Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/



Research Paper Open Access

A Multi-Modal Entity Resolution Architecture For Product Catalog Deduplication and Attribute Enrichment Across Heterogeneous E-Commerce Sources

Sandeep Ayloo

Independent Researcher, USA. ORCID ID: https://orcid.org/0009-0006-3632-9045

Abstract
Reconciling product records that describe the same real-world item across merchant feeds, marketplace listings, and supplier catalogs remains one of the most persistent data quality problems in e-commerce systems. Heterogeneity in schema, terminology, image quality, and completeness makes naive key-based matching unreliable at scale, while the combinatorial cost of pairwise comparison across millions of listings makes exhaustive matching computationally infeasible. We propose a multi-modal entity resolution architecture that combines text-based attribute normalization, image-based visual similarity matching, and a fusion/scoring layer to resolve product-level entities across heterogeneous sources. The architecture further incorporates pricing-signal reconciliation as a corroborating matching feature and a confidence-based human-in-the-loop review mechanism for ambiguous cases that fall inside an uncertain decision band. We discuss blocking and candidate-generation strategies suited to catalog-scale deduplication, compare matching-signal types and their tradeoffs, and outline a taxonomy of blocking approaches applicable to production catalog systems. We frame this as an architectural synthesis grounded in production experience rather than a benchmarked empirical study, and discuss precision/recall tradeoffs, failure modes, and open research directions. We conclude that no single signal, textual, visual, or price-based, is sufficient in isolation, and that principled fusion combined with selective human review is the most durable path to reliable, explainable catalog deduplication at scale.

Keywords: blocking; catalog deduplication; e-commerce data quality; entity resolution; human-in-the-loop; multi-modal matching.

1. Introduction

1.1 The Catalog Deduplication Problem

Large-scale e-commerce platforms rarely operate a single, authoritative source of product data. Instead, catalog data arrives from merchant-uploaded feeds, third-party marketplace listings, syndicated supplier catalogs, and manually curated internal records, each following its own schema conventions, attribute vocabularies, image quality standards, and update cadences. A single physical product, for example, a specific model of wireless headphone in a specific color and packaging configuration, may appear dozens of times across these sources with inconsistent titles, incomplete attribute sets, differently cropped or watermarked images, and prices that vary by currency, promotional state, or bundling. Reconciling these records into a single canonical entity, and enriching that entity with the union of available attributes, is the catalog deduplication problem that motivates this work.

The problem is a specific instance of the classical record linkage and entity resolution problem first formalized by Fellegi and Sunter [1], who introduced a probabilistic framework for deciding whether two records refer to the same underlying entity based on agreement and disagreement patterns across comparable fields. Subsequent decades of research, surveyed comprehensively by Christophides et al. [2], have extended this foundation along several axes: similarity functions beyond exact or edit-distance string comparison, learned rather than hand-tuned matching functions, and scalable candidate generation for datasets too large for exhaustive pairwise comparison. Catalog deduplication inherits all of these challenges and adds two of its own:

a genuinely multi-modal evidence surface (text, images, and price), and a production requirement for explainability and human oversight that many academic entity resolution benchmarks do not emphasize.

1.2 Why Single-Signal Matching Fails at Catalog Scale

A natural first approach to catalog deduplication is to match on structured identifiers, UPC, GTIN, MPN, or manufacturer part number, where such identifiers are present, consistent, and correctly populated. In practice, this assumption fails often enough that identifier-only matching is insufficient as a sole strategy. Marketplace sellers frequently omit identifiers, reuse a single identifier across variant products, or populate the field with placeholder or incorrect values. Falling back to text-only matching over titles and descriptions is more robust to missing identifiers but is vulnerable to the opposite failure mode: near-identical wording can describe genuinely different products (for example, different storage capacities of the same phone model), while genuinely identical products can be described in wording that shares almost no lexical overlap across sources with different copywriting conventions. Image-only matching addresses some of these lexical gaps but introduces its own failure modes, discussed in Section 5.3, including stock-photo reuse across unrelated variants and near-identical packaging across different product lines.

We argue that no single evidence channel, structured identifier, free text, image, or price, is sufficient in isolation for catalog-scale deduplication, and that the practical path forward is a fusion architecture that treats each channel as a partially reliable, partially independent signal to be combined rather than substituted for one another. This argument is developed across Sections 4 through 6 and is the central architectural thesis of the paper.

1.3 Contributions

This paper makes four contributions. First, we present a five-stage multi-modal entity resolution architecture that integrates text-based attribute normalization, image-based visual similarity, price-signal reconciliation, and a learned fusion and scoring layer into a coherent production pipeline (Section 3). Second, we develop a taxonomy of blocking and candidate-generation strategies suited to catalog-scale deduplication, organized into four tiers by mechanism and computational cost (Section 7). Third, we describe a confidence-banded human-in-the-loop review mechanism that routes only genuinely ambiguous candidate pairs to reviewers, and we discuss how reviewer decisions can be fed back into both the fusion model and blocking thresholds (Section 8). Fourth, we provide an honest discussion of precision/recall tradeoffs, failure modes, and evaluation discipline for production catalog systems, explicitly disclosing where our synthesis is grounded in production-systems experience rather than benchmarked empirical results (Section 9).

The remainder of the paper is organized as follows. Section 2 situates our architecture against the broader entity resolution literature, including deep learning-based entity matching, blocking research, multi-modal and visual product matching, and human-in-the-loop approaches. Section 3 presents the overall architecture. Sections 4 through 6 detail the text, image, and fusion/price components respectively. Section 7 develops the blocking taxonomy. Section 8 presents the human-in-the-loop review mechanism. Section 9 discusses tradeoffs, failure modes, and evaluation discipline. Section 10 outlines future directions, and Section 11 concludes.

2. Background and Related Work

2.1 Foundations of Record Linkage and Entity Resolution

The probabilistic record linkage framework of Fellegi and Sunter [1] remains the conceptual ancestor of nearly all subsequent entity resolution work, including the architecture presented here. Their formulation treats matching as a classification problem over field-agreement patterns, weighting agreements and disagreements by their informativeness, an idea that reappears in our fusion layer (Section 6) as learned feature weighting rather than hand-derived likelihood ratios. Christophides et al. [2] provide a comprehensive survey of the modern entity resolution pipeline as it has evolved for big data settings, decomposing the problem into blocking, matching, and clustering stages that map closely onto the architecture we describe in Section 3, and highlighting the scalability concerns that motivate our blocking taxonomy in Section 7.

2.2 Deep Learning-Based Entity Matching

A substantial line of work has applied neural representation learning to the matching step of entity resolution. Barlaug and Gulla [3] survey neural approaches to entity matching, characterizing the shift from hand-engineered similarity features toward learned representations of records. Ebraheem et al. [4] proposed DeepER, one of the earlier deep learning architectures for entity resolution, demonstrating that distributed representations of records could outperform traditional feature-engineering approaches on structured matching tasks. Mudgal et al. [5] conducted a systematic design-space exploration of deep learning

architectures for entity matching, distinguishing matching tasks by the type of textual attributes involved and showing that architecture choice interacts meaningfully with data characteristics, a finding consistent with our own argument in Section 4.3 that attribute sparsity requires deliberate architectural accommodation rather than a one-size-fits-all embedding approach.

More recent work has moved toward pre-trained language models as the backbone for entity matching. Li et al. [6] introduced Ditto, which reframes entity matching as a sequence-pair classification problem over pre-trained transformer representations, achieving strong results by injecting domain knowledge and data augmentation into the fine-tuning process. Li et al. [7] extended this line of work with a systematic study of transformer-based entity matching, examining robustness and efficiency tradeoffs relevant to production deployment. Most recently, Peeters and Bizer [16] and Peeters, Steiner, and Bizer [17] have explored large language models directly for entity matching, including fine-tuning strategies, suggesting that the representation-learning component of our text branch (Section 4.2) is likely to continue evolving toward larger pre-trained backbones even as the surrounding architectural pattern, normalize, embed, compare, fuse, remains stable. Konda et al. [8] describe Magellan, an entity matching management system that operationalizes much of this matching research into an end-to-end workflow tool, illustrating the broader industry need for systems-level tooling around matching algorithms rather than algorithms in isolation, which is the gap our architecture is intended to address for the catalog domain specifically.

2.3 Blocking and Candidate Generation

Because pairwise comparison across all records is quadratic in catalog size and therefore infeasible beyond modest dataset sizes, blocking, the process of cheaply grouping records into candidate buckets so that full comparison is restricted to within-bucket pairs, is a prerequisite for any production entity resolution system. Papadakis et al. [9] provide a comprehensive survey of blocking and filtering techniques, organizing the field along dimensions of schema-awareness, redundancy tolerance, and scalability that inform the four-tier taxonomy we present in Section 7.2. Thirumuruganathan et al. [10] specifically examine deep learning approaches to blocking, showing that learned embedding-based blocking can improve recall relative to hand-crafted blocking keys at the cost of additional training and inference overhead, a tradeoff we revisit in our discussion of learned-embedding blocking tiers.

2.4 Multi-Modal and Visual Product Matching

Visual similarity is a natural complementary signal for product matching because product images often carry information that text descriptions omit or describe inconsistently. Yang et al. [20] describe a similar-image retrieval system for product search in e-commerce, demonstrating that visual embedding retrieval can surface relevant candidate matches even where textual descriptions diverge substantially. Bardhan and Dutta [15] examine product matching at internet scale from a systems perspective, discussing the practical constraints, latency, throughput, infrastructure cost, that shape which matching signals are feasible to deploy in production, a concern central to our own architectural framing. More [19] describes deep learning approaches to product matching in a large-scale e-commerce setting; as an industry engineering blog post rather than a peer-reviewed source, it is cited here as corroborating evidence from a deployed industry-engineering perspective, distinct from the peer-reviewed academic sources above, illustrating that the architectural concerns we describe, heterogeneous sources, noisy attributes, need for human review, are widely observed in industry practice rather than specific to any one platform. An entity-matching system based on multimodal data for two major e-commerce stores in Mexico [21] specifically evaluates the combination of text and image signals for retail entity matching, providing supporting evidence for the multi-modal fusion premise that underlies Section 6 of this paper. The ProMap datasets for product mapping [18] provide benchmark data explicitly constructed around the product-matching setting, including both textual and visual product attributes, and represent the kind of resource against which future benchmarked evaluation of architectures like ours could be conducted, a point we return to in Section 9.4 and Section 10.

2.5 Human-in-the-Loop and Active Learning for Ambiguous Matches

Entity resolution systems deployed in production rarely achieve acceptable precision and recall simultaneously through fully automated decisioning alone, particularly near the decision boundary where evidence is genuinely ambiguous. Qian, Popa, and Sen [11] propose active learning strategies for large-scale entity resolution that prioritize which candidate pairs are most valuable for a human labeler to review, directly informing the reviewer-routing logic we describe in Section 8. The same authors [12] describe SystemER, a human-in-the-loop system explicitly designed around explainability for entity resolution decisions, which motivates our own emphasis in Section 8.3 on presenting reviewers with interpretable evidence rather than an opaque confidence score alone.

2.6 Text Embeddings and Approximate Nearest-Neighbor Retrieval

Two enabling technologies underlie the text-matching and blocking components of our architecture. Reimers and Gurevych [13] introduced Sentence-BERT, a siamese network architecture for producing semantically meaningful sentence embeddings efficiently comparable via cosine similarity, which is the representative embedding approach we assume for the text branch described in Section 4.2. Malkov and Yashunin [14] introduced Hierarchical Navigable Small World graphs, an approximate nearest-neighbor search structure that makes embedding-based retrieval computationally tractable at the scale of millions of catalog items, and which underlies the learned-embedding blocking tier described in Section 7.2. Together, these two technologies make embedding-based text matching and blocking practical at catalog scale, though, as we discuss in Section 4.3 and Section 7.3, they do not eliminate the need for complementary signals and careful recall-oriented design.

3. Overall Architecture

We propose a five-stage architecture for multi-modal catalog entity resolution, illustrated in Figure 1. The stages are: (a) source ingestion and normalization, in which heterogeneous feed formats are mapped into a common internal schema and field values are normalized; (b) multi-modal feature extraction, in which text, image, and price signals are independently processed into comparable representations; (c) blocking and candidate generation, in which the space of all possible record pairs is reduced to a computationally tractable set of plausible candidates; (d) fusion and scoring, in which signals from all modalities are combined into a single match confidence score, with a branch to human-in-the-loop review for ambiguous cases; and (e) canonical catalog record output, in which matched records are merged into a single deduplicated, attribute-enriched entity.

A defining feature of the architecture is that it is not a strictly feed-forward pipeline. Decisions made by human reviewers at stage (d) are fed back into two upstream components: the fusion and scoring model, which can be periodically retrained or recalibrated using reviewer-confirmed labels as ground truth, and the blocking stage's recall thresholds, which can be adjusted when reviewers surface systematic patterns of missed candidates. This feedback structure reflects a production reality: the confidence thresholds and learned weights governing an entity resolution system are not static artifacts fixed at initial deployment, but require ongoing maintenance as catalog composition, source mix, and product categories drift over time (see Section 9.3).

Each stage of the architecture is discussed in a dedicated section: source normalization and text matching in Section 4, image-based matching in Section 5, fusion and price reconciliation in Section 6, blocking in Section 7, and human-in-the-loop review in Section 8.

Figure 1. Five-layer multi-modal entity resolution architecture with feedback loops.

4. Text-Based Attribute Normalization and Matching

4.1 Schema and Vocabulary Reconciliation

Before any similarity computation can be meaningful, incoming records must be mapped from source-specific schemas into a shared internal representation. This involves field-level schema mapping (recognizing that one source's "brand" field corresponds to another source's "manufacturer" field), unit normalization (reconciling metric and imperial units, or differing currency and pack-size conventions), and vocabulary reconciliation (mapping synonymous category and attribute terms, for example, "notebook" and "laptop", onto a shared controlled vocabulary). This normalization stage is unglamorous relative to the learned matching components described later, but in our experience it materially determines the ceiling on downstream matching quality: no similarity function, however sophisticated, can compensate for attributes that were never rendered comparable in the first place.

4.2 Embedding-Based Textual Similarity

Once normalized, product titles, descriptions, and structured attributes are encoded into dense vector representations using sentence-embedding models in the tradition of Sentence-BERT [13], which produce embeddings whose cosine similarity correlates with semantic similarity of the underlying text. This embedding-based approach is preferred over purely lexical similarity metrics (edit distance, token overlap) because it captures paraphrase and synonym relationships that lexical metrics miss, for instance, recognizing that "wireless over-ear headphones" and "Bluetooth noise-cancelling headset" may describe overlapping or identical products despite minimal token overlap. Deep entity matching approaches such as Ditto [6] and its successors [7] extend this idea by fine-tuning transformer representations directly on matching-labeled pairs

rather than relying on general-purpose sentence embeddings, which can improve precision on domain-specific vocabulary at the cost of requiring labeled training data specific to the catalog domain.

In our architecture, textual similarity is computed both at the level of the full concatenated record (title plus description plus key attributes) and at the level of individual structured attributes where available (brand, model number, size, color), because aggregate similarity can mask disagreement on a single highly discriminative attribute, two products with 95% similar free text but a different stated capacity are not the same product. This attribute-decomposed comparison is passed forward as a feature vector to the fusion layer described in Section 6, rather than being collapsed into a single text-similarity scalar prematurely.

4.3 Handling Attribute Sparsity

Catalog data is characteristically sparse: individual source feeds frequently omit structured attributes that other sources populate, and free-text descriptions vary widely in completeness and quality. A text-matching approach that assumes complete attribute coverage will systematically underperform on the substantial fraction of real catalog records with missing fields. We address this by treating attribute presence itself as an informative signal, the fusion layer is trained to appropriately discount comparisons over attributes that are missing on one or both sides of a candidate pair, rather than treating a missing attribute as a disagreement, and by allowing the free-text similarity channel to serve as a fallback source of evidence when structured attributes are unavailable. This design choice is consistent with the design-space findings of Mudgal et al. [5], who show that matching architecture choice should be sensitive to the structural characteristics of the attributes involved rather than applied uniformly.

5. Image-Based Visual Similarity Matching

5.1 Role of the Visual Channel

Product images provide evidence that is frequently complementary to, and sometimes more reliable than, textual descriptions. Packaging, color, form factor, and physical configuration are often more immediately and unambiguously conveyed by an image than by text, particularly where source feeds use inconsistent or marketing-oriented copywriting. Visual similarity is therefore incorporated into our architecture as an independent matching signal rather than as a tie-breaker applied only after text-based matching is inconclusive, consistent with the multi-modal matching evidence reported for retail settings by the Mexican e-commerce entity-matching study [21] and the internet-scale product matching discussion of Bardhan and Dutta [15].

5.2 Visual Feature Extraction

Product images are encoded into fixed-length embedding vectors using convolutional or vision-transformer feature extractors, following the general approach demonstrated for product search by Yang et al. [20] in their SIR system for similar-image retrieval. These visual embeddings are compared using standard vector similarity, and, as with the text branch, the resulting similarity score is passed to the fusion layer rather than being used as a standalone match decision. Where a source record includes multiple images (e.g., multiple angles or packaging views), we compute similarity across the available image set and retain either the maximum pairwise similarity or an aggregated representative embedding, since a single matching image among several is often stronger evidence of a true match than dissimilarity among peripheral or contextual images (e.g., lifestyle photography included alongside the primary product shot).

5.3 Known Failure Modes of the Visual Channel

The visual channel has systematic failure modes that must be understood by any team relying on it as a matching signal. Stock-photo reuse is common: unrelated sellers of generically similar products (for example, generic phone cases or standard USB cables) frequently license or reuse the same manufacturer-supplied stock photography, producing high visual similarity between images despite the underlying products being distinct SKUs, colors, or bundle configurations. Packaging-driven false positives arise when different products within the same product line share highly similar box art or branding, causing visual embeddings to cluster more tightly on packaging style than on the specific product variant. Conversely, false negatives occur when the same physical product is photographed under different lighting conditions, against different backgrounds, cropped differently, or presented as a lifestyle image rather than an isolated product shot, all of which can depress visual similarity between images of a genuinely identical item. These failure modes are the basis for the visual-matching row of Table 1 and motivate our recommendation, discussed in Section 6.1, that visual similarity be fused with rather than substituted for text-based evidence.

Table 1. Comparison of matching-signal types across strength, failure mode, cost, heterogeneity sensitivity, and fusion weighting.

Signal Type	Primary Strength	Typical Failure Mode	Computational Cost Profile	Sensitivity to Source Heterogeneity	Recommended Fusion Weighting Scenario
Text/Attribute Matching	Captures structured attribute agreement (brand, model, size) and semantic paraphrase via embeddings	Sparse or missing attributes; inconsistent vocabulary across sources; near-identical wording for distinct variants (e.g., differing capacity)	Low to moderate, embedding generation is cheap relative to image encoding, and inverted-index/ANN retrieval scales well	High, most exposed to schema and vocabulary divergence across merchant feeds	Weight upward when structured attributes are well-populated and source vocabulary is consistent; weight downward for free-text-only, sparsely attributed listings
Visual/Image Matching	Captures physical configuration, packaging, and color cues that text often omits or misdescribes	Stock-photo reuse across unrelated SKUs; packaging-driven false positives; lighting/cropping-driven false negatives	High, image encoding and storage of embeddings is more compute- and storage-intensive than text at catalog scale	Moderate, less sensitive to schema divergence but sensitive to photography-standard divergence across sources	Weight upward when text signal is weak or attributes are sparse; weight downward for categories prone to stock-photo reuse (generic accessories)
Price-Signal Matching	Corroborates candidate matches with an independent, cheaply computed signal not derived from text or image content	Legitimate price variation (promotions, currency, bundling, condition) can appear as disagreement even for true matches	Very low, numeric comparison after currency/unit normalization is computationally trivial	Moderate, sensitive to promotional and regional pricing conventions but not to textual/visual heterogeneity	Use only as a corroborating feature with limited independent weight; never as a primary or standalone matching signal

6. Fusion, Scoring, and Price-Signal Reconciliation

6.1 Why Fusion Must Be Learned, Not Fixed

Given the complementary strengths and failure modes of text, image, and price signals summarized in Table 1, the central architectural question is how to combine them into a single match confidence score. A fixed linear combination with hand-tuned weights is tempting for its simplicity and interpretability, but our experience is that the relative reliability of each signal varies systematically by product category, source pair, and attribute completeness. A fixed weighting scheme optimal for one category will be miscalibrated for another. We therefore treat fusion as a learned function over the per-signal similarity features (text similarity, attribute-level agreement flags, visual similarity, price agreement) rather than a fixed rule, consistent with the general

shift toward learned matching functions documented across the entity matching literature [3]-[7]. This learned fusion layer is the component most directly analogous to the classical Fellegi-Sunter weighting of field agreements [1], updated to operate over continuous, multi-modal similarity features rather than discrete field-agreement patterns.

6.2 Price-Signal Reconciliation as a Corroborating Feature

Price is a weak signal in isolation, identical products legitimately carry different prices across sources due to promotions, regional pricing, bundling, and condition (new versus refurbished), but it is a useful corroborating feature precisely because it is cheap to compute and largely independent of the noise sources that affect text and image signals. After currency and unit normalization, we treat price agreement (or proximity within a category- and source-appropriate tolerance band) as an additional feature into the fusion layer rather than as a gating condition. In practice, near-identical pricing modestly increases confidence in a candidate match, while substantial price divergence should reduce confidence only mildly, since large price divergence is common even between true matches under promotional conditions. We explicitly caution against overweighting price, since doing so risks systematically rejecting true matches during promotional periods and systematically accepting false matches among generically low-cost commodity items whose prices cluster tightly regardless of underlying product identity.

6.3 Confidence Score Calibration

The output of the fusion layer is a single calibrated confidence score intended to approximate the probability that a candidate pair refers to the same underlying product. Calibration, ensuring that a stated confidence of, say, 0.9 corresponds empirically to a 90% match rate among pairs scored near that value, is essential because this score is the basis for the three-way decisioning described in Section 8.1. Calibration should be periodically re-validated using reviewer-confirmed outcomes as they accumulate (see Section 8.2), since drift in source composition or catalog category mix can silently invalidate calibration performed against an earlier data distribution, a concern we revisit in Section 9.3.

7. Blocking and Candidate Generation for Catalog Scale

7.1 The Scale Problem

A catalog of even a few million listings implies a pairwise comparison space on the order of trillions of candidate pairs, which is computationally infeasible to score exhaustively through the feature extraction and fusion pipeline described in Sections 4 through 6. Blocking addresses this by cheaply partitioning records into candidate buckets using coarse attributes, tokens, or learned representations such that only within-bucket pairs are passed forward for full comparison. The central tension in blocking design, as surveyed by Papadakis et al. [9], is between recall (not missing true matches by placing them in different blocks) and reduction ratio (the degree to which the candidate space is shrunk relative to exhaustive comparison); an overly aggressive blocking scheme can silently cap the achievable recall of the entire downstream pipeline regardless of how well the fusion layer is trained.

7.2 A Blocking Taxonomy for E-Commerce Catalogs

We organize blocking approaches suited to product catalog deduplication into four tiers, summarized in Table 2, ordered roughly by increasing computational sophistication and decreasing reliance on clean structured metadata. Categorical/attribute blocking groups records by coarse structured fields such as category, brand, or manufacturer, and is the cheapest and most interpretable tier but is only as reliable as the structured metadata itself, which, as discussed in Section 4.3, is frequently sparse or inconsistent. Token/sorted-neighborhood blocking groups records by shared tokens or n-grams in normalized titles, following classical blocking-key approaches, and tolerates missing structured attributes better than categorical blocking at the cost of being more sensitive to vocabulary divergence. Learned embedding-based blocking, following the deep learning approach to blocking described by Thirumuruganathan et al. [10] and made computationally practical at scale by approximate nearest-neighbor structures such as HNSW [14], groups records by proximity in a learned embedding space and can recover matches that share neither structured attributes nor lexical tokens, at the cost of requiring trained embedding models and additional infrastructure for approximate nearest-neighbor indexing. Hybrid/ensemble blocking combines multiple blocking tiers, for instance, taking the union of candidates generated by categorical, token-based, and embedding-based blocking, to improve recall beyond any single tier, at a proportional increase in downstream comparison volume and infrastructure complexity.

Table 2. Blocking strategy taxonomy for e-commerce catalog deduplication.

Tier	Mechanism	Recall Characteristics	Computational Cost	Data/Label Requirements	Recommended Use Case
Categorical/Attribute Blocking	Group by coarse structured fields (category, brand, manufacturer)	High precision of blocks but recall capped by metadata completeness/consistency	Very low, simple key-based grouping	None beyond existing structured metadata	Sources with reliably populated, consistent structured attributes
Token/Sorted-Neighborhood Blocking	Group by shared tokens/n-grams in normalized titles, sorted-window comparison	Better tolerance of missing attributes; degraded by vocabulary/synonym divergence	Low to moderate, tokenization and sorting at catalog scale	None beyond text normalization	Free-text-heavy sources with inconsistent structured metadata
Learned Embedding-Based Blocking	Group by proximity in a learned text/image embedding space via ANN retrieval (e.g., HNSW)	Recovers matches lacking shared attributes/tokens; recall depends on embedding model quality	Moderate to high, embedding generation plus ANN index build/maintenance	Requires trained embedding models; benefits from matching-labeled data for fine-tuning	Large, heterogeneous catalogs where lexical/attribute blocking under-recalls
Hybrid/Ensemble Blocking	Union of candidates from multiple blocking tiers above	Highest recall of the four tiers by construction	Highest, proportional to sum of constituent tiers' cost plus deduplication of overlapping candidates	Union of requirements from constituent tiers	Production systems where recall is prioritized and infrastructure budget allows multi-tier operation

7.3 Recall-Oriented Design Philosophy

Because the fusion and scoring layer described in Section 6 can only ever operate on candidate pairs that blocking has surfaced, we advocate a deliberately recall-oriented design philosophy at the blocking stage: it is preferable to accept a larger, noisier candidate set and rely on the fusion layer and human review to filter false positives, than to apply an overly restrictive blocking scheme that silently discards true matches before they are ever scored. This philosophy has direct infrastructure cost implications, since a larger candidate set increases downstream feature extraction and scoring volume, and is one of the reasons blocking threshold tuning is included as an explicit feedback target from human review decisions in Sections 3 and 8.2, reviewers who observe systematic patterns of missed matches provide the clearest available evidence that blocking recall, not merely fusion-layer precision, requires adjustment.

8. Confidence-Banded Human-in-the-Loop Review

8.1 Three-Way Decisioning

The calibrated confidence score produced by the fusion layer (Section 6.3) is used to partition candidate pairs into three decision bands: pairs scoring above a high-confidence threshold are auto-merged into a canonical record without human review; pairs scoring below a low-confidence threshold are auto-rejected as non-matches and discarded; and pairs falling inside the intermediate uncertain band are routed to a human

reviewer queue rather than being forced into an automated decision. This three-way structure reflects a deliberate acknowledgment that the fusion model's confidence score is most reliable at the extremes and least reliable near the decision boundary, and that forcing a binary decision on genuinely ambiguous evidence produces errors that are more costly to detect and correct after the fact than the cost of routing such cases to a reviewer up front.

8.2 Active Learning from Reviewer Decisions

Reviewer decisions on ambiguous pairs are not merely applied to resolve the individual candidate pair in question; they also serve as newly available labeled training data for the fusion model, consistent with the active learning framing for entity resolution proposed by Qian, Popa, and Sen [11]. Because pairs routed to review are, by construction, those the current model finds most ambiguous, they are disproportionately informative training examples relative to randomly sampled labeled data, and periodic retraining incorporating these labels can be expected to narrow the uncertain band over time as the fusion model becomes better calibrated on the specific ambiguity patterns the catalog actually exhibits. This retraining loop is one of the two feedback paths in the overall architecture.

8.3 Explainability for Reviewers

A raw confidence score alone is an unsatisfying and inefficient basis for human review, since it does not indicate which evidence drove the score or where the model's uncertainty originates. Following the explainability emphasis of SystemER [12], we present reviewers with the decomposed per-signal evidence, text similarity and attribute-level agreement, visual similarity with representative image thumbnails, and price comparison, alongside the aggregate confidence score, so that reviewers can quickly identify which signal(s) are driving ambiguity (for example, high textual similarity but low visual similarity, suggesting a possible stock-photo mismatch of the kind discussed in Section 5.3) and make a more efficient and more consistent decision than would be possible from the aggregate score alone.

8.4 Review Cost and Routing Considerations

Human review capacity is finite and materially more expensive per decision than any automated stage of the pipeline, which makes the width of the uncertain decision band a first-order operational cost lever: a wider band improves precision and recall of the overall system at the cost of higher review volume, while a narrower band reduces review cost at the risk of pushing more borderline decisions into automated auto-merge or auto-reject outcomes where they are more likely to be wrong. We recommend that this band width be set and periodically re-tuned based on observed reviewer throughput, the cost asymmetry between false merges and false splits discussed in Section 9.1, and category-specific difficulty patterns discussed in Section 9.2, rather than fixed at a single global value for the entire catalog.

9. Precision/Recall Tradeoffs and Failure Modes in Production Catalog Systems

9.1 Asymmetric Costs of False Merges and False Splits

False merges (incorrectly treating two distinct products as the same entity) and false splits (failing to merge records that describe the same product) carry different downstream costs that are frequently asymmetric and category-dependent. A false merge can corrupt a canonical record with attributes, reviews, or pricing history belonging to a genuinely different product, an error that is often difficult for downstream consumers of the catalog to detect and that can propagate into customer-facing surfaces before it is caught. A false split, by contrast, typically manifests as redundant or fragmented catalog entries, an inconvenience and a data quality blemish, but one that is usually more visible and more easily corrected than a silent false merge. In our experience, this asymmetry argues for a fusion and thresholding strategy that is conservative about auto-merging (favoring the uncertain band, and therefore human review, over an aggressive auto-merge threshold) even at some cost to overall automation rate, particularly in the early period after a new blocking or fusion configuration is deployed and its true error characteristics are not yet well understood.

9.2 Category and Source Heterogeneity in Matching Difficulty

Matching difficulty is not uniform across product categories or source pairs. Categories with highly standardized, densely populated structured attributes (for example, consumer electronics with manufacturer part numbers) tend to be comparatively tractable for text-based and blocking approaches, while categories dominated by loosely structured, seller-authored descriptions (for example, handmade or apparel goods with size and style variation) are more dependent on the visual channel and correspondingly more exposed to the visual failure modes discussed in Section 5.3. Similarly, matching within a single, consistently formatted source is generally easier than matching across sources with divergent schema and photography conventions. A

production system should therefore expect, and ideally measure, matching quality separately by category and source pair rather than assuming a single global precision/recall operating point is representative of system performance across the full catalog.

9.3 Drift and Maintenance

The fusion model's learned weights, the blocking stage's recall thresholds, and the confidence score calibration described in Section 6.3 are all fit against a particular historical distribution of catalog composition, source mix, and product categories. As new sources are onboarded, new product categories are introduced, or existing sources change their feed formats or photography standards, this distribution shifts, and previously well-calibrated thresholds and weights can silently degrade. We regard this drift as an expected, ongoing maintenance burden rather than a one-time deployment risk, and the feedback loops from human review into fusion retraining and blocking threshold tuning (Sections 3, 8.2, and 8.4) are the primary mechanisms by which the architecture is intended to remain adapted to a shifting catalog over time, rather than a static system requiring periodic full re-engineering.

9.4 Data and Evaluation Discipline

We wish to be explicit and conservative about the empirical status of the claims in this paper. This work is presented as an architectural and conceptual synthesis grounded in production-systems experience and the entity resolution literature, not as a benchmarked empirical study reporting precision, recall, or F1 measurements on a specific dataset. We have deliberately not fabricated or estimated numerical performance figures for the architecture as a whole, since we have not conducted a controlled empirical evaluation against a shared benchmark such as the ProMap product-mapping datasets [18], and presenting invented numbers would misrepresent the nature of the contribution. Readers seeking quantitative benchmarking of individual components, for example, transformer-based entity matching accuracy [6][7] or visual retrieval quality [20], should consult the empirical studies cited throughout Sections 2, 4, and 5, which report their own measured results under their own experimental conditions. We view rigorous, benchmarked empirical evaluation of the full fused architecture described here, ideally against publicly available product-matching datasets, as important future work, and we discuss this directly in Section 10.

10. Future Directions

Several directions follow naturally from the conceptual synthesis presented here. First, a controlled empirical evaluation of the full fused architecture against publicly available product-matching benchmarks, including the ProMap datasets [18], would allow the qualitative tradeoffs discussed in Sections 6 and 9 to be quantified and compared against alternative fusion strategies. Second, the growing body of work applying large language models directly to entity matching [16][17] suggests that the text branch described in Section 4 may increasingly incorporate LLM-based matching or explanation generation, potentially also improving the reviewer-facing explainability discussed in Section 8.3. Third, more principled treatment of category-specific and source-specific calibration, motivated by the heterogeneity discussed in Section 9.2, could replace the single global confidence threshold assumed in Section 8.1 with a hierarchy of category- or source-aware thresholds. Fourth, extending active learning strategies for entity resolution [11] to jointly optimize which candidate pairs are routed to review and which blocking regions warrant re-sampling could tighten the feedback loop between human review and blocking threshold tuning described in Sections 7.3 and 8.4 into a more unified active-learning framework rather than two loosely coupled feedback mechanisms.

11. Conclusion

We have presented a multi-modal entity resolution architecture for product catalog deduplication and attribute enrichment across heterogeneous e-commerce sources, combining text-based attribute normalization, image-based visual similarity, price-signal reconciliation, and a learned fusion and scoring layer, together with a confidence-banded human-in-the-loop review mechanism for ambiguous cases and a recall-oriented blocking taxonomy for catalog-scale candidate generation. Throughout, our central argument has been that no single evidence channel, textual, visual, or price-based, is sufficient in isolation: each carries systematic failure modes that are best addressed not by seeking a single dominant signal but by principled, learned fusion across modalities, combined with selective, well-targeted human review at the genuine margin of uncertainty. We have been deliberately conservative about the empirical status of this contribution, framing it as an architectural and conceptual synthesis grounded in production-systems experience and the entity resolution

literature rather than a benchmarked empirical study, and we regard rigorous empirical evaluation of the full fused architecture as the most valuable next step for this line of work.

Acknowledgments
Not applicable.

Author Contributions
The author confirms sole responsibility for conceptualization, methodology, investigation, and writing of this manuscript.

Conflicts of Interest
The author declares no conflict of interest.

References

[1] I.P. Fellegi and A.B. Sunter, "A Theory for Record Linkage," Journal of the American Statistical Association, vol. 64, no. 328, pp. 1183–1210, 1969.
[2] V. Christophides, V. Efthymiou, T. Palpanas, G. Papadakis, and K. Stefanidis, "An Overview of End-to-End Entity Resolution for Big Data," ACM Computing Surveys, vol. 53, no. 6, Article 127, pp. 1–42, 2021.
[3] N. Barlaug and J.A. Gulla, "Neural Networks for Entity Matching: A Survey," arXiv:2010.11075, 2021.
[4] M. Ebraheem, S. Thirumuruganathan, S. Joty, M. Ouzzani, and M. Tamer Ozsu, "DeepER: Deep Entity Resolution," arXiv:1710.00597, 2018.
[5] S. Mudgal, H. Li, T. Rekatsinas, A. Doan, Y. Park, G. Krishnan, R. Deep, E. Arcaute, and V. Raghavendra, "Deep Learning for Entity Matching: A Design Space Exploration," in Proc. 2018 ACM SIGMOD Int. Conf. on Management of Data, 2018, pp. 19–34.
[6] Y. Li, J. Li, Y. Suhara, A. Doan, and W.-C. Tan, "Deep Entity Matching with Pre-Trained Language Models," Proceedings of the VLDB Endowment, vol. 14, no. 1, pp. 50–60, 2020. arXiv:2004.00584.
[7] Y. Li, J. Li, Y. Suhara, J. Wang, W. Hirota, and W.-C. Tan, "Effective Entity Matching with Transformers," The VLDB Journal, vol. 32, pp. 1215–1235, 2023.
[8] P. Konda, S. Das, P. Suganthan G.C., A. Doan, et al., "Magellan: Toward Building Entity Matching Management Systems," Proceedings of the VLDB Endowment, vol. 9, no. 12, pp. 1197–1208, 2016.
[9] G. Papadakis, D. Skoutas, E. Thanos, and T. Palpanas, "Blocking and Filtering Techniques for Entity Resolution: A Survey," ACM Computing Surveys, vol. 53, no. 2, Article 31, 2020.
[10] S. Thirumuruganathan, H. Li, N. Tang, M. Ouzzani, Y. Govind, D. Paulsen, G. Fung, and A. Doan, "Deep Learning for Blocking in Entity Matching: A Design Space Exploration," Proceedings of the VLDB Endowment, vol. 14, no. 11, pp. 2459–2472, 2021.
[11] K. Qian, L. Popa, and P. Sen, "Active Learning for Large-Scale Entity Resolution," in Proc. 2017 ACM CIKM, 2017, pp. 1379–1388.
[12] K. Qian, L. Popa, and P. Sen, "SystemER: A Human-in-the-Loop System for Explainable Entity Resolution," Proceedings of the VLDB Endowment, vol. 12, no. 12, pp. 1794–1797, 2019.
[13] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in Proc. 2019 EMNLP-IJCNLP, 2019, pp. 3982–3992.
[14] Y.A. Malkov and D.A. Yashunin, "Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs," IEEE TPAMI, vol. 42, no. 4, pp. 824–836, 2020.
[15] S. Bardhan and A. Dutta, "E-Commerce Product Matching at Internet Scale," in Proc. 2022 13th Int. Conf. on E-business, Management and Economics (ICEME), 2022.
[16] R. Peeters and C. Bizer, "Entity Matching using Large Language Models," arXiv:2310.11244, 2023.
[17] R. Peeters, A. Steiner, and C. Bizer, "Fine-tuning Large Language Models for Entity Matching," arXiv:2409.08185, 2024.
[18] O. Kacha, D. Fiala, et al., "ProMap: Datasets for Product Mapping in E-Commerce," arXiv:2309.06882, 2023; extended in "Extended ProMap datasets for product mapping," Electronic Commerce Research, 2024.
[19] A. More, "Product Matching in eCommerce Using Deep Learning," Walmart Global Tech Engineering Blog, 2016.
[20] C. Yang et al., "SIR: Similar Image Retrieval for Product Search in E-Commerce," arXiv:2009.13836, 2020.
[21] "An Entity-Matching System Based on Multimodal Data for Two Major E-Commerce Stores in Mexico," Mathematics (MDPI), vol. 10, no. 15, Article 2564, 2022.