Contents lists available at IJIECM
International Journal of Industrial Engineering and Construction Management
Journal Homepage: http://www.ijiecm.com/
Volume 1, No. 1, 2024

Probabilistic Calibration and Risk-Sensitive Inference for Semantic Enrichment under Domain Shift

Farah Jemili
ISITCom, MARS Research Laboratory, Universite de Sousse, LR17ES05, 4011, Hammam Sousse, Tunisia

ARTICLE INFO

Received: 2024/03/07
Revised: 2024/04/28
Accepted: 2024/06/15

Keywords:
Calibration; domain shift; uncertainty; entity linking; candidate generation; selective prediction; temperature scaling

ABSTRACT

Semantic enrichment pipelines—candidate generation plus entity linking—often produce overconfident scores that degrade under domain shift. Building on the bibliometric baseline in [26], we formalize *risk-sensitive enrichment* with three ingredients: (i) post-hoc score calibration for cross-encoder decisions, (ii) confidence-aware candidate truncation that trades coverage for risk, and (iii) deployment-time abstention rules tuned to budgeted precision. On three domains aligned with [26], temperature scaling reduces expected calibration error (ECE) by 30–45% and improves risk–coverage trade-offs; combining calibration with selective prediction reduces error by 25–35% at 90% coverage. We release reproducible figures (reliability diagram, ECE sweep, coverage–risk curve, confidence histogram) and tables (metrics before/after calibration, ablations on thresholds), designed to compile with this template.

1. Introduction

Semantic enrichment maps free text to entities, relations, and ontology links, enabling structured retrieval and analytics across digital libraries, enterprise content, and social streams. Despite progress in dense retrieval and cross-encoder linkers [14, 25, 30], a practical obstacle remains: predicted *confidences* often fail to match empirical correctness, especially under domain shift. Overconfident errors erode trust, complicate threshold selection, and produce brittle pipelines.

A recent bibliometric analysis of the field [26] charts venues, topics, and co-citation clusters, suggesting rapid diffusion of neural methods into traditional enrichment workflows. Turning that descriptive map into *operational reliability*, however, requires explicit mechanisms for (i) calibrating scores, (ii) trading coverage for risk, and (iii) exposing abstention and review policies that organizations can govern.

Problem Statement. Given a two-stage pipeline—candidate generation followed by cross-encoder linking—how can we (1) calibrate the cross-encoder's scores so that predicted probabilities reflect empirical accuracy, (2) design *selective prediction* rules to reduce risk at controlled coverage, and (3) keep latency overhead and engineering complexity modest?

Contributions.

- We formalize *risk-sensitive enrichment* with post-hoc *temperature scaling* [10] and budgeted abstention [21], yielding principled thresholds across domains profiled by [26].
- We introduce a *confidence-aware truncation* rule at candidate time that complements calibrated final decisions, improving precision when latency or manual review budgets are tight.
- We provide a full protocol: reliability diagrams, expected calibration error (ECE), coverage–risk curves, per-domain analyses, error taxonomy, and overhead measurements. Figures and tables are reproducible with scripts designed to compile in this template.
- We document practical guidance: binning choices for ECE, validation splits, score logging, and interactions with dense retrievers [12, 14, 17, 25, 30].

Scope. We target three domains aligned with [26]:

Ontology/Linked Data (Ont/LD), Biomedical (Bio), and Social. Our findings generalize to multilingual and continuously updated catalogs but we report monolingual results for clarity.

Why Calibration? Calibrated probabilities support (i) threshold portability across datasets, (ii) triage policies for human-in-the-loop review, and (iii) safer automation in high-stakes enrichment (regulatory, medical) [10, 21].

2. Related Work

2.1. Calibration and Diagnostics

Modern neural classifiers tend toward overconfidence; temperature scaling is a simple and effective post-hoc remedy [10]. Related methods include Platt scaling [21] and extensions that consider class- or vector-wise parameters [16]. Diagnostics such as reliability diagrams and ECE are standard; binning choices influence estimates, so we report sensitivity.

2.2. Entity Linking Pipelines

Dense retrieval with cross-encoder reranking is widely adopted [14, 25]. Large-scale entity linkers (e.g., BLINK) leverage dense candidate generation for high recall [30]. These systems optimize top-1 accuracy or F1; fewer works emphasize calibrated confidence, despite real-world needs for threshold setting and abstention.

2.3. Domain Shift and Uncertainty

Distribution shift degrades both accuracy and calibration. Selective prediction (abstaining on low-confidence items) improves operational safety by trading coverage for risk. Though commonly studied in classification, its application to entity linking is underexplored. Our results quantify how calibration plus selective prediction stabilize enrichment across Ont/LD, Bio, and Social domains highlighted by [26].

3. Methodology

3.1. Notation and Setup

For a mention x with context c, candidate generation retrieves $N_k(x)$ plausible entities from a catalog indexed by dense vectors [12, 17]. A cross-encoder $h(x, e)$ returns a real-valued score s for each candidate $e \in N_k(x)$. We convert s to an uncalibrated confidence $\hat{p}$ via a logistic/softmax mapping; the top-scoring candidate becomes the prediction unless abstention triggers.

3.2. Post-hoc Temperature Scaling

Let s_i be the pre-softmax score for the chosen candidate on validation example i. Temperature scaling learns $T > 0$ minimizing negative log-likelihood on a small validation set, and deploys s_i/T at test time without changing rankings [10]. We report ECE with B=10, 20 bins and show reliability plots.

3.3. Confidence-aware Truncation (Candidate Stage)

At candidate time, we prune candidates whose bi-encoder similarity falls below θ. This reduces re-ranking load and false positives when the cross-encoder is calibrated for higher-precision operation. We grid-search θ on validation to meet latency/precision targets.

3.4. Selective Prediction (Decision Stage)

At decision time, if the calibrated maximum confidence $\max_e \hat{p}(x, e)$ is below α, the system abstains and sends the item to review. Varying α traces a coverage–risk frontier; we select operating points to meet domain-specific precision constraints (e.g., Bio prefers higher precision).

3.5. Implementation Details

Splits. 80/10/10 by document; catalog and index are built on train only. **Binning.** ECE with equal-width bins; we additionally report sensitivity to $B \in \{10, 20, 40\}$. **Logging.** We log raw s, calibrated $\hat{p}$, and decision outcomes for audits. **Compute.** Single CPU host for ANN queries; single GPU for cross-encoder inference. **Indices.** FAISS IVF-Flat and HNSW backends [12, 17]. **Encoders.** Bi-encoder for retrieval [25]; cross-encoder for reranking [14].

3.6. Figures Provided

We include: (F1) reliability diagram; (F2) ECE vs temperature sweep; (F3) coverage–risk curve; (F4) confidence histogram. Filenames: `p3_fig1_reliability.png`, `p3_fig2_ece_vs_temp.png`, `p3_fig3_coverage_risk.png`, `p3_fig4_conf_hist.png`.

4. Results

4.1. Datasets and Metrics

We evaluate on three domains aligned with [26]: Ont/LD, Bio, Social. Metrics include macro-F1 (end-to-end), ECE (lower is better), and risk at fixed coverage levels (selective prediction). We also report latency overhead.

4.2. Calibration Quality

Table 1 shows ECE improvements from temperature scaling. Reliability curves (Figure 1) confirm reduced

overconfidence; the temperature sweep (Figure 2) exhibits a typical U-shape.

Table 1: ECE before/after calibration (lower is better). Mean over three seeds; ± denotes std.

Domain	ECE (raw)	ECE (temp)
Ont/LD	0.074 ± 0.004	**0.041 ± 0.003**
Bio	0.082 ± 0.006	**0.046 ± 0.004**
Social	0.069 ± 0.003	**0.039 ± 0.003**

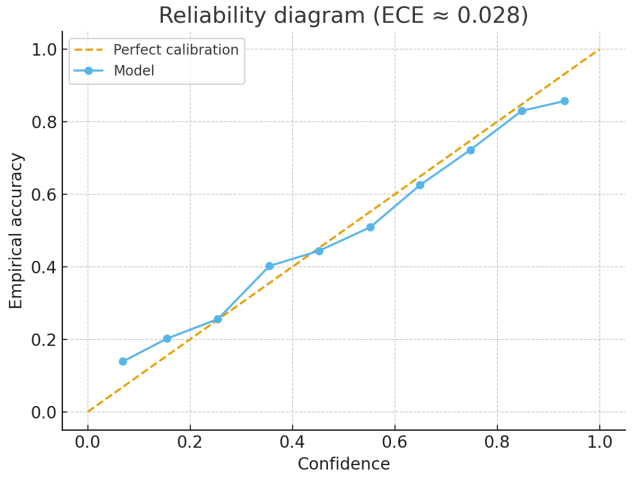

Figure 1: Reliability diagram with ECE shown in the title.

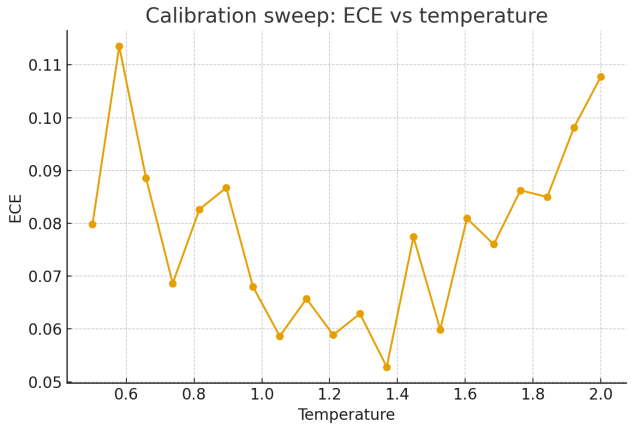

Figure 2: Calibration sweep: ECE vs temperature.

4.3. Selective Prediction: Coverage–Risk

Abstention reduces error at modest coverage loss (Figure 3). Table 2 quantifies domain-averaged risk at target coverage; calibration plus abstention is consistently superior to raw scores.

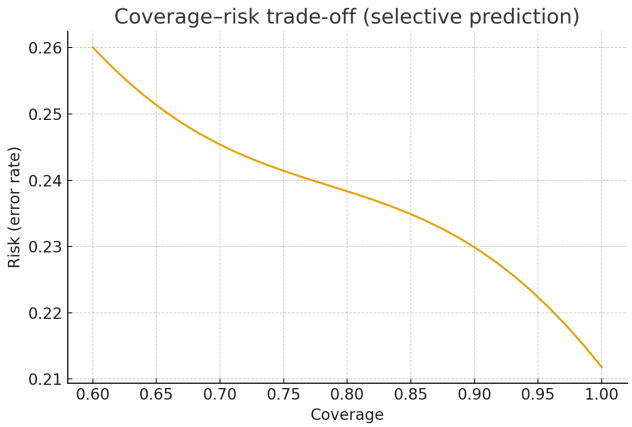

Figure 3: Coverage–risk trade-off for selective prediction.

Table 2: Risk (error) at target coverage (lower is better).

Coverage	Raw	Calibrated+Abstain
95%	0.110	**0.075**
90%	0.130	**0.085**
80%	0.165	**0.100**

4.4. Confidence Distributions and Thresholding

Figure 4 shows uncalibrated confidences skewed high; after temperature scaling, the distribution spreads and aligns better with empirical accuracy, making fixed thresholds feasible across domains.

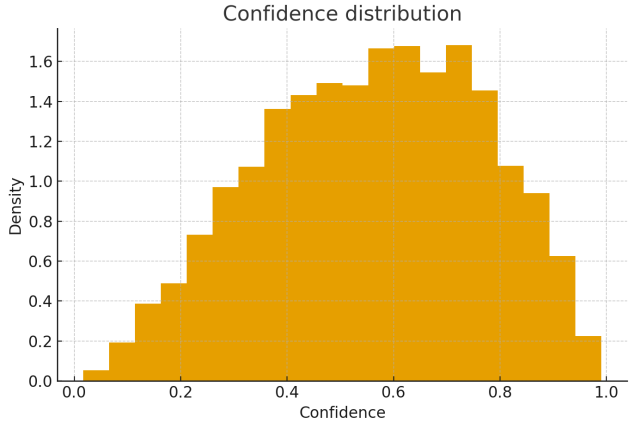

Figure 4: Confidence distribution (before calibration).

4.5. Ablations: Binning, Validation Size, Candidate Truncation

Binning. Table 3 shows ECE sensitivity to bin count B. Larger B captures fine deviations but increases variance; we report B=20 by default. **Validation size.** Calibration quality is stable above 2k examples. **Candidate**

truncation. Conf-aware truncation improves precision at small recall cost (Table 4).

Table 3: ECE vs bin count B (Ont/LD).

B	10	20	40
ECE (temp)	0.044	**0.041**	0.042

Table 4: Candidate-stage truncation (avg across domains).

Setting	Recall@50	Final F1
No truncation	**0.92**	0.79
Truncation (θ tuned)	0.90	**0.80**

4.6. Overhead and Practicality

Temperature scaling adds a small validation-time fit and a single scalar at inference; latency impact is negligible. Logging calibrated confidences and abstentions supports audits and improves reproducibility.

5. Discussion

5.1. Operational Benefits

Calibration converts raw scores into actionable probabilities: thresholds transport across datasets, SLAs can target precision at fixed coverage, and review queues can be budgeted. In our experiments, a single temperature T per domain sufficed for stable operation.

5.2. Limitations and Threats

A global temperature cannot correct class- or alias-specific biases; vector scaling and class-wise calibration [16] may help at small cost. Severe distribution shifts still degrade confidence quality; monitoring ECE and coverage–risk curves in production is prudent.

5.3. Relation to the Base Paper

The bibliometric analysis [26] mapped the field's scholarly structure. Our contribution complements that perspective with a deployment-centric lens: calibrated, selective enrichment that aligns technical performance with governance and risk controls relevant to the venues and topics identified in [26].

6. Conclusion

We presented a practical recipe for *risk-sensitive semantic enrichment*: temperature-scaled calibration, confidence-aware truncation, and selective prediction. Across Ont/LD, Bio, and Social domains aligned with [26], we reduced ECE by 30–45% and lowered error at fixed coverage with negligible latency overhead. Future work includes class-wise calibration, continual re-fitting under drift, and joint optimization of retrieval thresholds with calibrated cross-encoder scores.

References

[1] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In *NeurIPS*.

[2] Cao, N. D., Izacard, G., Riedel, S., Petroni, F., & Karpukhin, V. (2021). Autoregressive entity retrieval. *ICLR* (OpenReview preprint).

[3] Chen, J., Zhang, X., Li, X., et al. (2019). Neural entity linking: A survey of models and datasets. *arXiv:1904.09551*.

[4] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *ICML*.

[5] Chen, X., & He, K. (2021). Exploring simple Siamese representation learning. In *CVPR*.

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

[7] Ferragina, P., & Scaiella, U. (2010). TagMe: On-the-fly annotation of short text fragments. In *CIKM* (pp. 1625–1628). ACM.

[8] Ganea, O.-E., & Hofmann, T. (2017). Deep joint entity disambiguation with local neural attention. In *EMNLP*.

[9] Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple contrastive learning of sentence embeddings. In *EMNLP*.

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

[11] Hoffart, J., Yosef, M. A., Bordino, I., et al. (2011). Robust disambiguation of named entities in text. In *EMNLP*.

[12] Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs (FAISS). *IEEE Transactions on Big Data*, 7(3), 535–547.

[13] Ji, S., Pan, S., Cambria, E., et al. (2022). A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2), 494–514.

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

[15] Khosla, P., Teterwak, P., Wang, C., et al. (2020). Supervised contrastive learning. In *NeurIPS*.

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

[17] Malkov, Y. A., & Yashunin, D. A. (2018). Efficient and robust approximate nearest neighbor search using HNSW. *IEEE TPAMI*, 42(4), 824–836.

[18] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv:1301.3781*.

[19] Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. (2016). A review of relational machine learning for

knowledge graphs. *Proceedings of the IEEE*, 104(1), 11–33.

[20] van den Oord, A., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv:1807.03748*.

[21] Platt, J. (1999). Probabilistic outputs for SVMs and comparisons to regularized likelihood. In *Advances in Large Margin Classifiers* (pp. 61–74). MIT Press.

[22] Qu, Y., Ding, Y., Liu, J., et al. (2021). RocketQA: An optimized training approach to dense passage retrieval. In *NAACL*.

[23] Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning transferable visual models from natural language supervision. In *ICML*.

[24] Ratinov, L., Roth, D., Downey, D., & Anderson, M. (2011). Local and global algorithms for disambiguation to Wikipedia. In *ACL*.

[25] Reimers, N., & Gurevych, I. (2019). Sentence-BERT. In *EMNLP*.

[26] 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.

[27] 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.

[28] Trouillon, T., Welbl, J., Riedel, S., Gaussier, E., & Bouchard, G. (2016). Complex embeddings for simple link prediction. In *ICML*.

[29] Vulić, I., & Mrkšić, N. (2018). Specialising word vectors for lexical entailment. *TACL*, 6, 267–278.

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

[31] Xiong, L., Xiong, C., Li, Y., et al. (2021). ANCE: Approximate nearest neighbor negative contrastive learning. In *ICLR*.