https://doi.org/10.51408/issi2025_088

Exploring Novelty Differences between Industry and Academia: A Knowledge Entity-centric Perspective

Hongye Zhao[1], Yi Zhao[2], Chengzhi Zhang[3]

[1]zhaohongye_phd@njust.edu.cn, [2]yizhao93@njust.edu.cn, [3]zhangcz@njust.edu.cn
Department of Information Management, Nanjing University of Science and Technology,
Nanjing (China)

Abstract

Novel ideas drive innovation, and both academia and industry possess distinct strengths in advancing technological progress. The industrial sector, on the one hand, seeks to privatize knowledge to maintain appropriability, while on the other hand, it actively promotes open-sourcing of models and platform sharing. This paradox raises the question of whether industrial disclosures are less novel compared to those from academia. Some studies argue that academia tends to generate more novel ideas, while others suggest that industry researchers are more likely to drive new breakthroughs. Previous studies have been limited by data sources and inconsistent measures of novelty. To address these gaps, this study establishes a unified framework for calculating the novelty of papers and patent data in the field of Natural Language Processing (NLP), focusing on fine-grained knowledge entities. Additionally, a regression model is constructed to analyse the relationship between the type of institution and the novelty of their publications. The results show that academia demonstrates higher novelty in both patent and paper outputs. Notably, academic involvement significantly enhances the novelty of industrial patents. Furthermore, this study examines how team size impacts novelty in patents and papers, providing strategic recommendations for forming research teams. We release our data and associated codes at https://github.com/tinierZhao/entity_novelty.

Introduction

Academic research focuses on theoretical inquiry and the advancement of fundamental lence, aiming to expand human knowledge and drive disciplinary progress (Sauermann & Stephan, 2010). In contrast, the industrial sector emphasizes core competitiveness (Geisler, 1995), prioritizing economic returns and often safeguarding intellectual appropriability by restricting the disclosure of research outcomes (Arundel, 2001; Chirico et al., 2018).

Following this logic, the industry would typically choose to limit the disclosure of novel research outcomes to safeguard its competitive advantage. However, this traditional notion is being challenged in the field of artificial intelligence (AI), as the industry demonstrates a noticeably more open attitude. For example, leading tech companies have released cutting-edge technologies in algorithms and models, such as the BERT model (Devlin et al., 2019) and various other open-source large language models. Additionally, they have significantly lowered the barriers to adopt artificial intelligence technologies by offering application programming interfaces (APIs) and detailed technical documentation. This enables users to easily integrate these models into their own projects and supports further development and customization. Moreover, the industrial sector’s active participation in most active and popular AI conferences has spurred numerous disruptive innovations (Liang et al., 2024). While this openness may partially diminish the appropriability of knowledge, it offers substantial benefits. On the one hand, the public release of frontier research attracts a broader developer community, reducing long-term maintenance and development costs while generating economic returns through technology services (Homscheid et al., 2015). On the other hand, collaborations with prominent enterprises and academic institutions allow the industrial sector to access external knowledge, thereby maintaining its technological leadership and fostering product iteration and optimization through knowledge spillovers (Jiang et al., 2024), which in turn serves to broaden its market share (Hu et al., 2023; Tao et al., 2022). In this context, it remains uncertain whether the research outcomes from industry exhibit lower novelty compared to those from academia.

Evaluating the novelty of scientific and technical literature presents inherent challenges. Publications that introduce revolutionary technologies and lay the foundation for subsequent studies are rare (Arts et al., 2019; Arts et al., 2021). These works often go underappreciated initially, as they challenge existing conventions and may encounter resistance during the review process (Riera & Rodríguez, 2022). In contrast, studies that align with established theories are more likely to gain peer trust (Liang et al., 2022), putting highly novel research at a disadvantage in peer review (Koppman & Leahey, 2019; Wang et al., 2017). Even after publication, such research often faces delays in gaining recognition (Wang et al., 2017). Meanwhile, the growing volume of scientific and technical literature across disciplines has significantly increased the workload for reviewers (Shibayama et al., 2021).

In this context, the novelty of industrial disclosures compared to academic ones remains a topic of ongoing debate. As a key branch of artificial intelligence, NLP continues to experience rapid growth, with significant breakthroughs emerging from both academia and industry, despite a general slowdown in innovation across many fields (Park et al., 2023). Some scholars argue that academia contributes more novel ideas, while industry tends to adopt and refine academic advancements (Bikard & Marx, 2019). Subsequent studies further confirm academia's leadership in NLP innovation (Chen et al., 2024; Liang et al., 2024). However, Dwivedi et al. (2019) suggest that industry researchers are more likely to drive new AI technologies. The rise of pre-trained models such as Transformer (Vaswani et al., 2017) and GPT (Radford et al., 2018), along with the rapid development of large-scale language models like ChatGPT, Ahmed et al. (2023) highlights industry's dominance in computational resources, data, and talent.

To date, studies on the differences in the novelty of publications between academia and industry in NLP have primarily focused on papers. The limitation is not due to the availability of data. Instead, it occurs because the approaches for evaluating novelty vary considerably between patents and scientific papers. For scientific papers, novelty is typically measured through journal citation pair analysis. However, patents primarily cite other patents rather than academic papers (Ba et al., 2024), and they do not correspond to journal types. Therefore, the novelty of patents cannot be directly measured using citation journal pairs. Moreover, the classification codes commonly used in patents cannot be aligned with those used in scientific papers. As a result, previous studies have not fully incorporated patent data, leaving a gap in understanding the specific relationship between institutional types and the novelty of scientific and technical literature.

This study addresses the gap by using a unified novelty evaluation framework that leverages fine-grained knowledge entities to assess the novelty of publications across academia, industry, and their collaborations in NLP. We calculate the semantic distances between fine-grained knowledge entities and assess the difficulty of different entity combinations. Unlike previous studies, this research selects specific entity types based on the characteristics of the NLP field, reducing interference from certain types and enhancing the reliability of novelty assessments. While focused on NLP, the methodology is applicable to other domains, particularly those involving scientific papers and patents outcomes. It offers a general analytical framework for comparing novelty across academia and industry and evaluating the effectiveness cross-sector collaboration.

Specifically, we address the following two research questions:

RQ1: How to unify the novelty calculation method based on fine-grained knowledge entities for both papers and patents?

RQ2: Is there a difference in the novelty of scientific and technical literature between industry and academia?

The contributions of this paper are as follows:

First, we extend the entity-based novelty measurement method to the patent domain. By transferring the entity recognition model from papers to patents, we apply the same novelty measurement to both, enabling a unified assessment and supporting future data source expansion.

Second, our analysis confirms that academic outputs in the NLP field exhibit higher novelty. Additionally, our findings indicate that patents generated through collaboration between industry and academia exhibit a significant increase in novelty, highlighting the potential impact of cross-sector collaboration on innovation.

The code and data used in this study are open-sourced on GitHub and can be accessed via the following website: https://github.com/tinierZhao/entity_novelty

Related work

For the research questions proposed in this paper, we conducted a review of the scientific and technical literature on novelty measures, as well as the factors influencing novelty.

Novelty measures in the scientific and technical literature

The measurement of novelty not only helps to identify valuable innovations in advance, but also provides key insights for technological transfer and innovation. Currently, novelty is primarily measured through combinations, as Nelson and Winter (1982) argued, "the creation of novelty mainly involves the recombination of existing conceptual and physical materials." Traditional methods for measuring novelty include the use of journal pairs and classification code pairs to assess the novelty of literature. With the availability of large-scale data and the advancement of machine learning and natural language processing technologies, novelty measurement methods have been continuously innovated. The combination of other types of knowledge elements has gradually become an important approach for assessing novelty. Additionally, some studies have explored new avenues by treating novelty as a binary classification task, using classification or outlier detection methods to distinguish between novel and non-novel literature.

From a combination-based view, early methods primarily focused on citation references and classification codes. Uzzi et al. (2013) compared the observed and Monte Carlo-simulated frequencies of journal pairs to calculate z-score for each pair, using the lowest 10th percentile z score to indicate a paper's novelty and the median z score to indicate its conventionality. Lee et al. (2015) improved Uzzi's method in terms of computational difficulty by adopting a multi-year time window, which reduced the previous single-year window and calculated the commonness of citation pairs. Wang et al. (2017) measured novelty through the first-time combination of different citation journal pairs in a paper. Specifically, they constructed a co-citation matrix for the journals and used cosine similarity between the vectors of each journal to assess the difficulty of combining the journal pairs. However, while these methods are easy to understand and explain, they also face limitations such as self-citation and biased citing (MacRoberts & MacRoberts, 1996; Jeon et al., 2023; Anne, 2023). Additionally, as the number of papers analysed increases, costs and computational efficiency escalate sharply.

Regarding patent novelty measurement, early traditional methods focused on patent classification codes and backward citations (Ahuja & Lampert, 2001; Lee & Lee, 2019). However, citations merely describe existing technologies and fail to reflect the technology of the patent itself, often presenting incomplete and biased representations (Kuhn et al., 2020; Arts et al., 2021). Measuring technological novelty through patent IPC codes (Fleming, 2001) is overly broad and tends to capture interdisciplinarity rather than technological uncertainty.

With the continuous development of NLP technologies, tasks such as scientific terminology extraction (entities, keywords) and semantic embedding have matured, making the measurement of novelty based on scientific text content a more reasonable and effective approach. Liu et al. (2022) used the BioBERT model to calculate the semantics of biological entities, determining entity pair novelty based on semantic similarity. The novelty score for each paper is calculated as the proportion of novel entity pairs to the total possible entity pairs. Similarly, Chen et al. (2024) applied an entity similarity-based approach using S to evaluate the novelty of conference papers in the field of natural language processing. Luo et al. (2022) employed BERT word embeddings to measure novelty by assessing the novelty of research questions, methods, and their combinations. Arts et al. (2021) extracted keywords from patent titles and abstracts, calculating “new_ngram” and corresponding “new_ngram_reuse” to measure patent novelty. Wei et al. (2024) used the BERT model to extract innovative sentences from patent claims and distilled them into knowledge element triples, measuring novelty scores for the triples by projecting entities and relations into a common space, using a combination of word2vec and HGT.

Table 1. Related works of novelty measurement.

Author	*Domain and data*	*Method*
Uzzi et al. (2013)	17.9 million papers spanning all scientific fields	Monte Carlo + Journal pairs combinations
Wang et al. (2017)	785,324 Articles in 251 subject	Co-citation matrix + Journal pairs combinations
Liu et al. (2022)	98,981 coronavirus papers	BioBERT + Knowledge entities combinations
Chen et al. (2024)	14,812 ACL Anthology papers	SciBERT + Knowledge entities combinations
Wei et al. (2024)	1343 agricultural robots patents	BERT + Knowledge triples combinations
Luo et al. (2022)	204,224 papers in ACM database	BERT + Questions-Methods combinations
Arts et al. (2021)	1,302,956 patents spanning all fields	SnowBall + New_ngram combinations
Jeon et al. (2022)	1,877 medical image patents	Doc2Vec + Outlier detection binary classification
Jeon et al. (2023)	15,653 biomedical papers	FastText + Outlier detection binary classification
Zanella et al. (2021)	13,393 blockchain-related patents	Word2Vec + Outlier detection binary classification
Jang et al. (2023)	25,183 pairwise vehicle communication networks patents	RoBERTa + Explainable AI binary classification

From the perspective of binary classification. Jang et al. (2023) treated patent novelty as a classification task, using RoBERTa for semantic embedding of patent claims to develop a self-explainable novelty classification model. Jeon et al. (2022) embedded patent claims and used the local outlier factor (LOF) algorithm to calculate patent novelty. Their study showed that, although ELMo and BERT provide high-quality patent embedding vectors, they are less suitable for modeling the technological features of patents, particularly in single technical domains, compared to Doc2Vec. Jeon et al. (2023) trained a fastText model using paper titles in the biomedical field and applied the LOF algorithm to measure the novelty score of each paper. Zanella et al. (2021) combined cosine similarity and density-based anomaly detection to improve the identification of outliers within patent clusters. A detailed summary of the above works, including their data, methods, is provided in Table 1.

From the above-mentioned studies, the methods for measuring novelty have evolved from the early approaches relying on citation and classification codes to those based on text content analysis. Moreover, no unified framework yet exists for calculating the novelty of patents compared to scientific papers. In the following chapters, we will provide a detailed explanation of how to uniformly extract fine-grained knowledge entities from patents and papers, and how to calculate the novelty based on fine-grained knowledge entities.

Factors influencing the novelty of scientific and technical literature

Previous studies have explored the relationship between novelty from various perspectives, including institutional nature, team size, and author attributes within teams.

Regarding team size, existing research presents inconsistent findings. Uzzi et al. (2013) found that research teams are more likely to introduce novel combinations within familiar knowledge domains compared to single-author papers. Lee et al. (2015) identified an inverted U-shaped relationship between team size and novelty, with this effect largely driven by the interplay between team size and knowledge diversity. Wang et al. (2019) suggested that smaller teams are more likely to disrupt science and technology with new ideas, while larger teams tend to focus on existing ones. Shin et al. (2022), using Web of Science data, found that scientific collaboration negatively affects novelty, as collaborative research tends to remain within established fields. However, Wu et al. (2024) argued that collaboration fosters trust and problem-solving abilities, and that knowledge diversity enhances knowledge transfer and promotes the impact of science on technology. Conversely, some studies indicate that excessive team heterogeneity may reduce trust, hinder knowledge sharing, and obstruct innovation (Chen et al., 2015).

At the institutional level, academia tends to lead industry in terms of novelty at the paper level, generating more exploratory ideas, while industry is more likely to produce high-impact papers (Liang et al., 2024). Chen et al. (2024) measured the novelty in the NLP field, finding that academia and collaborative institutions tend to be more novel than industry, based on fine-grained combinations of knowledge entities. Other studies suggest that papers involving companies have a higher impact, and collaborations between industry and academia exhibit greater novelty (Jee & Sohn, 2023).

At the author attribute level within teams, teams with diversified expertise tend to produce more original work and have a long-term advantage in terms of impact (Zheng, Li, & Wang, 2022). Mori and Sakaguchi (2018) examined how differentiated knowledge among inventors enhances patent novelty using Japanese patents. Gender diversity within teams has also become a favored topic in recent years. Teams with gender diversity produce papers with higher novelty and greater impact compared to single-gender teams (Yang et al., 2022). Liu et al. (2024) explored the relationship between novelty and gender heterogeneity in doctoral theses, finding that female authors had lower average novelty scores than male authors, and male advisors were more likely to supervise students who produced theses with higher novelty. Notably, this gender difference was more pronounced in lower-prestige universities. Similarly, Chan and Torgler (2020) found that among elite scientists, female scientists tend to receive more citations than their male counterparts.

In this study, we explore the performance of different institutional types in terms of novelty in patents and papers, with a particular focus on comparing the relationship between novelty and team size, to uncover both consistencies and differences.

Methodology

This study aims to quantify the impact of different team compositions on the novelty of scientific and technical literature. The research framework in Figure 1 outlines three key steps:

First, dataset construction. We constructed an original dataset that includes scientific and technical literature in the NLP field, comprising papers and patents published between 2000 and 2022, and extracted author information and their affiliated institutions for each document.

Second, novelty assessment of scientific and technical literature. Fine-grained knowledge entities were extracted from both scientific papers and patents, with the knowledge from scientific papers being transferred to patents. To achieve this, we first employed an entity recognition model trained on scientific papers to perform preliminary entity extraction from patent texts. Subsequently, we conducted manual reviews and added annotations for tool-specific terms (such as software platforms) that are unique to patent texts. This iterative process continued until the model's performance converged. The difficulty of their combinations was measured based on the semantic distances between these entities (Liu et al., 2022; Chen et al., 2024). This approach was then used to assess the novelty of each document. Lastly, regression analysis. A regression model was employed to conduct statistical tests on the novelty of scientific and technical literature from different institutions (Chen et al., 2024). Additionally, we treated the top 10% of papers and patents each year as high-novelty documents and performed a robustness check of our results using binary logistic regression (Jeon et al. 2022).

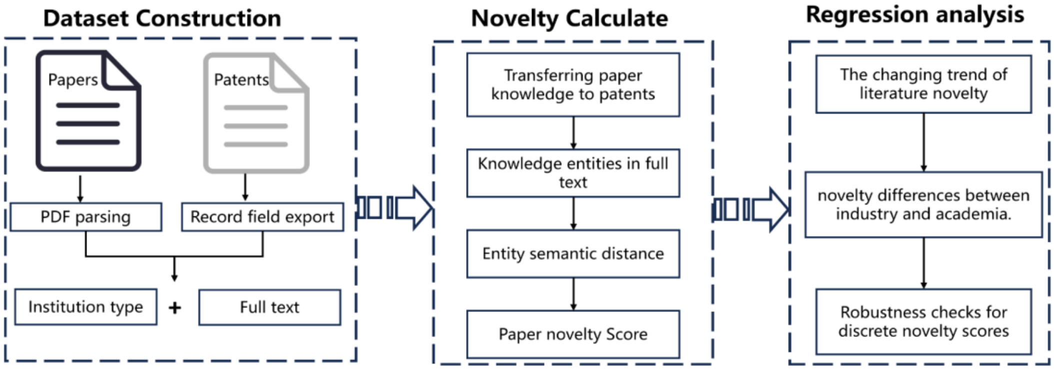
Dataset Construction
Papers
Patents
PDF parsing
Record field export
Institution type
+
Full text
Novelty Calculate
Transferring paper knowledge to patents
Knowledge entities in full text
Entity semantic distance
Paper novelty Score
Regression analysis
The changing trend of literature novelty
novelty differences between industry and academia.
Robustness checks for discrete novelty scores

Figure 1. Framework of this study.

Data collection

The paper data was collected from the ACL Anthology[1] website. We selected three representative conferences for our study: ACL (Annual Meeting of the Association for Computational Linguistics), EMNLP (Conference on Empirical Methods in Natural Language Processing), and NAACL (North American Chapter of the Association for Computational Linguistics). A total of 17,783 full-text papers from 2000 to 2022, were collected.

The patent data was collected from the United States Patent and Trademark Office (USPTO) through the patsnap[2] system. We conducted a search for patents within the time frame of 2000 to 2022, using the following query: CPC_GROUP: (G06F40[3]) AND APD: [20000101 TO 20221231] AND COUNTRY: (“US”). We focused on invention patents and filtered out those with legal statuses such as withdrawal, rejection, abandonment, application termination, or complete invalidation. Additionally, patents with the same priority were consolidated into families. Ultimately, a total of 25,305 patents were obtained.

Identification of the publishing institution type of the literature

By parsing the full-text PDFs and integrating data from the GRID and OpenAlex databases, we identified the authors and their institutions for 17,783 papers. For institutions not found through the search, we manually supplemented the data. Following Chen et al. (2024) and Xu et al. (2022), we categorized the institutions. In cases of multiple affiliations, we adopted the method of Hottenrott et al. (2021), considering the first-listed institution as the author's primary affiliation.

The patent data processing begins with extracting standardized applicant information from databases, where all non-personal names are presented in either Chinese or English. An edit distance algorithm, combined with a local dictionary, is then applied to normalize institutional names. Based on lexical features, two sets of keywords were defined: one for academic institutions and one for industrial organizations, covering both English and Chinese terms. The algorithm classifies institutions containing education-related terms (e.g., "edu," "univer") as academic, and those with company-related terms (e.g., "inc," "ltd," "lp") as industrial. This method ensures efficiency and accuracy, as the database provides standardized applicant fields. For unrecognized institutions, spacy[4] named entity recognition is used to determine whether the applicant is individual. For individual applicants appearing more than twice, we validate with ChatGPT to check for missed categorizations. Finally, the results are manually reviewed to correct and supplement the algorithm's output.

Specially, a paper is classified as "Academia" if all its authors are affiliated with academic institutions (such as universities or research institutes), as "Industry" if all authors are affiliated with industry institutions (such as companies or

[1] https://aclanthology.org/

[2] https://www.patsnap.com/

[3] CPC: G06F40, Handling natural language data

[4] https://pypi.org/project/spacy/

corporations), and as "Cooperation" if it involves authors from both academia and industry.

The specific institutional distribution for papers and patents is shown in Table 2.

Table 2. The institutional distribution of scientific and technical literature.

Institution Types	*Count*	*Ratio (%)*	*Count*	*Ratio (%)*
	Paper		Patent	
Academia	11,670	65.62	468	1.85
Industry	1,679	9.44	21732	85.97
Cooperation	4,315	24.26	69	0.27
Individual	0	0	2932	11.59
Other	119	0.67	104	0.41

Extraction of fine-grained knowledge entities from papers and patents

We adopt a combinatory perspective to assess the novelty of scientific and technical literature. Specifically, we analyze this based on the characteristics of the NLP field. NLP is a research domain centered around methods and data, with most studies typically involving the following key elements: 1) dataset construction or selection, often involving text resources such as corpora and dictionaries, which serve as the foundation for model training and validation; 2) method selection and application, which defines the strategies and steps for solving problems; 3) the choice of evaluation metrics, used to measure model performance and task quality; 4) the use of tools, including programming languages, software, and open-source tools required for implementing and testing NLP methods (Zhang et al., 2024; Pramanick et al., 2024). Based on this framework, we extract fine-grained knowledge entities from each patent and paper, covering the categories of Method, Tool, Metric, and Dataset.

In the fine-grained knowledge entity recognition task, we used the pre-trained SciBERT model. Due to differences in writing style and text structure between patents and papers, we trained separate entity recognition models for each type of document. Specifically, for papers, we adopted the framework proposed by Zhang et al. (2024). For patents, we initially applied a pre-trained model to annotate the patent sections, followed by re-annotation of the extracted entities according to the labelling rules. Additionally, for unique entities in patent texts, such as Storage medium, we performed extra annotation. After several rounds of iteration and adjustments, we obtained the patent entity recognition model (SciBERT + CRF), which achieved the following performance: Precision of 78.83%, Recall of 82.51%, and F1 score of 80.63%. Given that extracting entities only from titles and abstracts would miss many, we performed full-text extraction for both patents and papers. Paper data were extracted from PDFs, and the patent database was also exported in full text. For entity normalization, we used edit distance and semantic distance to cluster entities. Ultimately, we identified 22,871 entities in the papers and 9,523 entities in the patents.

Table 3. Top 5 entities in four types extracted from papers and patents.

Type	*Paper*		*Patent*	
	Entity	*Frequency*	*Entity*	*Frequency*
Method	BERT	4159	Neural network	3021
	Transformer	3844	Machine learning	1608
	N-gram	3733	N-gram	1365
	LSTM	3607	Language models	1160
	Attention Mechanism	3425	Deep learning	960
Tool	Pytorch	730	Computer system	11646
	MOSES	647	Storage medium	10413
	GIZA++	581	User interface	9323
	Python	430	Computer program	8738
	NLTK	333	Operating system	7636
Dataset	Wikipedia	3534	Emoji	306
	WordNet	2661	Email	122
	Twitter	1324	Soial meida	86
	Wall Street Journal	1005	World wide web	67
	Amazon Mechanical Turk	883	Twitter	43
Metric	Accuracy	10784	Accuracy	5278
	F_1	7802	Confidence	2500
	Precision	6024	Efficiency	2195
	Recall	5551	Relevance	1612
	Confidence	3832	Error	1453

Table 3 presents the top 5 entities in each category for both patents and papers. Due to the fact that patents are rarely evaluated on public datasets, the proportion of Dataset entities in patents is quite low, and as a result, the recognition performance for these entities is somewhat weaker. Additionally, a distinctive feature of patent terminology is its level of abstraction, particularly evident in the claims section. Unlike general discourse, which relies on precise wording to accurately convey content and avoid vague or overly broad terms, patent claims intentionally use generalized vocabulary (Codina-Filbà et al., 2016). This strategy enables companies to broaden the scope of their intellectual property protection, ensuring more extensive exclusivity over their innovations (Arinas, 2012; Ashtor, 2021). Furthermore, descriptions of Tool entities in patents tend to be more generalized, reflecting this situation.

Measurement of scientific and technical literature novelty

We explore the novelty of entity combinations through an analysis of the fine-grained knowledge entities extracted from scientific and technical literature. We draw on the work of Liu et al. (2024) in the field of scientific novelty assessment for biomedical papers. They treated biological entities as core elements of the research method and used the pre-trained Bio-BERT model to quantify the semantic distance between these entities to measure novelty. We applied this approach to evaluate the novelty of papers and patents in NLP, using pre-trained SciBERT to calculate the semantic similarity of entities for novelty measurement. Specifically, we extracted embeddings for each entity word from the “last_hidden_state”, removing [CLS] and [SEP] tokens. If an entity tokenizer contains multiple subwords, we averaged their embeddings. Cosine similarity was then used to calculate their semantic similarity. We labeled the top 10% of entities with the highest semantic distance as high-novelty entities. Finally, we analysed the frequency of these high-novelty entities in the text and measured the novelty of each paper based on their proportion in all entity combinations.

Furthermore, in domain-specific entity analysis, the ubiquity of certain entity types can cause inconsistencies between semantic distance and the actual difficulty of combining entities. For example, entities in the Metric category (such as accuracy, precision, recall, F_1 score, etc.) are often highly generic and strongly associated with most methods, but their semantic distance may not accurately reflect the actual situation. Due to their widespread use, these entities contribute little to novelty measurement and may even introduce noise. Therefore, we excluded entities of the Metric category from our analysis. For an entity pair $\left(e_i, e_j\right)$, the distance between the two is denoted as *D*, and $cosine\left(e_i, e_j\right)$ represents the semantic similarity between the entities. As shown in Equation (1):

$$D\left(e_i, e_j\right) = 1 - cosine\left(e_i, e_j\right) \tag{1}$$

Regression model for novelty comparison

To investigate the differences in novelty across various institutions, this study employs regression analysis to quantify and compare the novelty demonstrated in the scientific and technical literature produced by different institutions. The following sections provide a detailed description of the process of variable selection and the construction of the regression model.

Dependent variables: In the setting of independent variables, we first use the continuous novelty indicator (Novelty Score) calculated in the previous section for analysis. This indicator measures the proportion of novelty entity combinations in each paper or patent, with a score range from 0 to 1, where a higher score indicates greater novelty. Meanwhile, considering the uncertainty of novelty outcomes, we categorize the top 10% of papers and patents ranked by score each year as high novelty and construct a binary classification variable (Novelty Score 10%) for robustness checks.

Independent variables: This study defines the independent variables as the type of institution. After excluding institutions categorized as "other" and "individual", the remaining institutions are classified into three categories: academia, cooperation, and industry. Specifically, two binary variables—Academia and Cooperation, are defined. The Academia variable is set to 1 if the literature belongs to an academic institution, and the Cooperation variable is set to 1 for literature from cooperative institutions, with both variables set to 0 for literature from industry.

Control variables: In addition, the study considers several control variables to account for team characteristics. Specifically, it first considers the number of institutions (Institutions num), followed by the number of authors for papers and inventors for patents (Au/In num), in order to isolate the pure effect of institution type on the novelty of papers and patents. For patents, we also include the size of the patent family (Family size), which is commonly associated with welfare value and technological impact (Kabore & Park, 2019; Wu et al., 2015). Furthermore, the number of IPC classification codes at the subgroup level (IPC num) is controlled to account for the diversity of the patent's knowledge components (Sun et al., 2022). Finally, we include year as a dummy variable, using the publication year for papers and the application year for patents, to control for potential year-related differences that could affect the results. The summary statistics of the variables and the correlation coefficients between the variables are presented in Table 4. and Figure 2, respectively.

We found a strong correlation between the continuous and discrete forms of the dependent variable (novelty), while the correlations between the independent and dependent variables were weak. We then calculated the variance inflation factors (VIFs) for all explanatory variables to assess multicollinearity. The VIF for papers was 2.79 and for patents was 1.07, both below the threshold of 5 (Marcoulides & Raykov, 2019). These results indicate that multicollinearity has minimal impact on our model, ensuring the reliability of the estimates.

Table 4. Summary statistics of variables for regression analysis (N = 22,269 patents, N = 17,664 papers).

Variable	*Mean*	*Std. Dev.*	*Min*	*Max*	*Mean*	*Std. Dev.*	*Min*	*Max*
	Paper				Patent			
Novelty Score	0.10	0.07	0	0.51	0.11	0.09	0	0.75
Novelty Score 10%	0.10	0.30	0	1	0.10	0.30	0	1
IPC num	-	-	-	-	1.94	1.00	1	10
Family size	-	-	-	-	2.10	1.94	1	82
Au/In num	3.76	2.22	1	77	3.28	2.14	1	26
Institutions num	1.80	1.22	1	44	1.05	0.43	1	15
Academia	0.66	0.47	0	1	0.02	0.14	0	1
Cooperation	0.24	0.43	0	1	0.00	0.06	0	1

Note: The papers do not include IPC numbers or Family size, which are represented as '-'.

Regression analyses: Multivariable regression was conducted to examine how different types of institutions influence the novelty scores of the literature. As shown in Equation (2):

$$Novel_i = \alpha + \beta_1 Academia_i + \beta_2 Cooperation_i + Controls + Y_i + \varepsilon \qquad (2)$$

Where $Novel_i$ represents the novelty score of each literature i. The independent variables $Academia_i$ and $Cooperation_i$ indicating whether the literature is from an academic or cooperative institution, respectively. The variable Controls includes a set of control variables, Y_i denotes the publication year, and ε represents the error term in the model.

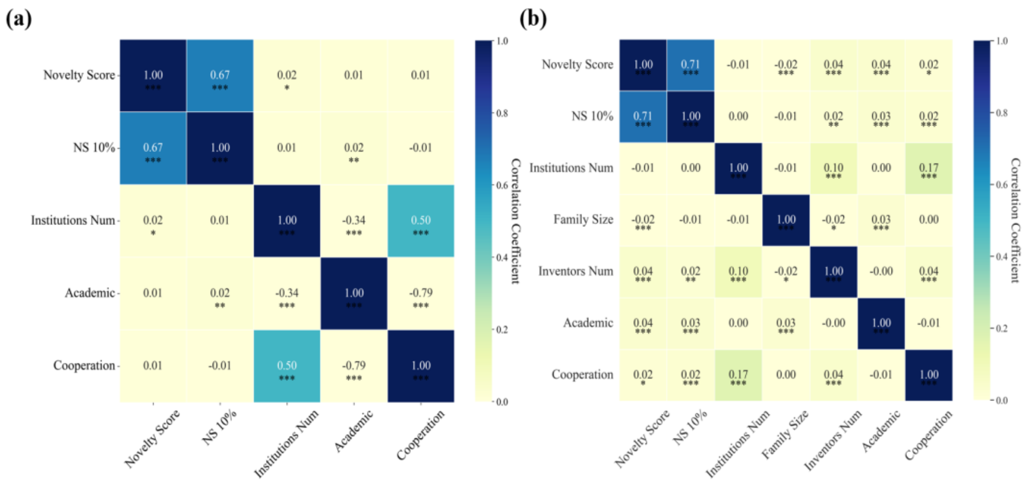

Figure 2. Pearson's rank correlation coefficient matrix (a) Correlation between variables in papers (b) Correlation between variables in patents.

Results

This study analyses papers published between 2000 and 2022 in three major NLP conferences and patents filed with the USPTO, focusing on the novelty differences across three types of publishing institutions: academia, industry, and collaboration. Our research not only compares the performance of different institution types in terms of novelty in literature, but also investigates the relationship between team size and novelty. The aim is to reveal how team size influences innovation across different types of scientific and technical literature.

Trends in publication volume of papers and patents

The field of NLP has experienced rapid growth, with a steady annual increase in patents and papers since 2000. The slight decrease in patent numbers in 2022 compared to 2021 is due to the America Invents Act (AIA), Section 35 U.S.C. § 122(b), which requires patents to be published 18 months after the earliest filing date, unless the applicant requests early publication. As of the retrieval date, some 2022 patents had not yet been published, which is common.

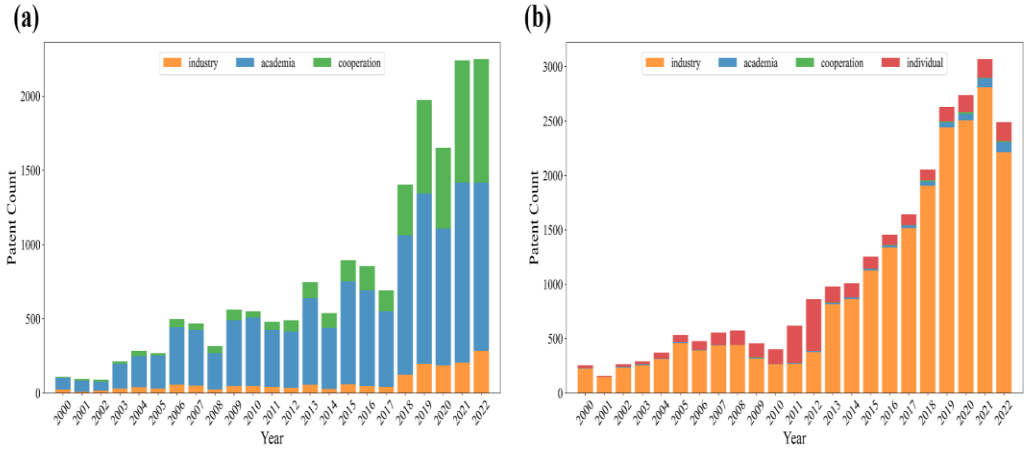

Figure 3. Annual publication volume of papers and patents. (a) Annual publication volume of papers (b) Annual publication volume of patents.

In addition, the distribution of patent numbers across institutions is more uneven compared to papers, with specific proportions detailed in the previous section on institutional distribution. Despite the concentration of the world's top higher education resources in the United States and the majority of government research funding directed towards universities, university-originated patents account for less than 4% of the total national patents, with corporate patents dominating the majority, followed by individual applications[5]. This phenomenon highlights the dominant role of industry in NLP patent filings. The annual publication volume of papers and patents is shown in Figure 3.

Trends in novelty changes of literature measured under a unified framework

In this section, we address RQ1. We first use the entity recognition models discussed in previous chapters to extract fine-grained knowledge entities from each paper and patent. Then, we leverage the pre-trained SciBERT model to obtain semantic vectors for the entities in both patents and papers. Next, we calculate the semantic distance between the entities to assess their novelty. The distribution of entity semantic distance-based novelty is shown in the Figure 4. The novelty score of each paper and patent is measured by the proportion of novel entities within the document.

[5] https://ncses.nsf.gov/pubs/nsb20204/invention-u-s-and-comparative-global-trends

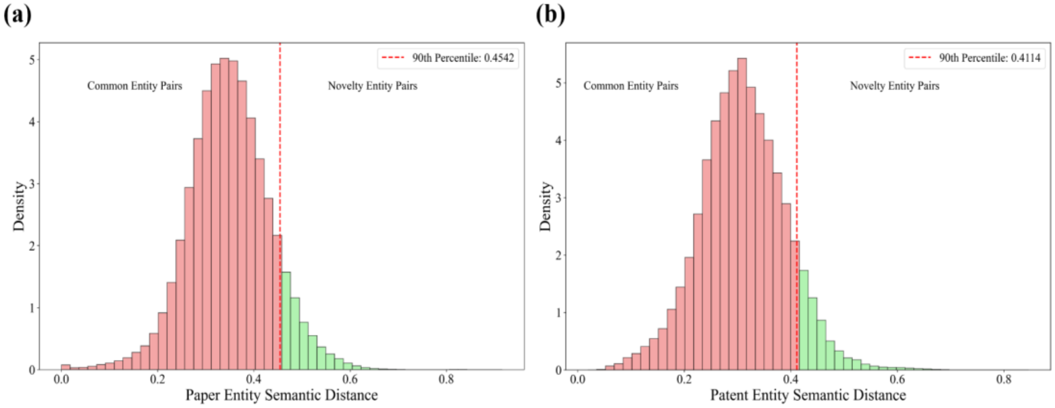

Figure 4. Semantic distance distribution of fine-grained knowledge entities (a) Semantic distance distribution of paper entities; (b) Semantic distance distribution of patent entities.

Based on the novelty of each patent and paper, we calculate the average novelty of patents and papers from each institution per year. As shown in Figure 5(a) and (b), the novelty of publications from various types of institutions in the NLP field generally exhibits an upward trend. Additionally, we observe that, the novelty trends of both papers and patents in industry are lower than those in academia and collaborations. This will be further explored in the next section.

Additionally, a six-year time window was employed, dividing the data into four intervals to assess differences over time, as shown in Figures 6. We conducted t-tests across different intervals to analyze the differences in novelty over time.

Although both paper and patent novelty trends exhibit upward growth, t the increase in novelty was more pronounced in the most recent time window (2018–2022) for patents, reflecting the rapid advancement of technological accumulation and application innovation. Although the t-tests in Figures 6(d) and 6(f) were not significant, this result is primarily due to the small number of patents related to academia and collaboration types. In contrast, the increase in novelty for NLP papers over the past six years was not significant. Several factors may contribute to this trend. First, this may be due to the gradual maturation of methodologies. Recent pre-trained models, in particular, show strong theoretical connections with earlier deep learning techniques. Second, the novelty measurement is based on the semantic distance calculated by SciBERT, whose training corpus primarily consists of Semantic Scholar papers before 2019. Consequently, it may have limited capacity to express fine-grained knowledge entities that appear in recent papers.

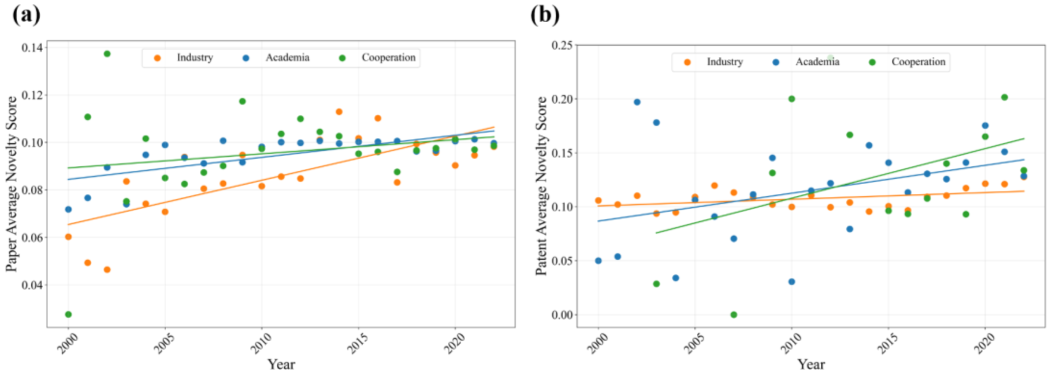

Figure 5. Trends in novelty changes of papers and patents (a) Average novelty of papers from different institutions (b) Average novelty of patents from different institutions.

Furthermore, in terms of collaboration types, Figures 6(c) and 6(f) exhibit different patterns. For papers, the novelty of collaboration types remains nearly constant across each window, while for patents, the novelty of collaboration types shows an upward trend. Although statistically insignificant (due to the small sample size). This highlights the different performances of collaboration types institutions in terms of patent and paper novelty. When it comes to industry and academia, we did not observe any significant differences in trends.

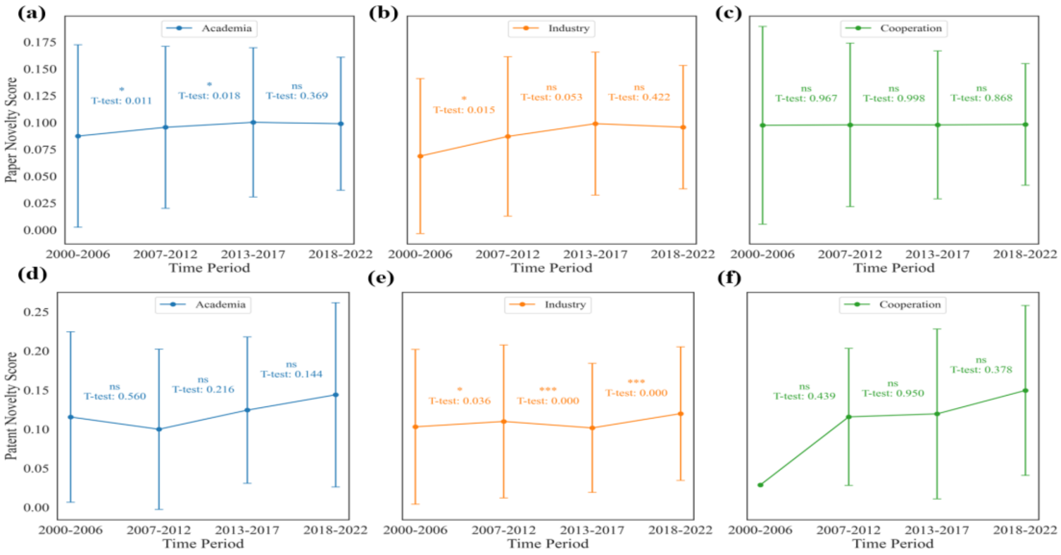

Figure 6. The differences in novelty across different time windows. (a) Novelty variation in academic papers over 6-year windows (b) Novelty variation in industry papers over 6-year windows (c) Novelty variation in cooperation papers over 6-year windows (d) Novelty variation in academic patents over 6-year windows (e) Novelty variation in industry patents over 6-year windows (f) Novelty variation in cooperation patents over 6-year windows.

Regression analysis of novelty differences across various type institutions

In this section, we focus on answering RQ2. Our preliminary analysis reveals the disparities in novelty among various types of institutions within both papers and patents, as shown in Figure 7. It is observed that academic and collaborative institutions exhibited higher novelty than industrial ones. Further, using t-tests, we found that the novelty differences between academic and collaborative institutions were not significant, with both exhibiting higher novelty than the industrial sector. To more accurately characterize the results and their reliability, we conducted regression analysis, controlling for year and institution count, to evaluate the novelty of different types of literature.

Further, we use institution type as the independent variable and introduce a series of control variables to explore the differences in novelty across different institution combinations. The regression results are shown in Tables 4 and 5.

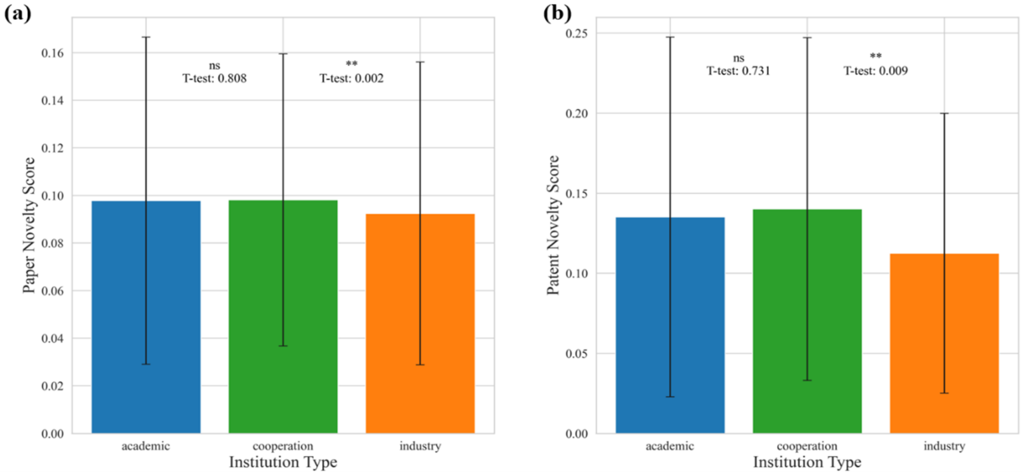

Figure 7. Box plot of novelty distribution. (a) Novelty differences across publishing institutions in the papers (b) Novelty differences across publishing institutions in the patents.

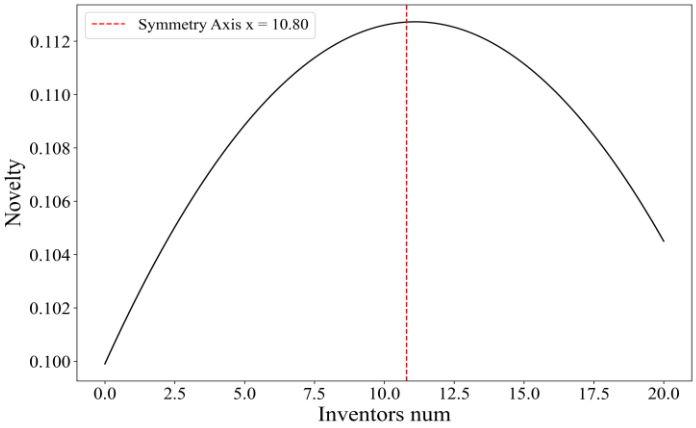

Figure 8. The relationship between the number of patent inventors and novelty. (The dashed line represents the axis of symmetry of the inverted U-shaped curve).

In the regression analysis, this study particularly focuses on the novelty performance of academia and industry in scientific papers and patents. To ensure a more focused analysis, other types of institutions were excluded. For patents, we controlled for year and institution type fixed effects, while also introducing various control variables to examine the relationship between institution type and novelty scores.

As shown in Table 5, Model (1), which includes only the independent variables, demonstrates that patents produced by academic and collaborative institutions exhibit significantly higher levels of novelty compared to those from industrial institutions. These differences are statistically significant at the 1% and 5% levels, respectively. Models (3) and (4) progressively incorporate control variables, yet the positive association between academic and collaborative institutions and patent novelty remains consistent and robust. This conclusion holds even after accounting for the number of inventors, IPC categories, and patent family size. Model (2) serves as the baseline model, exploring the relationship between team size (number of inventors) and patent novelty. The analysis reveals that the number of inventors is generally positively correlated with novelty, exhibiting a slight inverted U-shaped trend. The squared term of the number of inventors has a small but significant effect (β = 0.0001, p < 0.1). Further exploration confirms an inverted U-shaped relationship between team size and novelty. Figure 8 illustrates the trends in novelty as a function of inventor team size.

The regression analysis results at the paper level, presented in Table 6, reveal that in Model (1), which includes only the independent variables, indicates that academic papers and collaborative papers generally exhibit higher novelty. However, when the number of institutions is introduced as a control variable in Model (3), the novelty advantage of collaborative papers over industrial papers becomes statistically insignificant, suggesting that institutional factors mediate the observed effects of collaboration. Model (2) assesses the impact of the number of authors, revealing no significant correlation between the number of authors and paper novelty, in contrast to the notable role of inventor count in patents. Finally, Model (4), which includes all control variables, confirms the earlier conclusions: academic papers are still more novel compared to industrial papers, and the novelty of collaborative papers aligns more closely with that of industrial papers.

Table 5. Regression results for patent novelty.

Novelty **Variables**	(1) *Model 1*	(2) *Model 2*	(3) *Model 3*	(4) *Model 4*
Academic	0.020***		0.020***	0.021***
	(0.004)		(0.004)	(0.004)
Cooperation	0.025**		0.024**	0.0245**
	(0.011)		(0.011)	(0.011)
Family size				-0.001***
				(0.000)
Inventors num		0.002***		0.002***
		(0.001)		(0.001)

Inventors num sq		-0.000*		-0.000*
		(0.000)		(0.000)
Institutions num			-0.000	-0.001
			(0.001)	(0.001)
IPC num				0.001**
				(0.001)
Constant	0.104***	0.100***	0.105***	0.102***
	(0.006)	(0.006)	(0.006)	(0.009)
Year Fixed	Yes	Yes	Yes	Yes
Observations	22269	22,269	22,269	22,269
R-squared	0.015	0.014	0.015	0.017

Note: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 6. Regression results for paper novelty.

Novelty	(1)	(2)	(3)	(4)
Variables	*Model 1*	*Model 2*	*Model 3*	*Model 4*
Academic	0.005***		0.0054***	0.004***
	(0.002)		(0.002)	(0.002)
Cooperation	0.004**		0.0031	0.003
	(0.002)		(0.002)	(0.002)
Authors num		0.000		-0.000
		(0.000)		(0.000)
Institutions num			0.001	0.001
			(0.000)	(0.000)
Constant	0.063***	0.067***	0.063***	0.063***
	(0.007)	(0.006)	(0.007)	(0.007)
Year Fixed	Yes	Yes	Yes	Yes
Observations	17,644	17,644	17644	17,644
R-squared	0.005	0.004	0.0056	0.005

Note: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Robustness checks

We conducted a robustness check on the previous results to verify the reliability of the findings. Furthermore, we binarized the novelty scores by labeling the top 10% of papers with the highest novelty scores as "novel," while the remaining papers were labeled as "non-novel" (Jeon et al., 2022). Subsequently, we reanalyzed the data using logistic regression, and the results, as shown in Tables 7 and 8, were consistent with the previous findings.

Table 7. Regression results of patent novelty with novelty as a binary variable.

Novelty Variables	(1) *Model 1*	(2) *Model 2*	(3) *Model 3*	(4) *Model 4*
Academic	0.587***		0.587***	0.612***
	(0.127)		(0.127)	(0.127)
Cooperation	1.015***		1.02***	1.023***
	(0.287)		(0.295)	(0.297)
Family size				-0.022
				(0.013)
Inventors num		0.073***		0.070***
		(0.026)		(0.026)
Inventors num sq		-0.004		-0.003
		(0.002)		(0.002)
Institutions num			-0.006	-0.047
			(0.055)	(0.05)
IPC num				-0.003
				(0.024)
Constant	-2.171***	-2.313***	-2.164***	-2.101***
	(0.220)	(0.225)	(0.228)	(0.332)
Year Fixed	Yes	Yes	Yes	Yes
Pseudo R-squared	0.002	0.001	0.002	0.003

Note: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 8. Regression results of paper novelty with novelty as a binary variable.

Novelty Variables	(1) *Model 1*	(2) *Model 2*	(3) *Model 3*	(4) *Model 4*
Academic	0.256***		0.236**	0.221**
	(0.094)		(0.094)	(0.095)
Cooperation	0.120		0.04	0.027
	(0.103)		(0.019)	(0.110)
Authors num		0.006		-0.017
		(0.012)		(0.014)
Institutions num			0.046**	0.063**
			(0.020)	(0.026)
Constant	-2.469***	-2.259***	-2.510***	-2.483***
	(0.341)	(0.333)	(0.341)	(0.342)
Year Fixed	Yes	Yes	Yes	Yes
Pseudo R-squared	0.001	0.000	0.001	0.002

Note: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Discussion

This study adopted fine-grained knowledge entity analysis to evaluate the novelty of patents and papers within the NLP field. Based on previous entity-based novelty metrics, we further optimized the novelty measurement method. Through regression analysis, it was revealed that new ideas in the NLP field are continuously emerging (Zhang et al., 2024). Moreover, the level of novelty in academia surpasses that in industry when considering both papers and patents. This finding is consistent with the results of (Chen et al., 2024; Liang et al., 2024). Further research has found that, at the paper level, academic–industry collaborations struggle to replicate the novelty of academic teams and tend to resemble the work of industry teams (Liang et al., 2024).

As a catalyst, academia significantly promotes the enhancement of novelty, both in terms of filing patents individually and participating in patent research composition, thereby enabling the industry to disclose more innovative findings. This trend remains significant after controlling for the number of institutions and other relevant variables. This not only helps advance patent technologies to higher levels but also provides more competitive technological solutions for the industry. As Krieger et al. (2024) point out, scientific research enables companies to derive significantly more value from their inventions, and patents closer to science tend to exhibit higher novelty. In contrast, at the paper level, although academia overall performs with greater novelty, after controlling for the number of institutions, the impact of collaboration type and team size on the novelty of scientific papers is relatively small. This study only found an inverted U-shaped relationship between the size of collaborative teams and novelty in patents.

Implications

Theoretical implications: The theoretical significance of this study is reflected in the following three aspects: First, by transferring the paper entity recognition model knowledge to the patent entity recognition model and combining it with an entity-based novelty measurement method, this study achieves a unified measurement of novelty in both patents and papers. This provides a feasible framework for evaluating the novelty of paper and patent levels across a broader dataset. Second, this study provides new empirical evidence, revealing the novelty differences between academia and industry in the NLP field, both in patents and scientific papers, and highlights how novelty varies across different types of institutions. Finally, this study examines the relationship between team characteristics and novelty in the NLP field, particularly how team size impacts the novelty of research outcomes. It confirms that larger inventor teams, by combining diverse expertise, tend to innovate within familiar knowledge domains (Uzzi et al., 2013). However, when team size exceeds a certain threshold, increased coordination costs and communication challenges lead to incremental improvements rather than novel breakthroughs. This suggests that larger teams in the patent field may experience reduced innovation novelty, relying more on established solutions (Wu et al., 2019; Shin et al., 2022). This finding contributes to the understanding of research team formation and collaboration models in the NLP field.

Practical implications: The results of this study offer theoretical support for the distinct roles of academia and industry in technological innovation, while providing practical recommendations for optimizing research team composition and size. The findings show that academia generally exhibits higher novelty in both patents and papers, highlighting the importance of academic institutions' role in advancing fundamental research and innovation. Academia's openness and collaboration foster new ideas and support interdisciplinary efforts (Brescia et al., 2014). This study also reveals the impact of team size on novelty. In technology-intensive fields like NLP, larger inventor teams can drive innovation by integrating diverse expertise. While reasonable team size and interdisciplinary collaboration foster breakthroughs, overly large teams may increase coordination costs and dilute focus, reducing innovation efficiency. According to the regression analysis, the "threshold" for inventor teams appears to be around 10 members. For typical inventor teams, increasing team size helps improve patent novelty. However, for scientific papers, the number of authors does not directly affect innovation, indicating that novelty depends more on research depth and collaboration model than on team size or cross-sector collaboration work.

Limitations

Despite adjustments to the entity-based novelty measurement method and empirical analysis revealing novelty differences between academia and industry, this study has some limitations. First, while we classified entity relationships and quantified semantic distances, the removal of specific entity types remains coarse. Future research should refine entity distance measurements, especially for same-type and different-type entities, or incorporate discourse structure information. Additionally, there is some discrepancy between semantic distance and the difficulty of combining fine-grained knowledge entities. Future studies could explore combining graph representation learning with co-occurrence network topology to improve novelty assessment. Finally, although this study's dataset covers a wide range of patents and papers, the sample size in the NLP field is relatively limited. Additionally, the imbalanced distribution of institutions in the paper and patent data, especially in the patent data, may potentially affect the accuracy of the analysis results. In addition, although we found that the novelty of industry outputs is lower than that of academia, we did not further explore the reasons behind this. The study did not address whether the disclosure strategy of industry is more conservative, or if the research content itself lacks sufficient novelty. Finally, while we included several key factors that are easy to capture and control in the regression, other variables may have been overlooked, potentially influencing the study's outcomes.

Conclusion and future works

This study explores novelty differences between academia and industry. By extracting fine - grained knowledge entities and measuring paper novelty based on novel entity proportions, regression models analyse novelty differences in patents and papers from academia and industry.

Results show academia has a novelty advantage in both patents and papers, especially in patents. In scientific papers, the impact of collaboration type on novelty is insignificant when controlling for team size. There's an inverted U - shaped relationship between patent team size and novelty in the NLP field. For scientific papers with small inventor teams, increasing team size and cross - disciplinary collaboration can boost patent novelty.

Future research directions include: expanding the sample to the AI field to validate findings; using graph representation learning and entity connection frequency, instead of just semantic distance, to measure novelty; and exploring the mechanisms behind the greater patent novelty in academia - industry collaboration by examining factors like scientific - technical distance, institutional research backgrounds, and disclosure strategies.

Acknowledgments

This paper was supported by the National Natural Science Foundation of China (Grant No.72074113).

References

Ahmed, N., Wahed, M., & Thompson, N. C. (2023). The growing influence of industry in AI research. *Science*, 379(6635), 884–886.

Ahuja, G., & Lampert, C. M. (2001). Entrepreneurship in the large corporation: a longitudinal study of how established firms create breakthrough inventions. *Strategic Management Journal*, *22*(6–7), 521–543.

Arinas, I. (2012). How vague can your patent be? Vagueness strategies in US patents. Vagueness Strategies in US Patents. *HERMES-Journal of Language and Communication in Business*, *48*, 55-74.

Arts, S., Hou, J., & Gomez, J. C. (2019). Text Mining to Measure Novelty and Diffusion of Technological Inventions. In *Proceedings of the 1st Workshop on Patent Text Mining and Semantic Technologies*. Karlsruhe, Germany. https://doi.org/10.34726/pst2019.2

Arts, S., Hou, J., & Gomez, J. C. (2021). Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. *Research Policy*, *50*(2), 104144.

Arundel, A. (2001). The relative effectiveness of patents and secrecy for appropriation. *Research Policy*, *30*(4), 611–624.

Ashtor, J. H. (2021). Modeling patent clarity. *Research Policy*, *51*(2), 104415.

Anne, K. (2023). Data-drivenness, novelty, and interdisciplinarity in the study of criminology. In *Proceedings of the International Society for Scientometrics and Informetrics*, 2, 225–231. https://doi.org/10.5281/zenodo.8370929

Ba, Z., Meng, K., Ma, Y., & Xia, Y. (2024). Discovering technological opportunities by identifying dynamic structure-coupling patterns and lead-lag distance between science and technology. *Technological Forecasting & Social Change,* 200(6351), Article 123147.

Bikard, M., & Marx, M. (2019). Bridging Academia and industry: How geographic hubs connect university science and corporate technology. *Management Science*, *66*(8), 3425–3443.

Brescia, F., Colombo, G., & Landoni, P. (2016). Organizational structures of Knowledge Transfer Offices: an analysis of the world’s top-ranked universities. *The Journal of Technology Transfer,* 41(1), 132–151.

Chan, H. F., & Torgler, B. (2020). Gender differences in performance of top cited scientists by field and country. *Scientometrics*, *125*(3), 2421–2447.

Chen, C., Hsiao, Y., Chu, M., & Hu, K. (2015). The relationship between team diversity and new product performance: the moderating role of organizational slack. *IEEE Transactions on Engineering Management*, *62*(4), 568–577.

Chen, Z., Zhang, C., Zhang, H., Zhao, Y., Yang, C., & Yang, Y. (2024). Exploring the relationship between team institutional composition and novelty in academic papers based on fine-grained knowledge entities. *The Electronic Library, 42*(6): 905-930..

Chirico, F., Criaco, G., Baù, M., Naldi, L., Gomez-Mejia, L. R., & Kotlar, J. (2018). To patent or not to patent: That is the question. Intellectual property protection in family firms. *Entrepreneurship Theory and Practice*, *44*(2), 339–367.

Clarysse, B., Andries, P., Boone, S., & Roelandt, J. (2023). Institutional logics and founders' identity orientation: Why academic entrepreneurs aspire lower venture growth. Research Policy, 52(3), Article 104713.

Codina-Filbà, J., Bouayad-Agha, N., Burga, A., Casamayor, G., Mille, S., Müller, A., Saggion, H., & Wanner, L. (2016). Using genre-specific features for patent summaries. *Information Processing & Management*, *53*(1), 151–174.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. (pp. 4171–4186). Minneapolis, Minnesota.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., . . . Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, 57(2021), Article 101994.

Färber, M., & Tampakis, L. (2023). Analyzing the impact of companies on AI research based on publications. Scientometrics, 129(1), 31-63.

Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science*, *47*(1), 117–132.

Geisler, E. (1995). When whales are cast ashore: the conversion to relevancy of American universities and basic science. *IEEE Transactions on Engineering Management*, *42*(1), 3–8.

Homscheid, D., Kunegis, J., & Schaarschmidt, M. (2015). Private-Collective Innovation and Open Source Software: Longitudinal Insights from Linux Kernel Development. In *Conference on e-Business, e-Services and e-Society (pp. 299-313).* (pp. 299–313).

Hottenrott, H., Rose, M. E., & Lawson, C. (2021). The rise of multiple institutional affiliations in academia. Journal of the Association for Information Science and Technology, 72(8), 1039–1058.

Jang, H., Kim, S., & Yoon, B. (2023). An eXplainable AI (XAI) model for text-based patent novelty analysis. *Expert Systems Withwith Applications*, *231*, 120839.

Jee, S. J., & Sohn, S. Y. (2023). Firms' influence on the evolution of published knowledge when a science-related technology emerges: the case of artificial intelligence. *Journal of Evolutionary Economics*, 33(1), 209–247.

Jeon, D., Ahn, J. M., Kim, J., & Lee, C. (2022). A doc2vec and local outlier factor approach to measuring the novelty of patents. *Technological Forecasting and Social Change*, *174*, 121294.

Jeon, D., Lee, J., Ahn, J. M., & Lee, C. (2023). Measuring the novelty of scientific publications: A fastText and local outlier factor approach. *Journal of Informetrics*, *17*(4), 101450.

Johri, P., Khatri, S. K., Al-Taani, A. T., Sabharwal, M., Suvanov, S., & Kumar, A. (2021). Natural Language Processing: history, evolution, application, and future work. In *Proceedings of 3rd International Conference on Computing Informatics and Networks*. (pp. 365–375). Delhi, India.

Kabore, F. P., & Park, W. G. (2019). Can patent family size and composition signal patent value? *Applied Economics, 51*(60), 6476–6496.

Koppman, S., & Leahey, E. (2019). Who moves to the methodological edge? Factors that encourage scientists to use unconventional methods. *Research Policy, 48*(9), 103807.

Krieger, J. L., Schnitzer, M., & Watzinger, M. (2024). Standing on the shoulders of science. *Strategic Management Journal*, *45*(9), 1670–1695.

Kuhn, J., Younge, K., & Marco, A. (2020). Patent citations reexamined. *The RAND Journal of Economics*, *51*(1), 109–132.

Larivière, V., Macaluso, B., Mongeon, P., Siler, K., & Sugimoto, C. R. (2018). Vanishing industries and the rising monopoly of universities in published research. *PLoS ONE, 13*(8), e0202120.

Lee, C., & Lee, G. (2019). Technology opportunity analysis based on recombinant search: patent landscape analysis for idea generation. *Scientometrics*, *121*(2), 603–632.

Lee, Y., Walsh, J. P., & Wang, J. (2015). Creativity in scientific teams: Unpacking novelty and impact. *Research Policy, 44*(3), 684–697.

Liang, L., Zhuang, H., Zou, J., & Acuna, D. E. (2024). The complementary contributions of academia and industry to AI research. (arVix:2401.10268). *arXiv*.

Liang, Z., Mao, J., & Li, G. (2022). Bias against scientific novelty: A prepublication perspective. *Journal of the Association for Information Science and Technology*, *74*(1), 99–114.

Liu, M., Bu, Y., Chen, C., Xu, J., Li, D., Leng, Y., . . . Ding, Y. (2022). Pandemics are catalysts of scientific novelty: Evidence from COVID- 19. *Journal of the Association for Information Science and Technology*, *73*(8), 1065–1078.

Liu, M., Xie, Z., Yang, A. J., Yu, C., Xu, J., Ding, Y., & Bu, Y. (2024). The prominent and heterogeneous gender disparities in scientific novelty: Evidence from biomedical doctoral theses. *Information Processing & Management, 61*(4), 103743.

Luo, Z., Lu, W., He, J., & Wang, Y. (2022). Combination of research questions and methods: A new measurement of scientific novelty. *Journal of Informetrics*, *16*(2), 101282.

MacRoberts, M. H., & MacRoberts, B. R. (1996). Problems of citation analysis. *Scientometrics*, *36*(3), 435–444.

Marcoulides, K. M., & Raykov, T. (2019). Evaluation of variance inflation factors in regression models using latent variable modeling methods. *Educational and Psychological Measurement*, *79*(5), 874–882.

Martinez-Senra, A. I., Quintas, M. A., Sartal, A., & Vazquez, X. H. (2015). How Can Firms' Basic Research Turn Into Product Innovation? The Role of Absorptive Capacity and Industry Appropriability. *IEEE Transactions on Engineering Management*, *62*(2), 205–216.

Mori, T., & Sakaguchi, S. (2018). Collaborative knowledge creation: Evidence from Japanese patent data. (arXiv: 1908.01256). *arXiv*.

Nelson, R., & Winter, S. (1982). *An evolutionary theory of economic change*. Harvard University Press.

Park, M., Leahey, E., & Funk, R. J. (2023). Papers and patents are becoming less disruptive over time. *Nature, 613*(7942), 138–144.

Perkmann, M., & Walsh, K. (2009). The two faces of collaboration: impacts of university-industry relations on public research. *Industrial and Corporate Change, 18*(6), 1033–1065.

Pramanick, A., Hou, Y., Mohammad, S. M., & Gurevych, I. (2024). The Nature of NLP: Analyzing contributions in NLP papers. (arXiv: 2409.19505). *arXiv*.

Radford, A., & Narasimhan, K. (2018). Improving Language Understanding by Generative Pre-Training. *Preprint*. 1–12.

Riera, R., & Rodríguez, R. (2022). What if Peer-Review process is killing Thinking-Out-of-the-Box science? *Frontiers in Marine Science*, 9, Article 924469.

Sauermann, H., & Stephan, P. E. (2010). Twins or strangers? Differences and similarities between industrial and academic science. *National Bureau of Economic Research,* (No. w16113).

Shibayama, S., Yin, D., & Matsumoto, K. (2021). Measuring novelty in science with word embedding. *PLoS ONE, 16*(7), e0254034.

Shin, H., Kim, K., & Kogler, D. F. (2022). Scientific collaboration, research funding, and novelty in scientific knowledge. *PLoS ONE*, *17*(7), e0271678.

Sun, X., Chen, N., & Ding, K. (2022). Measuring latent combinational novelty of technology. *Expert Systems With Applications*, *210*, 118564.

Tao, A., Qi, Q., Li, Y., Da, D., Boamah, V., & Tang, D. (2022). Game Analysis of the Open-Source Innovation Benefits of Two Enterprises from the Perspective of Product Homogenization and the Enterprise Strength Gap. *Sustainability*, *14*(9), 5572.

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, 342(6157), 468–472.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need (arXiv:1706.03762). *arXiv*.

Wang, J., Veugelers, R., & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, *46*(8), 1416–1436.

Wei, T., Feng, D., Song, S., & Zhang, C. (2024). An extraction and novelty evaluation framework for technology knowledge elements of patents. *Scientometrics*.

Wu, K., Xie, Z., & Du, J. T. (2024). Does science disrupt technology? Examining science intensity, novelty, and recency through patent-paper citations in the pharmaceutical field. *Scientometrics*.

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, *566*(7744), 378–382.

Wu, M., Chang, K., Zhou, W., Hao, J., Yuan, C., & Chang, K. (2015d). Patent deployment Strategies and patent Value in LED industry. *PLoS ONE*, *10*(6), e0129911.

Xu, H., Bu, Y., Liu, M., Zhang, C., Sun, M., Zhang, Y., Meyer, E., Salas, E., & Ding, Y. (2022). Team power dynamics and team impact: New perspectives on scientific collaboration using career age as a proxy for team power. *Journal of the Association for Information Science and Technology*, *73*(10), 1489–1505.

Yang, Y., Tian, T. Y., Woodruff, T. K., Jones, B. F., & Uzzi, B. (2022). Gender-diverse teams produce more novel and higher-impact scientific ideas. *Proceedings of the National Academy of Sciences*, *119*(36), e2200841119.

Zanella, G., Liu, C. Z., & Choo, K. R. (2021). Understanding the trends in blockchain domain through an unsupervised systematic patent analysis. *IEEE Transactions on Engineering Management*, *70*(6), 1991–2005.

Zhang, H., Zhang, C., & Wang, Y. (2024). Revealing the technology development of natural language processing: A Scientific entity-centric perspective. *Information Processing & Management*, *61*(1), 103574.

Zheng, H., Li, W., & Wang, D. (2022). Expertise diversity of teams predicts originality and Long-Term impact in science and technology. (arXiv: 2210.04422). *arXiv*.