Information Processing and Management 64 (2027) 105142



Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman



Gender differences in early semantic trajectories and academic career persistence

Letian Hu [a,b], Alex J. Yang [a,b,*], Sanhong Deng [a,b,*]

[a] School of Information Management, Nanjing University, Nanjing, 210023, China
[b] International Joint Informatics Laboratory, Nanjing University, Nanjing, 210023, China

ARTICLE INFO

Keywords:
Semantic trajectories
Research careers
Academic attrition
Gender inequality

ABSTRACT

Gender disparities in academia are well documented in publication output, citations, and career attrition, yet relatively little is known about whether male and female scholars differ in the semantic development of their research. Using SPECTER2 embeddings derived from 67 million publications in OpenAlex, we characterize the semantic trajectories of 2.76 million scholars using three measures: semantic expansion (divergence from prior work), research coherence (consistency of prior research), and semantic alignment (continuity with prior research direction). On average, male scholars exhibit greater semantic expansion, whereas female scholars show higher research coherence and semantic alignment. These differences persist across career stages and productivity levels. Discrete-time hazard models further indicate that early semantic trajectories are associated with subsequent publication-based career persistence. Greater early semantic expansion is associated with a lower risk of dropout for both genders, whereas greater research coherence and semantic alignment are associated with a higher dropout risk, with stronger associations among female scholars. These patterns are consistently observed across major scientific disciplines. Together, our findings suggest that gender differences in academia are reflected not only in research productivity and career outcomes but also in the semantic organization of research development, and that early semantic trajectories are associated with subsequent patterns of career persistence.

1. Introduction

Scientific careers are shaped not only by how much scholars publish, but also by how their research evolves in the space of knowledge (Jia et al., 2017; Zeng et al., 2019). Over the past decades, studies of scholars' careers have largely focused on observable outcomes such as productivity, citation impact, collaboration patterns, and career length (Fortunato et al., 2018; Milojević et al., 2018; Sinatra et al., 2016; Wu et al., 2019). These studies have provided important insights into the accumulation of research advantages, career stratification, and the determinants of academic success. However, they frequently overlook a more substantive dimension of academic careers: how scholars navigate semantic space as they generate new work, maintain research continuity, or shift their research focus over time.

This limitation is particularly evident in studies of gender differences in scholarly careers. Extensive literature documents gender differences in scientific careers, including disparities in publication output, citation impact, promotion opportunities, collaboration

* Corresponding authors at: School of Information Management, Nanjing University, Nanjing, 210023, China.
E-mail addresses: alexjieyang@outlook.com (A.J. Yang), sanhong@nju.edu.cn (S. Deng).

https://doi.org/10.1016/j.ipm.2026.105142
Received 9 April 2026; Received in revised form 5 August 2026; Accepted 27 August 2026
0306-4573/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

structures, delayed recognition and career persistence (Box-Steffensmeier et al., 2015; Kwiek & Roszka, 2021; Larivière et al., 2013; Mihaljević-Brandt et al., 2016; Shen et al., 2022; Yang et al., 2025a). One particularly important aspect of this inequality is the higher rate of career attrition among female scholars. Previous research indicates that female scholars are more likely than their male counterparts to leave academic careers(Yang et al., 2025b), resulting in shorter career spans and lower cumulative publication output (Huang et al., 2020), and reduced representation in senior academic positions (Abramo et al., 2021). This attrition is typically attributed to structural disadvantages, workplace environment, familial pressures, and unequal access to recognition and advancement opportunities (Ecklund & Lincoln, 2011; Morgan et al., 2021; Régner et al., 2019; Ross et al., 2022). Most existing studies approach academic careers from an external perspective by quantifying publication output, citation impact, or collaboration patterns, while leaving the intellectual content and developmental structure of research largely unexplored. Consequently, little is known about potential gender differences in the semantic organization of research trajectories. It remains unclear whether male and female scholars differ in the breadth of their exploration, whether their work evolves along a continuous path or transitions across diverse topics, and whether such differences emerge during the early stages of their careers.

Recent advances in text and document embedding techniques have made it increasingly feasible to address these questions. By converting academic publications into high-dimensional semantic vectors, researchers move beyond coarse classifications such as disciplines, topics, or keywords to measure the relative proximity, divergence, and directional shifts among research outputs within a continuous semantic space (Arts et al., 2025; Kozlowski et al., 2019; Tshitoyan et al., 2019; Yang et al., 2026). This enables a more fine-grained perspective on scholarly development. Rather than asking only how productive a scholar is, this approach allows us to examine how far each publication departs from prior work, how coherent a sequence of publications remains, and how an individual's body of work evolves through semantic space. The perspective is particularly valuable for studying the early stages of academic careers, when researchers are still establishing their research identities and many long-term career differences begin to emerge.

Despite this potential, such semantic approaches are rarely integrated into studies of gender differences in academic careers. Specifically, there remains no systematic framework for measuring scholars' semantic trajectories in a way that jointly captures research exploration and internal continuity, or for using such measures to investigate gender differences in scholarly development.

This study aims to examine gender differences in the semantic development of research careers. Using semantic expansion, research coherence, and semantic alignment, we characterize scholars' semantic trajectories and examine how these vary by gender across career stages and productivity levels. We further investigate how early semantic trajectories are associated with dropout risk and whether the associations vary between male and female scholars.

This study addresses three central research questions:

(RQ1) How can the semantic trajectories of scholarly careers be measured in a systematic and interpretable manner?
(RQ2) What patterns characterize these semantic trajectories, and do such patterns vary by gender?
(RQ3) What relationship exists between early-career semantic trajectories and subsequent career persistence?

The study makes three main contributions. First, it introduces a semantic framework for analyzing scholarly careers by characterizing the semantic development of publications within a continuous embedding space. Second, it extends the literature on gender differences in academia by shifting attention from conventional performance indicators to the semantic organization and development of research. The findings suggest that gender differences are reflected not only in publication output and citation impact but also in the semantic organization of research trajectories. Third, using a large-scale publication dataset, the study combines semantic embeddings with statistical analysis to provide large-scale evidence on gender differences in research development. The consistency of the findings across multiple metrics and model specifications suggests that these patterns are not confined to a single measurement approach. These findings contribute to a better understanding of gendered patterns in scholarly career development and may inform the design of evaluation and support systems for early-career researchers.

The remainder of the paper is organized as follows. Section 2 reviews the relevant literature on semantic representations of science and gender differences in academic careers. Section 3 describes the dataset, the construction of semantic embeddings, the measurement of scholars' semantic trajectories, and the survival analysis of dropout risk. Section 4 examines the overall characteristics of semantic trajectories and their variation by gender. Section 5 presents survival analysis results on the relationship between early semantic trajectories and career dropout risk, with particular attention to gender heterogeneity. Section 6 discusses the findings, addresses the limitations of the study, and outlines directions for future research.

2. Background

2.1. From semantic similarity to trajectory

Document embedding models trained on scientific corpora enable researchers to move beyond discrete topic labels to estimate the relative proximity, divergence, and directional shifts of publications based on their textual content (Cohan et al., 2020). Unlike traditional bibliometric methods relying on citations, co-authorship, or disciplinary classifications, semantic representations embed titles, abstracts, or full texts into high-dimensional vectors. This approach facilitates a deeper exploration of substantive research content. For instance, Peng et al. (2021) demonstrated that neural embeddings of scholarly periodicals can reveal nuanced disciplinary and interdisciplinary relations and even uncover interpretable semantic axes of science. Huang et al. (2024) further showed that topic evolution can be analyzed through changes in semantic consistency in vector space, illustrating how semantic representations help trace the internal dynamics of research topics over time.

Semantic similarity provides a useful way of quantifying the relatedness between scientific publications and characterizing the structure of scientific knowledge (Boyack & Klavans, 2020). Although semantic similarity provides a pairwise and static measure of the proximity or divergence between two publications in semantic space, it cannot by itself capture the developmental trajectory of scholarly output over time (Singh et al., 2024). A scholarly career is not simply a collection of related papers, but an ordered sequence of research contributions produced within evolving intellectual, institutional, and social contexts (Foster et al., 2015). Accordingly, understanding scholarly development requires moving beyond isolated pairwise comparisons to examine the temporal order, cumulative direction, and shifting patterns of movement across knowledge space.

The concept of research trajectories offers a way of linking scientific outputs to the development of academic careers, as research behavior and output can evolve systematically across career stages (Tripodi et al., 2025). Rather than treating papers as isolated outputs, the trajectory perspective views each publication as part of an evolving sequence within knowledge space (Jia et al., 2017). From this perspective, an academic career operates as a trajectory characterized by both dynamic and structural features. New publications may depart from prior research, align with established directions, or contribute to a coherent thread of cumulative work (Hill et al., 2025; Liu et al., 2021). The approach enables researchers to distinguish several related yet distinct dimensions of academic development, such as the extent of semantic expansion, the degree of internal coherence, and the alignment between current research and previous directions.

This perspective proves particularly crucial for research on gender disparities. If male and female scholars encounter distinct structural constraints, incentives, or opportunities in sustaining their research careers, these differences likely manifest not only in scholarly output and academic recognition but also in the formation and development of their research trajectories (Kozlowski et al., 2022; Thelwall et al., 2019). Consequently, by analyzing the semantic trajectories of scholarly careers, this study connects external disparities with the internal evolution of academic work, offering a novel perspective on academic inequality.

2.2. Gender differences in scholarly careers

Gender differences extend across multiple dimensions of academic careers, including publication output, citation impact, collaboration patterns, promotion prospects, and long-term career persistence (Shen et al., 2022; Weisshaar, 2017). Most of these studies focus on observable career outcomes and their institutional determinants, emphasizing the roles of cumulative advantage, unequal access to resources, disparities in recognition, and inherent structural barriers within the academic system. From this perspective, gender inequality in academia is typically understood as a stratification issue concerning opportunities, evaluation, and career advancement. These inequalities have led to the survival challenges faced by female scholars. In the field of life sciences, Lerchenmueller and Sorenson (2018) found that during the transition from postdoctoral researcher to principal investigator, women's promotion rates were significantly lower than those of men. Differences in research output, combined with unequal returns on research achievements, collectively shaped career trajectories that differed by gender.

Huang et al. (2020) argue that there is actually no significant difference in the average annual output between male and female scholars. The sole major reason for the lag in women's total output and total citations is that female scholars have significantly shorter career lifespans, and their dropout rate is far higher than that of men. Similarly, Chan and Torgler (2020) found that in certain male-dominated fields, top female scholars perform at least as well as their male counterparts and are even more influential, suggesting that the underrepresentation of women among top scholars should not be conflated with a lack of ability. Yang et al. (2024) indicate that while women experience higher dropout rates than men in the early stages of their careers, in most fields, female researchers' citation impact is often comparable to, or even higher than that of their male counterparts. Their findings suggest that talent attrition in academia not only means a higher proportion of women are leaving the field, but also that the scientific community may be losing more women who would otherwise have gone on to achieve excellence (Alper, 1993).

The influence of gender differences also extends to the orientation and objectives of academic work. Zhang et al. (2021) found that male researchers more frequently prioritize and pursue research primarily designed to advance science, which typically receives higher citation impact. In contrast, female researchers often value and conduct research aimed at promoting societal progress, which generally receives more abstract views. Although female scholars tend to introduce more novel connections within knowledge networks, these contributions are often less widely adopted and receive lower academic career returns than comparable innovations introduced by male scholars (Hofstra et al., 2020).

While previous research provides valuable insights into the semantic representation of scientific publications and gender inequality in academic careers, a significant gap remains between these two areas of inquiry. Studies on gender disparities rarely examine the semantic evolution of research content. Conversely, research utilizing semantic embedding techniques typically focuses on document-level analysis rather than the longitudinal trajectories of individual scholars. This study bridges the gap by investigating gender differences within the semantic trajectories of scholars and employing semantic measures to characterize the temporal evolution of academic work.

3. Method

3.1. Dataset

We used the OpenAlex dataset (Priem et al., 2022), a large open-source catalog of scholarly publications maintained by OurResearch. OpenAlex combines data from multiple sources, including Microsoft Academic Graph, Crossref, PubMed, ORCID, and publisher websites. The version we downloaded in February 2025 contains over 250 million works published between 1666 and 2025. These

works include journal articles, conference papers, books, datasets, and dissertations, with rich metadata on authors, institutions, citations, concepts, and funders. OpenAlex provides strong entity disambiguation by assigning unique identifiers to authors and works based on names, publication histories, citation patterns, and ORCID records.

To obtain high-quality semantic embeddings, we included only publications with both a title and an abstract. We restricted the sample to non-retracted English-language journal articles and conference proceedings. We excluded publications from journals with an Impact Factor of zero or missing publisher information, as well as those with duplicate titles or abstracts. We also removed publications that lacked a title or had no listed authors. After applying these filters, the dataset consisted of 75,612,702 publications published between 1666 and 2024.

We further restricted the analysis to scholars who published at least five papers during the first five years of their academic careers, ensuring sufficient observations for estimating early semantic trajectories. To reduce potential author disambiguation errors, scholars with career lengths exceeding 70 years or >1000 publications were excluded. The final dataset comprised 2,768,621 scholars and 67,651,283 publications. The effective sample size varies slightly across analyses because of missing values and repeated observations arising from coauthored publications. Descriptive statistics are reported in Table 1.

To infer author gender, we used the gender probabilities provided in SciSciNet 2.0, which were generated using the statistical name-based inference model developed by Van Buskirk et al. (2023). This model estimates the probability that a first name is associated with a female or male gender based on large-scale demographic data and has been reported to achieve an overall accuracy of approximately 72–76% (Van Buskirk et al., 2023). Gender probabilities were available for 2,130,186 of the 2,768,621 scholars in our sample. Following the convention adopted in SciSciNet, authors with a predicted female probability greater than 0.5 were classified as female, whereas those with probabilities below 0.5 were classified as male. The final analytical sample included 10,897,167 publications authored by 817,617 female scholars and 20,994,838 publications authored by 1,312,569 male scholars.

As with all name-based gender inference methods, classification accuracy varies across cultural and linguistic contexts and is generally lower for names from regions such as East and South Asia. Consequently, some degree of gender misclassification is unavoidable. Because such errors are unlikely to be systematically associated with the semantic trajectory measures examined in this study, they are expected primarily to introduce random measurement error, which would tend to attenuate rather than inflate the estimated gender differences. Therefore, our findings should be interpreted as population-level statistical patterns rather than precise classifications of individual scholars.

3.2. Semantic representation of publications

Each publication was represented by a dense semantic embedding derived from its title and abstract. The resulting embeddings locate publications in a continuous semantic space, where distances between vectors reflect semantic relatedness.

We did not use large language models to generate text embeddings for this study. Existing research demonstrates that embedding quality does not scale monotonically with model size. Depending on the specific tasks and training configurations, small- to medium-sized embedding models can still achieve exceptional performance (Muennighoff et al., 2023). Given the scale of our corpus and the necessity for computationally efficient, stable, and comparable document representations, we used SPECTER2 (Singh et al., 2023), a specialized embedding model trained on scientific literature that incorporates both textual content and citation network information. This design enables the model to capture scientific relatedness beyond surface lexical similarity, which is well suited to tracing continuity, expansion, and redirection in research trajectories. Following data preprocessing steps such as tokenization and denoising, each publication's title and abstract is mapped to a fixed length embedding vector $\mathbf{e}_i \in \mathbb{R}^d$, where d denotes 768 dimensions. To ensure comparability across documents, all publication vectors undergo L2 normalization prior to subsequent calculations. This normalization projects all vectors onto a unit hypersphere, effectively reducing the bias introduced by text length variations on feature representation. It also allows the similarity between two publications to be directly measured by the cosine of the angle between their vectors. Given two publications i and j, with normalized semantic vectors $\mathbf{e}_i$ and $\mathbf{e}_j$, their semantic similarity is defined as cosine similarity:

$$\text{CosineSimilarity}(i,j) = \mathbf{e}_i \cdot \mathbf{e}_j$$

Although document embeddings usually generate cosine similarity values that are concentrated within a narrow range, our analytical focus is on research divergence rather than mere proximity in academic evolution. In line with this, we introduce the concept

Table 1
Descriptive statistics of the dataset.

Variables	Obs.	Mean	Std.	Min	10th pct.	Median	90th pct.	Max
Female	2,130,186	0.38	0.49	0	0	0	1	1
Dropout event	2,768,621	0.23	0.42	0	0	0	1	1
Career length	2,768,621	12.34	10.21	1	3	9	26	70
Productivity	2,768,621	24.27	39.09	5	5	12	53	995
Total citations	2,768,621	985.48	3203.56	0	17	231	2131	443,130
Semantic Expansion	2,768,621	0.11	0.03	0	0.08	0.11	0.15	0.26
Research Coherence	2,768,621	0.97	0.01	0.9	0.95	0.97	0.98	1
Semantic Alignment	2,768,621	0.92	0.02	0.8	0.89	0.92	0.94	1

Notes: gender information is available for 76.94% of scholars (2,130,186 of 2,768,621).

of cosine distance:

$$\text{CosineDistance}(i,j) = 1 - \text{CosineSimilarity}(i,j)$$

Cosine distance highlights directional shifts within the semantic space and eliminates the interference of vector magnitude under the condition of L2 normalization, as shown in Fig. 1. Within this measurement framework, a larger distance value indicates a broader span in research topics or methods and greater semantic deviation between two documents. Conversely, a smaller value signifies high continuity in the academic lineage and minimal semantic deviation.

3.3. Measuring scholars' semantic trajectories

For a given scholar, let $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_t$ denote the semantic embeddings of that scholar's publications ordered chronologically. For each latest paper t, the relevant reference set is the scholar's prior publication history $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{t-1}\}$. The core idea lies in comparing each new paper with the author's own cumulative research trajectory. In this way, these measures capture the scholar's trajectory of movement in semantic space relative to their own prior work, while also preserving, to some extent, semantic features relative to other research within the field.

(1) Semantic Expansion. To assess the extent to which a new publication departs from a scholar's previous work, we define semantic expansion (SE) as the average cosine distance between the focal paper and all earlier papers by the same scholar:

$$D_t = \frac{1}{t-1} \sum_{i=1}^{t-1} (1 - \mathbf{e}_t \cdot \mathbf{e}_i), \quad t \geq 2$$

where $\mathbf{e}_t \cdot \mathbf{e}_i$ is the cosine similarity between the focal publication and a prior publication. Because all embeddings are normalized, this measure captures the average angular departure of the focal paper from the scholar's existing semantic history. A larger value of D_t indicates that the new publication is, on average, more distant from the scholar's prior work, suggesting a greater movement into less familiar semantic regions. A smaller value indicates that the focal paper remains close to the scholar's established research profile.

To elaborate, if a scholar repeatedly publishes papers semantically similar to earlier works, each new publication occupies roughly the same region within the knowledge space, maintaining a low average distance from previous papers. Conversely, when a scholar publishes research that differs substantially in content, framework, or conceptual orientation from past outputs, the focal vector deviates more significantly from the historical dataset, resulting in a higher average cosine distance. In this sense, this metric reflects the degree of semantic deviation between a scholar and their own prior research.

It is important to note that SE captures relative semantic divergence rather than absolute novelty or substantive innovation. Within the semantic space, a focal paper situated at a greater distance from a scholar's prior body of work is not inherently more innovative than a semantically proximate one. This is because a larger semantic distance can also arise from variations in research topics, disciplinary contexts, terminology, or even writing styles. Within the context of this study, a higher value indicates the scholar explores directions increasingly dissimilar to their previous work.

(2) Research Coherence. To assess the degree of internal consistency in a scholar's prior research history, we define research coherence (RC) on the basis of the mean direction of all previous publication vectors:

$$\bar{\mathbf{h}}_{t-1} = \frac{1}{t-1} \sum_{i=1}^{t-1} \mathbf{e}_i$$

and then take the norm of this historical mean direction:

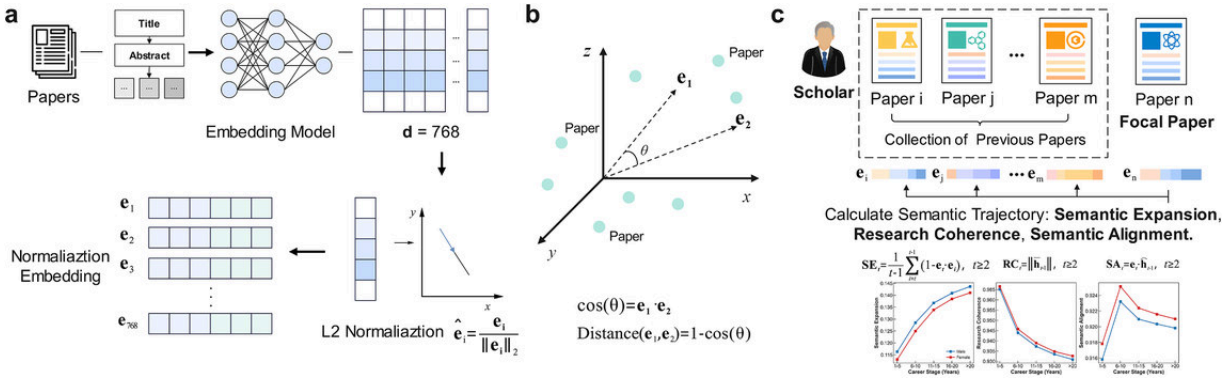


Fig. 1. Framework for measuring scholars' semantic trajectories. (a) Publications are encoded into 768-dimensional SPECTER2 embeddings and L2-normalized. (b) Publications are represented as points in a continuous semantic space, where distances reflect semantic relatedness. (c) Publications are ordered chronologically for each scholar. Each new publication is compared with the scholar's previous publications to compute paper-level semantic measures, which are aggregated into scholar-level trajectory measures.

$$C_t = \| \overline{\mathbf{h}}_{t-1} \|, \qquad t \geq 2$$

This metric reflects the degree of directional consistency among a scholar's early publications in the semantic space. Since each early publication is represented as a unit vector, the norm of their mean vector depends on how closely these vectors are concentrated in the semantic space. When the publications are semantically similar, their vectors point in comparable directions and reinforce one another, producing a mean vector with a relatively large norm. Conversely, when the publications are distributed across diverse semantic regions, their vectors point in less consistent directions and partially offset one another, resulting in a smaller norm.

Consequently, RC reflects the structural properties of a scholar's accumulated research history. A high value indicates that early publications form a concentrated and internally consistent body of work. Geometrically, these early vectors cluster within a relatively narrow region of the semantic space to establish a distinct historical direction. A low value suggests greater dispersion among early publications, indicating a broader, less focused, or more fragmented pattern of research development.

This metric differs conceptually from SE. While SE evaluates the relationship between a new focal paper and previous publications, RC assesses the internal structure of the early papers themselves. A scholar might possess a highly coherent research history yet produce new work that diverges significantly from past research. Alternatively, a scholar with a dispersed and weakly coherent research history might continuously publish papers situated relatively close to subsets of their previous work. By separating these two dimensions, our framework avoids conflating the maintenance of a focused research history with the production of papers similar to past outputs.

(3) Semantic Alignment. To assess whether a new publication follows the dominant direction established by a scholar's earlier work, we define semantic alignment (SA) as the cosine similarity between the focal paper and the normalized historical mean direction. Specifically, we first normalize the historical mean vector:

$$\widehat{\mathbf{h}}_{t-1} = \frac{\overline{\mathbf{h}}_{t-1}}{\| \overline{\mathbf{h}}_{t-1} \|}$$

provided that $\| \overline{\mathbf{h}}_{t-1} \| > 0$, and then compute

$$A_t = \mathbf{e}_t \cdot \widehat{\mathbf{h}}_{t-1}, \qquad t \geq 2$$

This metric reflects the directional relationship between the focal paper and the accumulated semantic orientation of a scholar's previous works. A higher value indicates that the new paper continues or follows the dominant thread implicit in the scholar's publication history. A lower value indicates weaker directional continuity, suggesting the focal paper deviates from the established trajectory or shifts toward a new direction.

The key contribution of SA lies in capturing directional features that average pairwise similarity alone cannot reveal. A new publication might share moderate similarity with many early papers yet fail to align with the dominant direction of the scholar's

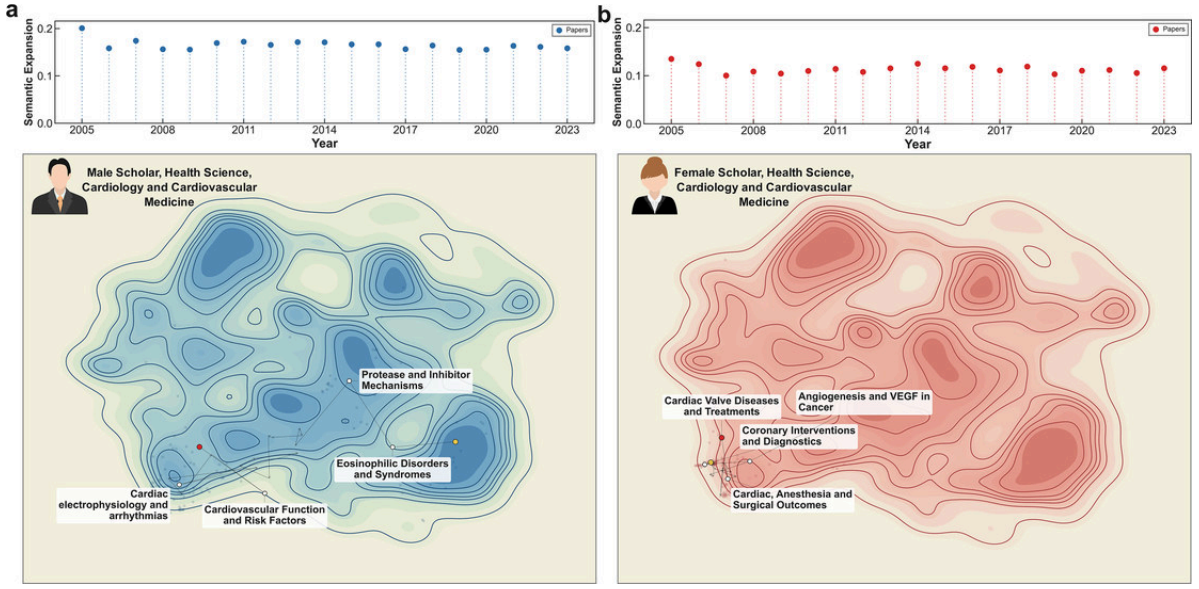


Fig. 2. Representative semantic trajectories of two scholars in Cardiology and Cardiovascular Medicine. (a) Example of a male scholar. The upper panel shows yearly mean Semantic Expansion (SE), and the lower panel visualizes the scholar's publications in a two-dimensional projection of the semantic space with UMAP. (b) Example of a female scholar from the same field. The upper panel presents yearly mean SE, and the lower panel shows the corresponding semantic projection. These examples illustrate different patterns of semantic expansion and topical concentration over time.

historical development. Similarly, a focal paper might not be particularly close to any single early publication, but its vector could still point in the same general direction as the cumulative trend of previous works. Therefore, SA identifies whether a scholar is consolidating an existing research thread rather than merely maintaining local proximity to a subset of early papers.

From the perspective of career development, this distinction is crucial. Academic trajectories often combine continuity and transition. Individual papers may vary in the extent to which they depart from a researcher's prior body of work, ranging from incremental continuation to substantial pivots into new research areas (Hill et al., 2025). Conversely, a scholar might gradually shift their research direction toward a new field while maintaining moderate proximity to past research. SA helps capture such directional continuity or divergence.

It should be noted that SE measures the extent to which a scholar's newly published paper deviates from the cumulative semantic profile of their prior work, thereby reflecting the degree of exploration in their research topics (Fig. 2). In contrast, SA assesses whether this deviation remains directionally consistent with the author's existing research trajectory. Geometrically, when comparing the focal paper's vector to the mean vector of previous publications, SE captures the semantic distance from this historical profile. Conversely, SA normalizes this relationship based on the strength of the historical direction, thereby reflecting the angular continuity with past work.

In conclusion, these three metrics provide a multidimensional description of the semantic trajectories of scholars. SE reflects the degree to which a new publication deviates from prior work. RC indicates the extent to which previous publications form a concentrated semantic history. SA reveals whether a focal paper follows the dominant direction of that established history. These dimensions are complementary rather than redundant. For example, a scholar may have a highly coherent prior research history but publish a focal paper that is both weakly aligned with that history and substantially distant from it, resulting in high RC, low SA, and high SE. Conversely, for a specific focal paper, the author's prior publications may constitute a relatively dispersed semantic history, yielding low RC. Nevertheless, if this focal paper aligns with a dominant or emerging trajectory within that heterogeneous body of early work, it can still exhibit high SA. The coexistence of these possibilities necessitates a trajectory-based framework that incorporates multiple metrics beyond a single similarity measure.

To characterize these semantic trajectories, we aggregate these paper-level metrics within specific observation windows. The analysis focuses on various stages of a scholar's career to calculate the author-level averages for SE, RC, and SA of the publications produced during each stage. These aggregated values summarize the general semantic characteristics of individual research trajectories and serve as the primary variables for subsequent comparative analyses. Because a scholar's first publication lacks historical data, the trajectory-based metrics remain undefined for this initial observation and are therefore treated as missing values during the aggregation process.

3.4. Survival analysis of dropout risk

To examine whether early semantic trajectories relate to later career persistence, we use a staged discrete-time hazard model to analyze dropout risk. Following prior publication-based studies of academic career persistence, a dropout event is identified when an author produces no new publication for five consecutive years after the last observed paper, and the year of the final publication is treated as the event time (Frandsen & Nicolaisen, 2024; Mihaljević-Brandt et al., 2016). Based on this design, the risk set begins in career year 6, and dropout risk is evaluated across six career stages: 6–10, 11–15, 16–20, 21–25, 26–30 and >30 years since first publication.

Building on existing research (Box-Steffensmeier et al., 2015), we construct the baseline model and subsequently introduce three semantic metrics and additional control variables, to test whether early research styles provide additional explanatory power for subsequent dropout risk.

$$\begin{aligned}\operatorname{logit}[P(EventInStage_{it}=1)] = {} & \alpha_{\mathrm{Stage}_{it}} + \delta_{\mathrm{Field}_i} + \beta_1 Gender_i + \beta_2 N_{papers,i}^{first5} + \beta_3 Cohort_i \\ & + \beta_4 FirstAuthorRatio_i + \beta_5 LastAuthorRatio_i \\ & + \beta_6 \log(1 + MeanCoauthor_i) + \beta_7 FundRatio_i + \beta_8 S_{ki}\end{aligned}$$

In the model, $EventInStage_{it} = 1$indicates that scholar i exits during career stage t. $\alpha_{\mathrm{Stage}_{it}}$denotes career stage fixed effects, and $\delta_{\mathrm{Field}_i}$ captures field fixed effects. $Gender_i$ is the scholar's gender variable, coded as 0 for men and 1 for women. $N_{papers,i}^{first5}$ is the number of publications produced in the first five career years, and $Cohort_i$ captures the scholar's first publication year. We further control for early-career authorship position, collaboration, and funding characteristics. Specifically, $FirstAuthorRatio_i$ and $LastAuthorRatio_i$ denote the shares of scholar i's first five-year publications in which the scholar is recorded as first author and last author respectively; $\log(1 + MeanCoauthor_i)$ denotes the mean number of coauthors across the scholar's first five-year publications; $FundRatio_i$ denotes the proportion of the scholar's publications during the first five career years that acknowledge research funding. A logarithmic transformation was applied to the mean number of coauthors because collaboration size is highly right-skewed. Finally, S_{ki} denotes the early-career semantic trajectory measure, $k$corresponds to SE, RC, or SA.

We further examine whether the associations between early semantic trajectories and dropout risk differ between male and female scholars. To achieve this, we estimate three interaction models, each incorporating a single semantic metric and its interaction with gender.

$$\begin{aligned}\text{logit}[P(EventInStage_{it}=1)] &= \alpha_{\text{Stage}_{it}} + \delta_{\text{Field}_i} + \beta_1 Gender_i + \beta_2 N^{first5}_{papers,i} \\ &+\beta_3 Cohort_i + \beta_4 FirstAuthorRatio_i + \beta_5 LastAuthorRatio_i \\ &+\beta_6 \log(1+MeanCoauthor_i) + \beta_7 FundRatio_i + \beta_8 S_{ki} + \beta_9 (Gender_i \times S_{ki})\end{aligned}$$

In this specification, β_8 represents the semantic coefficient for the reference gender group (male scholars), and β_9 captures the extent to which that coefficient differs by gender. For ease of interpretation, we convert the estimated coefficients into odds ratios and compute gender-specific semantic associations for each semantic metric.

Additionally, we re-estimate the gender-semantic interaction models separately within three scientific domains—Health Sciences, Life Sciences, Physical Sciences for robustness analysis. This domain-stratified analysis allows us to assess whether the estimated semantic and gender interaction terms are stable across disciplinary contexts. Furthermore, we conduct two additional robustness checks. First, recent work on hybrid retrieval has highlighted that different representation spaces may capture complementary dimensions of domain-specific information (Yu et al., 2026). To evaluate whether our findings rely on the choice of the text embedding model, we recalculate the semantic trajectory metrics using alternative models and re-estimated the interaction models within these alternative constructed semantic spaces. Second, to examine the sensitivity of our findings to the definition of dropout, we re-estimate the interaction models using alternative time windows. Detailed results are provided in the Supplementary note 1 and 2.

4. Profiling scholars' semantic trajectories and gender differences

4.1. Profiling semantic trajectories

This section describes the distributions of the three semantic trajectory measures—Semantic Expansion (SE), Research Coherence (RC), and Semantic Alignment (SA)—and examines how they vary across productivity, citation impact, career stage, and gender.

At the publication level, these three semantic metrics generally exhibit stable and interpretable distributions. As illustrated in Fig. 3a, SE, RC and SA all display regular unimodal distributions. Their overall distributional shapes remain largely similar across different publication periods, suggesting a general temporal trend where SE gradually shifts to the right, while RC and SA gradually shift to the left. This temporal stability indicates that the proposed metrics are not driven by isolated changes within narrow historical windows or the composition of scientific publication corpora, but rather capture relatively enduring structural characteristics of

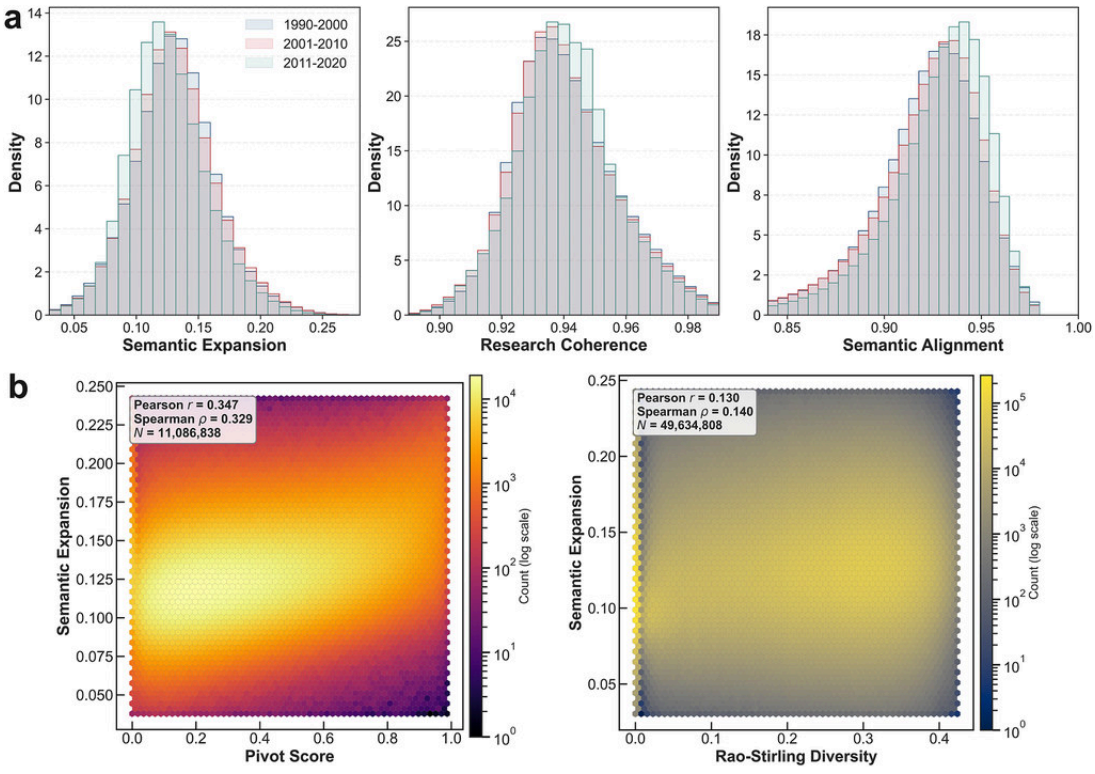


Fig. 3. Distributions of the three semantic measures and external validation of Semantic Expansion. (a) Distributions of Semantic Expansion (SE), Research Coherence (RC), and Semantic Alignment (SA) for publications published during 1990–2000, 2001–2010, and 2011–2020. (b) Relationships between publication-level SE and Pivot Score (left) and Rao–Stirling diversity (right). Color intensity represents the logarithm of observation density.

academic publications.

To assess SE, we compare it with existing bibliometric measures. At the publication level, SE correlates positively with the Pivot Score (Pearson $r = 0.347$, Spearman $\rho = 0.329$). This indicates that publications with higher SE more likely represent substantive shifts in a scholar's research direction, validating the proposed metric. In this sense, SE closely relates to the concept of pivoting (Hill et al., 2025), yet it directly observes this change within a continuous semantic space rather than identifying it through topic reclassification or disciplinary migration. Additionally, SE also correlates positively with Rao-Stirling diversity (Rao, 1982; Stirling, 2007). Overall, publications with high SE often involve broader knowledge combinations.

After the general analysis of these three metrics, we shift to an author-level analysis to explore their relationships with academic productivity. As illustrated in Fig. 4, the distribution of SE gradually shifts to the left slightly as annual productivity increases, indicating that scholars tend to exhibit lower SE during highly productive years. RC shows a similar pattern, indicating that scholars' research histories are more semantically dispersed in highly productive years. Meanwhile, SA shifts slightly to the right with increasing annual productivity, indicating that scholars maintain slightly stronger directional continuity with their preceding research trajectories during highly productive years.

We further observe that SE rises steadily with productivity ranking, suggesting that highly productive scholars, on average, tend to move further away from their prior research in the semantic space (Fig. 5a). Conversely, RC decreases as productivity ranking improves, indicating that highly productive scholars generally maintain less concentrated historical research trajectories compared to less productive scholars. SA also increases with annual productivity, although the increase is more modest than that observed for SE. Collectively, these three patterns suggest that highly productive scholars do not merely accumulate more papers within a narrow and internally concentrated research domain. Instead, while maintaining their established primary research frameworks, they tend to conduct broader expansion and exploration within the semantic space. When analyzing these three semantic metrics based on citation ranks, a similar stratification phenomenon emerges (Fig. 5b). As citation ranks increase, the average ranks of scholars in SE and SA rise steadily, whereas their average rank in RC declines. This pattern is highly consistent with the productivity-based findings.

The career trajectories of the three metrics further corroborate the overall productivity patterns mentioned above. From Fig. 5c, when observed by career stage, SE displays an upward trend over time, particularly among low and medium productivity scholars. In contrast, RC drops sharply after the early career stage and continues to decrease slowly thereafter. SA presents a more complex pattern. Among low productivity scholars, it decreases significantly with career age, whereas among high productivity scholars, it remains stable or even increases slightly. For highly productive scholars, broader exploration can coexist with relatively strong alignment to an evolving research direction. For less productive scholars, expansion more likely accompanies a decline in directional continuity. Therefore, a comprehensive interpretation of these three metrics reveals a nuanced structure of academic development where the breadth, coherence, and continuity of a scholar's research do not move in lockstep.

As shown in Fig. 5d, we also reveal the association between semantic trajectories and citation impact. Grouping scholars by citation level indicates that semantic expansion generally increases across all groups as career stages advance. However, the slope of this increase is steepest for less cited scholars and more gradual for highly cited individuals. RC exhibits a downward trend across all groups, with highly cited scholars often maintaining lower coherence in later stages compared to less cited scholars. This aligns with previous findings indicating that more impactful academic careers likely involve broader and less clustered research histories. SA indicates the most distinct stratification. Among highly cited scholars, it increases with career development and remains consistently higher than that of moderately and less cited scholars, whereas it declines significantly during the later stages for less cited scholars. This combination suggests that impactful academic development reflects not only broader semantic evolution but also the ability to maintain this evolution along a relatively coherent long-term direction. In other words, scholars who effectively reposition their research and expand directionally are more likely to achieve high impact.

4.2. Gender differences in semantic trajectories

The preceding analyses provide important context for examining gender differences in semantic trajectories. As shown in Fig. 6,

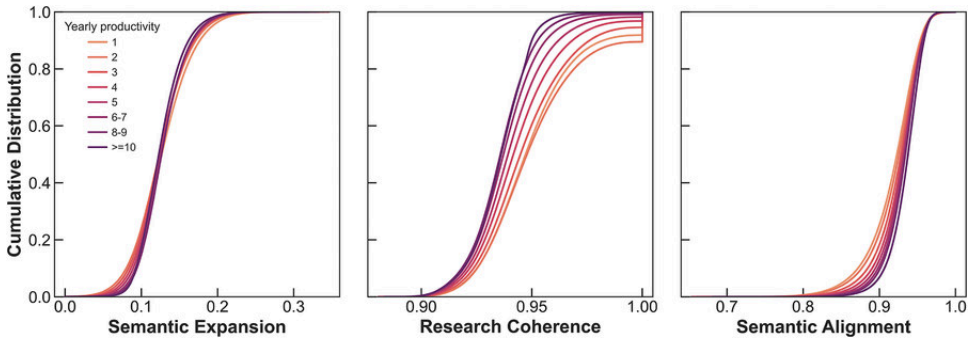


Fig. 4. Distributional shifts in semantic trajectory measures across yearly productivity groups. Cumulative distribution functions of Semantic Expansion, Research Coherence, and Semantic Alignment for publications grouped by the author's yearly productivity (1, 2, 3, 4, 5, 6–7, 8–9, and ≥10 papers).

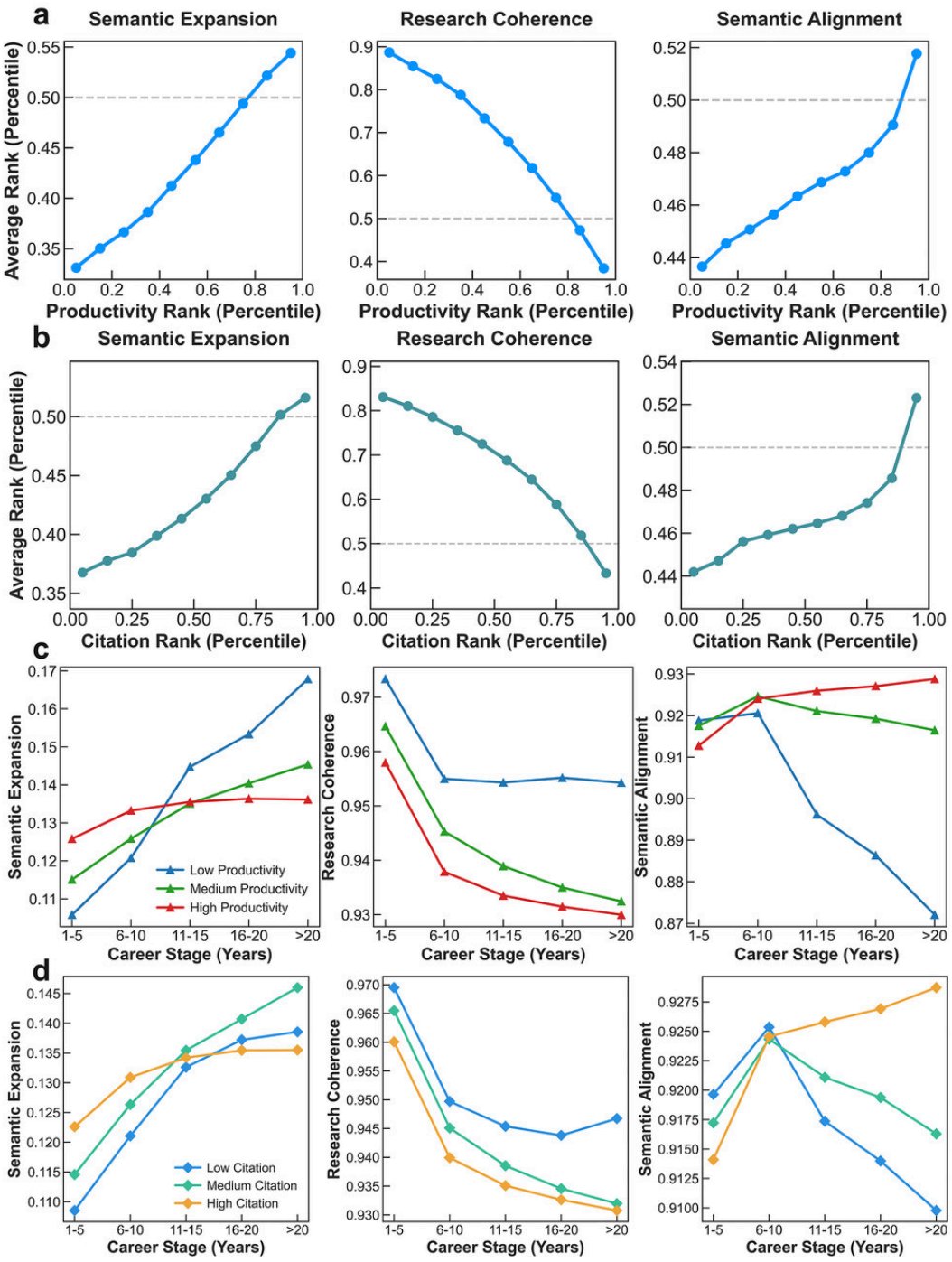


Fig. 5. Scholar-level semantic trajectory measures by productivity rank, citation rank, and career stage. (a) Mean ranks of Semantic Expansion (SE), Research Coherence (RC), and Semantic Alignment (SA) across productivity percentiles. (b) Mean ranks of the three semantic measures across citation percentiles. (c) Career-stage trajectories of SE, RC, and SA for low-, medium-, and high-productivity scholars. (d) Career-stage trajectories of SE, RC, and SA for low-, medium-, and high-citation scholars. Productivity and citation groups are defined by scholars' total counts, with the bottom 20% classified as low, the 20–95% range as medium, and the top 5% as high.

male and female scholars differ significantly on all three metrics. On average, male scholars show higher semantic expansion (SE), whereas female scholars show higher research coherence (RC) and semantic alignment (SA). These results indicate that gender disparities in academia appear not only in traditional outcomes such as publication counts and citations but also in the internal semantic structure of research careers. Overall, male scholars tend to explore a broader range of semantic space, while female scholars maintain

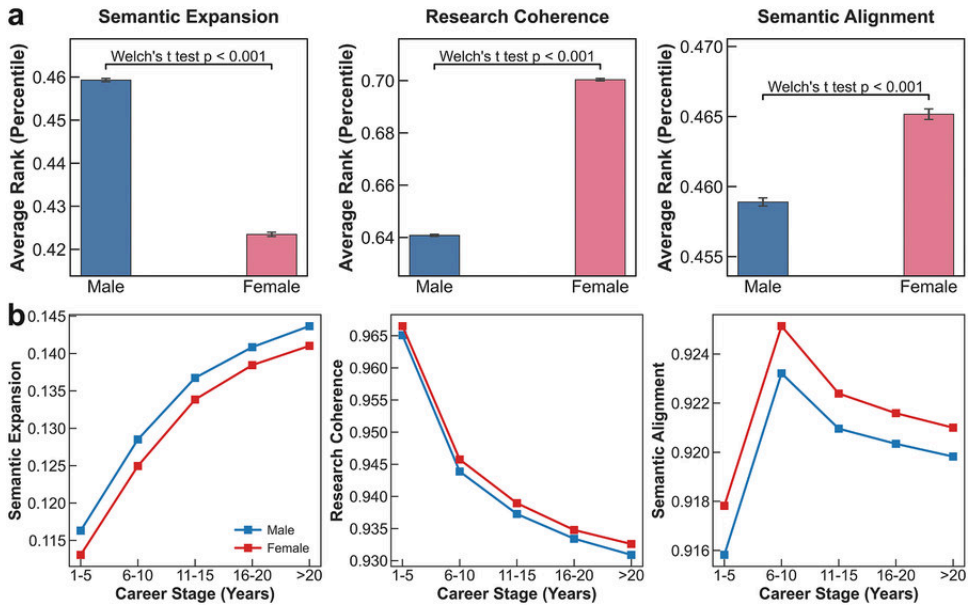


Fig. 6. Scholar-level semantic trajectory profiles by gender. (a) Average ranks of Semantic Expansion, Research Coherence, and Semantic Alignment for male and female scholars. Error bars indicate standard errors. All gender differences are significant by Welch's *t*-test ($p < 0.001$). (b) Gender differences in career-stage trajectories of the three semantic measures.

more concentrated research trajectories with stronger directional continuity between new and previous work.

These gender differences remain consistent across career stages. For both men and women, SE increases with career age, RC declines over time, and SA is highest in the early career before stabilizing or decreasing slowly. However, the relative positions of male and female scholars stay largely stable throughout most career stages. Male scholars consistently exhibit higher SE, while female scholars maintain slightly higher RC and SA. The persistence of these gaps suggests that gender differences in semantic trajectories are not limited to specific career transitions but reflect broader patterns in how academic careers develop. In general, women's publication histories are organized more tightly around a concentrated semantic core, whereas men show greater exploration in semantic space.

When scholars are grouped by productivity levels (Fig. 7), the gender difference in SE is smallest among low-productivity scholars and becomes more pronounced in the medium- and high-productivity groups, where men exhibit higher SE. For RC, female scholars score slightly higher than male scholars across nearly all productivity levels and career stages. This pattern indicates that the greater internal coherence of women's research histories is a stable feature rather than an outcome of differences in productivity. For SA, the gender gap is largest in the medium- and high-productivity groups, with female scholars showing stronger alignment with their prior research directions than male scholars in most career stages.

Overall, the three semantic measures exhibit systematic variation across productivity, citation impact, career stage, and gender. Male scholars consistently show higher SE, whereas female scholars show higher RC and SA across most career stages and productivity groups.

5. Early semantic trajectories and gendered dropout risk

5.1. Associations between early semantic trajectories and dropout risk

As illustrated in Fig. 8, Kaplan-Meier curves show that upon entering the risk set, female scholars consistently exhibit a lower survival probability than male scholars, a difference that is statistically significant ($p < 0.0001$, log-rank test). Although this gap is moderate in absolute terms, it persists throughout most of the observed career duration, indicating that women face a systematically higher risk of career dropout.

Table 2's results indicate a significant association between early semantic trajectories and later dropout risk ($p < 0.001$). Compared to the baseline model, incorporating each semantic metric improves the model fit, indicating that early semantic trajectories contain additional predictive information regarding long-term career persistence. Among these, early SE provided the strongest single semantic signal (AIC = 1,514,755). The three single-metric models offer the clearest substantive interpretations for each semantic dimension. Controlling for all other variables, SE is negatively associated with dropout risk, indicating that scholars whose early research spans a broader semantic scope are less likely to exit academia at later stages of their careers. In contrast, both RC and SA exhibit positive correlations with dropout risk. Consequently, scholars with highly concentrated early research paths or those who strictly follow established semantic directions face a higher probability of subsequent dropout.

Overall, the results indicate that early semantic trajectories provide additional explanatory information for later-stage career

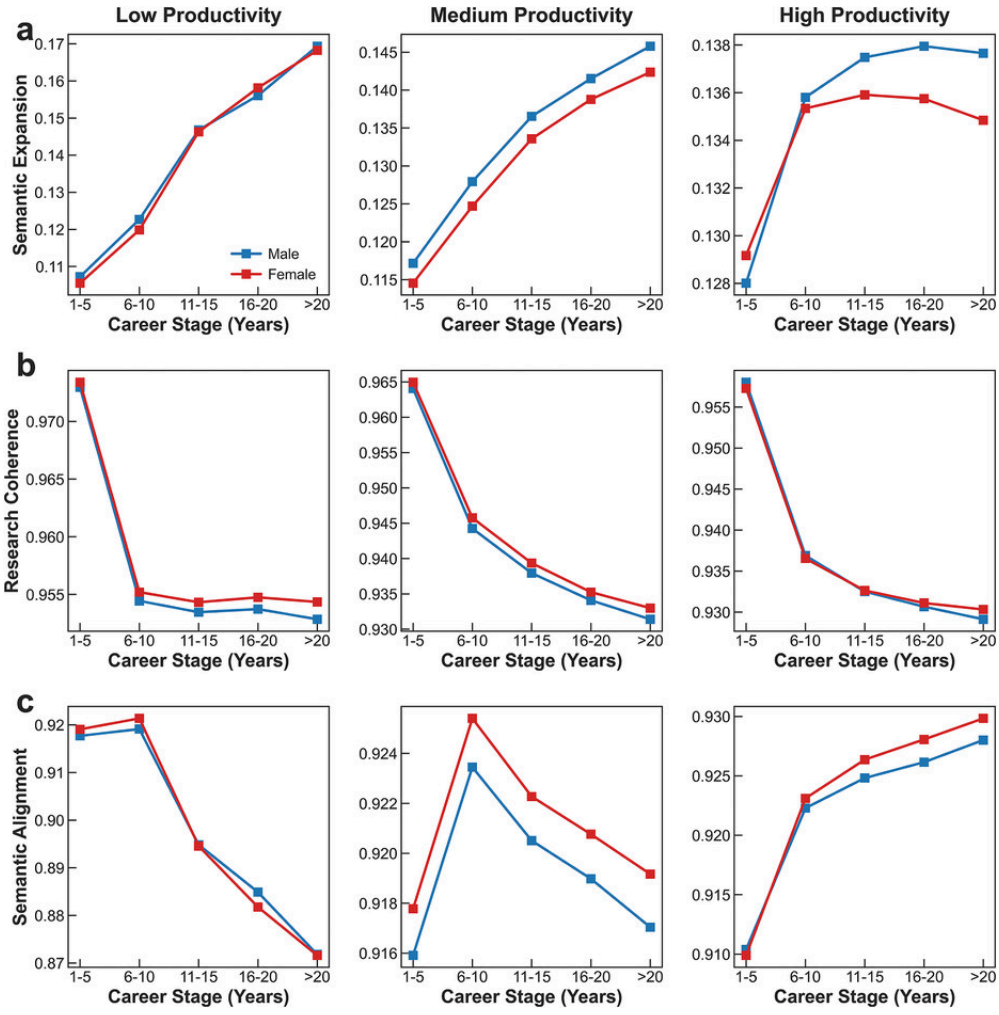


Fig. 7. Career-stage trajectories of the three semantic measures by gender and productivity group.

dropout risk beyond conventional career-related controls. Scholars' early exploratory research trajectories associate with career persistence, whereas the centralization of early paths and strong directional continuity associate with career vulnerability. In this sense, early semantic trajectories represent not only descriptive features of publication histories but also meaningful predictors of long-term academic career sustainability (Supplementary note 4).

5.2. Gender heterogeneity in the associations between early semantic trajectories and dropout risk

Having established the general link between early semantic trajectories and dropout risk, this section further explores the nuanced role played by gender.

Table 3's results indicate that the associations of these three semantic metrics are not consistent across gender groups. With male scholars as the reference group, the interaction coefficient indicates whether and by how much the association between each semantic metric and dropout risk differs for female scholars relative to male scholars. For SE, the interaction term is negative and statistically significant ($\beta = -0.0248$, $p < 0.001$). For RC, the interaction term is positive and significant ($\beta = 0.0371$, $p < 0.001$). For SA, the interaction term is also positive and significant ($\beta = 0.0190$, $p < 0.001$). Taken together, women exhibit a stronger negative correlation between SE and dropout risk, alongside stronger positive correlations for RC and SA.

To make the interaction results more interpretable, we further calculate the gender-specific associations of the three semantic metrics and present them as odds ratios (ORs) with 95% confidence intervals, as shown in Fig. 9. The results in Table 4 more directly elucidate the gender-specific associations between early semantic metrics and dropout risk.

Regarding SE, the odds ratios for both men and women are below 1, indicating that higher early SE is associated with a lower dropout risk. Notably, the estimated odds ratio for women (0.781) is lower than that for men (0.801). This suggests that the protective association of early SE is slightly stronger among female scholars. For RC, the odds ratios for both groups are greater than 1, confirming that higher RC in the early career stages correlates with an increased dropout risk. The estimated odds ratios are 1.268 for men and

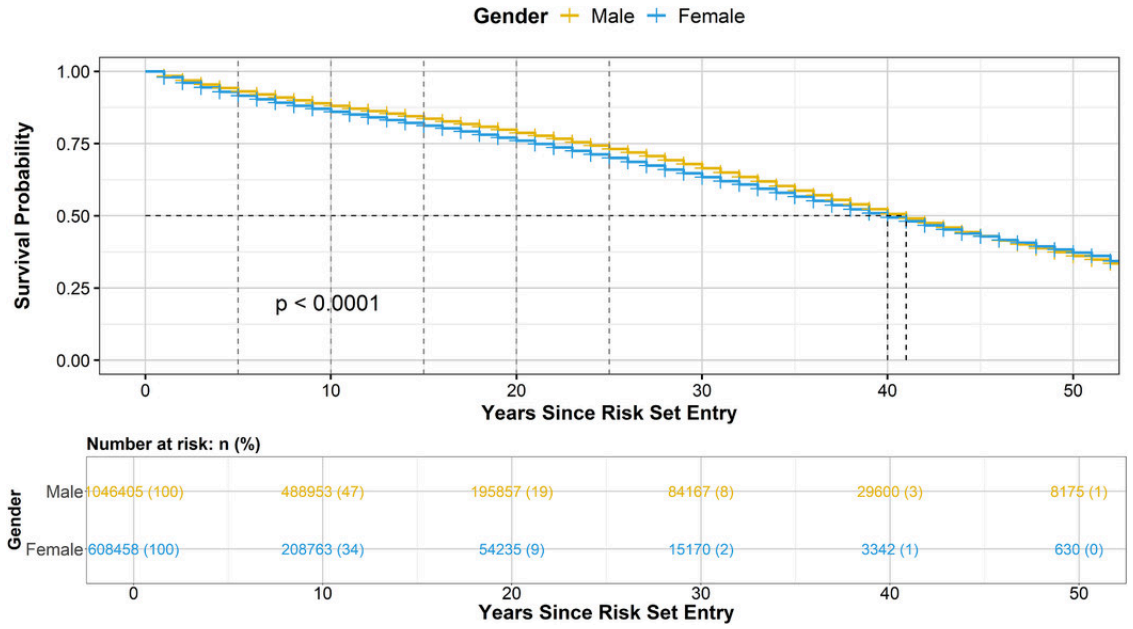


Fig. 8. Kaplan–Meier survival curves of career persistence by gender.

Table 2
Associations between early semantic trajectories and stage-specific dropout risk.

Variables	(1) Baseline	(2) SE	(3) RC	(4) SA
D.V.	EventInStage	EventInStage	EventInStage	EventInStage
Female = 1	0.2347***	0.2088***	0.2121***	0.2094***
	(0.0050)	(0.0051)	(0.0050)	(0.0051)
Productivity (z)	−0.2353***	−0.2379***	−0.0775***	−0.3162***
	(0.0040)	(0.0040)	(0.0039)	(0.0043)
Cohort (z)	−0.2683***	−0.2903***	−0.2849***	−0.2908***
	(0.0018)	(0.0018)	(0.0018)	(0.0018)
SE (z)		−0.2303***		
		(0.0022)		
RC (z)			0.2499***	
			(0.0026)	
SA (z)				0.2336***
				(0.0023)
First author ratio (z)	−0.1167***	−0.1386***	−0.1348***	−0.1386***
	(0.0025)	(0.0025)	(0.0025)	(0.0025)
Last author ratio (z)	0.0192***	0.0195***	0.0186***	0.0198***
	(0.0023)	(0.0023)	(0.0023)	(0.0023)
log(1 + Mean coauthor count) (z)	0.0351***	0.0399***	0.0394***	0.0396***
	(0.0030)	(0.0029)	(0.0029)	(0.0029)
Fund ratio (z)	−0.8374***	−0.8542***	−0.8498***	−0.8547***
	(0.0078)	(0.0079)	(0.0078)	(0.0079)
Stage FE	Yes	Yes	Yes	Yes
Field FE	Yes	Yes	Yes	Yes
Observations	3,604,030	3,604,030	3,604,030	3,604,030
AIC	1,524,765	1,514,755	1,516,020	1,515,014

Notes: robust standard errors are reported in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

1.316 for women. Consequently, as early research trajectories become more semantically concentrated, women face a higher risk of dropout. This finding also further corroborates that among the three semantic metrics, RC exhibits the highest sensitivity to gender differences. Similarly, SA exhibits a comparable pattern. The odds ratio is 1.255 for men and 1.279 for women. The slightly higher odds ratio for women implies that this risk-increasing association is also stronger among female scholars.

Overall, the gender-specific associations within each group are directionally consistent but vary in magnitude. SE is associated with a reduced dropout risk for both genders, exhibiting a stronger protective association for women. Conversely, RC and SA are positively correlated with dropout risk in both groups, with these associations being stronger among the female scholars. These findings suggest that early semantic trajectories are not only linked to overall career persistence, but also reveal gender differences in the strength of

Table 3
Gender heterogeneity in the associations of early semantic metrics on dropout risk.

Variables	(1) Gender × SE	(2) Gender × RC	(3) Gender × SA
D.V.	EventInStage	EventInStage	EventInStage
Female = 1	0.2074***	0.2056***	0.2095***
	(0.0051)	(0.0051)	(0.0051)
Productivity (z)	−0.2378***	−0.0779***	−0.3159***
	(0.0040)	(0.0039)	(0.0043)
Cohort (z)	−0.2900***	−0.2847***	−0.2906***
	(0.0018)	(0.0018)	(0.0018)
Semantic main effect	−0.2219***	0.2376***	0.2271***
	(0.0027)	(0.0031)	(0.0028)
Female × Semantic	−0.0248***	0.0371***	0.0190***
	(0.0047)	(0.0049)	(0.0047)
First author ratio (z)	−0.1387***	−0.1350***	−0.1387***
	(0.0025)	(0.0025)	(0.0025)
Last author ratio (z)	0.0195***	0.0186***	0.0198***
	(0.0023)	(0.0023)	(0.0023)
log(1 + Mean coauthor count) (z)	0.0396***	0.0392***	0.0394***
	(0.0029)	(0.0029)	(0.0029)
Fund ratio (z)	−0.8548***	−0.8501***	−0.8552***
	(0.0079)	(0.0078)	(0.0079)
Stage FE	Yes	Yes	Yes
Field FE	Yes	Yes	Yes
Observations	3,604,030	3,604,030	3,604,030
AIC	1,514,731	1,515,968	1,515,000
Δ Deviance vs. baseline	10,038.5	8,800.9	9,768.7

Notes: robust standard errors are reported in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

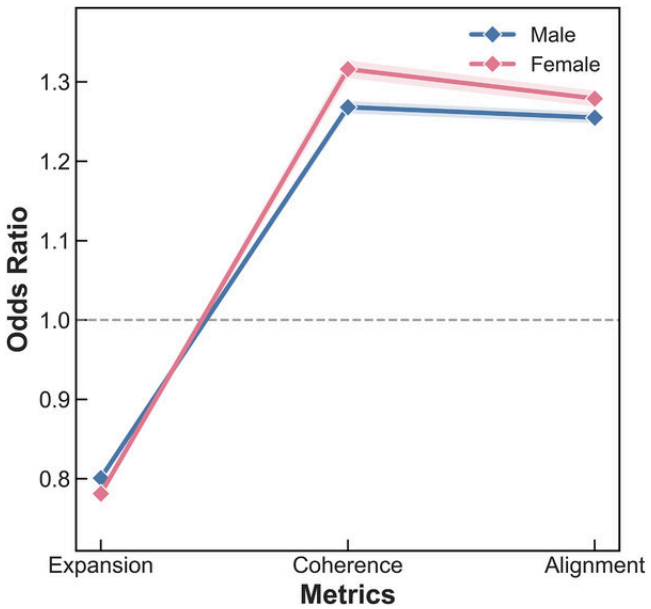


Fig. 9. Gender-specific odds ratios for early semantic metrics.

Table 4
Gender-specific associations of early semantic metrics on dropout risk.

Metrics	Male OR (95% CI)	Female OR (95% CI)
SE	0.801*** [0.797, 0.805]	0.781*** [0.776, 0.787]
RC	1.268*** [1.261, 1.276]	1.316*** [1.305, 1.327]
SA	1.255*** [1.248, 1.262]	1.279*** [1.269, 1.289]

Notes: OR and 95% confidence intervals are reported. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

these associations. Female scholars appear to benefit more from early semantic breadth, while also showing greater vulnerability to the risks associated with semantic concentration and directional continuity. Thus, these semantic metrics help identify gendered differences in how early research trajectories are associated with long-term career vulnerability.

5.3. Robustness analysis across domains

To test the robustness of our findings across disciplines, we re-estimate the interaction models separately for health, life and physical sciences. As shown in Table 5, the sample sizes across these disciplines are sufficiently large to enable meaningful robustness comparisons, ranging from 317,110 scholars in the life sciences to 574,670 scholars in the health sciences, with dropout event rates varying between 5.50 and 6.45%.

The domain-stratified estimates in Table 6 show patterns similar to those in the main analysis. Across all three disciplines, SE maintains a protective association with dropout risk for both male and female scholars, whereas RC and SA are still positively associated with dropout risk.

For SE, the interaction terms are negative and statistically significant in the health sciences ($\beta = -0.0669$, $p < 0.001$), life sciences ($\beta = -0.0346$, $p < 0.001$), and physical sciences ($\beta = -0.0490$, $p < 0.001$). The gender-stratified odds ratios presented in Table 7 offer a more direct interpretation of this pattern. Specifically, in the health sciences, the odds ratio is 0.847 for men and 0.793 for women; in the life sciences, it is 0.802 for men and 0.775 for women; and in the physical sciences, it is 0.733 for men and 0.698 for women. Consequently, across the disciplines and for both genders, a higher degree of early SE is associated with a lower dropout risk, and this negative correlation is consistently more pronounced among women. The magnitude of relevance varies across disciplinary fields, with the life sciences showing the smallest difference in the ratio, though the gender pattern remains consistent.

For RC, the interaction terms are positive and statistically significant in the health sciences ($\beta = 0.0683$, $p < 0.001$), life sciences ($\beta = 0.0285$, $p < 0.01$), and physical sciences ($\beta = 0.0533$, $p < 0.001$). The gender-stratified odds ratio is 1.192 for men and 1.277 for women in the health science, 1.286 and 1.324 in life sciences, 1.385 and 1.460 in physical sciences. The association between an early semantically concentrated research path and later dropout risk is not confined to a single disciplinary context. Instead, the stronger positive correlation observed among female scholars is validated across the health, life, and physical sciences.

The results for the SA remain consistent with the previous analysis. In the health sciences ($\beta = 0.0631$, $p < 0.001$), life sciences ($\beta = 0.0369$, $p < 0.001$), and physical sciences ($\beta = 0.0443$, $p < 0.001$), the interaction terms are positive and statistically significant, indicating that the positive association between strong alignment with prior research directions and later dropout risk is stronger for women than for men. The gender-stratified odds ratio is 1.186 for men and 1.263 for women in health sciences, 1.249 and 1.296 in life sciences, 1.377 and 1.440 in physical sciences. Stronger continuity along established semantic directions is associated with a higher dropout risk for both genders, and this positive correlation remains consistently more pronounced among female scholars.

Overall, the discipline-stratified analysis supports the robustness of the main findings. More importantly, the patterns of gender differences observed in the primary analysis remain stable when disaggregated by discipline. Although the magnitude of these associations varies across the health, life, and physical sciences, their directions and gender difference remain highly consistent. These results suggest that our findings are not solely driven by the disciplinary composition of the dataset. Rather, they reflect a broader cross-disciplinary pattern linking early semantic trajectories to gendered career vulnerability.

6. Discussion

6.1. Theoretical implications

This study advances research on academic careers and gender inequality by indicating that gender disparities extend beyond observable outputs such as publication volume and citations to the internal semantic organization of scholarly work. By conceptualizing academic careers as trajectories within a semantic space, our analysis reveals differences in how male and female scholars develop their research over time. We find that male scholars tend to exhibit higher SE, whereas female scholars incline toward maintaining higher RC and SA. These patterns indicate that male and female scholars follow divergent pathways in constructing and evolving their academic careers.

Simultaneously, comparable research performance does not necessarily translate into equal opportunities for research funding, scholarly recognition, promotion, or career rewards, which is a manifestation of the “productivity-impact paradox” (Huang et al., 2020; Ross et al., 2022; Weisshaar, 2017; Witteman et al., 2019). The higher coherence and alignment observed among female scholars may reflect research strategies formulated in evaluative environments characterized by greater career uncertainty, restricted access to resources, or where deviating from established research topics is perceived as high-risk. Parenthood-related caregiving demands can also disproportionately restrict women’s time and capacity for sustained research participation and are associated with

Table 5
Domain-stratified sample composition.

Domain	Scholars	Dropout events	Event rate (%)	Male	Female
Health Sciences	574,670	76,811	5.50	326,121	248,549
Physical Sciences	495,964	68,860	5.99	364,204	131,760
Life Sciences	317,110	50,131	6.45	184,149	132,961

Table 6
Domain-stratified gender heterogeneity in the associations between early semantic trajectories and dropout risk.

Domain	(1) Female × SE	(2) Female × RC	(3) Female × SA
Health Sciences			
Female × Semantic	−0.0669***	0.0683***	0.0631***
	(0.0074)	(0.0078)	(0.0075)
Observations	1,396,804	1,396,804	1,396,804
Life Sciences			
Female × Semantic	−0.0346***	0.0285**	0.0369***
	(0.0094)	(0.0102)	(0.0096)
Observations	777,270	777,270	777,270
Physical Sciences			
Female × Semantic	−0.0490***	0.0533***	0.0443***
	(0.0093)	(0.0097)	(0.0096)
Observations	1,149,468	1,149,468	1,149,468

Notes: robust standard errors are reported in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 7
Domain-stratified gender-specific associations of early semantic indicators on dropout risk.

Domain	Metric	Male OR (95% CI)	Female OR (95% CI)
Health Sciences	Semantic Expansion	0.847 [0.840, 0.855]	0.793 [0.784, 0.802]
	Research Coherence	1.192 [1.181, 1.205]	1.277 [1.261, 1.293]
	Semantic Alignment	1.186 [1.175, 1.197]	1.263 [1.249, 1.278]
Life Sciences	Semantic Expansion	0.802 [0.793, 0.812]	0.775 [0.764, 0.787]
	Research Coherence	1.286 [1.268, 1.305]	1.324 [1.301, 1.347]
	Semantic Alignment	1.249 [1.234, 1.264]	1.296 [1.276, 1.315]
Physical Sciences	Semantic Expansion	0.733 [0.727, 0.739]	0.698 [0.687, 0.709]
	Research Coherence	1.385 [1.371, 1.398]	1.460 [1.435, 1.486]
	Semantic Alignment	1.377 [1.365, 1.390]	1.440 [1.416, 1.464]

more adverse career consequences for mothers than for fathers in academia and STEM employment (Cech & Blair-Loy, 2019; Morgan et al., 2021). In this context, while a highly focused research agenda facilitates the establishment of a distinct scholarly identity, it also renders career progression more vulnerable to field-specific fluctuations in funding, job availability, and shifting preferences in academic evaluation. Correspondingly, the higher SE among male scholars does not necessarily merely represent a stronger personal inclination for exploration. It may also reflect gender differences in access to interdisciplinary collaboration, team-based resources, and institutional support.

The semantic trajectory perspective adopted in this study adds a new dimension to the classical "leaky pipeline" literature, which emphasizes the cumulative loss of women across successive stages of academic careers (Alper, 1993; Spoon et al., 2023). The results of this study indicate that similar forms of research development correspond to different degrees of career vulnerability across gender groups. Prior research shows that women may experience lower tenure returns from co-authored papers compared to their male counterparts, especially when they collaborate with male colleagues. (Sarsons, 2017). Such asymmetries in recognition and resource allocation suggest that similar patterns of research development may not generate equivalent career returns for female and male scholars. From this perspective, higher SE may be more strongly associated with persistence among women when expansion is accompanied by greater access to collaborative and institutional opportunities. By contrast, high coherence or alignment may be associated with greater career vulnerability when it reflects limited opportunities to redirect or broaden a research agenda.

Methodologically, the introduction of semantic trajectory offers a reproducible framework for analyzing the dynamic development of individual research programs. Recent studies have investigated gender disparities not only through traditional bibliometric indicators but also via text-based analysis, focusing on dimensions such as research topic transitions, gendered research content, and citation flows (Kim et al., 2022; Yang, 2025; Zhou et al., 2024). Unlike static pairwise similarity measures or discrete topic classifications, these metrics capture temporal order, cumulative direction, and movement within a continuous semantic space. Taking SE as an example, a higher SE indicates that a scholar's work connects previously separated knowledge domains, research topics, or even scholarly communities. When such semantically expanded work is later cited and absorbed by subsequent studies, SE may reflect the potential diffusion of scientific innovation. Consequently, these semantic trajectory metrics serve as a distinct complement to traditional bibliometric metrics. Traditional metrics reveal how much scholars have published, the degree of recognition their work has garnered, and with whom they collaborate, semantic trajectory metrics further explain what they are investigating and how their research agendas evolve over time.

Ultimately, by identifying gender differences in the boundaries of knowledge exploration and research pathways, this study advances the discourse on academic career inequality from disparities in external outcomes to internal structural differences within the knowledge production process itself.

6.2. Practical implications

The results have several implications for academic institutions, funding agencies, and evaluation practices. First, they suggest that career support programs and mentoring initiatives should pay attention not only to productivity and impact but also to the development of research agendas. Early academic conditions, including publication performance and access to collaboration opportunities, are closely associated with scholars' subsequent career trajectories and long-term academic outcomes (Duan et al., 2025; Li et al., 2019). Institutions and funding agencies should avoid assessing early-career scholars solely on short-term output or strict alignment with established topics. Instead, evaluation and promotion systems should be sufficiently inclusive to accommodate diverse research rhythms and exploratory pathways. Short-term shifts in research focus, interdisciplinary exploration, or early-career setbacks should not be automatically interpreted as evidence of instability or inadequate scholarly potential (Wang et al., 2019).

Second, the observed gender heterogeneity highlights the need to examine whether similar research trajectories receive comparable institutional support and career returns across genders. Rather than encouraging or discouraging particular research styles, institutions and review committees should consider whether differences in access to funding, collaboration opportunities, recognition, and mentoring make some trajectories more difficult to sustain for women. They should also assess whether apparently neutral evaluation procedures generate unequal judgments or resource allocations. This concern is consistent with prior evidence that apparently gender-equal funding outcomes can coexist with unequal underlying evaluations, suggesting that review systems may reproduce bias even when opportunities appear balanced (Bol et al., 2022). Institutions and funding agencies could use semantic trajectory metrics as supplementary diagnostic tools alongside traditional evaluation metrics, thereby supporting a more comprehensive understanding of scholars' research development. Greater awareness of these patterns may help reduce unintended gender disparities in evaluation and career development.

More broadly, this study underscores the value of open, large-scale bibliographic data and embedding models for monitoring and understanding inequalities in science. By making semantic trajectory metrics computable at scale, the approach developed here can be applied in future work to evaluate the effectiveness of diversity initiatives, track changes in research cultures over time, and inform evidence-based policies aimed at improving academic career sustainability for all scholars.

6.3. Limitations

This study has several limitations. First, although OpenAlex provides broad coverage of global scholarly production, its coverage and author-disambiguation procedures remain imperfect. Certain publication types, non-English-language outlets, and fields with limited database indexing may be underrepresented, while disambiguation errors may incorrectly merge or separate author profiles, particularly for scholars with common names, institutional mobility, or changes in publication identity. Second, the semantic embeddings are based solely on titles and abstracts. Full-text analysis could capture richer semantic information but was not feasible at this scale. As a result, the semantic trajectory metrics may underrepresent conceptual nuances that appear only in methods, results, or discussion sections. Third, the definition of dropout as five consecutive years without publications is reasonable but may misclassify scholars who temporarily pause publishing due to parental leave, administrative roles, or shifts to non-academic positions. Although our survival models account for right-censored observations, the long-term career status of scholars whose latest publications are highly concentrated near the end of the data-collection window may still require longer-term data tracking to be fully confirmed. Finally, this study is observational and identifies statistical associations rather than causal effects. Unobserved factors, such as institutional resources, mentoring quality, and family responsibilities, may jointly influence both the formation of semantic trajectories and career persistence.

6.4. Future research directions

Future work can extend this research in several directions. Full-text embeddings or multimodal representations that include figures, tables, and supplementary materials may provide deeper semantic content. Moreover, future studies could examine how semantic trajectories interact with collaboration networks, funding patterns, and institutional contexts. Comparative analysis across countries and academic systems would help determine the extent to which the observed gender patterns are universal or shaped by national policies and cultural norms. Detailed longitudinal data, natural experiments, or other research designs that permit stronger causal identification could further clarify whether changes in research trajectories influence career persistence, or whether both are driven by shared institutional and personal conditions. Finally, future research could also explore the phenomenon of career re-entry following a period of publication inactivity. Such "re-entry" can be conceptualized as a subsequent stage following academic dropout. However, these transitions may not be primarily driven by a scholar's prior research trajectory. They are likely influenced by external factors, including changes in employment status, institutional affiliations, family circumstances, or professional roles. Examining these transitions would provide a more comprehensive understanding of the interplay between semantic trajectories and the broader career experiences of scholars.

CRediT authorship contribution statement

Letian Hu: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Alex J. Yang:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Methodology, Funding acquisition, Formal analysis. **Sanhong Deng:** Writing – review & editing,

Supervision, Resources, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the National Social Science Fund of China (25ATQ008) and the National Natural Science Foundation of China (724B2015).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ipm.2026.105142.

Data availability

Data will be made available on request.

References

Abramo, G., Aksnes, D. W., & D'Angelo, C. A (2021). Gender differences in research performance within and between countries: Italy vs Norway. *Journal of Informetrics, 15*(2), Article 101144. https://doi.org/10.1016/j.joi.2021.101144

Alper, J. (1993). The pipeline is leaking women all the way along. *Science (New York, N.Y.), 260*(5106), 409–411. https://doi.org/10.1126/science.260.5106.409

Arts, S., Melluso, N., & Veugelers, R. (2025). Beyond citations: Measuring novel scientific ideas and their impact in publication text. *The Review of Economics and Statistics*, 1–33. https://doi.org/10.1162/rest_a_01561

Bol, T., de Vaan, M., & van de Rijt, A. (2022). Gender-equal funding rates conceal unequal evaluations. *Research Policy, 51*(1), Article 104399. https://doi.org/10.1016/j.respol.2021.104399

Box-Steffensmeier, J. M., Cunha, R. C., Varbanov, R. A., Hoh, Y. S., Knisley, M. L., & Holmes, M. A. (2015). Survival analysis of faculty retention and promotion in the social sciences by gender. *PLOS ONE, 10*(11), Article e0143093. https://doi.org/10.1371/journal.pone.0143093

Boyack, K. W., & Klavans, R. (2020). A comparison of large-scale science models based on textual, direct citation and hybrid relatedness. *Quantitative Science Studies, 1*(4), 1570–1585. https://doi.org/10.1162/qss_a_00085

Cech, E. A., & Blair-Loy, M. (2019). The changing career trajectories of new parents in STEM. *Proceedings of the National Academy of Sciences, 116*(10), 4182–4187. https://doi.org/10.1073/pnas.1810862116

Chan, H. F., & Torgler, B. (2020). Gender differences in performance of top cited scientists by field and country. *Scientometrics, 125*(3), 2421–2447. https://doi.org/10.1007/s11192-020-03733-w

Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D. (2020). SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 2270–2282). https://doi.org/10.18653/v1/2020.acl-main.207

Duan, Y., Memon, S. A., AlShebli, B., Guan, Q., Holme, P., & Rahwan, T. (2025). Postdoc publications and citations link to academic retention and faculty success. *Proceedings of the National Academy of Sciences, 122*(4), Article e2402053122. https://doi.org/10.1073/pnas.2402053122

Ecklund, E. H., & Lincoln, A. E. (2011). Scientists want more children. *PLOS ONE, 6*(8), Article e22590. https://doi.org/10.1371/journal.pone.0022590

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., & Barabási, A.-L. (2018). Science of science. *Science (New York, N.Y.), 359*(6379), Article eaao0185. https://doi.org/10.1126/science.aao0185

Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and innovation in scientists' research strategies. *American Sociological Review, 80*(5), 875–908. https://doi.org/10.1177/0003122415601618

Frandsen, T. F., & Nicolaisen, J. (2024). Defining the early career researcher: A study of publication-based definitions. *Learned Publishing, 37*(4), Article e1621. https://doi.org/10.1002/leap.1621

Hill, R., Yin, Y., Stein, C., Wang, X., Wang, D., & Jones, B. F. (2025). The pivot penalty in research. *Nature, 642*(8069), 999–1006. https://doi.org/10.1038/s41586-025-09048-1

Hofstra, B., Kulkarni, V. V., Munoz-Najar Galvez, S., He, B., Jurafsky, D., & McFarland, D. A (2020). The diversity–Innovation paradox in science. *Proceedings of the National Academy of Sciences, 117*(17), 9284–9291. https://doi.org/10.1073/pnas.1915378117

Huang, J., Gates, A. J., Sinatra, R., & Barabási, A.-L. (2020). Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences, 117*(9), 4609–4616. https://doi.org/10.1073/pnas.1914221117

Huang, S., Lu, W., Cheng, Q., Luo, Z., & Huang, Y. (2024). Evolutions of semantic consistency in research topic via contextualized word embedding. *Information Processing & Management, 61*(6), Article 103859. https://doi.org/10.1016/j.ipm.2024.103859

Jia, T., Wang, D., & Szymanski, B. K. (2017). Quantifying patterns of research-interest evolution. *Nature Human Behaviour, 1*(4), 0078. https://doi.org/10.1038/s41562-017-0078

Kim, L., Smith, D. S., Hofstra, B., & McFarland, D. A. (2022). Gendered knowledge in fields and academic careers. *Research Policy, 51*(1), Article 104411. https://doi.org/10.1016/j.respol.2021.104411

Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review, 84*(5), 905–949. https://doi.org/10.1177/0003122419877135

Kozlowski, D., Larivière, V., Sugimoto, C. R., & Monroe-White, T. (2022). Intersectional inequalities in science. *Proceedings of the National Academy of Sciences, 119*(2), Article e2113067119. https://doi.org/10.1073/pnas.2113067119

Kwiek, M., & Roszka, W. (2021). Gender disparities in International research collaboration: A study of 25,000 university professors. *Journal of Economic Surveys, 35*(5), 1344–1380. https://doi.org/10.1111/joes.12395

Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Bibliometrics: Global gender disparities in science. *Nature, 504*(7479), 211–213. https://doi.org/10.1038/504211a

Lerchenmueller, M. J., & Sorenson, O. (2018). The gender gap in early career transitions in the life sciences. *Research Policy, 47*(6), 1007–1017. https://doi.org/10.1016/j.respol.2018.02.009

Li, W., Aste, T., Caccioli, F., & Livan, G. (2019). Early coauthorship with top scientists predicts success in academic careers. *Nature Communications, 10*(1), 5170. https://doi.org/10.1038/s41467-019-13130-4

Liu, L., Dehmamy, N., Chown, J., Giles, C. L., & Wang, D. (2021). Understanding the onset of hot streaks across artistic, cultural, and scientific careers. *Nature Communications, 12*(1), 5392. https://doi.org/10.1038/s41467-021-25477-8

Mihaljević-Brandt, H., Santamaría, L., & Tullney, M. (2016). The effect of gender in the publication patterns in mathematics. *PLOS ONE, 11*(10), Article e0165367. https://doi.org/10.1371/journal.pone.0165367

Milojević, S., Radicchi, F., & Walsh, J. P. (2018). Changing demographics of scientific careers: The rise of the temporary workforce. *Proceedings of the National Academy of Sciences, 115*(50), 12616–12623. https://doi.org/10.1073/pnas.1800478115

Morgan, A. C., Way, S. F., Hoefer, M. J. D., Larremore, D. B., Galesic, M., & Clauset, A. (2021). The unequal impact of parenthood in academia. *Science Advances, 7*(9), Article eabd1996. https://doi.org/10.1126/sciadv.abd1996

Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2023). MTEB: Massive text embedding benchmark. In *Proceedings of the 17th conference of the european chapter of the association for computational linguistics* (pp. 2014–2037). https://doi.org/10.18653/v1/2023.eacl-main.148

Peng, H., Ke, Q., Budak, C., Romero, D. M., & Ahn, Y.-Y. (2021). Neural embeddings of scholarly periodicals reveal complex disciplinary organizations. *Science Advances, 7*(17), Article eabb9004. https://doi.org/10.1126/sciadv.abb9004

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv Preprint*, arXiv: 2205.01833.

Rao, C. R. (1982). Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology, 21*(1), 24–43. https://doi.org/10.1016/0040-5809(82)90004-1

Régner, I., Thinus-Blanc, C., Netter, A., Schmader, T., & Huguet, P. (2019). Committees with implicit biases promote fewer women when they do not believe gender bias exists. *Nature Human Behaviour, 3*(11), 1171–1179. https://doi.org/10.1038/s41562-019-0686-3

Ross, M. B., Glennon, B. M., Murciano-Goroff, R., Berkes, E. G., Weinberg, B. A., & Lane, J. I. (2022). Women are credited less in science than men. *Nature, 608*(7921), 135–145. https://doi.org/10.1038/s41586-022-04966-w

Sarsons, H. (2017). Recognition for group work: Gender differences in academia. *American Economic Review, 107*(5), 141–145. https://doi.org/10.1257/aer.p20171126

Shen, H., Xie, J., Ao, W., & Cheng, Y. (2022). The continuity and citation impact of scientific collaboration with different gender composition. *Journal of Informetrics, 16*(1), Article 101248. https://doi.org/10.1016/j.joi.2021.101248

Sinatra, R., Wang, D., Deville, P., Song, C., & Barabási, A.-L. (2016). Quantifying the evolution of individual scientific impact. *Science (New York, N.Y.), 354*(6312), Article aaf5239. https://doi.org/10.1126/science.aaf5239

Singh, A., D'Arcy, M., Cohan, A., Downey, D., & Feldman, S. (2023). SciRepEval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 5548–5566). https://doi.org/10.18653/v1/2023.emnlp-main.338

Singh, C. K., Tupikina, L., Lécuyer, F., Starnini, M., & Santolini, M. (2024). Charting mobility patterns in the scientific knowledge landscape. *EPJ Data Science, 13*(1), 12. https://doi.org/10.1140/epjds/s13688-024-00451-8

Spoon, K., LaBerge, N., Wapman, K. H., Zhang, S., Morgan, A. C., Galesic, M., … Clauset, A. (2023). Gender and retention patterns among US faculty. *Science Advances, 9*(42), eadi2205. https://doi.org/10.1126/sciadv.adi2205

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface, 4*(15), 707–719. https://doi.org/10.1098/rsif.2007.0213

Thelwall, M., Bailey, C., Tobin, C., & Bradshaw, N.-A. (2019). Gender differences in research areas, methods and topics: Can people and thing orientations explain the results? *Journal of Informetrics, 13*(1), 149–169. https://doi.org/10.1016/j.joi.2018.12.002

Tripodi, G., Zheng, X., Qian, Y., Murray, D., Jones, B. F., Ni, C., & Wang, D. (2025). Tenure and research trajectories. *Proceedings of the National Academy of Sciences, 122*(30), Article e2500322122. https://doi.org/10.1073/pnas.2500322122

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G., & Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature, 571*(7763), 95–98. https://doi.org/10.1038/s41586-019-1335-8

Van Buskirk, I., Clauset, A., & Larremore, D. B. (2023). An open-source cultural consensus approach to name-based gender classification. *Proceedings of the International AAAI Conference on Web and Social Media, 17*(1), 866–877. https://doi.org/10.1609/icwsm.v17i1.22195

Wang, Y., Jones, B. F., & Wang, D. (2019). Early-career setback and future career impact. *Nature Communications, 10*(1), 4331. https://doi.org/10.1038/s41467-019-12189-3

Weisshaar, K. (2017). Publish and perish? An assessment of gender gaps in promotion to tenure in academia. *Social Forces, 96*(2), 529–560. https://doi.org/10.1093/sf/sox052

Witteman, H. O., Hendricks, M., Straus, S., & Tannenbaum, C. (2019). Are gender gaps due to evaluations of the applicant or the science? A natural experiment at a national funding agency. *Lancet (London, England), 393*(10171), 531–540. https://doi.org/10.1016/S0140-6736(18)32611-4

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature, 566*(7744), 378–382. https://doi.org/10.1038/s41586-019-0941-9

Yang, A. J. (2025). Unraveling topic switching and innovation in science. *Information Processing & Management, 62*(4), Article 104171. https://doi.org/10.1016/j.ipm.2025.104171

Yang, A. J., Freeman, R. B., & Deng, S. (2026). A semantic atlas of journals: Structure, position, and dispersion. *Journal of the Association for Information Science and Technology, 77*(6), 831–847. https://doi.org/10.1002/asi.70067

Yang, A. J., Zhang, Y., Wang, Z., Wang, H., & Deng, S. (2025a). Quantifying delayed recognition of scientists. *Journal of Informetrics, 19*(3), Article 101688. https://doi.org/10.1016/j.joi.2025.101688

Yang, Y., Zhang, C., Xu, H., Bu, Y., Liu, M., & Ding, Y. (2024). Unveiling the loss of exceptional women in science. *Information Processing & Management, 61*(6), Article 103829. https://doi.org/10.1016/j.ipm.2024.103829

Yang, Y., Zhang, C., Xu, H., Bu, Y., Liu, M., & Ding, Y. (2025b). Gender differences in dropout rate: From field, career status, and generation perspectives. *Journal of Informetrics, 19*(1), Article 101615. https://doi.org/10.1016/j.joi.2024.101615

Yu, D., Bao, R., Ning, R., Peng, J., Mai, G., & Zhao, L. (2026). Spatial-RAG: Spatial retrieval augmented generation for real-world geospatial reasoning questions. In *Findings of the association for computational linguistics: ACL 2026* (pp. 11094–11112). https://doi.org/10.18653/v1/2026.findings-acl.539

Zeng, A., Shen, Z., Zhou, J., Fan, Y., Di, Z., Wang, Y., Stanley, H. E., & Havlin, S. (2019). Increasing trend of scientists to switch between topics. *Nature Communications, 10*(1), 3439. https://doi.org/10.1038/s41467-019-11401-8

Zhang, L., Sivertsen, G., Du, H., Huang, Y., & Glänzel, W. (2021). Gender differences in the aims and impacts of research. *Scientometrics, 126*(11), 8861–8886. https://doi.org/10.1007/s11192-021-04171-y

Zhou, S., Chai, S., & Freeman, R. B. (2024). Gender homophily: In-group citation preferences and the gender disadvantage. *Research Policy, 53*(1), Article 104895. https://doi.org/10.1016/j.respol.2023.104895