Descriptions are all you need: Semantic Vectorization of Hierarchical Medical Knowledge (ICD-11) via Large Language Models

Marco Lomele, Gleb Legotkin, Ilia Koldyshev,
Giorgio Caretti, Giovanni Mantovani, Leonardo Ruzzante

NLP Final Assignment

Bocconi University

May 2025

Introduction

Section 1

What is ICD-11, Really?

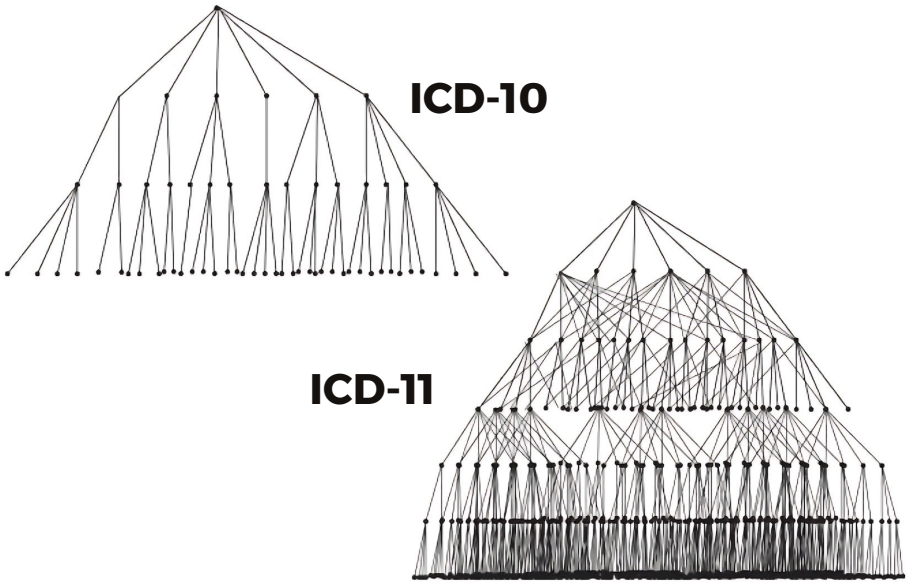
ICD-11: A Medical Knowledge Framework, Not Just a Code List

- Developed by the **WHO** as the global diagnostic standard
- Used in billing, epidemiology, clinical reporting, and ML models
- Contains over **40,000 disease entities**, organized into:
  - **Poly Hierarchies** (multiple parents)
  - **Extension codes** (e.g., severity, anatomy, cause)
  - **Linked relationships**, not just categories

ICD-11 vs ICD-10

Architectural Shift: Computational and Structural Challenges

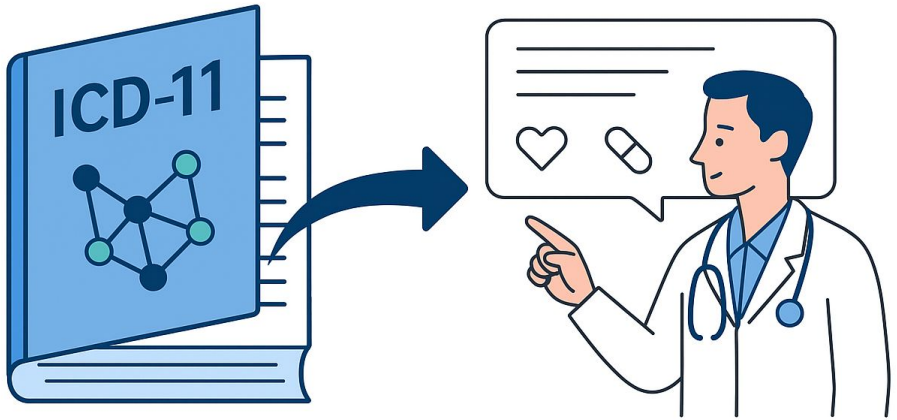
- ICD-11 introduces:
  - **Multiple inheritance**
  - **Post-coordination** with extension codes
  - **Web-like interrelations**, not just nested folders
- Why is ICD-11 hard to use in computational systems?
  - **Sparse** or absent natural language **descriptions**
  - **Hard to use** in semantic search, question answering, code prediction
  - Descriptions are **non-uniform**, often symbolic, and **fragmentary**



Project Objective

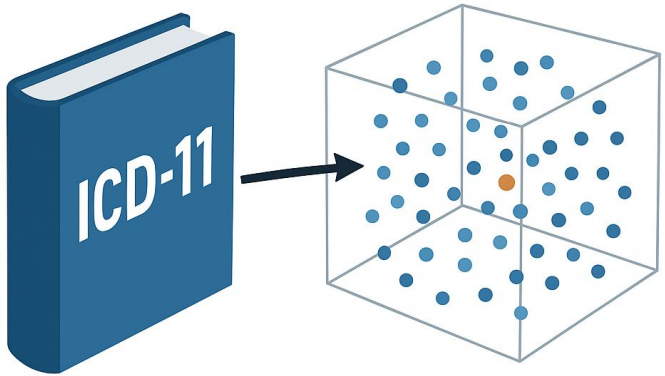
First Vectorization of ICD-11, as a foundation for interpretable natural language search engine and privacy-preserving medical ML

- Solves ICD-11's **lack** of complete **descriptions** and **semantic representations**
- Enables **natural language querying** of diseases, symptoms, and interventions
- Part of **Open Doctor**, a privacy-first ML project for real-time medical assistance, where all models run **locally**, keeping patient data **secure** and **compliant**



Roadmap of Our Paper

- Section 1 - **Introduction**
- Section 2 - **Related Works**: Literature review on ICD embeddings and LLMs
- Section 3 - **Experiment**: Data extraction, enrich ICD-11 descriptions, prompt design, text generation, validation
- Section 4 - **Results and Discussion**:
  - Comorbidity alignment
  - Symptom-disease matching
  - Hierarchical consistency
- Section 5 - **Conclusions**: Contributions, limitations, and future directions



Related Works

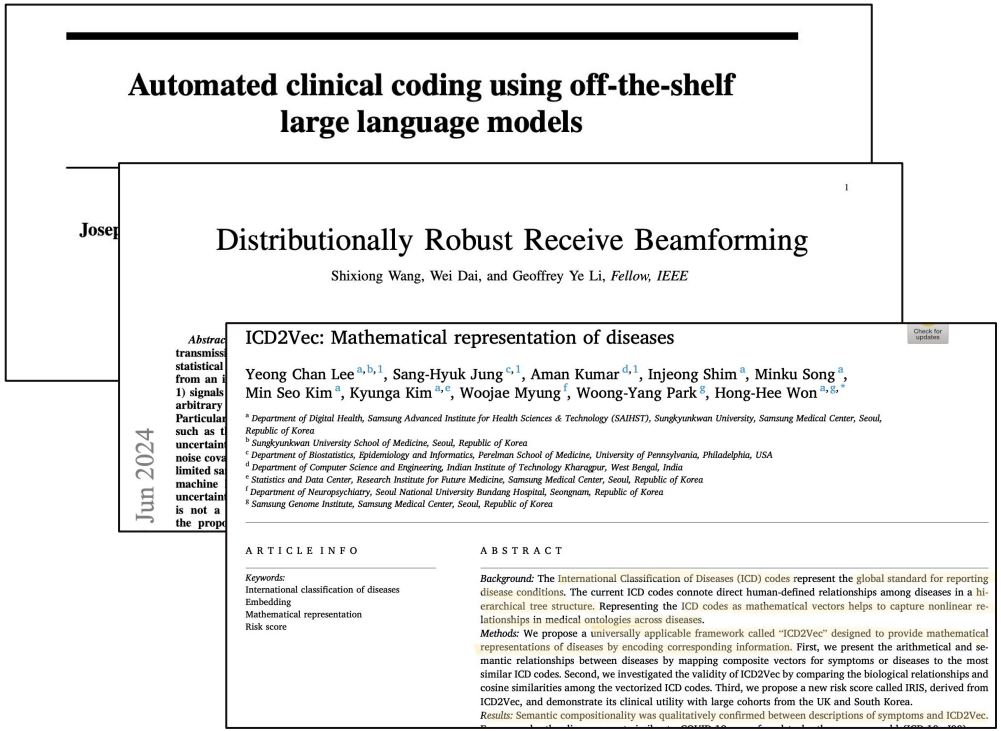
Section 2

Key Papers

What We Learned from Prior Literature

- **ICD2Vec** (Lee et al., 2023): **Good for ICD-10** comorbidities no fit for ICD-11
- **TF-IDF** (Feng et al., 2024): Simple and effective for retrieval; **fails** on semantics and **sparse descriptions**
- **WordNet / GPT-4** (Chen & Xu, 2020; Klotzman, 2024): Semantic enrichment works; **LLMs are closed-source**
- **PLM-ICD, DiLBERT, BigBird**: Powerful on ICD-9/10 clinical text; **not designed** for taxonomy embedding or **ICD-11 structure**

Common Limitation:
Focused on clinical documents, not on enriching the ICD-11 ontology itself



Automated clinical coding using off-the-shelf large language models

Distributionally Robust Receive Beamforming

Shixiong Wang, Wei Dai, and Geoffrey Ye Li, *Fellow, IEEE*

Jun 2024

ICD2Vec: Mathematical representation of diseases

Yeong Chan Lee [a,b,1], Sang-Hyuk Jung [c,1], Aman Kumar [d,1], Injeong Shim [a], Minku Song [a], Min Seo Kim [a], Kyunga Kim [a,e], Woojae Myung [f], Woong-Yang Park [g], Hong-Hee Won [a,g,*]

[a] Department of Digital Health, Samsung Advanced Institute for Health Sciences & Technology (SAIHST), Sungkyunkwan University, Samsung Medical Center, Seoul, Republic of Korea
[b] Sungkyunkwan University School of Medicine, Seoul, Republic of Korea
[c] Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA
[d] Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, West Bengal, India
[e] Statistics and Data Center, Research Institute for Future Medicine, Samsung Medical Center, Seoul, Republic of Korea
[f] Department of Neuropsychiatry, Seoul National University Bundang Hospital, Seongnam, Republic of Korea
[g] Samsung Genome Institute, Samsung Medical Center, Seoul, Republic of Korea

ARTICLE INFO

Keywords:
International classification of diseases
Embedding
Mathematical representation
Risk score

ABSTRACT

Background: The International Classification of Diseases (ICD) codes represent the global standard for reporting disease conditions. The current ICD codes connote direct human-defined relationships among diseases in a hierarchical tree structure. Representing the ICD codes as mathematical vectors helps to capture nonlinear relationships in medical ontologies across diseases.
Methods: We propose a universally applicable framework called "ICD2Vec" designed to provide mathematical representations of diseases by encoding corresponding information. First, we present the arithmetical and semantic relationships between diseases by mapping composite vectors for symptoms or diseases to the most similar ICD codes. Second, we investigated the validity of ICD2Vec by comparing the biological relationships and cosine similarities among the vectorized ICD codes. Third, we propose a new risk score called IRIS, derived from ICD2Vec, and demonstrate its clinical utility with large cohorts from the UK and South Korea.
Results: Semantic compositionality was qualitatively confirmed between descriptions of symptoms and ICD2Vec.

Experiment

Section 3

Data Extraction

Accessing Data via WHO API

WHO provides a public REST API for querying the ICD-11 hierarchy.

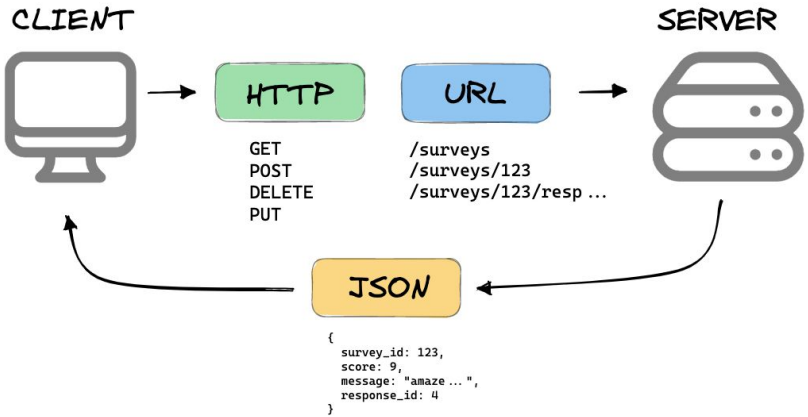
Endpoint example: https://icd.who.int/icdapi

Requires **API key** from WHO (free with registration).

Returns data in **JSON-LD format**, including:

- @id (URI of the entity)
- title, definition, synonyms
- Hierarchical links: child, parent, is-a

Crawling of the ICD-11 Structure
using a **Breadth-First Search (BFS)** approach



Data Generation

Original ICD-11 entries lacked sufficient detail

We enrich the descriptions via **Llama3-OpenBioLLM-70B model** trained on domain-specific corpora

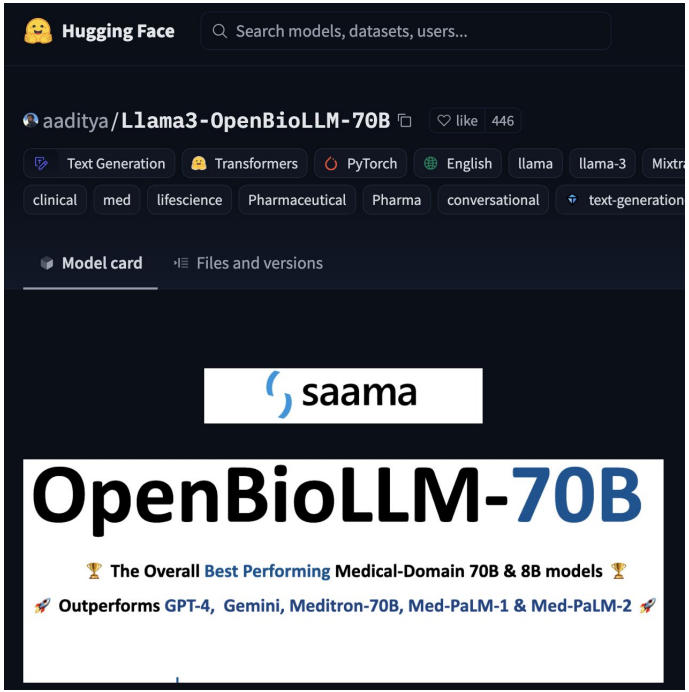
Generation Pipeline:

1) **System prompt** instructs LLM on format and scope
2) Disease names (from ICD-11) fed as **user queries**
3) Outputs saved and linked to original ICD codes

```
SYSTEM_MESSAGE = """
Your name is Llama3-OpenBioLLM-70B. You are an expert and experienced from the healthcare and biomedical domain with extensive medical knowledge.
Your mission is to provide comprehensive, technical, and accurate medical description descriptions of diseases and disease categories.
The user will input the name of a disease or the name of a category of diseases.
You will provide the description of the query.
Always structure the sentences of your response in this order: overview, causes, symptoms, transmission, diagnosis.
Write full sentences, using a concise and clear language.
"""
```

```
    response = client_nebius.chat.completions.create(
        model="aaditya/Llama3-OpenBioLLM-70B",
        temperature=temperature,
        messages=[
            {"role": "system", "content": SYSTEM_MESSAGE},
            {"role": "user", "content": f"Describe {query}?"}
        ],
        max_completion_tokens=800
    )

    return response.choices[0].message.content
```



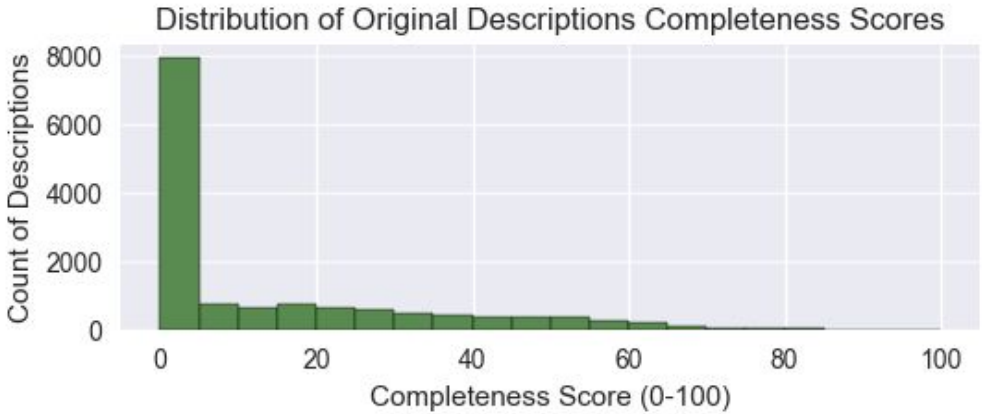
Linguistic Quality

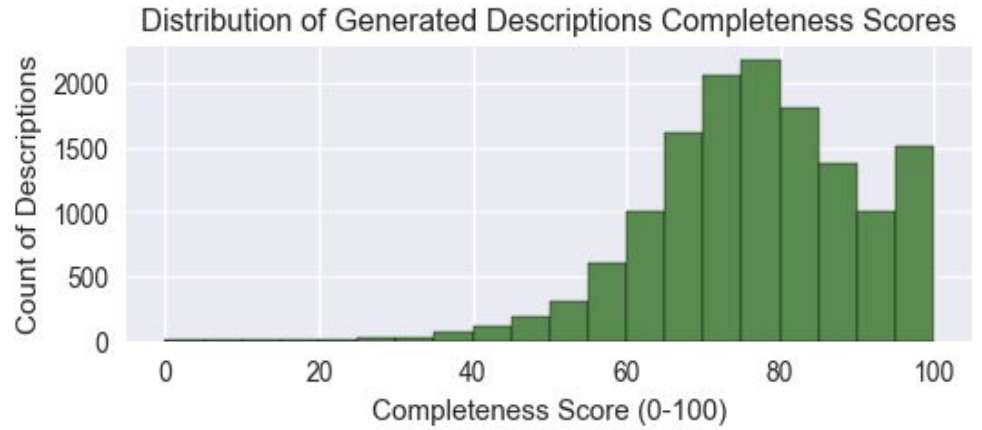
Original ICD-11 Descriptions

- **Sparse** and often **missing** (50% are empty)
- Average length: **154 characters**
  Low lexical diversity
- The mean **completeness score**: **14%**

LLM-Enriched Descriptions

- **No missing entries** (100% coverage)
  Average length: **800 characters**
- **High lexical diversity**
- The mean **Completeness score**: **78%**





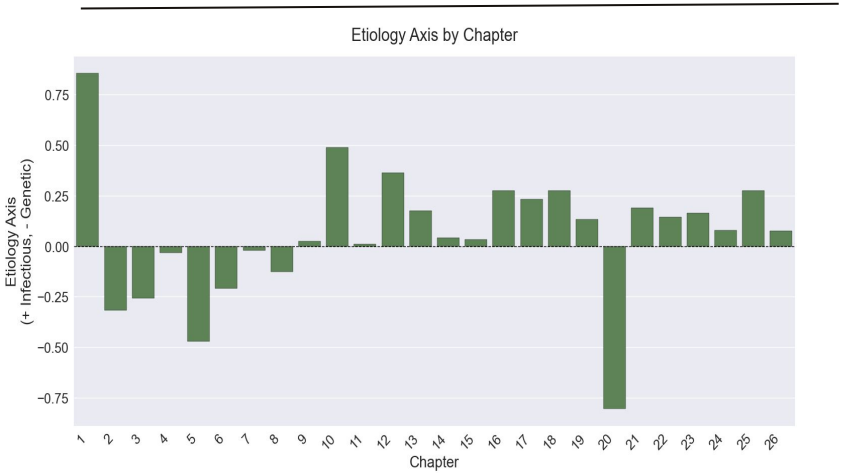
Medical Validation

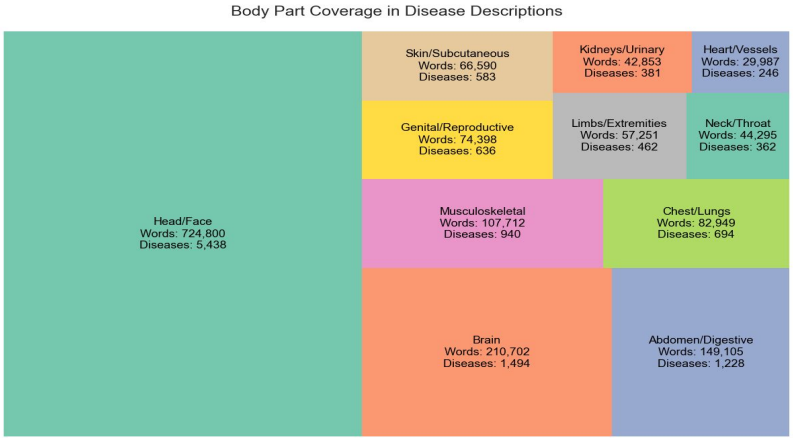
ICD-11 descriptions are medically accurate and meaningful across categories.

Built a custom **medical dictionary** with terms for: **Causes, Symptoms, Transmission and Diagnosis**
And assessed that **after the enrichment 6× more causal terms**

Medical Intuition has been captured:
Chapter 1 (Infectious Diseases): ↑ 25% content on *transmission*
Chapter 22 (External Causes): ↑ 62% focus on *causes*

Applied **vector clustering** to ensure that generated descriptions align with real-world disease categories





Hierarchical Analysis

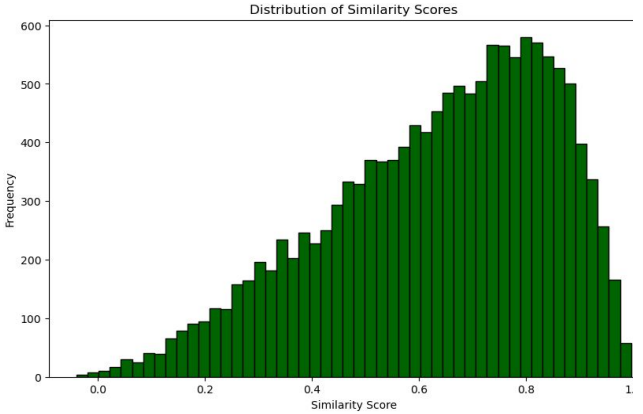
ICD-11 is organized as a multi-level tree

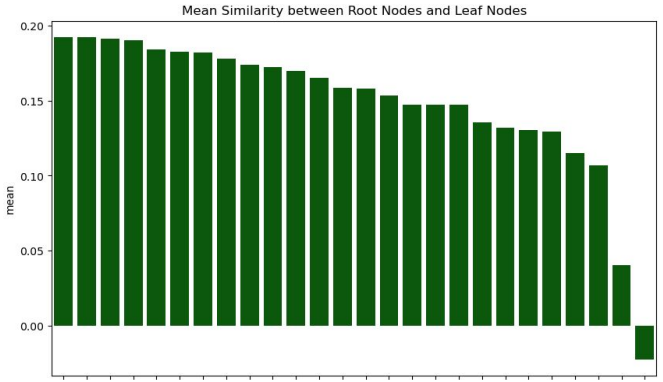
Text Information is passed through parent-child links. Diseases in the same branch tend to share causes, symptoms.

Graph features:
Average **depth**: 3.54 levels
Average **branching factor**: 4.25 children per node
all-MiniLM-L6-v2 embeddings on titles as proxy of enriched descriptions:

Average **Cosine similarity** between **parent-child pairs: 0.636**
Average **Cosine similarity** between **root and leaf nodes: 0.148**

The hierarchical structure enables easier **semantic search**, **clustering**, and **retrieval** in biomedical tasks





Embedding Models

Traditional Methods

Karen Spärck Jones

TF-IDF

1972, statistical model, corpus dependent

FastText

2015, ≈sub-word Word2Vec, variable parameters, custom corpora

Transformer Encoders

BERT

2018, 110M parameters, book corpus + english Wikipedia

PubMedBERT

2020, 108M parameters, PubMed corpus

GatorTron

2022, 8.9B parameter, clinical text + public medical data

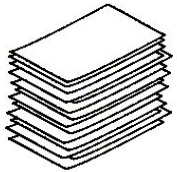
Bio-BERT

2018, PubMed abstracts + PMC full-text

BioClinicalBERT

2019, Clinical notes + MIMIC-III corpus

Experimental Setup

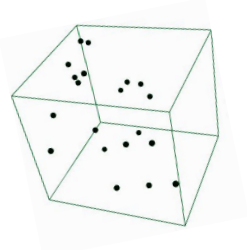


Raw Generated Descriptions

2. NLTK text preprocessing
- Light preprocessing.
- Lowercase, special characters removal, white space normalisation.

1. ICD-11 data appending
- Diagnostic synonyms.
- Related conditions.
- Inclusions & exclusions.
- Child & parent relationships.

3. Model specification
- TF-IDF: 3000 vocab, SVD reduction.
- FastText: skip-gram, window size 5.
- Transformers: pre-trained weights.

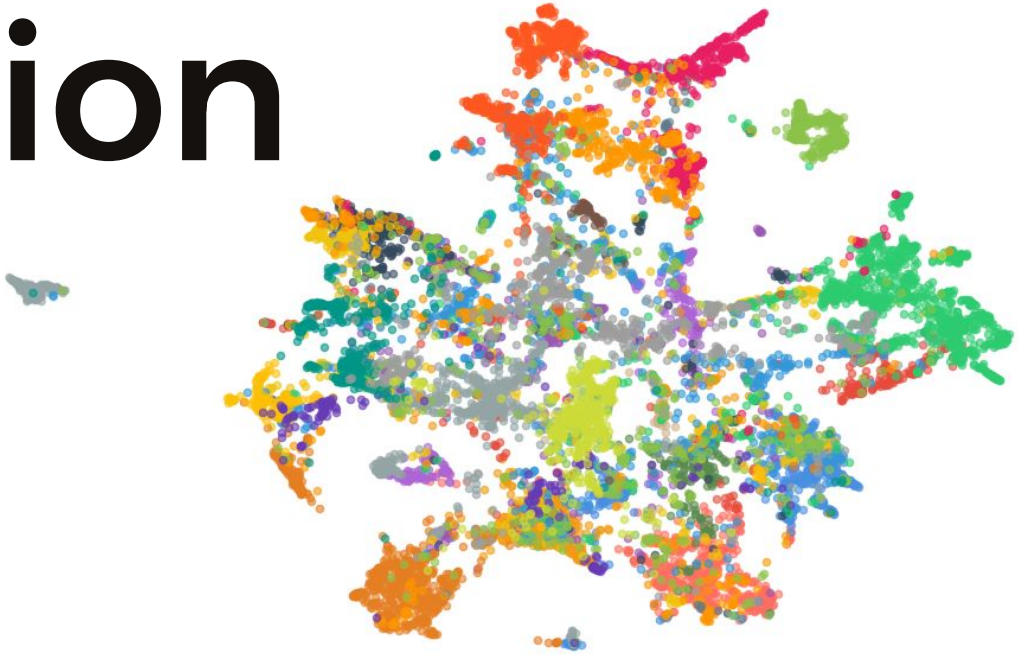


x 7

768-Dimensional Embeddings

Results & Discussion

Section 4

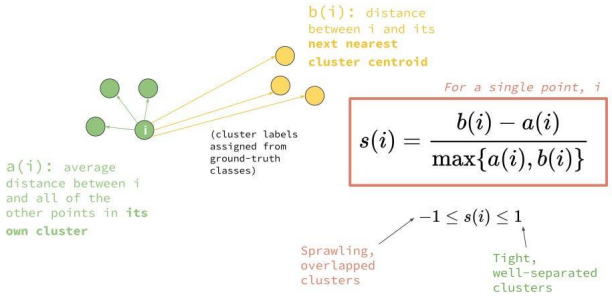


Intrinsic Evaluation

High-Dimensional Space

Model	Silhouette Score	Calinski-Harabasz Score	Davies-Bouldin Score
BERT	0.0859	171.5505	4.0329
BioBERT	0.0164	119.8868	4.1165
BioClinicalBERT	0.0839	123.0997	4.2670
PubMedBERT	0.0215	135.2616	3.8855
TF-IDF	0.0013	45.8275	5.6999
FastText	0.0134	230.7401	3.4326
GatorTron	0.0315	135.8351	3.8102

Table 2: Embedding Quality Metrics



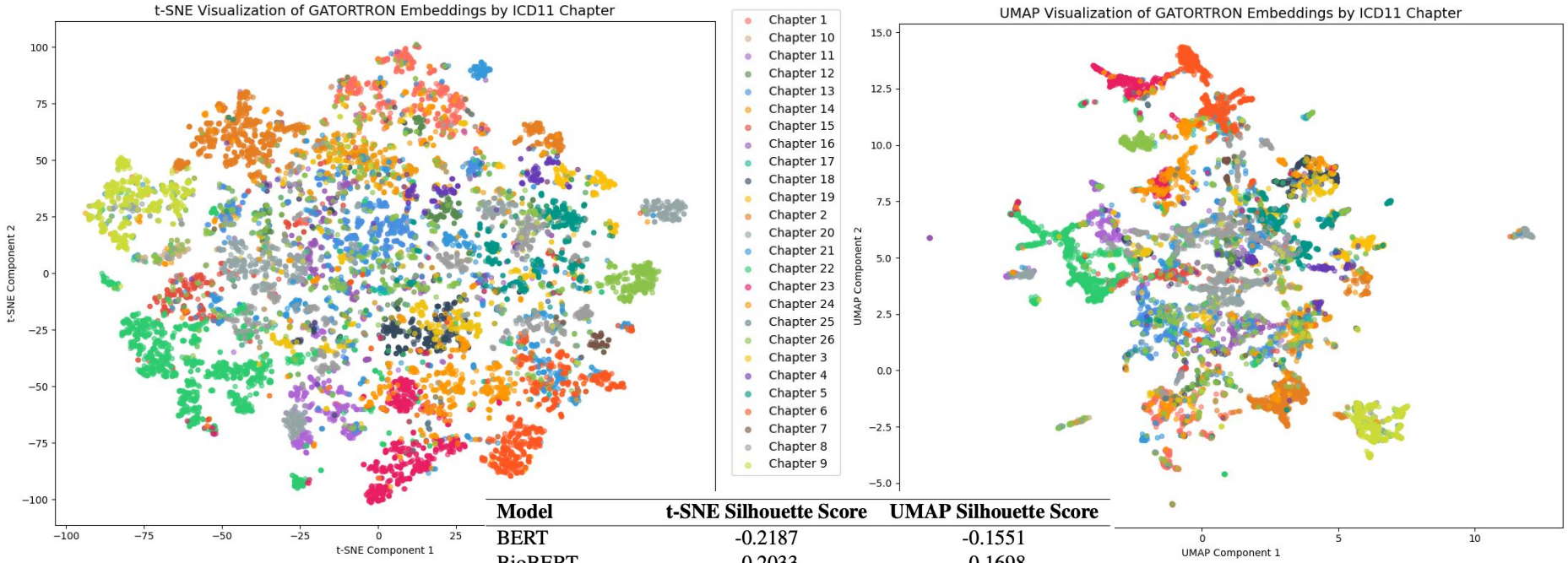
$$CH = \frac{BCSS/(k-1)}{WCSS/(n-k)}$$

BCSS = Between Cluster Separation
WCSS = Within Cluster Separation

$$DB = \frac{1}{k}\Sigma_{i=1}^{k} max\left(\frac{\Delta(X_i)+\Delta(X_j)}{\delta(X_i,X_j)}\right)$$

ΔX_i = intracluster distance
δ = intercluster distance

Low-dim. Visualisations



Model	t-SNE Silhouette Score	UMAP Silhouette Score
BERT	-0.2187	-0.1551
BioBERT	-0.2033	-0.1698
BioClinicalBERT	-0.1910	-0.1780
PubMedBERT	-0.1315	-0.0836
TF-IDF	-0.3374	-0.2890
FastText	-0.1195	-0.1406
GatorTron	-0.0738	-0.0936

Intrinsic Evaluation

Conversion quality.

Model	t-SNE Trustworthiness	t-SNE Continuity	UMAP Trustworthiness	UMAP Continuity
BERT	0.9327	0.9001	0.9164	0.9148
BioBERT	0.9128	0.9004	0.8982	0.9093
BioClinicalBERT	0.9180	0.8908	0.8944	0.9041
PubMedBERT	0.9234	0.8974	0.9068	0.9100
TF-IDF	0.7656	0.8148	0.7666	0.8202
FastText	0.9487	0.9273	0.9376	0.9323
GatorTron	0.9224	0.9052	0.9181	0.9193

Table 3: Dimensionality Reduction Quality Metrics

$$T(k) = 1 - \frac{2}{Nk(2N - 3k - 1)} \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i^k} (r(i,j) - k)$$

$\mathcal{N}_i^k$ Neighbours of point i in low-dimensional space and **not** in original k-neighbourhood.

$$C(k) = 1 - \frac{2}{Nk(2N - 3k - 1)} \sum_{i=1}^{N} \sum_{j \in \mathcal{M}_i^k} (s(i,j) - k)$$

$\mathcal{M}_i^k$ Neighbours of point i in original space **not** in low-dimensional k-neighbourhood.

Comorbidity Score

Comorbidity score quantifies the **mortality risk** from having **multiple diseases**

Comorbidity Benchmark Score: **correlation coefficient** between **cosine similarity** of diseases embeddings and their **comorbidity score**

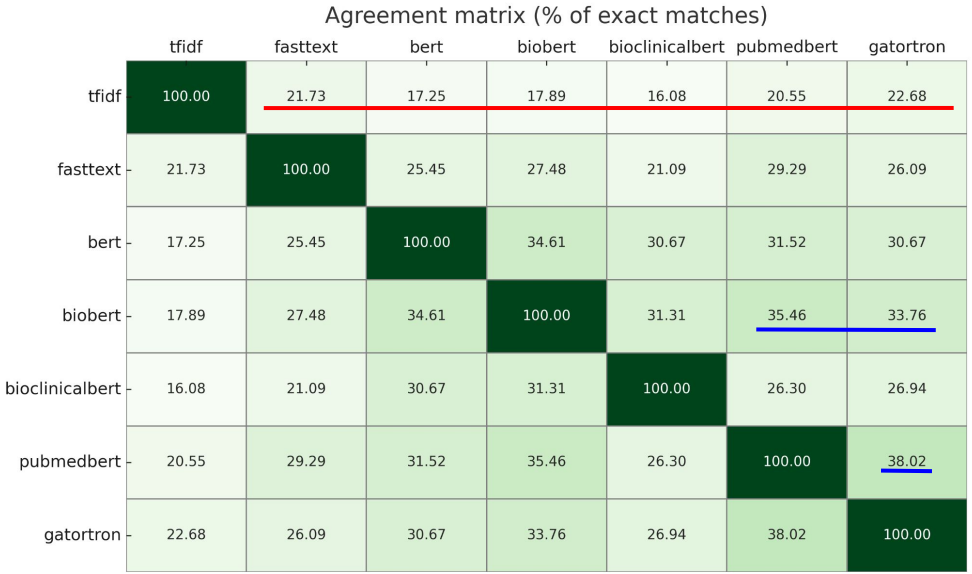
Studies suggest **strong positive correlation**

Model	Comorbidity Benchmark Score
TF-IDF	0.2801
FastText	0.0966
BERT	0.1332
BioBERT	0.1260
BioClinicalBERT	0.0972
PubMedBERT	0.1706
GatorTron	0.1886

Table 1: Comorbidity benchmark score across models

Symptom-Disease Matching

For each of 940 ICD-11 **symptoms**, we identify the **disease with the highest cosine similarity** in the embedding space



Model	Mean	Standard Deviation
TF-IDF	0.6158	0.1340
FastText	0.9754	0.0101
BERT	0.9600	0.0137
BioBERT	0.9791	0.0076
BioClinicalBERT	0.9768	0.0068
PubMedBERT	0.9952	0.0017
GatorTron	0.9648	0.0138

Table 5: Mean and standard deviation of cosine similarity of top disease across symptoms

Encyclopedia Definition Benchmark

We used definitions of diseases from the **Merriam-Webster Medical Dictionary** and embedded them using various language models.

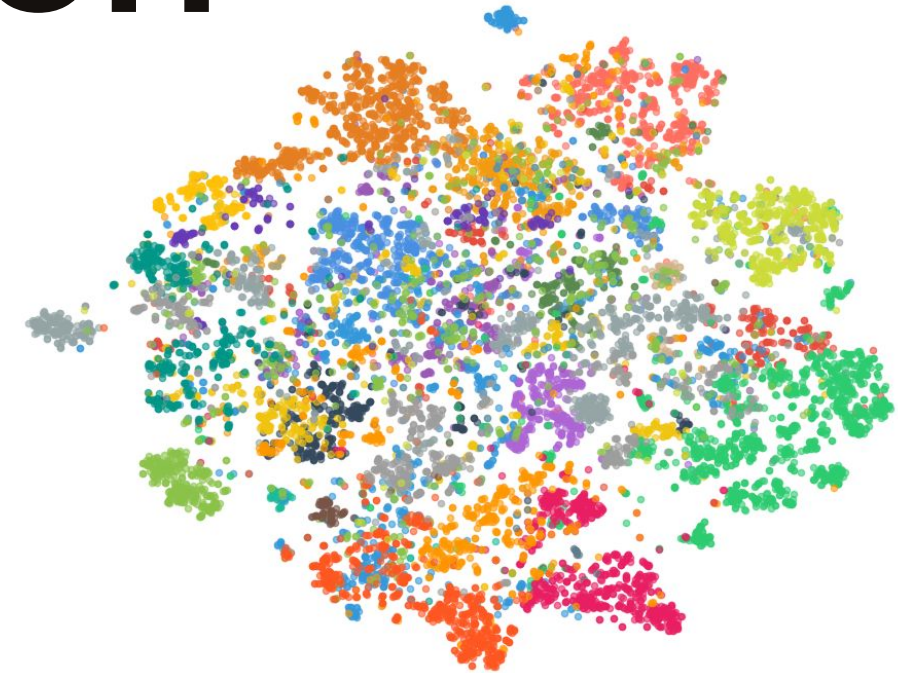
For each definition, we identified the **top three ICD-11 codes** based on cosine similarity between embeddings. We compared then the top predicted codes with the true ICD-11 code of the disease, counting **how many initial symbols matched**.

ICD-11 codes are hierarchical: matching initial symbols indicates a closer relationship between diseases.

Model	4 Symbols	3 Symbols	2 Symbols	1 Symbol	No Match
TF-IDF	1.06%	8.47%	15.61%	31.48%	68.52%
FastText	0.79%	2.12%	4.23%	16.67%	83.33%
BERT (base)	51.85%	64.81%	72.49%	83.07%	16.93%
BioBERT	66.14%	77.51%	82.01%	89.95%	10.05%
BioClinicalBERT	54.50%	65.34%	70.90%	82.54%	17.46%
PubMedBERT	75.13%	83.33%	86.24%	92.33%	7.67%
GatorTron	2.91%	11.11%	26.19%	50.79%	49.21%

Conclusion

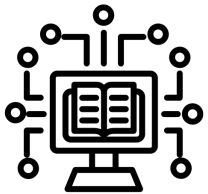
Section 5



Contributions

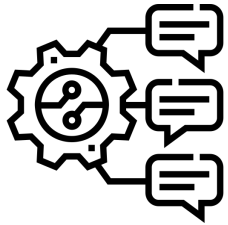
1. LLM-Augmented ICD-11 Dataset

- **Enriched** using **Llama3-OpenBioLLM-70B**: completeness ↑ **14** % → **78** %
- Validated **linguistically**, **semantically**, and **medically**



2. 7-model ICD-11 vector index

- **Seven embedding pipelines** spanning TF-IDF, FastText, Bert, BioBERT, BioClinicalBERT, PubMedBERT, GatorTron
- **PubMedBERT** excels at **deep semantics** & **hierarchy**, whereas **TF-IDF** leads in **comorbidity** and **FastText** maintains **robust performance** across embedding dimensions.



Non-Medical Queries

Prompt Benchmark: Everyday Language

Possible examples of **“Burns”** prompt:

- “i accidentally spilled boiling water on my arm and now it s red and blistering what should i do”
- “my kid touched a hot pan skin looks bad need help fast”
- “i fell asleep in the sun and now my whole back is bright red what now”

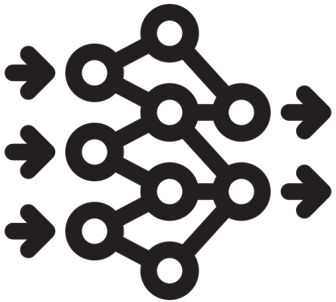
Model	Top-3 Matches	Acc.	Avg Sim.	Min Sim.	Max Sim.
BioClinicalBERT	21	7.0%	0.885	0.857	0.926
BERT	19	6.3%	0.665	0.580	0.752
BioBERT	14	4.7%	0.891	0.866	0.914
PubMedBERT	6	2.0%	0.970	0.958	0.980
TF-IDF	4	1.3%	0.467	0.360	0.638
FastText	0	0.0%	0.438	0.437	0.438
GatorTron	0	0.0%	0.939	0.885	0.964

Table 8: Model performance based on keyword matching between prompt text and retrieved disease titles

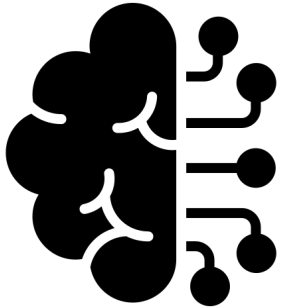
Key insight: current embeddings overfit to technical style.

Future Work

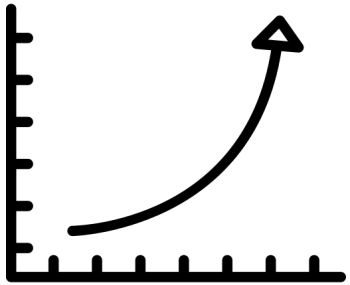
This limitation points to three directions for future work:



Integrate self-supervised Graph Neural Networks (GNNs) (e.g., link prediction) to model hierarchical and cross-disease relationships more effectively.



Leverage foundation models (GPT-4.1, Claude Sonnet 4, Grok 3) to generate richer, context-aware embeddings.



Investigate scaling laws by systematically varying embedding size, network capacity, and input length to boost performance.

Thank you

Questions? We are happy to answer!

Marco Lomele

Gleb Legotkin

Ilia Koldyshev

Giorgio Caretti

Giovanni Mantovani

Leonardo Ruzzante

Resources

Which have not been already cited on the report.

- Farshad K. Understanding Silhouette Score in Clustering. Medium Article 2024. https://farshadabdulazeez.medium.com/understanding-silhouette-score-in-clustering-8aedc06ce9c4
- Stasis S, Stables R, Hockman J. Semantically Controlled Adaptive Equalisation in Reduced Dimensionality Parameter Space. *Applied Sciences*. 2016; 6(4):116. https://doi.org/10.3390/app6040116.

Appendix.

Which have not been already appended in the report.