

Dynamic Knowledge Map of Artificial Intelligence Literature: Longitudinal Trend Analysis Using Latent Dirichlet Allocation

Aulia Khairunnisa^{1*}, Syalwa Salsabillah Siregar², Ken Ditha Tania³, Allsela Meiriza⁴, Zaqqi Yamani⁵,
Shanti Dewi Siawanta⁶

^{1,2,3,4,5}Information System, Faculty of Computer Science, Universitas Sriwijaya, Palembang, South Sumatra,
Indonesia

⁶School of Computer Science, Faculty of Science, University of Nottingham, Nottingham, Nottinghamshire, United
Kingdom

E-mail: ^{1*}09031382328151@student.unsri.ac.id, ²09031382328139@student.unsri.ac.id, ³kenya.tania@gmail.com,
⁴allsela_meiriza@yahoo.co.id, ⁵zaqqi_yamani@unsri.ac.id, ⁶psxss47@nottingham.ac.uk

(Received: 6 Mar 2026, revised: 2 May 2026, accepted: 4 May 2026)

Abstract

The rapid growth of Artificial Intelligence (AI) publications has increased the complexity of identifying thematic structures and understanding the evolution of research trends. Conventional citation-based bibliometric approaches are often limited in capturing semantic relationships within large-scale textual data. This study aims to analyze the knowledge structure and longitudinal dynamics of AI literature using a topic modeling approach within a knowledge mapping framework. The dataset consists of publication abstracts from the arXiv repository spanning 2015–2024, and the Latent Dirichlet Allocation (LDA) algorithm is employed with coherence-based evaluation to extract latent topics. The results indicate that the optimal model configuration consists of nine topic clusters, achieving a peak coherence score of 0.4605. Longitudinal analysis suggests a notable shift in research focus after 2021, marked by the rapid growth of Large Language Models (LLMs) and generative AI. This shift reflects a process of thematic integration, where foundational areas such as deep learning architectures are increasingly incorporated into more advanced domains. In addition, topics such as Graph Neural Networks (GNNs) exhibit relatively stable trends, which may indicate technological maturity rather than a decline in relevance. Overall, the findings provide data-driven insights into the evolving landscape of AI research and may support researchers and institutions in identifying emerging research directions within the scope of the analyzed dataset.

Keywords: Artificial Intelligence; Knowledge Mapping; Latent Dirichlet Allocation; Research Trend Mapping; Topic Modeling.

I. INTRODUCTION

Artificial Intelligence (AI) research has experienced an unprecedented surge in publication volume over the past decade, resulting in an increasingly complex and rapidly evolving scientific landscape. This exponential growth poses significant challenges in systematically identifying thematic structures and understanding the evolution of knowledge. Conventional bibliometric approaches, which primarily rely on citation-based indicators, are often insufficient to capture the semantic relationships and conceptual dynamics embedded within large-scale scientific texts [1], [2]. As a result, the rapid expansion of AI literature may lead to thematic fragmentation, making it difficult for researchers and institutions to identify strategic research directions and emerging innovation areas [1].

To address these limitations, content-based analytical approaches using text mining techniques have gained increasing attention. In particular, publication abstracts represent a highly efficient and information-dense data source, as they encapsulate the core objectives, methodologies, and findings of scientific studies [3]. By leveraging abstracts, it becomes possible to extract meaningful thematic patterns while minimizing noise from redundant or non-essential sections of full-text documents [4], [5]. This approach enables a more accurate representation of the conceptual structure of a research field.

Among various text mining techniques, topic modeling has been widely adopted to uncover latent thematic structures within large corpora [3]. Latent Dirichlet Allocation (LDA), a probabilistic generative model, has demonstrated robustness in identifying topic distributions and capturing macro-level



research trends across diverse domains [6], [7], [8]. Compared to more recent transformer-based approaches such as BERTopic, LDA offers greater stability and interpretability for longitudinal analysis. While transformer-based models excel in capturing fine-grained semantic relationships in short texts, their sensitivity to contextual embeddings and dynamic vocabulary shifts may introduce inconsistencies when applied to long-term thematic analysis [9], [10]. In contrast, LDA provides a consistent probabilistic framework that is particularly well-suited for tracking macro-level topic transitions over extended time horizons, such as decade-long research developments [4], [7].

Despite the widespread use of LDA in scientometric studies, most existing research focuses primarily on generating topic clusters without integrating them into a comprehensive knowledge mapping framework [1], [11]. Furthermore, previous studies often lack explicit longitudinal interpretation that connects thematic evolution with broader shifts in the AI research paradigm [12]. This indicates a clear research gap: the absence of an integrated approach that combines probabilistic topic modeling with dynamic and temporally structured knowledge mapping to support a more comprehensive and data-driven understanding of research evolution [1].

Therefore, this study proposes a systematic framework that integrates LDA-based topic modeling with longitudinal knowledge mapping, defined as a structured representation of semantic relationships among research topics, to analyze the evolution of AI research from 2015 to 2024, a period characterized by rapid advancements in deep learning and the emergence of large-scale language models (LLMs). Unlike prior studies that primarily generate static topic clusters, this study explicitly integrates coherence-optimized LDA with temporal knowledge mapping to enable a structured and longitudinal interpretation of research evolution, particularly in capturing how foundational domains progressively integrate into emerging paradigms. The approach incorporates rigorous preprocessing, domain-specific filtering, and coherence-based evaluation to ensure the interpretability and stability of extracted topics [10], [13].

The primary contributions of this study include the identification of latent thematic structures in AI literature using an optimized LDA model, as well as the examination of longitudinal topic dynamics to reveal shifts in research focus and underlying paradigm changes within the field. In addition, this study develops a structured knowledge mapping framework that represents relationships among research topics over time. Rather than providing prescriptive conclusions, this study aims to offer data-driven insights that may assist researchers and institutions in understanding the evolving landscape of AI research.

II. RESEARCH METHOD

This study adopts a Text Mining and Knowledge Discovery in Databases (KDD) framework designed to extract latent thematic structures from large-scale academic literature. As illustrated in Figure 1, the research workflow consists of

five systematic phases: Data Collection, Text Preprocessing, Feature Engineering, Model Evaluation and Topic Modeling, and Knowledge Mapping Analysis.

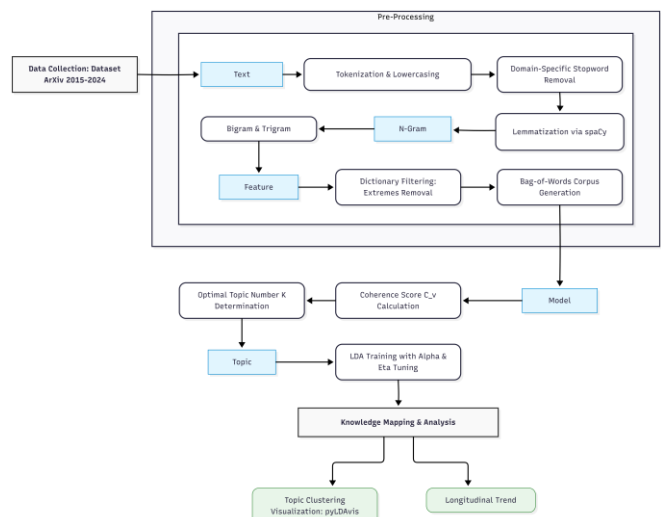


Figure 1. Research Architecture

A. Data Collection

The dataset utilized in this study comprises a comprehensive collection of scholarly abstracts extracted from the arXiv open-access repository [14]. arXiv was strategically chosen as the primary data source due to its established prominence as the leading preprint server for the rapid dissemination of cutting-edge research in the computational sciences [15]. The data extraction process specifically targeted three primary computer science sub-domains that represent the core pillars of contemporary algorithmic development: Artificial Intelligence (cs.AI), Machine Learning (cs.LG), and Computation and Language (cs.CL). The inclusion of cs.CL is particularly crucial to accurately capture the monumental advancements in Natural Language Processing (NLP) and large language models that have heavily influenced recent technological discourse [16].

The temporal scope of the collected data spans exactly one decade, from 2015 to 2024. This specific timeframe was deliberately established to capture the most dynamic era in the modern evolution of AI research, encompassing critical paradigm shifts, from the early maturation of deep learning architectures to the recent explosion of generative AI and foundational models [12].

Rather than analyzing full-text manuscripts, scholarly abstracts were selected as the primary corpus for the topic modeling algorithms. Abstracts inherently provide a dense, highly structured, and concise representation of a study's core objectives, proposed methodologies, and principal findings [3]. By focusing exclusively on abstracts, the computational model can effectively extract the most salient thematic signals without being overwhelmed by the syntactic noise, redundant literature reviews, or extensive technical appendices typically found in full texts [5]. This approach is highly optimal for scientometric text mining and has been widely validated as a robust strategy for mapping scientific knowledge structures

and research trends [4], [17]. Finally, to ensure the highest standard of data integrity, a rigorous initial data cleansing phase was executed. During this stage, any records containing missing values, duplicate entries, corrupted text elements, or incomplete abstract formats were systematically filtered out and excluded, guaranteeing that the subsequent models are trained exclusively on coherent and high-quality textual data [18].

B. Text Preprocessing

To guarantee that the topic modeling algorithm captures semantically meaningful latent structures rather than being misled by syntactical noise, a rigorous and multi-stage text preprocessing pipeline was implemented. Text preprocessing is universally recognized as a foundational component in Natural Language Processing (NLP) and Knowledge Discovery in Databases (KDD), as it directly dictates the quality of the feature matrix fed into the machine learning model [18]. The pipeline commenced with systematic case-folding to transform all text to lowercase, followed by tokenization to decompose complex academic paragraphs into granular, individual word units. During this phase, punctuation, numerical digits, and non-alphanumeric characters were completely eradicated to establish a standardized lexical baseline [19].

A highly critical step in academic scientometric text mining is the aggressive elimination of domain-specific stopwords. Relying solely on standard English stopword dictionaries, such as those provided by the NLTK library, is highly insufficient for scientific corpora [15]. Therefore, an extensive, custom-curated list of over 100 academic “noise” words including terms like *paper*, *propose*, *accuracy*, *experiment*, *framework*, and *result* was manually defined and filtered out. Failing to remove these high-frequency, domain-specific boilerplate terms inevitably forces topic modeling algorithms to erroneously cluster documents based on generic academic writing conventions and evaluation metrics, rather than their substantive technical or algorithmic content [9], [16].

The Finally, lemmatization was executed utilizing the advanced *spaCy* natural language processing library. Unlike simple stemming, which abruptly truncates word endings and often generates non-sensical tokens, lemmatization intelligently utilizes vocabulary and contextual morphological analysis to reduce words to their proper dictionary base forms (lemmas) for instance, dynamically converting variations like “learning”, “learned”, and “learns” into the singular semantic root “learn”. This semantic reduction is crucial for compressing the high-dimensional feature space and ensuring that conceptually identical terms are mathematically treated as a single entity during the subsequent corpus generation process [10], [20].

C. Feature Engineering: N-Gram Modeling and Corpus Generation

In the highly specialized domain of computer science, relying solely on single-word tokens (unigrams) often results in a severe loss of contextual specificity and semantic cohesion. For instance, isolated unigrams such as 'machine',

'deep', or 'neural' hold ambiguous meanings compared to their compounded academic forms like 'machine learning', 'deep learning', or 'neural network'. To preserve this essential semantic integrity, advanced N-Gram modeling, specifically Bigram and Trigram generation, was applied. This process statistically binds adjacent words that frequently co-occur across the corpus into unified multi-word expressions, fundamentally capturing deeper semantic representations than unigrams alone [3], [19]. The N-Gram phrase modeling and subsequent pipeline were programmatically executed utilizing the Gensim library, a highly optimized framework specifically designed for robust and scalable topic modeling on large corpora [10].

Following the N-Gram assimilation, a unified and comprehensive dictionary mapping all unique terms across the corpus was constructed. To optimize the computational efficiency of the algorithm and significantly improve topic coherence, a rigorous dimensionality reduction strategy, often referred to as extreme filtering, was executed. Terms appearing in fewer than 50 distinct documents were systematically purged to eliminate typographical errors, idiosyncratic terminology, and overly rare concepts that do not contribute to macro-level trend analysis. Conversely, terms appearing in more than 18% of the total corpus were also removed; these overly broad terms typically act as hidden domain-specific stop words that can dilute the distinctiveness of latent topics [18], [20].

The highly refined dictionary was subsequently transformed into a Bag-of-Words (BoW) matrix. The BoW architecture is a foundational feature extraction technique that mathematically represents the unstructured text data as a structured term-document frequency matrix, stripping away syntactic grammar while preserving the quantitative occurrence of essential terms [10], [18]. This mathematical matrix serves as the direct, quantitative input required by the generative probabilistic modeling algorithm to calculate semantic similarities, map word distributions, and successfully extract cohesive thematic clusters [5], [4].

The highly refined dictionary was subsequently transformed into a Bag-of-Words (BoW) matrix. The BoW architecture is a foundational feature extraction technique that mathematically represents the unstructured text data as a structured term-document frequency matrix, stripping away syntactic grammar while preserving the quantitative occurrence of essential terms [10], [18]. This mathematical matrix serves as the direct, quantitative input required by the generative probabilistic modeling algorithm to calculate semantic similarities, map word distributions, and successfully extract cohesive thematic clusters [5], [4].

D. Topic Modeling and Model Evaluation

The core analytical engine of this study is Latent Dirichlet Allocation (LDA), a three-level hierarchical Bayesian model originally introduced by Blei et al. [21] LDA operates on the premise that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

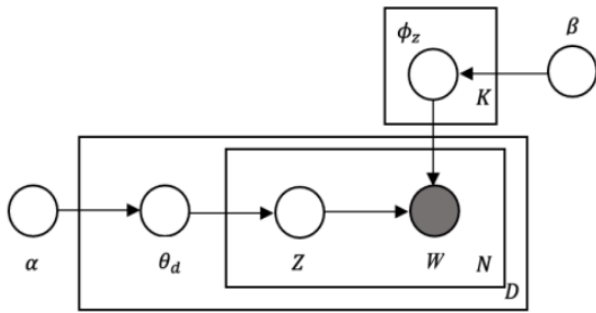


Figure 2. Latent Dirichlet Allocation [21].

To formally describe the semantic extraction, Figure 2 illustrates the graphical plate notation of the LDA generative process. The outer plate represents the total number of documents D , while the inner plate represents the sequence of N words within a specific document. The upper plate K denotes the total number of topics extracted from the corpus. The generative process for each document follows these probabilistic steps [2], [21]:

1. For each topic $k \in \{1, \dots, K\}$, a word distribution ϕ_k is drawn from a Dirichlet prior parameterized by β (denoted as η in algorithmic implementations): $\phi_k \sim \text{Dir}(\beta)$.
2. For each document $d \in \{1, \dots, D\}$, a topic distribution θ_d is drawn from a Dirichlet prior parameterized by α : $\theta_d \sim \text{Dir}(\alpha)$.
3. For each word w in document d :
 - a. A specific topic z is assigned by drawing from a multinomial distribution: $z \sim \text{Multinomial}(\theta_d)$.
 - b. The observable word w is then drawn from the corresponding topic-word distribution: $w \sim \text{Multinomial}(\phi_z)$.

Mathematically, the joint probability of a topic mixture θ , a set of N topics z , and a set of N words w given the priors α and β is formulated in Equation 1:

$$p(\theta, z, w | \alpha, \beta) = p(\theta | \alpha) \prod_{n=1}^N p(z_n | \theta) p(w_n | z_n, \beta) \quad (1)$$

To effectively mitigate the severe issue of topic overlap a pervasive challenge when analyzing highly intersected academic domains such as Artificial Intelligence and Machine Learning, the LDA model was trained utilizing strictly tuned hyperparameters. Specifically, both the document-topic density prior (α) and the topic-word density prior (β) were explicitly configured to an extremely low threshold of 0.01. From a mathematical perspective, lowering these Dirichlet priors forces the probabilistic mass towards the extreme corners of the geometric simplex, thereby generating a highly sparse distribution [9], [22]. This sparse architectural setting strictly compels each document to associate predominantly with only one or two major topics, significantly minimizing semantic noise and improving cluster separation [4], [9].

Furthermore, relying solely on predictive likelihood or perplexity to determine the optimal number of topics often yields clusters that are statistically sound but practically

incomprehensible to human researchers. Therefore, to ensure robust semantic interpretability, the optimal number of topics (K) was rigorously evaluated using the Topic Coherence Score based on the C_v metric [23], [24]. The core mechanism of the C_v metric relies heavily on calculating the Normalized Pointwise Mutual Information (NPMI) over a boolean sliding window to measure the co-occurrence of the top-weighted words (w_i, w_j) within a topic (Equation 2):

$$NPMI(w_i, w_j) = \frac{\log\left(\frac{P(w_i, w_j)}{P(w_i)P(w_j)}\right)}{-\log P(w_i, w_j)} \quad (2)$$

This equation mathematically measures the degree of semantic similarity, where higher joint probability $P(w_i, w_j)$ relative to independent probabilities strongly correlates with human contextual judgment. Through iterative execution across varying cluster sizes, the coherence evaluation validated that the model achieved its absolute optimal performance and maximum semantic stability at nine distinct topic clusters ($K = 9$), recording a peak Coherence Score of 0.4605 [10], [23].

The choice of LDA over contemporary transformer-based models like BERTopic is justified by its superior stability in capturing macro-level thematic shifts across decadal timelines. While transformer models offer high granularity, LDA provides a more robust probabilistic framework for longitudinal studies, ensuring that thematic evolution is mapped with consistent semantic boundaries over the ten-year observation period.

E. Knowledge Mapping and Trend Analysis

Following the rigorous training and validation of the LDA model, the multidimensional probabilistic results were visually rendered utilizing *pyLDAvis*, an advanced interactive visualization library specifically engineered for topic modeling interpretation. To effectively project the high-dimensional topic-word probability matrices into a human-interpretable two-dimensional Cartesian space, *pyLDAvis* employs Principal Component Analysis (PCA) combined with multidimensional scaling techniques [9], [10]. This Intertopic Distance Map mathematically plots the semantic boundaries and geometric distances between the generated clusters. Within this spatial representation, the proximity between centroids indicates semantic similarity, while the area of each circle proportionally represents the marginal prevalence of that specific topic within the entire overarching corpus [3], [16].

Subsequently, to transition from document-level probabilities to macro-level trend analysis, the dominant topic for each individual scholarly paper was systematically extracted. This operation was executed by isolating the specific topic index that yielded the maximum probability distribution weight (θ_d) for a given document [4], [14]. Once every manuscript was distinctly classified under its primary thematic cluster, the structured documents were chronologically aggregated based on their respective publication years to construct a comprehensive longitudinal time-series matrix.

This temporal mapping serves as a pivotal scientometric mechanism, as it quantitatively enables the precise identification of exponentially emerging trends, the detection of saturated or declining research areas, and the chronological tracing of fundamental paradigm shifts within the Artificial Intelligence domain over the observed decade [10], [12].

III. RESULT AND DISCUSSIONS

A. Model Evaluation and Optimal Topic Selection

The determination of the optimal number of topics (K) is a critical step in ensuring the model's interpretability and statistical validity. We evaluated the model's performance using the C_v coherence score across a range of K values from 7 to 16.

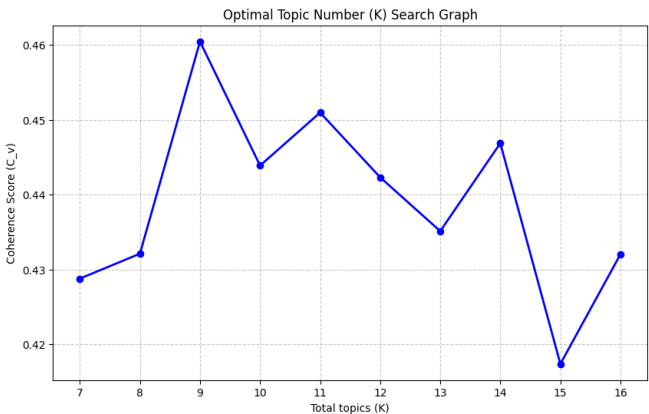


Figure 3. Coherence Score C_v Results Indicating K=9 as The Optimal Configuration.

As shown in Figure 3, the coherence score reaches its highest peak at $K = 9$ with a score of 0.4605. This peak indicates that at nine clusters, the words within each topic share the highest semantic similarity and are most distinct from other clusters. Beyond $K = 9$, the coherence score tends to decline and fluctuate, which suggests that increasing the number of topics further would lead to redundant or overly fragmented themes. Consequently, $K = 9$ was selected as the final model configuration.

B. Semantic Interpretation and Topic Labeling

The nine extracted topics represent the primary research pillars in the AI, ML, and NLP domains within the ArXiv dataset. Based on the salient terms (word distribution) provided by the LDA algorithm, we assigned descriptive labels to each topic.

The intertopic distance map (Figure 4) illustrates that the topics are well-distributed across the semantic space. The separation of Topic 8 (Graph) and Topic 0 (RL) from the central clusters indicates highly specialized sub-domains. In contrast, the proximity between Topic 1 (LLM) and Topic 5 (Transformer) accurately reflects the technical convergence where Transformer architectures serve as the backbone for modern Large Language Models (Table 1).

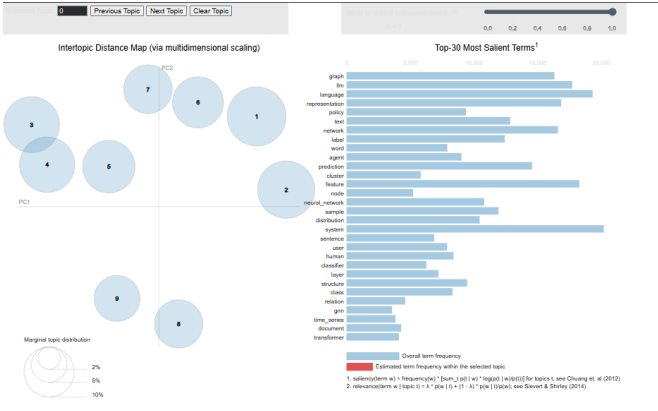


Figure 4. Intertopic Distance Map and Salient Terms for the $K = 9$ LDA Model.

Table 1. Identification of Research Topics and Representative Keywords

Topic ID	Proposed Label	Top Salient Keywords
0	Reinforcement Learning	policy, agent, environment, reward, action
1	LLMs & Cognitive Reasoning	llm, human, knowledge, large_language, reasoning
2	Predictive & Temporal Modeling	prediction, feature, time_series, temporal, patient
3	Robust Classification	label, sample, class, classifier, classification
4	Deep Learning Architectures	network, neural_network, layer, architecture, parameter
5	Transformer-based Generative AI	representation, transformer, attention, generation
6	Search & Optimization Systems	user, search, design, optimization, strategy
7	Text Mining & Linguistics	language, text, word, sentence, document
8	Graph Neural Networks (GNN)	graph, representation, node, gnn, structure

C. Longitudinal Research Trends Analysis (2015–2024)

The primary contribution of this study is to map research dynamics over the last decade. The temporal distribution of publication volumes for each topic reveals the "lifecycle" of AI sub-fields.



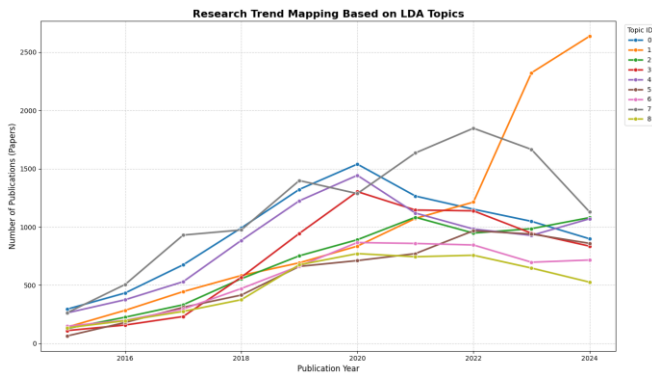


Figure 5. The Evolution of AI Research Topics From 2015 to 2024 Based on LDA Results.

Analysis of Figure 5 reveals several transformative shifts in the research landscape:

1. **The Meteoric Rise of LLMs (Topic 1):** Topic 1 (Orange line) exhibits the most dramatic growth in the entire dataset. While virtually non-existent before 2018, it began to skyrocket after 2021, reaching the highest publication volume by 2024. This trend clearly aligns with the "ChatGPT era" and the global shift toward generative AI.
2. **Saturation vs. Maturity in GNNs (Topic 8):** Analysis of Topic 8 (Graph Neural Networks) shows a stabilizing or slightly declining trend after 2022. It is crucial to distinguish between thematic saturation and implementation maturity. While the field may be reaching thematic saturation in terms of foundational 'breakthrough' papers, the technology itself has attained high implementation maturity. GNNs have transitioned from an experimental research 'hotspot' to a standard, robust methodology widely adopted for structured data analysis, explaining the shift in publication focus toward other emerging frontiers.
3. **The Stability of Foundations (Topic 7 & 6):** Topics such as Text Mining (Grey) and Optimization (Pink) show relatively flat or low-growth trajectories. Specifically, Topic 7 saw a significant peak in 2022 before stabilizing, reflecting its role as a fundamental but mature pillar of AI research. These remain foundational areas but are no longer the primary drivers of rapid research expansion compared to neural-based approaches.
4. **The Evolutionary Shift of Deep Learning (Topic 4):** Topic 4 (Deep Neural Architectures) exhibited a peak in volume around 2020 before beginning a downward trajectory. This trend does not signify technological obsolescence; rather, it indicates that foundational deep learning principles are being absorbed and integrated into more specialized domains. As the research community shifted focus toward Topic 1 (LLMs) and Topic 5 (Transformers), the 'general architectures' of the previous decade became the fundamental building blocks for modern generative systems. Thus, the decline in Topic 4 reflects a transition from general structural exploration to the specialized application of these architectures in large-scale reasoning models..

D. Practical Implications

The results confirm that the AI research community has shifted from general classification tasks (Topic 3) toward generative and reasoning-capable systems (Topic 1). For researchers and policymakers, these trends identify Topic 1 as the current "hotspot" for innovation, while Topic 8 remains a strong secondary pillar for structured data analysis.

IV. CONCLUSION

This study presents an analysis of the knowledge structure and longitudinal evolution of Artificial Intelligence (AI) research by integrating the Latent Dirichlet Allocation (LDA) algorithm within a knowledge mapping framework. Based on the analysis of arXiv publication abstracts from 2015 to 2024, nine primary thematic clusters were identified as a coherent representation of the AI research landscape. The findings indicate a substantial shift in research focus over the past decade, particularly the rapid emergence of Large Language Models (LLMs) and generative AI after 2021. This shift reflects a process of thematic integration rather than replacement, where foundational areas such as deep learning architectures are increasingly embedded within more advanced domains, especially transformer-based and LLM-driven research. Similarly, the relatively stable trend observed in Graph Neural Networks (GNNs) suggests technological maturity rather than a decline in relevance, indicating that these methods remain widely adopted despite fewer novel foundational contributions.

From a practical perspective, the findings provide data-driven insights that may assist researchers and academic institutions in understanding the distribution of research activity and identifying emerging directions within the AI landscape. Rapidly growing topics may indicate current areas of high research intensity, while stable domains may reflect mature methodologies that require more application-oriented exploration. However, this study is limited by its reliance on abstract-level data from a single repository (arXiv) and the use of a single modeling approach without direct comparison to alternative methods such as BERTopic. Future research may extend this work by incorporating multiple academic databases, analyzing full-text documents, and conducting comparative evaluations across different modeling techniques. Overall, the results suggest patterns and trends that may serve as a reference for understanding the evolving priorities and directions of AI research.

REFERENCES

- [1] A. Ayaz, A. K. Kabakuş, Ü. Özen, Ö. Alkan, and S. Aydın, "Topic modelling of contemporary management information systems research: a latent dirichlet allocation approach Topic Modelling of Contemporary Management Information Systems Research: A Latent Dirichlet Allocation Approach 1," 2025. [Online]. Available: <https://orcid.org/0000-0002-3814-3539>.

- [2] A. K. W. Weng, H. Y. Chang, K. K. Lai, and Y. B. Lin, "Topic Modeling on Peer Interaction in Online and Mobile Learning of Higher Education: 1993–2022," Aug. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/educsci14080867.
- [3] A. Abdelrazek, Y. Eid, E. Gawish, W. Medhat, and A. Hassan, "Topic modeling algorithms and applications: A survey," *Inf. Syst.*, vol. 112, p. 102131, Feb. 2023, doi: 10.1016/j.is.2022.102131.
- [4] Y. J. Jung and Y. Kim, "Research trends of sustainability and marketing research, 2010–2020: Topic modeling analysis," *Heliyon*, vol. 9, no. 3, Mar. 2023, doi: 10.1016/j.heliyon.2023.e14208.
- [5] X. Shu and Y. Ye, "Knowledge Discovery: Methods from data mining and machine learning," *Soc. Sci. Res.*, vol. 110, p. 102817, 2023, doi: <https://doi.org/10.1016/j.ssresearch.2022.102817>.
- [6] T. Dillan and D. H. Fudholi, "LDAViewer: An Automatic Language-Agnostic System for Discovering State-of-the-Art Topics in Research Using Topic Modeling, Bidirectional Encoder Representations From Transformers, and Entity Linking," *IEEE Access*, vol. 11, pp. 59142–59163, 2023, doi: 10.1109/ACCESS.2023.3285116.
- [7] R. K. Gupta, R. Agarwalla, B. H. Naik, J. R. Evuri, A. Thapa, and T. D. Singh, "Prediction of research trends using LDA based topic modeling," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 298–304, Jun. 2022, doi: 10.1016/j.gltp.2022.03.015.
- [8] Y. Pamungkas, "Leveraging Topic Modelling to Analyze Biomedical Research Trends from the PubMed Database Using LDA Method," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 13, no. 2, pp. 237–244, Jun. 2024, doi: 10.32736/sisfokom.v13i2.2117.
- [9] R. Egger and J. Yu, "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts," *Frontiers in Sociology*, vol. 7, May 2022, doi: 10.3389/fsoc.2022.886498.
- [10] Nursyahrina, S. Defit, and R. Sovia, "Metode BERTopic dan LDA untuk Analisis Tren Penelitian Bidang Ilmu Komputer," *Jurnal KomtekInfo*, pp. 332–341, Sep. 2024, doi: 10.35134/komtekinfo.v11i4.580.
- [11] R. G. Dalimunthe and R. A. Putri, "Pemetaan Topik Tugas Akhir Program Studi Ilmu Komputer Menggunakan Algoritma Latent Dirichlet Allocation," *Journal of Computer System and Informatics (JoSYC)*, vol. 5, no. 4, pp. 921–932, Aug. 2024, doi: 10.47065/josyc.v5i4.5759.
- [12] V. Maphosa and M. Maphosa, "Artificial intelligence in higher education: a bibliometric analysis and topic modeling approach," *Applied Artificial Intelligence*, vol. 37, no. 1, 2023, doi: 10.1080/08839514.2023.2261730.
- [13] A. Farea, S. Tripathi, G. Glazko, and F. Emmert-Streib, "Investigating the optimal number of topics by advanced text-mining techniques: Sustainable energy research," *Eng. Appl. Artif. Intell.*, vol. 136, Oct. 2024, doi: 10.1016/j.engappai.2024.108877.
- [14] Sumit Mishra, "arXiv Scientific Research Paper Dataset," Feb. 2025. Accessed: Mar. 04, 2026. [Online]. Available: <https://data.mendeley.com/datasets/mm6kst3krj/1>
- [15] U. Detthamrong *et al.*, "Topic Modeling Analytics of Digital Economy Research: Trends and Insights," *Journal of Scientometric Research*, vol. 13, no. 2, pp. 448–458, May 2024, doi: 10.5530/jscires.13.2.35.
- [16] D. Yu and B. Xiang, "Discovering topics and trends in the field of Artificial Intelligence: Using LDA topic modeling," *Expert Syst. Appl.*, vol. 225, p. 120114, 2023, doi: <https://doi.org/10.1016/j.eswa.2023.120114>.
- [17] L. Sun and Y. Yin, "Discovering themes and trends in transportation research using topic modeling," *Transp. Res. Part C Emerg. Technol.*, vol. 77, pp. 49–66, 2017, doi: <https://doi.org/10.1016/j.trc.2017.01.013>.
- [18] L. Zhu and D. Luo, "A Novel Efficient and Effective Preprocessing Algorithm for Text Classification," *Journal of Computer and Communications*, vol. 11, no. 03, pp. 1–14, 2023, doi: 10.4236/jcc.2023.113001.
- [19] R. P. F. Afidh and Syahrial, "Pemodelan Topik Menggunakan n-Gram dan Non-negative Matrix Factorization," *Jurnal Informasi dan Teknologi*, pp. 265–275, Sep. 2023, doi: 10.60083/jidt.v5i1.385.
- [20] I. V. de Mendizabal, V. Basto-Fernandes, E. Ezpeleta, J. R. Méndez, S. Gómez-Meire, and U. Zurutuza, "Multi-objective evolutionary optimization for dimensionality reduction of texts represented by synsets," *PeerJ Comput. Sci.*, vol. 9, 2023, doi: 10.7717/peerj-cs.1240.
- [21] D. M. Blei, A. Y. Ng, and J. B. Edu, "Latent Dirichlet Allocation Michael I. Jordan," 2003.
- [22] P. Upendra Dasu, E. A. Fox, and C. Xuan Wang Yinlin Chen, "Topic Modeling for Heterogeneous Digital Libraries: Tailored Approaches Using Large Language Models," 2024.
- [23] A. Amaro and F. Bacao, "Topic Modeling: A Consistent Framework for Comparative Studies," *Emerging Science Journal*, vol. 8, no. 1, pp. 125–139, Feb. 2024, doi: 10.28991/ESJ-2024-08-01-09.
- [24] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, Association for Computing Machinery, Feb. 2015, pp. 399–408. doi: 10.1145/2684822.2685324.