A Taxonomy of AI Research Areas: Unsupervised Embedding Clustering and LLM-Assisted Evaluation

Fevzi KILAS
Hacettepe University
fevzikilas@hacettepe.edu.tr

NERGIZ CAGILTAY
Hacettepe University
nergizcagiltay@hacettepe.edu.tr

Abstract

Artificial Intelligence (AI) research has expanded exponentially, creating a landscape so vast that traditional, manually curated taxonomies fail to capture its evolving nuances [4]. While unsupervised clustering of semantic embeddings offers a scalable solution, identifying the optimal granularity remains a challenge; standard geometric metrics (e.g., Silhouette Score) often fail to reflect semantic coherence in high-dimensional text embeddings [9]. This paper introduces a fully data-driven framework for constructing a dynamic taxonomy of AI. Starting from a bulk dataset of 2.91 million ArXiv papers, we developed a filtering pipeline to isolate 565,626 core AI publications. We propose a novel LLM-Assisted Semantic Validation methodology, rejecting reliance on geometric metrics in favor of a hybrid evaluation approach combining centroid-based and random sampling across varying k values. Our analysis identifies $k = 16$ as the optimal resolution for a Level-1 taxonomy, revealing distinct clusters for emerging fields such as *Scientific ML* and *Generative Audio* that were previously obscured. The resulting hierarchical taxonomy provides an empirical, literature-grounded map of the modern AI landscape.

1 Introduction

The volume of research in Artificial Intelligence (AI) has transcended human cognitive capacity for manual tracking. The arXiv repository alone receives thousands of submissions weekly, rendering static categorization systems (such as the classic `cs.AI`, `cs.CV`, `cs.CL`) insufficient for distinguishing modern sub-fields like *Prompt Engineering*, *Diffusion Models*, or *Neuromorphic Computing*. Researchers and policymakers lack a granular, dynamic map of the field.

Traditional bibliometric approaches relying on citation networks capture "influence" but often miss "content" [11, 2]. Conversely, standard topic modeling (e.g., LDA) struggles with the contextual depth required to differentiate subtle engineering domains [1]. Modern approaches utilizing dense vector embeddings offer a promising alternative [5]; however, they introduce a new challenge: *Model Selection*. In high-dimensional semantic spaces, research topics do not always form distinct, spherical clusters. Consequently, standard geometric metrics like the Elbow Method or Silhouette Coefficient often provide misleading signals regarding the optimal number of clusters (k) [10].

1. **Robust Data Pipeline:** We process a raw ArXiv metadata snapshot of 2.9 million papers [4], implementing a rigorous filtering funnel to curate a high-quality corpus of 565,626 AI-focused papers.
2. **Critique of Geometric Metrics:** We demonstrate that mathematical compactness in embedding space does not necessarily equate to semantic purity, necessitating a new evaluation paradigm.

3. **LLM-Assisted Validation:** We introduce a human-in-the-loop methodology where Large Language Models (LLMs), specifically GPT-4 [7], analyze both cluster centroids (dominant themes) and random samples (edge cases) to determine the semantic validity of the taxonomy.
4. **The Taxonomy:** We present a finalized, hierarchical map of AI (16 Clusters grouped into 5 Super-Categories), identifying the structural independence of fields like Audio Processing and the dominance of 3D Perception.

2 Related Work

Science mapping has traditionally relied on bibliometric methods such as citation networks and co-authorship graphs. Tools like VOSviewer [11] and CiteSpace [2] visualize scientific domains based on influence patterns but often overlook the semantic content of the publications. With the advent of deep learning, content-based approaches have gained traction. Neural topic modeling techniques, such as BERTopic [5], utilize embeddings from language models like BERT [8] to discover latent topics. However, these methods typically focus on micro-level topic extraction rather than macro-level field taxonomy construction.

Recent studies have begun exploring the use of Large Language Models (LLMs) for evaluation tasks, questioning whether they can replace human annotators [3]. Our work bridges these domains by applying unsupervised embedding clustering for structure discovery and integrating an LLM-assisted validation loop to ensure semantic coherence, a methodological combination not previously explored for large-scale AI taxonomy generation.

3 Data and Preprocessing

3.1 Data Source and Filtering

We utilized the ArXiv Public Dataset, containing metadata for approximately **2,918,474** papers [4]. To isolate relevant research, we applied a multi-stage filtering pipeline. First, we restricted the dataset to primary AI and Machine Learning categories: `cs.AI`, `cs.CL`, `cs.CV`, `cs.IR`, `cs.LG`, `cs.NE`, `cs.RO`, `cs.SI`, and `stat.ML`. Second, we performed content filtering to remove entries with null titles or abstracts. This process reduced the corpus to **565,626** high-quality documents, representing the core of modern AI literature.

3.2 Embedding Generation

For each paper, we concatenated the `Title` and `Abstract` to form a single textual input. To project these documents into a semantic vector space, we utilized the **Sentence Transformer** model `all-MiniLM-L6-v2` (hosted on Hugging Face) [8, 12]. This model transforms the raw text into **384-dimensional** dense vectors. This ensures that papers with semantically similar content (e.g., "object detection" and "semantic segmentation") are positioned proximally in the vector space, regardless of keyword overlap.

4 Methodology

4.1 Unsupervised Clustering

All generated embeddings lie in the same high-dimensional semantic vector space. We employed the K-Means clustering algorithm to partition this space [6]. Since the optimal number of research areas is unknown a priori, we conducted experiments across a range of $k \in [5, 22]$.

4.2 The Failure of Geometric Metrics

Initial attempts to select k using standard geometric metrics (e.g., Silhouette Score, Elbow Method) yielded inconclusive results. These metrics favor clusters with high intra-class compactness and high inter-class separation [9]. However, AI research topics lie on a continuous manifold; for instance, *Vision-Language Models* naturally bridge *Computer Vision* and *NLP*, creating geometric overlaps that penalize standard metrics despite representing valid, distinct research areas. Thus, we concluded that *geometric separation* is not a valid proxy for *semantic distinction* in this domain.

4.3 LLM-Assisted Semantic Validation

To overcome the limitations of geometric metrics, we devised a novel, agentic AI workflow for qualitative evaluation. For every cluster C_i generated in each k-experiment, we extracted two distinct sample sets:

1. **Centroid Samples (**$N = 50$**):** Papers geometrically closest to the cluster center. These represent the "ideal" or dominant theme of the cluster.
2. **Random Samples (**$N = 100$**):** Papers selected randomly from the cluster assignment. These are critical for detecting noise, impurities, and "outlier" topics that the algorithm incorrectly grouped.

These combined samples (150 abstracts per cluster) were processed through an Agentic AI workflow utilizing Large Language Models (LLMs) [7, 3]. The system was prompted to summarize the cluster's core theme, identify sub-topics, and flag semantic incoherence. Through this iterative analysis, $k = 16$ emerged as the optimal resolution. It successfully disentangled *Audio/Speech* from *NLP*, and *Reinforcement Learning Theory* from *Robotics Application*, separations that were conflated at lower k values.

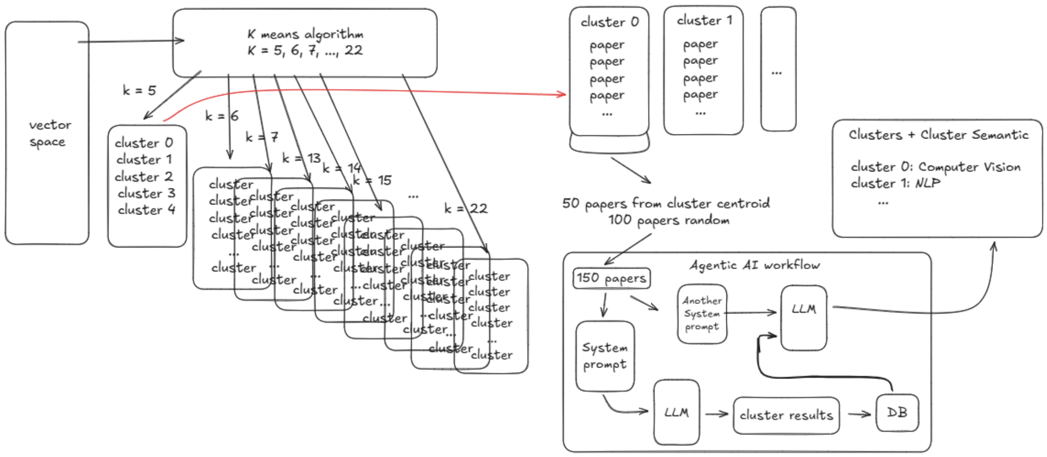

Figure 1: **Methodological Workflow.** The pipeline proceeds from raw embedding vectors to K-Means clustering (testing $k = 5$ to 22). For validation, a hybrid sampling strategy (Centroid + Random) is fed into an LLM-based agentic workflow to evaluate semantic coherence and assign labels.

5 Results: A Taxonomy of AI

Our analysis yielded 16 distinct semantic clusters. To facilitate navigation and provide a holistic view of the field, we aggregated these clusters into a hierarchical structure (Level 0.5) using a bottom-up approach.

5.1 Level 0.5: Super-Categories

The taxonomy is organized into the following high-level domains, as visualized in Figure 2:

- **Vision & Image:** Covering 2D/3D vision, medical imaging, and generative media.
- **NLP & Audio:** Text processing, retrieval, and speech intelligence.

- **Systems & Theory:** Mathematical foundations, optimization, and security.
- **Classical ML & Graph Networks:** Structured data, social computing, and graphs.
- **Decision Making & Logic:** Robotics, reinforcement learning, and symbolic AI.

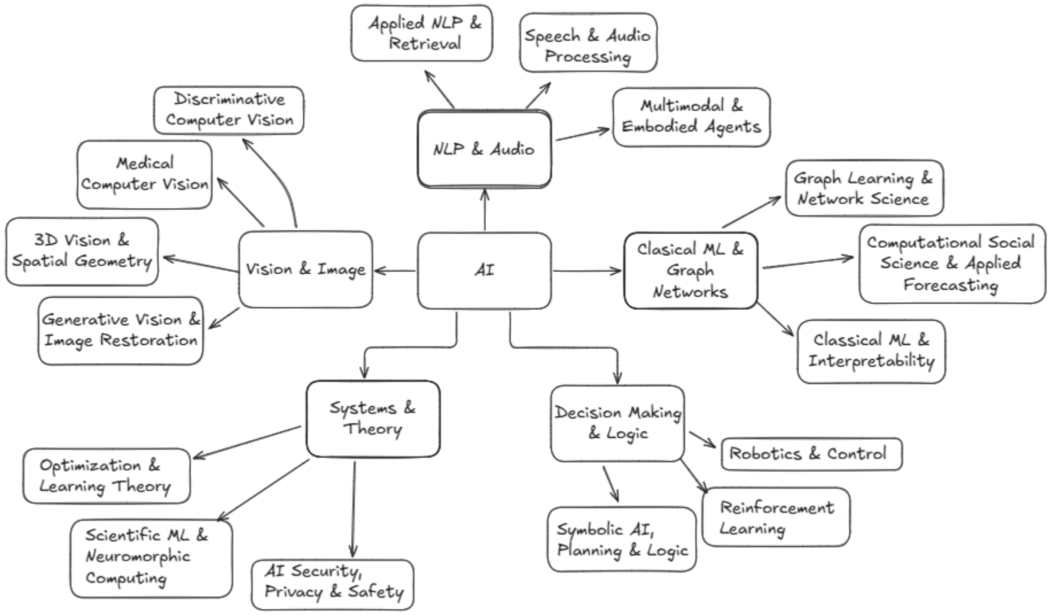
AI
NLP & Audio
Applied NLP & Retrieval
Speech & Audio Processing
Multimodal & Embodied Agents
Vision & Image
Discriminative Computer Vision
Medical Computer Vision
3D Vision & Spatial Geometry
Generative Vision & Image Restoration
Clasical ML & Graph Networks
Graph Learning & Network Science
Computational Social Science & Applied Forecasting
Classical ML & Interpretability
Systems & Theory
Optimization & Learning Theory
Scientific ML & Neuromorphic Computing
AI Security, Privacy & Safety
Decision Making & Logic
Robotics & Control
Reinforcement Learning
Symbolic AI, Planning & Logic

Figure 2: **Hierarchical AI Taxonomy.** The data-driven map of AI research areas, aggregating the 16 identified clusters into 5 semantic super-categories.

5.2 Level 1: The 16 Clusters

Table 1 presents the detailed breakdown of the 16 identified research areas alongside their content descriptions.

5.3 Key Findings and Emerging Trends

Our analysis revealed several structural insights into the current state of AI:

1. **Scientific ML is a Major Field (Cluster 0):** Often grouped under general theory, our analysis shows a massive, distinct cluster dedicated to Physics-Informed Neural Networks (PINNs) and Neuromorphic computing, indicating a convergence of AI and physical sciences.
2. **The Independence of Audio (Cluster 9):** Speech processing has structurally detached from NLP. While they share architectures (Transformers), Audio is now a distinct domain dominated by signal processing challenges and generative voice models.
3. **The Scale of 3D Perception:** The combination of Cluster 15 and Cluster 10 (which contains significant autonomous driving literature) represents one of the largest volumes of research, driven by the autonomous vehicle industry.
4. **NLP Segmentation:** We observe a clear split between *Applied NLP/Retrieval* (Cluster 10) and *Reasoning/Agents* (Cluster 1), reflecting the shift from simple text processing to complex agentic behaviors.

6 Conclusion

This paper presented a data-driven approach to mapping the AI research landscape. By moving beyond geometric clustering metrics and employing an LLM-assisted validation loop with random

Table 1: The 16 Semantic Clusters of AI Research (Level 1 Taxonomy)

Cluster Name	**Description & Key Topics**
Vision & Multimodal Perception	
Cluster 7: Discriminative Vision	2D Object Detection, Segmentation, Action Recognition.
Cluster 8: Generative Vision	Image Restoration, Inpainting, Super-Resolution, Diffusion Models.
Cluster 15: 3D Vision & Geometry	LiDAR, Point Clouds, SLAM, Gaussian Splatting, Depth Estimation.
Cluster 3: Medical Imaging	MRI/CT Segmentation, Tumor Detection, Computational Pathology.
Language & Audio	
Cluster 10: Applied NLP	RAG, NER, Sentiment Analysis, Text Mining, Information Retrieval.
Cluster 1: Multimodal Agents	VLM, Reasoning, Minecraft Agents, Embodied Planning.
Cluster 2: NLP Core	Machine Translation, Multilingual Modeling, Low-resource NLP.
Cluster 9: Speech & Audio	ASR, TTS, Music Generation, Speech Emotion Recognition.
Theory & Systems	
Cluster 12: Optimization Theory	Convergence Rates, Convex Optimization, SGD Dynamics.
Cluster 0: Scientific ML	Physics-Informed NN, Spiking Networks, Neuromorphic Computing.
Cluster 6: Security & Privacy	Adversarial Attacks, Differential Privacy, Deepfake Detection.
Cluster 11: Symbolic AI	Game Theory, Logic Programming, Constraint Satisfaction.
Structured Data & Applied	
Cluster 5: Graph Learning	Graph Neural Networks (GNN), Link Prediction, Molecular Graphs.
Cluster 4: Social & Forecasting	Time Series, Tabular Data, Social Media Analysis, Clinical Prediction.
Cluster 14: Classical ML	Interpretability (XAI), Random Forests, Feature Selection.
Robotics & Control	
Cluster 13: RL Algorithms	Deep Reinforcement Learning, Bandits, Exploration Strategies.
Cluster 2: Applied Robotics	Manipulation, UAV Control, Motion Planning, Sim2Real.

sampling, we constructed a taxonomy that accurately reflects the semantic structure of 565,626 papers. The resulting 16-cluster hierarchy highlights the emergence of new independent fields and provides researchers with a high-fidelity navigation tool for the ever-expanding universe of AI.

References

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine learning research*, 3(Jan):993–1022, 2003.

[2] Chaomei Chen. Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for information Science and Technology*, 57(3):359–377, 2006.

[3] Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*, 2023.

[4] Colin B. Clement, Matthew Bierbaum, Kevin P. O'Keeffe, and Alexander A. Alemi. On the use of arxiv as a dataset. *arXiv preprint arXiv:1905.00075*, 2019.

[5] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[6] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[7] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[8] 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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.

[9] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[10] Robert L Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267–276, 1953.

[11] Nees Jan Van Eck and Ludo Waltman. Software survey: Vosviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2):523–538, 2010.

[12] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020.