From Regulatory Approvals to Patents: Cross-Domain Linking for Cardiovascular Device Traceability

Qingqing Yang[1], **Haijiang Liu**[2,3]*, **Moyan Li**[1]*

[1] Innovation, Policy and Entrepreneurship (IPE) Thrust,
The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, 511453, China
[2] School of Computer Science and Technology,
Wuhan University of Science and Technology, Wuhan 430065, China
[3] Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System,
Wuhan University of Science and Technology, Wuhan 430065, China

Abstract

Linking FDA-approved medical devices to their underlying United States Patent and Trademark Office (USPTO) patents enables critical applications such as recall root-cause analysis, M&A-driven IP discovery, and technology trajectory mapping. However, this cross-domain entity linking task remains unexplored due to severe *semantic gaps*: FDA documents focus on clinical outcomes, while patents describe technical mechanisms, yielding minimal lexical overlap. We formalize medical device-patent linking as a challenging cross-domain entity linking problem characterized by label scarcity and domain shifts. Using cardiovascular devices as a high-impact, representative domain featuring diverse technologies, high recall rates, and abundant disclosures, we construct a benchmark with 434 devices, 698K patents, and 585 high-fidelity expert-verified pairs. To address these challenges, we propose Bridge-MedDevKG, a coarse-to-fine framework that integrates (1) **MedDevOnto**, a domain-specific ontology that anchors device concepts via three-tier UMLS normalization; (2) **Multi-signal candidate generation** fusing company affiliation, semantic similarity, and ontology-weighted entity overlap; and (3) **Heterogeneous reranking** with multi-signal scoring and XGBoost classification on hard negatives. Our approach achieves a conservative lower-bound recall of 91.6% on the gold standard with 50.9% noise reduction, substantially outperforming LLM baselines under comparable evaluation. The resulting MedDevKG provides 6.8M high-confidence links, laying a scalable foundation for regulatory-IP integration across medical specialties.

1 Introduction

Medical devices undergo a rigorous regulatory approval process before clinical deployment, while their underlying innovations are independently protected by their patents. These siloed regulatory and IP processes lack systematic cross-referencing. Automated linking would enable critical applications: *recall root-cause analysis* by tracing recalled devices to implicated patent families, *M&A (mergers and acquisitions)-driven IP discovery* across corporate acquisitions, and *technology trajectory mapping* for competitive intelligence.

This cross-domain entity linking task remains largely unexplored due to severe semantic gaps: the Food and Drug Administration (FDA) documents emphasize clinical outcomes (e.g., "treats coronary artery disease"), while the United States Patent and Trademark Office (USPTO) patents describe technical mechanisms (e.g., "helical coil with radiopaque markers"), yielding minimal lexical overlap (Jac-

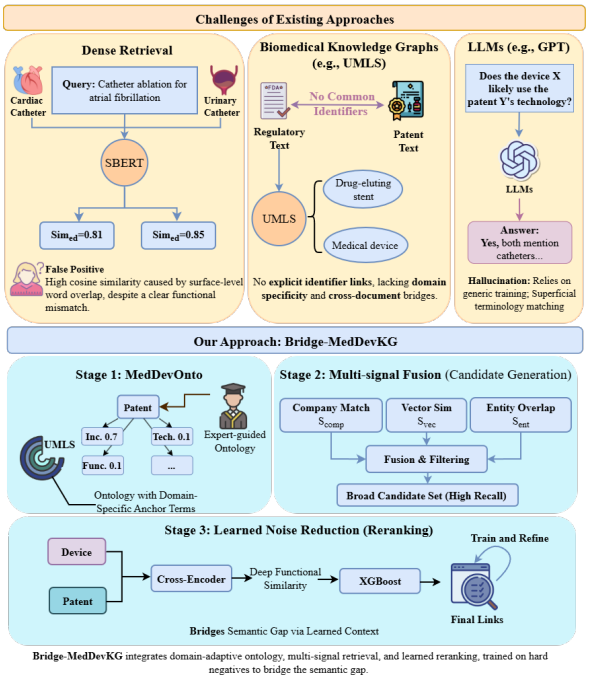

Figure 1: Comparison of approaches. (A) Existing methods struggle with cross-domain ambiguity. (B) Our **Bridge-MedDevKG** uses heterogeneous fusion to bridge the semantic gap where single-source methods fail.

*Co-corresponding Authors. E-mail: bill1103478225@outlook.com; moyanli@hkust-gz.edu.cn.

[0] https://github.com/myqqhub/Bridge-MedDevKG.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 34890–34906

card similarity 0.039 on verified pairs).

In this work, we bridge the semantic gap via an **ontology with domain-specific anchor terms** and **multi-signal reranking**. Despite existing approaches providing valuable tools for cross-domain matching, such as dense retrieval methods (Karpukhin et al., 2020; Reimers and Gurevych, 2019), biomedical knowledge graphs (Bodenreider, 2004; Chandak et al., 2023), and Large Language Models (LLMs) matching and ranking (Sun et al., 2023; Qin et al., 2024), when applied to medical device-patent linking, they often show limitations like **conflating** semantically similar but functionally distinct entities (e.g., cardiac vs. urinary catheters), **struggling** to elevate domain-critical terms (e.g., "stent") over generic ones (e.g., "device"), and **failing** to discriminate precise functional equivalence across regulatory and technical descriptions.

To address these challenges under realistic label scarcity, we propose **BRIDGE-MEDDEVKG**, a coarse-to-fine framework (Figure 1). It integrates **MEDDEVONTO**, an ontology with domain-specific anchor terms and three-tier UMLS concept normalization over the Unified Medical Language System (UMLS); **multi-signal candidate generation** fusing company affiliation, semantic similarity, and ontology-weighted entity overlap; and **heterogeneous reranking** with cross-encoder scoring and gradient-boosted classification on hard negatives.

Our contributions are:

1. The first formalization and benchmark for medical device-patent linking, with 585 high-fidelity expert-verified pairs under label scarcity.
2. MEDDEVONTO, identifying device-critical anchor terms as discriminative matching pivots and enabling cross-document entity resolution through three-tier UMLS concept normalization.
3. A label-efficient fusion pipeline achieving a conservative lower-bound recall of 91.6% with 50.9% noise reduction, outperforming LLM baselines under the reported evaluation settings.
4. MEDDEVKG, linking 434 cardiovascular devices to 698K patents via 6.8M high-confidence relationships.

2 Related Work

Cross-Domain Entity Linking. Entity linking (EL) connects mentions in text to canonical entries in a knowledge base. When source and target documents originate from different domains, the task becomes cross-domain entity linking, which presents unique challenges beyond standard EL (Shi et al., 2023) (formalized in §3.3 as C1–C4). Recent advances include joint representation learning across domains (Soliman, 2022), zero-shot transfer using BERT-based bi-encoders (Partalidou et al., 2022), pattern exploitation for improved domain transfer in NER (Blair and Bar, 2022), and pivot-based frameworks for cross-lingual settings (Rijhwani et al., 2019).

In our task, FDA approval documents describe clinical outcomes and safety profiles, whereas USPTO patents detail technical mechanisms and engineering specifications, resulting in fundamentally incompatible vocabularies that limit the ability of learned representations to bridge the gap.

Biomedical Knowledge Graphs. Structured knowledge resources have enabled significant advances in biomedical NLP. UMLS provides comprehensive cross-vocabulary normalization (Bodenreider, 2004), widely adopted for named entity recognition and relation extraction (Yuan et al., 2021). PrimeKG (Chandak et al., 2023) integrates multiple resources for precision medicine applications, while PKG 2.0 (Xu et al., 2025) connects scientific papers, patents, and clinical trials through citation and grant linkages. Cross-terminology link mining (Patel and Cimino, 2006) facilitates EHR integration via transitive relationship paths. In the patent domain, prior work extracts engineering knowledge graphs for design retrieval (Zuo et al., 2022; Siddharth et al., 2022), but focuses on *within-patent* analysis rather than cross-domain regulatory-IP linking.

Although these knowledge graphs capture structured knowledge of each field, they rely on *explicit cross-references*: citations, shared author identifiers, or grant numbers, to establish connections between entities, whereas *no standardized cross-reference exists between these two regulatory systems*. The relationship between a device and its underlying patents must therefore be inferred from indirect evidence such as company names, technical descriptions, and temporal patterns. Furthermore, standard entity matching typically assigns uniform importance to concepts within the same semantic type, failing to distinguish domain-critical terms (e.g., "drug-eluting stent") from generic ones (e.g., "medical device").[1]

[1] A concurrent study (Cunningham and Hall, 2025) maps

Retrieve-then-Rerank Architectures. Two-stage retrieve-then-rerank pipelines have become standard in information retrieval (Nogueira and Cho, 2019). Dense retrieval with learned representations addresses vocabulary mismatch (Karpukhin et al., 2020), while multi-aspect embeddings capture query-document relationships across multiple dimensions (Kong et al., 2022). For domain adaptation, disentangled representations separate domain-invariant and domain-specific features (Zhan et al., 2022), and regularized frameworks map heterogeneous domains onto shared latent spaces (Wang et al., 2009). Structure-aware approaches integrate document relationships with semantic content (Raman et al., 2022).

These methods achieve strong results when queries and documents occupy related semantic spaces. But in our case, with severe label scarcity, where training examples are limited, and domains are fundamentally misaligned, challenges remain.

LLMs for Information Retrieval. Large language models have demonstrated impressive zero-shot capabilities on standard IR benchmarks. Listwise ranking with GPT-4 achieves competitive results on established test collections (Sun et al., 2023), progressive training strategies bridge language modeling objectives with ranking tasks (Zhang et al., 2023), and uncertainty-aware approaches improve robustness (Zeng et al., 2024). Pairwise ranking prompting has also shown strong zero-shot and few-shot performance on diverse retrieval tasks (Qin et al., 2024).

However, when vocabulary disparity is severe, LLMs struggle to discriminate functional equivalence from surface similarity—a limitation particularly relevant for cross-domain matching where terminological conventions differ fundamentally.

Due to the nature of our task in terms of fundamentally different terminologies, indirect device-patent relationships inference, and distinguished terminology weighting for knowledge organization, we (1) build a domain-specific ontology with anchor terms and three-tier UMLS normalization that prioritizes clinically and technically critical concepts; (2) design a multi-signal fusion method combining structural, semantic, and knowledge-based evidence; and (3) construct a benchmark dataset of 585 expert-annotated device-patent pairs enabling systematic evaluation.

3 Task Definition

We formalize *medical device-patent linking* as a cross-domain entity linking task between the FDA-approved medical devices and the USPTO patents.

3.1 Problem Formulation

Let $\mathcal{D} = \{d_1, \ldots, d_n\}$ be the set of the FDA Premarket Approval (PMA) summaries for cardiovascular devices and $\mathcal{P} = \{p_1, \ldots, p_m\}$ the set of the USPTO patent abstracts in relevant medical classes. The goal is to predict a set of links $\mathcal{L} \subseteq \mathcal{D} \times \mathcal{P}$ such that $(d_i, p_j) \in \mathcal{L}$ if patent p_j protects technology embodied in device d_i. This yields a highly asymmetric one-to-many mapping (median 5, maximum 83 patents per device).

3.2 Data Construction and Benchmark

We focus on cardiovascular devices, selected as a high-impact representative showcase of the broader task. This subdomain accounts for the majority of life-sustaining device approvals, exhibits the highest recall rates, spans diverse technologies (stents, catheters, valves, pacemakers, ablation systems, grafts), experiences frequent M&A, and offers richer public patent disclosures than most other specialties, which enables reliable gold-standard construction while capturing the full spectrum of cross-domain challenges.

We construct the device corpus from 434 Class III cardiovascular PMA approvals (1976–2024), filtered by product codes and keywords, with manual exclusion of 29 non-cardiovascular entries (Appendix A for full criteria).

The patent corpus comprises 698,191 utility patents filtered by cardiovascular-relevant Cooperative Patent Classification (CPC) classifications (A61F2, A61M25, A61B5/6/8, etc.) with title keyword confirmation for manufacturing classes (Appendix A).

Company normalization uses a 29,758-entity dictionary covering subsidiaries and historical names to handle M&A complexity (Appendix B).

The gold standard $\mathcal{G}$ consists of 585 expert-verified device-patent pairs sourced from public corporate disclosures (litigation, virtual patent marking, SEC filings), covering 88 devices (20.3%). Table 1 summarizes key statistics.

device *classes* across CPC and CFR for ex-ante market exploration—a complementary but distinct task. We differ in granularity (instance-level device-to-patent linking for ex-post IP traceability), semantic gap (FDA clinical narratives vs. USPTO engineering claims; Jaccard = 0.039), and output (a complete linkage for all FDA cardiovascular devices verified on a 585-pair expert benchmark).

Table 1: Dataset and benchmark statistics from data source (device, patent documents, and company entities) to gold standard patent-device relations.

Component	Count
FDA Cardiovascular PMA Documents	434
USPTO Patents	698,191
Companies (normalized)	29,758
Gold-Standard Verified Pairs	585
Devices with Disclosures	88 (20.3%)
Patents per Device (median/max)	5 / 83

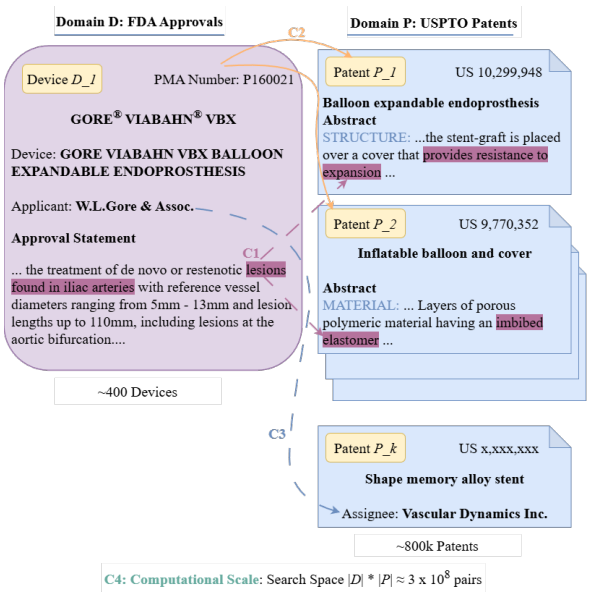

Figure 2: Overview of Cross-Domain Entity Linking challenges using GORE® VIABAHN® as a case study. Real Data Linkages: The device links to multiple patents (e.g., US 10,299,948 and US 9,770,352), demonstrating (C2) Granularity Asymmetry. The contrast between clinical indications ("lesions") and technical specifications ("resistance", "elastomer") highlights the (C1) Semantic Gap. Hypothetical Scenario: A fictional patent assignee is used to illustrate (C3) Corporate Structure challenges (M&A resolution). (C4) Scale: Represents the massive pair-wise search space.

3.3 Task Challenges

This task differs from standard entity linking in four aspects (Figure 2):

C1: Semantic Gap. FDA documents emphasize clinical outcomes (e.g., "treats coronary artery disease") while patents describe technical mechanisms (e.g., "helical coil with radiopaque markers"). Lexical overlap is minimal: average Jaccard similarity between device and patent vocabulary is 0.039.[2]

C2: Granularity Asymmetry. A single device integrates multiple patented technologies (median = 5, max = 83 patents per device), creating one-to-many alignment challenges.

C3: Corporate Structure Complexity. While only 10% of verified links involve explicit assignee-manufacturer mismatches, frequent M&A among major players requires cross-organization entity resolution.[3]

C4: Scale and Label Scarcity. With $|\mathcal{D}| \times |\mathcal{P}| \approx 3 \times 10^8$ candidate pairs, exhaustive evaluation is prohibitive, while expert-verified links cover only 88 devices (20.3%).

3.4 Evaluation Protocol

Let $\mathcal{D}$ and $\mathcal{P}$ denote the sets of devices and patents, respectively. Let $\mathcal{L}$ be the set of links returned by our pipeline. We use $\mathcal{C}$ to denote the candidate pool size at each stage.

Given the partial nature of disclosures, we adopt conservative metrics: (1) **Recall@Gold**: $|\mathcal{L} \cap \mathcal{G}|/|\mathcal{G}|$; (2) **Noise Reduction**: $(|\mathcal{C}| - |\mathcal{L}|)/|\mathcal{C}|$; (3) **FPR** (False Positive Rate): Reported specifically for LLM baseline comparisons.

4 Methodology

We propose BRIDGE-MEDDEVKG, a coarse-to-fine framework that maximizes recall while progressively reducing noise (Figure 3). Stages 1–2 construct a high-recall candidate pool using unsupervised and weakly supervised signals. Stage 3 applies learned reranking to bridge the semantic gap.

4.1 Stage 1: MedDevOnto - Domain-Specific Ontology with Anchor Terms

UMLS (Bodenreider, 2004) provides comprehensive biomedical concept normalization but assigns uniform importance to all semantic types, treating generic terms (e.g., "manufactured object") equivalently to device-critical ones (e.g., "stent"). Following middle-out ontology engineering (Uschold and Gruninger, 1996), we construct MEDDEVONTO by differentially weighting UMLS semantic types according to three domain-specific principles:

1. **Specificity**: specific device types outweigh generic categories.
2. **Discriminativeness**: terms distinguishing device families receive elevated weights.
3. **Anchor Stability**: high-frequency terms in both corpora serve as matching anchors.

Weights are assigned by UMLS semantic type and domain relevance:

- **High (1.0)**: Medical Device (T074), Therapeutic Procedure (T061), select diseases (T047).

[2]Computed over gold-standard pairs using unigram vocabularies after stopword removal.

[3]E.g., Abbott acquired St. Jude Medical (2016) and CardioMEMS (2014).

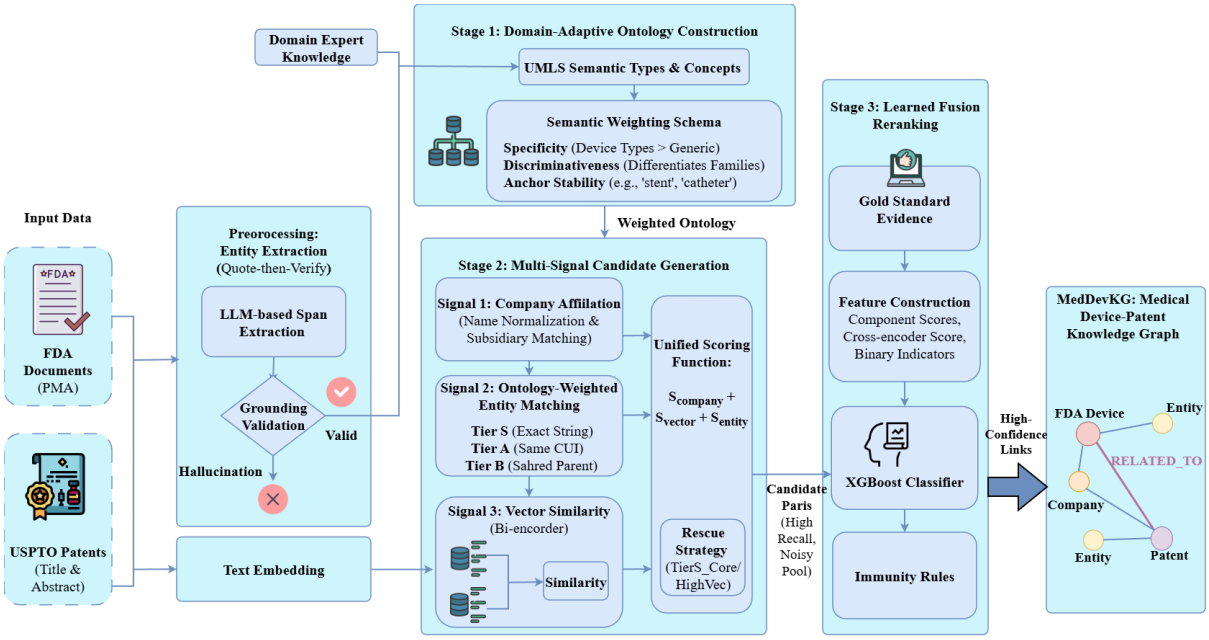

Figure 3: BRIDGE-MEDDEVKG pipeline: ontology with domain-specific anchor terms, multi-signal candidate generation, and learned reranking via cross-encoder + XGBoost.

- **Medium (0.5)**: Materials.
- **Low (0.1–0.3)**: Anatomy, generic manufactured objects.
- **Anchor terms** (stent, catheter, valve, etc.) receive weight 1.0 regardless of type.

We apply two-tier concept formalization: exact match after formalization (lowercasing, whitespace, and punctuation standardization), with fallback via syntactic head noun extraction, achieving 82.9% entity-to-UMLS mapping coverage (Appendix D).

Entity extraction and mapping use schema-constrained prompting with DeepSeek-V3, followed by Quote-then-Verify grounding (98.8% fidelity; Appendix C). This yields three-tier matching: exact string, same UMLS Concept Unique Identifier (CUI), shared parent concept.

4.2 Stage 2: Multi-Signal Candidate Generation

We fuse three complementary signals in a unified score $S(d, p)$:

$$S(d, p) = S_{\text{company}} + S_{\text{vector}} + S_{\text{entity}} \quad (1)$$

where $S_{\text{company}} \in \{0, 20\}$ is binary match using a 29,758-company dictionary; $S_{\text{vector}} \in [0, 65]$ is discretized Sentence-BERT cosine similarity; and S_{entity} is ontology-weighted overlap across three tiers (Tier S: exact string, Tier A: same CUI, Tier B: shared parent).

Pairs enter the candidate pool if they achieve a composite score $S(d, p) \geq 70$, or satisfy rescue conditions[4] for high-confidence single signals: Tier-S anchor match with core device terminology (weighted score ≥ 60), or vector similarity ≥ 0.88. Same-company pairs receive relaxed thresholds given the strong structural prior. All thresholds (company-match weight, composite score threshold, and rescue parameters) were selected on a held-out validation split comprising 20% of the devices with gold-standard labels, and were *not* tuned on the final evaluation set. This achieves 98.97% gold recall at candidate generation (579/585 verified pairs) (Appendix G).

4.3 Stage 3: Learned Noise Reduction

The candidate pool prioritizes recall but contains substantial noise. We apply heterogeneous feature fusion:

Cross-Encoder Scoring. We use BGE-M3 (BGE-reranker-v2-m3, 1024-token context) to compute deep contextual similarity:

$$S_{\text{cross}}(d, p) = \text{CrossEncoder}([d; \texttt{SEP}; p])$$

XGBoost Classification. Each candidate is represented by a 9-dimensional feature vector including all Stage-2 scores, cross-encoder signal (S_{cross}),the raw SBERT cosine similarity (sim_{raw}), and binary indicators $\mathbf{b} \in \{0, 1\}^3$. The binary vector $\mathbf{b} = [\mathbb{1}_{\text{core}}, \mathbb{1}_{\text{rescue}}, \mathbb{1}_{\text{org}}]$ captures structural and heuristic properties defined as:

[4] A rescue rule preserves below-threshold pairs when they contain strong single-signal evidence (Appendix §G.3).

$$\mathbb{1}_k = \begin{cases} 1 & \text{if condition } k \text{ is satisfied} \\ 0 & \text{otherwise} \end{cases}$$

where $\mathbb{1}_{\text{core}}$ denotes an exact match on the core device term, $\mathbb{1}_{\text{rescue}}$ indicates the pair was admitted via the high-precision rescue rule, and $\mathbb{1}_{\text{org}}$ signifies that the device and patent share the same company assignment.

The model is trained on gold positives and hard negatives (high-scoring non-gold pairs, $S \geq 70$, similarity > 0.8) with 5-fold cross-validation (F1 = 0.931).

Immunity Rules. High-confidence pairs (similarity ≥ 0.92 with company match, or Tier-S anchor match) bypass classification to prevent false negatives (Appendix F).

5 Experiments

5.1 Experimental Setup

We evaluate on the cardiovascular device-patent benchmark introduced in §3: 434 FDA PMA documents, 698,191 USPTO patents, and a gold standard of 585 verified pairs across 88 devices (Table 1).

Preprocessing. Device and patent texts are used as-is (approval summaries and patent abstracts). Entity extraction for ontology matching employs schema-constrained prompting with DeepSeek-V3, achieving 98.8% grounding fidelity via Quote-then-Verify (Appendix C). Company names are normalized using the 29,758-entity dictionary (Appendix B).

Implementation Details. For sentence embeddings, we utilized `all-mpnet-base-v2`. The cross-encoder reranker employed is `BGE-M3`, chosen for its 1024-token context capability (refer to Appendix H). XGBoost was implemented with 5-fold cross-validation, using gold positives alongside hard negatives; the probability threshold was optimized to achieve the maximum F1 score on validation, set at 0.22.

All experiments were conducted on a single A800 GPU, with candidate generation being parallelized for enhanced efficiency.

5.2 Baselines

We compare against representative methods from each major paradigm:

Retrieval Baselines We employ several retrieval baselines to evaluate the performance of the proposed methodologies. The lexical approaches utilized include Term Frequency-Inverse Document Frequency (**TF-IDF**) and **BM25** algorithms. For dense representations, we incorporate Sentence-BERT (**SBERT**) (Reimers and Gurevych, 2019) with the **`all-mpnet-base-v2`** model, as well as **BioBERT** (Lee et al., 2020) and **SapBERT** (Liu et al., 2021). Additionally, we implement a structural baseline that focuses on **company-only** matches following appropriate normalization procedures.

LLM Direct Classification We prompt state-of-the-art models to classify device-patent relevance using extracted entities and functional descriptions (structured chain-of-thought prompting; full details in Appendix E): GPT-4-turbo, GPT-4, Claude-3.5-Sonnet, Gemini-2.5-pro, DeepSeek-V3.

Cross-Encoder Baselines We also include standalone rerankers that operate independently without the integration of multi-signal fusion or XGBoost methodologies. The models under consideration include **MiniLM-L12**, **TinyBERT-L6**, **BGE-M3**, as well as large language model-based rerankers such as **Qwen3-4B** and **ERank-4B**.

6 Results

We investigate three evaluation questions:

- **EQ1**: How does our pipeline compare to retrieval and LLM baselines?
- **EQ2**: What drives the effectiveness of each component?
- **EQ3**: Does the framework generalize across device categories?

6.1 EQ1: Main Performance Comparison

Table 2 presents results for retrieval baselines and LLM direct classification compared to our pipeline, respectively.

Retrieval baselines struggle on cross-domain matching. No single retrieval method exceeds 50% R@500 on the 50K patent subset (Table 2, Group A). General-domain SBERT (39.3%) outperforms biomedical-specific models (BioBERT 35.6%, SapBERT 36.1%), suggesting that device documentation benefits from balanced engineering-clinical semantics rather than purely biomedical representations. Company matching achieves the

Table 2: **Main Benchmarking Results.** Comprehensive comparison of retrieval baselines, zero-shot LLMs, and our Bridge-MedDevKG framework on 585 expert-verified device-patent pairs. **R@Gold**: Recall on gold standard. **FPR**: False positive rate on balanced evaluation set. **Noise Red.**: Cumulative filtering ratio relative to full search space (300M pairs). Best results in **bold**. Our framework substantially outperforms retrieval and LLM baselines in recall while enabling tractable full-corpus processing.

Method Type	Model / Strategy	Links	R@10	R@100	R@Gold↑	FPR↓	Noise Red.↑
Group A: Retrieval Baselines *(50K patent subset)*							
Sparse	TF-IDF	—	5.6	16.5	35.5	—[a]	—[b]
	BM25 (Robertson et al., 2009)	—	4.7	13.3	28.8	—[a]	—[b]
Dense	BioBERT (Lee et al., 2020)	—	3.9	18.3	35.6	—[a]	—[b]
	SapBERT (Liu et al., 2021)	—	4.9	19.7	36.1	—[a]	—[b]
	SBERT (Reimers and Gurevych, 2019)	—	8.2	24.0	39.3	—[a]	—[b]
Structural	Company Name Matching	—	6.5	24.3	47.0	—[a]	—[b]
Group B: LLM Zero-shot Classifiers *(balanced evaluation set)*							
Commercial	Claude-3.5-Sonnet	—	—[e]	—[e]	29.6	11.0%	—[c]
	Gemini-2.5-Pro	—	—[e]	—[e]	50.2	10.1%	—[c]
	GPT-4 (Achiam et al., 2023)	—	—[e]	—[e]	52.4	27.1%	—[c]
	GPT-4-turbo (Achiam et al., 2023)	—	—[e]	—[e]	60.1	28.6%	—[c]
Open Source	DeepSeek-V3	—	—[e]	—[e]	23.5	**9.5%**	—[c]
Group C: Our Bridge-MedDevKG Framework *(full 698K corpus)*							
Baseline	All Candidates	∼300M	—[f]	—[f]	100.0	—	0.0%
Stage 2	Candidate Generation	13.9M	—[f]	—[f]	98.97	—[d]	95.4%
Stage 3	Full Pipeline[†]	6.8M	—[f]	—[f]	**91.61**	—[d]	**97.7%**

[†]Production configuration. Stage 3 achieves 50.9% incremental noise reduction (13.9M → 6.8M).

Metric applicability: [a]FPR requires explicit negative predictions; retrieval ranks without binary classification. [b]Noise reduction measures filtering efficiency; retrieval returns fixed top-K results. [c]LLMs evaluated on balanced sample, not full candidate space required for noise reduction. [d]Pipeline evaluated on full corpus; FPR requires same balanced set as LLM evaluation. [e]R@K metrics are for ranking tasks; LLMs perform binary classification without ranking. [f]R@K metrics are for ranking tasks; our pipeline outputs a filtered set, not a ranked list. **Design rationale:** Full-corpus evaluation for Group A requires 300M pairwise computations; Group B incurs prohibitive API costs ($> \$2,400$). Our coarse-to-fine architecture (Group C) is specifically designed to enable tractable full-corpus processing (Appendix L). Groups B and C use the same 585 gold pairs as ground truth, but differ in evaluation protocol: LLMs are evaluated as binary classifiers on a balanced pair set, whereas our pipeline filters the full candidate space. For completeness, few-shot LLM results are reported in Appendix E.

highest single-signal performance (47.0%), motivating its inclusion in multi-signal fusion. These results confirm a severe **vocabulary mismatch** between regulatory and technical documents.

LLMs exhibit strict recall-FPR tradeoff. Table 2 (Group B) shows that no LLM[5] achieves both high recall and low FPR simultaneously. GPT-4-turbo reaches the highest recall (60.1%) but with 28.6% FPR—classifying nearly one-third of negative pairs as matches. DeepSeek-V3 minimizes FPR (9.5%) but misses 76.5% of valid links. This pattern reveals that LLMs match surface terminology (e.g., both documents mention "catheter") without discriminating functional specificity (cardiac ablation vs. urinary catheter).

Structured fusion substantially outperforms end-to-end approaches. Our full pipeline achieves a conservative lower-bound R@Gold of 91.6% with 97.7% cumulative noise reduction (Table 2, Group C). This represents a +31.5% absolute improvement over the best LLM baseline (GPT-4-turbo). Table 2 (Group C) details progressive filtering: Stages 1–2 reduce the search space by 95.4% (from ∼300M to 13.9M candidates) while preserving 98.97% gold recall; Stage 3 eliminates an additional 50.9% of remaining noise, yielding 91.61% final recall—a deliberate tradeoff favoring precision for downstream regulatory applications.

6.2 EQ2: Component Analysis

We analyze contributions at two levels: candidate construction (Stages 1–2) and learned reranking (Stage 3).

Candidate Construction Ablation. Table 3 quantifies signal contributions. Vector similarity is most critical (−36.2% when removed), providing the primary semantic bridge between regulatory and technical vocabularies. Entity matching contributes +11.3%, supporting the value of device-critical anchor term identification and three-tier UMLS normalization over generic matching. Company matching adds +2.4%, while the rescue strategy recovers +1.5% edge cases with strong single-signal evidence.

[5]For completeness, few-shot LLM results are reported in Appendix E

Table 3: **Stage-2 core scoring ablation** Vector similarity dominates recall, while ontology-weighted entity overlap provides complementary gains; company and rescue mainly recover edge cases.

Configuration	R@Gold	Δ
Core scoring	70.77%	—
w/o Company	68.38%	−2.39
w/o Vector	34.53%	−36.24
w/o Entity	59.49%	−11.28
w/o Rescue	69.23%	−1.54

θ=70 + sim-rescue (sim $\geq$ 0.88).

Table 4: Learned fusion vs. rule-based thresholds. Threshold baselines use score-only admission without rescue rules.*Fixed thresholds fail to balance recall and noise, motivating learned fusion.*

Strategy	R@Gold	Noise Red.
Threshold θ=60	99.83%	0%
Threshold θ=70	69.23%	0%
Threshold θ=80	68.38%	12.3%
Threshold θ=90	68.38%	28.7%
Stage 3 (Learned)	**91.6%**	**50.9%**

Learned Fusion vs. Rule-Based Thresholds. Table 4 demonstrates why learned fusion outperforms fixed thresholds. Lowering the threshold to θ=60 achieves near-perfect recall (99.8%) but retains all noise. Raising to θ=80 or θ=90 reduces recall without proportional noise reduction (both plateau at 68.4%). In contrast, Stage 3's learned approach achieves *both* higher recall (+20.8% over the Stage-2 core scoring configuration at θ=70 with sim-rescue; Table 3) *and* substantial noise reduction (50.9%) by capturing non-linear signal interactions. For instance, a pair with moderate similarity (0.82) but strong company match and anchor term overlap may be valid, while one with high similarity (0.90) but no supporting evidence may be spurious—patterns that fixed thresholds cannot capture. Rule-based thresholds can also reduce noise, but only by sacrificing recall; Stage 3 improves this trade-off rather than replacing heuristic filtering altogether.

Cross-Encoder Selection. Table 5 compares cross-encoder architectures.[6] Here, *ms* denotes average inference latency in milliseconds per pair. Traditional cross-encoders substantially outperform LLM-based rerankers (0.795 vs. 0.401 ROC-AUC). We deploy **BGE-M3**[7] despite slightly lower ROC-AUC for two reasons: (1) higher PR-AUC (0.602 vs. 0.508), indicating superior positive-class performance in imbalanced settings; (2) 1024-token context window prevents truncation of long patent abstracts. All model differences are statistically significant ($p < 0.001$, bootstrap test).

Table 5: Cross-encoder comparison.*Traditional cross-encoders outperform LLM-based rerankers, with BGE-M3 offering the best trade-off between accuracy and context length.*

Model	ROC	PR-AUC	F1	ms
MiniLM-L12	**.795**	.508	**.590**	3.2
TinyBERT-L6	.772	**.609**	.583	4.2
BGE-M3†	.760	.602	.574	8.4
Qwen3-4B	.401	.203	.402	41.1
ERank-4B	.519	.302	.403	33.3

† Deployed model (1024-token context).

XGBoost Feature Analysis. Five-fold cross-validation yields F1 = 0.931 $\pm$ 0.012 and ROC-AUC = 0.991 $\pm$ 0.004, confirming stable generalization. Feature importance analysis reveals `score_total` (22.1%), `score_entity` (20.0%), and `is_same_company` (18.2%) as top contributors[8], while `ai_score` contributes only 4.0%. A direct feature ablation under matched retraining conditions confirms this: removing `ai_score` reduces gold recall by 1.03 percentage points, indicating a refinement rather than primary role.

Entity Weighting Validation. Table 6 validates our claim that *which* entities to match matters more than *how* to weight them. Once anchor terms are identified, specific numeric weights have a limited impact (<2% variation). Gold pairs exhibit significantly higher weighted entity overlap than random pairs (mean 47.3 vs. 12.1, $p < 0.001$, Welch's t-test), confirming that entity overlap remains discriminative for filtering.

6.3 EQ3: Generalization Across Categories

To verify robustness, we stratify the 88 devices with gold-standard links by primary clinical function. Table 7 shows recall remains stable across functionally distinct categories (89.7–92.3%), indicating that the anchor-term-based ontology layer

[6]The net contribution of the cross-encoder to the full pipeline is a +1.03% gain in gold recall (confirmed by ablating `ai_score` under matched retraining) at an inference cost of 8.4 ms per pair—a favourable cost–benefit trade-off given that cross-encoder scoring operates only on the reduced 13.9M candidate pool rather than the full 300M search space.

[7]Although TinyBERT-L6 achieves slightly higher PR-AUC, that comparison is based on shorter effective inputs and does not provide the same coverage for long, technically dense patents.

[8]Feature importance analysis reveals that semantic signals remain dominant: score_total and score_entity both exceed the company feature. This pattern suggests that the model is not driven primarily by company-structural signals from disclosed patents.

Table 6: Entity weighting sensitivity. *We observe limited variation in recall across different entity weighting strategies when anchor entities are fixed, while their effects on filtering non-gold pairs are discussed in the text.*

Strategy	R@Gold	Δ
Expert-Guided (Core=60, Other=6)	70.77%	—
Uniform High (All=60)	70.94%	+0.17
Uniform Mid (All=30)	70.94%	+0.17
Uniform Low (All=6)	69.06%	−1.71
Binary (Existence Only)	70.94%	+0.17

Table 7: Performance by device category. *Recall remains stable across major cardiovascular device types.*

Category	Gold Pairs	R@Gold
Stents	142	92.3%
Catheters	98	90.8%
Valves	65	91.5%
Pacemakers/ICDs	54	89.7%

and multi-signal fusion generalize without overfitting to specific device types.

7 Conclusion

We presented **Bridge-MedDevKG**, a framework for cross-domain entity linking between FDA-approved medical devices and USPTO patents. Our contributions include: (1) the first formalization and benchmark for device-patent linking with expert-verified evaluation pairs; (2) **MedDevOnto**, which identifies device-critical anchor terms and enables three-tier UMLS-based cross-document entity resolution; and (3) a heterogeneous signal fusion pipeline that significantly outperforms LLM baselines in both recall and noise reduction. Our experiments demonstrate that LLM direct classification struggles in zero-shot and few-shot settings on this task, validating the need for structured multi-signal approaches.

The resulting **MedDevKG** enables practical applications in recall surveillance, M&A IP discovery, and technology trajectory analysis (see Appendix J for case studies).

Limitations

Evaluation Protocol Constraints. Our evaluation relies on corporate disclosures (litigation records, SEC filings) as ground truth. Despite exhaustive manual search across all FDA PMA-approved cardiovascular devices, we could only identify 585 verifiable device-patent pairs for 88 devices. This scarcity reflects the inherent opacity of medical device IP landscapes: companies strategically disclose only a fraction of relevant patents. Consequently, pairs absent from the gold standard may still be valid associations, making our recall estimates conservative lower bounds. This also renders standard precision metrics less interpretable—predicted links not in the gold standard are not necessarily false positives.

Benchmark Scale and Scope. Our gold standard of 88 devices with 585 pairs, though it represents a reasonably comprehensive publicly verifiable collection for this task, remains limited for fine-grained statistical analysis. The 18 anchor terms and UMLS semantic type weights are calibrated specifically for cardiovascular devices; other therapeutic areas (e.g., orthopedics, neurology) exhibit distinct lexical patterns and may require domain-specific recalibration of both ontology weights and signal fusion parameters.[9]

Methodological Design Choices. Our framework incorporates heuristic components (similarity thresholds, immunity rules, rescue strategies) alongside learned models. While these pragmatic choices optimize real-world performance, they introduce hyperparameters requiring tuning for new domains. Our Stage 3 reranker is designed primarily for noise reduction rather than recall improvement; the modest contribution of cross-encoder scores (4.0% feature importance) relative to structural signals reflects this design intent—semantic similarity serves as a refinement signal atop robust entity-matching features. Additionally, our ablation on ontology weighting (Table 6) suggests that anchor term identification may matter more than specific weight assignments; minimal-supervision alternatives could improve transferability. Our sensitivity analysis and per-category results (Table 7; recall 89.7–92.3% across stents, catheters, valves, and pacemakers/ICDs) suggest that this heuristic layer is compact rather than brittle within the cardiovascular domain; nevertheless, cross-specialty validation remains future work.

Recall-Noise Trade-off. Our Stage 3 reranking trades 7.4% recall for 50.9% noise reduction, reflecting an explicit design choice. Applications requiring exhaustive coverage should use Stage 2 outputs directly (98.97% recall), accepting higher noise in exchange for near-complete retrieval.

[9]The core pipeline remains largely device-agnostic, including multi-stage candidate generation, multi-signal fusion, XGBoost reranking, and immunity-style filtering. Domain transfer mainly requires recalibration of a compact peripheral layer, including anchor terms and score thresholds, for specialties such as orthopedics or neurology.

Future Directions. Key extensions include: (1) refined evaluation metrics—among the 49 unrecovered gold pairs, expert audit reveals that 60.9% are weak associations (tangentially related patents disclosed for legal completeness), 30.4% are genuine semantic failures, and 8.7% reflect questionable gold labels; future work should develop stratified evaluation protocols that distinguish association strength levels and incorporate expert-annotated link quality scores, enabling more accurate assessment of true system recall; (2) multilingual and international expansion—our current focus on English USPTO-FDA documents leaves international regulatory-IP linking unexplored; extending to multilingual patent corpora (EPO, CNIPA, JPO) and harmonizing cross-jurisdictional device classifications would enable global IP landscape analysis; (3) cross-domain validation on orthopedics and neurology devices; (4) precision estimation via expert annotation of sampled predictions; (5) temporal M&A modeling for dynamic corporate genealogies; and (6) part-whole compositional reasoning for multi-component devices.

Ethics Statement

This research adheres to ethical standards in biomedical informatics and knowledge graph construction. All datasets are derived from publicly available, appropriately licensed sources: FDA Premarket Approval (PMA) approval statements and USPTO patent records, both accessible through official government databases. No personally identifiable information is involved, as our analysis focuses exclusively on institutional entities (companies, regulatory bodies) and technical documentation.

We develop scalable tools for cross-domain entity linking to promote transparency in the medical device ecosystem. Through systematic device-patent mapping, we aim to facilitate recall root-cause analysis, technology trajectory tracking, and intellectual property landscape understanding—applications that benefit patients, researchers, and regulatory bodies alike.

We recognize several limitations and potential risks. First, our gold standard is constructed from corporate disclosures, which may reflect strategic rather than comprehensive patent associations; thus, pairs absent from the gold standard should not be assumed incorrect. Second, the knowledge graph captures correlational relationships between devices and patents, not causal or legal dependencies—it should complement, not replace, expert judgment in regulatory or litigation contexts. Third, automated entity extraction and linking may propagate errors from source documents or introduce biases from the underlying language models.

The resulting MEDDEVKG is intended for research and analytical purposes. We caution against using it as the sole basis for legal, regulatory, or investment decisions without independent verification. Future work should incorporate additional validation mechanisms and broader stakeholder input to enhance reliability and fairness.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Philip Blair and Kfir Bar. 2022. Improving few-shot domain transfer for named entity disambiguation with pattern exploitation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6797–6810.

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

Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67.

Colleen Cunningham and David Hall. 2025. Linking medical device technologies and product markets. Working paper, May 2025.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.

Weize Kong, Swaraj Khadanga, Cheng Li, Shaleen Kumar Gupta, Mingyang Zhang, Wensong Xu, and Michael Bendersky. 2022. Multi-aspect dense retrieval. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3178–3186.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment

pretraining for biomedical entity representations. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 4228–4238.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

Eleni Partalidou, Despina Christou, and Grigorios Tsoumakas. 2022. Improving zero-shot entity retrieval through effective dense representations. In *Proceedings of the 12th Hellenic Conference on Artificial Intelligence*, pages 1–5.

Chintan O Patel and James J Cimino. 2006. Mining cross-terminology links in the umls. In *AMIA Annual Symposium Proceedings*, volume 2006, page 624.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, and 1 others. 2024. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1504–1518.

Natraj Raman, Sameena Shah, and Manuela Veloso. 2022. Structure and semantics preserving document representations. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 780–790.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, page 3982. Association for Computational Linguistics.

Shruti Rijhwani, Jiateng Xie, Graham Neubig, and Jaime Carbonell. 2019. Zero-shot neural transfer for cross-lingual entity linking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6924–6931.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Jiyun Shi, Zhimeng Yuan, Wenxuan Guo, Chen Ma, Jiehao Chen, and Meihui Zhang. 2023. Knowledge-graph-enabled biomedical entity linking: a survey. *World Wide Web*, 26(5):2593–2622.

L Siddharth, Lucienne TM Blessing, Kristin L Wood, and Jianxi Luo. 2022. Engineering knowledge graph from patent database. *Journal of Computing and Information Science in Engineering*, 22(2):021008.

Hassan Soliman. 2022. Cross-domain neural entity linking. *arXiv preprint arXiv:2210.15616*.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agents. *arXiv preprint arXiv:2304.09542*.

Mike Uschold and Michael Gruninger. 1996. Ontologies: Principles, methods and applications. *The knowledge engineering review*, 11(2):93–136.

Bo Wang, Jie Tang, Wei Fan, Songcan Chen, Zi Yang, and Yanzhu Liu. 2009. Heterogeneous cross domain ranking in latent space. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 987–996.

Jian Xu, Chao Yu, Jiawei Xu, Vetle I Torvik, Jaewoo Kang, Mujeen Sung, Min Song, Yi Bu, and Ying Ding. 2025. Pubmed knowledge graph 2.0: Connecting papers, patents, and clinical trials in biomedical science. *Scientific Data*, 12(1):1018.

Zheng Yuan, Yijia Liu, Chuanqi Tan, Songfang Huang, and Fei Huang. 2021. Improving biomedical pretrained language models with knowledge. *arXiv preprint arXiv:2104.10344*.

Yifan Zeng, Ojas Tendolkar, Raymond Baartmans, Qingyun Wu, Lizhong Chen, and Huazheng Wang. 2024. Llm-rankfusion: Mitigating intrinsic inconsistency in llm-based ranking. *arXiv preprint arXiv:2406.00231*.

Jingtao Zhan, Qingyao Ai, Yiqun Liu, Jiaxin Mao, Xiaohui Xie, Min Zhang, and Shaoping Ma. 2022. Disentangled modeling of domain and relevance for adaptable dense retrieval. *arXiv preprint arXiv:2208.05753*.

Longhui Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, Meishan Zhang, and Min Zhang. 2023. Rankinggpt: Empowering large language models in text ranking with progressive enhancement. *CoRR*.

Haoyu Zuo, Yuan Yin, and Peter Childs. 2022. Patentkg: Patent knowledge graph extraction for engineering design. *Proceedings of the Design Society*, 2:821–830.

A Data Collection and Filtering

A.1 Patent Corpus Construction

We construct the patent corpus from USPTO PatentsView (1976–October 2024) through multi-stage filtering:

Step 1: Patent Type. Retain utility and reissue patents; exclude design patents lacking technical claims.

Step 2: Assignee Type. Retain corporate / institutional assignees (USPTO codes 2, 3, 6–15); exclude individual inventors to focus on commercially relevant IP.

Step 3: CPC Classification. We filter patents using a domain-specific schema of CPC prefixes defined in Table 8. To ensure precision, candidates falling under broad Manufacturing categories (e.g., B23P, C25D) are retained only if their titles contain cardiovascular-specific keywords (e.g., "stent," "catheter," or "valve") matching our regex constraints.

Table 8: Complete CPC classification schema for cardiovascular patents.

Category	CPC Codes
Core Prostheses	A61F2 (e.g., A61F2/82 stents, A61F2/24 valves)
Catheters	A61M25 (e.g., A61M25/10 balloons)
Diagnostics	A61B5, A61B6, A61B8, A61B17, A61B34
Materials	A61L31, A61L27
Manufacturing*	B23P15, B21D53, C25D5, C23C14, C22C38

* Requires keyword matching (regex) to exclude general industrial patents.

A.2 Device Corpus Construction

Cardiovascular Filtering. We apply dual-layer filtering to FDA PMA approvals:

1. **Keyword matching**: Device name or trade name contains cardiovascular terms: *cardio, vascular, coronary, atrial, heart, stent, valve, artery, aortic, mitral, pacemaker, defibrillator, ablation, angioplasty, graft, catheter, atherectomy, embolectomy, oximeter, electrode, annuloplasty, cannula, occluder, arrhythmia*
2. **Product code matching**: FDA product codes in cardiovascular categories (DXY, LWS, NKE, PAQ, NPT, NIQ, MIH, LJP, MIP, MAJ, etc.)

Manual Exclusion. We curate an exclusion list of 29 non-cardiovascular devices that passed initial filters (Table 9).

Table 9: Non-cardiovascular device exclusions by category.

Category	Count	Example PMA
Orthopedic/Spine	13	P000028 (Cervical Cage)
Neuro/Spinal	3	P810033 (Epidural Electrode)
Gynecology	4	P010013 (Endometrial Ablation)
Dental	4	P040013 (Dental Bone Graft)
Ophthalmology	4	P080030 (Glaucoma Implant)
Other	1	P010020 (Fecal Incontinence)

A.3 Final Statistics

After filtering: 698,191 patents from 11.2M USPTO records (6.2% retention); 434 FDA PMA cardiovascular devices after excluding 29 non-cardiovascular entries.

B Company Normalization

Medical device industry M&A creates complex corporate genealogies that complicate assignee-manufacturer matching. We construct a normalization dictionary through three sources:

SEC Filings. Extract subsidiary relationships from 10-K annual reports of major device manufacturers (Medtronic, Abbott, Boston Scientific, Johnson & Johnson, Edwards Lifesciences, etc.).

Historical Records. Track corporate name changes and acquisitions. Key examples:

- Abbott ← St. Jude Medical (2016, $25B)
- Abbott ← CardioMEMS (2014)
- Medtronic ← Covidien (2015, $50B)
- Boston Scientific ← Guidant (2006, $27B)
- Endologix ← TriVascular (2016, $211M)

Abbreviation Expansion. Map common abbreviations and trading names:

- J&J → Johnson & Johnson
- BSC → Boston Scientific Corporation
- MDT → Medtronic plc

Statistics. The resulting dictionary covers 29,758 companies. Among these, 84 are *bridging companies*—organizations/companies with both USPTO patents and FDA device approvals—enabling direct manufacturer-assignee linkage.

Commercial Reference Data. We also use Compustat-based standardization records to reconcile historical naming variants across public-company filings. Because Compustat is a proprietary commercial database, the full normalization dictionary cannot be released publicly.

C Entity Extraction Details

C.1 Extraction Schema

We define a rigorous schema to bridge the linguistic gap between regulatory and technical documents. Unlike standard named entity recognition (NER), our schema explicitly incorporates software and digital functions, reflecting the complexity of modern cardiovascular devices:

- **COMPONENT (Hardware & Software)**:
  - *Physical*: Stents, catheters, balloons, sensors, valves.
  - *Digital*: AI algorithms, mapping software, control units, user interfaces.
- **MECHANISM (Clinical & Digital)**:

– *Clinical*: Ablation, angioplasty, pacing, hemostasis.
– *Functional*: Signal analysis, image reconstruction, remote monitoring.

C.2 Prompting Strategy

We employ **Schema-Guided Instruction Prompting** with DeepSeek-V3. Rather than relying solely on few-shot examples, we construct a structured system prompt containing three critical constraint layers to ensure standardization at the source:

1. In-Context Standardization. The model is instructed to perform implicit normalization during extraction. For instance, engineering descriptions like "expandable prosthesis" are mapped to "Stent," and "RF thermal heating" is mapped to "Cardiac Ablation" (as seen in the *Patent Translation Rules* section of our prompt).

2. The Anti-Super-Node Rule. To prevent the knowledge graph from degenerating into generic nodes, we enforce a specificity constraint: generic terms (e.g., "System," "Device," "Method") are forbidden unless modified (e.g., "Stent Delivery System"). This directly addresses the granularity challenge (§3, C2).

3. Domain-Specific Role Play. We use distinct system personas:
- **FDA Extractor**: Acts as a "Precision Extractor" focusing on identifying clinical indications and standardized product codes.
- **Patent Decoder**: Acts as a "Translator" to convert obfuscated engineering embodiments into standardized clinical functional terms.

C.3 Quote-then-Verify Validation

To ensure extraction fidelity, each extracted span must be grounded in the source document. We apply multi-tier matching:

1. **Exact match**: Span appears verbatim in source.
2. **Case-insensitive**: Span matches after lowercasing.
3. **Fuzzy match**: Levenshtein similarity ≥ 0.85.

Non-grounded spans are rejected. Table 10 summarizes validation results.

C.4 Final Entity Statistics

The extraction process yields:
- 21,002 unique COMPONENT entities
- 12,057 unique MECHANISM entities

Table 10: Quote-then-Verify validation statistics.

Match Type	Count	Percentage
Exact match	700,054	97.9%
Case-insensitive	1,686	0.2%
Fuzzy match	4,785	0.7%
Rejected (hallucination)	8,383	1.2%
Total validated	**706,525**	**98.8%**

- 714,907 total entity mentions across all documents
- 384-dimensional Sentence-BERT embeddings for all documents

D UMLS Mapping Details

D.1 Mapping Procedure

We map extracted entities to UMLS concepts using a two-tier strategy:

Tier 1: Exact Match. After surface normalization (lowercasing, whitespace standardization, punctuation removal), we match against 2.35M UMLS terms.

Tier 2: Head Noun Extraction. For multi-word terms failing exact match, we extract the syntactic head noun using spaCy and attempt matching. Example: "bi-directional steerable catheter" $\rightarrow$ "catheter" $\rightarrow$ C0085590.

D.2 Coverage Statistics

Table 11: UMLS mapping statistics.

Match Tier	Count	Percentage
Tier 1 (Exact)	2,999	9.2%
Tier 2 (Head Noun)	24,096	73.7%
No Match	5,580	17.1%
Total Coverage	**27,095**	**82.9%**

D.3 Semantic Type Distribution

Table 12 shows the distribution of mapped concepts across UMLS semantic types (TUIs).

Table 12: Top 10 UMLS semantic types in mapped entities.

TUI	Semantic Type	Count
T073	Manufactured Object	5,913
T074	Medical Device	5,899
T169	Functional Concept	3,355
T061	Therapeutic/Preventive Procedure	2,170
T170	Intellectual Product	1,690
T058	Health Care Activity	1,410
T080	Qualitative Concept	894
T059	Laboratory Procedure	835
T081	Quantitative Concept	758
T060	Diagnostic Procedure	663

E LLM Classification Study

E.1 Experimental Setup

Sampling. From 85 devices with gold-standard links, we construct an evaluation set stratified by:[10]

- **Gold positives**: Verified device-patent pairs from corporate disclosures
- **Hard negatives**: High-similarity pairs (sim $>$ 0.8) not in gold standard
- **Random negatives**: Randomly sampled non-gold pairs

Hard negatives are sampled from high-scoring non-gold pairs ($S \geq 70$, sim > 0.8) to challenge the classifier with difficult cases.

Entities failing Quote-then-Verify validation (1.2%) are discarded before downstream matching.

Prompt Engineering. We iteratively refine prompts through three strategies:

1. Direct yes/no classification
2. Chain-of-thought reasoning
3. Structured criteria for functional matching

All models use the final structured prompt requiring explicit reasoning about component overlap, functional alignment, and temporal plausibility.

Models Evaluated. GPT-4, GPT-4-turbo, Claude-3.5-Sonnet, Gemini-2.5-pro, Gemini-2.5-flash, Qwen2.5-72B, Qwen3-32B, GLM-4-32B, Kimi-K2, DeepSeek-V3.

E.2 Full Results

We first evaluate few-shot prompting effectiveness on a held-out balanced test set (30 gold positive pairs and 30 hard/random negative pairs, n=60, Table 13).

Table 13: Few-shot classification performance (Recall / FPR) on the held-out balanced sample.

Model	0-shot	1-shot	3-shot
GPT-4-turbo	56.7 / 6.7	66.7 / 10.0	66.7 / 13.3
DeepSeek-V3	40.0 / 0.0	53.3 / 6.7	66.7 / 6.7

Our pipeline achieves **91.6% recall** on the full gold-standard set. Even the best few-shot LLM result (66.7%) remains 24.9 percentage points below our pipeline. Note that 0-shot figures differ slightly from those in Table 2 due to the smaller sample size (n=60), but the performance gap remains consistent. This confirms that the gap primarily stems from severe cross-domain vocabulary mismatch (Jaccard similarity = 0.039) rather than suboptimal prompting.

Table 14 presents complete results (0-shot) across all evaluated models.

Table 14: LLM direct classification results (0-shot, 85 devices).

Model	Recall (%)	Error (%)	FPR (%)
GPT-4-turbo	60.1	39.9	28.6
GPT-4	52.4	47.6	27.1
Kimi-K2	54.7	45.3	20.5
Gemini-2.5-pro	50.2	49.8	10.1
Gemini-2.5-flash	45.8	54.2	12.3
Qwen2.5-72B	42.1	57.9	15.7
Qwen3-32B	38.9	61.1	14.2
GLM-4-32B	35.6	64.4	13.8
Claude-3.5-Sonnet	29.6	58.5	11.0
DeepSeek-V3	23.5	76.5	9.5

E.3 Error Analysis

Models exhibit a clear recall-FPR tradeoff: aggressive models (GPT-4-turbo) achieve higher recall but suffer from excessive false positives; conservative models (DeepSeek-V3) minimize FPR but miss the majority of valid links.

Qualitative analysis reveals that LLMs rely heavily on surface terminology matching—e.g., identifying "catheter" in both documents—without discriminating functional specificity. A "cardiac ablation catheter" and "urinary catheter" may both be classified as related to a catheter patent, despite entirely different clinical applications.

F Immunity Rules Details

Immunity rules prevent false negatives on high-confidence matches that the XGBoost classifier might incorrectly reject due to threshold conservatism:

- **High-similarity with company match** (sim $\geq$ 0.92, same company): Strong semantic alignment confirmed by organizational evidence makes false positives unlikely.
- **Tier-S anchor match**: Exact string match on domain-critical anchor terms indicates precise functional correspondence.

Ablation (Table 4) shows removing immunity rules drops recall from 91.6% to 85.27% (-6.3%) while only improving noise reduction by 4.4%.

G Threshold Selection

G.1 Gold Pair Similarity Distribution

We analyze the similarity distribution of 585 expert-verified device-patent pairs (Table 15). The distribution is concentrated in the high-similarity region,

[10] Three of the 88 disclosed devices lacked complete document representations required for structured prompt construction.

with all verified pairs exceeding 0.83 cosine similarity.

Table 15: Gold pair similarity distribution.

Statistic	Value
Mean	0.885
Std	0.021
Min / Max	0.832 / 0.940
50th percentile	0.885
75th percentile	0.898
95th percentile	0.924

G.2 Sensitivity Analysis of Stage-2 Heuristics

We examine the sensitivity of Stage-2 candidate generation to the composite score threshold θ without rescue rules. Table 16 shows that lower thresholds ($\theta \leq 65$) retain near-complete gold recall but admit substantially more candidates, whereas stricter thresholds sharply reduce recall. We therefore select $\theta = 70$ as a practical operating point for Stage 2, to be complemented by the rescue strategy described below.

Table 16: Score threshold sensitivity (no rescue).

θ	Hits	Recall
60	584	99.83%
65	584	99.83%
70	**405**	**69.23%**
75	383	65.47%
80	383	65.47%
90	237	40.51%

G.3 Rescue Strategy Contribution

To recover valid pairs with weak composite scores but strong individual signals, we apply layered rescue rules (Table 17). This design improves recall without relaxing the global admission threshold for all pairs. Using sim ≥ 0.83 directly would substantially expand the candidate set and introduce excessive noise, making downstream reranking intractable.

Table 17: Rescue rule contribution ($\theta = 70$).

Configuration	Recall	Δ
$\theta \geq 70$ only	69.23%	—
+ Entity ≥ 60 (Core)	69.23%	+0.00%
+ sim ≥ 0.88	70.77%	+1.54%
+ sim ≥ 0.83 (oracle lower bound)	100.00%	+29.23%

The empirical lower bound of 0.83 is justified by the gold-pair distribution: 100% of verified pairs have similarity above this value (min=0.832). In practice, we apply a strict rescue threshold of 0.88, while same-company pairs receive additional tolerance to cover the 0.83–0.88 interval given the strong structural prior. This strategy achieves 98.97% recall (579/585) at candidate generation and provides additional evidence that the selected Stage-2 calibration is reasonably robust.

G.4 Missed Pairs Analysis

The 6 missed gold pairs (1.03%) all have `kg_score=NULL`, indicating missing document embeddings rather than threshold failures. These correspond to two FDA documents (P990071, P990054) lacking vector representations in the original dataset.

H Cross-Encoder Evaluation Details

H.1 Evaluation Protocol

We evaluate cross-encoder models on 2,672 samples: 672 gold positives and 2,000 hard negatives (high-similarity non-gold pairs). For each model, we compute:

- **ROC-AUC**: Area under receiver operating characteristic curve
- **PR-AUC**: Area under precision-recall curve
- **Best F1**: Maximum F1 score across thresholds
- **Latency**: Average inference time per pair (ms)

H.2 Full Results

Table 18 presents complete cross-encoder comparison results.

Table 18: Complete cross-encoder comparison (2,672 samples).

Model	ROC-AUC	PR-AUC	F1	ms/pair
Traditional Cross-Encoders				
MiniLM-L12	.795	.508	.590	3.2
TinyBERT-L6	.772	.609	.583	4.2
MiniLM-L6	.736	.513	.521	2.8
BGE-reranker-large	.748	.536	.556	8.4
BGE-reranker-base	.697	.512	.502	4.9
BGE-M3†	.760	.602	.574	8.4
GTE-ModernBERT	.720	.497	.523	17.0
LLM-based Rerankers				
E2Rank-4B	.547	.302	.409	42.1
ERank-4B	.519	.302	.403	33.3
mxbai-rerank-large	.527	.276	.405	16.8
mxbai-rerank-base	.413	.213	.403	7.9
Jina-Reranker-v3	.444	.242	.404	26.1
Qwen3-Reranker-4B	.401	.203	.402	41.1

† Deployed model.

H.3 Statistical Significance

We perform bootstrap significance tests (1000 iterations) comparing all models against MiniLM-L12. All differences are statistically significant ($p < 0.001$).

H.4 Model Selection Rationale

Despite MiniLM-L12 achieving the highest ROC-AUC (0.795), we deploy BGE-M3 for production

due to its 1024-token context window. Patent abstracts frequently exceed 512 tokens; truncation degrades performance on long documents. BGE-M3's slight ROC-AUC reduction (0.760) is offset by improved handling of full-length abstracts.

I Gold Standard Construction

I.1 Data Sources

We construct the gold standard by exhaustively searching public disclosures for all 434 cardiovascular devices:

1. **Patent litigation filings**: PACER database, ITC Section 337 investigations
2. **Virtual patent marking**: Manufacturer websites listing patents covering specific products
3. **SEC filings**: 10-K risk factors, 8-K material events mentioning IP
4. **Investor presentations**: Quarterly earnings calls, analyst day materials
5. **FDA submissions**: PMA summary documents citing prior art

I.2 Coverage Statistics

- Devices searched: 434
- Devices with disclosed patents: 88 (20.3%)
- Total verified pairs: 585
- Patents per device: median = 5, max = 83

The low disclosure rate (20.3%) reflects strategic corporate practices—companies selectively disclose patents for litigation or marketing purposes rather than comprehensive IP mapping.

I.3 Quality Assurance

All pairs are derived from legally binding corporate disclosures (patent litigation filings, SEC regulatory submissions, official virtual patent marking pages). These sources carry legal accountability, providing high-confidence ground truth without requiring additional manual verification.

J Case Studies

J.1 Case A: Recall Root-Cause Analysis

The Biosense Webster Cardiac Ablation Catheter (PMA P030031) experienced multiple FDA recalls citing "bi-directional navigation catheter irrigation path" failures. Our knowledge graph links this device to four patents:

- US7377906: "Steering mechanism for bi-directional catheter" (2008)
- US7591799: "Bi-directional catheter steering control" (2009)
- US8021327: "Irrigated bi-directional catheter" (2011)
- US8348888: "Multi-directional steering apparatus" (2013)

The recall explicitly implicates the subsystem these patents protect. This enables *proactive surveillance*: querying devices sharing this patent family identifies additional inspection candidates before failures occur.

J.2 Case B: M&A-Driven IP Discovery

The Ovation Stent Graft System (PMA P120006) was originally manufactured by TriVascular Technologies. Our gold standard includes 82 associated patents. Following Endologix's 2016 acquisition ($211M), many patents transferred ownership while the FDA approval remained under the TriVascular name.

Our entity matching identifies persistent technical concepts ("endovascular graft," "polymer fill," "low-profile delivery") across corporate boundaries, automatically surfacing M&A-driven IP relationships without requiring manual corporate genealogy tracking.

J.3 Case C: Technology Trajectory Mapping

For Shockwave Medical's Intravascular Lithotripsy (IVL) System (PMA P200039), we recover 19 linked patents spanning 2014–2018:

- 2014: Foundational electrode designs (US8728091)
- 2015: Energy delivery control mechanisms (US9011463)
- 2016–2017: Catheter integration patents
- 2018: Multi-source architecture (US10039561)

This trajectory enables: (1) patent expiration forecasting for competitive intelligence, (2) R&D gap analysis identifying uncovered technical domains, and (3) prior art mapping for freedom-to-operate assessments.

K Error Analysis

The 49 unrecovered gold pairs (8.4%) exhibit: average similarity 0.848 (below 0.92 immunity threshold), average fusion probability 0.294, and 88% have company match but lack confirming signals. Expert audit categorizes:

- **Weak links (60.9%)**: Tangentially related patents disclosed for legal completeness
- **Genuine failures (30.4%)**: Semantic drift or part-whole mismatch

• **Annotation errors (8.7%)**: Incorrect gold labels
The high proportion of weak links suggests our conservative metrics underestimate true recall.

L Computational Cost

We report the computational resources required by our coarse-to-fine pipeline to support full-corpus processing. Table 19 summarizes the end-to-end cost for entity extraction and embedding-based retrieval, which together constitute the dominant computational components of the system.

Entity extraction is executed once over the full patent and medical device corpus using parallel CPU processing. Despite its computational intensity, the task is embarrassingly parallel and completes within approximately 72 hours on 40 CPU workers, incurring a total cost of around CNY 1,600. The 72-hour entity extraction is a one-time initialization cost over the 698K-document corpus; subsequent per-device inference completes in minutes.

Embedding computation and associated ablation experiments are conducted on a single A800 GPU and complete within 8 hours at a marginal cost of approximately CNY 100. Notably, subsequent stages operate only on the reduced candidate set produced by the earlier stages, avoiding full pairwise cross-encoder evaluation over the entire corpus.

In total, the full pipeline completes in under 80 hours with an estimated cost of CNY 1,700, making full-corpus evaluation tractable in contrast to exhaustive cross-encoder scoring, which would require hundreds of millions of comparisons and incur prohibitive computational and API costs.

Table 19: Computational cost breakdown. *The full pipeline is computationally tractable, avoiding exhaustive cross-encoder evaluation over hundreds of millions of pairs.*

Stage	**Hardware**	**Time**	**Cost (CNY)**
Entity extraction (698K docs)	40× CPU parallel	72 hrs	∼1,600
Embedding + Ablations	A800 GPU	8 hrs	∼100
Total	—	∼80 hrs	∼1,700