UNIVERSITÀ DEGLI STUDI
DI MILANO

FACOLTÀ DI SCIENZE POLITICHE,
ECONOMICHE E SOCIALI

Samet Cagan
Matriculation: 985869
Master in Data Science and Economics

Quantifying Thematic Drift in Medical AI (2015–2025)

Abstract

In this project, I looked at how Medical AI research has been changing over time, and whether journal "aims and scope" statements still match what researchers are actually publishing. To do this, I collected over 10,500 arXiv papers from 2015 to 2025 in major Computer Science areas and filtered them using medical and clinical keywords. I used a Transformer model (S-PubMedBERT-MS-MARCO `https://huggingface.co/pritamdeka/S-PubMedBert-MS-MARCO`) to turn each paper's title and abstract into embeddings to compare papers by meaning. The biggest result is the rapid rise of Generative AI / LLMs. This area grew up around 30% of the papers in the dataset. At the same time, topics like Clinical Informatics take up a smaller share than before.

Contents

List of Figures

List of Tables

1 Introduction

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.

In this project, I focus on thematic drift. By thematic drift, I mean the situation where a field's real research output starts to move away from its official mission statement—either because new topics appear faster than the scope can be updated, or because older priorities become less dominant over time. Instead of relying on opinions or a few example papers, I wanted to measure this drift in a structured way using a large dataset.

To do that, my goals are straightforward. First, I built a longitudinal corpus of Medical AI-related papers from arXiv covering 2015–2025. Next, I represent each paper (title and abstract) as a dense semantic vector using a domain-adapted Transformer model, so I can compare papers based on meaning rather than just keywords. Then, I measure how closely each paper matches a predefined "Aims & Scope" statement by computing an alignment score. Finally, I use unsupervised clustering to uncover the main research pillars in the dataset, label them using TF-IDF keywords, and track how their "market share" changes year by year.

1.1 Literature Review

This project is related to (i) semantic text representations for scientific documents, (ii) topic discovery methods for large corpora, and (iii) bibliometric analyses of research evolution.

Transformer-based sentence embedding methods such as Sentence-BERT enable semantic comparison beyond keyword overlap, which is important when domain vocabulary is dense or heterogeneous [2]. For topic discovery, embedding-based clustering can be complemented by keyword-based topic labeling; BERTopic is a closely related approach that combines embeddings with class-based TF–IDF to produce interpretable topic representations [1].

1.2 Research Question and Problem Definition

1.2.1 Research question

Does the thematic content of Medical AI research (as observed in a longitudinal arXiv corpus from 2015–2025) remain aligned with a fixed "Aims & Scope" reference statement, and how does this alignment and the internal topic composition change over time?

1.2.2 Problem definition

Let $D = \{d_i\}_{i=1}^{N}$ be the set of papers, where each paper d_i is represented by its title and abstract. Let s be the "Aims & Scope" reference statement. Let $f(\cdot)$ be the embedding function (S-PubMedBERT), producing vectors in R^{768}. The paper embedding is $\mathbf{p}_i = f(d_i)$ and the scope embedding is $\mathbf{a} = f(s)$. The alignment score for paper i is:

$$\text{align}(i) = \cos(\mathbf{p}_i, \mathbf{a}) = \frac{\mathbf{p}_i \cdot \mathbf{a}}{\|\mathbf{p}_i\| \, \|\mathbf{a}\|} \tag{1}$$

To study drift, alignment scores are analyzed over time (e.g., yearly distributions or yearly means). To characterize topic structure, embeddings are clustered and topic prevalence is tracked by year using normalized topic shares.

2 Methodology

In this project, I built a reproducible pipeline to study how Medical AI research on arXiv has shifted from 2015–2025. The workflow involves collecting papers, cleaning text, converting to semantic embeddings, and running alignment and clustering analyses.

2.1 Metrics and Experimental Protocol

2.1.1 Alignment score

For each paper, I compute cosine similarity between the paper embedding and the embedded Aims & Scope reference (Eq. 1). Higher values indicate stronger thematic alignment.

2.1.2 Topic market share

After clustering, I measure topic prevalence by year-normalized shares:

$$\text{share}(k, y) = \frac{n_{k,y}}{\sum_{k'} n_{k',y}} \tag{2}$$

where $n_{k,y}$ is the number of papers assigned to topic k in year y.

2.1.3 Experimental protocol

Embeddings are computed on CPU. Clustering is performed in the original embedding space using KMeans ($k = 3$). UMAP is used only for 2D visualization. Cluster labels are assigned using TF–IDF top terms within each cluster.

2.2 Data Curation

I used the arXiv API to create my corpus by targeting core CS categories where Medical AI is commonly published (cs.CV, cs.LG, cs.CL, cs.AI) and then filtered for medical relevance using keywords (e.g., medical, clinical, health, EHR, patient) in the title/abstract. Each record stores the arXiv id/URL, title, abstract, published date, year, and categories. In total, I collected 10,592 unique papers (2015 – 2025).

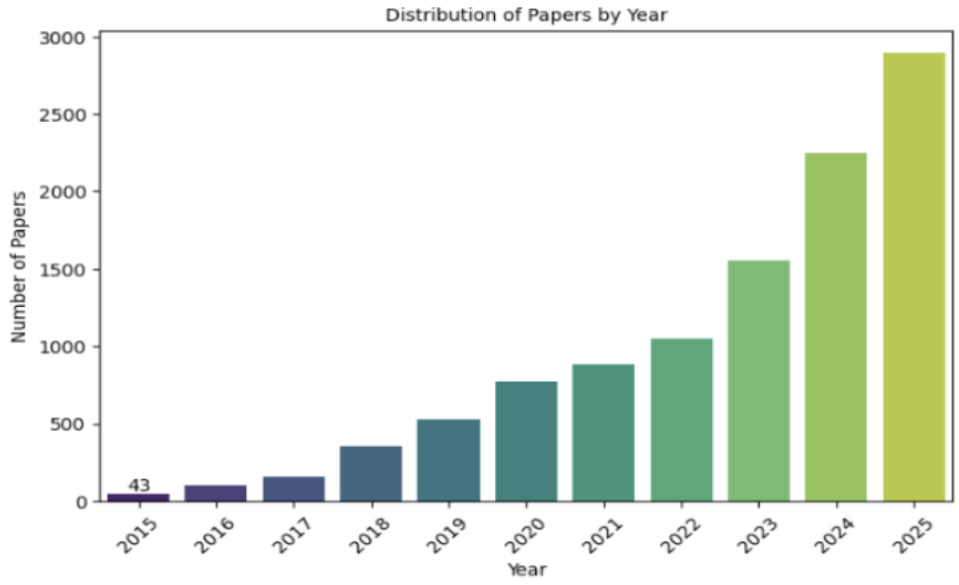

Figure 1: Distribution of arXiv Medical AI papers per year (2015-2025).

2.3 Text Processing

To prepare the papers for embedding, I combined the title and abstract into one text field. I did this because the title usually contains the most important keywords (often specific medical terms), while the abstract provides the full context needed to understand what the paper is actually about. My cleaning step was intentionally minimal because Transformer models work best when they keep natural sentence structure. I cleaned titles and abstracts separately and then concatenated them. The preprocessing included:

- Unicode normalization (NFKC)
- Lowercasing, removing punctuation
- URL removal, hyphen and whitespace handling

A key decision is that I did not remove stopwords 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. After cleaning and merging, the text lengths stayed within the model's limits: the average length was about 205 words, the maximum was about 402 words, and only one paper exceeded 400 words, which fits well under the embedding model's 512-token constraint.

2.4 Model Selection

I selected **S-PubMedBERT**—a sentence embedding model trained on biomedical literature—instead of a general-purpose alternative such as **all-MiniLM-L6-v2**. This choice was motivated by the nature of the corpus, which contains dense clinical terminology interleaved with AI-specific concepts. In this setting, a standard model can overweight biomedical vocabulary and dilute the signal associated with methodological distinctions.

S-PubMedBERT captures biomedical semantics more reliably and, in practice, helped preserve the *technical* structure of the papers. As a result, downstream clustering was better aligned with AI methodology (e.g., separating work focused on *image segmentation* from work centered on *language generation*) rather than being driven primarily by clinical terms.

To generate document representations, I computed embeddings on CPU with a batch size of 32. The resulting embedding matrix had shape (10,592, 768), indicating that all 10,592 papers were successfully mapped into 768-dimensional vectors. As a sanity check prior to clustering, I computed the L2 norm of each embedding. The minimum observed norm was approximately 14.78, confirming that no zero vectors were produced and that the embedding output was valid for subsequent clustering.

2.5 Defining the "Ground Truth"

To measure thematic drift, I defined a short Aims & Scope mission statement representing what a Medical AI journal claims to publish. I embedded this mission statement using the S-PubMedBERT model so that both papers and the mission statement live in the same vector space.

To scientifically measure if the field was drifting, I needed a fixed reference point, so I defined a standard "Aims & Scope" mission statement that represents what a typical Medical AI journal claims to publish. I ran this text through the model to create a "target vector," effectively establishing a mathematical center of gravity for the field. With this target established, I calculated an alignment score for every single paper using cosine similarity, which measures the angle between the paper embedding and the aim vector.

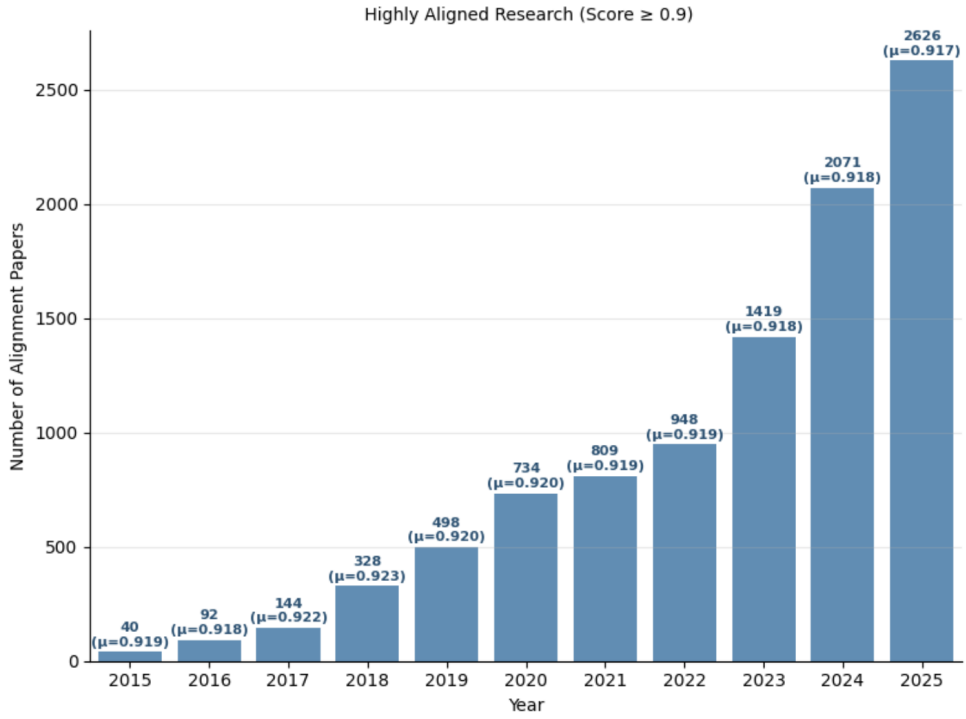

Figure 2: Alignment score of arXiv Medical AI papers per year (2015-2025).

2.6 Unsupervised Clustering (UMAP + KMeans + TF-IDF)

Alignment scoring tells me whether papers broadly match the mission, but it does not explain what topics make up the field. To discover the main research pillars, I used unsupervised clustering on the paper embeddings.

First, I applied KMeans directly in the embedding space to group papers into topics. In my notebook, I set the number of clusters to 3, which produced a simple and interpretable breakdown of the field. Because embeddings are 768-dimensional and hard to visualize, I used UMAP to project the vectors into 2D for plotting (UMAP is for visualization; clustering done in the embedding space). To name and interpret each cluster, I used TF–IDF keyword extraction on the text within each cluster and then labeled the clusters based on their strongest keywords. After labeling the clusters, I measured topic trends by computing each cluster's year-normalized share ("market share") so I could compare growth and decline fairly across years.

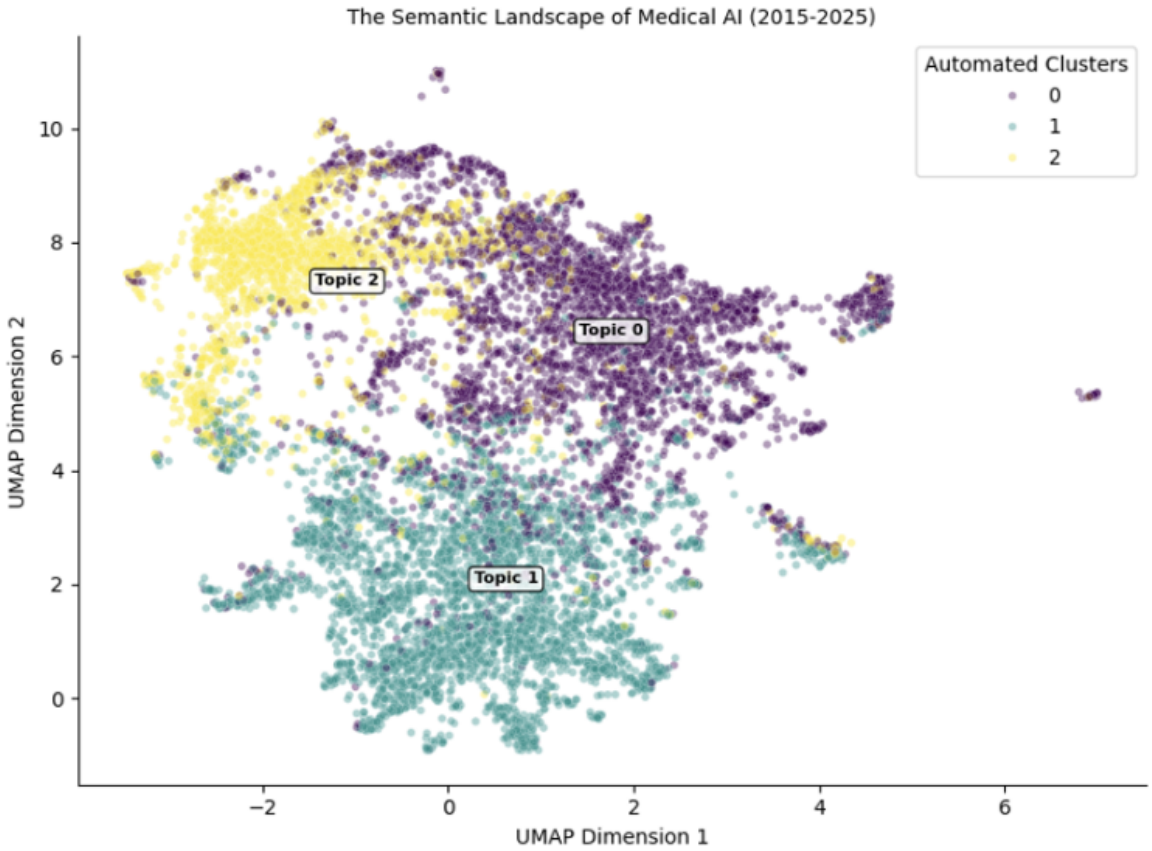

Figure 3: Paper embeddings colored by KMeans cluster (3 topics).

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

3 Results

I calculated an alignment score for every single paper using cosine similarity. The results showed my dataset was very accurate: 91.7% of the papers' scores were above 0.90, confirming that the vast majority of the research was very close to the target.

3.1 Topic Evolution and Trend Analysis

After defining the three clusters, I wanted to see how the field's focus changes across time. To do this fairly (because the total number of papers increases a lot each year), I measured topic "market share" by normalizing cluster counts within each year. This way, each year adds up to 100%, and the trends show changes in attention, not just growth in arXiv volume.

Visual snapshots of the trends confirm how quickly the balance shifts year to year.

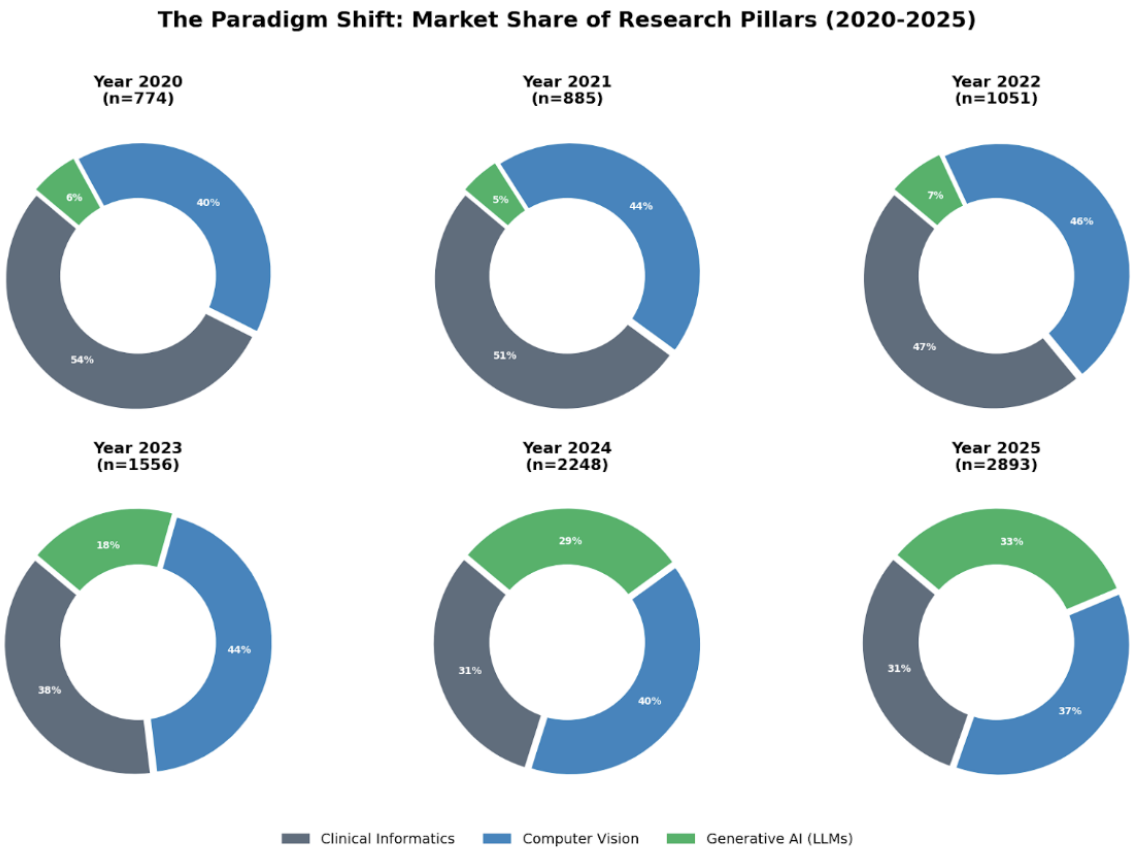

Figure 4: Donut-chart grid showing topic shares for 2020-2025.

The most significant result is the rise of LLMs and the relative decline of Clinical Informatics.

- **Initial share (2019):** 3.6%
- **Current share (2025):** 32.6%
- **Net growth:** +29.0 percentage points

4 Conclusion and Limitations

4.1 Conclusion

Based on the analyses in this project, I conclude that thematic drift is real and measurable in Medical AI research over the 2015–2025 period. Even though the overall alignment scores against the "Aims & Scope" mission statement stay high, the clustering and trend analysis show that the content of the field has shifted in a major way. Medical AI is no longer dominated only by traditional clinical informatics (EHRs, structured data mining) or even mainly by medical imaging. Instead, a large and fast-growing share of the field is now centered on Generative AI and foundation models, especially LLM-based methods applied to clinical text and workflows.

In other words, the field has moved toward a more mixed identity. It is no longer just an "informatics-style" area focused on structured clinical data; it has become a hybrid intelligence space that combines classical clinical informatics, computer vision for imaging, and foundation-model approaches for language and multimodal reasoning.

4.2 Limitations and Future Work

Limitations. This study is based on arXiv papers and a keyword-based medical filter, which may exclude relevant work or include borderline cases. The analysis uses titles and abstracts rather than full text, and topic modeling with $k = 3$ provides a deliberately coarse taxonomy.

Future work. A direct extension is to apply the same pipeline to a specific journal using the journal's official Aims & Scope statement and the journal's published abstracts. Future iterations could also test alternative topic granularities (larger k or BERTopic), evaluate clustering stability across random seeds, and perform deeper qualitative review of the lowest-alignment outliers.

References

[1] Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint* arXiv:2203.05794.

[2] Reimers, N., & Gurevych, I. (2019, November). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982–3992).

[3] Picascia, S., Montanelli, S., Salini, S., & Verzillo, S. (2025). The Atlas of Data Science Research. *IEEE Access*.

[4] Hassan-Montero, Y., Guerrero-Bote, V. P., & De-Moya-Anegón, F. (2014). Graphical interface of the Scimago Journal and Country Rank: an interactive approach to accessing bibliometric information. *El profesional de la información*, 23(3).