AI Integration in the Workforce:
A Multi-Industry Sentiment and
Technology Trend Analysis

Monica Ko

Agenda

- Executive Summary

- Article clean-up and filtering

- Topic Detection

- Entity extraction (people, organizations, technologies)

- Topic-level sentiment analysis

- Entity-level sentiment analysis

- Conclusion

Executive Summary

This project analyzes 200K+ AI-related articles (2022–2025) to understand how AI is reshaping industries, jobs, and regions through entity sentiment and timeline tracking.

Key Findings

• **Industries Affected**: Technology, finance, and healthcare show both strong AI adoption and signs of disruption.

• **Jobs at Risk**: Jobs like teaching and law are facing automation pressures. Human-centered roles, such as UX design and nursing, are less affected.

• **Tech Drivers**: ChatGPT, Bard, and Generative AI peaked in early 2023, triggering sentiment shifts.

• **Global Focus**: The US and India dominate discussions, reflecting regulatory and innovation leadership.

• **Action Needed**: As major technology firms drive innovation, growing concerns around ethics and job loss highlight the need for policy, reskilling, and collaboration.

Data Overview

This dataset includes AI-related news articles from various global sources. Article volume peaked in 2023, reflecting increased interest after significant AI advancements. Key domains are rawpixel.com, einpresswire.com, and nasdaq.com, highlighting media, press releases, and financial news.
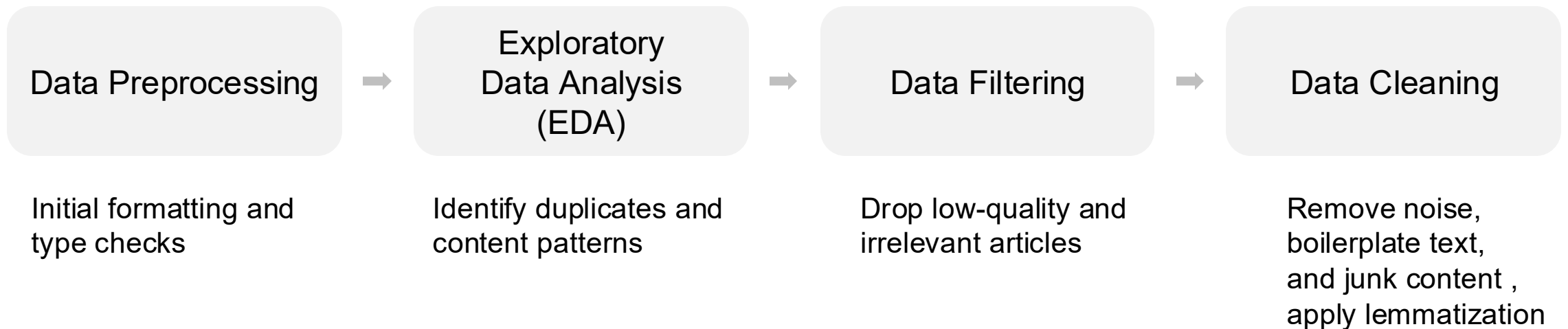
- Date Period: 2022/1/1 – 2025/4/28
- Data Columns: url, date, language, title, text
- Data Shape: 200,083
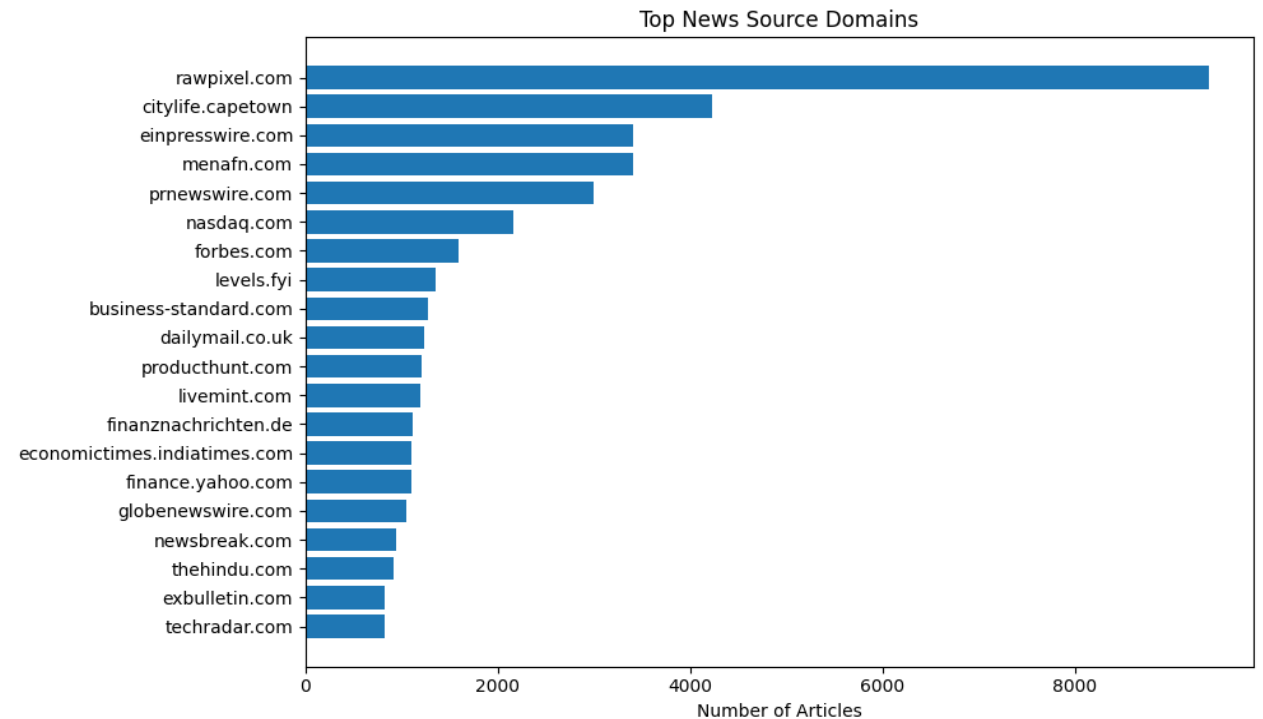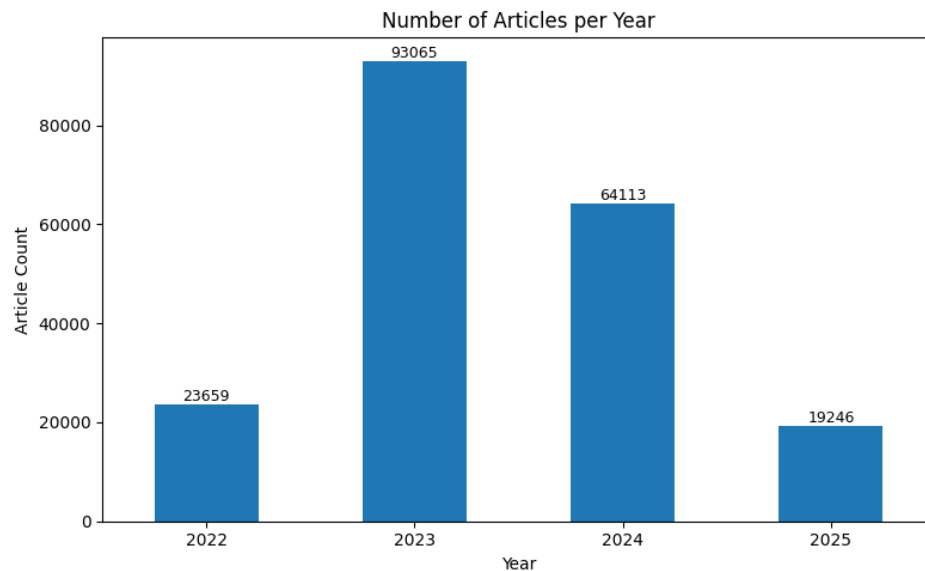
Article Clean-up and Filtering Workflow

Through exploratory data analysis, this project has identified duplicate articles that share identical titles and content, differing only by their URL suffixes. This process aims to eliminate low-quality sources, filter out irrelevant text, and apply lemmatization to standardize the data for further analysis.

Data Processing Workflow:

| Data Preprocessing	→	Exploratory Data Analysis (EDA)	→	Data Filtering	→	Data Cleaning
Initial formatting and type checks		Identify duplicates and content patterns		Drop low-quality and irrelevant articles		Remove noise, boilerplate text, and junk content , apply lemmatization

Data Processing Workflow - Article Filtering

The initial inspection showed that several top domains (e.g., rawpixel.com, citylife.capetown) contained irrelevant or low-quality content, including AI-generated images and adult material. These sources were filtered out to focus the dataset on meaningful AI-related news, and duplicate articles with identical titles and content will also be removed. There are 185,916 articles remaining.
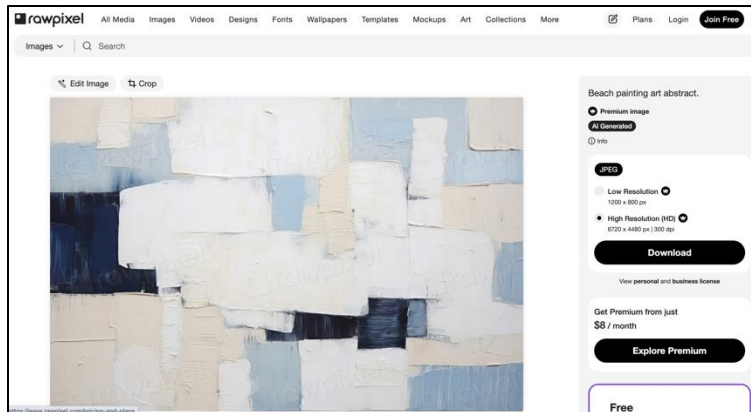
rawpixel.com
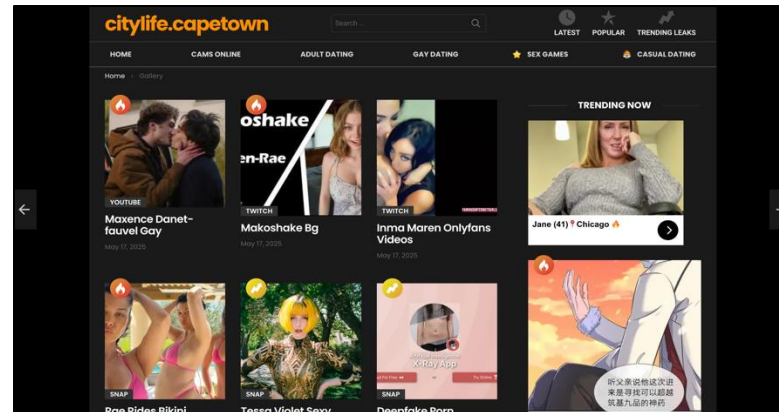
citylife.capetown
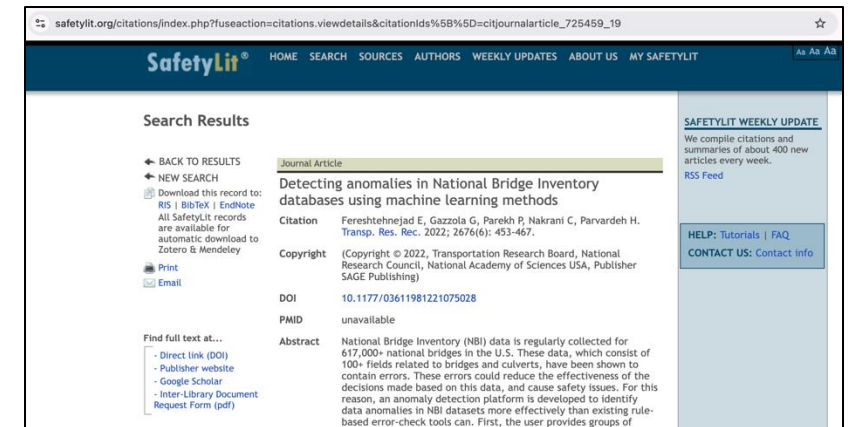
Same articles with different URL suffixes example

Data Processing Workflow – Text Cleaning Pipeline

The text cleaning pipeline consists of two steps: (1) removing noise such as boilerplate, dates, HTML tags, and junk words; and (2) preparing for topic modeling through lowercasing, lemmatization, and stopword removal to produce clean, meaningful tokens.

Step 1: Remove Noise and Artifacts

- Decode HTML entities (e.g., & → &)
- Clean escape characters (\', \", \\uXXXX, \\xXX, \\n, etc.)
- Remove date patterns
    - e.g., [May 20, 2023], March 10, 2023
- Remove common boilerplate and generic phrases
    - e.g., Read more, Copyright, Subscribe
- Strip HTML tags and URLs
- Remove non-ASCII characters
- Remove custom junk words
    - e.g., cnn, bbc, click, watch, video, etc.
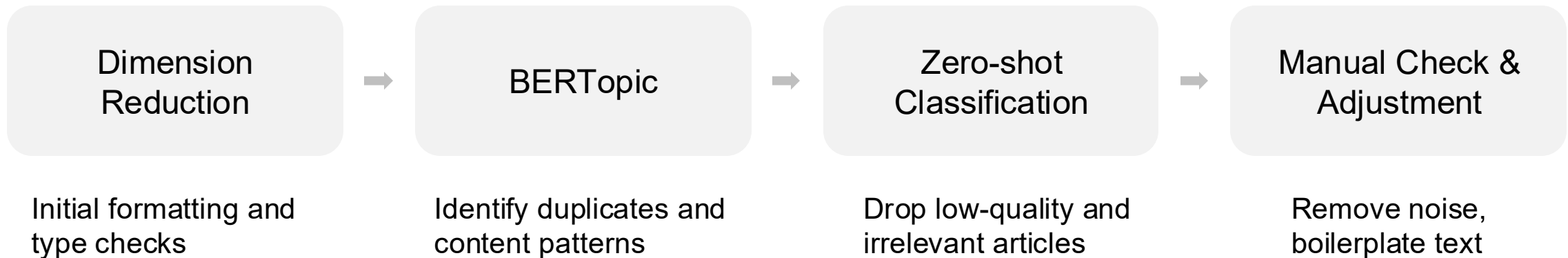- Normalize whitespace

Step 2: Prepare for Topic Modeling

- Convert text to lowercase
- Remove punctuation and digits
- Remove single-character tokens
- Apply spacy lemmatization
    - Keep only alphabetic and meaningful word roots
- Remove stopwords
- Return a clean, space-separated string of tokens

Topic Detection

To identify key themes in the AI-related news corpus, this project used a BERTopic pipeline. First, the process reduced dimensionality for high-dimensional embeddings, then performed clustering and topic extraction. The final stage organized the topics into predefined industry categories using zero-shot classification and refined the labels to ensure clarity and relevance to real-world contexts.

Topic Detection Workflow:

Dimension Reduction	→	BERTopic	→	Zero-shot Classification	→	Manual Check & Adjustment
Initial formatting and type checks		Identify duplicates and content patterns		Drop low-quality and irrelevant articles		Remove noise, boilerplate text

Topic Detection - Rationale for Component Settings

1. Embedding Model

This project used **all-MiniLM-L6-v2** for embeddings due to its balance of speed and accuracy. It provides high-quality results for large corpora and is for robust semantic understanding.

2. Dimensionality Reduction

UMAP reduces to 5 dimensions for efficient clustering while preserving semantic structure. With n_neighbors set to 30 for balanced context and min_dist at 0.0 for compact clusters, the approach enhances topic coherence and interpretability.

3. Clustering Model

HDBSCAN identifies dense clusters in the reduced space, with a min_cluster_size of 50 ensuring each topic includes enough articles to be statistically meaningful. This approach minimizes noise and prevents non-informative topics.

4. Text Vectorization

CountVectorizer is used to compute word frequencies while removing common English stop words. This allows the model to focus on domain-relevant content terms, which directly improves topic keyword extraction and clarity.

5. Topic Word Weighting

TF-IDF enhances topic representations by emphasizing words that are uniquely important to a specific topic compared to others. This produces cleaner, more discriminative keyword sets that align with industry-relevant themes.

6. Topic Labeling

To further improve topic interpretability, this project employ KeyBERT-inspired scoring, which leverages embedding similarity between words and topic centroids. This results in more semantically meaningful and readable topic labels.

Topic Detection - Zero-Shot Classification for Industry

This project generated over 500 topics, using zero-shot classification (Model: facebook/bart-large-mnli) to categorize them into 13 industries based on keywords. This model was chosen for its strong performance in natural language inference and its ability to assign semantic labels without task-specific training.

Workflow Steps

1. Input Data

- The Representation column contains keyword lists for each topic.

2. Preprocessing

- Convert stringified keyword lists to actual lists using ast.literal_eval.

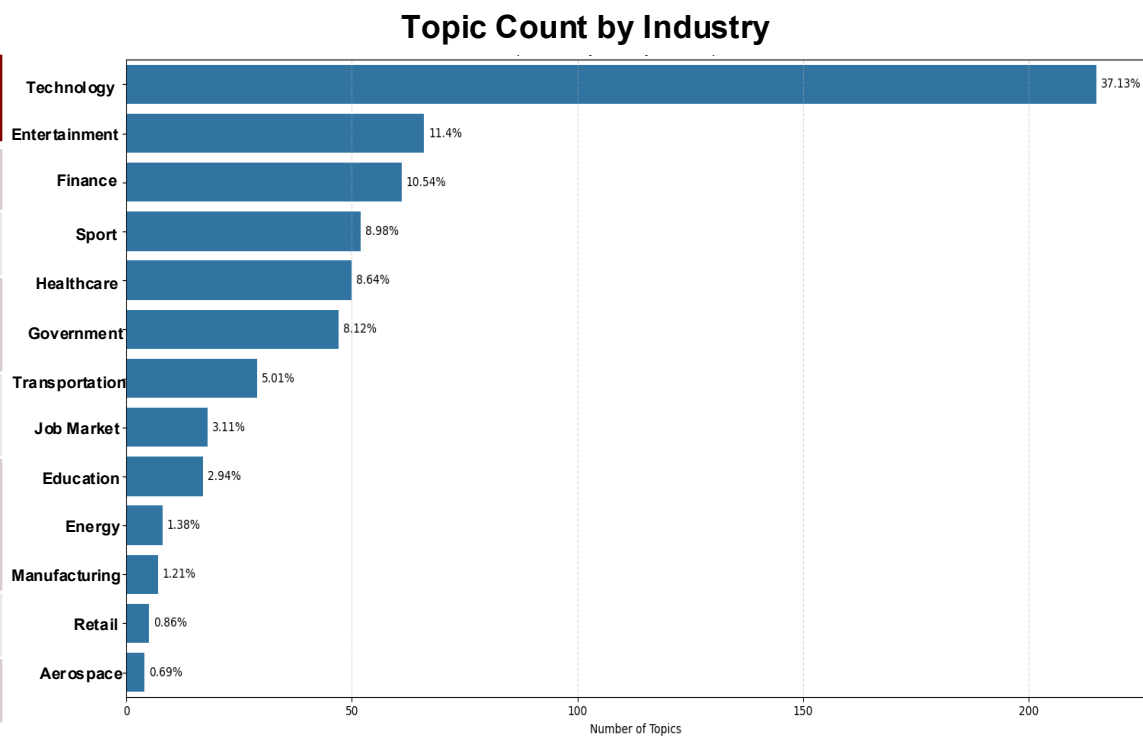- Join keywords into a readable sentence as the input prompt.

3. Zero-shot Classification

- Use the following industry labels:

Technology, Finance, Healthcare, Education, Retail, Transportation, Aerospace, Government, Job Market, Energy, Entertainment, Sport, Manufacturing

Topic Detection - Result

This project used BERTopic to identify 578 topics from AI-related news articles. To manage the high dimensionality, we applied zero-shot classification, which consolidated these topics into 13 industries. The most common categories were Technology, Entertainment, Finance, and Sports, providing a structured framework for future sentiment and trend analysis.

Topic	Name	Predicted Industry
-1	-1_tech_industry_technology_feature	Technology
0	0_industry_testimonial_idaho illinois_nevada	Manufacturing
1	1_transportation_auto transportation_heavy industry_service invest	Transportation
2	2_market global_qatar kuwait_industry region_bahrain qatar	Finance
3	3_sportspromotionsnd todayag_contentweatherwatch livevideokfyr_livevideokfyr sportspromotionsnd_contentweatherwatch	Sport
4	4_elon musk_musk xai_musk openai_musk	Technology
5	5_openai chatgpt_use chatgpt_chatgpt plus_chatgpt	Technology

Topic Count by Industry

Number of Topics

- Technology — 37.13%
- Entertainment — 11.4%
- Finance — 10.54%
- Sport — 8.98%
- Healthcare — 8.64%
- Government — 8.12%
- Transportation — 5.01%
- Job Market — 3.11%
- Education — 2.94%
- Energy — 1.38%
- Manufacturing — 1.21%
- Retail — 0.86%
- Aerospace — 0.69%

Industry Impact Analysis

In 2023, generative AI coverage peaked due to significant advancements in technology, healthcare, and finance. While the transportation sector saw a brief spike in interest, it quickly declined. In contrast, manufacturing, retail, and the job market showed steady growth, reflecting increasing interest in AI.

Number of Articles per Industry per Year
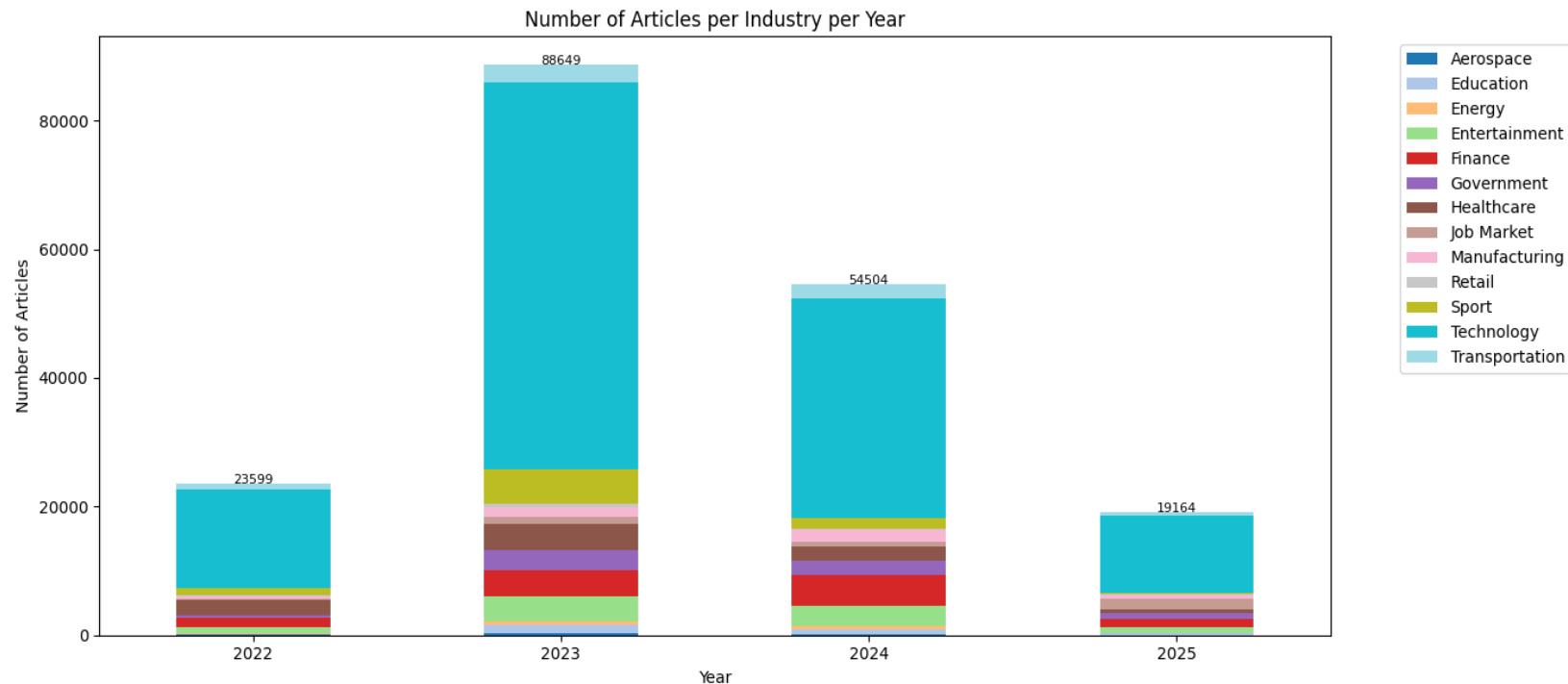
Entity extraction (people, organizations, technologies)

This project used Spacy for entity extraction. Named entities were normalized and grouped into people, organizations, and technologies to identify major actors and AI-related tools shaping the landscape. Below are top 10 results:

	People Name
1	Elon Musk
2	Sam Altman
3	Donald Trump
4	Joe Biden
5	Taylor Swift
6	Narendra Modi
7	Kamala Harris
8	Mark Zuckerberg
9	Satya Nadella
10	Rishi Suna

	Organizations Name
1	OpenAI
2	Google
3	Microsoft
4	Meta
5	Apple
6	NVIDIA
7	Amazon
8	LG
9	Reuters
10	FCC

	Technologies Name
1	ChatGPT
2	Bard
3	Generative AI
4	LLaMA
5	Ultragear
6	Meta AI
7	Galaxy AI
8	Gmail
9	Vision Pro
10	Metaverse

Entity extraction (people)

	People Name
1	Elon Musk
2	Sam Altman
3	Donald Trump
4	Joe Biden
5	Taylor Swift
6	Narendra Modi
7	Kamala Harris
8	Mark Zuckerberg
9	Satya Nadella
10	Rishi Sunak

Key Insight & Interpretation:

• **Elon Musk, Sam Altman, Mark Zuckerberg, Satya Nadella**: These tech CEOs and AI pioneers are mentioned a lot due to their roles in developing and deploying AI.

• **Donald Trump, Joe Biden, Narendra Modi, Kamala Harris, Rishi Sunak**: The dataset period features major elections like the 2022 UK elections and the 2024 US presidential election. Political figures are discussed regarding national strategies and regulation, reflecting heightened public discourse as elections approach.

• **Taylor Swift**: Based on URL lookback, it seems this artist will appear in many articles discussing AI in the music field, although some appear to be advertisements.

The integration of AI involves both tech leaders and politicians, showing it is a societal concern that intersects with governance and global competition. The prominence of these individuals highlights their influence in shaping direction of AI public and policy discussions.

Entity extraction (organizations)

	Organization Name
1	OpenAI
2	Google
3	Microsoft
4	Meta
5	Apple
6	NVIDIA
7	Amazon
8	LG
9	Reuters
10	FCC

Key Insight & Interpretation:

• **OpenAI, Google, Microsoft, Meta, Apple, NVIDIA, Amazon**: These companies are leading the development of AI models, cloud infrastructure, and AI-powered products. Their frequent mentions reflect advancements, product announcements, and strategic partnerships.

• **LG**: As a major electronics and hardware manufacturer, LG is investing in AI for smart devices, vision systems, and home automation.

• **Reuters**: A major global news agency, this company is commonly cited in AI-related news dissemination and media coverage.

• **FCC**: The U.S. Federal Communications Commission appears in discussions on regulation, AI in communications, and data privacy, especially in light of growing concerns about algorithmic transparency and AI governance.

Entity extraction (technologies)

	Technology Name
1	ChatGPT
2	Bard
3	Generative AI
4	LLaMA
5	Ultragear
6	Meta AI
7	Galaxy AI
8	Gmail
9	Vision Pro
10	Metaverse

Key Insight & Interpretation:

• **ChatGPT / Bard / LLaMA / Meta AI / Generative AI**

These AI models have dominated the public discourse since 2022. Their frequent mentions reflect advancements in natural language processing, competition among tech giants, and widespread integration across applications.

• **Ultragear / Galaxy AI / Vision Pro**

These are AI-enhanced consumer technologies tied to product lines from LG (Ultragear), Samsung (Galaxy AI), and Apple (Vision Pro). Their visibility stems from media coverage of device launches and AI-driven feature upgrades.

• **Metaverse**

The Metaverse trend still receives attention within the data time period due to its overlap with AI, virtual avatars, immersive experiences, and long-term visions from Meta and other firms.

Sentiment Analysis

To analyze industry sentiment toward AI integration, a hybrid pipeline was developed. First, 500 articles were labeled using GPT-4 and lexicon-based heuristics. A BERT model was fine-tuned using this combined dataset and deployed to infer sentiment across articles.

Sentiment Analysis Workflow:

Labeled Data	→	Sentiment Model Training	→	Evaluate Performance	→	Prediction on Full Dataset
GPT-4 labeling & lexicon-based labeling		Pre-trained model bert-base-uncased		sentiment-tagged using trained model		Confusion Matrix Accuracy F-1 score
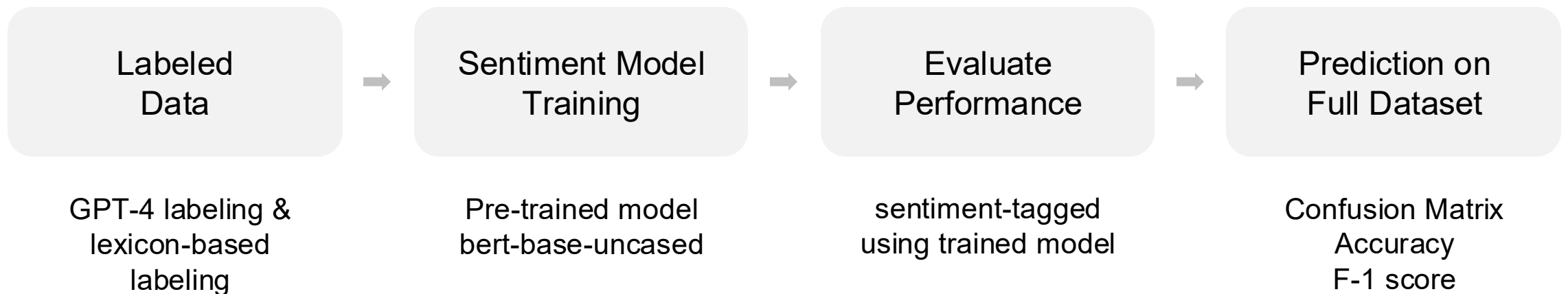
Sentiment Analysis – Performance Evaluation

A BERT-based model was fine-tuned on GPT-labeled news articles to classify sentiment as positive, neutral, or negative, achieving 71% accuracy. It effectively identified positive sentiment but often misclassified neutral sentiment, indicating some ambiguity in the news tone. The model was subsequently used for timeline and industry-level analysis on the full dataset.

Confusion Matrix - BERT Sentiment
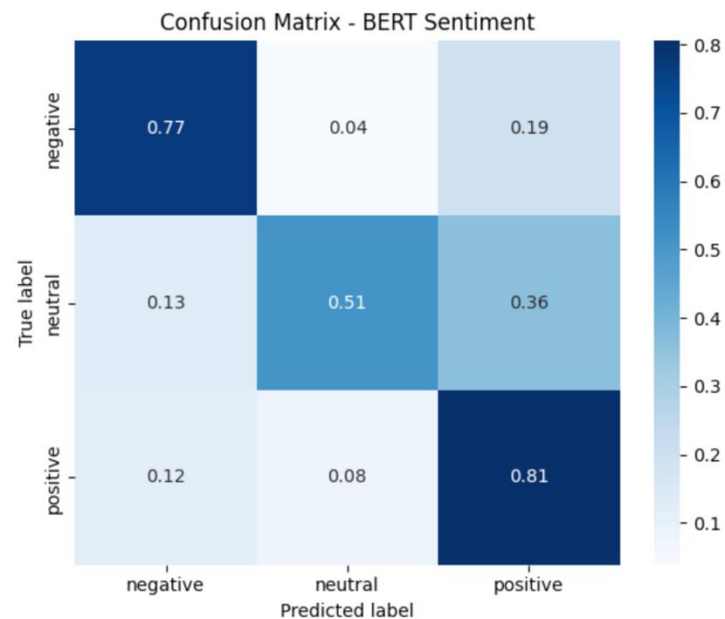
	precision	recall	F1-score	support
negative	0.57	0.77	0.66	26
neutral	0.77	0.51	0.62	47
positive	0.74	0.81	0.77	77
accuracy			0.71	150
Macro avg	0.69	0.70	0.68	150
Weighted avg	0.72	0.71	0.70	150

Sentiment Analysis over time

Between 2022 and 2025, over 60% of AI-related news articles expressed positive sentiment, peaking in mid-2023 with the rise of tools like ChatGPT. Despite a simultaneous peak in negative sentiment, it later stabilized, suggesting that while public concerns remained, confidence in AI continued to grow.
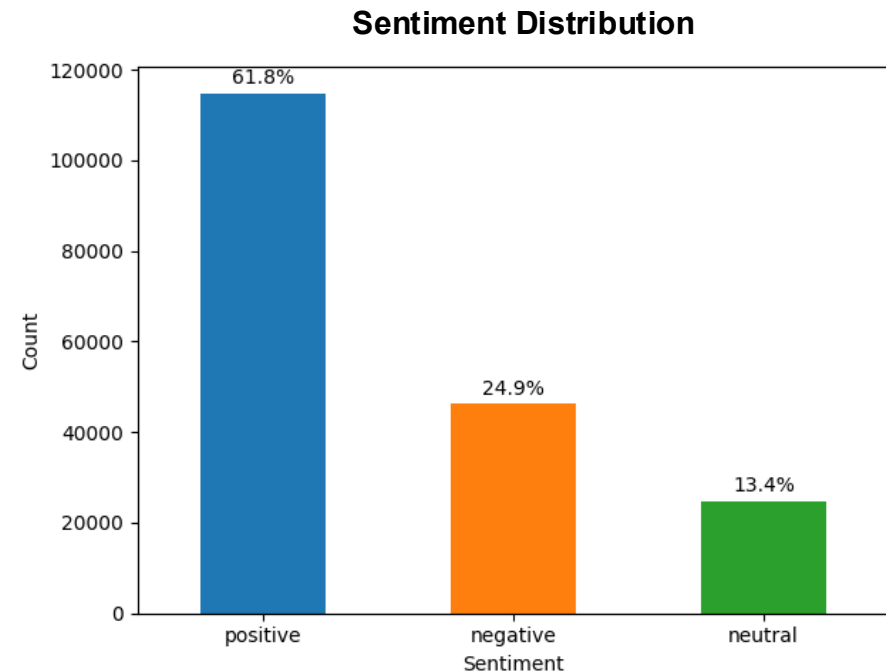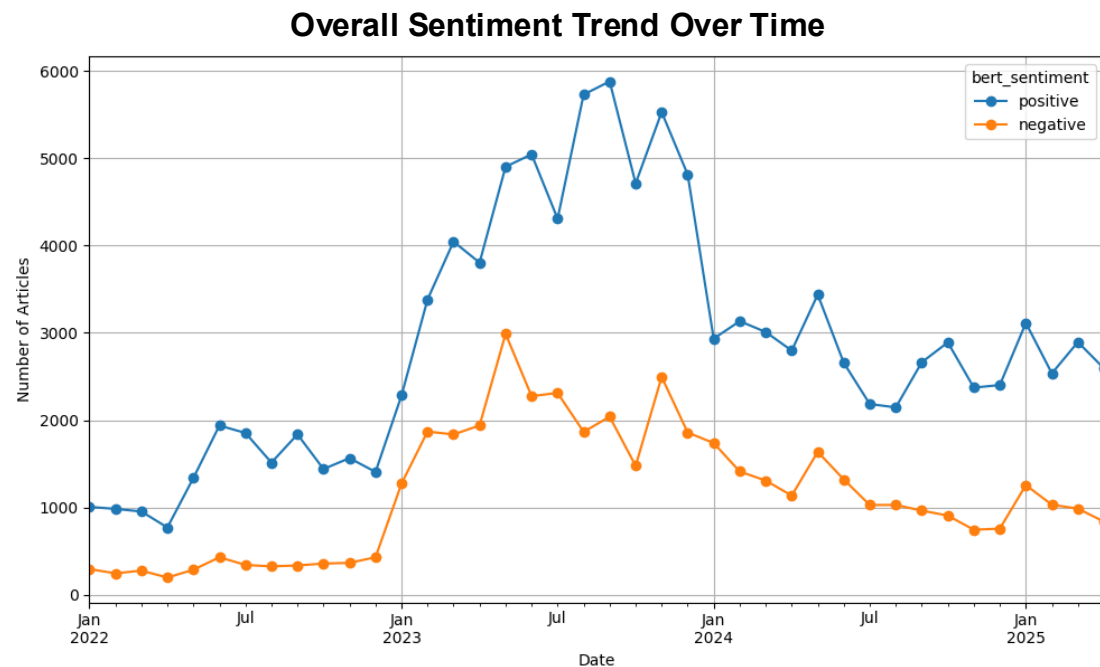
Topic-level sentiment analysis

bert_sentiment	negative	neutral	positive	net_positive
Predicted_Industry				
Manufacturing	0.693304	0.101786	0.204911	-0.488393
Entertainment	0.587008	0.111424	0.301568	-0.285440
Sport	0.542496	0.078202	0.379302	-0.163194
Education	0.491831	0.098031	0.410138	-0.081693
Job Market	0.322824	0.380859	0.296317	-0.026507
Government	0.350896	0.267729	0.381375	0.030479
Transportation	0.370947	0.016372	0.612681	0.241734
Healthcare	0.255093	0.170384	0.574523	0.319430
Energy	0.199536	0.246713	0.553751	0.354215
Retail	0.233615	0.098309	0.668076	0.434461
Aerospace	0.135762	0.238411	0.625828	0.490066
Technology	0.188645	0.108459	0.702896	0.514251
Finance	0.042954	0.342763	0.614283	0.571329

Negative Sentiment

Negative sentiment often stems from fears of job loss, automation, and diminished human oversight. Industries like manufacturing, entertainment, and sports show more skepticism toward AI.

Neutral Sentiment

Neutral sentiment reflects factual reporting or balanced views. Industries like education and job market with mixed reactions to AI acknowledge both its benefits and concerns, contributing to this middle ground.

Positive Sentiment

Positive sentiment thrives in sectors like finance, technology, and aerospace, where AI enhances productivity, safety, and innovation.

Topic-level sentiment analysis

AI Sentiment Trend - Technology

AI Sentiment Trend - Finance

AI Sentiment Trend – Job Market
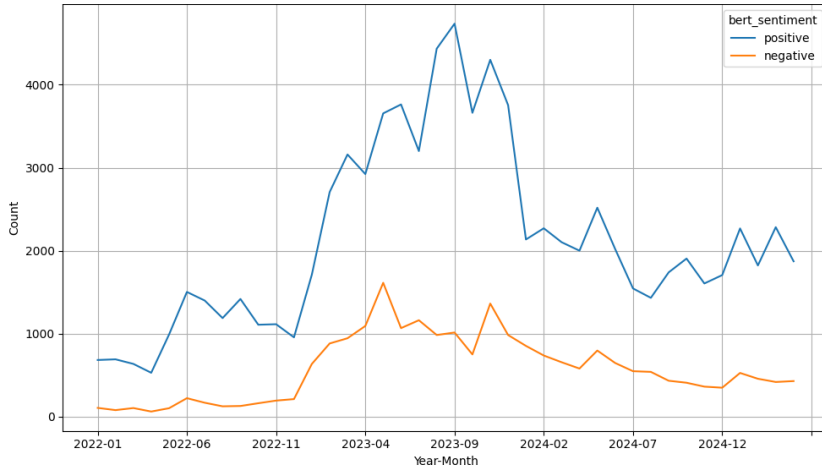
Technology:

Positive sentiment peaked in mid-2023 with the widespread adoption of generative AI tools like ChatGPT. It slightly decreased afterward but stayed consistently higher than negative sentiment.
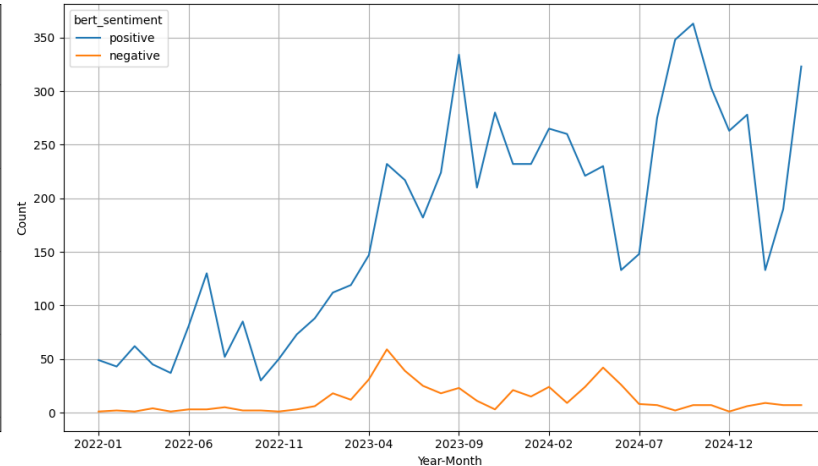
Finance:

Since early 2023, positive sentiment has gradually increased, highlighting growing interest in AI-powered trading, robo-advisors, and fraud detection, while negative sentiment has stayed low and stable.

Job Market:

Public sentiment varies, reflecting concerns about AI's impact on employment. A rise in negativity was noted in late 2024 due to layoff announcements or regulatory discussions.

Topic-level sentiment analysis

AI Sentiment Trend - Manufacturing

AI Sentiment Trend - Transportation

AI Sentiment Trend - Government
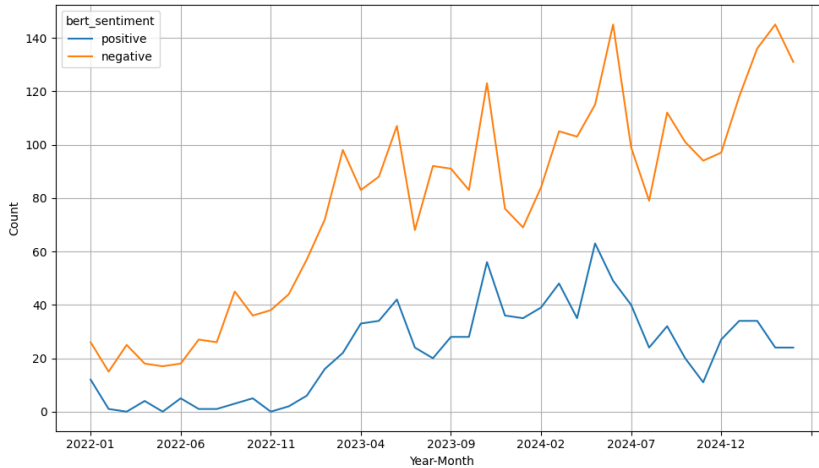
Manufacturing:

Sentiment remained predominantly negative throughout the timeline, reflecting concerns about automation, job displacement, or industrial disruption caused by AI.
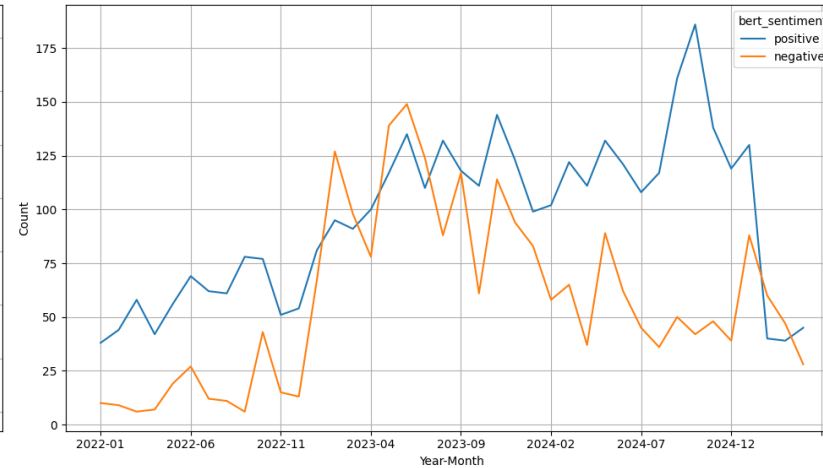
Transportation:

Sentiment was primarily positive during 2023–2024, reflecting optimism about AI-driven innovations such as autonomous vehicles and logistics optimization.
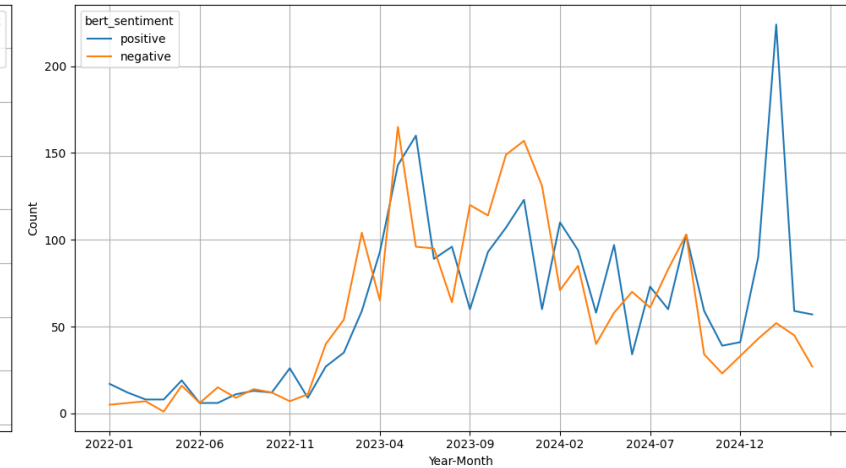
Government:

Sentiment displayed more volatility with several spikes in both directions. This indicates mixed discourse surrounding AI regulation, surveillance, and public trust issues.

Entity-level sentiment analysis

Elon Musk and Sam Altman received predominantly positive sentiment, while political figures like Donald Trump and Joe Biden had mixed reactions due to ongoing regulatory debates.

OpenAI and Google also had positive sentiment, whereas platforms like Twitter and the FCC faced criticism because they are platforms with discussions and news. Geographically, the United States received a lot of negative sentiment, followed by India and the Middle East, while most regions exhibited overall positive sentiment.

Entity-level sentiment analysis - jobs

Roles like financial analysts, project managers, and writers are often discussed in the context of AI. Jobs requiring higher cognitive skills or communication tend to have more negative sentiment due to concerns about automation and job displacement. In contrast, data scientists and engineers have a positive outlook as they lead AI development. However, project managers are experiencing negative sentiment due to the pressures of adapting to AI-driven workflows.

Top Jobs by Sentiment
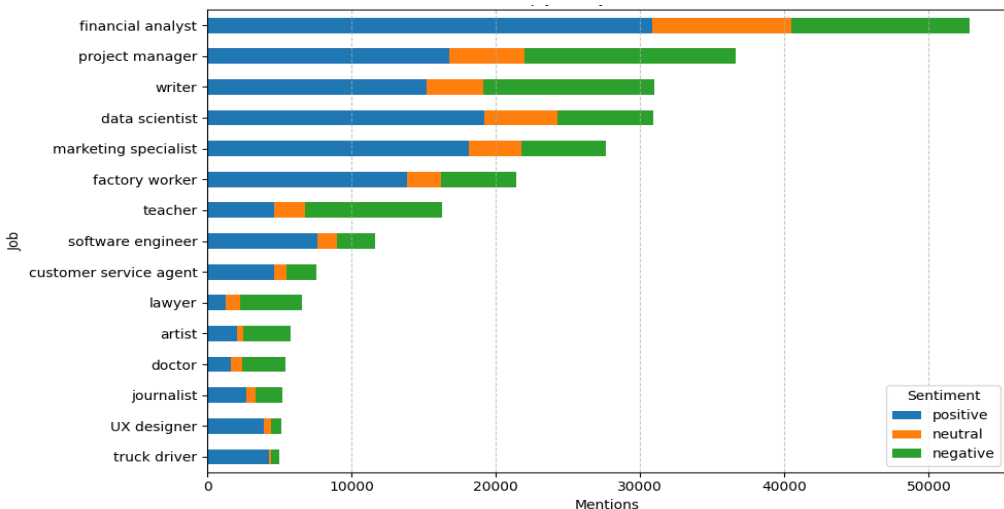
Sentiment Trend: Financial Analyst, Project Manager

Sentiment Trend Over Time - Financial Analyst
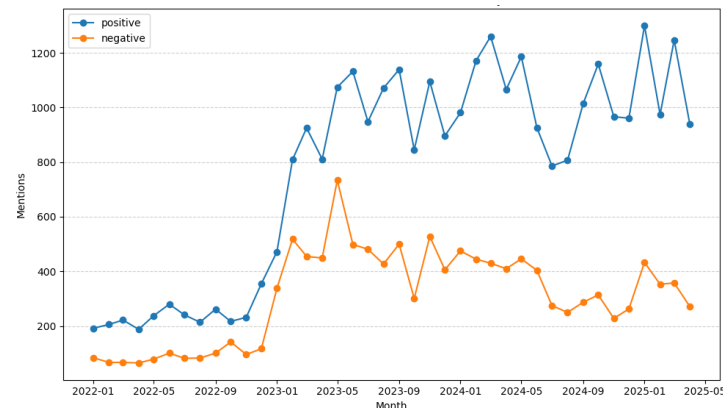
Sentiment Trend Over Time - Project Manager
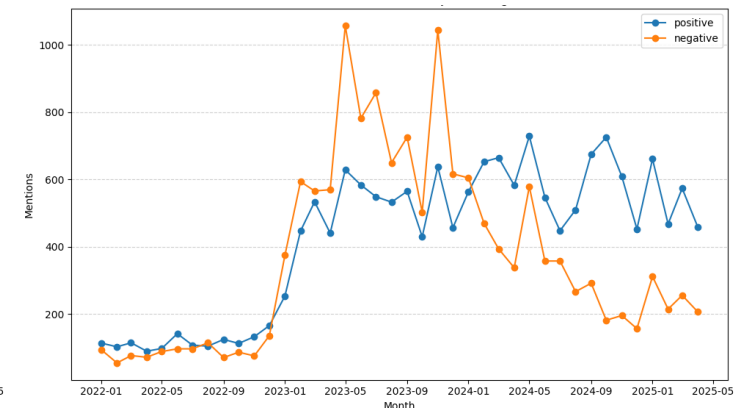
Entity-level sentiment analysis - people

In late 2023, Sam Altman saw a notable rise in mentions, likely due to leadership changes at OpenAI and global AI policy debates, generating both strong support and criticism.

Donald Trump also draws attention, especially regarding AI regulation, surveillance, and employment concerns, reflected in a higher level of negative sentiment.
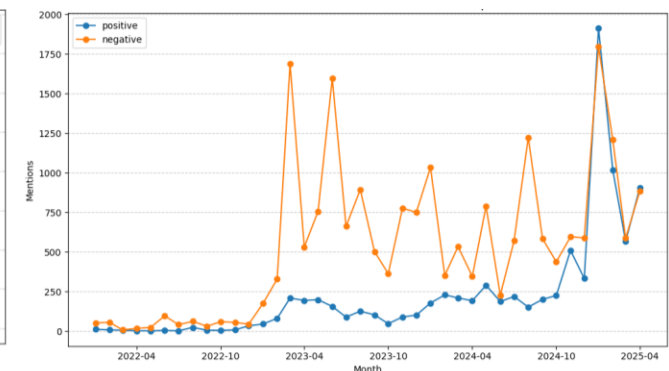

Top People Mentions Over Time
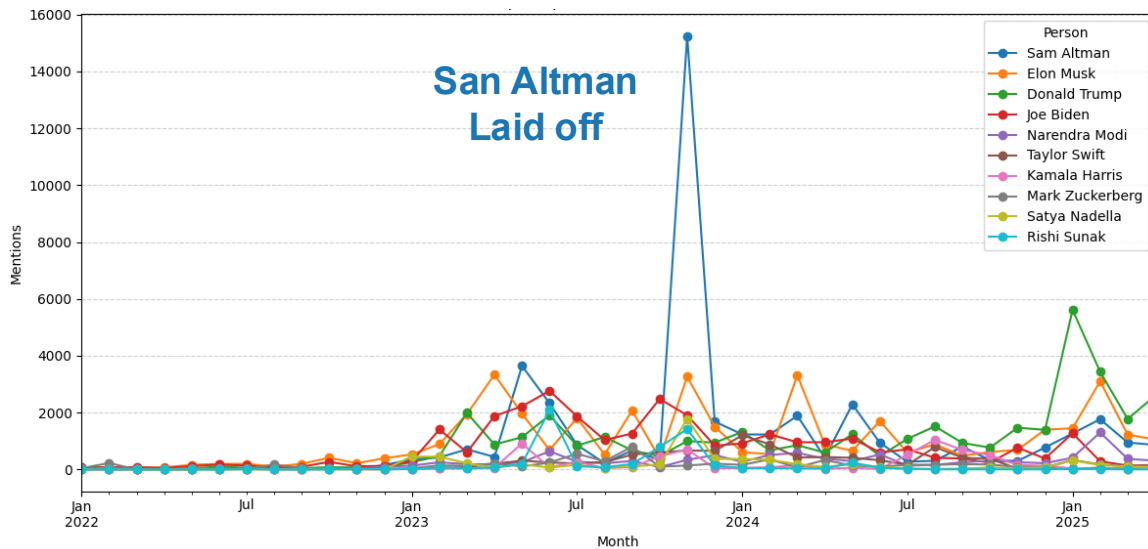
Sentiment Trend: San Altman, Donald Trump

Sentiment Trend Over Time - San Altman   Sentiment Trend Over Time - Donald Trump
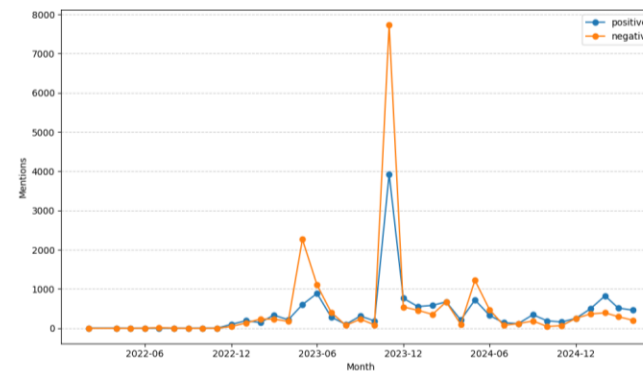
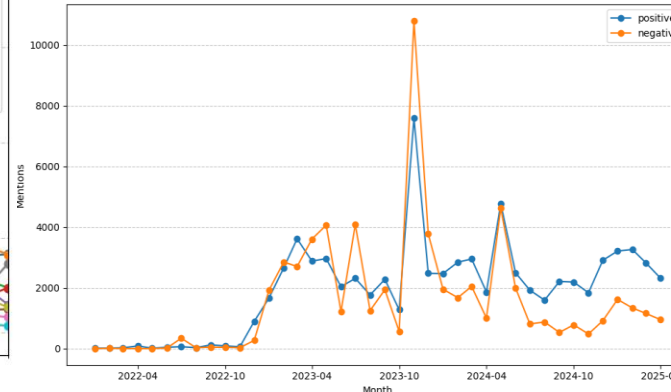# Entity-level sentiment analysis - organizations

Mentions of AI organizations saw a significant increase following major product launches in late 2022 and early 2023. OpenAI experienced the largest surge in positive sentiment due to ChatGPT. In contrast, Google and Microsoft received mixed reactions. These trends highlight the importance of effective communication and innovation in shaping the future of AI.
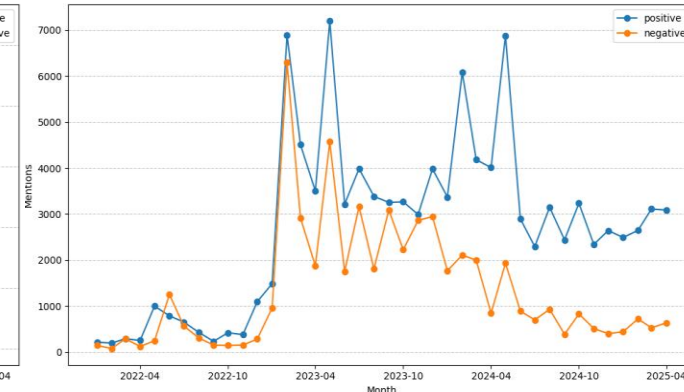

Top Organizations Mentions Over Time
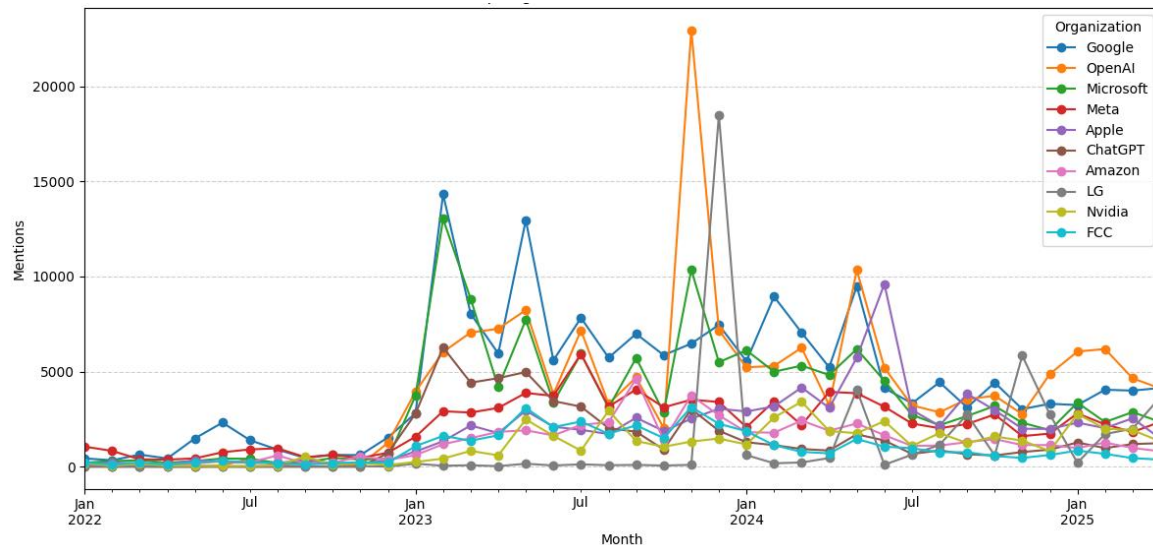
Sentiment Trend: OpenAI, Google

Sentiment Trend Over Time - OpenAI

Sentiment Trend Over Time - Google

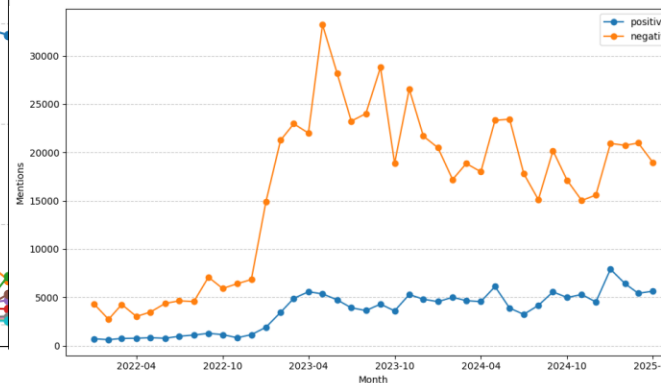# Entity-level sentiment analysis - locations

The United States leads discussions on artificial intelligence, emphasizing its crucial role in AI development and policymaking. Since 2023, negative sentiment has risen, likely due to concerns about AI's impact on jobs, the need for regulation, and geopolitical issues. In early 2025, China saw an increase in mentions, accompanied by a more balanced sentiment trend. Deepseek appears to have gained positive attention. This rise may be linked to global competition for AI leadership and concerns regarding surveillance technology.
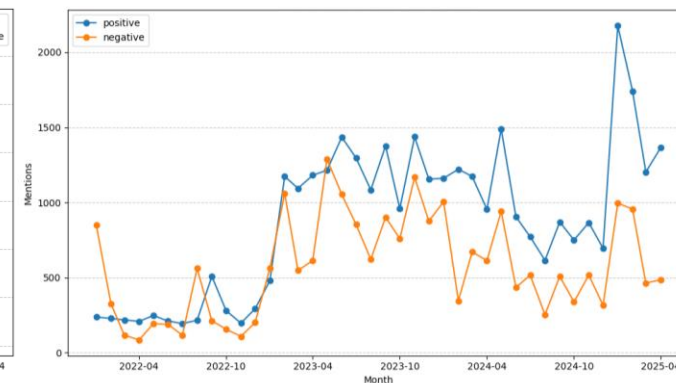

Top Locations Mentions Over Time
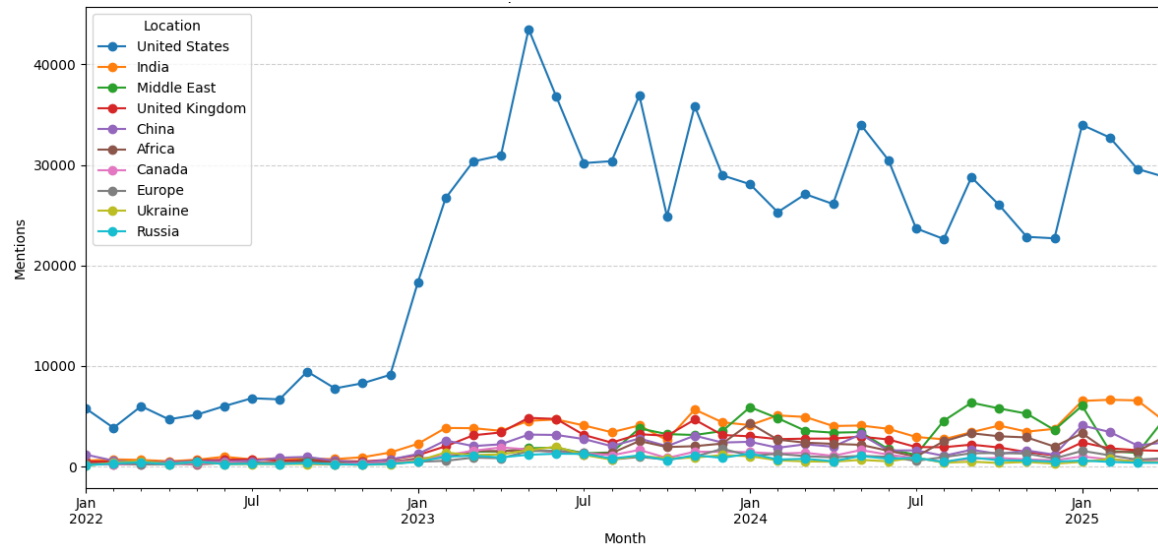
Sentiment Trend: US, China
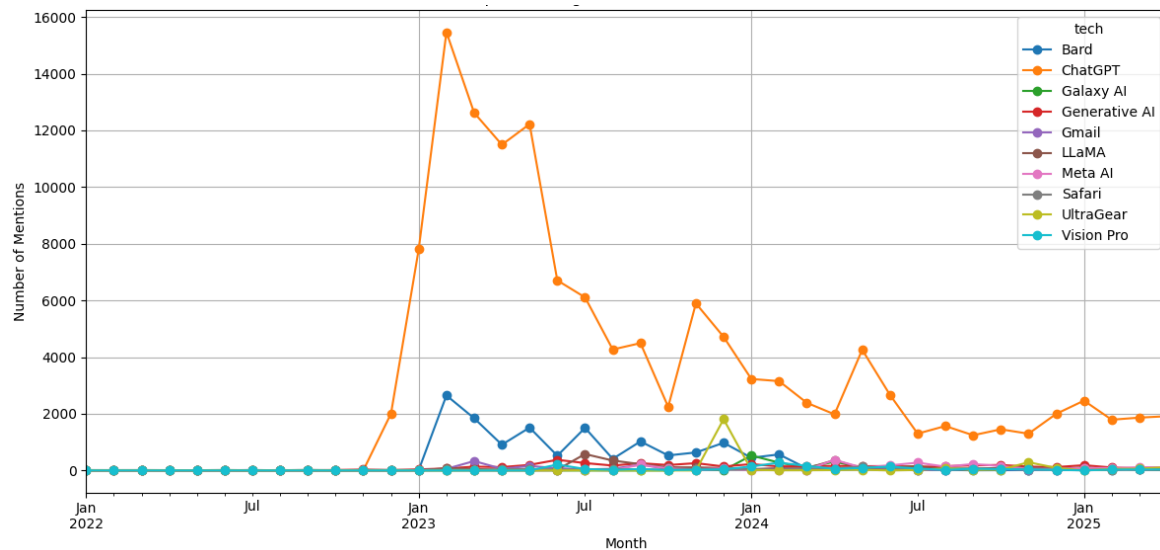
Sentiment Trend Over Time - US

Sentiment Trend Over Time - China

Entity-level sentiment analysis - technology

Technologies like ChatGPT, Bard, and Generative AI saw significant interest in early 2023, generating both excitement and concern about their rapid deployment. LLAMA from Meta maintained positive sentiment, reflecting support for open-source AI research. Apple's Vision Pro gained attention around its launch, indicating rising interest in spatial computing and AI interfaces. These trends point to innovations driving workforce transformation and guide institutional investment and policy focus.
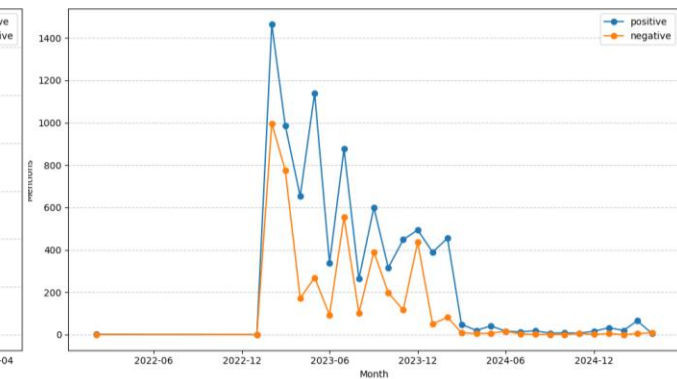

Top Technologies Mentions Over Time

Sentiment Trend: Generative AI, Bard

Sentiment Trend Over Time - Generative AI
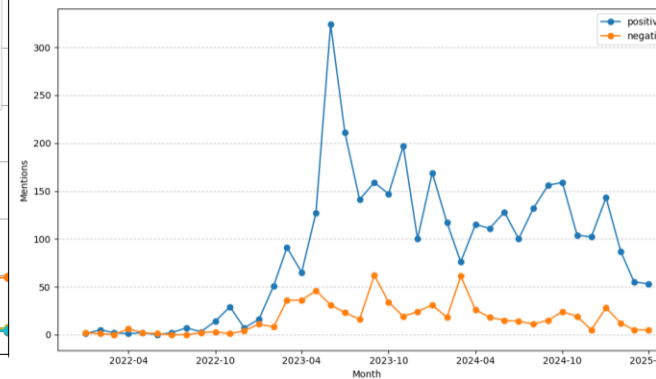
Sentiment Trend Over Time - Bard

Conclusion

Industries most impacted by AI (2022–2025):

• **Technology, Finance, Retail, Healthcare, and Transportation** show high AI-related activity.

• **Job roles** like project manager, writer, and customer service agent face increasing automation risk based on negative sentiment signals.

Why are certain jobs more exposed:

• Jobs with **structured, repeatable tasks** (e.g., lawyers) or **text generation** are easier to automate.

• Occupations needing **human intuition, creativity, or empathy** (e.g., UX designers) show lower disruption risk.

Who is investing & why:

• Companies like **OpenAI, Google, Microsoft, Meta, NVIDIA** are driving AI deployment through LLMs, cloud platforms, and productivity tools.

• Their innovations (e.g., **ChatGPT, Bard**) fuel adoption across sectors, making them key ecosystem builders.

Applications resistant to current AI:

• **High-stakes decision-making**, **emotional labor**, and **cross-context reasoning** (e.g., healthcare diagnostics, diplomacy, early education) remain difficult to fully automate.

Actionable Recommendations

For Companies:

• **Adopt AI-Human Collaboration Models**: Utilize AI to enhance, rather than replace, human roles in areas such as finance, marketing, and project management.

• **Reskill Workforce**: Invest in continuous training for employees in data literacy, AI operations, and prompt engineering.

• **Monitor Emerging Tools**: Track developments like ChatGPT, Bard, and LLaMA to remain competitive.

For Academic Institutions:

• **Update Curriculum**: Integrate AI ethics, explainable AI, and application-specific AI modules into programs.

• **Promote Interdisciplinary Research**: Encourage collaboration across CS, business, health, and law to explore AI impacts.

• **Support Open Research**: Lead in transparent model development and AI policy experimentation.

For Governments & Policymakers:

• **Establish AI Governance Frameworks**: Define standards for safe and ethical deployment of AI in high-risk sectors.

• **Fund Public AI Infrastructure**: Build accessible resources for startups, educators, and researchers.

• **Encourage Regional AI Innovation Hubs**: Support public-private partnerships focused on AI use in job creation, education, and healthcare.