# A Comprehensive Analysis of AI Research Topics Using LDA and LDAvis: Insights from arXiv Data

No Author Given

No Institute Given

**Abstract.** The objective of this paper is to examine the current developments and research trends in the field of international Artificial Intelligence (AI) and to provide valuable references for future research in Computer Science. This study employs a Latent Dirichlet Allocation (LDA) Topic Model and utilizes the Jieba library for text preprocessing of the abstract sections of academic literature. The optimal number of topics is determined by calculating perplexity. The model training yields both "document-topic" and "topic-word" distributions. Based on the training results, topic identification and strength calculations are performed, and a threshold is set to isolate prevalent topics. These topics are analyzed alongside relevant literature to conduct a comprehensive hot topic analysis and predict future research directions. The model's evaluation identifies six prominent topics in Computer Science, including deep learning, reinforcement learning, and natural language processing, among others. The LDA model demonstrates high accuracy in identifying key topics, aiding researchers in understanding the evolution of the field and in guiding future research efforts.

**Keywords:** Text Analysis · Literature Summarizing · Latent Dirichlet Allocation · Research Hotspots.

## 1 Introduction

Artificial Intelligence (AI) has witnessed rapid development in recent years, marked by significant breakthroughs in machine learning algorithms[7]. Systems based on deep learning now possess the capability to perform complex tasks in pattern recognition and prediction[12]. Advances in computing hardware, such as the emergence of GPUs and specialized AI chips, have provided the computational power necessary to process vast amounts of data for machine learning[6]. The abundance of data generated by the Internet, social media, and sensors serves as a rich source of training material for AI algorithms, addressing the critical need for large datasets to drive AI advancements[13].

The evolving field of computer vision has enabled computers to analyze images and video with increasing efficiency and accuracy. Simultaneously, natural language processing (NLP) has made significant strides, empowering machines to undertake various language-related tasks, including voice interaction, translation, and information extraction. The field of robotics is also progressing, integrating AI, computer vision, and other related technologies, resulting in the realization of perception and manipulation in increasingly complex environments[9]. Moreover, AI development has benefited from multidisciplinary collaboration, drawing insights from information science, cognitive science, neuroscience, psychology, control theory, and linguistics[5].

Given the complexity of AI research across diverse fields, it is essential to analyze AI advancements within this intricate landscape. This paper proposes using the Latent Dirichlet Allocation (LDA) model to predict international research hotspots and development trends in computer science, with a focus on AI. By conducting a thematic analysis, this study aims to identify research hotspots, understand the current state of AI development, and anticipate future research directions. Through the application of the LDA model to analyze the prominent topics in international AI research, based on existing global research data, this paper seeks to provide valuable references for future studies in the field.

## 2 Literature Review

The origins of Artificial Intelligence (AI) as a sub-discipline of computer science can be traced back to the 1950s, a period marked by early exploratory efforts to emulate human cognition and simulate intelligence through programming[2]. Notable achievements from this era include the development of proof-of-concept machines and generalized problem solvers, which laid the groundwork for foundational AI research. The 1960s saw rapid advancements in AI, characterized by the emergence of cognitive simulation. During this time, researchers introduced neural network models with learning capabilities and developed expert systems that leveraged knowledge bases to perform reasoning, achieving a level of intelligence comparable to experts in specific, narrow domains. The successes of this period generated considerable interest in AI's potential to have far-reaching impacts.

However, the 1970s witnessed a decline in AI research, largely due to the severe limitations of contemporary computer hardware, leading to what is now known as the "AI winter." During this downturn, funding and progress in AI significantly diminished[1]. Faced with the challenges posed by limited computational power, researchers shifted their focus towards symbolic AI and knowledge representation, which were more feasible given the available resources. In the 1980s, expert systems, reliant on knowledge bases for intelligent reasoning, emerged as a prominent direction in AI research. This period also saw parallel advancements in robotics, and by the mid-to-late 1980s, connectionism began to supplant symbolic AI as the dominant paradigm.

The 1990s marked another challenging period for AI, as research once again hit a bottleneck, leading to a second "AI winter" characterized by reduced enthusiasm[17]. During this time, neural networks and probabilistic statistical methods regained attention, becoming focal points of research. Machine learning emerged as a significant direction in AI research, establishing itself as a key paradigm by the late 1990s[8].

The beginning of the 21st century started in a third major wave of AI research, driven by advancements in statistical learning and deep learning[18]. This new wave of AI has seen widespread industrial applications and has achieved superhuman performance in tasks such as speech recognition and image processing. AI technology is currently in a period of rapid growth and unprecedented progress. In recent years, Kaplan et al. reviewed the evolution of AI technology, from early expert systems to contemporary deep learning[11]. Young et al. examined the development of techniques such as word vector representations, attention mechanisms, and graph neural networks, along with their applications in machine translation and question-answering systems[19]. Pinto et al. explored advancements in deep reinforcement learning, highlighting its role in endowing machines with planning and control capabilities[15]. However, despite these contributions, there remains a lack of systematic overviews and summaries of research hotspots across the entire AI field.

Topic modeling is a class of statistical models designed to perform unsupervised learning on large-scale text corpora, uncovering latent topic information within document collections[4]. As a prominent topic modeling method, the Latent Dirichlet Allocation (LDA) model leverages Bayesian theory to automatically identify multiple topics within a document by analyzing word co-occurrence relationships, assigning a set of high-frequency keywords to each topic.

In the context of analyzing AI research hotspots, the LDA model offers significant advantages, including its ability to fully automate the extraction of topic information from large document sets without the need for manual annotation or predefined dictionaries. For instance, Blei et al. utilized the LDA model to analyze the titles and abstracts of scientific literature, revealing the evolution of relationships among various sub-disciplines within computer science[3]. Mohammad et al. applied LDA to explore changes in research hotspots within the field of data mining[14]. Despite the LDA model's demonstrated potential in academic trend analysis, its application in predicting research hotspots specifically within AI remains limited.

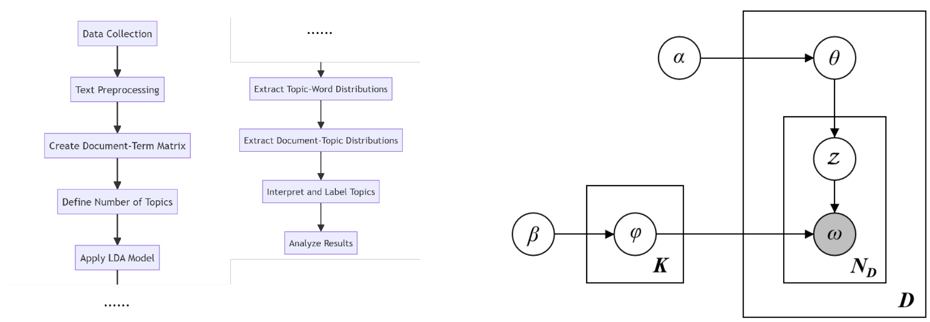
In this study, the LDA topic model is employed to identify research hotspots in the field of computer AI, using articles from arXiv as the text corpus. This analysis aims to uncover the prominent topics in AI research and their evolutionary trends. The findings of this study will provide scholars with a deeper understanding of the discipline's development and offer critical support for decision-makers in formulating research strategies.

## 3 Methods

In this paper, we preprocess the text of AI-related literature from the arXiv website and subsequently perform LDA topic modeling. The optimal number of topics in the model is determined using the perplexity index. Research topics are identified based on the probability distribution of "topic-word" pairs, and hot topics are confirmed by calculating topic intensity. The research results for each hot topic are then visualized and analyzed to explore the current focal points of AI research. The specific steps are as follows: (1) collecting abstracts from literature in the field of AI research; (2) text preprocessing; (3) determining the optimal number of topics by calculating perplexity; and (4) identifying topics based on the keywords associated with each topic.

### 3.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a topic modeling algorithm introduced in 2001 by David M. Blei et al.[4], building upon earlier work in Latent Semantic Analysis (LSA) and Probabilistic Latent Semantic Analysis (PLSA). The fundamental idea behind topic modeling is that documents are composed of a mixture of several topics, each represented by a probability distribution over words. LDA is a text mining technique used for unsupervised machine learning to identify potentially hidden topic information within large collections of documents or corpora .




(a) Flowchart of Modeling (b) Model Structure

Fig. 1: LDA Model

A topic model is a generative model that simulates the cognitive process of a human author composing a document by defining a probabilistic sampling process. The generation process of a document in this model can be described as follows: first, a topic is randomly selected based on the document's topic distribution; next, a word is drawn from the word distribution associated with the selected topic. This process is repeated for each word in the document until the entire document is generated. LDA formalizes this generative process for each document in the corpus: it first samples a topic from the document's topic distribution, then samples a word from the word distribution corresponding to that topic, and continues this process until all words in the document have been processed.

LDA employs a T-dimensional latent random variable, $\theta$, to represent the topic probability distribution of a document, which follows a Dirichlet distribution. Estimating the values of the latent variables in a topic model requires approximate inference algorithms. The most commonly used statistical inference methods include Expectation-Maximization, Variational Inference, and Gibbs Sampling. By solving the Gibbs Sampling algorithm, the distributions of global topics $Z$ and words $W$ can be obtained. Posterior probability distributions, such as $\theta$ and $Z$, are derived through inference algorithms, and model parameters, such as $\alpha$ and $\beta$, can be estimated based on the text's vocabulary. This allows for the extraction of implicit correlations between documents and topics, thereby defining a fixed set of topic-document relationships within a corpus.

The LDA topic model is highly regarded for its robust dimensionality reduction capabilities, strong probabilistic foundation, and scalability, making it particularly effective for processing large-scale text data. As a result, LDA has gained widespread attention in industry and is increasingly applied to the analysis and processing of natural language texts.

Another important hyperparameter of the LDA model, $K$, representing the optimal number of topics, is determined by calculating the perplexity of the model. The formula for calculating perplexity is shown in Equation (1), where $D$ is the test set, $M$ is the number of documents in the test corpus, $P(w_d)$ is the probability of the text, and $N_d$ is the number of words in the document.

$$Perplexity(D_t) = exp[-\frac{\sum_{i=1}^{M} logp(w_d)}{\sum_{i=1}^{M} N_d}] \tag{1}$$

The text probability $P(w_d)$ can be determined using Equation (2), and $p(Z|D)$ represents the distribution of text d on that topic $Z$.

$$P(w_d) = \sum_z P(z, w_d) = \sum_z P(z|d)P(w_d|z) \tag{2}$$

The perplexity measures the accuracy of the LDA topic model in predicting the samples, so theoretically the smaller the perplexity indicates the higher the accuracy of the model prediction. Therefore, the best number of topics should be the number of topics with the lowest degree of perplexity or perplexity inflection point.

### 3.2 LDAvis

LDAvis is a topic visualization method introduced by Sievert and Shirley in 2014[16]. LDAvis facilitates the selection of feature words that represent topics, as well as the evaluation of the association between these feature words and the corresponding topics. The LDAvis visualization graph is particularly useful for observing the interconnections between topics in a holistic manner. Essentially, LDAvis explores the relationships between documents and topics, as well as between topics and words. By using multidimensional scaling, LDAvis projects these relationships into a low-dimensional space, enabling contrast and analysis. The connection between topics and words in LDAvis is determined by two key attributes: the frequency of word occurrence and the uniqueness of the word within the topic. The strength of this connection is measured on a scale from 0 to 1, with the optimal value depending on the specific research question being addressed.

## 4 Experiment

### 4.1 Data

The experimental data for this study is sourced from a dataset containing arXiv papers available on the Kaggle platform[10]. This dataset comprises 1.7 million academic papers and includes various features related to each paper, such as the title, author, category, abstract, and full-text PDF. The arXiv dataset provides a metadata file in JSON format, which includes relevant entries for each paper as follows:

- id: The paper's access address, which can be used to retrieve the paper.
- submitter: The individual who submitted the paper.
- authors: The authors of the paper.
- title: The title of the paper.
- comments: Information about the number of pages, as well as additional details such as graphs and charts.
- journal-ref: Information about the journal in which the paper was published.
- doi: The Digital Object Identifier(DOI) of the paper.
- abstract: The abstract of the paper.
- categories: The categories or tages that the paper is associated with in arXiv.
- versions: The versions of the paper.

### 4.2 Text Preprocessing and LDA Parameters Settings

The downloaded dataset is first organized by filtering AI-related papers using an editor, and then converting the relevant entries into Excel files for subsequent processing. Regular expressions are employed to extract titles, keywords, and abstract information from the document metadata, forming the corpus source for the LDA model. The Jieba component in Python is utilized to perform segmentation on the preprocessed files, while pre-defined stopword lists are loaded to remove stopwords from the corpus, resulting in the document-word matrix.

Next, the LDA model is constructed using the Scikit-learn package in Python. Prior to model construction, the optimal number of topics must be determined. In this paper, the optimal number of topics is identified by calculating the perplexity of the topic model. The process involves the following steps:

1. Setting the range for the number of topics to [0, 9] with a step size of 1. The parameters $\alpha$ (document-topic distribution prior) and $\beta$ (topic-word distribution prior) are set to their default values.
2. Calculating the perplexity for each number of topics, with the number of topics corresponding to the lowest perplexity being selected as the optimal value.

The program yields two primary results: the document-topic distribution and the topic-word distribution. The following figures illustrate the data processing stages: the original corpus (Fig. 2), the corpus after Jieba segmentation and stopword removal (Fig. 2), and the trend in perplexity as the number of topics increases (Fig. 3).

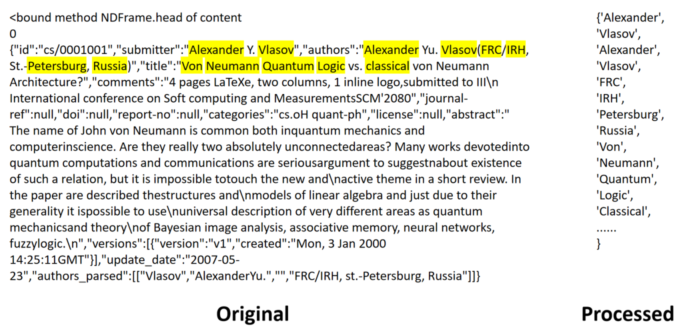



Fig. 2: Corpus

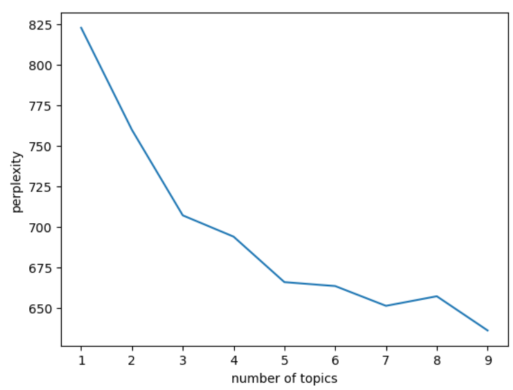



Fig. 3: Perplexity fluctuation with topic numbers

As shown in Fig. 3, the perplexity decreases progressively as the number of topics increases, which aligns with our initial theoretical expectations. However, while lower perplexity indicates that a higher number of topics is more suitable for modeling the existing documents, an excessive number of topics can lead to overfitting, thereby diminishing the model's ability to generalize. To avoid this, I selected the inflection point on the curve—corresponding to six topics—for modeling the corpus.

## 5 Results & Analysis

After theme modeling, the top six themes are listed in the following three tables. These tables lists the 9 most frequently occurring keywords under each theme. Through the words we can identify what areas each of the 6 themes belongs to, thus helping us to summarize and generalize the theme heat trends.

Table 1: Keywords 1-3 From Different Topics

| Topic | Keyword 1 | Keyword 2 | Keyword 3 |
|---|---|---|---|
| Topic 1 | networks | convolutional | classification |
| Topic 2 | learning | gradients | approximate |
| Topic 3 | segmentation | tracking | graph |
| Topic 4 | word | language | syntactic |
| Topic 5 | Bayesian | probabilistic | chains |
| Topic 6 | feature | vector | decision |

Table 2: Keywords 4-6 From Different Topic

| Topic | Keyword 4 | Keyword 5 | Keyword 6 |
|---|---|---|---|
| Topic 1 | detection | semantic | transfer |
| Topic 2 | simulation | Markov | agents |
| Topic 3 | image | video | action |
| Topic 4 | entity | text | translation |
| Topic 5 | stochastic | Gaussian | variational |
| Topic 6 | unsupervised | clustering | dimensionality |

Table 3: Keywords 7-9 From Different Topic

| Topic | Keyword 7 | Keyword 8 | Keyword 9 |
|---|---|---|---|
| Topic 1 | feature | recognition | analysis |
| Topic 2 | hereditary | game | deep |
| Topic 3 | clouds | reconstruction | remote |
| Topic 4 | sentiment | dialog | corpora |
| Topic 5 | inference | random | optimization |
| Topic 6 | ensemble | regression | fuzzy |

The LDA model categorizes words into topics based on their co-occurrence but does not inherently provide named themes for these topics. However, by analyzing the keywords and leveraging pre-existing classifications on arXiv, we can interpret and assign themes to the identified topics.

Theme 1 includes terms such as convolutional, semantic, transfer, feature, and recognition. These terms are closely associated with computer vision and deep learning. Specifically, "convolutional" refers to convolutional neural networks used for image classification and object detection, while "semantic" and "transfer" pertain to semantic understanding and transfer learning techniques. "Feature" and "recognition" relate to feature extraction and pattern recognition. Collectively, these terms reflect the theme of "Deep Learning," which encompasses current research hotspots in image processing and semantic understanding, including visual understanding models and algorithms based on deep learning.

Theme 2 features keywords such as learning, gradients, approximation, simulation, Markov, agents, hereditary, game, and deep. Terms like "simulation" relate to learning and modeling methods, while "Markov" and "agents" denote Markov decision processes and intelligence. "Hereditary" and "game" correspond to genetics and game theory, alongside "deep" learning. This theme pertains to "Reinforcement Learning," focusing on training methods, decision modeling, and game theory solutions in reinforcement learning. Specifically, "gradients," "approximation," and "simulation" relate to policy gradients, approximate dynamic programming, and simulation environments, with Markov decision processes serving as mathematical frameworks for reinforcement learning.

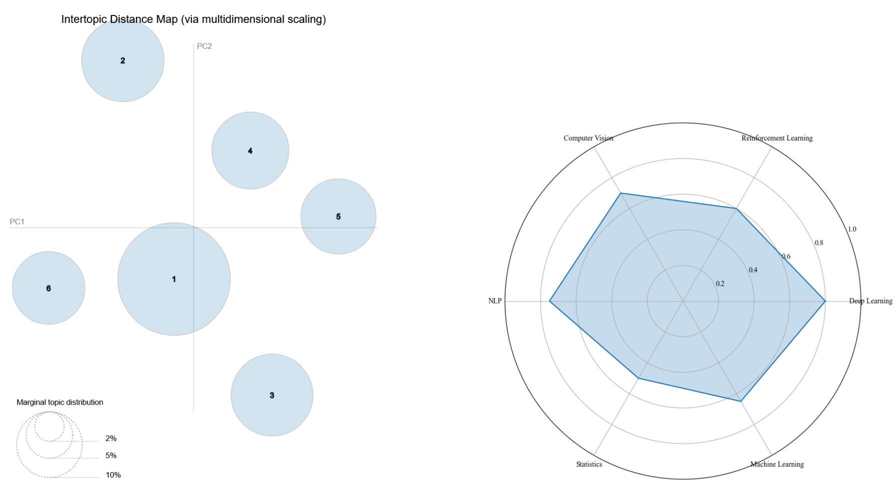
Theme 3 includes terms such as segmentation, tracking, graph, image, video, action, clouds, reconstruction, and remote. "Segmentation," "tracking," "image," and "video" relate to image and video analysis techniques, while "action" pertains to motion recognition. "Reconstruction" refers to three-dimensional reconstruction, "clouds" relates to point clouds, and "remote" pertains to remote sensing applications. These terms collectively address the themes of image/video understanding and 3D reconstruction in computer vision, summarizing the theme as "Computer Vision."

Theme 4 comprises words such as word, language, syntactic, entity, text, translation, sentiment, dialog, and corpora. Terms like "word," "language," "syntactic," and "text" are linked to language understanding and text analysis in natural language processing (NLP). "Translation" refers to machine translation, "sentiment" to sentiment analysis, "dialog" to dialog systems, and "corpora" to text corpora. These terms are associated with language and text analysis tasks, summarizing the theme as "Natural Language Processing."

Theme 5 features terms like Bayesian, probabilistic, chains, stochastic, Gaussian, variational, inference, random, and optimization. These terms are related to probabilistic graphical modeling and variational Bayesian inference. Specifically, "Bayesian," "probabilistic," "stochastic," and "Gaussian" pertain to statistical methods, while "chains" refer to Markov chains, and "variational inference" denotes a method in Bayesian statistics. The theme can be summarized as "Statistics and Probability," which forms a critical theoretical foundation for machine learning and AI research.

Theme 6 includes keywords such as feature, vector, decision, unsupervised, clustering, dimensionality, ensemble, regression, and fuzzy. Terms like "feature" and "vector" are related to feature engineering, "decision" and "regression" to supervised learning, while "unsupervised" and "clustering" pertain to unsupervised learning. "Dimensionality reduction" is a technique used for preprocessing, and "ensemble" refers to methods that aggregate multiple models to improve performance. "Fuzzy" logic is related to fuzzy systems and soft computing. This theme encompasses directions in feature engineering, supervised learning, unsupervised learning, dimensionality reduction, and ensemble learning, and is summarized as "Machine Learning."

Fig. 4 illustrates the topics using LDAvis and radar chart, providing a visual representation of the keyword distribution across different topics within the corpus.




(a) LDAvis Mapping

(b) Radar Chart

Fig. 4: Topic Modeling Visualization

## 6 Conclusion

In this study, LDA topic modeling was employed to mine topics from papers in the field of artificial intelligence research. The LDA model automatically identifies topic information within a collection of papers and extracts representative keywords based on word frequencies. Unlike manually specified topics, LDA modeling offers data-driven, unsupervised advantages, providing a more objective reflection of topic distributions in the papers.

By analyzing 25 representative keywords for each theme, this study summarizes the development hotspots and trends in the AI research field:

1. Deep Learning: The first theme indicates that deep learning remains a prominent research area. Keywords such as convolutional neural networks, image classification, and object detection highlight ongoing advancements in deep learning methods. Concepts like transfer learning and augmentation learning demonstrate the expansion and refinement of deep learning techniques.
2. Reinforcement Learning: The second theme focuses on reinforcement learning, emphasizing its advancements in decision-making and control through


Fig. 5: Wordcloud of Keywords

simulated environments and reward mechanisms. Significant progress in gaming and robotics is noted, reflecting the evolution of reinforcement learning methods.

3. Computer Vision: The third theme reflects developments in computer vision, including improvements in image classification and object detection. New application areas such as frequency analysis and medical imaging also emerge as significant. This theme underscores computer vision's continuing importance as an application area for deep learning.
4. Natural Language Processing: The fourth theme features keywords related to language modeling and text analysis, including word vectors, text classification, machine translation, and dialog systems. This theme highlights the role of deep learning in advancing natural language processing.
5. Statistics and Probability: The fifth theme includes terms related to probabilistic modeling and inference, such as Bayesian methods, probabilistic graphical models, and variational inference. These components are crucial for probabilistic modeling, with Bayesian methods showing increased potential in conjunction with deep learning advancements.
6. Machine Learning: The sixth theme covers classical machine learning methods, including support vector machines and decision trees. These techniques provide foundational support for more complex deep learning methods, emphasizing their continued relevance as the field progresses.

Overall, the study indicates that deep learning continues to be a central research focus, with significant progress in applications such as image classification and speech recognition. Reinforcement learning, computer vision, and natural language processing remain key areas of development, driven by innovations in deep learning. Bayesian methods and classical machine learning techniques continue to offer essential theoretical and engineering support. As the field evolves, these traditional techniques will likely maintain their importance.

## 7 Discussion & Limitation

LDAvis is a visualization tool that enhances the understanding of topic distributions and the significance of keywords through scatterplots based on topic-related data. This visualization allows for a deeper comprehension of the relationships between topics, revealing similarities and intersections that might not be apparent from a simple list of topics. By transforming abstract theme distributions into intuitive visual representations, LDAvis aids users in grasping the relationships between themes and evaluating the effectiveness of topic modeling. The combination of LDA modeling and LDAvis thus provides robust support for both quantitative analysis and thematic visualization. Future research will leverage LDAvis to further enhance the interpretability of text theme analysis.

This study primarily relies on arXiv data, which, while extensive, may not fully encapsulate the entire spectrum of AI research. Publications in other venues, such as traditional journals or conference proceedings not archived in arXiv, are not represented in our analysis. This limitation potentially introduces a bias towards certain types of research or researcher demographics that preferentially use arXiv. The timeframe of the analyzed data is not explicitly defined in our methodology. This omission hampers the ability to contextualize our findings within specific periods of AI development and may limit the relevance of our conclusions to current trends.

While the LDA model objectively identifies word clusters, the process of interpreting and labeling these clusters as coherent topics introduces an element of subjectivity. This interpretation could be influenced by the researchers' backgrounds and perspectives.

To address these limitations and extend the scope of this research, we propose to implement a time-series analysis to track the evolution of AI research topics, identifying emerging trends and declining areas of interest, and expand the dataset to include publications from diverse sources, such as IEEE, ACM, and major AI conferences, to provide a more comprehensive view of the field.

By addressing these limitations and pursuing these future research directions, we aim to provide a more nuanced, dynamic, and comprehensive understanding of the AI research landscape. This extended analysis would not only benefit academic researchers in identifying promising areas of study but also inform policymakers and industry leaders about the trajectories of AI development.

## References


1. Alberto P., Beatriz P., E.M.O.E.R.F.P.J.O., R., I.: Neural networks: An overview of early research, current frameworks and new challenges. Neurocomputing **214**, 242–268 (2016). https://doi.org/https://doi.org/10.1016/j.neucom.2016.06.014, https://www.sciencedirect.com/science/article/pii/S0925231216305550
2. Andrzej G., K.P., Lambert, W.C.: A history of artificial intelligence. Clinics in Dermatology **42**(3), 221–229 (2024). https://doi.org/https://doi.org/10.1016/j.clindermatol.2023.12.016, https://www.sciencedirect.com/science/article/pii/S0738081X23002687, dermatology and Artificial Intelligence
3. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: International Conference on Machine Learning (ICML). pp. 113–120 (2006)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. In: Conference on Neural Information Processing Systems (NeurIPS). pp. 601–608 (2001)
5. Chen, Z., Yadollahpour, A.: A new era in cognitive neuroscience: the tidal wave of artificial intelligence (AI). BMC Neuroscience **25**(1) (2024)
6. G., D.: How cutting-edge computer chips are speeding up the ai revolution. Nature **630**(8017), 544–546 (Aug 2024). https://doi.org/10.1038/d41586-024-01544-0, https://www.nature.com/articles/d41586-024-01544-0
7. H., X.: The promising future of cognitive science and artificial intelligence. In: Nat Rev Psychol 2, 202 (2023). https://doi.org/https://doi.org/10.1038/s44159-023-00170-3
8. Haton, J.P.: A brief introduction to artificial intelligence. IFAC Proceedings Volumes **39**(4), 8–16 (2006)
9. Intel: Robots and artificial intelligence overview, https://www.intel.com/content/www/us/en/robotics/artificial-intelligence-robotics.html
10. Kaggle: arxiv dataset (Aug 2024), https://www.kaggle.com/datasets/Cornell-University/arxiv
11. Kaplan, A., Haenlein, M.: Siri, siri, in my hand: Who's the fairest in the land? on the interpretations, illustrations, and implications of artificial intelligence. Business Horizons **62** (11 2018). https://doi.org/10.1016/j.bushor.2018.08.004
12. LeCun, Y., B.Y..H.G.: Deep learning. In: Nature 521, 436–444 (2015). https://doi.org/https://doi.org/10.1038/nature14539
13. Liu, J., Kong, X., Xia, F., Bai, X., Wang, L., Qing, Q., Lee, I.: Artificial intelligence in the 21st century. IEEE Access **6**, 34403–34421 (2018). https://doi.org/10.1109/ACCESS.2018.2819688
14. Mohammad, S.M., Yang, T.: Tracking Sentiment in Mail: How Genders Differ on Emotional Axes. In: Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA). pp. 70–79 (2011)
15. Pinto, L., Davidson, J., Gupta, A.: Supervision via competition: Robot adversaries for learning tasks. In: IEEE International Conference on Robotics and Automation (ICRA). pp. 1601–1608 (2017)
16. Sievert, C., Shirley, K.: Ldavis: A method for visualizing and interpreting topics. In: Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. Association for Computational Linguistics (2014)
17. Wang, H., Raj, B., Xing, E.P.: On the Origin of Deep Learning. arXiv **abs/1702.07800** (2017)

18. Xie, Z., Li, M., Li, J., Duan, X., Ouyang, Z.: Feature analysis of multidisciplinary scientific collaboration patterns based on PNAS. EPJ Data Science **7**(1) (2018)
19. Young, T., Hazarika, D., Poria, S., Cambria, E.: Recent Trends in Deep Learning Based Natural Language Processing. IEEE Computational Intelligence Magazine **13**(3), 55–75 (2018)