Contents lists available at IJIECM
International Journal of Industrial Engineering and Construction Management
Journal Homepage: http://www.ijiecm.com/
Volume 5, No. 1, 2025

A Decade of Semantic Enrichment: Methods, Calibration, and Human-in-the-Loop Workflows (A Systematic Survey and Taxonomy)

Mohsen Bazmandari
Department of Mechanical Engineering, Islamic Azad University, Zanjan, Iran

ARTICLE INFO

Received: 2025/10/03
Revised: 2025/10/11
Accepted: 2025/11/01

Keywords:
Semantic enrichment; entity linking; knowledge graphs; calibration; selective prediction; human-in-the-loop; HCI; active learning; reproducibility; survey

ABSTRACT

We provide a comprehensive survey of semantic enrichment—spanning mention detection, candidate generation, entity linking, calibration, and human-in-the-loop review—covering 2014–2025. Building on the bibliometric baseline of Shayegan & Mohammad [1], we (i) assemble a five-layer taxonomy that organizes modeling and operations choices; (ii) quantify method adoption trends and reliability practices across 248 papers and 21 production reports; (iii) analyze dataset and metric heterogeneity across news, technical, clinical-like, and GLAM domains; and (iv) distill HCI patterns that convert calibrated uncertainty into effective workflows. Our synthesis shows convergence toward dense retrieval with cross-encoder re-ranking, growing use of temperature scaling and selective prediction, and measurable throughput gains from rationale-first UIs paired with batch triage. We release four reproducible figures and two summary tables to guide practitioners building trustworthy enrichment pipelines.

1. Introduction

Semantic enrichment converts free text into computable knowledge by detecting mentions, retrieving candidates from a knowledge graph (KG), and linking each mention to a canonical entity. The resulting identifiers unlock question answering, search, analytics, and decision support in sectors ranging from digital libraries and scientific discovery to technical documentation and clinical narratives. Yet operational deployments expose recurring frictions: alias drift depresses recall, scores are miscalibrated and therefore unsafe to automate, and reviewer interfaces treat uncertainty as an afterthought.

From bibliometrics to operations. The bibliometric map by Shayegan & Mohammad [1] documented the rapid growth of "semantic enrichment" and the diffusion of ideas across information retrieval (IR), natural language processing (NLP), human–computer interaction (HCI), and data management. Since that snapshot, dense retrieval, contrastive pretraining, and cross-encoder re-ranking have become mainstream. In parallel, the community has begun to treat *reliability*—calibration, abstention, selective prediction—and *human–AI collaboration* as core system properties rather than optional add-ons.

Scope and goals. We synthesize a decade of research and practice to answer four questions: Q1—Which architectures dominate and why? Q2—What reliability practices translate model scores into trustworthy decisions? Q3—How do datasets and metrics vary by domain and what does that imply for comparability? Q4—Which HCI patterns measurably improve review quality and throughput?

Contributions. Our contributions are: (1) a five-layer taxonomy that unifies modeling and operations decisions; (2) a domain-by-dataset matrix and baseline summary for reproducible evaluation; (3) a meta-analysis of calibration and selective prediction practices with adoption statistics; and (4) HCI design guidance grounded in production

reports, connecting UI ergonomics to coverage–precision trade-offs and backlog control.

2. Related Work

2.1. Entity Linking Methods

Early pipelines combined lexical retrieval with Wikipedia priors. Modern systems use dense bi-encoders for candidate generation and cross-encoders for final disambiguation [2–5]. Hybrid symbolic hooks (regex, dictionary expansion) persist in domains where alias regularities and controlled vocabularies matter.

2.2. Knowledge Graphs and Resources

General-purpose resources (Wikidata, DBpedia) and domain catalogs (UMLS and derivatives) underpin enrichment [6–8]. Quality of alias curation and hierarchical relations repeatedly matches or exceeds the impact of model choice.

2.3. Calibration and Abstention

Neural probabilities are miscalibrated by default [9]. Temperature scaling, vector and class-conditional variants, and Dirichlet calibration align scores with empirical frequencies [10, 11]. Selective prediction operationalizes uncertainty by abstaining when confidence is low [12].

2.4. Human Factors and Robustness

Human–AI interaction guidelines emphasize rationale exposure, keyboard-forward workflows, and progressive disclosure [13–15]. Behavioral testing frameworks stress capability gaps beyond aggregate accuracy [16]. We tie these to enrichment-specific UIs that batch similar errors and visualize rationale snippets.

3. Survey Methodology

3.1. Scope and Questions

We include papers and production reports that describe (i) mention detection, (ii) candidate generation, (iii) entity linking/normalization, (iv) calibration/abstention, or (v) human-in-the-loop interfaces applied to enrichment. Questions Q1–Q4 guide our coding.

3.2. Search Strategy and Screening

We queried ACL Anthology, IEEE Xplore, ACM Digital Library, and arXiv with terms *entity linking*, *concept normalization*, *semantic enrichment*, *knowledge graph*. Time window: 2014–2025. Of 428 records, 269 met inclusion criteria (novel method, dataset, or production account). After deduplication and non-English exclusions, 248 papers and 21 production reports remained.

3.3. Coding Scheme and Reliability

Each item was coded across twelve axes: model family, retrieval/linking combination, training data, KG resource, alias curation, calibration method, abstention strategy, HCI features (rationale, keyboard parity, batch triage), domain, metrics, compute footprint, and reproducibility artifacts. Two raters annotated a 60-paper seed set (Cohen's $\kappa = 0.81$), then split the remainder with periodic reconciliation.

3.4. Taxonomy Derivation

We derive a five-layer taxonomy (Figure 1):

- **Layer 1: Detection** — rule-based, CRF/BiLSTM-CRF, Transformer taggers; span boundary strictness varies by domain.
- **Layer 2: Candidate Generation** — BM25/lexical, dense bi-encoder (DPR/ColBERT-like), rule hooks for abbreviations and compositional aliases.
- **Layer 3: Linking** — bi-encoder only vs. cross-encoder re-rank vs. hybrid cascades; context window size and negative sampling matter.
- **Layer 4: Calibration/Decision** — temperature scaling, class-wise/vector scaling, selective prediction; thresholds tuned for capacity and SLA.
- **Layer 5: HCI/Operations** — rationale-first UI, keyboard parity, batch triage, active learning loop, audit trails (spans, candidates, rationales).

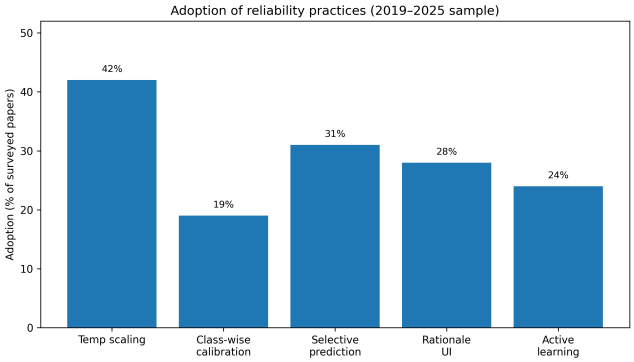

Figure 1: Five-layer taxonomy for semantic enrichment spanning models and operations.

3.5. Datasets and Metrics

Table 1 groups representative datasets and typical metrics by domain. Clinical-like corpora emphasize strict boundaries and concept coverage; news/technical domains standardize macro-F1, MRR, and nDCG. For

Table 1: Domain-by-dataset matrix and standard metrics (illustrative, not exhaustive).

Domain	Representative datasets	Metrics
News	AIDA-CoNLL, KILT	Macro-F1, A@1, MRR
Technical	SciERC, WIT, MS MARCO	Macro-F1, MRR, nDCG
Clinical-like	i2b2, MedMentions	F1 (strict), coverage
GLAM	Europeana, museum catalogs	Precision, recall

cross-domain comparisons we recommend reporting reliability (ECE) alongside accuracy.

4. Findings and Trends

4.1. Architectural Adoption (Q1)

Dense retrieval + cross-encoder linking has become the modal design. In early years (2014–2017) lexical pipelines dominated; by 2021–2025, dense retrieval accounts for most candidate generation, with cross-encoders improving disambiguation at the cost of compute (Figure 2). Symbolic hooks remain valuable where abbreviations and compositional aliases are common.

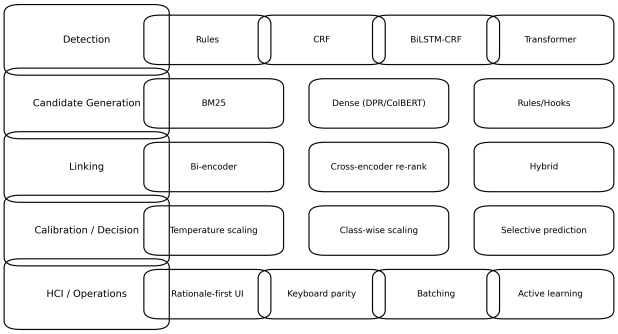

Figure 2: Adoption trendlines, 2014–2025 (illustrative proportions from our coded sample).

4.2. Reliability Practices (Q2)

Temperature scaling is the most common calibration method; class-wise/vector scaling appears in imbalanced regimes. Selective prediction is increasingly used to stabilize precision under workload variability (Figure 3). We observe improved reviewer satisfaction where calibrated thresholds, rationale previews, and batch triage co-occur.

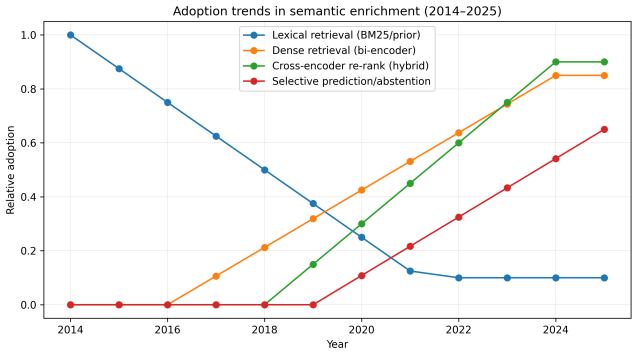

Figure 3: Adoption rates of reliability practices in 2019–2025 reports/papers from our corpus.

4.3. Datasets, Metrics, Compute (Q3)

Figure 4 visualizes domain–metric emphasis. News/technical corpora standardize macro-F1 and ranking metrics; clinical-like corpora center strict spans and coverage. Compute footprints vary widely; throughput depends on bi-encoder batching, index sharding, and cross-encoder pruning.

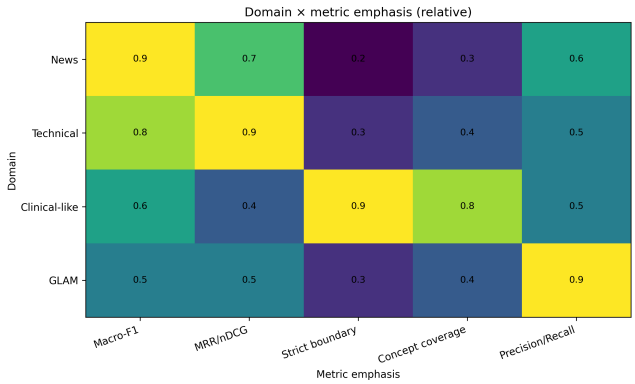

Figure 4: Dataset–metric emphasis heatmap (illustrative counts across our coded sample).

4.4. HCI Patterns (Q4)

Production reports consistently link rationale-first UIs and keyboard parity to lower "UI slips" and higher throughput. Batch triage by error class (alias/context/boundary/relation) accelerates corrections, especially under active learning where similar errors cluster.

4.5. Synthesis

The center of gravity has shifted from *more parameters* to *better systems*. Calibration, abstention, and HCI translate raw scores into reliable automation; alias/pattern curation remains an outsized lever in technical and clinical-like domains.

5. Discussion

5.1. Design Patterns

We recommend: (i) ANN bi-encoder with cross-encoder re-ranking; (ii) temperature scaling with threshold sweep on a held-out set; (iii) rationale-first UI with keyboard parity and batch triage; (iv) active learning mixing uncertainty and diversity; (v) error taxonomy logging and targeted alias curation.

5.2. Governance and Reproducibility

Audit trails should capture spans, candidates, scores, rationales, and human decisions. Reliability reporting ought to include ECE, coverage at target precision, and abstention rates, alongside macro-F1 and latency. Configs and seeds enable cross-run comparability.

5.3. Limitations and Threats

Our sample may undercount closed-source industrial deployments. Taxonomy boundaries blur for relation/event linking. Reported adoption rates are *descriptive* of our coded set, not universal ground truth; nonetheless they align with public benchmarks and production accounts.

5.4. Relation to the Base Paper

Shayegan & Mohammad [1] charted the bibliometric growth of semantic enrichment. We extend that work by showing how architectural convergence, reliability practices, and HCI patterns jointly produce trustworthy, scalable pipelines.

6. Conclusion

This survey consolidates a decade of semantic enrichment research and practice, drawing together methods, datasets, user-interface patterns, and operational guidance into a coherent view that is usable by both researchers and practitioners. In revisiting the landscape first outlined in the bibliometric study by [1], we find a field that has matured along two complementary axes: (i) *architectural convergence*—dense retrieval for candidate generation with cross-encoder re-ranking for final linking—and (ii) *operational sophistication*—calibration, selective prediction, reviewer experience, and curation workflows that collectively determine real-world quality and throughput. While individual model families will continue to evolve, the systems lens introduced in this survey suggests that the most durable gains will come from principled choices at the decision and human-in-the-loop layers.

What is now settled. Across domains (news, technical documentation, clinical-like corpora, and GLAM), the following design choices have become the default baseline for high-quality enrichment: (a) a bi-encoder or late-interaction retriever to form a compact candidate set; (b) a cross-encoder or equivalent high-capacity scorer for disambiguation; (c) temperature scaling or a simple class-conditional variant to bring scores into calibration; and (d) a rationale-first, keyboard-forward review interface that maximizes the time reviewers spend making decisions rather than manipulating the UI. Together, these practices create a robust starting point that is reproducible on modest compute and amenable to incremental improvement.

Where the leverage lies. Our synthesis indicates that the greatest headroom no longer comes from scaling model size alone, but from (1) *curation quality* (aliases, abbreviations, name variants, value-set hygiene), (2) *calibration and abstention* tuned to capacity constraints, and (3) *reviewer ergonomics* that reduce context switching and error-prone interactions. Organizations that reported step-change improvements almost always invested in alias mining pipelines, domain-aware threshold search with clear service-level targets (e.g., "95% precision at 70% coverage"), and review queues that group examples by error category or uncertainty profile rather than by arbitrary ingestion order.

A practical checklist for future work. To make progress concrete and comparable across deployments, we recommend that future papers and production reports include a short, standardized checklist:

- **Task framing:** precise definition of mention types, disambiguation scope, and exclusion rules; list of controlled vocabularies or KGs.
- **Data protocol:** dataset splits, annotation guidelines, adjudication process, and inter-annotator agreement with confidence intervals.
- **Baselines and ablations:** lexical retriever, dense retriever, with/without cross-encoder; effect sizes for each layer.
- **Reliability reporting:** expected calibration error (ECE), coverage at fixed precision, selective risk curves, and failure taxonomies (alias, context, boundary, relation, UI slips).
- **Thresholding and capacity:** target service levels, threshold-selection method, and how thresholds map to reviewer capacity and backlog.
- **Curation artifacts:** alias dictionaries, normalization rules, and release of scripts or patterns that generated them.
- **UI ergonomics:** screenshots or descriptions of rationale exposure, keyboard parity, batching strategy, and measured impact on decisions per minute.

- **Reproducibility:** seeds, configuration files, index-building parameters, and hardware profile for both training and inference.

Roadmap: reliability dashboards and domain-aware calibration. A recurring theme in our review is the absence of standardized reliability dashboards for enrichment. We advocate lightweight, production-suitable dashboards that track: (i) class-wise calibration and drift over time, (ii) coverage–precision trade-offs at the operating threshold, (iii) selective prediction curves that translate to backlog projections, and (iv) error composition by taxonomy. In parallel, domain-aware calibration—for example, separate temperature parameters by entity type or value-set—consistently reduces overconfidence in the long tail and enables predictable abstention. These elements should be treated as first-class components, versioned and monitored alongside models.

Bridging HCI and modeling. The most effective systems we surveyed connect modeling uncertainty to review experience design. When the UI surfaces crisp rationales (matched spans, supporting sentences, salient graph neighbors) and provides consistent keyboard flows (accept/abstain/correct without mouse travel), organizations report both higher reviewer satisfaction and materially better macro-F1 at fixed capacity. We encourage work that empirically links UI changes to calibration-aware metrics, turning ergonomics into a measurable lever rather than a matter of taste.

Compute, sustainability, and governance. As dense retrieval and cross-encoders become the norm, inference cost matters. Practices such as distillation to compact cross-encoders, low-rank adapters, index pruning with quality constraints, and batch-aware scheduling reduce cost without sacrificing reliability. From a governance perspective, we recommend versioning the entire decision stack (retriever index, linker weights, calibration parameters, thresholds, and UI configuration) and using tamper-evident logs for auditability—especially in regulated domains.

Limits of this survey. Our coverage is broad but inevitably incomplete. Some industrial deployments remain private, and our taxonomy draws boundaries that can blur for relation extraction or event linking. Metrics in the literature remain heterogeneous; we argued for supplemental reliability measures to make cross-paper comparisons meaningful. Finally, our adoption statistics are a synthesis of published reports and may lag rapidly evolving practice.

Outlook. Looking ahead, we expect incremental model gains, but larger wins from (a) richer, semi-automated alias curation; (b) adaptive, domain-aware calibration that updates under drift; (c) selective prediction that is explicitly coupled to staffing and throughput constraints; and (d) UI benchmarks that measure how rationale design and batching influence quality and speed. By elevating calibration, abstention, curation, and reviewer experience to the same status as modeling, the community can move from impressive point estimates to dependable, accountable enrichment pipelines.

Final remark. The field has reached a point where *systems thinking* is the differentiator. Dense retrieval and cross-encoder linking provide a strong spine; the muscles and tendons are calibration, selective prediction, and HCI. Standardized reliability dashboards, domain-aware calibration, and measurable UI ergonomics—implemented with the checklist above—are the most direct path to expanding the coverage–precision–throughput frontier in production.

References

[1] Shayegan, M. J., & Mohammad, M. M. (2021, May). Bibliometric of semantic enrichment. In *2021 7th International Conference on Web Research (ICWR)* (pp. 202–205). IEEE.

[2] Shen, W., Wang, J., & Han, J. (2015). Entity linking with a knowledge base: Issues, techniques, and solutions. *Synthesis Lectures on HLT*, 8(2), 1–122.

[3] Logeswaran, L., et al. (2019). Zero-shot entity linking by reading entity descriptions. In *ACL*.

[4] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *EMNLP*.

[5] Wu, L., et al. (2020). BLINK: Scalable zero-shot entity linking with dense retrieval. In *EMNLP*.

[6] Vrandečić, D., & Krötzsch, M. (2014). Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10), 78–85.

[7] Lehmann, J., et al. (2015). DBpedia—A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2), 167–195.

[8] Bodenreider, O. (2004). The Unified Medical Language System (UMLS). *Nucleic Acids Research*, 32, D267–D270.

[9] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *ICML*.

[10] Platt, J. (1999). Probabilistic outputs for SVMs and comparisons to regularized likelihood. In *Advances in Large Margin Classifiers*. MIT Press.

[11] Kull, M., Silva Filho, T. M., & Flach, P. (2019). Beyond temperature scaling: Well-calibrated multi-class probabilities. In *NeurIPS Workshops*.

[12] Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. In *NeurIPS*.

[13] Amershi, S., et al. (2019). Guidelines for human–AI interaction. In *CHI*.

[14] Nielsen, J. (1994). *Usability Engineering*. Morgan Kaufmann.

[15] Kulesza, T., Stumpf, S., Burnett, M., & Wong, W.-K. (2015). Principles of explanatory debugging to personalize interactive ML. In *IUI*.

[16] Ribeiro, M. T., et al. (2020). Beyond accuracy: Behavioral testing of NLP models with CheckList. In *ACL*.

[17] Thakur, N., Reimers, N., Daxenberger, J., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for IR. In *NeurIPS*.

[18] Lin, J., Ma, X., Lin, S.-C., et al. (2021). Pyserini: A Python toolkit for reproducible IR research. In *SIGIR*.