

FACULDADE DE CIÊNCIAS DA UNIVERSIDADE DO PORTO
FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Normalization and Efficient Search: Algorithmic Solutions for Geographic Data

João Macedo Lima



Mestrado em Inteligência Artificial

Supervisor: Prof. Rosaldo Rossetti

Second Supervisor: Gonalo Almeida

June 29, 2026

Normalization and Efficient Search: Algorithmic Solutions for Geographic Data

João Macedo Lima

Mestrado em Inteligência Artificial

Approved in oral examination by the committee:

President: Prof. Henrique Lopes Cardoso

Examiner: Prof. Mónica Faria

Supervisor: Prof. Rosaldo Rossetti

June 29, 2026

Resumo

Os dados de endereços geográficos funcionam como uma interface textual entre utilizadores finais e bases de dados espaciais, mas a sua entropia, manifestada através de abreviações, inconsistências estruturais e variações translinguísticas, cria um "Human-Database Gap" entre as entradas do mundo real e os requisitos rígidos das bases de dados relacionais. Assim, a questão central desta investigação é: em que medida pode a IA colmatar esta lacuna a uma escala multinacional, e a que custo computacional?

As métricas clássicas de strings carecem da consciência semântica necessária para equiparar "R. Sta. Catarina" a "Rua de Santa Catarina", enquanto que os geocodificadores mais recentes continuam a depender de índices de palavras-chave para a geração de candidatos, levando-os a perder correspondências semanticamente equivalentes devido a incompatibilidades léxicas. Por outro lado, as alternativas modernas de deep learning, embora superiores em generalização semântica, introduzem a sua própria falha: a recuperação vetorial sem restrições desencadeia alucinações espaciais, recuperando localizações geograficamente distantes mas lexicalmente semelhantes.

Para colmatar estas lacunas, esta dissertação segue uma metodologia de três fases, avaliada contra uma base de dados multinacional de 25 milhões de registos e 50.000 consultas empresariais reais. A Fase 1 avalia componentes isolados, onde arquiteturas de parsing e modelos de correspondência semântica foram avaliados. A Fase 2 integra estes componentes em paradigmas completos de ponta a ponta, avaliados através de uma metodologia "LLM-as-a-Judge". Por fim, a Fase 3 culmina na solução proposta, o Hybrid Neuro-Symbolic Framework: um parser Bi-LSTM fine-tuned combinado com recuperação híbrida em Elasticsearch (k-NN denso com guardrails geográficos booleanos determinísticos) e um ranker XGBoost com Destilação de Conhecimento.

O framework final alcançou superioridade de recuperação em todos os dados europeus, incluindo taxas de vitória de 98% em Espanha e 91% em França, provando empiricamente que o raciocínio semântico pode ser comprimido numa estrutura estatística de alta velocidade sem perda de precisão (MAE do XGBoost de 3,72% relativamente ao modelo teacher Cross-Encoder). No entanto, a avaliação revelou também dois princípios arquiteturais: os sistemas Neuro-Simbólicos exibem dependência do parser, evidenciada pelo colapso de desempenho nos dados dos EUA; e a incorporação de consultas híbridas booleano-vetoriais no Elasticsearch introduz um bottleneck de recuperação de 331ms, demonstrando que o custo computacional da precisão neural acaba por migrar para a camada de persistência da base de dados. Por fim, este trabalho contribuiu com evidência empírica que o "Human-Database Gap" não pode ser resolvido por um único algoritmo, mas apenas através do co-design de flexibilidade semântica, restrições espaciais determinísticas e arquitetura de base de dados.

Palavras-chave: Resolução de Entidades Geográficas • Normalização de Endereços • Correspondência Semântica • Pesquisa Aproximada do Vizinho Mais Próximo • Destilação de Conhecimento • Inteligência Artificial Neuro-Simbólica

Abstract

Geographic address data acts as a textual interface between end users and spatial databases, but its entropy, manifested through abbreviations, structural inconsistencies, and cross-linguistic variations, creates a "Human-Database Gap" between real-world input and the strict requirements of relational databases. The core question of this research is: to what extent can AI bridge this gap at a multinational scale, and at what computational cost?

Classical string metrics lack the semantic awareness required to "R. Sta. Catarina" with "Rua de Santa Catarina", while state-of-the-art geocoders still rely on legacy keyword indices for candidate generation, leading them to miss semantically equivalent matches due to lexical mismatches. On the other hand, modern deep learning alternatives, although they are superior in semantic generalization, introduce their own failure: unrestrained vector retrieval triggers spatial hallucinations, retrieving geographically distant but lexically similar locations.

To address these gaps, this dissertation follows a three-phase methodology, evaluated against a 25-million-record multinational database and a 50,000 real-world enterprise queries. Phase 1 benchmarks isolated components, benchmarking parsing architectures and semantic matching models in isolation. Phase 2 assembles these into complete end-to-end paradigms, evaluated via an LLM-as-a-Judge methodology. Finally, Phase 3 culminates in our proposed solution, the Hybrid Neuro-Symbolic Framework: a fine-tuned Bi-LSTM parser combined with hybrid Elasticsearch retrieval (dense k-NN with deterministic Boolean geographic guardrails) and a Knowledge-Distilled XGBoost ranker.

The final framework achieved retrieval superiority across all European data, including a 98% win rates in Spain and 91% in France, empirically proving that deep semantic reasoning can be compressed into a high-speed statistical structure without accuracy loss (XGBoost MAE of 3.72% relative to the Cross-Encoder teacher). However, the evaluation also uncovered two architectural findings: Neuro-Symbolic systems exhibit parser dependency, evidenced by the performance collapse for US data; and embedding hybrid Boolean-vector queries into Elasticsearch introduces a 331ms retrieval bottleneck, demonstrating that the computational cost of neural precision ends up migrating to the database persistence layer. Therefore, this work contributes with empirical evidence that the Human-Database Gap can only be closed through the co-design of semantic flexibility, spatial constraints, and database architecture.

Keywords: Geographic Entity Resolution • Address Normalization • Semantic Matching • Approximate Nearest Neighbor Search • Knowledge Distillation • Neuro-Symbolic AI

UN Sustainable Development Goals

The United Nations Sustainable Development Goals (SDGs) provide a global framework for achieving a better and more sustainable future for all. It includes 17 goals¹ to address the world's most pressing challenges, including poverty, inequality, climate change, environmental degradation, peace, and justice.

The specific Sustainable Development Goals mentioned have the following names:

SDG 9 Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation.

SDG 11 Make cities and human settlements inclusive, safe, resilient and sustainable.

SGD	Target	Contribution	Performance Indicators and Metrics
9	9.1	Transitioning from brittle, rule-based legacy systems to scalable, AI-driven semantic architectures enhances the resilience and reliability of enterprise geographic data infrastructure.	Improvements in macro F1-Scores across unstructured and underrepresented spatial classes.
	9.5	Developing a knowledge distillation pipeline that compresses heavy cross-encoders into high-speed statistical models promotes AI innovation while significantly reducing the computational cost of enterprise searches.	Mean Squared Error (MSE) minimization between teacher/student models and reduction in median query latency (ms).
11	11.2	Highly accurate multinational address matching resolves severe structural noise, streamlining urban logistics, reducing failed deliveries, and minimizing the environmental footprint of misrouted transportation.	Decisive retrieval win-rates (%) and reductions in mutual failure rates across highly chaotic real-world queries.

¹<https://sdgs.un.org/goals>

Declaration of honour

I declare that this dissertation is my original work and has not previously been submitted or assessed in any other course or curricular unit, whether at this or any other institution.

References to other authors (statements, ideas, thoughts), as well as the use of published works fully complies with the rules of attribution and are appropriately cited in the text and bibliography, in accordance with reference standards. I am aware that plagiarism and self-plagiarism constitute academic misconduct.

I further declare that I have not used Artificial Intelligence tools to produce any part(s) of this dissertation or, when such tools were used, their use and results are clearly indicated and were carefully reviewed for the detection of errors, inaccuracies, unfounded attributions, or other forms of content and argument bias, as well as for the determination of authorship.

I further declare that I have included studies carried out with multiple co-authors, having specified, with transparency and rigour, the contribution of each co-author, provided information on the possible existence of any other dissertation in which the same article has been included, reproduced the work in full with the permission of the journal in which it was published, and obtained the explicit written consent of all co-authors for the inclusion of the article.

João Macedo Lima



Acknowledgements

This dissertation would not have been possible without the support, guidance, and encouragement of a number of outstanding individuals, to whom I owe a sincere debt of gratitude.

First and foremost, I would like to express my deepest appreciation to my supervisor, Prof. Rosaldo Rossetti, for his academic-oriented guidance, academic discipline, and constant availability throughout this research. His knowledge and constructive feedback were fundamental in developing both the direction and the quality of this work. Besides that, I am equally grateful to ‘JTA The Data Scientists’ for providing the real-world environment, data, and resources to make this research possible. In particular, I wish to extend my sincere thanks to Gonalo Almeida, whose mentorship and consistent guidance throughout this journey proved invaluable.

On a personal level, I would also like to express my profound gratitude to my family. To my father, whose enthusiasm for technology and encouragement to pursue a vocation in this field planted the seed that ultimately led to this dissertation (this work is, in no small part, a reflection of the path you inspired me to take). To my mother and brother, thank you for your unconditional support, patience, and encouragement during the more demanding moments of this journey. Your belief in me was a reliable source of strength.

Finally, to my girlfriend Lara, and to my closest friends - thank you. The demands of a master’s dissertation are rarely confined to working hours, and your encouragement and ability to keep me focused made all the difference.

Joo Macedo Lima

“We shouldn’t be looking for heroes, we should be looking for good ideas.”

Noam Chomsky

Contents

1	Introduction	1
1.1	Context and Problem Statement	1
1.2	Motivation and Contributions	2
1.3	Research Questions	3
1.4	Document Structure	4
2	Literature Review	5
2.1	Background	5
2.1.1	Classical String Metrics and Record Linkage	5
2.1.2	Probabilistic Sequence Modeling for Parsing	6
2.1.3	The Deep Learning Shift: From Recurrence to Attention	7
2.1.4	High-Dimensional Vector Indexing	8
2.1.5	Generative AI and Retrieval-Augmented Generation	9
2.2	Related Work	10
2.2.1	Review Protocol	10
2.2.2	Neural Address Parsing and Geocoding	11
2.2.3	Deep Entity Resolution and Matching Architectures	13
2.2.4	Vector Blocking and Efficient Indexing	14
2.2.5	Gap Analysis	15
2.3	Summary	16
3	Methodological Approach	18
3.1	Research Design	18
3.2	Data Engineering and Ground Truth Construction	19
3.3	Phase 1: Preliminary Component Benchmarking	25
3.4	Phase 2: End-to-End Paradigms	28
3.5	Phase 3: The Hybrid Neuro-Symbolic Framework	35
4	Results and Discussion	40
4.1	Phase 1: Isolated Component Benchmarks	40
4.1.1	Parsing Efficacy	40
4.1.2	Semantic Matching and Feature Engineering	45
4.2	Phase 2: End-to-End Paradigm Showdown	50
4.2.1	Deterministic Parsing at Scale	50
4.2.2	Quantitative End-to-End Showdown (Paradigm 1 vs. Paradigm 2)	53
4.2.3	Domain Adaptation and Cross-Cultural Robustness	55
4.2.4	Qualitative Audit of Retrieval Outputs	60
4.2.5	The Operational Bottleneck and Hardware Profiling	62

4.3	Phase 3: The Neuro-Symbolic Framework	64
4.3.1	Schema Adaptation and the Hybrid Retrieval Trade-off	64
4.3.2	Knowledge Distillation and Feature Intelligence	67
4.3.3	Operational Viability (Speed vs. Accuracy)	69
5	Conclusion	73
5.1	Final Remarks	73
5.2	Limitations and Future Work	74
	References	76
A	Exploratory Data Engineering and Baseline Infrastructure	82
A.1	Exploratory Data Analysis	82
A.2	PostGIS Spatial Indexing and architecture Benchmarks	82
A.3	Baseline Performance and Accuracy Results	82
B	Feature Engineering Inventory	86

List of Figures

2.1	PRISMA flow diagram illustrating the literature search and selection process. . .	12
3.1	High-level block diagram of the iterative research design, progressing from data engineering to the final Hybrid Neuro-Symbolic Framework.	19
3.2	The three-stage data transformation pipeline, showing how unstructured text's parsed into structured JSON and subsequently aligned into strict BIO token sequences using the Longest-Span-First algorithm.	21
3.3	The multi-stage routing architecture for populating the ground truth database. The diagram illustrates the standard deterministic lexical routing (solid lines) alongside the probabilistic fuzzy and spatial fallback mechanisms (dashed lines) for orphaned records.	23
3.4	The progressive architectural evaluation framework for the isolated parsing component, advancing from context-blind lexical classifiers to deep attention-based sequence models.	26
3.5	Architecture of the Siamese Neural Network for semantic matching. The dual-encoder setup handles address pairs simultaneously, applying mean pooling and a fully connected classification head to predict semantic equivalence.	28
3.6	The end-to-end architecture of Paradigm 1. The pipeline utilizes a CRF for morphological parsing, BM25 Boolean queries for sparse retrieval, and an Isotonic-calibrated XGBoost classifier for the final threshold-based candidate matching. . .	30
3.7	The end-to-end architecture of Paradigm 2. The pipeline transitions entirely to deep learning, utilizing an XLM-RoBERTa token classifier for parsing, a Bi-Encoder for dense vector k-NN retrieval, and a Cross-Encoder for final semantic matching. . .	32
3.8	The end-to-end architecture of Paradigm 2.1. This fully domain-adapted pipeline utilizes a lightweight Bi-LSTM network for parsing, alongside a custom fine-tuned Bi-Encoder and Cross-Encoder to optimize both semantic precision and computational latency.	33
3.9	The ontological schema adaptation process, demonstrating how the native 10-tag Deepparse sequence output was systematically mapped and fine-tuned into the strict 7-tag enterprise geographic schema.	36
3.10	The Neuro-Symbolic retrieval architecture, demonstrating the simultaneous execution of dense vector k-NN searches and deterministic Boolean geographic filters within the Elasticsearch index.	38
3.11	The Knowledge Distillation architecture used in the Hybrid Matcher. The heavy Cross-Encoder (Teacher) generates target probabilities offline, which are then used to train the highly efficient XGBoost Regressor (Student) using a Neuro-Symbolic feature space.	39

4.1	Per-component Macro F1-Score heatmap across the evaluated parsing architectures, showing the performance degradation on rare less common classes (e.g., Floor Info, Building Name).	42
4.2	Performance ranking of traditional string comparison algorithms on the binary matching dataset. The chart illustrates the "glass ceiling" of isolated mathematical metrics when confronted with structural geographic noise.	45
4.3	Top 10 feature importances driving the XGBoost semantic matcher. The classification is heavily dependent on interaction features and exact component-wise matches, showing a strong dependency on upstream parsing accuracy.	47
4.4	Receiver Operating Characteristic (ROC) curves comparing the 49-feature XGBoost matcher against the fine-tuned Siamese Network. The XGBoost model displays near-perfect discriminative capacity ($AUC = 0.999$) when supplied with cleanly parsed feature vectors.	48
4.5	Prediction confidence distribution of the Siamese Network. The clear bifurcation between True Matches (peaking near 1.0) and Hard Negatives (peaking near 0.0) demonstrates autonomous semantic separation without reliance on pre-parsed geographic components.	49
4.6	Global LLM-as-a-Judge win rates comparing Paradigm 1 (Lexical/Statistical) and Paradigm 2 (Dense Semantic) across 50,000 operational queries. The high "Both Failed" rate establishes the baseline noise floor of unstandardized geographic data.	54
4.7	End-to-end global retrieval win rates across successive paradigm evaluations. The transition to domain-adapted models (P2.1 and P2.0) causes a regression, nearly doubling the "Both Failed" rate.	56
4.8	Decisive cross-cultural win rates of the newer paradigm in head-to-head matchups. The 50% parity line indicates where the newer architecture mathematically outperforms the older baseline.	56
4.9	End-to-end latency breakdown by paradigm (median execution time per query). The chart illustrates the time distribution across the Parsing, Retrieval, and Matching modules.	63
4.10	Feature Importance (Gain) for the distilled Semantic XGBoost model. The overwhelming reliance on <code>street_levenshtein</code> demonstrates that the model utilizes deterministic character-level distance to approximate the token-to-token cross-attention of the neural Teacher model.	68
4.11	Final LLM-as-a-Judge retrieval win rates (Global and Cross-Cultural) comparing Paradigm 2 against the Neuro-Symbolic Paradigm 3. The 60.3% mutual failure rate visually captures the strict enforcement of the new geographic guardrails.	70
4.12	Final end-to-end latency profiling comparing Paradigm 2 against Paradigm 3. While the Knowledge Distillation successfully eliminated the heavy matching phase, the complex boolean/vector queries caused a severe latency spike in the Elasticsearch retrieval layer.	72
A.1	Summary of data quality and completeness metrics for the baseline dataset.	82
A.2	Comprehensive analysis of the deterministic address normalization and validation phase.	83
A.3	Database query performance benchmarks comparing execution times across PostgreSQL (GiST, SP-GiST, B-Tree) and Elasticsearch indices.	84
A.4	Geographic distribution and density of raw OpenStreetMap address nodes across continental Portugal.	85

List of Tables

2.1	Systematic Comparison of Limitations in State-of-the-Art Approaches	15
3.1	Distribution of multinational records across the enterprise master database and the curated evaluation dataset.	24
3.2	Summary of the evaluation metrics established for the isolated component benchmarking, categorized by their respective tasks and analytical objectives.	29
4.1	Comprehensive parsing performance (F1-Score) across all isolated architectures. The table details both the overall sequence metrics and the specific component-level extraction capabilities.	41
4.2	Qualitative comparison of parsing architectures on highly unstructured Portuguese addresses, illustrating the limitations of statistical models on rare components and abbreviations.	44
4.3	Performance comparison of Supervised Machine Learning classifiers on the binary semantic matching dataset, utilizing the 49-dimensional manually engineered feature space.	46
4.4	Performance comparison of Deep Learning semantic matching architectures. The table contrasts zero-shot contextual embeddings against fine-tuned contrastive learning.	47
4.5	End-to-end deterministic parsing performance comparison between Paradigm 1 (CRF) and Paradigm 2 (RoBERTa) across the 50,000 multinational evaluation queries. Metrics display token-level F1-Scores alongside strict Exact Match (EM) rates.	50
4.6	End-to-end deterministic parsing performance comparison between Paradigm 2.0 (RoBERTa) and Paradigm 2.1 (Bi-LSTM) across evaluation queries. Metrics display token-level F1-Scores alongside strict Exact Match (EM) rates.	57
4.7	Qualitative examples of the LLM-as-a-Judge evaluations, illustrating how dense vector retrieval (Paradigms 2, 2.0 and 2.1) successfully handles severe structural noise that paralyzed the sparse lexical index (Paradigm 1), which retrieves identically-spelled locations on the wrong continent.	61
4.8	End-to-end deterministic parsing performance for Paradigm 3 (Fine-Tuned Bi-LSTM) across evaluation queries. Metrics display token-level F1-Scores alongside strict Exact Match (EM) rates.	64
A.1	Key Performance Indicators (KPIs) for the baseline deterministic retrieval system.	83
B.1	Exhaustive catalog of the 49 engineered features.	86

List of Acronyms

AI	Artificial Intelligence
ANN	Approximate Nearest Neighbor
BCE	Binary Cross-Entropy
BERT	Bidirectional Encoder Representations from Transformers
Bi-LSTM	Bidirectional Long Short-Term Memory
BIO	Begin-Inside-Outside
BLISS	Billion scale Index using Iterative Re-partitioning
B-VSS	Blocking with Vector Similarity Search
CNN	Convolutional Neural Network
CRF	Conditional Random Fields
DIAL	Deep Indexed Active Learning
DPR	Dense Passage Retriever
EDA	Exploratory Data Analysis
EM	Expectation-Maximization
ER	Entity Resolution
ETL	Extract, Transform, Load
FLAP	Framework for Linking Free-text Addresses to the Ordnance Survey Unique Property Reference Number database
GIR	Geographic Information Retrieval
GiST	Generalized Search Tree
GNN	Graph Neural Network
GraSP	Graph-based Nearest Neighbor Search with Subgraph Sampling and Pruning
GSP	Geo-Semantic-Parsing
HEM	Heterogeneous Entity Matching
HMM	Hidden Markov Models
HNSW	Hierarchical Navigable Small World
IR	Information Retrieval
LLM	Large Language Model
LSH	Locality Sensitive Hashing
LTR	Learning-To-Rank

MAE	Mean Absolute Error
MEMM	Maximum Entropy Markov Models
MSE	Mean Squared Error
NER	Named Entity Recognition
NLP	Natural Language Processing
OCR	Optical Character Recognition
OOV	Out-of-Vocabulary
OSM	OpenStreetMap
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RAG	Retrieval-Augmented Generation
RNN	Recurrent Neural Networks
ROC-AUC	Receiver Operating Characteristic Area Under the Curve
SBERT	Sentence-BERT
Seq2Seq	Sequence-to-Sequence
SLR	Systematic Literature Review
SQL	Structured Query Language
SVM	Support Vector Machines
TF-IDF	Term Frequency-Inverse Document Frequency
TGFR	Transformer-Gather, Fuzzy-Reconsider

Chapter 1

Introduction

1.1 Context and Problem Statement

In an increasingly digitized world, geographic information has act as a cornerstone of the global economy, powering sectors ranging from logistics to financial services. However, the utility of these data is fundamentally constrained by their quality and, in the centre of this ecosystem, lies address data, the primary textual representation of physical locations. Therefore, as organizations accumulate vast amounts of geographic information, they are faced with a serious challenge: the gap separating human language and database rigidity. Addresses are often entered manually, containing variations, local abbreviations, and structural inconsistencies; thereby, **Traditional deterministic systems** (based on rigid rules and regular expressions) fail to scale against this variation. Besides that, this scenario presents a classic **Artificial Intelligence problem**: the need for a system capable of understanding semantic intent and context within noisy unstructured text, connecting the gap between Natural Language Processing (Natural Language Processing (**NLP**)) and Geographic Information Retrieval (Geographic Information Retrieval (**GIR**)).

Although the significance of precise geographic data is acknowledged, existing solutions for address normalization and search typically fall short when applied to enterprise-scale data streams. The central problem handled in this thesis is the incompatibility between the heterogeneous, unstructured quality of real-world address input and the rigid, structured requirements of enterprise databases. This misfit manifests itself primarily as a dual challenge of linguistic ambiguity and computational scalability. From a linguistic perspective, traditional methods, such as simple string comparisons (for example, Levenshtein distance) or standard Structured Query Language (**SQL**) queries, are insufficient to handle the intrinsic noise of address data. Addresses lack standardization and are prone to considerable variation; therefore, a rigid rule-based system can easily struggle to equate abbreviated forms like "R. Sta. catarina" with their canonical form "Rua de Santa Catarina". Furthermore, standard algorithms lack the probabilistic awareness required to resolve semantic ambiguity, such as distinguishing whether the token "Washington" refers to a state, a city, or a street based on context.

In the specific context of the partner organization, these language limitations create a critical

bottleneck for Entity Resolution against a Master Database. The organization maintains a high-volume "Master Record" database of validated addresses, but when new, noisy address streams arrive from clients or partners, the current system fails to reliably link this input to existing records. This failure leads to high false negative rates, where valid addresses are rejected or duplicated because the system cannot infer that a noisy input corresponds to an existing record. Moreover, attempting to resolve these ambiguities by comparing every incoming string against millions of records using complex fuzzy logic becomes computationally excessive, creating severe scalability difficulties. Consequently, this thesis addresses the need for an AI-driven framework able of closing this gap, allowing probabilistic address matching that is both semantically robust and computationally effective for large-scale databases.

1.2 Motivation and Contributions

The motivation for this research relies on replacing brittle rule-based legacy systems with strong pipelines. For the partner company and the wider industry (logistics, fintech, and so many others), the inability to faithfully match an input address to a database record results in operational flaws such as failed deliveries, duplicate customer records, and failed fraud detection checks. Given this, we can say that there is a clear need for a solution that integrates **modern Artificial Intelligence (AI) techniques** like probabilistic parsing (e.g. Conditional Random Fields or deep learning models like transformers) and semantic search (e.g. Vector Embeddings) with efficient data structures. Therefore, this research intends to prove that an AI-centric approach is able to significantly beat traditional string matching in both accuracy (Recall) and operational throughput. The main objective of this dissertation is to design, implement, and evaluate an AI-driven solution for the normalization and efficient matching of geographic address data against large-scale databases. This main goal requires a transition from rigid rule driven systems to a probabilistic system capable of understanding the semantic structure of unstandardised user input. Therefore, to achieve this, we begin through dissecting the State of the Art and analysing the evolution from traditional Information Retrieval to modern Neural Search. This conceptual foundation informs the implementation of probabilistic normalization models, where different architectures are deployed to structure raw address strings. Additionally, beyond parsing, the research tackles the retrieval bottleneck by benchmarking efficient indexing strategies, specifically contrasting the performance of traditional Inverted Indexes against high-dimensional Vector Indexes to enable semantic similarity search at scale. Consequently, these components will eventually culminate in the construction of a prototype matching engine, a pipeline capable of ingesting raw real addresses data and query it in real time. Finally, the final phase will focus on the practical validation of this architecture within an enterprise context. This involves benchmarking the different architectures (whether they use AI or traditional methods) using rigorous metrics (such as F1-Score or Query latency), thereby showing its practical value in reducing entity resolution failures.

Although this dissertation utilizes geographic data as a primary case study, the core research focuses on the wider challenge of Entity Resolution (Entity Resolution (ER)) over unstructured

text streams. It is expected that the architecture proposed in this work, combining Probabilistic Parsing for structure extraction with Vector Embeddings for semantic retrieval, will offer a contribution to the field of Neural Entity Resolution that goes beyond the immediate scope of the partner company. Scientifically, this work advances the field of Neural Entity Resolution by trying to address semantic hallucinations inherent to pure deep learning retrieval. It empirically proves that deep semantic reasoning can be compressed into high-speed statistical structures (for example, via Knowledge Distillation). In addition to that, this research delivers the **Hybrid Neuro-Symbolic Framework**, which is a robust and scalable framework that closes the gap between the probabilistic adaptability of modern AI and the deterministic inflexibility required by enterprise relational databases. Furthermore, while the immediate beneficiary of this framework is the partner organization, this architecture is designed to be transferable to concrete, real-world scenarios across various industries. Specific applicational use cases include :

- Automating delivery routing in logistics by resolving highly corrupted, user-entered addresses;
- Consolidating fragmented product catalogues in e-commerce by matching multi-vendor inventories containing wildly varying descriptions;
- Deduplicating national citizen registries and demographic census data within government digital services by cross-referencing legacy, unstandardised public records against centralized administrative hierarchies;
- Enhancing urban infrastructure management and emergency dispatch systems in Smart Cities by accurately mapping confusing incident reports to precise municipal spatial coordinates;
- Standardizing municipal land registries and property tax assessments by normalizing unstructured historical property deeds into a unified national cadastre database.

1.3 Research Questions

The development of this system is driven by two fundamental inquiries that tackle the limitations identified in the problem statement. The first interrogates the qualitative shift from symbolic to semantic processing, while the second examines the quantitative consequences of this shift on system performance. Specifically, this dissertation seeks to answer:

- **RQ1 (Effectiveness):** To what extent does modeling addresses as **probabilistic semantic entities** outshine traditional **deterministic string matching** in resolving ambiguity as well as minimizing false negatives?
- **RQ2 (Efficiency):** How does the transition from **lexical indexing** (e.g. Inverted Indexes) to **high-dimensional vector indexing** impact the trade-off between query latency and retrieval precision in enterprise-scale databases?

1.4 Document Structure

Following this introduction, the dissertation establishes its theoretical grounding in Chapter 2 (Background and Literature Review), which reviews the evolution of NLP for geospatial data, Vector Search, and Probabilistic Data Cleaning, and identifies the gaps this work aims to fill. Additionally, the architectural vision is formalized in Chapter 3 (Methodology Approach), which details the system design, the fundamental reasons/justifications for specific AI models and the characteristics of the data sources. Consequently, the validity of the proposed solution is assessed in Chapter 4 (Results and Discussion), where the experimental results are presented to benchmark the system against traditional baselines. Finally, Chapter 5 (Conclusion) synthesizes the contributions of this research and details potential directions for future work.

Chapter 2

Literature Review

2.1 Background

This section establishes the theoretical foundations of the dissertation, it traces the trajectory of entity resolution from deterministic string metrics to probabilistic record linkage, and finally to modern neural architectures. Furthermore, it details the mathematical principles of high-dimensional vector retrieval and formalizes the Retrieval-Augmented Generation (**RAG**) architecture.

2.1.1 Classical String Metrics and Record Linkage

Before the arrival of deep learning, the normalization of geographic data relied on defining metric spaces where the similarity between two strings could be quantified mathematically.

The fundamental measure in this domain is the **Levenshtein Distance**, introduced by Vladimir Levenshtein in 1965 [31]. It is defined as the minimum number of single-character operations (insertions, deletions, or substitutions) required to transform string a into string b . Mathematically, the distance is calculated using **Dynamic Programming** where a matrix D of size $(|a| + 1) \times (|b| + 1)$ is constructed and $D_{i,j}$ represents the distance between the first i characters of a and the first j characters of b :

$$D_{i,j} = \min \begin{cases} D_{i-1,j} + 1 & \text{(deletion)} \\ D_{i,j-1} + 1 & \text{(insertion)} \\ D_{i-1,j-1} + 1_{(a_i \neq b_j)} & \text{(substitution)} \end{cases} \quad (2.1)$$

The computational complexity of this algorithm is $O(|a| \cdot |b|)$, which becomes extremely expensive for large-scale comparisons. Furthermore, Levenshtein fails to efficiently model **transpositions** (e.g. "Main Srt" vs. "Main Str"). Consequently, the **Damerau-Levenshtein** distance [9] addresses this by extending the operator set to include adjacent character swaps, making it a more robust metric for real-world data entry errors.

On the other hand, for short strings like proper names, edit distance can be too coarse. **Matthew Jaro** [23] proposed a metric based on the number of matching characters (m) and transpositions (t) within a sliding window. Besides that, **William Winkler** [57] extended this by introducing

a *prefix scale* p , based on the observation that errors are unlikely to occur at the beginning of a string. In addition to that, the Jaro-Winkler similarity is defined as:

$$Sim_{JW} = Sim_J + p \cdot l \cdot (1 - Sim_J) \quad (2.2)$$

where l is the length of the common prefix (up to 4 characters). This metric is particularly effective for toponym resolution, where "Washington" and "Washingtn" share a strong prefix.

Besides that, string metrics quantify similarity, but determining whether two records refer to the same entity requires a decision framework. **Fellegi and Sunter** [13] formalized this in their landmark "Theory for Record Linkage". The model classifies pairs into Matches (M) and Non-Matches (U) using a likelihood ratio R :

$$R(\gamma) = \frac{P(\gamma|M)}{P(\gamma|U)} = \frac{m(\gamma)}{u(\gamma)} \quad (2.3)$$

where $m(\gamma)$ is the probability that a field agrees given the pair is a match (True Positive) and $u(\gamma)$ is the probability of random agreement (False Positive).

However, when M and U are unknown, the **Expectation-Maximization (EM)** algorithm [10] is typically used to estimate the m and u probabilities. This iterative process allows the model to "learn" the distinguishing power of different fields (e.g. matching on "Zip Code" is less significant than matching on "Street Name") without labeled training data.

2.1.2 Probabilistic Sequence Modeling for Parsing

Before moving towards the matching part, raw address strings must first be parsed into structured components. This is formally defined as a **Sequence Labeling** problem where, given an input sequence of tokens $X = (x_1, x_2, \dots, x_n)$, the goal is to predict the optimal sequence of semantic tags $Y = (y_1, y_2, \dots, y_n)$.

Initially, early probabilistic approaches relied on Hidden Markov Models (**HMM**). As generative models, HMMs learn the joint probability $P(X, Y)$ but are constrained by strong independence assumptions: the current label depends only on the previous label and the current observation depends only on the current label. This renders HMMs incapable of utilizing global features, for instance, realizing that if a string ends with "USA", the first token is likely a street number rather than a country code. So, in order to overcome these limitations, **Lafferty et al.** [29] introduced Conditional Random Fields (**CRF**), which are discriminative models that directly model the conditional probability $P(Y|X)$.

Unlike HMMs, CRFs normalize the probability globally over the entire sequence, effectively solving the "Label Bias Problem" found in earlier discriminative models like Maximum Entropy Markov Models (**MEMM**). Besides that, the probability of a label sequence Y given observation X is defined as:

$$P(Y|X) = \frac{1}{Z(X)} \exp \left(\sum_{j=1}^n \sum_{k=1}^K \lambda_k f_k(y_{j-1}, y_j, X, j) \right) \quad (2.4)$$

In this formulation, f_k represents feature functions that capture specific patterns (e.g. returning 1 if x_j is a number and y_j is a "Zip Code"), which are weighted by learned parameters λ_k . Also, the expression is normalized by the Partition Function $Z(X)$, a constant that sums over all possible label sequences to ensure that all probabilities sum to 1. This global perspective establishes CRFs as the "gold standard" for non-neural parsing, enabling the incorporation of rich, overlapping features across the entire input string.

2.1.3 The Deep Learning Shift: From Recurrence to Attention

Although CRFs improved accuracy over HMMs, they ultimately rely on manual feature engineering. The advent of Deep Learning fundamentally shifted this paradigm, replacing hand-crafted features with learned dense vector representations.

Initially, **Sutskever et al.** [51] revolutionized text processing with the **Sequence-to-Sequence (Seq2Seq)** architecture, which utilizes an Encoder-Decoder structure based on Recurrent Neural Networks (**RNN**). Prior to this, **Hochreiter and Schmidhuber** [20], had also introduced the **Long Short-Term Memory** as a direct solution to the vanishing gradient problem inherent by plain RNNs, by equipping each unit with a gated memory mechanism. In this framework (the Seq2Seq), the Encoder processes the input token-by-token, updates its hidden state at each step and the final hidden state acts as a fixed-length vector intended to contain a semantic summary of the entire sequence, which is then used by the Decoder to generate output tags.

However, this architecture suffers from a critical flaw known as the "**Information Bottleneck**". This means that compressing a complex address (such as a long, multi-line international string) into a single fixed vector inevitably results in a loss of detail. Additionally, the sequential nature of RNNs often causes the model to "forget" the beginning of the string by the time it processes the end, limiting performance on long sequences.

Consequently, new techniques started emerging to try to tackle these issues, such as **transformers and self-attention mechanisms**. **Vaswani et al.** [52] resolved the bottleneck issue by discarding recurrence entirely in favor of the **Transformer** architecture. The core innovation is the **Self-Attention Mechanism**, which mimics a database retrieval process. For every token, the model generates three vectors: **Query (Q)**, a **Key (K)**, and a **Value (V)**. The attention score determines how much focus to place on other parts of the sentence:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (2.5)$$

This mechanism allows the model to dynamically look at relevant parts of the input regardless of distance. For example, when encoding the word "Paris," the model can "attend" to the word "France" appearing later in the string to disambiguate it from "Paris, Texas". This global, non-sequential processing enables the generation of **Contextual Embeddings** (e.g., Bidirectional Encoder Representations from Transformers (**BERT**)), where the vector representation of a word varies based on its surrounding semantic context.

On the other hand, we started to see as well **Neural Entity Resolution Architectures**. So, in the domain of matching these contextual embeddings, two distinct architectures have emerged, presenting a trade-off between speed and accuracy [47]. Therefore, we have the **Cross-Encoder** where its approach maximizes accuracy by allowing full interaction between inputs. The two strings being compared (e.g, Address A and Address B) are concatenated into a single input sequence (for example, [CLS] A [SEP] B) and passed through the Transformer. This allows the self-attention mechanism to compare every token in A with every token in B directly. Although this yields the highest precision, it is computationally prohibitive ($O(N^2)$) for large datasets because every possible pair must be processed individually. So, in order to address this scalability constraint, **Bi-Encoders** (Siamese Networks) decouple the processing. In this architecture, each string is processed by a separate (but weight-shared) Transformer to produce an independent fixed embedding (u and v). Similarity is then calculated using a computationally efficient distance metric, such as Cosine Similarity:

$$\text{Score}(A, B) = \cos(u, v) = \frac{u \cdot v}{\|u\| \|v\|} \quad (2.6)$$

Although this lack of token-level interaction results in slightly lower accuracy compared to Cross-Encoders, it allows embeddings to be pre-computed and indexed. This capability ends up being crucial for enterprise applications, as it enables the use of Approximate Nearest Neighbor (**ANN**) search algorithms like Hierarchical Navigable Small World (**HNSW**) to retrieve candidates from millions of records in milliseconds.

2.1.4 High-Dimensional Vector Indexing

The shift from symbolic matching to Bi-Encoders fundamentally alters the retrieval landscape. We transition from comparing discrete strings to searching for neighbours in a continuous, high-dimensional vector space ($\mathbb{R}^d$, where $d \geq 128$). This shift necessitates indexing structures capable of navigating complex geometric relationships rather than simple lexical prefixes.

The primary challenge in this topic is the "**Curse of dimensionality**", a term coined by **Richard Bellman** [3] to describe the exponential increase in volume as dimensions are added. In the context of nearest-neighbour search, this manifests as the "**Concentration of Measure**" phenomenon: as dimensionality increases, the distance between the query point and its nearest neighbour approaches the distance to its farthest neighbour. This renders traditional spatial partitioning structures like **k-d Trees** [4] ineffective. Besides that, while trees offer logarithmic search in low dimensions ($d < 20$), their performance degrades to a brute-force linear scan ($O(N)$) in high-dimensional semantic spaces, as the algorithm must visit almost every branch to guarantee exactness.

On the other hand, to achieve scalability across millions of records, it becomes necessary to accept an Approximate Nearest Neighbor (ANN) result, trading a negligible amount of recall for logarithmic query speed. Historically, two major algorithmic families attempted to solve this before the current graph-based era. Early attempts like Ball-Trees tried to adapt partitioning to

spherical clusters but still struggled with the "curse" in high dimensions. Subsequently, Locality Sensitive Hashing (**LSH**), introduced by **Indyk and Motwani** in 1998 [21], proposed projecting high-dimensional vectors into low-dimensional "buckets" using random hash functions. Although it is theoretically elegant and fast ($O(1)$), LSH requires significant memory overhead (maintaining multiple hash tables) to achieve acceptable recall on dense embeddings.

Finally, considering all of these different challenges, we have now the most modern techniques. The current state-of-the-art relies on Proximity Graphs, an approach grounded in the "Small World" theory popularized by Watts and Strogatz in 1998 [55], which posits that in certain networks, any node can be reached from any other in a small number of hops. Besides that, **Malkov and Yashunin** [37] operationalized this for vector search with the Hierarchical Navigable Small World (HNSW) algorithm.

Crucially, HNSW solves the local minima problem inherent in greedy graph traversal by adapting the concept of Skip Lists [43]. This algorithm constructs a multi-layer structure where the top layers are sparse, containing only long-range links that act as "express ways" for zooming across the vector space. Additionally, the search process employs a coarse-to-fine traversal: it begins at the top layer, taking large steps toward the query and, upon reaching a local minimum, drops to lower and denser layers to refine the search radius. This hierarchical architecture allows HNSW to achieve $O(\log N)$ query complexity with high recall, establishing it as the standard for enterprise vector search.

2.1.5 Generative AI and Retrieval-Augmented Generation

Finally, the most recent paradigm shift in Natural Language Processing involves the transition from purely discriminative models to generative architectures. Large Language Models, such as GPT-4, operate as probabilistic engines trained to predict the next token in a sequence $P(w_t | w_{1:t-1})$. While these models exhibit emergent reasoning capabilities, their deployment in enterprise contexts is hindered by two fundamental limitations: **knowledge cut-offs**, where the model's static weights remain unaware of private or real-time data and **hallucination**, where the model fills information gaps with statistically plausible but factually incorrect fabrications.

Therefore, to resolve these reliability issues, **Lewis et al.**[32] introduced Retrieval-Augmented Generation (RAG) in 2020. This architecture fundamentally hybridizes two distinct forms of memory: **parametric memory**, stored in the pre-trained weights of the Large Language Model (**LLM**) (representing implicit general knowledge), and **non-parametric memory**, represented by an external and updatable vector index (containing explicit specific knowledge). The workflow typically employs a **Dense Passage Retriever (DPR)** [27] to fetch relevant context vectors z from an index (such as HNSW) based on an input query x . The Generator then conditions its output y on both the original query and the retrieved context:

$$P(y|x) \approx \sum_{z \in \text{top-k}(x)} P_{\eta}(y|x, z) P_{\theta}(z|x) \quad (2.7)$$

By explicitly conditioning generation on retrieved evidence, this framework allows the system to "ground" its reasoning in verifiable database records. This effectively combines the scalability of vector search with the semantic understanding of LLMs, ensuring that answers remain factually consistent with the underlying enterprise data.

2.2 Related Work

The problem of resolving unstructured geographic references to structured database records sits at the intersection of Natural Language Processing (NLP), Information Retrieval (IR), and Database Management. This section reviews the state-of-the-art across these three dimensions: neural address parsing and geocoding, deep entity resolution architectures, and scalable vector indexing techniques

2.2.1 Review Protocol

First of all, we needed to ensure a rigorous evaluation of the state-of-the-art. Therefore, this study followed a Systematic Literature Review (SLR) methodology aligned with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines. This review process was designed to answer the research questions regarding the trade-offs between deep learning architectures and vector scalability in geospatial entity resolution.

Firstly, the comprehensive search was conducted across four primary scientific databases: **Scopus**, **Web of Science**, **IEEE Xplore**, and **ACM Digital Library**. So, in order to capture the interdisciplinary nature of the research problem, bridging Natural Language Processing, Geocoding, and Database Systems, we utilized a Boolean search string designed to retrieve studies at the intersection of these fields:

("Entity Resolution" OR "Record Linkage" OR "Address Matching") AND ("Deep Learning" OR "Transformer" OR "Large Language Model") AND ("Vector Database" OR "HNSW" OR "Approximate Nearest Neighbor") AND ("Geocoding" OR "Toponym Resolution")

On the other hand, we needed to ensure the selected studies were strictly relevant to the integration of Deep Learning and High-Dimensional Vector search. Therefore, distinct eligibility criteria were established (prior to the search) using the Parsifal management tool. So, the **Inclusion Criteria** selected studies that met the following requirements:

- **Timeframe:** Peer-reviewed articles and conference proceedings published between 2019 and 2025.
- **Technology:** Studies employing Deep Learning, Transformers, Dense Embeddings, or LLMs.
- **Domain Task:** Research focusing on Entity Resolution, Record Linkage, Address Matching, Geocoding, or Toponym Resolution.

- **Baselines:** Studies using traditional Machine Learning (e.g., Random Forest, Support Vector Machines (**SVM**)) were included only if used as a baseline for comparison against neural methods.

Conversely, the Exclusion Criteria filtered out studies based on the following factors:

- **Publication Type:** Posters, keynotes, editorials, or abstract-only entries were excluded to ensure methodological rigor.
- **Data Pipeline Focus:** Studies focused solely on data acquisition, web crawling, or Extract, Transform, Load (**ETL**) automation rather than the entity matching logic itself.
- **Indexing Mechanics:** Research on Learned Indexes or Recursive Model Indexes designed purely for storage optimization (e.g., B-Tree replacements) rather than semantic vector retrieval.
- **Task Incompleteness:** Studies focused solely on Named Entity Recognition (**NER**) or Information Extraction without a subsequent Resolution/Matching phase.

Finally, the selection process followed the standard PRISMA workflow. Initially, duplicate records were removed, and the remaining titles and abstracts were screened for relevance. Subsequently, the results and the full-text articles, respectively, were assessed against the eligibility criteria defined above. This rigorous filtering process resulted in the selection of 36 primary studies for detailed data extraction and synthesis. The complete flow of information through the different phases of the systematic review is illustrated in Figure 2.1.

2.2.2 Neural Address Parsing and Geocoding

The first challenge in any address matching pipeline is processing and interpreting the raw input string and it is crucial for structuring unstructured text and capturing semantic relationships. So, for this task we can find many different and diverse techniques (fuzzy matching, machine learning or even deep learning). Therefore, we can say that we have the traditional methods that rely on rule-based parsers and more modern techniques that are described in more recent literature which has largely shifted towards deep learning to handle the variability of global addresses. For example, **Beauchemin and Yassine** introduced Deepparse [2], a Seq2Seq architecture using Bidirectional Long Short-Term Memory (**Bi-LSTM**)s and BPEmb embeddings. Unlike C-based libraries, Deepparse allows for fine-tuning on new countries without defining complex regular expressions, achieving over 99% accuracy on unseen multinational data.

Beyond simple parsing, the task of **Geocoding** (Toponym Resolution) requires resolving ambiguous place names (e.g. "Paris, Texas" vs. "Paris, France") and that represents a very important task to better construct semantic relationships and to improve address matching. **Wang and Hu** [53] conducted a systematic evaluation of neural geoparsers (such as DM_NLP and UniMelb) on the EUPEG platform, concluding that while they excel in clean text, they struggle significantly

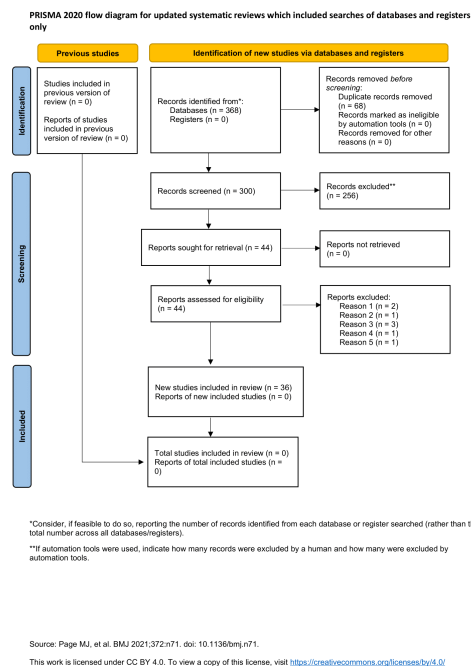


Figure 2.1: PRISMA flow diagram illustrating the literature search and selection process.

with noisy social media data. Consequently, in order to address this ambiguity, **Fize et al.** [14] proposed a model based on toponym pairs, using LSTM units to capture spatial co-occurrences (e.g. using "France" as context to resolve "Southwest"). Similarly, **Yan et al.** [58] also developed LGGeoCoder, which integrates global context embeddings (topic modeling) with Convolutional Neural Network (CNN)-based character extraction to better handle data sparsity in place names.

Furthermore, recent works explored hybrid retrieval methods for this task in order to find more efficient and precise techniques. **Zhang and Bethard** [66] proposed GeoNorm, a "generate-and-rank" framework that first uses Lucene (Information Retrieval) to find candidates and then a Transformer-based reranker to select the correct entity based on metadata like population. Additionally, **Nizzoli et al.** [41] introduced Geo-Semantic-Parsing (GSP), which traverses a knowledge graph horizontally to expand candidate sets, significantly improving recall over standard baselines. Also, in the domain of search ranking, **Zhang et al.** [61] demonstrated that a fast attention-based Learning-To-Rank (LTR) model could process heterogeneous map queries with 3 ms latency, outperforming BERT-based rankers in speed while maintaining high relevance.

Besides all the previous techniques, specific linguistic domains have also been explored. **Molina-Villegas et al.** [38] applied word embeddings (Word2Vec) to Mexican Spanish news, proving that semantic spaces can capture geographic relationships (e.g. "Paris - France + Italy = Rome"), and **Cheng and Chen** [7] focused on road location conversion, using deep learning to interpret qualitative directions (e.g. "north section of Jewett Avenue"). Most recently, **Lin et al.** [35] introduced RACCOON, a Retrieval-Augmented Generation (RAG) approach that utilizes Large Language Models (LLMs) like GPT-4 to generate coordinates from news articles, although they noted issues

with population bias in the underlying LLMs. Finally, **Weckmüller et al.** [56] further extended this concept to "Geospatial Semantic Search" proposing modular workflows that combine text embeddings with spatial aggregation for urban analysis.

2.2.3 Deep Entity Resolution and Matching Architectures

Once a location is parsed, it must be linked to a unique record and this is the domain of Entity Resolution (ER). While traditional baselines like **Framework for Linking Free-text Addresses to the Ordnance Survey Unique Property Reference Number database (FLAP)** [62] and **Adaptive Fuzzy String Matching** [28] rely on Random Forests and iterative human-in-the-loop string metrics (Levenshtein, Jaccard), the field has rapidly moved toward Deep Learning due to multiple limitations presented by the traditional methods for example, semantic generalization or ambiguity handling.

Regarding this field, we found that **Foxcroft et al.** [16] pioneered the use of Name2Vec, applying Doc2Vec to name matching to capture character-level structural similarities. Also, **Zeakis et al.** [60] performed a comprehensive analysis of pre-trained embeddings (BERT, RoBERTa, FastText) for ER, finding that while Sentence-BERT (**SBERT**) variants offer the best blocking recall, static embeddings like GloVe are faster for initialization. Additionally, this was supported by **Duarte and Oliveira** [12], who achieved 99.6% accuracy on Portuguese addresses by training a Transformer bi-encoder from scratch using a multiple negatives ranking loss.

So, we can affirm that the current state-of-the-art focuses on "Deep Entity Resolution" frameworks that handle heterogeneity. **Moslemi et al.** [39] surveyed the challenges of Heterogeneous Entity Matching (**HEM**), highlighting that models trained on clean data often fail to generalize to "dirty" test sets. Therefore, another problem emerged and, to tackle this issue, **Joshi et al.** [24] proposed ReLink, a hybrid system using BERT features and XGBoost for industrial complete-linkage. Additionally, **Sharifi and Ahmadzadeh** [50] introduced Transformer-Gather, Fuzzy-Reconsider (**TGFR**) (Transformer-Gather, Fuzzy-Reconsider), a scalable framework that uses Transformers for retrieval but falls back to fuzzy logic for efficient verification on CPUs.

Besides the focus on Entity Resolution that handles heterogeneity, more advanced architectures attempt to unify the blocking and matching steps. **Dou et al.** [11] proposed MutualER, a joint training framework where the "Blocker" and "Matcher" teach each other via knowledge transfer, significantly outperforming pipeline approaches. Meanwhile, **Ning et al.** [40] introduced SC-Block++, which applies supervised contrastive learning with adaptive flood regularization to prevent overfitting in the blocking phase. Also, **Jain et al.** [22] proposed Deep Indexed Active Learning (**DIAL**) (Deep Indexed Active Learning), utilizing a committee of encoders to improve recall in low-resource settings, and (for tabular data) **Rauf et al.** [46] developed TableDC, a deep clustering method using Mahalanobis distance to handle high-density tabular embeddings.

However, the dominance of neural methods has been challenged by recent studies where its' authors argue that their superiority is often overstated due to flawed evaluation benchmarks. **Leone et al.** [30] critically re-evaluated neural entity alignment methods (such as RDGCN and BootEA), showing they often rely on unrealistic assumptions (like '1-to-1 matching) and that a well-tuned

probabilistic baseline (Paris+) could still outperform complex Graph Neural Network (GNN)s in efficiency and accuracy. Conversely, for generative approaches, **Arora and Dell** [1] released LinkTransformer to democratize LLM-based linkage for social sciences but with heavy hardware requirements. These have driven research into efficiency-focused frameworks and that is why **Li et al.** [33] proposed an uncertainty reduction framework that selectively queries the LLM only for ambiguous pairs, while **Saengsiripaiboon et al.** [49] improved efficiency via batch prompting strategies.

2.2.4 Vector Blocking and Efficient Indexing

Parallel to the advancements made in ER, Geocoding and in address parsing, we have also observed the raise of different new challenges. The application of Deep Learning to ER creates a scalability bottleneck: comparing high-dimensional vectors is computationally expensive ($O(N^2)$). This has led to new research where the main goal is to improve its efficiency, for example, the adoption of Approximate Nearest Neighbor (ANN) search for the blocking phase. In **Ma et al.**'s work [36], the author provided a comprehensive survey of Vector Databases, categorizing technologies into tree-based, hash-based (LSH) and graph-based (HNSW) indexing. In addition, **Karapiperis and Verykios** [26] reviewed the evolution of partitioning techniques, tracing the shift from K-Means clustering to modern indexes like Google's ScaNN. In a separate work, **Karapiperis et al.** [25] proposed LSBlock, a hybrid blocking system combining MinHash (lexical) and SentenceBERT (semantic) with HNSW indexing to balance recall and precision. Later, this approach was rigorously evaluated by **Zhang et al.** [63] in Blocking with Vector Similarity Search (B-VSS) (Blocking with Vector Similarity Search), demonstrating that graph-based indices like HNSW significantly outperform LSH in terms of blocking quality (pair completeness) for deep embeddings. Finally, **Wang and Zhang** [54] further studied the generalization of these pre-trained blocking models, finding that neural IR models like Contriever generalize better to out-of-distribution data.

Besides that, we can also conclude that the optimization of these indices is an active research area. **Gupta et al.** [18] introduced Billion scale Index using Iterative Re-partitioning (BLISS), a "learning-to-index" algorithm that achieves billion-scale scalability by learning optimal partitioning buckets, requiring 100x less RAM than HNSW. On the other hand, **Zhang et al.** [64] proposed Graph-based Nearest Neighbor Search with Subgraph Sampling and Pruning (GraSP), a graph optimization technique that prunes redundant edges based on query distribution (speeding up search by 2x), while **Qin** [44] developed a cost model to predict the efficiency of these graph-based searches based on topological features. Finally, for system integration, **Zhang et al.** [65] presented VBASE, a system unifying vector search with relational databases via "relaxed monotonicity," enabling complex queries like vector joins up to 7,000x faster than brute force.

2.2.5 Gap Analysis

Despite the significant advancements detailed in the previous subsections, a critical analysis of the literature reveals that existing solutions tend to operate in isolated silos. As summarized in Table 2.2, current state-of-the-art frameworks typically optimize for either linguistic precision (Geocoding), matching accuracy (Entity Resolution), or retrieval speed (Vector Indexing), but rarely integrate all three for global-scale data.

Table 2.1: Systematic Comparison of Limitations in State-of-the-Art Approaches

Category	Reference Studies	Primary Technique	Critical Limitation (The Gap)
Deep (Joint)	ER MutualER [11], SC-Block++ [40]	Joint Blocker-Matcher Training	Incompatibility: Relies on dense vectors; cannot integrate with the discrete reasoning of Generative LLMs.
Deep (Hybrid)	ER TGFR [50], Re-Link [24]	Transformer + Fuzzy/Manual	Rigidity: Falls back to manual rules or fuzzy logic for final verification, losing semantic nuance.
Geocoding (RAG)	RACCOON [35], GeoNorm [66]	Retrieval-Augmented Generation	Recall Bottleneck: Relies on legacy keyword search (Lucene/Elastic) for candidates, missing semantic matches.
Geocoding (Global)	Deepparse [2], LGGeoCoder [58]	Seq2Seq / Grid Classification	Precision: Parses text but lacks a resolution mechanism (Deepparse) or loses geometric precision (LGGeoCoder).
Vector Indexing	B-VSS [63], GraSP [64]	HNSW / Graph Optimization	Complexity: High-performance indexing requires complex tuning and lacks “out-of-the-box” semantic awareness.
Vector Blocking	LSBlock [25], BLISS [18]	Hybrid Blocking / Learned Index	Sensitivity: Struggle to balance load (partition size) with recall; “Near-neighbor” property degrades at scale.
Baselines	FLAP [62], Link-Transformer [1]	Random Forest / Toolkit	Robustness: Sensitive to OCR errors (LinkTransformer) or completely lacks semantic understanding (FLAP).

2.2.5.1 Gap 1: The Disconnect Between Neural Parsing and Vector Retrieval

While vector indexing has revolutionized general information retrieval, specialized geocoding architectures have largely failed to adopt these modern standards. For example, State-of-the-art frameworks like GeoNorm [66] and RACCOON [35] utilize advanced Transformer/LLM rerankers but still rely on traditional keyword-based indices (Lucene, ElasticSearch) for candidate generation. Zhang & Bethard [66] explicitly note that if the inverted index misses a candidate due to lexical mismatch (e.g. "St." vs. "Street"), the downstream neural model never sees it. Furthermore, even "global" approaches like LGGeoCoder [58] revert to rigid grid classifications that lose geometric precision, while Deepparse [2] focuses entirely on segmentation without a mechanism

for entity linking. So, we can say that there is a clear lack of architectures that apply Deep Vector Blocking (like HNSW) specifically to the domain of toponym resolution to solve this recall bottleneck.

2.2.5.2 Gap 2: The Trade-off Between "Joint Training" and "Generative Reasoning"

The literature presents a dichotomy between efficient discriminative models and powerful generative models, with no clear bridge between them. On one side, architectures like MutualER [11] and SC-Block++ [40] achieve high accuracy by jointly training the blocker and matcher. However, this relies on differentiable dense vectors, making them incompatible with the discrete token outputs of modern Large Language Models (LLMs). On the other hand, generative approaches like LinkTransformer [1] and RACCOON [35] treat the LLM as a "black box," leading to high latency and Optical Character Recognition (OCR) sensitivity. Consequently, hybrid attempts like TGFR [50] tries to bridge this but fall back to deterministic fuzzy logic for verification, reintroducing the very semantic rigidity that neural models were meant to solve.

2.2.5.3 Gap 3: Scalability vs. Semantic Robustness in Heterogeneous Data

A major divide exists between "fast" indexing methods and "robust" matching methods. Benchmarks like B-VSS [63] and GraSP [64] prove that graph-based indices (HNSW) offer superior scalability. However, papers like LSBLOCK [25] and BLISS [18] reveal that these indices struggle with "Recall vs. Balance" trade-offs when applied to messy, high-dimensional text data. Conversely, traditional baselines like FLAP [62] and ReLink [24] scale well but require extensive manual feature engineering or lack semantic understanding (e.g. not matching "Home" with "House"). Therefore, we can say that there is a distinct lack of Hybrid Architectures that leverage the speed of HNSW for blocking while using the generalization power of an LLM (only on hard candidates) to handle global heterogeneity without the prohibitive cost of full-scan approaches.

2.3 Summary

This chapter established the empirical and theoretical foundations required to address the challenges of large-scale address normalization and matching. By combining a systematic literature review with a comprehensive set of preliminary benchmarks, we have defined both the boundaries of current state-of-the-art solutions and the specific architectural constraints of the target problem.

On one side, the **Preliminary Experiments** provided critical data-driven insights into the "Accuracy-Latency" trade-off. We demonstrated that while Deep Learning models (such as BERT) offer superior semantic generalization for parsing ($F1 = 0.993$), they present bigger inference costs. Conversely, statistical models like **CRF** proved to be the optimal baseline for structure extraction, offering near-perfect accuracy on standard components with negligible latency (0.06ms). Similarly, for the matching phase, **XGBoost** with engineered features emerged as a "sweet spot,"

outperforming both traditional fuzzy logic and raw embeddings in pairwise classification efficiency.

On the other hand, the **Literature Review** confirmed that the field is undergoing a fundamental shift from deterministic string metrics to Neural Entity Resolution and Vector Indexing. However, our **Gap Analysis** revealed that current solutions remain fragmented. State-of-the-art geocoders often rely on outdated keyword search for candidate generation, while modern neural linkers struggle to balance the "reasoning" power of LLMs with the computational efficiency required for global-scale data. Furthermore, there is a distinct lack of hybrid architectures that effectively integrate the scalability of graph-based indices (HNSW) with the precision of discriminative matchers.

To conclude, these findings collectively become crucial to formulate or final framework that will be proposed in the following chapter. Rather than relying on a single monolithic model, the proposed architecture will leverage the complementary strengths identified here. This approach will aim to bridge the identified gaps by delivering a system that is robust to semantic ambiguity without sacrificing the performance requirements of enterprise data streams.

Chapter 3

Methodological Approach

3.1 Research Design

The primary goal of this research is to construct a scalable, highly accurate system that connects the inconsistency between unstructured, high-entropy real-world geographic data and the rigid, structured requirements of enterprise databases. Given the computational complexity of resolving millions of addresses and the high variability of multinational linguistic patterns, solving this challenge required a systematic, iterative research design. However, rather than deploying a deep learning model from the outset, this methodological approach was structured as a progression of empirical experiments, and the complete trajectory of this research design can be seen in Figure 3.1. This strategy allowed for the isolation of individual algorithmic behaviors, the accurate measurement of accuracy-latency trade-offs, and the systematic scaling of the data environment. Before formal modeling, an exploratory data engineering phase was conducted to parse raw OpenStreetMap (OSM) files, perform an **Exploratory Data Analysis**, and test initial database persistence architectures (some comprehensive results and additional information of this preliminary phase can be found in Appendix A). Additionally, our research was conducted through multi-experimental phases, where, initially, we focused on a tightly controlled environment using only Portuguese OpenStreetMap data. The initial objective at this stage was to favor accuracy over operational latency of the core components of entity resolution, our parsing and matching phases, in absolute isolation and very specific environments (without introducing multiple languages). Consequently, for parsing, the evaluation progressed from traditional rule-based algorithms and external libraries to Machine Learning classifiers, Conditional Random Fields (CRFs), and Deep Learning sequence models. Simultaneously, the matching phase benchmarked traditional fuzzy string metrics against Supervised Machine Learning models using comprehensive feature engineering, and Deep Learning matchers such as BERT and Siamese Networks.

Following the isolated benchmarks, the research transitioned to an operational showdown in which the individual components were assembled into complete end-to-end pipelines (hereafter referred to as Paradigms). To evaluate our different paradigms, we used a 50k multinational dataset constructed from real address data given by the partner organization. This second phase com-

pared a purely statistical pipeline using inverted indexing against a Deep Learning baseline and domain-adapted pipelines utilizing dense vector indexing. Furthermore, to address the scale and subjectivity of end-to-end resolution, an LLM-as-a-Judge methodology was implemented to rigorously evaluate the final output from each paradigm (while tracking both retrieval precision and the compounding computational debt of high-dimensional vector search). For the parsing component, we also implemented a more deterministic evaluation by constructing a dataset (using the 50k multinational dataset) of parsed addresses, allowing us to benchmark the parsing architectures for each paradigm. Finally, after synthesizing all our empirical findings from the previous stages, a final architecture was proposed. By identifying that pure neural architectures could suffer from vector space collapse and prohibitive latency, while traditional statistical methods failed to generalize over heterogeneous data, a **Hybrid Neuro-Symbolic Framework** (our last paradigm) was developed. This final phase details the specific ontological schema (for parsing), the dual-mode retrieval strategy enforcing hard lexical constraints prior to dense vector search, and the combination of a Semantic XGBoost ranker to achieve enterprise-grade scalability and accuracy.

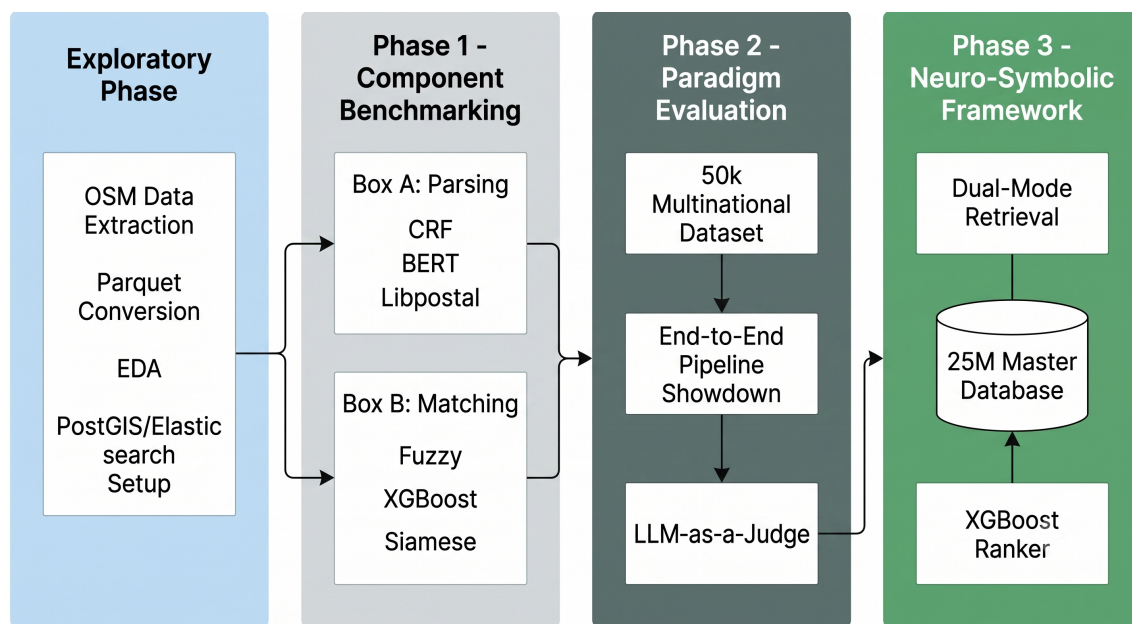


Figure 3.1: High-level block diagram of the iterative research design, progressing from data engineering to the final Hybrid Neuro-Symbolic Framework.

3.2 Data Engineering and Ground Truth Construction

The foundation of any robust geographic entity resolution system is heavily dependent on the quality, scale, and structure of its underlying data. Since this research required both a massive master database for retrieval and highly particular datasets for training models, multiple data engineering steps were performed to meet these specific requirements. The primary data source was OpenStreetMap (OSM) and, given the massive scale of the global OSM database, handling the

raw data required memory-efficient streaming pipelines. Initially, we focused only on extracting the Portuguese geographic subset and, after extracting it, an **Exploratory Data Analysis (EDA)** was conducted to quantify the structural integrity of the real-world geographic records. This initial assessment revealed a critical insight into the “Human-Database Gap” at the source level, where the vast majority of the geographic data is inherently fragmented. Specifically, the EDA demonstrated that only 2.87% of the Portuguese addresses in the dataset are fully complete (contain all tags from an OSM address, such as house number, city, postal code, among others), as shown in the additional information in Appendix A. Therefore, this high-entropy reality (missing information, structural noise, or colloquial formatting) clearly justified the need for refined probabilistic parsing and semantic-matching architectures capable of addressing all these problems.

Furthermore, before progressing to deep learning and dense vector retrieval, it was necessary to establish a baseline understanding of traditional geographic indexing and persistence architectures. Therefore, an exploratory database environment was constructed combining PostgreSQL and Elasticsearch. In this initial setup, PostgreSQL served as the geometric source of truth, and the spatial data was indexed using the **Generalized Search Tree (GiST)** indexes, which optimize bounding-box queries for physical coordinate resolution. Similarly, an initial Elasticsearch instance (using Docker) was also deployed to handle basic token-based and fuzzy candidate retrieval. After combining everything, this exploratory architecture provided a foundational performance baseline for traditional deterministic spatial queries, illustrating the limitations of relying purely on lexical indexes and strict spatial bounding boxes before transitioning to the final, highly scalable vector-based paradigms (also, the execution latencies of these spatial indexes are formally recorded in Appendix A). Additionally, while OSM data natively contains spatial coordinates for each record, real-world enterprise addresses rarely include this explicit geometric data, which makes text-based resolution a strict operational necessity.

Having quantified the structural sparsity of the raw geographic data and established baseline performance using traditional techniques, the next essential step was to generate a ground-truth dataset capable of training probabilistic parsing models for our experiments. Traditional manual annotation presents severe scalability bottlenecks, whereas purely rule-based synthetic generation often fails to capture the authentic language forms and semantic ambiguities inherent in human-generated text. Therefore, an automated annotation pipeline powered by an LLM was developed using **Google’s Geminin 3.1 Flash model**. In our script, we processed 110,000 raw addresses extracted from our Portuguese OSM dataset, and this generation was also constrained by a Pydantic schema (to enforce strict output formatting) with a temperature setting of 0.0, guaranteeing deterministic, highly consistent component extraction. In the end, our schema incorporated a three-tier geographic hierarchy (State, City, Other) alongside granular fields such as Road, House Number, Floor Info, and PO Box. However, despite the large volume of processed records, natural geographic data shows severe class imbalance, where specific components like street names or postal codes almost appear universally, while elements like building names or floor identifiers are exceedingly rare. Therefore, a specialized data augmentation strategy was also implemented, where approximately 2,000 additional addresses were synthetically generated and injected into the training

pool, specifically targeting underrepresented components like “Floor Info” and “PO Box”. While the LLM pipeline successfully generated highly accurate structured JSON annotations, sequence labelling architectures (like CRFs and Bidirectional LSTMs) require token-level annotations in the strict Begin-Inside-Outside (**BIO**) (Begin-Inside-Outside) format. Therefore, a custom “Longest-Span-First” matching algorithm was implemented to intelligently handle overlapping and nested components, as illustrated in Figure 3.2, sorting the extracted address components by text length in descending order before matching them with the tokenized input string. This prioritization makes sure that longer, multi-word spans are correctly tagged with appropriate **B-ROAD** and **I-ROAD** sequence tags before shorter substrings can incidentally match, consequently preserving the exact token alignment required for solid sequence modelling.

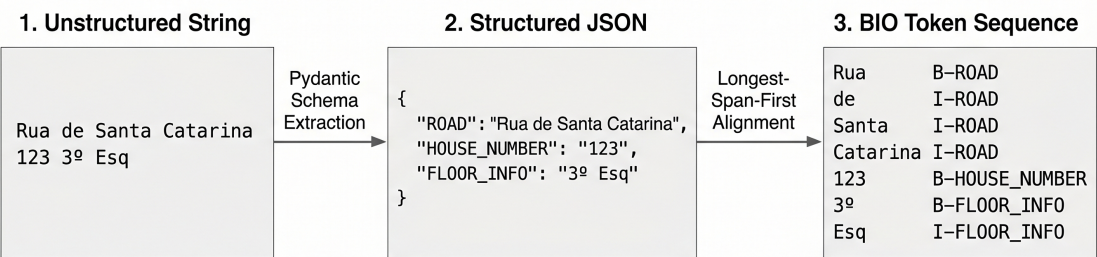


Figure 3.2: The three-stage data transformation pipeline, showing how unstructured text’s parsed into structured JSON and subsequently aligned into strict BIO token sequences using the Longest-Span-First algorithm.

In parallel to this parsing dataset, the supervised matching models required a distinct ground-truth corpus to teach them to distinguish between true semantic equivalence and superficial string similarity. Therefore, a comprehensive synthetic data generation pipeline was also developed to construct a large-scale, balanced binary classification dataset. For the positive class (matching pairs), an error injection algorithm was designed to introduce realistic Portuguese-specific noise into authentic baseline addresses. This included methodically applying abbreviation variations (for example, “Rua” to “R.”), generating manual typographical errors, and introducing structural reordering and partial component omissions. Conversely, the negative class (non-matching pairs) was explicitly engineered to present “hard negatives” rather than obvious mismatches. These challenges included pairs of same-city / different-streets, phonetically similar street names in distinct locations, and identical streets with differing house numbers. This complex generation process

yielded strictly balanced datasets, including a primary partition of 80/20 for the train/test split, making sure that the models were forced to learn fine-grained semantic distinctions (this dataset generation script was also reused in other future matching experiences for different data).

With the isolated training data firmly established, the attention shifted towards the construction of the enterprise-scale master database that would serve as the retrieval foundation for all subsequent end-to-end paradigms. The architecture of this ground truth database was anchored by a hierarchical administrative tree, provided by the enterprise partner, which structured country-specific administrative divisions down to the most granular micro-level without addresses or postal code data. Therefore, to populate our hierarchical tree, the parsed addresses from OpenStreetMap were integrated alongside a global postal code dataset (GeoNames) [42] to maximize spatial coverage. During these processes, as shown in Figure 3.3, global postal codes were anchored to the tree by normalizing administrative names (removing accents, standardized casing and stripping geographic stop-words) and evaluating candidate nodes using a combined similarity metric that weighted Jaccard similarity, Levenshtein distance, and bigram/trigram overlaps. This allowed us to traverse the hierarchy and add postal codes to the deepest valid administrative node possible. Following this, the OSM addresses were processed using the LibPostal library to standardize component labels and expand language-specific abbreviations. After this initial normalization, we started by inserting the addresses, and if an address contained a recognized postal code, it was nested directly within that node. Otherwise, if the postal code was missing, the system utilized the previously defined combined similarity scores to perform a hierarchical fuzzy search based on city, county, or district names. Finally, to capture addresses that failed lexical routing and defaulted to the root country node, a spatial fallback mechanism was added. This system converted the existing latitude and longitude coordinates into 3D Cartesian space and constructed a spatial KD-Tree [4], geometrically relocating addresses to the closest valid administrative node within a strict 0.75km threshold. Furthermore, to evaluate the scalability of the proposed architectures in varied linguistic landscapes, this master database was expanded to include data from five different countries: Portugal, France, Germany, Spain, and the United States. However, to preserve operational viability within hardware storage constraints, a strict cap of 5 million addresses per country was enforced, resulting in a final, highly curated 25-million-record ground truth database. Also, for these paradigm-level experiments, **Elasticsearch** was selected as the sole persistence layer due to its native support for storing, indexing, and retrieving high-dimensional dense vector embeddings, which is an essential requirement for our deep learning models. In parallel to the master database, we also had to prepare a distinct evaluation dataset (that did not have data from the same sources we used to train our models) to thoroughly test and benchmark our paradigms against authentic, high-entropy user input. Therefore, a sample of 50,000 real-world unstandardised addresses (10,000 for each of the five target countries) was curated from the internal data streams of the enterprise partner, as summarized in Table 3.1. Given this, to also prepare this dataset for deterministic evaluation, an internal parsing tool was initially applied, followed by a secondary LLM-powered review phase utilizing the Gemini 3.1 flash model. In this step, the LLM acted as an automated auditor, reviewing and correcting the assigned tags to ensure strict, deterministic

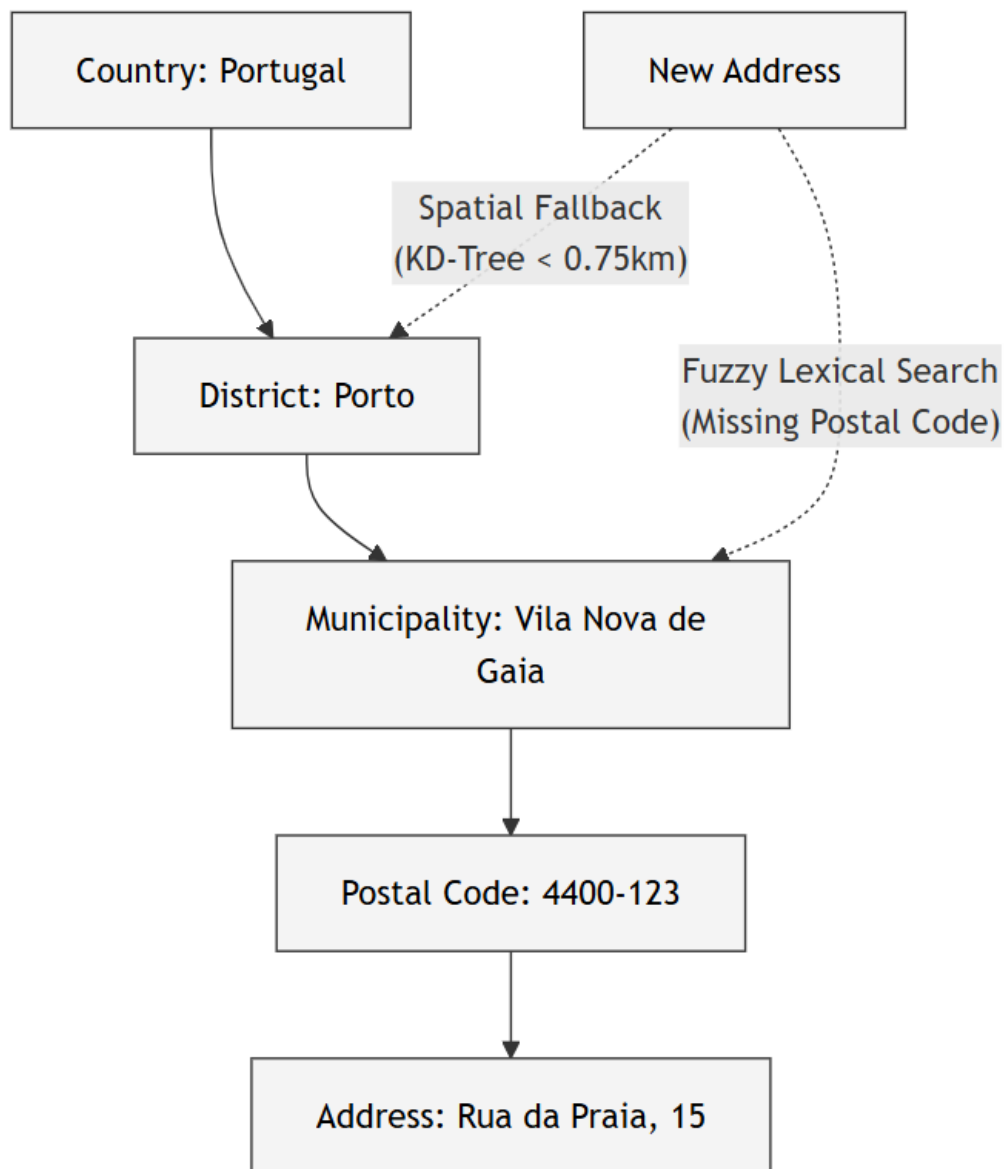


Figure 3.3: The multi-stage routing architecture for populating the ground truth database. The diagram illustrates the standard deterministic lexical routing (solid lines) alongside the probabilistic fuzzy and spatial fallback mechanisms (dashed lines) for orphaned records.

compliance with the established tag ontological schema (different schemas were used during our experiments). In addition to that, to guarantee clarity across the progressive architectural evaluations, it is important to explicitly define how the ontological schema progressed alongside the paradigms:

- **Phase 1 Schema:** A custom, highly granular BIO-schema used exclusively for the Portuguese OSM dataset, which included micro-components like B-ROAD, B-POSTALCODE, and B-FLOOR_INFO.
- **Paradigms 1 & 2 Schema:** A 9-tag evaluation schema used for the global dataset, introducing macro-geography terms such as SUBURB, DISTRICT, and COUNTY.
- **Paradigms 2.0 & 2.1 Schema:** The native 10-tag Deepparse schema, which removed CITY in favor of MUNICIPALITY and PROVINCE.
- **Paradigm 3 Schema:** The final, strict 7-tag enterprise mapping designed to collapse granular multinational tags into standard database fields (for example, mapping MUNICIPALITY to CITY).

Table 3.1: Distribution of multinational records across the enterprise master database and the curated evaluation dataset.

Country	Master Database (Cap)	Evaluation Queries
Portugal (PT)	5,000,000	10,000
Spain (ES)	5,000,000	10,000
France (FR)	5,000,000	10,000
Germany (DE)	5,000,000	10,000
United States (US)	5,000,000	10,000
Total	25,000,000	50,000

After populating the master database and the evaluation dataset, the final data engineering step involved preparing the specific training corpora for the end-to-end paradigm evaluations. For the initial operational baselines, Paradigms 1 and 2, the models were trained on a multinational dataset that encompasses all five target countries. This allowed us to evaluate whether the models were capable of handling cross-cultural language forms and country-specific formatting conventions right from the outset. Given that, to train our parsing models for these initial paradigms, a dedicated pipeline processed up to 50,000 OSM records per country, and random noise injection was applied to each record in order to prevent the models from merely memorizing positional biases (for example, assuming a number at the end of a string is always a postal code). In this way, we were able to try to reproduce realistic human data entry by randomly dropping components, applying structural typos (swapping or dropping adjacent characters), altering delimiters (removing commas or adding double spaces), and occasionally completely randomizing the component order. The resulting chaotic addresses were then dynamically tokenized, and a custom algorithm mapped

the character spans to their corresponding BIO tags, yielding a strong CoNLL-formatted dataset ready for sequence labelling. In addition to the parsing data, a binary classification dataset was constructed to train the supervised matching components. As we did for previous experiments, we followed the same logic of creating positive and negative pairs by introducing asymmetric missing data, country-specific abbreviation variations, minor typos into the baseline addresses, altering digits in a valid postal code, or pairing perfectly identical city and postal code components with a completely different valid street name from the same local area. Furthermore, a feature extraction layer calculated 16 distinct distance measures for every generated pair, including global and component-wise Levenshtein distances, Jaccard indices, and phonetic Metaphone matches. Furthermore, to construct the domain-adapted pipelines (Paradigms 2.0 and 2.1), a globally recognized multinational dataset was introduced, the Deepparse dataset [59]. Since this dataset natively contains addresses from dozens of countries already structurally labelled for sequence parsing, it was used directly to train the advanced sequence models. Additionally, to train their corresponding matching components, the exact same synthetic data generation logic (injecting controlled noise and pairing hard geographic negatives) was applied to the Deepparse records. Finally, training our ultimate **Hybrid Neuro-Symbolic Framework** (Paradigm 3) required an entirely new specialized data curation process to conform to its strict ontological schema and dual-mode retrieval logic. The details of this final data preparation phase will be detailed along with the architectural breakdown of our Paradigm 3 in the following sections of this methodology.

3.3 Phase 1: Preliminary Component Benchmarking

In parallel to our data preparations and before assembling the end-to-end entity resolution paradigms, it was fundamental to evaluate the core algorithmic components in absolute isolation (our Parsing and Semantic matching phases). Therefore, our main objective with this preliminary phase was to systematically measure how different model architectures processed unstructured Portuguese geographic data, enabling the accurate measurement of their semantic capabilities and linguistic blind spots. By benchmarking these components individually, we could empirically justify which architectures would be worth trying in our multinational, end-to-end pipelines in Phase 2.

The parsing evaluation was designed as a progressive architectural hierarchy, illustrated in Figure 3.4, beginning with naive classification approaches and culminating in deep attention-based sequence models. Initially, standard **Machine Learning classifiers**, like Random Forests, Support Vector Machines or XGBoost, were trained to predict the tag of each token independently, depending exclusively on isolated lexical features. These baseline techniques were included to test the hypothesis of whether isolated token classification could succeed without sequence awareness, and their primary goal was to measure their baseline capacity to handle multi-word geographic entities (such as “Rua Nossa Senhora da Conceição”) without maintaining internal tag states (I-ROAD). In addition, **Conditional Random Fields** (CRFs) were implemented by modelling the conditional probability of neighbouring tags and learning transition matrices in order to evaluate their capacity to capture the linear syntactic structure of the addresses, establishing our primary

statistical sequence baseline. In parallel with the CRF, the industry-standard **Libpostal** library was also evaluated. Although it is known that Libpostal uses a globally trained statistical model and performs exceptionally well on standard formats, this phase was designed to benchmark its globally trained model against our synthetically generated Portuguese dataset. The main objective was to systematically assess whether an out-of-the-box global library could accurately parse micro-administrative divisions without country-specific fine-tuning. Furthermore, to go beyond statistical probabilities and capture deep semantic context, **Deep Learning** sequence models were also evaluated, specifically a **BiLSTM-CRF** network and a fine-tuned BERT Token Classifier. Although traditional CRF relied heavily on manual feature engineering established during the data preparation phase (such as capitalization rules and digit flags), the neural architectures generated dense contextual embeddings for each token. Therefore, the **BERT** architecture was introduced, employing its bidirectional self-attention mechanism, to test its capacity to generalize on rare and underrepresented components compared to the rigid feature engineering of statistical models. Ultimately, the methodology of this isolated phase was intended to measure the exact trade-offs (focused primarily on predictive performance rather than computational latency) between the rigid speed of statistical models and the semantic depth of transformer-based sequence modelling, dictating which architectures would be worth advancing to our Phase 2 pipelines.

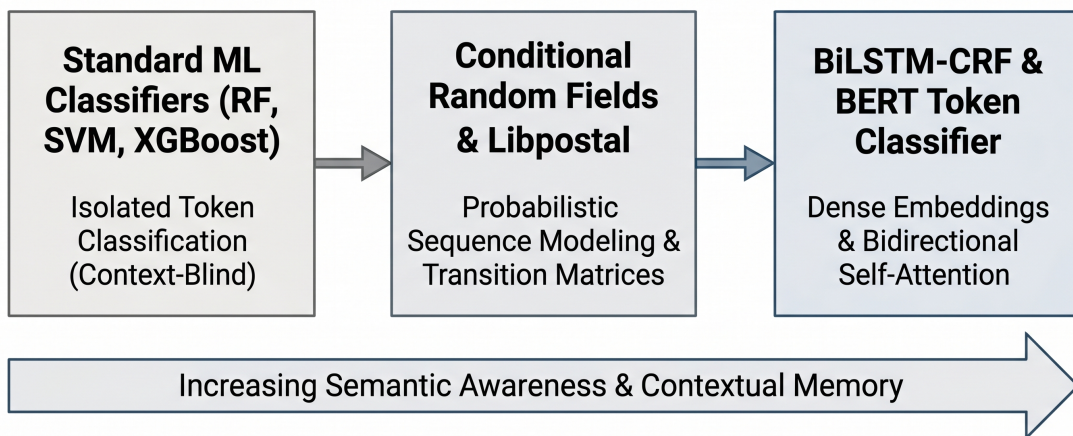


Figure 3.4: The progressive architectural evaluation framework for the isolated parsing component, advancing from context-blind lexical classifiers to deep attention-based sequence models.

Following the evaluation of the parsing components, the isolated benchmarking shifted to the second essential pillar of our pipelines, **Semantic Matching**. The main goal in this phase was to measure the capacity of different architectures to determine whether two structurally unique address strings referred to the same physical location. Therefore, to carefully test this, the models were evaluated against our previously mentioned binary classification datasets, and our evaluation

was structured across three different strategies: **Traditional String Metrics**, **Feature-Engineered Supervised Learning**, and **Deep Learning**. The benchmarking began with the evaluation of traditional, purely mathematical string comparison algorithms. This baseline generation included character-based edit distances (such as Levenshtein and Jaro-Winkler), token-based overlap metrics (for example, Term Frequency-Inverse Document Frequency (**TF-IDF**)), and phonetic algorithms. The main methodological goal here was to establish an algorithmic baseline and systematically assess the “glass ceiling” of these more established approaches. In addition, these metrics were tested to determine the threshold at which purely distance-based algorithms fail when confronted with targeted structural noise, extreme abbreviation variations, and complex word reordering. However, to bridge the gap between simple string overlap and true semantic equivalence, the evaluation advanced to supervised machine learning, using different classifiers (for example, XGBoost [6]) powered by a highly dimensional feature extraction framework. Instead of relying on a single string metric, a comprehensive set of 49 distinct features was engineered for every pair of addresses, and the metrics that were calculated are described in detail in Appendix B. This included global fuzzy distances, phonetic algorithms (such as Metaphone) to capture auditory similarities, and, crucially, component-specific metrics. By computing independent similarities (for example, for the street, city, and postal code separately), this model was designed to evaluate whether a mathematically dense, manually engineered feature space could adequately model human geographic reasoning. Finally, to test the limits of automated feature extraction and deep semantic understanding, neural matching architectures were also introduced. Initially, out-of-the-box, pre-trained **BERT** embeddings were evaluated to measure their zero-shot capability to group geographically equivalent strings in a dense vector space. Subsequently, a custom **Siamese Network**, illustrated in Figure 3.5, was constructed and fine-tuned specifically adapted to our geographic data. This architecture featured a dual-encoder setup in which both addresses in a candidate pair were processed concurrently using identical transformer weights. Additionally, the resulting contextual token embeddings were condensed using mean pooling, and the concatenated sentence-level representations were passed through a three-layer fully connected classification head, trained through Binary Cross-Entropy (**BCE**) loss. The fundamental goal of testing this Siamese architecture was to determine whether deep learning could autonomously capture latent spatial semantics.

Ultimately, there was a need to empirically compare and benchmark the different architectures for each isolated component. Because real-world geographic data suffer from class imbalance, standard accuracy metrics might heavily bias the results toward ubiquitous components (such as street names and cities) while masking serious failures on less frequent classes. Therefore, sequence parsing models were evaluated primarily using Precision, Recall, and the Macro-Averaged F1-Score. By calculating the F1-Score independently for each sequence tag and averaging them unweighted, the Macro F1-Score ensured that a model’s ability to extract a rare “PO Box” tag was weighted equally to its capacity to extract a common “Postal Code”. In this way, it effectively penalized the models that collapsed under the pressure of underrepresented classes, as summarized in the evaluation framework in Table 3.2. Furthermore, for the semantic matching compo-

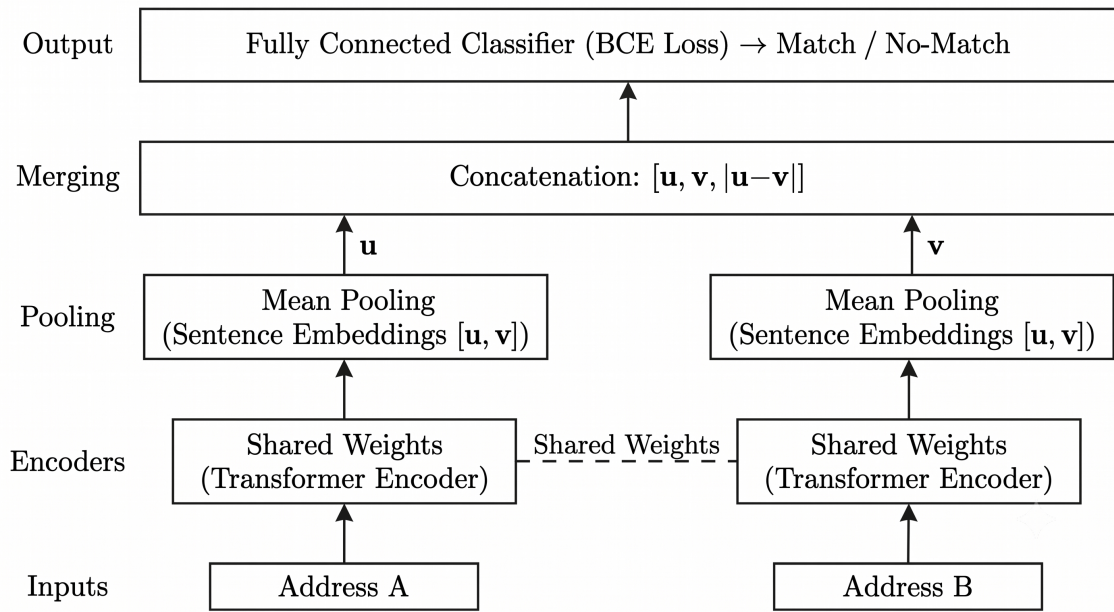


Figure 3.5: Architecture of the Siamese Neural Network for semantic matching. The dual-encoder setup handles address pairs simultaneously, applying mean pooling and a fully connected classification head to predict semantic equivalence.

nents, the evaluation focused on binary classification metrics where the primary metric used was the **Receiver Operating Characteristic Area Under the Curve (ROC-AUC)**. Considering that the matching architectures (especially the XGBoost and the Siamese Network) output continuous probability scores rather than binary decisions, the ROC-AUC provided a comprehensive measure of each model's ability to separate true geographic matches from deceptively similar false positives across all possible confidence thresholds. In addition to that, the Binary F1-Score was calculated at the optimal threshold to measure the harmonic mean of precision & recall in a strict operational setting.

3.4 Phase 2: End-to-End Paradigms

Although isolated benchmarking in Phase 1 provided important insights into the theoretical capabilities of individual components, geographic entity resolution in an enterprise setting is fundamentally a pipeline problem. For example, a highly accurate semantic matcher is operationally useless if the retrieval mechanism fails to surface the correct candidates from the database in the first place. Therefore, Phase 2 transitioned the research from isolated metrics to end-to-end pipeline evaluation. In this phase, the selected parsers and matchers were integrated with an Elasticsearch persistence layer to process the 50,000 evaluation queries (from our previously explained evaluation dataset). In addition to that, all paradigms executed their searches against the full 25-million record master database (strictly enforcing the 5-million address cap across all target countries that we have previously explained). The main goal of our methodology from this last phase was to construct distinct retrieval paradigms, ranging from traditional statistical pipelines (more

Table 3.2: Summary of the evaluation metrics established for the isolated component benchmarking, categorized by their respective tasks and analytical objectives.

Pipeline Component	Primary Metric	Methodological Justification
Sequence Parsing	Macro F1-Score	Prevents ubiquitous tags (e.g., City) from masking failures on rare tags (e.g., Floor Info) due to unweighted class averaging.
	Precision & Recall	Quantifies the trade-off between false-positive boundary expansions and missed component extractions.
Semantic Matching	ROC-AUC	Evaluates the model’s discriminative capacity across all confidence thresholds against hard negative pairs.
	Binary F1-Score	Measures strict classification performance at the optimal operational threshold.

focused on speed retrieval) to advanced deep learning frameworks (more focused on pure accuracy and semantic meaning), and evaluate their functional feasibility, cross-cultural robustness, and computational effectiveness.

Our first pipeline, designated as Paradigm 1 and illustrated in Figure 3.6, was engineered to establish itself as our efficiency baseline. It deliberately avoided deep deep neural architectures, instead relying on sparse retrieval, probabilistic sequence modeling, and manually engineered features. This allowed us to systematically benchmark the maximum capabilities of traditional, computationally lightweight techniques before justifying the transition to refined Neural Networks. Because Elasticsearch Boolean queries cannot inherently traverse deep, multi-level JSON trees, the 25-million-record master database was transformed. Therefore, a custom flattening algorithm traversed the hierarchical geographic nodes, concatenating the specific tags (street, house number, postal code, city, and broader administrative divisions) into a unified flat document structure. Then, these records were ingested into an Elasticsearch index configured with a custom lexical analyser using ASCII folding, lowercasing, and edge n-gram token filters (ranging from 3 to 10 characters) to natively support sub-word matching and minor typographical errors. Once the master database was properly indexed, the operational flow of the pipeline could be executed. Therefore, when an unstructured query entered the system, it was first processed by the probabilistic parsing module. Specifically, the **Conditional Random Field (CRF)**, trained using the L-BFGS optimization algorithm, was deployed by relying on the handcrafted morphological features established in Phase 1 (for example, regular expressions of digits and capitalization flags). This CRF segmented the raw query string and extracted the strict BIO-formatted address components, and with the query successfully parsed, the next critical step was candidate retrieval. Given the computational impossibility of running pairwise matching against the entire 25-million record database, a rapid block-

ing phase was executed. The components extracted by the CRF were then dynamically formulated into an Elasticsearch Boolean *should* query. This query leveraged the traditional BM25 [48] scoring algorithm against the previously constructed inverted edge n-gram index effectively filtering the massive database down to a strict top- K candidate list (specifically, the 20 most similar lexical records). Finally, the retrieval phase passed these top- K candidates to the semantic matching component powered by the XGBoost classifier. For each candidate pair, our system computed the 49-dimensional feature vector, which incorporated global fuzzy string distances, phonetic overlaps, and component-wise metrics. Crucially, to ensure that the model could definitely output a binary “Match” or “No-Match” decision rather than returning an arbitrary similarity score, the XGBoost probabilities were calibrated using Isotonic Regression via cross-validation. In this way, it ensured that the output closely mirrored the true statistical confidence, allowing the pipeline to enforce a strict confidence threshold. Ultimately, if the top-ranked candidate retrieved by Elasticsearch and scored by XGBoost exceeded this predefined threshold, the pipeline successfully resolved the address; otherwise, the system was programmed to systematically abstain.

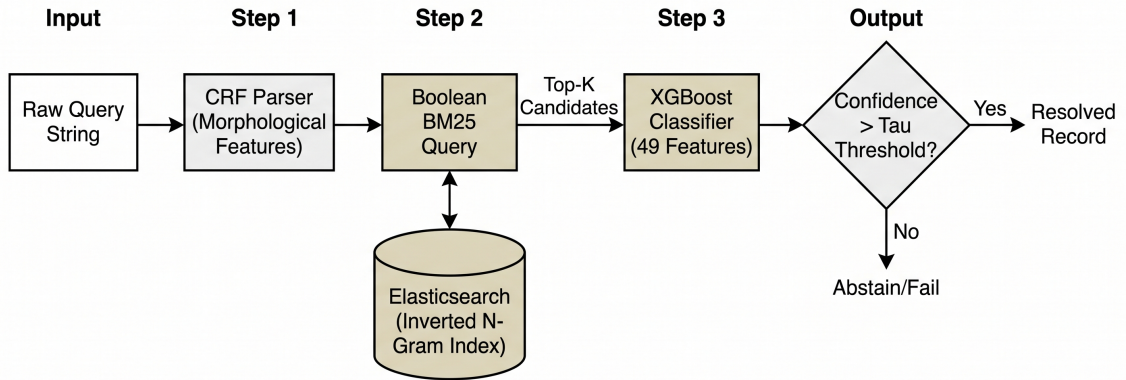


Figure 3.6: The end-to-end architecture of Paradigm 1. The pipeline utilizes a CRF for morphological parsing, BM25 Boolean queries for sparse retrieval, and an Isotonic-calibrated XGBoost classifier for the final threshold-based candidate matching.

Following the establishment of the statistical baseline, the methodological focus shifted towards the deployment of Paradigm 2, the **Deep Semantic Standard**. The primary objective of this architecture was to systematically test whether replacing deterministic rules and handcrafted features with continuous dense vector representations could actually overcome the lexical “glass ceiling” observed in traditional systems. Therefore, the entire underlying persistence layer required a fundamental transformation. Instead of relying on an inverted n-gram index, the 25-million capped master database was converted into a continuous semantic space. For this step, a

multilingual Bi-Encoder (*paraphrase-multilingual-MiniLM-L12-v2*) was used to generate a 384-dimensional dense vector for every geographic record. Then, these embeddings were ingested into an Elasticsearch instance explicitly configured for dense vector storage, utilizing cosine similarity to enable highly efficient **Approximate Nearest Neighbor** (ANN) searches across the massive multinational corpus. Furthermore, in this paradigm, the operational execution completely replaced the probabilistic and statistical components of the previous pipeline with deep transformer-based architectures. When an unstructured evaluation query entered the system, it was first routed to a neural parsing module powered by a fine-tuned *xlm-roberta-base* model [8]. Unlike the previous CRF (reliant on rigid handcrafted morphological regexes), this token classifier leveraged bidirectional self-attention and sub-word tokenization to generate contextualized embeddings, extracting the required BIO sequence tags directly from the latent linguistic structure of the query address. Following the parsing phase, the semantic retrieval (our blocking phase) was executed. In this step, the raw query string was embedded by the identical Bi-Encoder used during the database preparation. Therefore, instead of relying on a Boolean BM25 lexical match, the system executed a k-Nearest Neighbors (k-NN) vector search within Elasticsearch, fetching the top- K candidates located closest to the query in the 384-dimensional latent space. In addition to that, this step was methodologically designed to bypass traditional lexical failures, allowing the pipeline to retrieve geographically relevant candidates even when faced with extreme structural noise. Finally, to continue trying to improve the precision of geographic matching of the retrieved top- K candidates, Paradigm 2 discarded the 49-dimensional manual feature engineering used by XGBoost in favor of a neural Cross-Encoder. By using another instance of the *xlm-roberta-base* model, this architecture processed the original evaluation query and the database candidate simultaneously. Unlike the Bi-Encoder, which encodes strings in strict isolation, the Cross-Encoder concatenated the two strings, allowing the transformer’s self-attention heads to compute deep cross-sequence semantic interactions at every layer. Then, the resulting sequence output was passed through a linear classification head, producing a raw logit that was subsequently converted via a sigmoid function into a final confidence probability of 0.0 to 1.0. Also, like in the preceding paradigm, if this final probability exceeded the predefined threshold, the pipeline recorded a successful match.

Although the initial deep semantic standard (Paradigm 2) successfully transitioned the retrieval architecture into a continuous vector space, its reliance on synthetically generated OpenStreetMap data inherently might have limited its capacity to model authentic cross-cultural linguistic anomalies. Therefore, the methodology evolved to incorporate strict domain adaptation in order to achieve true multinational robustness and mitigate the risk of vector space collapse (where dense embeddings lose their discriminative power when exposed to massive out-of-distribution data). This required abandoning the synthetic datasets in favour of the globally recognized multilingual Deepparse dataset, leading to the development of our next two challenger pipelines, Paradigms 2.0 and 2.1. The paradigm 2.0 was engineered to refine the Transformer-based approach through rigorous domain adaptation. The sequence parsing module was completely overhauled by fine-tuning a new *xlm-roberta-base* token classifier exclusively on the new Deepparse corpus. This training phase used mixed-precision (FP16) and gradient accumulation to efficiently process the

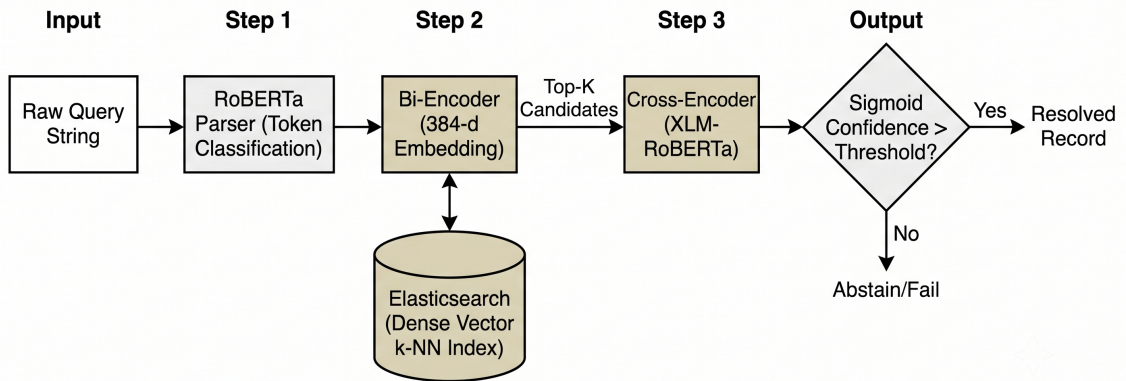


Figure 3.7: The end-to-end architecture of Paradigm 2. The pipeline transitions entirely to deep learning, utilizing an XLM-RoBERTa token classifier for parsing, a Bi-Encoder for dense vector k-NN retrieval, and a Cross-Encoder for final semantic matching.

massive dataset mapping the sub-word token output to a highly specific 10-tag multinational geographic schema. Although it maintained the off-the-shelf multilingual Bi-Encoder for its initial k-NN candidate retrieval against the database, the downstream semantic matching module was significantly upgraded. A new Cross-Encoder, initialized from the *ms-marco-MiniLM-L-12-v2* architecture, was fine-tuned specifically on millions of matching pairs generated from the Deepparse data, where the main goal was to provide a much stricter domain-aware semantic ranking. In parallel, Paradigm 2.1, illustrated in Figure 3.8, was introduced as a direct challenger, designed to explicitly test whether the quadratic computational latency of Transformer attention mechanisms could be bypassed without sacrificing extraction accuracy. Therefore, it replaced the computationally heavy RoBERTa parser with the Deepparse library’s native Bidirectional Long Short-Term Memory (Bi-LSTM) network, powered by a lightweight FastText embeddings. Additionally, this paradigm also pushed domain adaptation to its absolute limit by also replacing the off-the-shelf retriever. This time, it utilized a custom Bi-Encoder that had been explicitly fine-tuned on the Deepparse corpus using contrastive loss, ensuring that the entire pipeline (from the Bi-LSTM parsing and custom Bi-Encoder retrieval down to the custom Cross-Encoder matching) was seamlessly adapted to the identical multinational data distribution. Therefore, by executing both of these domain-adapted paradigms against our evaluation queries, the methodology established a rigorous framework to empirically isolate whether the future of geographic entity resolution lies in massive self-attention models or highly optimized, domain-specific recurrent networks. Ultimately, it is also critical to note a shift in the query formulation logic introduced during this domain adaptation phase. While Paradigm 2 executed its vector search by directly embedding the raw, unparsed input string, Paradigms 2.0 and 2.1 formulated their input queries differently. For

these updated pipelines, the raw string was first processed by their respective sequence models, and the respective extracted structured tags were subsequently utilized to construct the retrieval query. This shift was implemented to test whether pre-structuring the input prior to embedding would tighten the geographic clustering within the dense vector space.

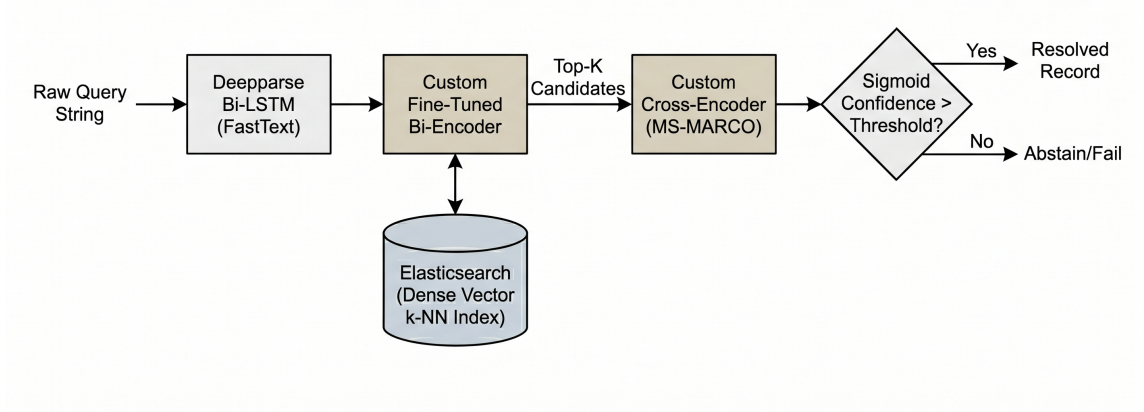


Figure 3.8: The end-to-end architecture of Paradigm 2.1. This fully domain-adapted pipeline utilizes a lightweight Bi-LSTM network for parsing, alongside a custom fine-tuned Bi-Encoder and Cross-Encoder to optimize both semantic precision and computational latency.

Furthermore, in order to ensure a rigorous and fair comparison across all the different architectures (spanning statistical baselines, raw semantic embeddings, and domain-adapted neural networks), a unified evaluation strategy was established. Since the core challenge of geographic entity resolution involves both extracting structure and querying a master database, relying on a single metric is insufficient. Given this, every paradigm was subjected to a strict tridimensional evaluation protocol using our evaluation dataset (the 50,000-record sample of highly unstructured, real world evaluation queries) against the 25-million capped master database. Therefore, the first phase of this protocol consisted of a deterministic parsing benchmark. Given that our evaluation dataset was annotated with clean, formatted strings rather than strict token-level BIO tags, traditional sequence labelling metrics could not be used. Instead, parsing performance was evaluated at the individual component level using two string-comparison metrics. Firstly, the system calculated an **Exact Match (EM)** rate to measure absolute precision. Let $\hat{y}$ represent the predicted string and y represent the ground truth string. The Exact Match is defined as:

$$EM = \begin{cases} 1, & \text{if } \hat{y} = y \\ 0, & \text{otherwise} \end{cases} \quad (3.1)$$

Then, a **Token-Overlap F1-Score** was also calculated to account for partial semantic credit. By calculating precision and recall based on the intersection of tokens between the predicted string and the ground truth, the framework could accurately penalize models for dropping critical words. Therefore, both strings are tokenized into multisets of words. Let P represent the multiset of predicted tokens and G represent the multiset of ground truth tokens. The token-level metrics are

computed as follows:

$$Precision = \frac{|P \cap G|}{|P|} \quad (3.2)$$

$$Recall = \frac{|P \cap G|}{|G|} \quad (3.3)$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (3.4)$$

Where the intersection $|P \cap G|$ accounts for the overlapping token frequencies between the prediction and the ground truth. Additionally, these metrics were also computed globally and then disaggregated across the five distinct target countries to empirically evaluate the cross-cultural robustness of each parser. In parallel, the second dimension addresses the end-to-end retrieval and matching efficacy. A critical methodology challenge also arose in this step due to the real-world evaluation queries lacking definitive, pre-linked database identifiers mapping them to the golden records. Therefore, the solution presented was to perform an automated “LLM-as-a-judge” methodology (which was also validated by **Zheng et al.** [67]). In this framework, the raw input query alongside the top retrieved database candidate from two competing paradigms were passed to a Large Language Model (in this case, Google’s Gemini 3.1 Flash) configured with a strict temperature of 0.0 for deterministic reasoning. The enforcement of this zero temperature setting combined with a constrained Pydantic output schema was a deliberate design choice to mitigate the position and verbosity bias, and stochastic inconsistency (that was also documented as systematic failure modes of LLM evaluators by **Jiang et al.** [17]). This way, constrained by a strict ontological schema, the LLM acted as an impartial GIS auditor instructed to account for phonetic typos, missing geographic hierarchy, and abbreviation variations. Given this, the judge at the end systematically outputted one of four verdicts: a **Win** for the first or the second frameworks, a **Tie** for perfectly equal records, or a **Mutual Failure** if neither systems retrieved a geographically compatible record. This way, this approach was able to provide a highly scalable, empirical win-rate while precisely quantifying the “noise floor” where all standard retrieval architectures collapse. Finally, our third and last phase of this evaluation protocol, focused on evaluating the operational feasibility through granular hardware and latency profiling. As established in previous sections of our document, computational efficiency can be the primary bottleneck for deploying deep learning in production. Therefore, the pipeline natively measured and logged the execution time across the three distinct execution stages for every query: the **parsing time**, **retrieval time**, and **matching time**. By tracking these metrics sequentially, the evaluation established a definitive assessment about the trade-off between semantic accuracy and operational latency. Finally, to accurately contextualize the latency profiling within this evaluation protocol, the experiments were executed in two distinct hardware environments. The lightweight statistical baseline (Paradigm 1) was deliberately executed on a standard consumer-grade CPU to simulate localized, low-resource deployment constraints. In contrast, due to the severe computational demands of processing complex self-attention layers, the deep semantic pipelines (Paradigms 2, 2.0, and 2.1) required the use of dedicated high-performance GPU hardware (given by the enterprise partner).

Ultimately, this tridimensional protocol provided the empirical foundation necessary to expose the limitations of both statistical baseline and the deep semantic challengers, thereby justifying the development of our ultimate and final architectural solution: the **Hybrid Neuro-Symbolic Framework**.

3.5 Phase 3: The Hybrid Neuro-Symbolic Framework

While the deep semantic methodologies evaluated in Phase 2 successfully demonstrated the capability of continuous vector spaces to capture geographic nuances, their purely probabilistic nature introduced an unacceptable margin of error for enterprise-grade entity resolution. By allowing a neural network to autonomously guess the geographic hierarchy without deterministic safety nets, inherently led to spatial hallucinations (particularly when confronted with the “noise floor” of highly corrupted data). Therefore, the final methodological evolution of this research culminated in our final Paradigm 3 where its architecture was designed to bridge the gap between two traditionally opposing philosophies. It leverages the “Neuro” capabilities of fine-tuned deep learning models to understand unstructured human text, while enforcing the “Symbolic” rigidity of deterministic geographic hierarchies and mathematical string rules to govern the actual database retrieval. This way, the construction of this final Paradigm was executed across three distinct architectural layers, including ontological schema adaptation, constrained semantic retrieval, and knowledge distillation matching.

The foundation of the framework began with the parsing module. While the previous domain-adapted pipeline (Paradigm 2.1) successfully utilized the Deepparse library’s Bidirectional LSTM network, its native output produced a 10-tag geographical schema (including different micro-divisions such as Municipality, and Province). However, according to the partner organization, our database should strictly adhere to a more generalized 7-tag ontological schema consisting only of *StreetNumber*, *StreetName*, *Unit*, *City*, *State*, *Country*, and *PostalCode*. Therefore, a comprehensive schema adaptation pipeline needed to be implemented in order to integrate the deep learning parser into this deterministic environment. By using the master training dataset of over 500,000 synthetically noised addresses (mentioned in a previous section of our methodology), a deterministic mapping logic, illustrated in Figure 3.9, was possible to be engineered to funnel the tags into the enterprise structure. For example, micro-administrative divisions such as Municipalities were clustered into the “City” tag, while broader Districts and Regions were systematically re-mapped to the “State” tag. Therefore, with this new mapping established, the underlying Fast-Text Bi-LSTM model was initialized and a targeted fine-tuning phase was executed. By freezing the deep linguistic embeddings and exclusively retraining the sequence classification head over multiple epochs (5) with a high batch size (256), the model was explicitly taught to project its rich semantic understanding directly into the rigid 7-tag enterprise schema. This ensured that the first layer of the Neuro-Symbolic Framework could extract operationally compliant structured data without sacrificing the low-latency inference speeds inherent to recurrent neural networks.

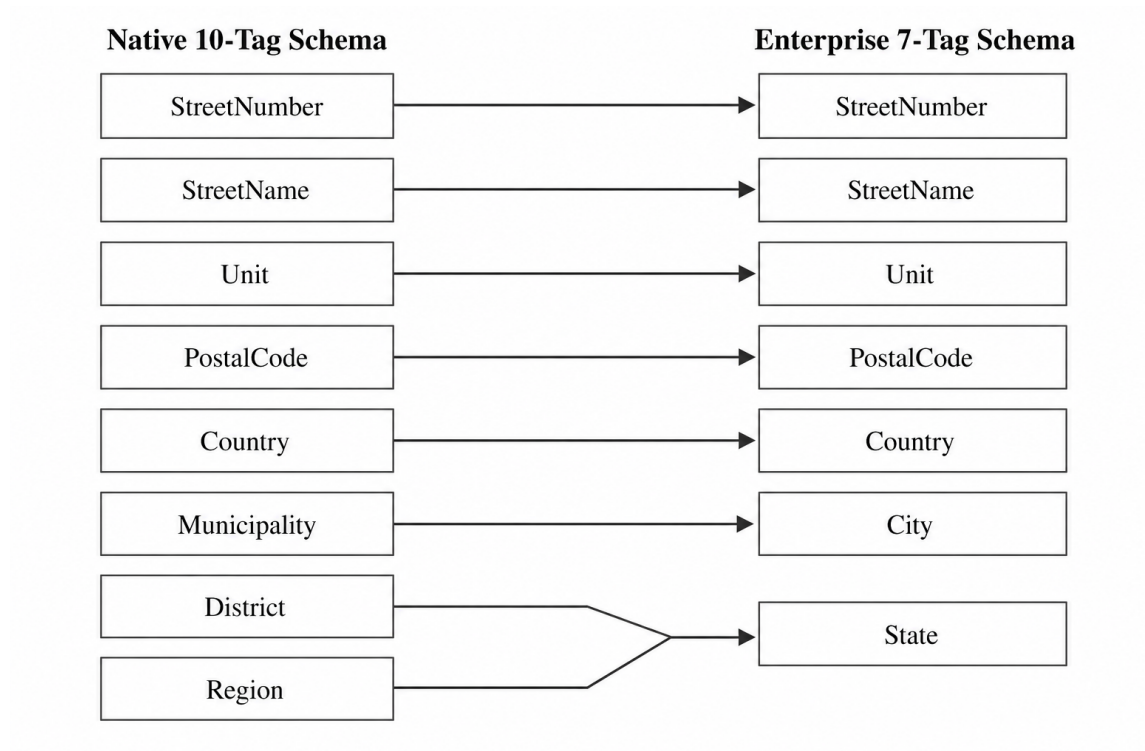


Figure 3.9: The ontological schema adaptation process, demonstrating how the native 10-tag Deepparse sequence output was systematically mapped and fine-tuned into the strict 7-tag enterprise geographic schema.

Given that the sequence parser successfully aligned to the enterprise schema, the methodology advanced to the second architecture layer. In this **constrained semantic retrieval**, the underlying Elasticsearch persistence layer required a specialized hybrid configuration. Rather than simply storing records purely as flat text or isolated embeddings, the database was engineered to store the 384-dimensional dense vectors directly alongside strict categorical **keyword** fields representing the structured geographic hierarchy (such as City, State, Country, and PostalCode). During this data flattening process, UUIDs were explicitly generated for every record in order to guarantee that addresses sharing identical postal codes would not overwrite one another during bulk indexing. Given this, the operational retrieval phase of Paradigm 3 constitutes the core “Symbolic” component of the framework. In the previous baseline and deep semantic paradigm, deterministic geographic filters were intentionally omitted. That design choice was a strict methodological requirement necessary to empirically benchmark the raw, unassisted recall capabilities of isolated BM25 lexical searches and pure k-NN vector searched. Having established those absolute baselines, Paradigm 3 was designed to fuse both paradigms. Therefore, when the structured tags are extracted by the fine-tuned Bi-LSTM parser, they are not merely passed as context; they are dynamically injected into a complex Elasticsearch Boolean query. Specifically, while the raw query string is converted into a dense vector to execute an Approximate Nearest neighbor search, the parsed categorical tags (such as the extracted City, Country, or Postal Code) are simultaneously applied as rigid *match clauses* within a Boolean *should* or *must* array, as conceptualized in Figure 3.10. Therefore, by combining the semantic vector similarity (the “Neuro” component) with these deterministic categorical filters (the “Symbolic” component), the retrieval pipeline drastically reduces the high-dimensional search space. This fusion primary objective is to inherently eliminate spatial hallucinations (such as retrieving a semantically identical street name located on the wrong continent) by forcing the neural vector search to operate strictly within the mathematically defined geographic boundaries extracted by the parser.

Finally, the final architecture layer of Paradigm 3 addressed the most significant operational bottleneck identified during the previous evaluations: the computational latency of deep learning in this last component. While deep transformer-based matchers often achieved state-of-the-art semantic precision by calculating complex token-level interactions, their quadratic complexity rendered them unscalable for high frequency enterprise querying. Therefore, there was the need to effectively compress the deep semantic reasoning of the neural network into a highly efficient statistical model. This matching layer was built on the Knowledge Distillation framework introduced by **Hinton et al.** [19], in which a large teacher model transfers its learned reasoning into a smaller and faster student. It exposes the soft probability distribution of its outputs rather than binary hard labels, in order to enable the student to learn inter-class relationships that hard labels alone can’t encode. Therefore, in this phase of our final architecture, the computationally heavy Cross-Encoder acted as a “Teacher” model. It was deployed offline to evaluate many address candidate pairs (it was used the same dataset as previous paradigms, we just added a column for each address with the model’s output) by generating a continuous probability score (the target score) for each pair that represented the absolute semantic truth. Subsequently, an XGBoost regressor was

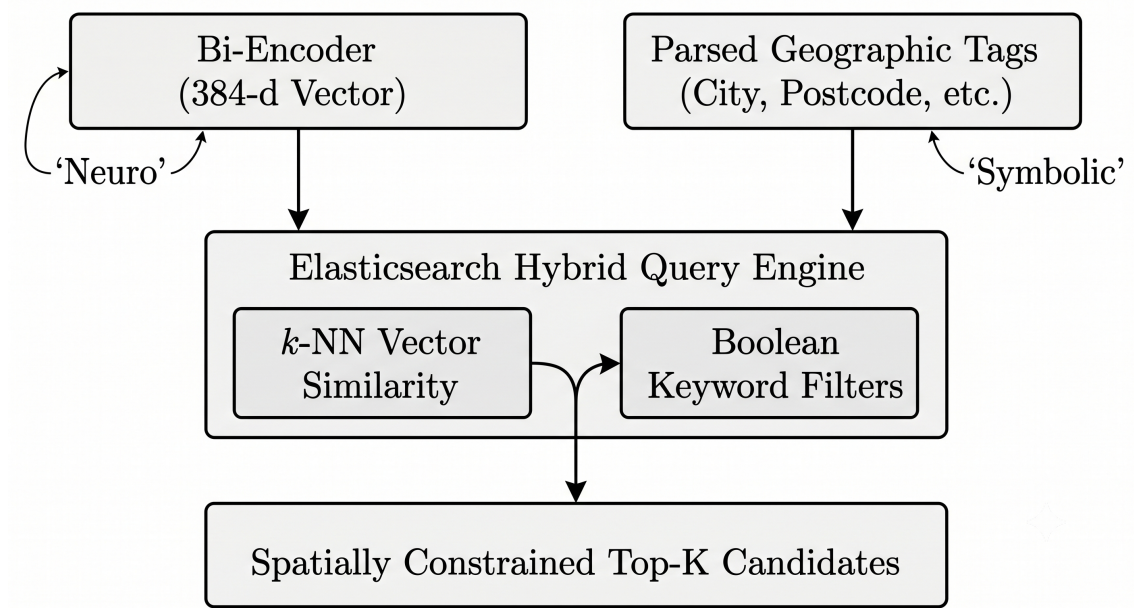


Figure 3.10: The Neuro-Symbolic retrieval architecture, demonstrating the simultaneous execution of dense vector k -NN searches and deterministic Boolean geographic filters within the Elasticsearch index.

instantiated to act as the “Student” model. Unlike the Cross-Encoder, which processed raw text tokens through massive self-attention layers, the XGBoost student was trained on a lightweight, extracted feature vector. This feature space was explicitly designed as a **Neuro-Symbolic** bridge by fusing the continuous cosine similarity score generated by the retrieval Bi-Encoder with traditional, deterministic string metrics calculated between the original query and the database candidate. Also, these symbolic features included global and component specific (like in previous experiments with machine learning models for the matching phase) Levenshtein distances, Jaccard overlaps, Jaro-Winkler similarities, and phonetic Metaphone matches. Furthermore, having been configured with a logistic regression objective, the XGBoost model was explicitly trained to predict the continuous probability output of the Teacher model, rather than attempting to predict a binary label. By evaluating the Mean Squared Error (**MSE**) and Mean Absolute Error (**MAE**) alongside early stopping mechanisms, the model learned to mathematically approximate the neural network’s complex decision boundaries. Finally, this distillation process allowed the final hybrid framework to execute pairwise candidate matching at the microsecond speeds of a tree-based classifier, while fundamentally retaining the deep semantic geographic resilience originally learned by the massive transformer.

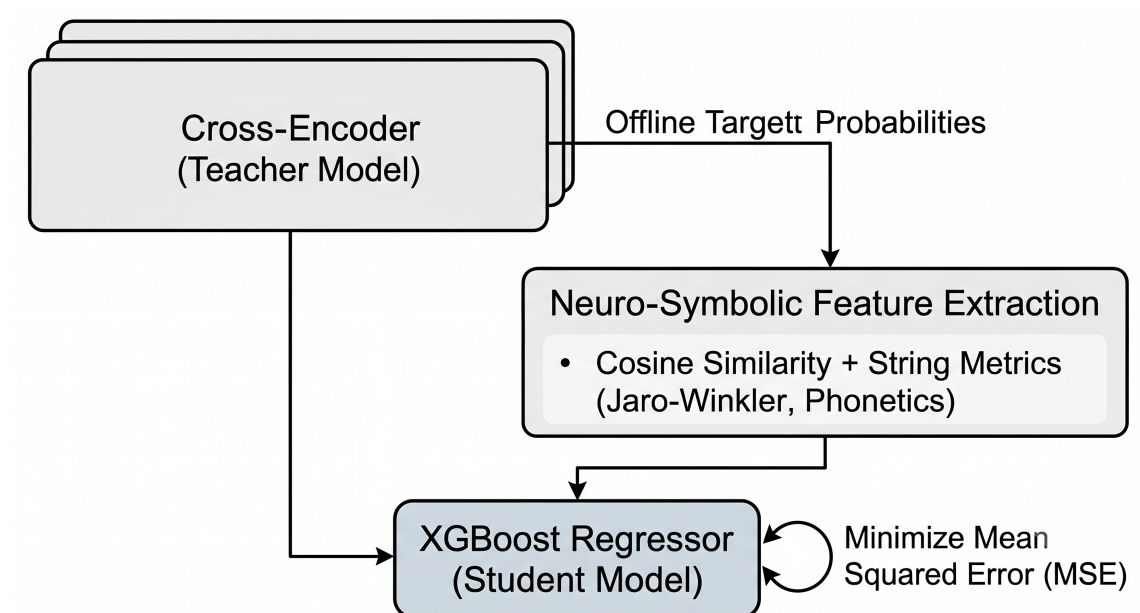


Figure 3.11: The Knowledge Distillation architecture used in the Hybrid Matcher. The heavy Cross-Encoder (Teacher) generates target probabilities offline, which are then used to train the highly efficient XGBoost Regressor (Student) using a Neuro-Symbolic feature space.

Chapter 4

Results and Discussion

This chapter presents the empirical results of the architectures and evaluation protocols defined in Chapter 3 through systematically quantifying the capacity of both statistical and deep learning models developed to resolve high-entropy geographic data. Therefore, to ensure clear progression, the findings are structured in accordance with the three-phase chronological methodology. In Section 4.1, the results of Phase 1 will be presented, in which the parsing and semantic-matching components were evaluated in absolute isolation. Subsequently, Sections 4.2 and 4.3 will transition to the multinational functional scale. Using the previously mentioned 50k evaluation dataset, these sections analyse the LLM-as-a-Judge win rates and end-to-end hardware profiling. Ultimately, these results empirically justify the transition to the final Hybrid Neuro-Symbolic Framework.

4.1 Phase 1: Isolated Component Benchmarks

4.1.1 Parsing Efficacy

The objective of the isolated parsing benchmark was to quantify the theoretical capacities of multiple sequence modelling architectures when confronted with the high-entropy reality of unstructured Portuguese geographic data. As it was established in the methodology, parsing performance was measured using Precision, Recall, and Macro-Average F1-Score to ensure that universal components (for instance, Street Names) did not mathematically obscure serious failures in other classes. In addition to that, the overall predictive scoreboard, summarizing the performance of the five evaluated architectures, is presented in Table 4.1.

A thorough analysis of these results immediately exposes the core limitations of context-blind statistical approaches. The baseline Machine Learning classifier, which predicted tags based solely on isolated lexical features, achieved severely degraded results, including a Macro F1-Score of 0.374. Besides that, without the architectural capacity to track sequence transitions, the model’s component-level precision collapsed on multi-word entities, registering an F1-Score of 0.000 on sequential sequence tags such as I-ROAD, I-CITY, and I-OTHER. Therefore, this confirms the hypothesis that geographic parsing should not be treated as isolated token classifications and that spatial entities are relational and sequential.

Table 4.1: Comprehensive parsing performance (F1-Score) across all isolated architectures. The table details both the overall sequence metrics and the specific component-level extraction capabilities.

Metric / Component	ML Baseline	Libpostal	CRF	BiLSTM	BERT
Overall Performance					
Macro F1-Score	0.374	0.829	0.991	0.811	0.993
Precision	0.354	0.833	0.991	0.845	0.993
Recall	0.468	0.866	0.992	0.803	0.993
Accuracy	0.468	0.866	0.992	0.803	0.993
Per-Component F1-Scores					
B-BUILDING_NAME	0.012	0.000	0.575	0.000	0.621
B-CITY	0.371	0.720	0.997	0.001	0.998
B-COUNTRY	0.000	0.000	0.000	0.000	0.000
B-FLOOR_INFO	0.107	0.000	0.444	0.207	0.432
B-HOUSE_NUMBER	0.971	0.961	0.992	0.922	0.994
B-OTHER	0.313	0.003	0.980	0.558	0.985
B-POSTCODE	0.999	0.989	1.000	0.909	1.000
B-PO_BOX	0.000	0.000	0.000	0.000	0.000
B-ROAD	0.424	0.965	0.998	0.970	0.998
B-STATE	0.000	0.000	0.000	0.000	0.000
I-BUILDING_NAME	0.000	0.000	0.605	0.000	0.536
I-CITY	0.000	0.564	0.998	0.000	0.998
I-FLOOR_INFO	0.000	0.000	0.600	0.222	0.316
I-HOUSE_NUMBER	0.000	0.656	0.820	0.525	0.829
I-OTHER	0.000	0.004	0.979	0.577	0.986
I-POSTCODE	0.000	0.000	0.000	0.000	0.000
I-PO_BOX	0.000	0.000	0.000	0.000	0.000
I-ROAD	0.000	0.963	0.997	0.964	0.997
O	0.283	0.106	0.646	0.055	0.751

Conversely, the introduction of linear sequence tracking via the Conditional Random Field (CRF) established itself as a remarkably strong statistical baseline by achieving an overall Macro F1-Score of 0.991. Its success was largely driven by the heavily engineered morphological feature space implemented during the data preparation phase (such as regex flags for digits, capitalization heuristics, and common Portuguese prefixes), allowing to successfully capture structured addresses. However, the benchmark also evaluated the Libpostal library that is the current industry standard for statistical geographic parsing. Although it is robust across environments, when tested in a very specific environment, Libpostal underperformed our custom CRF (yielding a Macro F1-Score of 0.829). Although Libpostal performed really well on standard components like B-ROAD (0.965), it ended up failing on rare tags such as B-BUILDING_NAME, B-FLOOR_INFO, and B-PO_BOX. This discrepancy reveals a critical finding: out-of-the-box global models can underperform when tasked with handling localized, colloquial formatting or micro-administrative schemas lacking targeted, country-specific fine-tuning.

Finally, to visually expose these performance differences across the entire tag schema and to bridge the gap between rigid morphological features and true contextual comprehension, the evaluation transitioned to deep learning architectures. The architectural divide is mapped in the Per-Component Heatmap (Figure 4.1).

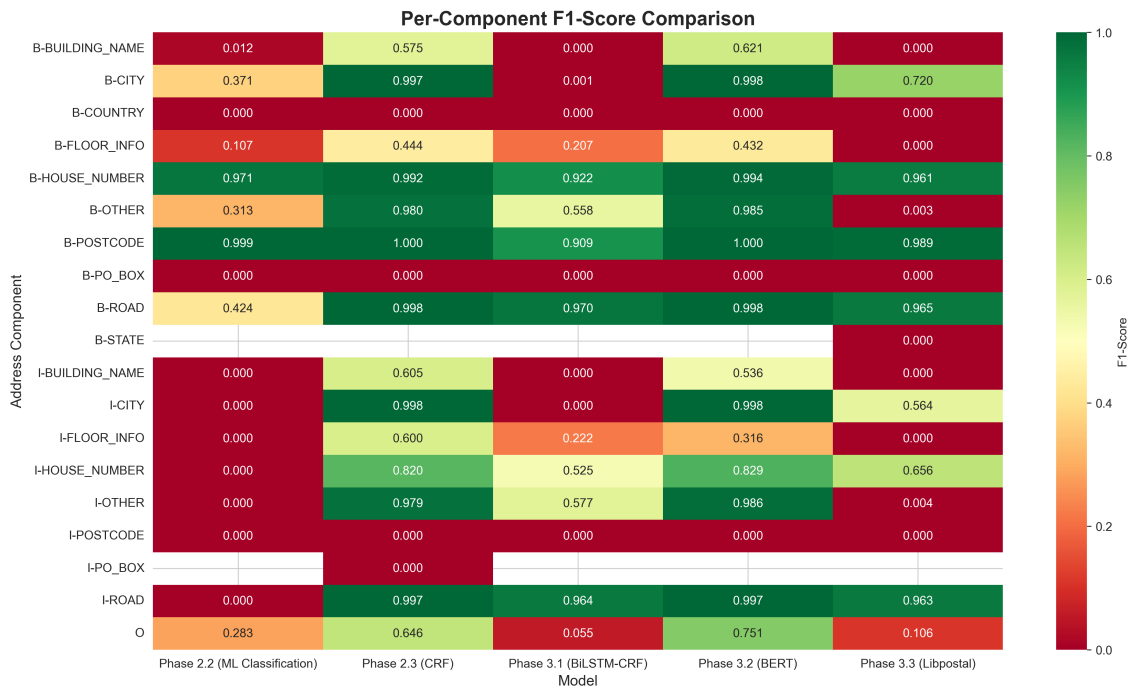


Figure 4.1: Per-component Macro F1-Score heatmap across the evaluated parsing architectures, showing the performance degradation on rare less common classes (e.g., Floor Info, Building Name).

This heatmap visually represents the inherent weakness of context-blind systems. The statistical ML Baseline is characterized by severe fragmentation, with high F1-Scores strictly isolated to standard B- tags (the start of an entity), while sequential I- tags collapse into a zero-score void.

Furthermore, while the CRF and Libpostal columns show dense clusters of high accuracy for universal geographic anchors (such like roads and postal codes), their heatmaps rapidly dissipate to absolute zero for unstructured, rare classes. Against this visual backdrop, the transformer-based BERT model emerged as the absolute superior architecture among our deep learning models (and overall as well), achieving an overall F1-Score of 0.993. This plot confirms that BERT's advantage did not stem merely from parsing standard roads but from its near-uniform "heat" distribution along highly unstructured minority components. Using its bidirectional self-attention mechanisms, BERT exhibited an exceptional capacity to generalize on highly unstructured, rare components. For example, it achieved an F1-Score of 0.621 on `B-BUILDING_NAME`, vastly outperforming Libpostal (0.000). However, the deep learning architecture also revealed a highly specific architectural vulnerability related to the BiLSTM. As starkly visible in the heatmap as a sudden performance void, it suffered a catastrophic, localized failure, leading to the complete collapse of some components, such as `CITY`. This detected anomaly is diagnosed as a fundamental limitation of traditional word-level embeddings when processing out-of-vocabulary (Out-of-Vocabulary (OOV)) regional names. Unlike BERT, which uses sub-word tokenization to mathematically reconstruct unknown words from known syllables, the BiLSTM-CRF struggled to map unseen Portuguese municipalities into its continuous vector space, defaulting to incorrect transition probabilities. Additionally, in order to translate these quantitative indicators into operational insights as well as understand when these models practically fail, a qualitative error analysis was conducted on highly chaotic real-world inputs, as detailed in Table 4.2.

This qualitative breakdown definitively illustrates the boundaries and blind spots of both probabilistic and neural parsing. In Example 1, Libpostal managed to successfully identify the severe abbreviation "V.N.G." (Vila Nova de Gaia) as a `B-CITY`, but fails completely on the floor information ("3º Dto"), erroneously classifying it as a house number. Conversely, the custom CRF correctly parses the floor information due to its engineered regex features, but entirely misses "V.N.G." (○) because the abbreviation fell outside its programmed morphological scope. Furthermore, Example 2 exposes in what structural noise fractures traditional sequence logic. Confronted with a building name and a PO Box, the CRF forces the string into a standard "Road + Number" mold, tagging "Centro Empresarial" as a `ROAD` and "1234" as a `HOUSE_NUMBER`. On the other hand, Libpostal simply ignored the components altogether (○) while BERT, using deep semantic attention, correctly understands that "Centro Empresarial Lionesa" is a `BUILDING_NAME`, but its attention window bleeds into the PO Box, misclassifying "Apartado" as part of the building. Finally, Example 3 reveals the most major limitation of self-governing parsing architectures. Confronted with a complex micro-geographic string completely lacking standard morphological anchors (like "Lugar da Bouça Freguesia de Campanhã"), every single model presented hallucinations. We could clearly observe that, while desperately trying to map the text into standard hierarchical tags, the models confidently misclassified the string as a `ROAD` or `CITY` rather than the ground-truth `OTHER`.

Therefore, these qualitative and quantitative discussions yielded a definitive conclusion for the research design. It proved that relying purely on probabilistic sequence models or neural networks

Table 4.2: Qualitative comparison of parsing architectures on highly unstructured Portuguese addresses, illustrating the limitations of statistical models on rare components and abbreviations.

Raw Token	CRF	Libpostal	BERT	Ground Truth
<i>Example 1: R. Eng. Adelino Amaro da Costa 15 3° Dto 4400-134 V.N.G.</i>				
R.	B-ROAD	B-ROAD	B-ROAD	B-ROAD
Eng.	I-ROAD	I-ROAD	I-ROAD	I-ROAD
Adelino	I-ROAD	I-ROAD	I-ROAD	I-ROAD
Amaro	I-ROAD	I-ROAD	I-ROAD	I-ROAD
da	I-ROAD	I-ROAD	I-ROAD	I-ROAD
Costa	I-ROAD	I-ROAD	I-ROAD	I-ROAD
15	B-HOUSE_NUMBER	B-HOUSE_NUMBER	B-HOUSE_NUMBER	B-HOUSE_NUMBER
3°	B-FLOOR_INFO	I-HOUSE_NUMBER	B-FLOOR_INFO	B-FLOOR_INFO
Dto	I-FLOOR_INFO	I-HOUSE_NUMBER	I-FLOOR_INFO	I-FLOOR_INFO
4400-134	B-POSTCODE	B-POSTCODE	B-POSTCODE	B-POSTCODE
V.N.G.	O	B-CITY	B-OTHER	B-CITY
<i>Example 2: Centro Empresarial Lionesa Apartado 1234 Matosinhos</i>				
Centro	B-ROAD	O	B-BUILDING_NAME	B-BUILDING_NAME
Empresarial	I-ROAD	O	I-BUILDING_NAME	I-BUILDING_NAME
Lionesa	I-ROAD	O	I-BUILDING_NAME	I-BUILDING_NAME
Apartado	I-ROAD	O	I-BUILDING_NAME	B-PO_BOX
1234	B-HOUSE_NUMBER	O	B-HOUSE_NUMBER	I-PO_BOX
Matosinhos	B-CITY	B-CITY	B-CITY	B-CITY
<i>Example 3: Lugar da Bouça Freguesia de Campanhã Porto</i>				
Lugar	B-ROAD	B-ROAD	B-ROAD	B-OTHER
da	I-ROAD	I-ROAD	I-ROAD	I-OTHER
Bouça	I-ROAD	I-ROAD	I-ROAD	I-OTHER
Freguesia	I-ROAD	B-CITY	I-ROAD	B-OTHER
de	I-ROAD	I-CITY	I-ROAD	I-OTHER
Campanhã	I-ROAD	I-CITY	I-ROAD	I-OTHER
Porto	B-CITY	O	I-ROAD	B-CITY

without deterministic safety nets can inherently lead to spatial hallucinations on edge cases. No model, regardless of its F1-Score, could flawlessly parse high-entropy inputs in absolute isolation. Therefore, this realization directly influenced the next methodological steps, demonstrating that accurate entity resolution could not be solved by parsing alone. Instead, these parsed components needed to be cross-referenced and anchored against a definitive geographic database, fundamentally justifying the transition to the end-to-end operational retrieval pipelines evaluated in Phase 2.

4.1.2 Semantic Matching and Feature Engineering

Following the evaluation of the parsing models, the focus rapidly shifted to the semantic matching components. The primary goal of this phase was to evaluate the capability of different algorithmic architectures to accurately classify whether two structurally diverse geographic strings refer to the same physical location. Therefore, this evaluation was executed, as previously mentioned, using the synthetically noised, binary classification dataset (encompassing abbreviation variations, typographical errors, and same-city hard negatives). Given this, the evaluation began by establishing a deterministic baseline using traditional mathematical string comparison algorithms, as illustrated in the performance ranking in Figure 4.2.

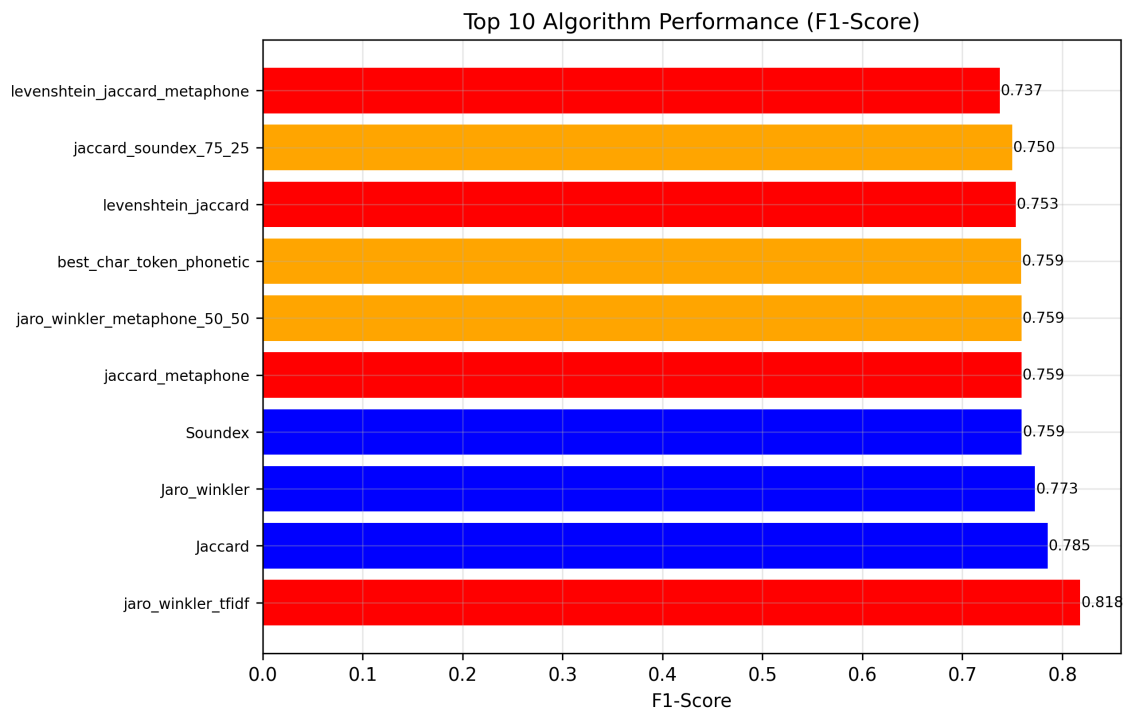


Figure 4.2: Performance ranking of traditional string comparison algorithms on the binary matching dataset. The chart illustrates the "glass ceiling" of isolated mathematical metrics when confronted with structural geographic noise.

As we can observe, algorithms such as raw Levenshtein distance and Jaro-Winkler similarity failed to surpass an F1-Score of 0.70. Although these metrics proved effective at penalizing minor

typographical errors, they still lacked semantic awareness. For example, a traditional edit-distance algorithm heavily penalizes the structural difference between "Rua" and "R.", or incorrectly assumes that two completely different streets sharing the exact same postal code and city are a match due to high character overlap. Therefore, to close this gap without immediately defaulting to deep learning, the methodology advanced to Supervised Machine Learning, deploying classification models powered by the 49-dimensional manually engineered feature space. The comparative results of these statistical models are detailed in Table 4.3.

Table 4.3: Performance comparison of Supervised Machine Learning classifiers on the binary semantic matching dataset, utilizing the 49-dimensional manually engineered feature space.

Model Architecture	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.961	0.957	0.959	0.992
Random Forest	0.992	0.996	0.994	0.999
XGBoost	0.996	0.996	0.996	0.999

The empirical outcomes demonstrated a profound leap in accuracy when transitioning from isolated string metrics to a highly dimensional feature space. Although the linear Logistic Regression model performed well (achieving a highly robust F1-score of 0.959), it struggled to capture the complex, non-linear relationships among disparate geographic features. On the other hand, the tree-based ensembles exhibited impressive results, with XGBoost [6] emerging as the definitive champion, achieving an F1-Score of 0.996. Therefore, to deconstruct the mechanics driving this exceptional performance and understand how XGBoost broke the lexical "glass ceiling", an analysis of the model's feature importance was conducted, as illustrated in Figure 4.3.

As we can see, the importance matrix offers a fundamental insight into automated geographic reasoning. The model completely bypassed global, traditional string metrics (such as a generic Levenshtein distance across the entire address) in favor of non-linear interaction features ('interact_phase2_comp', 'interact_score_mean') and highly granular, component-specific metrics ('comp_road_exact_match', 'comp_city_similarity'). Given this, it is possible to conclude that the overwhelming dominance of these features indicate that XGBoost synthesizes holistic string similarities with strict mathematical checks on individual geographic levels. For instance, the model autonomously learned that a high global similarity combined with a low 'comp_house_number_exact_match' likely represents a "hard negative" (for example, adjacent buildings on the exact same street). However, while the model attained mathematical dominance in this isolated test, it relies entirely on human-engineered feature extraction. Given that, in order to evaluate whether Deep Learning could autonomously deduce these spatial rules without specific feature engineering, the benchmark transitioned to pure semantic embeddings, as summarized in Table 4.4.

The Deep Learning evaluation revealed a severe, yet highly informative, baseline failure. The out-of-the box (zero-shot) BERT model (evaluating semantic equivalence purely via cosine similarity of contextual embeddings) suffered a catastrophic collapse in Precision (0.497). Given this, it is possible to conclude that the zero-shot model was completely fooled by geographic "hard

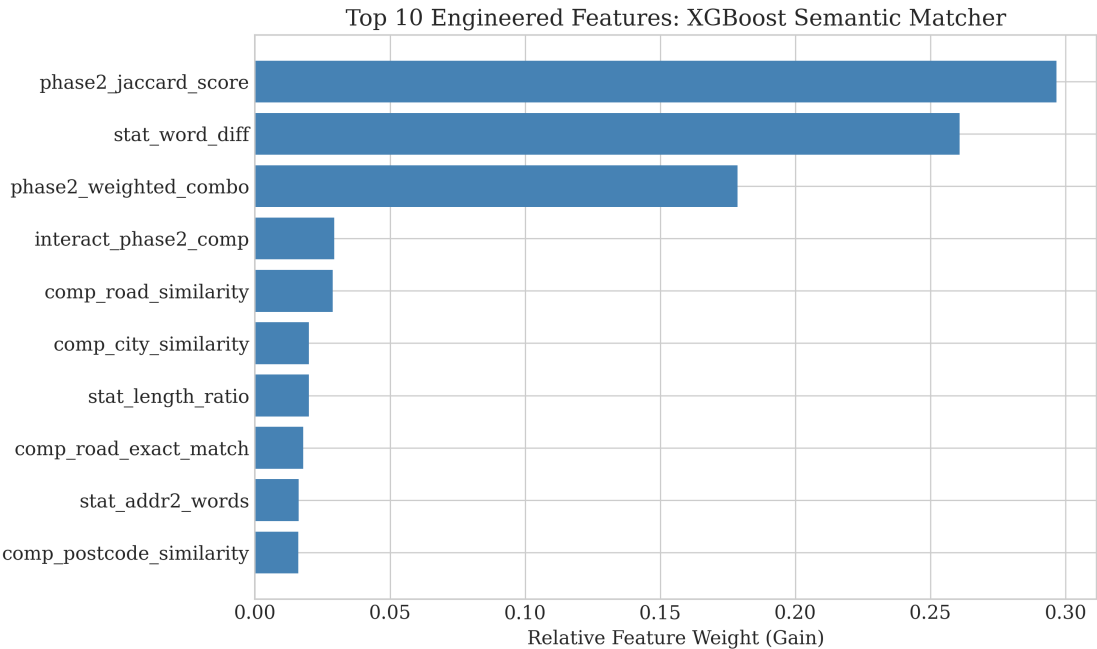


Figure 4.3: Top 10 feature importances driving the XGBoost semantic matcher. The classification is heavily dependent on interaction features and exact component-wise matches, showing a strong dependency on upstream parsing accuracy.

Table 4.4: Performance comparison of Deep Learning semantic matching architectures. The table contrasts zero-shot contextual embeddings against fine-tuned contrastive learning.

Model Architecture	Precision	Recall	F1-Score	ROC-AUC
BERT (Zero-Shot / Cosine Similarity)	0.497	0.997	0.663	0.644
Siamese Network (Contrastive Fine-Tuning)	0.953	0.988	0.970	0.995

negatives" because standard transformer embeddings group lexically similar sentences together. Therefore, when presented with two addresses sharing the same street and city but contrasting by a single house digit, pre-trained BERT interpreted them as highly semantically equivalent, resulting in massive false-positive matching. This failure explicitly validated the need for particular domain adaptation. By deploying the dual-encoder Siamese Network and fine-tuning it with a contrastive loss function, the deep learning architecture was forced to learn fine-grained spatial distinctions, achieving a highly robust F1-Score of 0.970. Furthermore, to comprehensively summarize this architecture-related divide and definitively compare the champions for both paradigms (XGBoost vs. Siamese Network), their performance covering all possible classification thresholds was measured using the Receiver Operating Characteristic (ROC) curve (Figure 4.4).

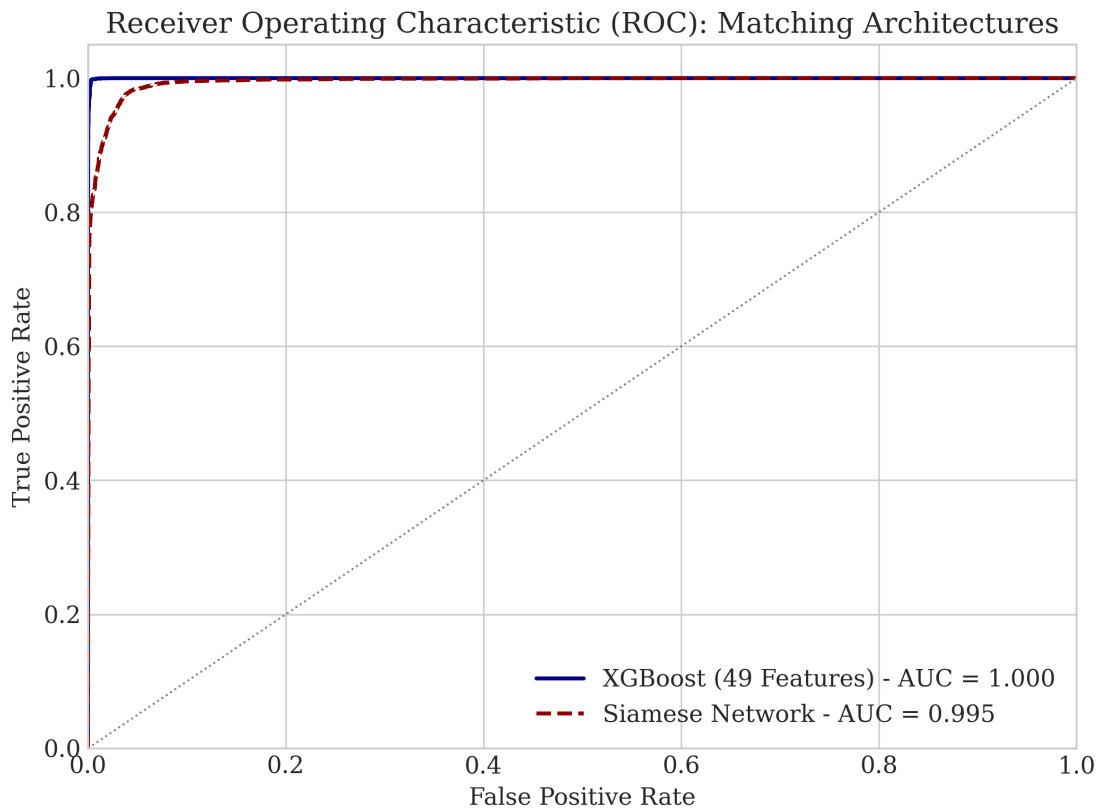


Figure 4.4: Receiver Operating Characteristic (ROC) curves comparing the 49-feature XGBoost matcher against the fine-tuned Siamese Network. The XGBoost model displays near-perfect discriminative capacity (AUC = 0.999) when supplied with cleanly parsed feature vectors.

From the previous plot, it is clear that the highly engineered XGBoost model mathematically outperformed the deep learning Siamese Network, though by a very small margin. This aligns with the findings of **Qin et al.** [45], who demonstrated that gradient boosted decision trees retain a superior inductive bias over neural rankers when applied to unstructured and tabular feature spaces. This way, it is possible to conclude that XGBoost's advantage in this tightly controlled environment is a theoretically expected output rather than an anomaly. Therefore, in a tightly

controlled environment where addresses are cleanly separated, tree-based ensembles capturing explicit geographic rules are unparalleled. However, this super-detailed analysis uncovers a severe systemic vulnerability. XGBoost's triumph is strictly conditional; it implicitly assumes that its 49-dimensional feature vector is populated correctly. Therefore, if the upstream parsing module (evaluated previously in Part A) fails to extract the street or city from a disorderly raw string, critical features like 'comp_road_exact_match' will collapse to null values, paralysing the model. Conversely, in strong contrast, the Siamese Network handles raw, unparsed strings, capturing latent spatial semantics without any dependence on explicit components, and the behavioural reality of this robust semantic understanding is visualized in the model's prediction confidence distribution (Figure 4.5).

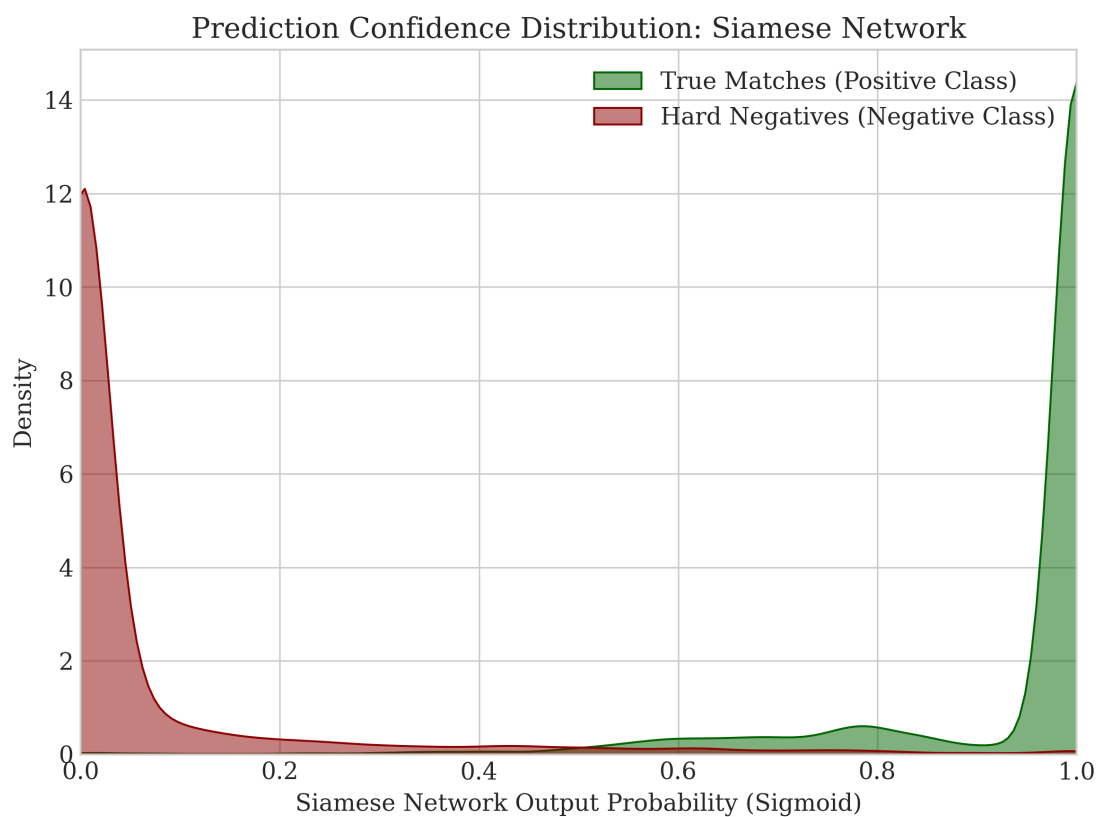


Figure 4.5: Prediction confidence distribution of the Siamese Network. The clear bifurcation between True Matches (peaking near 1.0) and Hard Negatives (peaking near 0.0) demonstrates autonomous semantic separation without reliance on pre-parsed geographic components.

The density distribution illustrates a strong bifurcation where the network autonomously forces True Matches toward a 1.0 probability and Hard Negatives toward a 0.0 probability. Unlike a tree-based model, which draws rigid mathematical boundaries, the continuous nature of the Siamese Network's vector space results in a slight density spread (explaining the marginal drop in its AUC). Therefore, since it analyses the address holistically rather than component-by-component, it can occasionally be fooled by deceptive noise. However, this complete processing can also be a great

asset, as it is theoretically immune to the cascading parsing failures that destroy the XGBoost architecture. Consequently, it is fundamentally less prone to cascading errors originating from the upstream parsing module.

Ultimately, the deep discussion of these Phase 1 results presented a serious operational dilemma. On one hand, the statistical approach (XGBoost) offers maximum precision but is structurally fragile to upstream parsing errors. Conversely, the deep semantic approach (Siamese Network) is structurally robust to missing components but slightly less crisp in its treatment of explicit geographic boundaries. This difference proved that isolated benchmarking, even though theoretically informative, is insufficient to declare an enterprise-ready solution. These exact findings established the absolute imperative for the next methodological step: moving into Phase 2, where those theoretical trade-offs were exposed to compounding errors in complete, end-to-end retrieval pipelines queried against a disorderly 25-million-record master database.

4.2 Phase 2: End-to-End Paradigm Showdown

4.2.1 Deterministic Parsing at Scale

Although the isolated-component benchmarks in Section 4.1 established the theoretical upper bounds of the models on tightly controlled data, evaluating these architectures within an operational end-to-end framework poses a significant challenge. In phase 2, all the parsers from the many Paradigms were subjected to the 50,000 multinational evaluation queries. In this initial analysis, we will focus only on the parsers of Paradigms 1 and 2. Unlike the controlled Phase 1 environment, these production-scale data contain highly unstructured syntax, cross-border formatting anomalies, and erratic token omissions. Table 4.5 presents the granular multi-country comparative breakdown of token-level F1-Scores and strict Exact Match (EM) rates for both architectures.

Table 4.5: End-to-end deterministic parsing performance comparison between Paradigm 1 (CRF) and Paradigm 2 (RoBERTa) across the 50,000 multinational evaluation queries. Metrics display token-level F1-Scores alongside strict Exact Match (EM) rates.

Geographic Scope / Tag	Paradigm 1 (CRF)	Paradigm 2 (RoBERTa)	Winner
GLOBAL DATASET			
HOUSE_NUM	F1: 0.240 (EM: 24.0%)	F1: 0.520 (EM: 38.0%)	Paradigm 2
STREET	F1: 0.166 (EM: 16.6%)	F1: 0.117 (EM: 11.4%)	Paradigm 1
SUBURB	F1: 0.925 (EM: 92.5%)	F1: 0.900 (EM: 90.0%)	Paradigm 1
DISTRICT	F1: 0.960 (EM: 96.0%)	F1: 0.957 (EM: 95.7%)	Paradigm 1
CITY	F1: 0.009 (EM: 0.9%)	F1: 0.151 (EM: 11.4%)	Paradigm 2
COUNTY	F1: 0.985 (EM: 98.5%)	F1: 0.897 (EM: 89.7%)	Paradigm 1
STATE	F1: 0.389 (EM: 38.9%)	F1: 0.389 (EM: 38.9%)	TIE

Continued on next page

Table 4.5 - continued from previous page

Geographic Scope / Tag	Paradigm 1 (CRF)	Paradigm 2 (RoBERTa)	Winner
POSTCODE	F1: 0.123 (EM: 12.3%)	F1: 0.251 (EM: 25.0%)	Paradigm 2
COUNTRY	F1: 0.000 (EM: 0.0%)	F1: 0.000 (EM: 0.0%)	TIE
UNITED STATES			
HOUSE_NUM	F1: 0.210 (EM: 21.0%)	F1: 0.505 (EM: 47.8%)	Paradigm 2
STREET	F1: 0.177 (EM: 17.7%)	F1: 0.009 (EM: 0.6%)	Paradigm 1
SUBURB	F1: 0.991 (EM: 99.1%)	F1: 0.989 (EM: 98.9%)	Paradigm 1
DISTRICT	F1: 0.995 (EM: 99.5%)	F1: 0.995 (EM: 99.5%)	TIE
CITY	F1: 0.004 (EM: 0.4%)	F1: 0.043 (EM: 3.7%)	Paradigm 2
COUNTY	F1: 0.998 (EM: 99.8%)	F1: 0.705 (EM: 70.5%)	Paradigm 1
STATE	F1: 0.034 (EM: 3.4%)	F1: 0.034 (EM: 3.4%)	TIE
POSTCODE	F1: 0.108 (EM: 10.8%)	F1: 0.328 (EM: 32.7%)	Paradigm 2
COUNTRY	F1: 0.000 (EM: 0.0%)	F1: 0.000 (EM: 0.0%)	TIE
GERMANY			
HOUSE_NUM	F1: 0.154 (EM: 15.4%)	F1: 0.769 (EM: 49.3%)	Paradigm 2
STREET	F1: 0.115 (EM: 11.5%)	F1: 0.502 (EM: 49.6%)	Paradigm 2
SUBURB	F1: 0.968 (EM: 96.8%)	F1: 0.879 (EM: 87.8%)	Paradigm 1
DISTRICT	F1: 0.977 (EM: 97.7%)	F1: 0.977 (EM: 97.7%)	TIE
CITY	F1: 0.004 (EM: 0.4%)	F1: 0.237 (EM: 14.5%)	Paradigm 2
COUNTY	F1: 0.995 (EM: 99.5%)	F1: 0.854 (EM: 85.4%)	Paradigm 1
STATE	F1: 0.394 (EM: 39.4%)	F1: 0.394 (EM: 39.4%)	TIE
POSTCODE	F1: 0.071 (EM: 7.1%)	F1: 0.143 (EM: 14.2%)	Paradigm 2
COUNTRY	F1: 0.000 (EM: 0.0%)	F1: 0.000 (EM: 0.0%)	TIE
FRANCE			
HOUSE_NUM	F1: 0.247 (EM: 24.7%)	F1: 0.632 (EM: 42.5%)	Paradigm 2
STREET	F1: 0.129 (EM: 12.9%)	F1: 0.029 (EM: 2.4%)	Paradigm 1
SUBURB	F1: 0.954 (EM: 95.4%)	F1: 0.925 (EM: 92.5%)	Paradigm 1
DISTRICT	F1: 0.950 (EM: 95.0%)	F1: 0.950 (EM: 95.0%)	TIE
CITY	F1: 0.009 (EM: 0.9%)	F1: 0.141 (EM: 9.7%)	Paradigm 2
COUNTY	F1: 0.988 (EM: 98.8%)	F1: 0.988 (EM: 98.8%)	TIE
STATE	F1: 0.490 (EM: 49.0%)	F1: 0.490 (EM: 49.0%)	TIE
POSTCODE	F1: 0.068 (EM: 6.8%)	F1: 0.179 (EM: 17.7%)	Paradigm 2

Continued on next page

Table 4.5 - continued from previous page

Geographic Scope / Tag	Paradigm 1 (CRF)	Paradigm 2 (RoBERTa)	Winner
COUNTRY	F1: 0.000 (EM: 0.0%)	F1: 0.000 (EM: 0.0%)	TIE
SPAIN			
HOUSE_NUM	F1: 0.265 (EM: 26.5%)	F1: 0.349 (EM: 26.4%)	Paradigm 2
STREET	F1: 0.183 (EM: 18.3%)	F1: 0.014 (EM: 1.0%)	Paradigm 1
SUBURB	F1: 0.912 (EM: 91.2%)	F1: 0.910 (EM: 91.0%)	Paradigm 1
DISTRICT	F1: 0.944 (EM: 94.4%)	F1: 0.944 (EM: 94.4%)	Paradigm 1
CITY	F1: 0.009 (EM: 0.9%)	F1: 0.083 (EM: 7.0%)	Paradigm 2
COUNTY	F1: 0.964 (EM: 96.4%)	F1: 0.964 (EM: 96.4%)	Paradigm 1
STATE	F1: 0.353 (EM: 35.3%)	F1: 0.353 (EM: 35.3%)	TIE
POSTCODE	F1: 0.153 (EM: 15.3%)	F1: 0.315 (EM: 31.3%)	Paradigm 2
COUNTRY	F1: 0.000 (EM: 0.0%)	F1: 0.000 (EM: 0.0%)	TIE
PORTUGAL			
HOUSE_NUM	F1: 0.323 (EM: 32.3%)	F1: 0.350 (EM: 24.1%)	Paradigm 2
STREET	F1: 0.225 (EM: 22.5%)	F1: 0.033 (EM: 3.1%)	Paradigm 1
SUBURB	F1: 0.799 (EM: 79.9%)	F1: 0.798 (EM: 79.8%)	Paradigm 1
DISTRICT	F1: 0.935 (EM: 93.5%)	F1: 0.918 (EM: 91.8%)	Paradigm 1
CITY	F1: 0.017 (EM: 1.7%)	F1: 0.252 (EM: 22.2%)	Paradigm 2
COUNTY	F1: 0.979 (EM: 97.9%)	F1: 0.979 (EM: 97.9%)	TIE
STATE	F1: 0.675 (EM: 67.5%)	F1: 0.675 (EM: 67.5%)	TIE
POSTCODE	F1: 0.214 (EM: 21.4%)	F1: 0.292 (EM: 28.9%)	Paradigm 2
COUNTRY	F1: 0.000 (EM: 0.0%)	F1: 0.000 (EM: 0.0%)	TIE

A complete assessment of these results immediately exposes systemic performance collapse when models are forced outside their initial native domain. In Phase 1, the custom CRF and transformer achieved almost perfect Macro F1-Scores exceeding 0.990 on Portuguese data. However, when exposed to the global 50,000-query dataset, the performance on high-variance micro-components, specifically `HOUSE_NUM`, `STREET`, `CITY`, and `POSTCODE`, deteriorated severely across both paradigms. The most alarming empirical discovery is the absolute collapse of the `CITY`. Globally, the CRF achieved a nearly non-existent `CITY` F1-Score of 0.009 (EM: 0.9%), while RoBERTa reached only 0.151 (EM: 11.4%). This catastrophic failure is a direct consequence of localized training bias, where the underlying sequence models were trained on localized data structures that did not explicitly teach multinational high-entropy city variations. In addition to that, this behaviour was also predicted by the concept of capacity dilution formalized by [Conneau et al. \[8\]](#). Since *xlm-roberta-base* distributed its fixed parameter budget across 100

languages simultaneously, the administrative vocabulary is also proportionally reduced, resulting in out-of-domain blindness when confronted with highly specific geographic entities, such as obscure municipal names. For example, when confronted with international municipality naming conventions, the models lacked the semantic vocabulary to map them, mapping them incorrectly or dropping them into the sequence altogether. In addition, it is observable that the `COUNTRY` component achieved an F1-Score of 0. Although this tag was present in the evaluation data, it almost did not appear in our training data. Therefore, it also demonstrates a huge need to adapt our training data in order to make our deep learning models learn these types of tags. Furthermore, despite this shared degradation, the data expose an intriguing architectural divide in how different models handle structural noise. Paradigm 2 consistently outperformed the CRF on highly variable numeric and structural anchors where RoBERTa achieved a `HOUSE_NUM` F1-Score of 0.520 and a `POSTCODE` F1-Score of 0.251, being significantly better than the CRF (0.240 and 0.123, respectively). Besides that, this trend is even more magnified in the German (DE) sub-dataset, where RoBERTa reached a great `HOUSE_NUM` F1 of 0.769 and a `STREET` F1 of 0.502, leaving the CRF far behind at 0.154 and 0.115. This phenomenon emphasizes the mathematical superiority of transformer-based sub-word tokenization when dealing with localized compounding languages, because German addresses frequently concatenate street names and suffixes. Therefore, it is possible to conclude that the rigid regex-dependent feature rules of CRF failed to find distinct boundaries in these concatenated strings, while RoBERTa’s bidirectional self-attention efficiently leveraged the surrounding context to isolate the core numeric and alphabetic transitions. On the other hand, Paradigm 1’s CRF maintained an unexpected advantage on administrative macro-components. Across the global dataset, the CRF successfully preserved high accuracy on `SUBURB` (0.925), `DISTRICT` (0.960), and `COUNTY` (0.985), consistently out-performing RoBERTa. This shows that when textual tags are relatively stable, large, and repeat standard linguistic blocks, feature-engineered conditional probabilities are remarkably resilient. Since macro-regional indicators undergo less micro-level syntax reshuffling than street lines and house numbers, the CRF’s pre-programmed token transition rules acted as a stabilizing anchor.

To conclude, this highly detailed parsing discussion yields a critical conclusion that directly shaped the trajectory of this thesis: autonomous, out-of-domain sequence parsing at a global scale is fundamentally broken. The fact that both models cratered below 15% on global city matching proved that these extracted strings could not be trusted to generate clean component-wise inputs for traditional relational databases. If a system depends on a parser to cleanly extract a city name before executing a search, the sequential failures exposed in Table 4.5 would paralyze the entire pipeline or trick it into retrieving terrible results (a completely wrong address). Therefore, this empirical breakdown provided great insight into the retrieval mechanisms (and other parsing architectures and strategies) tested in the following subsections.

4.2.2 Quantitative End-to-End Showdown (Paradigm 1 vs. Paradigm 2)

Having established the severe degradation of isolated parsing of these first 2 Paradigms when exposed to authentic global data, the evaluation proceeded to the end-to-end retrieval showdown.

The principal objective of this phase was to answer one core architectural question to whether a strictly lexical pipeline (Paradigm 1: CRF + BM25 + XGBoost) or a deep semantic pipeline (Paradigm 2: RoBERTa + Bi-Encoder + Cross-Encoder) yield superior geographic resolution. Therefore, the automated LLM-as-a-Judge protocol deterministically evaluated the top-1 retrieved record for each query in both paradigms. The global win-rate distribution is presented in Figure 4.6.

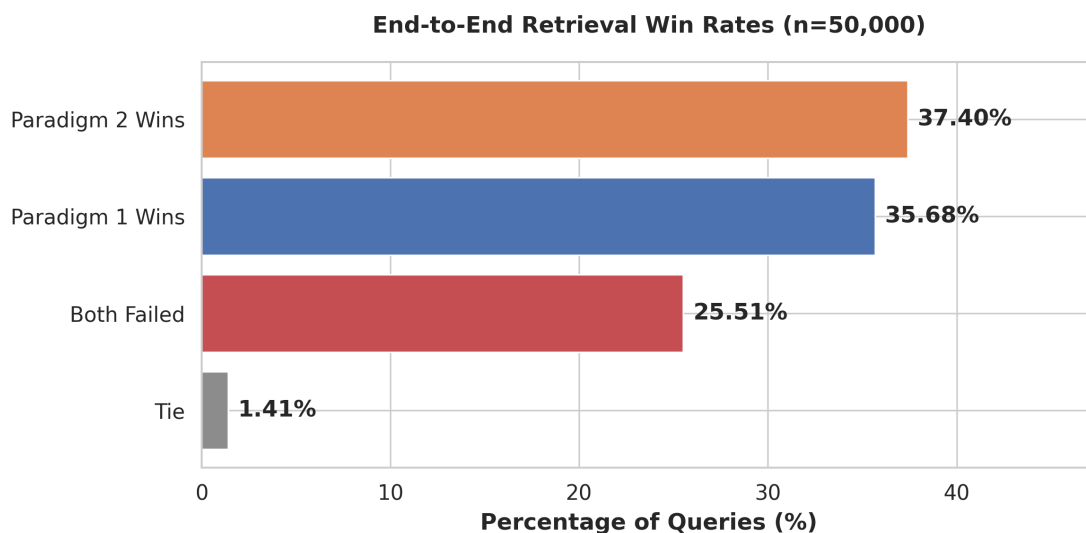


Figure 4.6: Global LLM-as-a-Judge win rates comparing Paradigm 1 (Lexical/Statistical) and Paradigm 2 (Dense Semantic) across 50,000 operational queries. The high "Both Failed" rate establishes the baseline noise floor of unstandardized geographic data.

In-depth analyses of these win rates revealed several important insights into the behavioural realities of geographic retrieval at scale. Initially, Paradigm 2 emerged as the superior architecture, winning 37.40% of the operational queries compared to Paradigm 1's 35.68%. Although the margin of victory appears relatively narrow (1.72%), the foundational mechanics of how Paradigm 2 achieved this are meaningful. As demonstrated in the previous section, the parsing modules for both paradigms suffered severe failures on global data, particularly with `CITY` and `COUNTRY` tags. However, because Paradigm 2 utilizes a Dense Vector index (Bi-Encoder), it is fundamentally free from the rigid structural boundaries of the parser. Instead, it retrieves candidates by embedding the holistic sequence into a continuous vector space, allowing it to bypass fractured syntax, severe abbreviations, and cross-lingual spelling variations that instantly broke the BM25 inverted index of Paradigm 1. On the other hand, Paradigm 1's ability to maintain a highly competitive 35.68% win rate underscores the enduring value of sparse lexical indexing. Although BM25 failed abruptly on misspelled or abbreviated tokens, it is mathematically unmatched at "exact-match" precision. For example, when a user inputs a relatively clean, unambiguous address, Paradigm 1 retrieves the exact matching string with extreme prejudice. Besides that, the deep learning paradigm's dense vector space occasionally groups two adjacent, highly similar streets or house numbers too closely together, causing the Cross-Encoder to misclassify a neighboring property as the correct target.

Ultimately, this confirms that while deep learning introduces essential fuzziness, it occasionally sacrifices the crisp, deterministic boundaries required for exact geographic pinpointing.

Furthermore, perhaps the most academically significant data point in this evaluation is the "Both Failed" classification, which accounted for a massive 25.51% of all queries. Rather than viewing this as an architectural flaw, this metric establishes the empirical "noise floor" of real-world enterprise data. More than a quarter of the evaluation queries were so deeply corrupted, truncated, or ambiguous (for example, lacking a city identifier while searching for a highly ubiquitous street name) that they lacked the mathematical entropy required to isolate a single point within a 25-million record database. Besides that, we need to consider the fact that it was expected to see a percentage this high, and it is also expected to see this percentage as high or even higher in the next experiments. This may be due to the cap we needed to apply to each country, based on the number of addresses in our database (in Elasticsearch). While creating this evaluation dataset allowed to perform these direct comparisons, it also created the problem where we are not sure if the input query we are sending to the database even exists or not. Therefore, the way we constructed our ground truth database (the logic to complete our hierarchic tree and the addresses cap) directly affected the results, specifically in the "BOTH FAILED" metric. Finally, the exceptionally low "Tie" rate (1.41%) directly informs the ultimate trajectory of this research. A tie indicates that the two paradigms retrieved the exact same valid record (or a very similar one). The fact that this occurred in less than 2% of the cases mathematically proves that lexical search and semantic search are fundamentally orthogonal systems, where they succeed and fail in entirely different data stratifications. Additionally, we can conclude that this specific quantitative realization dictated the methodological necessity of Phase 3. If Paradigm 1 and Paradigm 2 rarely overlap in their victories, it theoretically implies that combining their retrieval mechanisms into a single Hybrid search could capture the win-shares of both.

4.2.3 Domain Adaptation and Cross-Cultural Robustness

While Paradigm 2 established the baseline superiority of dense vector retrieval over sparse lexical indexing, its initial training was restricted. Therefore, in order to evaluate whether exposing these models to highly diverse multinational geographic structures could improve end-to-end retrieval, Paradigms 2.0 (RoBERTa) and 2.1 (BiLSTM + FastText) were introduced. These models went through rigorous domain adaptation via the multinational Deepparse dataset, and to quantify the impact of this adaptation, Figure 4.7 illustrates the overall paradigm shifts, while Figure 4.8 isolates the decisive win rates (excluding ties and mutual failures) in the five target countries.

An analysis of the cross-cultural baseline (the blue bars in Figure 4.8) reveals that Paradigm 2's global victory over Paradigm 1 was not geographically uniform. Paradigm 2 demonstrated overwhelming dominance in the United States (64%), Spain (53%), and Germany (52%). However, Paradigm 1 achieved a distinct advantage in Portugal (where P2 only achieves a 43% decisive win rate) and France (44%). This difference indicates that Paradigm 1's highly engineered CRF features were heavily biased towards Portuguese morphological rules during development, making it highly effective locally. However, rigid morphological rules lack the semantic elasticity required

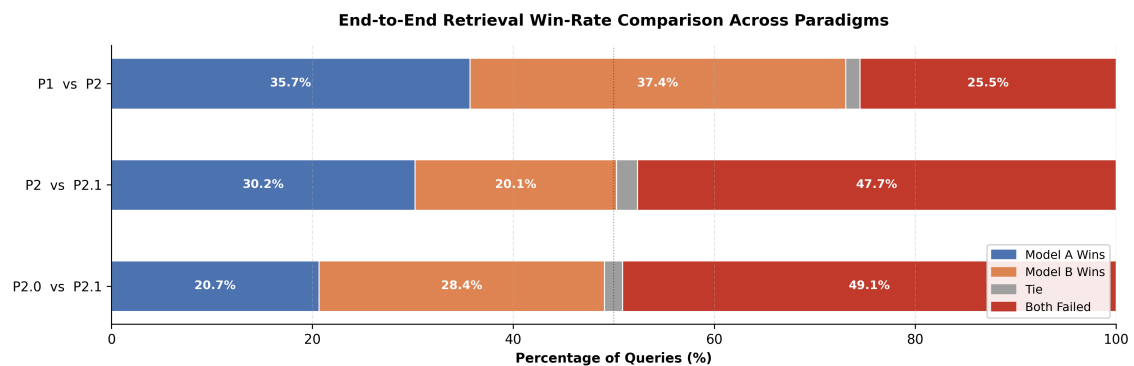


Figure 4.7: End-to-end global retrieval win rates across successive paradigm evaluations. The transition to domain-adapted models (P2.1 and P2.0) causes a regression, nearly doubling the "Both Failed" rate.

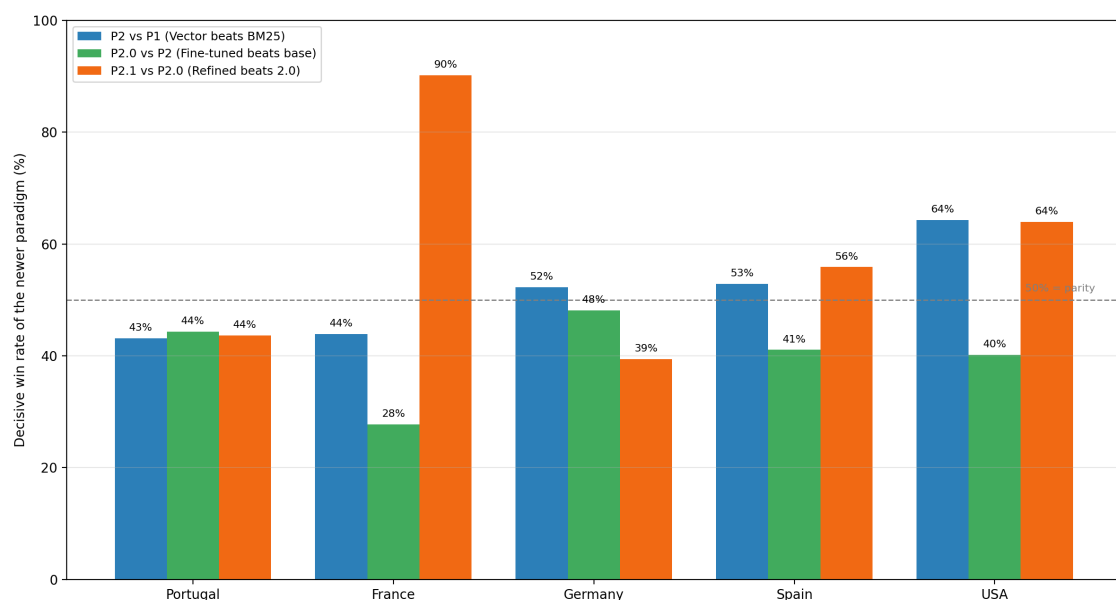


Figure 4.8: Decisive cross-cultural win rates of the newer paradigm in head-to-head match-ups. The 50% parity line indicates where the newer architecture mathematically outperforms the older baseline.

to generalize to foreign addressing schemas, allowing the Dense Vector space of Paradigm 2 to easily dominate in out-of-domain countries like the USA and Germany. Therefore, in order to artificially inject this cross-cultural elasticity, Paradigm 2.0 fine-tuned the RoBERTa pipeline on the multinational Deepparse dataset. However, the empirical results that came from this modification revealed results contrary to what we were expecting. Besides that, as it is possible to observe in the global stacked chart (Figure 4.7), transitioning from the base P2 model to the fine-tuned pipelines caused the overall "Both Failed" rate to nearly double, rocketing from 25.5% to almost 50%. Additionally, the green bars in Figure 4.8 confirm the severity of this regression: the fine-tuned Paradigm 2.0 decisively lost to the base Paradigm 2 across *every single country*, collapsing to a dismal 28% win rate in France and 40% in the USA. Additionally, in Paradigms 2.0 and 2.1 we also performed a slight change in how we formulated the query. In paradigm 2 we were performing the query with the raw input, while in the next two paradigms we were utilizing the parsed tags to construct our input query. Therefore, the performance of our parsers might have affected some of our end-to-end results and made the pipeline underperform instead of improving. Also, we also have to be aware of possible LLM hallucinations and errors along this process.

Despite this overall pipeline regression, a head-to-head comparison between the two domain-adapted architectures revealed a considerable insight into model weights and linguistic robustness. Our lightweight BiLSTM + FastText architecture (the Deepparse architecture from Paradigm 2.1) decisively defeated the heavy RoBERTa transformer. Globally, Paradigm 2.1 won 28.4% of the queries compared to Paradigm 2.0's 20.7%, and the orange bars in Figure 4.8 highlight this advantage, with the lightweight model achieving an astonishing 90% decisive win rate in France and 64% in the USA. This unexpected victory demonstrates the morphological resilience of FastText's sub-word n -gram embeddings [5]. Although standard transformer tokenizers (like RoBERTa's WordPiece) can struggle when encountering heavily abbreviated, out-of-vocabulary foreign street names, FastText reconstructs meaning by breaking words into character n -grams. Consequently, it handles the high-entropy structural noise of French and American addresses with far greater agility than the rigid embedding layers of a heavy transformer.

Table 4.6: End-to-end deterministic parsing performance comparison between Paradigm 2.0 (RoBERTa) and Paradigm 2.1 (Bi-LSTM) across evaluation queries. Metrics display token-level F1-Scores alongside strict Exact Match (EM) rates.

Geographic Scope / Tag	Paradigm 2.0 (RoBERTa)	Paradigm 2.1 (Bi-LSTM)	Winner
GLOBAL DATASET			
STREET_NUM	F1: 0.863 (EM: 84.4%)	F1: 0.883 (EM: 84.0%)	Paradigm 2.1
STREET_NAME	F1: 0.271 (EM: 26.8%)	F1: 0.863 (EM: 66.8%)	Paradigm 2.1
UNIT	F1: 0.925 (EM: 92.5%)	F1: 0.944 (EM: 94.4%)	Paradigm 2.1
SUBURB	F1: 0.893 (EM: 89.2%)	F1: 0.925 (EM: 92.5%)	Paradigm 2.1

Continued on next page

Table 4.6 - continued from previous page

Geographic Scope / Tag	Paradigm 2.0 (RoBERTa)	Paradigm 2.1 (Bi-LSTM)	Winner
DISTRICT	F1: 0.935 (EM: 93.5%)	F1: 0.960 (EM: 96.0%)	Paradigm 2.1
POSTCODE	F1: 0.985 (EM: 97.8%)	F1: 0.753 (EM: 59.9%)	Paradigm 2.0
MUNICIPALITY	F1: 0.657 (EM: 63.5%)	F1: 0.639 (EM: 34.4%)	Paradigm 2.0
PROVINCE	F1: 0.663 (EM: 66.2%)	F1: 0.145 (EM: 10.1%)	Paradigm 2.0
COUNTY	F1: 0.933 (EM: 93.3%)	F1: 0.985 (EM: 98.5%)	Paradigm 2.1
COUNTRY	F1: 0.990 (EM: 98.0%)	F1: 0.000 (EM: 0.0%)	Paradigm 2.0
UNITED STATES			
STREET_NUM	F1: 0.965 (EM: 94.9%)	F1: 0.964 (EM: 94.7%)	Paradigm 2.0
STREET_NAME	F1: 0.184 (EM: 18.3%)	F1: 0.941 (EM: 82.2%)	Paradigm 2.1
UNIT	F1: 0.815 (EM: 81.5%)	F1: 0.871 (EM: 87.1%)	Paradigm 2.1
SUBURB	F1: 0.967 (EM: 96.6%)	F1: 0.991 (EM: 99.1%)	Paradigm 2.1
DISTRICT	F1: 0.986 (EM: 98.6%)	F1: 0.995 (EM: 99.5%)	Paradigm 2.1
POSTCODE	F1: 0.997 (EM: 99.6%)	F1: 0.626 (EM: 55.0%)	Paradigm 2.0
MUNICIPALITY	F1: 0.674 (EM: 67.1%)	F1: 0.640 (EM: 10.3%)	Paradigm 2.0
PROVINCE	F1: 0.881 (EM: 87.9%)	F1: 0.077 (EM: 2.8%)	Paradigm 2.0
COUNTY	F1: 0.996 (EM: 99.6%)	F1: 0.998 (EM: 99.8%)	Paradigm 2.1
COUNTRY	F1: 0.999 (EM: 99.9%)	F1: 0.000 (EM: 0.0%)	Paradigm 2.0
GERMANY			
STREET_NUM	F1: 0.974 (EM: 96.9%)	F1: 0.970 (EM: 96.2%)	Paradigm 2.0
STREET_NAME	F1: 0.643 (EM: 63.9%)	F1: 0.869 (EM: 78.2%)	Paradigm 2.1
UNIT	F1: 0.993 (EM: 99.3%)	F1: 0.982 (EM: 98.2%)	Paradigm 2.0
SUBURB	F1: 0.947 (EM: 94.7%)	F1: 0.968 (EM: 96.8%)	Paradigm 2.1
DISTRICT	F1: 0.969 (EM: 96.9%)	F1: 0.977 (EM: 97.7%)	Paradigm 2.1
POSTCODE	F1: 0.996 (EM: 99.4%)	F1: 0.796 (EM: 56.6%)	Paradigm 2.0
MUNICIPALITY	F1: 0.844 (EM: 82.0%)	F1: 0.628 (EM: 33.1%)	Paradigm 2.0
PROVINCE	F1: 0.696 (EM: 69.5%)	F1: 0.453 (EM: 41.1%)	Paradigm 2.0
COUNTY	F1: 0.989 (EM: 98.9%)	F1: 0.995 (EM: 99.5%)	Paradigm 2.1
COUNTRY	F1: 0.996 (EM: 98.8%)	F1: 0.000 (EM: 0.0%)	Paradigm 2.0
FRANCE			
STREET_NUM	F1: 0.949 (EM: 94.0%)	F1: 0.959 (EM: 94.9%)	Paradigm 2.1
STREET_NAME	F1: 0.126 (EM: 12.1%)	F1: 0.883 (EM: 61.9%)	Paradigm 2.1

Continued on next page

Table 4.6 - continued from previous page

Geographic Scope / Tag	Paradigm 2.0 (RoBERTa)	Paradigm 2.1 (Bi-LSTM)	Winner
UNIT	F1: 0.991 (EM: 99.1%)	F1: 0.973 (EM: 97.3%)	Paradigm 2.0
SUBURB	F1: 0.918 (EM: 91.7%)	F1: 0.954 (EM: 95.4%)	Paradigm 2.1
DISTRICT	F1: 0.947 (EM: 94.7%)	F1: 0.950 (EM: 95.0%)	Paradigm 2.1
POSTCODE	F1: 0.994 (EM: 98.9%)	F1: 0.849 (EM: 68.0%)	Paradigm 2.0
MUNICIPALITY	F1: 0.679 (EM: 66.3%)	F1: 0.624 (EM: 41.3%)	Paradigm 2.0
PROVINCE	F1: 0.596 (EM: 59.3%)	F1: 0.035 (EM: 0.8%)	Paradigm 2.0
COUNTY	F1: 0.959 (EM: 95.9%)	F1: 0.988 (EM: 98.8%)	Paradigm 2.1
COUNTRY	F1: 0.993 (EM: 98.3%)	F1: 0.000 (EM: 0.0%)	Paradigm 2.0
SPAIN			
STREET_NUM	F1: 0.781 (EM: 73.7%)	F1: 0.781 (EM: 70.9%)	Paradigm 2.0
STREET_NAME	F1: 0.196 (EM: 19.3%)	F1: 0.793 (EM: 56.2%)	Paradigm 2.1
UNIT	F1: 0.881 (EM: 88.1%)	F1: 0.912 (EM: 91.2%)	Paradigm 2.1
SUBURB	F1: 0.855 (EM: 85.4%)	F1: 0.912 (EM: 91.2%)	Paradigm 2.1
DISTRICT	F1: 0.900 (EM: 89.9%)	F1: 0.944 (EM: 94.4%)	Paradigm 2.1
POSTCODE	F1: 0.988 (EM: 97.5%)	F1: 0.787 (EM: 58.1%)	Paradigm 2.0
MUNICIPALITY	F1: 0.563 (EM: 51.1%)	F1: 0.608 (EM: 42.4%)	Paradigm 2.1
PROVINCE	F1: 0.460 (EM: 45.6%)	F1: 0.102 (EM: 3.8%)	Paradigm 2.0
COUNTY	F1: 0.861 (EM: 86.1%)	F1: 0.964 (EM: 96.4%)	Paradigm 2.1
COUNTRY	F1: 0.995 (EM: 98.7%)	F1: 0.000 (EM: 0.0%)	Paradigm 2.0
PORTUGAL			
STREET_NUM	F1: 0.650 (EM: 62.9%)	F1: 0.743 (EM: 63.5%)	Paradigm 2.1
STREET_NAME	F1: 0.207 (EM: 20.5%)	F1: 0.831 (EM: 55.9%)	Paradigm 2.1
UNIT	F1: 0.944 (EM: 94.4%)	F1: 0.985 (EM: 98.5%)	Paradigm 2.1
SUBURB	F1: 0.778 (EM: 77.7%)	F1: 0.799 (EM: 79.9%)	Paradigm 2.1
DISTRICT	F1: 0.875 (EM: 87.5%)	F1: 0.935 (EM: 93.5%)	Paradigm 2.1
POSTCODE	F1: 0.955 (EM: 94.3%)	F1: 0.713 (EM: 62.2%)	Paradigm 2.0
MUNICIPALITY	F1: 0.528 (EM: 51.2%)	F1: 0.699 (EM: 45.0%)	Paradigm 2.1
PROVINCE	F1: 0.688 (EM: 68.8%)	F1: 0.056 (EM: 2.2%)	Paradigm 2.0
COUNTY	F1: 0.859 (EM: 85.9%)	F1: 0.979 (EM: 97.9%)	Paradigm 2.1
COUNTRY	F1: 0.980 (EM: 96.2%)	F1: 0.000 (EM: 0.0%)	Paradigm 2.0

In addition to what we have already seen, it was needed to empirically validate what architectures also performed better in the parsing component (which also had an impact in the end-to-end

results as we have previously mentioned). The results, detailed in Table 4.6, expose a profound architectural dichotomy between micro-level street accuracy and macro-level geographic context.

The data confirms the hypothesis regarding sub-word embeddings, where Paradigm 2.1 (Bi-LSTM + FastText) achieved absolute dominance on micro-geography. Globally, it parsed the `StreetName` component with an F1-Score of 0.863, significantly outperforming RoBERTa's devastatingly low 0.271. Furthermore, this catastrophic failure by RoBERTa is heavily localized in Romance languages, where in France its `StreetName` Exact Match collapsed to a mere 12.1%, compared to FastText's 61.9%. Also, standard transformer tokenizers can aggressively shatter unknown or hyphenated foreign street names into meaningless fragments, destroying the sequence logic; while FastText, conversely, bypasses vocabulary limitations by mathematically interpolating character n -grams, successfully grouping chaotic street strings. However, the table also reveals the critical vulnerability of recurrent neural networks (RNNs). Although Paradigm 2.1 excelled at the beginning of the address string, its performance disintegrated at the end of the sequence. It registered an 0.000 F1-Score on the `Country` tag globally; this was due to the fact that the Deepparse wasn't trained to identify this tag. Besides that, it struggled immensely with `Province` achieving an F1-Score of 0.145. On the other hand, RoBERTa achieved a near-perfect `Country` F1 of 0.990 and a `PostalCode` F1 of 0.985. This perfectly illustrates the "forgetting problem" inherent to LSTMs: as the recurrent network processes a long, 15-token address, the hidden state degrades, causing it to lose context by the time it reaches the final macro-geographic tags. Conversely, RoBERTa, empowered by global self-attention mechanisms, views the entire sequence simultaneously, effortlessly identifying macro-anchors regardless of their position in the string. Furthermore, this parsing dichotomy perfectly explains the end-to-end retrieval results observed in Figure 4.8. Even though RoBERTa (Paradigm 2.0) was far superior at identifying Countries and Postal Codes, it decisively lost the retrieval showdown to the Bi-LSTM, where, in a 25-million record database, macro-geography provides very low discriminative entropy. Therefore, knowing the Country or PostalCode narrows the search space, but failing to accurately extract the `StreetName` makes it mathematically impossible to isolate the correct building.

Ultimately, these combined parsing and retrieval insights solidify the fundamental conclusion of Phase 2. While Paradigm 2 demonstrated superior end-to-end retrieval (validating the OSM training data), Paradigms 2.0 and 2.1 suffered due to a more complex, parsed-based querying methodology, despite benefiting from the higher data variability of the Deepparse dataset. Therefore, we decided to train our models for our next final solution using a combination between the OSM and Deepparse data in order to create the best dataset possible for our final framework. Additionally, we also observed that the Deepparse architecture (even without any fine-tuning) appeared to be the clear winner, therefore, we decided to fine-tune this architecture for our final solution in order to adapt it to our restricted 7-tag schema and to start identifying the `COUNTRY` tag.

4.2.4 Qualitative Audit of Retrieval Outputs

While the quantitative analysis gave us great insight, a qualitative analysis is also important in order to obtain more information specifically about where these systems ended up failing. Therefore,

Table 4.7: Qualitative examples of the LLM-as-a-Judge evaluations, illustrating how dense vector retrieval (Paradigms 2, 2.0 and 2.1) successfully handles severe structural noise that paralyzed the sparse lexical index (Paradigm 1), which retrieves identically-spelled locations on the wrong continent.

Raw User Query	Paradigm 1 Retrieved	Paradigm 2 Retrieved	Paradigm 2.0 Retrieved	Paradigm 2.1 Retrieved
Schlachthofstraße 4 Weimar, Thuringia 99423, DE	78962-3303 715, 78962 weimar, united states, tex, mitchell county, colorado city	schillerstraße 4a, 56075 koblenz, germany, rhineland palatinate	puschkinstraße 4, 98527 suhl, germany, thuringia	riedstraße 4, 99988 südeichsfeld, ger- many, thuringia, unstrut hainich kreis
Esmarchstrasse 51 Kiel, 24105, DE	dry hollow road 3854 3854, 53812 kieler, united states, state of wisconsin, portage county, town of grant	hauptstraße 90a, 65760 eschborn, ger- many, he, regierungs- bezirk darmstadt, main taunus kreis, eschborn	auf der schwann 102, 66953 pirmasens, germany, rhineland palatinate	ziegelstraße 51, 23795 bad segeberg, germany, sh, kreis segeberg, trave land, stipsdorf
Cranachstr. 43 Berlin, Berlin 12157, DE	cranston 2920-5539, 2920-5539 providence, united states, rhode island and providence plantations, providence, cranston	schichauweg 56, 12307 berlin, ger- many, berlin	württembergische str. 42-43, 10707 berlin, germany, berlin	berlewitzweg 43, 12557 berlin, ger- many, berlin

Table 4.7 extracts three highly chaotic queries from the German (DE) evaluation subset, comparing the top-1 retrieved record from the statistical baseline (Paradigm 1) against the semantic vector spaces of Paradigms 2, 2.0, and 2.1. An analysis of these examples exposes some major architectural vulnerabilities inherent to sparse lexical indexing. In all three examples, Paradigm 1 (BM25) confidently retrieves an identically spelled or lexically similar location located on an entirely different continent, completely ignoring the explicitly provided country code ("DE"). For example, in the first query targeting Weimar, Germany, Paradigm 1 finds a location in Weimar, Texas, while in the second query targeting Kiel, Germany, it found "Kieler" in Wisconsin. Meanwhile, in order to understand why this occurs, one must deconstruct the mathematics of Term Frequency-Inverse Document Frequency (TF-IDF) scoring underlying BM25. Sparse indexes have zero intrinsic awareness of physical space, national borders, or geographic hierarchies, and calculate relevance purely on the basis of token rarity. Therefore, since tokens like "Weimar" or "Kiel" are mathematically rare within our database, BM25 assigns them a massive weight. Conversely, standard tokens like "DE", or "Germany" are highly frequent and mathematically penalized as "stopwords".

In profound contrast, deep semantic pipelines (Paradigms 2, 2.0, and 2.1) demonstrated robust latent spatial awareness. Even when the user query is noisy, every vector-based paradigm successfully restricted its top-1 retrieval to the correct country, and sometimes the city as well. This behavior proves the theoretical advantage of contrastive learning in geographic retrieval, where, during training, the models inherently learned that tokens like "Schlachthofstraße," "Thuringia," and "DE" heavily co-occur in the same spatial dimension. Therefore, the resulting dense vector represents a holistic geographic cluster, even if the dense index cannot find the exact street, it

drops the "pin" in the correct national or regional cluster, completely avoiding the intercontinental failures of the lexical index. Additionally, the finite capacity of the database directly impacted the absolute retrieval limits, further inflating the 'Both Failed' metric. Furthermore, these qualitative examples provided early visual evidence of the impact of domain adaptation, which will be quantitatively dissected in the next section. Although the baseline RoBERTa model (Paradigm 2) successfully mapped the query to Germany, it frequently missed the micro-region (such as retrieving Rhineland-Palatinate instead of Thuringia for Query 1). In contrast, Paradigms 2.0 and 2.1, which were fine-tuned globally on the Deepparse dataset, demonstrated a tighter geographic grouping. As you can see in Query 3, targeting "12157 Berlin", both Paradigm 2.0 and 2.1 successfully anchored their predictions strictly within Berlin, proving that exposure to diverse multinational training data significantly tightened the geographic clustering of their vector spaces. However, this qualitative audit also revealed the ultimate limitation of pure dense vector search. Although Paradigms 2, 2.0, and 2.1 are geographically safe and often precise, they can also be inaccurate at the street level. For example, in Query 1, instead of retrieving "Schlachthofstraße," the dense models retrieve "Schillerstraße" or "Puschkinstraße". Since transformer models analyze semantic similarity rather than exact character matching, they can often retrieve roads that sound structurally similar or share geographic proximity, failing to enforce the strict deterministic character matching that users require for physical navigation.

4.2.5 The Operational Bottleneck and Hardware Profiling

Throughout Phase 2, the evaluation has been highly focused on precision, geographic robustness, and adaptation to the cross-cultural domain. However, assessing an architecture for enterprise deployment requires evaluating its computational efficiency under load. Therefore, to quantify the operational viability of these pipelines, a detailed hardware latency profile was recorded for the 50,000 queries, where Figure 4.9 illustrates the mean execution time (in milliseconds) per query, decomposed into the three sequential architectural stages: Parsing, Retrieval, and Semantic Matching.

An initial surface-level observation of Figure 4.9 presents a seemingly contradictory result: the highly complex, multi-million parameter deep semantic pipelines (Paradigms 2, 2.0, and 2.1) executed queries slightly faster (137 ms) than the lightweight statistical baseline of Paradigm 1 (153 ms). However, a rigorous discussion of this metric requires contextualizing the specific hardware environments in which these models were executed. As outlined in the methodology, the deep learning pipelines necessitated the use of dedicated, high-performance GPU hardware to process the heavy transformer embeddings (RoBERTa and the Cross-Encoder). On the other hand, the statistical baseline (Paradigm 1) was deliberately executed on a standard consumer-grade CPU laptop to test localized low-resource deployment constraints. "Viewed through this hardware context, the data reveals a severe computational bottleneck inherent to deep semantic retrieval. Despite operating on vastly superior parallelized GPU compute, Paradigm 2 still required 137 ms per query. Deconstructing this timeline exposes the exact architectural limitation. Although vector parsing (24 ms) was relatively fast, the dense bi-encoder retrieval (84 ms) and the final semantic

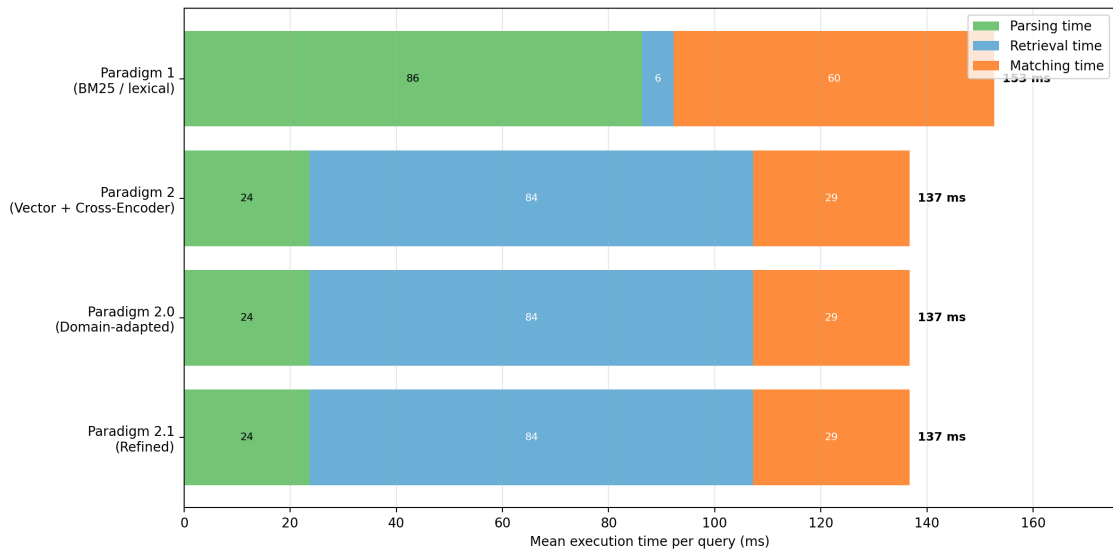


Figure 4.9: End-to-end latency breakdown by paradigm (median execution time per query). The chart illustrates the time distribution across the Parsing, Retrieval, and Matching modules.

matching stage (which relies on the heavy Cross-Encoder to re-rank the retrieved candidates) were the ones which took longer. Even on a GPU, processing a single batch of retrieved candidates through a Cross-Encoder requires calculating full self-attention across every query-candidate pair, establishing a hard, unyielding computational floor. Comparing this to Paradigm 1 exposes the sheer efficiency of statistical models. Despite running on a vastly inferior, single-threaded laptop CPU, it shows that calculating lots of metrics for our statistical models (Parsing and Matching phases) can take time and delay the Paradigm’s end-to-end output. Furthermore, the Elasticsearch BM25 retrieval was executed in a mere 6 ms, mathematically proving that inverted index lookups are orders of magnitude faster than nearest-neighbor dense vector searches (84 ms) when hardware constraints are equalized.

In conclusion, these latency metrics dictate the final architectural requirements. While deep learning models offer superior semantic resolution, they are operationally restrictive without dedicated hardware. Conversely, statistical models provide high-speed efficiency but require rigorous upstream feature engineering. Therefore, to synthesize the strengths of both paradigms while remaining well within acceptable latency boundaries, the final proposed framework (Phase 3) will decouple the architecture by using a high-speed XGBoost model for semantic matching, while explicitly injecting parsed components directly into the retrieval query to maintain deterministic geographic boundaries.

4.3 Phase 3: The Neuro-Symbolic Framework

4.3.1 Schema Adaptation and the Hybrid Retrieval Trade-off

The severe retrieval regressions observed in Phase 2 demonstrated that isolated domain adaptation is fundamentally counterproductive if the parser’s output schema misaligns with the downstream database index. Therefore, to try to rectify the 50% mutual failure rate caused by the 10-tag Deepparse schema, this final phase initiated with deterministic schema adaptation. The highest-performing micro-geography parser from Phase 2 (the Bi-LSTM + FastText architecture) was re-trained to explicitly collapse the granular multinational tags into the rigid 7-tag database structure (mapping Municipality, District, and Suburb into CITY; and Province and County into STATE). Table 4.8 presents the granular parsing performance of this newly fine-tuned Paradigm 3 model across the 50,000-query evaluation set.

Table 4.8: End-to-end deterministic parsing performance for Paradigm 3 (Fine-Tuned Bi-LSTM) across evaluation queries. Metrics display token-level F1-Scores alongside strict Exact Match (EM) rates.

Geographic Scope / Tag Paradigm 3 (Fine-Tuned Bi-LSTM)	
GLOBAL DATASET	
STREET_NUM	F1: 0.890 (EM: 85.3%)
STREET_NAME	F1: 0.894 (EM: 70.5%)
UNIT	F1: 0.857 (EM: 83.2%)
CITY	F1: 0.633 (EM: 23.0%)
STATE	F1: 0.468 (EM: 38.5%)
POSTCODE	F1: 0.642 (EM: 60.9%)
COUNTRY	F1: 0.898 (EM: 88.7%)
UNITED STATES	
STREET_NUM	F1: 0.975 (EM: 97.4%)
STREET_NAME	F1: 0.939 (EM: 78.1%)
UNIT	F1: 0.846 (EM: 80.6%)
CITY	F1: 0.716 (EM: 21.2%)
STATE	F1: 0.290 (EM: 4.6%)
POSTCODE	F1: 0.172 (EM: 17.2%)
COUNTRY	F1: 0.608 (EM: 60.7%)
GERMANY	
STREET_NUM	F1: 0.948 (EM: 94.0%)

Continued on next page

Table 4.8 - continued from previous page

Geographic Scope / Tag	Paradigm 3 (Fine-Tuned Bi-LSTM)
STREET_NAME	F1: 0.899 (EM: 79.4%)
UNIT	F1: 0.953 (EM: 95.2%)
CITY	F1: 0.622 (EM: 23.6%)
STATE	F1: 0.488 (EM: 43.1%)
POSTCODE	F1: 0.785 (EM: 77.7%)
COUNTRY	F1: 0.925 (EM: 92.5%)
FRANCE	
STREET_NUM	F1: 0.958 (EM: 94.2%)
STREET_NAME	F1: 0.849 (EM: 54.4%)
UNIT	F1: 0.936 (EM: 93.2%)
CITY	F1: 0.494 (EM: 22.7%)
STATE	F1: 0.514 (EM: 49.1%)
POSTCODE	F1: 0.862 (EM: 83.8%)
COUNTRY	F1: 0.995 (EM: 99.3%)
SPAIN	
STREET_NUM	F1: 0.816 (EM: 74.3%)
STREET_NAME	F1: 0.890 (EM: 73.9%)
UNIT	F1: 0.805 (EM: 76.7%)
CITY	F1: 0.644 (EM: 21.9%)
STATE	F1: 0.417 (EM: 34.7%)
POSTCODE	F1: 0.758 (EM: 64.5%)
COUNTRY	F1: 0.977 (EM: 94.3%)
PORTUGAL	
STREET_NUM	F1: 0.755 (EM: 66.9%)
STREET_NAME	F1: 0.897 (EM: 66.8%)
UNIT	F1: 0.748 (EM: 70.5%)
CITY	F1: 0.690 (EM: 25.3%)
STATE	F1: 0.631 (EM: 60.8%)
POSTCODE	F1: 0.635 (EM: 61.3%)
COUNTRY	F1: 0.984 (EM: 97.0%)

A comprehensive analysis of these results immediately validates the efficacy our targeted

schema adaptation. As we can see from the table, the Paradigm 3 parser successfully resurrected this critical macro-component, achieving a global `COUNTRY` F1-Score of 0.898 (EM: 88.7%), and reaching near-perfect accuracy in countries like France (EM: 99.3%) and Portugal (EM: 97.0%). Crucially, this macro-geographic awareness was also attained without sacrificing the architecture’s inherent strength in micro-geography. The model maintained absolute dominance over highly entropic structural anchors, parsing `STREET_NAME` with a global F1 of 0.894 and `STREET_NUM` at 0.890. However, the data also reveals the inherent semantic friction of compressing complex international hierarchies into simplified database schemas. The global performance on `CITY` (F1: 0.633) and `STATE` (F1: 0.468) exposes the linguistic difficulty of mapping disparate administrative boundaries into monolithic tags. For example, in the United States, the `STATE` Exact Match plummeted to 4.6%, while `POSTCODE` collapsed to 17.2%. This localized degradation occurs because forcing the Bi-LSTM to artificially merge "Counties" and "Provinces" into a single `STATE` output occasionally confuses the sequence transition probabilities, especially when dealing with noisy, unformatted strings where postal codes are misplaced or alphanumeric. Besides that, we also have to consider that the Ground Truth annotations we have in our evaluation dataset were made by an LLM which could have assigned "wrong" tag values in dubious cases or even had some hallucinations. Therefore, this could have had some impact in our parsing results, making our parser "underperform" when it was not the case.

Despite these localized classification struggles, this adapted 7-tag output proved critical for the next architectural evolution: the Hybrid Retrieval Engine. In Phase 2, the dense vector index (Bi-Encoder) operated autonomously, which successfully bypassed structural typos but frequently triggered severe "spatial hallucinations". Instead, to cure this vulnerability, Paradigm 3 conceptually decoupled semantic fuzziness from geographic boundaries. As implemented in the retrieval pipeline, the dense k -NN vector search was executed simultaneously alongside strict Elasticsearch boolean constraints (utilizing the newly parsed `CITY`, `POSTCODE`, and `COUNTRY` tags as explicit should/must filters). This neuro-symbolic retrieval logic introduced a profound operational trade-off. By injecting these parsed tags as deterministic filters, the pipeline established rigid "geographic guardrails" where the vector search was mathematically restricted. However, establishing these safety boundaries inherently increased (or sustained) the "Both Failed" retrieval rate. If a user submitted a highly corrupted query where the `COUNTRY` or `CITY` was illegible, and the parser subsequently failed to extract it, the strict Elasticsearch filters cleanly rejected the query rather than allowing the dense vector to make a wild, unregulated guess. Conversely, in the context of enterprise Geographic Information Systems, this shift is not a regression; it is a mandatory safety mechanism. A production API serving logistics or navigation cannot afford the liability of routing a user to the wrong continent, city, state, or country. Therefore, the schema adaptation in Paradigm 3 successfully engineered a retrieval pool that is smaller but mathematically guaranteed to be geographically safer.

4.3.2 Knowledge Distillation and Feature Intelligence

Having secured a geographically safe pool of retrieved candidates through the hybrid vector-lexical search, the final architectural hurdle was to rank these candidates. Although we have seen that the Cross-Encoder was faster on a GPU environment compared to a XGBoost (that had to calculate a lot of features), the final decision was to deploy Knowledge Distillation by training a lightweight XGBoost regressor to map deterministic string comparisons to the continuous probability logits of the heavy Cross-Encoder. Given our results (whether from our experiments or paradigm comparisons), the conclusion was that the pure performance was very similar, and the XGBoost had the advantage of being easy to train and easily corrected, improved or updated according to real-world enterprise needs that might appear in the long term.

The empirical evaluation of this distillation process reveals a resounding mathematical success. When evaluated against a holdout test set, the XGBoost student achieved a Mean Absolute Error (MAE) of 0.0372 and a Mean Squared Error (MSE) of 0.0041. These metrics dictate that, on average, the predictions of the lightweight tree ensemble deviate from the massive multi-million parameter transformer by only 3.72%. This exceptionally low MSE proves that the XGBoost model did not merely memorize binary match/non-match labels. It successfully approximated the deep, continuous semantic manifold of the Cross-Encoder by compressing gigabytes of neural intelligence into a high-speed statistical structure. Therefore, to fundamentally understand how the student model engineered this approximation and to open the "black box" of its decision-making process, an analysis of its internal feature weighting was conducted, where Figure 4.10 illustrates the relative gain (importance) of the extracted neuro-symbolic features driving the final semantic ranker.

By analysing the importance matrix, a profound and initially counter-intuitive architectural behavior was revealed, where the deep semantic feature (`bi_encoder_cosine`) was assigned as the lowest importance weight of all evaluated metrics. While strict, deterministic string distances, specifically `street_levenshtein` (approx. 0.33) and `postcode_jaccard` (approx. 0.14), overwhelmingly dominated the decision tree. At first glance, the dismissal of the dense vector cosine score might suggest a rejection of semantic intelligence. However, when contextualized within the pipeline's operational flow, it reveals a highly sophisticated mathematical proxy. The Bi-Encoder is utilized in the preceding retrieval step to fetch the top 50 nearest Neighbours. Consequently, the candidate pool passed to the XGBoost model is already highly semantically saturated, and every candidate share a very high cosine similarity with the query. Therefore, since the variance in cosine similarity is so low among these top candidates, the metric loses its discriminative entropy during the re-ranking phase. Ultimately, the model had to autonomously discover how to separate the "true match" from a pool of highly similar "semantic Neighbours". So, in order to do this, it mimicked the specific mechanics of its Teacher, where the neural Cross-Encoder excels because it calculates exact token-to-token cross-attention, allowing it to heavily penalize a single-digit difference in a street name or postal code—something a holistic Bi-Encoder often overlooks. The XGBoost independently deduced that the mathematical equivalent

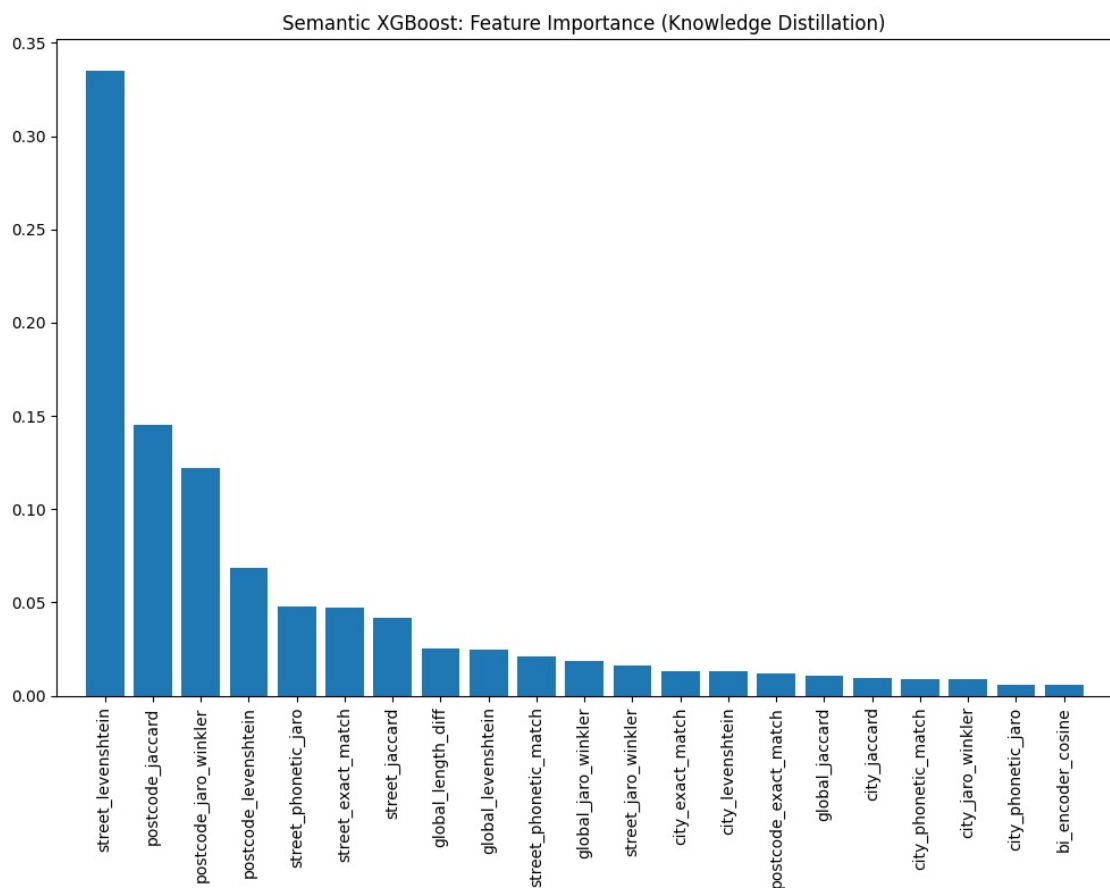


Figure 4.10: Feature Importance (Gain) for the distilled Semantic XGBoost model. The overwhelming reliance on `street_levenshtein` demonstrates that the model utilizes deterministic character-level distance to approximate the token-to-token cross-attention of the neural Teacher model.

of cross-attention is strict edit distance applied to micro-geography. Therefore, by heavily weighting `street_levenshtein` and `postcode_jaro_winkler`, the tree ensemble successfully replicates the strict, localized penalization behaviour of the transformer. In addition to that, this is precisely the behaviour predicted by **Hinton et al.** [19]. The soft probability targets produced by a teacher encode fine-grained structural preferences that hard binary labels can't convey, and which the student subsequently reconstructs through the nearest mathematical proxy available in its own feature space. In conclusion, this specific discovery defines the ultimate "Neuro-Symbolic" triumph of this research. It empirically proves that while continuous vector semantics (Neural) are required to bypass structural noise and find the correct regional neighbourhood (Retrieval), discrete mathematical string metrics (Symbolic) are vastly superior for isolating the exact physical property within that neighbourhood (Matching). Besides that, this bidirectional translation between neural and tree-based representations has been formally studied in the **Tree-Network-Tree** framework [34]. Here was demonstrated that gradient boosted trees and neural networks can serve as complementary knowledge carriers, and each capable of encoding aspects of a decision boundary that the other can't efficiently represent on its own.

4.3.3 Operational Viability (Speed vs. Accuracy)

The culmination of this research rests on evaluating the final Neuro-Symbolic architecture (Paradigm 3) against the pure deep learning baseline (Paradigm 2). Therefore, to definitively determine whether the integration of a distilled XGBoost matching engine and hybrid Elasticsearch retrieval successfully improved the overall accuracy without creating latency bottlenecks, a final end-to-end evaluation was conducted. This assessment juxtaposes the ultimate LLM-as-a-Judge retrieval win rates against the physical hardware latency required to execute them. Although the latency observed in the previous phase was very well within the boundaries of an acceptable latency, we need to make sure that our final framework (even if it is slower) keeps within acceptable latency values. Therefore, Figure 4.X illustrates the global and cross-cultural decisive win rates between Paradigm 3 and Paradigm 2. Globally, Paradigm 3 secured a decisive victory, winning 35.1% of the queries compared to Paradigm 2's marginal 4.5%. However, this overwhelming global victory is accompanied by a massive 60.3% "Both Failed" rate.

As discussed in a previous section, this high mutual failure rate is not a regression, but the direct mathematical consequence of the newly implemented geographic guardrails. Since Paradigm 3 utilizes the parsed micro and macro tags as strict must and should clauses within the Elasticsearch query, it refuses to guess on highly corrupted inputs. If the user input lacks the entropy required to satisfy these deterministic filters, the system formally fails. Although Paradigm 2 would occasionally secure a "lucky" semantic match on these noisy queries, it did so at the unacceptable cost of spatial hallucinations. On the other hand, Paradigm 3 trades this reckless semantic guessing for deterministic enterprise safety, deliberately absorbing a higher failure rate to guarantee that whatever it does retrieve is geographically constrained. While the strict geographic guardrails explain the inflated mutual failure rate, they do not entirely account for the catastrophic collapse of Paradigm 2's independent win rate (dropping to a mere 4.5% globally). This drastic

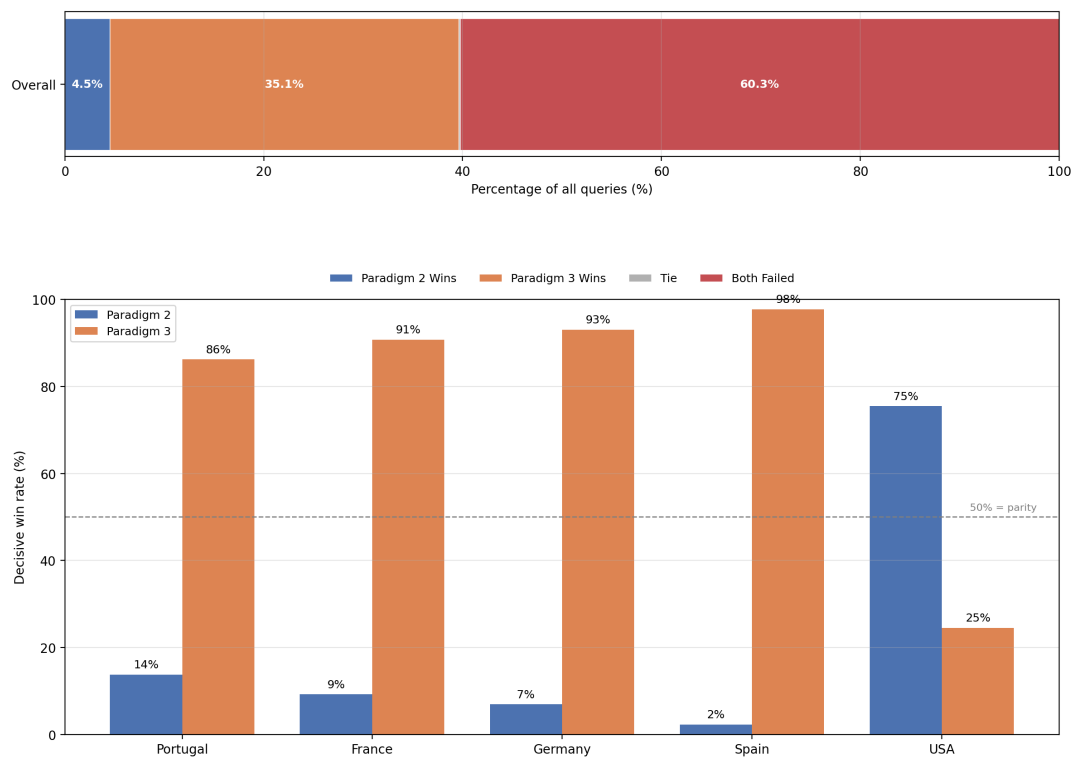


Figure 4.11: Final LLM-as-a-Judge retrieval win rates (Global and Cross-Cultural) comparing Paradigm 2 against the Neuro-Symbolic Paradigm 3. The 60.3% mutual failure rate visually captures the strict enforcement of the new geographic guardrails.

reduction is the result of a fundamental architectural supersession. In previous evaluations against Paradigm 1, Paradigm 2 secured a high win rate because its dense vector space easily handled the typos and abbreviations that paralyzed sparse lexical indexes. However, because Paradigm 3 is a hybrid system, it actively integrates that exact same dense vector space. Consequently, Paradigm 3 absorbed all of Paradigm 2’s semantic strengths and, whenever a query contains solvable structural noise, Paradigm 3 retrieves the same vector candidate pool as Paradigm 2, but successfully isolates the exact building using the superior, distilled precision of the XGBoost matching layer. In a head-to-head comparison, Paradigm 2 is mathematically stripped of its comparative advantage; the only scenarios where Paradigm 2 can still secure an isolated victory are the rare edge cases where Paradigm 3’s parser completely misclassifies a macro-tag (causing the deterministic filter to falsely reject the query), leaving Paradigm 2 to win by default via unconstrained semantic guessing.

However, isolating these win rates by country reveals a profound architectural vulnerability regarding domain adaptation. Paradigm 3 achieved total dominance across European datasets, crushing Paradigm 2 in Spain (98%), Germany (93%), France (91%), and Portugal (86%). Since the Bi-LSTM parser accurately extracted European street names and macro-geography, the Elasticsearch filters worked perfectly, feeding the XGBoost model a pristine pool of highly relevant candidates. Conversely, in the United States, this dynamic violently inverted, where Paradigm 2 defeated Paradigm 3 by a staggering 75% to 25% margin. This localized collapse is the direct consequence of parser dependency. As shown previously in the Phase 3 parsing evaluation, the parser struggled immensely with US addressing schemas, achieving an Exact Match of only 4.6% on US States and 17.2% on US Postal Codes. Therefore, since Paradigm 3 relies on these extracted tags to formulate its Elasticsearch query, parsing failures actively poisoned the retrieval pool. The sparse lexical filters forced the database to search for incorrect or empty geographic boundaries. On the other hand, Paradigm 2, which bypasses the parser entirely and queries the database using the raw, unparsed continuous vector, was completely immune to this structural fragility, allowing it to easily win the US benchmark. This proves that Neuro-Symbolic hybrid architectures are exceptionally powerful, but they possess a single point of failure: they are only as robust as the deterministic parser generating their query parameters.

Although the accuracy metrics validated the hybrid approach, the hardware latency profiling (Figure 4.12) introduces a severe unexpected operational bottleneck. The primary justification for distilling the Cross-Encoder into an XGBoost model was to provide a faster, simpler and easier to train matching model (which seemed to be the best option in long term). However, the empirical profiling reveals that Paradigm 3 (378 ms) is substantially slower end-to-end than Paradigm 2.

A granular deconstruction of the execution timeline perfectly isolates the architectural shift. The Knowledge Distillation was an undeniable success, where the XGBoost matching phase executed in 35 ms (which includes the computational overhead of extracting traditional string features like Levenshtein and Jaro-Winkler across 50 candidates), remaining highly competitive with the GPU-bound Cross-Encoder (29 ms) while requiring zero GPU VRAM. Furthermore, the optimized Bi-LSTM parser executed in an exceptionally fast 12 ms. Additionally, the fatal bottleneck

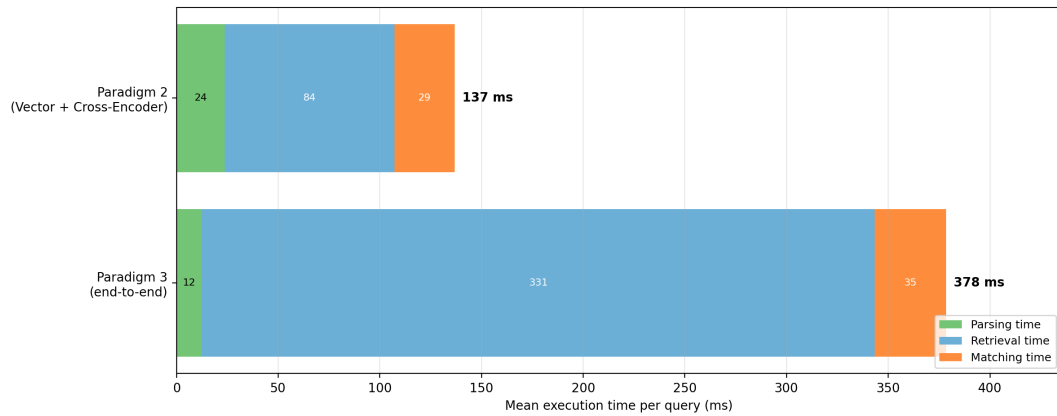


Figure 4.12: Final end-to-end latency profiling comparing Paradigm 2 against Paradigm 3. While the Knowledge Distillation successfully eliminated the heavy matching phase, the complex boolean/vector queries caused a severe latency spike in the Elasticsearch retrieval layer.

migrated entirely to the database layer, where the retrieval time for Paradigm 3 exploded to 331 ms (compared to 84 ms in Paradigm 2). This massive latency spike is intrinsic to the mechanics of Hybrid Search within Elasticsearch. In Paradigm 2, the database only performed an Approximate Nearest Neighbor (k -NN) vector search. Meanwhile, in Paradigm 3, the database was forced to execute a complex boolean query (intersecting must and should sparse lexical filters) simultaneously alongside the k -NN vector search. Calculating high-dimensional vector similarities while simultaneously evaluating sparse inverted indices forces the database engine into highly intensive CPU and I/O operations, increasing the total latency of this intermediate phase.

In conclusion, the results of Paradigm 3 provided the definitive conclusion to this research. The Neuro-Symbolic framework successfully solved the core problems of pure deep learning: it eliminated dangerous spatial hallucinations by enforcing deterministic geographic guardrails, and it proved that heavy transformer logic can be successfully distilled into lightweight, CPU-friendly tree ensembles without losing semantic intelligence. However, it simultaneously uncovered some realities of enterprise systems engineering. By decoupling the semantic intelligence from the retrieval engine, the pipeline inherited strict parser dependency, causing a few failures in out-of-domain regions like the US. Furthermore, while the GPU bottleneck was successfully eliminated, the computational latency was simply transferred to the database cluster, resulting in a 331 ms hybrid retrieval. Ultimately, this proves that geographic entity resolution cannot be solved by a single "silver bullet" algorithm. It is a delicate and perpetual balancing act between semantic elasticity, deterministic safety, and database architecture constraints.

Chapter 5

Conclusion

5.1 Final Remarks

As geographic information becomes an increasingly critical pillar of the global digital economy, its operational utility remains fundamentally constrained by the quality and structure of textual address data. This dissertation set out to solve the profound "Human-Database Gap" such as the systemic misalignment between chaotic, high-entropy real-world user inputs and the rigid structural requirements of enterprise databases. Therefore, in order to address this challenge, this research undertook a rigorous, three-phase methodological journey to design, benchmark, and deploy a highly scalable geographic entity resolution pipeline against a multinational database. Progressing from isolated component benchmarking (Phase 1) to end-to-end deep learning evaluations (Phase 2), the research ultimately culminated in the development and validation of a Hybrid Neuro-Symbolic Framework (Phase 3). Central to the development of this framework was the need to evaluate the theoretical effectiveness of modern Artificial Intelligence in geographic contexts. Therefore, the first Research Question (RQ1) investigated the extent to which modeling addresses as probabilistic semantic entities outperform traditional deterministic string matching in resolving ambiguity and minimizing false negatives. The empirical results provided a definitive, albeit highly nuanced, affirmative answer. Deep semantic models, in this case, Transformer-based Bi-Encoders and Cross-Encoders, vastly outperformed traditional character-based metrics (such as Levenshtein distance) by autonomously learning to bypass severe structural noise, cross-lingual abbreviations, and out-of-vocabulary anomalies. However, end-to-end evaluations uncovered a critical caveat, where unrestrained semantic elasticity directly causes dangerous spatial hallucinations, often retrieving highly similar street names on entirely wrong continents. Consequently, this research concludes that while semantic modelling is undeniably superior for handling linguistic ambiguity, it is operationally insufficient on its own. True enterprise-grade effectiveness is only achieved when semantic fuzziness is mathematically constrained by deterministic geographic guardrails.

Furthermore, resolving geographic entities at an enterprise scale is not merely a linguistic challenge, but a severe systems engineering problem. Thus, the second Research Question (RQ2)

interrogated how the transition from lexical indexing (for example, Inverted Indexes) to high-dimensional vector indexing impacts the trade-off between query latency and retrieval precision. The findings demonstrate that dense vector indexing successfully cures the lexical recall bottleneck, allowing systems to retrieve candidates that traditional BM25 searches miss entirely. However, this transition fundamentally shifts the operational bottleneck rather than eliminating it. The implementation of Knowledge Distillation proved that the heavy GPU-bound latency of deep semantic matching could be successfully bypassed, where the distilled XGBoost model evaluated candidate pairs in just 35 ms on a standard CPU. However, integrating this semantic intelligence directly into the database via hybrid Elasticsearch queries (combining dense vector k -NN searches with sparse Boolean filters) triggered a massive latency spike of 331 ms. Ultimately, this research concludes that achieving neural retrieval precision at a massive scale imposes a severe I/O and CPU tax on the database architecture, dictating a perpetual, delicate balance between algorithmic accuracy and hardware constraints.

5.2 Limitations and Future Work

Demonstrating academic rigor requires a transparent acknowledgment of the boundaries and limitations inherent in the proposed framework. The most critical architectural vulnerability exposed during the final evaluation was severe parser dependency. Because the Neuro-Symbolic retrieval engine heavily relies on extracted tags to formulate its deterministic Boolean queries, parsing failures actively poison the downstream candidate pool. This flaw was catastrophically evident in the United States dataset, where the parser's inability to accurately extract macro-geographic components (like US States and Postal Codes) caused the hybrid pipeline to decisively lose against the parser-independent pure deep learning baseline. Additionally, the operational constraints of the evaluation environment introduced inherent data limitations. The strict enforcement of a 25-million record cap on the enterprise master database inevitably meant that some legitimate evaluation queries may not have physically existed within the index. Therefore, it is possible to conclude that this limitation artificially inflated the "Both Failed" noise floor, complicating the absolute measurement of algorithmic recall. Finally, relying on a Large Language Model (Gemini 3.1 Flash) as an automated "LLM-as-a-Judge" to determine the end-to-end win rates introduces potential evaluation vulnerabilities. Although the model was constrained by a strict prompt schema and zero-temperature deterministic parameters, LLMs remain theoretically susceptible position and verbosity bias, and stochastic hallucinations [17] when interpreting highly ambiguous geographic strings, representing an unavoidable limitation of automated semantic auditing at scale.

The limitations identified in this research naturally illuminate the path for future work in Geographic Information Retrieval. Therefore, the first step would be to address the severe parser dependency that crippled the pipeline in out-of-domain regions like the United States, where subsequent research must prioritize targeted region-specific data. By heavily augmenting the multinational training corpora with localized, high-entropy administrative variations, the sequence parsing models can be trained to learn the specific macro-geographic structures that currently evade them,

thereby fortifying the accuracy of the downstream retrieval pool. Furthermore, solving the 331 ms retrieval bottleneck requires a fundamental re-evaluation of the underlying persistence layer. Future implementations should investigate migrating away from generalized search engines like Elasticsearch towards dedicated, vector-native databases. As an intermediate optimization before a full infrastructure migration, would be replacing the rigid BM25 Boolean filters with a learned sparse model such as **SPLADE** [15]. This generates inverted-index-compatible representations with built-in semantic term expansion, which could reduce lexical rigidity while preserving the lookup speeds of sparse retrieval. In the longer term, dedicated vector-native infrastructures explicitly designed to execute complex high-dimensional vector searches simultaneously with scalar metadata filtering could represent the definitive path towards resolving the severe CPU and I/O friction that currently limits the scalability of hybrid semantic architectures. Finally, the rapid evolution of Generative AI presents an opportunity to completely reimagine the parsing and normalization phases. Rather than relying strictly on discriminative sequence classifiers, future architectures should explore the integration of Large Language Models (LLMs) via Retrieval-Augmented Generation (RAG) pipelines directly into the active workflow. Therefore, by leveraging the deep zero-shot reasoning capabilities of modern generative models to semantically map chaotic user inputs into rigid ontological schemas prior to indexing, future systems could theoretically bridge the Human-Database gap with unprecedented flexibility, pushing the boundaries of automated Entity Resolution even further.

References

- [1] Abhishek Arora and Melissa Dell. *LinkTransformer: A Unified Package for Record Linkage with Transformer Language Models*. June 24, 2024. arXiv: [2309.00789\[cs\]](#).
- [2] David Beauchemin and Marouane Yassine. *Deepparse : An Extendable, and Fine-Tunable State-Of-The-Art Library for Parsing Multinational Street Addresses*. Nov. 20, 2023. arXiv: [2311.11846\[cs\]](#).
- [3] Richard E. Bellman. *Dynamic Programming*. Princeton, NJ, USA: Princeton University Press, 1957.
- [4] Jon Louis Bentley. “Multidimensional binary search trees used for associative searching”. In: *Communications of the ACM* 18.9 (Sept. 1975), pp. 509–517. ISSN: 0001-0782. DOI: [10.1145/361002.361007](#).
- [5] Piotr Bojanowski et al. “Enriching Word Vectors with Subword Information”. In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 135–146. ISSN: 2307-387X. DOI: [10.1162/tacl_a_00051](#).
- [6] Tianqi Chen and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA, USA: ACM, Aug. 13, 2016, pp. 785–794. ISBN: 978-1-4503-4232-2. DOI: [10.1145/2939672.2939785](#).
- [7] Ruozhen Cheng and Jing Chen. “A location conversion method for roads through deep learning-based semantic matching and simplified qualitative direction knowledge representation”. In: *Engineering Applications of Artificial Intelligence* 104 (Sept. 2021), p. 104400. ISSN: 09521976. DOI: [10.1016/j.engappai.2021.104400](#).
- [8] Alexis Conneau et al. “Unsupervised Cross-lingual Representation Learning at Scale”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020, pp. 8440–8451. DOI: [10.18653/v1/2020.acl-main.747](#).
- [9] Frederick J Damerau. “A technique for computer detection and correction of spelling errors”. In: *Communications of the ACM* 7.3 (1964), pp. 171–176.
- [10] Arthur P Dempster, Nan M Laird, and Donald B Rubin. “Maximum likelihood from incomplete data via the EM algorithm”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977), pp. 1–22.

- [11] Wenzhou Dou et al. “Enhancing Deep Entity Resolution with Integrated Blocker-Matcher Training: Balancing Consensus and Discrepancy”. In: *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. CIKM ’24: The 33rd ACM International Conference on Information and Knowledge Management. Boise ID USA: ACM, Oct. 21, 2024, pp. 508–518. ISBN: 979-8-4007-0436-9. DOI: [10.1145/3627673.3679843](#).
- [12] André V. Duarte and Arlindo L. Oliveira. “Improving Embeddings for High-Accuracy Transformer-Based Address Matching Using a Multiple in-Batch Negatives Loss”. In: *2023 International Conference on Machine Learning and Applications (ICMLA)*. 2023 International Conference on Machine Learning and Applications (ICMLA). Jacksonville, FL, USA: IEEE, Dec. 15, 2023, pp. 120–127. ISBN: 979-8-3503-4534-6. DOI: [10.1109/ICMLA58977.2023.00025](#).
- [13] Ivan P Fellegi and Alan B Sunter. “A theory for record linkage”. In: *Journal of the American Statistical Association* 64.328 (1969), pp. 1183–1210.
- [14] Jacques Fize, Ludovic Moncla, and Bruno Martins. “Deep Learning for Toponym Resolution: Geocoding Based on Pairs of Toponyms”. In: *ISPRS International Journal of Geo-Information* 10.12 (Dec. 2, 2021), p. 818. ISSN: 2220-9964. DOI: [10.3390/ijgi10120818](#).
- [15] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. “SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking”. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. Virtual Event, Canada: ACM, July 11, 2021, pp. 2288–2292. ISBN: 978-1-4503-8037-9. DOI: [10.1145/3404835.3463098](#).
- [16] Jeremy Foxcroft, Adrian d’Alessandro, and Luiza Antonie. “Name2Vec: Personal Names Embeddings”. In: *Advances in Artificial Intelligence*. Ed. by Marie-Jean Meurs and Frank Rudzicz. Vol. 11489. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 505–510. ISBN: 978-3-030-18304-2. DOI: [10.1007/978-3-030-18305-9_52](#).
- [17] Jiawei Gu et al. *A Survey on LLM-as-a-Judge*. 2024. arXiv: [2411.15594\[cs.CL\]](#).
- [18] Gaurav Gupta et al. “BLISS: A Billion scale Index using Iterative Re-partitioning”. In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. KDD ’22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Washington DC USA: ACM, Aug. 14, 2022, pp. 486–495. ISBN: 978-1-4503-9385-0. DOI: [10.1145/3534678.3539414](#).
- [19] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. *Distilling the Knowledge in a Neural Network*. 2015. arXiv: [1503.02531\[cs.NE\]](#).
- [20] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-Term Memory”. In: *Neural Computation* 9.8 (Nov. 15, 1997), pp. 1735–1780. ISSN: 0899-7667. DOI: [10.1162/neco.1997.9.8.1735](#).
- [21] Piotr Indyk and Rajeev Motwani. “Approximate nearest neighbors: towards removing the curse of dimensionality”. In: *Proceedings of the thirtieth annual ACM symposium on Theory of computing*. 1998, pp. 604–613.
- [22] Arjit Jain, Sunita Sarawagi, and Prithviraj Sen. “Deep Indexed Active Learning for Matching Heterogeneous Entity Representations”. In: *Proceedings of the VLDB Endowment* 15.1 (Sept. 2021), pp. 31–45. ISSN: 2150-8097. DOI: [10.14778/3485450.3485455](#).

- [23] Matthew A Jaro. “Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida”. In: *Journal of the American Statistical Association*. Vol. 84. 1989, pp. 414–420.
- [24] Salil Rajeev Joshi, Arpan Somani, and Shourya Roy. “ReLink: Complete-Link Industrial Record Linkage Over Hybrid Feature Spaces”. In: *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. 2021 IEEE 37th International Conference on Data Engineering (ICDE). Chania, Greece: IEEE, Apr. 2021, pp. 2625–2636. ISBN: 978-1-7281-9184-3. DOI: [10.1109/ICDE51399.2021.00293](https://doi.org/10.1109/ICDE51399.2021.00293).
- [25] Dimitrios Karapiperis, Christos Tjortjis, and Vassilios Verykios. “LSBlock: A Hybrid Blocking System Combining Lexical and Semantic Similarity Search for Record Linkage”. In: *Advances in Databases and Information Systems*. Ed. by Panos K. Chrysanthis et al. Vol. 16043. Series Title: Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2026, pp. 131–146. ISBN: 978-3-032-05280-3. DOI: [10.1007/978-3-032-05281-0_9](https://doi.org/10.1007/978-3-032-05281-0_9).
- [26] Dimitrios Karapiperis and Vassilios S. Verykios. “Scaling Entity Resolution with K-Means: A Review of Partitioning Techniques”. In: *Electronics* 14.18 (Sept. 11, 2025), p. 3605. ISSN: 2079-9292. DOI: [10.3390/electronics14183605](https://doi.org/10.3390/electronics14183605).
- [27] Vladimir Karpukhin et al. “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.
- [28] Aaron R. Kaufman and Aja Klevs. “Adaptive Fuzzy String Matching: How to Merge Datasets with Only One (Messy) Identifying Field”. In: *Political Analysis* 30.4 (Oct. 2022), pp. 590–596. ISSN: 1047-1987, 1476-4989. DOI: [10.1017/pan.2021.38](https://doi.org/10.1017/pan.2021.38).
- [29] John Lafferty, Andrew McCallum, and Fernando CN Pereira. “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”. In: *Proceedings of the 18th International Conference on Machine Learning (ICML)*. 2001, pp. 282–289.
- [30] Manuel Leone et al. “A critical re-evaluation of neural methods for entity alignment”. In: *Proceedings of the VLDB Endowment* 15.8 (Apr. 2022), pp. 1712–1725. ISSN: 2150-8097. DOI: [10.14778/3529337.3529355](https://doi.org/10.14778/3529337.3529355).
- [31] Vladimir I Levenshtein. “Binary codes capable of correcting deletions, insertions, and reversals”. In: *Soviet Physics Doklady* 10.8 (1966), pp. 707–710.
- [32] Patrick Lewis et al. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”. In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 9459–9474.
- [33] Huahang Li et al. *On Leveraging Large Language Models for Enhancing Entity Resolution: A Cost-efficient Approach*. Sept. 12, 2024. arXiv: [2401.03426\[cs\]](https://arxiv.org/abs/2401.03426).
- [34] Jiawei Li et al. “TNT: An Interpretable Tree-Network-Tree Learning Framework using Knowledge Distillation”. In: *Entropy* 22.11 (Oct. 24, 2020), p. 1203. ISSN: 1099-4300. DOI: [10.3390/e22111203](https://doi.org/10.3390/e22111203).
- [35] Jonathan Lin et al. *RACCOON: A Retrieval-Augmented Generation Approach for Location Coordinate Capture from News Articles*. Jan. 20, 2025. arXiv: [2501.11440\[cs\]](https://arxiv.org/abs/2501.11440).
- [36] Le Ma et al. *A Comprehensive Survey on Vector Database: Storage and Retrieval Technique, Challenge*. June 16, 2025. arXiv: [2310.11703\[cs\]](https://arxiv.org/abs/2310.11703).
- [37] Yury A. Malkov and Dmitry A. Yashunin. “Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.4 (2020), pp. 824–836.

- [38] Alejandro Molina-Villegas et al. “Geographic Named Entity Recognition and Disambiguation in Mexican News using word embeddings”. In: *Expert Systems with Applications* 176 (Aug. 2021), p. 114855. ISSN: 09574174. DOI: [10.1016/j.eswa.2021.114855](https://doi.org/10.1016/j.eswa.2021.114855).
- [39] Mohammad Hossein Moslemi et al. *Heterogeneity in Entity Matching: A Survey and Experimental Analysis*. Aug. 11, 2025. arXiv: [2508.08076\[cs\]](https://arxiv.org/abs/2508.08076).
- [40] Arthur Ning, Flavius Frasinca, and Tarmo Robal. “SC-Block++: A Blocking Algorithm Based on Adaptive Flood Regularization”. In: *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*. SAC ’25: 40th ACM/SIGAPP Symposium on Applied Computing. Catania International Airport Catania Italy: ACM, Mar. 31, 2025, pp. 1258–1266. ISBN: 979-8-4007-0629-5. DOI: [10.1145/3672608.3707946](https://doi.org/10.1145/3672608.3707946).
- [41] Leonardo Nizzoli et al. “Geo-semantic-parsing: AI-powered geoparsing by traversing semantic knowledge graphs”. In: *Decision Support Systems* 136 (Sept. 2020), p. 113346. ISSN: 01679236. DOI: [10.1016/j.dss.2020.113346](https://doi.org/10.1016/j.dss.2020.113346).
- [42] Opendatasoft and GeoNames. *All Postal Code - All countries (Geonames)*. May 2026. URL: <https://public.opendatasoft.com/explore/assets/geonames-postal-code/> (visited on 05/21/2026).
- [43] William Pugh. “Skip lists: a probabilistic alternative to balanced trees”. In: *Communications of the ACM* 33.6 (June 1990), pp. 668–676. ISSN: 0001-0782. DOI: [10.1145/78973.78977](https://doi.org/10.1145/78973.78977).
- [44] Yuting Qin. “Understanding Indexing Efficiency for Approximate Nearest Neighbor Search in High-dimensional Vector Databases”. MA thesis. MIT, 2024.
- [45] Zhen Qin et al. “Are Neural Rankers still Outperformed by Gradient Boosted Decision Trees?” In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=UtlvF_q_vC.
- [46] Hafiz Tayyab Rauf, Andre Freitas, and Norman W. Paton. *TableDC: Deep Clustering for Tabular Data*. May 28, 2024. arXiv: [2405.17723\[cs\]](https://arxiv.org/abs/2405.17723).
- [47] Nils Reimers and Iryna Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. 2019, pp. 3982–3992.
- [48] Stephen Robertson and Hugo Zaragoza. “The Probabilistic Relevance Framework: BM25 and Beyond”. In: *Foundations and Trends in Information Retrieval* 3.4 (2009), pp. 333–389. ISSN: 1554-0669. DOI: [10.1561/1500000019](https://doi.org/10.1561/1500000019).
- [49] Smith Saengsiripaiboon, Eakasit Pacharawongsakda, and Duangjai Jitkongchuen. “Enhancing Efficiency in Entity Resolution Strategies for Batch Prompting”. In: *2025 6th International Conference on Big Data Analytics and Practices (IBDAP)*. 2025 6th International Conference on Big Data Analytics and Practices (IBDAP). Chiang Mai, Thailand: IEEE, Aug. 1, 2025, pp. 31–35. ISBN: 979-8-3315-9474-9. DOI: [10.1109/IBDAP65587.2025.11145850](https://doi.org/10.1109/IBDAP65587.2025.11145850).
- [50] Mohammadreza Sharifi and Danial Ahmadzadeh. *Transformer-Gather, Fuzzy-Reconsider: A Scalable Hybrid Framework for Entity Resolution*. Oct. 24, 2025. arXiv: [2509.17470\[cs\]](https://arxiv.org/abs/2509.17470).
- [51] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. “Sequence to Sequence Learning with Neural Networks”. In: *Advances in Neural Information Processing Systems*. Vol. 27. 2014.
- [52] Ashish Vaswani et al. “Attention is All You Need”. In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017.

- [53] Jimin Wang and Yingjie Hu. *Are We There Yet? Evaluating State-of-the-Art Neural Network based Geoparsers Using EUPEG as a Benchmarking Platform*. July 15, 2020. arXiv: [2007.07455\[cs\]](#).
- [54] Runhui Wang and Yongfeng Zhang. “Pre-trained Language Models for Entity Blocking: A Reproducibility Study”. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, June 2024, pp. 8720–8730. DOI: [10.18653/v1/2024.naacl-long.483](#).
- [55] Duncan J Watts and Steven H Strogatz. “Collective dynamics of ‘small-world’ networks”. In: *Nature* 393.6684 (1998), pp. 440–442.
- [56] Dominik Weckmüller, Alexander Dunkel, and Dirk Burghardt. “Embedding-Based Multilingual Semantic Search for Geo-Textual Data in Urban Studies”. In: *Journal of Geovisualization and Spatial Analysis* 9.2 (Dec. 2025), p. 31. ISSN: 2509-8810, 2509-8829. DOI: [10.1007/s41651-025-00232-5](#).
- [57] William E Winkler. “String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage”. In: *Proceedings of the Section on Survey Research Methods*. American Statistical Association. 1990, pp. 354–359.
- [58] Zheren Yan et al. “The Integration of Linguistic and Geospatial Features Using Global Context Embedding for Automated Text Geocoding”. In: *ISPRS International Journal of Geo-Information* 10.9 (Aug. 24, 2021), p. 572. ISSN: 2220-9964. DOI: [10.3390/ijgi10090572](#).
- [59] Marouane Yassine and David Beauchemin. *Structured Multinational Address Data*. <https://huggingface.co/datasets/deepparse/worldwide-addresses>. 2026.
- [60] Alexandros Zeakis et al. *Pre-trained Embeddings for Entity Resolution: An Experimental Analysis [Experiment, Analysis & Benchmark]*. Apr. 24, 2023. arXiv: [2304.12329\[cs\]](#).
- [61] Chiqun Zhang et al. “Fast Attention-based Learning-To-Rank Model for Structured Map Search”. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. Virtual Event Canada: ACM, July 11, 2021, pp. 942–951. ISBN: 978-1-4503-8037-9. DOI: [10.1145/3404835.3462904](#).
- [62] Huayu Zhang et al. “FLAP: a framework for linking free-text addresses to the Ordnance Survey Unique Property Reference Number database”. In: *Frontiers in Digital Health* 5 (Nov. 28, 2023), p. 1186208. ISSN: 2673-253X. DOI: [10.3389/fdgth.2023.1186208](#).
- [63] Kaiyu Zhang et al. “High-Dimensional Nearest Neighbor Search-Based Blocking in Entity Resolution”. In: *Web Information Systems and Applications*. Ed. by Cheqing Jin et al. Vol. 14883. Series Title: Lecture Notes in Computer Science. Singapore: Springer Nature Singapore, 2024, pp. 215–226. ISBN: 978-981-97-7706-8. DOI: [10.1007/978-981-97-7707-5_19](#).
- [64] Minjia Zhang, Wenhan Wang, and Yuxiong He. “GraSP: Optimizing Graph-based Nearest Neighbor Search with Subgraph Sampling and Pruning”. In: *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. WSDM ’22: The Fifteenth ACM International Conference on Web Search and Data Mining. Virtual Event AZ USA: ACM, Feb. 11, 2022, pp. 1395–1405. ISBN: 978-1-4503-9132-0. DOI: [10.1145/3488560.3498425](#).

- [65] Qianxi Zhang et al. “VBASE: Unifying Online Vector Similarity Search and Relational Queries via Relaxed Monotonicity”. In: *17th USENIX Symposium on Operating Systems Design and Implementation (OSDI 23)*. Boston, MA: USENIX Association, July 2023, pp. 377–395. ISBN: 978-1-939133-34-2. URL: <https://www.usenix.org/conference/osdi23/presentation/zhang-qianxi>.
- [66] Zeyu Zhang and Steven Bethard. “Improving Toponym Resolution with Better Candidate Generation, Transformer-based Reranking, and Two-Stage Resolution”. In: *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*. Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023). Toronto, Canada: Association for Computational Linguistics, 2023, pp. 48–60. DOI: [10.18653/v1/2023.starsem-1.6](https://doi.org/10.18653/v1/2023.starsem-1.6).
- [67] Lianmin Zheng et al. “Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena”. In: *Advances in Neural Information Processing Systems*. Vol. 36. Curran Associates, Inc., 2023.

Appendix A

Exploratory Data Engineering and Baseline Infrastructure

A.1 Exploratory Data Analysis

Data Quality Issues Summary				
Issue Category	Elements Affected	Severity	Description	Priority
Incomplete Addresses	117,694	High	Missing street or house number	Critical
Street Prefix Variations	1	Medium	10 prefixes with variations	High
Non-standard Postal Codes	8,346	Medium	Format corrections needed	High
Complex House Numbers	39,474	Low	Non-simple number formats	Medium
Potential City Duplicates	86	Low	Similar city names detected	Low
Unmatched Street Patterns	66,759	Medium	Streets without recognized prefixes	Medium

Figure A.1: Summary of data quality and completeness metrics for the baseline dataset.

A.2 PostGIS Spatial Indexing and architecture Benchmarks

A.3 Baseline Performance and Accuracy Results



Figure A.2: Comprehensive analysis of the deterministic address normalization and validation phase.

Table A.1: Key Performance Indicators (KPIs) for the baseline deterministic retrieval system.

Metric	Recorded Value	Notes
Reliability	100% Success Rate	Zero system failures recorded.
Performance	165.9 ms	Average response time (95th percentile: 235.3 ms).
Overall Accuracy	22.4%	Top-1 exact match accuracy.
Typo Accuracy	0.0%	Complete failure on character-level noise.
Complex Query Accuracy	0.0%	Complete failure on multi-field anomalies.
Postcode-Only Accuracy	40.0%	Highest accuracy subset (structured data).
Normalization Throughput	2,351 records/sec	Baseline processing speed.

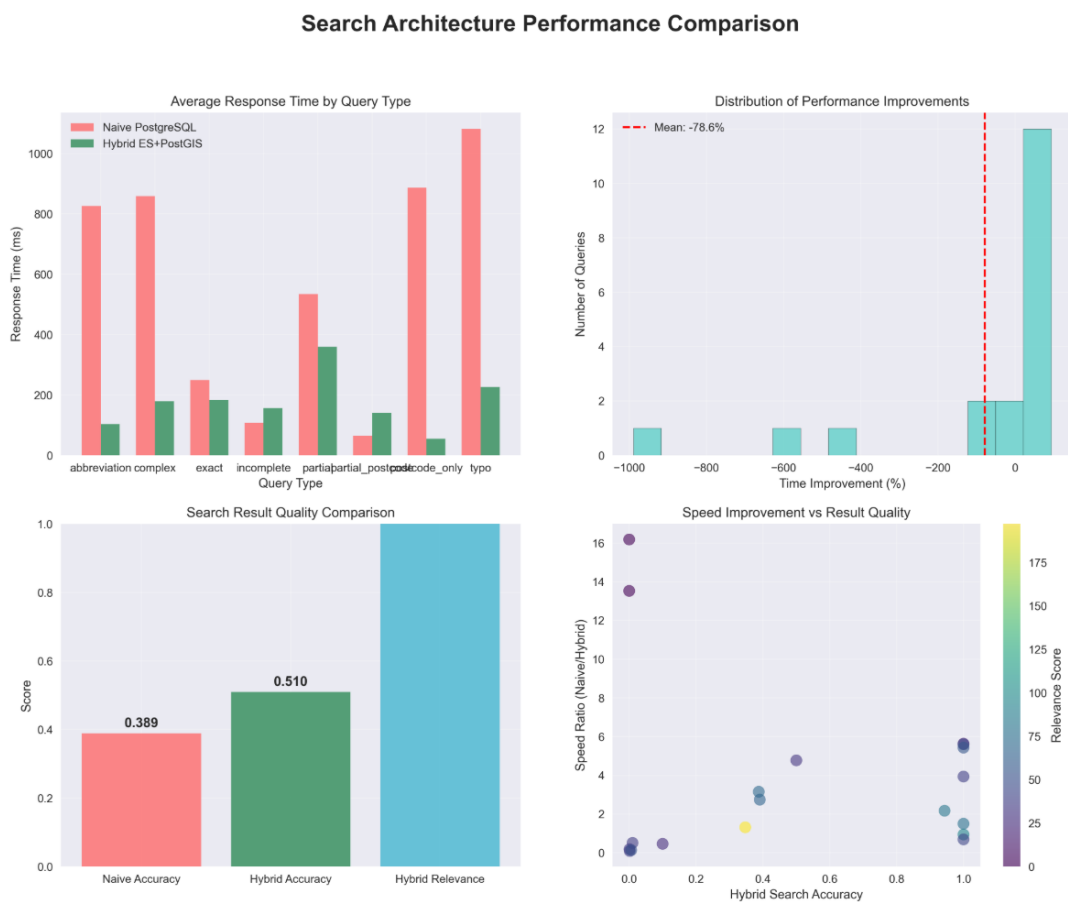


Figure A.3: Database query performance benchmarks comparing execution times across PostgreSQL (GiST, SP-GiST, B-Tree) and Elasticsearch indices.

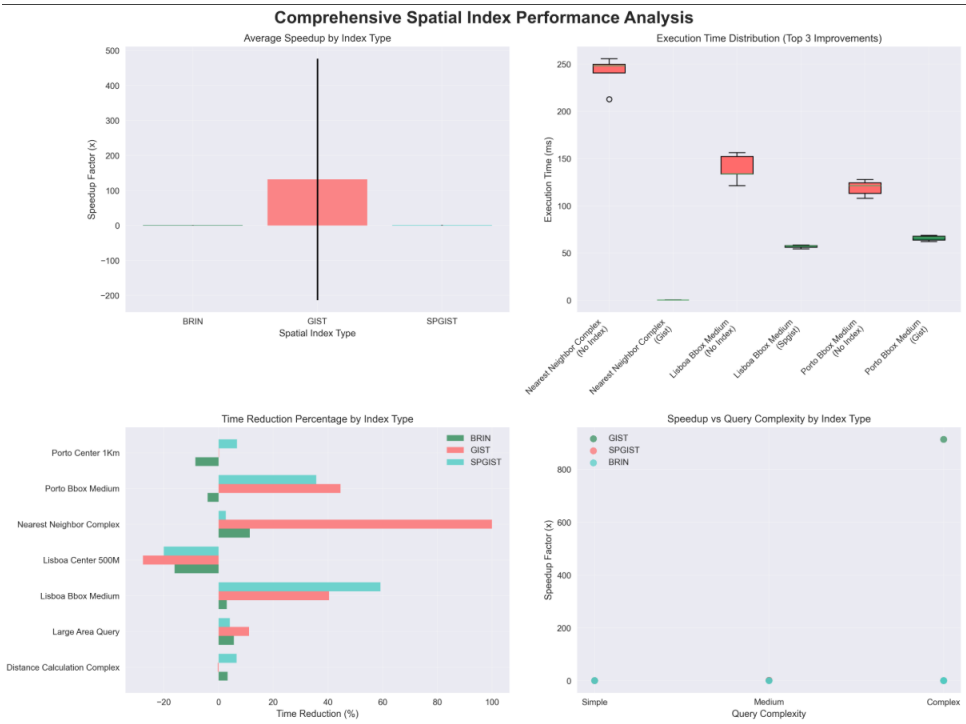


Figure A.4: Geographic distribution and density of raw OpenStreetMap address nodes across continental Portugal.

Appendix B

Feature Engineering Inventory

This appendix provides an exhaustive, line-by-line technical catalog of the 49-dimensional feature space engineered across Phase 3.2 and Phase 3.3 of this research work. The variables are structured by architectural category, explicitly mapping out the data points utilized by downstream supervised machine learning models.

Table B.1: Exhaustive catalog of the 49 engineered features.

Category	Feature Variable Name	Mathematical Logic / Description
Named Entity Recognition	ner_loc_similarity	Jaccard similarity of extracted location/GPE entities between the raw input strings.
	ner_road_loc_similarity	Jaccard similarity of location entities found specifically inside parsed road sub-components.
	ner_city_loc_similarity	Jaccard similarity of location entities found inside parsed city sub-components.
	ner_address1_loc_count	Absolute count of unique GPE/LOC entities identified in address 1.
	ner_address2_loc_count	Absolute count of unique GPE/LOC entities identified in address 2.
	ner_entity_count_diff	Absolute difference between the entity counts: $ \text{count}_1 - \text{count}_2 $.
	ner_has_location_entities	Binary flag indicating if at least one of the address strings contained an entity.
Continued on next page		

Category	Feature Variable Name	Mathematical Logic / Description
	ner_both_have_locations	Binary flag indicating if both address strings concurrently contained entities.
	ner_shared_entities	Count of identical location entities intersecting across both input strings.
	ner_total_unique_entities	Count of unique total entities found across the union of both strings.
Global Baseline Similarities	phase2_jaro_winkler_score	Jaro-Winkler string similarity measuring holistic character-level alignments.
	phase2_levenshtein_score	Normalized Levenshtein distance: $1.0 - (\text{edit_dist} / \max(\text{len}_1, \text{len}_2))$.
	phase2_jaccard_score	Jaccard similarity calculated over character bigram tokens.
	phase2_hybrid_avg	Simple arithmetic mean of Jaro-Winkler, Levenshtein, and Jaccard metrics.
	phase2_weighted_combo	Scalar combo weight: $0.5(\text{JW}) + 0.3(\text{Jaccard}) + 0.2(\text{Levenshtein})$.
Component-wise Alignments	comp_road_similarity	Jaro-Winkler score evaluating the parsed road name field.
	comp_house_number_similarity	Jaro-Winkler score evaluating the parsed house number field.
	comp_city_similarity	Jaro-Winkler score evaluating the parsed city field.
	comp_postcode_similarity	Jaro-Winkler score evaluating the parsed postal code field.
	comp_normalized_similarity	Jaro-Winkler score evaluating the entirely normalized address sequence.
	comp_road_exact_match	Binary indicator of a strict case-insensitive token match on road names.
	comp_house_number_exact_match	Binary indicator of a strict case-insensitive match on house numbers.
	comp_city_exact_match	Binary indicator of a strict case-insensitive match on city fields.
<i>Continued on next page</i>		

Category	Feature Variable Name	Mathematical Logic / Description
	comp_postcode_exact_match	Binary indicator of a strict case-insensitive match on postcodes.
	comp_normalized_exact_match	Binary indicator of a strict case-insensitive match on full normalized structures.
	comp_avg_similarity	Row-wise mean across all component similarities.
	comp_max_similarity	Highest similarity value seen among the sub-component comparisons.
	comp_min_similarity	Lowest similarity value seen among the sub-component comparisons.
Textual Statistics	stat_addr1_length	Total character length count of raw address string 1.
	stat_addr2_length	Total character length count of raw address string 2.
	stat_length_diff	Absolute character count difference: $ \text{len}_1 - \text{len}_2 $.
	stat_length_ratio	Symmetrical length bounding ratio: $\min(\text{len}_1, \text{len}_2) / \max(\text{len}_1, \text{len}_2)$.
	stat_addr1_words	Total word count (whitespace tokens) of address string 1.
	stat_addr2_words	Total word count (whitespace tokens) of address string 2.
	stat_word_diff	Absolute token count difference between the strings.
	stat_addr1_digits	Absolute count of numeric characters (0 – 9) present inside address 1.
	stat_addr2_digits	Absolute count of numeric characters (0 – 9) present inside address 2.
	stat_digits_match	Binary indicator specifying if both strings contain the identical frequency of digits.
	interact_phase2_comp	Product interaction score mapping global baseline similarity to component average.
Continued on next page		

Multi-Modal Interactions

Category	Feature Variable Name	Mathematical Logic / Description
	interact_phase2_ner	Product interaction score mapping global baseline similarity to overall NER metrics.
	interact_comp_ner	Product interaction score mapping component averages directly to NER similarities.
	interact_all_three	Core alignment consensus product: $\text{Baseline} \times \text{Component} \times \text{NER}$.
	interact_phase2_high	Binary indicator evaluating if the baseline score exceeds confidence thresholds (> 0.7).
	interact_comp_high	Binary indicator evaluating if the component average exceeds confidence thresholds (> 0.7).
	interact_ner_high	Binary indicator evaluating if the primary NER similarity exceeds thresholds (> 0.7).
	interact_all_high	Strict voting compliance flag when all three key metrics sit safely above 0.7.
	interact_majority_high	Majority voting flag proving at least 2 out of 3 features passed high confidence (> 0.7).
	interact_score_variance	Cross-modal row variance across Baseline, Component, and NER scores.
	interact_score_mean	Cross-modal row mean evaluating baseline, component, and entity metrics.