Measuring Thematic Drift in Medical AI: Alignment of Publications with Aims & Scope (2015–2025)

Samet Cagan, MSc DSE, University of Milan
samet.cagan@studenti.unimi.it

KEY WORDS:
Thematic Drift,
Semantic Similarity,
S-Bert,
Embeddings,
Cosine Similarity,
K-Means,
TF-IDF
NLP

Why Medical context

Medical AI has changed a lot over the last decade, and the focus of the field keeps shifting as new methods become popular. Around 2015, a big part of Medical AI work centered on machine learning using electronic health records (EHRs) and structured clinical data.

Then the field moved into deep learning for medical imaging, especially in areas like radiology and pathology. Most recently, there has been a major shift toward Generative-AI, especially large language models (LLMs) that work on text, clinical notes, and multimodal data.

Core Objective

Build a reproducible NLP pipeline that:

1. Measures thematic alignment of published research against a reference mission statement (Aims & Scope)
2. Identifies major semantic pillars in the literature using unsupervised clustering
3. Tracks how the market share of these pillars changes over time to quantify thematic drift

Research Questions & Hypotheses

1. **RQ — Thematic Alignment**

   Does Medical AI research remain aligned with the journal's intended scope over time?
   Metric: cosine similarity between each paper embedding and the Aims & Scope embedding.

   Hypothesis: Overall alignment stays high across years.

2. **RQ — Thematic Drift (Composition Change)**

   How does the internal topic composition of Medical AI research change from 2015 to 2025?
   Metric: yearly topic distribution using normalized topic market share.

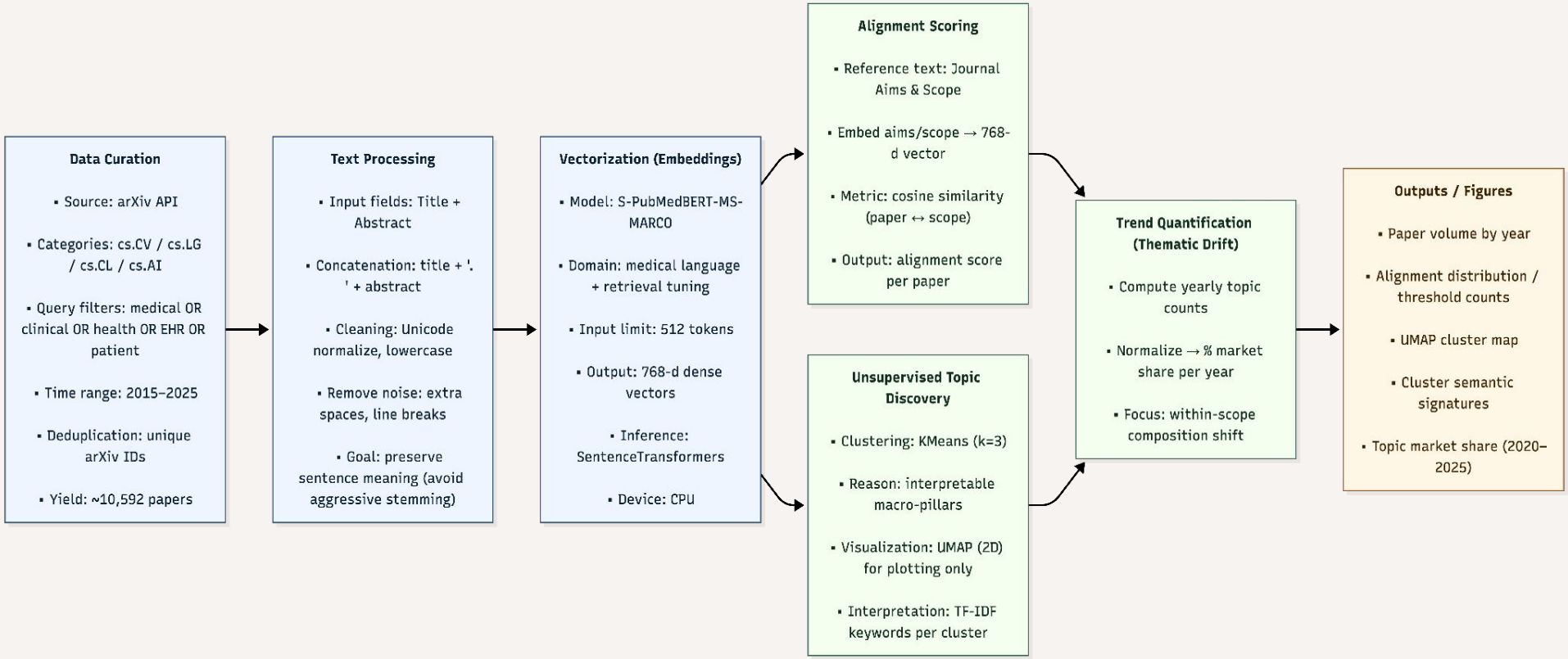
   Hypothesis: Topic proportions shift significantly over time.

3. **RQ — Topic Shift (Emerging Dominant Pillar)**

   Is there evidence of an emerging dominant research pillar, such as LLM-based methods?
   Metric: growth trend of the GenAI/LLM cluster share over time.

   Hypothesis: The GenAI/LLM topic shows the strongest growth

Pipeline


Data Curation
▪ Source: arXiv API
▪ Categories: cs.CV / cs.LG / cs.CL / cs.AI
▪ Query filters: medical OR clinical OR health OR EHR OR patient
▪ Time range: 2015–2025
▪ Deduplication: unique arXiv IDs
▪ Yield: ~10,592 papers
Text Processing
▪ Input fields: Title + Abstract
▪ Concatenation: title + '. ' + abstract
▪ Cleaning: Unicode normalize, lowercase
▪ Remove noise: extra spaces, line breaks
▪ Goal: preserve sentence meaning (avoid aggressive stemming)
Vectorization (Embeddings)
▪ Model: S-PubMedBERT-MS-MARCO
▪ Domain: medical language + retrieval tuning
▪ Input limit: 512 tokens
▪ Output: 768-d dense vectors
▪ Inference: SentenceTransformers
▪ Device: CPU
Alignment Scoring
▪ Reference text: Journal Aims & Scope
▪ Embed aims/scope → 768-d vector
▪ Metric: cosine similarity (paper ↔ scope)
▪ Output: alignment score per paper
Unsupervised Topic Discovery
▪ Clustering: KMeans (k=3)
▪ Reason: interpretable macro-pillars
▪ Visualization: UMAP (2D) for plotting only
▪ Interpretation: TF-IDF keywords per cluster
Trend Quantification (Thematic Drift)
▪ Compute yearly topic counts
▪ Normalize → % market share per year
▪ Focus: within-scope composition shift
Outputs / Figures
▪ Paper volume by year
▪ Alignment distribution / threshold counts
▪ UMAP cluster map
▪ Cluster semantic signatures
▪ Topic market share (2020–2025)

Data Curation

- I curated the dataset from arXiv using an automated loader. I restricted papers to core AI categories and medical keywords, extracted title/abstract metadata, and stored results in a DataFrame with publication year.

Why arXiv?

- Open and easy to query programmatically
- Provides consistent metadata: title, abstract, date, categories
- This makes it well-suited for studying topic evolution over time.

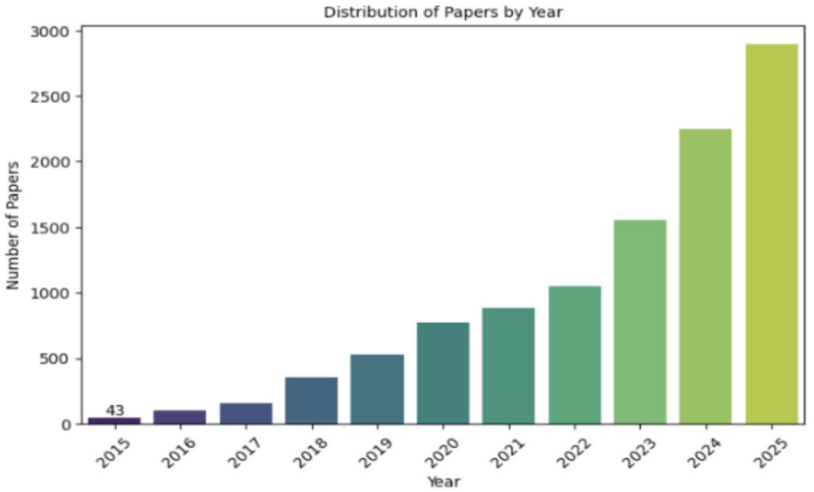
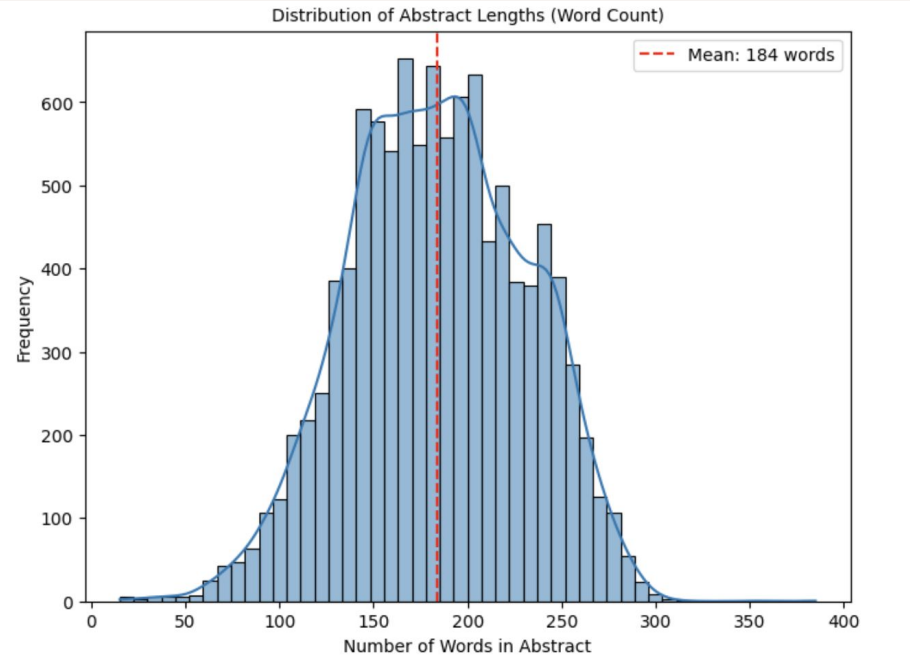


Figure 1: Distribution of arXiv Medical AI papers per year (2015-2025).

Text Processing

- I processed the text by combining title and abstract;
- Removing formatting noise,
- normalizing Unicode, and lowercasing.

- A key decision is that I did not remove stop words or numbers. Since I am using a Transformer embedding model, the model benefits from complete sentence structure, and removing common words can reduce the meaning of the text.



Model Selection

Aspect	BERT-base	S-PubMedBERT
Training domain	General text	Biomedical literature (PubMed-style language)
Strength	Broad language coverage	Strong biomedical terminology + context
Embedding suitability	Not optimized for sentence/document similarity by default	Designed for semantic similarity (sentence-transformers style)
Impact on clustering	Risk of noisy or shallow topic separation	Cleaner separation into meaningful pillars (CV / EHR / LLMs)
Why it matters here	We need stable semantic geometry for cosine similarity + topic drift	Preserves biomedical meaning while keeping technical structure

Defining “Ground Truth”

“””The Journal publishes high-quality research at the intersection of clinical medicine and artificial intelligence. The scope includes the application of machine learning, deep learning, computer vision, and natural language processing to medical diagnostics, patient outcome prediction, treatment optimization, and healthcare informatics. We prioritize studies that demonstrate clinical utility and methodological novelty in processing medical images, electronic health records, and genomic data. “””

- Created a reference **Aims & Scope** mission text
- Converted it into a dense embedding **Reference Vector** (field “center of gravity”)
- Embedded each paper (title + abstract) → **Paper Vector**
- Computed **Alignment Score** with **cosine similarity** *(angle between vectors)*
- Higher score → paper is semantically closer to the aim&scope

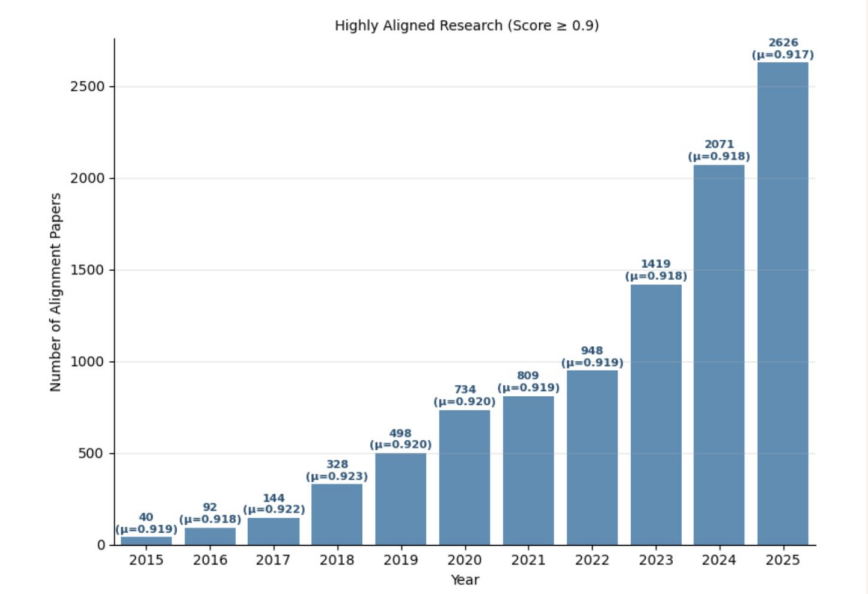


Figure 2: Alignment score of arXiv Medical AI papers per year (2015-2025).

Unsupervised Clustering (UMAP + KMeans + TF-IDF)

- Applied **KMeans** directly on the **768D embedding vectors** to group papers into semantic topics
- Set **3 cluster**, (more cluster caused similar clusters)
- Used **UMAP** to project embeddings into **2D** for visualization only *(clustering remains in the original embedding space)*
- I used **TF–IDF** keyword extraction on the text within each cluster and then labeled the clusters based on their strongest keywords.

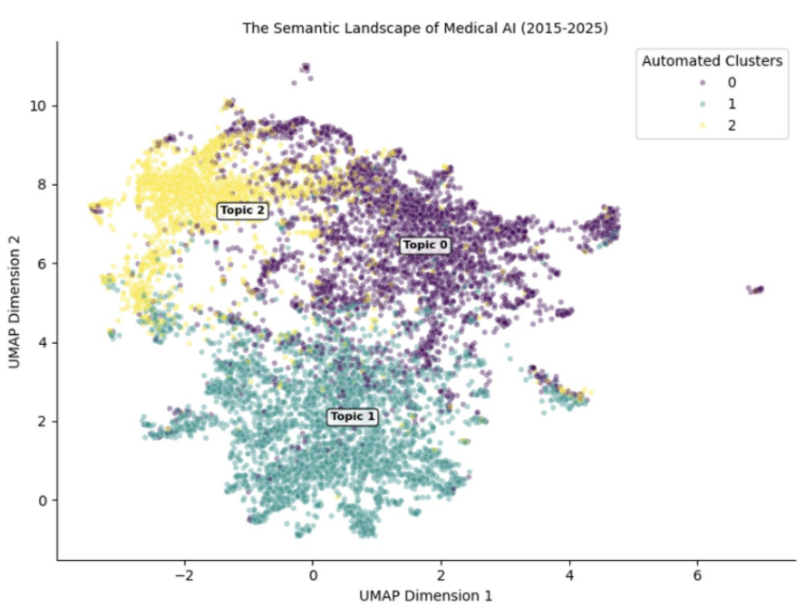
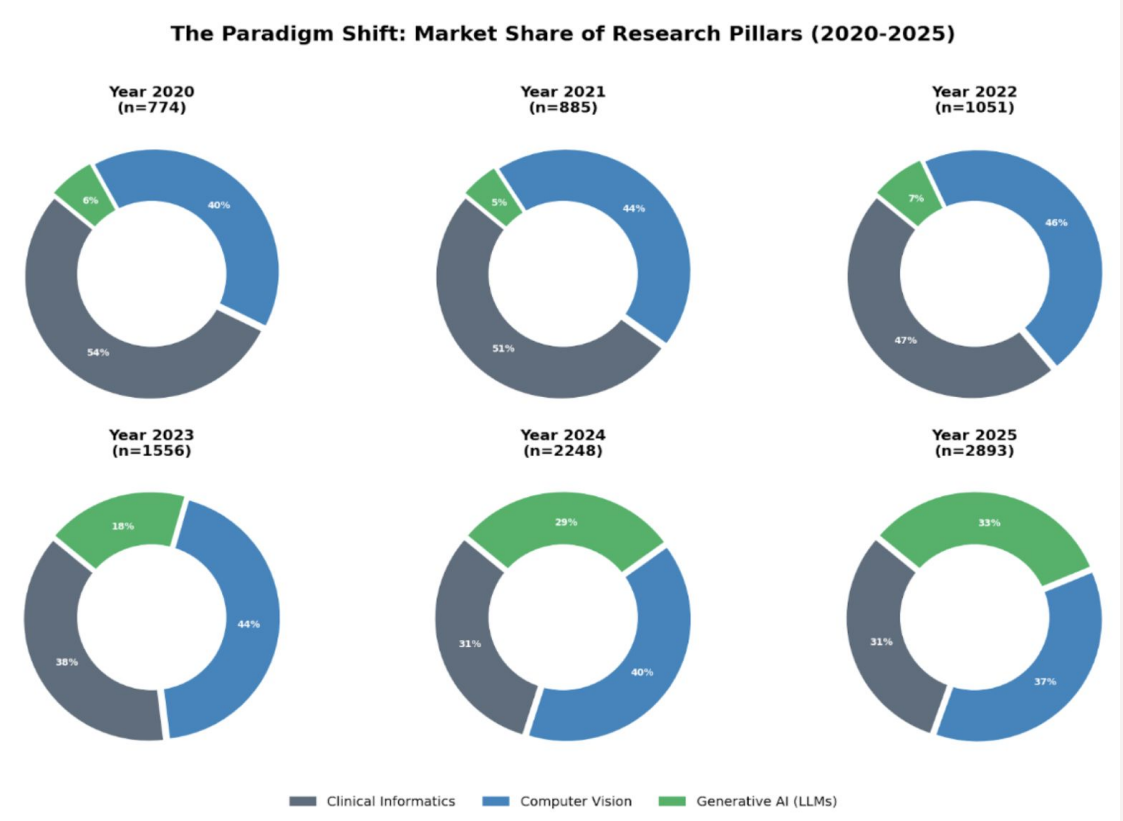


Figure 3: Paper embeddings colored by KMeans cluster (3 topics).

Table 1: Cluster semantic signatures (TF-IDF top terms)

Topic	Label	Top Keywords
Topic 0	Clinical Informatics	data, clinical, health, learning, medical, model, patient, based, patients
Topic 1	Computer Vision	medical, image, segmentation, learning, data, images, model, performance
Topic 2	GenAI (LLMs)	medical, models, clinical, language, llms, large, data, performance

Topic Evolution and Trend Analysis



- After defining the 3 semantic clusters, I checked how Medical AI priorities change over time.
- Since publication volume grows each year, I used topic **market share** instead of raw counts.
- This highlights **real shifts in attention** rather than simple growth in output.
- The analysis shows which topics are **rising** and which are becoming **less dominant**.
- This snapshots offer a quick visual of how topic dominance shifts between 2020 and 2025.
- A rapidly growing share is now driven by **Generative AI / foundation models**

Limitations

- The dataset is based on **arXiv**, which may not fully represent journal publications.
- The **keyword-based medical filter** may include borderline cases or miss relevant papers.
- The analysis uses **title + abstract**, not full-text content.
- I used k = 3, so the topics are broad and high-level rather than very detailed.

Future Work

- Apply the same pipeline to a **specific journal corpus** using its official Aims & Scope.
- Test **higher topic granularity** (larger k) or **BERTopic** for finer themes.
- Evaluate **clustering robustness** across random seeds and model variants.