



## Measuring Inconsistencies in Research and Development Descriptions in Annual Reports of Listed Companies

Jing Li<sup>1</sup> , Jing Qi<sup>2</sup> , Yanchun Zhu<sup>1\*</sup> , Wei Zhang<sup>3</sup> , Yicen Jiang<sup>4</sup>

<sup>1</sup> Business School, Beijing Normal University, 100875 Beijing, China

<sup>2</sup> Visitor Services Department, China Science and Technology Museum, 100101 Beijing, China

<sup>3</sup> School of Information, Central University of Finance and Economics, 100081 Beijing, China

<sup>4</sup> Business School, Renmin University of China, 100872 Beijing, China

\* Correspondence: Yanchun Zhu (kddzw@163.com)

**Received:** 06-15-2024

**Revised:** 08-15-2024

**Accepted:** 08-21-2024

**Citation:** Li, J., Qi, J., Zhu, Y. C., Zhang, W., & Jiang, Y. C. (2024). Research on measuring inconsistencies in R&D descriptions in the annual reports of listed companies. *J. Organ. Technol. Entrep.*, 2(3), 136-150. <https://doi.org/10.56578/jote020301>.



© 2024 by the author(s). Published by Acadlore Publishing Services Limited, Hong Kong. This article is available for free download and can be reused and cited, provided that the original published version is credited, under the CC BY 4.0 license.

**Abstract:** The descriptions of Research and Development (R&D) activities in the annual reports of listed companies provide crucial insights into a company's internal governance, external competitiveness, and long-term sustainability strategies. However, R&D disclosures in China's securities market are largely semi-mandatory, often leading listed companies to adopt either "self-enhancement" or "self-suppression" approaches in their R&D activity descriptions. This inconsistency between actions and disclosures erodes trust among market participants and exacerbates information asymmetry. Based on annual report data from Chinese manufacturing firms, this study assesses the intensity of R&D activity disclosures from both accounting data and textual information using Latent Dirichlet Allocation (LDA) topic modeling and principal component analysis (PCA). Normalized difference metrics are applied to quantify the level of inconsistency between these two dimensions. Empirical findings reveal a prevalent degree of inconsistency in R&D activity descriptions within manufacturing firms' annual reports. Furthermore, both accounting and textual disclosure intensities have increased over time, with inconsistency levels initially rising and then showing a marked decline. The findings offer a theoretical basis for enhancing and standardizing R&D activity descriptions and disclosures, serving as a resource for government, companies, third-party agencies, and investors.

**Keywords:** Research and Development (R&D) activity description; Information disclosure; Inconsistency measurement; Listed companies

### 1. Introduction

The R&D of listed companies is linked to corporate value (Fedorova et al., 2023) and market reaction (Cheng et al., 2022), serving as the core competitiveness for a company's survival and development. R&D encompasses corporate R&D investments, which to a certain extent reflect innovation capability and technological strength, thus forming the core competitiveness of a company (Cheng et al., 2023). The R&D information disclosed by listed companies mainly falls into two categories: one includes data-based indicators such as R&D expenditures and patent applications, while the other consists of textual narrative information in sections like "Company Business Overview," "Management Discussion and Analysis", "Principal Business", and "Corporate Development Strategy", which pertain to R&D activities. The former displays the company's R&D investments and innovation outputs through specific numerical information. However, as the amount of R&D investments rapidly increases, and with the frontier nature, complexity, and uncertainty of R&D activities rising, the limitations of standardized data indicators become prominent, resulting in a decrease in their correlation with market value (Ciftci & Zhou, 2016). The latter interprets corporate R&D information in narrative form, compensating for the limitations of standardized indicators that are uniform, require high investor expertise, and disclose limited information, thereby conveying the hard-to-quantify and observable R&D value of the company to investors more effectively. Although both data indicators and textual information provide potentially valuable insights into R&D activities, quantitative



## Measuring Inconsistencies in R&D Descriptions in Annual Reports of Listed Companies

Jing Li<sup>1</sup>, Jing Qi<sup>2</sup>, Yanchun Zhu<sup>1\*</sup>, Wei Zhang<sup>3</sup>, Yicen Jiang<sup>4</sup>

<sup>1</sup> Business School, Beijing Normal University, 100875 Beijing, China

<sup>2</sup> Visitor Services Department, China Science and Technology Museum, 100101 Beijing, China

<sup>3</sup> School of Information, Central University of Finance and Economics, 100081 Beijing, China

<sup>4</sup> Business School, Renmin University of China, 100872 Beijing, China

\* Correspondence: Yanchun Zhu (kddzw@163.com)

**Received:** 06-15-2024

**Revised:** 08-15-2024

**Accepted:** 08-21-2024

**Citation:** Li, J., Qi, J., Zhu, Y. C., Zhang, W., & Jiang, Y. C. (2024). Research on measuring inconsistencies in R&D descriptions in the annual reports of listed companies. *J. Organ. Technol. Entrep.*, 2(3), 136-150. <https://doi.org/10.56578/jote020301>.



© 2024 by the author(s). Published by Acadlore Publishing Services Limited, Hong Kong. This article is available for free download and can be reused and cited, provided that the original published version is credited, under the CC BY 4.0 license.

**Abstract:** The descriptions of Research and Development (R&D) activities in the annual reports of listed companies provide crucial insights into a company's internal governance, external competitiveness, and long-term sustainability strategies. However, R&D disclosures in China's securities market are largely semi-mandatory, often leading listed companies to adopt either "self-enhancement" or "self-suppression" approaches in their R&D activity descriptions. This inconsistency between actions and disclosures erodes trust among market participants and exacerbates information asymmetry. Based on annual report data from Chinese manufacturing firms, this study assesses the intensity of R&D activity disclosures from both accounting data and textual information using Latent Dirichlet Allocation (LDA) topic modeling and principal component analysis (PCA). Normalized difference metrics are applied to quantify the level of inconsistency between these two dimensions. Empirical findings reveal a prevalent degree of inconsistency in R&D activity descriptions within manufacturing firms' annual reports. Furthermore, both accounting and textual disclosure intensities have increased over time, with inconsistency levels initially rising and then showing a marked decline. The findings offer a theoretical basis for enhancing and standardizing R&D activity descriptions and disclosures, serving as a resource for government, companies, third-party agencies, and investors.

**Keywords:** Research and Development (R&D) activity description; Information disclosure; Inconsistency measurement; Listed companies

### 1. Introduction

The R&D of listed companies is linked to corporate value (Fedorova et al., 2023) and market reaction (Cheng et al., 2022), serving as the core competitiveness for a company's survival and development. R&D encompasses corporate R&D investments, which to a certain extent reflect innovation capability and technological strength, thus forming the core competitiveness of a company (Cheng et al., 2023). The R&D information disclosed by listed companies mainly falls into two categories: one includes data-based indicators such as R&D expenditures and patent applications, while the other consists of textual narrative information in sections like "Company Business Overview," "Management Discussion and Analysis", "Principal Business", and "Corporate Development Strategy", which pertain to R&D activities. The former displays the company's R&D investments and innovation outputs through specific numerical information. However, as the amount of R&D investments rapidly increases, and with the frontier nature, complexity, and uncertainty of R&D activities rising, the limitations of standardized data indicators become prominent, resulting in a decrease in their correlation with market value (Ciftci & Zhou, 2016). The latter interprets corporate R&D information in narrative form, compensating for the limitations of standardized indicators that are uniform, require high investor expertise, and disclose limited information, thereby conveying the hard-to-quantify and observable R&D value of the company to investors more effectively. Although both data indicators and textual information provide potentially valuable insights into R&D activities, quantitative

R&D investment information can be difficult for investors to understand without supplementary explanation of the development potential underlying the figures (Cheng et al., 2023). Textual narrative information, especially in the Chinese language context, carries richer content than data indicators and holds higher research value (Zeng et al., 2018).

In recent years, to regulate the content of information disclosure by listed companies and improve the quality of disclosures, the China Securities Regulatory Commission has successively introduced more standardized requirements for R&D information disclosure in annual reports of listed companies. Numerous policy releases have gradually made R&D disclosure reforms more stringent and refined. However, companies still maintain significant autonomy in terms of textual disclosures in their annual reports; whether to disclose, how much to disclose, and the manner of disclosure remain at the discretion of the company. Consequently, textual information often contains varying degrees of “language inflation,” causing textual descriptions of R&D activities in annual reports to be subject to different extents of “language inflation,” where stakeholders may benefit from practices such as exaggerating or embellishing innovation through text (Zhou & Lu, 2021).

Existing research primarily relies on textual data from annual reports of listed companies to construct empirical models, examining the impact of R&D textual information on aspects such as primary market pricing efficiency (Wu & Zhao, 2024), corporate value (Feng et al., 2022; Fedorova et al., 2023), patent applications (Chen et al., 2022), R&D investments (Liu et al., 2022; Liu et al., 2023), financing constraints (Xu et al., 2020), stock price crash risk (Yu & Xiao, 2022), analyst forecasts (Xu & Zhu, 2019), and competitiveness (Lakhal & Dedaj, 2020) from various perspectives including disclosure level, tone, readability, and similarity. However, few studies focus on the relationship between textual descriptions and quantitative data indicators, overlooking the consistency between the two. The inconsistency between textual information and data indicators regarding R&D activities in corporate annual reports intensifies information asymmetry in the market and disrupts the order of the capital market.

Furthermore, academic research has predominantly concentrated on either financial accounting data or textual disclosure information from a single perspective, with a bias toward examining the influence of fundamental financial indicators on R&D investment levels. There is a lack of literature that integrates both R&D textual and accounting data perspectives to comprehensively study the descriptions of R&D activities in annual reports.

Based on the above background, this paper takes Chinese manufacturing listed companies as the research subject and conducts quantitative measurements on R&D activity descriptions in annual reports from the perspectives of textual information and accounting data, comparing and analyzing the differences and inconsistencies between the two. This provides a policy basis for enhancing the effective implementation of China’s technological innovation, market fairness, transparency, and long-term stability.

## **2. Current State of Related Research**

### **2.1 Inconsistencies in R&D Information Disclosure**

Koh & Reeb (2015) refute the notion that companies not disclosing R&D expenditures are not engaged in innovative activities, demonstrating that a significant portion of such companies are “pseudo-blank” companies—companies that do not disclose R&D expenditures but still engage in innovation activities and apply for patents. In terms of narrative text, Mazzi et al. (2019), through analyzing annual reports of listed companies, found that companies tend to report more about R&D in the narrative sections of their annual reports compared to the financial statements. Tsalavoutas et al. (2014) conducted an extensive survey on the R&D reporting practices of listed companies worldwide and found that over half of the companies did not list any R&D assets or expenses separately in their financial statements. However, a substantial number of these companies used extensive R&D-related terminology in their annual reports. Li & Yao (2020) applied text analysis techniques to extract R&D text information from annual reports, suggesting that R&D narrative descriptions carry informational content and that companies tend to selectively disclose R&D information that favors the company. These findings indicate that disclosed R&D information does not correspond one-to-one with actual R&D activities, and there are inconsistencies between narrative text and data indicators in terms of whether R&D activities are disclosed, the extent of disclosure, and R&D activity levels. Chen et al. (2023) used the Word2vec model to construct a technological innovation dictionary, using the proportion of keywords in company disclosure texts as a measure of subjective technological innovation, and obtained objective technological innovation using seven principal component-based objective indicators. They found inconsistencies between subjective and objective technological innovation, which affected the listing outcomes on the science and technology innovation board. Bellstam et al. (2021) argued that data indicators, such as R&D expenditures and patent outputs, focus on product-related innovation, neglecting other forms of innovation. Their proposed LDA-based text innovation measure filled this gap, assessing a broader scope of company innovation. Their study found that text-based innovation measures were positively related to sales growth, while patent counts showed a negative correlation with sales growth under fixed effects, and there was no significant relationship between R&D intensity and sales growth.

In summary, the academic field has conducted considerable research on the two distinct forms of R&D information. While previous studies have recognized the phenomenon and risks of “language inflation” in R&D narratives within annual reports, there is still a lack of research on the measurement and impact of inconsistencies between R&D narrative descriptions and data indicators. Quantifying these inconsistencies between narrative text and data indicators and analyzing their effects is a topic that requires urgent further exploration.

## 2.2 Inconsistency Measurement Methods

Measuring inconsistencies between numerical data and textual information has always been a challenging task, involving the identification and quantification of structured and unstructured data. Current research in this field primarily focuses on scenarios such as mixed reviews on e-commerce platforms (Wu et al., 2024), medical Q&A systems (Xia et al., 2023), legal document review (Li et al., 2024), fake news detection (Zhang & Li, 2021), and financial report audits (Ali et al., 2023). Methods used for measuring inconsistencies in areas like tabular data-text matching, score-rating correlations, and review-text relevance provide valuable references for measuring inconsistencies between accounting data and textual descriptions in annual reports.

In the field of tabular-text Q&A, methods based on attention mechanism models, multimodal fusion, and knowledge enhancement are typically employed for recognition and measurement. Li et al. (2024) constructed a dialogue-level semantic graph and used a multi-relational graph convolutional neural network to capture semantic information, achieving multi-class detection of inconsistencies between tabular data and abstract text using multiple classifiers. Xia et al. (2023) innovatively optimized traditional machine learning algorithms by proposing a Seq2Seq-based model that utilized TAPAS and other methods specialized in table-text Q&A tasks, proving its efficiency and accuracy across financial statement and mathematical report datasets.

For rating-review text matching, studies have often focused on the quality and credibility of UGC to ensure the usefulness of reviews (Liu & Gao, 2024). For instance, Shan et al. (2018) used sentiment analysis methods to extract the sentiment of reviewers from text and used Pearson correlation coefficients and box plots to examine the correlation and inconsistency between product ratings and review sentiment. Hazarika et al. (2021) used sentiment mining techniques to measure the discrepancy between review sentiment and app ratings.

In the context of measuring inconsistencies between R&D accounting data and text descriptions, similarity models (Atabuzzaman et al., 2021), sentiment analysis models (Bigne et al., 2023), and topic modeling techniques (Bai et al., 2019) are commonly applied. Although machine learning and deep learning models and methods are continuously optimized and iterated, improving the efficiency and accuracy of data processing and text recognition, there remains no unified and standardized metric for comprehensively measuring inconsistencies in R&D activity descriptions (Biswas et al., 2022). Therefore, this study considers the use of quantitative approaches for both accounting data and textual information, combining topic modeling, bag-of-words models, and deep learning methods with subjective and objective indicators. By normalizing the difference in numerical data, this approach measures the degree of inconsistency between R&D data and textual intensity, facilitating the quantification of inconsistency levels and addressing the gap in establishing comprehensive inconsistency metrics for R&D activity descriptions. This provides a scientifically credible empirical basis for future improvements in R&D disclosure standards in annual reports.

## 3. Model Assumptions

### 3.1 LDA Topic Model

In this study, the LDA topic model is employed to quantify R&D-related textual data from the annual reports of publicly listed manufacturing companies. By sampling topic information from the documents of each year and identifying the topics most similar to R&D innovation, the LDA model enables the extraction of loadings on the most relevant topics, thus obtaining a quantitative composite score for R&D intensity at the textual information level.

The basic concept of the LDA topic model is to represent each document as a multinomial distribution over topics, which in turn are multinomial distributions over words. When the document-topic and topic-word distributions share a conjugate Dirichlet distribution, unknown parameters can be estimated through statistical sampling. This bag-of-words-based topic model, encompassing prior distributions, sample information, and posterior distributions, measures the degree of content similarity under the criterion of topic similarity.

Since the LDA model disregards word order, words and topics are assumed to be drawn independently, yielding a probability distribution over words. After obtaining the probability distributions of each topic, these are compared with reference documents on R&D activities to calculate the Kullback-Leibler (KL) divergence, thereby assessing the distributional differences between topics and selecting the topic that best aligns with R&D innovation. Based on the KL divergence values, the topic with the minimum value is selected as the most relevant to R&D innovation. Subsequently, the loading of each document  $d$  on topic  $t$  is calculated for each fiscal year. By summing

and averaging the topic loadings of annual documents, a quantitative composite score of R&D intensity at the textual information level is obtained.

### 3.2 PCA

PCA is a statistical technique used to transform high-dimensional data into a lower-dimensional space while preserving as much of the original information as possible. It achieves this by projecting data onto a set of orthogonal axes (the “principal components”) to reduce dimensionality.

This study utilizes PCA to quantify R&D-related financial data from the financial statements of publicly listed manufacturing companies. Key indicators—such as R&D expenditure, the proportion of R&D expenditure, the number of R&D personnel, the proportion of R&D personnel, and the number of patents filed—are subjected to dimensionality reduction, resulting in a quantitative composite score of R&D intensity at the accounting data level.

### 3.3 Research Hypotheses

From both accounting measurement and disclosure regulations, as well as differences in disclosure quality, R&D disclosures in the annual reports of Chinese listed companies exhibit certain variances. Under International Financial Reporting Standards (IFRS), R&D expenditures are often classified into capitalized and expensed forms, leading to differences in classification and presentation. Additionally, some studies have found that the strength of internal controls and audit supervision also affects the accounting records and classifications of R&D activities (Chang et al., 2019). R&D narratives in the annual reports frequently cover management’s strategic positioning, innovation achievements, and anticipated impacts of R&D activities. Particularly in recent years, with increased emphasis on innovation and independent R&D in manufacturing, textual disclosures often carry optimistic and positive tones, manifesting a “self-promotion” phenomenon (Kabuye et al., 2019). Furthermore, companies may selectively disclose R&D data and information based on certain competitive strategies, and under the subjectivity and uncertainty inherent in textual information, sensitive information and core R&D progress are often downplayed, leading to discrepancies in R&D narratives (Cheng et al., 2022; Gordon et al., 2020).

Based on this, Hypothesis 1 is proposed:

Hypothesis H1: There is an inconsistency between accounting data and textual descriptions of R&D activities in the annual reports of Chinese listed manufacturing companies.

In recent years, with the refinement of regulatory environments and standards, Chinese accounting standards have increasingly aligned with international norms, encouraging companies to adopt more rigorous and detailed R&D activity measurement and classification (Li et al., 2019). In a period marked by rapid technological innovation, some scholars (Zhou & Lu, 2021) have suggested that market pressures and growing investor attention have led companies to continually improve the quality and transparency of their R&D disclosures. Other studies (Banerjee, 2022; Zhou et al., 2023) have found that, after recognizing the importance of R&D, certain listed manufacturing companies have strengthened both internal R&D mechanisms and external disclosure efficiency, thereby reducing information asymmetry, enhancing the authenticity, accuracy, and completeness of data and text disclosures, and actively improving R&D disclosure quality to assume social responsibility and establish a favorable market image.

Accordingly, Hypothesis 2 is proposed:

Hypothesis H2: The degree of inconsistency in R&D disclosures in the annual reports of listed manufacturing companies follows a trend of initial increase, followed by a significant decrease over time (by year).

## 4. Empirical Analysis

### 4.1 Research Design

#### 4.1.1 Data sources

The sample consists of data from the annual reports of Chinese manufacturing companies listed on the stock exchange from 2012 to 2022. R&D-related financial data and patent information were obtained from the CSMAR (Guo Tai An) database, while R&D-related textual data were sourced from the CNRDS (China Research Data Service Platform). The initial total sample includes 18,249 annual reports from 1,659 listed companies. To reduce statistical errors and ensure scientific accuracy in the empirical analysis, the following sample screening steps were conducted:

(1) Exclusion of ST and \*ST companies: These companies report abnormal financial conditions, such as negative net profits in two consecutive years or a per-share net asset value below the face value of their stock. These anomalies may lead to distortion and uncertainty in their annual reports and financial statements, potentially causing bias and interference in the results, so these companies were excluded.

(2) Exclusion of samples with missing or obviously erroneous variables: Some companies have incomplete or



extreme data, such as missing R&D expenditure, continuous three-year missing asset data, or abnormal financial performance outliers. These data were excluded to minimize the inconvenience and errors caused by incomplete and extreme data.

After screening, the final sample includes 13,618 annual reports from 1,238 manufacturing companies, covering R&D-related accounting and textual data from 2012 to 2022. The collected data includes R&D expenditure, the proportion of R&D expenditure to revenue, the number of patents filed, the number of R&D personnel, the proportion of R&D personnel, and management discussions and analyses (MD&A) related to R&D activities. Prior to empirical analysis, outliers in continuous variables were trimmed to mitigate the influence of extreme values. All data analysis, text processing, and related estimation tests were carried out using Python and R software.

#### 4.1.2 Variable definition and measurement

##### (1) R&D Activity Intensity – Accounting Data

###### R&D Expenditure

This refers to the total monetary expenditure allocated to R&D activities within a fiscal year, as disclosed in the financial statement under “Total R&D Expenditure”. This expenditure typically includes, but is not limited to, staff salaries and benefits, direct materials, equipment depreciation, and leasing costs directly related to R&D activities. Given the large potential variations in the magnitude of R&D expenditure, the natural logarithm of this value is used in the analysis.

###### R&D Expenditure as a Proportion of Revenue

This key indicator reflects the intensity and focus of a company’s R&D activities, measured as  $(\text{Total R\&D Expenditure}/\text{Total Revenue}) \times 100\%$ . It provides insight into the company’s investment in technological innovation and reliance on R&D for sustained competitiveness and growth.

###### Number of R&D Personnel

The total number of personnel engaged in R&D activities within the company, as disclosed in the MD&A section of the annual report, including full-time and part-time employees such as engineers, scientists, technicians, and others directly involved in R&D. Similar to R&D expenditure, the natural logarithm of this value is used to handle potential large differences in personnel numbers across companies.

###### Proportion of R&D Personnel

This is calculated as  $(\text{Number of R\&D Personnel}/\text{Total Number of Employees}) \times 100\%$ , based on data disclosed in sections such as MD&A or financial report footnotes. It reflects the company’s investment in human capital for innovation and its long-term R&D potential and competitive advantage.

###### Number of Patents Filed

This measures the total number of patents filed with the national intellectual property office or other international patent bodies within a fiscal year, as disclosed in the financial report footnotes. It includes various types of patents, such as inventions, utility models, and designs. It reflects the company’s active pursuit of technological innovations, indicating the quantity of innovations under consideration for protection.

##### (2) R&D Activity Intensity – Textual Data

###### R&D Activity Textual Intensity

To measure the intensity of textual descriptions related to R&D activities, the study analyzes the “Management Discussion and Analysis,” “Company Operating Status,” and “Future Development Outlook” sections of the annual reports. Specifically, the number of times key innovation-related terms appear in the text is counted, and the theme loading of these innovation-related words is evaluated. The ratio of “the number of innovation-related keywords” to the “total number of words in the text” is calculated, along with the corresponding theme loading for these keywords. This provides a composite quantitative measure of the textual intensity of R&D activity descriptions, reflecting the importance placed on R&D activities and the depth of information disclosure in the annual reports.

#### 4.1.3 Model design and steps

This study builds upon the research ideas on corporate R&D investment found in existing literature (Zhou & Lu, 2021; Zhou et al., 2023), and adopts current measurement methods for corporate innovation texts (Bellstam et al., 2021; Ye et al., 2024). The specific steps of the research method are as follows:

##### (1) Accounting Data Processing

PCA is used to reduce the dimensionality and extract features from the R&D activity intensity-related accounting data. The steps include: First, calculating the correlation matrix for each indicator’s sample data. Next, performing eigenvalue decomposition and selecting the principal components based on the ranking of eigenvalues. Finally, projecting the original data onto a new space, obtaining the principal component scores through a linear combination. To mitigate the effects of extreme values, the principal component scores are normalized. The cumulative score for each year’s accounting data on R&D activities is calculated to quantify the intensity of R&D activity descriptions for each year.

## (2) Text Information Processing

Following the R&D dictionary theme words from previous studies (Cheng et al., 2022; Feng et al., 2022; Liu et al., 2022), this study employs the LDA model to perform topic modeling and compute the corresponding topic loadings to quantify the R&D activity intensity-related textual data. The steps include: First, transforming documents into a polynomial distribution of topics and words to build a document-term frequency matrix. Next, constructing a probabilistic model based on the document corpus and maximizing the likelihood function to iteratively calculate the optimal number of topics. Then, calculating the cosine similarity, Jaccard similarity, and KL divergence with reference to innovative theme documents to identify the most similar topics. Finally, obtaining the topic loadings for each document in the identified topics. The overall intensity of the R&D activity disclosure in the text is measured by the topic loadings. The cumulative score for each year's R&D activity description text is calculated to quantify the intensity of R&D activity descriptions for each year.

## (3) Comprehensive Comparison

Normalization and difference comparisons are used to measure the level of inconsistency in the R&D activity descriptions across different years. The steps include: First, standardizing the accounting and text indicator values for each year using the Z-score formula to normalize them to the same scale, facilitating comparison. Then, calculating the difference in normalized indicators for each adjacent year to quantify the changes in the inconsistency of R&D activity descriptions over time. Finally, analyzing the trend changes and fluctuations by calculating the standard deviation and coefficient of variation of the differences. This allows for a comprehensive and systematic analysis of the causes, potential impacts, and trends of the inconsistencies, considering both internal and external factors such as corporate environment and policies. This detailed analysis aims to provide a scientific measurement and in-depth interpretation of the inconsistencies in the R&D activity descriptions in the annual reports of manufacturing companies listed in China.

## 4.2 Empirical Analysis and Results

### 4.2.1 Descriptive statistical analysis

Descriptive statistics of the collected 13,618 annual reports from 1,238 Chinese manufacturing companies (2012-2022) are shown in Table 1. The natural logarithms of the manufacturing industry's R&D expenditure in China (in 10,000 yuan) from 2012 to 2022 range from a maximum of 14.9 to a minimum of 0.67, with a standard deviation of 3.48, indicating a significant disparity in R&D expenditure among companies. Similarly, there are noticeable differences in the number and proportion of R&D personnel, with a disparity of over 9 times between the maximum and minimum values. The number of patents filed in a given year varies greatly, with a maximum of 2,251 and a minimum of 0, reflecting significant variability. This shows that the companies in the selected sample have varying levels of reliance on technological R&D and innovation. The uneven development of R&D activities could be attributed to some companies lacking an early innovation mindset and being less inclined to invest heavily in R&D. From the perspective of R&D text disclosures, the standard deviation is 3.18, and the average is 14.37, further demonstrating substantial variation and imbalance.

These results suggest that there are significant differences in the importance attached to R&D activities and the level of disclosure in both accounting data and textual information among Chinese manufacturing companies. Therefore, studying the level of inconsistency in these disclosures is of theoretical and practical significance.

**Table 1.** Descriptive statistics analysis

Variable Name	Observations	Mean	Standard Deviation	Max Value	Min Value
Ln R&D Investment Amount (10,000 yuan)	13618	8.21	3.48	14.90	0.67
R&D Investment as % of Revenue (%)	13618	4.73	4.66	77.36	0.00
Ln R&D Personnel Count (persons)	13618	8.04	1.95	10.65	0.69
R&D Personnel Proportion (%)	13618	18.41	3.73	94.49	0.00
Number of Patents Filed	13618	154.00	13.67	2251	0.00
Ln R&D Text Frequency Proportion (%)	13618	14.37	3.18	26.88	4.31
Principal Component Scores for Each Year	11	20.57	4.69	26.89	13.11
Optimal Number of LDA Topics	13618	5.09	8.61	19.00	4.00
KL Divergence for Topics	13618	2.30	0.93	4.51	1.63
Total Topic Loadings for Each Year	11	1140.68	614.08	2199.53	482.08

### 4.2.2 LDA topic modeling analysis

#### (1) Data Preprocessing

First, data containing missing values, text with fewer than 100 words, or large amounts of corrupted data were removed. Python was used to detect the language of the text and standardize the storage format. A specialized

dictionary for the manufacturing industry was built to avoid incorrect word segmentation due to professional terms. Additionally, a standard list of stopwords was merged to assist in word segmentation and stopword removal. Finally, using Python, the processed R&D activity description documents were converted into the necessary text feature matrix format for LDA, which is a document-term frequency vector and then input into the LDA model, thereby finishing all tasks of the data preprocessing stage.

#### (2) Model Construction

In the construction of the LDA model, four important parameters need to be determined: the number of topics  $K$ , the document-topic distribution hyperparameter  $\alpha$ , the topic-word distribution hyperparameter  $\beta$ , and the number of iterations.

First, when determining the number of topics  $K$ , based on experience and estimation, the maximum number of topics was set to 20. Referring to Ye et al. (2024), the perplexity and coherence for each topic count were calculated by maximizing the likelihood function. After fully considering the topic interpretability and avoiding overfitting and underfitting, the optimal number of topics was selected based on the minimum perplexity and highest coherence.

Next, for the hyperparameters  $\alpha$  and  $\beta$ , after multiple experiments and referring to relevant literature, the commonly used default values were adopted,  $\alpha(\text{doc\_topic\_prior}) = 1/K$ ,  $\beta(\text{topic\_word\_prior}) = 1/K$ . In addition, to ensure the effectiveness of the model, the number of iterations was set to 500.

Once these key parameters were confirmed, the document-topic probability matrix and topic-word distribution were output, and further visualized and displayed using pyLDAvis and wordcloud. This concluded the model construction phase.

#### (3) Quantification of R&D Text Scores

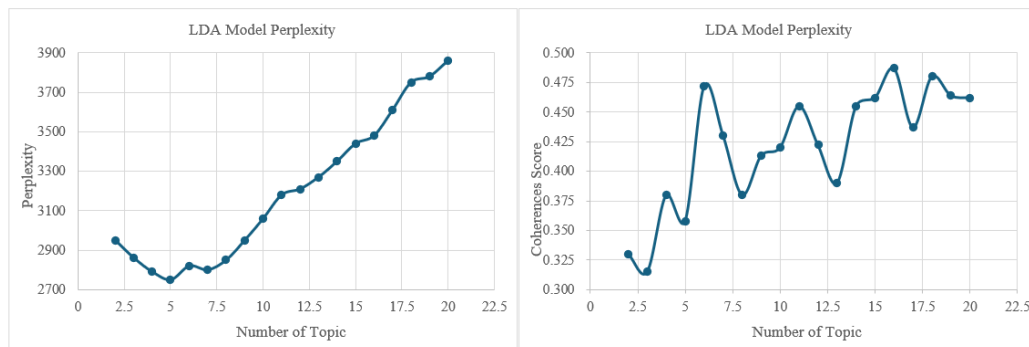
After determining the optimal number of topics and visualizing the topic characteristic words, the similarity to reference documents was calculated to identify the topics most relevant to R&D innovation activities.

First, referring to previous research literature (Cheng et al., 2022; Feng et al., 2022; Liu et al., 2022), this study also selects the authoritative textbook in the field of innovation management, *Innovation Management: Winning with Continuous Competitive Advantage (3rd Edition)* by Chen Jin and Zheng Gang (2016), as the benchmark reference document. By calculating the similarity metrics between each topic and the authoritative reference document, the most similar research and innovation topic is accurately identified. For the similarity metric, this paper uses the KL divergence, i.e., relative entropy, as the calculation indicator. By measuring the degree of difference between two probability distributions, the most similar topic ID is obtained. A smaller KL divergence value indicates a higher similarity to the word distribution in the foundational textbook of research and innovation descriptions, which enhances the accuracy and comprehensiveness of the topic distribution.

Then, after determining the most similar topic, the topic load (i.e., topic probability or topic weight) for each document is calculated. The higher the value, the stronger the association between the content of the document and the research and innovation-related topics. The document topic load for each accounting year is then weighted and averaged to generate a comprehensive score for the research activity description intensity in the annual report for each accounting year.

Finally, the variation and trend of the research activity description intensity in the annual reports for each year are observed, and the inconsistency with the comprehensive score of the accounting data is compared. This forms the basis for further normalization and difference measurement of the inconsistency between the two, thus achieving the core research objective of measuring the inconsistency in the research activity descriptions in the annual reports of listed companies. This concludes the task of quantifying the research text intensity score.

#### (4) Result Analysis



**Figure 1.** Trend of perplexity and coherence in R&D Text LDA for 2018

The LDA modeling analysis was conducted for R&D-related texts in the annual reports of Chinese manufacturing companies listed from 2012 to 2022. For each accounting year, the optimal number of topics related

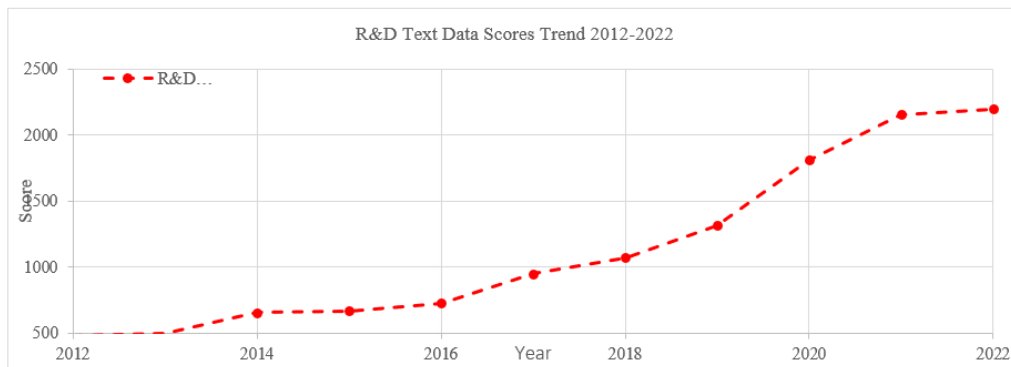


to R&D innovation was determined through perplexity and coherence, and the document-topic probability matrix and topic-word distributions were output. For example, in 2018, the perplexity and coherence scores for various topic numbers were evaluated, as shown in Figure 1.

Based on the trends shown in the figure, it can be observed that in 2018, the optimal number of topics for the lowest perplexity was concentrated between 5 and 7. The optimal number of topics for the highest coherence appeared at topic numbers 6 and 16. Therefore, considering both the minimum perplexity and the highest coherence, the optimal number of topics for R&D activity descriptions in 2018 was determined to be 6. Once the optimal number of topics was identified, the document-topic probability for 2018 was output. To further understand the key words within each topic, the LDA model was retrained and optimized with the selected number of topics, and the corresponding topics were displayed.

In the resulting topic vocabulary list, the top 6 weighted and frequent feature words for each topic were selected. The key words highlighted the focus of the documents on areas such as technology, finance, production, sales, and R&D. Words like “innovation,” “product,” and “technology” frequently appeared and were consistently ranked highly among the feature words related to R&D activity descriptions across the years.

After constructing the topic model and determining the optimal number of topics and high-frequency terms, the KL divergence between each topic and authoritative R&D innovation literature (Zheng & Chen, 2016) was calculated. The topic with the smallest KL divergence was chosen as the most relevant to R&D innovation. Finally, to obtain a comprehensive score for R&D activity descriptions in each document, the topic loadings for the selected most relevant R&D innovation theme were calculated and used as the final score for R&D activity text descriptions. To observe the overall trend of R&D descriptions over the years, the cumulative topic loadings for each year were summed and presented by year. The final trend is shown in Figure 2.

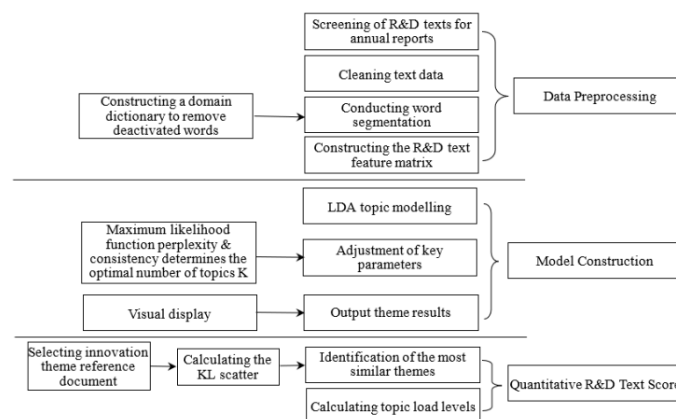


**Figure 2.** Trend of R&D activity text description intensity from 2012-2022

The trend in Figure 2 shows that over time, the intensity of R&D activity descriptions in the annual reports of Chinese listed companies has been steadily increasing. Notably, since the announcement of relevant R&D innovation strategies in 2018, there has been a significant rise in the intensity of these descriptions. This reflects the growing emphasis on R&D investment and innovation in the context of the new technological innovation era.

#### 4.2.3 PCA

The PCA process is designed as follows in Figure 3.



**Figure 3.** PCA process

(1) Data Preprocessing

This experiment primarily selects accounting data related to R&D from the financial reports of listed manufacturing companies in China for the years 2012 to 2022. The selected data includes five indicators: R&D expenditure, the ratio of R&D expenditure to main business income, the number of R&D personnel, the ratio of R&D personnel to total employees, and the number of patents applied for during the year. First, data with missing values, such as undisclosed R&D personnel, missing R&D expenditure ratios, or large numbers of missing patent application data, were removed. Second, the data for each indicator were standardized to eliminate the impact of differing units and scales, ensuring that each indicator was equally weighted in the analysis. Third, KMO and Bartlett tests were performed to evaluate and confirm that the data were suitable for PCA and other dimensionality reduction statistical methods. This step ensured the validity and accuracy of the PCA method. Thus, the data preprocessing phase was completed.

(2) Model Construction

After obtaining the standardized data, the covariance matrix and correlation coefficient matrix were constructed to compute their eigenvalues, which were then used to determine the direction and number of principal components.

First, the covariance matrix was constructed by calculating the covariance between the variables, which revealed and assessed the linear relationships among the five indicators mentioned above. Next, the covariance matrix was subjected to eigenvalue decomposition, yielding the eigenvalues and corresponding eigenvectors. The eigenvalues represented the variance of each principal component, while the eigenvectors defined the direction of the principal components. Third, the eigenvalues were sorted in descending order, and the number of principal components was determined using the “explained variance ratio” and the “scree plot.” Based on previous studies, this paper selected the minimum number of principal components whose cumulative contribution rate reached 80%. From the scree plot, two principal components were chosen before the eigenvalues dropped rapidly. This completed the model construction phase.

(3) Quantification of R&D Accounting Data Scores

First, a score matrix was constructed by multiplying the selected principal component eigenvectors with the standardized original data matrix, thereby obtaining the scores of each observation for each principal component. Then, by analyzing the factor loadings of the principal components and explaining the significance of each principal component, the scores were used as the quantified composite score of the R&D accounting data intensity. This was done by summing the scores for each fiscal year and averaging them, resulting in the quantified intensity level of R&D accounting data for Chinese manufacturing listed companies from 2012 to 2022. This allowed for a consistency comparison and difference analysis with the previously obtained R&D text investment intensity scores. Thus, the task of quantifying the R&D accounting data scores was completed.

On an annual basis, PCA experiments were conducted on the R&D-related accounting data in the annual reports of Chinese manufacturing listed companies from 2012 to 2022. The principal component numbers and specific scores were determined through covariance matrices, eigenvalues, and eigenvectors. The relevant results are shown in Table 2, Table 3, and Table 4.

**Table 2.** KMO and Bartlett’s test of sphericity

<b>KMO Sampling Adequacy</b>		<b>0.81</b>
Bartlett’s Sphericity Test	Chi-Square Value	2678.39
	Degrees of Freedom	319
	Significance	0.01

To test whether the selected raw data of the R&D activity-related accounting indicators are suitable for PCA, KMO and Bartlett’s sphericity tests were first conducted. According to the results in Table 2, the KMO value is greater than or equal to 0.5, and the Bartlett test shows a significance level (P) less than 0.05, both of which meet the application requirements for PCA, indicating that the results are scientifically reliable.

**Table 3.** Correlation matrix of indicators

	<b>RD1 (R&amp;D Personnel)</b>	<b>RD2 (R&amp;D Personnel Ratio)</b>	<b>RD3 (R&amp;D Investment)</b>	<b>RD4 (R&amp;D Investment Ratio)</b>	<b>RD5 (Patents)</b>
RD1 (R&D Personnel)	1.00	0.07	0.78	-0.01	0.53
RD2 (R&D Personnel Ratio)	0.07	1.00	-0.01	0.05	0.01
RD3 (R&D Investment)	0.78	-0.01	1.00	0.01	0.66
RD4 (R&D Investment Ratio)	-0.01	0.05	0.01	1.00	-0.01
RD5 (Patents)	0.53	0.01	0.66	-0.01	1.00

Based on the standardized data, a correlation matrix between the variables was calculated. According to the results in Table 4, the number of R&D personnel (RD1) and R&D investment (RD3) show a high positive correlation, and the number of patents applied for (RD5) also shows a moderate positive correlation with the above two indicators. Based on this, some highly correlated indicators can be merged. Meanwhile, there are no extremely high or low correlations in the overall correlation matrix, indicating that PCA dimensionality reduction can be performed.

**Table 4.** Eigenvalues and variance contribution of principal components

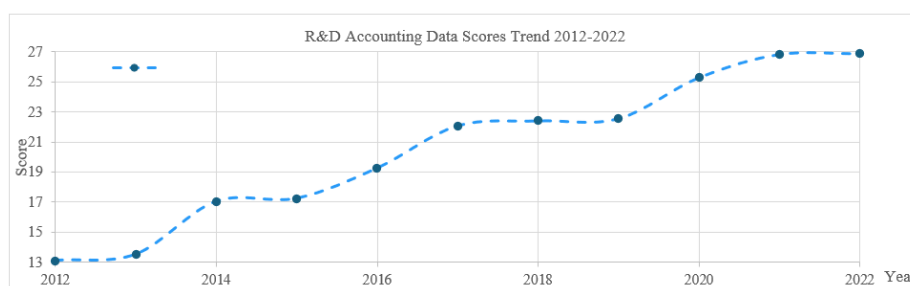
Principal Component	Eigenvalue	Variance Contribution (%)	Cumulative Variance Contribution (%)
Comp.1	2.32	46.35	46.35
Comp.2	1.06	21.13	67.48
Comp.3	0.95	18.96	86.44
Comp.4	0.48	9.67	96.11
Comp.5	0.19	3.89	100

By calculating the eigenvalues and eigenvectors, we can determine the variance and direction of each principal component. From the cumulative variance contribution, we can decide the minimum number of principal components required. Based on the results in Table 4, the variance contribution of Comp.1 is 46.35%, and the variance contribution of Comp.2 is 21.13%, with a cumulative variance contribution of over 65%. These two principal components capture most of the information in the original variables. Therefore, we choose two principal components to reflect the R&D activity intensity.

**Table 5.** Principal component decomposition calculation

Principal Component	Name	RD1 (R&D Personnel)	RD2 (R&D Personnel Ratio)	RD3 (R&D Investment)	RD4 (R&D Investment Ratio)	RD5 (Patents)
f1	Absolute R&D Capacity	0.58		0.61		0.54
f2	Relative R&D Capacity		0.72		0.69	

Next, principal component decomposition and factor loading calculations were performed, and the results are shown in Table 5. According to the results in the table, the first principal component represents absolute R&D capability:  $Comp1 = 0.58RD1 + 0.61RD3 + 0.538RD5$ ; the second principal component represents relative R&D capability:  $Comp2 = 0.718RD2 + 0.694RD4$ . Based on the contribution rates and weights of the principal components, the R&D strength score can be calculated as:  $NumRD = 0.463 / (0.463 + 0.211) * Comp1 + 0.211 / (0.463 + 0.211) * Comp2$ . The final score is obtained by summing and averaging the values for each accounting year, followed by normalization:  $RD\_num = (NumRD - MinNumRD) / (MaxNumRD - MinNumRD)$ . The resulting trend of the R&D activity accounting data strength score for manufacturing listed companies over the years is shown in Figure 4.



**Figure 4.** R&D activity accounting data intensity trend from 2012 to 2022

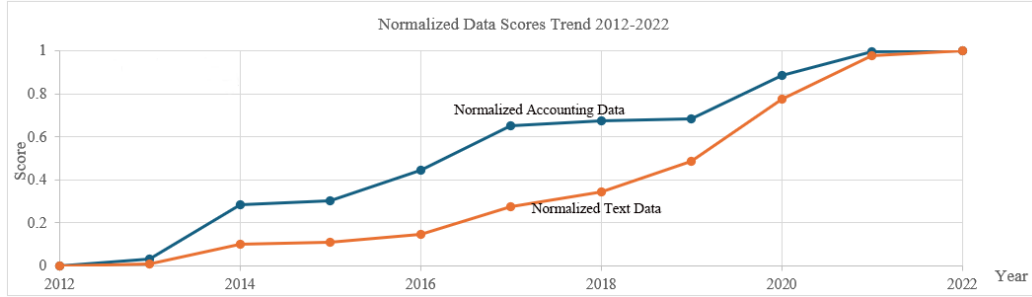
Figure 4 shows that the intensity of R&D activity accounting data remained stable overall, with a clear upward trend. Notably, after the revision of accounting standards in 2016, companies significantly increased their disclosure of R&D data in financial reports. This demonstrates the growing emphasis on detailed reporting of R&D expenditures in the financial statements of listed manufacturing companies.

#### 4.2.4 Inconsistency measurement and analysis

After obtaining the R&D activity text description intensity from the LDA model and the R&D activity

accounting data intensity from PCA, we perform normalization of both sets of scores to compare and measure their inconsistency.

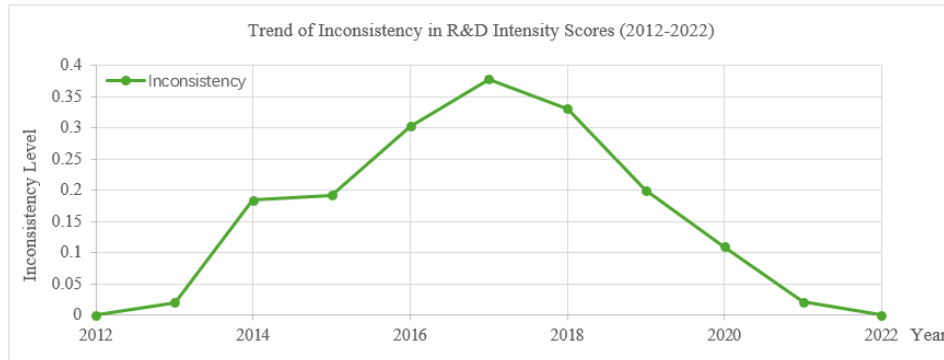
Since the scales of the R&D activity description text intensity and the R&D activity accounting data intensity differ, normalization is applied. The resulting normalized scores for both the R&D text description intensity and the accounting data intensity are shown in Figure 5.



**Figure 5.** Trends of R&D activity description accounting and text scores from 2012 to 2022

Figure 5 shows that both the accounting data and text information scores for R&D activity descriptions exhibit a steady upward trend year by year. The trends for both are generally similar, but there still exists some degree of difference and inconsistency between them. This supports the Hypothesis 1: “There is inconsistency between the accounting data and the text information description of R&D activities in the annual reports of manufacturing listed companies”.

Furthermore, to calculate and measure the specific degree and trend of inconsistency, the normalized scores of both data types were subtracted to obtain the final inconsistency difference between the accounting and text descriptions of R&D activities, as shown in Figure 6.



**Figure 6.** Inconsistency in R&D activity descriptions between accounting data and text

Figure 6 shows that the inconsistency between the disclosure of R&D accounting data and the intensity of textual descriptions follows a trend of rising first and then declining over time. The inconsistency peaked in 2017 at 0.37, after which it significantly decreased, reaching its lowest value of 0.02 in 2022. This supports Hypothesis 2: “The level of inconsistency in R&D activity disclosures in manufacturing companies’ annual reports increases initially and then significantly decreases over time.”

To quantify the differences, we used two metrics: Mean Absolute Deviation (MAD) and Mean Squared Error (MSE).

$$MAD = \frac{1}{N} \sum_{i=1}^N (Score\ difference_i) \quad (1)$$

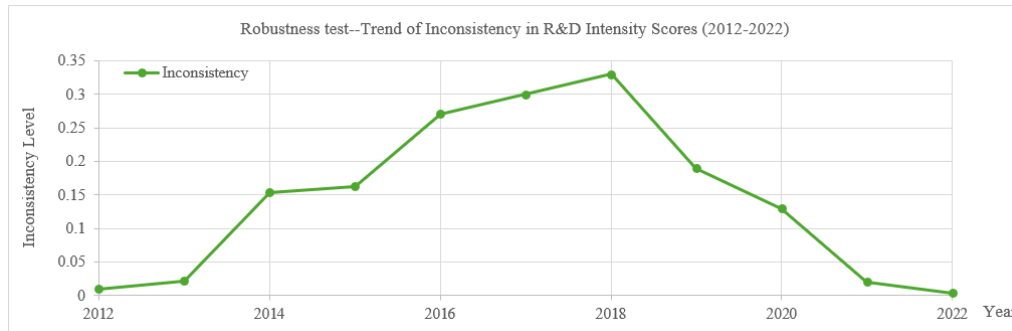
$$MSE = \frac{1}{N} \sum_{i=1}^N (Score\ difference_i)^2 \quad (2)$$

where,  $N$  is the number of companies in a given fiscal year.

The MAD and MSE values were found to be 0.08 and 0.01, respectively, indicating a certain level of inconsistency between the textual and accounting data descriptions of R&D activities. This further verifies the validity of Hypothesis 1 and Hypothesis 2.

#### 4.2.5 Robustness test

A robustness test was performed using the indicator replacement method, and the results of the inconsistency trend after variable replacement are shown in Figure 7.



**Figure 7.** Inconsistency trend in R&D activity descriptions after robustness check

First, in the text processing section, cosine similarity was used to replace KL divergence for calculating the similarity between the topics generated by the LDA model and the authoritative reference document topics. The R&D innovation theme IDs for each year obtained from this calculation were fully consistent with the previous experimental results. Additionally, the average of the total loadings of each document on this theme was calculated and used as the new quantification score for the strength of the R&D activity text, which was then used in subsequent computations and analysis. Next, in the accounting data processing section, the number of patents applied for was replaced with the maximum value from the number of patents authorized or granted, and PCA was performed again. The number of principal components and their respective weights did not show significant changes, which indicates the scientific validity and reliability of the selected indicators. Finally, the results after these substitutions were also normalized and compared by taking the difference between the scores. The MAD and MSE were calculated to observe the degree of inconsistency between the two.

Figure 7 shows that whether different similarity metrics or different accounting data indicators were used in the robustness test with variable replacement, the strength of R&D activity descriptions in both accounting data and text data still showed a continuous upward trend over time. There remained some degree of inconsistency between the two, with the level of inconsistency initially rising and then significantly decreasing, which is consistent with the previous experimental findings. This confirms the reliability and scientific validity of the experimental results and conclusions, enhancing their credibility.

## 5. Conclusion

This paper used LDA topic modeling, KL divergence, and cosine similarity calculations to examine the text-based intensity of R&D activities in the annual reports of manufacturing companies from 2012 to 2022. By calculating the similarity with authoritative R&D innovation reference documents, the most relevant R&D activity topics were identified. The loadings on these topics were then used to determine the intensity of R&D activities within companies, which was used to measure the inconsistency in R&D activity descriptions between text and accounting data in the annual reports. The empirical results show that: (1) The disclosure and intensity of R&D activity information in the text of Chinese manufacturing companies' annual reports are generally low, with little variation in overall levels, but over time, the text descriptions of R&D innovation show a continuous upward trend. (2) The disclosure and intensity of R&D activity data in the accounting section of Chinese manufacturing companies' annual reports are generally high, with significant variation in overall levels, and over time, companies show a continuous increase in their investment and disclosure of R&D innovation accounting indicators. (3) There is a common phenomenon of inconsistency in the description of R&D activities in the annual reports of Chinese manufacturing companies, but over time, the degree of inconsistency first increases and then significantly decreases.

The conclusions of this study enrich the research on the disclosure of R&D information by listed companies in China, particularly filling the gap in the measurement of inconsistency between financial and non-financial information disclosure. It provides some reference and practical insights for the future measurement of corporate R&D investment disclosures: (1) From the government level, create a transparent and fair disclosure environment for R&D activities, with clear and standardized description guidelines. (2) From the corporate level, enhance internal control and governance structures, continuously improve the importance of consistency in R&D activity disclosures. (3) From the third-party agency level, increase supervision of the standards for R&D information disclosure, and establish and provide the best industry standards and practice guidelines. (4) From the investor



level, deepen the ability to mine R&D activity information and strengthen the attention and understanding of the consistency of R&D investment descriptions.

Due to limitations in technical methods and empirical techniques, this study still has certain shortcomings and can be further explored and improved in the future in the following ways: (1) In this paper, only five relevant indicators from two categories—R&D investment and R&D outcomes—were considered, which cover the necessary core R&D indicators. However, other indicators, such as the relative changes in R&D funding over the years or non-material R&D investments like innovation training, were not considered, which may lead to omissions. Future research should expand the scope of indicators and search for more scientific and reasonable comprehensive accounting data indicators for R&D activities, thus establishing a more universal and comprehensive financial information measurement model. (2) This paper combines topic modeling with similarity calculation to measure the comprehensive score of R&D text. However, with the continuous iteration and development of deep learning, natural language processing, and other models and technologies, future studies could explore the use of pre-trained language models like BERT and FinBERT to extract complex structures and deep semantic features from text data, which would allow for a more comprehensive measurement of the text strength and quality of R&D activity descriptions in corporate annual reports. (3) This paper only selected the annual reports of Chinese manufacturing companies over the past 11 years as sample data, without considering specific classifications based on regions, company size, etc., thus the research conclusions have some limitations. Future studies could include data from more industries and further classify companies based on regional (East, Central, West) and operational scale factors. This would provide a deeper, more comprehensive, and macro-level exploration of the inconsistencies and their degree of change in R&D activity descriptions in corporate annual reports, thus laying a theoretical foundation and providing policy guidance for better regulation of R&D disclosure practices in the future.

### Data Availability

The data supporting our research results are included within the article or supplementary material.

### Conflicts of Interest

The authors declare no conflict of interest.

### References

- Ali, S. M., Deußer, T., Houben, S., Hillebrand, L., Metzler, T., & Sifa, R. (2023). Automatic consistency checking of table and text in financial documents. *Proc. of Northern Lights Deep Learn. Workshop, 4*, 1-9. <https://doi.org/10.7557/18.6816>.
- Atabuzzaman, M., Shajalal, M., Ahmed, M. E., Afjal, M. I., & Aono, M. (2021). Leveraging grammatical roles for measuring semantic similarity between texts. *IEEE Access, 9*, 62972-62983. <https://doi.org/10.1109/ACCESS.2021.3074747>.
- Bai, X., Dong, Y., & Hu, N. (2019). Financial report readability and stock return synchronicity. *Appl. Econ., 51*(4), 346-363. <https://doi.org/10.1080/00036846.2018.1495824>.
- Banerjee, P. (2022). Nature of financial constraints and R&D intensity. *Finance Res. Lett., 50*, 103184. <https://doi.org/10.1016/j.frl.2022.103184>.
- Bellstam, G., Bhagat, S., & Cookson, J. A. (2021). A text-based analysis of corporate innovation. *Manage. Sci., 67*(7), 4004-4031. <https://doi.org/10.1287/mnsc.2020.3682>.
- Bigne, E., Ruiz, C., Perez-Cabañero, C., & Cuenca, A. (2023). Are customer star ratings and sentiments aligned? A deep learning study of the customer service experience in tourism destinations. *Serv. Bus., 17*(1), 281-314. <https://doi.org/10.1007/s11628-023-00524-0>.
- Biswas, B., Sengupta, P., Kumar, A., Delen, D., & Gupta, S. (2022). A critical assessment of consumer reviews: A hybrid NLP-based methodology. *Decis. Support Syst., 159*, 113799. <https://doi.org/10.1016/j.dss.2022.113799>.
- Chang, K., Zeng, Y., Wang, W., & Wu, X. (2019). The effects of credit policy and financial constraints on tangible and research & development investment: Firm-level evidence from China's renewable energy industry. *Energy Policy, 130*, 438-447. <https://doi.org/10.1016/j.enpol.2019.04.005>.
- Chen, C., Gu, J., & Luo, R. (2022). Corporate innovation and R&D expenditure disclosures. *Technol. Forecasting Social Change, 174*, 121230. <https://doi.org/10.1016/j.techfore.2021.121230>.
- Chen, N., Han, C., & Wei, L. (2023). The impact of subjective and objective inconsistencies in scientific and technological innovation attributes on the listing of enterprises on the science and technology innovation board. *Procedia Comput. Sci., 221*, 493-500. <https://doi.org/10.1016/j.procs.2023.08.006>.
- Cheng, J., Huang, D., & Wu, Y. (2023). Managerial perception on competition and strategic R&D decisions. *Econ.*

- Modell.*, 123, 106281. <https://doi.org/10.1016/j.econmod.2023.106281>.
- Cheng, X. S., Wu, Q., Xiu, W. X., & Liu, Y. (2022). Enterprise R&D investment fluctuation and information disclosure: A study based on investors' innovation tolerance. *Econ. Res. J.*, 57(6), 191-208.
- Ciftci, M. & Zhou, N. (2016). Capitalizing R&D expenses versus disclosing intangible information. *Rev. Quant. Finance Accounting*, 46, 661-689. <https://doi.org/10.1007/s11156-014-0482-0>.
- Fedorova, E., Drogovoz, P., Popova, A., & Shiboldenkov, V. (2023). Impact of R&D, patents and innovations disclosure on market capitalization: Russian evidence. *Kybernetes*, 52(12), 6078-6106. <https://doi.org/10.1108/K-08-2021-0760>.
- Feng, C., Fay, S., & Kashmiri, S. (2022). The value relevance of descriptive R&D intensity. *J. Bus. Res.*, 139, 1394-1407. <https://doi.org/10.1016/j.jbusres.2021.10.033>.
- Gordon, E. A., Hsu, H. T., & Huang, H. (2020). Peer R&D disclosure and corporate innovation: Evidence from American depository receipt firms. *Adv. Accounting*, 49, 100471. <https://doi.org/10.1016/j.adiac.2020.100471>.
- Hazarika, B., Chen, K., & Razi, M. (2021). Are numeric ratings true representations of reviews? A study of inconsistency between reviews and ratings. *Int. J. Bus. Inf. Syst.*, 38(1), 85-106. <https://doi.org/10.1504/IJBIS.2021.118637>.
- Kabuye, F., Bugambiro, N., Akugizibwe, I., Nuwasiima, S., & Naigaga, S. (2019). The influence of tone at the top management level and internal audit quality on the effectiveness of risk management practices in the financial services sector. *Cogent Bus. Manage.*, 6(1), 1704609. <https://doi.org/10.1080/23311975.2019.1704609>.
- Koh, P. S. & Reeb, D. M. (2015). Missing R&D. *J. Accounting Econ.*, 60(1), 73-94. <https://doi.org/10.1016/j.jacceco.2015.03.004>.
- Lakhal, N. & Dedaj, B. (2020). R&D disclosures and earnings management: The moderating effects of IFRS and the global financial crisis. *J. Financial Reporting Accounting*, 18(1), 111-130. <https://doi.org/10.1108/JFRA-10-2019-0129>.
- Li, F., Deng, K., Fan, M., Teng, C., & Ji, D. H. (2024). Examining dialogue consistency based on chapter-level semantic graph. *Data Anal. Knowl. Discovery*, 8(5), 18-28. <https://doi.org/10.11925/infotech.2096-3467.2023.0431>.
- Li, X., Qiao, P., & Zhao, L. (2019). CEO media exposure, political connection and Chinese firms' stock price synchronicity. *Inter. Rev. Econ. Finance*, 63, 61-75. <https://doi.org/10.1016/j.iref.2018.08.006>.
- Li, Y. Q. & Yao, Y. (2020). R&D narrative disclosure: Is talk more really useless? – Textual analysis based on analyst forecast. *Accounting Res.*, 2, 26-42. <https://doi.org/10.3969/j.issn.1003-2886.2020.02.003>.
- Liu, Y. & Gao, Q. (2024). Navigate through massive and mixed online reviews: The role of central and peripheral cues in consumers' information search. *Inter. J. Hum. Comput. Interact.*, 1-22. <https://doi.org/10.1080/10447318.2024.2336701>.
- Liu, Z., Hu, K., & Hussain, A. (2023). R&D disclosure and corporate innovation: Mediating role of financing structure. *Finance Res. Lett.*, 56, 104106. <https://doi.org/10.1016/j.frl.2023.104106>.
- Liu, Z., Rao, P., Wang, X., & Yue, R. (2022). Peer innovation tone and corporate innovation adjustments: Evidence from China. *Asia Pac. J. Atmos. Sci.*, 51(6), 945-980. <https://doi.org/10.1111/ajfs.12407>.
- Mazzi, F., Slack, R., Tsalavoutas, I., & Tsigkas, F. (2019). Country-level corruption and accounting choice: Research & development capitalization under IFRS. *Br. Accounting Rev.*, 51(5), 100821. <https://doi.org/10.1016/j.bar.2019.02.003>.
- Shan, G., Zhang, D., Zhou, L., Suo, L., Lim, J., & Shi, C. (2018). Inconsistency investigation between online review content and ratings. In *Twenty-Fourth Americas Conference on Information Systems. AMCIS 2018 Proceedings*. New Orleans, USA. <https://aisel.aisnet.org/amcis2018/eBusiness/Presentations/23>
- Tsalavoutas, I., André, P., & Dionysiou, D. (2014). Worldwide application of IFRS 3, IAS 38 and IAS 36, related disclosures, and determinants of non-compliance. *ACCA Res. Rep.*, 134. [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2603572](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2603572)
- Wu, D. & Zhao, Q. (2024). IPO underpricing and corporate innovation: Evidence from China. *Econ. Innov. New Technol.*, 33(6), 866-888. <https://doi.org/10.1080/10438599.2023.2258355>.
- Wu, H., Guo, G., Yang, E., Luo, Y., Chu, Y., Jiang, L., & Wang, X. (2024). PESI: Personalized explanation recommendation with sentiment inconsistency between ratings and reviews. *Knowledge-Based Syst.*, 283, 111133. <https://doi.org/10.1016/j.knosys.2023.111133>.
- Xia, Y., Li, F., Liu, Q., Jin, L., Zhang, Z., Sun, X., & Shao, L. (2023). Reasonfuse: Reason path driven and global–Local fusion network for numerical table-text question answering. *Neurocomputing*, 516, 169-181. <https://doi.org/10.1016/j.neucom.2022.09.04>.
- Xu, K., Geng, C., Wei, X., & Jiang, H. (2020). Financing development, financing constraint and R&D investment of strategic emerging industries in China. *J. Bus. Econ. Manage.*, 21(4), 1010-1034. <https://doi.org/10.3846/jbem.2020.12727>.
- Xu, W. Z. & Zhu, Z. H. (2019). Will the text message in annual reports be used by analysts in their prediction? *Contemp. Finance Econ.*, 1, 131-141. <https://doi.org/10.13676/j.cnki.cn36-1030/f.2019.01.012>.

- Ye, Q., Cai, J., & Zhang, Q. (2024). Measuring corporate innovation based on LDA topic model. *Sci. Technol. Prog. Policy*, 41(2), 90-98. <https://doi.org/10.6049/kjbydc.2022040376>.
- Yu, Z. & Xiao, X. (2022). Innovation information disclosure and stock price crash risk-based supervision and insurance effect path analysis. *Aust. Econ. Pap.*, 61(3), 534-590. <https://doi.org/10.1111/1467-8454.12258>.
- Zeng, Q. S., Zhou, B., Zhang, C., & Chen, X. Y. (2018). Annual report tone and insider trading. *Manage. World*, 34(9), 143-160. <https://doi.org/10.19744/j.cnki.11-1235/f.2018.09.012>.
- Zhang, G. B. & Li, J. (2021). Detecting social media fake news with semantic consistency between multi-model contents. *Data Anal. Knowl. Discovery*, 5(5), 21-29. <https://doi.org/10.11925/infotech.2096-3467.2020.0884>.
- Zhou, M. S. & Lu, H. Z. (2021). Venture capital backed companies and innovation information: Words or actions. *J. Manage. Sci. Chin.*, 24(11), 78-97. <https://doi.org/10.19920/j.cnki.jmsc.2021.11.005>.
- Zhou, M., Xu, Y., & Xu, G. (2023). Can narrative innovation information disclosure boost R&D investment? Empirical evidence from a-share listed companies. *East Chin. Econ. Manage.*, 37(10), 104-115. <https://doi.org/10.19629/j.cnki.34-1014/f.221210002>.