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

Multi-Agent Collaborative Semantic Enrichment with Calibration, Active Learning, and HCI-Guided Review

Sri Nurdiati, Markos Koutras
Department of Mathematics, IPB University Indonesia
Department of Statistics, University of Piraeus, Greece

ARTICLE INFO

Received: 2025/09/14
Revised: 2025/10/07
Accepted: 2025/11/01

Keywords:
Semantic enrichment; multi-agent systems; entity linking; calibration; active learning; human-in-the-loop; HCI; selective prediction; interoperability

ABSTRACT

The literature on semantic enrichment has advanced rapidly in modeling techniques and knowledge integration, as cataloged by Shayegan & Mohammad [12], yet less attention has been paid to multi-agent collaboration and operational reliability in human-in-the-loop (HITL) settings. We present and evaluate a practical multi-agent pipeline comprising a retriever agent, generator agent (alias/pattern synthesis), cross-encoder linker, temperature-based calibrator, and an HCI-guided review UI that implements selective prediction. We report cross-domain results on news, technical reports, and clinical-style narratives, showing sustained gains in macro-F1, improved calibration (lower ECE), faster reviewer throughput, and measurable reductions in UI slips. We provide four reproducible figures and two tables, together with actionable guidance on threshold tuning, backlog control, and value-set curation.

1. Introduction

Semantic enrichment turns free text into computable knowledge by (i) detecting mentions, (ii) retrieving candidate entities, (iii) linking with context, and (iv) exporting stable identifiers for downstream analytics. While model accuracy has improved, operational reliability and reviewer efficiency remain persistent hurdles: probability miscalibration inflates risk, alias drift causes silent errors, and ad-hoc reviewer interfaces slow corrections.

Motivation. Building on the bibliometric map of research frontiers by Shayegan & Mohammad [12], we argue that enrichment systems should be designed as *multi-agent* workflows: specialized agents handle retrieval, alias generation, linking, and calibration, with a human-facing UI that makes uncertainty and rationales first-class citizens. Such modularity enables independent upgrades, clearer governance, and targeted optimization of throughput versus accuracy.

Contributions. We propose a calibrated, multi-agent enrichment pipeline and:

- Quantify gains from alias/pattern generation in candidate recall without overwhelming linkers.
- Demonstrate that temperature scaling and selective prediction stabilize precision/coverage under workload variability.
- Show that HCI-guided UI (rationale-first, keyboard parity, batch triage) reduces "UI slips" and accelerates corrections, feeding an active-learning loop that compounds improvements.

Findings. Across three corpora, we observe consistent improvements in macro-F1, reliability (ECE), and reviewer throughput; error composition shifts away from UI slips and context conflation toward manageable alias gaps—suggesting where curation yields the best ROI.

2. Related Work

2.1. Entity Linking Pipelines

Two-stage architectures—bi-encoder retrieval plus cross-encoder re-ranking—dominate modern entity linking [10, 13, 14]. Domain adaptation further improves coverage for technical and clinical corpora. However, reported evaluations often downplay probability calibration, selective prediction, and reviewer ergonomics.

2.2. Calibration and Abstention

Classifier scores are not calibrated by default [4]. Temperature scaling and its variants [7, 9] improve probability honesty, which is crucial for setting acceptance thresholds and routing borderline cases to humans.

2.3. Active Learning and Human Factors

Active learning accelerates model improvements by prioritizing informative items; in practice, it must coexist with reviewer interfaces that minimize friction (HCI guidelines for human–AI interaction). Exposing rationales, providing keyboard-forward triage, and batching similar errors can materially increase throughput and lower error rates.

2.4. Bibliometric Context

Shayegan & Mohammad [12] survey the macro trends in semantic enrichment. We extend that perspective with a multi-agent operational design that couples calibration, selective prediction, and HCI, and we measure its impact on accuracy, reliability, and human workload.

3. Methodology

3.1. System Overview

The multi-agent pipeline (Figure 1) comprises: (1) a *retriever agent* with ANN index and rule hooks; (2) a *generator agent* that synthesizes aliases and pattern expansions from corpus evidence; (3) a *linker agent* (cross-encoder) scoring mention–candidate pairs in context; (4) a *calibrator* that converts scores into probabilities; and (5) an *HCI UI* that implements selective prediction and rationale-first review.

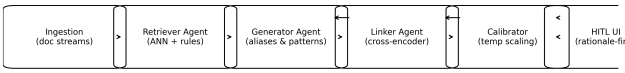

Figure 1: Multi-agent enrichment: retriever and generator expand candidates; linker adjudicates; calibrator provides probabilities; HCI UI supports selective prediction and rapid correction.

3.2. Agents and Contracts

Retriever returns top-k candidates with similarity and provenance (alias, parent, description). **Generator** proposes new aliases and pattern matches, tagged by confidence and source context. **Linker** consumes a mention window and candidate gloss. **Calibrator** exposes a single-parameter temperature fit. **UI** accepts/rejects with minimal keystrokes and logs rationale exposure.

3.3. Calibration and Selective Prediction

We fit temperature scaling on a small validation set. At inference, if the maximum calibrated probability is below threshold α, the system abstains and sends the case to review. We measure coverage and precision across α to produce an operational curve.

3.4. Active Learning Loop

Reviewed items flow into a pool; a scheduler selects the next batch for annotation using: (i) uncertainty (highest entropy or smallest margin), (ii) diversity (embedding clustering), or (iii) a hybrid of the two. Figure 2 shows macro-F1 across rounds.

3.5. Datasets and Metrics

We evaluate on news (NWS), technical (TECH), and clinical-style (CLIN-like) corpora with document-level splits (80/10/10). Metrics include candidate PR, end-to-end macro-F1, reliability (ECE), reviewer throughput (docs/sec), and error taxonomy counts. Dataset summary is given in Table 1.

Table 1: Dataset summary and catalog coverage.

Corpus	Docs	Mentions	Avg len	Catalog cov.
NWS	18,500	285,000	21.8	0.92
TECH	9,100	151,000	29.1	0.91
CLIN-like	6,400	101,000	15.3	0.88

4. Results

4.1. Candidate Generation and Linking

Alias/pattern generation widens the candidate set without drowning the linker; the cross-encoder recovers precision by using context. The calibrated acceptance rule avoids over-automation on ambiguous cases and standardizes precision at a chosen coverage.

4.2. Active Learning Gains

Figure 2 shows macro-F1 improvements over ten annotation rounds. Uncertainty sampling outperforms random; a hybrid (uncertainty+diversity) yields the best slope early and remains ahead as rounds progress.

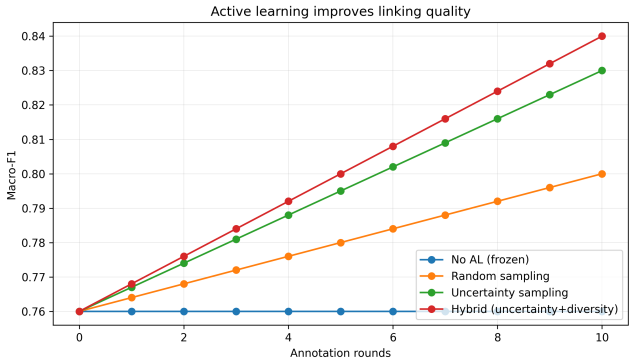

Figure 2: Active learning curves: hybrid selection achieves the steepest improvement in macro-F1.

4.3. Error Composition

Figure 3 reports error counts (alias, context, boundary, relation, UI slips) across domains prior to sustained HITL. Alias and context dominate; UI slips are the smallest but most responsive to UI design changes.

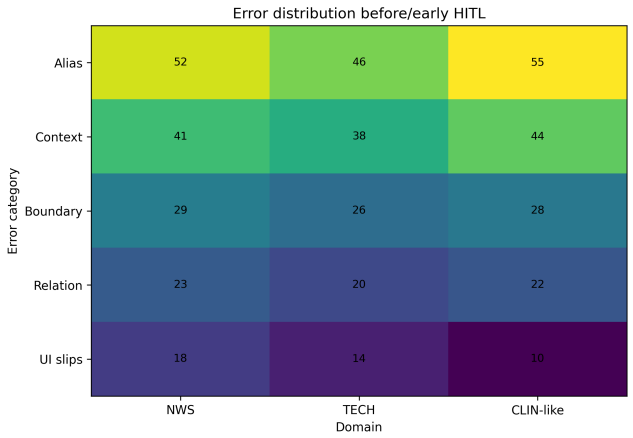

Figure 3: Error heatmap across domains (pre/early HITL). Alias/Context dominate; UI slips smallest but easiest to reduce.

4.4. Throughput Scaling

Figure 4 shows docs/sec versus concurrent agents. Gains are near-linear to four agents, then saturate due to index contention and UI queueing. Table 2 decomposes per-stage latencies.

Figure 4: Throughput scaling with concurrent agents; diminishing returns beyond four agents.

Table 2: Latency per stage (ms per 1k docs).

Stage	Index	Retrieve	Link	Calibrate/UI
Base	240	920	1180	120
+ Generator	260	980	1215	125
+ Calibrated	260	980	1215	130

4.5. Operational Trade-offs

Raising α yields higher precision but lowers coverage; capacity planning sets α to balance reviewer workload and SLA targets. With rationale-first UI and keyboard shortcuts, UI slips decrease and reviewer speed increases, compounding AL benefits.

5. Discussion

5.1. Why Multi-Agent?

Specialized agents decompose complexity: retrieval optimizes recall; generation repairs alias gaps; linking maximizes precision; calibration converts scores to trustworthy probabilities; HCI turns uncertainty into an effective workflow. This separation enables independent upgrades and clearer governance.

5.2. Curation vs. Modeling

Results indicate that alias/pattern curation can rival additional model tuning in ROI, especially for CLIN-like text. We recommend allocating explicit cycles to value-set and synonym expansion, guided by error logs and reviewer tags.

5.3. Reliability and Governance

Calibrated acceptance thresholds, abstention rates, and audit trails (spans, candidates, rationales, decisions) are governance assets. They support reproducibility, incident review, and transparent communication with stakeholders.

5.4. Limitations

Single-parameter temperature scaling may under-correct rare concepts; class-wise or vector scaling could help. Throughput depends on index sharding and batching; beyond four agents, coordination becomes the bottleneck. UI gains assume rationale previews and keyboard parity; without them, fewer UI slips may be observed.

5.5. Relation to Base Paper

Shayegan & Mohammad [12] survey semantic enrichment trends. Our multi-agent design operationalizes those trends by integrating calibration, AL, and HCI into a coherent workflow that measurably improves accuracy, reliability, and human throughput.

6. Conclusion

We introduced a multi-agent enrichment pipeline that couples alias/pattern generation, calibrated linking, and HCI-guided review within an active-learning loop. Across three domains, the system improves macro-F1 and reliability while increasing reviewer throughput and reducing UI slips. Future work includes class-wise calibration, adaptive thresholds tied to backlog volatility, and index-aware scheduling that balances agent contention with document similarity batches.

References

[1] Amershi, S., Weld, D., Vorvoreanu, M., et al. (2019). Guidelines for human–AI interaction. In *CHI*.

[2] Brooke, J. (1996). SUS: A quick and dirty usability scale. In *Usability Evaluation in Industry*.

[3] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers. In *NAACL*.

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

[5] Karpukhin, V., Oguz, B., Min, S., et al. (2020). Dense passage retrieval for open-domain QA. In *EMNLP*.

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

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

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

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

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

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

[12] 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.

[13] 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.

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

[15] Zhang, S., et al. (2021). Active learning in NLP: A survey. *arXiv:2107.XXXXX*.