APPLIED DATA SCIENCE

ONTOLOGY ALIGNMENT CLASSIFIER

BASED ON NLP

Era Alcani
Francesco Dal Cero
Francesco Mastrosimone

Politecnico di Torino
1859

Repertorio
UNLOCKING PATIENT PREFERENCES

TABLE OF CONTENTS

PROJECT OVERVIEW

01

PROJECT CONTEXT AND SOLUTION

THE CHALLENGE: SEMANTIC INCONSISTENCY

Context : Healthcare evidence is scattered across isolated "data silos".
Problem : Different terms (e.g., Fatigue vs. Asthenia) prevent meaningful integration of this data.

THE SOLUTION: AI SEMANTIC ENGINE

Goal : Develop an AI-powered engine to automatically harmonize fragmented data into structured evidence.
Impact : Replaces manual "brittle scripts" with a scalable, automated pipeline.

OBJECTIVES AND RESEARCH PERSPECTIVE

THE GOAL: AUTOMATED HARMONIZATION

To replace manual, brittle scripts with a robust framework that transforms raw attributes into high-confidence clinical evidence.

THE RESEARCH OBJECTIVES

SYSTEM LEVEL: Build a reusable framework for interpretable ontology alignment.

RESEARCH LEVEL: Evolve from simple string matching to capturing actual clinical meaning and context.

RESEARCH HYPOTHESIS

02

RESEARCH HYPOTHESIS

Efficiency and Ambiguity (H1)

Question: Do we always need complex deep learning models to align clinical terms?
Hypothesis: No. Model complexity should match semantic ambiguity: deterministic rules handle clear cases, while deep models are needed only for ambiguous or synonymous terms.

Architecture (H2)

Question: How do we balance high accuracy with scalability?
Hypothesis: Efficiency demands a "Filter-then-Rank" architecture. Scalability is achieved by filtering candidates with a fast retriever before applying expensive deep semantic scoring.

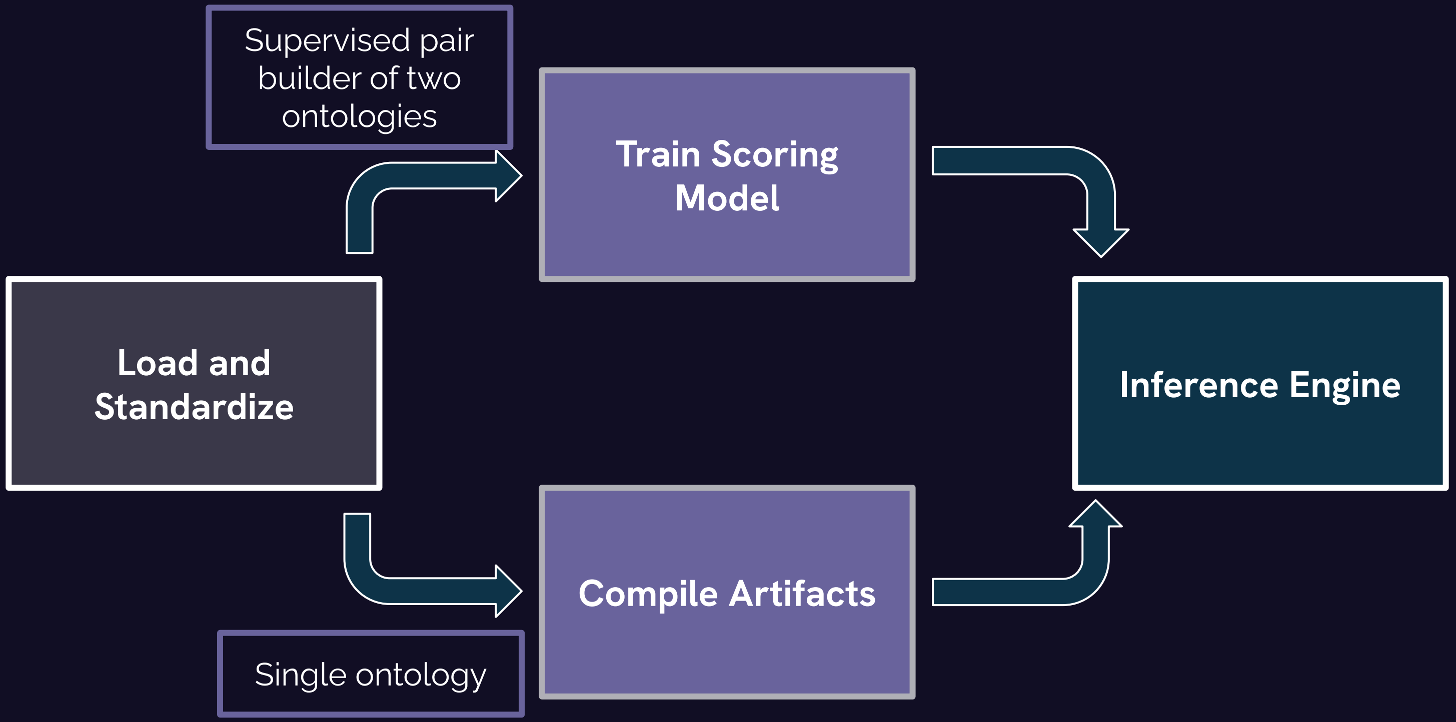
Data Strategy (H3)

Question: Is random training data sufficient for clinical accuracy?
Hypothesis: No. To generalize well, the model must be explicitly trained on "Hard Negatives"—tricky, similar-looking concepts—rather than just random non-matches.

SYSTEM DESIGN

03

SYSTEM DESIGN - MODULAR OVERVIEW



- Reusable across studies – build once, run many times

SYSTEM WORKFLOWS & ENTRY POINTS

1

Training

- Ontology loading & standardization
- Reference alignment loading
- Training dataset construction
- Model fine-tuning

From standardized ontology →

Offline Preprocessing

- Unified ontology representation
- Exact-match lookup structures
- Candidate retrieval indices
- Semantic representations (for hybrid retrieval)

2

3

Inference

Exact match shortcut → Candidate selection → Neural scoring

RETRIEVAL PIPELINE: PRE-SCORING DECISIONS

H1

Retrieval operates on offline-compiled ontology artifacts

Exact Match Shortcut

- Constant-time lookup on labels & synonyms
- Deterministic resolution for unambiguous cases
- Maximal score, no neural model invocation

Hybrid Candidate Selection

- Lexical retrieval via subword inverted index
- IDF-weighted lexical evidence
- Semantic retrieval via embedding similarity
- Unified top-k candidate pool for neural scoring

Hybrid retrieval mitigates structural candidate selection errors that cannot be recovered by downstream neural models

NEURAL SCORING: FINAL ALIGNMENT DECISION

H2

Joint encoding of attribute and candidate class

Full contextual interaction via Transformer attention

Produces a continuous semantic alignment score

Final ranking and top-1 match selection

Neural scoring focuses on precision, after recall has been ensured by hybrid retrieval

DATA ENGINEERING

04

ONTOLOGY LOADER FRAMEWORK

Load & standardize one or more ontologies into a unified representation

Ontology A

Ontology B

Ontology Loader

Unified Representation

- IRIs & hierarchy
- Labels & synonyms
- Definitions & descriptions
- Structural axioms

- Symbolic retrieval & exact match
- Rich textual input for model

DATASET CONSTRUCTION

H3

Convert ontologies and reference alignments into a supervised dataset. Each sample is a concept–concept pair labeled as match or non-match.

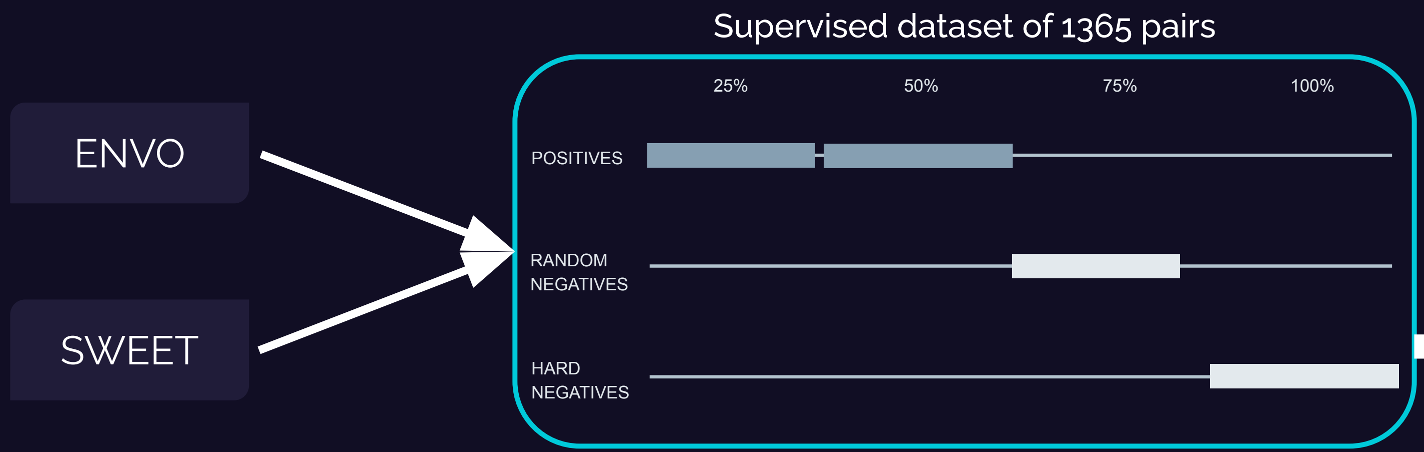
A training sample can be:

- **POSITIVE** = gold-standard alignment
- **NEGATIVE** = unrelated concept pair
- **HARD NEGATIVE** = semantically confusing non-match

EVALUATION & RESULTS

05

EVALUATION SETUP



1. **Retrieval quality** (Recall-oriented)
Is the correct ontology class present among the retrieved candidates?

2. **Final decision quality** (Precision-oriented)
Is the top-ranked prediction selected by the neural scorer correct?

136 attributes from SWEET ontology

MODEL COMPARISON

- All models trained on the same dataset with **hard negatives enabled**
- Models are evaluated under two retrieval modes: **Lexical** and **Hybrid**
- Model selection in based on **Precision**

Domain-specific pretraining consistently improves semantic alignment performance.
SciBERT achieves the best overall results and is selected as the reference model.

Lexical retrieval yields higher precision on this dataset due to strong lexical overlap.
Hybrid retrieval improves robustness by recovering semantically related candidates when lexical signals fail.

ABLATION STUDY: HARD NEGATIVES

H3

Ablation setup (SciBERT)
Same model, same training pipeline, same splits.
Only change: **Hard Negatives ON vs OFF**.

Retrieval Mode	**Precision HN**	**Precision NO HN**
Lexical	0.6667	0.6522
Hybrid	0.6232	0.5942

Accuracy HN	**Recall HN**	**Accuracy NO HN**	**Recall NO HN**
Threshold = 0.999345		Threshold = 0.997826	
0.9562	0.9565	0.8905	0.8551

Key result

Hard negatives improve **top-1 precision**, with a larger gain under **Hybrid retrieval**, where candidates are more semantically confusable.

Secondary analysis (thresholding)

When calibrated for binary decisions, hard negatives improve **match/no-match reliability**, especially **Recall** (fewer false negatives).

THRESHOLDING & BINARY DECISION (Secondary Analysis)

METHODOLOGY (THE HOW)

SETUP:
- Focuses on pairwise attribute–class classification
- Raw output: Continuous similarity scores (logits/probabilities).

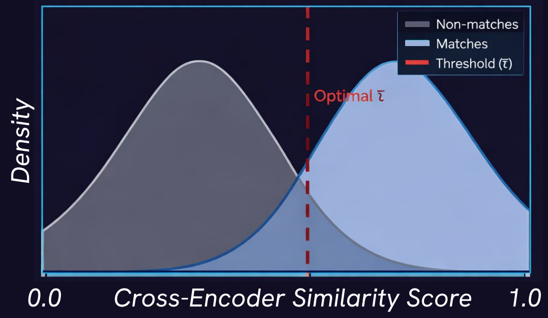
BINARY CONVERSION:
- Scores are converted into hard decisions (Match / No-Match).

THRESHOLD SELECTION (τ):
- Optimized on Validation Set.
- Objective: Maximize Accuracy.
- Applied uniformly to Test Set.

PURPOSE & CONTEXT (THE HOW)

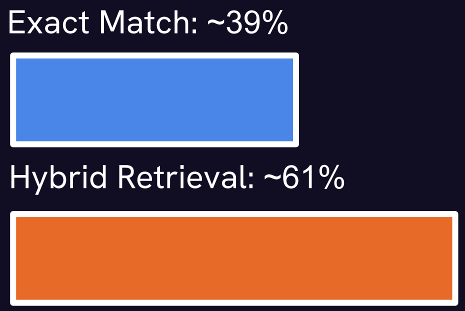
GOAL: Calibration & Interpretability
- Assess the discriminative power of the model in isolation
- Understand model confidence distribution

⚠ **SCOPE LIMITATION**
- Not used for end-to-end inference.



END-TO-END INFERENCE: EXACT vs HYBRID RETRIEVAL

RETRIEVAL SOURCE DISTRIBUTION



PERFORMANCE BREAKDOWN

EXACT MATCH

+ High precision for identical labels.
- Very limited coverage; fails on synonyms or rephrasing.

HYBRID RETRIEVAL

+ Higher recall via semantic embeddings.
+ Ensures the correct class is "caught" for the neural model.

KEY INSIGHT

While deterministic shortcuts provide easy wins, Hybrid Retrieval is essential for a scalable, robust, and domain-agnostic alignment system.

CONCLUSIONS

06

CONCLUSIONS

WHAT WE VALIDATED

We empirically evaluated a modular ontology alignment pipeline combining exact matching, hybrid candidate retrieval, and neural semantic scoring. Our results show that retrieval design and data construction critically shape downstream alignment performance.

KEY INSIGHTS FROM EVALUATION

Domain-specific cross-encoders outperform generic models in semantic ranking tasks. Hard negative mining improves robustness on semantically ambiguous cases. Hybrid retrieval mitigates structural errors that cannot be recovered by neural scoring alone.

FINAL VERDICT

Accurate ontology alignment does not rely on a single component. It emerges from a carefully engineered interaction between retrieval strategies, data design, and neural models, each addressing complementary sources of error.

THANK YOU

ANY QUESTIONS?