Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Semantic Novelty, Influence, and Patent Persistence in Robotics Innovation: Embedding-Native Indicators and Maintenance-Fee Validation

Chong Guan
Singapore University of Social Sciences
Yuchao Cheng
jackchengyc@suss.edu.sg

Singapore University of Social Sciences

Research Article

Keywords: robotics patents, transformer embeddings, semantic novelty, patent analytics, maintenance-fee lapse, semantic influence

Posted Date: September 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10480137/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Patent documents provide a systematic record of inventive activity, but large-scale technology mapping often remains disconnected from behavioural indicators of patent persistence. This study develops a transformer-based patent-analytics framework for examining the structure, novelty, influence, and maintenance behaviour of robotics patents. Using 19,796 United States robotics utility grants published from 2006 to 2025, the study represents Derwent World Patents Index (DWPI) patent abstracts with SciBERT embeddings and applies k-means clustering to identify five technological domains: robotic mechanisms and actuation, articulated motion and manipulation, sensing, perception and autonomous control, intelligent systems and field robotics, and navigation and environment interaction. The analysis then derives embedding-native indicators of backward semantic novelty, forward semantic influence, and cluster typicality, and links them to dated failure-to-pay maintenance-fee events in a discrete-time fee-window panel. The corpus shows strong growth in robotics patenting and a compositional movement toward sensing, autonomy, and system-facing domains. Novelty is highest in sensing, perception and autonomous control, while forward semantic influence is highest in intelligent systems and field robotics. Maintenance-fee models show that forward semantic influence is associated with a significantly lower likelihood of fee lapse. Novelty alone is not associated with retention, but the novelty × typicality interaction and a concave novelty term indicate that novelty matters when it remains technically legible within a recognisable robotics domain. The study contributes a reproducible framework for linking transformer-based semantic indicators to observable patent-maintenance behaviour.

1. Introduction

Robotics has become a foundational technology for contemporary automation systems. Industrial robots, service robots, field robots, surgical systems, autonomous vehicles, inspection platforms, and collaborative robots increasingly combine mechanical embodiment with sensing, perception, control, and software-based decision making. As robotic systems diffuse across manufacturing, healthcare, logistics, agriculture, domestic services, and infrastructure inspection, understanding the structure and evolution of robotics innovation has become a central priority for technology intelligence and scientometric analysis (International Federation of Robotics, 2025).

Patents are a natural empirical source for studying robotics innovation. They provide dated and structured records of inventive activity across organisations and inventors, and have long been used to study technological trajectories, knowledge diffusion, and economic value (Griliches, 1990; Hall et al., 2005; Trajtenberg, 1990). Yet conventional patent-landscape methods are limited when technologies are semantically complex. Classification codes such as the International Patent Classification (IPC) and Cooperative Patent Classification (CPC) provide useful formal categories, but they do not fully capture how robotics inventions combine actuation, manipulation, perception, planning, mobility, human interaction, and deployment contexts (Leydesdorff et al., 2014; Momeni & Rost, 2016). Keyword-based approaches face similar limits because the same terms may appear across multiple subsystems, while emerging combinations may not fit predefined taxonomies.

Recent advances in transformer-based language models provide a way to analyse patent text directly in a high-dimensional semantic space. Domain-adapted models such as SciBERT can represent technical documents based on contextual meaning rather than surface keyword overlap (Beltagy et al., 2019). Text-based similarity measures also offer a way to capture technological relatedness that is not fully represented by classification systems (Arts et al., 2018). However, semantic maps alone remain descriptive. They identify what clusters exist and how large they are, but they do not show whether patents in different semantic positions are subsequently built upon or whether owners make different maintenance decisions about them.

This study addresses that gap by linking transformer-based patent-text indicators to maintenance-fee behaviour. It asks whether semantic position within the robotics patent landscape is associated with observable decisions to keep patents in force. The analysis begins by mapping robotics patents using SciBERT embeddings and k-means clustering. It then derives three patent-level indicators from the same embedding space. Backward semantic novelty measures distance from earlier patents in the robotics corpus, following the broader logic of patent-text novelty measurement (Kelly et al., 2021; Verhoeven et al., 2016). Forward semantic influence measures whether later patents appear near the focal patent in the semantic space. Cluster typicality measures how centrally a patent sits within its assigned robotics domain. These indicators are then linked to dated failure-to-pay maintenance-fee events observed at the three United States fee-window decisions.

The empirical setting is a corpus of 19,796 United States robotics utility grants published between 2006 and 2025. The study uses Derwent World Patents Index (DWPI) abstract text, 768-dimensional SciBERT embeddings, k-means cluster assignments, exact nearest-neighbour search, and a patent-window panel of maintenance-fee decisions. The design distinguishes semantic mapping from behavioural validation: the text-derived indicators are not treated as direct measures of patent value, but are tested against a real legal-economic outcome — whether owners continue paying maintenance fees (Pakes, 1986; Lanjouw et al., 1998; Bessen, 2008; Serrano, 2010).

Guided by this design, the study addresses four research questions. First, what semantic domains structure United States robotics patenting between 2006 and 2025? Second, how do these domains differ in backward semantic novelty and forward semantic influence? Third, how do novelty, influence, and typicality combine into descriptive novelty regimes? Fourth, are these text-derived indicators associated with maintenance-fee lapse behaviour after accounting for fee window, grant year, cluster, and legal characteristics?

The results show that robotics patenting expanded strongly over the study period and became more broadly distributed across component-facing and system-facing domains. Sensing, perception, and autonomous control exhibit the highest backward novelty, while intelligent systems and field robotics exhibit the strongest forward semantic influence. In fee-window models, patents with stronger forward influence are significantly less likely to lapse. Novelty alone is not significant once fixed effects and legal controls are included. However, the significant novelty × typicality interaction indicates that novelty is associated with a lower likelihood of fee lapse when patents remain anchored in a recognisable technological domain. This pattern is consistent with the idea that valuable inventive positions are not simply different; they are distinctive while remaining technically legible.

The paper makes three contributions. First, it extends patent-based technology mapping by combining transformer-based clustering with embedding-native novelty, influence, and typicality indicators. Second, it provides a behavioural validation design by linking those indicators to maintenance-fee lapse events rather than relying solely on text-internal coherence. Third, it offers substantive evidence on robotics innovation, showing how growth across robotics domains differs from patent-level influence and persistence. The contribution is therefore both methodological and empirical: the paper provides a replicable framework for measuring semantic positions in patent text and demonstrates it in the robotics domain.

2. Literature review and conceptual framework

2.1 Patent analytics and technology mapping

Patent statistics have long been used as indicators of technological activity, innovation trajectories, and economic value (Griliches, 1990). Early patent-analytics work emphasised counts, classifications, citations, and maintenance behaviour as observable traces of inventive output and owner valuation (Hall et al., 2005; Pakes, 1986; Trajtenberg, 1990). Classification-based patent maps remain useful because they provide stable formal categories, but they can be too coarse for technologies that cut across multiple engineering subsystems (Leydesdorff et al., 2014). Robotics inventions often integrate hardware, sensing, control software, navigation, and task-specific applications, making the field difficult to capture through formal classifications alone.

Technology mapping in robotics has therefore moved toward text-based, semantic, and network-based approaches. Patent abstracts contain concise descriptions of technical contributions, while Derwent World Patents Index (DWPI) abstracts add standardised and technically enriched summaries. Semantic methods can recover technological neighbourhoods that are not fully visible through classification codes alone (Arts et al., 2018; Qiu & Wang, 2022). Patent-based robotics studies have also used co-patenting networks to map the geographical organisation of R&D collaboration in the field, reinforcing the value of patent records for studying robotics innovation structure (Scherngell et al., 2023). In this study, the mapping layer uses transformer-based patent-text representations to identify the broad semantic architecture of robotics patenting. The paper then moves beyond mapping by asking whether patents' semantic positions are associated with later semantic referencing and maintenance-fee decisions.

2.2 Transformer embeddings and embedding-native patent indicators

Transformer models such as BERT and SciBERT represent documents using contextual language patterns rather than isolated terms (Devlin et al., 2019; Beltagy et al., 2019). For technical corpora, SciBERT is suitable because it was pretrained on scientific text and uses a vocabulary adapted to scientific language. When patent abstracts are embedded into a common semantic space, similarities between patents can be used to construct indicators of technological relatedness, novelty, and influence. This builds on a growing literature that uses patent text itself to measure technological similarity and innovation rather than relying solely on classifications or citations (Arts et al., 2018; Kelly et al., 2021).

The present study defines backward semantic novelty as distance from earlier patents in a trailing priority-year window, and forward semantic influence as the frequency with which a patent appears among later patents' nearest prior semantic neighbours. These indicators are conceptually related to earlier text-based measures of technological novelty and influence, but are computed using transformer embeddings rather than bag-of-words representations (Kelly et al., 2021). They are also explicitly corpus-bounded: novelty is measured relative to observed robotics patents, not relative to all global prior art. This distinction is consistent with work showing that novelty can be measured in multiple ways depending on whether the reference set is technological recombination, knowledge origin, or textual proximity (Verhoeven et al., 2016). Because semantic novelty is defined relative to a comparison set, its interpretation depends on the size and density of the prior patent pool. The relevant construct in this study is therefore not absolute distance across all years, but a patent's semantic position relative to other patents in the same cohort.

2.3 Novelty, typicality, and technological legibility

Novelty is not necessarily valuable in isolation. Research on technological search, categories, and optimal distinctiveness suggests that actors and audiences often respond favourably to offerings that are sufficiently distinctive to stand out but sufficiently recognisable to be interpreted within an existing category (Deephouse, 1999; Zuckerman, 1999; Zhao et al., 2017). Related evidence shows that high-impact work often combines unusual combinations with conventional grounding (Uzzi et al., 2013), while technological recombination can increase both novelty and uncertainty (Fleming, 2001). In patent text, this implies that semantic novelty may matter differently depending on whether the patent remains legible within a recognised technological domain.

The study captures this legibility through cluster typicality. Conceptually, typicality reflects how centrally a patent is positioned within its assigned robotics domain. A highly typical patent is recognisably located within an established semantic area, whereas a peripheral patent is more distant from the core of that domain. The central expectation is therefore not simply that more novel patents are more valuable or persistent. Rather, novelty may be more likely to be associated with retention when it remains technically legible. In other words, novel-but-legible patents may be more likely to be maintained than patents that are novel but too peripheral to an established robotics domain.

2.4 Maintenance-fee lapse as behavioural validation

Maintenance-fee behaviour provides an external behavioural criterion for patent-text indicators. Patent owners face renewal costs at specified points in the patent life cycle. When owners fail to pay maintenance fees, they indicate that the expected private value of keeping the patent in force may not justify the cost, subject to organisational, legal, and portfolio considerations. This logic underpins a long literature using patent-renewal data to infer private patent value (Pakes, 1986; Schankerman & Pakes, 1986; Lanjouw et al., 1998; Bessen, 2008). Maintenance-fee lapse is therefore not a perfect measure of patent value, but it is a behavioural outcome linked to owner decisions.

The present study uses dated failure-to-pay maintenance-fee events rather than raw legal status. This distinction is essential because a raw dead/alive field conflates fee-related abandonment with natural term expiry. Natural expiry does not imply low value; it can simply reflect the end of the patent term. The

analysis therefore treats fee-lapse events as the outcome, while natural expiry and administrative discontinuation are censored. This allows semantic indicators to be tested against an outcome that is closer to owner maintenance decisions, while recognising that renewal behaviour also varies by owner type, technology field, and portfolio strategy (Bessen, 2008; Serrano, 2010).

Taken together, the literature suggests that patent-text indicators become more informative when semantic mapping, novelty measurement, and behavioural validation are considered jointly. Classification and citation approaches provide important signals, but they do not fully capture the semantic position of robotics inventions within an evolving technological landscape. Transformer embeddings make it possible to locate patents within that landscape, while maintenance-fee lapse behaviour provides an external behavioural criterion for assessing whether those semantic positions correspond to owner decisions. The next section operationalises this framework by constructing a SciBERT-based robotics patent map, deriving embedding-native indicators of novelty, influence, and typicality, and linking those indicators to maintenance-fee lapse behaviour in a discrete-time fee-window panel.

3. Methodology

3.1 Research design

This study develops a semantic patent-analytics design for examining the structure, novelty, influence, and maintenance behaviour of robotics patents. The analysis proceeds in four phases. Phase 1 constructs the semantic map of robotics innovation by representing each patent abstract using SciBERT embeddings and identifying major technological domains through k-means clustering. Phase 2 derives patent-level indicators of backward semantic novelty, forward semantic influence, and cluster typicality from the same SciBERT embedding space. Phase 3 classifies patents into descriptive novelty regimes that combine novelty, influence, and typicality. Phase 4 tests whether the text-derived indicators are associated with maintenance-fee lapse behaviour using discrete-time fee-window models.

The design separates three types of evidence. The first layer maps the technological structure of robotics patenting. The second layer measures how individual patents are positioned within that structure. The third layer evaluates whether those semantic positions are associated with observable maintenance-fee decisions. This design isolates the informational value of patent text by testing whether text-derived indicators are associated with maintenance-fee behaviour, even without citation data, assignee information, or manually coded patent-value labels. Instead, it tests whether embedding-native indicators derived from patent text are associated with a legally observable economic behaviour: whether owners continue paying maintenance fees to keep patents in force.

3.2 Patent corpus and record filtering

Robotics-related patents were retrieved using robot-related terminology and DWPI technology class filters to improve recall while limiting the corpus to robotics-relevant records. The initial corpus contained 19,823 patent records published between 2006 and 2025.

Because the maintenance-fee analysis requires standard utility-grant status, 27 non-standard records with C1, C2, or E kind codes were excluded. The main analytical sample therefore contains 19,796 B1/B2 utility grants. These records form the basis for the semantic-indicator and fee-window analyses. The excluded records are not used in the main models because reissue, reexamination, and certificate records do not follow the same maintenance-fee interpretation as ordinary utility grants. Table 1 summarises the corpus construction and event audit.

The principal textual input is the DWPI abstract. DWPI abstracts were used because they provide standardised and technically enriched summaries of patent disclosures. The analysis uses titles and abstracts rather than full claims or specifications to maintain a consistent and scalable patent-level text representation across the full corpus. Claims and specifications may contain richer legal and technical detail, but they vary substantially in length and structure, making them less suitable for the present corpus-level embedding design.

Table 1
Corpus construction and event audit.

metric	value
Total records	19,823
B1/B2 utility grants retained	19,796
Non-B1/B2 excluded	27
Priority year parsed	19,823
Publication date parsed	19,823
Fee-lapse events (dated failure-to-pay)	2,731
Fee-lapse events inside windows 1–3	2,729
Admin/discontinuation dated exits	24
Dead total	3,264
Dead via fee lapse	2,730
Dead, natural/term expiry	533

Note: Event counts are reported before construction of the patent-window panel. Fee-lapse events outside expected maintenance-fee windows are excluded from the panel; natural expiry and administrative discontinuation are treated as censoring events.

3.3 Date reconstruction and temporal variables

Several date fields required harmonisation before temporal indicators could be computed. For B1/B2 records, publication date is treated as the grant-publication date and is used to construct maintenance-fee windows. Priority-date fields were parsed explicitly because records used mixed date formats. When multiple priority dates were present, the earliest valid date was retained. To avoid residual day-month ambiguity, the analysis uses priority year rather than exact priority date.

The analysis uses different temporal variables for different purposes. Publication year is used to describe patenting growth because recent priority-year counts are affected by grant-lag truncation. Priority year is used for novelty and influence calculations because it is closer to the timing of inventive activity. Grant-publication date is used for maintenance-fee windows because United States maintenance-fee deadlines are defined relative to grant date.

The Derwent export snapshot is treated as current to 8 March 2026, which provides the censoring cutoff for fee-window construction. A 30-day buffer before the cutoff is applied to avoid treating windows opening immediately before the data boundary as observed non-events.

3.4 SciBERT embedding and semantic clustering

Patent abstracts were represented using SciBERT, specifically the allenai/scibert_scivocab_uncased checkpoint, a BERT-base transformer model pretrained on scientific text (Beltagy et al., 2019). Each patent abstract was tokenised with the SciBERT vocabulary and encoded using pretrained weights in inference mode. Document-level vectors were constructed through attention-masked mean pooling over the final hidden states, consistent with sentence- and document-embedding practice in transformer-based text representation (Reimers & Gurevych, 2019). This produced a 768-dimensional embedding for each patent. The embeddings were then L2-normalised so that cosine similarity could be computed through the inner product of unit vectors, following standard information-retrieval practice for vector-space similarity measurement (Manning et al., 2008).

K-means clustering was applied to the full set of L2-normalised SciBERT embeddings using k-means + + initialisation (Arthur & Vassilvitskii, 2007). The number of clusters was fixed at k = 5 following interpretive comparison of alternative solutions. Because high-dimensional nearest-neighbour and distance-based structure can be affected by distance concentration (Beyer et al., 1999), the cluster solution was assessed primarily through semantic coherence, representative keywords, and centroid-proximate patents rather than through geometric compactness alone.

After clustering, each cluster was interpreted by examining representative keywords and patents closest to the cluster centroid. These labels are reported in the Results section. The resulting cluster assignments provide the semantic architecture for the subsequent novelty, influence, typicality, and fee-lapse analyses. No keyword-based reconstruction or quota-based reassignment was performed. All downstream semantic indicators are computed directly from the same SciBERT embedding space used for clustering, ensuring consistency between the technology-map layer and the patent-level indicator layer.

3.5 Backward semantic novelty

Backward semantic novelty measures the extent to which a focal patent differs from earlier robotics patents in the observed corpus. For each patent i, we identify patents with earlier priority years within a five-year trailing window. We then compute cosine similarities between patent i's SciBERT vector and all prior patents in that window. The main novelty measure is defined as:

$$Novelty_i = 1 - \frac{1}{k} \sum_{j \in NN_k(i)} cos\left(e_i, e_j\right), \ k = 10 \qquad (1)$$

where e_i is the L2-normalised SciBERT embedding of patent i, $NN_k(i)$ is the set of patent i's k nearest prior semantic neighbours in the five-year trailing priority window, and $cos(e_i, e_j)$ is the cosine similarity between patent i and prior patent j.

Higher values indicate that the patent is less similar to its nearest prior semantic neighbours and therefore more novel within the observed robotics patent corpus. The novelty indicator is explicitly corpus-bounded: it measures novelty relative to earlier patents in the robotics corpus, not absolute novelty relative to all global prior art. A patent that imports a mature technique from outside robotics may therefore appear novel within this corpus if that technique is not already represented in earlier robotics patents.

To reduce boundary effects from the early years of the corpus and ensure that each focal patent is evaluated against a sufficiently populated five-year prior robotics window, the main novelty-valid analysis begins with priority year 2010. Because recent priority years are affected by grant-lag truncation, cross-year novelty levels near the right edge of the series are interpreted cautiously.

3.6 Forward semantic influence

Forward semantic influence measures whether later patents appear near a focal patent in the semantic space. For each later patent l, we identify its k nearest prior semantic neighbours within the five-year trailing priority window. A focal patent i receives one count each time it appears among those nearest prior neighbours. Forward semantic influence is defined as:

$$Influence_i = \sum_l 1(i \in NN_k(l)), \quad k = 10 \qquad (2)$$

where l indexes later patents, $NN_k(l)$ is the set of patent l's k nearest prior semantic neighbours, and ($1(i \in NN_k(l))$ is an indicator equal to one when patent i appears among later patent l's nearest prior neighbours.

The resulting count captures how often later robotics patents are located near the focal patent in the SciBERT space. For this reason, the term semantic influence is used throughout.

Because forward influence requires sufficient future observation time, influence statistics are restricted to patents with priority years up to 2018. This conservative cutoff allows focal patents to receive later semantic-neighbour counts within the five-year forward-observation structure. Since $Influence_i$ is skewed, the validation models use $\log(1 + Influence_i)$, standardised to mean zero and unit variance.

3.7 Cluster typicality

Cluster typicality measures how centrally a patent is located within its assigned robotics domain. For each of the five SciBERT k-means clusters, the L2-normalised centroid of all patents assigned to that cluster is computed. A patent's typicality is then calculated as its cosine similarity to the centroid of its own cluster:

$$\text{Typicality}_i = \cos\left(e_i, c_{g(i)}\right) \qquad (3)$$ where $e_i$is the L2-normalised SciBERT embedding of patent i, $g(i)$denotes the cluster assigned to patent i, and $c_{g(i)}$is the L2-normalised centroid of that cluster.

Higher typicality indicates that the patent is more recognisably located within its semantic domain, while lower typicality indicates a more peripheral position. The preferred specification focuses on novelty, typicality, their interaction, and forward influence.

3.8 Novelty-regime classification

To provide descriptive insight into how novelty and influence combine, patents were assigned to novelty regimes using within-year percentile ranks. Backward novelty, forward influence, and typicality were each converted to priority-year percentile ranks. Patents were then classified into descriptive regimes: incremental core, influential core, peripheral novel, dormant novel, breakthrough-like, and routine/peripheral other.

Because the regime taxonomy uses forward semantic influence, the main regime analysis is restricted to priority years 2010–2018. The lower bound is set by the requirement that each focal patent have a sufficiently populated prior window for novelty measurement, while the upper bound is set by the forward-observability cutoff used for the influence measure. Regime shares after 2018 are not interpreted because later patents have incomplete forward-observation windows.

These regimes are not latent classes and are not treated as theoretically fixed invention types. They are descriptive categories that summarise how novelty, influence, and domain typicality combine within priority-year cohorts. The regime classification is used to support descriptive interpretation and to compare maintenance-fee lapse patterns across the first two fee windows. Regime-level results for the third fee window are not interpreted because several cells are small.

The regime taxonomy also clarifies the theoretical relevance of the novelty × typicality interaction. Backward novelty is based on distance from prior neighbours, while typicality is based on proximity to a cluster centroid. Highly novel and highly typical patents therefore occupy a relatively narrow part of the semantic space. The interaction term tests whether novelty is associated with retention primarily when it remains legible within a recognisable technological domain.

3.9 Maintenance-fee lapse outcome

The behavioural validation layer uses maintenance-fee lapse events. The raw Dead/Alive field is not used as the dependent variable because it conflates fee-related abandonment with natural term expiry. Instead, failure-to-pay maintenance-fee events are extracted from the Estimated Expiration Date Lapse Factor field. Records containing dated failure-to-pay maintenance-fee language are coded as fee-lapse events, and the event date is parsed from the text string.

This procedure identifies 2,731 dated failure-to-pay events in the full corpus. Of these, 2,729 fall within the expected three maintenance-fee windows; two dated failure-to-pay events fall outside those windows and are excluded from the fee-window panel. After applying the observability rules and data-cutoff

buffer, 2,727 fee-lapse events enter the final patent-window panel. Natural term expiry and administrative discontinuation are treated as censoring events: patents leave the risk set at those dates without being coded as fee-lapse outcomes.

The fee-window panel is constructed at the patent-window level. Each patent can contribute up to three observations: window 1, window 2, and window 3. A patent enters a window only if that window is observable before the data cutoff. Patents exit the risk set after fee lapse, natural term expiry, administrative discontinuation, or right-censoring at the data cutoff. This construction avoids misclassifying natural expiry as owner abandonment and avoids treating windows opening at the data boundary as observed non-events.

3.10 Model specification

The main validation models are discrete-time logit models estimated on the patent-window panel. The dependent variable equals one if patent i experiences a maintenance-fee lapse in fee window w, and zero if the patent remains at risk without lapsing in that window. The preferred model includes backward novelty, squared novelty, cluster typicality, the novelty × typicality interaction, forward influence, IPC breadth, terminal-disclaimer status, patent-term adjustment, fee-window fixed effects, grant-year fixed effects, and cluster fixed effects.

$$\text{logit}\left\{P\left(Y_{iw}=1\right)\right\}=\alpha\ +\beta_1\text{Novelty}_i+\beta_2\text{Novelty}_i^2+\beta_3\text{Typicality}_i+\beta_4\left(\text{Novelty}_i\times\ \text{Typicality}_i\right)+\beta_5\log\left(1+\text{Influence}_i\right)+X_i\gamma\ +$$

4

where $X_i$contains IPC breadth, terminal-disclaimer status, and patent-term adjustment; $\delta_w$are fee-window fixed effects; $\lambda_{g(i)}$are cluster fixed effects; and $\tau_t$are grant-year fixed effects.

Forward influence enters the models as $log(1\ +\ forward\ influence)$. IPC breadth is measured as the number of distinct IPC subclasses. Terminal disclaimer is included as a binary indicator. Patent-term adjustment is measured in days. Standard errors are clustered at the patent level. All continuous predictors are standardised, so reported odds ratios (OR) are interpreted per standard-deviation change.

The model sequence proceeds as follows. M1 includes only fee-window, grant-year, and cluster fixed effects. M2 adds novelty and typicality. M3 adds forward influence and is estimated only on patents with sufficient forward-observation time. M4 adds the squared novelty term. M5 adds legal and classification controls: IPC subclass breadth, terminal-disclaimer status, and patent-term-adjustment days. M6, the preferred theoretical specification, adds the novelty × typicality interaction. Because M2 and M3 are estimated on different samples, likelihood-ratio tests are not used across that boundary.

The models are interpreted as validation-grade associations between semantic indicators and maintenance-fee decisions. They are not causal estimates of patent value. Assignee and owner-type controls are unavailable in the current export, and ownership characteristics may influence maintenance-fee behaviour.

3.11 Robustness and reproducibility

Novelty robustness was assessed using eighteen operational specifications. The main specification uses a five-year trailing window, 10 nearest neighbours, and mean top-k similarity. The robustness grid combines three trailing-window lengths of three, five, and ten years; three nearest-neighbour counts of 5, 10, and 20; and two similarity summaries: mean top-k similarity and maximum similarity. These eighteen specifications correspond to twelve distinct novelty rankings because the maximum-similarity variant is invariant to neighbour count. Each alternative novelty measure was compared with the main specification using Spearman rank correlation.

The analysis was executed through a single scripted pipeline that constructs the cleaned patent corpus, merges the SciBERT cluster assignments, aligns the 768-dimensional embedding matrix, computes exact nearest-neighbour indicators, constructs the fee-window panel, estimates the validation models, and exports the tables and figures. Input files were recorded with SHA-256 hashes, and the discrete-time logit models were estimated in statsmodels using patent-clustered covariance estimators.

The robustness checks reported in this paper are internal to the SciBERT embedding space used in the semantic mapping stage. Cross-model embedding robustness is therefore not claimed; this limitation is addressed explicitly in the limitations section.

4. Results

4.1 Growth and structural composition of robotics patenting

The final corpus contains 19,796 B1/B2 utility grants after excluding 27 non-standard records from the original 19,823-record corpus. Annual patenting increased substantially over the publication-year window. As shown in Fig. 1 and Table 3, robotics utility-grant publications rose from 241 in 2006 to 2,756 in 2025, indicating more than a tenfold increase over the study period.

Table 2
Selected publication-year growth anchors.

Publication year	Patents
2006	241
2010	284
2015	736
2020	1,560
2024	2,270
2025	2,756

The SciBERT k-means solution identifies five robotics domains: robotic mechanisms and actuation; articulated motion and manipulation; sensing, perception, and autonomous control; intelligent systems and field robotics; and navigation and environment interaction. These clusters capture distinct but interdependent layers of robotics systems, spanning mechanical embodiment, manipulation, perception, control, navigation, and deployment. Table 3 reports the size and share of each cluster in the retained utility-grant corpus.

Table 3
SciBERT k-means clusters in the retained B1/B2 utility-grant corpus.

Cluster	Cluster label	Patents	Share (%)
0	Robotic mechanisms and actuation	3,429	17.3%
1	Articulated motion and manipulation	3,050	15.4%
2	Sensing, perception and autonomous control	3,158	16%
3	Intelligent systems and field robotics	5,234	26.4%
4	Navigation and environment interaction	4,925	24.9%

The largest cluster is intelligent systems and field robotics, with 5,234 patents, or 26.4% of the retained corpus. Navigation and environment interaction follows closely with 4,925 patents, or 24.9%. Robotic mechanisms and actuation accounts for 17.3%, sensing, perception, and autonomous control for 16.0%, and articulated motion and manipulation for 15.4%. The corpus is therefore not concentrated in one robotics subsystem; rather, it spans component-facing, perception-oriented, and system-facing technological domains.

The temporal composition of these domains changes substantially over the study period. As shown in Fig. 2 and Table 4, the 2006 composition was dominated by navigation and environment interaction, which accounted for 47.7% of patents, while sensing, perception, and autonomous control accounted for only 1.7%. By 2024, the corpus was more evenly distributed across the five domains: intelligent systems and field robotics accounted for 29.2%, sensing, perception, and autonomous control for 20.9%, navigation and environment interaction for 20.8%, robotic mechanisms and actuation for 19.2%, and articulated motion and manipulation for 9.9%.

This contrast shows that the major compositional change is not simply the growth of system-facing clusters, but especially the rise of sensing, perception, and autonomous-control patents from a very small early share to a substantial share of later robotics patenting. This compositional evolution motivates the patent-level indicator analysis. Cluster shares describe which parts of the robotics landscape expanded, while novelty, influence, and typicality measure how individual patents are positioned within the semantic space.

Table 4
Cluster composition in 2006 and 2024.

Year	Cluster label	Patents	Share (%)
2006	Robotic mechanisms and actuation	17	7.1%
2006	Articulated motion and manipulation	67	27.8%
2006	Sensing, perception and autonomous control	4	1.7%
2006	Intelligent systems and field robotics	38	15.8%
2006	Navigation and environment interaction	115	47.7%
2024	Robotic mechanisms and actuation	436	19.2%
2024	Articulated motion and manipulation	224	9.9%
2024	Sensing, perception and autonomous control	475	20.9%
2024	Intelligent systems and field robotics	662	29.2%
2024	Navigation and environment interaction	473	20.8%

4.2 Backward semantic novelty

Backward novelty was computed as one minus the mean cosine similarity to the ten nearest prior patents in a five-year trailing priority window. The main novelty series begins in priority year 2010 to reduce early-window boundary effects. The measure should therefore be interpreted as novelty within the observed robotics patent corpus.

Figure 3 and Table 5 report the priority-year novelty series. Mean novelty declines gradually over the novelty-valid series, from 0.0693 in 2010 to 0.0614 in 2019. This decline is largely inseparable from the growth of the prior comparison window itself. The prior window expands from approximately 2,000 patents for the 2010 cohort to more than 9,000 patents by the 2021 cohort; with a fixed top-10 nearest-neighbour rule, denser prior windows mechanically increase the likelihood of close semantic neighbours and therefore reduce measured novelty. We therefore do not interpret the level trend as evidence of declining inventiveness in robotics. Instead, the novelty measure is used primarily through within-year rankings and standardised predictors, limiting the influence of cross-cohort prior-pool growth on the validation models.

Novelty differs substantially across the five clusters, as shown in Fig. 4 and Table 6A. Sensing, perception, and autonomous control has the highest mean novelty at 0.0727, followed by robotic mechanisms and actuation at 0.0647. Articulated motion and manipulation has a mean novelty of 0.0613, navigation and environment interaction has 0.0608, and intelligent systems and field robotics has the lowest mean novelty at 0.0600. A Kruskal–Wallis test indicates that novelty differs systematically across clusters ($H = 2{,}248.8$, $p < .001$).

A further measurement issue is the anisotropy of mean-pooled SciBERT embeddings. Because cosine similarities are concentrated in a narrow range, the novelty measure also occupies a compressed interval: the 5th–95th percentile range is approximately 0.048–0.085, with a standard deviation of about 0.012. The analysis therefore does not rely on absolute novelty magnitudes. Instead, it uses within-year rankings and standardised predictors. Section 4.7 reports robustness checks showing high rank stability across alternative novelty specifications.

Table 5
Backward novelty by priority year.

Priority year	Patents	Mean novelty	Median novelty	Mean typicality
2010	551	0.0693	0.0673	0.9430
2011	610	0.0681	0.0659	0.9454
2012	678	0.0665	0.0647	0.9452
2013	757	0.0655	0.0632	0.9462
2014	1,066	0.0654	0.0638	0.9462
2015	1,300	0.0634	0.0622	0.9470
2016	1,556	0.0642	0.0620	0.9460
2017	1,977	0.0631	0.0615	0.9465
2018	2,074	0.0623	0.0603	0.9460
2019	2,188	0.0614	0.0593	0.9458
2020	1,613	0.0615	0.0595	0.9453
2021	1,247	0.0619	0.0600	0.9447
2022	693	0.0629	0.0614	0.9432
2023	264	0.0645	0.0622	0.9413
2024	94	0.0651	0.0618	0.9419
2025	10	0.0678	0.0639	0.9334

Table 6
A. Novelty and typicality by cluster, novelty-valid sample, priority year ≥ 2010.

Cluster	Cluster label	Patents	Mean novelty	Median novelty	Mean typicality
0	Robotic mechanisms and actuation	3,078	0.0647	0.0628	0.9445
1	Articulated motion and manipulation	2,230	0.0613	0.0604	0.9484
2	Sensing, perception and autonomous control	2,980	0.0727	0.0706	0.9368
3	Intelligent systems and field robotics	4,541	0.0600	0.0587	0.9486
4	Navigation and environment interaction	3,849	0.0608	0.0595	0.9482

Table 6
B. Forward semantic influence by cluster, influence-valid cohort, priority years 2010–2018.

Cluster	Cluster label	Patents	Mean influence	Median influence
0	Robotic mechanisms and actuation	1,855	11.34	5
1	Articulated motion and manipulation	1,703	11.55	6
2	Sensing, perception and autonomous control	1,566	9.43	5
3	Intelligent systems and field robotics	2,879	15.06	6
4	Navigation and environment interaction	2,566	13.10	6

4.3 Forward semantic influence

Forward semantic influence captures how often a patent appears among later patents' nearest prior semantic neighbours. The measure is restricted to priority years up to 2018 to ensure sufficient forward-observation time; the cluster statistics in Table 6B are reported on the common 2010–2018 analytical cohort used for influence-valid patent-level analyses. Unlike backward novelty, which measures distance from prior patents, forward influence measures whether later patents are semantically located near the focal patent.

Forward influence is skewed. Cluster-level means range from 9.43 to 15.06, while medians range from 5 to 6. This indicates that most patents appear only infrequently among later patents' nearest prior neighbours, while a smaller number become more central semantic reference points for later robotics patents. For this reason, the validation models use the log-transformed measure, log(1 + forward influence).

The cluster pattern for forward influence differs from the cluster pattern for novelty. Intelligent systems and field robotics has the highest mean forward influence at 15.06, followed by navigation and environment interaction at 13.10 and articulated motion and manipulation at 11.55. Robotic mechanisms and actuation has a mean influence of 11.34. Sensing, perception, and autonomous control has the lowest mean influence at 9.43, despite having the highest backward novelty. Figure 5 brings these indicators together by showing cluster-level profiles of backward novelty, forward influence, and typicality. Thus, the most novel semantic domain is not necessarily the domain around which later patents cluster semantically. This distinction between novelty and influence is central to the subsequent validation models.

4.4 Novelty regimes

The novelty-regime classification combines within-year rankings of novelty, influence, and typicality. Because influence requires forward observability, regime-share results are reported for priority years 2010–2018 only. The categories are descriptive, not latent classes. They summarise whether patents are novel or routine, central or peripheral, and influential or not within their priority-year cohort.

Figure 6 shows regime shares across the 2010–2018 regime-analysis window, and Table 7 reports the 2010 composition as an anchor. The largest categories are incremental core, influential core, peripheral novel, and routine/peripheral other. Breakthrough-like and dormant-novel regimes are smaller because they require more selective combinations of high novelty and high influence or high novelty without influence. In 2010, incremental-core patents accounted for 27% of the cohort, influential-core patents for 22.7%, peripheral-novel patents for 24.7%, and routine/peripheral-other patents for 25.2%. Breakthrough-like and dormant-novel patents together accounted for less than 1% in that year.

The sparsity of the breakthrough-like and dormant-novel categories is expected rather than anomalous. Backward novelty is based on distance from prior neighbours, while typicality is based on proximity to a cluster centroid; high novelty and high typicality therefore occupy a relatively narrow part of the semantic space. This structural tension is also why the novelty × typicality interaction is theoretically informative in the fee-lapse models: it distinguishes novelty that remains legible within a recognised technological domain from novelty that is more peripheral.

The regime results support two descriptive points. First, novelty and influence are not interchangeable. Some patents are semantically novel but do not become reference points for later patents, while others are semantically central and influential without being highly novel. Second, the robotics landscape contains both core incremental development and more peripheral novelty. The fee-lapse models therefore examine not only whether novelty matters, but whether novelty matters conditionally on domain typicality.

Table 7
Novelty-regime composition in 2010, influence-valid regime-analysis window.

Regime	Patents	Share (%)
Breakthrough-like	1	0.2%
Dormant novel	1	0.2%
Incremental core	149	27%
Influential core	125	22.7%
Peripheral novel	136	24.7%
Routine/peripheral other	139	25.2%

4.5 Maintenance-fee lapse structure

As shown in Fig. 7 and Table 8, the fee-window panel contains 17,717 patent-window observations from 11,046 patents and includes 2,727 fee-lapse events. The lapse rate rises across the three maintenance-fee windows. Window 1 contains 11,046 patents and 1,224 events, corresponding to a lapse rate of 11.1%. Window 2 contains 4,920 patents and 952 events, corresponding to a lapse rate of 19.3%. Window 3 contains 1,751 patents and 551 events, corresponding to a lapse rate of 31.5%.

This rising pattern is consistent with the economic logic of maintenance-fee decisions. Patents that survive to later windows face higher renewal costs and have aged further from their grant dates, so owners may be more selective about which patents they continue to maintain. Importantly, these lapse events are defined only from dated failure-to-pay maintenance-fee text. Natural expiry and administrative discontinuation are censored rather than treated as fee-lapse outcomes. Window 3 covers older grant cohorts only and is therefore interpreted more cautiously than windows 1 and 2.

Table 9 shows that lapse rates also vary descriptively by cluster. At window 1, sensing, perception, and autonomous control has the lowest lapse rate at 8.8%, while articulated motion and manipulation and robotic mechanisms and actuation have higher rates of 12.7% and 12.5%, respectively. At window 2, sensing, perception, and autonomous control again has a relatively low lapse rate of 16.8%, while navigation and environment interaction has 20.5%. At window 3, navigation and environment interaction has the highest lapse rate at 37.4%, while intelligent systems and field robotics has a lower rate of 23.7%. These cluster differences are descriptive and are later modelled with cluster fixed effects.

Table 10 reports regime-level lapse rates descriptively for windows 1 and 2 only because several window-3 regime cells are small. Breakthrough-like patents show the lowest descriptive lapse rates in windows 1 and 2, at 10.3% and 11.1%, respectively. Dormant-novel patents and peripheral-novel patents show higher lapse rates in the first two windows. These descriptive patterns are consistent with the interpretation that novelty without subsequent semantic influence or domain legibility is not necessarily associated with retention.

Table 8
Fee-window panel summary.

Fee window	Risk set	Events	Min grant year	Max grant year	Lapse rate (%)
1	11,046	1,224	2006	2022	11.1%
2	4,920	952	2006	2018	19.3%
3	1,751	551	2006	2014	31.5%

Table 9
Fee-lapse rates by cluster and window.

Cluster	Cluster label	Fee window	Risk set	Events	Lapse rate (%)
0	Robotic mechanisms and actuation	1	1,771	222	12.5%
0	Robotic mechanisms and actuation	2	674	136	20.2%
0	Robotic mechanisms and actuation	3	185	46	24.9%
1	Articulated motion and manipulation	1	2,170	275	12.7%
1	Articulated motion and manipulation	2	1,110	229	20.6%
1	Articulated motion and manipulation	3	468	142	30.3%
2	Sensing, perception and autonomous control	1	1,354	119	8.8%
2	Sensing, perception and autonomous control	2	482	81	16.8%
2	Sensing, perception and autonomous control	3	114	39	34.2%
3	Intelligent systems and field robotics	1	2,719	286	10.5%
3	Intelligent systems and field robotics	2	1,087	185	17%
3	Intelligent systems and field robotics	3	321	76	23.7%
4	Navigation and environment interaction	1	3,032	322	10.6%
4	Navigation and environment interaction	2	1,567	321	20.5%
4	Navigation and environment interaction	3	663	248	37.4%

Table 10
Fee-lapse rates by novelty regime for windows 1 and 2.

Regime	Fee window	Risk set	Events	Lapse rate (%)
Breakthrough-like	1	39	4	10.3%
Breakthrough-like	2	9	1	11.1%
Dormant novel	1	26	4	15.4%
Dormant novel	2	8	2	25%
Incremental core	1	2,199	256	11.6%
Incremental core	2	602	123	20.4%
Influential core	1	1,717	175	10.2%
Influential core	2	535	105	19.6%
Peripheral novel	1	2,012	239	11.9%
Peripheral novel	2	594	138	23.2%
Routine/peripheral other	1	2,021	224	11.1%
Routine/peripheral other	2	603	110	18.2%

4.6 Validation models: semantic indicators and fee-lapse behaviour

The validation models test whether semantic indicators are associated with maintenance-fee decisions after controlling for fee window, grant year, and cluster. M1 includes only fixed effects. M2 adds novelty and typicality. M3 adds forward influence and is restricted to the influence-valid cohort. M4 adds a quadratic novelty term. M5 adds legal and classification controls. M6 adds the novelty × typicality interaction and is the preferred specification.

Table 11 reports the estimation samples for the model sequence. M1 and M2 are estimated on 10,500 patent-window observations from 8,014 patents. M3 through M6 are estimated on 10,183 observations from 7,697 patents because forward influence requires sufficient future observation time. For this reason, fit statistics are interpreted within sample only, and likelihood-ratio tests are not used across the M2/M3 boundary. Although the full fee-window panel contains 17,717 observations, the validation-model sequence begins with the novelty-valid sample because the main theoretical predictors require a sufficiently populated prior window.

In M5, forward semantic influence is significantly associated with a lower likelihood of fee lapse. The OR is 0.876 (p = .0046), indicating that patents with higher forward semantic influence are less likely to lapse at observable maintenance-fee decisions. Backward novelty is not significant in M5 (OR = 1.064, p = .384), nor is typicality (OR = 1.076, p = .269). Thus, novelty alone is not associated with retention once fixed effects and controls are included.

The preferred M6 specification reveals a more informative pattern. Table 12 reports the preferred estimates, and Fig. 8 plots the corresponding ORs. Forward influence remains significant, with an OR of 0.853 (95% CI: 0.777–0.937, p = .0008). The novelty × typicality interaction is also significant, with an OR of 0.887 (95% CI: 0.824–0.956, p = .0017). The squared novelty term is significant and below one, with an OR of 0.900 (95% CI: 0.834–0.971, p = .0068). In contrast, the novelty main effect remains non-significant (OR = 1.074, p = .318), while typicality is marginal (OR = 1.128, p = .072).

These estimates suggest that novelty is not associated with maintenance-fee retention in the abstract. Rather, novelty is associated with retention primarily when it remains anchored in a recognisable technological domain. Patents that are both novel and legible within their assigned semantic domain are less likely to lapse, while novelty that is too peripheral appears less consistently retained. This interpretation should be treated as a validation-grade association, not a causal estimate of patent value.

The effect size is modest but meaningful for a text-only indicator. Against the window-1 baseline lapse rate of approximately 11.1%, a one-standard-deviation increase in log forward influence corresponds to roughly a 1.5-percentage-point lower lapse probability. This effect is estimated without citation data, assignee characteristics, claim counts, or market-value measures. The result therefore supports the usefulness of semantic indicators as behavioural validation signals, while also highlighting the need for future work using assignee and citation controls.

Table 11
Model estimation samples.

Sample	Observations	Patents	Events
fee_window_panel_all	17,717	11,046	2,727
novelty_valid_models_M1_M2	10,500	8,014	1,426
novelty_and_influence_valid_models_M3_M6	10,183	7,697	1,393

Table 12
Preferred M6 discrete-time logit estimates.

Term	Odds ratio (OR)	95% CI low	95% CI high	p-value
Novelty	1.0739	0.9336	1.2354	0.318
Typicality	1.1281	0.9893	1.2863	0.072
Novelty × typicality	0.8875	0.8237	0.9562	0.002
Novelty²	0.8999	0.8337	0.9714	0.007
Forward influence	0.8527	0.7767	0.9361	< .001
IPC breadth	0.9969	0.9414	1.0558	0.917
Terminal disclaimer	1.0224	0.8047	1.2992	0.856
Patent-term adjustment	1.0257	0.9678	1.0872	0.392

4.7 Robustness checks

The novelty indicator is robust to alternative operational choices. Eighteen operational specifications were computed, corresponding to twelve distinct novelty rankings because the maximum-similarity variant is invariant to neighbour count within a given prior-window length. These specifications combine three-, five-, and ten-year prior windows; 5, 10, and 20 nearest neighbours; and mean versus maximum similarity summaries. Spearman rank correlations with the main novelty measure range from 0.932 to 1.000. This indicates that the novelty ranking is not an artefact of a single temporal window, neighbour count, or similarity-summary rule. Figure 9 reports the rank stability of the alternative novelty specifications against the main specification, while Table 13 reports the corresponding summary statistics and Spearman correlations.

Table 13
Novelty robustness grid.

Window years	Top-k	Variant	N valid	Mean	SD	Spearman vs main
3	5	mean_topk	16,678	0.0622	0.0120	0.9883
3	5	max_sim	16,678	0.0577	0.0119	0.9316
3	10	mean_topk	16,678	0.0647	0.0123	0.9930
3	10	max_sim	16,678	0.0577	0.0119	0.9316
3	20	mean_topk	16,678	0.0676	0.0127	0.9893
3	20	max_sim	16,678	0.0577	0.0119	0.9316
5	5	mean_topk	16,678	0.0610	0.0118	0.9947
5	5	max_sim	16,678	0.0567	0.0117	0.9371
5	10	mean_topk	16,678	0.0635	0.0120	1.000
5	10	max_sim	16,678	0.0567	0.0117	0.9371
5	20	mean_topk	16,678	0.0662	0.0124	0.9963
5	20	max_sim	16,678	0.0567	0.0117	0.9371
10	5	mean_topk	16,678	0.0602	0.0116	0.9883
10	5	max_sim	16,678	0.0560	0.0115	0.9318
10	10	mean_topk	16,678	0.0625	0.0119	0.9948
10	10	max_sim	16,678	0.0560	0.0115	0.9318
10	20	mean_topk	16,678	0.0651	0.0122	0.9931
10	20	max_sim	16,678	0.0560	0.0115	0.9318

4.8 Summary of findings

Taken together, the results provide three layers of evidence. First, the corpus shows substantial growth in robotics patenting from 2006 to 2025 and a semantic structure organised around five robotics domains. Second, the SciBERT-based indicators distinguish backward novelty from forward semantic influence: sensing, perception, and autonomous control is the most novel domain within the observed corpus, while intelligent systems and field robotics is the most semantically influential. Third, the maintenance-fee models show that patents with higher semantic influence are less likely to lapse, and that novelty is associated with retention primarily when it remains technically legible within a recognised robotics domain.

5. Discussion

This study moves from semantic technology mapping to behavioural evidence in a large robotics patent corpus. The results show that transformer-based semantic positions are not merely descriptive: they are associated with subsequent maintenance-fee decisions. They also show that semantic position is multidimensional. Backward novelty, forward influence, and cluster typicality capture different properties of patents. Novelty identifies distance from earlier patents, influence identifies whether later patents are located near the focal patent in semantic space, and typicality identifies whether a patent remains legible within an established robotics domain.

5.1 From robotics growth to semantic structure

The first finding is the strong growth and compositional change of robotics patenting. Annual utility-grant publications rose from 241 in 2006 to 2,756 in 2025. The early corpus was dominated by navigation and environment interaction, while sensing, perception, and autonomous control accounted for only a very small share. By 2024, the corpus was more evenly distributed, and sensing, perception, and autonomous-control patents accounted for more than one-fifth of annual patents. This shift is consistent with the increasing role of perception, sensing, and software-enabled autonomy in robotics systems.

The five-cluster structure shows that robotics innovation is not a homogeneous field. Mechanical actuation, manipulation, perception, field robotics, and navigation form related but distinct technological regions. These regions correspond to the functional architecture of robotics systems: embodiment, manipulation, perception, control, mobility, and deployment. In this sense, the observed movement from component-facing domains toward sensing, autonomy, and system-facing domains resembles an architectural reconfiguration of how robotics inventions combine subsystems, rather than merely a change in component emphasis (Henderson & Clark, 1990). The cluster map therefore provides a meaningful foundation for the patent-level indicators that follow.

5.2 Novelty and influence are different constructs

The results show that backward novelty and forward semantic influence should not be conflated. Sensing, perception, and autonomous control has the highest backward novelty but the lowest mean forward semantic influence. In contrast, intelligent systems and field robotics has the lowest mean backward novelty but the highest forward semantic influence.

This inversion matters for patent analytics. A patent can be novel relative to the prior robotics corpus without becoming a semantic reference point for later patents. Conversely, a patent can be less novel in nearest-neighbour distance but still sit in a domain near which later inventions repeatedly cluster. Technology mapping based only on novelty would therefore miss an important dimension of how patent positions persist in subsequent inventive activity. The results support treating novelty and semantic influence as separate indicators rather than alternative measures of the same construct.

5.3 Novelty, legibility, and patent retention

The most theoretically informative result is the interaction between novelty and typicality. Novelty alone is not associated with retention once fixed effects and controls are included. However, the negative novelty × typicality interaction, together with the nonlinear novelty term, indicates that novelty is more strongly associated with retention when it remains anchored within a recognisable robotics domain. Novel-but-legible patents therefore appear to occupy a more persistent position: they are different enough to stand apart but recognisable enough to be interpreted within an established technological category.

This pattern is consistent with optimal-distinctiveness reasoning. Patents that are too routine may offer little differentiation, while patents that are too peripheral may lack recognisable technological anchoring (Zhao et al., 2017). It is also consistent with evidence that impactful knowledge production often combines novelty with conventional grounding (Uzzi et al., 2013). In the present context, the result suggests that semantic novelty becomes more behaviourally meaningful when it is interpreted together with domain typicality.

This interpretation should be handled carefully. The design does not prove that patent owners consciously value optimal distinctiveness, nor does it causally estimate patent value. Rather, the observed pattern indicates that text-derived novelty has behavioural relevance only when conditioned on domain typicality. This strengthens the measurement contribution by showing that semantic indicators become more informative when interpreted within a technological category structure.

5.4 Maintenance-fee behaviour as validation, not a complete value model

The fee-lapse layer is best interpreted as behavioural validation rather than a complete economic model of patent value. Maintenance-fee lapse is an owner decision that reflects expected value, cost, portfolio strategy, and organisational resources. The models show that semantic influence and novel-but-legible positioning are associated with maintenance behaviour, but they do not claim that these text-derived indicators are the sole or causal determinants of patent retention.

This distinction is especially important because assignee and owner-type variables are not available in the current export. Large firms, universities, start-ups, individuals, and government agencies may differ systematically in renewal behaviour. If assignee data were available, owner fixed effects or entity-type controls would strengthen the validation design. In the present study, the fee-window models are therefore deliberately framed as validation-grade associations.

5.5 Embedding-space interpretation

The results should be interpreted within the SciBERT semantic space used throughout the study. This is appropriate because the same embedding space supports clustering, novelty, influence, and typicality. The technology-map layer and patent-level indicators are therefore internally consistent. At the same time, mean-pooled SciBERT embeddings are anisotropic, and raw cosine distances are compressed. The analysis therefore relies on within-year rankings,

standardised predictors, exact nearest-neighbour search, and robustness checks across alternative operational specifications. These checks support the stability of the indicator rankings within the SciBERT space, but the findings should not be read as claims about absolute semantic distance.

6. Contributions and implications

6.1 Contributions to patent analytics

This paper makes three contributions to the measurement of technological change from patent text. First, it extends patent text analytics from descriptive mapping toward behaviourally anchored measurement. Where established approaches map technological structure through keyword co-occurrence or probabilistic topic models such as latent Dirichlet allocation (Wang et al., 2014), this study links transformer-based patent mapping to embedding-native, patent-level measures of novelty, influence, and typicality derived directly from a patent's position in semantic space rather than from citations (Kelly et al., 2021). Rather than stopping at clusters, the analysis then tests whether those measures are associated with maintenance-fee lapse behaviour. Because renewal and lapse decisions are well established as revealed indicators of patent value (Bessen, 2008), this grounds text-internal indicators in an external, economically consequential decision made by patent holders.

Second, it contributes conceptually by showing that semantic novelty is not self-evidently valuable or persistent. Novelty becomes more informative when interpreted relative to domain typicality. This connects patent-text indicators with category and optimal-distinctiveness reasoning: distinctive inventions may be more likely to persist when they remain recognisable within a technological domain (Deephouse, 1999; Uzzi et al., 2013; Zhao et al., 2017; Zuckerman, 1999).

Third, it provides substantive evidence on innovation in robotics, a domain in which patent-based topic analysis remains comparatively underdeveloped (Kumari et al., 2021). Applying the framework across robotics domains, the study shows that domain growth is distinct from the influence and persistence of the underlying patents: rapidly expanding areas are not necessarily those containing the most influential or most durable inventions. This distinction speaks to a broader literature on how the growth of a technological area is interpreted as a signal of its emergence and importance (Rotolo et al., 2015).

Underpinning all three contributions, the study improves measurement transparency by explicitly addressing several common sources of distortion in patent-text indicators: corpus-boundary effects, prior-pool growth, embedding anisotropy, forward-observation censoring, and contaminated legal-status outcomes. Together, the contributions offer a replicable framework for measuring semantic positions in patent text, a conceptual account of when novelty is likely to persist, and behaviourally anchored empirical evidence from the robotics domain.

6.2 Contributions to robotics technology mapping

For robotics research, the paper provides a large-scale semantic map of United States robotics utility grants from 2006 to 2025. The five-cluster structure captures both component-facing and system-facing domains, showing that robotics patenting expanded substantially and that sensing, perception, and autonomous control became much more prominent over time.

The findings also show that growth, novelty, and influence do not move identically across robotics domains. The sensing, perception, and autonomous-control cluster is the most novel within the observed corpus, whereas intelligent systems and field robotics is the most semantically influential. This distinction matters for technology intelligence: the domains generating the greatest semantic distance from prior robotics patents are not necessarily the domains around which later invention consolidates.

6.3 Practical implications

The framework can support technology intelligence, R&D portfolio analysis, and patent landscaping. Organisations can use semantic influence indicators to flag patents that later inventions build near in semantic space, especially when citation data are incomplete or delayed. They can also use novelty and typicality jointly to distinguish patents that are merely unusual from patents that are novel while remaining recognisable within a technological domain.

For R&D managers and patent analysts, the approach provides a screening tool for identifying patents that occupy strategically meaningful semantic positions. High forward influence can help identify patents that become reference points for later inventive activity, while the combination of novelty and typicality can help distinguish novel-but-legible patents from patents that are semantically peripheral. This is useful for portfolio review, competitor monitoring, and identifying areas where emerging inventions are consolidating around recognisable technological directions.

For policymakers and analysts, the approach provides a way to monitor emerging technological structure without relying solely on classification codes or manual taxonomies. In robotics, the rise of sensing, perception, and autonomous-control patents and the influence of intelligent systems and field robotics suggest that the field is increasingly organised around integrated system capability rather than only component design.

7. Limitations and future research

The study has several limitations. First, the semantic indicators are computed from DWPI titles and abstracts rather than full claims or complete patent specifications. Abstracts provide a consistent and scalable representation of inventions across a large corpus, but claims may contain additional information about scope, novelty, and legal boundaries. Future work could compare abstract-based, claim-based, and full-specification embeddings to assess how textual granularity affects semantic novelty, influence, and typicality measures.

Second, the analysis is internal to the SciBERT embedding space used for semantic mapping and downstream indicator construction. This consistency is appropriate for the present design, but it also means that cross-model embedding robustness is not claimed. Mean-pooled SciBERT embeddings are

anisotropic, and although the study uses within-year rankings, standardised predictors, exact nearest-neighbour search, and robustness checks across alternative window and neighbour specifications, it does not test whether the findings replicate under contrastively trained patent embeddings such as PaECTER (Ghosh et al., 2024). Future work should examine whether the forward-influence and novelty × typicality patterns persist across alternative patent-specific or contrastively trained embedding models.

Third, the study measures forward semantic influence through nearest-neighbour relations in embedding space rather than through formal patent citations. This is appropriate for developing a text-derived influence indicator, but it leaves open the question of convergent validity with citation-based measures. When cited-reference fields are available, future work should test whether forward semantic influence correlates with forward patent-citation influence and whether the two indicators capture complementary forms of technological impact.

Fourth, the analysis is limited to granted United States robotics utility patents. The United States patent system captures domestic and international actors seeking protection in a major technology market, but it does not represent the full global robotics patent universe. Future research could incorporate European, Chinese, Japanese, and Korean patent records to examine whether the same semantic domains, novelty patterns, and maintenance associations appear across patent systems.

Fifth, the novelty measure is corpus-bounded. It measures distance from earlier robotics patents in the observed corpus, not absolute novelty relative to all global prior art. A patent may appear novel within robotics because it imports concepts from another field. This is not a flaw for the present research question, which concerns within-domain robotics innovation, but it means that the findings should be interpreted as evidence of corpus-relative semantic novelty rather than universal prior-art novelty.

Sixth, assignee and owner-type variables are unavailable. This limits the interpretation of maintenance-fee models because renewal decisions may vary across large corporations, small firms, universities, government entities, and individual inventors (Bessen, 2008; Serrano, 2010). Future work should incorporate assignee fixed effects, owner-type controls, portfolio-size measures, and ownership-transfer information. Such extensions would allow stronger claims about patent value, portfolio strategy, and owner-level heterogeneity.

Finally, maintenance-fee lapse is an imperfect outcome. It captures a real behavioural decision, but it does not fully measure patent value. Maintenance decisions may reflect expected value, organisational routines, portfolio policies, legal strategy, or administrative error. The outcome is therefore best understood as an external validation signal rather than a complete valuation measure.

8. Conclusion

This study develops and applies a semantic patent-analytics framework for robotics innovation. Using SciBERT embeddings for 19,796 United States robotics utility grants, the analysis maps the technological structure of robotics into five domains and derives patent-level measures of backward novelty, forward semantic influence, and cluster typicality. These indicators are linked to dated maintenance-fee lapse events in a discrete-time fee-window panel.

The results show that robotics patenting grew strongly from 2006 to 2025 and became more broadly distributed across component-facing and system-facing domains. Sensing, perception, and autonomous control exhibits the highest backward novelty, while intelligent systems and field robotics exhibits the highest forward semantic influence. In the validation models, forward influence is associated with a lower likelihood of fee lapse. Novelty alone is not associated with retention, but novelty combined with domain typicality is. This indicates that patent persistence is linked not simply to being different, but to being novel in a technically legible way.

The paper advances patent analytics by linking transformer-based semantic indicators to observable maintenance behaviour. It shows that semantic maps can be used not only to describe technological domains, but also to construct indicators that correspond to owner maintenance decisions. For robotics, the findings highlight the growing importance of sensing, autonomy, and system-facing domains, while also showing that semantic influence and novel-but-legible positioning are associated with patent retention.

Declarations

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Funding

This research received no external funding.

Author Contribution

G.C. and C.Y. jointly conceived and designed the study, performed the analysis, and wrote the manuscript. Both authors reviewed and approved the final manuscript.

Data Availability

The Derwent World Patents Index (DWPI) data used to construct the text embeddings are proprietary and were accessed under licence from Clarivate; these data cannot be redistributed by the authors and are available from Clarivate subject to a valid subscription. The derived, de-identified datasets and the analysis code that support the findings of this study are available from the corresponding author (C.Y.) upon reasonable request.

References

1. Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, 1027–1035.
2. Arts, S., Cassiman, B., & Gomez, J. C. (2018). Text matching to measure patent similarity. Strategic Management Journal, 39(1), 62–84.
3. Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 3615–3620.
4. Bessen, J. (2008). The value of U.S. patents by owner and patent characteristics. Research Policy, 37(5), 932–945.
5. Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When is nearest neighbor meaningful? Proceedings of the International Conference on Database Theory, 217–235.
6. Deephouse, D. L. (1999). To be different, or to be the same? It is a question (and theory) of strategic balance. Strategic Management Journal, 20(2), 147–166.
7. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT, 4171–4186.
8. Fleming, L. (2001). Recombinant uncertainty in technological search. Management Science, 47(1), 117–132.
9. Ghosh, M., Erhardt, S., Rose, M. E., Buunk, E., & Harhoff, D. (2024). PaECTER: Patent-level representation learning using citation-informed transformers. arXiv:2402.19411.
10. Griliches, Z. (1990). Patent statistics as economic indicators: A survey. Journal of Economic Literature, 28(4), 1661–1707.
11. Hall, B. H., Jaffe, A., & Trajtenberg, M. (2005). Market value and patent citations. RAND Journal of Economics, 36(1), 16–38.
12. Henderson, R. M., & Clark, K. B. (1990). Architectural innovation: The reconfiguration of existing product technologies and the failure of established firms. Administrative Science Quarterly, 35(1), 9–30.
13. International Federation of Robotics. (2025). World Robotics 2025: Industrial robots. IFR Statistical Department.
14. Kelly, B., Papanikolaou, D., Seru, A., & Taddy, M. (2021). Measuring technological innovation over the long run. American Economic Review: Insights, 3(3), 303–320.
15. Kumari, R., Jeong, J. Y., Lee, B. H., Choi, K. N., & Choi, K. (2021). Topic modelling and social network analysis of publications and patents in humanoid robot technology. *Journal of Information Science*, *47*(5), 658–676.
16. Lanjouw, J. O., Pakes, A., & Putnam, J. (1998). How to count patents and value intellectual property: The uses of patent renewal and application data. Journal of Industrial Economics, 46(4), 405–432.
17. Leydesdorff, L., Kushnir, D., & Rafols, I. (2014). Interactive overlay maps for US patent (USPTO) data based on International Patent Classification. Scientometrics, 98(3), 1583–1599.
18. Manning, C. D., Raghavan, P., & Schutze, H. (2008). Introduction to information retrieval. Cambridge University Press.
19. Momeni, A., & Rost, K. (2016). Identification and monitoring of possible disruptive technologies by patent-development paths and topic modeling. Technological Forecasting and Social Change, 104, 16–29.
20. Pakes, A. (1986). Patents as options: Some estimates of the value of holding European patent stocks. Econometrica, 54(4), 755–784.
21. Qiu, Z., & Wang, Z. (2022). Technology forecasting based on semantic and citation analysis of patents: A case of robotics domain. IEEE Transactions on Engineering Management, 69(4), 1216–1236.
22. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 3982–3992.
23. Rotolo, D., Hicks, D., & Martin, B. R. (2015). What is an emerging technology?. *Research policy*, *44*(10), 1827–1843.
24. Schankerman, M., & Pakes, A. (1986). Estimates of the value of patent rights in European countries during the post-1950 period. Economic Journal, 96(384), 1052–1076.
25. Scherngell, T., Schwegmann, K., & Zahradnik, G. (2023). The geographical dynamics of global R&D collaboration networks in robotics: Evidence from co-patenting activities across urban areas worldwide. PLOS ONE, 18(4), e0281353.
26. Serrano, C. J. (2010). The dynamics of the transfer and renewal of patents. RAND Journal of Economics, 41(4), 686–708.
27. Trajtenberg, M. (1990). A penny for your quotes: Patent citations and the value of innovations. RAND Journal of Economics, 21(1), 172–187.
28. Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. Science, 342(6157), 468–472.
29. Verhoeven, D., Bakker, J., & Veugelers, R. (2016). Measuring technological novelty with patent-based indicators. Research Policy, 45(3), 707–723.
30. Wang, B., Liu, S., Ding, K., Liu, Z., & Xu, J. (2014). Identifying technological topics and institution-topic distribution probability for patent competitive intelligence analysis: a case study in LTE technology. *Scientometrics*, *101*(1), 685–704.

31. Zhao, E. Y., Fisher, G., Lounsbury, M., & Miller, D. (2017). Optimal distinctiveness: Broadening the interface between institutional theory and strategic management. Strategic Management Journal, 38(1), 93–113.
32. Zuckerman, E. W. (1999). The categorical imperative: Securities analysts and the illegitimacy discount. American Journal of Sociology, 104(5), 1398–1438.

Figures

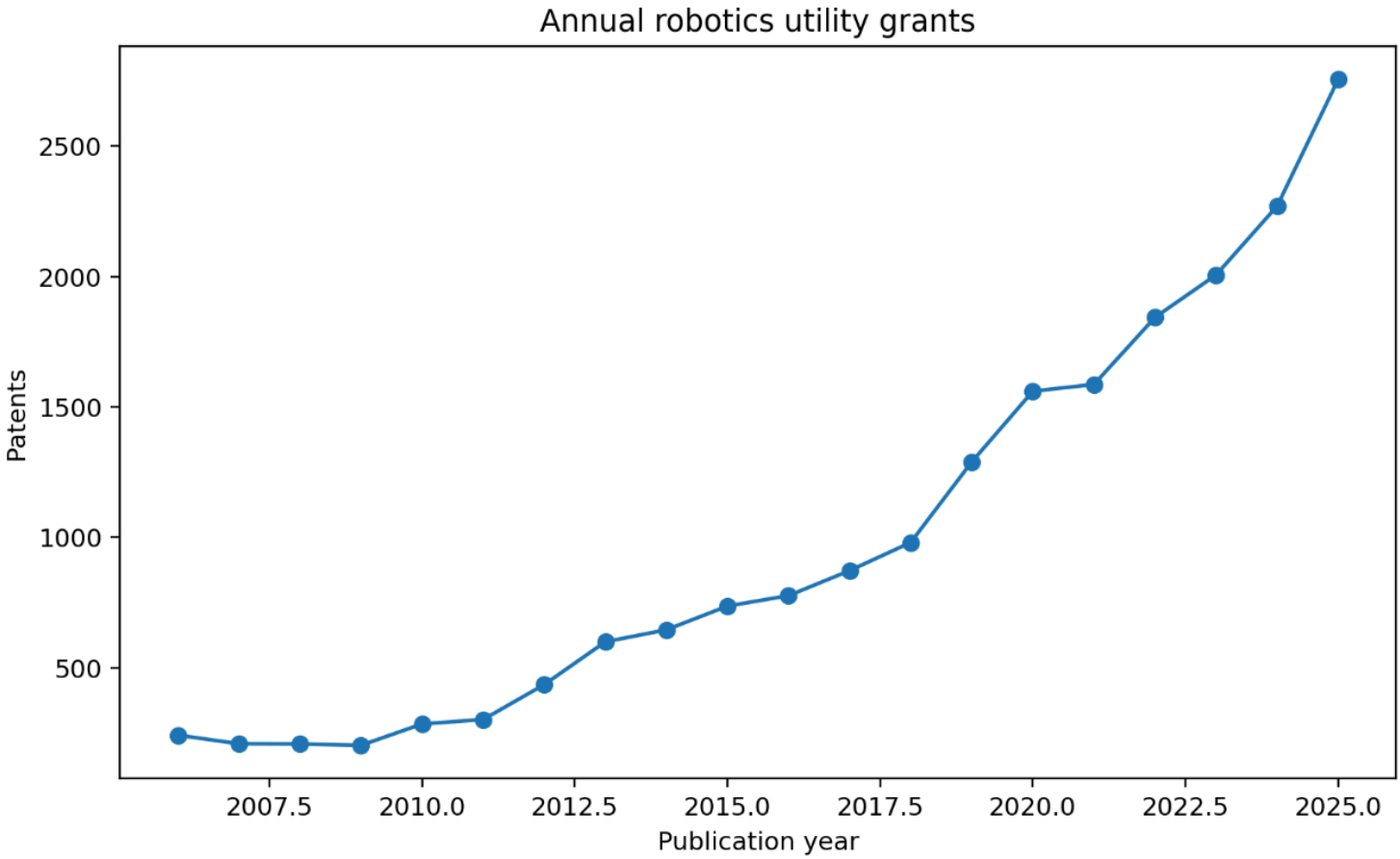


Figure 1

Annual robotics utility grants by publication year

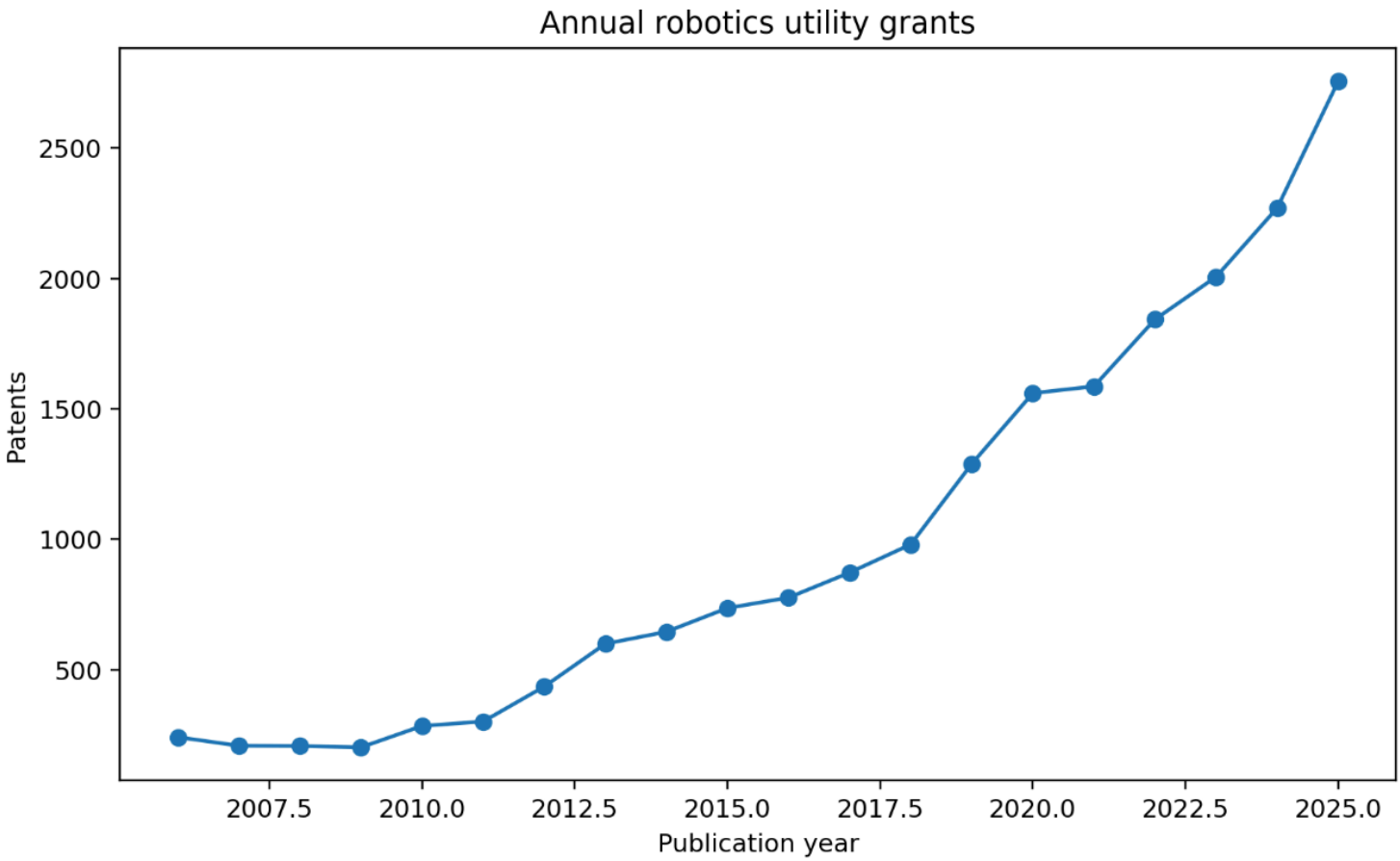


Figure 1

Annual robotics utility grants by publication year

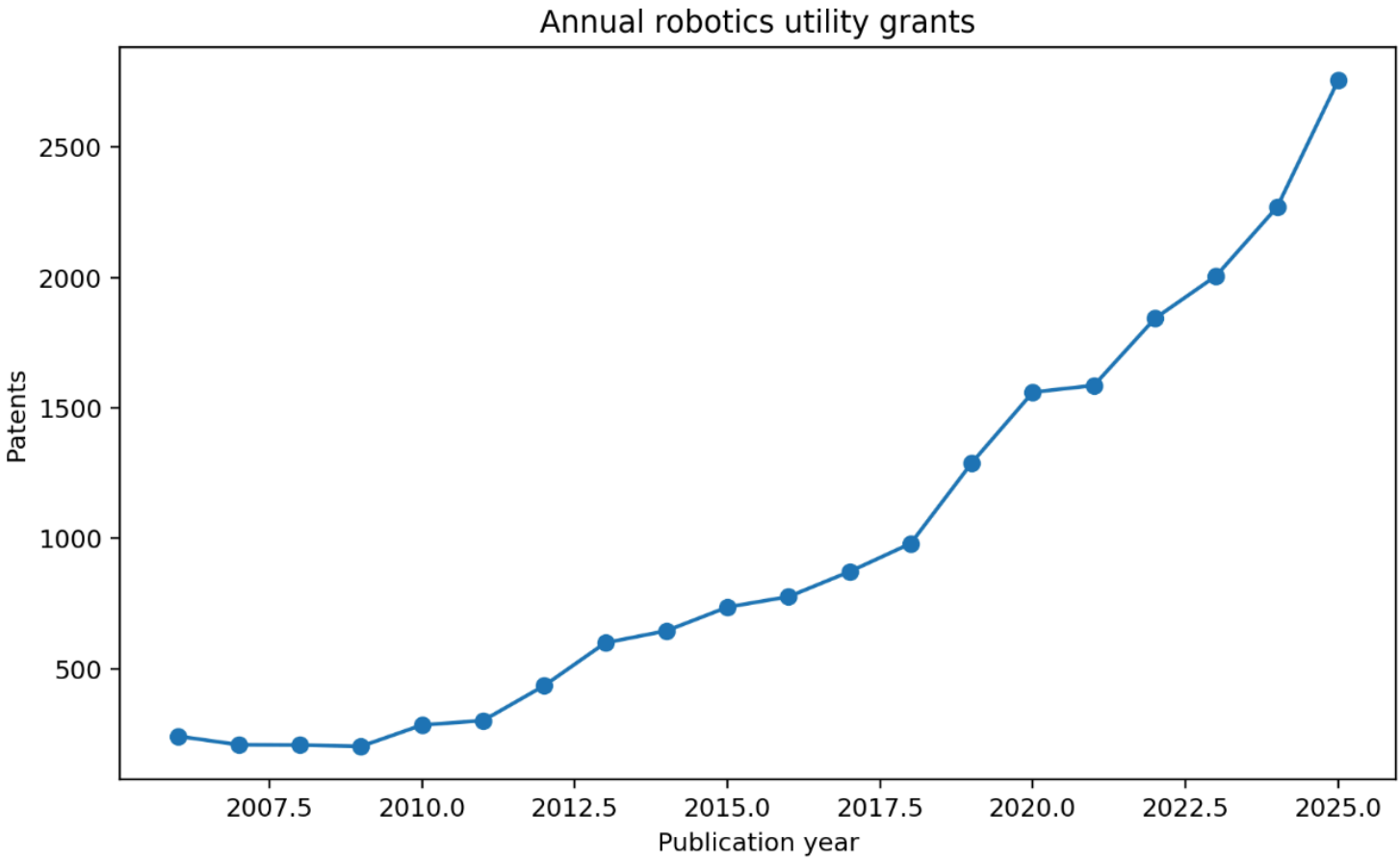


Figure 1

Annual robotics utility grants by publication year

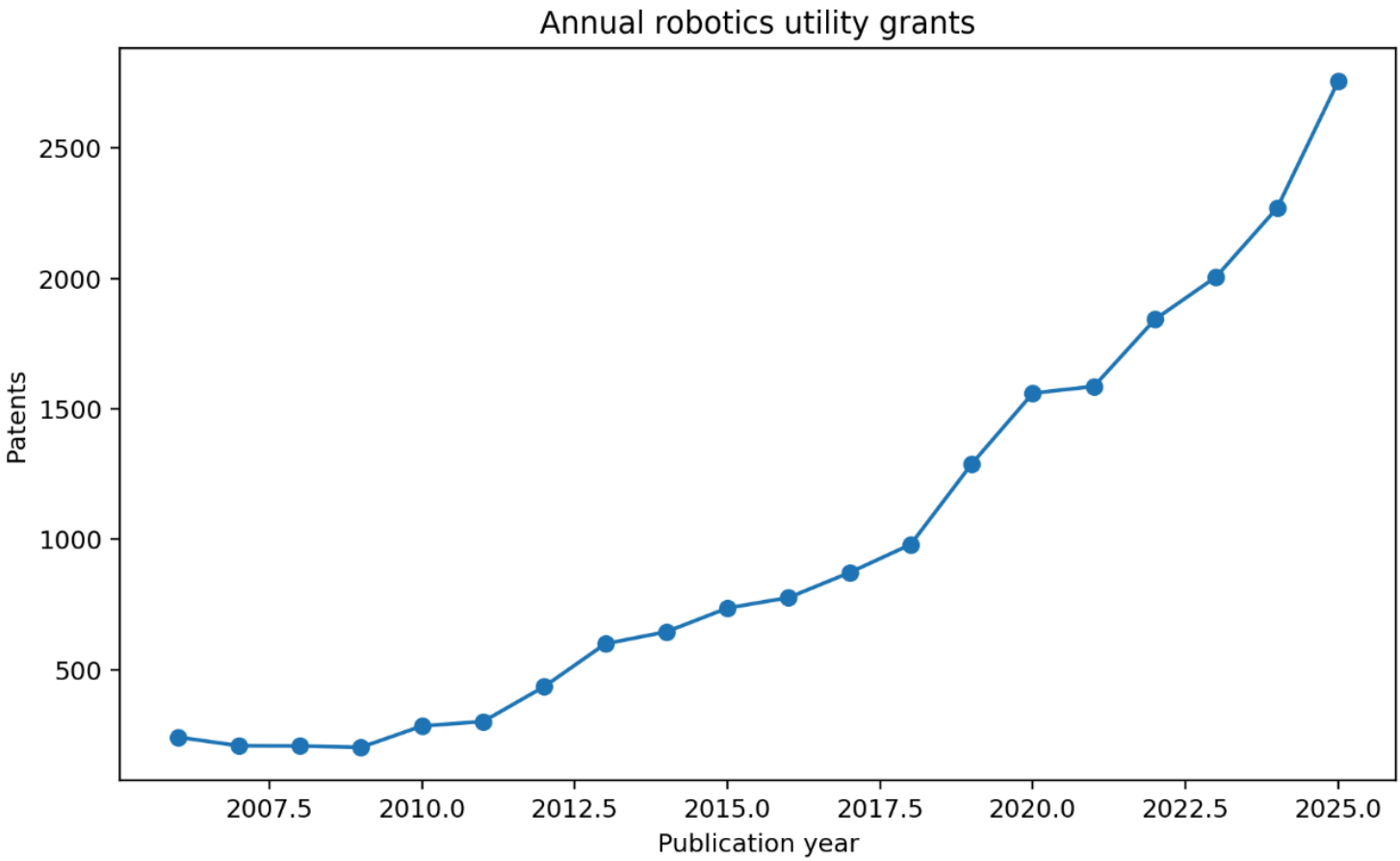


Figure 1

Annual robotics utility grants by publication year

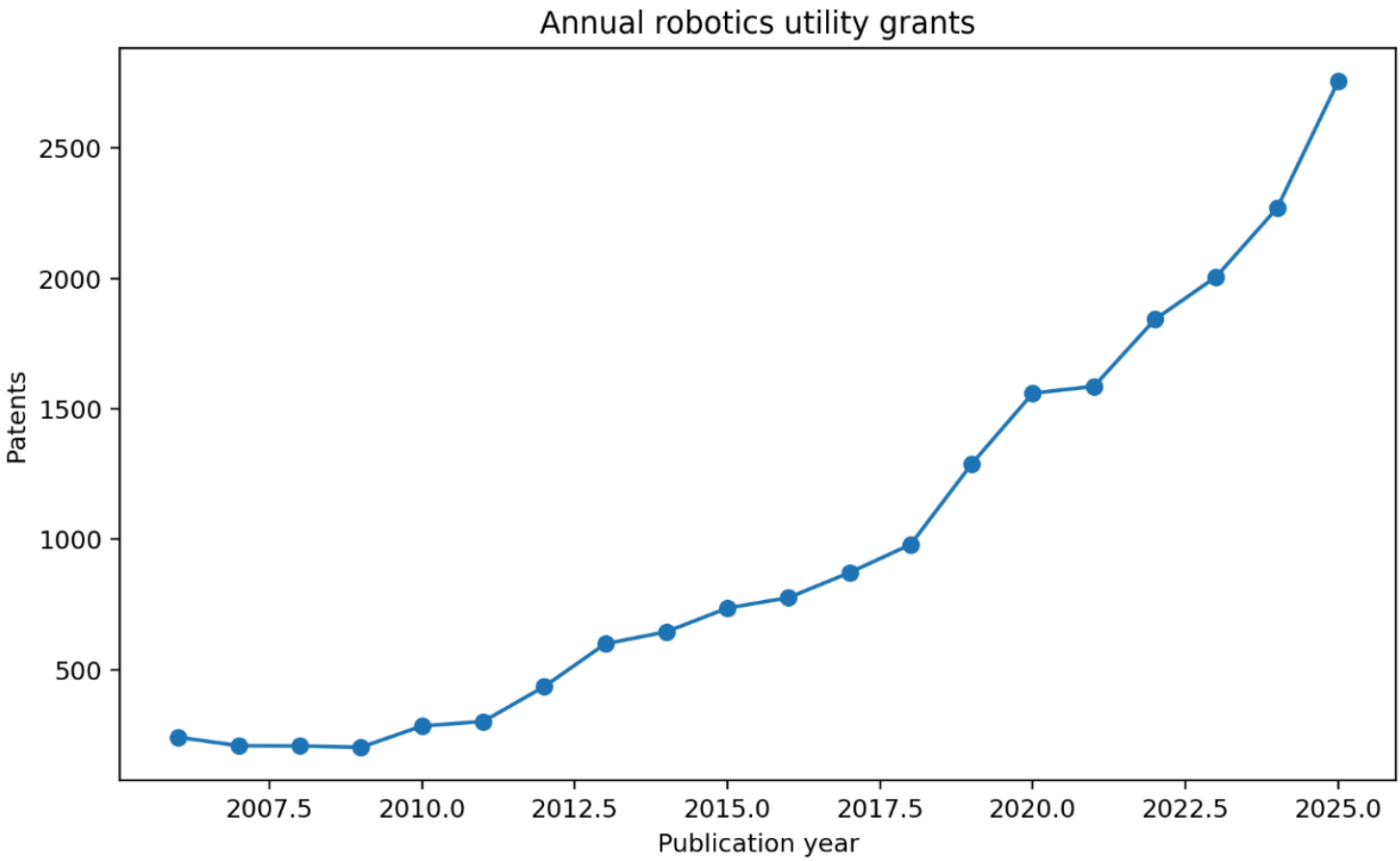


Figure 1

Annual robotics utility grants by publication year

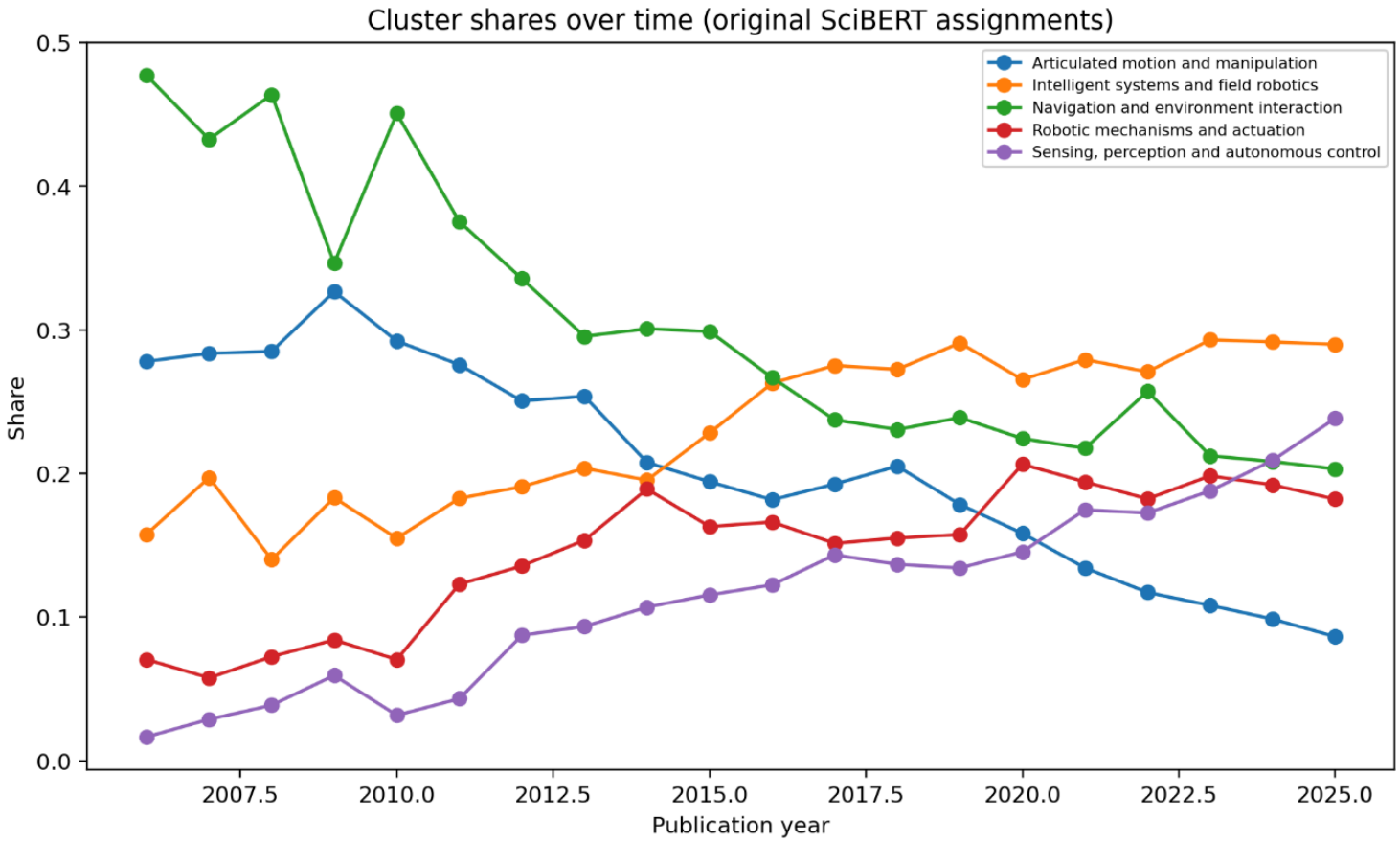


Figure 2

SciBERT cluster shares by publication year

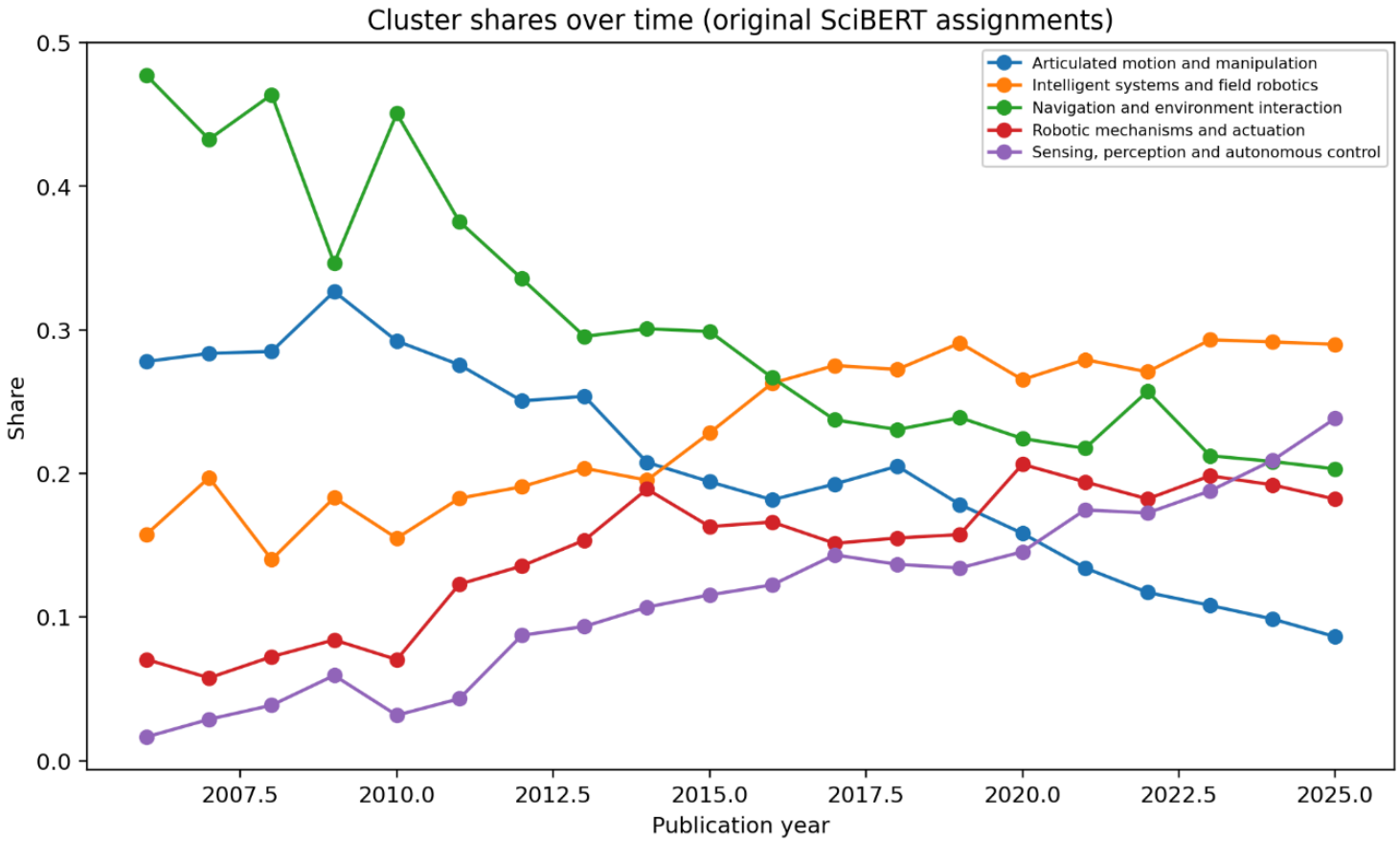


Figure 2

SciBERT cluster shares by publication year

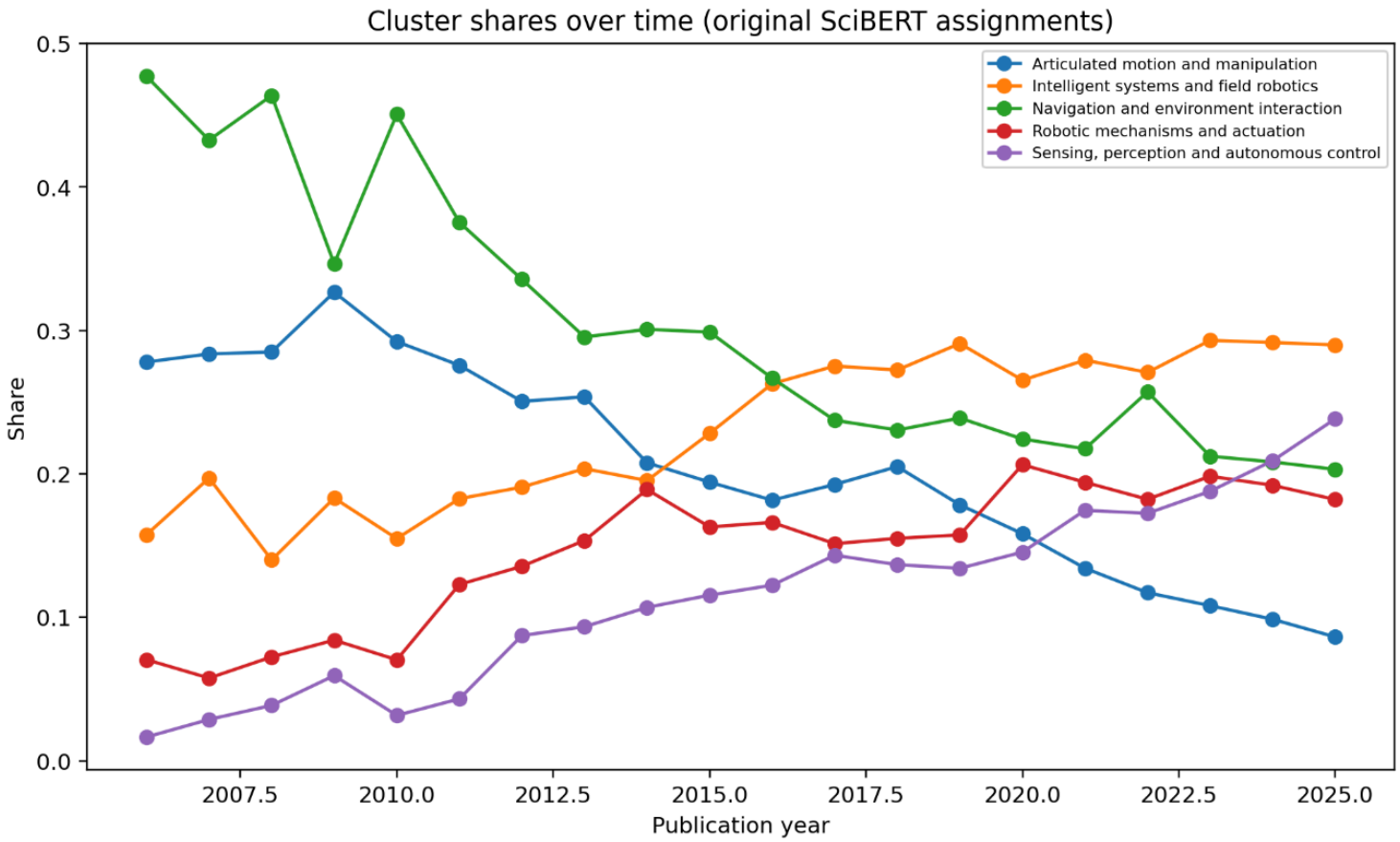


Figure 2

SciBERT cluster shares by publication year

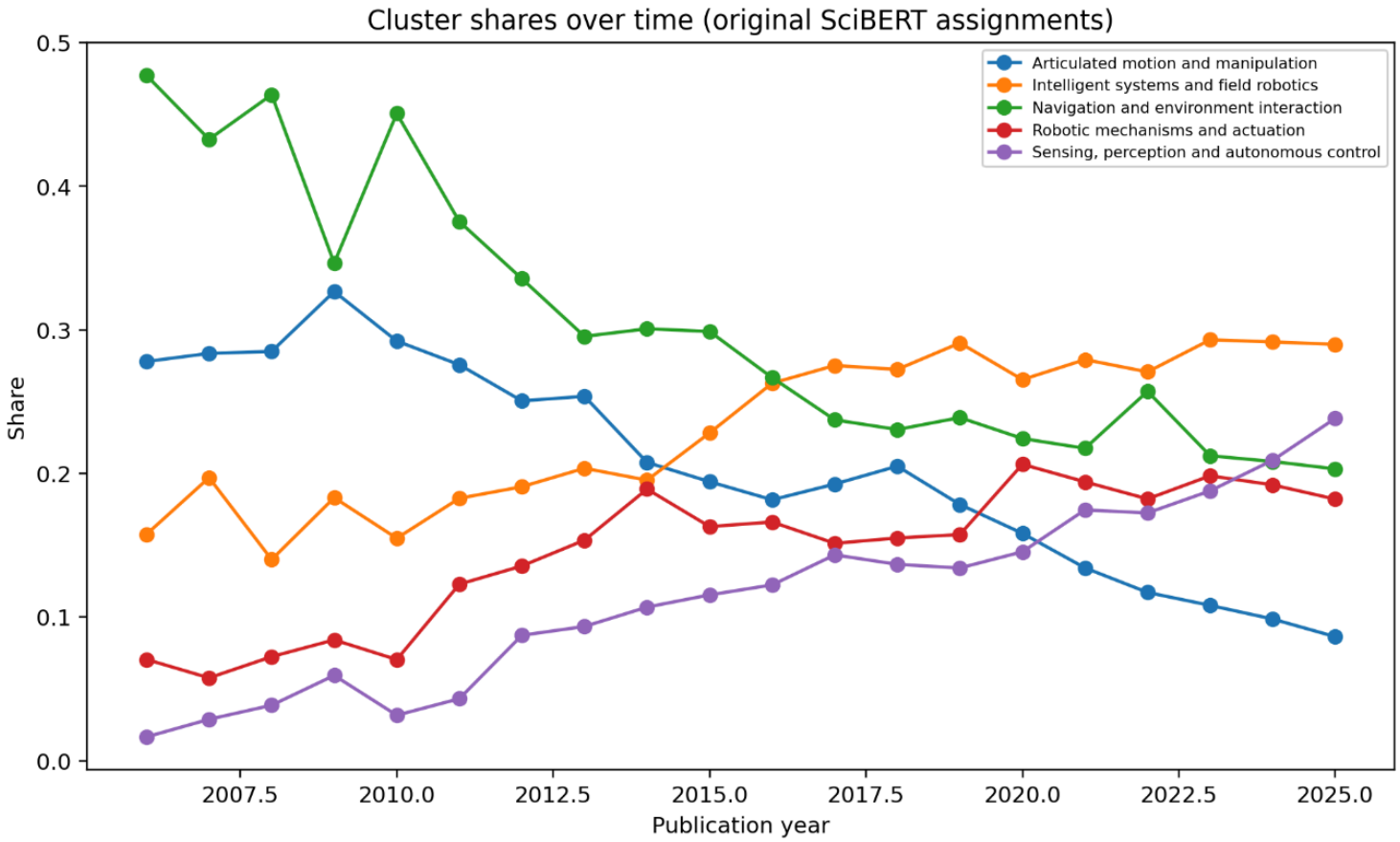


Figure 2

SciBERT cluster shares by publication year

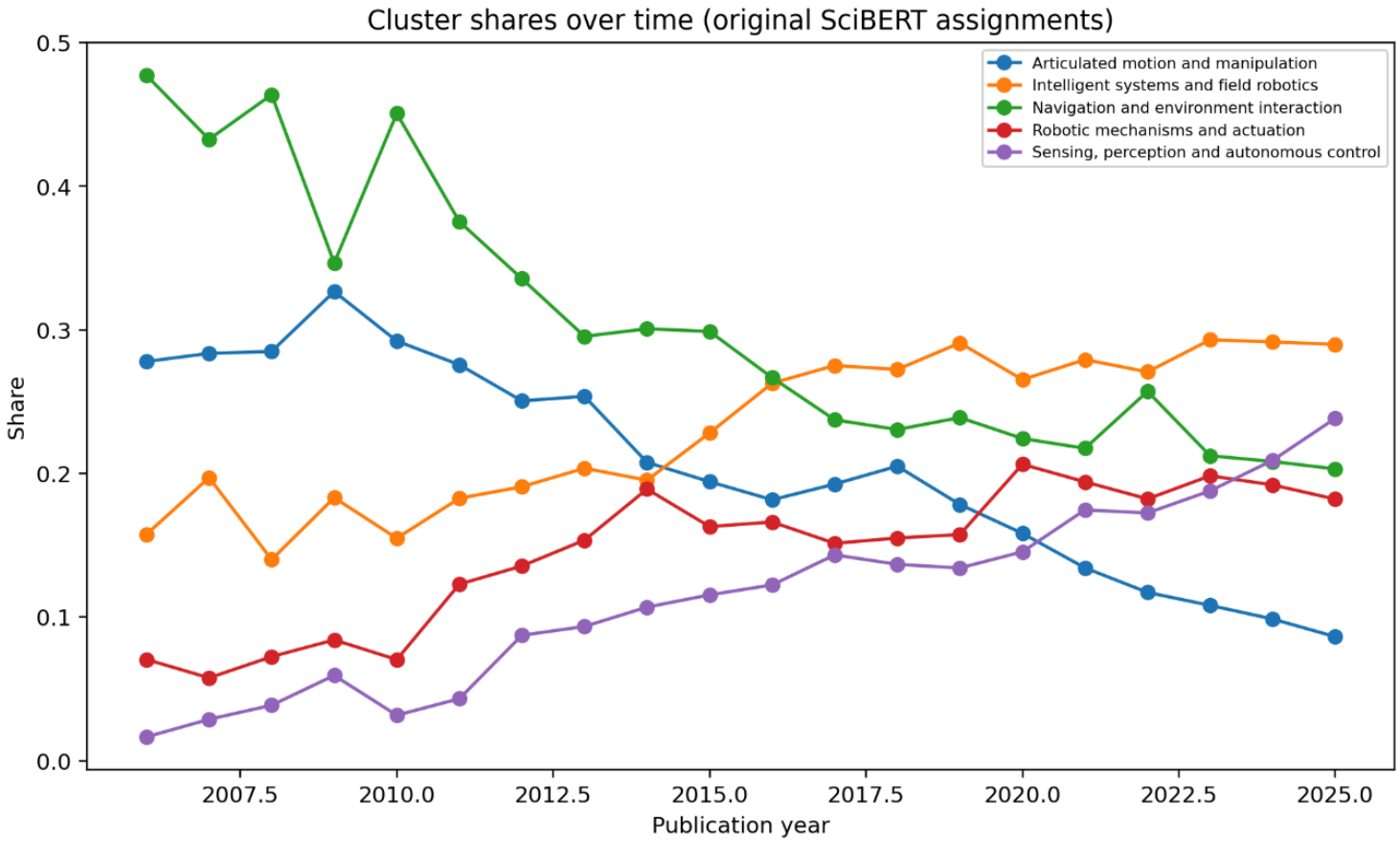


Figure 2

SciBERT cluster shares by publication year

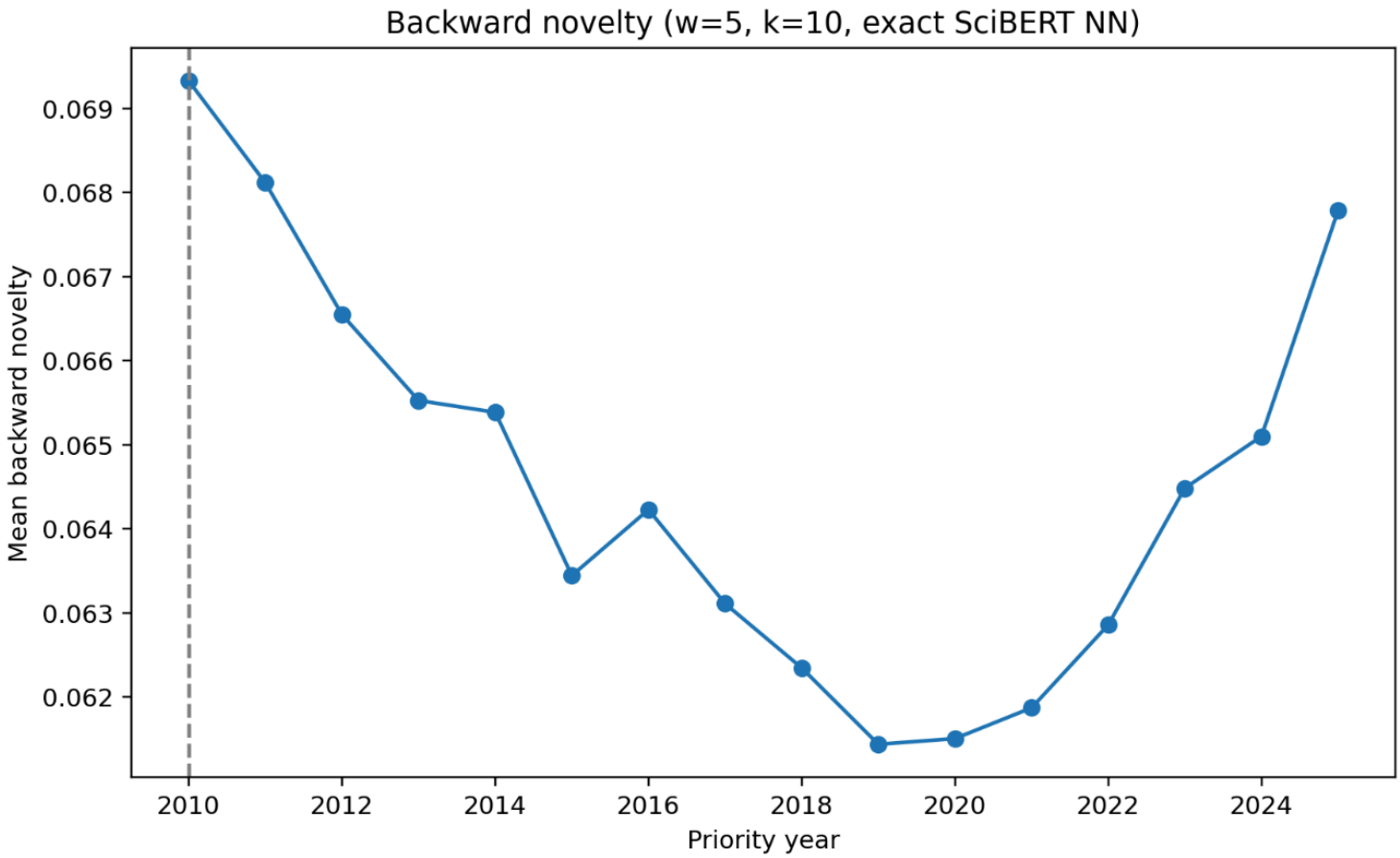


Figure 3

Mean backward novelty by priority year

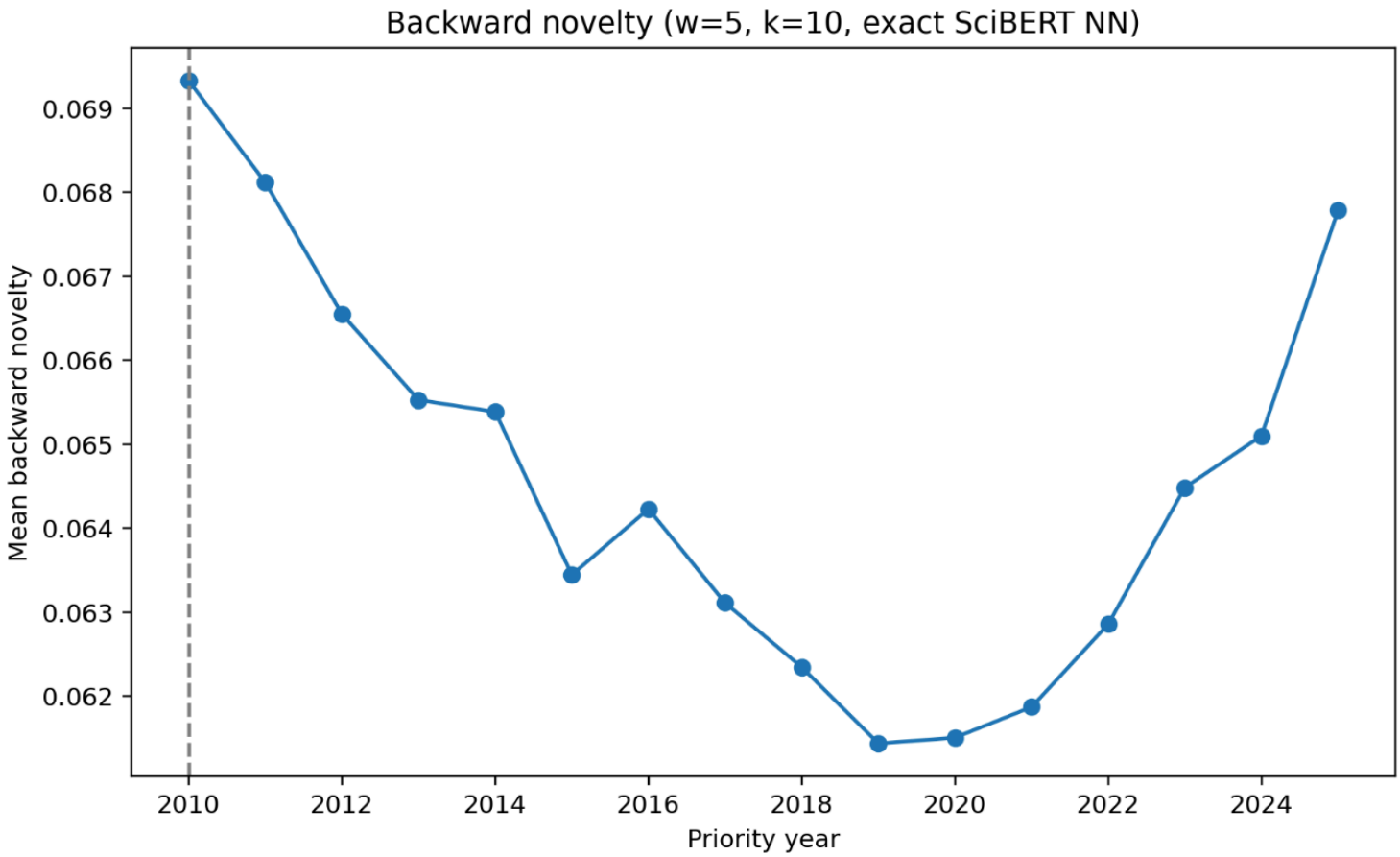


Figure 3

Mean backward novelty by priority year

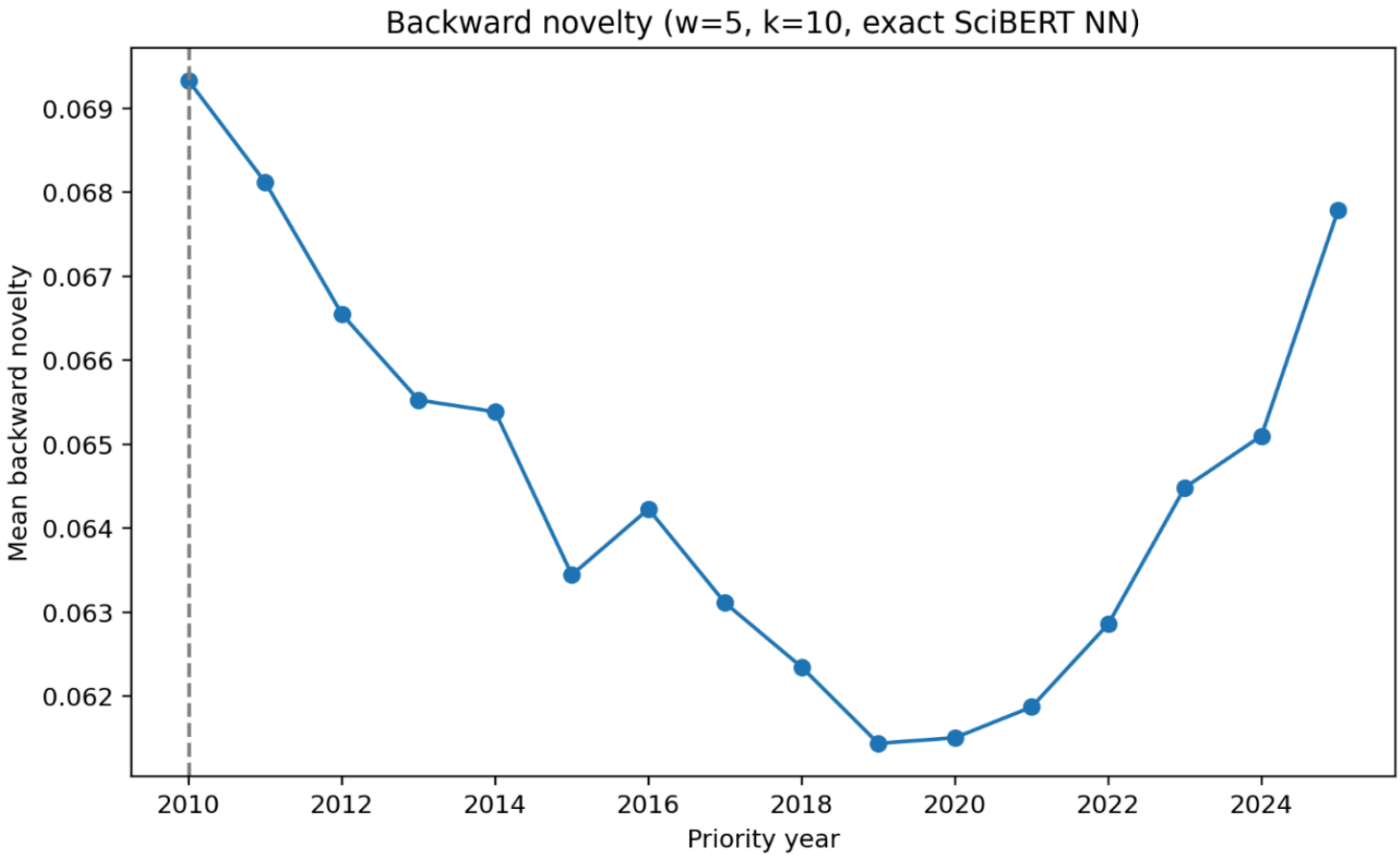


Figure 3

Mean backward novelty by priority year

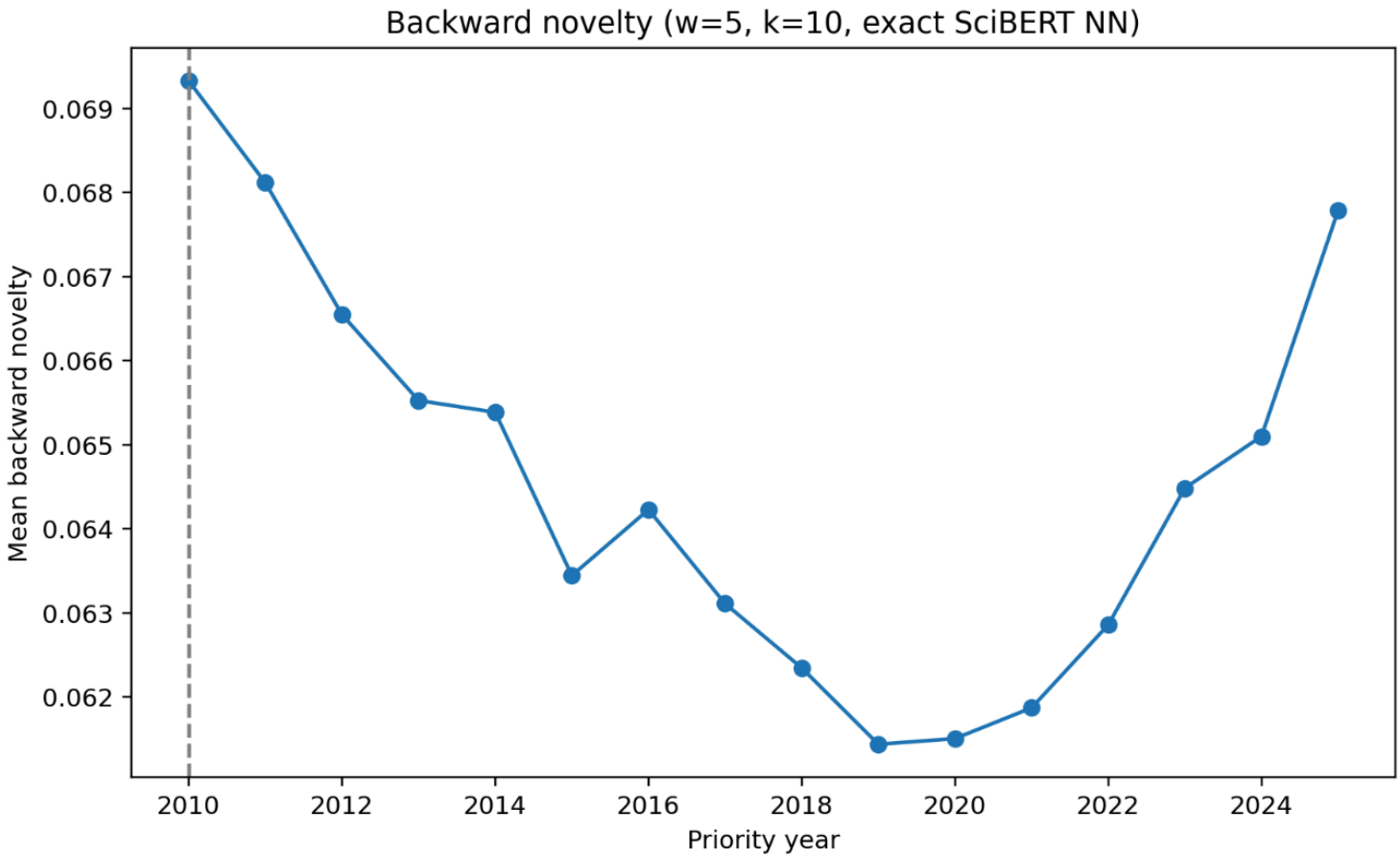


Figure 3

Mean backward novelty by priority year

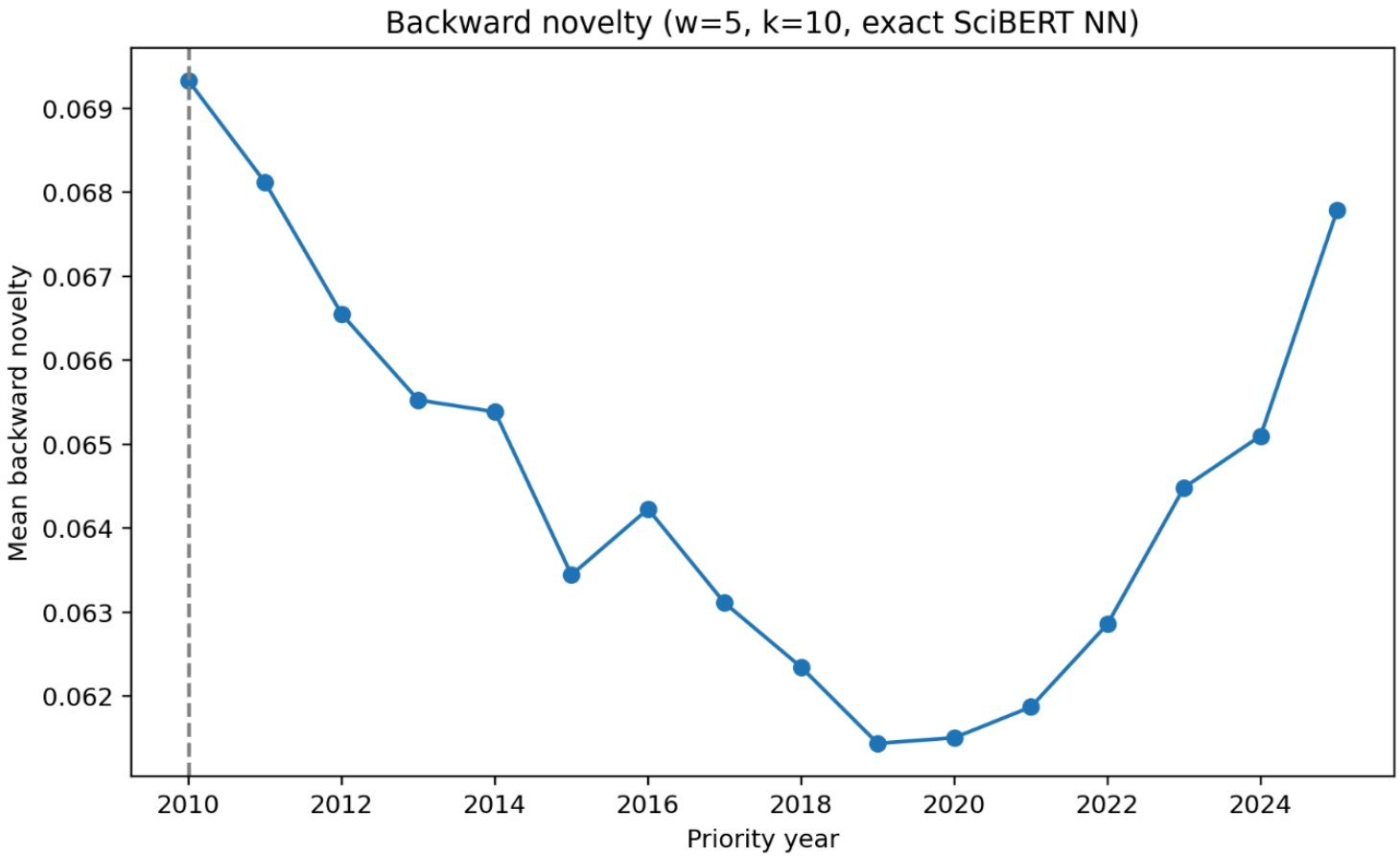


Figure 3

Mean backward novelty by priority year

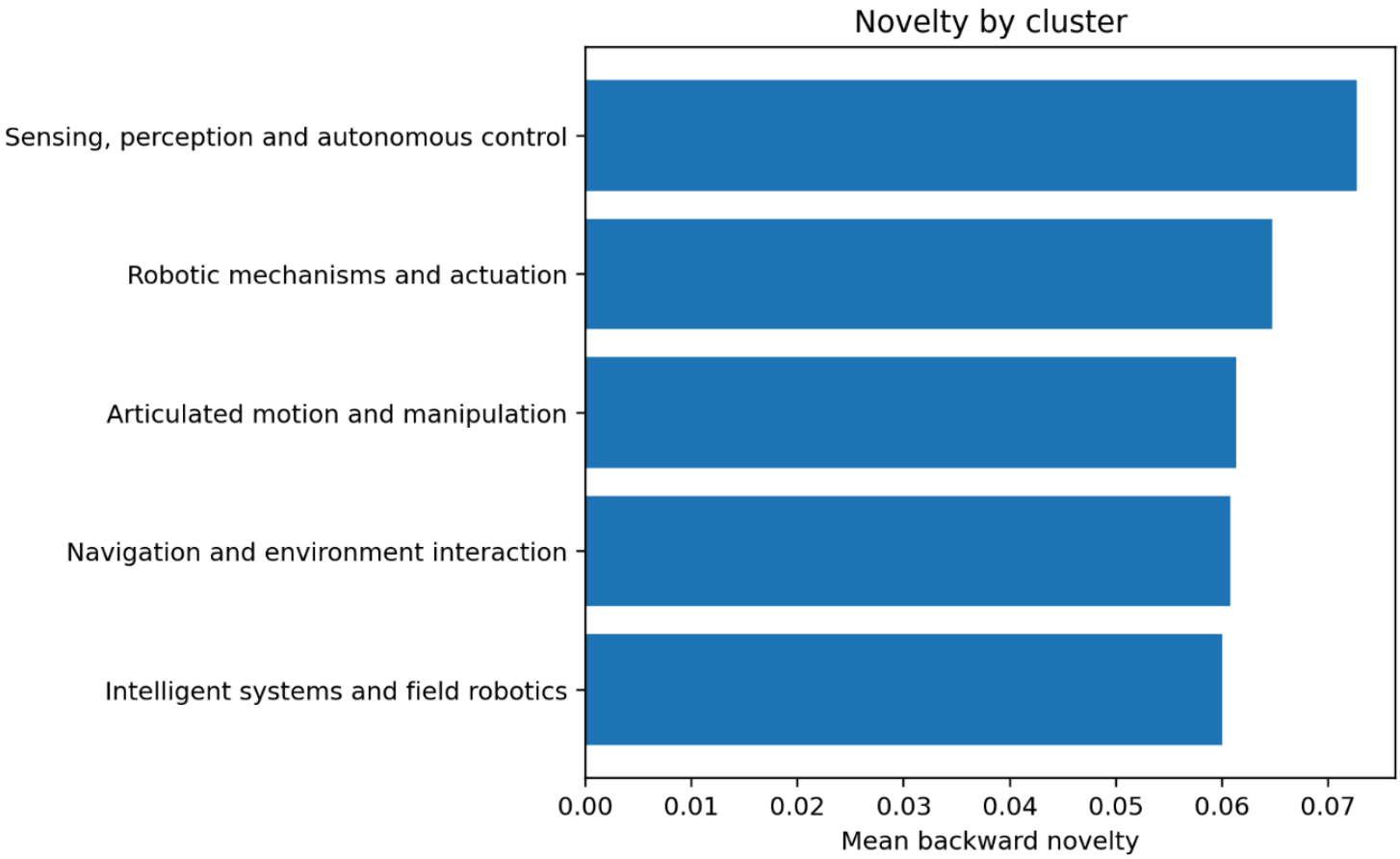


Figure 4

Mean backward novelty by SciBERT cluster

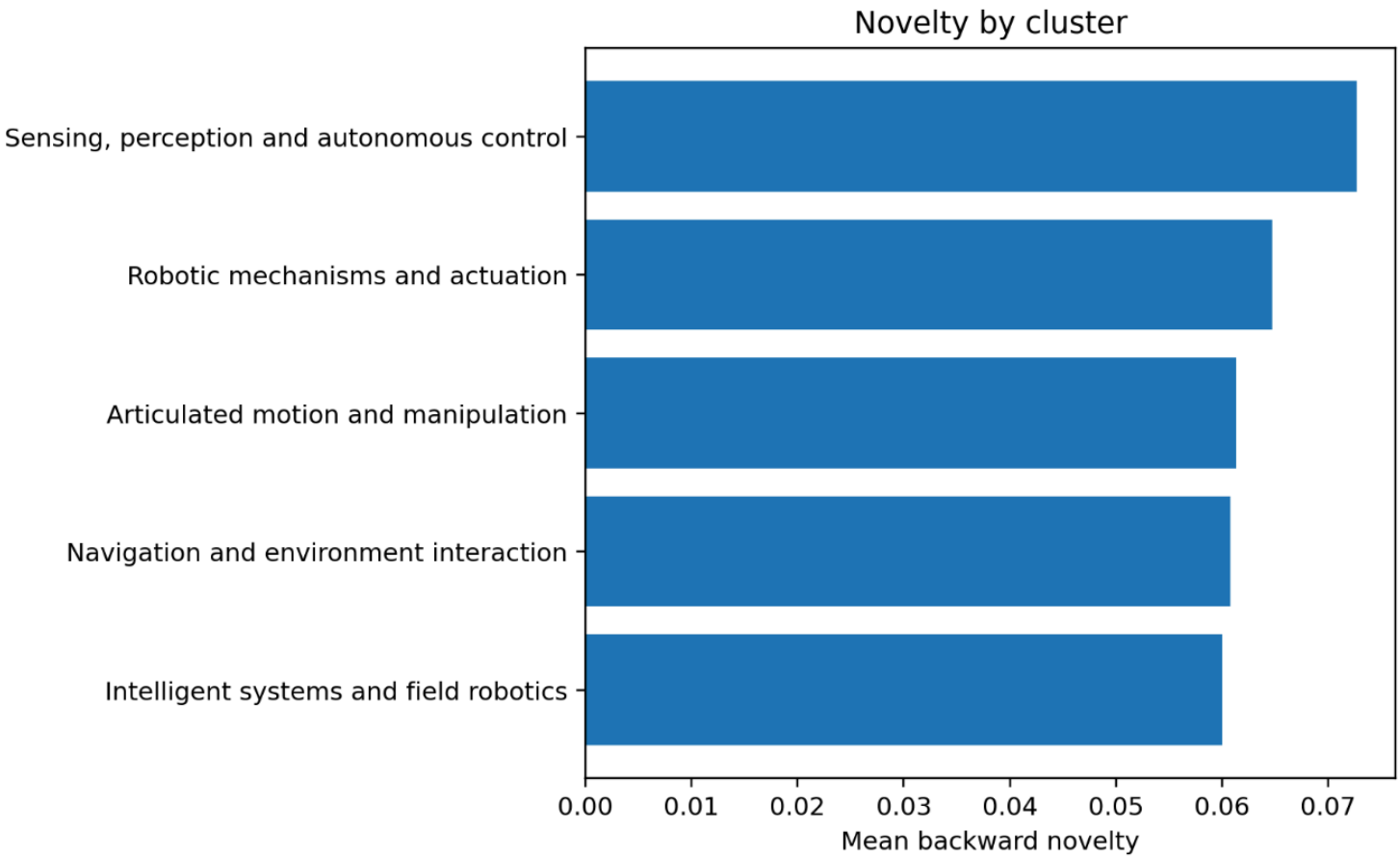


Figure 4

Mean backward novelty by SciBERT cluster

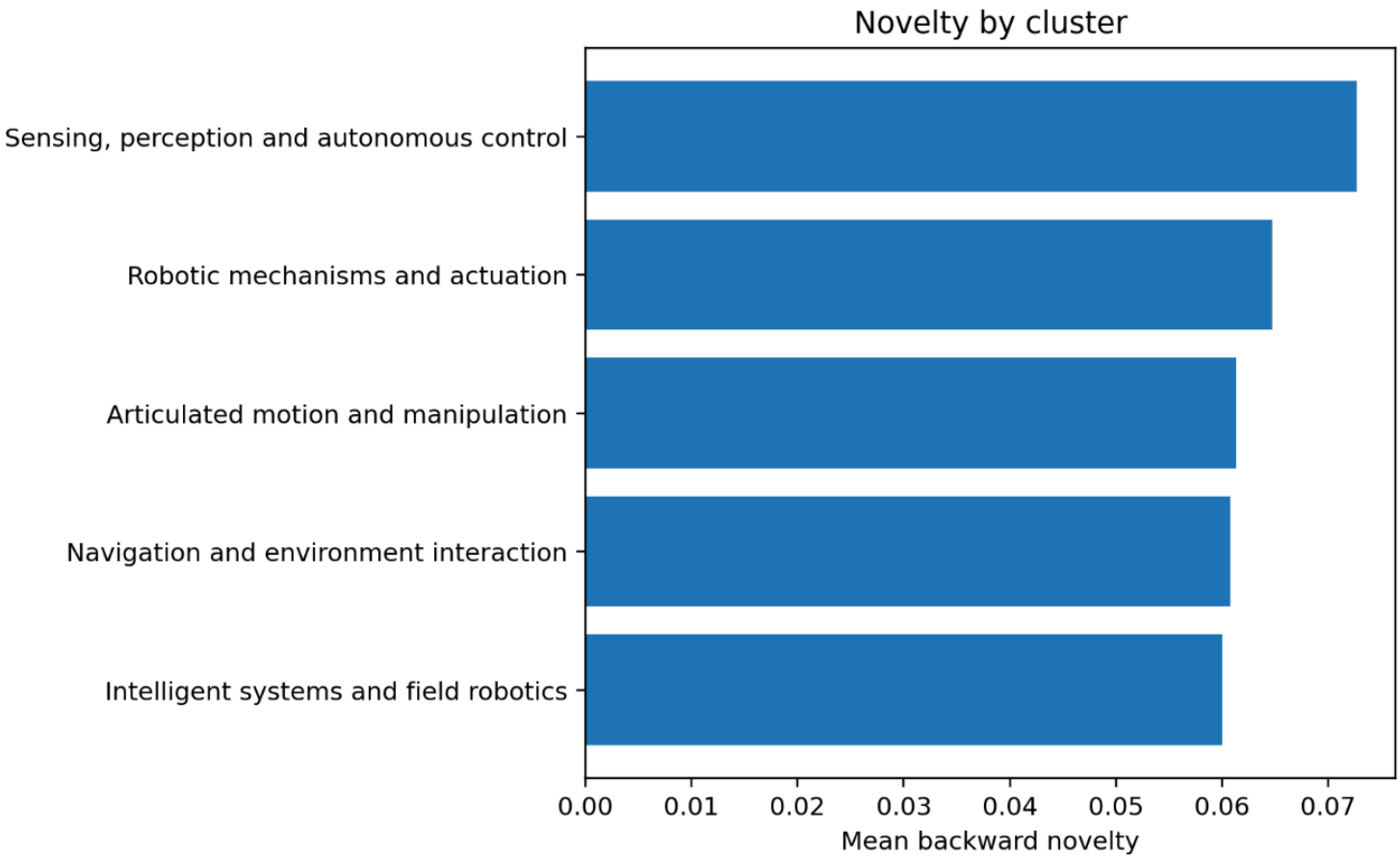


Figure 4

Mean backward novelty by SciBERT cluster

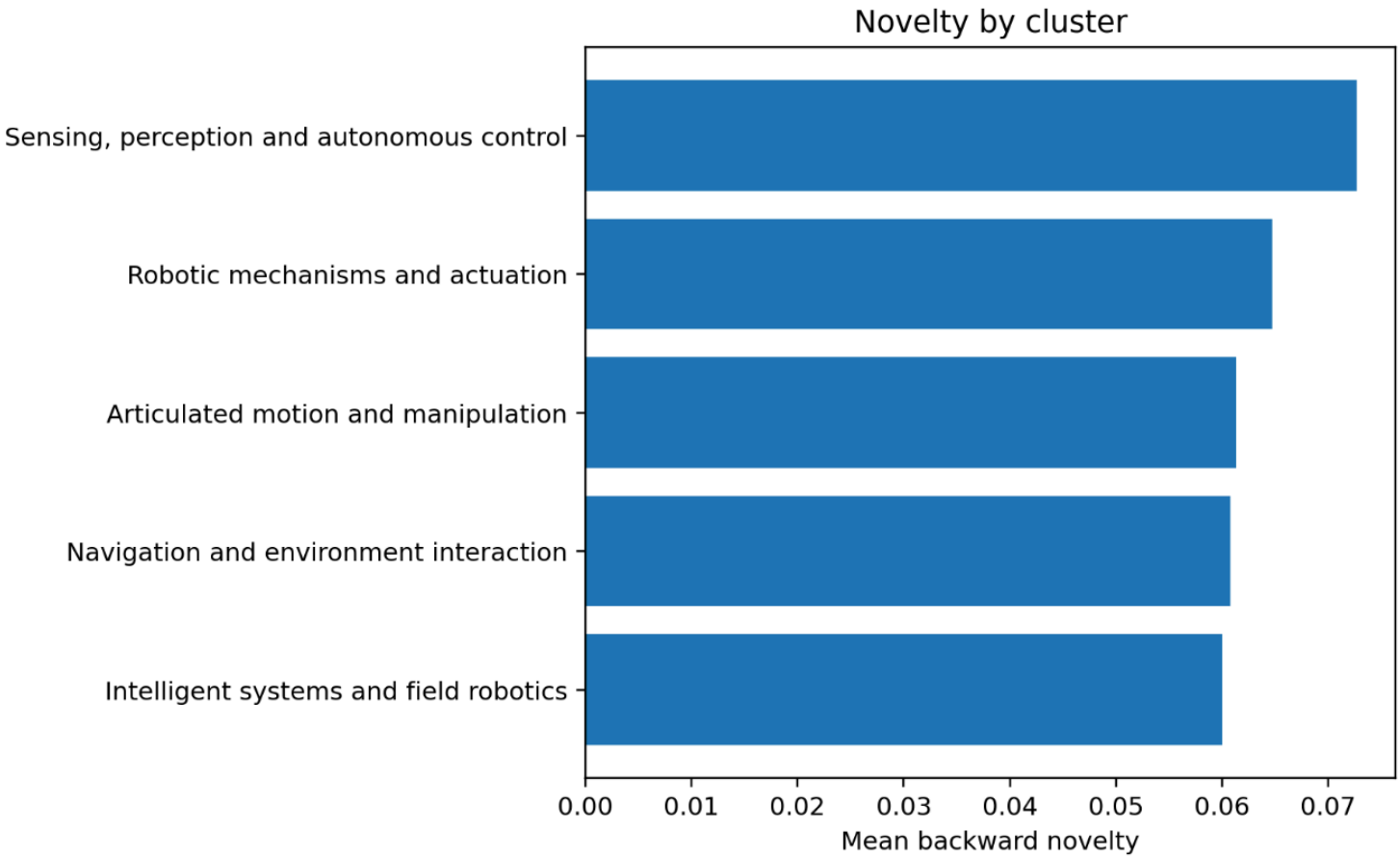


Figure 4

Mean backward novelty by SciBERT cluster

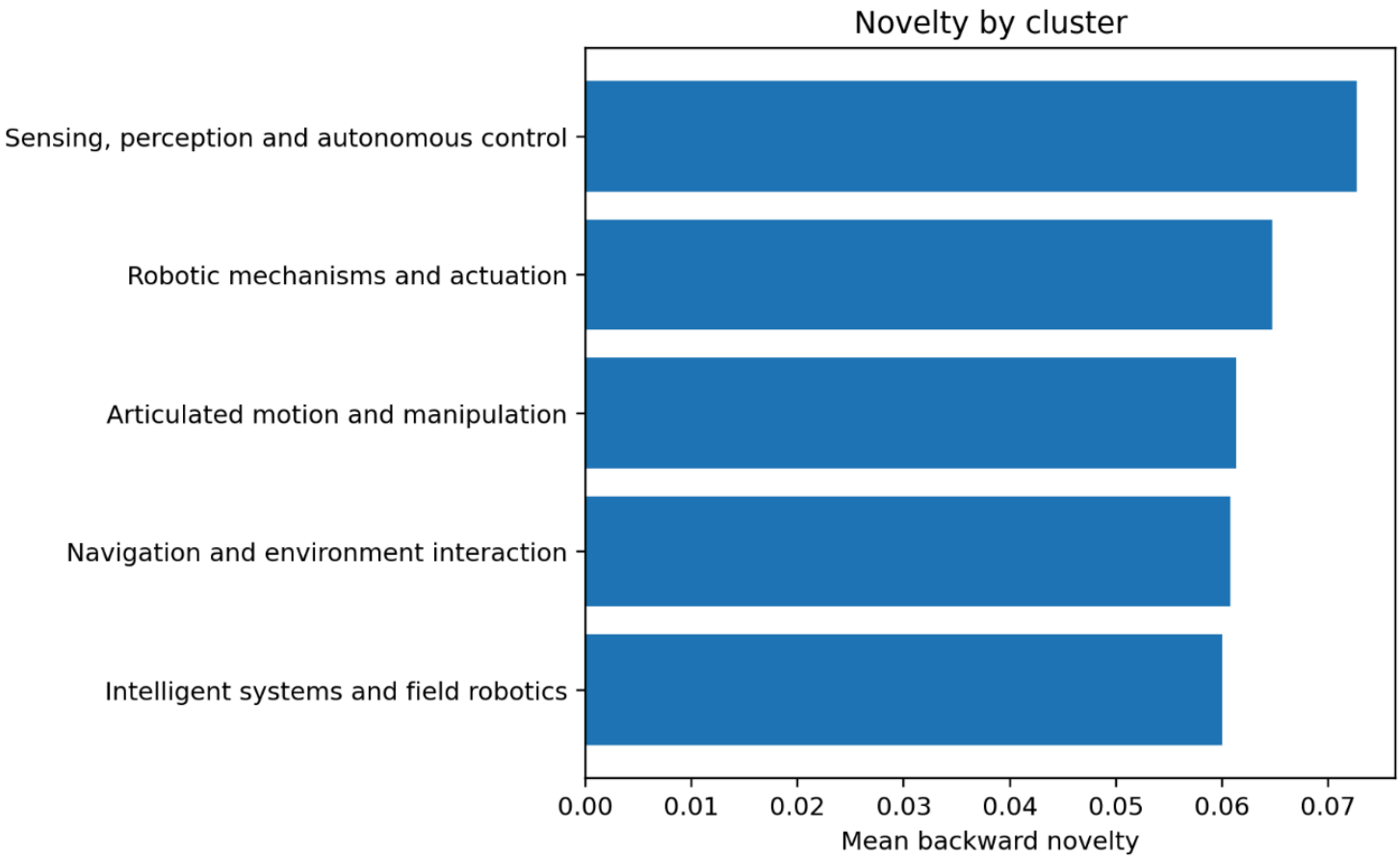


Figure 4

Mean backward novelty by SciBERT cluster

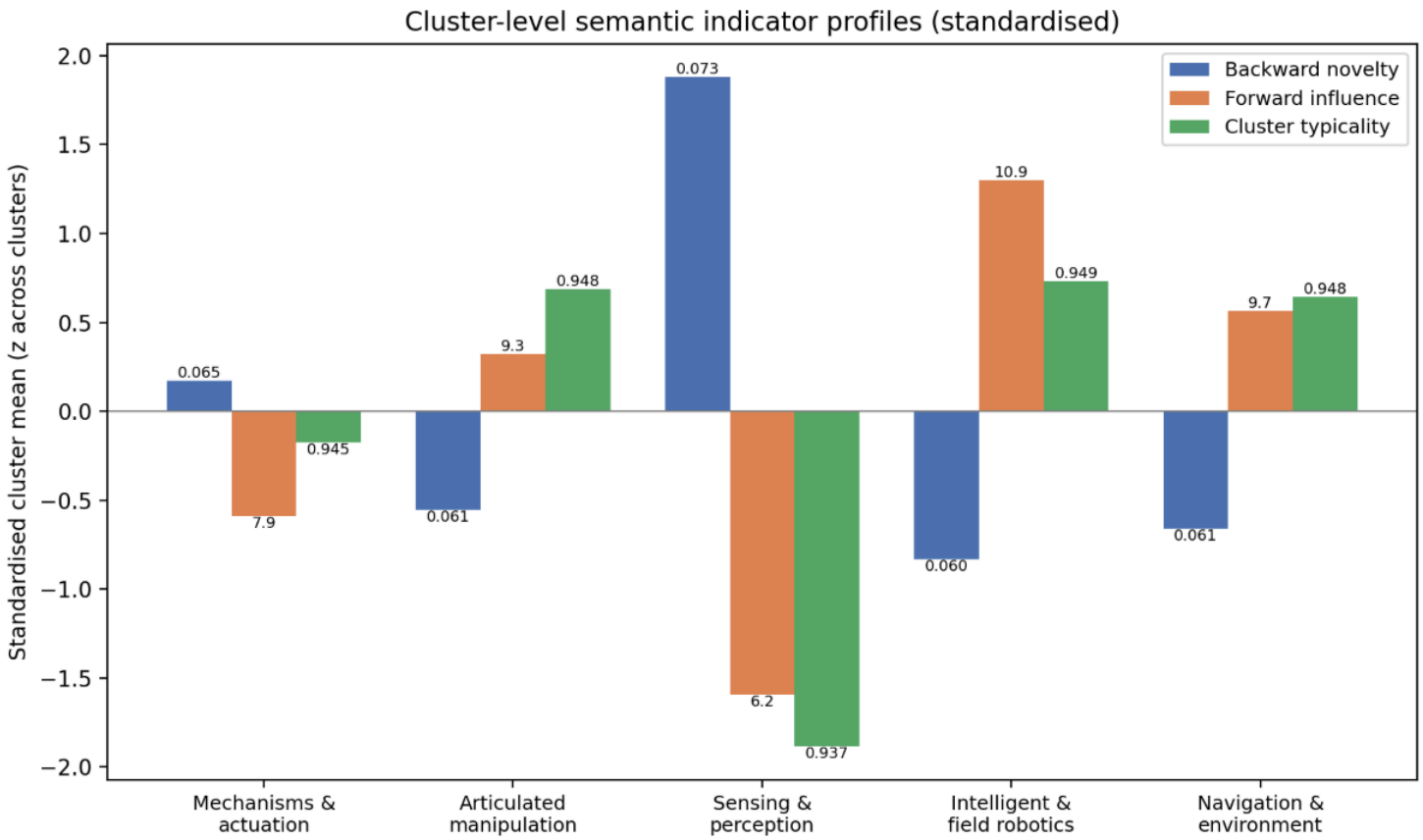


Figure 5

Cluster-level profiles of backward novelty, forward influence, and cluster typicality. Bars show cluster means standardised across the five clusters; numeric labels report the corresponding raw cluster means.

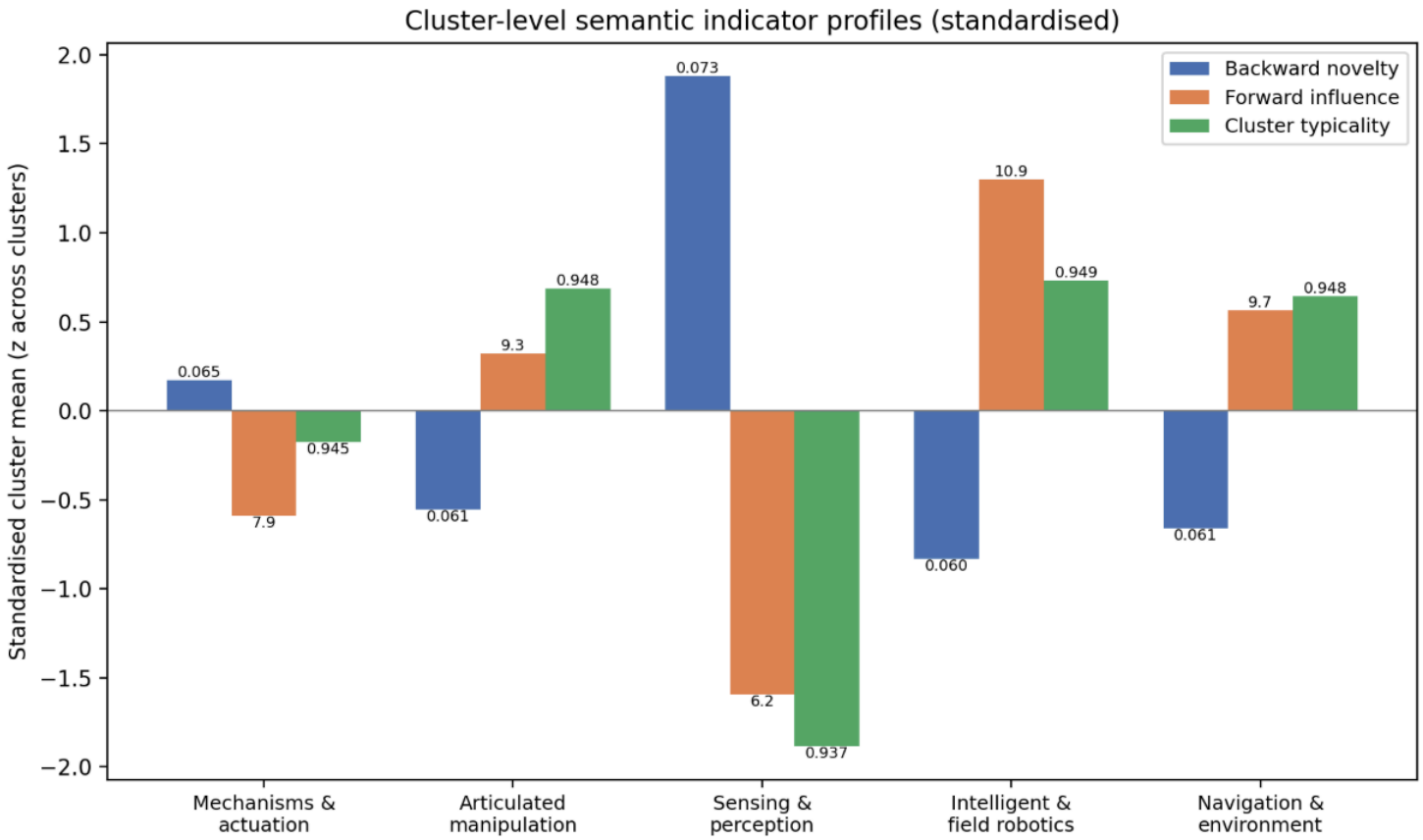


Figure 5

Cluster-level profiles of backward novelty, forward influence, and cluster typicality. Bars show cluster means standardised across the five clusters; numeric labels report the corresponding raw cluster means.

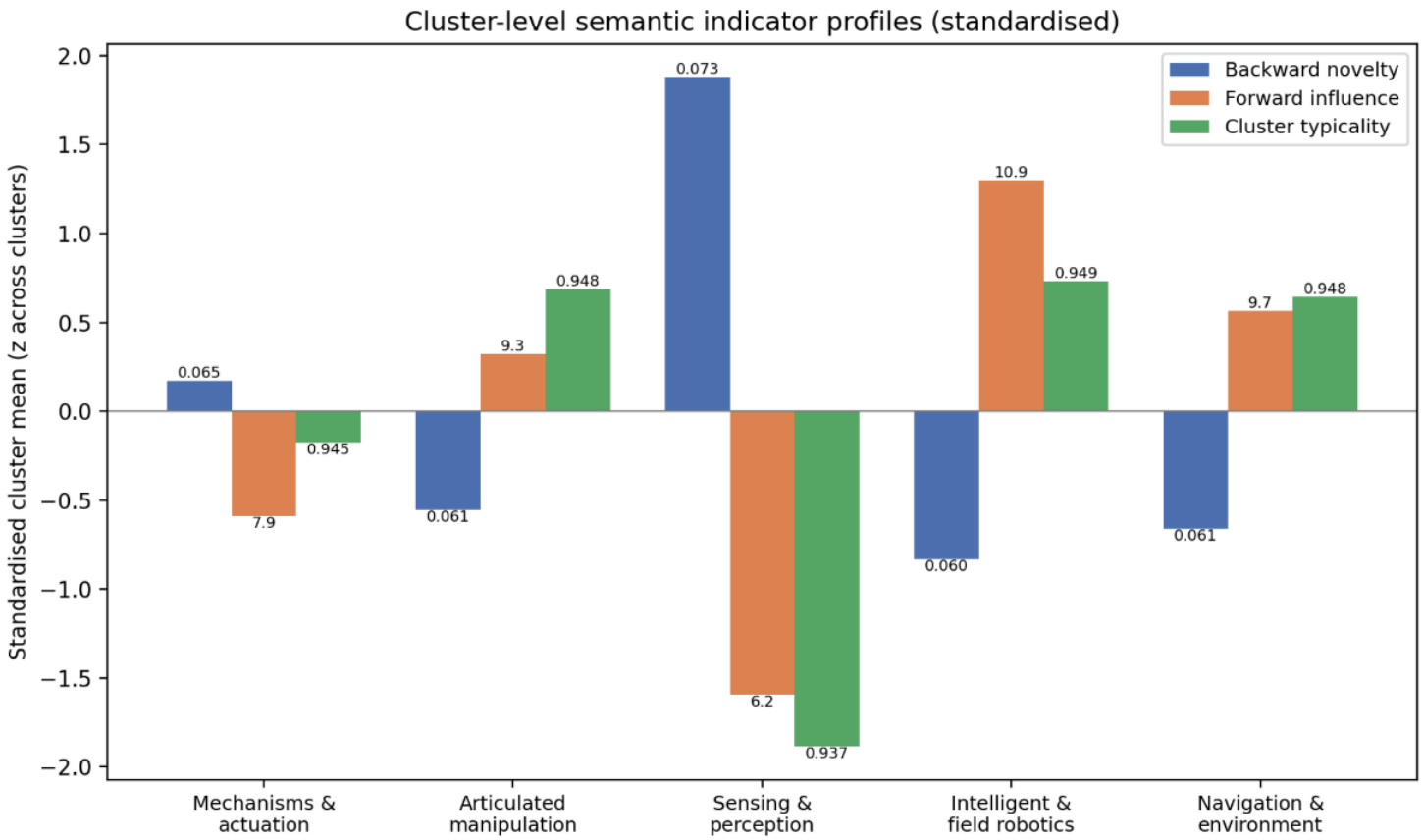


Figure 5

Cluster-level profiles of backward novelty, forward influence, and cluster typicality. Bars show cluster means standardised across the five clusters; numeric labels report the corresponding raw cluster means.

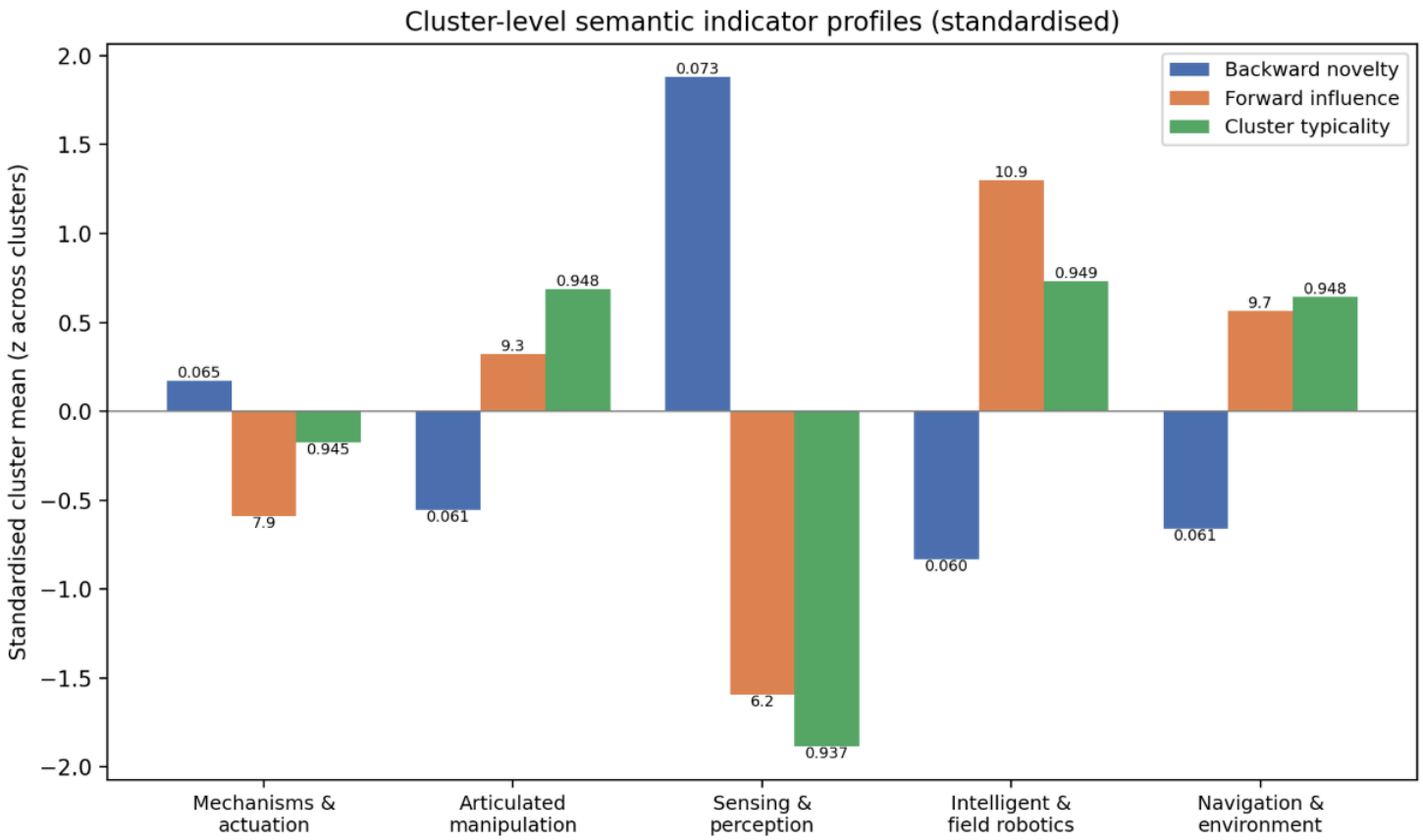


Figure 5

Cluster-level profiles of backward novelty, forward influence, and cluster typicality. Bars show cluster means standardised across the five clusters; numeric labels report the corresponding raw cluster means.

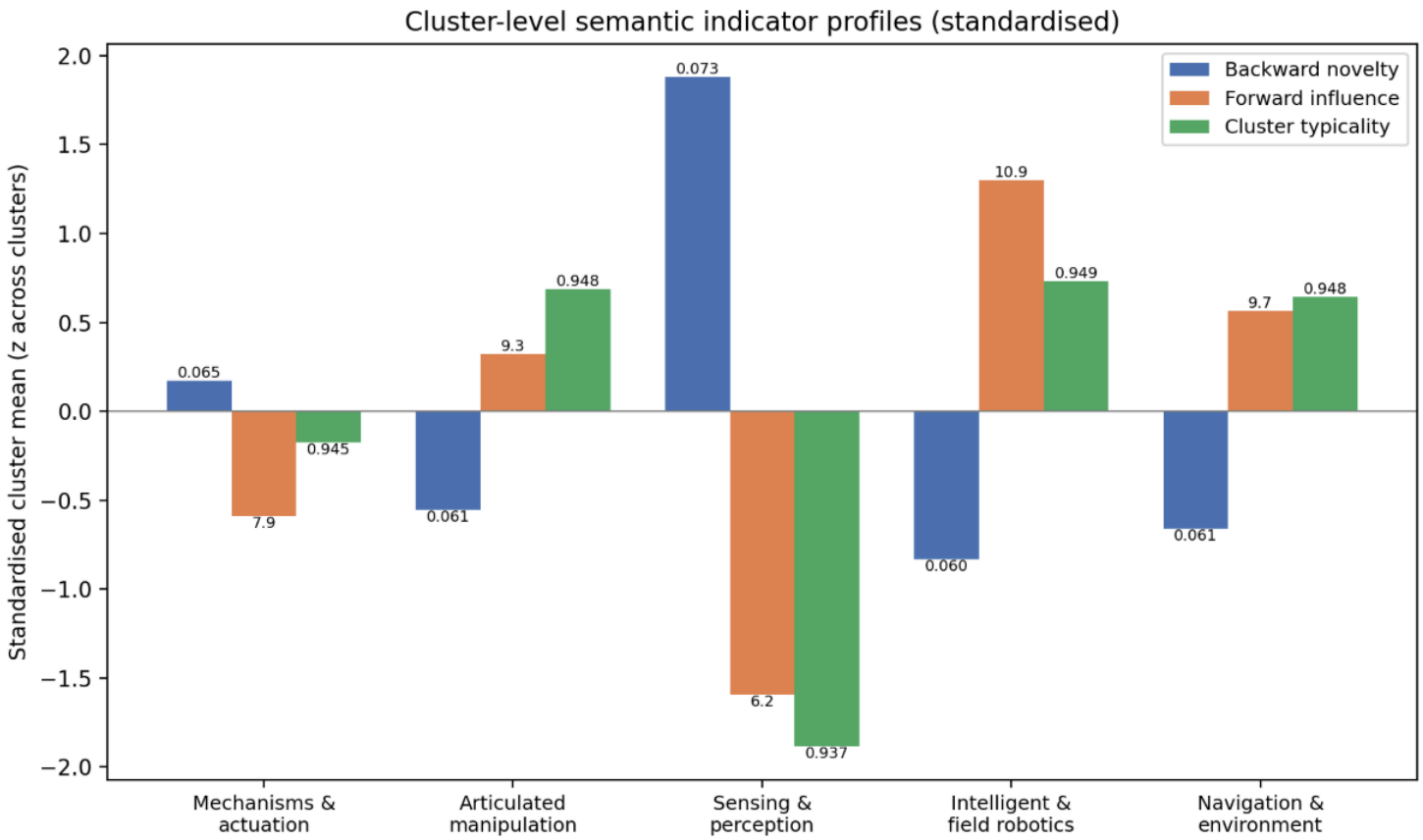


Figure 5

Cluster-level profiles of backward novelty, forward influence, and cluster typicality. Bars show cluster means standardised across the five clusters; numeric labels report the corresponding raw cluster means.

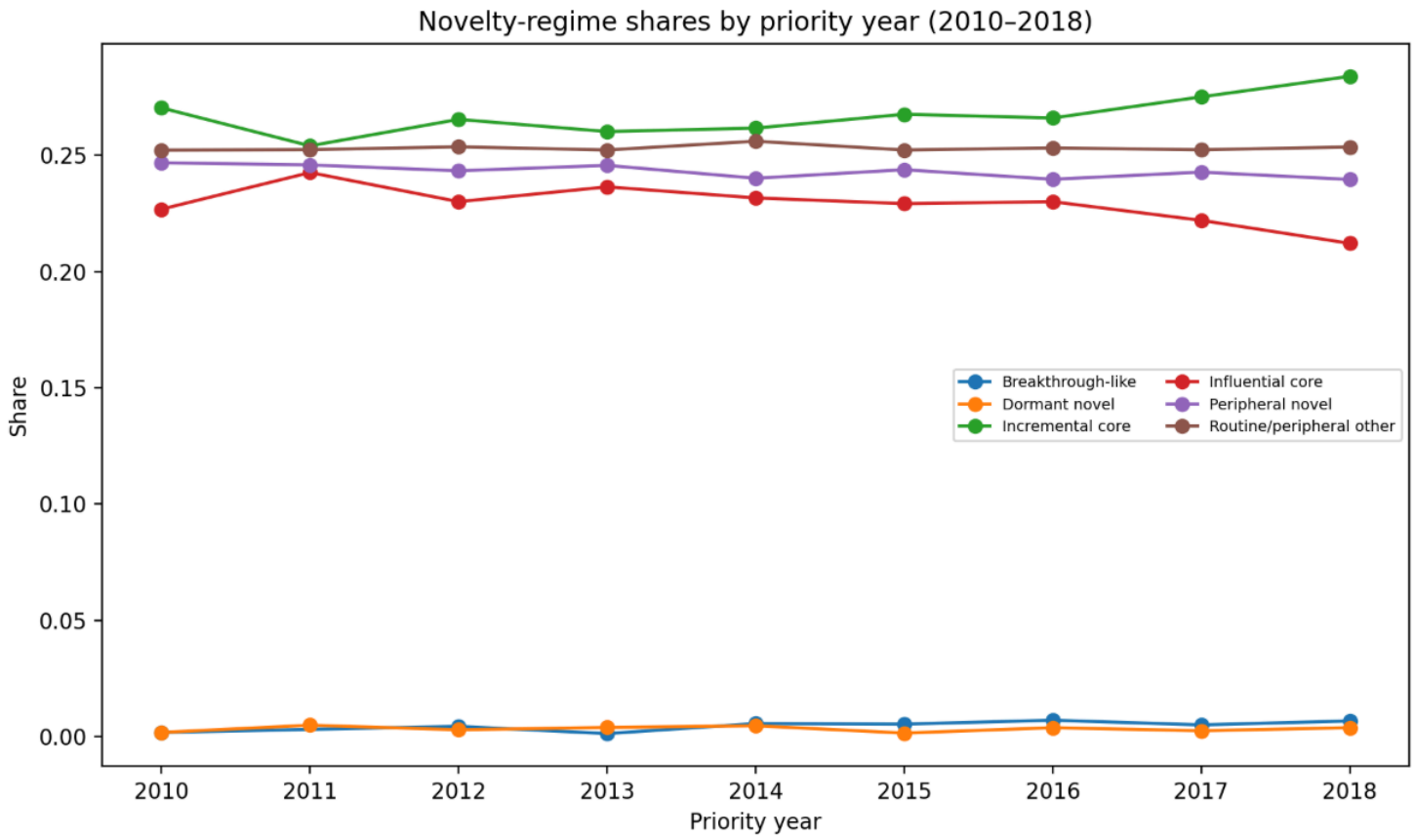


Figure 6

Novelty-regime shares by priority year, 2010–2018 influence-valid cohorts

Novelty-regime shares by priority year (2010–2018)
Share
Priority year
0.00
0.05
0.10
0.15
0.20
0.25
2010
2011
2012
2013
2014
2015
2016
2017
2018
Breakthrough-like
Dormant novel
Incremental core
Influential core
Peripheral novel
Routine/peripheral other

Figure 6

Novelty-regime shares by priority year, 2010–2018 influence-valid cohorts

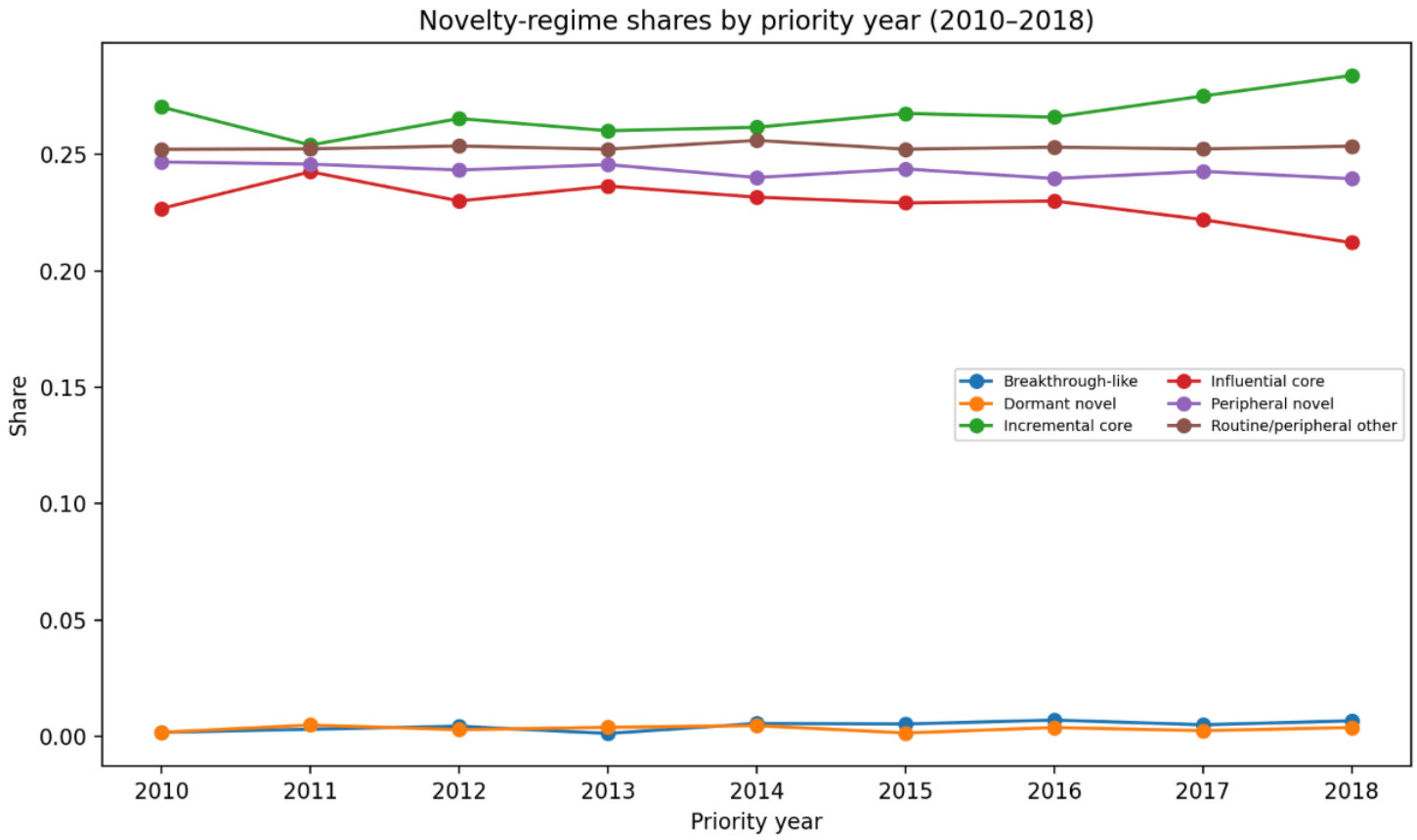


Figure 6

Novelty-regime shares by priority year, 2010–2018 influence-valid cohorts

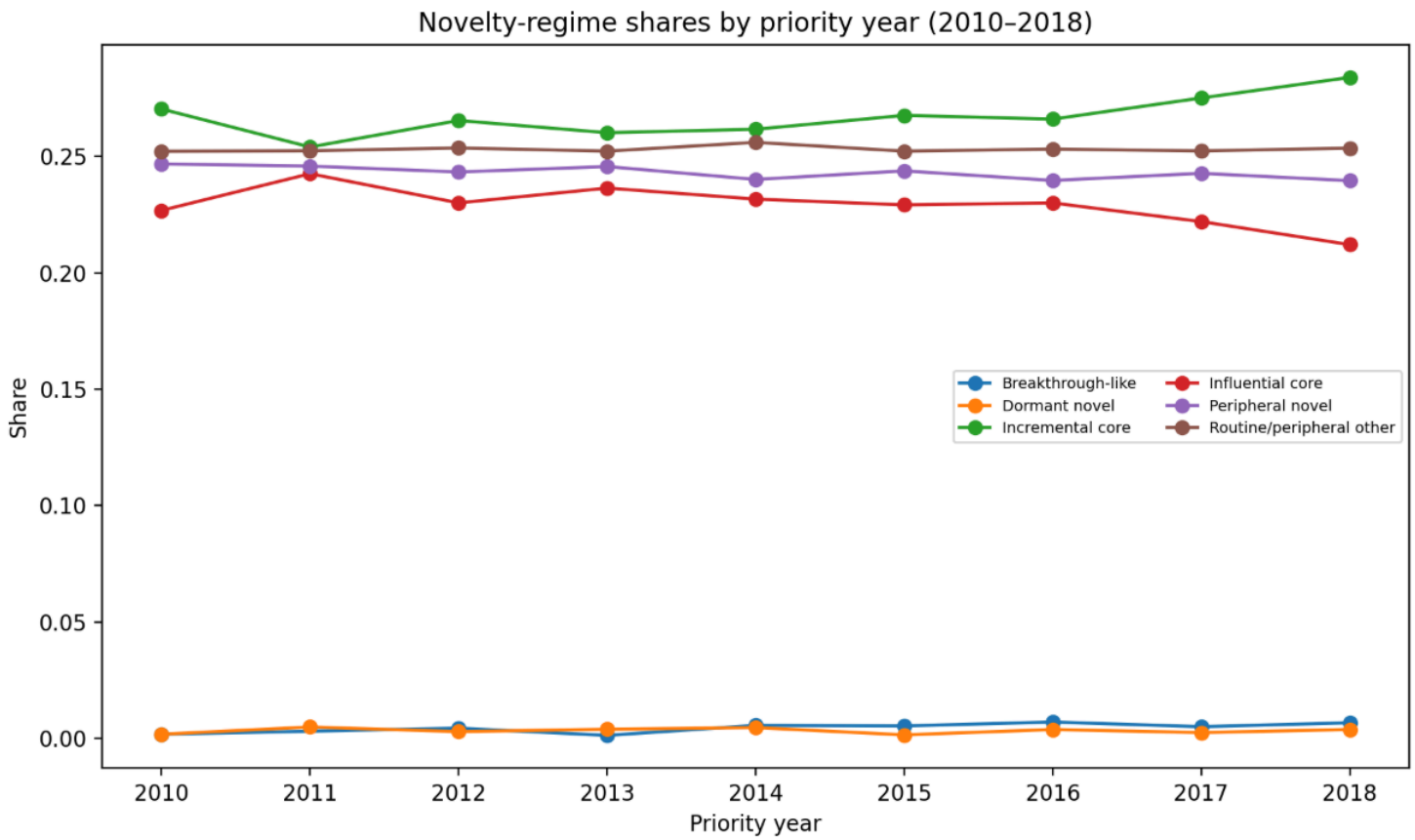


Figure 6

Novelty-regime shares by priority year, 2010–2018 influence-valid cohorts

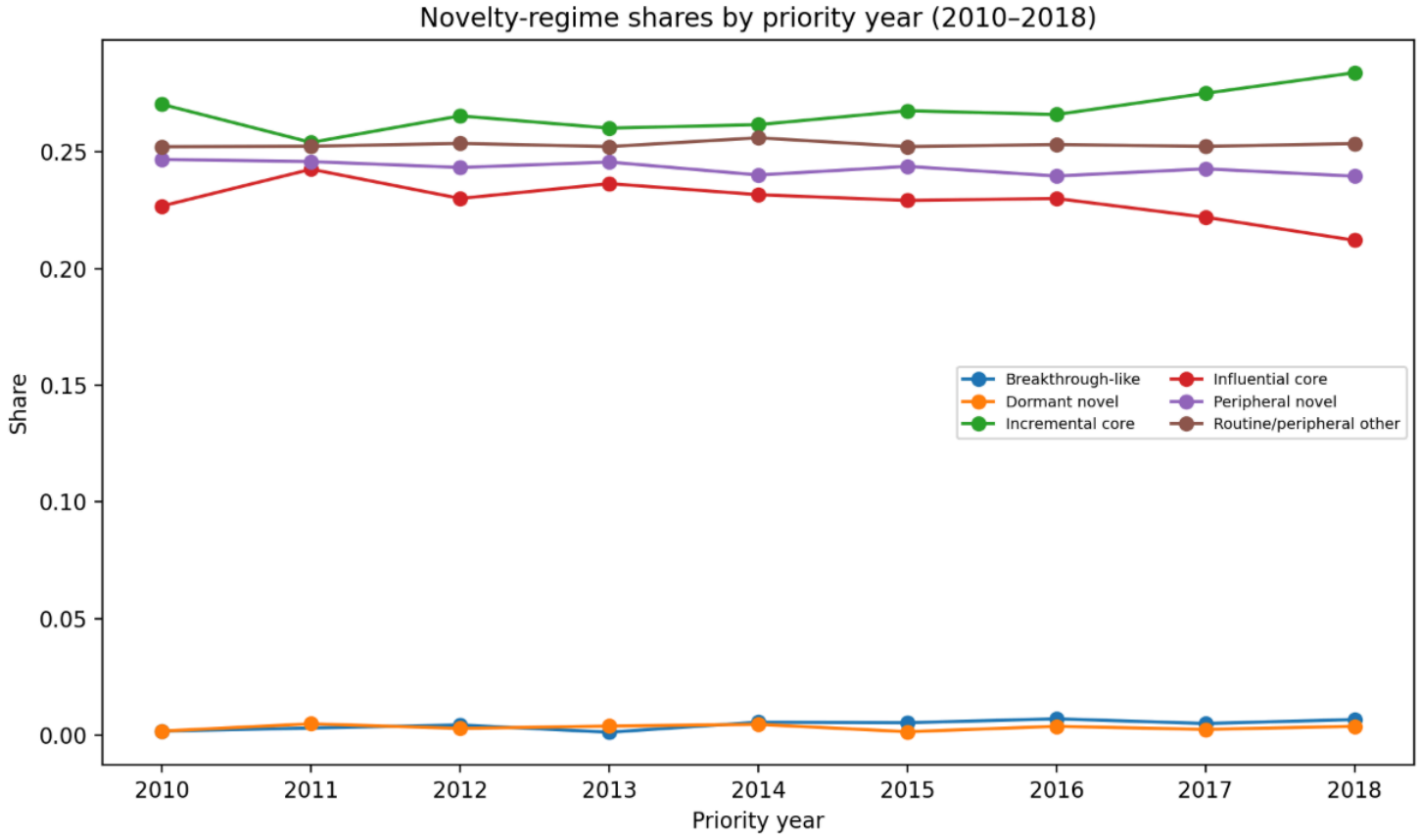


Figure 6

Novelty-regime shares by priority year, 2010–2018 influence-valid cohorts

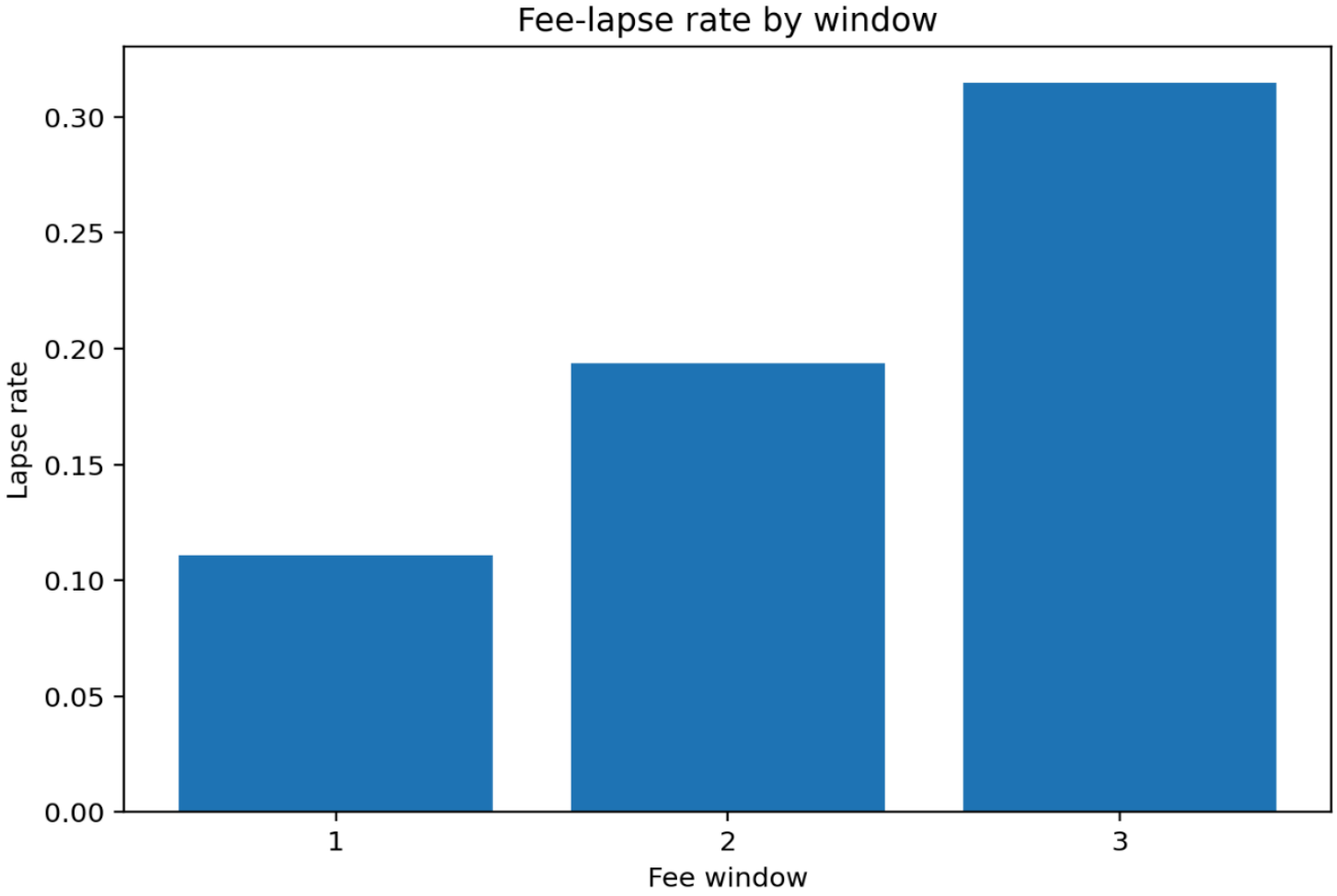


Figure 7

Maintenance-fee lapse rate by fee window

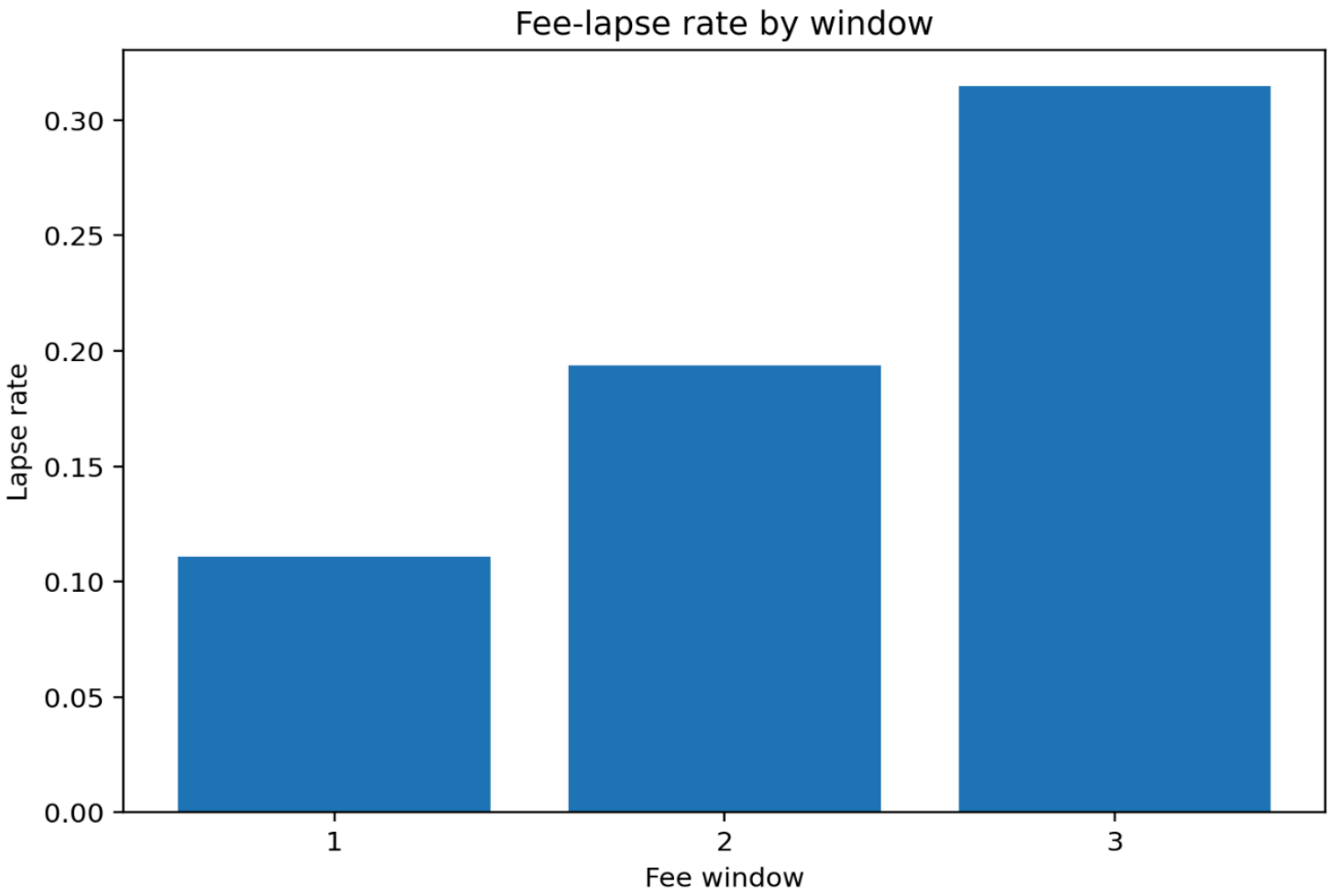


Figure 7

Maintenance-fee lapse rate by fee window

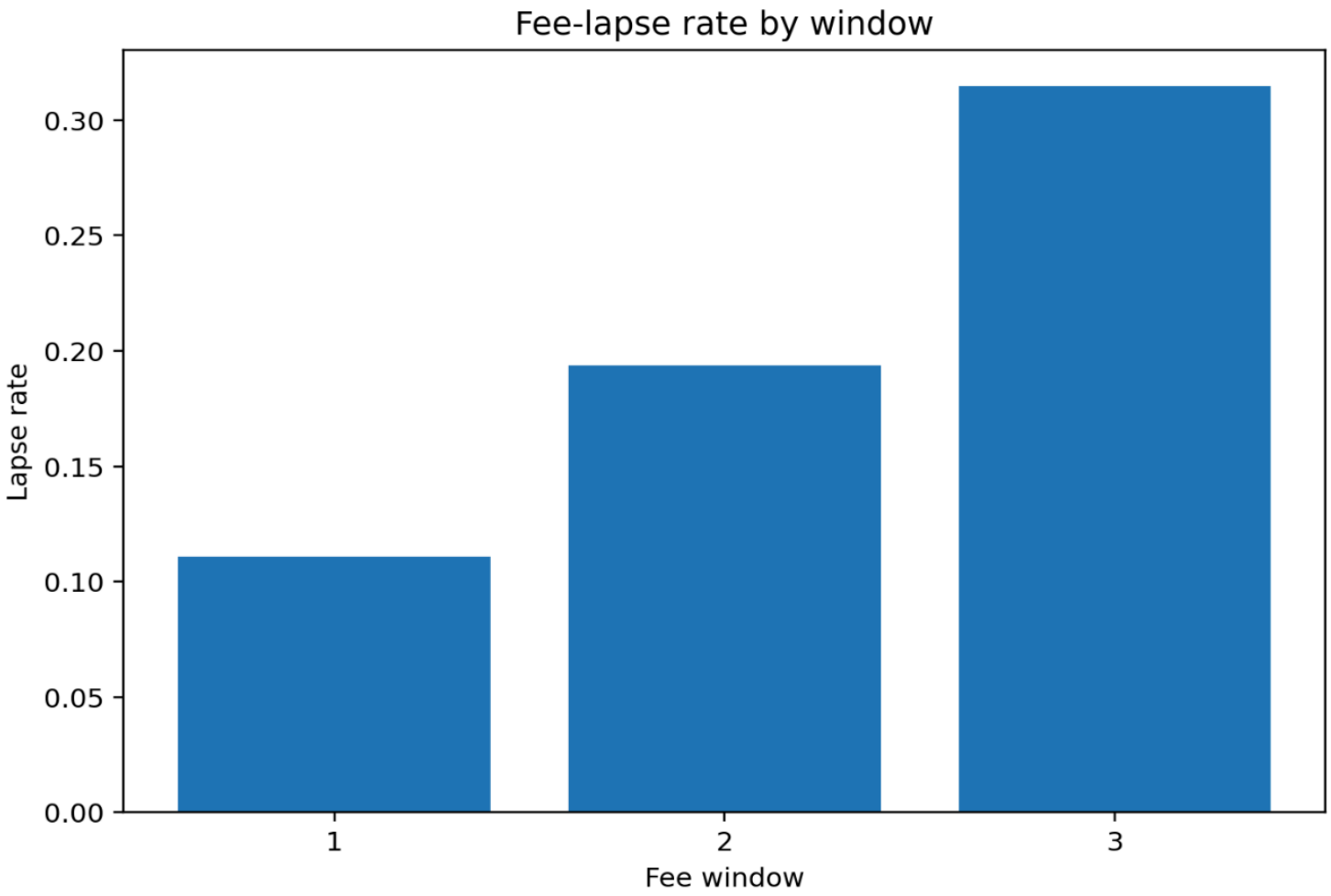


Figure 7

Maintenance-fee lapse rate by fee window

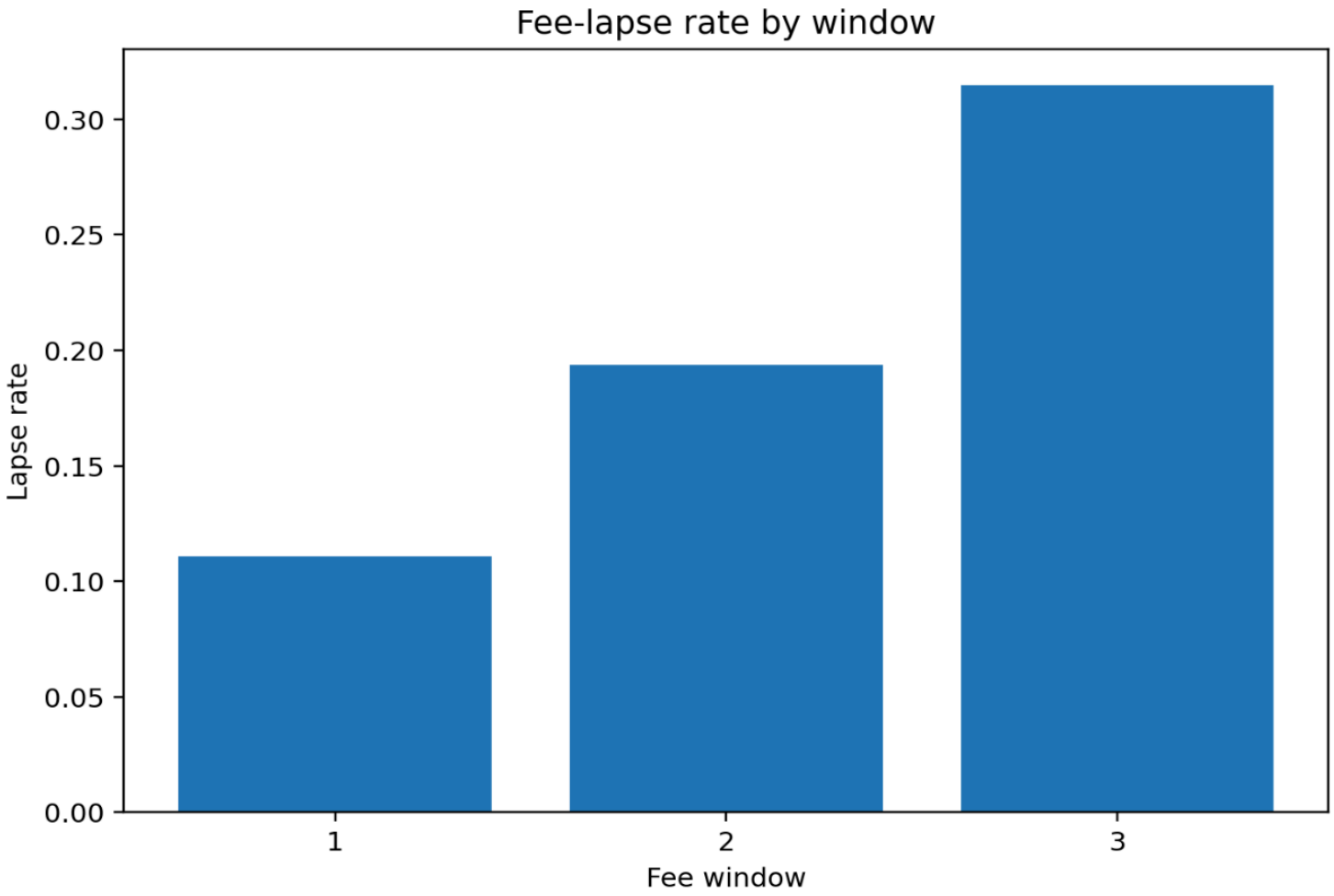


Figure 7

Maintenance-fee lapse rate by fee window

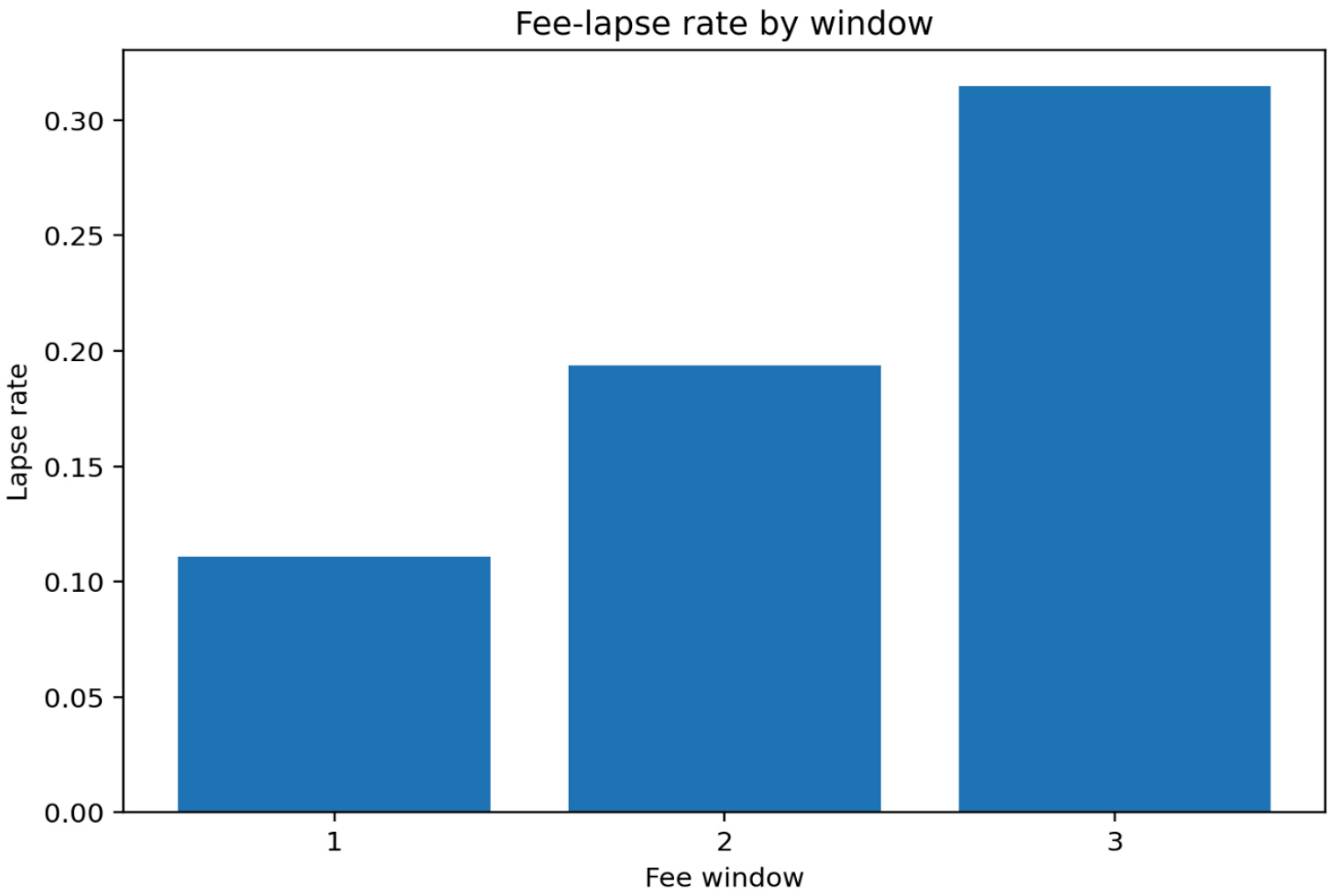


Figure 7

Maintenance-fee lapse rate by fee window

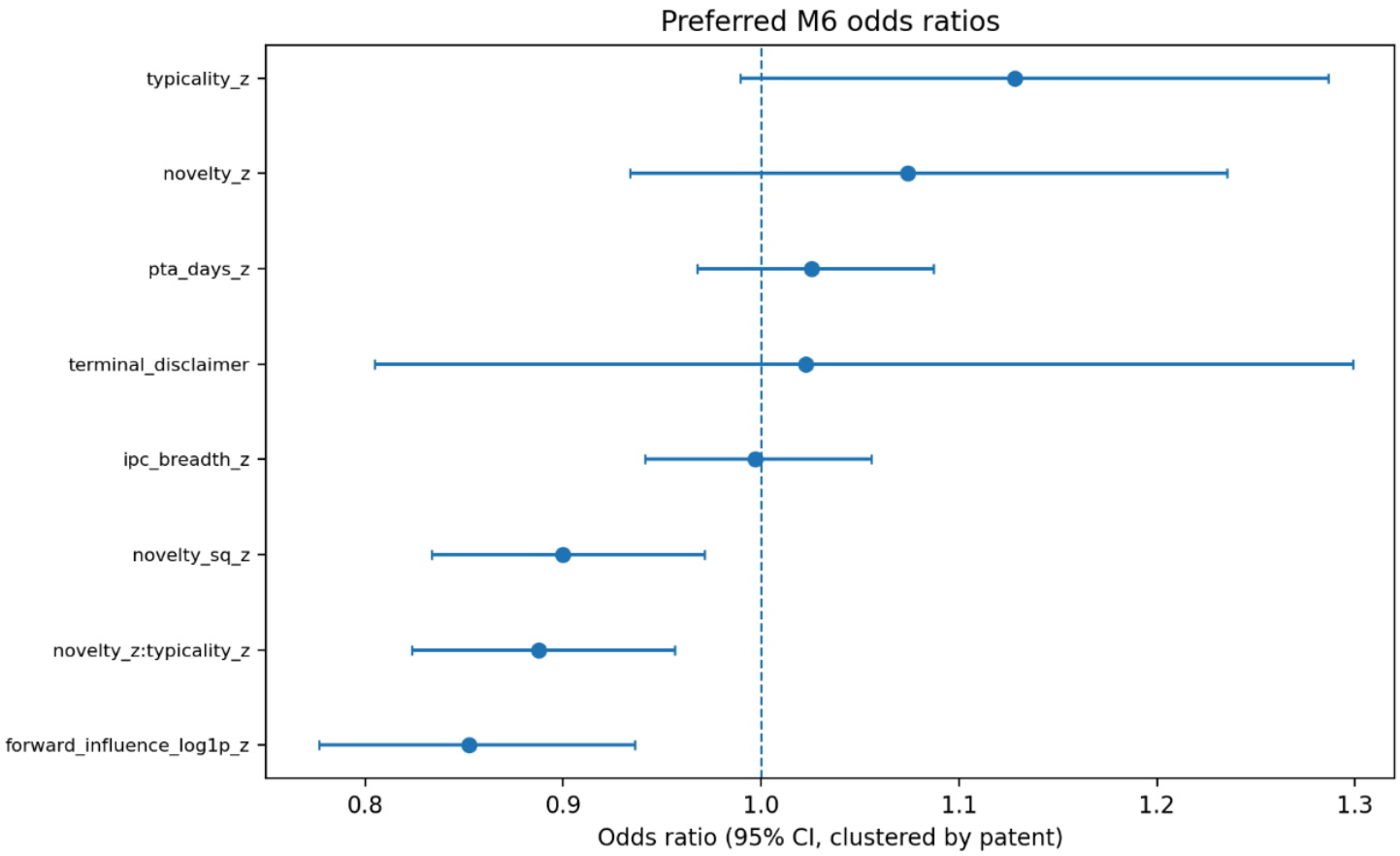


Figure 8

Preferred M6 odds ratios

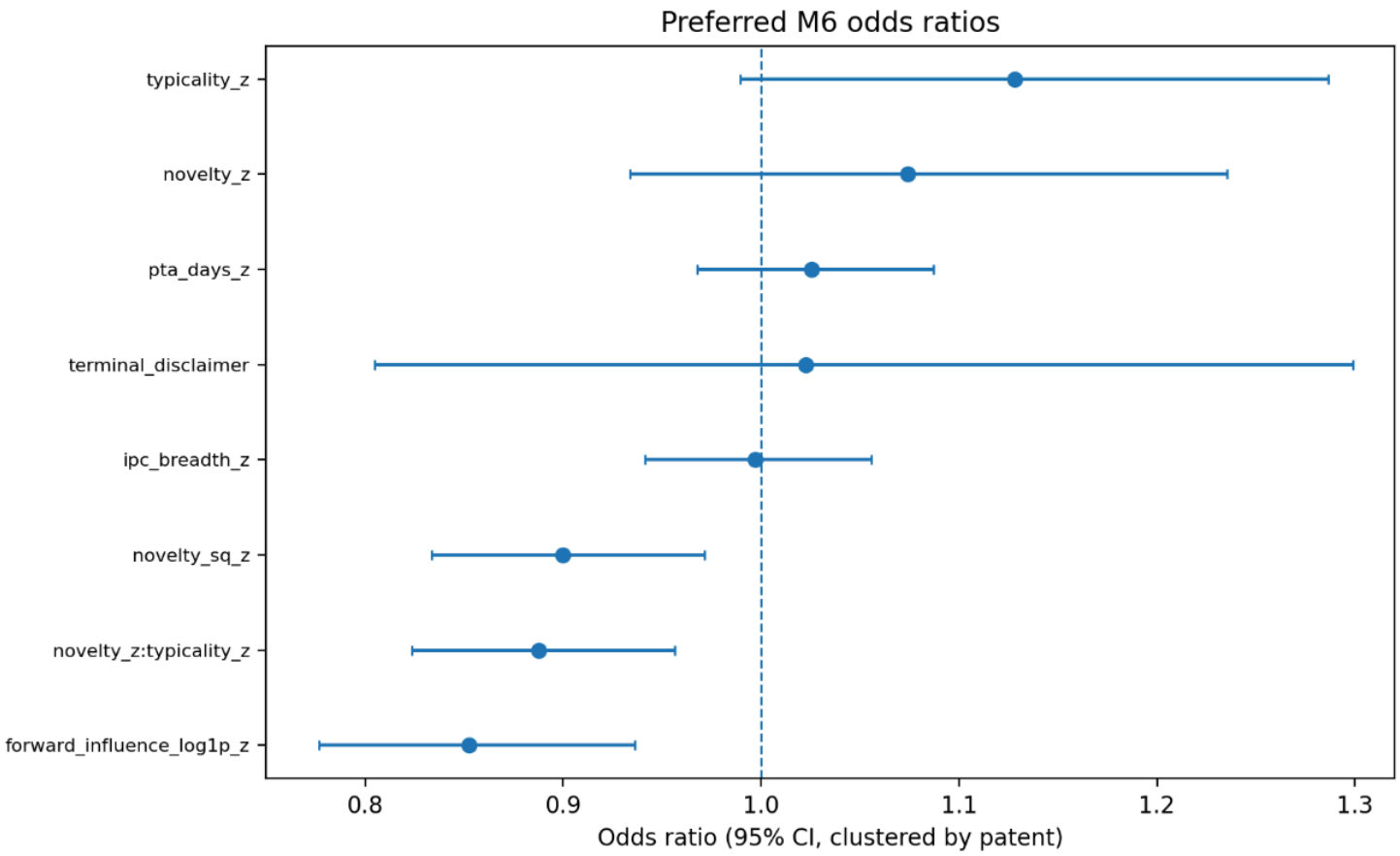


Figure 8

Preferred M6 odds ratios

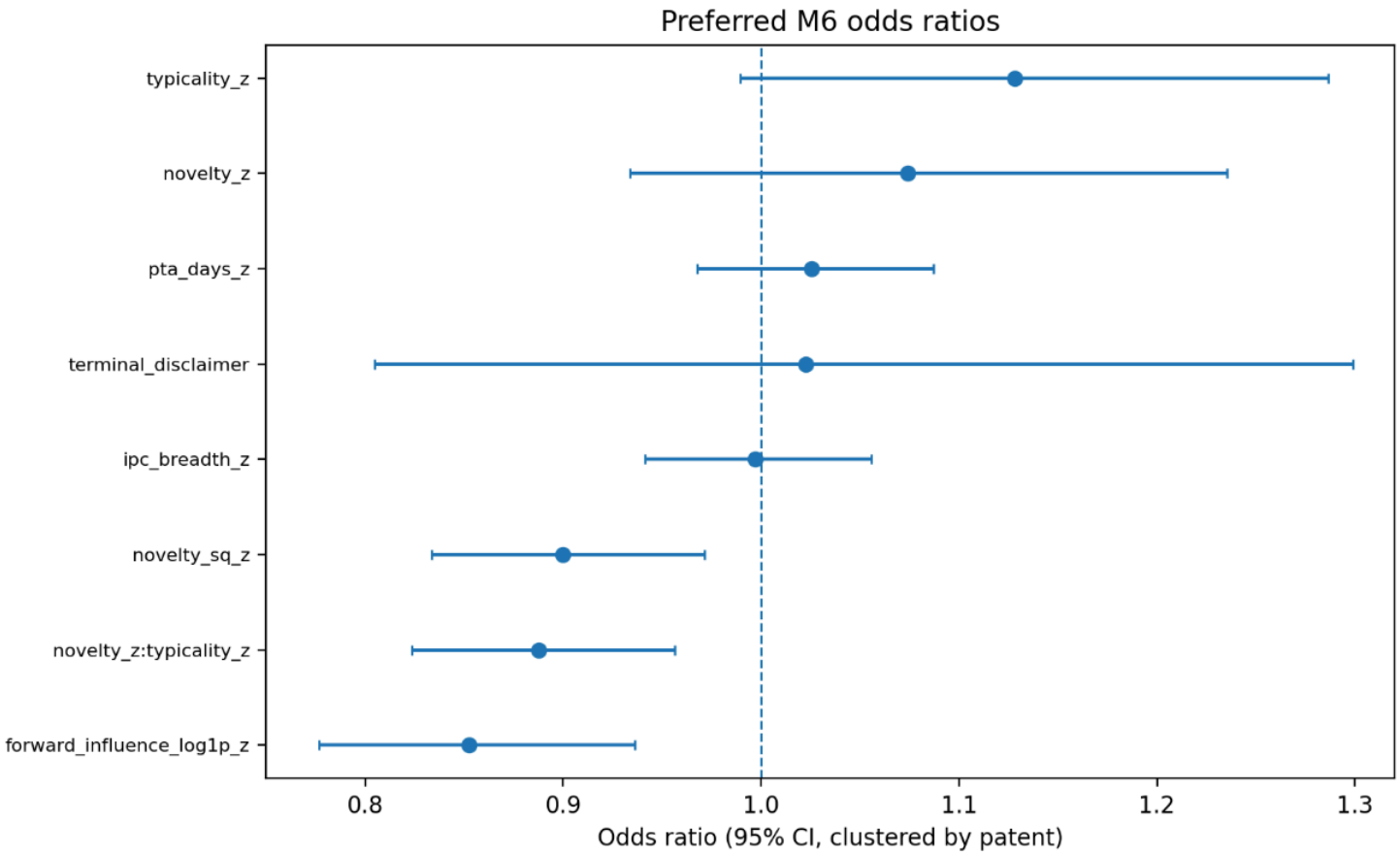


Figure 8

Preferred M6 odds ratios

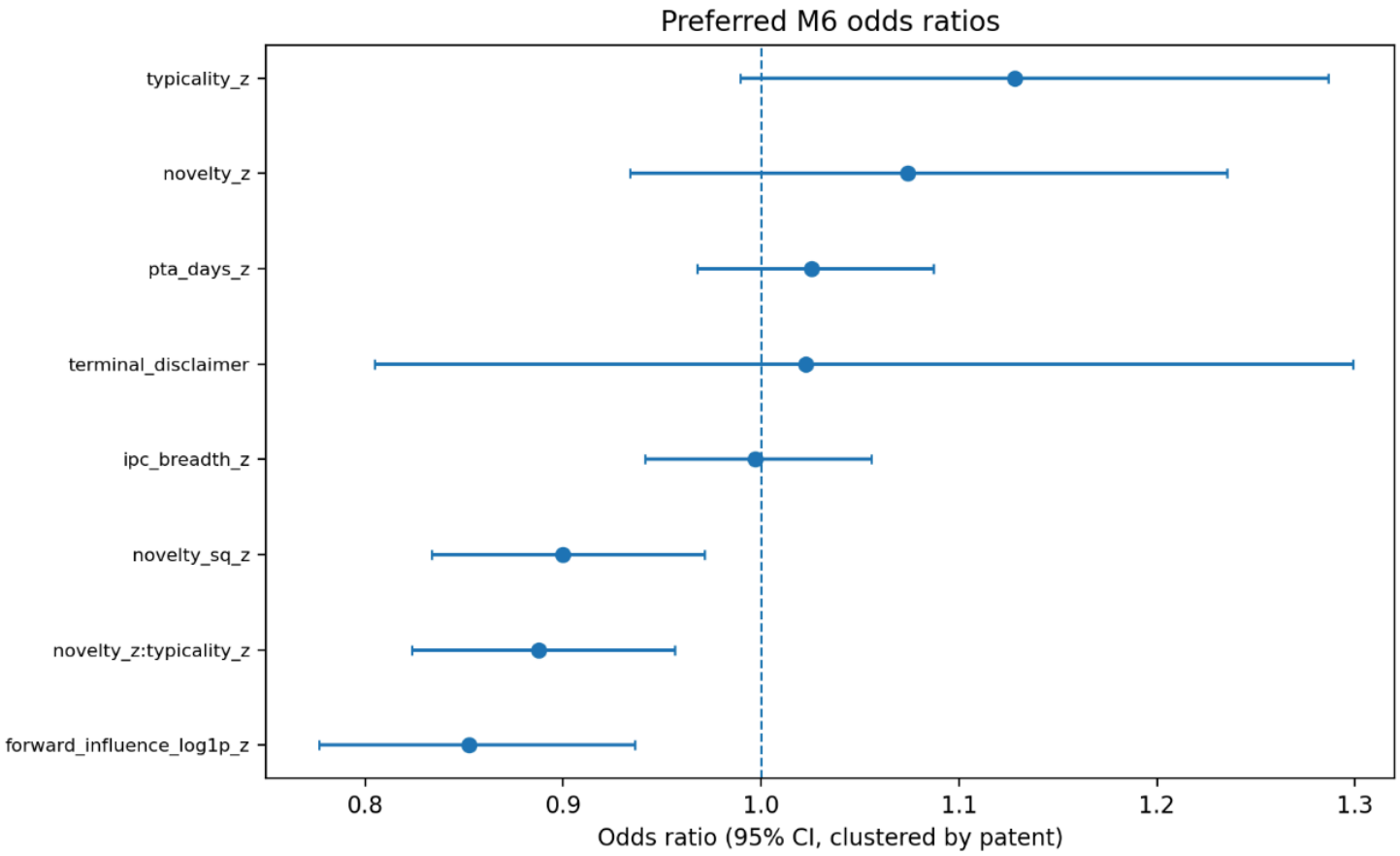


Figure 8

Preferred M6 odds ratios

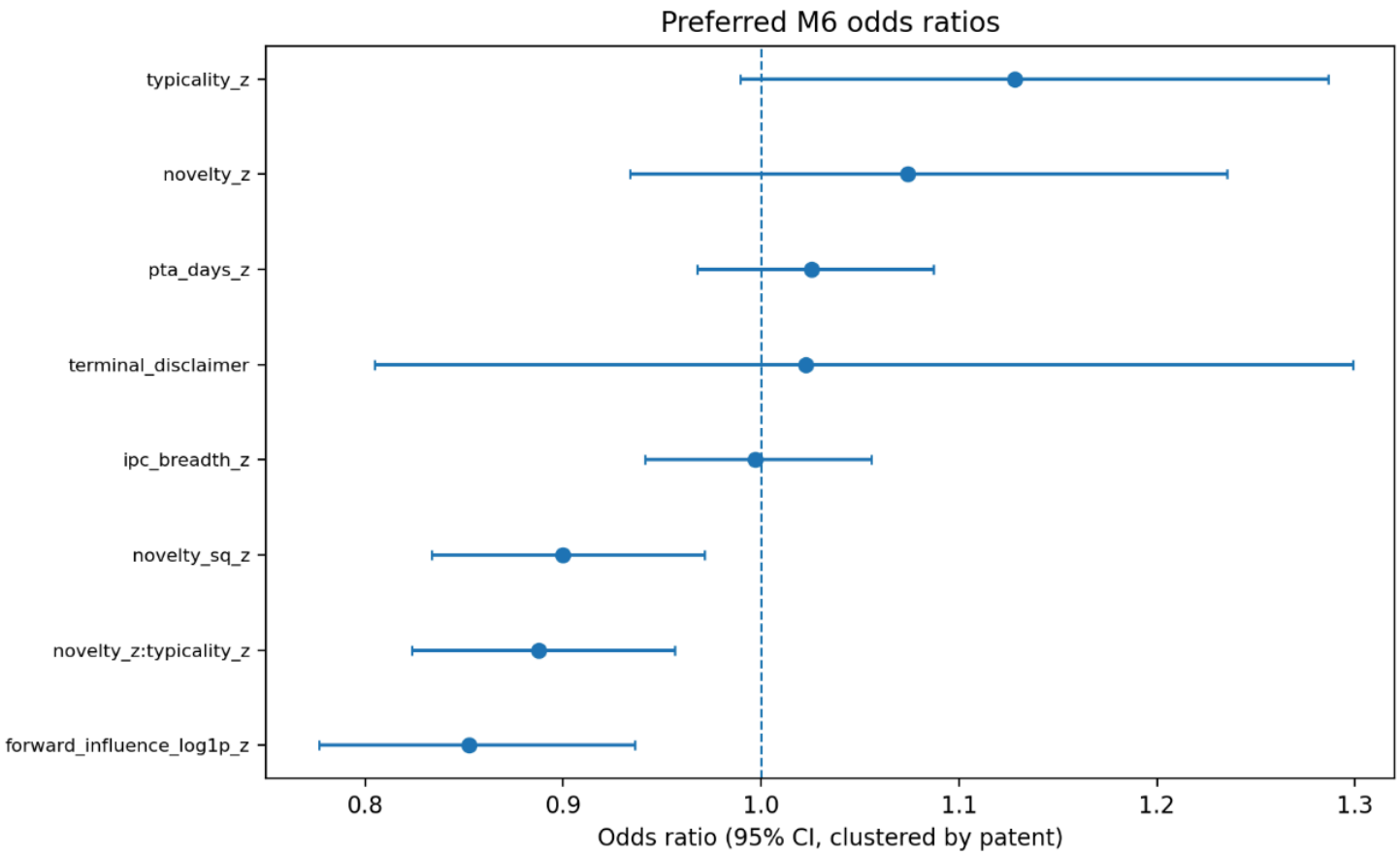


Figure 8

Preferred M6 odds ratios

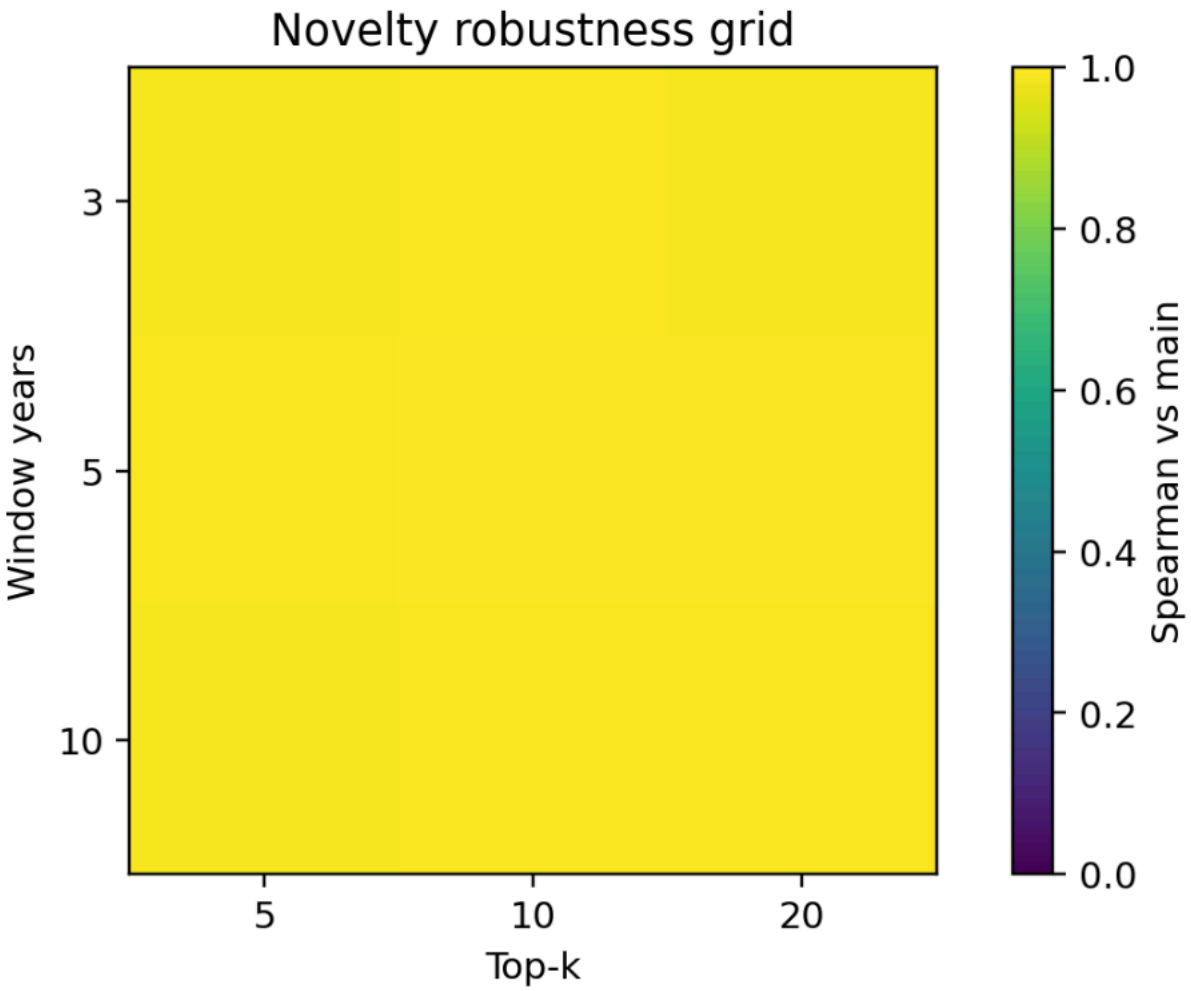


Figure 9

Novelty robustness grid: rank stability against the main specification

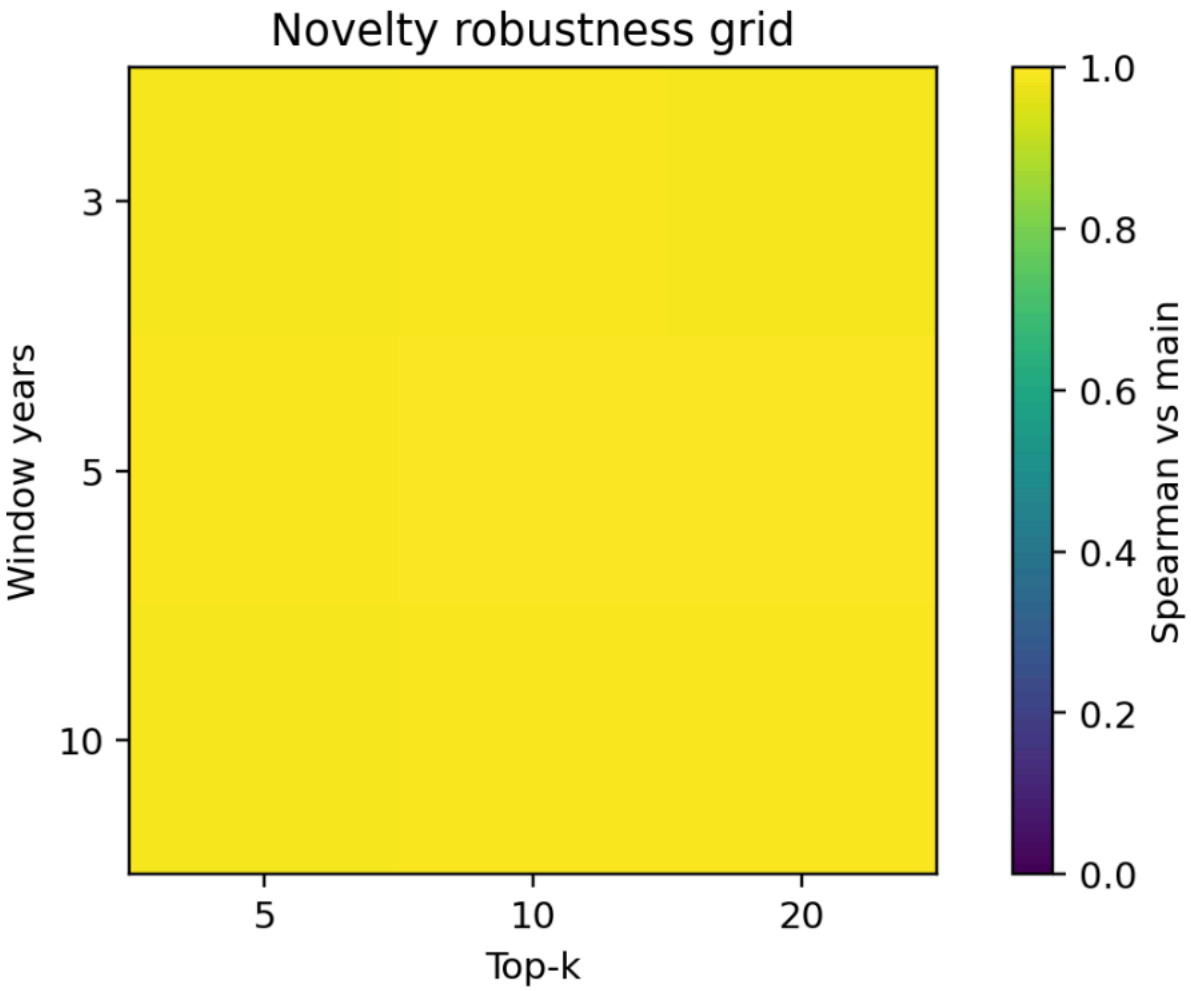


Figure 9

Novelty robustness grid: rank stability against the main specification

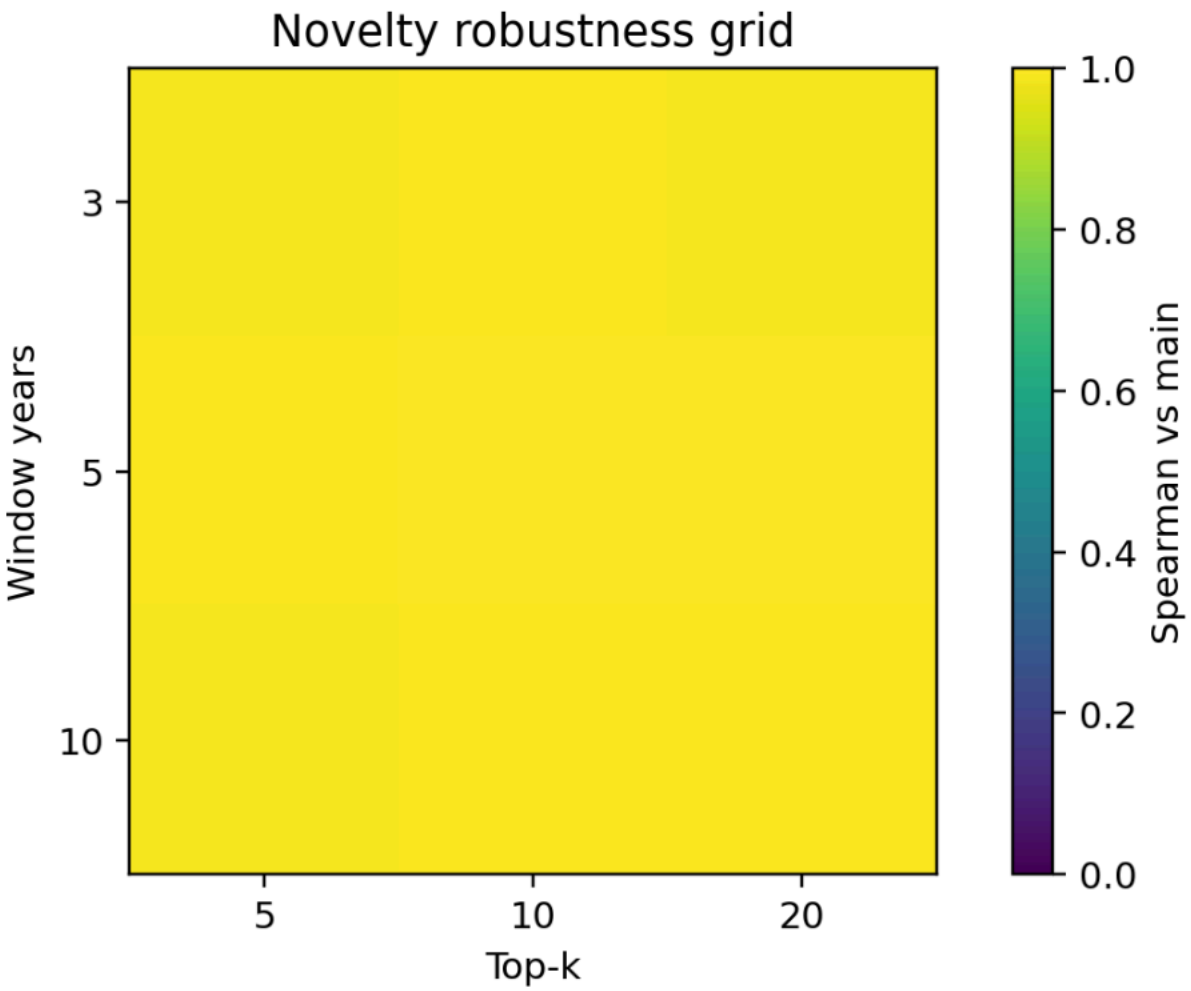


Figure 9

Novelty robustness grid: rank stability against the main specification

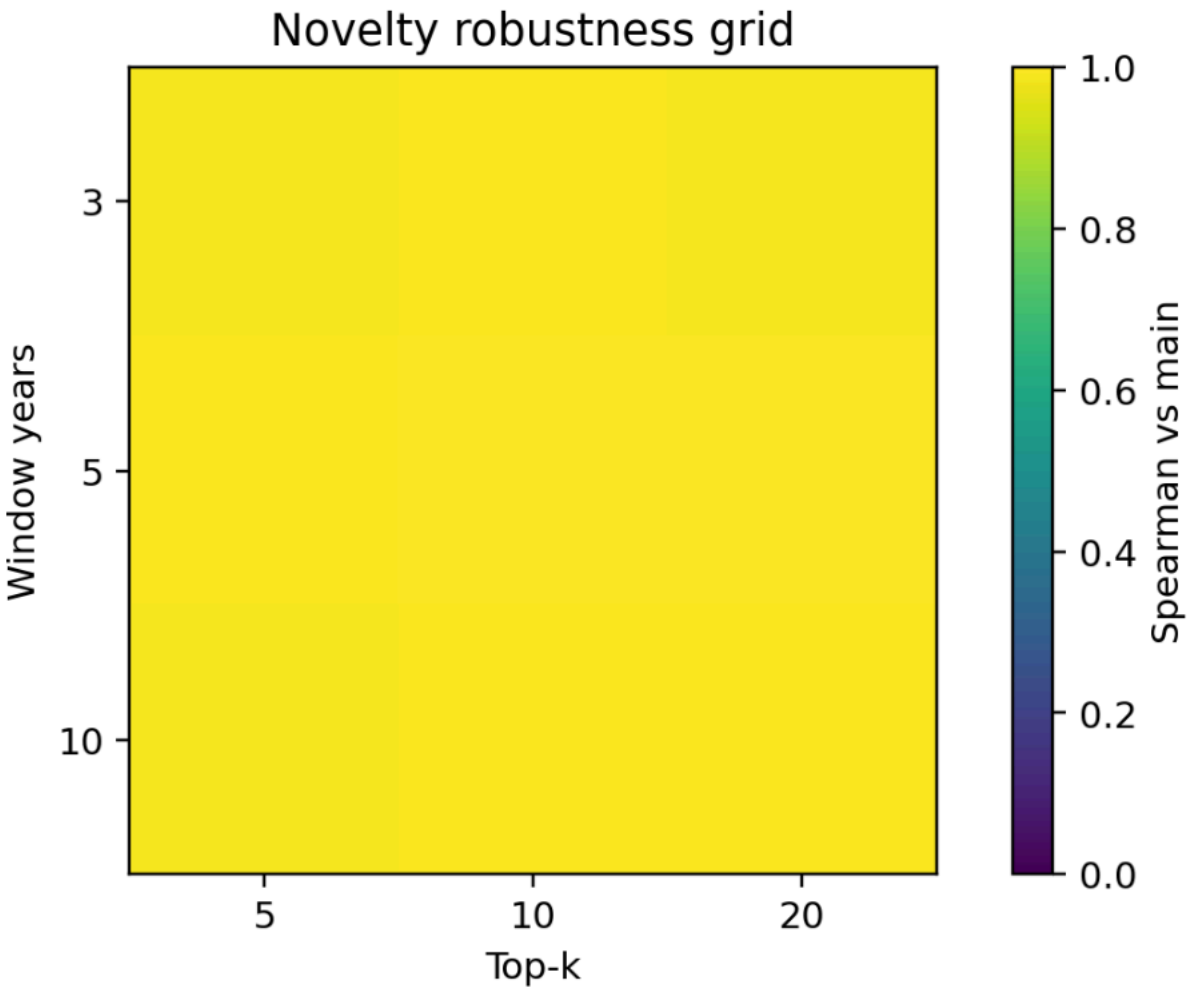


Figure 9

Novelty robustness grid: rank stability against the main specification

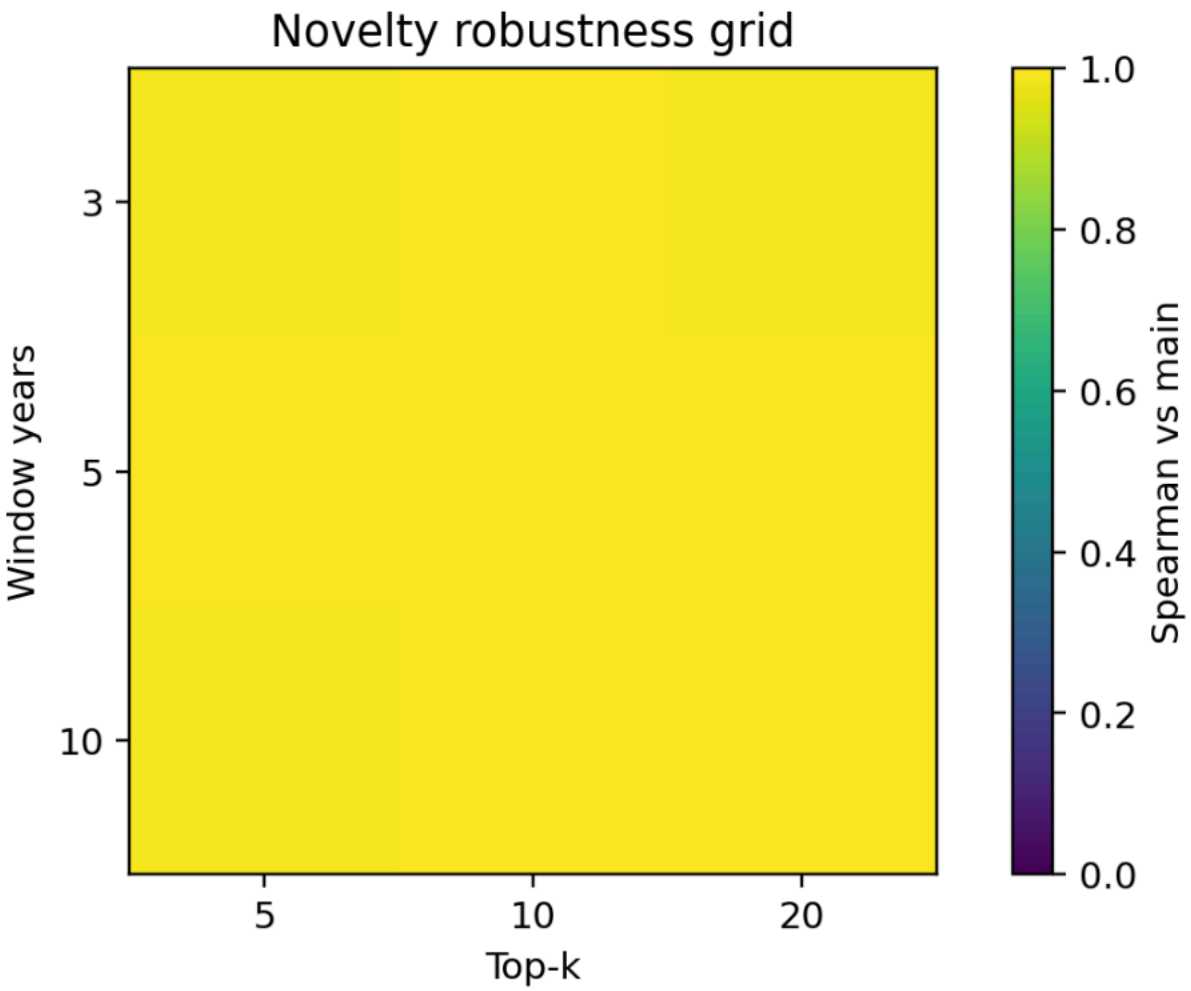


Figure 9

Novelty robustness grid: rank stability against the main specification