Analyzing and discussing various trends in AI conferences: ICML, ICLR and NeurIPS

Aryan Solanki
IIT Gandhinagar
23110049@iitgn.ac.in

Nupoor Assudani
IIT Gandhinagar
23110224@iitgn.ac.in

Rishabh Jogani
IIT Gandhinagar
23110276@iitgn.ac.in

Aditya Jain
IIT Gandhinagar
23110016@iitgn.ac.in

ABSTRACT

This study presents a comprehensive analysis of machine learning research trends from 2014 to 2024, drawing on publications from ICML, NeurIPS, and ICLR. We first quantify the growth of industry-affiliated contributions relative to academic outputs, noting that by 2024 industry-authored papers approach one-fourth of academic volumes. Next, we reveal a paradigm shift in NeurIPS abstracts—from statistical and probabilistic methods (2014–2016) to deep learning architectures, attention mechanisms, and adversarial robustness (2017–2019). We then examine collaboration patterns, reporting a median of four authors per paper and a maximum of 76 co-authors on a single work. We also document increasing interdisciplinary influence: titles referencing biology, physics, or neuroscience rise markedly after 2018. Finally, we uncover regional and institutional specializations: U.S. institutions emphasize fairness, causal inference, and graph learning; Chinese institutions focus on federated and semi-supervised learning and adversarial attacks; corporate research (e.g., Google, Meta, Microsoft) prioritizes large language models, self-supervised learning, and optimization. **All codes and notebooks are available on GitHub.**

1 DATA SUMMARY

In this study, we utilize two comprehensive datasets to analyze publication trends across major machine learning conferences:

(1) **ICML, NeurIPS, and ICLR Paper Dataset**
Source: GitHub Repository by Marten Lienen
This dataset contains metadata for papers from three prominent conferences:
  - **ICML (International Conference on Machine Learning):** 2017–2024
  - **NeurIPS (Conference on Neural Information Processing Systems):** 2006–2024
  - **ICLR (International Conference on Learning Representations):** 2018–2024 (excluding 2020)

  The dataset includes columns for the conference name, year, paper title, author, and author affiliation. This metadata supports analysis of temporal and institutional trends in machine learning research.

(2) **All NeurIPS Papers (1987–2019) Dataset**
Source: Kaggle Dataset by Rohit Swami
This dataset provides a more detailed view of NeurIPS publications from 1987 to 2019, including:
  - Year of publication
  - Paper title
  - Authors
  - Abstracts
  - Full text content of papers

  This richer content enables deeper semantic analyses such as topic modeling, syntactic structure analysis, and keyword frequency studies.

2 KEY INSIGHTS

We conducted an exploratory data analysis of two datasets (see GitHub) to validate our hypotheses. Here are the key takeaways:

- **Leading Authors:** Sergey Levine, Yoshua Bengio, and Stefano Ermon are the top three most frequently mentioned authors, with 209, 169, and 147 mentions respectively.
- **Dominant Institutions:** Stanford University, Google, and Carnegie Mellon University lead in publication volume, each contributing over 1,700 papers.
- **Publication Peaks by Author:** Sergey Levine set a record in 2024 with 54 publications, highlighting accelerated author productivity.
- **Institutional Yearly Highs:** Google dominated outputs from 2016–2023, while Tsinghua University surged to 461 papers in 2024.
- **Collaboration Intensifies:** The median number of authors per paper is 4.00; the maximum of 76 authors appeared on a single paper ("CVQA: Culturally-diverse Multilingual Visual Question Answering Benchmark"), indicating growing team-based research.
- **Affiliation Growth Trend:** Google's yearly paper count grew from 32 in 2016 to 345 in 2023, reflecting expanding corporate research efforts.
- **Implication:** Rising author and affiliation counts underscore deeper collaboration networks and increasingly complex challenges in top ML conferences.

3 HYPOTHESIS

3.1 First Hypothesis

> *Over the years 2017 through 2024, the number of ICML publications authored by industry-affiliated researchers has steadily increased, and by 2024, it has reached approximately one-fourth the number of publications authored by academia-affiliated researchers.*

Motivation. The balance between academic and industry contributions in a top-tier conference like ICML reflects the shifting landscape of machine learning research. In recent years, industry labs have become major players in publishing impactful work. Analyzing their presence helps reveal trends in research ownership, collaboration, and the influence of commercial interests.

Approach. To investigate this hypothesis, we a dataset of ICML publications spanning 2017 to 2024. A crucial part of the analysis involves classifying each author's affiliation as either *Industry* or *Academia*. This is done in two stages:

- **Heuristic Annotation:** A rule-based approach is used initially, we match keywords (e.g., university, , institute, and common company names) to annotate a majority of affiliations.
- **LLM-Assisted Annotation (OpenAI):** For affiliations that remain ambiguous we use an OpenAI api to semantically classify them. This ensures a complete annotation.

After annotating the affiliations, we compute and visualize the yearly distribution of publications by affiliation type.

3.2 Second Hypothesis

> *The most common words in the abstracts of NeurIPS papers from 2017–2019 show a noticeable evolution compared to those from 2014–2016, reflecting the rapidly changing landscape of machine learning research in just a few years.*
>
> *While earlier abstracts (2014–2016) commonly referenced statistical/probabilistic learning, clustering, and kernel methods, the later years (2017–2019) exhibit a stronger emphasis on deep learning architectures, reinforcement learning, and emerging trends such as attention mechanisms and adversarial robustness.*

Motivation. This hypothesis is important because it highlights the shifts in machine learning research trends over a short period, providing insights into the evolving focus areas and methodologies. Understanding these changes can provide explainability for future research directions (post 2019), funding priorities, and explain current research motivation. It also highlights how quickly the field is evolving, with topics like deep learning and reinforcement learning leading the way.

Approach. To examine the linguistic shift in NeurIPS abstracts between 2014–2016 and 2017–2019, we conduct a comparative textual analysis with three techniques: (1) *Log-odds ratio* with additive smoothing to identify discriminative terms across the two time periods, (2) extraction and frequency analysis of *n-grams* (bigrams and trigrams) to capture key phrases and trends, and (3) *network analysis* of word co-occurrence graphs to uncover thematic clusters and semantic drift. These methods collectively enable both low word level comparison and higher level topic structure analysis. Based on the results, we assess the extent to which the focus of ML research has shifted from classical statistical methods to modern deep learning paradigms is supported by the textual evidence.

3.3 Third Hypothesis

> *The frequency of interdisciplinary terms from domains such as biology, physics, and neuroscience in titles has increased from 2018 to 2024 in ICLR, NeurIPS, and ICML, indicating growing cross-domain influence in ML research.*

Approach. To examine the rise of interdisciplinary influence in machine learning research, we analyzed paper titles from ICLR, NeurIPS, and ICML spanning the years 2018 to 2024. Our goal was to track the frequency of terms originating from three scientific domains: biology, physics, and neuroscience.

Seed Dictionaries. We manually made domain-specific seed dictionaries containing representative keywords from each field. For example, the biology dictionary (`bio_seeds`) included terms from molecular biology, genetics, cell biology, physiology, and ecology, such as:

```
genome, protein, dna, enzyme, mitochondria,
adaptation, ecosystem, cell, hormone, ...
```

Similar dictionaries were constructed for physics and neuroscience. To reduce false positives from the ML terminology that may semantically overlap with these domains, we explicitly excluded ambiguous terms such as `language`, `neuron`, `attention`, `memory`, `learning`, `activation`, and others.

(1)*Tokenization and Matching.* We removed duplicate titles across venues and tokenized each title using a simple word-level tokenizer. Each token was then compared against the seed dictionaries to determine if it belonged to one or more target domains. A title was marked as a "hit" for a domain if it contained at least one seed word from that domain. (2)*Frequency Aggregation.* For each year, we calculated the proportion of titles containing biology, physics, or neuroscience terms, yielding the `bio_ratio`, `physics_ratio`, and `neuro_ratio`, respectively.

3.4 Fourth Hypothesis

> *Between 2018 and 2024, paper titles from US, China, and Corporate affiliations in ICLR and ICML disproportionately emphasize different machine learning subfields, revealing regional and institutional specialization in research focus.*
>
> *Specifically, titles from US institutions are more likely to highlight areas such as Fairness, Causal Inference, and Graph Learning; Chinese institutions tend to focus on Federated Learning, Semi-supervised Learning, and Adversarial Attacks; while Corporate-affiliated papers (e.g., from Google, Meta, Microsoft) emphasize topics like Large Language Models, Self-supervised Learning, and Optimization.*

Motivation. This hypothesis is important because it reveals how *research priorities in machine learning differ across regions and institutions*, reflecting broader socio-technical influences. By identifying distinctive thematic focuses—such as fairness and optimization in US academia, vision and distribution estimation in Chinese research, and large language models and efficiency in industry—this analysis helps map the *divergent trajectories* of ML innovation. Such insights are critical for *collaboration strategies, policy-making, and anticipating future research trends* across global AI ecosystems.

Approach. To examine regional and institutional distinctions in NeurIPS submissions, we apply two complementary techniques: (1) *Pairwise log-odds ratio* with informative Dirichlet priors to identify discriminative words between any two groups (e.g., US vs. China), and (2) *Multiclass log-odds classification* to find globally distinctive terms across all groups. To validate these findings at a thematic level, we use *Latent Dirichlet Allocation (LDA)* to uncover dominant topics per group. Together, these methods offer both *lexical-level specificity and abstract topic-level insights*, supporting the hypothesis of differentiated research agendas in the ML community.

4 ANALYSIS AND SETTLING THE HYPO

4.1 First Hypothesis

Results. The number of ICML publications authored by industry-affiliated researchers has steadily increased from 375 in 2017 to 1837 in 2024, reflecting nearly a fivefold growth. However, the proportion of industry to academia-affiliated publications in 2024 is approximately 20.2% (1837 out of 9095), falling short of the hypothesized one-fourth ratio. While industry contributions have grown consistently, the data reveals that academia has expanded at an even faster pace, leading to a gradual decline in the relative share of industry publications over time.

Year	Academia	Industry	Industry / Total (%)
2017	955	375	28.2%
2018	1503	491	24.6%
2019	1870	617	24.8%
2020	2824	903	24.2%
2021	3269	1013	23.7%
2022	3626	1038	22.3%
2023	5460	1483	21.4%
2024	9095	1837	16.8%

Table 1: Yearly ICML publication counts by affiliation and the percentage of industry-affiliated papers out of the total.

4.2 Second Hypothesis

1. Log-Odds Discriminative Vocabulary

The discriminative vocabulary obtained using log-odds ratio reveals a clear thematic divergence between the two eras:

- **Era B (2017–2019)** is characterized by terms such as *adversarial, deep, different, gradient, learning, network, networks, neural, performance, policy, reinforcement, state-of-the-art, tasks, trained, training*. These reflect the rise of deep learning, reinforcement learning, and performance-centric approaches.
- **Era A (2014–2016)** displays a focus on terms like *algorithm, clustering, computational, data, general, inference, kernel, large, linear, matrix, probabilistic, problem, problems, sparse, statistical*. This vocabulary aligns with traditional ML approaches, structured models, and theoretical analysis.

This suggests a clear shift from probabilistic and kernel-based methods to deep, task-specific learning frameworks.

2. Co-occurrence Network Structure

The co-occurrence networks (Figure **??** and Figure **??**) further support the observed lexical and thematic transitions.

- **Era A** shows a network with strong connections around terms like *sparse*, *matrix*, *linear*, *inference*, and *kernel*, which are central to classical ML paradigms such as matrix factorization and probabilistic graphical models.
- **Era B** displays denser interconnections around *deep*, *neural*, *training*, *performance*, and *adversarial*, indicating an ecosystem dominated by deep learning, training optimization, and neural architectures.

This evolution of term connectivity hints at the growing complexity and integration of methods in modern ML research.

3. N-gram Trends

The bigram and trigram analysis further substantiates the hypothesis, offering insights into evolving research topics:

- **Era B** featured popular phrases such as *graph neural networks (GNNs)*, *neural architecture search (NAS)*, *generative adversarial networks (GANs)*, *non-convex optimization problems*, *reinforcement learning algorithms*, *latent representations*, and *continual learning*.
- **Era A** favored n-grams such as *matrix decomposition*, *chain graph*, *object proposal*, *activity set*, *guided policy search*, *gaussian graphical models*, *nearest neighbour classification*, *lifted inference algorithms*, and *sparse logistic regression.*

These results highlight the transition from foundational, theory-heavy problems to modern, neural, and representation-focused techniques.

Conclusion

The analysis across log-odds vocabulary, co-occurrence networks, and n-gram patterns provides strong evidence for a thematic shift in ML research. Era A emphasizes theoretical and structured approaches, while Era B marks a definitive move toward data-driven, neural-centric, and application-oriented methodologies.

4.3 Third Hypothesis

Year	Biology	Physics	Neuroscience	Total Titles
2018	30	25	3	1966
2019	34	38	13	2701
2020	33	38	14	2982
2021	63	64	24	4377
2022	100	83	17	5229
2023	168	134	45	7033
2024	248	206	60	9466

Table 2: Year-wise count of titles containing terms from biology, physics, and neuroscience.

Analysis. We observe a clear upward trend in the use of interdisciplinary terms across all three domains. From 2018 to 2024, biology-related terms saw a more than eightfold increase, while physics and neuroscience terms also rose substantially. This indicates a growing cross-pollination between machine learning and fields like biology, neuroscience, and physics—suggesting that ML research is becoming increasingly integrative and application-driven.

4.4 Fourth Hypothesis

To evaluate our hypothesis that *US research is theory-driven*, *Chinese research is application-focused*, and *Industry is product-oriented*, we analyze the vocabulary and topic distributions across groups using log-odds ratios and LDA.

1. Pairwise Log-Odds Discriminative Vocabulary

Comparison	Top Distinctive Terms
US vs China	linear, bandits, fairness, poisoning, convergence *(Theory-heavy)* vs. graph, heterogeneous, spiking, kernel *(Structural/data-specific)*
US vs Industry	robust, markov, convergence, provable, approximation *(Theoretical)* vs. attacks, analysis, data *(Security, practical)*
China vs Industry	graph, detection, domain, gaussian, bounds *(Applications, structural)* vs. robust, method, streaming *(Practical focus)*

Table 3: Top 10 log-odds discriminative terms across group comparisons

Conclusion: This supports the hypothesis. US leans theoretical, China focuses on data and application domains, and Industry emphasizes robustness, implementation, and scalability.

2. Unique Vocabulary Per Group

Group	Top Unique Words (Log-Odds)
US	harbor, interpolants, semidefinite, chatgpt, remote *(Mathematical foundations + real-world bridge)*
China	energyguided, underwater, glmb, symbol, linearised *(Application-specific and niche domains)*
Industry	trillions, wavenet, musicrl, streamingqa, hyperprompt *(Product/scale-oriented keywords)*

Table 4: Unique high log-odds words per group

Conclusion: Strongly supports the hypothesis—vocabulary reveals that Industry is pushing deployable innovation, China is domain-specific, and the US blends theory with emerging ideas.

3. LDA Topics

Group	Top Topics (Condensed)
US	RL, optimization, convergence, contrastive learning, LLMs
China	Multi-label learning, graph models, diffusion, image modeling, time series
Industry	LLMs, generative models, efficient scaling, human-in-the-loop, video/data processing

Table 5: Top topics via LDA per group

Conclusion: US continues to emphasize theory (RL, optimization), China shows strong alignment with application-driven domains, and Industry focuses on deployability, generation, and efficiency—consistent with the hypothesis.

5 DATA VISUALIZATION

5.1 General Insights

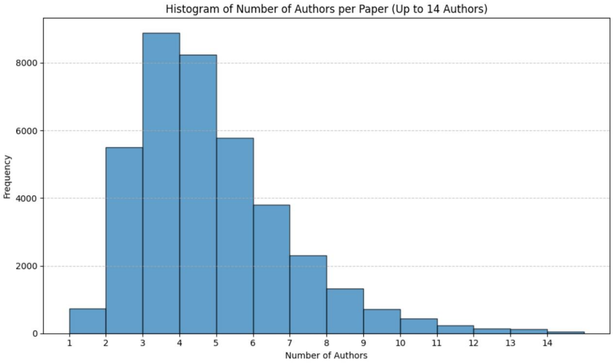

Figure 1: Number of authors per paper.

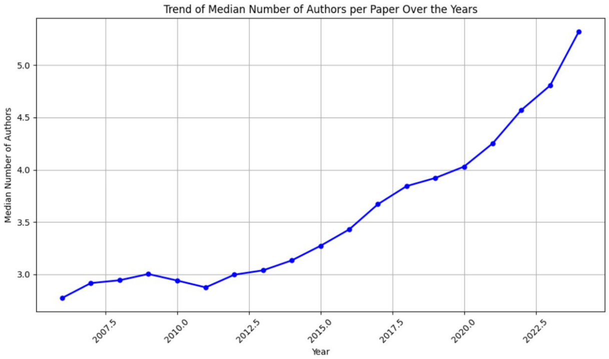

Figure 2: Number of authors per paper over years.

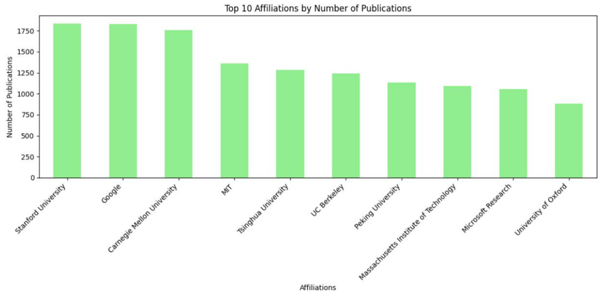

Figure 3: Histogram of affiliations by number of publications.

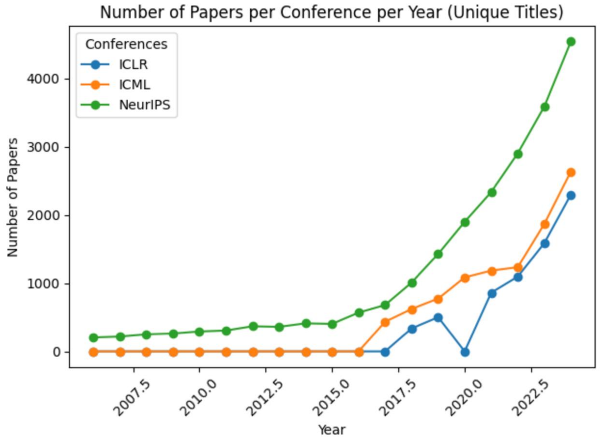

Figure 4: Number of papers per conference per year. (Note: Our dataset has some data missing as mentioned in Dataset summary)

5.2 Hypothesis specific Insights

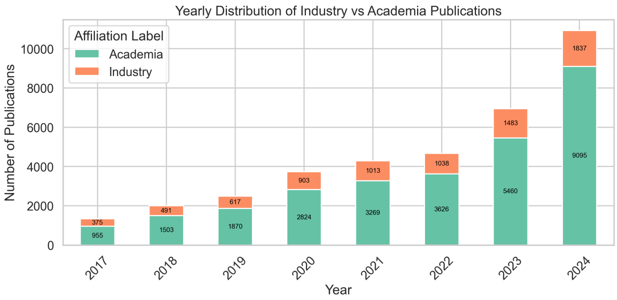

Figure 5: Number of academic and industry affiliated ICML publications per year from 2017 to 2024.

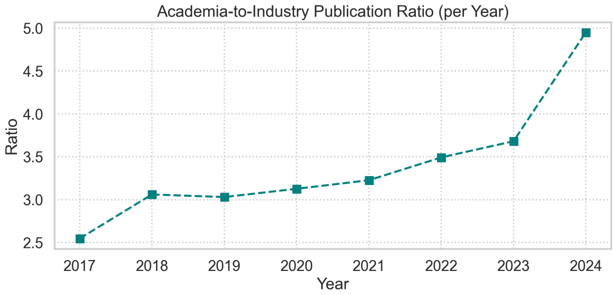

Figure 6: Ratio of academic to industry-affiliated ICML publications per year from 2017 to 2024.

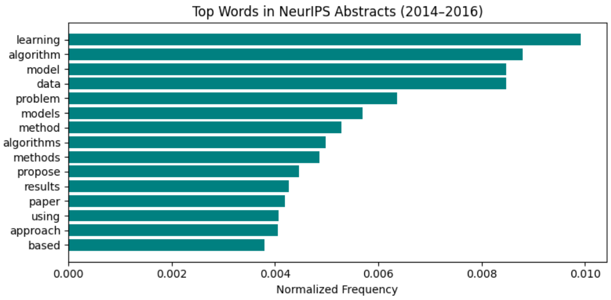

Figure 7: Most frequently occurring words in the abstracts of NeurIPS publications from 2014 to 2016.

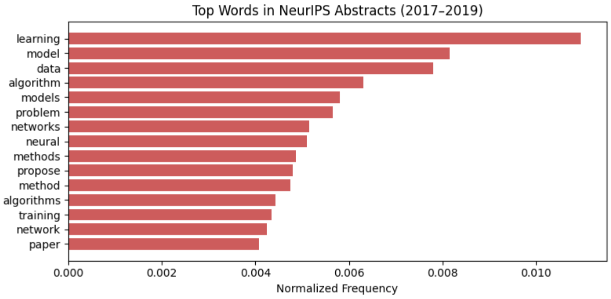

Figure 8: Most frequently occurring words in the abstracts of NeurIPS publications from 2017 to 2019.

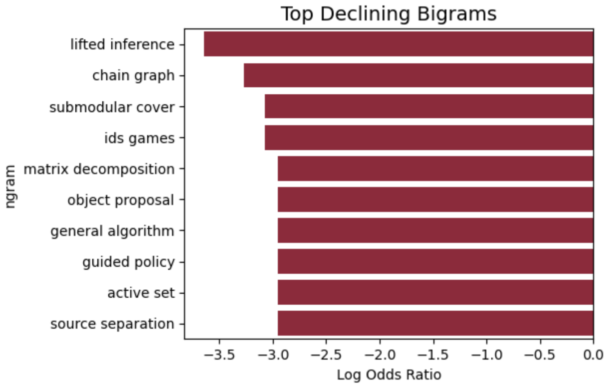

Figure 11: Top declining bigrams in NeurIPS paper abstracts from 2014–2016 to 2017–2019. These bigrams saw the greatest drop in frequency, suggesting a shift away from certain topics or terminology.

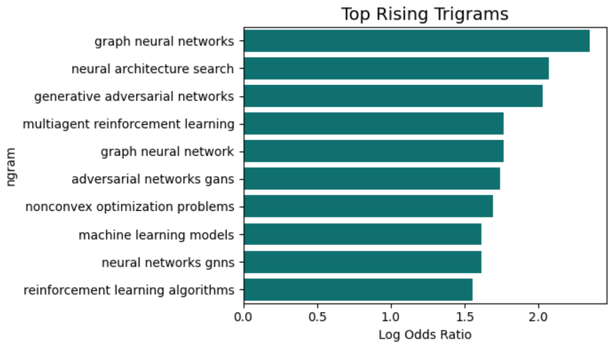

Figure 12: Top rising trigrams in NeurIPS paper abstracts from 2014–2016 to 2017–2019. These trigrams show the most significant increase in usage, indicating emerging research themes during the later period.

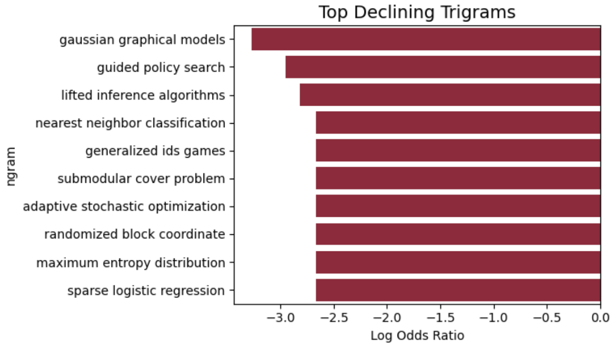

Figure 13: Top declining trigrams in NeurIPS paper abstracts from 2014–2016 to 2017–2019. These trigrams saw the greatest drop in frequency, suggesting a shift away from certain topics or terminology.

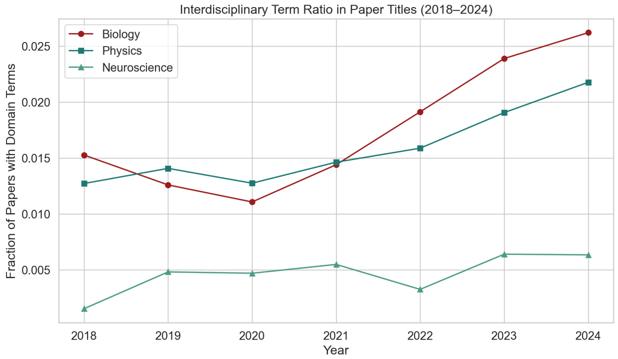

Figure 14: Fraction of papers (2018–2024) containing domain-specific terms from biology, physics, and neuroscience in their titles.

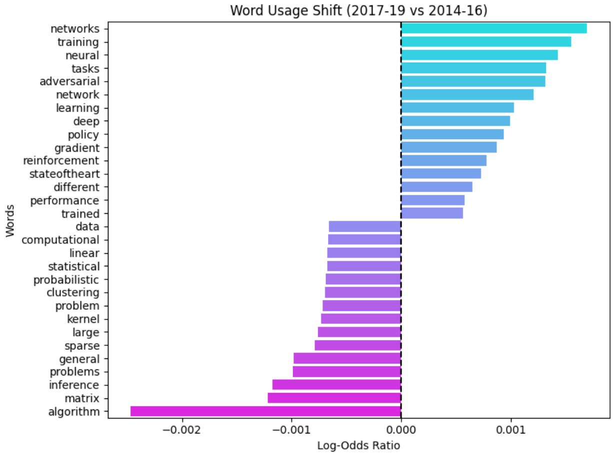

Figure 9: Log-odds ratio of word usage in NeurIPS abstracts, comparing papers from 2017–2019 to those from 2014–2016. Positive values (right) indicate words that became more prominent in the later period, while negative values (left) highlight terms that were more common in the earlier period.

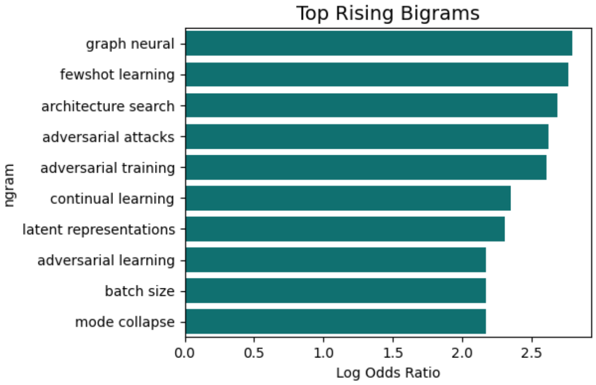

Figure 10: Top rising bigrams in NeurIPS paper abstracts from 2014–2016 to 2017–2019. These bigrams show the most significant increase in usage, indicating emerging research themes during the later period.