1	121	881	1,853	1,142	3,474	217
		KG-LLM-3.5 / **KG-Gen**-4o	124	890	1,254	1,137	1,154	233
		KG-LLM-5.1 / **KG-Gen**-4o	168	2,717	1,715	5,454	1,585	338

Table 6: Statistics for ReDocRED-EA, BioREDirect-EA, and REDFM-EA. Each row corresponds to one EA setting and Text2KG configuration. #G. is the number of evaluated graph pairs; #E$_s$/#E$_t$ and #T$_s$/#T$_t$ are source/target entity and triple totals; #QEnt. is the number of evaluated query entities.

under this cost function is $\mathrm{GED}(G_{\mathrm{src}}, G_{\mathrm{tgt}}) = \min_{\pi \in \Pi(G_{\mathrm{tgt}} \to G_{\mathrm{src}})} C(\pi)$, where $\Pi(G_{\mathrm{tgt}} \to G_{\mathrm{src}})$ denotes the set of edit paths that transform G_{tgt} into a graph matching G_{src}.

Because exact GED is computationally expensive, `optimize_edit_paths` returns an iterator over candidate edit paths. For each graph pair, we run the iterator under a wall-clock timeout of 3,000 seconds and take the minimum-cost path observed within the time limit as our approximate edit path $\hat{\pi}$. We report $\widehat{\mathrm{GED}}(G_{\mathrm{src}}, G_{\mathrm{tgt}}) = C(\hat{\pi})$. Appendix C.2 verifies that the resulting approximate GED is stable under longer time budgets.

For comparability across graph pairs of different sizes, we report normalized GED. Let $C_\tau(\hat{\pi})$ be the total cost contributed by operation type τ in $\hat{\pi}$. We normalize node-operation costs by the total number of source and target entities, and edge-operation costs by the total number of source and target triples:

$$\mathrm{nGED}(G_{\mathrm{src}}, G_{\mathrm{tgt}}) = \sum_{\tau \in \{\mathrm{ns,nd,ni}\}} \frac{C_\tau(\hat{\pi})}{|E_{\mathrm{src}}| + |E_{\mathrm{tgt}}|} + \sum_{\tau \in \{\mathrm{es,ed,ei}\}} \frac{C_\tau(\hat{\pi})}{|T_{\mathrm{src}}| + |T_{\mathrm{tgt}}|}.$$

The six stacked components in the GED figures 3, 4 correspond to these six normalized operation costs. The edit-type breakdown is used as an interpretable diagnostic. Node operations capture entity-label variation and entity-coverage differences, while edge operations capture missing or spurious triples and relation-label differences. For example, when transforming G_{tgt} into G_{src}, edge deletions indicate triples present in G_{tgt} but absent from G_{src}, whereas edge insertions indicate triples missing from G_{tgt} but present in G_{src}.

B.2 Implementation Details

We summarize the implementations and hyperparameter settings used in our experiments for both Text2KG and EA. We primarily follow the authors' public code releases and keep the default configurations whenever available, with minor adjustments required for our document-level benchmark setting and for API/model availability. All non-default choices are documented below to support reproducibility. We run each method once with a fixed random seed (or deterministically). All experiments are conducted on a Linux server equipped with an Intel Xeon E5-1683 v4 CPU and an NVIDIA Tesla V100 (PCIe, 32 GB) GPU.

B.2.1 Text2KG methods

We instantiate LLM-generated graphs using three Text2KG pipelines: EDC (Zhang and Soh, 2024), KG-LLM (Carta et al., 2024), and KG-Gen (Mo et al., 2025).

EDC We use the authors' implementation[2] with two LLMs: `gpt-3.5-turbo-0125` (default) and `gpt-5.1-2025-11-13` (stronger variant). For each run, the same LLM is used across EDC's three stages: OpenIE, schema definition, and schema canonicalization. We keep the sentence-embedding model fixed to `intfloat/e5-mistral-7b-instruct` and run schema canonicalization in `Target Alignment` mode using the 32 REDFM relation definitions (Huguet Cabot et al., 2023) as the target schema.

KG-LLM We use the authors' implementation[3]. Because the OpenIE model used in the original paper, `gpt-3.5-turbo-0301`, is no longer available, we replace it with `gpt-3.5-turbo-0125`. We additionally run an ablation using `gpt-5.1-2025-11-13`.

KG-Gen We use the authors' implementation[4] with the same LLM as in the original paper, `gpt-4o-2024-08-06`. We enable clustering (`clustering=True`) as part of the default generation workflow.

B.2.2 EA methods

We evaluate seven EA methods: ERAlign (Yoon et al., 2025), NameEA, FGWEA (Tang et al., 2023), LightEA (Mao et al., 2022), SUDA (Xu et al., 2025), LEA (Long et al., 2025), and LLM4EA (Chen et al., 2024). Unless explicitly stated otherwise, we use each method's recommended hyperparameters for evaluation.

Candidate space. To ensure a fair evaluation under our partial-query setting, we define prediction candidates as all entities in the corresponding target graph. Several public EA implementations construct the prediction candidate space from the file containing query entity pairs. This assumption is harmless when all graph entities are included

[2] https://github.com/clear-nus/edc/tree/main, commit hash: 02aad17

[3] https://github.com/SandroGT/KG-LLM-Prompting/tree/main, commit hash: d613a73

[4] https://github.com/stair-lab/kg-gen, commit hash: 63726f4

as evaluated entities, because the query-pair file then covers the full graph entity sets. However, it is inappropriate for our benchmark. In our benchmark, exact-match pairs are excluded from evaluation, Step 5 derives configuration-specific query sets, and generated graphs may contain entities that are not gold-aligned query entities. Using the query-pair file to define prediction candidates would therefore remove target-side distractors and artificially simplify the EA task.

For NameEA, FGWEA, LightEA, LEA, and LLM4EA, we modify the data-loading or evaluation interface so that the alignment score matrix is constructed over all entities in the corresponding source and target graphs. ERAlign and SUDA already construct prediction candidates from all entities in the target graph in their evaluation workflow, so no analogous candidate-space modification was needed for these methods. The query-pair file is used only to specify which source entities are evaluated and what their gold target matches are. This change affects only the candidate space used for prediction and evaluation; the model scoring functions and method-specific hyperparameters are otherwise kept consistent with the public implementations.

NameEA We implement NameEA as an embedding-only baseline following the name-based baseline used in FGWEA (Tang et al., 2023). It disables graph-structure and relation signals and ranks target candidates by cosine similarity between L2-normalized entity-name embeddings from `sentence-transformers/LaBSE` (Feng et al., 2022). Because NameEA uses only entity names, it serves as a lexical/semantic-name baseline for assessing how much of the benchmark can be solved without graph-structural information.

FGWEA We use the authors' implementation[5] and follow the default configuration unless noted otherwise. Since our benchmarks do not provide auxiliary attributes beyond entity and relation names, we enable name, relation and structure signals while disabling attribute features (i.e., `-use_name 1`, `-use_rel 1`, and `-use_stru 1`). We set `gw_iter=0` to disable the Gromov–Wasserstein refinement stage, which did not run reliably on some small document-level graph pairs.

[5] https://github.com/squareRoot3/FusedGW-Entity-Alignment/tree/master, commit hash: 2af3734

LightEA We use the authors' implementation[6] in the unsupervised LightEA-L configuration. LightEA employs Sparse Sinkhorn Iteration with a top-K neighborhood candidate parameter (K, default $K = 500$) for efficient matching. In our document-level KGs, many instances contain extremely sparse nodes (often with degree 1), and the default $K = 500$ was not compatible with these small graphs, causing the implementation to fail on some instances. Therefore, we set the neighborhood candidate size to $K = 1$ to ensure that LightEA runs reliably on all graph pairs.

ERAlign We use the authors' implementation[7]. All options follow the default configuration provided in the released codebase.

SUDA We use the authors' implementation[8] with the default open-set EA pipeline. Since the default dangling-candidate size $K = 27$ caused unstable execution on some document-level graphs, we set $K = 1$ for all experiments. SUDA first performs dangling entity detection and then ranks candidates only for entities predicted as matchable. In the authors' original evaluation code, ranking metrics are computed only over these predicted-matchable entities. In our benchmark, however, every evaluated query entity has a gold counterpart after Step 5 projection and exact-match filtering. Therefore, if SUDA predicts an evaluated query entity as dangling, we count it as an alignment error: H@1 and H@3 are set to 0, and MRR is computed with infinite rank. This affects only evaluation accounting; SUDA's model scores and hyperparameters are otherwise unchanged.

LEA We use the authors' implementation[9] with `gpt-4o-mini-2024-07-18`, matching the original paper. Because our benchmarks do not provide auxiliary attributes, we run LEA under the `name+rel` information setting.

LLM4EA We use the authors' implementation[10] with `gpt-3.5-turbo-0125`, matching the

[6] https://github.com/MaoXinn/LightEA/tree/main, commit hash: dd8e2a3

[7] https://github.com/SoojinY/ERAlign/tree/main, commit hash: 2839f1a

[8] https://github.com/skyccong/SUDA/tree/main, commit hash: cbf6480

[9] https://github.com/Longmeix/LEA/tree/main, commit hash: c511638

[10] https://github.com/chensyCN/llm4ea_official/tree/master, commit hash: ab7dbc8

Setting	T2KG	#QE.	#Add.	#Expand.	Add. Share
H-L	A-3.5	446	39	485	8.04%
	A-5.1	792	21	813	2.58%
Ave.	-	619	30	649	4.62%

Table 7: Manual inspection of additional alignable pairs on ReDocRED-EA H-L setting after exact-match pair removal. #QE. denotes the number of query entities constructed by Step 5. #Add. denotes manually identified additional alignable pairs involving entities newly introduced by Text2KG. #Expand. denotes the expanded gold set obtained by adding these pairs to the Step 5 gold under the final evaluation protocol. "Add. Share" is the percentage of Add. in Exp.

original paper. The default LLM-query budget (`-budget 0.1`) can select no query nodes on small graph pairs, so we set `-budget 1` to ensure valid execution on every instance.

C Additional Analyses and Results

C.1 Analysis of Additional Alignable Pairs

Counting protocol. We manually inspect representative generated-graph settings to estimate whether additional alignable pairs omitted by the Step 5 filtering strategy in Section 4.1 affect the main results. Specifically, we analyze two ReDocRED-EA H-L configurations, A-3.5 and A-5.1, and inspect the generated target graphs to identify alignable pairs that are not included in the validated H-H gold $M^{\text{H-H}}$. These include generated aliases or canonicalized entities that are semantically alignable to source-side entities but are not realized with the exact H-H target string.

Consistent with the main evaluation, we remove exact-match source-target pairs from the evaluated query set before counting additional alignable pairs. Table 7 summarizes the manual inspection results. Averaged over A-3.5 and A-5.1, the additional alignable pairs account for only 4.62% of the expanded gold set. Thus, the Step 5 gold still provides a substantial evaluation set in these representative settings, and the manually identified additional pairs form only a small fraction of the expanded gold.

We further evaluate whether incorporating these additional alignable pairs changes the empirical conclusions. Table 8 reports H@1 under the expanded gold set, and the evaluation preserves the same overall trends as the main results. In particular, adding these pairs does not make the generated-graph setting substantially easier, nor does it reverse the relative trends among methods. These results suggest that the main conclusions are not primarily driven by the omission of additional alignable pairs outside $M^{\text{H-H}}$.

C.2 GED Diagnostics

Additional Results. Described in Section 4.3, we use normalized GED as a compact diagnostic of graph-level discrepancy. Figure 4 shows the GED breakdowns on BioREDirect-EA and REDFM-EA, respectively. Across datasets, H-H has the smallest graph discrepancy, whereas settings involving LLM-generated graphs introduce larger edit costs. This indicates that the difficulty of aligning LLM-generated KGs is not merely a matter of renamed entities: generated graphs also differ in which entities and triples are extracted. In particular, insertion and deletion components reflect coverage and granularity mismatches between the compared graphs, which are not captured by surface-form variation alone.

Timeout sensitivity. In our setting, the best-so-far approximate GED produced by `optimize_edit_paths` stabilizes quickly as the timeout increases. To verify robustness, we additionally record the best GED value attained at intermediate checkpoints (60, 300, 600, ..., 2,700, 3,000 seconds) for every graph pair and aggregate results by setting. Fig. 5 plots time budget versus approximate GED. Across all settings, the averaged best-so-far GED is nearly flat over time: longer timeouts yield only negligible improvements. More importantly, the total improvement from 60 to 3,000 seconds is much smaller than the GED gaps observed across EA settings (e.g., H-L vs. L-L or L-L vs. L1-L2). Therefore, our qualitative conclusions based on GED—including cross-setting comparisons and edit-type breakdown analyses—are not sensitive to the timeout choice within this range.

C.3 Entity-Name Discrepancy Analysis

This section provides additional details for the entity-name discrepancy analysis in Section 5.4. The goal of this analysis is to characterize what kinds of non-identical surface forms appear in the verified H-H gold alignments and how EA performance differs across these variation types.

Category definitions. We define nine mutually exclusive entity-name discrepancy categories. Each category captures a primary type of name

Text2KG	Gold	NameEA	FGWEA	LightEA	SUDA	ERAlign	LEA	LLM4EA	Ave.
A-3.5	Step 5 only	94.0	88.5	89.7	44.8	80.4	81.3	63.9	77.5
	Expanded	92.4	84.6	87.7	42.7	78.6	80.3	61.8	75.5
A-5.1	Step 5 only	93.1	81.3	88.0	58.8	66.4	77.7	74.5	77.1
	Expanded	91.6	76.3	86.3	55.6	65.7	77.4	73.5	75.2

Table 8: Calibration evaluation with supplemental generated-only alignments on the ReDocRED-EA H-L setting. The "Step 5 only" rows reproduce the strict-gold evaluation under the final exact-match-filtered protocol, and the "Expanded" rows add manually identified generated-only alignable pairs.

Category	Definition	Typical patterns
Orthographic	Differences are mainly due to orthography or formatting within the same language and naming convention.	Case, spacing, punctuation, hyphenation, apostrophes, Unicode normalization, or minor diacritic differences.
Acronym	One name is an abbreviated form of the other.	Full name vs. acronym, abbreviation, or initialism.
Rephrasing/reordering	The two names use same-language lexical variants or reorder equivalent name components.	Alternative titles, conventional retitling, word-order changes, or name-order reversal.
Name shortening	The difference is mainly due to a shorter or more expanded form of a personal name.	Initials, added or removed middle names, given-name expansion, surname-only forms, or suffixes.
Common vs official	One side uses a common, informal, or alias-like name, while the other uses an official, legal, or canonical name.	Common names, stage names, pen names, nicknames, legal names, official names, or canonical titles.
Formal designator	The difference is mainly the presence or absence of a formal title, role label, or organization-form marker.	*Sir*, *Prince*, *Inc.*, *Ltd.*, *AB*, *Group*.
Parenthetical type	One name adds an explicit type label or parenthetical disambiguator while preserving the same referent.	Parenthetical markers or typed names such as "(band)", "(artist)", or "(play)".
Cross-lingual variant	The difference is mainly caused by translation, transliteration, romanization, or another form of cross-lingual transfer.	Translated names or titles, endonym/exonym pairs, transliterations, romanizations, or language-transfer-based spellings.
Scope qualifier	The pair is treated as equivalent in the benchmark, but one side adds or removes a qualifier or mildly shifts geographic, administrative, or conceptual scope.	Place or administrative qualifiers, contextual qualifiers, or mild scope/granularity shifts still annotated as equivalent.

Table 9: Definitions of entity-name discrepancy categories.

variation between a source entity and its verified H-H target counterpart. Table 9 lists the category definitions and representative examples.

LLM-based category assignment. We assign each evaluated H-H query pair to one of the nine categories using an LLM-as-a-judge classifier. Specifically, we use `gpt-5.4-2026-03-05` and provide the source entity name, the target entity name, the category definitions, and representative examples as input. The model is instructed to select the single category that best describes the primary surface-form difference between the two names. When multiple categories could apply, the model selects the category corresponding to the most salient difference. The category assignment is used only for analysis and does not affect benchmark construction, gold alignment derivation, or EA evaluation.

Category distributions. Figure 6 shows the distribution of entity-name discrepancy categories in the evaluated H-H query pairs. We compute these distributions after exact-match filtering, because exact-match entity pairs are excluded from evaluation in our experiments. The figure shows that the verified H-H alignments contain a range of non-trivial name variations rather than being reducible to exact or case-normalized matches. It also provides context for the category-level H@1 results: categories with smaller support should be interpreted as diagnostic slices, while recurring cat-

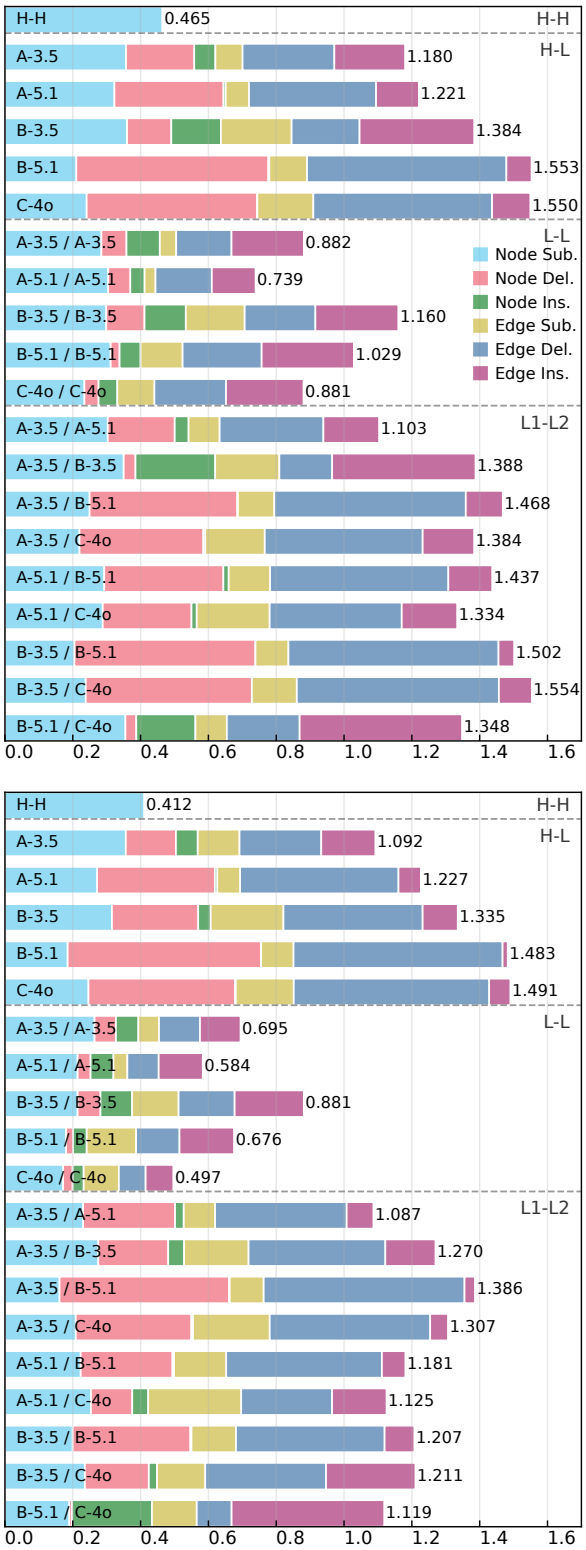

Figure 4: Normalized GED breakdown on BioREDirect-EA and REDFM-EA. The top panel shows BioREDirect-EA, and the bottom panel shows REDFM-EA. Each bar reports the normalized GED, decomposed by edit operation type.

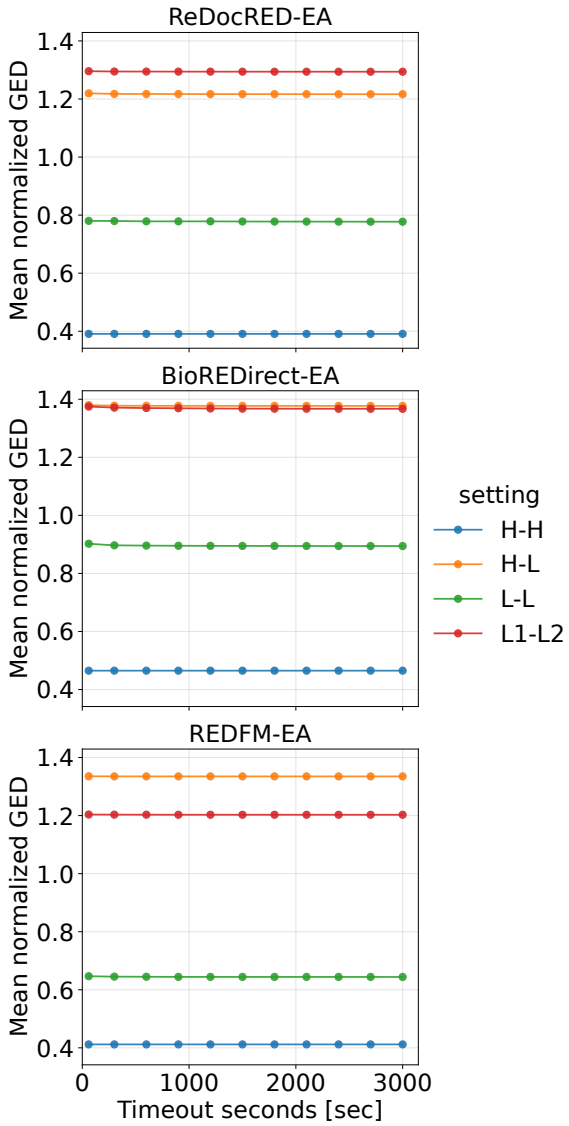

Figure 5: Timeout sensitivity of approximate normalized GED.

egories indicate broader sources of surface-form variation in the benchmark.

Additional Results. Additional results on BioREDirect-EA and REDFM-EA further support the same conclusions. Table 10 reports category-level H@1 on the two datasets. Across both datasets, H-H remains higher than the generated-graph settings for almost all entity-name discrepancy categories, showing that Text2KG-induced discrepancies make alignment harder even when the underlying type of entity-name discrepancy is fixed. The advantage of L-L over L1-L2 is also consistent across categories, indicating that using the same Text2KG pipeline reduces alignment difficulty beyond dataset-level averages.

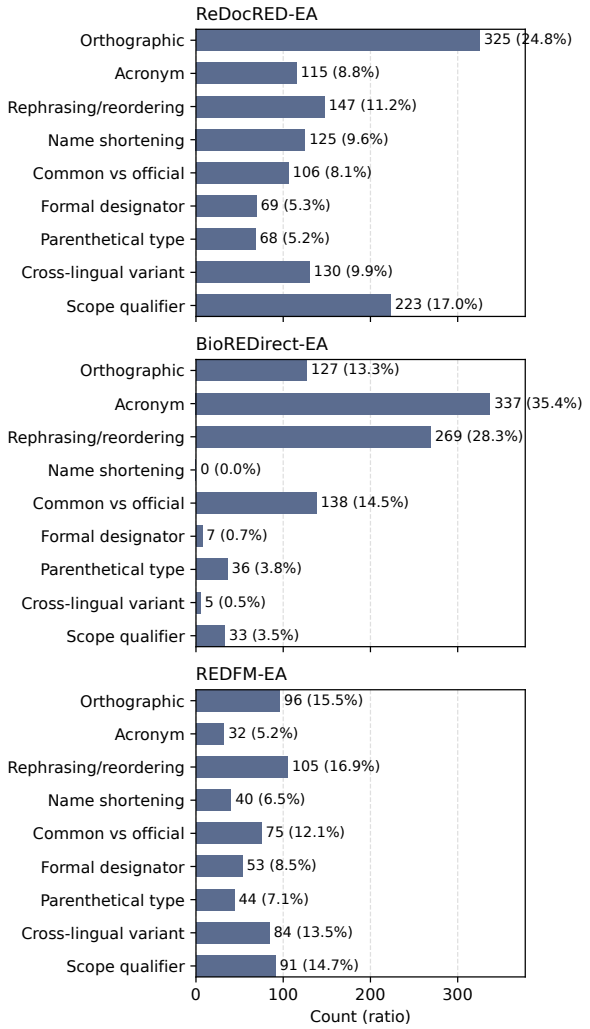

Figure 6: Distribution of entity-name discrepancy categories in evaluated H-H query pairs. The top, middle, and bottom panels correspond to ReDocRED-EA, BioREDirect-EA, and REDFM-EA, respectively.

Category	H-H	H-L	L-L	L1-L2
BioREDirect-EA				
Orthographic	91.9	76.0	79.5	70.5
Acronym	71.2	44.9	46.1	39.0
Rephrasing/reordering	88.8	67.3	69.5	61.9
Name shortening	–	–	–	–
Common vs official	68.4	45.1	49.1	38.6
Formal designator	91.8	82.6	81.3	68.6
Parenthetical type	87.2	70.4	74.8	58.9
Cross-lingual variant	90.7	80.7	86.3	68.6
Scope qualifier	89.5	71.7	77.8	64.8
REDFM-EA				
Orthographic	95.9	83.2	86.6	81.4
Acronym	75.3	48.5	60.3	51.4
Rephrasing/reordering	90.5	73.7	75.6	66.3
Name shortening	97.1	85.8	88.6	87.3
Common vs official	86.8	68.1	71.5	64.2
Formal designator	93.6	78.3	81.4	75.9
Parenthetical type	91.5	81.7	84.4	80.9
Cross-lingual variant	87.0	67.7	70.8	64.8
Scope qualifier	89.7	80.0	82.1	78.1

Table 10: Hits@1 by entity-name discrepancy category on BioREDirect-EA and REDFM-EA. Generated-graph settings are averaged over the corresponding Text2KG configurations. "–" indicates that no evaluated pairs fall into the category.

BioREDirect-EA is especially challenging for categories involving strong terminological variation. Acronym and common-vs-official variants obtain low scores even in H-H, and their performance drops substantially in generated-graph settings: for example, acronym pairs decrease from 71.2 H@1 in H-H to 44.9 in H-L and 39.0 in L1-L2, while common-vs-official pairs decrease from 68.4 to 45.1 and 38.6, respectively. This confirms that the lower aggregate performance on BioREDirect-EA is not merely a dataset-level effect, but is tied to specialized surface forms whose lexical realizations are difficult to normalize, especially after Text2KG generation.

REDFM-EA shows the same qualitative trends but with generally higher absolute scores. Orthographic variants, name shortening, parenthetical types, and scope qualifiers remain comparatively easier, whereas acronym, common-vs-official, and cross-lingual variants are harder. In particular, acronym pairs remain the most difficult category under generated-graph settings, with H@1 dropping from 75.3 in H-H to 48.5 in H-L and 51.4 in L1-L2. These results indicate that surface-form variation interacts with Text2KG-induced graph discrepancies across domains: simple orthographic or shortening variants are often handled well, but abbreviations, official/common-name alternations, and cross-lingual variants remain difficult when generated graphs alter naming, coverage, and structural evidence.

C.4 Entity-Type Analysis

This section analyzes whether EA performance differs across entity types. Unlike the entity-name discrepancy analysis in Appendix C.3, which categorizes how an entity name changes between the source and target sides, this analysis uses the entity-type annotations inherited from the original base datasets. Because ReDocRED (Tan et al., 2022), BioREDirect (Lai et al., 2025), and REDFM (Huguet Cabot et al., 2023) define entity types under different annotation schemas, we report the results separately for each dataset rather

Dataset	Category	NameEA	FGWEA	LightEA	SUDA	ERAlign	LEA	LLM4EA
ReDocRED-EA	Orthographic	99.9	88.8	97.3	86.2	77.8	80.7	73.8
	Acronym	69.1	71.8	63.0	75.4	22.3	84.6	73.3
	Rephrasing/reordering	92.5	80.4	91.4	84.3	64.5	84.4	67.7
	Name shortening	97.8	89.4	99.1	88.9	84.8	80.5	84.1
	Common vs official	78.4	77.4	85.0	77.8	58.6	76.7	69.6
	Formal designator	96.8	89.5	96.3	82.7	80.4	73.7	80.9
	Parenthetical type	96.1	88.6	96.3	84.3	75.3	72.4	75.6
	Cross-lingual variant	97.2	83.5	78.2	83.0	45.7	81.2	72.7
	Scope qualifier	93.9	84.3	91.0	77.1	54.5	77.2	78.5
BioREDirect-EA	Orthographic	97.0	67.2	97.0	74.1	68.2	81.0	71.6
	Acronym	56.1	47.0	44.1	49.7	11.3	70.8	73.3
	Rephrasing/reordering	92.7	59.3	76.2	72.3	43.0	83.0	76.6
	Name shortening	–	–	–	–	–	–	–
	Common vs official	43.6	43.2	52.7	47.5	20.9	71.0	73.1
	Formal designator	100.0	66.3	98.3	76.5	62.9	87.3	76.4
	Parenthetical type	92.4	68.5	89.3	77.4	35.9	59.3	87.0
	Cross-lingual variant	100.0	78.3	100.0	64.8	45.0	87.8	95.0
	Scope qualifier	96.9	58.9	95.7	73.1	57.1	78.7	71.2
REDFM-EA	Orthographic	100.0	73.1	97.1	85.7	83.7	86.7	81.4
	Acronym	52.8	58.9	52.4	62.1	21.3	88.2	76.3
	Rephrasing/reordering	91.7	69.1	89.6	76.0	59.7	89.1	60.6
	Name shortening	99.6	79.6	99.6	82.5	87.5	85.5	93.8
	Common vs official	72.3	66.5	81.3	72.5	58.6	77.0	80.5
	Formal designator	99.1	72.6	93.7	82.5	70.7	82.5	74.9
	Parenthetical type	95.6	77.4	94.3	77.6	78.9	82.6	85.8
	Cross-lingual variant	88.0	67.0	69.3	77.5	45.3	84.0	77.0
	Scope qualifier	91.3	72.7	93.0	78.2	71.4	84.5	86.3

Table 11: Hits@1 by entity-name discrepancy category. Categories describe the variation between each source entity and its validated H-H counterpart. Each column is averaged over all T2KG configurations. "–" indicates that no evaluated pairs fall into the category.

than merging types across datasets.

For each evaluated H-H query pair, we assign the entity type of the original source-side entity in the base dataset. When an entity is rewritten in Step 2, the rewritten target-side entity inherits the same entity type as the original entity, because the rewriting is intended to preserve referential identity. For H-L, L-L, and L1-L2, we evaluate the same projected gold correspondences as in the main experiments and aggregate Hits@1 over the corresponding Text2KG configurations. We also report the number and percentage of evaluated query pairs for each entity type to indicate the support of each slice.

Table 12 shows that the degradation in generated-graph settings is not confined to a single entity type. Across datasets, H-H generally remains higher than H-L, L-L, and L1-L2 for most entity types, indicating that Text2KG-induced mismatch affects multiple semantic classes. The BioREDirect-EA results are particularly informative: gene entities form the largest slice and are also the most difficult type, while disease, chemical, and variant entities obtain higher scores. This suggests that the lower aggregate performance on BioREDirect-EA is not uniform across specialized entities, but partly reflects type-dependent difficulty within the domain. In ReDocRED-EA and REDFM-EA, person- and time/date-like entities tend to be easier, whereas organization, location, media, or concept-like entities are often more affected by generated-graph settings. These patterns should be interpreted as diagnostic evidence rather than standalone conclusions, because some entity types have limited support.

C.5 Graph-Size Sensitivity Analysis

Table 13 reports H@1 by graph-size bin across all three datasets. For generated-graph settings, we sort graph pairs within each Text2KG configuration by the total number of nodes and edges and split them into approximately equal-sized small (S), medium (M), and large (L) bins. H-H graph pairs are split analogously. Across datasets, the main trends remain stable across size bins: H-H is consistently higher than settings involving generated graphs, and L-L generally remains higher than L1-L2. This indicates that the main perfor-

Type	#Pair (%)	H-H	H-L	L-L	L1-L2
ReDocRED-EA					
LOC	529 (35.4%)	89.9	73.2	80.1	73.2
PER	300 (20.1%)	94.7	80.9	88.0	81.9
ORG	259 (17.3%)	89.0	73.0	77.6	68.6
MISC	237 (15.9%)	94.7	77.5	84.3	73.7
TIME	168 (11.3%)	95.1	82.7	90.4	83.4
Overall	1,493 (100%)	91.9	76.2	82.9	75.5
BioREDirect-EA					
gene	390 (38.8%)	73.4	51.0	50.7	41.5
disease	309 (30.7%)	85.5	60.2	65.1	52.9
chemical	179 (17.8%)	83.2	61.4	62.5	55.9
variant	128 (12.7%)	82.6	61.1	63.5	57.5
Overall	1,006 (100%)	79.9	57.1	61.2	50.9
REDFM-EA					
LOC	243 (34.3%)	89.7	74.8	76.9	70.1
Concept	125 (17.7%)	91.6	74.6	75.6	73.5
ORG	97 (13.7%)	88.2	72.0	77.6	72.1
PER	82 (11.6%)	92.2	80.6	81.7	77.1
MEDIA	81 (11.4%)	90.3	71.6	74.8	66.3
DATE	29 (4.1%)	97.2	85.8	82.6	79.5
Others	51 (7.2%)	91.3	72.9	78.7	73.6
Overall	708 (100%)	90.3	74.9	77.9	72.0

Table 12: Hits@1 by entity type across the three benchmarks. #Pair reports the number and percentage of evaluated H-H query pairs assigned to each type. H-L, L-L, and L1-L2 results are averaged over the corresponding Text2KG configurations. Entity types are dataset-specific.

mance degradation observed in generated-graph settings is not driven solely by a particular graph-size regime. At the same time, BioREDirect-EA and REDFM-EA show somewhat larger drops for larger graph pairs in some settings, suggesting that graph size can modulate difficulty even when it does not change the overall conclusions.

C.6 Full Results

We report the full configuration-level H@1 results for all three datasets without averaging over Text2KG configurations. These tables expose the variation across individual Text2KG pipelines and model choices, while preserving the same query filtering and instance-weighted aggregation protocol used in the main experiments.

To verify that our conclusions are not specific to Hits@1, we also report Hits@3 (H@3) and Mean Reciprocal Rank (MRR). For each EA setting $s \in \{\text{H-H, H-L, L-L, L1-L2}\}$, we aggregate instance-level scores by weighting each graph pair by the number of evaluated query entities:

$$\mathrm{H@3}^s = \frac{1}{\sum_{i=1}^{I_s} |M_i^s|} \sum_{i=1}^{I_s} \left[\mathrm{H@3}_i \times |M_i^s|\right], \quad (1)$$

$$\mathrm{MRR}^s = \frac{1}{\sum_{i=1}^{I_s} |M_i^s|} \sum_{i=1}^{I_s} \left[\mathrm{MRR}_i \times |M_i^s|\right], \quad (2)$$

where I_s is the number of evaluated graph pairs under setting s, M_i^s is the gold alignment set for pair i, and $\mathrm{H@3}_i$ and MRR_i are the instance-level Hits@3 and MRR, respectively.

The H@3 and MRR results show the same qualitative trends as H@1: performance decreases when at least one side is an LLM-generated graph, L-L is generally easier than L1-L2, and methods that handle unmatched or noisy candidates remain competitive under Text2KG-induced mismatch. These results indicate that the main findings are not artifacts of using H@1 as the primary metric.

Dataset	Setting	S	M	L	All
ReDocRED-EA	H-H	93.3	93.0	90.3	91.9
	H-L	76.2	76.5	76.0	76.2
	L-L	81.3	83.4	83.2	82.9
	L1-L2	74.1	75.2	76.3	75.5
BioREDirect-EA	H-H	85.3	82.0	77.1	79.9
	H-L	62.8	56.7	54.4	57.1
	L-L	65.2	61.6	57.6	61.2
	L1-L2	56.3	50.7	47.1	50.9
REDFM-EA	H-H	91.0	90.2	90.1	90.3
	H-L	75.8	75.3	74.1	74.9
	L-L	79.2	79.9	75.5	77.9
	L1-L2	75.2	72.1	69.5	72.0

Table 13: Hits@1 by graph-size bin across datasets. For generated-graph settings, graph pairs are sorted by the total number of nodes and edges within each Text2KG configuration and split into approximately equal-sized small (S), medium (M), and large (L) bins. H-H graph pairs are split analogously. "All" reports the aggregate score over all graph pairs.

Setting	T2KG	NameEA	FGWEA	LightEA	SUDA	ERAlign	LEA	LLM4EA	Ave.
H-H	-	93.3	99.8	98.8	77.2	99.2	85.4	89.5	91.9
H-L	**A**-3.5	94.0	88.5	89.7	44.8	80.4	81.3	63.9	77.5
	A-5.1	93.1	81.3	88.0	58.8	66.4	77.7	74.5	77.1
	B-3.5	95.2	90.5	92.4	51.8	77.7	78.8	65.8	78.9
	B-5.1	91.2	75.9	84.5	63.7	61.0	73.6	56.5	72.4
	C-4o	92.5	81.9	88.0	57.9	63.5	74.7	67.6	75.1
L-L	**A**-3.5 / **A**-3.5	94.3	86.6	90.9	59.5	85.5	81.8	78.0	82.4
	A-5.1 / **A**-5.1	92.7	81.2	86.8	64.5	87.9	84.1	85.6	83.2
	B-3.5 / **B**-3.5	95.7	82.4	91.2	67.1	88.3	82.1	76.0	83.2
	B-5.1 / **B**-5.1	91.8	74.6	87.0	68.3	88.1	79.8	66.4	79.4
	C-4o / **C**-4o	92.8	90.9	90.9	70.8	90.8	84.0	84.8	86.4
L1-L2	**A**-3.5 / **A**-5.1	93.0	69.3	83.3	66.2	71.3	76.8	74.8	76.4
	A-3.5 / **B**-3.5	95.4	72.8	86.6	52.7	70.4	77.7	67.7	74.8
	A-3.5 / **B**-5.1	92.0	58.7	80.2	68.9	67.1	64.9	74.2	72.3
	A-5.1 / **B**-5.1	91.4	71.2	83.9	61.7	71.4	72.3	68.5	74.3
	A-3.5 / **C**-4o	92.3	68.4	85.7	60.2	72.3	73.2	73.3	75.1
	A-5.1 / **C**-4o	92.5	78.7	86.3	48.6	81.3	77.1	73.9	76.9
	B-3.5 / **B**-5.1	92.4	62.4	82.7	66.6	71.7	71.0	70.1	73.9
	B-3.5 / **C**-4o	92.9	78.4	89.0	68.5	71.6	76.3	77.4	79.2
	B-5.1 / **C**-4o	92.2	85.0	87.4	37.6	88.8	78.9	64.8	76.4
Ave.	-	93.0	78.9	87.7	60.8	77.7	77.6	72.7	78.3

Table 14: Full **Hits@1** results on **ReDocRED-EA**. T2KG labels use compact notation: **A**=EDC, **B**=KG-LLM, **C**=KG-Gen; numbers denote the underlying LLM (gpt-3.5/4o/5.1).

Setting	T2KG	NameEA	FGWEA	LightEA	SUDA	ERAlign	LEA	LLM4EA	Ave.
H-H	-	76.5	96.1	87.8	38.9	84.2	83.0	92.9	79.9
H-L	**A**-3.5	75.1	55.6	66.9	31.6	62.1	73.4	68.1	61.9
	A-5.1	70.7	52.7	62.2	33.3	52.6	72.0	72.3	59.4
	B-3.5	76.7	50.9	64.2	27.0	54.7	76.7	62.0	58.9
	B-5.1	70.9	30.1	55.9	34.3	35.4	63.1	51.6	48.8
	C-4o	67.5	49.6	59.7	30.6	47.2	70.4	69.9	56.4
L-L	**A**-3.5 / **A**-3.5	75.4	52.2	68.0	30.3	71.3	76.8	64.9	62.7
	A-5.1 / **A**-5.1	68.6	46.5	67.4	38.8	77.4	76.5	77.9	64.7
	B-3.5 / **B**-3.5	72.1	27.9	65.1	27.9	53.5	83.7	66.7	56.7
	B-5.1 / **B**-5.1	76.2	27.4	68.3	38.4	70.7	74.4	64.2	59.9
	C-4o / **C**-4o	67.9	47.3	61.7	35.6	69.7	81.1	71.7	62.1
L1-L2	**A**-3.5 / **A**-5.1	73.8	43.8	61.7	31.7	54.1	76.3	63.8	57.9
	A-3.5 / **B**-3.5	80.8	51.3	61.5	23.1	57.7	76.9	59.0	58.6
	A-3.5 / **B**-5.1	71.8	24.7	52.3	32.2	33.1	67.2	49.1	47.2
	A-5.1 / **B**-5.1	71.9	23.4	53.1	28.6	33.3	56.8	57.6	46.4
	A-3.5 / **C**-4o	69.5	41.8	55.6	30.2	49.6	73.5	62.7	54.7
	A-5.1 / **C**-4o	66.3	39.4	56.7	29.9	49.9	71.9	68.5	54.6
	B-3.5 / **B**-5.1	67.7	6.5	43.5	25.8	25.8	59.7	58.1	41.0
	B-3.5 / **C**-4o	64.8	18.2	46.6	20.5	39.8	64.8	60.2	45.0
	B-5.1 / **C**-4o	72.2	38.0	62.0	18.1	56.1	74.3	51.3	53.1
Ave.	-	71.8	41.2	61.0	30.3	53.9	72.6	64.6	56.5

Table 15: Full **Hits@1** results on **BioREDirect-EA**. T2KG labels use compact notation: **A**=EDC, **B**=KG-LLM, **C**=KG-Gen; numbers denote the underlying LLM (gpt-3.5/4o/5.1).

Setting	T2KG	NameEA	FGWEA	LightEA	SUDA	ERAlign	LEA	LLM4EA	Ave.
H-H	-	92.6	98.9	97.3	78.4	85.8	89.0	90.2	90.3
H-L	**A**-3.5	90.7	74.6	86.8	55.4	74.0	88.6	77.2	78.2
	A-5.1	89.0	72.1	83.5	66.6	73.2	85.2	77.8	78.2
	B-3.5	87.0	66.3	83.5	60.4	64.2	77.7	73.6	73.3
	B-5.1	86.7	50.4	77.5	64.8	59.6	76.5	76.5	70.3
	C-4o	88.8	62.2	83.5	60.5	69.8	80.0	76.2	74.4
L-L	**A**-3.5 / **A**-3.5	90.1	69.5	84.8	60.5	84.0	90.9	82.5	80.3
	A-5.1 / **A**-5.1	89.2	72.9	85.6	60.4	85.3	86.8	77.3	79.6
	B-3.5 / **B**-3.5	87.1	58.6	83.9	58.6	83.9	81.2	71.4	75.0
	B-5.1 / **B**-5.1	86.5	46.7	79.6	60.2	84.1	81.4	62.4	71.5
	C-4o / **C**-4o	89.2	70.3	87.4	71.0	91.9	86.9	85.1	83.1
L1-L2	**A**-3.5 / **A**-5.1	86.5	63.1	82.5	64.7	71.4	87.7	73.3	75.6
	A-3.5 / **B**-3.5	85.2	59.1	83.0	59.7	68.2	79.5	70.8	72.2
	A-3.5 / **B**-5.1	84.1	42.0	78.4	64.1	62.9	80.0	72.0	69.1
	A-5.1 / **B**-5.1	87.5	52.0	78.6	60.7	69.9	78.9	69.5	71.0
	A-3.5 / **C**-4o	87.6	55.6	83.6	62.4	72.4	82.8	77.9	74.6
	A-5.1 / **C**-4o	89.8	64.6	82.8	52.0	77.7	80.2	76.3	74.8
	B-3.5 / **B**-5.1	86.6	41.1	75.3	58.4	61.5	73.6	67.7	66.3
	B-3.5 / **C**-4o	88.7	55.9	81.8	54.7	72.5	80.2	76.8	72.9
	B-5.1 / **C**-4o	88.2	60.0	79.7	41.4	83.4	82.8	63.3	71.3
Ave.	-	88.1	61.8	83.0	60.7	74.8	82.5	74.9	75.1

Table 16: Full **Hits@1** results on **REDFM-EA**. T2KG labels use compact notation: **A**=EDC, **B**=KG-LLM, **C**=KG-Gen; numbers denote the underlying LLM (gpt-3.5/4o/5.1).

Setting	T2KG	NameEA	FGWEA	LightEA	SUDA	ERAlign	LEA	LLM4EA	Ave.
H-H	-	97.9	100.0	99.8	77.8	99.9	95.8	98.0	95.6
H-L	**A**-3.5	99.4	96.6	99.0	46.0	96.6	93.8	82.3	87.7
	A-5.1	97.8	94.7	96.3	60.0	89.7	91.5	86.7	88.1
	B-3.5	98.6	96.9	98.2	52.7	95.5	93.4	82.9	88.3
	B-5.1	97.1	91.4	93.2	66.4	88.9	88.0	67.6	84.7
	C-4o	97.1	94.0	95.0	59.2	87.4	88.7	78.9	85.8
L-L	**A**-3.5 / **A**-3.5	99.4	98.3	97.7	60.7	95.7	93.7	89.6	90.7
	A-5.1 / **A**-5.1	97.9	96.1	97.0	65.9	95.8	94.4	94.3	91.6
	B-3.5 / **B**-3.5	99.2	95.5	97.9	67.4	98.4	94.4	89.2	91.7
	B-5.1 / **B**-5.1	97.1	91.4	93.8	69.2	94.8	94.3	77.5	88.3
	C-4o / **C**-4o	97.0	97.2	96.3	71.3	95.8	93.1	92.3	91.9
L1-L2	**A**-3.5 / **A**-5.1	97.6	89.9	93.0	69.3	88.0	90.8	83.2	87.4
	A-3.5 / **B**-3.5	98.6	89.0	95.4	54.4	89.8	91.5	84.6	86.2
	A-3.5 / **B**-5.1	97.2	81.1	90.8	72.4	87.1	83.0	79.3	84.4
	A-5.1 / **B**-5.1	96.7	88.7	91.1	64.0	88.4	90.3	79.3	85.5
	A-3.5 / **C**-4o	96.9	88.0	93.6	63.3	90.1	87.8	83.1	86.1
	A-5.1 / **C**-4o	96.7	91.9	95.1	49.3	93.0	91.3	84.9	86.0
	B-3.5 / **B**-5.1	96.8	83.2	93.1	69.4	89.4	89.2	77.7	85.5
	B-3.5 / **C**-4o	96.6	90.3	94.5	69.2	92.4	91.7	86.7	88.8
	B-5.1 / **C**-4o	96.7	96.8	95.0	37.7	95.3	91.1	77.1	84.2
Ave.	-	97.6	92.6	95.3	62.3	92.6	91.4	83.8	87.9

Table 17: Full **Hits@3** results on **ReDocRED-EA**. T2KG labels use compact notation: **A**=EDC, **B**=KG-LLM, **C**=KG-Gen; numbers denote the underlying LLM (gpt-3.5/4o/5.1).

Setting	T2KG	NameEA	FGWEA	LightEA	SUDA	ERAlign	LEA	LLM4EA	Ave.
H-H	-	91.8	99.2	97.9	41.3	94.4	95.6	98.8	88.4
H-L	**A**-3.5	89.5	85.0	83.3	35.6	83.4	90.1	86.5	79.1
	A-5.1	85.6	80.9	73.0	37.2	75.3	90.7	87.2	75.7
	B-3.5	92.5	81.8	83.0	29.6	83.6	96.2	82.3	78.4
	B-5.1	85.3	52.0	68.3	39.9	57.8	82.7	62.2	64.0
	C-4o	82.8	71.0	72.5	34.8	71.5	87.3	81.7	71.6
L-L	**A**-3.5 / **A**-3.5	89.5	78.5	81.1	32.9	91.3	90.8	84.7	78.4
	A-5.1 / **A**-5.1	84.4	75.1	80.2	39.7	91.5	89.2	90.7	78.7
	B-3.5 / **B**-3.5	97.7	62.8	88.4	32.6	86.0	95.3	81.0	77.7
	B-5.1 / **B**-5.1	88.4	47.0	81.1	40.2	87.8	89.6	84.3	74.1
	C-4o / **C**-4o	83.3	73.1	76.6	36.3	87.8	94.3	83.8	76.5
L1-L2	**A**-3.5 / **A**-5.1	89.2	69.6	75.4	34.6	77.3	93.3	81.5	74.4
	A-3.5 / **B**-3.5	93.6	76.9	83.3	26.9	85.9	94.9	83.3	77.8
	A-3.5 / **B**-5.1	86.2	44.3	62.6	36.2	59.4	89.1	61.3	62.7
	A-5.1 / **B**-5.1	84.4	43.8	63.0	30.2	59.4	77.6	69.1	61.1
	A-3.5 / **C**-4o	84.7	65.5	71.6	35.3	72.1	90.5	73.8	70.5
	A-5.1 / **C**-4o	81.8	67.2	70.4	31.9	73.4	88.4	79.7	70.4
	B-3.5 / **B**-5.1	85.5	14.5	58.1	29.0	50.0	77.4	66.1	54.4
	B-3.5 / **C**-4o	78.4	45.5	61.4	25.0	58.0	88.6	68.2	60.7
	B-5.1 / **C**-4o	84.8	63.3	71.7	18.1	75.5	89.9	65.7	67.0
Ave.	-	87.0	64.9	75.1	33.4	76.1	89.6	78.6	72.1

Table 18: Full **Hits@3** results on **BioREDirect-EA**. T2KG labels use compact notation: **A**=EDC, **B**=KG-LLM, **C**=KG-Gen; numbers denote the underlying LLM (gpt-3.5/4o/5.1).

Setting	T2KG	NameEA	FGWEA	LightEA	SUDA	ERAlign	LEA	LLM4EA	Ave.
H-H	-	98.2	99.8	99.7	79.2	97.4	98.5	99.3	96.0
H-L	**A**-3.5	97.0	95.2	95.8	58.4	94.0	98.5	95.6	90.6
	A-5.1	96.3	91.0	93.8	69.7	93.0	96.2	89.4	89.9
	B-3.5	96.1	89.1	94.4	64.2	90.2	93.6	87.9	87.9
	B-5.1	94.3	76.2	87.8	70.3	82.3	92.3	83.6	83.8
	C-4o	96.2	84.1	93.0	63.6	89.0	94.1	86.9	86.7
L-L	**A**-3.5 / **A**-3.5	97.1	90.9	97.5	61.7	97.1	99.2	94.7	91.2
	A-5.1 / **A**-5.1	96.3	91.9	96.1	61.1	95.1	96.6	90.5	89.7
	B-3.5 / **B**-3.5	96.8	84.4	95.7	60.8	95.2	94.6	89.1	88.1
	B-5.1 / **B**-5.1	95.2	73.7	90.7	61.3	94.4	95.8	81.5	84.7
	C-4o / **C**-4o	96.5	87.4	95.5	71.5	98.0	95.7	91.2	90.8
L1-L2	**A**-3.5 / **A**-5.1	95.2	84.5	94.0	68.7	90.5	97.2	89.0	88.4
	A-3.5 / **B**-3.5	95.5	81.2	93.8	63.1	91.5	94.9	86.9	86.7
	A-3.5 / **B**-5.1	93.1	68.2	88.2	66.9	84.9	93.1	81.5	82.3
	A-5.1 / **B**-5.1	95.4	76.7	91.1	62.9	88.9	94.6	83.0	84.6
	A-3.5 / **C**-4o	94.4	76.8	91.6	66.0	92.0	97.2	88.3	86.6
	A-5.1 / **C**-4o	96.8	85.5	95.2	53.1	93.3	94.9	87.6	86.6
	B-3.5 / **B**-5.1	94.4	72.3	87.9	63.2	84.4	91.8	77.0	81.6
	B-3.5 / **C**-4o	96.0	81.4	90.3	55.9	90.7	95.5	88.0	85.4
	B-5.1 / **C**-4o	96.6	81.4	91.8	42.0	94.6	93.5	82.5	83.2
Ave.	-	95.9	83.6	93.2	63.2	91.8	95.4	87.7	87.2

Table 19: Full **Hits@3** results on **REDFM-EA**. T2KG labels use compact notation: **A**=EDC, **B**=KG-LLM, **C**=KG-Gen; numbers denote the underlying LLM (gpt-3.5/4o/5.1).

Setting	T2KG	NameEA	FGWEA	LightEA	SUDA	ERAlign	LEA	LLM4EA	Ave.
H-H	-	95.8	99.9	99.3	77.5	99.6	90.8	93.8	93.8
H-L	**A**-3.5	96.7	92.8	94.1	45.4	88.5	88.0	75.3	83.0
	A-5.1	95.7	88.3	92.4	59.5	78.6	84.9	82.1	83.1
	B-3.5	97.0	94.1	95.4	52.2	87.0	86.5	76.4	84.1
	B-5.1	94.3	84.3	89.5	65.2	75.6	82.0	65.0	79.4
	C-4o	95.1	88.2	92.2	58.7	76.7	82.8	75.7	81.3
L-L	**A**-3.5 / **A**-3.5	96.9	92.4	94.4	60.1	91.2	88.3	85.1	86.9
	A-5.1 / **A**-5.1	95.5	88.7	91.9	65.2	92.2	89.8	90.4	87.7
	B-3.5 / **B**-3.5	97.5	89.3	94.6	67.2	93.5	88.6	83.5	87.7
	B-5.1 / **B**-5.1	94.6	83.7	91.1	68.8	91.8	87.4	74.3	84.5
	C-4o / **C**-4o	95.3	94.4	94.0	71.2	93.8	89.2	89.5	89.6
L1-L2	**A**-3.5 / **A**-5.1	95.5	79.8	89.2	67.9	80.9	84.7	80.6	82.7
	A-3.5 / **B**-3.5	97.2	82.3	91.4	53.6	81.3	85.2	77.6	81.2
	A-3.5 / **B**-5.1	94.6	70.9	86.1	70.7	78.0	75.9	78.5	79.2
	A-5.1 / **B**-5.1	94.2	80.5	88.3	63.1	80.8	82.1	75.7	80.7
	A-3.5 / **C**-4o	95.0	79.4	90.4	61.9	82.2	82.0	80.0	81.6
	A-5.1 / **C**-4o	95.0	86.0	91.2	49.1	87.8	84.8	81.2	82.1
	B-3.5 / **B**-5.1	94.9	74.1	88.3	68.1	81.5	80.5	75.9	80.5
	B-3.5 / **C**-4o	95.3	85.6	92.4	68.9	82.4	84.5	83.4	84.7
	B-5.1 / **C**-4o	94.9	91.0	91.6	37.7	92.4	85.7	73.7	81.0
Ave.	-	95.6	86.3	91.9	61.6	85.8	85.2	79.9	83.7

Table 20: Full **MRR** results on **ReDocRED-EA**. T2KG labels use compact notation: **A**=EDC, **B**=KG-LLM, **C**=KG-Gen; numbers denote the underlying LLM (gpt-3.5/4o/5.1).

Setting	T2KG	NameEA	FGWEA	LightEA	SUDA	ERAlign	LEA	LLM4EA	Ave.
H-H	-	84.9	97.7	92.8	40.3	89.8	89.4	96.0	84.4
H-L	**A**-3.5	83.5	71.0	76.7	34.4	74.7	82.7	78.1	71.6
	A-5.1	79.7	68.1	70.5	35.9	65.9	82.0	80.4	68.9
	B-3.5	85.1	68.1	75.8	28.3	70.9	86.3	73.6	69.7
	B-5.1	79.4	45.2	64.8	37.9	50.2	74.4	60.6	58.9
	C-4o	77.1	63.1	68.7	33.7	61.6	80.3	77.5	66.0
L-L	**A**-3.5 / **A**-3.5	83.8	67.1	77.0	31.9	82.1	84.9	76.0	71.8
	A-5.1 / **A**-5.1	78.3	63.0	75.8	39.3	85.1	83.7	85.0	72.9
	B-3.5 / **B**-3.5	83.5	50.0	77.6	30.8	70.5	89.7	76.3	68.3
	B-5.1 / **B**-5.1	83.1	41.3	76.2	39.1	80.5	83.5	74.8	68.4
	C-4o / **C**-4o	77.5	62.7	71.4	36.1	79.8	88.1	79.4	70.7
L1-L2	**A**-3.5 / **A**-5.1	82.3	59.8	71.0	33.3	67.8	85.4	74.4	67.7
	A-3.5 / **B**-3.5	87.3	66.5	73.5	24.8	72.8	85.8	72.5	69.0
	A-3.5 / **B**-5.1	80.2	39.1	61.2	34.5	48.7	78.8	60.0	57.5
	A-5.1 / **B**-5.1	79.6	37.9	62.0	29.6	49.7	70.1	66.4	56.5
	A-3.5 / **C**-4o	79.0	56.5	66.3	32.7	63.6	82.7	71.3	64.6
	A-5.1 / **C**-4o	75.9	57.2	66.2	30.9	63.9	81.1	76.0	64.4
	B-3.5 / **B**-5.1	77.8	17.0	54.2	28.8	43.9	71.2	65.2	51.2
	B-3.5 / **C**-4o	74.2	36.7	57.7	23.7	52.8	77.8	67.2	55.7
	B-5.1 / **C**-4o	80.4	54.1	69.4	18.1	67.5	83.1	62.4	62.2
Ave.	-	80.6	56.1	70.4	32.2	67.1	82.1	73.7	66.0

Table 21: Full **MRR** results on **BioREDirect-EA**. T2KG labels use compact notation: **A**=EDC, **B**=KG-LLM, **C**=KG-Gen; numbers denote the underlying LLM (gpt-3.5/4o/5.1).

Setting	T2KG	NameEA	FGWEA	LightEA	SUDA	ERAlign	LEA	LLM4EA	Ave.
H-H	-	95.6	99.3	98.5	78.8	91.8	93.8	94.5	93.2
H-L	**A**-3.5	94.2	84.9	91.9	56.9	84.4	93.7	86.1	84.6
	A-5.1	93.0	82.3	89.3	68.2	83.9	90.8	85.0	84.7
	B-3.5	91.7	78.4	89.7	62.3	78.0	86.2	82.1	81.2
	B-5.1	91.0	65.4	84.4	68.0	72.3	85.0	82.1	78.3
	C-4o	92.8	74.6	89.0	62.3	80.5	87.6	82.6	81.3
L-L	**A**-3.5 / **A**-3.5	93.9	81.2	90.9	61.0	90.7	94.8	88.9	85.9
	A-5.1 / **A**-5.1	93.1	82.9	91.1	60.8	90.6	91.9	85.0	85.0
	B-3.5 / **B**-3.5	92.1	73.0	90.1	59.8	90.3	88.4	81.6	82.2
	B-5.1 / **B**-5.1	91.2	62.8	86.1	60.7	89.9	88.7	73.4	79.0
	C-4o / **C**-4o	93.1	80.1	91.9	71.2	95.2	92.1	89.3	87.6
L1-L2	**A**-3.5 / **A**-5.1	91.5	75.1	88.7	66.7	81.9	92.8	82.5	82.8
	A-3.5 / **B**-3.5	90.7	72.0	89.1	61.3	80.0	87.7	80.5	80.2
	A-3.5 / **B**-5.1	89.2	58.3	84.5	66.0	74.6	87.4	78.4	76.9
	A-5.1 / **B**-5.1	91.8	66.5	85.7	61.9	80.3	86.8	78.1	78.7
	A-3.5 / **C**-4o	91.7	68.7	88.6	64.2	82.8	90.0	84.6	81.5
	A-5.1 / **C**-4o	93.5	76.6	89.3	52.5	86.0	87.8	83.0	81.3
	B-3.5 / **B**-5.1	91.0	58.5	82.9	60.9	74.6	83.7	75.3	75.3
	B-3.5 / **C**-4o	92.7	70.1	87.3	55.5	82.1	87.9	83.3	79.8
	B-5.1 / **C**-4o	92.6	72.9	87.2	41.7	89.3	89.2	74.7	78.2
Ave.	-	92.3	74.2	88.8	62.0	84.0	89.3	82.6	81.9

Table 22: Full **MRR** results on **REDFM-EA**. T2KG labels use compact notation: **A**=EDC, **B**=KG-LLM, **C**=KG-Gen; numbers denote the underlying LLM (gpt-3.5/4o/5.1).