# From Words to Networks: Media-driven Insights into the G60 Sci-Tech Corridor

Jingyao Zhong*
Yufei Wang*
{jingyaochung,0241136002}@shisu.edu.cn
Shanghai International Studies University
Songjiang District, Shanghai, China

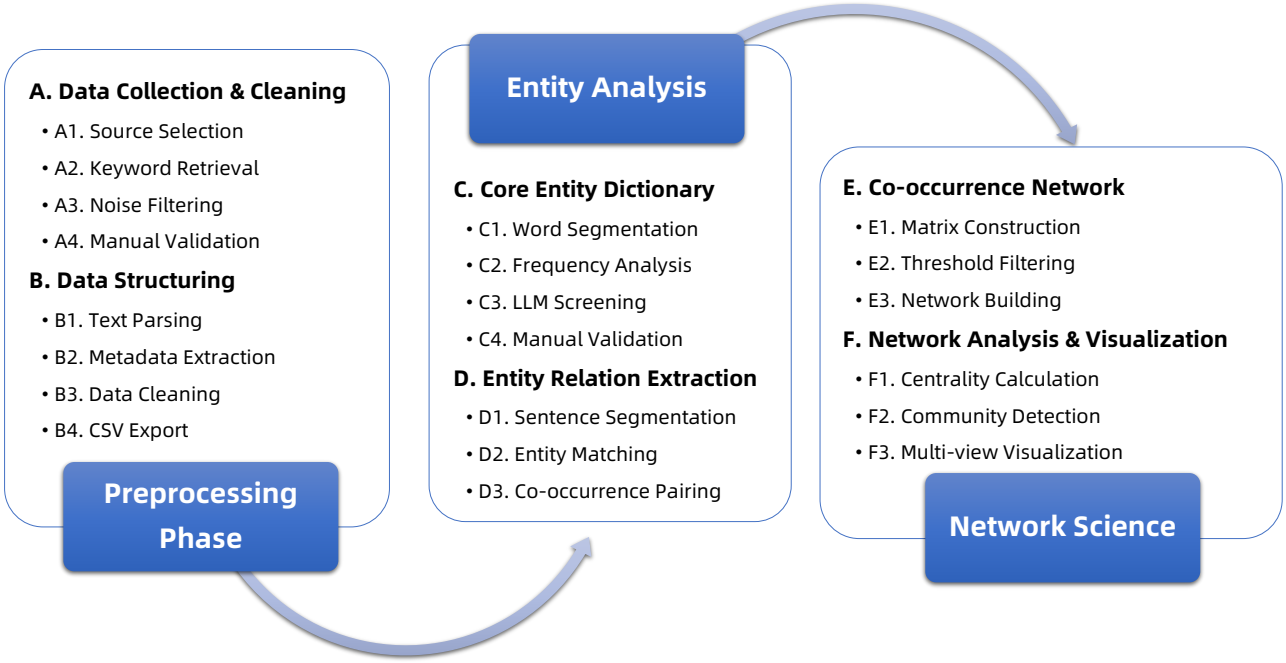


**Figure 1: Overview of the data-driven workflow for constructing and analyzing the G60 Sci-Tech Corridor innovation network from news media. The pipeline consists of multi-stage data collection, entity lexicon generation, co-occurrence network construction, and multi-dimensional visualization, enabling systematic exploration of regional innovation dynamics.**

## Abstract

This paper presents a comprehensive, data-driven framework for analyzing the innovation ecosystem of the Yangtze River Delta G60 Science and Technology Innovation Corridor through large-scale news media text mining and multidimensional visualization. We construct a longitudinal dataset of 3,786 news articles (2021–2025) from authoritative Chinese media, systematically filtered and structured for analysis. Leveraging a Large Language Model (LLM), we generate a core lexicon of 60 entities across industry, technology, and capital dimensions, and extract their co-occurrence relationships to build a weighted, undirected network. Advanced network analysis (including Louvain community detection and centrality metrics) and a suite of visualization techniques—such as force-directed graphs, Sankey diagrams, heatmaps, and geographic maps—reveal the dynamic interplay, temporal evolution, and spatial distribution of innovation actors. Our results highlight the centrality of technology-capital synergy, the dominance of core cities, and the emergence of new innovation clusters, providing actionable insights for regional policy and communication. The full dataset and code for data processing and visualization are available at: https://github.com/agenttomzhong/G60-News-Network-Visualization

## CCS Concepts

• **Human-centered computing** → **Visual analytics**; **Visualization design and evaluation methods**.

*Both authors contributed equally to this research.

## Keywords

data visualization, news mining, entity co-occurrence, Yangtze River Delta, G60 Sci-Tech Corridor, network visualization, data-driven analysis

# 1 Introduction

## 1.1 Background and Motivation

The G60 Science & Technology Innovation Corridor in the Yangtze River Delta has become a flagship region for national innovation and industrial upgrading, driven by coordinated policy support, infrastructure investment, and targeted enterprise deployment. Spanning nine cities and leveraging clusters in advanced manufacturing, biotechnology, and new energy, the Corridor aims to foster cross-regional collaboration and accelerate technology transfer. Despite a growing body of quantitative studies based on patents, firm financials, and regional economic indicators, the evolutionary trajectory of its innovation ecosystem—as embedded in thousands of news reports—remains underexplored. News media act as "information sensors" for regional innovation, not only chronicling policy iterations, technological breakthroughs, and industry alliances, but also reshaping public discourse by linking diverse actors and events into semantic networks. Traditional text-mining methods, such as term-frequency or topic models, struggle to capture the latent, multi-dimensional relationships and temporal dynamics present in large-scale corpora. To bridge this gap, a data-visualization–driven approach is therefore required—one that transforms unstructured text into interactive, quantitative network structures, enabling researchers and policymakers to trace the Corridor's innovation narrative, identify emerging clusters, and monitor shifts in media attention over time.

## 1.2 Main Contributions

Our work makes the following key contributions:

(1) **A longitudinal news dataset (2021–2025):**
We systematically collected twelve categories of sources, including policy releases and corporate announcements, resulting in a temporal corpus of 3,786 G60-related articles. This dataset provides a solid foundation for media-driven regional innovation studies.
(2) **Dictionary generation via LLMs:**
We leveraged the Qwen3-Plus model to perform frequency analysis on high-occurrence terms, and refined the results through manual validation to accurately extract 60 entities across three dimensions: industry, technology, and capital.
(3) **Multi-dimensional visualization scheme:**
   a. **Network topology analysis:** Applied the Louvain algorithm to detect thematic communities of innovation.
   b. **Temporal evolution views:** Employed Sankey diagrams, sunburst charts, and line-matrix plots to trace the co-occurrence evolution of entities over time.
   c. **Multi-scale association mapping:** Integrated nine chart types — including heatmaps, scatter plots, and chord diagrams — to enable interactive exploration from micro-level entity pairs to macro-level industrial patterns.

# 2 Related Work

## 2.1 Studies on the G60 Corridor

Existing studies on the G60 Science & Technology Innovation Corridor primarily address policy effects, regional restructuring, and technological ecology. Sun and Chen [5] employ a difference-in-differences (DID) model on firm financial and patent-transfer data to demonstrate that the G60 policy significantly boosts firms' innovation inputs and outputs by facilitating factor mobility, alleviating financing constraints, and enhancing market competition. Li et al. [2] propose a dual "power–factor" restructuring framework from the regional innovation system perspective, highlighting how cross-jurisdictional institutional innovation and the four-dimensional integration of technology, industry, finance, and talent drive high–quality integration. Wang and Liu [6] apply technology niche theory and find that both niche breadth expansion and intercity niche overlap significantly improve innovation efficiency, while exhibiting negative spatial spillovers from core to peripheral cities. However, these works rely on structured data and traditional econometric methods, overlooking the rich, unstructured information embedded in news media.

## 2.2 News Text Mining and Entity Network Analysis

News media serve as "information sensors" in regional innovation ecosystems, recording dynamic events and influencing public perception through entity co-occurrence and discourse networks [1]. Early approaches leveraged term-frequency metrics (e.g., TF–IDF) or shallow learning algorithms to extract keywords and construct co-occurrence networks for topic evolution analysis [3]. More recently, transformer-based models such as BERT improved extraction accuracy but lacked interpretability. Pakhale [4] introduces a dynamic dictionary-generation framework using large models that iteratively refines high-frequency lexicons, balancing flexibility and precision. Yet, current research often targets single domains (e.g., financial news) and fails to integrate news-based network analysis with the multi-dimensional mechanisms of innovation corridors.

## 2.3 Gaps and Positioning

Despite significant advances in policy evaluation, institutional innovation, and technological ecology for the G60 Corridor, three limitations remain:

(1) **Limited data sources:** Heavy reliance on structured data (e.g., patents, fiscal statistics) neglects the latent associations in unstructured news texts.
(2) **Methodological constraints:** Traditional term-frequency and deep-learning approaches each have shortcomings, necessitating novel methods that combine large-model–driven dynamic dictionaries with network analysis.

(3) **Visualization scope:** Most studies employ single-chart representations and lack a systematic framework integrating network topology, temporal evolution, and multi-scale mapping.

To address these gaps, this paper constructs a four-year longitudinal news dataset, employs large models for dynamic dictionary generation, and develops a data-visualization–driven framework—spanning entity co-occurrence networks, Louvain community detection, and multi-view visualizations (including Sankey, sunburst, and chord diagrams)—to advance media-driven studies of regional innovation corridors.

## 3 Data Collection and Dataset Construction

This section provides a detailed account of the multi-stage process for acquiring, filtering, and structuring the data used in this study. The workflow begins with sourcing raw information and concludes with the construction of a finalized, analysis-ready dataset.

### 3.1 Data Acquisition and Filtering Strategy

The empirical data for this research were systematically sourced from the WiseSearch (Huike) News Research Database, a premier all-media information consolidation platform in China. This database ensures the authoritativeness and timeliness of the data by aggregating content from over 1,200 print publications and more than 8,000 online media outlets. The selected temporal scope, from June 2021 to June 2025, strategically covers the critical implementation phase of China's 14th Five-Year Plan.

To capture a holistic and multi-faceted view, our data collection strategy targeted three principal categories of media sources:

- **Official Local Government Channels:** Publications from governmental media platforms, such as the "Shanghai Release" platform[1].
- **Mainstream State Media Outlets:** Authoritative national reports from premier news organizations, such as Xinhua News Agency[2] and People's Daily[3].
- **Major Portal Websites:** Widely-circulated content from high-traffic commercial news portals, including Tencent News[4] and Toutiao[5].

A rigorous, multi-stage filtering protocol was designed and executed to guarantee the relevance and accuracy of the final sample.

(1) **Stage 1: Keyword-Based Retrieval.** The initial data retrieval was performed using a set of core keywords central to the research, namely: "G60 Sci-Tech Innovation Corridor," "Yangtze River Delta integration," and "technological innovation." To enhance the comprehensiveness and recall of the search, this list was expanded with supplementary terms like "industrial upgrading," "technology transfer," and "innovation ecosystem."
(2) **Stage 2: Automated Noise Reduction.** The retrieved articles then underwent an automated cleaning phase. This process involved applying filters to exclude articles containing keywords indicative of irrelevant topics, such as "real estate listings," "housing prices," and other commercial promotions, thereby performing a first-pass noise reduction.
(3) **Stage 3: Manual Validation.** As a final and crucial step, the remaining articles were subjected to a thorough manual content review. Each article was individually assessed by researchers to confirm its contextual relevance to the study's objectives. This human-in-the-loop validation was essential for ensuring the high quality and thematic consistency of the data.

This meticulous acquisition and filtering process yielded a final corpus of 3,786 relevant news articles, which formed the basis for the dataset construction.

### 3.2 Dataset Structuring and Finalization

The 3,786 news articles, initially existing as a collection of semi-structured raw text files from the database, were then programmatically processed to create a clean and structured dataset. This was achieved using a custom Python script leveraging the `pandas` and `re` (regular expressions) libraries.

The primary task of the script was to parse each raw text file. It employed regular expressions to programmatically identify logical boundaries between articles and to locate and extract key metadata fields from the unstructured text. Specifically, the script was designed to extract the following three fields: the publication date, the media source, and the news headline.

Concurrently with extraction, the script performed an initial cleaning by identifying and removing non-analytical content embedded within the files. This included database-specific artifacts (e.g., "Text Snapshot," "Article ID"), navigational text (e.g., breadcrumb links), and standard legal disclaimers. This step ensured that the core content of the news reports was cleanly isolated.

The extracted and cleaned information for each article was then systematically organized into a `pandas` DataFrame. This process culminated in the creation of a single, consolidated, and analysis-ready dataset, which was saved in the CSV (Comma-Separated Values) format. The final dataset consists of 3,786 rows, where each row represents a unique news article, and four structured columns: `date` (the publication date), `source` (the media outlet), `title` (the news headline), and `text` (the full, clean body of the article). This structured CSV file constitutes the final dataset for this study, upon which all subsequent analyses are based.

## 4 Methodology

Following the construction of the structured news dataset, our methodology transitions from textual data to network structures, aligning with the paper's "From Words to Networks" framework. This process is systematically organized into three main phases: (1) construction of a core entity lexicon using a Large Language Model (LLM), (2.1) extraction of entity co-occurrence relationships from the text corpus, (2.2) construction of a co-occurrence network, and (3) application of network analysis and visualization techniques.

[1] www.shanghai.gov.cn
[2] www.xinhuanet.com
[3] www.people.com.cn
[4] news.qq.com
[5] www.toutiao.com

## 4.1 LLM-based Core Entity Lexicon Construction

To identify the most salient concepts within the G60 Sci-Tech Corridor discourse, we employed a hybrid approach combining quantitative frequency analysis with advanced LLM-based semantic filtering.

First, to prepare the textual data for analysis, the full text from the 'cleaned_text' column of our dataset was pre-processed. We utilized the `jieba` library in its "precise mode" (`jieba.cut`) for initial word segmentation. During this tokenization, a custom stop word list was applied to remove common, non-informative terms, and an initial custom dictionary (if available) was loaded to enhance segmentation accuracy for domain-specific vocabulary. Following this, an annual word frequency table was generated from the entire pre-processed text corpus. The top 300 most frequent terms were selected as candidates for core entities.

We then utilized a powerful LLM (`Qwen3[7]`) via its API to perform a sophisticated classification task. A detailed system prompt was engineered to instruct the model to screen and categorize these candidate terms into three distinct, predefined dimensions critical to the innovation ecosystem:

- **Industry Entities:** Terms representing specific industrial sectors or fields (e.g., "integrated circuit," "new energy vehicles").
- **Technology Entities:** Terms related to technological innovations, standards, or R&D activities (e.g., "patent," "industrial internet").
- **Capital Entities:** Terms associated with financial instruments, funding, and investment mechanisms (e.g., "fund," "sci-tech innovation board").

The prompt specified that the LLM should select the top 20 most representative entities for each category, providing a brief justification for each selection. This process, combining the statistical significance of high-frequency words with the semantic understanding of an LLM, resulted in a curated and validated core entity lexicon of 60 key terms. This lexicon, stored as an `entity_dict.csv` file, then served a crucial role: it was incorporated as a new custom dictionary into the `jieba` segmentation tool. This dynamic update mechanism ensures that 'jieba' accurately recognizes these newly identified core entities as single tokens in subsequent processing steps, ensuring the lexicon's continued relevance.

## 4.2 Co-occurrence Network Construction

The core of our methodology lies in transforming the identified entities into a network of relationships based on their co-occurrence within the news articles.

*4.2.1 Entity Relationship Extraction.* The process begins by preparing the text of each article. The full text from the 'cleaned_text' column of our dataset was first segmented into individual sentences. Each sentence was then re-tokenized using the `jieba` library. Critically, this re-tokenization step now leveraged the **updated custom dictionary containing the LLM-curated entity lexicon** to ensure accurate and consistent segmentation of all identified domain-specific terms.

With the text fully tokenized, we systematically scanned each article to identify the co-occurrence of entities. An entity pair was considered to have a co-occurrence relationship if two entities from our lexicon appeared together within the scope of a single news article. To focus the analysis on the interplay between different dimensions of the innovation ecosystem, we imposed a crucial constraint: only pairs of entities belonging to **different categories** were recorded (e.g., an Industry-Technology pair, a Technology-Capital pair, or an Industry-Capital pair). This approach filters out intra-category connections to specifically highlight the cross-domain interactions central to our research questions.

*4.2.2 Network Model Formulation.* The extracted co-occurrence pairs were aggregated across the entire corpus to build a symmetric co-occurrence matrix, where the value of each cell $(i, j)$ represents the total frequency that entity $i$ and entity $j$ appeared together in the same article.

Based on this matrix, we constructed an **undirected and weighted network** $G = (V, E, W)$, where:

- $V$ is the set of 60 core entities, which constitute the nodes of the network.
- $E$ is the set of edges, where an edge $(v_i, v_j)$ exists if the co-occurrence frequency between entity $v_i$ and $v_j$ is greater than a predefined threshold. This thresholding, guided by Zipf's law, helps to remove noise from low-frequency, potentially spurious connections while preserving meaningful weak ties.
- $W$ is the set of edge weights, where the weight $w_{ij}$ of an edge $(v_i, v_j)$ is equal to their co-occurrence frequency, signifying the strength of their association in the media discourse.

## 4.3 Network Analysis and Visualization

Building on the constructed entity co-occurrence network, we employed a series of network analysis techniques to uncover the structural and dynamic characteristics of the G60 innovation ecosystem as reflected in news media. First, we calculated node centrality metrics—including degree and weighted degree centrality—to identify core entities that serve as hubs bridging industry, technology, and capital. Next, we applied the Louvain community detection algorithm to reveal modular structures within the network, enabling the identification of closely linked thematic clusters and the examination of cross-domain interactions.

For visualization, we adopted a multi-perspective approach to enhance interpretability. The global network topology was visualized using a force-directed layout, where node size encodes centrality and edge thickness represents co-occurrence strength, intuitively highlighting key actors and their relationships. To capture temporal evolution, parallel coordinate plots were used to track changes in entity importance and community structure across years. This integrated analysis and visualization framework not only makes complex relationships explicit, but also supports the dynamic tracing of innovation trends and the diagnosis of structural features within the regional innovation network.

# 5 Visualization Design

## 5.1 Visualization Objectives

The visualization objective of this paper is to systematically deconstruct the implicit hot topics, focal points, and ecological evolution logic in news reports about the Yangtze River Delta G60 Science and Technology Innovation Corridor through diversified word frequency visualizations and entity co-occurrence graphs. Specifically, it aims to achieve three functions: first, making entity relationships explicit by constructing nodes and arranging them in a two-dimensional plane to reveal associations; second, enabling the tracing of temporal dynamics by incorporating data evolution from 2021 to 2025 into the charts; and third, deconstructing community structures to uncover the influence of informal collaborative networks on regional innovation synergy based on semantic similarity and entity co-occurrence strength. Through this design, the visualization system will address the limitations of traditional quantitative methods in mining unstructured textual relationships, providing dynamic narrative mapping support for policy evaluation.

In terms of element design, the study focuses on three core dimensions: core entities are distinguished by node size and color gradients, with their influence quantified by betweenness centrality; evolutionary trends are presented using Sankey diagrams to display cross-year thematic shifts, supplemented by heat curves to mark technological trends; and innovation collaboration intensity between cities is mapped via chord diagrams, with modularity indices used to evaluate community stability. This multi-scale visualization approach intuitively reveals the G60 Corridor's innovation upgrade path from "factor aggregation" to "ecological coupling."

For community partitioning, the Louvain algorithm is applied to categorize news entities into three major types—capital, technology, and industry—based on empirical observations of the G60 Corridor's innovation elements. Capital entities (e.g., investment, funds) reflect resource allocation mechanisms; technology entities (e.g., AI, semiconductors) reveal knowledge production dynamics; and industry entities (e.g., new energy vehicles, biopharma) indicate commercialization pathways. This classification allows the visualization system to examine synergies and imbalances—such as whether capital flows align with technology maturity curves or whether industrial upgrades receive sustained technological support. By analyzing inter-community flows and barriers, it can diagnose structural blockages or policy inefficiencies in the regional innovation ecosystem.

## 5.2 Color, Layout, and Interaction Overview

In the data visualization design for network analysis of news texts related to the Yangtze River Delta G60 Science and Technology Innovation Corridor, this study enhances chart readability and information communication effectiveness through optimized color schemes, layout designs, and interaction logic. The color scheme follows principles of semantic association and visual hierarchy. First, primary hues are assigned based on entity categories: industrial, technological, and capital entities are distinguished by different color tones for rapid classification. Second, lightness gradients within the same color family reflect entity influence—high-frequency or core entities use high-saturation colors, while secondary entities are marked with lighter tones to establish clear visual hierarchy. In interactive displays, attention is guided by highlighting related nodes (e.g., using high-contrast colors) and reducing the opacity of unrelated elements. For temporal evolution visualizations (e.g., Sankey diagrams), progressive colors mark cross-period changes of the same entity, ensuring visual continuity.

The visual design employs a multi-view coordination approach to balance information density and readability. Force-directed algorithms arrange nodes, with highly interconnected entities naturally clustering at the visual center and peripheral nodes radiating outward, intuitively reflecting the hierarchical structure of the innovation network. Sankey diagrams are divided into four annual segments, with nodes layered horizontally to clearly display temporal changes in entity relationships. Supplementary views—such as line charts and stacked bar charts—respectively illustrate nested relationships between entities and communities as well as frequency distributions of different entity categories, optimizing visual presentation.

Through meticulously designed color encoding, spatial arrangement, and interactive features, this study develops a scientifically robust data visualization framework that provides strong visual support for analyzing the G60 Corridor's innovation ecosystem. All visualizations are programmatically implemented rather than constructed as a complete system.

## 5.3 Overview of Main Visualization Types

The main types of overview visualizations (see illustrative examples below) include:

a. **Line chart:** Used to represent the temporal evolution of entities.
*Figure 2 Example of a Line Chart.*

b. **Force-directed graph:** Designed to reveal the global structure of the entity network.
*Figure 3 Example of a Force-directed Chart.*

c. **Parallel coordinate plot:** Used to illustrate the evolution of communities across multiple dimensions.
*Figure 4 Example of a Parallel Coordinate Graph.*

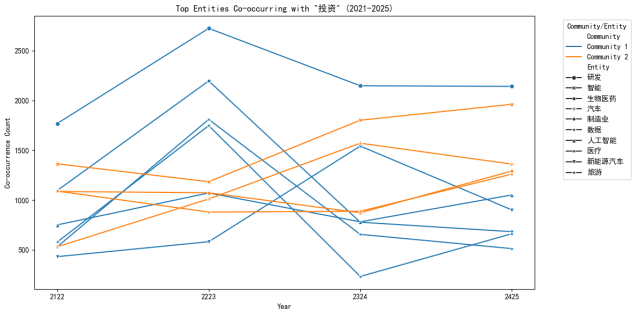


**Figure 2: Example of a Line Chart.**

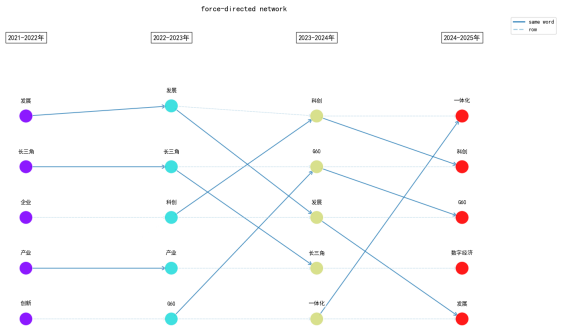


**Figure 3: Example of a Force-directed Chart.**

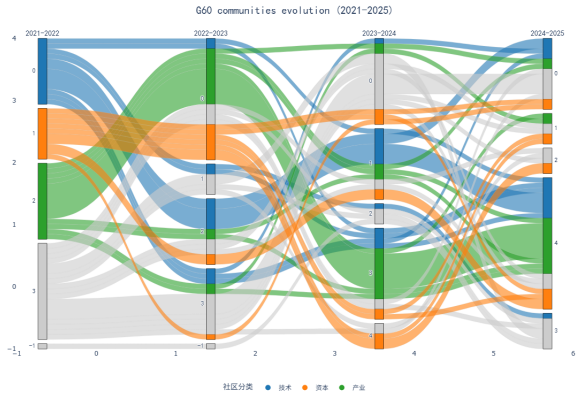


**Figure 4: Example of a Parallel Coordinate Graph.**

## 6 Results

### 6.1 Entity Co-occurrence Network and Community Structure

The co-occurrence network diagram illustrates the co-occurrence relationships between different entities. This diagram is generated based on data read from a CSV file, where each entity is assigned a distinct color according to its category (industry, technology, capital), and node positions are determined by the entities' co-occurrence frequency. To establish edges in the network, the code employs an improved connection logic: if two adjacent entities belong to the same category, an edge with a weight of 0.5 is added between them, highlighting strong intra-category connections. The network layout is optimized using the `spring_layout` algorithm to ensure visual appeal and readability. Finally, an interactive co-occurrence network diagram is generated using the Plotly library. The diagram includes detailed node information such as entity name, category, frequency, and description. Through color and size encoding, it intuitively displays co-occurrence relationships and the relative importance of entities.

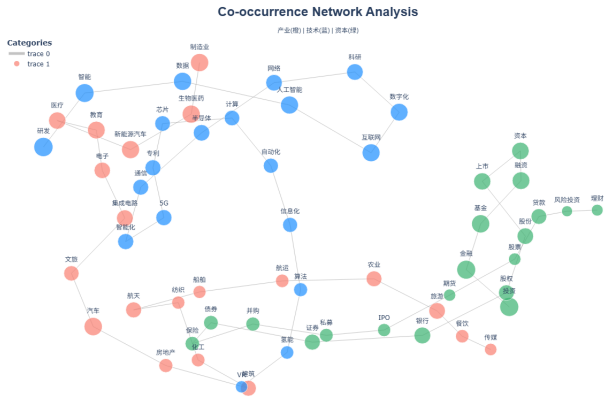


**Figure 5: Co-occurrence network diagram of G60 entities.**

The stacked bar chart displays the total word frequency of key entities across three communities, with each community represented by differently colored segments. This allows for a visual comparison of the proportional differences between communities in the overall distribution and the influence of their core entities. From the chart, it is evident that entities in the technology community appear most frequently in the filtered G60 news texts, exceeding 25,000 mentions. The capital community ranks second, while the industry community has the lowest frequency. In terms of entity distribution within communities, terms such as "manufacturing," "biopharmaceuticals," "healthcare," and "new energy" appear most frequently, dominating the industry community. The technology community is primarily characterized by high-frequency terms like "R&D," "intelligent," "artificial intelligence," and "data." Meanwhile, the capital community's core entities include "investment," "finance," "fund," and "financing."

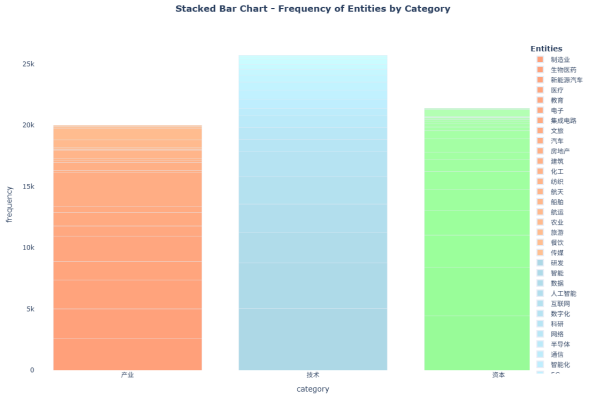


**Figure 6: Stacked bar chart of G60 entity word frequency distribution.**

The figure presents the word frequency distribution of various industries and technology categories within the G60 region. The radar chart uses distance from the center to represent the word

frequency of each entity, clearly highlighting key focus areas such as manufacturing and automotive industries, as well as technology fields like R&D and intelligent systems. This indicates these sectors serve as crucial pillars and development priorities for the G60 region. Within the technology community, high-frequency entities include ”R&D”, ”intelligent”, ”internet”, and ”digitalization”. The industry community shows an uneven distribution, with ”automotive”, ”education”, ”healthcare”, and ”manufacturing” emerging as predominant entities. In the capital community, ”investment” and ”finance” demonstrate notably higher frequencies.

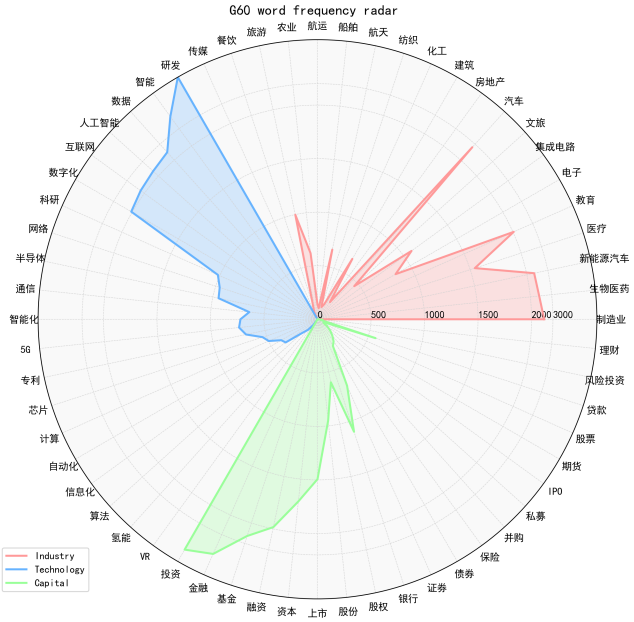


**Figure 7: Radar chart of G60 entity word frequency distribution.**

## 6.2 Temporal Evolution and Co-occurrence Patterns

The heatmap displays a four-year co-occurrence matrix heatmap (2021-2025) showing the co-occurrence frequency between different entities. This heatmap was generated by merging datasets from four years and using a pivot table to summarize the co-occurrence counts for each entity pair.

Specifically, the process began by reading annual co-occurrence matrix data files. These datasets were then merged, with year labels added to each entry. Next, a pivot table was created to aggregate co-occurrence values by entity pair, and the indexes were rearranged to improve heatmap visualization. Finally, the heatmap was plotted using the Seaborn library with a ”YlGnBu” color map. Only color gradients were shown (without numerical values) to avoid data overlap and ensure readability. This approach helps visually reveal co-occurrence patterns and intensity between entities across different years.

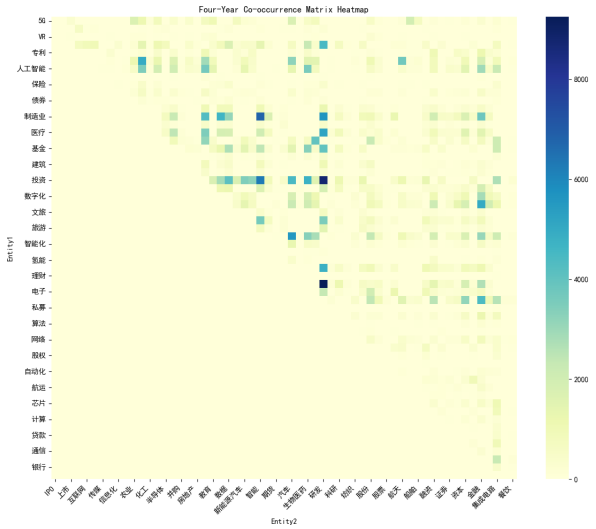


**Figure 8: Heatmap of G60 entity co-occurrence frequencies (2021-2025).**

The data show that key co-occurring word pairs include “manufacturing-intelligent,” “investment-intelligent,” “investment-R&D,” and “biopharmaceuticals-R&D.” These results indicate that, over the four-year period, intelligent manufacturing represented a leading technological frontier. Capital flows were primarily directed toward intelligent technologies, R&D, and biopharmaceutical fields. This reflects the interdependence and collaborative development trends among these sectors.

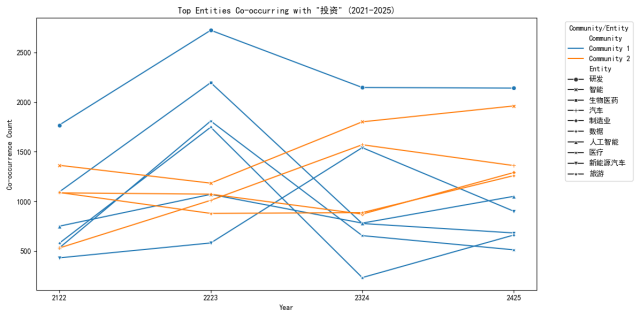


**Figure 9: Line chart of entities co-occurring with “investment” (2021–2025).**

## 6.3 Temporal Evolution and Co-occurrence Patterns

The images visually reflect the spatial characteristics of media coverage through geographic maps. The program reads four CSV files (covering data from 2021 to 2025), extracts and merges the place name frequency information within them. First, a list of place names in the Yangtze River Delta region and their coordinates are defined, and the word frequency data is processed based on these

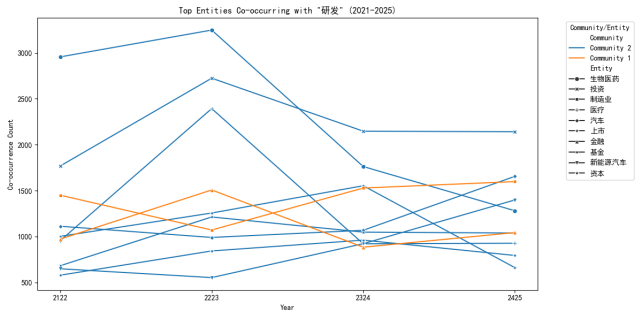


**Figure 10: Line chart of entities co-occurring with "R&D" (2021–2025).**

place names. Next, GeoPandas is used to read and merge the geographic JSON files of the Yangtze River Delta region, generating map boundaries. On the map, the word frequencies of place names are converted into different colors using Matplotlib's colormap, and the place names are labeled at their corresponding locations. Finally, a scale bar, north arrow, and color bar are added to enhance the map's readability. To further refine the analysis, we also separately plotted the distribution of place name frequencies in the Shanghai area, focusing only on information related to Shanghai, and processed and displayed it using the same methods.

These two images show the distribution of place name frequencies in the Yangtze River Delta region and the Shanghai area during different time periods. Through the visual representation of word frequencies, we can understand the importance and frequency of occurrence of various cities and regions in the text data. The Yangtze River Delta region map helps identify which areas are frequently mentioned within the overall scope, while the Shanghai area map provides a more detailed local perspective, supporting deeper regional research.

From an overall perspective of the Yangtze River Delta region map, high-frequency areas in G60 news texts are widely distributed across the Yangtze River Delta. Among them, Shanghai, as a core city, occupies a prominent position in the word frequency distribution, demonstrating its dominant role in the text data. Other major cities such as Nanjing, Suzhou, and Hangzhou also show relatively high word frequencies, reflecting their significance. Additionally, some smaller cities and regions also exhibit certain word frequency distributions, indicating their presence in specific textual content. The color bar displays different levels of word frequency, making the distinction between high-frequency and low-frequency words more apparent. In the Shanghai area map, Songjiang District, Lingang, and Zhangjiang show higher word frequencies, likely due to their greater activity in economic, cultural, and other aspects. In contrast, some suburban and more remote areas have relatively lower word frequencies, reflecting their lower attention in the text data. The color bar further highlights the differences in word frequency, making the varying levels of importance across different regions clear at a glance.

Parallel coordinate plots are a multidimensional data visualization technique that can intuitively display the relationships and changing trends among various dimensions in high-dimensional

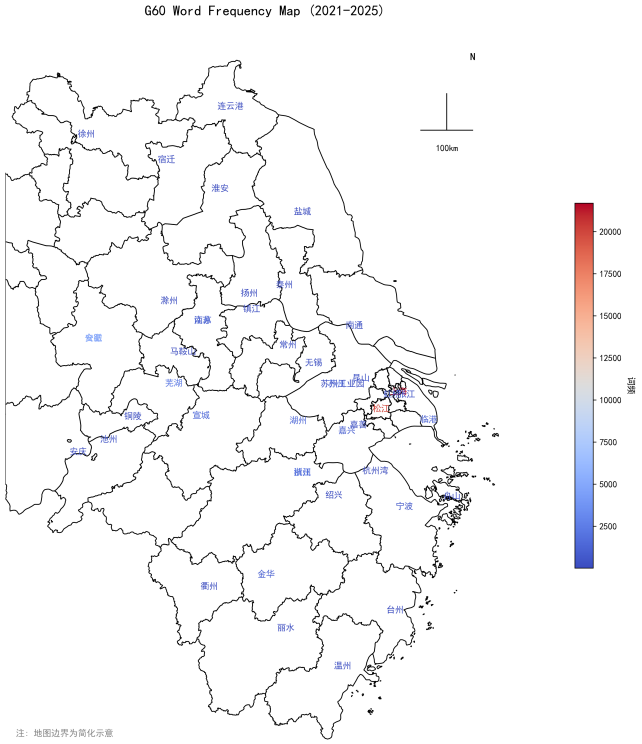


**Figure 11: Word frequency distribution in Yangtze River Delta region (2021-2025).**

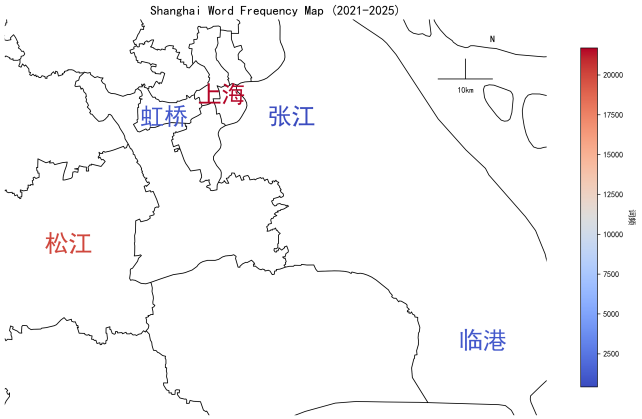


**Figure 12: Detailed word frequency distribution in Shanghai.**

data. In this study, we employ parallel coordinate plots to analyze the evolutionary relationships among three types of entities—industry, technology, and capital—in the Yangtze River Delta G60 Science and Technology Innovation Corridor over four years. As shown in Figure X, each vertical axis represents a year, and the horizontal polylines connect the word frequencies of the same entity across different years. The dynamic Sankey diagram is generated using the Plotly library, allowing users to adjust the displayed year

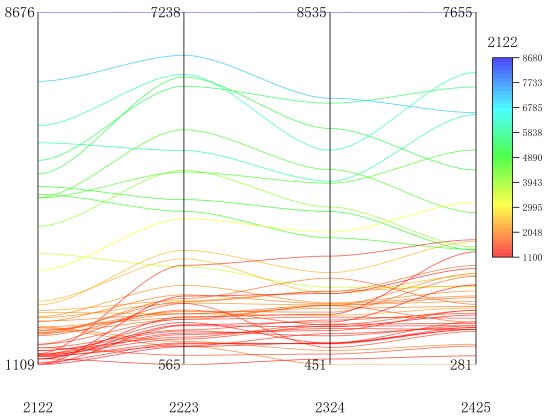


**Figure 13: Parallel coordinates visualization of tri-dimensional entity evolution (2021-2025).**

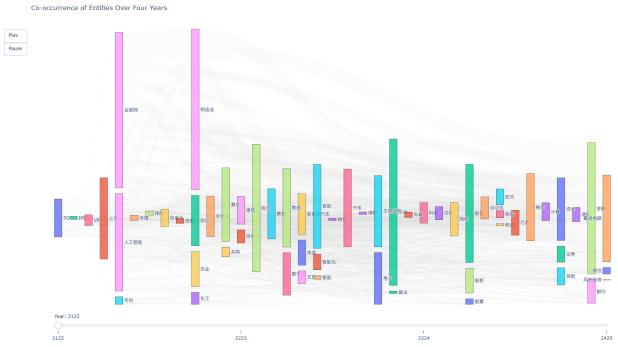


**Figure 14: Interactive Sankey diagram depicting evolving entity co-occurrence networks (2021-2025).**

through the settings below, visualizing the co-occurrence relationships between entities and their changing trends across different years. This illustrates the temporal evolution of media coverage hotspots. The Sankey diagram reveals the co-occurrence relationships among entities and their variation patterns over time. By visualizing the co-occurrence matrix, researchers can intuitively identify which entities frequently appear together in different time periods, thereby uncovering their interconnections and relative importance. This dynamic presentation helps identify long-term stable relationships as well as evolving trends.

From the Sankey diagram, it can be observed that certain entities maintain consistently high co-occurrence frequencies throughout the entire period, indicating stable associations between them. For example, entities such as "R&D" and "manufacturing" exhibit persistently high co-occurrence rates from 2021 to 2025, likely due to their close relevance in specific domains or themes. Meanwhile, entities like "shares" and "financing" show fluctuating co-occurrence frequencies across different years, reflecting shifts in their relationships over time. The color scale and link thickness clearly indicate the intensity of different co-occurrence frequencies, facilitating a better understanding of interaction patterns among entities. Overall, the Sankey diagram provides an intuitive and effective method for exploring and analyzing co-occurrence relationships and trending topics within multidimensional data.

## 7 Discussion

### 7.1 Method Effectiveness and Limitations

This study systematically deconstructs news coverage of the Yangtze River Delta G60 Science and Technology Innovation Corridor through multidimensional visualization techniques, achieving significant methodological effectiveness. By employing visualization tools including co-occurrence networks, Sankey diagrams, and heatmaps, the research not only renders entity relationships explicitly visible but also captures the temporal evolution of hotspot topics while deciphering the community structure of the innovation network. Co-occurrence network graphs combined with node encoding techniques reveal complex interactions among three core entities: industry, technology, and capital, with particular emphasis on synergistic triads such as R&D, investment, and manufacturing. Temporal analysis tools demonstrate the rising prominence of emerging concepts like the digital economy, corroborating the measurable impact of policy inflection points on media discourse. Community structure analysis confirms the centrality of technology clusters within the innovation network, as well as the strong interdependencies between subfields including biomedicine, smart technologies, and capital flows. These findings provide robust empirical validation for the "innovation-driven" development model of the G60 region.

However, several methodological limitations merit consideration. At the data collection stage, reliance on publicly available news reports may introduce media source bias, disproportionately reflecting narratives from governmental or large enterprise sources while overlooking the innovation dynamics of small and medium-sized enterprises. Geographic coverage exhibits spatial heterogeneity, with word frequencies for core cities like Shanghai and Suzhou substantially higher than those in peripheral regions, potentially underestimating marginal cities' contributions to regional synergy. From a technical perspective, entity recognition based on pre-trained models may produce inaccuracies in extracting specialized terminology, particularly for technical entities where misidentification risks persist. The determination of co-occurrence relationships relies solely on textual proximity, which fails to distinguish between positive and negative semantic associations, potentially inflating the perceived strength of certain entity connections. Furthermore, interpreting complex visualizations such as Sankey diagrams and parallel coordinate plots may require specialized training to interpret effectively, which could limit direct application by non-technical stakeholders including policymakers.

### 7.2 Implications for Policy and Communication

The research findings offer actionable insights for regional innovation policy formulation and media communication strategies. Regarding innovation policy, the strong correlation between capital and technology communities suggests policy resources should prioritize cutting-edge R&D fields such as biomedicine and artificial intelligence, while enhancing "industry-academia-research" collaboration through targeted instruments like special funds. Heatmap analyses revealing high co-occurrence between manufacturing and intelligence domains indicate strategic opportunities for developing intelligent manufacturing clusters along the G60 corridor, fostering industrial chain complementarity between high-frequency regions like Songjiang and surrounding cities. The study advocates establishing a dynamic monitoring mechanism using visualization tools to track emerging fields like the digital economy, thereby preventing excessive concentration of policy resources on traditional industries.

For media communication practices, the current disproportionate focus on core cities including Shanghai and Suzhou requires recalibration, with expanded coverage of innovation cases from Anhui, Zhejiang, and other peripheral regions to cultivate a more balanced regional innovation narrative. Media practitioners can adopt the "continuity" logic demonstrated in Sankey diagrams, emphasizing the long-term accumulation process inherent in technological breakthroughs within science reporting to enhance public understanding of the complete innovation chain from R&D to industrialization.

### 7.3 Comparison with Existing Literature

Compared to existing literature, this study has made valuable expansions in both methodology and empirical findings. Methodologically, unlike traditional word frequency statistical techniques, the innovative combination of various visualization methods employed in this study can more comprehensively reveal the spatiotemporal evolutionary characteristics of entity relationships, effectively compensating for the limitations of static quantitative analysis. While existing literature often focuses on single-dimensional analysis, this study creatively integrates three key elements—industry, technology, and capital—aligning more closely with the basic framework of innovation ecosystem theory.

In terms of empirical findings, the research not only verifies existing literature's conclusion about Shanghai being a core node, but also discovers through temporal analysis the new phenomenon of its radiating effects weakening after 2023, while observing the upward trend in word frequency of peripheral cities like Wuhu. These findings form an interesting contrast with the "capital-dominant" hypothesis proposed by earlier research, with actual data better supporting the "technology-capital dual-drive" development model in the G60 region.

At the policy implication level, this study breaks through the limitation of existing literature's overemphasis on hardware infrastructure construction, highlighting through visual analysis the crucial role of informal collaboration networks in regional innovation, and adding a new dimension of "innovation community cultivation" to innovation policy, which holds important reference value for improving regional innovation governance systems.

## 8 Conclusion and Future Work

### 8.1 Conclusion

This study systematically reveals the implicit characteristics of innovation ecosystems in news coverage of the Yangtze River Delta G60 Science and Technology Innovation Corridor by constructing a multidimensional visual analysis framework. The research finds that three types of entities—industry, technology, and capital—form a synergistic network centered on "R&D-investment-manufacturing," with technology communities playing a dominant role in the innovation ecosystem. Fields such as biomedicine and artificial intelligence show high coupling with capital flows. Temporal analysis captures the rising trajectory of emerging concepts like the "digital economy," confirming the actual impact of policy inflection points on innovation ecosystems. Geographical analysis reveals significant spatial imbalances in media coverage, with core cities like Shanghai and Suzhou occupying dominant positions. These findings not only validate the "innovation-driven" development logic of the G60 region but also provide empirical evidence for regional innovation policy formulation.

Visual analysis demonstrates unique methodological value in news text research. By transforming unstructured text into dynamic network diagrams, this method overcomes the limitations of traditional quantitative analysis, achieving explicit representation of complex relationships and intuitive presentation of temporal evolution. Network visualization can not only deconstruct hidden entity relationships and community structures in texts but also capture the evolutionary patterns and focus preferences of media coverage, providing new analytical perspectives for policy evaluation and communication research. This research paradigm integrating computer science, information visualization, and social sciences enhances both the interpretability of textual data and the decision-making reference value of research outcomes, opening new pathways for regional innovation ecosystem studies.

### 8.2 Future Work

Future research could deepen the application of visual analysis in news texts through three directions. First, integrating sentiment analysis and topic modeling techniques could quantify reporting attitudes and identify implicit themes, enabling a more comprehensive understanding of innovation discourse systems. Second, expanding into multimodal data analysis by using computer vision to extract innovation elements from news images and correlating them with textual network analysis would establish a more three-dimensional framework for innovation communication research. Finally, developing real-time monitoring systems based on dynamic visualizations to create innovation early-warning mechanisms—setting thresholds for key indicators like technology term frequency growth and regional collaboration density—would provide immediate data support for policy adjustments and accelerate the translation of research findings into practical applications.

## 9 Authors' Contributions

Jingyao.Zhong was responsible for data collection, data preprocessing, network construction, the overall visualization framework design, manuscript writing, and LaTeX typesetting.

Yufei.Wang was responsible for the practical implementation of visualizations, figure generation, manuscript writing, and data analysis.

## References


[1] Tsvetanka Georgieva-Trifonova and Miroslav Dechev. 2021. Applying text mining methods to extracting information from news articles. *IOP Conference Series: Materials Science and Engineering* 1031 (01 2021), 012054. doi:10.1088/1757-899X/1031/1/012054

[2] Jinghua Li, Liran Li, Jingyu Li, and Jie Lyu. 2025. Research on Regional Restructuring from the Perspective of Regional Innovation Systems: A Case Study of the Yangtze River Delta G60 Science and Technology Innovation Corridor. *Science and Technology Management Research* 45, 7 (2025), 113–123. doi:10.3969/j.issn.1000-7695.2025.7.012

[3] k. Norvag and R. Oyri. 2005. News Item Extraction for Text Mining inWeb Newspapers. In *International Workshop on Challenges in Web Information Retrieval and Integration*. 195–204. doi:10.1109/WIRI.2005.27

[4] Kalyani Pakhale. 2023. Comprehensive Overview of Named Entity Recognition: Models, Domain-Specific Applications and Challenges. arXiv:2309.14084 [cs.CL] https://arxiv.org/abs/2309.14084

[5] Ruidong Sun and Liu Chen. 2024. The Innovative Effects of Regional Industrial Coordination Policies: Evidences from the G60 S&T Innovation Valley of Yangtze River Delta. *East China Economic Management* 38, 10 (2024), 36–45. doi:10.19629/j.cnki.34-1014/f.240630012

[6] Caijun Wang and Gaoping Liu. 2024. Impact of Technology Niche on the Innovation Efficiency of G60 S&T Innovation Corridor in the Yangtze River Delta. *Science and Technology Management Research* 44, 9 (2024), 83–91. doi:10.3969/j.issn.1000-7695.2024.9.010

[7] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. arXiv:2505.09388 [cs.CL] https://arxiv.org/abs/2505.09388