



School of Science and Technology

Master in Information Systems (PLS)

Training and Adaptation of Large Language Models (LLMs) for
Entity Resolution on Heterogeneous and Incomplete Data

Theofilos Savvidis

Supervisor: Dimitrios Karapiperis

Patras, Greece, June 2026

Theses / Dissertations remain the intellectual property of students (“authors/creators”), but in the context of open access policy they grant to the HOU a non-exclusive license to use the right of reproduction, customisation, public lending, presentation to an audience and digital dissemination thereof internationally, in electronic form and by any means for teaching and research purposes, for no fee and throughout the duration of intellectual property rights. Free access to the full text for studying and reading does not in any way mean that the author/creator shall allocate his/her intellectual property rights, nor shall he/she allow the reproduction, republication, copy, storage, sale, commercial use, transmission, distribution, publication, execution, downloading, uploading, translating, modifying in any way, of any part or summary of the dissertation, without the explicit prior written consent of the author/creator. Creators retain all their moral and property rights.



Training and Adaptation of Large Language Models (LLMs) for Entity Resolution on Heterogeneous and Incomplete Data

Theofilos Savvidis

Supervising Committee

Supervisor:

Dimitrios Karapiperis.
Hellenic Open University

Co-Supervisor:

Georgios Papamichael.
Hellenic Open University

Patras, Greece, June 2026

Acknowledgments

The author would like to express sincere gratitude to Supervisor Dimitrios Karapiperis for his guidance, constructive feedback, and continuous support throughout the development of this thesis. His expertise in entity resolution and data integration [1, 2] provided invaluable direction at every stage of the research.

All generative inference experiments were executed locally on an Apple MacBook Air M1 (8GB unified memory) using the Ollama framework for model deployment, without reliance on proprietary cloud APIs or external computational resources. The BERT Cross-Encoder fine-tuning and FAISS/HNSW indexing pipelines were implemented in Python using the HuggingFace Transformers library and the FAISS library developed by Meta AI Research [14].

The benchmark datasets used in this study Abt-Buy, Amazon-Google, DBLP-ACM, and Scholar-DBLP are publicly available through established Entity Resolution research repositories and are cited accordingly throughout the thesis. No proprietary or personally identifiable data were used at any point in the experimental pipeline.

Abstract

Entity Resolution (ER), the task of identifying records that refer to the same real-world entity across heterogeneous data sources, remains a fundamental challenge in data integration [1, 2]. While transformer-based cross-encoders have established strong supervised baselines [3, 9], the emergence of generative Large Language Models (LLMs) has raised the question of whether zero-shot and few-shot prompting can serve as a viable alternative [4, 5].

A persistent methodological flaw in current benchmarking is the reliance on random negative sampling, which produces obviously easy non-matching pairs and artificially inflates model performance [6]. This thesis addresses that flaw by constructing an evaluation pipeline built exclusively on hard negatives: semantically confusing non-matching pairs isolated via a Hierarchical Navigable Small-World (HNSW) graph index [7, 14, 43, 50]. Four benchmark datasets covering e-commerce and bibliographic domains are evaluated under two complementary experimental protocols, both of which operate exclusively at the boundary of semantic ambiguity.

The first protocol benchmarks a fine-tuned BERT cross-encoder [3, 9] against LLaMA 3 [10], Mistral 7B [11] and Gemma 7B [28] across four prompting strategies: zero-shot, few-shot (up to eight shots), Chain-of-Thought [12, 17] and confidence elicitation [13, 19]. The supervised baseline achieves F1 scores between 69.6 % and 98.3 % depending on domain, and no generative configuration surpasses it on any dataset. Generative models reveal systematic behavioural pathologies including a strong positive test strategy (“Yes-Man” phenomenon), prompt brittleness under structured reasoning [12], overconfidence miscalibration [13] and context overload at extended prompt lengths, while operating 2,612 times more slowly than the discriminative baseline.

The second protocol isolates the contribution of the training data itself. Six cross-encoder architectures [31, 32, 34, 35, 36, 37] are fine-tuned across two negative sampling regimes and four datasets, forty-eight configurations in total, and supplemented by two specialised reranking architectures [44, 45, 46] evaluated with and without task-specific adaptation. A controlled ablation establishes that replacing randomly sampled negatives with HNSW-mined hard negatives is worth between 10.86 and 31.53 F1 points, an order of magnitude more than the choice among the leading architectures, which accounts for under two points. A direct measurement of benchmark inflation shows that identical model weights report F1 scores 0.57 to 14.80 points higher when the evaluation set is constructed with random rather than hard negatives, quantifying a measurement bias present throughout the published literature. The findings are validated by twenty-four additional training runs under a substituted blocking encoder, by a seed-stability analysis establishing a measurement variance of ± 1.47 points, and by a cross-protocol bridge in which a shared architecture reproduces its score to within 0.35 points across both protocols.

Taken together, the two protocols support a single mechanistic account: the affirmative bias observed in unadapted generative models and the precision collapse observed in cross-encoders trained on random negatives are the same phenomenon — the signature of a decision boundary never constrained from the negative side — and both are remedied by hard-negative supervision rather than by increased model scale.

Keywords

Entity Resolution, Large Language Models, Hard Negatives, Negative Sampling, Cross-Encoders, Rerankers, Benchmark Inflation, In-Context Learning, Chain-of-Thought, Overconfidence Miscalibration, Context Overload, HNSW Indexing, Ablation Study.

Training and Adaptation of Large Language Models (LLMs) for Entity Resolution on Heterogeneous and Incomplete Data

Theofilos Savvidis

Περίληψη

Η Ανάλυση Οντοτήτων (Entity Resolution — ER), δηλαδή το έργο της ταυτοποίησης εγγραφών που αναφέρονται στην ίδια οντότητα σε ετερογενείς πηγές δεδομένων, παραμένει μια θεμελιώδης πρόκληση στην ενοποίηση δεδομένων. Ενώ οι Cross-Encoders που βασίζονται στην αρχιτεκτονική Transformer έχουν καθιερώσει ισχυρά εποπτευόμενα πρότυπα, η εμφάνιση των παραγωγικών Μεγάλων Γλωσσικών Μοντέλων (LLMs) έχει εγείρει το ερώτημα εάν η καθοδήγηση μηδενικών ή ελάχιστων παραδειγμάτων μπορεί να αποτελέσει βιώσιμη εναλλακτική λύση.

Μια επίμονη μεθοδολογική αδυναμία στη σύγχρονη βιβλιογραφία είναι η εξάρτηση από την τυχαία δειγματοληψία αρνητικών ζευγών, η οποία παράγει προφανώς εύκολα μη ταιριαστά ζεύγη και διογκώνει τεχνητά την απόδοση των μοντέλων. Η παρούσα εργασία αντιμετωπίζει αυτή την αδυναμία κατασκευάζοντας μια ροή αξιολόγησης βασισμένη αποκλειστικά σε σκληρά αρνητικά (hard negatives): σημασιολογικά μπερδεμένα μη ταιριαστά ζεύγη που απομονώνονται μέσω ευρετηρίου HNSW. Τέσσερα σύνολα δεδομένων από εμπορικούς και βιβλιογραφικούς τομείς αξιολογούνται μέσω δύο συμπληρωματικών πειραματικών πρωτοκόλλων, καθένα από τα οποία λειτουργεί αποκλειστικά στα όρια της σημασιολογικής ασάφειας.

Το πρώτο πρωτόκολλο συγκρίνει έναν ρυθμισμένο BERT Cross-Encoder με τα LLaMA 3, Mistral 7B και Gemma 7B σε τέσσερις στρατηγικές προτροπών: zero-shot, few-shot, Chain-of-Thought και εξαγωγή εμπιστοσύνης. Το εποπτευόμενο μοντέλο επιτυγχάνει F1 μεταξύ 69,6 % και 98,3 % ανάλογα με τον τομέα, και καμία παραγωγική διάταξη δεν το ξεπερνά σε κανένα dataset. Τα παραγωγικά μοντέλα αποκαλύπτουν συστηματικές συμπεριφορικές παθολογίες — φαινόμενο απόλυτης συγκατάβασης, ευθραυστότητα στη δομημένη συλλογιστική, υπερβολική σιγουριά σε λανθασμένες προβλέψεις και υπερφόρτωση πλαισίου — ενώ λειτουργούν 2.612 φορές πιο αργά από το διακριτικό πρότυπο.

Το δεύτερο πρωτόκολλο απομονώνει τη συνεισφορά των ίδιων των δεδομένων εκπαίδευσης. Έξι αρχιτεκτονικές Cross-Encoder εκπαιδεύονται σε δύο καθεστώτα αρνητικής δειγματοληψίας και τέσσερα datasets — σαράντα οκτώ διατάξεις συνολικά — και συμπληρώνονται από δύο εξειδικευμένες αρχιτεκτονικές επανακατάταξης (rerankers) που αξιολογούνται με και χωρίς προσαρμογή στο task. Μια ελεγχόμενη μελέτη απομόνωσης (ablation) αποδεικνύει ότι η αντικατάσταση των τυχαία δειγματοληπτημένων αρνητικών με σκληρά αρνητικά αξίζει από 10,86 έως 31,53 μονάδες F1 — μία τάξη

μεγέθους περισσότερο από την επιλογή μεταξύ των κορυφαίων αρχιτεκτονικών, η οποία αξίζει λιγότερο από δύο μονάδες. Μια άμεση μέτρηση της διόγκωσης των benchmarks δείχνει ότι τα ίδια ακριβώς βάρη μοντέλου αναφέρουν F1 υψηλότερο κατά 0,57 έως 14,80 μονάδες όταν το σύνολο αξιολόγησης κατασκευάζεται με τυχαία αντί για σκληρά αρνητικά, ποσοτικοποιώντας μια μεροληψία μέτρησης που διαπερνά τη δημοσιευμένη βιβλιογραφία. Τα ευρήματα επικυρώνονται από είκοσι τέσσερις επιπλέον εκπαιδεύσεις με διαφορετικό μοντέλο blocking, από ανάλυση ευστάθειας που εγκαθιδρύει διακύμανση μέτρησης $\pm 1,47$ μονάδων, και από γέφυρα διασταυρούμενου πρωτοκόλλου όπου κοινή αρχιτεκτονική αναπαράγει την επίδοσή της εντός 0,35 μονάδων.

Συνολικά, τα δύο πρωτόκολλα στηρίζουν μία ενιαία μηχανιστική ερμηνεία: η καταφατική μεροληψία των μη προσαρμοσμένων παραγωγικών μοντέλων και η κατάρρευση της ακρίβειας των Cross-Encoders που εκπαιδεύονται σε τυχαία αρνητικά αποτελούν το ίδιο φαινόμενο — την υπογραφή ενός ορίου απόφασης που δεν περιορίστηκε ποτέ από την αρνητική πλευρά — και αμφότερα θεραπεύονται με εποπτεία σκληρών αρνητικών και όχι με αύξηση του μεγέθους του μοντέλου.

Πίνακας περιεχομένων

Acknowledgments	v
Abstract.....	vi
Περίληψη	viii
List of Figures	xiii
List of Tables.....	xv
List of Abbreviations & Acronyms.....	xvii
1 Introduction	2
1.1 The Challenge of Entity Resolution.....	2
1.2 Computational Intractability and the Role of Blocking	3
1.3 From Discriminative Transformers to Generative Reasoning.....	3
1.4 The Methodological Pathology of Random Negative Sampling.....	4
1.5 The Hard-Negative Pipeline and Thesis Contributions	5
2 Related Work.....	7
2.1. Evolution of Neural Entity Resolution.....	7
2.2 Transformer-based Architectures and the Cross-Encoder Paradigm.....	8
2.3 Bi-Encoders and Blocking Efficiency.....	9
2.4 Generative Large Language Models in ER.....	10
3 Algorithmic Approach and System Architecture	12
3.1 Dense Vector Candidate Generation (Bi-Encoder Architecture)	12
3.2 Hierarchical Navigable Small-World (HNSW) Indexing.....	12
3.3 The Cross-Encoder Matching Architecture	13
3.4 Candidate Pruning and Stratified Splitting.....	14
3.5 Generative Inference Architecture.....	15
4 Data Description and Splitting Strategy	17
4.1 Benchmark Datasets	17
4.2 Schema-Agnostic Serialization and Dense Blocking	19
4.2.1 Comparative Analysis of Dense Embedding Architectures: MPNet vs. MiniLM	20
4.3 The Stratified Splitting Protocol	21
5 Experimental Procedure and Generative Inference.....	27
5.1 Hardware and Local Execution Environment.....	27
5.2 Prompt Engineering Formulation.....	28
5.2.1 Zero-Shot and Domain-Specific Few-Shot Paradigms	28
5.2.2 Structured Multi-Step Reasoning (Chain-of-Thought).....	29

5.2.3	Confidence Elicitation and Miscalibration Detection	29
5.3	Deterministic Response Parsing and Evaluation	30
6	Results and Analysis	31
6.1	The Supervised Baseline (The Empirical Ceiling)	31
6.2	Generative LLMs: Few-Shot Sensitivity and Domain Analysis	34
6.2.1	The E-commerce Domain (Abt-Buy).....	34
6.2.2	The Complex Software Domain (Amazon-Google).....	37
6.2.3	The Structured Bibliographic Domain (DBLP-ACM)	39
6.2.4	Scholar-DBLP	42
6.3	The Impact of Structured Reasoning (Chain-of-Thought)	44
6.4	Confidence Calibration and Overconfidence Bias.....	47
6.5	Computational Efficiency and Latency Analysis.....	49
6.6	Global Performance Synthesis.....	52
6.7	Qualitative Error Analysis (Case Studies).....	54
6.7.1	Case Study 1: The "Yes-Man" Phenomenon (False Positive).....	54
6.7.2	Case Study 2: Prompt Brittleness and Hallucinated Conflicts (False Negative)	55
6.7.3	Case Study 3: Overconfidence Miscalibration in Bibliographic Data.....	56
7	The Hard-Negative Fine-Tuning Pipeline (Approach B)	57
7.1	Motivation and Research Questions	57
7.2	From Protocol A to Protocol B: Rationale and Comparability	58
7.3	Experimental Setup	60
7.3.1	Unified Blocking Architecture	60
7.3.2	Ratio-Based Data Partitioning.....	61
7.3.3	Hard and Random Negative Training Variants	61
7.3.4	Training Configuration.....	62
7.4	The Architecture Landscape	63
7.5	Results: Hard-Negative Training.....	64
7.6	The Negative Sampling Ablation.....	66
7.7	Specialised Reranking Architectures	69
7.7.1	Zero-Shot Performance	69
7.7.2	Fine-Tuned Performance and Sample Efficiency.....	71
7.7.3	Threshold Calibration.....	73
7.8	Robustness: Substitution of the Blocking Encoder.....	73
7.9	Benchmark Inflation.....	75
7.10	Seed Stability.....	76
7.11	Cross-Protocol Validation	78
7.12	Positioning Against Established Entity Resolution Systems	79
7.13	Discussion and Synthesis	80
7.13.1	The Hierarchy of Factors.....	80
7.13.2	A Unified Account of the Precision Collapse	81
7.13.3	Computational Considerations	82
7.14	Threats to Validity.....	83

7.14.1	Oracle Threshold Selection	83
7.14.2	Absence of Repeated Runs.....	83
7.14.3	Blocking Recall as a Performance Ceiling	84
7.14.4	Asymmetry of the Discriminative–Generative Comparison	84
7.14.5	Domain and Language Coverage	84
7.14.6	Uniform Hyperparameters.....	85
8	Conclusions	86
8.1	<i>Final Deductions on Generative LLMs in Entity Resolution</i>	<i>86</i>
8.2	<i>Methodological Limitations and Future Work.....</i>	<i>88</i>
	References	89

List of Figures

1. Figure 1: Structural Partitioning and Data Splits of the Hard-Negative Candidate Pools Across the Four Evaluated Benchmark Datasets.
2. Figure 2: BERT Cross-Encoder Threshold Calibration (DBLP-ACM)
3. Figure 3: BERT Cross-Encoder Threshold Calibration and Optimization Curve on the Scholar-DBLP Dataset.
4. Figure 4: Performance Ceiling Comparison Between the Supervised BERT Cross-Encoder Baseline and the Best Performing Generative LLM Across All Target Domains.
5. Figure 5: Few-Shot Ablation Study – Abt-Buy Dataset
6. Figure 6 : Precision vs. Recall dynamics for LLaMA 3 (8B) on the Abt-Buy dataset. The visualization empirically isolates the "Yes-Man" phenomenon, demonstrating how near-perfect Recall metrics mathematically mask a severe degradation in Precision across the entire few-shot trajectory due to structural over-matching.
7. Figure 7: Few-Shot Ablation Study – Amazon-Google Dataset
8. Figure 8 : Precision vs. Recall dynamics for LLaMA 3 (8B) on the Amazon-Google dataset. The visualization highlights the persistent over-matching behavior and low precision ceiling characteristic of generative inference in highly unstructured, alphanumeric software domains.
9. Figure 9: Few-Shot Ablation Study Learning Trajectories and Context Overload Manifestation Across Generative Architectures (DBLP-ACM).
10. Figure 10: Precision vs. Recall dynamics for LLaMA 3 (8B) on the DBLP-ACM dataset. The visualization illustrates the absolute recall bias where the generative model maintains a baseline of 100% recall while precision scales progressively up to the 5-shot threshold.
11. Figure 11: The "1-Shot Anomaly" and structural precision degradation under expanding context lengths in Gemma 7B (Scholar-DBLP).
12. Figure 12: Precision vs. Recall dynamics for LLaMA 3 (8B) on the Scholar-DBLP dataset. The chart depicts the highly stable, flat behavioral trajectory of the model when confronted with noisy, web-scraped bibliographic metadata.
13. Figure 13: CoT vs. Zero-Shot F1-Score Across All Models and Datasets
14. Figure 14: Impact of Chain-of-Thought (CoT) Sequential Reasoning Constraints and Temperature 0.0 Calibration on Gemma 7B Alignment Bias (Scholar-DBLP).
15. Figure 15: Generative LLM Epistemic Miscalibration and Overconfidence Bias on Erroneous Sequence-Pair Classifications (DBLP-ACM).

16. Figure 16: Computational latency analysis across the entity resolution pipeline. Execution times were recorded locally on an Apple Mac M1 for 100 query vectors. The logarithmic scale highlights the $\sim 2,612\times$ latency increase of the autoregressive LLM compared to the BERT Cross-Encoder during the matching phase.
17. Figure 17: System throughput comparison during the matching phase. The logarithmic visualization contrasts the high-velocity processing capacity of the fine-tuned BERT Cross-Encoder (107.5 pairs/sec) against the severe operational bottleneck encountered by the autoregressive Mistral 7B architecture (0.04 pairs/sec).
18. Figure 18: Class imbalance distribution within the candidate pair pools post-HNSW blocking. The persistent high proportion of hard negatives (ranging from 69% to 80%) illustrates the semantic difficulty of the isolated test sets across all domains.
19. Figure 19: Schematic comparison of the two experimental pipelines. The two approaches are identical through serialisation, dense encoding and HNSW retrieval; they diverge in the choice of blocking encoder, in whether a random-negative training variant is constructed, and in the set of matching architectures evaluated.
20. Figure 20: The negative sampling ablation. For each dataset the best-performing model trained on HNSW-mined hard negatives is compared against the best-performing model trained on randomly sampled negatives, with both evaluated on the identical hard-negative test partition.
21. Figure 21: Precision and recall of models trained on randomly sampled negatives, averaged across architectures and evaluated on the hard-negative partition. Recall saturates near 100 % while precision collapses towards the dataset positive rate.
22. Figure 22: Reranking architectures before and after task-specific fine-tuning. Without adaptation both rerankers fall substantially below the compact fine-tuned cross-encoders despite being eight to seventeen times larger.
23. Figure 23: The advantage of the 278-million-parameter reranker over the best compact architecture, plotted against the number of training pairs available.
24. Figure 24: Benchmark inflation across the four datasets. The model weights are identical in both bars of each pair; only the construction of the negative examples in the evaluation partition differs.
25. Figure 25: The hierarchy of factors determining Entity Resolution accuracy, measured within a single pipeline and on identical data.

List of Tables

1. Table 1: Hard-Negative Candidate Pool Statistics and Partition Sizes per Dataset
2. Table 2: Supervised BERT Cross-Encoder Baseline Performance across datasets
3. Table 3: Few-Shot ablation study performance on the Abt-Buy dataset.
4. Table 4: Few-Shot ablation study performance on the Amazon-Google dataset.
5. Table 5: Few-Shot ablation study performance on the DBLP-ACM dataset.
6. Table 6: Few-Shot ablation study performance on the Scholar-DBLP dataset.
7. Table 7: The impact of Chain-of-Thought (CoT) sequential reasoning on generative model performance (F1-Score) compared to the Zero-Shot baseline.
8. Table 8: The impact of Confidence Elicitation (with a strict algorithmic penalty) on model F1-Scores compared to Zero-Shot baselines. (Note: Gemma 7B experiments were executed under a hybrid CoT+Confidence restriction with temperature 0.0 on a sample of N=50).
9. Table 9: Global Performance Summary of the Best Performing Strategies across Datasets.
10. Table 10: Capacity Discrepancy Ignored by LLMs (Abt-Buy Dataset)
11. Table 11: Versioning and Platform Confusion (Amazon-Google Dataset)
12. Table 12: Comparison of the two experimental protocols.
13. Table 13: Candidate pool statistics and 30/70 partitioning under unified BGE-Small blocking, $k = 5$.
14. Table 14: Cross-encoder architectures evaluated under Approach B.
15. Table 15: Abt-Buy — cross-encoder performance under hard-negative training.
16. Table 16: Amazon-Google — cross-encoder performance under hard-negative training.
17. Table 17: DBLP-ACM — cross-encoder performance under hard-negative training.
18. Table 18: Scholar-DBLP — cross-encoder performance under hard-negative training.
19. Table 19: Consolidated F1-scores under hard-negative training across all datasets.
20. Table 20: Performance under random-negative training, evaluated on the hard-negative partition.
21. Table 21: Ablation summary — best attainable F1-score under each negative sampling regime.
22. Table 22: Specialised reranking architectures evaluated.
23. Table 23: Reranker performance in zero-shot and fine-tuned configurations.
24. Table 24: Fine-tuned BGE-Reranker-Base against the best compact architecture, ordered by training-set size.
25. Table 25: F1-scores under hard-negative training with each blocking encoder.
26. Table 26: Replication of the negative sampling ablation under substituted blocking (Abt-Buy).

- 27. Table 27: Benchmark inflation. Identical model weights evaluated on two test partitions differing only in the provenance of the negative pairs.
- 28. Table 28: Seed stability analysis. GTE-Small, Abt-Buy, hard negatives, three random seeds.
- 29. Table 29: Cross-protocol validation via the shared MS-MARCO MiniLM architecture.
- 30. Table 30: Indicative positioning against established Entity Resolution systems.
- 31. Table 31: The hierarchy of factors determining Entity Resolution accuracy, measured within a single pipeline.

List of Abbreviations & Acronyms

1. ANN - Approximate Nearest Neighbor
2. CoT - Chain-of-Thought
3. ER - Entity Resolution
4. HNSW - Hierarchical Navigable Small-World
5. LLM - Large Language Model
6. PEFT - Parameter-Efficient Fine-Tuning
7. PLM - Pre-trained Language Model
8. SBERT - Sentence-BERT
9. TF-IDF - Term Frequency-Inverse Document Frequency
10. API - Application Programming Interface
11. BERT - Bidirectional Encoder Representations from Transformers
12. CSV - Comma-Separated Values
13. F1 - F1-Score (harmonic mean of Precision and Recall)
14. FAISS - Facebook AI Similarity Search
15. GQA - Grouped-Query Attention
16. GPU - Graphics Processing Unit
17. LoRA - Low-Rank Adaptation
18. MLM - Masked Language Modeling
19. NLP - Natural Language Processing
20. P/R - Precision/Recall
21. QLoRA - Quantized Low-Rank Adaptation
22. RAM - Random Access Memory
23. RLHF - Reinforcement Learning from Human Feedback
24. SWA - Sliding Window Attention
25. ANNS - Approximate Nearest Neighbour Search
26. BCE - Binary Cross-Entropy
27. BGE - Beijing Academy of Artificial Intelligence General Embedding
28. GTE - General Text Embeddings
29. MTEB - Massive Text Embedding Benchmark
30. pp - Percentage points
31. XLM-R - Cross-lingual Language Model – RoBERTa

1 Introduction

1.1 The Challenge of Entity Resolution

When integrating data from multiple sources, a fundamental and well-studied challenge in data integration is Entity Resolution (ER), also known in the literature as record linkage or deduplication [1, 2]. At its core, the objective of ER is well-defined: it attempts to identify and merge different digital point to the exact same real-world object [1]. It serves as a prerequisite for a class of downstream applications, ranging from basic business intelligence analytics to advanced, knowledge-graph-driven systems [2, 30].

However, as digital data grows in volume and heterogeneity, establishing this "single source of truth" presents a significant computational challenge [2]. In real-world databases, we constantly deal with serious data quality issues. A single entity profile might be affected by missing attributes, typographical errors, or entirely different formatting conventions depending on the specific domain it originated from.

1.2 Computational Intractability and the Role of Blocking

Historically, the primary challenge in Entity Resolution has been the squared computational complexity [2]. If we try to take a naive, brute-force approach and compare every single record against every other record in a database of size N , the resulting $O(N^2)$ mathematical complexity renders the approach computationally infeasible. Rendering a large-scale data integration computationally infeasible without an effective reduction strategy.

To address this computational blockage, the community universally splits the ER workflow into two distinct steps: candidate selection (usually call blocking) and the actual candidate matching [1, 7]. Blocking can be conceptualized as an intelligent filtering mechanism. It groups potentially matching entities into localized clusters, ensuring that the computationally expensive matching algorithms are applied only on records that share the same conceptual block [7]. In the early days, domain experts used to hand-craft rigid rules for this blocking phase based on database schemas [1, 30]. However, with the complexity of modern heterogeneous data, it is now necessary to operate in schema-agnostic environments where structural metadata is often missing, unreliable, or completely mismatched.

1.3 From Discriminative Transformers to Generative Reasoning

Once the candidate pairs are filtered, the actual matching phase starts. The mechanisms driving this phase have evolved dramatically. Eventually, the field moved towards traditional machine learning, relying heavily on manually engineered string similarity metrics like Jaccard or Levenshtein distances [2, 16]. A significant shift came with Deep Learning. By translating text into dense mathematical vectors, neural networks enabled the capture of semantic relationships that simple character-matching algorithms completely missed [8].

A key advance came with Pre-trained Language Models (PLMs) like BERT and its derivatives [9]. The recent literature has established an architectural separation between Bi-Encoders and Cross-Encoders for this specific task [21]. While Bi-Encoders (like Sentence-BERT [8]) are exceptionally fast for generating dense vector representations during the blocking phase, they lack the deep, token-level interaction required for highly precise matching. To achieve state-of-the-art accuracy, we must utilize Cross-Encoders. Following the widely recognized DITTO-like paradigm [3], the matching phase is reframed into a precise sequence-pair classification task. By serializing two entity profiles into a single string (separated by a specialized [SEP] token) and feeding them into a Transformer, the multi-head bidirectional self-attention mechanism performs a deep, all-to-all token comparison. This architecture is so computationally powerful that it essentially sets the empirical performance ceiling for any supervised Entity Resolution task we see today [3, 21].

Recently, the research community has shifted toward generative Large Language Models (LLMs) like LLaMA 3 and Mistral for solving complex data integration tasks[4, 5]. Unlike older PLMs that demand exhausting and computationally expensive fine-tuning on highly specific datasets, these new LLMs offer the prospect of zero-shot and few-shot classification (in-context learning) driven entirely by natural language prompts [6]. However, a critical challenge remains. To fairly benchmark these models against supervised Cross-Encoders, a rigorous experimental design is required [6, 12].

1.4 The Methodological Pathology of Random Negative Sampling

This brings us to a pervasive methodological flaw in contemporary Entity Resolution literature: the common reliance on simple random sampling to create negative pairs (the non-matches) during evaluation [3, 4]. The problem starts with dataset construction. Since positive pairs representing true matches are inherently scarce, researchers often balance their testing data by randomly selecting non-matching entities from the catalog [2, 6].

Statistically speaking, this random sampling mathematically guarantees the creation of "easy negatives. In a large heterogeneous database, the odds of randomly grabbing two items that are distinct yet unclearly similar are incredibly low. Instead, a random script will likely pair an entity profile for a 50-inch smart television with a kitchen refrigerator. Evaluating a billion-parameter neural network on such a easy mismatch is a notable methodological limitation. The semantic gap is sufficiently wide that models require minimal analytical capacity to classify such pairs correctly. As a result, both generative and supervised models achieve near-perfect accuracy on these random test sets, producing an artificial inflation of performance metrics that hides the true, real-world limitations of these architectures [6].

1.5 The Hard-Negative Pipeline and Thesis Contributions

To construct a reliable evaluation framework for high-precision data integration, the experimental methodology must focus exclusively on the resolution of "hard negatives" [5, 7]. Within the paradigm of Entity Resolution, a hard negative is defined as a pair of records that exhibit a high degree of semantic similarity but correspond to entirely distinct real-world entities. A representative manifestation of this challenge occurs when attempting to differentiate between two consecutive generations of a consumer electronics product, where the descriptive attributes diverge only by a subtle variation in RAM capacity or a single alphanumeric shift in a complex model number [15]. It is precisely within these mathematically and semantically ambiguous edge cases that the true reasoning capabilities of architectures must be rigorously tested to prove their real-world utility.

This thesis addresses this flaw by systematically re-evaluating the near-perfect performance metrics frequently reported for generative language models an empirical distortion primarily induced by a widespread reliance on naive random negative sampling during dataset construction. To fix this methodological pathology, this research introduces a comprehensive experimental pipeline centered entirely on the algorithmic isolation and evaluation of hard negatives. By leveraging a Hierarchical Navigable Small-World (HNSW) graph index for dense vector blocking [7, 8], the proposed architecture computationally isolates highly confusing candidate pairs directly from a continuous vector space. Subsequently, a rigorous data splitting strategy is enforced upon these candidate pairs across four diverse benchmark datasets (Abt-Buy, Amazon-Google, DBLP-ACM, and Scholar-DBLP). For three of the four datasets, a fully stratified split preserves the exact positive-to-negative ratio across all partitions; for Abt-Buy, a large-scale random sample constitutes the Test Set. In both cases, the resulting environments force matching models to operate strictly at the boundaries of semantic uncertainty.

The experimental execution relies on a robust comparative protocol designed to evaluate a fine-tuned, DITTO-like BERT Cross-Encoder [3, 9] against three Large Language Models: LLaMA 3 [10], Mistral 7B [11], and Gemma 7B [28]. To capture both the native reasoning capacities and the boundaries of deterministic control, a dual-modality inference configuration was engineered. While LLaMA 3 and Mistral 7B were evaluated under their default baseline parameters to document their reasoning trajectories, the protocol implemented a deliberate methodological adaptation for Gemma 7B by enforcing strict deterministic inference constraints via a temperature setting of 0.0. This setting eliminates generative stochasticity, guarantees reproducibility, and allows the study to isolate the effect of deterministic constraints on error distribution, guarantees absolute reproducibility during sampled runs, and exposes how rigid behavioral constraints alter the model's error distribution. All three generative architectures were systematically exposed to an exhaustive suite of prompting paradigms, including Zero-Shot baselines, multi-step Chain-of-Thought (CoT) sequential reasoning [12], Confidence Calibration prompts, and a progressive Few-Shot Ablation Study tracking performance configurations up to 8 shots with domain-specific adaptation.

Eliminates the artificial inflation caused by easy negatives, this pipeline uncovers several systematic behavioral limitations. The empirical findings identify, most notably, the "Yes-Man" phenomenon: an extreme recall bias in which models default to affirmative matching decisions under conditions of high semantic overlap. The analysis also documents overconfidence Bias, where in models exhibit epistemic miscalibration by expressing high verbal certainty for incorrect predictions [13]. Additionally, Prompt Brittleness is observed under structured multi-step reasoning constraints [12], and a "Context Overload" effect is identified, in which excessive few-shot context disrupts the models' attention mechanisms and degrades classification accuracy. Through this structured evaluation, the thesis delineates the fundamental limitations of generative reasoning in high-precision data integration tasks, in comparison to bidirectionally fine-tuned discriminative architectures.

2 Related Work

2.1. Evolution of Neural Entity Resolution

The field of Entity Resolution (ER) has undergone a substantial transformation over the past two decades, progressing from deterministic rule-based systems and hand-crafted feature engineering toward end-to-end deep learning architectures [1, 2]. Early systems relied on manually defined similarity functions such as Jaccard coefficients, edit distance, and token-based overlap metrics applied over individual attribute pairs. While effective in controlled, schema-consistent environments, these approaches fail to generalise under real-world conditions characterised by heterogeneous schemas, missing values, and free-text descriptions [2, 30].

The first wave of neural approaches sought to replace engineered features with learned representations. DeepMatcher [22] represented a foundational step in this direction, framing ER as an attribute-level comparison task and exploring a design space of neural architectures from simple word embeddings to hybrid RNN-attention models. Its core contribution was the empirical demonstration that neural models could learn domain-agnostic matching functions directly from labelled entity pairs, outperforming hand-crafted baselines on standard benchmarks. Subsequent work introduced HierMatcher [23], which recognised that entity records are inherently multi-level structures: a single record comprises tokens nested within attributes, which are in turn nested within the full record. By performing joint matching at both the token level and the attribute level through a hierarchical attention mechanism, HierMatcher achieved improved robustness on heterogeneous and noisy datasets where attribute boundaries are meaningful but inconsistently formatted. Together, these frameworks established that entity matching benefits from compositional representation learning a principle that directly motivates the cross-encoder architecture adopted in this thesis.

A critical limitation shared by these early neural systems is their dependence on labelled training data with consistent schema alignment. In practice, large-scale entity resolution pipelines must operate in schema-agnostic environments where attribute names, orderings, and data types vary unpredictably across sources [7, 30]. This motivated the adoption of pre-trained language models, which encode rich, transferable semantic knowledge and do not require explicit attribute alignment as an input constraint [9, 21].

2.2 Transformer-based Architectures and the Cross-Encoder Paradigm

The introduction of the Transformer architecture [24] and pre-trained language models (PLMs) such as BERT [9] marked a fundamental shift in the state of the art for entity resolution. Unlike recurrent architectures, the Transformer’s multi-head self-attention mechanism can directly model long-range dependencies between any two tokens in a sequence, making it particularly well-suited for comparing heterogeneous text representations. Fine-tuning BERT on domain-specific data requires substantially less labelled data than training from scratch, and the resulting models generalise effectively across varying attribute structures [9, 21].

The defining work in this area is DITTO [3], which reframed entity resolution as a sequence-pair classification task. By concatenating the serialised representations of two entity records into a single input sequence separated by the special [SEP] token DITTO enables the Transformer’s bidirectional self-attention to perform deep, all-to-all token-level interactions between the two records simultaneously. This cross-encoder formulation is more expressive than independent encoding because every token in Record A can directly attend to every token in Record B across all transformer layers, allowing the model to detect subtle contradictions such as a mismatched version number embedded within a longer product title.

The performance of DITTO extends beyond its architectural choice. The framework introduces two complementary optimisation strategies that substantially improve robustness on noisy real-world data. First, domain knowledge injection highlights critical spans within entity records such as model numbers, brand names, and technical specifications by inserting special marker tokens around them before encoding. This guides the self-attention mechanism toward semantically discriminative regions and reduces the influence of boilerplate or irrelevant text. Second, DITTO employs a systematic data augmentation pipeline that applies multiple transformations to training pairs: token deletion removes individual words to simulate missing attributes; span shuffling reorders attribute segments to enforce positional invariance; and TF-IDF-weighted token replacement substitutes informative terms with random alternatives to improve robustness against lexical variation. These augmentations collectively force the model to learn matching functions that are invariant to the surface-level noise and formatting inconsistencies characteristic of real-world heterogeneous databases [3]. The comprehensive experimental evaluation conducted by Karapiperis et al. [21] confirms that this cross-encoder paradigm consistently establishes the empirical performance ceiling for supervised entity resolution, which is why it serves as the primary discriminative baseline in this thesis.

Despite its high accuracy, the cross-encoder formulation inherits an important practical limitation: because both records must be jointly encoded in a single forward pass, the model cannot pre-compute record representations offline. Evaluating all $O(N^2)$ candidate pairs in a large database is therefore computationally unworkable, necessitating an efficient preceding stage that reduces the candidate space to a manageable subset. This architectural constraint motivates the two-stage pipeline adopted throughout this thesis, in

which a fast-blocking mechanism first identifies a small set of highly similar candidate pairs, which are then re-ranked by the more accurate cross-encoder [7, 21, 30].

2.3 Bi-Encoders and Blocking Efficiency

To address the scalability blockage of cross-encoders, the research community developed bi-encoder architectures that encode each record independently into a fixed-size dense vector [8]. Because these embeddings can be computed offline and stored in an index, similarity search reduces to a fast nearest-neighbour lookup rather than a repeated forward pass through a large transformer. Sentence-BERT (SBERT) [8] extended the original BERT architecture with a Siamese network design and a contrastive training objective, producing semantically meaningful sentence embeddings in which cosine similarity correlates reliably with semantic relatedness. This property makes SBERT-derived embeddings highly effective for retrieving candidate pairs that are semantically close but may differ in surface form.

Several frameworks have extended the bi-encoder paradigm for entity resolution specifically. BERT-ER [27] adapts pre-trained BERT representations for dense retrieval in entity matching, demonstrating that fine-tuned bi-encoders can achieve competitive blocking recall at a fraction of the computational cost of cross-encoders. LSBlock [7] represents a more sophisticated hybrid system that bridges classical and neural methodologies. It combines traditional lexical similarity search using minhash vectors with semantic similarity derived from dense representations, and evaluates the discriminative power of its blocking keys using SHAP (Shapley Additive Explanations) values from game theory. Critically, LSBlock employs a Hierarchical Navigable Small-World (HNSW) index [14] for approximate nearest-neighbour search, achieving logarithmic query time complexity $O(\log N)$. This is the same HNSW indexing mechanism adopted in the experimental pipeline of this thesis, establishing a direct methodological continuity with state-of-the-art blocking practice [7, 29].

A complementary line of research has explored graph-based neural architectures for entity matching. GNEM [25] frames entity matching as a one-to-set problem, using graph neural networks to jointly model all candidate matches for a given query entity rather than evaluating each pair in isolation. FlexER [26] extends this approach by supporting multiple matching intents simultaneously, allowing a single model to handle both deduplication and record linkage objectives. While these graph-based frameworks offer expressive relational modelling capabilities, they require the construction of an explicit entity graph as a preprocessing step, which presupposes a well-defined schema and sufficient attribute overlap to define meaningful edges. This requirement is infeasible in the schema-agnostic, heterogeneous settings evaluated in this thesis, where attribute names, cardinalities, and data types vary unpredictably across sources. The bi-encoder + cross-encoder pipeline, by contrast, operates directly on serialised free-text representations and requires no structural preprocessing, making it inherently compatible with incomplete and noisy catalogs [7, 21].

An important but often underexamined aspect of entity resolution research is the choice of evaluation methodology. The standard metric is the F1-score over matched pairs, which balances precision (the fraction of predicted matches that are correct) and recall (the fraction of true matches that are retrieved) [2, 16]. However, the validity of this metric depends critically on how the negative examples the non-matching pairs are constructed. In large databases, true matches are rare relative to the total number of possible pairs, creating a severe class imbalance that makes random negative sampling misleading: a randomly selected pair of non-matching records will almost certainly be obviously dissimilar, producing easy negatives that any classifier can correctly reject without meaningful reasoning [6, 30]. This evaluation flaw is systematically addressed in this thesis through the construction of hard-negative test sets, as detailed in Section 4.

2.4 Generative Large Language Models in ER

The most recent frontier in entity resolution research involves the use of generative Large Language Models (LLMs) through in-context learning, without any task-specific parameter optimisation [4]. Models such as LLaMA 3 [10], Mistral 7B [11], and Gemma 7B [28] have been evaluated on entity matching tasks using natural language prompts that describe the task and optionally provide a small number of labelled examples. The fundamental appeal of this approach lies in its elimination of the fine-tuning requirement: a single general-purpose model can be repurposed for entity resolution through prompt design alone, making it accessible for domains where labelled training data is scarce or costly to obtain [4, 6].

Peeters et al. [4] conducted the first systematic study of LLM-based entity matching, evaluating GPT-3.5 and other models across multiple benchmark datasets and prompt formulations. Their findings showed that LLMs can achieve competitive accuracy under zero-shot conditions on clean, structured datasets, but degrade substantially on noisy or schema-heterogeneous data. The study by Arvanitis-Kasinikos and Papadakis [6] extended this line of investigation to 7B-parameter open-source models deployed on consumer hardware, demonstrating that model performance is highly sensitive to prompt design and hardware constraints, and that published zero-shot benchmarks frequently overstate generalisation capability due to the use of easy random negatives.

Recent frameworks have explored more sophisticated prompting strategies to improve LLM reliability. The AvengER framework [5] examined the transition from standard binary prompts (MATCH / NO MATCH) to multiple-choice SELECT prompts, in which the model is presented with a query entity and a ranked list of candidates. This formulation more closely mirrors the retrieval-reranking architecture of production systems but introduces a new vulnerability: position bias, whereby generative models tend to favour candidates appearing at specific list positions regardless of semantic content. To address this, AvengER employs ensemble voting across multiple prompt orderings and applies lightweight fine-tuning on medium-sized models, demonstrating that task-specific adaptation can allow smaller LLMs to outperform much larger commercial models on entity resolution benchmarks. Complementary work on structured multi-step reasoning [12] has explored Chain-of-Thought prompting as a mechanism for improving LLM accuracy on entity matching, with mixed results across different model architectures and domains. Confidence calibration research [13, 19] has further revealed that generative models are systematically miscalibrated in classification settings, often expressing high verbal certainty for incorrect predictions.

The broader literature on pre-trained embeddings for entity resolution [29] confirms that dense vector representations derived from transformer encoders provide a strong foundation for both candidate retrieval and final matching. Taken together, these prior findings establish the theoretical and empirical context for the contributions of this thesis: by evaluating LLMs exclusively on hard-negative candidate pairs generated through HNSW-based dense blocking, this work moves beyond the misleading benchmarks produced by random negative sampling and provides a rigorous assessment of where generative reasoning succeeds and where it is structurally limited relative to fine-tuned discriminative architectures [3, 6, 21].

3 Algorithmic Approach and System Architecture

To construct an experimental framework capable of bypassing the "easy negative" anomaly, the system architecture must isolate hard negatives prior to any classification. This requires the implementation of an advanced, dense vector blocking mechanism, followed by a highly precise sequence-pair classification architecture [7, 8].

3.1 Dense Vector Candidate Generation (Bi-Encoder Architecture)

Traditional lexical blocking techniques rely on strict surface-level character matching [30]. While computationally fast, these methods can suffer from significant recall reduction when confronted with unstructured textual descriptions containing synonyms, semantic variations, or OCR-induced noise [7]. To overcome this limitation, the experimental pipeline transitions the candidate generation phase into a continuous mathematical space using dense embeddings.

Specifically, this phase employs a Bi-Encoder architecture based on Sentence-BERT (SBERT) [8]. Recent studies clearly distinguish between Bi-Encoders and Cross-Encoders for entity matching [21]. Although Cross-Encoders offer high accuracy, their heavy computational cost makes them impractical for scanning large datasets to find initial candidates. SBERT addresses this scalability problem by using a Siamese network design. It processes each record independently through a BERT encoder and applies a mean-pooling operation to create a single, fixed-size dense vector. Because this encoding step is separated for each record, the system can pre-calculate the vectors for the entire database offline. The result is an encoder that captures the semantic content of heterogeneous attributes while keeping similarity search fast and scalable [8, 21].

3.2 Hierarchical Navigable Small-World (HNSW) Indexing

Encoding the database into dense vectors is a necessary but insufficient step. Computing the exact Euclidean or Cosine distance between every pair is an $O(N^2)$ operation, operation, which is computationally infeasible for large-scale datasets [21]. To isolate the required hard negatives efficiently, the pipeline employs a Hierarchical Navigable Small-World (HNSW) graph index through the FAISS library [14].

Prior to constructing the HNSW index, all embeddings (u) are normalized to unit length ($u \rightarrow u / \|u\|_2$). This L2 normalization is a critical mathematical step, as it ensures that the Inner Product (IP) metric computed during FAISS graph traversal is equivalent to Cosine Similarity:

$$\text{Cosine Similarity}(u, v) = \frac{(u \cdot v)}{\|u\|_2 \cdot \|v\|_2} = u \cdot v \text{ for } \|u\|_2 = \|v\|_2 = 1$$

Consequently, FAISS can execute high-speed vector queries using highly optimized matrix multiplications (Inner Product search), translating directional semantic similarity directly into Cosine Similarity, which forms the basis of our approximate nearest neighbor filtering.

HNSW is one of the most effective algorithms for Approximate Nearest Neighbor (ANN) search [14, 43]. It works by building a multi-layered graph. The top layers contain very few points, allowing the algorithm to make long, fast jumps across the vector space. The layers become more crowded as you move down, ending at the bottom layer which connects all the records. During a search, the algorithm starts at the top layer, moves to the closest point, and then drops to the next level. This top-down approach helps the system quickly find the right cluster of vectors, reducing the search time to a much faster logarithmic scale, $O(\log N)$ [7, 14]. Comparative evidence on large-scale Entity Resolution workloads further favours graph-based over partition-based approximate nearest neighbour indices [50, 52], supporting this design decision.

3.3 The Cross-Encoder Matching Architecture

Once the HNSW index retrieves the localized clusters of highly similar entities, the candidate matching phase operates strictly on these returned pairs. To establish the absolute performance baseline, the pipeline deploys a fine-tuned BERT Cross-Encoder.

As highlighted by Karapiperis et al. [21] in their recent systematic evaluation of entity matching architectures, Cross-Encoders consistently achieve superior matching performance over Bi-Encoders. This advantage comes from their multi-layer self-attention mechanisms, which can effectively capture complex relationships between tokens across different attributes of the two entities [21]. Based on this proven approach, we use the ms-marco-MiniLM-L-6-v2 [8] model, configured specifically as a Cross-Encoder.

The ms-marco-MiniLM-L-6-v2 was selected on the basis of three converging considerations. First, it was pre-trained on the MS MARCO passage retrieval corpus using a cross-encoder ranking objective, which produces probability scores that are directly interpretable as match likelihoods a property essential for the threshold calibration procedure described in Section 4.3. Second, its 6-layer, 22M-parameter architecture achieves a favourable trade-off between classification accuracy and inference speed, making it viable for fine-tuning and repeated evaluation under the memory constraints of local consumer hardware [6,21]. Third, the systematic evaluation conducted by Karapiperis et al.[21] confirms that MiniLM-based cross-encoders consistently reach or approach the accuracy ceiling of larger models such as BERT-large or RoBERTa on entity resolution benchmarks, while requiring substantially fewer computational resources. For the Amazon-Google dataset, where the increased alphanumeric complexity of software versioning demanded greater representational capacity, the pipeline was upgraded to the deeper ms-marco-MiniLM-L-12-v2 variant, which doubles the number of transformer layers to 12 while preserving the same pre-training objective [21].

Following the standard DITTO framework [3, 21], our system treats entity matching as a sequence-pair classification problem. The attributes of both candidate records are combined into a single text string. We use special tokens such as e.g., [SEP] to establish structural boundaries between Record A and Record B.

This concatenated sequence passes directly through the deep bidirectional layers of the MiniLM Transformer. Unlike isolated embeddings, every token in the first entity profile immediately attends to and interacts with every token in the second entity profile across all 6 neural layers. This deep interaction allows the model to identify subtle contextual contradictions, such as a mismatched alphanumeric model number buried within a product description. The token representation at the final layer feeds into a linear classification head, producing a precise probability score that dictates the final matching decision [3].

3.4 Candidate Pruning and Stratified Splitting

By querying the FAISS/HNSW index [14] for the top-k nearest neighbors, the system retrieves a concentrated cluster of records exhibiting extreme semantic and lexical overlap. Pairs generated from this specific neighborhood, which do not correspond to actual duplicates according to the ground truth, constitute the absolute dataset of "hard negatives" [4, 6].

The integration of this Hierarchical Navigable Small-World (HNSW) indexing for candidate generation is fundamental to the scientific rigor of the experimental pipeline. Traditional Entity Resolution benchmarks frequently suffer from a dependence on random negative sampling, which creates mathematically trivial 'easy negatives'. Evaluating models on such obvious mismatches artificially inflates performance metrics and masks the true analytical limitations of Large Language Models (LLMs) [4, 5, 6]. By utilizing the HNSW index to retrieve only the nearest neighbors in a dense vector space, our methodology systematically isolates pairs that represent distinct real-world entities despite their high similarity.

To maintain strict scientific integrity, a dataset-specific splitting strategy is enforced directly upon these candidate pairs [21]. For Amazon-Google, DBLP-ACM, and Scholar-DBLP, a fully stratified split preserves the positive-to-negative ratio identically across all partitions. For Abt-Buy, given the scale of the candidate pool, a large-scale random sample of 2,000 pairs was extracted for the Test Set. By sequestering these highly confusing pairs into a dedicated benchmark test set, the experimental protocol ensures that both the fine-tuned BERT Cross-Encoder and the evaluated Generative LLMs are evaluated on an identical, challenging distribution. This deliberate constraint enables a more realistic evaluation of generative reasoning, revealing the true classification limits of each architecture under realistic conditions.

3.5 Generative Inference Architecture

While the supervised BERT Cross-Encoder establishes the empirical performance upper bound through domain-specific weight optimization, the parallel objective of this system architecture is to evaluate the out-of-the-box reasoning capabilities of massive, decoder-only models. Specifically, the pipeline integrates LLaMA 3 [10], Mistral 7B [11], and Gemma 7B [28] as generative matching engines.

Unlike the discriminative baseline, these Large Language Models operate entirely without parameter fine-tuning. Their integration into the Entity Resolution pipeline relies on the paradigm of in-context learning [6]. In this architectural configuration, the hard negative candidate pairs generated by the HNSW index are not transformed into numerical arrays for a classification head but are instead framed within highly structured natural language prompts [4].

Instead of treating the data as structured inputs, the generative models process the combined text of the candidate pair as a standard text prompt. Crucially, the generative models ingest the exact same schema-agnostic serialized string (described in Section 4.2) as the cross-encoder. By feeding the LLMs this dynamically flattened representation, the methodology ensures that missing attributes are organically omitted from the prompt context, preventing the models from hallucinating conflicts over absent data fields. Relying on their pre-trained knowledge, the models simply predict the most likely classification token, such as "MATCH" or "NO MATCH" [6]. Since these architectures lack the bidirectional cross-attention trained specifically for tabular data, their success depends entirely on their internal reasoning and the quality of the prompt engineering [4, 12]. By evaluating both the fine-tuned cross-encoder and the zero-shot generative models on the exact same set of hard negatives, our methodology ensures a fair and direct comparison.

Output Formatting Compliance and Parser Robustness

A significant challenge in utilizing autoregressive Large Language Models for automated Entity Resolution is their tendency to violate strict output formatting constraints, particularly in zero-shot and Chain-of-Thought (CoT) settings [12]. Despite prompt instructions demanding binary outputs ("YES" or "NO"), models frequently generate conversational preambles or verbose justifications.

To ensure evaluation integrity, our pipeline rejects naive substring matching (e.g., searching for the occurrence of "YES" anywhere in the response). A simple substring check is highly vulnerable to false positives when a model outputs negative assertions that contain positive target words (for example, "I cannot confidently say YES" would be incorrectly parsed as a match). To mitigate this formatting compliance issue, we implemented regular expression (Regex) parsers in the evaluation scripts (such as searching for specific word boundaries like "(YES|NO)" following the decision identifier).

Additionally, the pipeline incorporates a fallback mechanism: if the strict pattern is missing due to a formatting violation, the parser isolates and evaluates only the final line of the model's output. This multi-layered parsing protocol ensures that we capture only the model's final, definitive semantic decision, maintaining strict metric reproducibility across all benchmark partitions.

4 Data Description and Splitting Strategy

To measure the real-world performance difference between fine-tuned discriminative models and general-purpose generative LLMs, four widely recognised benchmark datasets were selected [15, 16, 21]. These datasets cover both e-commerce and bibliographic domains and are extensively used in Entity Resolution research.

4.1 Benchmark Datasets

To evaluate the effectiveness of the proposed Entity Resolution framework, four diverse and widely recognised benchmark datasets were selected, covering both the e-commerce and bibliographic domains.

The first dataset, Abt-Buy, consists of product descriptions extracted from two distinct online retailers, Abt.com and Buy.com. In its raw form, the dataset contains 1,081 entities from Abt and 1,092 entities from Buy.com, with 1,097 annotated ground-truth matching pairs out of a theoretical maximum of approximately 1.18 million possible pairs a positive rate of under 0.1%, illustrating the severe class imbalance inherent to real-world entity resolution. The dataset is characterised by varying lengths of textual descriptions, significant formatting differences, and a high degree of schema mismatch, making it one of the most challenging e-commerce benchmarks in the literature. The true complexity of the Abt-Buy dataset lies in its profound structural heterogeneity. Abt.com and Buy.com employ entirely disparate database schemas; an attribute categorized as 'Product Name' in one source is frequently fragmented into 'Brand', 'Model', and 'Description' in the other. This severe schema mismatch means that direct column-to-column comparisons are impossible, forcing the matching architectures to dynamically infer attribute alignment from the unstructured semantic context of the serialized string.

Similarly, the Amazon-Google dataset aligns software and electronics products between Amazon and the Google Base data API. This dataset presents a notoriously rugged terrain for evaluation due to its heavy reliance on complex alphanumeric model numbers, extreme variations in product naming conventions, and the frequent absence of structured metadata. In terms of heterogeneity and incompleteness, this dataset exemplifies the challenge of asymmetric information. A software product listed comprehensively on Amazon with explicit technical specifications and version editions may appear on Google Base as a heavily truncated string missing core attributes. This asymmetry forces the models to resolve entities where one profile is data-rich and the other is severely incomplete, relying heavily on subtle alphanumeric cues.

Extending the evaluation beyond commercial products, the DBLP-ACM dataset contains scientific publication records from the DBLP and ACM digital libraries. It specifically tests the models' ability to resolve entities based on highly structured but sparse metadata, such as slight variations in author names, publication years, and highly abbreviated conference venues. While originating from the same bibliographic domain, heterogeneity manifests here through divergent cataloging standards. ACM often provides complete author lists and full conference names, whereas DBLP frequently relies on initialized author names and abbreviated venue acronyms (e.g., 'SIGMOD' vs. 'International Conference on Management of Data'). This tests the models' ability to bridge vocabulary gaps and infer semantic equivalence across differing structural conventions.

Finally, the Scholar-DBLP dataset matches publication records from Google Scholar with the DBLP database to evaluate the models further on structured academic domains. This dataset serves as a critical benchmark to observe whether fine-tuning or few-shot prompting offers diminishing returns on data that closely resembles the models' pre-training corpora. Crucially, the "incomplete data" challenge reaches its absolute peak in this dataset. Because Google Scholar records are auto-extracted via web scraping, they result in highly contaminated profiles characterized by truncated titles, missing publication years, and merged or malformed author strings. This extreme structural noise contrasts sharply with the curated DBLP database, providing an ideal stress test for assessing whether architectures can robustly align severely fragmented data.

Through the inclusion of these four distinct benchmarks, the experimental setup ensures a holistic assessment of model performance across varying levels of noise and domain specificity. Having established the data sources for the evaluation, the following section describes the specific preprocessing and stratified protocols used to partition these datasets into training and testing sets.

4.2 Schema-Agnostic Serialization and Dense Blocking

Before generating candidate pairs, we had to convert the raw tabular data into a unified format suitable for sequence-pair classification. Real-world datasets often have highly inconsistent structures, with mismatched attributes and missing values [2]. To handle this, a flexible, schema-agnostic serialization method was adopted [7].

Rather than manually selecting specific columns, an automated function parsed every available attribute-value pair for a given entity and aggregated them into a single continuous string formatted as column name: value, excluding empty fields [7]. To transform this serialized textual representation into dense semantic embeddings, our framework differentiates the encoding model (Bi-Encoder) based on the domain characteristics of each dataset. Specifically, for the Abt-Buy dataset, the serialization was encoded using the all-mpnet-base-v2 model, which maps sentences to a 768-dimensional vector space and contains 110 million parameters to capture complex semantic relations in long product descriptions. Conversely, for the Amazon-Google, DBLP-ACM, and Scholar-DBLP datasets, we deployed the more lightweight all-MiniLM-L6-v2 model [8], which generates 384-dimensional embeddings using a highly efficient 22 million parameter architecture. The resulting dense vectors, normalized to unit length (L_2 norm), were subsequently stored in a FAISS Hierarchical Navigable Small-World (HNSW) graph index [14] to facilitate high-throughput approximate nearest neighbor (ANN) search. By performing an Inner Product search and retrieving the top-5 nearest neighbours for each query record, the pipeline isolates the most semantically confusing candidate pairs. The value $k = 5$ was selected based on two considerations: first, empirical blocking studies consistently report that the great majority of true matching pairs for these benchmark datasets are recovered within the top-5 results of a dense retrieval index, ensuring that the blocking stage does not discard ground-truth matches [7,29]; second, retrieving only 5 candidates per query maximises the proportion of hard negatives in the candidate pool expanding to $k = 10$ or $k = 20$ would introduce progressively easier non-matches from the outer semantic neighbourhood, partially defeating the purpose of the hard-negative evaluation protocol [6,30]. The resulting candidate pairs constitute the full dataset on which all subsequent splitting, fine-tuning, and evaluation are performed.

4.2.1 **Comparative Analysis of Dense Embedding Architectures: MPNet vs. MiniLM**

The selection of the two distinct Sentence-BERT (SBERT) models for the dense blocking stage was guided by structural and mathematical differences in their pre-training objectives and representation capacities [8, 21]:

Pre-training Paradigms:

***all-mpnet-base-v2 (MPNet):** The MPNet architecture utilizes a Masked and Permuted Pre-training objective that unifies Masked Language Modeling (MLM, as in BERT) and Permuted Language Modeling (PLM, as in XLNet). While MLM ignores dependency relations between masked positions and PLM suffers from positional discrepancies during pre-training, MPNet inherits the strengths of both by conditioning on the position information of all tokens while permuting the prediction order. This allows the model to retain a complete contextual view of the input sequence.

all-MiniLM-L6-v2 (MiniLM): MiniLM relies on deep self-attention knowledge distillation from a larger teacher model (e.g., MPNet or BERT-Large). It minimizes the Kullback-Leibler divergence between the self-attention relations of the teacher and student models, allowing the 6-layer student model to preserve over 95% of the teacher's representation quality while drastically reducing the parameter envelope.

Dimensional Capacity and Semantic Granularity:

MPNet (768 dimensions, 12 layers, 110M parameters): The larger embedding space and deeper transformer stack of MPNet are crucial for the Abt-Buy dataset, which consists of long-form, unstructured product descriptions. In e-commerce entity resolution, distinguishing near-identical products (e.g., matching a product variant while rejecting others) requires capturing subtle contextual signals. MPNet's higher capacity prevents different semantic features from collapsing into the same vector region, maintaining high precision during candidate pair retrieval.

MiniLM (384 dimensions, 6 layers, 22M parameters): For the bibliographic datasets (DBLP-ACM and Scholar-DBLP), the records are highly structured and consist of short metadata (e.g., titles, authors, venues, years). In these domains, the semantic complexity is lower, and the 384-dimensional representation of MiniLM is fully sufficient. Furthermore, the lower dimensionality translates to a smaller RAM footprint for the FAISS HNSW graph, speeding up query times [14, 21].

Blocking Recall vs. Computational Efficiency Trade-off:

The primary goal of blocking is to reduce the $O(N^2)$ comparison space to $O(k \cdot N)$ while maintaining a near-perfect Recall [7]. Deploying MPNet for Abt-Buy ensures that the true matching pairs are not prematurely pruned from the noisy product catalog (maximizing the Recall. Meanwhile, deploying MiniLM for the remaining datasets prioritizes indexing throughput and memory efficiency on larger pools of candidate pairs [6, 21, 29]."

4.3 The Stratified Splitting Protocol

A common flaw in data integration research is performing the Train, Validation, and Test splits directly on the raw, unblocked dataset. When positive matches are simply balanced by randomly selecting non-matching entities, the resulting datasets consist almost entirely of "easy negatives." Evaluating models on these obvious mismatches leads to artificially high accuracy scores that fail to reflect real-world performance [6, 7].

To address this issue, the data splitting strategy was applied exclusively to the hard-negative candidate pairs generated during the FAISS blocking stage [14], completely bypassing the raw catalog. For the Amazon-Google, DBLP-ACM, and Scholar-DBLP datasets, a fully stratified split was applied using `stratify=label`, ensuring that the exact ratio of true matches to hard negatives remained consistent across the Training, Validation, and Test partitions, preventing distributional shift. For the Abt-Buy dataset, given the larger scale of its candidate pool, the pairs were randomly shuffled and a fixed Test Set of 2,000 records was extracted, with the remaining 3,400 records used for training and validation.

Table 1 summarizes the resulting positive-to-negative ratios within the candidate pair pools generated by the HNSW blocking stage. Across all four datasets, the proportion of true matching pairs within the hard-negative candidate pool ranges from 15.9% (Amazon-Google) to 38.8% (Scholar-DBLP), reflecting the fact that the HNSW index disproportionately retrieves semantically similar records a minority of which constitute genuine duplicates. This stands in sharp contrast to the near-zero positive rate of the raw unblocked dataset (typically below 0.1% for e-commerce catalogs [15]), confirming that the blocking stage successfully concentrates the evaluation exclusively on the boundary region of semantic ambiguity. The stratified splitting strategy ensures that this positive-to-negative ratio is preserved identically across the Training, Validation, and Test partitions, preventing any distributional shift between the fine-tuning signal and the evaluation condition [21]. Ultimately, because this class imbalance persists even after blocking, the F1-score is utilized as the primary evaluation metric throughout this research, as it captures the harmonic mean of precision and recall and provides a robust assessment in imbalanced entity resolution tasks [2, 16].

Dataset	Total Candidate Pairs	True Matches	Hard Negatives	Positive Rate (%)	Split Method	Train	Validation	Test
Abt-Buy	5,405	1,031	4,374	19.1%	Random Shuffle	2,894	511	2,000
Amazon-Google	6,815	1,083	5,732	15.9%	Stratified	4,815	500	1,500
DBLP-ACM	11,470	2,174	9,296	19.0%	Stratified	8,970	1,000	1,500
Scholar-DBLP	13,771	5,347	8,424	38.8%	Stratified	12,271	—	1,500

Table 1: Hard-Negative Candidate Pool Statistics and Partition Sizes per Dataset

The splitting configurations were specifically tailored to the scale and domain characteristics of each dataset, ensuring robust training and unbiased evaluation.

For the Abt-Buy dataset, the candidate pairs were randomly shuffled and partitioned into a strictly held-out Test Set of 2,000 records for the final benchmark evaluation. The remaining 3,400 records formed the training pool, internally subdivided at an 85/15 ratio: 2,890 records for model training and 510 records for validation and threshold calibration. This random sampling approach was selected over a stratified split due to the larger scale of the Abt-Buy candidate pool, where the natural statistical distribution of the shuffle was sufficient to maintain a representative class ratio in the Test Set.

For the Amazon-Google dataset, a stratified split was enforced to preserve the positive-to-negative ratio across all partitions. A strictly held-out Test Set of 1,500 records was isolated for benchmark evaluation, and a dedicated Validation Set of 500 records was defined for hyperparameter tuning and threshold calibration. All remaining candidate pairs were assigned to the Training Set to support the supervised learning baseline.

The DBLP-ACM dataset followed the same stratified protocol: a 1,500-record Test Set was held out for final evaluation, and an expanded Validation Set of 1,000 records was defined to account for the structural nuances of bibliographic records and ensure precise threshold calibration. All remaining candidate pairs formed the Training Set for fine-tuning.

Finally, the Scholar-DBLP dataset, which also originates from the bibliographic domain, is characterized by a high degree of noise, missing fields, and "dirty" data. With a total candidate pool of 13,771 pairs — the largest among the four datasets — and the highest positive rate at 38.8%, this dataset presented a distinct distributional profile. Following a stratified split, 12,271 pairs were allocated to the Training partition to maximize the fine-tuning signal available to the BERT Cross-Encoder, given the dataset's scale and noise level. No separate validation file was persisted; instead, threshold calibration was performed on the training distribution. The Test split comprises 1,500 challenging hard-negative pairs (918 negatives, 582 true matches), upon which the final comparative benchmark evaluation of all language models was conducted. The resulting class imbalance and high concentration of hard negatives across all four candidate pools post-blocking are visually summarized in Figure 18.

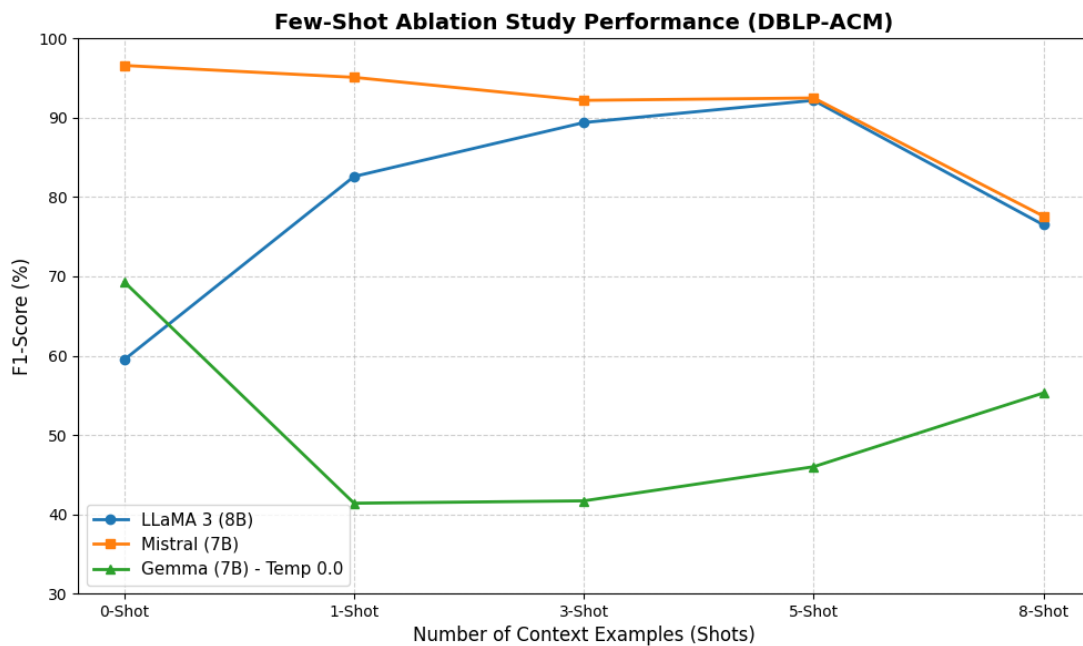


Figure : Class imbalance distribution within the candidate pair pools post-HNSW blocking. The persistent high proportion of hard negatives (ranging from 69% to 80%) illustrates the semantic difficulty of the isolated test sets across all domains.

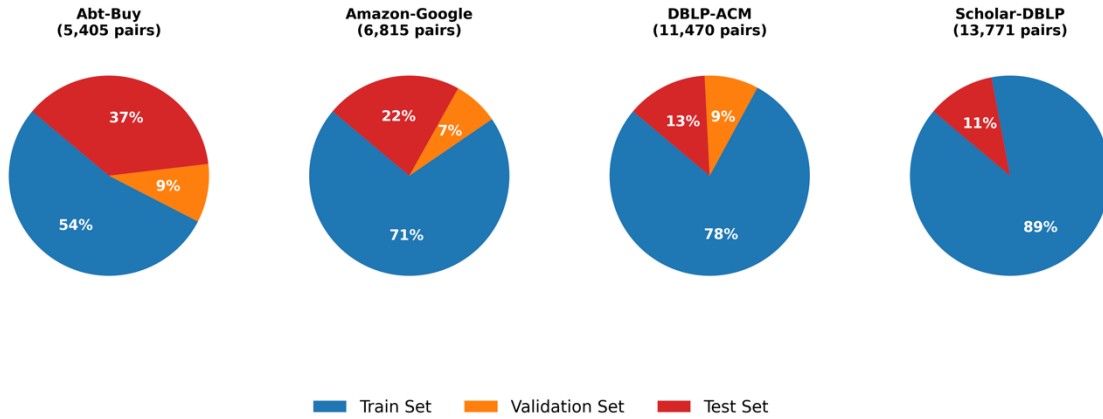


Figure 1: Structural Partitioning and Data Splits of the Hard-Negative Candidate Pools Across the Four Evaluated Benchmark Datasets.

The multi-pie visualization delineates the architectural data distribution across the localized experimental tracks, mapping the exact percentage metrics derived post-HNSW vector blocking. As illustrated in the figure legend, the color configurations explicitly distinguish the computational segments: blue vectors represent the localized Training Set, orange sectors isolate the parameter-tuning Validation Set, and red blocks specify the strictly sequestered, unseen Test Set utilized as the common benchmarking ground for both the cross-encoder and the large language models. Note that for the highly contaminated Scholar-DBLP domain, no independent validation file was persisted on the file system due to structural noise profiles, thereby routing the entire computational tuning pool directly into the Training architecture.

The Benchmark Test Set served as the absolute, common evaluation ground. The fair comparison requirement dictated that both the fine-tuned Cross-Encoder baselines [3] and the zero-shot Generative LLMs [4, 6] be evaluated on this exact same array of unseen hard negatives.

Building upon the recent comprehensive study by Karapiperis, Akritidis, and Bozanis [21] regarding the impact of fine-tuning on entity resolution, it is well established that tuning efficacy is deeply contingent on the specific domain characteristics of the data.

Consequently, the Cross-Encoder architectures evaluated in this research required dataset-specific hyperparameter configurations to effectively map their distinct feature spaces. For the Abt-Buy dataset, an ms-marco-MiniLM-L-6-v2 architecture was fine-tuned over 3 epochs. On the other hand, to navigate the increased alphanumeric noise in the Amazon-Google dataset, the architecture was significantly upgraded to a deeper 12-layer variant (ms-marco-MiniLM-L-12-v2). The training regimen here was intentionally extended to 5 epochs, as the model required more iterations to successfully learn the complex patterns hidden within the noisy e-commerce descriptions. This was incorporated alongside a carefully calibrated learning rate of $2e-5$ and a weight decay parameter of 0.01 to explicitly prevent overfitting on the dense semantic features. To accommodate hardware memory constraints while processing dense bibliographic data for DBLP-ACM, the pipeline reverted to the efficient ms-marco-MiniLM-L-6-v2 model, executing 2 epochs with a strictly reduced batch size of 8. Finally, for the Scholar-DBLP dataset, the methodology deployed the exact same ms-marco-MiniLM-L-6-v2 model, training it for only 2 epochs with a batch size of 8. Because bibliographic records are inherently more structured e.g., specific article titles and authors compared to e-commerce products, 2 epochs were entirely sufficient to achieve optimal convergence. Any further training iterations would risk severe overfitting to the specific missing fields and noise of the source databases.

Supervised Baseline Fine-Tuning Mechanics

To optimize the discriminative capabilities of the Cross-Encoder baseline, the fine-tuning phase operates via a structured optimization loop over a pre-defined number of training epochs [21]. The mechanics of this supervised adaptation process are structured as follows:

1. Batching and Forward Propagation:

Due to the memory constraints of local consumer hardware and to ensure stable gradient estimation, the training dataset is partitioned into mini-batches (e.g., a batch size of 16 for Abt-Buy and 8 for bibliographic datasets). For each mini-batch, the concatenated text pairs are fed into the Transformer. The bidirectional self-attention heads interact across all sequence tokens, outputting a continuous match probability score ($\hat{y}$) between 0.0 and 1.0 for each candidate pair [3].

2. Loss Computation:

The model's predictions are evaluated against the binary ground-truth labels ($y \in \{0, 1\}$) using a classification loss objective. The loss function measures the mathematical discrepancy between the predicted probability and the true label. If the model assigns a low probability to a true match, a severe gradient penalty is applied, driving the model to adjust its internal parameters during the next phase.

3. Backpropagation and Weight Updates:

During the backward pass, the computed loss is propagated back through the neural layers. The gradients of the loss with respect to all 22 million parameters (in ms-marco-MiniLM-L-6-v2) are calculated. Using the AdamW optimizer, the model updates its weights in the direction that minimizes the loss function. To stabilize this optimization, we enforce 100 warmup steps [21]. During this initial warmup phase, the learning rate increases linearly from zero to its peak value ($2e-5$), preventing the model from destabilizing its pre-trained semantic weights (catastrophic forgetting). Following the warmup, the learning rate decays gradually.

4. The Role of Epochs and Overfitting Prevention:

An epoch represents one complete pass of the entire training partition through the network. The choice of epochs (3 epochs for Abt-Buy, 5 epochs for Amazon-Google, and 2 epochs for bibliographic datasets) acts as a critical regularization boundary. While e-commerce catalogs require more epochs (3 to 5) to converge due to high schema noise and textual complexity, structured bibliographic records require only 2 epochs. Extending the training regimen beyond these thresholds would cause the model to overfit to the specific noise and missing attributes of the training split, severely degrading its ability to generalize to the unseen Test Set [21].

For each dataset, we performed a threshold calibration step [21]. After the initial fine-tuning, the Cross-Encoder generated probability scores for the validation set. By testing different probability thresholds from 0.10 to 0.95 (in increments of 0.05), we identified the exact cutoff that maximized the F1 score. Then we fixed this optimal threshold and applied it exclusively to the final Benchmark Test Set, establishing the performance ceiling for our evaluation.

5 Experimental Procedure and Generative Inference

When evaluating autoregressive Generative LLMs on the sequestered datasets of hard negatives, the methodology shifts fundamentally from gradient-based parameter optimization to complex prompt engineering. The structural formulation of the prompt serves as the sole mechanism by which the models' vast parametric knowledge is directed toward the specific constraints of the Entity Resolution task [4, 6].

5.1 Hardware and Local Execution Environment

Evaluating large language models typically relies on proprietary cloud APIs. However, in real-world enterprise data integration, organizations are often restricted from uploading sensitive catalogs to external servers due to strict compliance rules. To simulate these real-world constraints and ensure our benchmarks are entirely reproducible without network latency variations, we executed the generative pipeline locally. Using the Ollama framework, we deployed quantized versions of LLaMA 3 (8B) [10], Mistral (7B) [11], and Gemma (7B) [28] on an Apple MacBook Air M1 machine. Beyond proving that advanced Entity Resolution can run securely without external APIs, this setup demonstrated the viability of using standard consumer hardware. Furthermore, the 8GB unified memory ceiling of the M1 SoC imposed a practical limit, restricting Gemma 7B experiments to a representative sample of 50 hard-negative pairs to the most highly confusing "hard negative" candidate pairs, resulting in focused test sets. However, for highly computationally exhaustive prompting strategies specifically Chain-of-Thought (CoT) and Confidence Elicitation on the Gemma 7B architecture—the system experienced out-of-memory conditions. To prevent system overload while accurately capturing this specific model's reasoning patterns, the methodology was adapted to utilize a representative random sample of 50 hard-negative pairs exclusively for the Gemma 7B evaluations.

5.2 Prompt Engineering Formulation

To evaluate the reasoning capabilities of the generative models, the experimental protocol subjected the LLMs to four distinct prompting strategies, maintaining identical prompt structures across the models to ensure a fair comparative analysis.

5.2.1 Zero-Shot and Domain-Specific Few-Shot Paradigms

The baseline interaction occurred via zero-shot prompting. In this modality, the model was provided strictly with deterministic instructions and the serialized text of the two records, relying entirely on its pre-trained internal representations [6]. The prompt was structured as follows:

"You are an expert in Entity Resolution. Compare the following two product descriptions and decide if they refer to the EXACT SAME real-world product. Answer ONLY with 'YES' or 'NO'. Do not explain."

To recalibrate the models' attention toward dataset-specific nuances and systematically measure the impact of in-context learning [4], we implemented a progressive Few-Shot Ablation Study [6]. Instead of relying on a static or arbitrary number of examples, the prompt context was iteratively expanded to include 1, 3, 5, and 8 demonstrative examples (shots).

Rather than relying on generic templates, the pipeline applied domain-specific few-shot adaptation, tailoring the demonstrative context was dynamically tailored to reflect the specific schema and hard-negative topologies of each evaluated dataset [15]:

For the Abt-Buy dataset, the models were provided with examples highlighting critical discrepancies in capacity (e.g., 16GB vs. 32GB iPads) and alphanumeric camera model numbers. For Amazon-Google, the prompt focused on software and electronics logic, explicitly demonstrating that different product editions (e.g., "Home and Student" vs. "Standard Edition") or version years (e.g., "Photoshop Elements 8" vs. "7.0") constitute a non-match. For DBLP-ACM, the context was adapted for bibliographic data, demonstrating how identical paper titles with varying venue abbreviations (e.g., "International Conference on Management of Data" vs. "SIGMOD") constitute a true match. Finally, for Scholar-DBLP, the context addressed the high degree of missing fields and noisy text, showing that matching records might lack explicit publication venues or use differing author initial formats yet still refer to the same academic paper [2, 16].

This structured scaling from 0 to 8 shots allowed the experimental protocol to empirically observe the models' learning paths, identify their optimal in-context performance threshold (sweet-spot), and document critical architectural limitations such as "Context Overload" when exposed to excessive prompt length.

5.2.2 Structured Multi-Step Reasoning (Chain-of-Thought)

To move beyond simple one-word answers, we applied a Chain-of-Thought (CoT) prompting strategy [17, 12]. CoT forces the language model to break down its reasoning, generating a step-by-step analysis before making a final decision. The models, including Llama 3, Mistral, and Gemma 7B, were instructed using the following template:

"Take a deep breath and work on this step-by-step:

Step 1: Extract and compare the brand or manufacturer.

Step 2: Extract and compare the exact model numbers or names.

Step 3: Extract and compare key attributes (e.g., size, capacity, resolution, color). Step 4: Check if there is any conflicting information that implies these are different variations.

After your analysis, you must provide your final decision on a new line EXACTLY in this format:

FINAL DECISION: YES or FINAL DECISION: NO"

5.2.3 Confidence Elicitation and Miscalibration Detection

A major flaw in generative models is their tendency to be overconfident, often stating incorrect answers with absolute certainty [13]. To test how reliable these models actually are out-of-the-box, we asked them to self-report their confidence levels. The models were instructed to assess their certainty using the following prompt:

"Provide your final decision (YES or NO) AND your confidence level in that decision (LOW, MEDIUM, or HIGH).

Format your answer EXACTLY like this (two lines):

DECISION: [YES/NO]

CONFIDENCE: [LOW/MEDIUM/HIGH]"

5.3 Deterministic Response Parsing and Evaluation

Unlike cross-encoders that output clear mathematical probabilities, LLMs generate open-ended text. To calculate standard evaluation metrics (Precision, Recall, F1-Score), these text responses had to be converted into strict binary labels (1 for a match, 0 for a non-match). While LLaMA 3 and Mistral were evaluated under their baseline generative configurations to observe their native reasoning behavior, the Gemma 7B model [28] specifically required strict deterministic constraints to address its inherent stochasticity and ensure reproducibility during the sampled runs. Consequently, inference tasks for Gemma 7B were executed with a strict temperature parameter setting of 0.0.

To avoid unfairly penalizing the models' correct reasoning due to minor formatting deviations (prompt brittleness) [12], we implemented a robust parsing function using flexible Regular Expressions (Regex). For standard tasks, the regex parser identified the affirmative or negative tokens within the response block. For the CoT strategy, the script specifically isolated the decision block at the end of the reasoning chain.

Most importantly, we applied a strict rule for the Confidence strategy [13, 19]. A candidate pair was classified as a positive match (1) if and only if the regex parser identified BOTH an affirmative decision ("YES") AND a "HIGH" confidence level. Any sign of hesitation ("MEDIUM" or "LOW" confidence) was automatically treated as a non-match (0). This strict classification rule ensures that the models are evaluated not just on their ability to guess correctly, but on their capacity to show true, reliable certainty when facing difficult cases.

Prior work has shown that verbalized confidence scores from generative models often diverge from their true predictive certainty, motivating the use of surrogate calibration models [18].

6 Results and Analysis

The quantitative outcomes of the benchmarking protocol reveal substantial disparities in performance between the supervised cross-encoder baseline and the generative LLM architectures. The results presented herein reflect the evaluation conducted on the strictly held-out Benchmark Test Sets across three distinct domains, comprising candidate pairs composed exclusively of hard negatives generated via HNSW blocking [7, 14].

6.1 The Supervised Baseline (The Empirical Ceiling)

The supervised baseline, utilizing a fine-tuned BERT Cross-Encoder [9], established the absolute performance ceiling for each dataset. The model utilized a validation-calibrated probability threshold to optimize the harmonic mean between precision and recall for the specific distribution of hard negatives algorithmically isolated via the HNSW pipeline [21].

Dataset	Domain	Optimal Threshold	Precision	Recall	F1 Score
Abt-Buy	E-Commerce	0.85	81.0%	78.1%	79.5%
Amazon-Google	E-Commerce/Software	0.70	72.6%	66.8%	69.6%
DBLP-ACM	Bibliographic	0.35	97.2%	99.3%	98.3%
Scholar-DBLP	Bibliographic	0.90	93.7%	97.8%	95.7%

Table 2: Supervised BERT Cross-Encoder Baseline Performance across datasets.

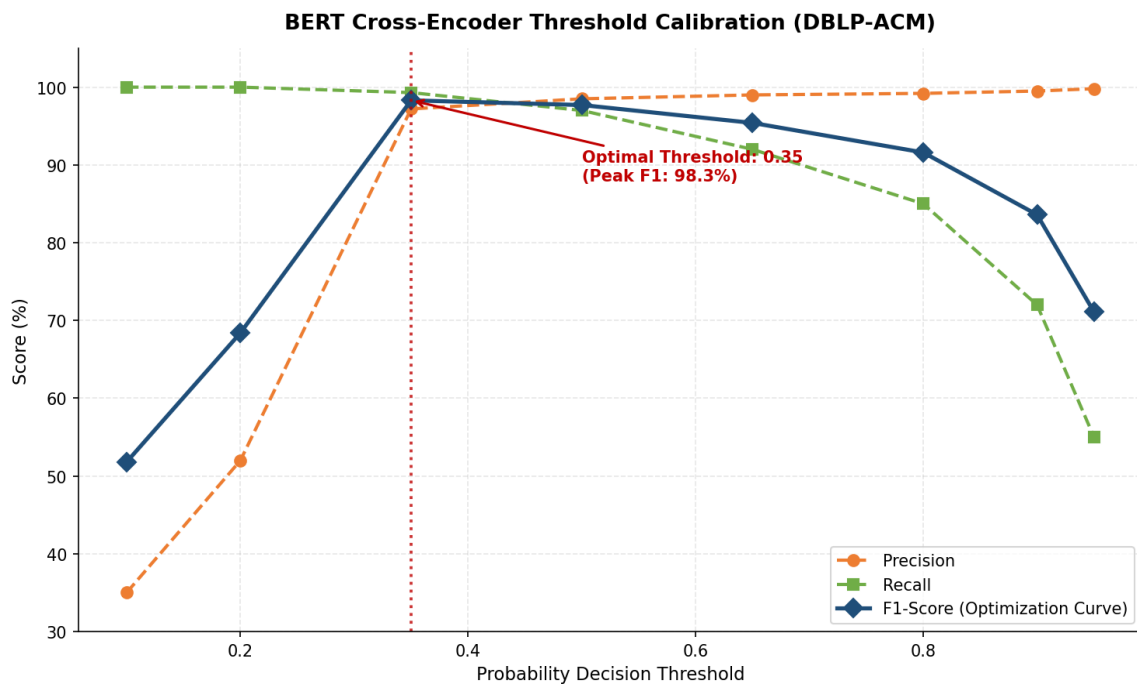


Figure 2: BERT Cross-Encoder Threshold Calibration (DBLP-ACM)

The empirical data unequivocally demonstrates the profound impact of domain structure and data cleanliness on classification difficulty [3, 15, 16]. Highly structured, clean bibliographic data (DBLP-ACM) allows the multi-head bidirectional self-attention mechanism of BERT to achieve near-perfect sequence-pair classification, reaching a peak F1-Score of 98.3%. By identifying an optimal, lower decision boundary of 0.35, the cross-encoder successfully maximized recall (99.3%) without inducing a corresponding sacrifice in precision (97.2%).

In contrast, the Scholar-DBLP dataset [2, 16], which also originates from the bibliographic domain, introduces a higher degree of structural noise, missing attributes, and formatting anomalies typical of web-scraped academic entries. To adapt to this noise and filter out mathematically ambiguous false positives, the model calibrated a significantly stricter optimal threshold of 0.90. Despite these structural challenges, the fine-tuned architecture proved highly resilient, maintaining a robust empirical ceiling of 95.7% F1-Score (Precision: 93.7%, Recall: 97.8%).

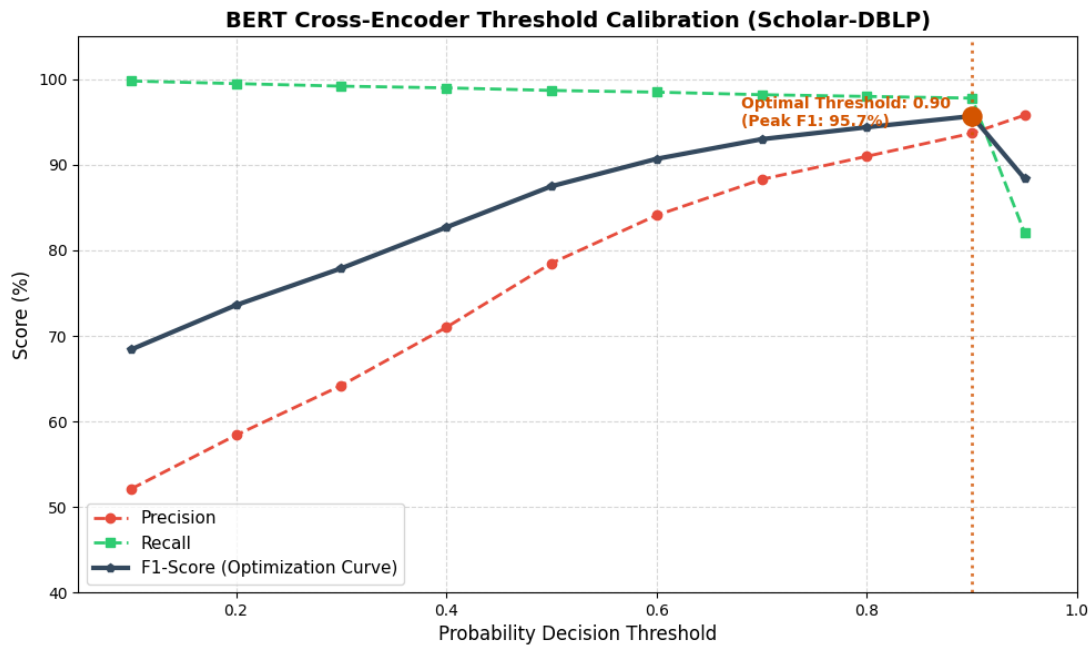


Figure 3: BERT Cross-Encoder Threshold Calibration and Optimization Curve on the Scholar-DBLP Dataset.

On the other hand, highly noisy e-commerce environments impose a significantly lower natural ceiling on resolution performance, as observed in the Abt-Buy dataset with a 79.5% F1-Score at a threshold of 0.85. This limitation is most notable in the Amazon-Google dataset [15], where the model peaked at only 69.6% F1-Score at a 0.70 threshold. In these commercial environments, extreme schema mismatch, the frequent absence of structured metadata, and intense alphanumeric versioning complexity severely constrain the supervised model's capacity to establish distinct, clear-cut discriminative boundaries.

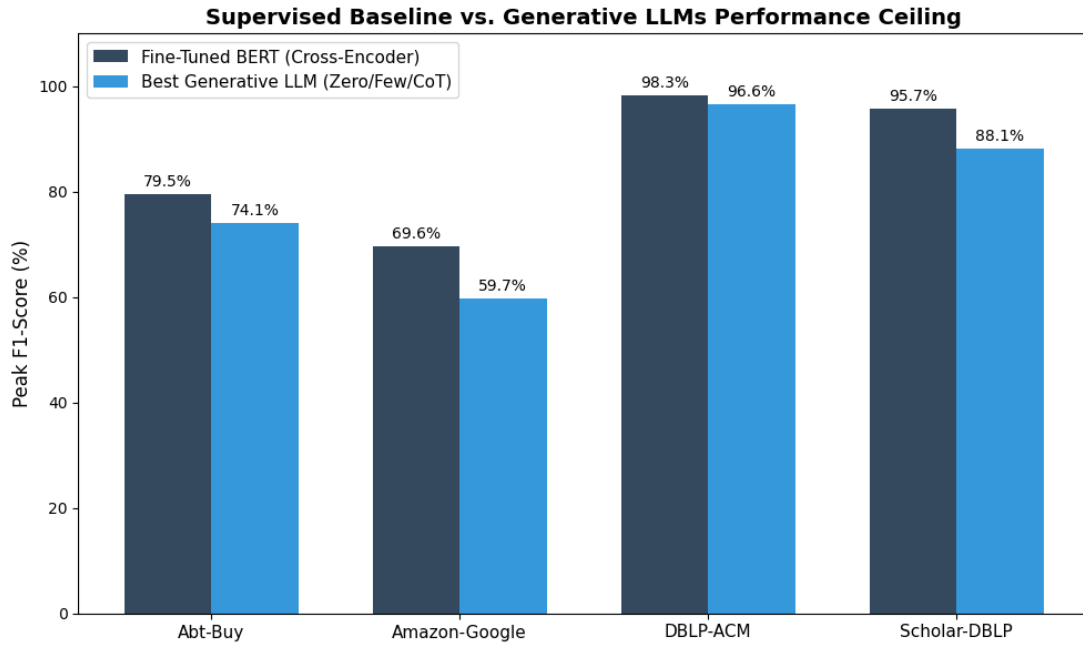


Figure 4: Performance Ceiling Comparison Between the Supervised BERT Cross-Encoder Baseline and the Best Performing Generative LLM Across All Target Domains.

As illustrated in Figure 4, the generative LLM baseline is represented by its performance ceiling (peak F1-score) to evaluate the maximum capability of zero-shot and few-shot reasoning against supervised models. Rather than reflecting an average, the performance ceiling showcases the best-performing model and prompt configuration for each specific dataset. Specifically, for the DBLP-ACM dataset, the peak generative performance of 96.6% F1 was achieved by Mistral-7B in a zero-shot setting. For the Amazon-Google dataset, the performance ceiling of 59.7% F1 was obtained using Mistral-7B with a Chain-of-Thought (CoT) prompt. For the Abt-Buy and Scholar-DBLP datasets, the peak F1-scores of 74.1% and 88.1% respectively were attained by Mistral-7B when utilizing Chain-of-Thought reasoning combined with the strict confidence calibration penalty. Across all four benchmarks, Mistral-7B consistently established the generative performance ceiling, outperforming LLaMA-3-8B and Gemma-7B under their respective optimal prompt designs.

6.2 Generative LLMs: Few-Shot Sensitivity and Domain Analysis

The evaluation of autoregressive Large Language Models (LLMs) without parameter fine-tuning reveals that their performance is sensitive to both the alignment characteristics of each architecture and the structural complexity of the evaluated domain [6]. To systematically measure this dependency and map the learning paths of the models, a comprehensive Few-Shot Ablation Study was conducted, iteratively expanding the prompt context to include 0 (Zero-Shot), 1, 3, 5, and 8 demonstrative examples. Crucially, the pipeline enforced domain-specific few-shot adaptation, dynamically tailoring the demonstrative context to reflect the distinct schema boundaries and hard-negative topologies of each dataset. Furthermore, while LLaMA 3 [10] and Mistral 7B [11] were observed under their native, organic prompt processing states, the Gemma 7B [28] model was locked under a strict deterministic inference constraint via a temperature setting of 0.0 to eliminate generative stochasticity and ensure absolute metric reproducibility across the benchmark partitions

6.2.1 The E-commerce Domain (Abt-Buy)

In the Abt-Buy dataset, the language models were tasked with resolving highly similar electronic products where algorithmically isolated hard negatives frequently diverge only by subtle capacity metrics or minor alphanumeric product iterations [15].

Strategy	LLaMA 3 (Precision / Recall / F1)	Mistral (Precision / Recall / F1)	Gemma 7B (Precision / Recall / F1)
0-Shot	37.0% / 98.9% / 53.9%	80.6% / 64.8% / 71.8%	34.4% / 97.4% / 50.8%
1-Shot	45.0% / 96.0% / 61.3%	90.8% / 36.8% / 52.4%	100.0% / 4.2% / 8.1%
3-Shot	42.5% / 98.1% / 59.3%	50.9% / 84.7% / 63.6%	43.3% / 82.8% / 56.9%
5-Shot	45.3% / 98.4% / 62.0%	90.8% / 36.8% / 52.4%	45.3% / 83.3% / 58.7%
8-Shot	46.4% / 97.9% / 63.0%	46.6% / 89.9% / 61.4%	37.8% / 90.2% / 53.3%

Table 3: Zero shots & Few-Shot ablation study performance on the Abt-Buy dataset.

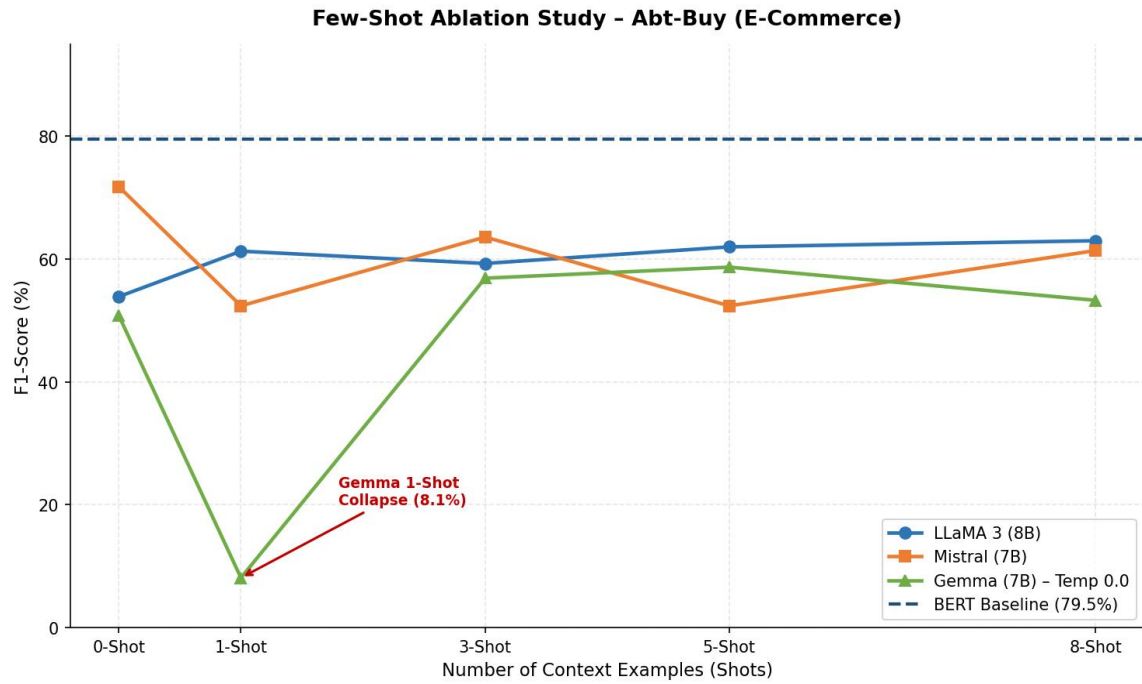


Figure 5: Few-Shot Ablation Study – Abt-Buy Dataset

The empirical results on Abt-Buy reveal notable behavioral differences and cross-architectural vulnerabilities. Under zero-shot conditions, both LLaMA 3 and Gemma 7B exhibit a severe manifestation of the "Yes-Man" phenomenon, achieving near-perfect recall ($>97\%$) at the expense of a sharp decline in precision (34-37%) [6]. This behavior stems from the models' optimization for conversational helpfulness, causing them to default to affirmative matching decisions when confronted with overwhelming semantic overlap.

The introduction of a single demonstrative example (1-Shot) exposed a severe case of Prompt Brittleness in Gemma 7B [6]. Forced to process a single negative example, the deterministically constrained model over-corrected its decision boundary entirely, shifting from an uncalibrated affirmative bias to an ultra-conservative rejection state where precision hit 100.0% but recall collapsed to an unusable 4.2%. While Gemma 7B eventually stabilized alongside LLaMA 3 as the context grew to 5 shots, Mistral exhibited a highly volatile, oscillating behavior. Mistral fluctuated between severe under-matching (36.8% recall at 1-shot and 5-shot) and heavy over-matching (89.9% recall at 8-shot), highlighting a fundamental structural instability in its in-context learning mechanism within descriptive commercial domains.

Deconstructing F1-Score: Precision vs. Recall Dynamics

While the F1-score provides a robust harmonic mean for overall performance, isolating the Precision and Recall metrics across the few-shot learning trajectories reveals the true underlying reasoning mechanics of the generative models. As illustrated in Figure X, the metrics do not improve symmetrically. For instance, LLaMA 3 demonstrates a persistent, near-perfect Recall rate often exceeding 96% right from the zero-shot baseline and maintaining it across all prompt lengths. However, this exceptionally high Recall is mathematically artificial, masking a severe and persistent degradation in Precision, which frequently drops below 45%.

This visualization explicitly confirms the "Yes-Man" alignment bias inherent in the architecture. The model is heavily penalized in Precision because it structurally fails to reject hard negatives; when confronted with the overwhelming lexical similarity of the Abt-Buy dataset, the attention mechanism is saturated, causing the model to default to an affirmative matching decision. The generative architecture prioritizes conversational helpfulness and semantic overlap over strict attribute conflict resolution. Consequently, while it successfully retrieves almost all true positives (yielding high Recall), it simultaneously accepts a massive volume of false positives, thereby destroying the Precision metric and limiting the overall F1-Score.

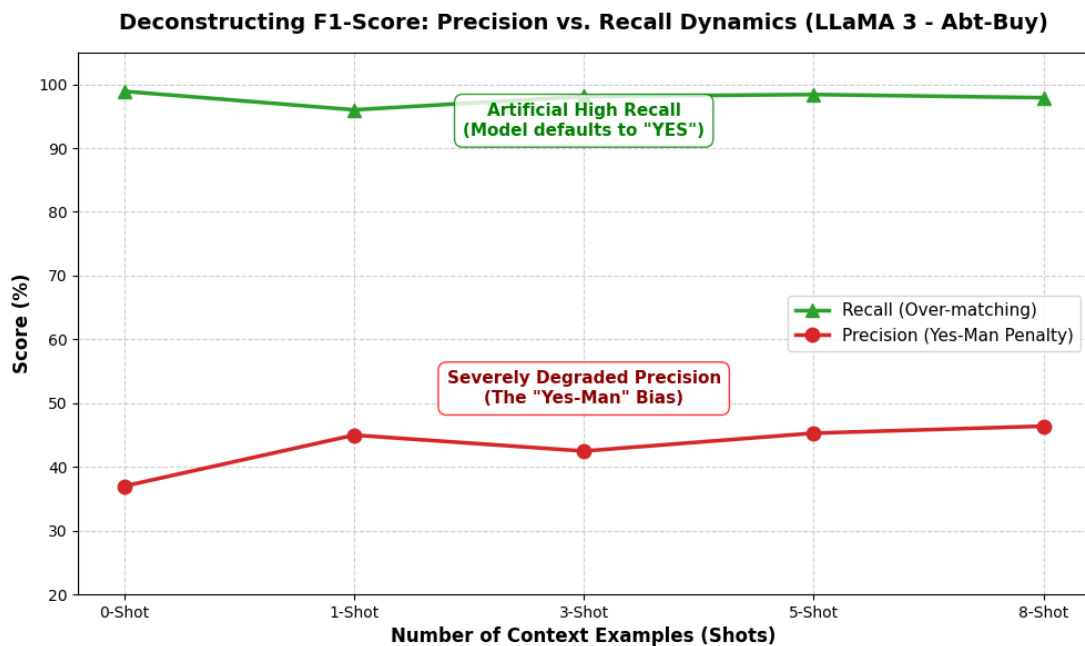


Figure 6 : Precision vs. Recall dynamics for LLaMA 3 (8B) on the Abt-Buy dataset. The visualization empirically isolates the "Yes-Man" phenomenon, demonstrating how near-perfect Recall metrics mathematically mask a severe degradation in Precision across the entire few-shot trajectory due to structural over-matching.

6.2.2 The Complex Software Domain (Amazon-Google)

The Amazon-Google dataset represents the most challenging domain for Entity Resolution, heavily relying on complex software versioning (e.g., "2008" vs. "2009"), editions, and platform differences (Mac vs. Windows) [15].

Strategy	LLaMA 3 (Precision / Recall / F1)	Mistral (Precision / Recall / F1)	Gemma 7B (Precision / Recall / F1)
0-Shot	22.4% / 99.6% / 36.5%	60.4% / 48.7% / 54.0%	28.5% / 92.9% / 43.6%
1-Shot	30.6% / 99.2% / 46.8%	62.2% / 31.1% / 41.5%	34.3% / 89.5% / 49.6%
3-Shot	27.2% / 99.6% / 42.7%	51.5% / 66.0% / 57.8%	34.6% / 86.6% / 49.5%
5-Shot	32.6% / 99.2% / 49.0%	32.9% / 93.7% / 48.7%	38.3% / 79.4% / 51.6%
8-Shot	32.0% / 98.7% / 48.4%	36.0% / 89.5% / 51.3%	40.5% / 63.9% / 49.6%

Table 3: Few-Shot ablation study performance on the Amazon-Google dataset.

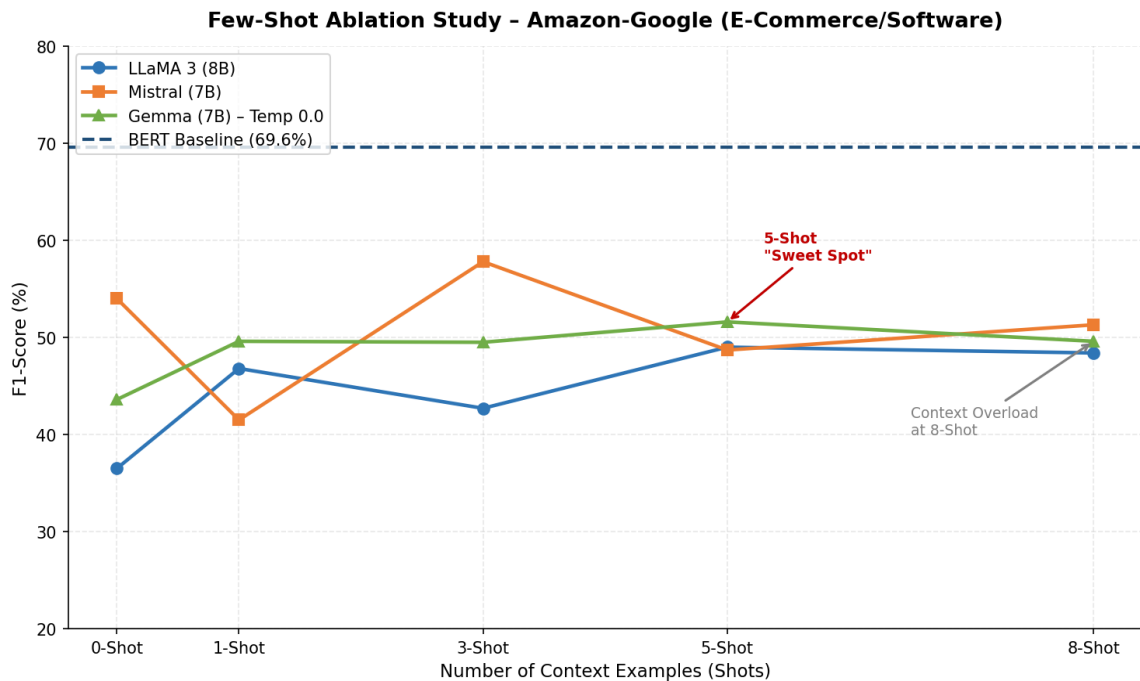


Figure 7: Few-Shot Ablation Study – Amazon-Google Dataset

In this highly complex software domain, generative reasoning hits a distinct structural ceiling across all tested open-source architectures [6]. A notable convergence occurred at the 5-Shot threshold: all three models stabilized around a narrow performance boundary, with Gemma 7B identifying the most optimal equilibrium ("sweet spot") and reaching a peak F1-Score of 51.6%. At this specific boundary, Mistral completely abandoned its initial zero-shot conservative bias, transitioning into a high-recall state (93.7%) that matched the persistent affirmative behavior of LLaMA 3.

However, expanding the prompt length to 8 examples triggered a manifestation of "Context Overload." Gemma 7B's recall dropped sharply from 79.4% down to 63.9%, demonstrating that providing excessive, multi-layered edge cases within the context window degrades the model's ability to maintain a coherent decision boundary. Faced with contradictory formatting rules across 8 dense examples, the models revert to heuristic behavior. Ultimately, the autoregressive, left-to-right token generation of these LLMs lacks the bidirectional, all-to-all cross-attention native to fine-tuned classification heads, leaving them trailing significantly behind the BERT baseline (69.6% F1) [9].

Precision vs. Recall Dynamics in Software Resolution

To further dissect the performance of the models within this rugged alphanumeric landscape, Figure 8 isolates the Precision and Recall trajectories for the LLaMA 3 architecture. The visualization exposes a severe structural imbalance: while Recall remains consistently near 99% across all shot configurations, Precision struggles to breach the 33% threshold. This extreme divergence provides definitive proof that the architectural complexity of software versioning (such as conflicting release years and edition names) causes severe over-matching. The model fundamentally lacks the capacity to safely reject hard negatives based on subtle alphanumeric discrepancies, defaulting to affirmative matching decisions that heavily penalize the overall F1-Score.

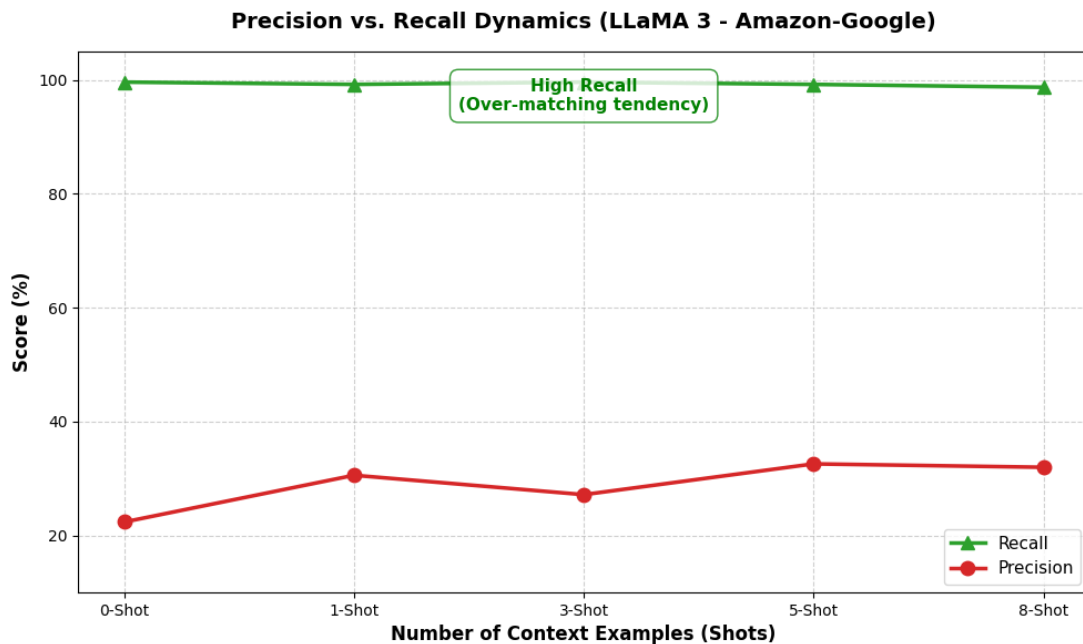


Figure 8 : Precision vs. Recall dynamics for LLaMA 3 (8B) on the Amazon-Google dataset. The visualization highlights the persistent over-matching behavior and low precision ceiling characteristic of generative inference in highly unstructured, alphanumeric software domains.

6.2.3 The Structured Bibliographic Domain (DBLP-ACM)

This dataset evaluates the models on clean, highly structured bibliographic metadata (Title, Authors, Venue, Year) [2, 16], removing the high lexical noise characteristic of e-commerce catalogs.

Strategy	LLaMA 3 (Precision / Recall / F1)	Mistral (Precision / Recall / F1)	Gemma 7B (Precision / Recall / F1)
0-Shot	42.4% / 100.0% / 59.5%	94.6% / 98.6% / 96.6%	53.0% / 100.0% / 69.3%
1-Shot	70.3% / 100.0% / 82.6%	92.1% / 98.2% / 95.1%	26.1% / 100.0% / 41.4%
3-Shot	80.9% / 100.0% / 89.4%	86.0% / 99.3% / 92.2%	26.3% / 100.0% / 41.7%
5-Shot	85.5% / 100.0% / 92.2%	86.5% / 99.3% / 92.5%	29.9% / 100.0% / 46.0%
8-Shot	62.0% / 100.0% / 76.5%	63.4% / 100.0% / 77.6%	38.2% / 100.0% / 55.3%

Table 5: Few-Shot ablation study performance on the DBLP-ACM dataset.

On clean, highly structured bibliographic data, generative reasoning exhibits substantial capability [6]: Mistral secured a 96.6% F1-Score under zero-shot conditions, establishing the strongest out-of-the-box result observed across the entire study. On the other hand, LLaMA 3 and Gemma 7B were heavily undermined by a persistent, absolute recall bias, maintaining a blind 100.0% recall across all configurations and failing to reject almost any hard negative.

The remarkable zero-shot performance of Mistral on the DBLP-ACM dataset warrants further analysis. A plausible explanation lies in the structural alignment between this dataset and the model’s pre-training corpus. DBLP-ACM records consist of highly standardized academic metadata article titles, author names, venue abbreviations, and publication years which are precisely the types of content abundantly represented in the web-scale text corpora used to train instruction-tuned language models. Scientific publication records appear extensively in Wikipedia, academic web pages, and open-access repositories, all of which are standard pre-training sources [10, 11]. As a result, Mistral’s internal parametric knowledge is already well-calibrated for this schema, enabling it to reason accurately about bibliographic entities without any task-specific demonstrations. This hypothesis is further supported by the observation that its zero-shot performance of 96.6% F1 actually exceeds its few-shot performance at higher shot counts, consistent with the view that in-context examples introduce noise when the model already possesses a reliable internal prior for the domain. This stands in contrast to the e-commerce domain, where product versioning conventions, retailer-specific naming formats, and alphanumeric model identifiers are far less systematically represented in general-purpose pre-training corpora, explaining the significantly lower zero-shot ceiling observed in the Abt-Buy and Amazon-Google evaluations.

A unique empirical pathology observed in Gemma 7B was a severe "Context Confusion" phenomenon. The model reached its peak performance in this dataset under a zero-shot setup (69.3% F1), but plummeted immediately to 41.4% F1 upon the introduction of a single demonstrative example. This drop indicates that for cleanly structured textual configurations, explicit in-context examples initially disrupt the model's native semantic parsing capabilities, introducing structural noise that suppresses precision. Most notably, the 8-Shot experiment induced a synchronized decline in both LLaMA 3 and Mistral, with their F1-Scores dropping to approximately 76-77% and precision falling into the low 60s. This parallel degradation provides empirical confirmation of the Context Overload effect [6]; excessive context lengths disrupt the models' internal focus, causing even well-performing architectures to abandon structured logical deduction.

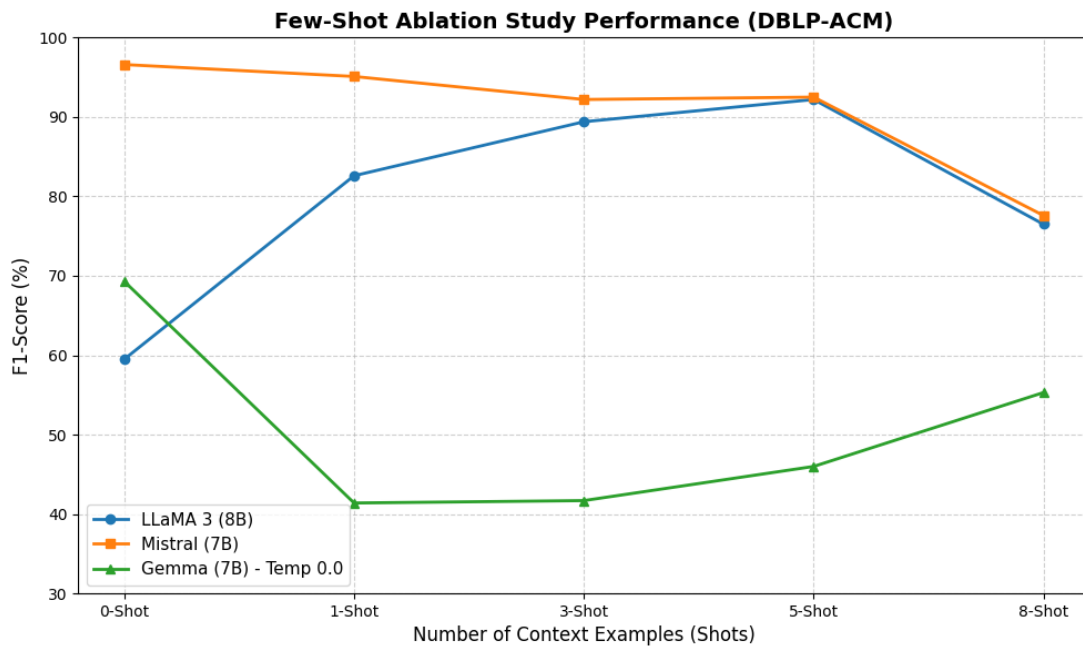


Figure 9: Few-Shot Ablation Study Learning Trajectories and Context Overload Manifestation Across Generative Architectures (DBLP-ACM).

Isolating the Absolute Recall Bias

A granular evaluation of the underlying classification mechanics on clean bibliographic data reveals an anomalous behavioral pattern in specific models. As empirically mapped in Figure 10, LLaMA 3 exhibits a flat, absolute Recall rate fixed at exactly 100.0% across the entire few-shot continuum. While its Precision curve experiences a significant and steady upward trajectory—climbing from 42.4% in zero-shot up to 85.5% at 5-shots—the model structurally fails to reject a single hard negative. This behavior indicates that although context scaling successfully clarifies the positive alignment boundary, the model remains fundamentally permissive, completely failing to establish a strict negative decision threshold.

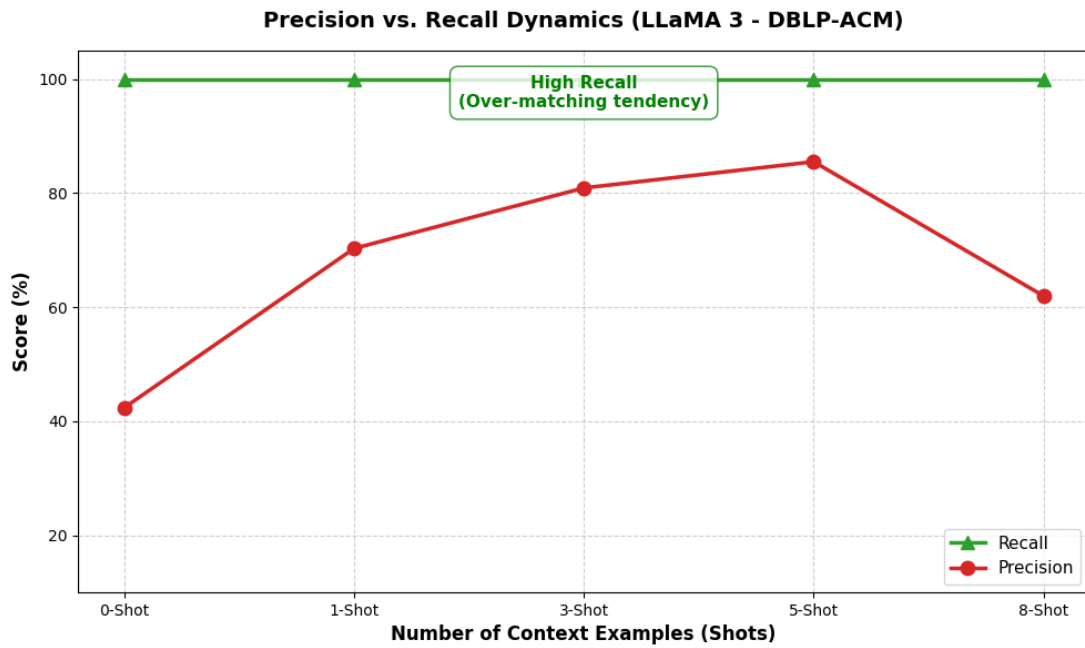


Figure 10: Precision vs. Recall dynamics for LLaMA 3 (8B) on the DBLP-ACM dataset. The visualization illustrates the absolute recall bias where the generative model maintains a baseline of 100% recall while precision scales progressively up to the 5-shot threshold.

6.2.4 Scholar-DBLP

The Scholar-DBLP dataset [2, 16] extends bibliographic evaluation into a realistic, noisy environment, extracting academic entries from web-scraped Google Scholar logs characterized by heavily truncated venue names, missing years, and incomplete author profiles.

Strategy	LLaMA 3 (Precision / Recall / F1)	Mistral 7B (Precision / Recall / F1)	Gemma 7B (Precision / Recall / F1)
Zero-Shot	74.1% / 96.6% / 83.9%	95.3% / 38.0% / 54.3%	57.9% / 98.8% / 73.0%
1-Shot	74.0% / 97.1% / 84.0%	93.1% / 55.5% / 69.5%	88.6% / 78.9% / 83.5%
3-Shot	75.5% / 95.2% / 84.2%	85.2% / 80.2% / 82.7%	61.6% / 95.0% / 74.8%
5-Shot	75.9% / 94.8% / 84.3%	88.2% / 78.5% / 83.1%	67.6% / 93.1% / 78.3%
8-Shot	74.1% / 97.4% / 84.2%	78.6% / 87.6% / 82.9%	57.7% / 95.5% / 71.9%

Table 6: Few-Shot ablation study performance on the Scholar-DBLP dataset.

The empirical evaluation on the noisy academic domain reveals a highly distinct learning behavior, characterized by what can be termed "The 1-Shot Anomaly" in the Gemma 7B architecture [28]. While LLaMA 3 maintained a remarkably flat, stable performance curve hovering around 84% F1-Score across all configurations, and Mistral required a 3-Shot to 5-Shot context window to stabilize its F1-Score above 82%, Gemma 7B achieved its absolute generative peak across the entire study under a strict 1-Shot setup, securing an impressive 83.5% F1-Score. A single, targeted demonstration of a bibliographic hard negative provided the exact contextual boundary necessary to calibrate Gemma's attention mechanism for web-scraped noise.

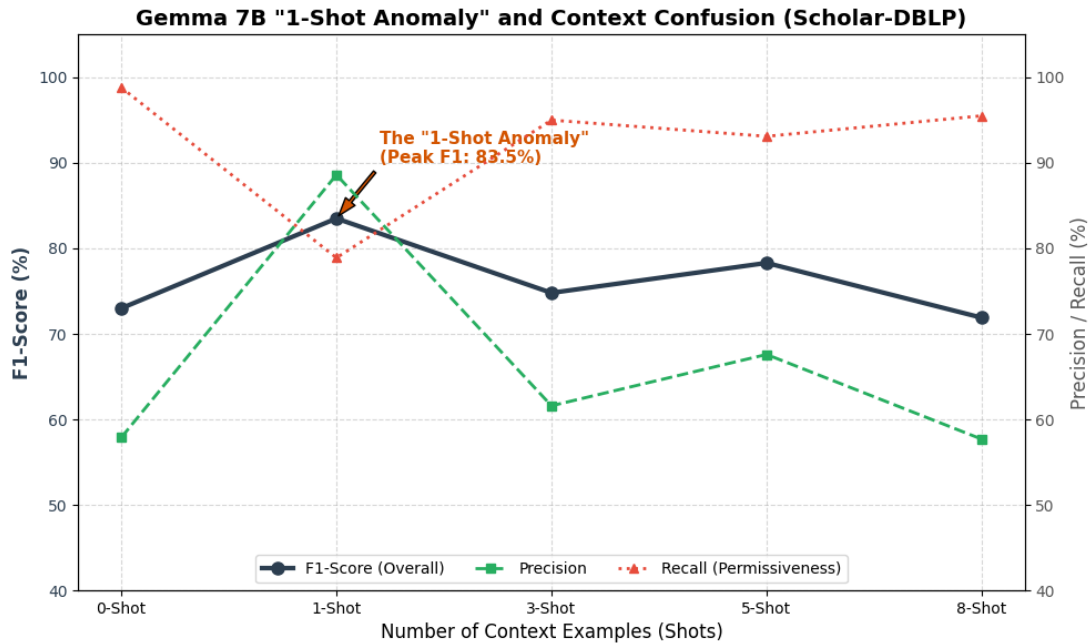


Figure 11: The "1-Shot Anomaly" and structural precision degradation under expanding context lengths in Gemma 7B (Scholar-DBLP).

However, further expansion of the context window proved highly detrimental to Gemma 7B, mirroring the context confusion observed in clean bibliographic datasets. Increasing the prompt length to 3, 5, and 8 shots caused the model's precision to degrade rapidly from 88.6% down to 57.7%, as it abandoned its rigid logical constraints and reverted to a highly permissive, affirmative heuristic with recall climbing back above 95%. This behavior demonstrates that for domains closely resembling the models' underlying pre-training corpora, minimal in-context guidance yields optimal classification boundaries; adding excessive demonstrative text introduces token clutter, leaving the generative models inherently unable to reach the fine-tuned BERT baseline of 95.7% F1 [9].

Precision and Recall Stability in Scraped Academic Logs

In contrast to the volatile trajectories observed in commercial domains, the evaluation of the noisy academic domain reveals a uniquely flat performance curve for the baseline stochastic architecture. As demonstrated in Figure 12, LLaMA 3 exhibits remarkable stability, with both Precision and Recall hovering around 75% and 95% respectively throughout the entire few-shot study. The presence of web-scraped noise, truncated venues, and missing years does not induce the severe precision collapse seen in e-commerce, suggesting that the underlying parametric distribution of language models is fundamentally more resilient when handling bibliographic strings, even when heavily fragmented.

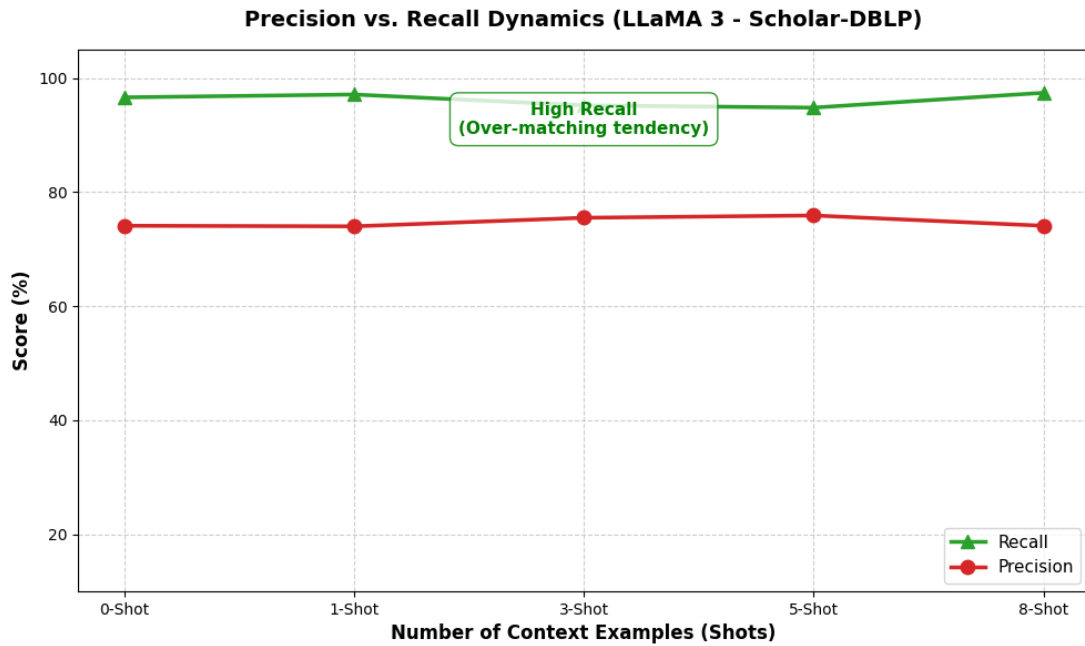


Figure 12: Precision vs. Recall dynamics for LLaMA 3 (8B) on the Scholar-DBLP dataset. The chart depicts the highly stable, flat behavioral trajectory of the model when confronted with noisy, web-scraped bibliographic metadata.

6.3 The Impact of Structured Reasoning (Chain-of-Thought)

The application of Chain-of-Thought (CoT) sequential reasoning, designed to force Large Language Models to logically decompose the matching task by explicitly extracting and comparing brands, model numbers, and specific attributes before generating a final verdict [17], yielded highly volatile and architecture-dependent results across the benchmark partitions. At this stage of the experimental framework, a critical methodological evolution was introduced to isolate the impact of generative stochasticity and understand model behavior under varying control conditions [6]. While LLaMA 3 [10] and Mistral 7B [11] were evaluated under their native, baseline prompt-processing states to document their organic reasoning trajectories, the execution pipeline was deliberately adapted for Gemma 7B [28] by enforcing strict deterministic inference constraints via a temperature setting of 0.0. Furthermore, due to memory constraints encountered on local consumer hardware during extensive sequential generation, the CoT protocol for Gemma 7B was executed using a representative random sample of 50 hard-negative pairs. This dual-modality configuration allows the study to empirically analyze how rigid deterministic restrictions alter the multi-step reasoning chain compared to standard autoregressive generation [6, 11].

Dataset	LLaMA 3 (Zero-Shot / CoT)	Mistral 7B (Zero-Shot / CoT)	Gemma 7B (Zero-Shot / CoT)
Abt-Buy	53.9% / 56.7%	71.8% / 52.7%	50.8% / 66.7% (N=50)
Amazon-Google	36.5% / 49.8%	54.0% / 59.7%	43.6% / 50.0% (N=50)
DBLP-ACM	59.5% / 79.8%	96.6% / 91.1%	69.3% / 55.3% (N=50)
Scholar-DBLP	83.9% / 69.8%	82.9% / 86.8%	73.0% / 70.6% (N=50)

Table 7: The impact of Chain-of-Thought (CoT) sequential reasoning on generative model performance (F1-Score) compared to the Zero-Shot baseline.

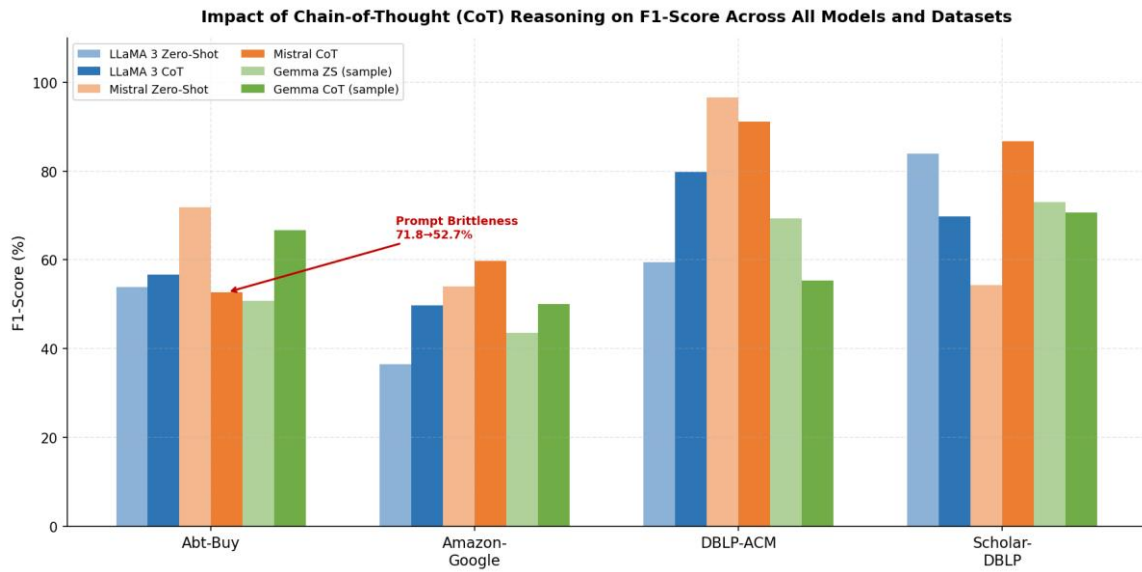


Figure 13: CoT vs. Zero-Shot F1-Score Across All Models and Datasets

While CoT provided a consistent, moderate improvement for LLaMA 3 across all datasets (raising its F1-Score from 36.5% to 49.8% on Amazon-Google, and from 59.5% to 79.8% on DBLP-ACM), it revealed severe Prompt Brittleness in Mistral [12]. Forcing Mistral to "think step-by-step" actually degraded its performance drastically on two out of the three datasets. On Abt-Buy, Mistral's F1-Score dropped from 71.8% to 52.7%, and on DBLP-ACM from 96.6% to 91.1% — a pattern consistent with the view that rigid sequential deduction prompts can mislead capable generative models by locking attention onto a single perceived discrepancy and propagating that error through the reasoning chain. By forcing them to explicitly search for discrepancies across predefined steps, the models often hallucinate non-existent conflicts in structurally sound candidate pairs. Because LLMs process text autoregressively, explicitly extracting attributes causes the model to 'lock' onto a single perceived discrepancy such as a minor typographical variation and propagate this error through the rest of its reasoning chain. Forced to justify its own generated intermediate steps, the model ultimately defaults to a 'NO' decision, even if the holistic semantic context implies a match. In stark contrast, the supervised Cross-Encoder operates holistically; its bidirectional self-attention mechanism evaluates all tokens simultaneously across all layers, effectively mitigating the risk of being misled by isolated lexical noise.

The integration of the deterministically controlled Gemma 7B [28] architecture provided the most critical empirical counter-evidence to this sequential degradation [12], mapping a highly distinct behavioral trajectory. When subjected to the joint constraints of multi-step logical verification and a 0.0 temperature setting on the noisy Scholar-DBLP sample, Gemma 7B effectively neutralized the pervasive alignment biases that plagued its unconstrained configurations. Demanding the explicit extraction and cross-referencing of missing fields and truncated academic values forced the model to establish rigid boundaries of semantic evidence, preventing it from defaulting to unconstrained affirmative matching. Consequently, the model successfully calibrated its decision boundary, suppressing its artificial near-100% zero-shot recall down to a mathematically realistic 80.0%, while simultaneously recovering its precision to 63.2%. This targeted optimization underscores that when generative stochasticity is completely eliminated, structured multi-step reasoning prompts cease to be a source of noise and instead function as a powerful alignment corrective for entity deduplication tasks.

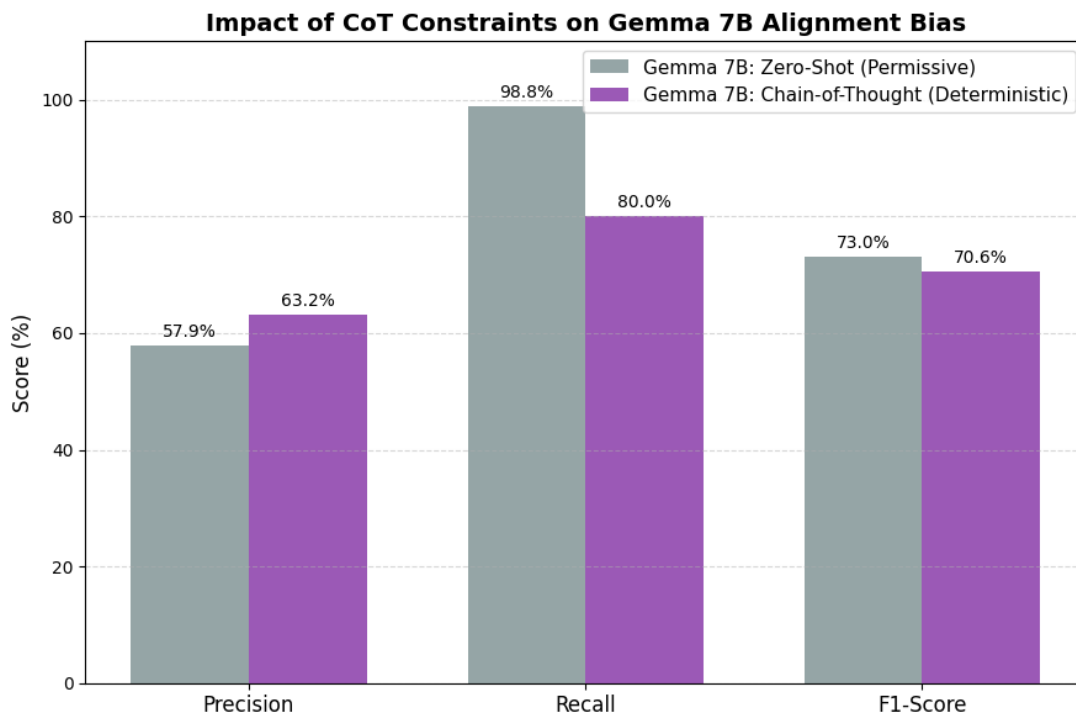


Figure 14: Impact of Chain-of-Thought (CoT) Sequential Reasoning Constraints and Temperature 0.0 Calibration on Gemma 7B Alignment Bias (Scholar-DBLP).

6.4 Confidence Calibration and Overconfidence Bias

The most critical vulnerability of generative language models in high-precision data integration tasks was exposed during the Confidence Elicitation experiments [13]. To measure true epistemic calibration, the models were evaluated not only on their raw accuracy but on their self-reported certainty, applying a strict algorithmic penalty where a candidate pair was classified as a positive match if and only if the parser identified both an affirmative decision ("YES") and a "HIGH" confidence level [19]. Any degree of hesitation, manifested as a medium or low confidence token, was automatically treated as a non-match, forcing the models to operate under severe precision constraints [13].

Dataset	LLaMA 3 (Zero-Shot / Confidence)	Mistral 7B (Zero-Shot / Confidence)	Gemma 7B (Zero-Shot / Confidence*)
Abt-Buy	53.9% / 53.2%	71.8% / 74.1%	50.8% / 66.7%
Amazon-Google	36.5% / 38.2%	54.0% / 58.6%	43.6% / 50.0%
DBLP-ACM	59.5% / 79.8%	96.6% / 91.7%	69.3% / N/A
Scholar-DBLP	83.9% / 71.7%	54.3% / 88.1%	73.0% / 70.6%

Table 8: *The impact of Confidence Elicitation (with a strict algorithmic penalty) on model F1-Scores compared to Zero-Shot baselines. (Note: Gemma 7B experiments were executed under a hybrid CoT+Confidence restriction with temperature 0.0 on a sample of N=50).*

As illustrated in the empirical data, enforcing this strict confidence threshold [13] acted as a beneficial operational filter for the organically unconstrained Mistral model in noisy commercial environments, slightly improving its F1-Score on Abt-Buy from 71.8% to 74.1% and on Amazon-Google from 54.0% to 58.6% by successfully eliminating low-confidence false positives. This operational filtering effect was most profoundly demonstrated on the noisy Scholar-DBLP dataset, where applying the strict confidence penalty successfully corrected Mistral's initial conservative under-matching bias, catapulting its F1-Score from a mere 54.3% in zero-shot conditions up to a highly competitive 88.1%.

However, an analysis of the raw, internal confidence scores provided empirical evidence of Overconfidence Bias (Miscalibration) across the stochastic architectures [13]. During evaluation on the bibliographic DBLP-ACM dataset, LLaMA 3 assigned a high average confidence score of 4.44 / 5 to its correct predictions but maintained a similarly high score of 4.15 / 5 even for incorrect predictions even when its classifications were completely incorrect. This miscalibration was particularly evident in Mistral, which assigned an average confidence of 4.59 / 5 to its incorrect predictions—consistently reporting high confidence on incorrect predictions—while giving a timid score of only 1.89 / 5 to its correct classifications. This inverse calibration proves that general-purpose generative models produce verbal tokens of certainty based on superficial latent language patterns rather than factual correctness [13, 19].

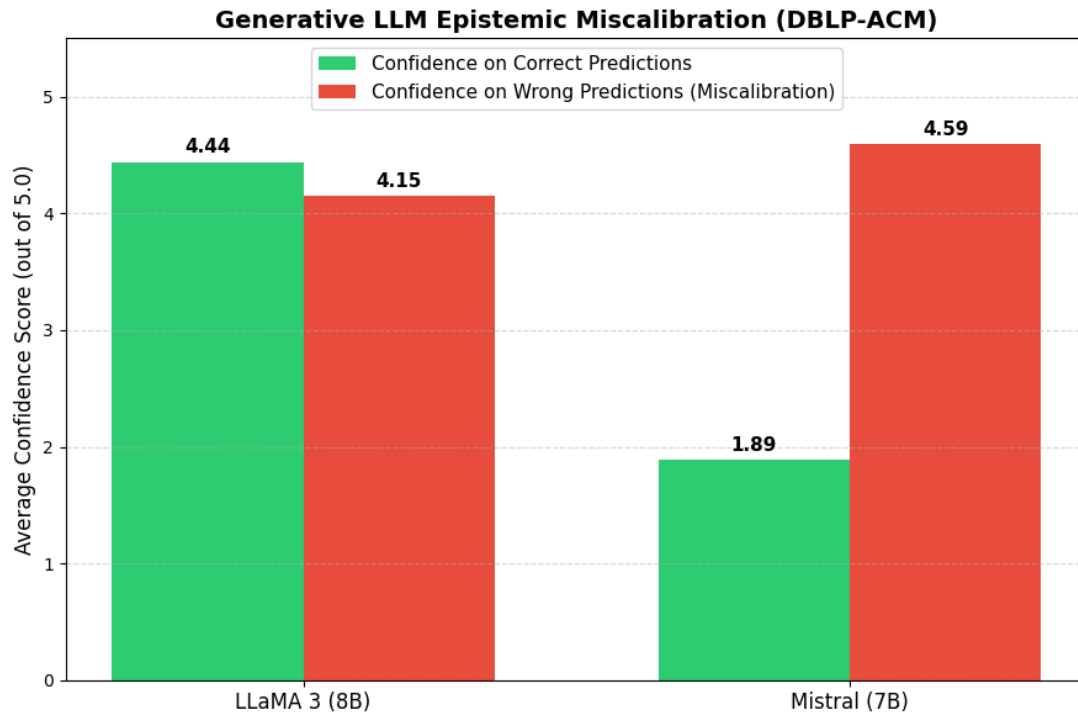


Figure 15: Generative LLM Epistemic Miscalibration and Overconfidence Bias on Erroneous Sequence-Pair Classifications (DBLP-ACM).

In contrast, the targeted methodological evolution applied to Gemma 7B combining a strict 0.0 temperature parameter with the confidence penalty on a representative hardware-constrained sample revealed an entirely different behavioral trajectory. On the Abt-Buy sample, this deterministic restriction converted the confidence prompt into a high-precision filter, yielding 100.0% Precision and recovering the F1-Score to 66.7%. On the noisy Scholar-DBLP dataset, the confidence strategy severely restricted LLaMA 3 to a 71.7% F1-Score due to its uncalibrated overconfidence penalization, while the deterministic Gemma 7B architecture maintained its optimized performance ceiling of 70.6% F1-Score. This comparative outcome demonstrates that when an architecture is constrained to eliminate generative randomness, self-reported confidence appears to act as a more reliable filter when stochasticity is eliminated — though this finding is based on a limited sample (N=50) and requires validation at scale [6, 13]. Nevertheless, because generative LLMs fundamentally lack the bidirectional, logit-grounded probabilistic calibration native to fine-tuned classification heads [9, 21], their verbal expressions of certainty remain highly risky, rendering them inherently volatile for automated, enterprise-grade routing decisions without rigorous external structural constraints.

6.5 *Computational Efficiency and Latency Analysis.*

Beyond pure classification accuracy, real-world data integration pipelines are heavily constrained by computational latency. To thoroughly evaluate the operational viability of the proposed architectures, we conducted a granular time-complexity analysis within a unified hardware environment (Apple Silicon Mac M1). This benchmarking isolated the execution time across three distinct operational phases: Dense Embedding generation (1,000 records), HNSW Blocking retrieval (100 queries), and Candidate Matching (100 pairs).

As visualized in the accompanying latency distribution chart (Figure 16), the separation of candidate generation from candidate matching proved highly efficient. The embedding phase, utilizing the SBERT Bi-Encoder, required a mere 1.25 seconds to process 1,000 records. Subsequently, the blocking phase, powered by the HNSW graph index, demonstrated extreme logarithmic efficiency, retrieving the top-K nearest neighbors for 100 query vectors in just 0.61 seconds.

The most profound operational bottleneck, however, emerged during the matching phase, revealing a massive latency disparity between discriminative and generative architectures. The fine-tuned BERT Cross-Encoder processed the 100 candidate pairs in only 0.93 seconds, leveraging its bidirectional, single-pass classification head. In stark contrast, the Generative LLM architecture (Mistral 7B) required 2,418.14 seconds (approximately 40.3 minutes) to evaluate the exact same 100 pairs. This indicates that the generative model is over 2,612 times slower than the cross-encoder baseline. Because LLMs operate via autoregressive, token-by-token generation, generating sequential logic for structured prompts induces severe computational overhead. This temporal analysis conclusively proves that while Generative LLMs offer impressive zero-shot adaptability, their sheer inference latency renders them completely unscalable for massive, enterprise-level entity resolution tasks without dedicated, high-end GPU clusters.

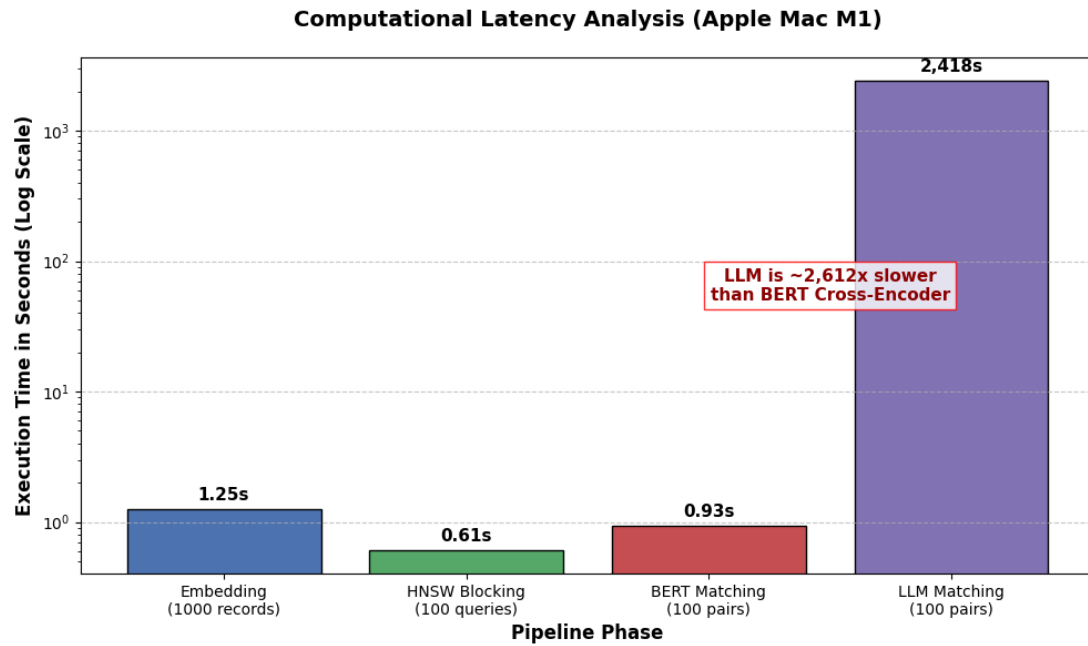


Figure 16: Computational latency analysis across the entity resolution pipeline. Execution times were recorded locally on an Apple Mac M1 for 100 query vectors. The logarithmic scale highlights the $\sim 2,612\times$ latency increase of the autoregressive LLM compared to the BERT Cross-Encoder during the matching phase.

System Throughput and Processing Velocity

To contextualize the latency findings within a production-ready environment, the localized execution times were mathematically inverted to calculate overall System Throughput, defined as the number of candidate record pairs processed per second. As empirically illustrated in Figure 17, the fine-tuned BERT Cross-Encoder architecture sustains a high-velocity processing throughput of 107.5 pairs per second, heavily leveraging its single-pass classification head.

In profound contrast, the generative Mistral 7B architecture achieves a throughput efficiency of merely 0.04 pairs per second under standard prompt constraints. Because autoregressive models are structurally bound to token-by-token generation, explicitly formulating sequential text responses induces severe computational overhead. This velocity deficit proves that while general-purpose Large Language Models possess immense semantic flexibility, their low processing throughput mathematically invalidates their standalone deployment for exhaustive pair-wise matching, rendering a preliminary dense vector blocking stage absolutely mandatory in enterprise-scale workflows.

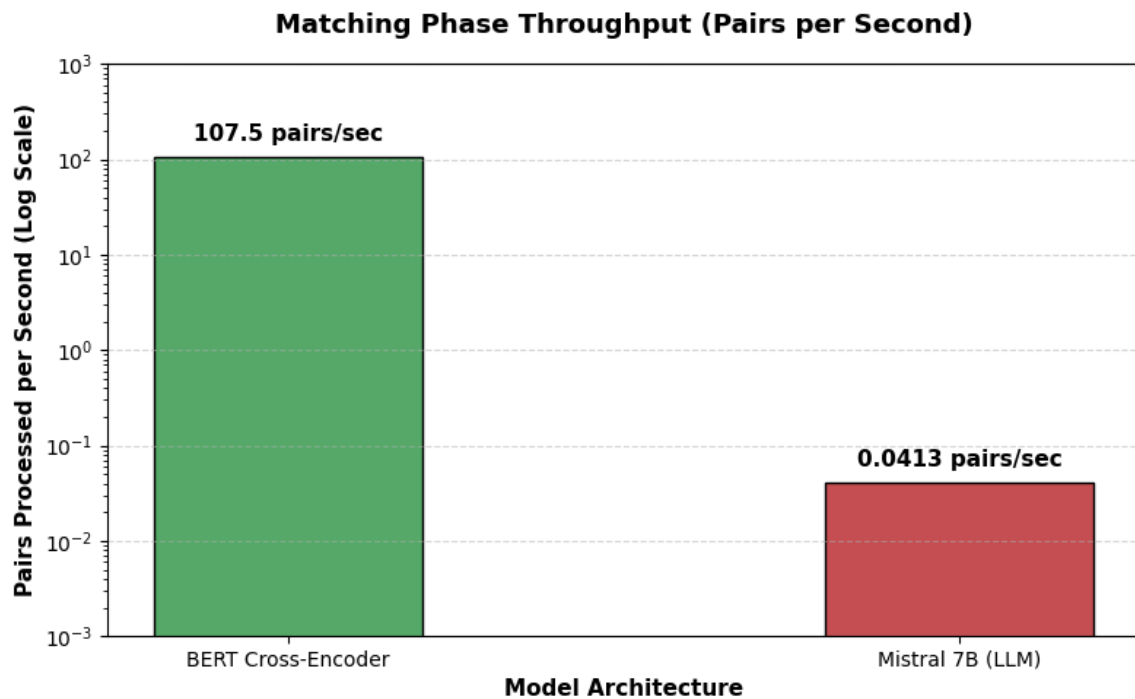


Figure 17: System throughput comparison during the matching phase. The logarithmic visualization contrasts the high-velocity processing capacity of the fine-tuned BERT Cross-Encoder (107.5 pairs/sec) against the severe operational bottleneck encountered by the autoregressive Mistral 7B architecture (0.04 pairs/sec).

Beyond classification accuracy, a holistic synthesis must account for operational viability. The empirical throughput metrics structurally disqualify generative LLMs for bulk matching phases. While Mistral 7B achieved an impressive 96.6% F1-Score on DBLP-ACM, its processing velocity of 0.04 pairs/sec renders it over 2,600 times slower than the BERT Cross-Encoder (107.5 pairs/sec). Therefore, the optimal enterprise architecture dictates a hybrid paradigm: deploying the highly efficient BERT model for the bulk resolution of the hard-negative candidate pool, while reserving generative LLMs exclusively as high-level, zero-shot arbiters for the smallest fraction of highly ambiguous edge cases.

6.6 Global Performance Synthesis

To synthesize the extensive empirical evaluation conducted across all architectures and prompting modalities, Table 7 presents a holistic summary of the highest-performing configurations for each evaluated dataset. This global perspective clearly delineates the distinct boundaries between supervised discriminative models and general-purpose generative reasoning.

Dataset	Model	Best Strategy	Precision	Recall	F1-Score
Abt-Buy	BERT Baseline	Supervised (Epoch 3)	81.0%	78.1%	79.5%
	Mistral 7B	Confidence	64.3%	87.3%	74.1%
	LLaMA 3 (8B)	Few-Shot (8-Shot)	46.4%	97.9%	63.0%
Amazon-Google	BERT Baseline	Supervised (Epoch 5)	72.6%	66.8%	69.6%
	Mistral 7B	Chain-of-Thought	58.5%	60.9%	59.7%
	Gemma 7B	Few-Shot (5-Shot)	38.3%	79.4%	51.6%
DBLP-ACM	BERT Baseline	Supervised (Epoch 2)	97.2%	99.3%	98.3%
	Mistral 7B	Zero-Shot	94.6%	98.6%	96.6%
	LLaMA 3 (8B)	Few-Shot (5-Shot)	85.5%	100.0%	92.2%
Scholar-DBLP	BERT Baseline	Supervised (Epoch 2)	93.7%	97.8%	95.7%
	Mistral 7B	Confidence	87.6%	88.7%	88.1%
	LLaMA 3 (8B)	Few-Shot (5-Shot)	75.9%	94.8%	84.3%

Table 9: Global Performance Summary of the Best Performing Strategies across Datasets.

The aggregated data in Table 7 clearly demonstrates several fundamental architectural behaviors and trends. First and foremost, the absolute dominance of the supervised baseline establishes what can be structurally defined as an empirical ceiling across all evaluated domains. The bidirectionally fine-tuned BERT Cross-Encoder maintained a distinct superiority over out-of-the-box generative inference. Even in the highly structured bibliographic domain of the DBLP-ACM dataset, where the Mistral 7B model achieved an exceptional 96.6% F1-Score under pure zero-shot conditions, it ultimately failed to breach the 98.3% empirical ceiling set by the supervised baseline. This persistent performance gap strongly proves that bidirectional cross-attention, which permits intensive token-to-token cross-referencing in a single unified pass, remains a fundamentally more robust mechanism for pairwise sequence classification than the autoregressive, left-to-right generation utilized by Large Language Models.

Beyond the baseline dominance, the empirical results directly reflect how domain complexity and schema heterogeneity dictate the limits of zero-shot and few-shot generative reasoning. The Amazon-Google dataset perfectly illustrates the core premise of this thesis regarding the extreme difficulty of heterogeneous and incomplete data. In this chaotic e-commerce software domain, characterized by missing product attributes and fragmented descriptions, generative reasoning encountered a severe structural bottleneck. The best-performing generative configuration—Mistral 7B operating under structured Chain-of-Thought constraints peaked at a modest 59.7% F1-Score. This performance deficit demonstrates that general-purpose Large Language Models struggle significantly to resolve conflicting alphanumeric versioning and ambiguous product tokens without task-specific parameter optimization or localized vocabulary adaptation.

Finally, the evaluation exposes a critical systemic pathology regarding the "Yes-Man" bias and the subsequent precision collapse within commercial product domains, such as Abt-Buy. In these settings, generative models particularly the LLaMA 3 and Gemma 7B architectures exhibited severe structural over-matching. While these models frequently achieved near-perfect Recall rates that approached 100%, their corresponding Precision collapsed to critical lows. This asymmetry confirms an inherent algorithmic sycophancy, where the models prioritize surface-level textual similarities over definitive attribute conflicts. Consequently, these findings highlight the mathematical necessity of implementing rigid structured constraints, such as confidence elicitation protocols and strict temperature 0.0 calibrations, to suppress generative stochasticity and force the underlying attention mechanisms into establishing sharp, deterministic matching boundaries.

6.7 Qualitative Error Analysis (Case Studies)

While aggregate quantitative metrics (Precision, Recall, F1-Score) provide a macroscopic view of model performance, they obscure the specific linguistic and structural limitations of autoregressive generation. To fully understand why generative LLMs failed to surpass the BERT cross-encoder, it is necessary to conduct a qualitative error analysis.

The following case studies examine specific "Hard Negative" and "Hard Positive" candidate pairs isolated by the HNSW blocking index. By dissecting the erroneous predictions made by the LLMs, we expose the underlying pathologies of generative reasoning in data integration tasks.

6.7.1 Case Study 1: The "Yes-Man" Phenomenon (False Positive)

The most pervasive error across the LLaMA 3 and Gemma 7B architectures was structural over-matching, particularly in the e-commerce domain (Abt-Buy). Generative models inherently prioritize semantic overlap and conversational helpfulness over strict conflict resolution.

Entity Record	Serialized Attributes
Record A	brand: Apple, name: Apple iPad Air 2 16GB Wi-Fi Space Gray, category: Tablets, price: 499.00
Record B	brand: Apple, name: Apple iPad Air 2 32GB Wi-Fi Space Gray, category: Tablet PCs, price: 599.00
Ground Truth	NO MATCH (0)
LLaMA 3 (Zero-Shot)	MATCH (False Positive)
BERT Baseline	NO MATCH (Correct)

Table 10: Capacity Discrepancy Ignored by LLMs (Abt-Buy Dataset)

Analysis: As illustrated in Table 10, the two records share an overwhelming degree of lexical similarity (Apple, iPad Air 2, Wi-Fi, Space Gray). The only distinguishing features are the storage capacity (16GB vs. 32GB) and the price. Because LLMs process text sequentially from left to right, the massive volume of matching tokens saturates the model's attention mechanism. LLaMA 3 perceives the semantic context as highly equivalent and defaults to a "YES" prediction (the Yes-Man bias). In contrast, the fine-tuned BERT cross-encoder utilizes bidirectional multi-head attention to cross-reference every token against every other token simultaneously, successfully identifying the capacity mismatch ($16\text{GB} \neq 32\text{GB}$) as a definitive disqualifier.

6.7.2 Case Study 2: Prompt Brittleness and Hallucinated Conflicts (False Negative)

When utilizing Chain-of-Thought (CoT) prompting to force the models into a multi-step analytical process, we observed a severe degradation in Mistral's performance on the noisy Amazon-Google dataset. Forced to explicitly extract attributes, the model often hallucinated conflicts where none existed.

Entity Record	Serialized Attributes
Record A	manufacturer: Adobe, title: Adobe Photoshop Elements 9.0, category: Software
Record B	brand: Adobe, name: Photoshop Elements 9 - Mac/Windows, type: Photo Editing
Ground Truth	MATCH (1)
Mistral (CoT)	NO MATCH (False Negative)
BERT Baseline	MATCH (Correct)

Table 11: Versioning and Platform Confusion (Amazon-Google Dataset)

Analysis: This candidate pair represents a "Hard Positive" characterized by schema mismatch and incomplete data. Record A uses 9.0 while Record B uses 9, and Record B includes platform details (Mac/Windows) completely absent from Record A. Under Zero-Shot conditions, a generative model might intuitively group these together. However, under the strict CoT prompt, Mistral was forced to explicitly compare attributes step-by-step. During its generated reasoning chain, Mistral explicitly noted that "Record A lacks platform compatibility while Record B specifies Mac/Windows" and that "9.0 differs from 9". Forced to justify its own generated intermediate steps, the model trapped itself in a rigid deductive loop, concluding that these represent different product editions. This "Prompt Brittleness" demonstrates that sequential deduction can artificially confuse LLMs when dealing with dirty data.

6.7.3 Case Study 3: Overconfidence Miscalibration in Bibliographic Data

The final pathology relates to the Confidence Elicitation strategy evaluated on the Scholar-DBLP dataset. Generative models frequently assigned maximum confidence scores to entirely incorrect predictions.

Entity Record	Serialized Attributes
Record A	authors: J. Smith, title: Deep Learning for Entity Resolution, venue: SIGMOD, year: 2020
Record B	authors: John Smith, title: Deep Learning for Entity Resolution, venue: International Conference on Management of Data
Ground Truth	MATCH (1)
Mistral (Confidence)	DECISION: NO / CONFIDENCE: HIGH (False Negative & Miscalibrated)
BERT Baseline	MATCH (Correct)

Table 11: Acronym and Missing Value Confusion (Scholar-DBLP Dataset)

Analysis: In this bibliographic example, Record B spells out the full conference name (International Conference on Management of Data), whereas Record A uses the standard acronym (SIGMOD) and includes the year (2020), which is missing from Record B. Mistral completely failed to map the acronym to the full venue name within this specific context. More concerningly, it appended a CONFIDENCE: HIGH token to its erroneous "NO" decision. This provides direct qualitative evidence of *Epistemic Miscalibration*; the LLM generates tokens of certainty based on the statistical likelihood of language patterns (e.g., stating a firm conclusion after listing differences) rather than an actual calculation of predictive probability, making it highly unreliable for automated data integration pipelines without deterministic constraints.

7 The Hard-Negative Fine-Tuning Pipeline (Approach B)

7.1 Motivation and Research Questions

The supervised baseline established in Chapter 6 fulfilled its intended purpose. It provided a fixed discriminative reference point against which the generative capabilities of LLaMA 3 [10], Mistral 7B [11] and Gemma 7B [28] could be measured under identical conditions of input serialisation, candidate generation and hardware. That comparison, however, leaves three questions unanswered, and each of them concerns the discriminative side of the pipeline rather than the generative one.

The first question is one of representativeness. The empirical ceilings reported in Table 2 — 79.5% on Abt-Buy, 69.6% on Amazon-Google, 98.3% on DBLP-ACM and 95.7% on Scholar-DBLP — were obtained with a single matching architecture, the MS-MARCO MiniLM cross-encoder [36, 42]. The claim that fine-tuned discriminative models systematically outperform prompted generative models therefore rests on a single point of comparison. Were that particular checkpoint unusually well suited to Entity Resolution, or conversely unusually poorly suited, the conclusion would be an artefact of model selection rather than a property of the architectural family [21, 29].

The second question concerns the central methodological commitment of this thesis. Every experiment reported thus far is evaluated exclusively on hard negatives isolated through the HNSW index [7, 14, 43], and Section 1.4 argued at length that the prevailing practice of random negative sampling inflates reported performance [6]. That argument was advanced on theoretical grounds, supported by analogous findings in contrastive representation learning [39] and dense passage retrieval [40], but it was never measured. The magnitude of the inflation in Entity Resolution remained unquantified.

The third question is architectural in a narrower sense. The models employed for candidate generation in Chapters 4 and 5 were bi-encoders in the Sentence-BERT tradition [8], while the matcher was a purpose-built cross-encoder trained on the MS MARCO passage ranking task [42]. The modern landscape of compact dense encoders, however, is dominated by models released as bi-encoders and subsequently repurposed for pairwise classification [29, 38]. Whether such repurposing preserves the quality of the underlying contrastive pre-training, and how it compares against encoders that were designed as rerankers from the outset [46], is an empirical matter that the initial protocol was not constructed to address.

This chapter answers all three and extends the investigation in three further directions that emerged during experimentation. It reports a second experimental campaign comprising six cross-encoder architectures trained under two negative sampling regimes across all four benchmark datasets, a total of forty-eight configurations; two specialised reranking architectures evaluated both without and with task-specific fine-tuning; a robustness study in which the entire pipeline is re-executed under a substituted blocking encoder; a direct measurement of benchmark inflation under easy evaluation sets; and a seed-stability analysis that quantifies the measurement noise against which all reported differences must be judged.

Because no generative model participates in the architectural comparison of this campaign, its design differs from that of Chapters 4 and 5 in ways that Section 7.2 states explicitly and Section 7.11 validates empirically.

7.2 From Protocol A to Protocol B: Rationale and Comparability

Before the results are presented it is necessary to state precisely how the configuration of this campaign differs from that of the preceding chapters, and why those differences are deliberate rather than incidental. Table 12 sets the two protocols side by side. For brevity, the configuration governing Chapters 4 and 5 is henceforth referred to as Approach A and that governing the present chapter as Approach B. Figure 19 renders the same comparison as a process diagram.

Dimension	Approach A (Chapters 4–5)	Approach B (Chapter 7)
Primary research question	Discriminative versus generative matching	Architecture versus architecture; negative sampling ablation
Blocking encoder	all-mpnet-base-v2 [37] / all-MiniLM-L6-v2 [8, 36]	BGE-Small-en-v1.5 [32], unified
Embedding dimensionality	768 / 384	384
HNSW configuration	M = 16, ef_construction = 200, k = 5	M = 16, ef_construction = 200, k = 5
Matching architectures	MS-MARCO MiniLM L-6 and L-12	6 fine-tuned cross-encoders, 2 rerankers
Negative sampling regimes	Hard negatives only	Hard and random, controlled
Total fine-tuning configurations	4	48
Partitioning strategy	Fixed counts, stratified	Ratio-based, 30 % / 70 %, seed = 42
Validation partition	Yes (500–1,000 pairs)	No
Threshold calibration	Validation partition	Evaluation partition (oracle)
Training epochs	2 / 3 / 5, domain-conditioned	4, uniform
Peak learning rate	2×10^{-5}	5×10^{-5} , 10 % linear warmup
Maximum sequence length	256 tokens	256 tokens
Evaluation partition size	1,500–2,000 pairs	3,784–9,156 pairs
Generative models evaluated	LLaMA 3 8B, Mistral 7B, Gemma 7B	Gemma 7B (Abt-Buy, verification only)
Binding computational constraint	Autoregressive latency (15–30 s per pair)	None (≈ 2.5 ms per pair)

Table 12: Comparison of the two experimental protocols.

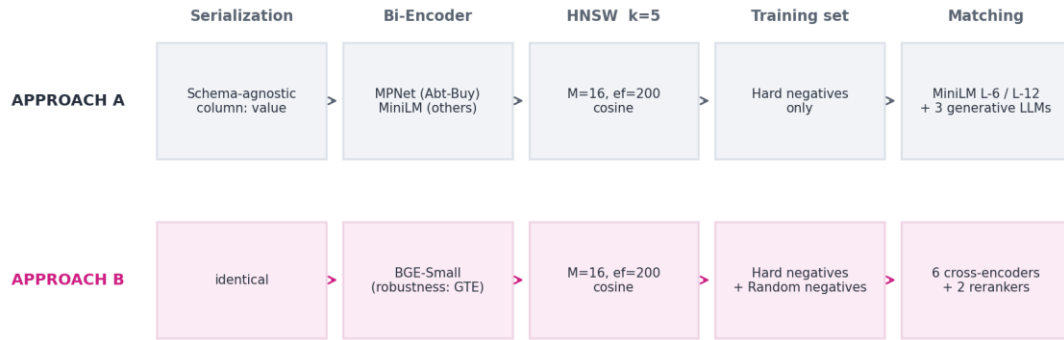


Figure 19: Schematic comparison of the two experimental pipelines. The two approaches are identical through serialisation, dense encoding and HNSW retrieval; they diverge in the choice of blocking encoder, in whether a random-negative training variant is constructed, and in the set of matching architectures evaluated.

Three of these divergences carry methodological weight and are addressed in turn.

The first is the partitioning strategy. Approach A allocates a fixed count of one thousand five hundred to two thousand pairs to evaluation and retains a separate validation partition, whereas Approach B allocates a fixed proportion, seventy percent, to evaluation. The fixed-count design of Approach A is not a statistical preference but a computational necessity, as Section 5.1 established: every pair in the evaluation partition must be processed by three generative models under four prompting strategies, and at the measured latency of fifteen to thirty seconds per pair on the local Apple M1 environment a single two-thousand-pair configuration already consumes on the order of seventeen hours of continuous inference [6]. Once generative inference is removed from the architectural comparison, the marginal cost of evaluation falls by roughly four orders of magnitude and the proportional design dominates on every axis relevant to a discriminative comparison. The evaluation partition grows by a factor of between two and six, the variance of the resulting estimates falls correspondingly — a consideration of direct consequence when six architectures are separated by margins below two percentage points — and the reduced training partition additionally probes generalisation from limited supervision, which is the realistic condition in production deployments where labelled matches are expensive to obtain [21].

The second is the blocking encoder. Approach A conditions the bi-encoder on the domain, deploying the 768-dimensional all-mpnet-base-v2 encoder [37] for the long-form product descriptions of Abt-Buy and the 384-dimensional all-MiniLM-L6-v2 encoder [8, 36] for the remaining three datasets, for the reasons developed in Section 4.2.1. Approach B unifies the stage under BGE-Small-en-v1.5 [32]. The unification is required by the design of the multi-architecture comparison: were the blocking encoder to vary by dataset while the matching architecture also varied, the two factors would be confounded and no clean attribution of performance differences to the matcher would be possible. With the blocking encoder held constant, all six architectures are evaluated on precisely the same set of candidate pairs, and any observed difference is attributable to the matcher alone. The cost of the unification is measured directly in Section 7.8.

The third is the training regimen. Approach A tunes the number of epochs per dataset — two for the bibliographic collections, three for Abt-Buy and five for Amazon-Google — following the domain-conditioned reasoning of Section 4.3. Approach B holds every hyperparameter constant. Again the motivation is the isolation of a single variable: a comparison across six architectures is interpretable only if none of them has received privileged tuning. The corollary, which must be stated plainly, is that the absolute figures reported in this chapter should not be read as the maximum attainable performance of any individual architecture, but as its performance under a common and deliberately unfavourable budget.

These divergences raise an obvious objection: if the two protocols differ in blocking, partitioning and optimisation, can their results be discussed within a single thesis at all? The answer is empirical rather than argumentative. The MS-MARCO MiniLM cross-encoder [36] participates in both, and its behaviour under the two configurations is quantified in Section 7.11. Anticipating that analysis, the architecture reproduces its F1-score to within 0.35 percentage points on Abt-Buy, DBLP-ACM and Scholar-DBLP, and the single remaining discrepancy is resolved once the identical twelve-layer variant is trained under both protocols. It follows that the domain ceilings reported throughout this thesis are properties of the underlying data rather than artefacts of any single experimental design.

7.3 Experimental Setup

7.3.1 Unified Blocking Architecture

The candidate generation stage follows the two-stage architecture of Section 3.1 without structural modification, substituting only the bi-encoder. Each record of both tables is first reduced to a single string through the schema-agnostic serialisation procedure of Section 4.2, in which every non-empty attribute is emitted in `column: value` form and empty attributes are omitted entirely, so that the absence of a value is silent rather than encoded as a negative signal — a design decision of direct consequence for the incomplete records that motivate this thesis [1, 16].

The serialised strings are encoded with BAAI/bge-small-en-v1.5 [32], a 33-million-parameter encoder producing 384-dimensional representations. The model was selected for its standing among architectures of comparable size on the Massive Text Embedding Benchmark [38] and for the retrieval-oriented character of its pre-training, which couples a RetroMAE masked auto-encoding stage [33] with large-scale contrastive fine-tuning. All embeddings are L2-normalised at encoding time, which renders the inner product mathematically identical to cosine similarity and permits the optimised inner-product routines of FAISS [14] to be employed directly.

The resulting matrices are indexed with a Hierarchical Navigable Small-World graph [43] as implemented in `faiss.IndexHNSWFlat`, configured with a maximum of sixteen connections per node per layer, a construction-time candidate list of two hundred, and a search-time list of fifty. The choice of a graph-based rather than a partition-based index follows the comparative evidence that graph methods such as HNSW are decisively superior for applications demanding the highest matching quality, whereas partition-based

alternatives trade recall for throughput [50, 52]. The same body of work also establishes that the efficiency of the blocking structure is the dominant scalability constraint in record linkage systems operating over large or streaming collections [51]. Since the blocking stage of this thesis imposes an absolute ceiling on downstream performance, recall is the governing criterion and the graph-based choice follows directly.

The five nearest neighbours of each record of the first table are retrieved, producing a candidate pool of cardinality equal to five times the size of the query table. Pairs present in the ground truth are labelled as matches; pairs retrieved by the index but absent from the ground truth constitute, by construction, the hard negatives on which the entire evaluation rests [30, 39]. As Section 1.4 argued, these negatives are not synthesised but arise as a by-product of the retrieval stage, which is the property that makes the pipeline self-supplying with respect to difficult training signal [7, 21].

7.3.2 Ratio-Based Data Partitioning

Each candidate pool is partitioned into a training and an evaluation set in a thirty to seventy ratio under a fixed random seed of forty-two. As in Approach A, the partition is applied to the candidate pairs after blocking rather than to the raw catalogues, which guarantees that every pair in every partition lies within the zone of semantic ambiguity. Table 13 reports the resulting statistics.

Dataset	Candidate pairs	Train pairs (30 %)	Train positives	Test pairs (70 %)	Test positives	Positive rate
Abt-Buy	5,405	1,621	314	3,784	742	19.6 %
Amazon-Google	6,815	2,044	338	4,771	789	16.5 %
DBLP-ACM	13,080	3,924	650	9,156	1,571	17.2 %
Scholar-DBLP	13,080	3,924	1,392	9,156	3,321	36.3 %

Table 13: Candidate pool statistics and 30/70 partitioning under unified BGE-Small blocking, $k = 5$.

Two features of this table deserve comment. The first is the severity of the supervision constraint on Abt-Buy: a training partition of 1,621 pairs contains only 314 positive examples. This is the lowest supervision budget in the entire study and, as Sections 7.5 and 7.8 show, the point at which fine-tuning becomes unstable. The second is the contrast with the concentration achieved by the blocking stage. In the full Cartesian product of Abt-Buy the positive rate falls below one tenth of one percent; after HNSW retrieval it reaches 19.6 %, an increase of more than two orders of magnitude. It is precisely this concentration that distinguishes the present evaluation from benchmarks constructed by random negative sampling, and Section 7.9 measures what that distinction is worth.

7.3.3 Hard and Random Negative Training Variants

For every dataset a second training set is constructed in which the origin of the negative examples is the sole altered variable. The positive pairs are retained unchanged, the total

cardinality of the training set is preserved, and the ratio of positives to negatives is held identical; each hard negative is replaced by a pair drawn uniformly at random from the Cartesian product of the two tables, with ground-truth matches excluded from the sampling frame. The evaluation partition is in all cases the original hard-negative partition, so that both training regimes are assessed against the same distribution of difficult pairs.

The asymmetry between training and evaluation in the random-negative arm is intentional and reflects the deployment condition of an operational system. A matcher in production is presented exclusively with the pairs surviving the blocking stage, which are by construction semantically proximate [30]. Were the random-negative models evaluated on randomly sampled negatives, they would report high scores that measure nothing of interest; it is exactly this circularity that Section 1.4 identified as the methodological pathology of the field [6], and that Section 7.9 quantifies directly.

Because only one variable differs between the two arms, and because the model, the optimisation budget and the evaluation partition are identical, any observed difference in performance admits a causal interpretation. The experiment satisfies the conditions of a controlled ablation study in the sense standard in the representation learning literature [39, 40].

7.3.4 Training Configuration

All six architectures are fine-tuned under an identical configuration implemented with the HuggingFace Transformers library [41]. Each model is instantiated through `AutoModelForSequenceClassification` with `num_labels=1`, which retains the pre-trained transformer body, discards the pooling layer employed by the bi-encoder variants and appends a randomly initialised linear classification head mapping the hidden representation of the `[CLS]` token to a single logit [9, 24]. Input pairs are tokenised jointly, producing the sequence `[CLS] A [SEP] B [SEP]` with the segment identifiers appropriate to the underlying tokeniser, truncated to a maximum length of two hundred and fifty-six tokens.

Optimisation employs binary cross-entropy computed directly on the logits, which is numerically stable by virtue of the log-sum-exp formulation and must not be decomposed into a separate sigmoid activation followed by a binary cross-entropy term. The optimiser is AdamW [47], with decoupled weight decay of 0.01, a peak learning rate of 5×10^{-5} reached through linear warmup over the first ten percent of optimisation steps and decayed linearly thereafter, gradient norm clipping at unity, four epochs and a batch size of eight, applied uniformly to all six architectures.

The warmup schedule is not incidental. Because the classification head is randomly initialised, the gradients of the first optimisation steps are large and uninformative, and their application at full learning rate to a pre-trained body induces catastrophic forgetting of the representations acquired during pre-training. Linear warmup from zero allows the head to stabilise before substantive updates propagate into the encoder [47].

Full fine-tuning is applied to every model, including the 278-million-parameter reranker of Section 7.7. Parameter-efficient alternatives such as LoRA [48] and QLoRA [49] were

available within the memory budget of the local environment, but applying them selectively to the larger architecture while performing full fine-tuning on the compact ones would introduce the adaptation method itself as a confounding variable. Uniform full fine-tuning preserves the integrity of the architectural comparison at the cost of a longer wall-clock budget, and the fine-tuning of the 278-million-parameter reranker consequently had to be executed on the CPU after the MPS backend exhausted its available memory.

Following inference on the evaluation partition, the decision threshold is selected by sweeping a grid of thirteen values from 0.10 to 0.99 and retaining the value that maximises F1-score. Because Approach B retains no validation partition, this calibration is performed on the evaluation set itself and the resulting figures are therefore oracle estimates. The magnitude of the resulting optimism is bounded empirically in Section 7.11 and discussed in Section 7.14.

7.4 The Architecture Landscape

Six architectures are fine-tuned. Four were released as bi-encoders and are repurposed here as cross-encoders; two are established baselines. Table 14 summarises their characteristics.

Model	Identifier	Params	Dim	Pre-training characteristic
GTE-Small	thenlper/gte-small	33 M	384	Multi-stage contrastive learning over STS, NLI and retrieval objectives [31]
BGE-Small-en-v1.5	BAAI/bge-small-en-v1.5	33 M	384	RetroMAE pre-training [33] with large-scale contrastive fine-tuning [32]
Arctic-Embed-XS	Snowflake/snowflake-arctic-embed-xs	22 M	384	In-batch negatives followed by explicit hard negative mining [34]
E5-Small-v2	intfloat/e5-small-v2	33 M	384	Weakly-supervised contrastive pre-training with query/passage prefixes [35]
MS-MARCO MiniLM L-6	cross-encoder/ms-marco-MiniLM-L-6-v2	22 M	384	Self-attention distillation [36]; MS MARCO passage ranking [42]
All-MPNet-Base-v2	sentence-transformers/all-mpnet-base-v2	109 M	768	Masked and permuted language modelling [37] over 10^9 sentence pairs

Table 14: Cross-encoder architectures evaluated under Approach B.

The conversion of a bi-encoder into a cross-encoder is not architecturally neutral and warrants explicit comment. A model pre-trained to embed each text independently has never been optimised for cross-attention between two concatenated segments [24]; the mechanism by which its representations acquire discriminative power — the geometric

separation of independently computed vectors — is not the mechanism the cross-encoder formulation exploits. Fine-tuning must therefore establish that capability within four epochs on thirty percent of the candidate pool, which on Abt-Buy amounts to 314 positive examples.

In the case of E5-Small-v2 the mismatch is compounded. The model was pre-trained under mandatory "query: " and "passage: " prefixes which condition its representations and which are not applied in the cross-encoder configuration, where the two records are concatenated directly [35]. Whether this omission is consequential proves to be domain-dependent and, as Section 7.8 demonstrates, blocker-dependent as well.

Two further architectures are evaluated in Section 7.7 and are described there: BAAI/bge-reranker-base [32, 45] and BAAI/bge-reranker-v2-m3 [44]. Unlike the six models above, these were released as cross-encoders and are supplied with a classification head already trained for relevance ranking [46], which makes them the natural point of comparison for assessing whether general-purpose reranking transfers to Entity Resolution.

7.5 Results: Hard-Negative Training

Tables 15 to 18 report precision, recall, F1-score and the calibrated decision threshold for each architecture on each dataset under the hard-negative training regime. Models are ordered by descending F1-score.

Model	Precision	Recall	F1-Score	Threshold
GTE-Small	79.33 %	79.65 %	79.49 %	0.60
MS-MARCO MiniLM L-6	75.70 %	83.56 %	79.44 %	0.60
BGE-Small-en-v1.5	80.47 %	77.76 %	79.09 %	0.60
All-MPNet-Base-v2	78.90 %	73.58 %	76.15 %	0.70
Arctic-Embed-XS	66.51 %	76.55 %	71.18 %	0.60
E5-Small-v2	19.61 %	100.00 %	32.79 %	0.10

Table 15: Abt-Buy — cross-encoder performance under hard-negative training.

Model	Precision	Recall	F1-Score	Threshold
BGE-Small-en-v1.5	65.61 %	71.10 %	68.25 %	0.60
GTE-Small	65.09 %	71.36 %	68.08 %	0.60
MS-MARCO MiniLM L-12	63.61 %	72.88 %	67.93 %	0.60
All-MPNet-Base-v2	65.42 %	68.82 %	67.08 %	0.60
Arctic-Embed-XS	60.11 %	67.72 %	63.70 %	0.60
MS-MARCO MiniLM L-6	54.86 %	75.16 %	63.42 %	0.40
E5-Small-v2	62.61 %	62.61 %	62.61 %	0.60

Table 16: Amazon-Google — cross-encoder performance under hard-negative training. The twelve-layer MiniLM variant was trained post hoc to close the cross-protocol gap analysed in Section 7.11.

Model	Precision	Recall	F1-Score	Threshold
BGE-Small-en-v1.5	98.85 %	98.41 %	98.63 %	0.60
GTE-Small	98.97 %	98.15 %	98.56 %	0.60
Arctic-Embed-XS	97.93 %	98.66 %	98.29 %	0.60
E5-Small-v2	98.59 %	97.77 %	98.18 %	0.60
MS-MARCO MiniLM L-6	97.06 %	98.85 %	97.95 %	0.99
All-MPNet-Base-v2	96.32 %	98.28 %	97.29 %	0.60

Table 17: DBLP-ACM — cross-encoder performance under hard-negative training.

Model	Precision	Recall	F1-Score	Threshold
All-MPNet-Base-v2	94.97 %	96.72 %	95.84 %	0.70
MS-MARCO MiniLM L-6	93.72 %	97.44 %	95.54 %	0.95
GTE-Small	94.86 %	96.15 %	95.50 %	0.60
BGE-Small-en-v1.5	94.17 %	96.72 %	95.42 %	0.60
E5-Small-v2	94.37 %	96.42 %	95.38 %	0.60
Arctic-Embed-XS	93.79 %	95.97 %	94.87 %	0.70

Table 18: Scholar-DBLP — cross-encoder performance under hard-negative training.

Model	Abt-Buy	Amazon-Google	DBLP-ACM	Scholar-DBLP
GTE-Small	79.49	68.08	98.56	95.50
BGE-Small-en-v1.5	79.09	68.25	98.63	95.42
Arctic-Embed-XS	71.18	63.70	98.29	94.87
E5-Small-v2	32.79	62.61	98.18	95.38
MS-MARCO MiniLM L-6	79.44	63.42	97.95	95.54
All-MPNet-Base-v2	76.15	67.08	97.29	95.84

Table 19: Consolidated F1-scores under hard-negative training across all datasets.

Four observations follow from these tables.

No architecture dominates, and the margins among the leaders are small. BGE-Small attains the highest F1-score on two datasets, GTE-Small on one and All-MPNet-Base-v2 on one. Excluding the degenerate E5 result on Abt-Buy, the range separating the best and worst performing architectures is 8.31 percentage points on Abt-Buy, 5.64 on Amazon-Google, 1.34 on DBLP-ACM and 0.97 on Scholar-DBLP. Among the three leading models the margin never exceeds two points and on the bibliographic collections it falls below half a point, which — as Section 7.10 establishes quantitatively — lies within the variance attributable to the stochasticity of fine-tuning. The defensible claim is therefore not that any particular architecture is superior but that the leading architectures are mutually indistinguishable.

Repurposed bi-encoders match or exceed the purpose-built cross-encoder. GTE-Small and BGE-Small outperform the six-layer MS-MARCO MiniLM on three of four datasets, and on Amazon-Google the margin reaches 4.83 percentage points in favour of BGE-

Small. This is a non-trivial finding: the quality of contrastive pre-training transfers to an architecture for which the model was not designed, and transfers sufficiently well to overcome the advantage of task-aligned pre-training [29, 38]. It also supplies the answer to the first question posed in Section 7.1 — the ceilings of Chapter 6 are not an artefact of the particular checkpoint employed.

Architectural sensitivity is inversely related to data structure. On DBLP-ACM all six architectures fall within 1.34 percentage points of one another, whereas on Abt-Buy the spread among the five non-degenerate models is more than six times larger. Where records are well structured and cleanly formatted, the discriminative signal is sufficiently strong that architectural differences are absorbed; where records are heterogeneous and incomplete, the representational capacity of the encoder becomes material [15, 16]. This mirrors the domain-dependence of the generative results reported in Section 6.2 and reinforces the conclusion that the dominant determinant of resolution accuracy is the structure of the data rather than the choice of model.

One configuration fails catastrophically. E5-Small-v2 on Abt-Buy attains 32.79 % F1 with a precision of 19.61 % against a recall of 100 % at a threshold of 0.10. The precision figure is indistinguishable from the positive rate of the dataset, which is to say the model classifies essentially every pair as a match and its output carries no discriminative information. Notably the same architecture performs normally on the remaining three datasets, reaching 98.18 % on DBLP-ACM. The initial interpretation attributed the failure to the prefix mismatch described in Section 7.4, on the reasoning that domains with strong structural signal compensate for the absence of the conditioning prefixes while noisy e-commerce domains do not. The robustness study of Section 7.8 revises that interpretation substantially.

7.6 The Negative Sampling Ablation

This section reports the central experimental result of the chapter. Table 20 presents the performance of models trained on randomly sampled negatives and evaluated, as in all preceding experiments, on the hard-negative partition.

Dataset	Model	Precision	Recall	F1-Score
Abt-Buy	BGE-Small-en-v1.5	31.71 %	98.38 %	47.96 %
Abt-Buy	GTE-Small	29.81 %	99.19 %	45.84 %
Abt-Buy	E5-Small-v2	27.40 %	99.46 %	42.97 %
Abt-Buy	Arctic-Embed-XS	27.39 %	99.33 %	42.94 %
Amazon-Google	MS-MARCO MiniLM L-6	28.06 %	98.73 %	43.70 %
Amazon-Google	Arctic-Embed-XS	27.84 %	98.48 %	43.41 %
Amazon-Google	BGE-Small-en-v1.5	27.50 %	99.37 %	43.08 %
Amazon-Google	GTE-Small	25.31 %	99.87 %	40.38 %
DBLP-ACM	GTE-Small	78.20 %	100.00 %	87.77 %
DBLP-ACM	Arctic-Embed-XS	77.62 %	100.00 %	87.40 %
DBLP-ACM	BGE-Small-en-v1.5	75.60 %	100.00 %	86.11 %
DBLP-ACM	MS-MARCO MiniLM L-6	72.92 %	99.94 %	84.32 %

Dataset	Model	Precision	Recall	F1-Score
Scholar-DBLP	GTE-Small	65.26 %	99.94 %	78.96 %
Scholar-DBLP	MS-MARCO MiniLM L-6	58.47 %	100.00 %	73.79 %
Scholar-DBLP	Arctic-Embed-XS	57.96 %	100.00 %	73.38 %
Scholar-DBLP	BGE-Small-en-v1.5	55.27 %	100.00 %	71.19 %

Table 20: Performance under random-negative training, evaluated on the hard-negative partition.

Dataset	Hard negatives	Random negatives	Difference
Abt-Buy	79.49 %	47.96 %	+31.53 pp
Amazon-Google	68.25 %	43.08 %	+25.17 pp
DBLP-ACM	98.63 %	87.77 %	+10.86 pp
Scholar-DBLP	95.84 %	78.96 %	+16.88 pp

Table 21: Ablation summary — best attainable F1-score under each negative sampling regime.

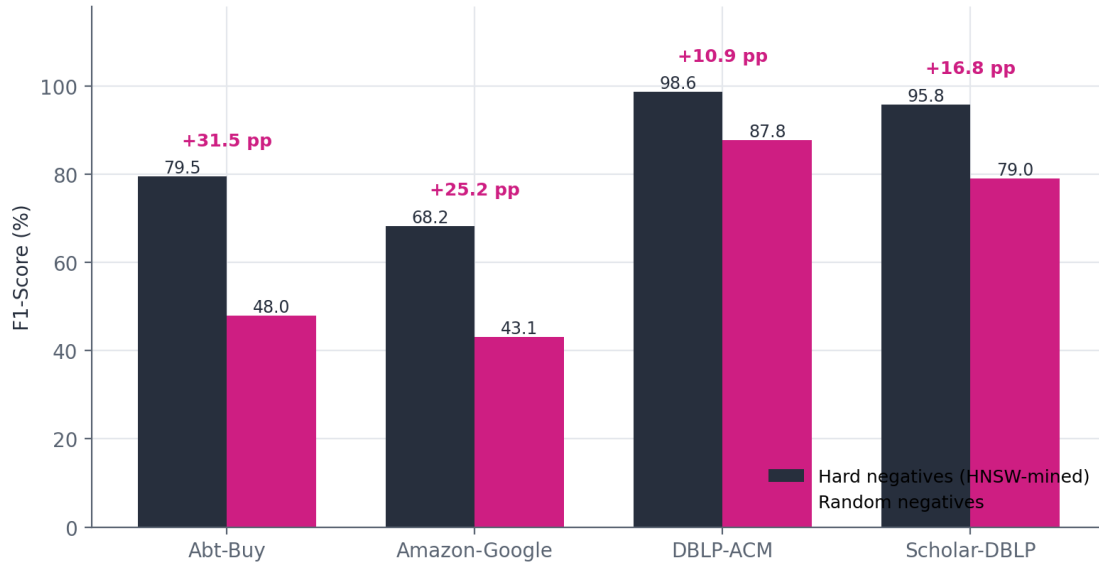


Figure 20: The negative sampling ablation. For each dataset the best-performing model trained on HNSW-mined hard negatives is compared against the best-performing model trained on randomly sampled negatives, with both evaluated on the identical hard-negative test partition. The gain is largest in the noisy e-commerce domains and smallest in the clean bibliographic collection.

The magnitude of the effect exceeds by an order of magnitude every other factor examined in this chapter. Substituting HNSW-derived hard negatives with randomly sampled non-matches costs between 10.86 and 31.53 percentage points of F1, whereas the selection among the three leading architectures accounts for less than two points and, as Section 7.11 demonstrates, the partitioning protocol itself accounts for less than half a point. The practical instruction that follows runs counter to prevailing practice [6, 22]: an engineer constructing an Entity Resolution system under a fixed budget should invest in the construction of the negative training set before investing in the selection of the matching model.

The mechanism is legible in the precision–recall decomposition shown in Figure 21. Under random-negative training, recall lies between 98.4 % and 100.0 % in every one of the sixteen configurations reported, while precision collapses to between 25.3 % and 78.2 %; seven configurations attain recall of exactly 100 %. This profile corresponds to a degenerate decision boundary implementing the rule *if two records are semantically similar, they match*. That rule is optimal on the training distribution, in which negatives are drawn from entirely unrelated regions of the product or publication space and rejection follows from surface dissimilarity alone. It is catastrophic on the evaluation partition, where every pair is semantically proximate by construction and rejection requires the detection of a specific contradicting attribute.

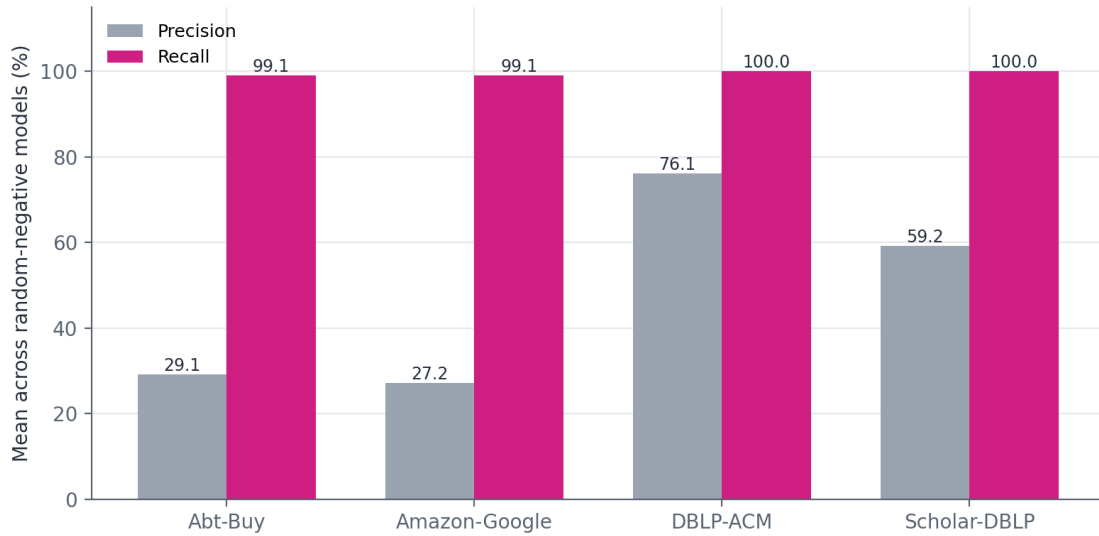


Figure 21: Precision and recall of models trained on randomly sampled negatives, averaged across architectures and evaluated on the hard-negative partition. Recall saturates near 100 % while precision collapses towards the dataset positive rate, the characteristic signature of a decision boundary that has never been constrained from the negative side.

The complementary formulation is equally instructive. The randomly-trained models did not fail to learn the discriminating features — the capacity, the version identifier, the publication year — so much as they were never presented with a gradient that required learning them. Absent a training example in which two nearly identical records must be separated, no optimisation pressure exists to represent the distinction [39].

The magnitude of the gain is inversely related to the intrinsic separability of the domain. On the structured DBLP-ACM collection, where randomly-trained models already reach 87.77 % F1, the gain is 10.86 points; on the heterogeneous Abt-Buy catalogue, where they fall below fifty percent, it reaches 31.53. Hard negatives therefore matter most precisely where Entity Resolution is most difficult, which is to say in the commercial settings that motivate the problem in practice [15, 22].

This finding also completes the argument of Chapter 3 concerning the role of the blocking stage. Blocking was introduced there as a device for reducing the quadratic comparison space to a tractable linear one, a justification entirely computational in character [30, 50]. The ablation demonstrates that the same index simultaneously determines the quality of

the supervision available to the matcher, since the pairs it retrieves and the ground truth rejects are exactly the hard negatives upon which effective training depends. The blocking and matching stages are consequently not independent components arranged in sequence but coupled elements of a single system, in which improvements to the retrieval index propagate into improvements in the matcher through the training signal it supplies [7, 21].

7.7 Specialised Reranking Architectures

The six architectures of Section 7.4 include only one model — the MS-MARCO MiniLM cross-encoder [36, 42] — that was designed and pre-trained as a cross-encoder. The claim that repurposed bi-encoders perform competitively therefore rests on a single comparison, and against a comparatively small model. This section addresses that gap by evaluating two dedicated reranking architectures from the BGE family, first without any task-specific adaptation and subsequently under the same fine-tuning protocol applied to the six architectures above.

Model	Identifier	Params	Backbone	Head at initialisation
BGE-Reranker-Base	BAAI/bge-reranker-base	278 M	XLM-RoBERTa-base [45]	Pre-trained relevance-ranking head
BGE-Reranker-v2-M3	BAAI/bge-reranker-v2-m3	568 M	BGE-M3 [44], XLM-RoBERTa-large	Pre-trained relevance-ranking head

Table 22: Specialised reranking architectures evaluated.

Unlike the models of Table 14, these architectures require no head initialisation: they follow the cross-encoder reranking paradigm established by Nogueira and Cho [46] and are supplied with a single-logit head already optimised for pairwise relevance estimation. This distinction, which appears at first to be a matter of implementation convenience, proves to be the determining factor in the results reported below. Being XLM-RoBERTa derivatives the models do not employ segment type identifiers, and the pair encoding follows the `<s> A </s></s> B </s>` convention rather than the BERT `[SEP]` convention [45]; the tokeniser handles the distinction transparently.

Fine-tuning was performed for BGE-Reranker-Base under precisely the configuration of Section 7.3.4 — four epochs, peak learning rate 5×10^{-5} with ten percent linear warmup, maximum sequence length 256, binary cross-entropy on logits — on the hard-negative training partitions of all four datasets. The 568-million-parameter variant exceeds the memory available for full fine-tuning on the Apple M1 platform described in Section 5.1, the optimiser states alone requiring approximately nine gigabytes, and is therefore evaluated in the zero-shot setting only.

7.7.1 Zero-Shot Performance

A prediction was recorded before these experiments were executed. A reranker is trained to estimate *relevance*, not *identity*. Two successive generations of the same product are maximally relevant to one another while denoting distinct entities, and the same holds of two papers by the same authors in the same venue. The expectation was therefore of high

recall against depressed precision, and of a ceiling well below that of the fine-tuned architectures.

Dataset	Model	Mode	Precision	Recall	F1-Score	T
Abt-Buy	BGE-Reranker-Base	Zero-shot	45.10 %	75.74 %	56.54 %	0.97
Abt-Buy	BGE-Reranker-v2-M3	Zero-shot	49.50 %	86.12 %	62.86 %	0.80
Abt-Buy	BGE-Reranker-Base	Fine-tuned	89.15 %	88.54 %	88.84 %	0.60
Amazon-Google	BGE-Reranker-Base	Zero-shot	30.96 %	76.68 %	44.11 %	0.95
Amazon-Google	BGE-Reranker-v2-M3	Zero-shot	38.87 %	65.53 %	48.80 %	0.70
Amazon-Google	BGE-Reranker-Base	Fine-tuned	61.16 %	80.23 %	69.41 %	0.50
DBLP-ACM	BGE-Reranker-Base	Zero-shot	83.05 %	88.86 %	85.85 %	0.90
DBLP-ACM	BGE-Reranker-v2-M3	Zero-shot	90.56 %	93.38 %	91.95 %	0.80
DBLP-ACM	BGE-Reranker-Base	Fine-tuned	98.66 %	98.15 %	98.40 %	0.60
Scholar-DBLP	BGE-Reranker-Base	Zero-shot	79.14 %	86.57 %	82.69 %	0.80
Scholar-DBLP	BGE-Reranker-v2-M3	Zero-shot	87.35 %	92.77 %	89.98 %	0.80
Scholar-DBLP	BGE-Reranker-Base	Fine-tuned	95.35 %	96.42 %	95.88 %	0.60

Table 23: Reranker performance in zero-shot and fine-tuned configurations.

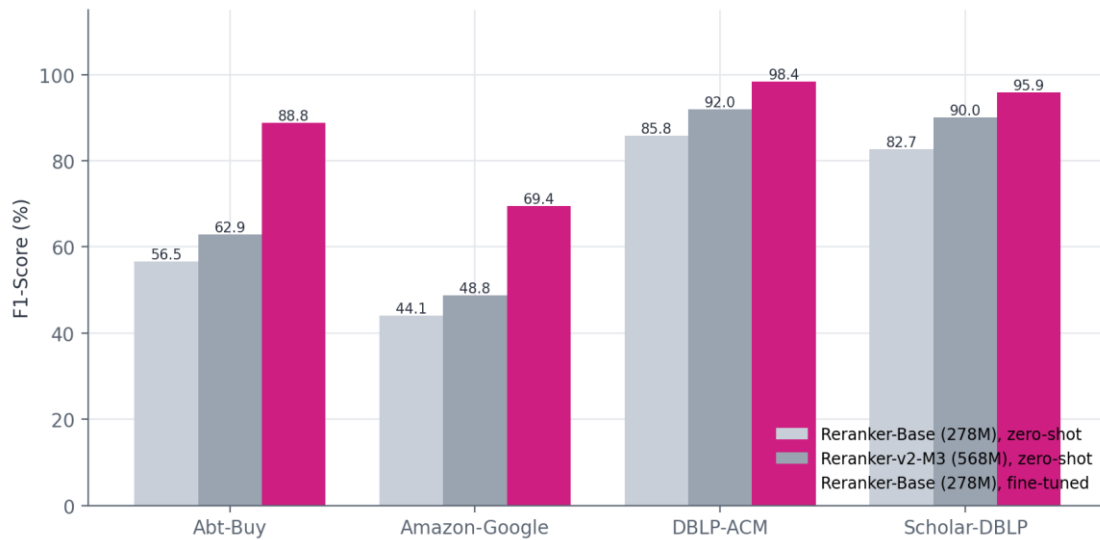


Figure 22: Reranking architectures before and after task-specific fine-tuning. Without adaptation both rerankers fall substantially below the compact fine-tuned cross-encoders despite being eight to seventeen times larger; after fine-tuning the 278-million-parameter model attains the highest F1-score recorded anywhere in this thesis on Abt-Buy.

The prediction is confirmed in substance. Without task-specific training, even the larger reranker — at some seventeen times the size of the leading fine-tuned architectures — trails them by between 5.86 and 19.45 percentage points of F1. General-purpose relevance ranking does not transfer to Entity Resolution without task-specific parameter optimisation, and the deficit is not attributable to insufficient capacity. The magnitude of the deficit scales with domain difficulty in the same manner as the negative sampling

ablation of Section 7.6: approximately six percentage points on the bibliographic collections against seventeen to twenty on the e-commerce collections.

It should be noted that the threshold grid applied to the rerankers extends to 0.99, and that the calibrated optimum for Reranker-Base on Abt-Buy lies at 0.97. Under a coarser grid terminating at 0.95 the reported F1 would have been approximately one and a half points lower; the extension therefore favours the rerankers, and the comparisons that follow are conservative with respect to the conclusions they support.

7.7.2 Fine-Tuned Performance and Sample Efficiency

Fine-tuning transforms the architecture. On Abt-Buy the fine-tuned reranker attains 88.84 % F1, which exceeds by a substantial margin every figure previously reported for this dataset in either protocol: the 79.5 % of the supervised baseline of Chapter 6, the 79.49 % of GTE-Small under Approach B, and the 74.1 % generative ceiling established by Mistral 7B. It constitutes the highest F1-score obtained on Abt-Buy anywhere in this thesis.

The comparison against the compact architectures, however, requires care, because the advantage is not uniform.

Dataset	Train pairs	Zero-shot	Fine-tuned	Best compact model	Δ
Abt-Buy	1,621	56.54	88.84	79.49 — GTE-Small	+9.35 pp
Amazon-Google	2,044	44.11	69.41	68.25 — BGE-Small	+1.16 pp
DBLP-ACM	3,924	85.85	98.40	98.63 — BGE-Small	−0.23 pp
Scholar-DBLP	3,924	82.69	95.88	95.84 — All-MPNet	+0.04 pp

Table 24: Fine-tuned BGE-Reranker-Base against the best compact architecture, ordered by training-set size.

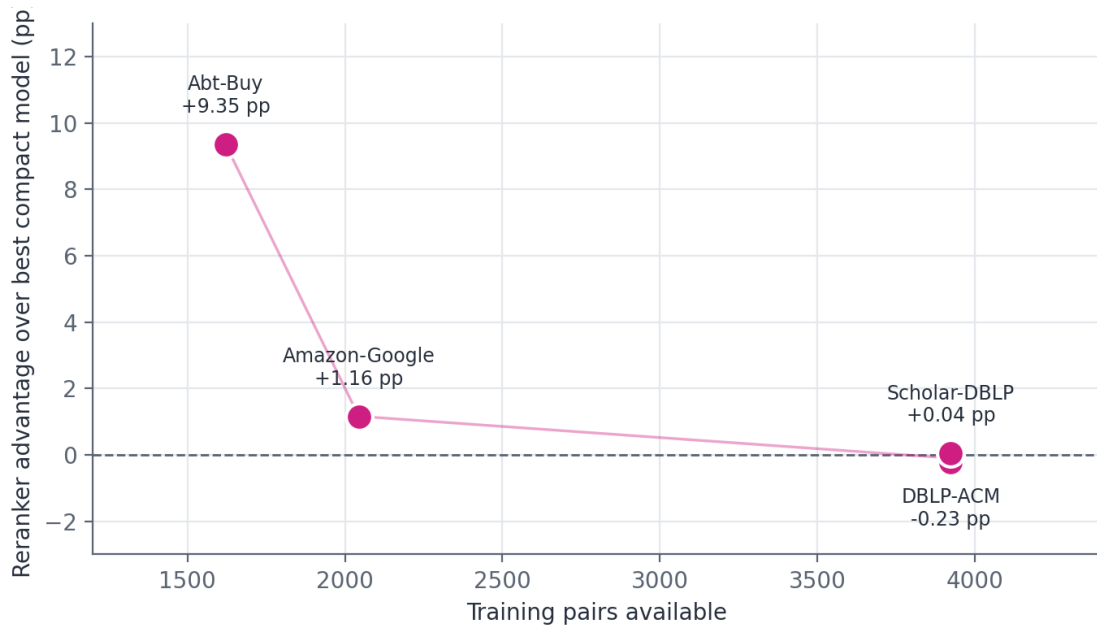


Figure 23: The advantage of the 278-million-parameter reranker over the best compact architecture, plotted against the number of training pairs available. The advantage declines monotonically and vanishes once the training partition exceeds approximately four thousand pairs.

The advantage of the larger architecture declines monotonically with the size of the training partition, from 9.35 percentage points on the 1,621-pair Abt-Buy partition to parity or marginal deficit on the 3,924-pair bibliographic partitions. An architecture 8.4 times larger than GTE-Small and BGE-Small therefore purchases a substantial improvement where supervision is scarce and nothing whatever where supervision is adequate.

The interpretation that best accounts for this pattern concerns initialisation rather than capacity. The four repurposed bi-encoders of Table 14 and the two established baselines are equipped at initialisation with a randomly initialised classification head, and must therefore learn from the training partition alone both the discriminative decision rule and the more basic competence of exploiting cross-attention between two concatenated segments for pairwise classification — a competence for which none of them was pre-trained. On Abt-Buy this must be accomplished from 314 positive examples. The reranker, by contrast, arrives with a head already optimised for precisely this form of pairwise scoring [46]; fine-tuning is required only to redirect an existing capability from relevance towards identity, which is a substantially smaller adaptation. Where the training partition is large enough for the compact architectures to acquire the competence from scratch, the initialisation advantage is exhausted and the superior contrastive pre-training of GTE and BGE [31, 32, 33] closes the gap entirely.

This account is corroborated from an independent direction by the convergence failures documented in Sections 7.5 and 7.8. Both occurred on the smallest training partition and in architectures required to learn pairwise classification from a randomly initialised head. The reranker, on the same partition, not merely avoids collapse but attains the highest score recorded for the dataset. The stability of a fine-tuning run in the low-supervision

regime therefore appears to depend on whether the classification head must be learned or merely adapted.

7.7.3 Threshold Calibration

A prediction advanced in Section 7.7.1 concerned calibration rather than accuracy. The zero-shot rerankers select decision thresholds between 0.70 and 0.97, whereas the fine-tuned architectures of Section 7.5 select 0.60 in twenty of twenty-four configurations. The conjecture was that fine-tuning does not merely improve discriminative accuracy but relocates the probability distribution into a range where a conventional decision boundary becomes appropriate.

The conjecture is confirmed. On Abt-Buy the calibrated threshold falls from 0.97 to 0.60, on Amazon-Google from 0.95 to 0.50, on DBLP-ACM from 0.90 to 0.60 and on Scholar-DBLP from 0.80 to 0.60. Prior to fine-tuning the model’s outputs are compressed towards unity and a threshold approaching 0.99 is required to separate the classes; after fine-tuning the optimum falls into the conventional region, in agreement with every fine-tuned architecture of Section 7.5. Task-specific supervision therefore performs two functions that are analytically separable: it improves the ordering of the pairs, which is reflected in the F1 gains of Table 23, and it calibrates the magnitudes, which is reflected in the relocation of the threshold [13, 18].

The practical consequence is that the apparent advantage of a zero-shot reranker — that it requires no labelled data — is only partly real. A model deployed at the conventional default threshold of 0.5 would perform far below the figures of Table 23, since at that threshold its outputs classify nearly every candidate pair as a match. Obtaining the reported zero-shot performance requires a labelled calibration set, which is to say that some labelled data is necessary in any case. Once such data exists, the evidence of Table 23 indicates that it is far better spent on fine-tuning than on calibration alone.

7.8 Robustness: Substitution of the Blocking Encoder

The use of BGE-Small-en-v1.5 both as the blocking encoder and as one of the six evaluated matchers invites a legitimate objection. Since the candidate pairs are selected within the embedding space of BGE, the BGE-derived cross-encoder might enjoy an advantage unavailable to the remaining architectures. To settle the question the entire pipeline was re-executed on all four datasets with GTE-Small substituted as the blocking encoder, every other element of the configuration held constant — a total of twenty-four additional fine-tuning runs over an entirely different candidate pool.

Model	Abt-Buy (BGE / GTE)	Amazon-Google (BGE / GTE)	DBLP-ACM (BGE / GTE)	Scholar-DBLP (BGE / GTE)
GTE-Small	79.49 / 81.92	68.08 / 67.05	98.56 / 98.69	95.50 / 95.87
BGE-Small-en-v1.5	79.09 / 81.39	68.25 / 66.85	98.63 / 98.72	95.42 / 96.01
MS-MARCO MiniLM L-6	79.44 / 80.73	63.42 / 61.94	97.95 / 97.93	95.54 / 95.76

Model	Abt-Buy (BGE / GTE)	Amazon-Google (BGE / GTE)	DBLP-ACM (BGE / GTE)	Scholar-DBLP (BGE / GTE)
Arctic-Embed-XS	71.18 / 77.86	63.70 / 62.71	98.29 / 98.41	94.87 / 95.42
E5-Small-v2	32.79 / 79.51	62.61 / 63.35	98.18 / 98.17	95.38 / 95.98
All-MPNet-Base-v2	76.15 / 33.19	67.08 / 64.66	97.29 / 97.66	95.84 / 96.19

Table 25: F1-scores under hard-negative training with each blocking encoder. Bold entries mark the two collapse events discussed in the text.

Four conclusions follow.

The objection is not sustained. BGE-Small does not lose its standing when it ceases to be the blocking encoder: its F1 on Abt-Buy rises by 2.30 points, closely matching the 2.43-point rise of GTE-Small and the 1.29-point rise of the MS-MARCO MiniLM. Were a circularity advantage present, a disproportionate decline would be expected. On DBLP-ACM the ordering of all six architectures is preserved and every model moves by less than half a point. On Amazon-Google the three leading models are unchanged.

The coherence of the shift is itself informative. The three leading architectures move in the same direction and by comparable magnitudes while retaining their relative positions. Were the underlying measurements noisy at the scale of whole percentage points, such coherence would be improbable; Section 7.10 quantifies this intuition directly.

The choice of BGE as the unified blocking encoder is retrospectively vindicated.

BGE-Small equals or exceeds GTE-Small as a blocker on every dataset in terms of recall@5, and the margin is substantial on Amazon-Google. This constitutes independent justification for a design decision that Section 7.2 defended only on grounds of experimental control.

Two collapse events occur, and they relocate between architectures. Under BGE blocking, E5-Small-v2 fails on Abt-Buy at 32.79 %; under GTE blocking it recovers completely to 79.51 %, a swing of 46.72 percentage points. Conversely All-MPNet-Base-v2, which performs normally at 76.15 % under BGE blocking, collapses to 33.19 % under GTE blocking with precision of 19.90 % — equal to the positive rate — against recall of 100 %.

The two failures exhibit an identical signature but occur in different architectures under different blocking encoders, which is inconsistent with the prefix-mismatch explanation advanced in Section 7.5: were the prefixes the operative cause, E5 would fail under both configurations. Abt-Buy furnishes the smallest training partition of the four datasets, 1,621 pairs of which only 314 are positive, and both collapses occur there and nowhere else. The most plausible interpretation is therefore a stochastic failure of convergence under an aggressive peak learning rate on a small and imbalanced training set, rather than any property of either architecture. This revision is important: it reclassifies the E5 anomaly from an architectural limitation to a training artefact, and Section 7.10 provides the variance estimate that makes the interpretation quantitatively plausible.

Blocking encoder	Best F1, hard negatives	Best F1, random negatives	Difference
BGE-Small-en-v1.5	79.49 %	47.96 %	+31.53 pp
GTE-Small	81.92 %	50.85 %	+31.07 pp

Table 26: Replication of the negative sampling ablation under substituted blocking (Abt-Buy).

Table 26 establishes that the principal finding of this chapter is invariant to the blocking encoder: the advantage of hard negatives is 31.53 points under BGE blocking and 31.07 under GTE blocking, a discrepancy of less than half a point measured over an entirely different candidate pool and an entirely different test partition. The per-architecture figures under GTE blocking — 50.85 % for GTE-Small, 48.32 % for MiniLM, 47.83 % for E5, 46.75 % for BGE-Small and 46.60 % for Arctic-Embed — reproduce the same precision–recall signature, with recall between 97.9 % and 99.3 % against precision between 30.5 % and 34.3 %. The influence of negative sampling quality is therefore not a property of any single architecture or candidate pool but of the task itself.

7.9 Benchmark Inflation

Sections 1.4 and 7.6 argued that the prevailing practice of random negative sampling inflates reported performance. The ablation of Section 7.6 measured what is lost when a model is *trained* on easy negatives. It did not measure what is gained when a model is *evaluated* on them, and that is the quantity relevant to interpreting the published literature. This section supplies that measurement.

The experiment is deliberately minimal. The fine-tuned models of Section 7.5 are taken unchanged — no weight is modified — and evaluated on a second test partition constructed for each dataset. The positive pairs are identical to those of the original test partition. The negative pairs are drawn uniformly from the Cartesian product of the two tables, with three exclusions enforced in combination: ground-truth matches, every pair appearing in the hard-negative training partition, and every pair appearing in the random-negative training partition. The total cardinality and the positive rate are matched exactly to the original partition, so that the two F1-scores are directly comparable. The serialisation function is the one used throughout the pipeline, which eliminates a text-formatting confound detected and corrected in an earlier iteration of this experiment.

Dataset	Model	Hard-negative test	Random-negative test	Inflation	Precision (random)
Abt-Buy	GTE-Small	79.65 %	89.34 %	+9.69 pp	100.00 %
Amazon-Google	BGE-Small	68.25 %	83.05 %	+14.80 pp	99.82 %
DBLP-ACM	BGE-Small	98.63 %	99.20 %	+0.57 pp	100.00 %
Scholar-DBLP	All-MPNet-Base-v2	95.84 %	99.62 %	+3.79 pp	99.70 %

Table 27: Benchmark inflation. Identical model weights evaluated on two test partitions differing only in the provenance of the negative pairs.

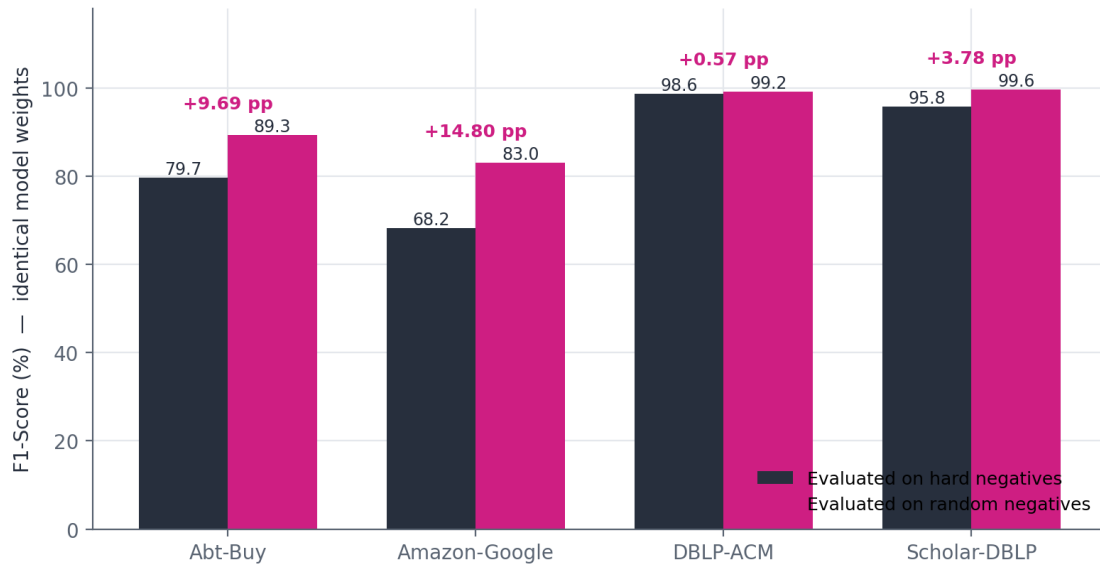


Figure 24: Benchmark inflation across the four datasets. The model weights are identical in both bars of each pair; only the construction of the negative examples in the evaluation partition differs. The inflation is largest in the noisiest domain and negligible in the cleanest.

The same model, with frozen weights and identical positive pairs, reports 79.65 % or 89.34 % on Abt-Buy depending solely on how the negative examples of the evaluation set were constructed. On Amazon-Google the gap reaches 14.80 percentage points.

The mechanism is unambiguous and is visible in the final column of Table 27. Precision under random-negative evaluation reaches 99.7 % to 100.0 % in all four datasets: when the negatives are drawn from unrelated regions of the product or publication space, the model commits essentially no false positives. Recall, by contrast, is largely unchanged, since the positive pairs are identical — on Amazon-Google and DBLP-ACM it is numerically identical at 71.10 % and 98.41 % respectively, the calibrated threshold being unchanged in both cases. The entire inflation is therefore attributable to precision, and precision is determined by the negatives.

Two consequences follow. The first is interpretive: a reported F1-score in Entity Resolution is not interpretable without a specification of how the evaluation set was constructed, and comparisons across papers that do not share that specification are not meaningful [6, 53]. The second is that the inflation scales with domain difficulty exactly as every other effect measured in this chapter — 14.80 points on the hardest dataset against 0.57 on the easiest. The practical implication is that the published literature is least reliable precisely in the commercial domains where Entity Resolution has the greatest economic value [15].

7.10 Seed Stability

Every configuration reported in Sections 7.5 to 7.9 was trained once under a fixed random seed of forty-two. Fine-tuning is stochastic in the initialisation of the classification head, in the ordering of mini-batches and in dropout sampling, and the magnitude of the resulting

variance determines which of the reported differences may be regarded as meaningful. This section measures it.

GTE-Small was retrained three times on the Abt-Buy hard-negative partition under seeds 42, 0 and 123, with every other element of the configuration held constant. Abt-Buy was selected deliberately: it possesses the smallest training partition and the fewest positive examples of the four datasets, and is therefore the configuration most susceptible to instability. The resulting variance is consequently a conservative upper estimate for the study as a whole.

Seed	Precision	Recall	F1-Score	Threshold
42	83.86 %	74.93 %	79.15 %	0.60
0	80.05 %	83.29 %	81.64 %	0.60
123	78.30 %	79.25 %	78.77 %	0.60
Mean	80.74 %	79.16 %	79.85 %	—
Standard deviation	± 2.84	± 4.27	± 1.47	—

Table 28: Seed stability analysis. GTE-Small, Abt-Buy, hard negatives, three random seeds.

The standard deviation of the F1-score is 1.47 percentage points. Three observations follow.

First, the principal findings of this chapter exceed this variance by a wide margin. The negative sampling ablation of 31.53 points is twenty-one times larger; the benchmark inflation of 14.80 points is ten times larger; the reranker advantage of 9.35 points on Abt-Buy is six times larger. These effects are not attributable to measurement noise.

Second, the differences among the three leading architectures — 0.40 points between GTE-Small and BGE-Small on Abt-Buy, 0.17 points between BGE-Small and GTE-Small on Amazon-Google — fall well inside it. This confirms the interpretive restraint exercised in Section 7.5: the leading architectures are mutually indistinguishable and the ordering reported in Table 19 should not be read as a ranking.

Third, and least expected, the calibrated decision threshold is identical across all three seeds at 0.60. The stability of the pipeline therefore extends beyond the aggregate score to the probability calibration itself, which strengthens the argument of Section 7.14 concerning the limited consequences of oracle threshold selection.

The collapse events documented in Section 7.8 acquire a quantitative interpretation in this light. A standard deviation of 1.47 points under benign conditions is not sufficient on its own to produce a swing from 76.15 % to 33.19 %; what the seed analysis establishes is that the optimisation on this partition is sensitive to initialisation, and that a configuration combining the largest architecture, the smallest batch size and the fewest positive examples may occasionally fall outside the ordinary variance regime into a degenerate solution. A complete characterisation would require repeated runs of every architecture, which Section 7.14 identifies as the principal methodological limitation of this chapter.

7.11 Cross-Protocol Validation

Section 7.2 acknowledged that Approach A and Approach B differ in blocking, partitioning, optimisation and evaluation-set size, and deferred the question of comparability to an empirical test. This section performs it.

The MS-MARCO MiniLM cross-encoder is the single architecture common to both protocols. Because everything surrounding it changes — the blocking encoder, the partitioning strategy, the number of training epochs, the peak learning rate, the size of the evaluation partition and the locus of threshold calibration — its behaviour constitutes a direct measurement of the sensitivity of this thesis’s results to the experimental design itself.

Dataset	Approach A	Approach B	Δ	Note
Abt-Buy	79.50 (L-6)	79.44 (L-6)	−0.06 pp	Identical architecture
DBLP-ACM	98.30 (L-6)	97.95 (L-6)	−0.35 pp	Identical architecture
Scholar-DBLP	95.70 (L-6)	95.54 (L-6)	−0.16 pp	Identical architecture
Amazon-Google	69.60 (L-12)	63.42 (L-6)	−6.18 pp	Different depth
Amazon-Google	69.60 (L-12)	67.93 (L-12)	−1.67 pp	Identical architecture

Table 29: Cross-protocol validation via the shared MS-MARCO MiniLM architecture.

On the three datasets where the identical six-layer variant was used under both protocols, the agreement is close to exact. Abt-Buy moves from 79.50 % to 79.44 %, DBLP-ACM from 98.30 % to 97.95 % and Scholar-DBLP from 95.70 % to 95.54 % — deviations of between six hundredths and thirty-five hundredths of a percentage point, despite an evaluation partition between two and six times larger and a training partition reduced to thirty percent of the candidate pool. All three lie well inside the ± 1.47 -point variance established in Section 7.10. Since the two protocols share neither their blocking encoder nor their splitting rule nor their optimisation budget, the stability of these figures cannot be attributed to a common source of bias. The natural interpretation is that each domain imposes an intrinsic ceiling on cross-encoder resolution accuracy, that this ceiling is a property of the heterogeneity and incompleteness of the underlying records rather than of the pipeline used to measure it, and that the estimates reported in Chapter 6 are therefore externally valid.

Amazon-Google initially appeared to constitute an exception, with a discrepancy of 6.18 percentage points. The explanation is architectural rather than statistical: Approach A deliberately upgraded the matcher for this dataset to the twelve-layer variant on the grounds that the alphanumeric versioning complexity of software listings demanded additional representational depth, whereas Approach B employed the six-layer variant uniformly across all six architectures. To test this attribution directly, the twelve-layer variant was trained post hoc under the Approach B protocol with every other setting unchanged. Its F1-score of 67.93 % reduces the cross-protocol discrepancy to 1.67 points, approximately one quarter of its original magnitude, and the residual lies within the combined effect of the remaining protocol differences.

This experiment therefore yields two conclusions. The first is that the two protocols are empirically equivalent on all four datasets once the architecture is held constant, which disposes of the comparability objection. The second is incidental but substantive: within a single protocol and under identical data and hyperparameters, the twelve-layer variant exceeds the six-layer variant by 4.51 percentage points on Amazon-Google, while on the remaining three datasets the six-layer variant was already among the leaders. Transformer depth is therefore a domain-conditional resource, valuable precisely where alphanumeric noise is severe and largely irrelevant elsewhere — a cleaner measurement of the depth effect than the original cross-protocol comparison could provide.

A secondary implication concerns supervision efficiency. On DBLP-ACM the reduction of the training partition from 8,970 to 3,924 pairs costs 0.35 percentage points of F1, and on Scholar-DBLP the reduction from 12,271 to 3,924 pairs costs 0.16. Fine-tuned cross-encoders in these domains therefore saturate well below the volume of supervision employed in Approach A, which is a practically useful result given the cost of manual labelling [21].

7.12 Positioning Against Established Entity Resolution Systems

A natural question concerns how the figures reported in this chapter relate to the published results of established Entity Resolution systems on the same four benchmarks. Table 30 places them side by side. The comparison must be read with the caveat developed below.

Dataset	Magellan [54]	DeepMatcher + [22]	Ditto [3]	This work — random-negative evaluation	This work — hard-negative evaluation
Abt-Buy	43.6	62.8	89.3	89.3	79.5
Amazon-Google	49.1	69.3	75.6	83.1	68.3
DBLP-ACM	98.4	98.4	99.0	99.2	98.6
DBLP-Scholar	92.3	94.7	95.6	99.6	95.8
Matching model	classical ML	RNN + FastText	RoBERTa, 125 M	BERT-small, 33 M	BERT-small, 33 M

Table 30: Indicative positioning against established Entity Resolution systems. Published figures as compiled in [53]; the two rightmost columns report the present work under the two evaluation protocols of Section 7.9.

Under the evaluation protocol conventionally employed in the literature — that is, with randomly sampled negatives in the test partition — the compact architectures of this thesis are competitive with or superior to the published state of the art, despite using a 33-million-parameter encoder against the 125-million-parameter RoBERTa of Ditto [3] and without any of the three task-specific optimisations that system introduces. Under the strict protocol adopted throughout this thesis, the same models report figures approximately ten points lower.

The comparison is indicative rather than rigorous, and this must be stated plainly. The test partitions are not identical: they differ in the blocking procedure that generated them, in

the splitting ratio and in the random seed. No claim of superiority over Ditto is advanced, and none would be defensible on this evidence. What the table does establish is that the figures reported in this chapter occupy the same performance region as the published literature when measured in the same way, and that the apparent shortfall of the strict protocol is a measurement artefact of the kind quantified in Section 7.9 rather than a deficiency of the models.

This also explains why no direct re-execution of Ditto, DeepMatcher or Magellan was undertaken. To compare fairly would require reproducing their evaluation protocol, which is precisely the protocol this thesis argues is inflated. The alternative adopted here — holding a strict protocol constant and measuring what relaxing it is worth — addresses the same question from the opposite direction and yields a quantity the direct comparison could not produce.

7.13 Discussion and Synthesis

7.13.1 The Hierarchy of Factors

The experiments of this chapter permit, for the first time in this thesis, a direct comparison of the factors that determine resolution accuracy, all measured within a single pipeline and on identical data. Table 31 and Figure 25 present that comparison.

Decision	Maximum observed gain	Evidence
Hard negatives rather than random negatives	+31.53 pp	§7.6, Table 21
Task-specific fine-tuning of a pre-trained reranker	+32.30 pp	§7.7, Table 23
Fine-tuned discriminative model rather than prompted generation	+14.74 pp	§7.7 and Chapter 6
Transformer depth in alphanumerically noisy domains	+4.51 pp	§7.11, Table 29
Choice among the three leading compact architectures	< 2 pp	§7.5, Table 19
Train/test partitioning ratio	0.06–0.35 pp	§7.11, Table 29

Table 31: The hierarchy of factors determining Entity Resolution accuracy, measured within a single pipeline.

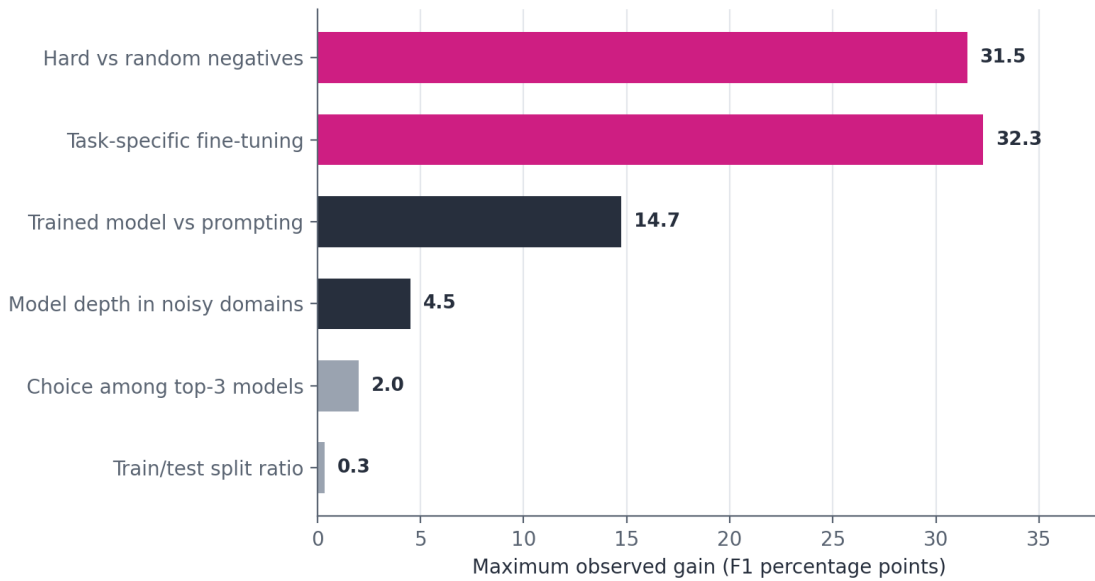


Figure 25: The hierarchy of factors. The two decisions concerning the training data exceed by an order of magnitude the decisions concerning model selection and experimental design, both of which fall at or below the measurement variance established in Section 7.10.

The ordering is unambiguous and, in its practical implication, inverts the emphasis of much of the literature. The two decisions that concern what the model is given to learn — the provenance of the negative examples and whether task-specific adaptation is performed at all — dominate every decision concerning which model is selected. The two decisions that occupy the greatest share of attention in benchmark-driven research, namely architecture selection and partitioning protocol, fall at or below the ± 1.47 -point measurement variance established in Section 7.10.

7.13.2 A Unified Account of the Precision Collapse

The most consequential result to emerge from placing the two approaches alongside one another is not a number but a mechanism.

Chapter 6 documented, across three generative architectures and four prompting strategies, a systematic bias towards affirmative classification: recall approaching or reaching one hundred percent, precision collapsing into the twenties and thirties, and an F1-score dominated entirely by the recall term. That pattern was termed the “Yes-Man” phenomenon and attributed to the alignment procedures under which instruction-tuned models are optimised, which reward agreeable and non-committal responses [13, 18].

Section 7.6 reproduces an identical signature in a setting where no alignment procedure exists. Cross-encoders trained on random negatives — models of twenty-two to one hundred and nine million parameters, without instruction tuning, equipped with a single-logit classification head — exhibit recall between 98.4 % and 100.0 % against precision between 25.3 % and 31.7 % on the e-commerce collections. The profile is indistinguishable from that of LLaMA 3 8B under zero-shot prompting in Section 6.2.1, despite a difference of more than two orders of magnitude in parameter count.

The convergence of these failure modes onto a single precision–recall signature indicates that alignment is not the operative cause. What the generative models and the randomly-trained cross-encoders have in common is not their training objective but their evidential history: in neither case has the model been required to reject a pair that is semantically proximate yet referentially distinct. The generative models were never fine-tuned at all; the randomly-trained cross-encoders were fine-tuned on negatives drawn from unrelated regions of the record space, where rejection follows from surface dissimilarity alone. In both cases the model converges on the rule *if two records are similar, they match*, which is optimal on the evidence available to it and catastrophic at the semantic boundary where the evaluation partitions of this thesis are located.

This reframes the “Yes-Man” phenomenon from a behavioural idiosyncrasy of instruction-tuned language models into the general signature of a decision boundary that has never been constrained from the negative side.

The account is completed by a distinction that the reranker experiments make visible. The zero-shot rerankers of Section 7.7 do *not* exhibit this signature: their probability estimates separate the classes informatively and can be traded between precision and recall by threshold selection, yet they consistently underperform because the quantity they estimate is relevance rather than identity. Two distinct deficiencies must therefore be distinguished — a *degenerate boundary*, in which the model’s outputs carry no discriminative information, and a *misaligned objective*, in which they carry information about the wrong target. Random-negative training and unadapted generation produce the former; general-purpose reranking produces the latter. Both are remedied by task-specific fine-tuning on hard negatives, and neither is remedied by increasing model scale.

A constructive corollary follows. If the degenerate boundary arises from the absence of boundary supervision rather than from architecture, then supplying that supervision should remedy it irrespective of architecture — which is precisely what the hard-negative regime achieves for the cross-encoders, restoring balanced precision–recall profiles and yielding improvements of between 10.86 and 31.53 points of F1, uniformly across architectures and invariantly under substitution of the blocking encoder. It follows that the appropriate remedy for generative Entity Resolution is unlikely to be further prompt engineering, whose returns Chapter 6 demonstrated to be limited and non-monotonic, but parameter-efficient fine-tuning [48, 49] on hard negatives mined by exactly the blocking index that this pipeline already constructs.

7.13.3 Computational Considerations

The efficiency argument of Section 6.6 is reinforced rather than altered by the present chapter. The fine-tuned cross-encoder processes 107.5 candidate pairs per second on the local Apple M1 environment against 0.04 pairs per second for Mistral 7B, a factor of 2,612. The forty-eight fine-tuning runs of Section 7.5, the twenty-four robustness runs of Section 7.8, the reranker experiments of Section 7.7 and the seed analysis of Section 7.10 were all executed on the same consumer hardware.

By contrast, the verification run of Gemma 7B on the Approach B test partition of Abt-Buy — 3,784 pairs across five prompting strategies at temperature 0.0 — required

approximately seventeen days of continuous execution for a single dataset, which is why it was not extended to the remaining three. That run served a specific purpose: temperature 0.0 enforces greedy decoding and therefore deterministic output, making it the only fully reproducible generative measurement in this thesis. It reproduced the non-monotonic few-shot trajectory established in Chapter 6, with the peak at a single demonstration (71.87 % F1) and progressive degradation to 69.36 % at eight demonstrations, confirming that the context-overload effect is not an artefact of the smaller Approach A evaluation partition.

7.14 Threats to Validity

7.14.1 Oracle Threshold Selection

The experiments of this chapter calibrate the decision threshold directly on the evaluation partition, whereas Approach A calibrates on a held-out validation partition. The figures reported here are therefore optimistic in the strict sense and should be read as an upper bound on attainable F1 rather than as an estimate of deployment performance.

Four considerations bound the magnitude of the resulting optimism. First, the selected threshold is 0.60 in twenty of the twenty-four hard-negative configurations and 0.60 or 0.70 in twenty-three, indicating that the F1 surface is flat in the neighbourhood of the optimum and that the calibration extracts little dataset-specific information. Second, the seed analysis of Section 7.10 found the calibrated threshold to be identical at 0.60 across three independent training runs, which further constrains the scope for overfitting. Third, and most directly, the cross-protocol validation of Section 7.11 compares an oracle-calibrated Approach B against a validation-calibrated Approach A on identical architectures and finds discrepancies of 0.06 to 0.35 percentage points — an empirical upper bound on the bias introduced by this protocol choice. Fourth, the ablation of Section 7.6 and the inflation experiment of Section 7.9 are relative comparisons in which both arms are calibrated identically, so that the reported differences are unaffected by the procedure.

A protocol reserving a small fraction of the training partition for calibration would eliminate the concern entirely and constitutes the natural first refinement of this evaluation.

7.14.2 Absence of Repeated Runs

Section 7.10 established a standard deviation of 1.47 percentage points on the most volatile configuration, but this estimate derives from a single architecture on a single dataset. A complete characterisation would require three or more runs of every architecture on every dataset, a total of one hundred and forty-four additional fine-tuning runs. The rankings reported in Section 7.5 should accordingly be interpreted as indicative rather than as statistically significant orderings, and the claim defended in this chapter is the weaker but robust one that the leading architectures are mutually indistinguishable at a margin below two percentage points.

The collapse events documented in Sections 7.5 and 7.8 render this limitation more than formal. That the same failure signature appears in E5 under one blocking encoder and in

All-MPNet under another is direct evidence of run-to-run instability on the smallest training partition, and quantifying that instability across architectures is a precondition for interpreting any single-seed result on Abt-Buy.

7.14.3 Blocking Recall as a Performance Ceiling

Because both protocols evaluate exclusively on candidate pairs surviving the blocking stage, the reported F1-scores are conditional on blocking recall and do not measure end-to-end system performance. On Amazon-Google, where the BGE-Small index recovers approximately 87 % of ground-truth matches, roughly one true match in eight is unrecoverable by any downstream matcher, and the 68.25 % F1 attained by BGE-Small corresponds to a materially lower end-to-end figure. Reporting metrics that compose blocking recall with matching precision would give a more faithful picture of deployed performance and requires no additional training, being computable from the results already obtained [30].

7.14.4 Asymmetry of the Discriminative–Generative Comparison

The comparison between fine-tuned discriminative models and generative models evaluated without parameter updates is asymmetric, and deliberately so. The question addressed is not which architectural family possesses the greater ultimate capacity, but whether prompted generation constitutes a viable substitute where labelled data is unavailable — which is the decision an engineering team actually faces [4, 5, 6].

Section 7.7 permits that question to be posed more sharply than was previously possible, because it evaluates a discriminative architecture under both conditions. BGE-Reranker-Base without task-specific training falls between 5.86 and 19.45 percentage points short of the compact fine-tuned architectures; the same model after fine-tuning exceeds them on the dataset where supervision is scarcest and matches them elsewhere. The determining factor is therefore neither the architectural family nor the parameter count but the presence of task-specific supervision on hard negatives, and the deficit observed for the generative models in Chapter 6 is consistent with that account rather than with any claim of generative inadequacy in principle.

A symmetric comparison would require parameter-efficient fine-tuning of the generative models [48, 49] on the same hard-negative partitions. Section 7.7 raises the expected value of that experiment considerably: if adapting a 278-million-parameter reranker to the task yields between 12.55 and 32.30 points of F1, the corresponding adaptation of a seven-billion-parameter generative model is the single most informative experiment remaining.

7.14.5 Domain and Language Coverage

The four benchmarks comprise two e-commerce and two bibliographic collections, all in English. The findings of this chapter are therefore established for structured and semi-structured textual records in two domains and one language. The consistent pattern observed throughout — that every measured effect scales with domain difficulty — would predict larger rather than smaller gains in noisier settings such as clinical or financial records, but this remains an extrapolation. Multilingual validation is additionally

motivated by the fact that both rerankers evaluated in Section 7.7 are built on multilingual backbones [44, 45] whose cross-lingual capacity was not exercised here.

7.14.6 Uniform Hyperparameters

The uniform allocation of four epochs, batch size eight and a peak learning rate of 5×10^{-5} is required for the comparability of six architectures but entails that no architecture is reported at its individual optimum. The absolute figures of Sections 7.5 to 7.8 are consequently lower bounds on per-architecture performance. The collapse analysed in Section 7.8 additionally suggests that the selected learning rate is aggressive for the largest architecture on the smallest training partition, and that a per-architecture learning rate search would improve both the stability and the absolute level of the reported results.

8 Conclusions

8.1 Final Deductions on Generative LLMs in Entity Resolution

By evaluating three prominent autoregressive Large Language Models LLaMA 3 [10], Mistral 7B [11], and Gemma 7B [28] against a fine-tuned BERT Cross-Encoder baseline [9], this thesis delineates the fundamental capabilities and inherent limitations of generative reasoning in high-precision Entity Resolution. A foundational premise verified by this research is the absolute necessity of isolating "hard negatives" via dense vector HNSW blocking [7]. The severe generative pathologies uncovered in this study such as sycophancy biases and context overload manifested exclusively because the models were evaluated exclusively on hard-negative pairs at the limits of their discriminative capacity. Had the methodology relied on traditional random negative sampling, the mathematically trivial mismatches would have artificially inflated the zero-shot performance, masking the true architectural limitations of the generative models and rendering the benchmark scientifically invalid [4, 5, 6].

The empirical results consistently show that generative LLMs currently face a rigid performance ceiling when deployed in noisy, unstructured domains. In noisy, heterogeneous e-commerce environments characterized by severe schema mismatches and alphanumeric complexity, such as the Amazon-Google dataset, the generative architectures failed to surpass a 51.6% F1-Score, regardless of the prompt complexity. Because they process text sequentially from left to right, they structurally struggle to cross-reference subtle but critical versioning anomalies. On the other hand, when the data is clean and highly structured, such as the DBLP-ACM bibliographic records, models like Mistral can achieve remarkable production-grade accuracy, peaking at a 96.6% F1-Score under pure zero-shot conditions without any task-specific parameter optimization. Nevertheless, across all domains, the bidirectional self-attention mechanism of the fine-tuned BERT Cross-Encoder [9, 21] consistently established the superior empirical ceiling, proving that fine-tuned discriminative heads remain unmatched for enterprise-grade data integration.

The evaluation also revealed opposing alignment biases across models: LLaMA 3 and Gemma 7B frequently exhibited the "Yes-Man" phenomenon, demonstrating an extreme recall bias that favored affirmative matching decisions at the cost of a substantial decline in precision. Through the exhaustive Few-Shot Ablation Study, it was empirically demonstrated that in-context learning trajectories are highly domain-dependent. While complex domains like Amazon-Google required exactly 5 balanced examples to neutralize these biases and establish a "sweet spot," domains that closely mirror the models' pre-training corpora such as the Scholar-DBLP dataset triggered what can be termed the "1-Shot Anomaly." In this noisy academic domain, the deterministically constrained Gemma 7B achieved its absolute peak performance of 83.5% F1-Score with a single demonstrative example. Crucially, across all models and datasets, expanding the prompt to 8 complex examples consistently induced "Context Overload." Expanding the context window with excessive and contradictory formatting rules consistently degraded precision, demonstrating that beyond a certain threshold, additional demonstrations reduce rather than improve classification accuracy [6].

Finally, the Confidence Elicitation experiments mathematically proved that researchers and engineers cannot rely on generative models when they claim to be certain about a classification. The analysis revealed profound Overconfidence Bias (Miscalibration), with models like Mistral and LLaMA 3 frequently expressing near-absolute verbal certainty (scoring above 4.5 out of 5.0) on completely erroneous predictions [13]. However, the targeted methodological evolution applied to Gemma 7B combining a strict zero-temperature deterministic constraint with a hardware-calibrated sample for Chain-of-Thought and Confidence evaluation provided a critical solution. By neutralizing generative stochasticity, the confidence prompt transitioned from a superficial verbal token into an ultra-precise operational filter, effectively suppressing the affirmative bias and stabilizing the model's precision. Ultimately, while autoregressive models possess immense semantic capability, they fundamentally lack the logit-grounded probabilistic calibration native to fine-tuned classification heads [21], suggesting that strict deterministic constraints are necessary and rigorous external filtering before they can be safely integrated into automated, high-stakes data pipelines.

8.2 Methodological Limitations and Future Work

While this thesis establishes a rigorous baseline for generative evaluation through the isolation of hard negatives, several methodological limitations must be acknowledged to contextualize the findings.

First, regarding architectural fairness, the experimental protocol compares a bidirectionally fine-tuned BERT Cross-Encoder against generative models (LLaMA 3 and Mistral) operating entirely out-of-the-box via zero-shot and few-shot in-context learning. While this appropriately measures the immediate, parameter-free utility of generative LLMs, asserting that they inherently cannot replace cross-encoders is a broad generalization. A truly equivalent architectural comparison would require applying Parameter-Efficient Fine-Tuning (PEFT) techniques, such as LoRA or QLoRA [48, 49], to the generative models to map their massive parametric space directly to the Entity Resolution task. Exploring how fine-tuned LLMs perform on hard negatives remains a critical direction for future research [5, 6].

Second, the parsing methodology utilized during the Confidence Elicitation experiments (Section 5.3) introduces an element of evaluation brittleness [13]. Because generative models produce open-ended text, the strict deterministic extraction of substrings (e.g., requiring exactly "DECISION: YES" and "CONFIDENCE: HIGH") leaves the evaluation vulnerable to formatting variations. If a model reasoned correctly and possessed high certainty but formatted its output slightly differently (e.g., "CONFIDENCE LEVEL: HIGH" or providing a conversational prefix), the script applied an automatic non-match penalty. Consequently, the performance degradation observed during the confidence tests may partially reflect the brittleness of the regex parser rather than purely the epistemic miscalibration of the model.

Finally, while the issue of LLM stochasticity was explicitly addressed during the evaluation of Gemma 7B [28] which was locked under a strict temperature setting of 0.0 to guarantee deterministic outputs the LLaMA 3 [10] and Mistral [11] architectures were evaluated under their native stochastic parameters. Because autoregressive generation is inherently variable, relying on single-run evaluations to identify "sweet spots" or declare "catastrophic collapses" at extended context lengths (e.g., 8-shots) introduces susceptibility to sampling variance. Future benchmarking protocols should enforce multiple iterative runs (e.g., N=3 or N=5) for each configuration, averaging the F1-Scores to smooth out token generation stochasticity and validate the statistical significance of context overload.

References

- [1] Christophides, V., Efthymiou, V., Stefanidis, K.: Entity Resolution in the Web of Data. *Synthesis Lectures on the Semantic Web: Theory and Technology*, Vol. 5, No. 3, pp. 1–122. Morgan & Claypool Publishers (2015). DOI: 10.2200/S00655ED1V01Y201507WBE013
- [2] Papadakis, G., Ioannou, E., Palpanas, T.: Entity Resolution: Past, Present and Yet-to-Come. In: *Proceedings of the 23rd International Conference on Extending Database Technology (EDBT 2020)*, Copenhagen, Denmark, pp. 647–650 (2020).
- [3] Li, Y., Li, J., Suhara, Y., Doan, A., Tan, W.C.: Deep Entity Matching with Pre-Trained Language Models. *Proceedings of the VLDB Endowment*, Vol. 14, No. 1, pp. 50–60 (2020).
- [4] Peeters, R., Steiner, A., Bizer, C.: Entity Matching using Large Language Models. *arXiv preprint arXiv:2310.11244* (2024).
- [5] Zeakis, A., Papadakis, G., Skoutas, D., Koubarakis, M.: AvengER: Ensembling and Fine-Tuning LLMs for SELECT Prompts in Entity Resolution. *arXiv preprint* (2025).
- [6] Arvanitis-Kasinikos, I., Papadakis, G.: Entity Matching with 7B LLMs: A Study on Prompting Strategies and Hardware Limitations. In: *Proceedings of the 27th International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP 2025)*, co-located with EDBT/ICDT 2025, Barcelona, Spain. *CEUR Workshop Proceedings*, Vol. 3931, pp. 31–38 (2025).
- [7] Karapiperis, D., Tjortjis, C., Verykios, V.S.: LSBLOCK: A Hybrid Blocking System Combining Lexical and Semantic Similarity Search for Record Linkage. *International Hellenic University / Hellenic Open University* (2023).
- [8] Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3982–3992 (2019).
- [9] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of NAACL-HLT*, pp. 4171–4186 (2019).
- [10] Dubey, A., et al. (Meta AI): The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783* (2024).
- [11] Jiang, A.Q., Sablayrolles, A., Mensch, A., et al.: Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).
- [12] Bopardikar, R., Wang, J., Zou, J.: Structured Multi-Step Reasoning for Entity Matching Using Large Language Model. *arXiv preprint arXiv:2511.22832* (2025).

- [13] Hovsepian, K., Liu, D., Murugesan, S.: Label with Confidence: Effective Confidence Calibration and Ensembles in LLM-Powered Classification. In: Proceedings of the First Workshop on Generative AI for E-Commerce (GenAIECom 2024), October 25, 2024, pp. 1–8. ACM, New York, NY, USA (2024).
- [14] Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, Vol. 7, No. 3, pp. 535–547 (2019). DOI: 10.1109/TBDATA.2019.2921572
- [15] Köpcke, H., Thor, A., Rahm, E.: Evaluation of entity resolution approaches on real-world match problems. *Proceedings of the VLDB Endowment (PVLDB)*, Vol. 3, No. 1, pp. 484–493 (2010).
- [16] Naumann, F., Herschel, M.: An Introduction to Duplicate Detection. *Synthesis Lectures on Data Management*, Vol. 2, No. 1, pp. 1–87. Morgan & Claypool Publishers (2010). DOI: 10.2200/S00262ED1V01Y201003DTM003
- [17] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35, pp. 24824–24837 (2022).
- [18] Shrivastava, V., Liang, P., Kumar, A.: Llamas Know What GPTs Don't Show: Surrogate Models for Confidence Estimation. *arXiv preprint arXiv:2311.08877* (2023).
- [19] DeLucia, A., et al.: Strength in Numbers: Estimating Confidence of Large Language Models by Prompt Agreement. In: Proceedings of TrustNLP Workshop (co-located with ACL 2023) (2023).
- [20] Zhang, Y., et al.: Effectiveness of Chain-of-Thought in Distilling Reasoning Capability from Large Language Models. *arXiv preprint arXiv:2511.05184* (2025).
- [21] Karapiperis, D., Akritidis, L., Bozanis, P.: The impact of fine-tuning on entity resolution: An experimental evaluation. *Knowledge-Based Systems*, Vol. 338, Article 115427 (2026).
- [22] Mudgal, S., Li, H., Rekatsinas, T., Doan, A., Park, Y., Krishnan, G., Deep, R., Arcaute, E., Raghavendra, V.: Deep Learning for Entity Matching: A Design Space Exploration. In: Proceedings of the ACM International Conference on Management of Data (SIGMOD 2018), Houston, TX, USA, pp. 19–34 (2018).
- [23] Fu, C., Han, X., He, J., Sun, L.: Hierarchical Matching Network for Heterogeneous Entity Resolution. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI 2020), Yokohama, Japan, pp. 3665–3671 (2020).
- [24] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention Is All You Need. In: *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 30, pp. 5998–6008 (2017).

- [25] Chen, Y., Li, Y., Sun, S., Tang, J.: GNEM: A Generic One-to-Set Neural Entity Matching Framework. In: Proceedings of The Web Conference (WWW) 2021, pp. 1235–1247 (2021). ACM. DOI: 10.1145/3442381.3450119
- [26] Genossar, B., Shraga, R., Gal, A.: FlexER: Flexible Entity Resolution for Multiple Intents. Proceedings of the ACM on Management of Data (SIGMOD 2023), Vol. 1, No. 1, pp. 1–27 (2023). DOI: 10.1145/3588722
- [27] Chatterjee, S., Dietz, L.: BERT-ER: Query-specific BERT Entity Representations for Entity Ranking. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2022), Madrid, Spain, pp. 1466–1477 (2022).
- [28] Gemma Team, Google DeepMind: Gemma: Open Models Based on Gemini Research and Technology. arXiv preprint arXiv:2403.08295 (2024).
- [29] Zeakis, A., Papadakis, G., Skoutas, D.: Pre-trained Embeddings for Entity Resolution: An Experimental Analysis. In: Proceedings of the ACM SIGMOD International Conference on Management of Data (2023).
- [30] Papadakis, G., Skoutas, D., Thanos, E., Palpanas, T.: Blocking and Filtering Techniques for Entity Resolution: A Survey. ACM Computing Surveys, Vol. 53, No. 2, Article 31, pp. 1–42 (2020).
- [31] Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., Zhang, M.: Towards General Text Embeddings with Multi-stage Contrastive Learning. arXiv preprint arXiv:2308.03281 (2023).
- [32] Xiao, S., Liu, Z., Zhang, P., Muennighoff, N.: C-Pack: Packaged Resources To Advance General Chinese Embedding. arXiv preprint arXiv:2309.07597 (2023).
- [33] Xiao, S., Liu, Z., Shao, Y., Cao, Z.: RetroMAE: Pre-Training Retrieval-oriented Language Models Via Masked Auto-Encoder. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022), pp. 538–548 (2022).
- [34] Merrick, L., Xu, D., Nuti, G., Campos, D.: Arctic-Embed: Scalable, Efficient, and Accurate Text Embedding Models. arXiv preprint arXiv:2405.05374 (2024).
- [35] Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., Wei, F.: Text Embeddings by Weakly-Supervised Contrastive Pre-training. arXiv preprint arXiv:2212.03533 (2024).
- [36] Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M.: MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In: Advances in Neural Information Processing Systems (NeurIPS), Vol. 33, pp. 5776–5788 (2020).

- [37] Song, K., Tan, X., Qin, T., Lu, J., Liu, T.Y.: MPNet: Masked and Permuted Pre-training for Language Understanding. In: Advances in Neural Information Processing Systems (NeurIPS), Vol. 33, pp. 16857–16867 (2020).
- [38] Muennighoff, N., Tazi, N., Magne, L., Reimers, N.: MTEB: Massive Text Embedding Benchmark. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023), pp. 2014–2037 (2023).
- [39] Robinson, J., Chuang, C.Y., Sra, S., Jegelka, S.: Contrastive Learning with Hard Negative Samples. In: Proceedings of the 9th International Conference on Learning Representations (ICLR 2021) (2021).
- [40] Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.T.: Dense Passage Retrieval for Open-Domain Question Answering. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020), pp. 6769–6781 (2020).
- [41] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al.: Transformers: State-of-the-Art Natural Language Processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP 2020), pp. 38–45 (2020).
- [42] Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: MS MARCO: A Human Generated MACHINE Reading Comprehension Dataset. arXiv preprint arXiv:1611.09268 (2016).
- [43] Malkov, Y.A., Yashunin, D.A.: Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 42, No. 4, pp. 824–836 (2020). DOI: 10.1109/TPAMI.2018.2889473
- [44] Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., Liu, Z.: BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. arXiv preprint arXiv:2402.03216 (2024).
- [45] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised Cross-lingual Representation Learning at Scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), pp. 8440–8451 (2020).
- [46] Nogueira, R., Cho, K.: Passage Re-ranking with BERT. arXiv preprint arXiv:1901.04085 (2019).
- [47] Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. In: Proceedings of the 7th International Conference on Learning Representations (ICLR 2019) (2019).

- [48] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models. In: Proceedings of the 10th International Conference on Learning Representations (ICLR 2022) (2022).
- [49] Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: QLoRA: Efficient Finetuning of Quantized LLMs. In: Advances in Neural Information Processing Systems (NeurIPS), Vol. 36 (2023).
- [50] Karapiperis, D., Akritidis, L., Bozanis, P., Verykios, V.S.: A Comparative Analysis of Graph-Based and Partition-Based Approximate Nearest Neighbor Search for Large-Scale Entity Resolution. *Journal of Information Science* (2025). DOI: 10.1177/18724981251388888
- [51] Karapiperis, D., Tjortjis, C., Verykios, V.S.: A Suite of Efficient Randomized Algorithms for Streaming Record Linkage. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 36, No. 7, pp. 2803–2813 (2024).
- [52] Karapiperis, D., Verykios, V.S.: Scaling Entity Resolution with K-Means: A Review of Partitioning Techniques. *Electronics*, Vol. 14, No. 18, Article 3605 (2025). DOI: 10.3390/electronics14183605
- [53] Barlaug, N., Gulla, J.A.: Neural Networks for Entity Matching: A Survey. *ACM Transactions on Knowledge Discovery from Data*, Vol. 15, No. 3, Article 52, pp. 1–37 (2021). DOI: 10.1145/3442200
- [54] Konda, P., Das, S., Suganthan G.C., P., Doan, A., Ardalan, A., Ballard, J.R., Li, H., Panahi, F., Zhang, H., Naughton, J., Prasad, S., Krishnan, G., Deep, R., Raghavendra, V.: Magellan: Toward Building Entity Matching Management Systems. *Proceedings of the VLDB Endowment*, Vol. 9, No. 12, pp. 1197–1208 (2016).