Reliance on Scientific Ideas in Patenting

Sam Arts[1]
Department of Management, Strategy and Innovation
Faculty of Economics and Business, KU Leuven
sam.arts@kuleuven.be

Nicola Melluso
Department of Business and Management
LUISS University
nmelluso@luiss.it

Preliminary and incomplete, please do not circulate

ABSTRACT

Scientific ideas drive firm innovation, yet measuring firms' reliance on science remains a challenge. Prior research has often relied on patent-to-paper citations, whose primary purpose is to delineate the legal scope of a patent rather than to document all scientific ideas on which it builds. We develop a novel text-based approach that identifies when patents reuse specific ideas from the scientific literature. We validate this method and show that patents frequently build on prior scientific ideas without citing the original source paper, particularly when the paper is older or less similar in content. Finally, we demonstrate that reliance on scientific ideas positively correlates with the private value of patents, even after controlling for traditional patent-to-paper citations and within the subset of patents without such citations. Together, these results suggest that traditional citation-based measures underestimate the role of science and illustrate the complementary value of our text-based approach and open dataset for studying how scientific knowledge contributes to firm innovation.

[1] Corresponding author: Sam Arts, Department of Management, Strategy and Innovation, Faculty of Economics and Business, KU Leuven, email: sam.arts@kuleuven.be, mailing address: Hendrik Conscienceplein 8 - box 15520, 2000 Antwerp, Belgium, phone: +32 3 201 18 36.

1. Introduction

Scientific ideas fuel technological progress and economic growth in industry (Bush, 1945; Mokyr, 2002). For instance, Stanley Cohen's scientific discovery of epidermal growth factor, a cell-growth protein that earned him the Nobel Prize, was foundational for the development of cancer therapies and blockbuster drugs in the pharmaceutical industry. Measuring and studying how firms draw on scientific knowledge to develop new technologies is therefore a central concern in strategy and innovation research (Cohen & Levinthal, 1990; Fleming & Sorenson, 2004). To date, most empirical work has relied on two primary data sources: firm surveys (e.g., Mansfield, 1991; Cohen et al., 2002) and patent citations to scientific publications (e.g., Narin et al. 1997; Krieger et al. 2024). While surveys provide direct and valuable insights, they are typically constrained by small, cross-sectional samples. More recently, patent citations to scientific papers (PCS) have emerged as the standard large-scale proxy for science–technology linkages (Marx & Fuegi, 2020). Yet, like any data source, PCS have limitations, offering only a partial view of how firms draw on scientific knowledge to drive innovation performance and build competitive advantage (Roach & Cohen, 2013).

These limitations primarily stem from the fact that the main purpose of patent citations is to delineate the legal scope of an invention, not to comprehensively document all sources of prior knowledge—such as scientific ideas—that contributed to the development of the patented technology (Jaffe et al., 1993). As a result, PCS underestimate the role of science in firm innovation (Type II error). Inventor surveys reveal that nearly 30% of patents drawing on scientific ideas cite no scientific publications (Callaert et al., 2014), and firm-level surveys suggests that PCS capture only about half of science's actual contribution to firms' innovation performance (Roach & Cohen, 2013). At the same time, PCS are also prone to Type I error: inventors report that roughly half of the scientific publications they do cite were not important for the development of the patented technology (Callaert et al., 2014).[2] Finally, PCS rest on a linear, one-way model of innovation in which science is treated solely as an upstream input to technology. This view neglects the dynamic, two-way relationship in which science and technology may advance together through iterative feedback loops (Rosenberg, 1976; Stokes, 1997; Barbosu & Teodoridis, 2025). As Paula Stephan (2012: 84) observes, "scientific

[2] This limitation of citations as indicator of knowledge flows extends beyond patent-to-paper citations. In the context of paper-to-paper citations, a survey of 9,000 academics across 15 scientific fields found that more than half of the papers they cited were judged to have had only minor or very minor intellectual influence (Teplitskiy et al., 2022).

research leads to advances in technology, but it is new technology that often brings about advances in science."

In this paper, we develop and validate a new method to measure a patent's reliance on scientific ideas and make the underlying data publicly available. Rather than relying on citations, our approach analyzes the full text of all U.S. patents to identify the scientific ideas embedded in each invention. We build on Arts et al. (2025), who processed the complete historical corpus of scientific publications across all domains to identify novel scientific ideas—captured as newly introduced noun phrases such as *microrna, atomic force microscope,* or *polymerase chain reaction*—that first appeared in the scientific literature.[3] By processing the text of both scientific publications and all U.S. patents, we identify, for each patent, the reuse of scientific ideas originally pioneered in the academic literature, as well as ideas that first appeared in patents before diffusing into scientific discourse.

To validate our approach, we first show that patents with clear ties to science—such as those assigned to universities or invented by scientists—score higher on our metrics of reliance on scientific ideas, even in a sample of patents without patent-to-paper citations. We further demonstrate that patents reusing a given scientific idea are more likely to cite the original paper in which the idea first appeared, relative to matched control patents. To assess whether patent-to-paper citations accurately capture a patent's reliance on scientific ideas, we construct different datasets where there is strong indication that a patent builds on a specific scientific idea introduced in a particular paper. These include cases where inventors had previously cited the paper in patents or publications, or had co-authored the paper themselves. Across all settings, citation rates remain relatively low—ranging from 31% to 68%—suggesting that patents often rely on scientific ideas without citing their original source. In line with the legal function of patent references to delineate the scope of protection, we also find that patents are more likely to cite the specific scientific papers whose ideas they reuse when those papers are more recent or when their content is more closely aligned with the patent.

Finally, to demonstrate the value of our new approach for strategy and innovation research, we revisit the long-standing question of whether firm inventions that build on

[3] Arts et al. (2025) validate this approach in multiple ways. They show, for instance, that the noun-phrases identify the first appearance of major scientific ideas in the literature and that, in a subset of cases, the detected ideas align with Nobel Prize–winning contributions as independently documented in the official Nobel Prize materials.

science have greater private value (Fleming & Sorenson, 2004; Krieger et al., 2024). We find that patents drawing on scientific ideas yield higher private value, a result that holds both after controlling for traditional patent-to-paper citations and when restricting the sample to patents without such citations. This finding underscores how our text-based approach and open dataset provide a powerful complement to patent-to-paper citations in understanding how science fuels firm innovation.

2. Identifying patent's reliance on scientific ideas

2.1 Data and sample

Our analysis combines data from multiple sources. For patent data, we use PatentsView, from which we extract the full text—including the title, abstract, claims, summary, detailed description, and drawing descriptions—of all U.S. patents granted between January 1976 and June 2023 (n=7,699,586). For scientific publications, we use the January 2024 snapshot of OpenAlex, as processed and cleaned by Arts et al. (2025), which includes 75,295,921 papers and conference proceedings published between 1901 and 2023.[4] For brevity, we refer to both as papers throughout the manuscript. We supplement these sources with additional data: assignee data from DISCERN 2.0 (Arora et al. 2021) and PATSTAT, patent-to-paper citations from Marx & Fuegi (2020), inventor–author linkages from Scharfmann et al. (2024), and patent commercial value estimates from Kogan et al. (2017).

To identify scientific ideas, we begin with the processed text of scientific publications from Arts et al. (2025), which yields 27,079,343 distinct noun phrases.[5] Although they use a baseline dictionary and apply preprocessing to remove boilerplate language and generic terms unlikely to represent meaningful scientific ideas, their method remains inherently imperfect. To further refine the identification of genuine scientific ideas, we use a LLM to assign each noun phrase a confidence score from 1 to 3, reflecting the likelihood that it captures a true scientific idea. In the remainder of the paper, we exclude phrases with a score of 1 (8.9% of the sample). These phrases are rated as unlikely to represent a scientific idea, and validation

[4] Arts et al. (2025) retain papers with an English title and abstract, complete publisher information, a valid impact factor for the publication year, non-empty titles, at least one listed author, and no duplicate titles or abstracts.

[5] A noun phrase consists of a single noun or a sequence of words centered on a noun (the head), along with modifiers that refine its meaning. It is widely used in computational text analysis to capture distinct concepts in natural language, making it a practical unit for identifying scientific ideas or concepts in the text of scientific papers. Arts et al. (2025) retain noun phrases that appear at least twice in papers published between 1901 and 2023.

exercises confirm that excluding them reduces false positives without reducing coverage of true ideas (see Appendix B for details).[6]

We then process the full text of all patents using the same preprocessing and extraction procedures applied to the scientific corpus. Of the noun phrases identified in scientific publications, 29.32% also appear in patents, while 70.68% appear only in scientific papers. Among noun phrases that appear in both papers and patents, 57.13% (n = 4,535,841) first appear in a scientific paper, and 42.87% (n = 3,403,386) first appear in a patent, as determined by the publication date of the paper and the filing date of the patent. As such, a notable share of ideas are first disclosed in patents and only subsequently appear in the scientific literature. Among all phrases first introduced in patents, 23.8% (n = 621,807) subsequently appear in a scientific paper for the first time after the patent's filing date but before it is granted. Only 1.7% first appear in a paper co-authored by at least one inventor of the pioneering patent, suggesting that only a very small minority are likely to represent patent–paper pairs.

To give one example, *interleukin-7* was first introduced in a patent (US 4,965,195) filed by Immunex (which was later acquired by Amgen for 16 billion USD) in 1988 describing how to clone and produce this immune-system protein, which proved essential for understanding T-cell and lymphocyte development. The patent gave Immunex strong control over interleukin-7-based therapeutics. The discovery has since influenced both scientific immunology research and technological development in clinical therapies for immune reconstitution, as reflected in the reuse of *interleukin-7* in 997 subsequent scientific papers and 1,536 patents. In line with prior work examining how patents and firms rely on scientific ideas, the remainder of the paper focuses on ideas first introduced in the scientific literature and later reused in technology development (Iara et al., 2018). However, unlike patent-to-paper citations—which are inherently unidirectional—our data and code also enable analysis of the reverse pathway, where new technologies drive scientific progress (e.g., Rosenberg, 1976; Barbosu & Teodoridis, 2025).

Table A.1 in Appendix A lists the top ten noun phrases that first appeared in scientific literature and are most frequently reused in patents, grouped by the decade in which they were first adopted in patents. Early decades are dominated by materials and chemical sciences (e.g.,

[6] We provide open access to the full set of noun phrases along with their associated confidence scores.

polypropylene, *polyethylene_glycol*), later decades by computing and electronics sciences (e.g., *liquid_crystal_display*, *field-programmable_gate_array*), and the most recent decades by life sciences (e.g., *imatinib*, *cas9*) and advanced information sciences and artificial intelligence (e.g., *generative_adversarial_network*, *smart_contract*). This progression reflects the historical shift from chemistry and materials toward computing and electronics, life sciences, and advanced information sciences including artificial intelligence.

2.2 Descriptive statistics

We measure a focal patent's reliance on scientific ideas or concepts by identifying noun phrases that were first introduced in scientific papers and later reused in the focal patent's full text. Based on this, we construct three measures: *Science Phrases* as the total number of unique scientific noun phrases in the patent, *Science Phrases Share* as their proportion relative to all unique noun phrases in the patent, and *Science Phrases Bin* which is a binary indicator equal to one if the patent reuses at least one noun phrase first introduced in the scientific literature.[7] For each focal patent, we also calculate *Science References* as the number of distinct scientific papers cited by the focal patent on the front page or in the body text of the patent (Marx and Fuegi, 2020; Bryan et al., 2020), and *Science References Bin* which is a binary indicator equal to one if the patent cites at least one scientific paper.[8]

'Insert Table 1'

Table 1 reports summary statistics at the patent level. The full text of the average (median) patent contains 255 (179) distinct noun phrases, of which 32 (9) were first introduced in the scientific literature before being reused in the focal patent. Accordingly, *Science Phrases Share* for the average (median) patent is 0.081 (0.053), and 96% of all patents reuse at least one scientific noun phrase. By comparison, patents cite only 3 scientific

[7] One might argue that longer patents naturally contain more *Science Phrases*, potentially biasing the measure. However, our analyses control for patent text length, and results remain robust when using *Science Phrases Share*, which corrects for length by construction. A related concern is that multiple distinct noun phrases in a patent may capture the same underlying scientific concept. Here too, *Science Phrases Share* mitigates this issue by construction.

[8] To measure the number of science references made by a patent, we rely on the same population of scientific papers from OpenAlex that we use to construct our measure of patents' reliance on scientific ideas. Because OpenAlex is very broad and includes working papers and obscure journals, we follow Arts et al. (2025) and restrict the sample to journal articles and conference proceedings that (i) have an English title and abstract, (ii) include complete publisher information, (iii) are linked to a valid impact factor in the publication year, (iv) contain non-empty titles, (v) list at least one author, and (vi) are not duplicate records. Importantly, the correlation between our restricted measure of science references and an unrestricted count using the full OpenAlex sample (as in Marx and Fuegi, 2020) is 0.998, and all results reported in this paper hold under both approaches.

papers on average (median = 0), and only 23% of patents cite at least one scientific paper. The correlation between *Science Phrases* and *Science References* is 0.58, suggesting that while the two measures are positively related, they also capture distinct dimensions of a patent's reliance on science. As shown in Figure A.1 in Appendix, even patents without any science references, representing 78% of the full population, exhibit substantial variation in *Science Phrases*, suggesting that *Science References* may underestimate a patents' reliance on scientific ideas. In fact, 95% of patents without any scientific references reuse at least one idea (noun phrase) from the scientific literature, and these patents reuse an average of 14 scientific ideas (median = 10).

Notably, the average (median) age of reused scientific ideas—measured as the number of years between the publication of the paper in which the phrase first appeared and the patent's filing year—is 46 (46) years. This is substantially older than the average (median) age of cited scientific papers, measured as the number of years between the publication of the cited paper and the patent's filing year, which is only 10 (9) years. Thus, patents tend to cite more recent scientific papers, even though the underlying inventions often also rely on older, foundational scientific ideas without citing the original source publications.

Table A.2 lists, for selected industries, the top five U.S. public firms whose technology development most heavily relies on science, as measured by the average number of distinct scientific phrases reused per patent, along with each firm's five most frequently reused scientific phrases. In the Chemicals and Allied Products industry (including Biotech), Genentech's most reused science phrases—*polypeptide*, *amino_acid_sequence*, *plasmid*—trace directly to molecular biology research underlying its protein-based therapies, while Amgen's *monoclonal_antibody* and *polyethylene_glycol* reflect reliance on immunology research and polymer chemistry for modifying and stabilizing biologic drugs. Ionis Pharmaceuticals' top reused science phrases—*oligonucleotide*, *phosphorothioate*, *nucleoside*—originate in nucleic acid chemistry and underpin its RNA-targeted therapeutics. In the Measuring, Analyzing, and Controlling Instruments industry, Hologic's heavy reliance on *nucleic_acid_amplification* and *oligonucleotide* reflects its use of foundational genetic testing research to develop molecular diagnostics, while Thermo Fisher and Waters build on analytical chemistry (e.g., *mass_spectrometry* and *liquid_chromatography*) to develop instruments for chemical and biochemical analysis, and Bio-Rad's heavy reuse of *peptide* and *oligonucleotide* reflects its reliance on protein chemistry and nucleic acid research to develop

reagents, assay kits, and instruments for life sciences laboratories. In the Petroleum Refining industry, ExxonMobil's *copolymer*, *polyolefin*, and *molecular_sieve* derive from polymer science and catalysis research, which underpin its development of high-performance plastics, chemical products, and catalytic processes for fuel refining.

3. Validation text-based metrics

We validate our new method using two complementary approaches. First, we assess face validity by testing whether patents with clear links to science—such as those that cite scientific publications, are assigned to universities, or list scientist–inventors—exhibit greater reliance on scientific ideas according to our metrics. Second, we examine whether patents that reuse a given scientific idea are more likely to cite the original paper in which the idea first appeared, compared to matched control patents.

3.1 Face validity

As a first validation exercise, we compare groups of patents that should, on average, rely more heavily on scientific knowledge—namely, those that cite at least one scientific publication, are assigned to a university, or list a scientist among their inventors—to the rest of the patent population. Table 2 confirms this pattern. Patents in these groups are more likely to contain at least one scientific idea, reuse a larger number of scientific ideas, and exhibit a higher share of scientific content. For example, patents with science references reuse on average 91 *Science Phrases* and have a 15% *Science Phrases Share*, compared to 14 and 6% for patents without science references. University patents show a similar contrast (103 versus 29 *Science Phrases* and 16% versus 8% *Science Phrases Share* on average) as do patents with scientist–inventors (49 versus 15 *Science Phrases* and 10% versus 6% *Science Phrases Share* on average). While *Science Phrases* and *Science References* perform similarly in distinguishing university from non-university patents, as indicated by comparable Mann–Whitney test statistics in Table 2, *Science Phrases* substantially outperform *Science References* in distinguishing patents with and without scientist–inventors. Importantly, as illustrated in Table A.3, even within the subset of patents without any science references, university patents and those with scientist–inventors still exhibit significantly higher *Science Phrases* and *Science Phrases Share*. This indicates that our text-based measure uncovers meaningful residual variation in patents' reliance on science that patent-to-paper citations fail to detect.

'Insert Table 2'

3.2 Patents' reliance on scientific ideas and citations to source papers

As a second validation, we randomly sample 100,000 scientific ideas (noun phrases) first introduced in scientific papers and subsequently reused in at least one patent. We then identify all patents reusing these ideas, yielding 7,139,906 patent–reused scientific idea pairs. For each pair, we select a randomly matched control patent that does not reuse the specific scientific idea but is filed and granted in the same years and assigned to the same technology subclass as the patent reusing the scientific idea. We find that patents that reuse a given scientific idea are 67 times more likely to cite the scientific paper in which the idea was first introduced than matched control patents (0.340% vs. 0.005%).[9] This pattern confirms that reusing a given scientific idea from the literature is correlated with citing the scientific paper in which the idea first appeared, validating that our text-based method captures a patent's reliance on specific scientific ideas. Nevertheless, only a small fraction of patents reusing a scientific idea cite the corresponding source paper. This result is consistent with Arts et al. (2025), who show that only a minority of papers relying on scientific ideas from other papers include a citation to those source papers. Interestingly, Figures A.2 and A.3 reveal substantial heterogeneity in citation rates across technologies and industries. Citation rates are highest in chemistry, human necessities, and physics, and lowest in fields such as mechanical engineering. At the industry level, industries such as drugs, laboratory instruments, and surgical and medical devices stand out with the highest citation rates, while industries such as motor vehicles and equipment exhibit much lower citation rates.

4. Validation Patent-to-Paper Citations

4.1 Do patent-to-paper citations accurately capture reliance on scientific ideas?

To assess how well patent-to-paper citations capture a patent's actual reliance on scientific ideas, we construct three datasets at the paper–patent level in which there is strong indication that the focal patent draws on a specific scientific idea introduced in the corresponding paper. We then examine how often the patent cites that source paper.

[9] Besides directly citing the source paper, a patent may also cite it indirectly by referencing a paper or patent that cites the pioneering paper (e.g., Ahmadpoor and Jones, 2017). We find that patents reusing a given scientific idea are 20 times more likely to cite a patent that cites the source paper (0.301% vs. 0.015%), and 17 times more likely to cite a paper that cites the source paper (1.252% vs. 0.075%).

First, following Lampe (2021), we identify from the full population of U.S. patents all cases in which a patent *p*, invented by inventors *i*, builds on a scientific idea (i.e., reuses a noun phrase) that was first introduced in a paper previously cited by any earlier patent co-invented by at least one of the inventors *i*. This results in 1,913,536 patent–paper observations where the focal patent reuses a scientific idea from the focal paper that at least one of the inventors had cited in a prior patent and thus is arguably aware of. Interestingly, only 68% of these cases include an explicit citation to the original paper.

Second, we conduct a parallel analysis based on citations in inventors' prior scientific publications, rather than in their prior patents. Unlike patent-to-paper citations, which may be added by patent attorneys or examiners, citations in scientific papers are typically included by the authors themselves and thus more directly reflect their knowledge of the cited work. We identify from the full population of U.S. patents all cases in which a patent *p*, invented by inventors *i*, reuses a scientific idea that was first introduced in a paper previously cited by any earlier scientific paper co-authored by one of the inventors *i*. This yields 456,751 patent–paper observations in which the focal patent builds on a scientific idea introduced in the focal paper previously cited by one of its inventors in their earlier scientific work. Yet, only 31% of these observations include a citation to the original paper.

Finally, we identify all cases in which a patent *p*, invented by inventors *i*, builds on a scientific idea that was first introduced in a paper co-authored by one of the inventors *i*. In these cases, the inventors should clearly be aware of the paper, having written it themselves. This results in 655,825 patent–paper observations. Even in this setting, where the paper is (co-)authored by the same inventor(s) and the idea is reused in a subsequent patent, the citation rate is only 35%, suggesting that in 65% of the cases the patent builds on the scientific idea without citing the source paper.[10] To give one example, *adiponectin*—a protein hormone with antidiabetic and anti-inflammatory properties—was first discovered and published by Masashi Maeda, Iichiro Shimomura, and Yuji Matsuzawa. This scientific

[10] Remember that we are restricting the analysis to new ideas (noun phrases) that first appear in the scientific literature before being reused in patents. One might argue that inventors strategically avoid citing their own scientific work to reduce the risk of patent rejection. However, the grace period under U.S. patent law allows inventors to publicly disclose their invention—such as through a scientific publication or conference presentation—up to 12 months before filing a patent application without jeopardizing patentability. As a second strategy to avoid the risk that their own scientific publication becomes prior art that could block patentability, inventors may choose to file a patent application shortly before publishing the related scientific work. In such cases, the new idea (noun phrase) would typically appear first in a patent before entering the scientific literature. However, among all phrases first introduced in patents, only 1.7% subsequently first appear in a paper co-authored by at least one inventor of the pioneering patent, suggesting that only a very small minority are likely to represent patent–paper pairs.

breakthrough transformed our understanding of fat tissue as an active endocrine organ and catalyzed a wave of scientific research on obesity, insulin resistance, and cardiovascular disease. It also spurred technological advances in diagnostics and therapeutics targeting metabolic disorders. The underlying scientific idea—captured by the noun phrase *adiponectin*—was subsequently reused in 23,688 scientific papers and 2,481 patents. Despite being co-authors of the original scientific discovery, the scientists did not cite their own paper in their first patent on adiponectin, filed a few years after the paper publication.

Taken together, these results show that even when there is strong indication that a patent draws directly on a specific scientific idea—and the inventor is arguably aware of the original paper—the patent often does not cite it. In line with prior survey evidence from inventors and R&D lab managers (Roach & Cohen, 2013; Callaert et al., 2014), our findings suggest that patent-to-paper citations underestimate the extent to which patents rely on science. This pattern reinforces the view that patents do not necessarily cite all the scientific work on which they build, perhaps especially when the cited work is not legally required to delineate the scope of protection.

4.2 When do patents cite the scientific papers they build on?

Building on our previous finding that patents often do not cite the scientific papers they rely on, we now examine the conditions under which patents are more likely to cite these source papers. Given that the primary objective of patent references is to delineate the legal scope of protection—rather than to acknowledge all scientific ideas upon which a patent builds—patents arguably prioritize citations to prior patents over scientific publications (Jaffe et al., 1993; Roach & Cohen, 2013). When patents do cite scientific papers, they are perhaps more common when the paper's content is closely aligned with the patent, making it more relevant for defining legal boundaries, or when the paper is more recent, since newer work better reflects the current state of the art while older scientific ideas may already be incorporated or cited by prior patents.

To test this idea, we use the same random sample of 100,000 scientific ideas (noun phrases) first introduced in scientific papers and later reused in at least one patent. For each idea, we identify all reusing patents, resulting in 7,139,906 patent–reused scientific idea pairs covering 2,468,508 unique patents. Using all these patent–reused scientific idea pairs, we then estimate the likelihood that the patent cites the original paper as a function of the time lag between the paper's publication and the patent's filing date. As shown in Figure 1, the

predicted probability of citation declines steeply over time, falling to nearly zero for papers published more than 25 years earlier. Across the full population of U.S. patents, the scientific ideas reused within a patent are on average 46 years old at the time of filing, and in 75% of patents the average reused idea is at least 37 years old. This helps explain why many scientific papers are not cited, even when their ideas are incorporated into downstream patents.[11]

'Insert Figure 1'

Next, we replicate the exact same analysis but predict the likelihood of citation as a function of the semantic instead of temporal distance between the patent and the paper on whose scientific idea(s) the patent builds. Specifically, we compute embedding vectors for each patent and paper based on their titles and abstracts, and calculate cosine similarity for all 7,139,906 patent–paper observations (Gonash et al., 2024). As shown in Figure 2, the predicted probability of citation declines steeply as the semantic distance between the patent and the paper increases. This pattern further supports the idea that patents tend to cite scientific papers only when the underlying content is closely aligned and thus more likely to be viewed as legally relevant.[12]

'Insert Figure 2'

5. Patents' Reliance on Scientific Ideas and Private Commercial Value

To demonstrate the value of our new approach and open dataset for strategy and innovation research, we revisit a long-standing research question: Are firm inventions that build on science associated with greater private value (Fleming & Sorenson, 2004; Krieger et al. 2024)? We follow the estimation strategy of Krieger et al. (2024) and rely on the KPSS measure of patent value, which captures the abnormal stock market returns experienced by a firm around the patent grant date (Kogan et al., 2017). To measure a patent's reliance on

[11] As previously discussed, besides directly citing the source paper, a patent may also cite it indirectly by referencing a paper or patent that cites the pioneering paper (e.g., Ahmadpoor & Jones, 2017). We find consistent results when using indirect citation as the outcome variable: both the predicted probability of citing a patent that cites the pioneering paper and the predicted probability of citing a paper that cites the pioneering paper decline sharply as the time lag between the patent's filing and the pioneering paper's publication increases. Results, not reported here to save space, are available from the authors upon request.

[12] We find consistent results when using indirect citation as the outcome: both the predicted probability of citing a patent that cites the pioneering paper and the predicted probability of citing a paper that cites the pioneering paper decrease sharply as the semantic distance between the focal patent and the source paper increases. Results, not reported here to save space, are available from the authors upon request.

scientific ideas, we incorporate both the traditional metric (*Science References*) and our newly developed text-based metric (*Science Phrases*).

We follow Krieger et al. (2024) and sample 1,331,880 U.S. patents granted to 7,112 public U.S. firms between 1980 and 2010. A key feature and limitation of the KPSS data is that abnormal stock returns are observed at the firm-week level, rather than at the level of individual patents. Because firms often have multiple patents granted in the same week, these patents are all associated with a single stock market reaction, and each patent is assigned the average return for that firm-week. Following the preferred estimation strategy of Krieger et al. (2024), we also aggregate the independent variables to the firm-week level to match the unit of observation, computing their averages across all patents granted to the same firm within a given week. We then estimate weighted regressions at the firm-week level, using the number of patents granted as weights.

'Insert Table 3'

Table 3 reports results from specifications estimated without firm fixed effects (columns 1–4) and with firm fixed effects (columns 5–8). The inclusion of firm fixed effects accounts for persistent, time-invariant heterogeneity across firms, such as differences in industry, R&D strategy, or firms' overall science reliance. Across all models, we find that *Science Phrases* is positively and significantly associated with higher patent value. A one standard deviation increase in *Science Phrases* corresponds to a 25.5% increase in patent value without firm fixed effects (column 2) and an 18.2% increase with firm fixed effects (column 6). In line with prior work, *Science References* also show a positive association: a one standard deviation increase is linked to an 11.6% increase in patent value without firm fixed effects (column 1) and a 15.3% increase with firm fixed effects (column 5). Columns 4 and 8 re-estimate the correlation between *Science Phrases* and patent value for the subsample of firm-weeks in which none of the firm's patents contain any science references. Interestingly, the effect of *Science Phrases* remains positive, significant at $p<0.001$, and large in magnitude in this restricted sample: a one standard deviation increase corresponds to a 22.9% increase in patent value without firm fixed effects (column 4) and a 7.5% increase with firm fixed effects (column 8). This suggests that *Science Phrases* capture remaining variation in a patent's reliance on science that is correlated with private patent value, even when no patent-to-paper citations are present.

Notably, the marginal effect of *Science Phrases* is larger than that of *Science References* in all specifications, and it remains statistically and economically significant when both metrics are included jointly (columns 3 and 7). When *Science References* is added to a model that already includes *Science Phrases*, the marginal effect of *Science Phrases* slightly decreases (from 25.5% to 22.7% without firm fixed effects, and from 18.2% to 14.4% with firm fixed effects), but remains economically meaningful and statistically significant at $p < 0.01$. In contrast, when *Science Phrases* is added to a model that already includes *Science References*, the marginal effect of *Science References* drops more sharply and loses statistical significance in the model without firm fixed effects (from 11.6% to 8.6%, and p-value from 0.01 to 0.07), while only slightly decreasing in the model with firm fixed effects (15.3% to 13.7%).

We conduct several robustness checks. One potential concern is that, even though all regressions control for patent text length, longer patents may mechanically contain more *Science Phrases*, which could bias the measure. A related concern is that multiple distinct noun phrases within a patent may be synonyms and capture the same underlying scientific idea, artificially inflating *Science Phrases*. Both issues are mitigated by using *Science Phrases Share,* i.e. the proportion of phrases reused from the scientific literature relative to all unique noun phrases in the patent. As shown in Table A.4 in Appendix, our main results remain consistent when using *Science Phrases Share* instead of *Science Phrases*. Second, a key limitation is that abnormal stock returns are observed at the firm-week level, which provides a noisy proxy for the private value of individual patents when multiple patents are granted to the same firm in a given week. This concern is especially relevant because patent value is highly skewed and unlikely to be evenly distributed across patents within a firm-week. To address this issue, we followed the preferred specification of Krieger et al. (2024) and use the firm-week rather than the patent as our unit of analysis. Nevertheless, our results remain robust when using the patent as the unit of observation (Table A.5) and when restricting the sample to single-patent firm-weeks, where stock market reactions provide a cleaner signal of private patent value (Table A.6). Finally, private patent value based on stock market reactions is observable only for patents assigned to U.S. public firms and represents just one proxy for patent value. To broaden the analysis, Table A.7 replicates the results for the full population of U.S. patents, including those assigned to non-U.S. or privately held firms, using forward citations as an alternative measure of patent value (Harhoff et al., 1999).

Taken together, these results reaffirm that firm inventions building on science are more valuable (Fleming & Sorenson, 2004; Krieger et al., 2024). Our patent-level findings align with firm-level survey evidence showing that patent-to-paper citations underestimate the true impact of science on firm innovation performance (Roach & Cohen, 2013). By detecting the reuse of scientific ideas even when uncited, our text-based approach complements citation-based indicators, broadens the lens on science–technology linkages, and uncovers firm performance effects that traditional measures overlook.

6. Discussion and Conclusion

How do firms turn scientific knowledge into technological progress and competitive advantage? Strategy research has long emphasized that access to and absorption of scientific ideas is central to firm innovation. Yet despite this wide recognition by scholars, firms and policymakers, much of the science that underpins firm technologies leaves no trace in the conventional metric of patent-to-paper citations. This gap raises a central question: to what extent do firms rely on scientific ideas that remain invisible in traditional citation data, and does this hidden reliance matter for the value of firm innovation?

We address this question by developing a novel text-based measure that identifies when patents reuse specific scientific ideas from the academic literature. Unlike patent-to-paper citations, which primarily serve legal functions, this method captures the underlying scientific ideas on which patents are built, regardless of whether a citation is provided. In doing so, we open the black box of uncited science and trace its role in firm innovation. Our findings show that firms frequently build on scientific ideas without citing the original source paper—particularly when the science is older or less similar in content. Crucially, this hidden reliance is economically meaningful: patents that draw more heavily on scientific ideas are associated with higher private value, even after controlling for patent-to-paper citations and within the subset of patents without such citations. Put differently, science shapes firm innovation in ways that conventional citation indicators miss.

This study makes three contributions to strategy and innovation research. First, it develops and validates a scalable complementary measure of firms' reliance on scientific ideas at the level of individual patents. Second, it shows that uncited science is systematically associated with private patent value, revealing a blind spot in citation-based approaches. Third, it introduces an open dataset that links the complete corpus of all U.S. patents with the corpus of all scientific publications until 2023. This resource enables researchers to trace idea-

level connections between science and technology at scale—capturing flows not only from science to technology but also, for the first time, from technology to science. Such bidirectional linkages cannot be observed using patent-to-paper citations, which are inherently unidirectional.

The open method and dataset we provide enables strategy and innovation scholars to revisit foundational questions and to address new ones about when, how, and why scientific ideas translate into firm innovation and competitive advantage. For instance: How do firms access and absorb external scientific knowledge in the process of innovation (Fleming & Sorenson, 2004)? How does reliance on science interact with firms' internal capabilities (Cohen & Levinthal, 1990)? And under what conditions does it translate into superior innovation performance (Krieger et al., 2024)? Importantly, our dataset also creates opportunities to study at scale the two-way relationship in which science and technology co-evolve through iterative feedback loops (Rosenberg, 1976; Stokes, 1997; Barbosu & Teodoridis, 2025). While patent-to-paper citations are inherently unidirectional and thus only capture flows from science to technology, the text-based idea linkages in our dataset allow researchers to trace connections in both directions. In principle, one could study flows from technology to science via paper-to-patent citations, but such citations are very rare (Hsiao & Torvik, 2020).

Naturally, our study has limitations. Not all technologies are patented, and patenting propensities vary across industries and firms. Moreover, inventors and attorneys may strategically frame or draft patent texts—just as they may strategically cite prior art—which could affect how scientific ideas are reflected in our metric. Finally, firms may draw on science in ways not captured by either patent citations or our text-based measures. These limitations underscore that our approach should be viewed as a complement, rather than a substitute, for existing measures of science–technology linkages.

In sum, this paper shows that scientific ideas—cited or uncited—are a critical foundation of firm innovation and competitive advantage. By uncovering hidden reliance on science, our text-based approach expands the empirical toolkit available to strategy and innovation scholars and opens new avenues for understanding how science shapes firm performance.

REFERENCES

Ahmadpoor, M., & Jones, B. F. (2017). The dual frontier: Patented inventions and prior scientific advance. *Science*, *357*(6351), 583-587.

Arora, A., Belenzon, S., & Sheer, L. (2021). Matching patents to Compustat firms, 1980–2015: Dynamic reassignment, name changes, and ownership structures. *Research Policy*, *50*(5), 104217.

Arts, S., Melluso, N., & Veugelers, R. (2025). Beyond citations: Measuring novel scientific ideas and their impact in publication text. *Review of Economics and Statistics*, forthcoming.

Barbosu S., Teodoridis, F., "*Catalysts of Discovery: How Technologies are Shaping the Future of Science and Innovation*", Palgrave MacMillan, forthcoming.

Bryan, K. A., Ozcan, Y., & Sampat, B. (2020). In-text patent citations: A user's guide. *Research Policy*, *49*(4), 103946.

Bush, V. (1945). *Science, the Endless Frontier: A Report to the President* (Washington: United States Government Printing Office, 1945).

Callaert, J., Pellens, M., & Van Looy, B. (2014). Sources of inspiration? Making sense of scientific references in patents. *Scientometrics*, *98*(3), 1617-1629.

Cohen, W. M., & Levinthal, D. A. (1990). Absorptive capacity: A new perspective on learning and innovation. *Administrative science quarterly*, *35*(1), 128-152.

Cohen, W. M., Nelson, R. R., & Walsh, J. P. (2002). Links and impacts: the influence of public research on industrial R&D. *Management Science*, *48*(1), 1-23.

Fleming, L., & Sorenson, O. (2004). Science as a map in technological search. *Strategic Management Journal*, *25*(8-9), 909-928.

Ghosh, M., Erhardt, S., Rose, M. E., Buunk, E., & Harhoff, D. (2024). PaECTER: Patent-level representation learning using citation-informed transformers. *arXiv preprint arXiv:2402.19411*.

Harhoff, D., Narin, F., Scherer, F. M., & Vopel, K. (1999). Citation frequency and the value of patented inventions. *Review of Economics and Statistics*, 81(3), 511-515.

Hsiao, T. K., & Torvik, V. I. (2020). Technology footprints in scientific discovery: Citation contexts of paper-to-patent citations. *Proceedings of the Association for Information Science and Technology*, 57(1), e337.

Iaria, A., Schwarz, C., & Waldinger, F. (2018). Frontier knowledge and scientific production: Evidence from the collapse of international science. *Quarterly Journal of Economics*, 133(2), 927-991.

Jaffe, A. B., Trajtenberg, M., & Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly journal of Economics*, *108*(3), 577-598.

Kogan, L., Papanikolaou, D., Seru, A., & Stoffman, N. (2017). Technological innovation, resource allocation, and growth. *Quarterly journal of Economics*, *132*(2), 665-712.

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

Lampe, R. (2012). Strategic citation. *Review of Economics and Statistics*, *94*(1), 320-333.

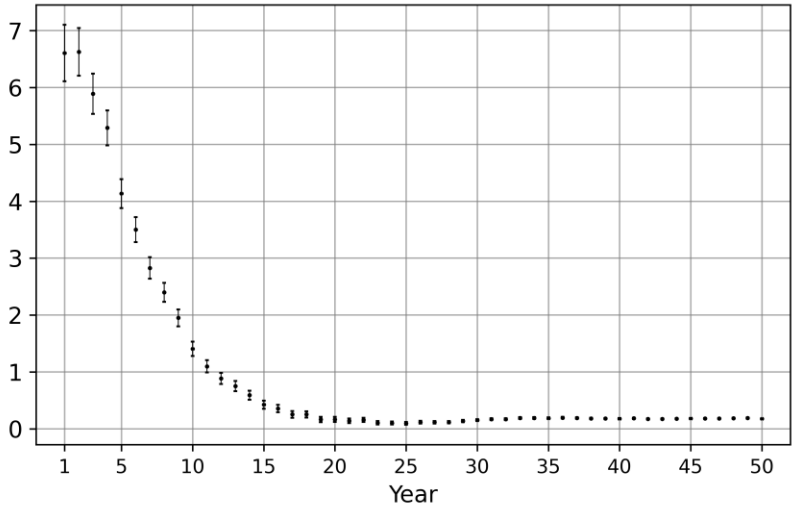
Mansfield, E. (1991). Academic research and industrial innovation. *Research Policy*, *20*(1), 1-12.

Marx, M., & Fuegi, A. (2020). Reliance on science: Worldwide front-page patent citations to scientific articles. *Strategic Management Journal*, *41*(9), 1572-1594.

Mokyr, J. (2011). The gifts of Athena: Historical origins of the knowledge economy. In *The gifts of Athena*. Princeton University Press.

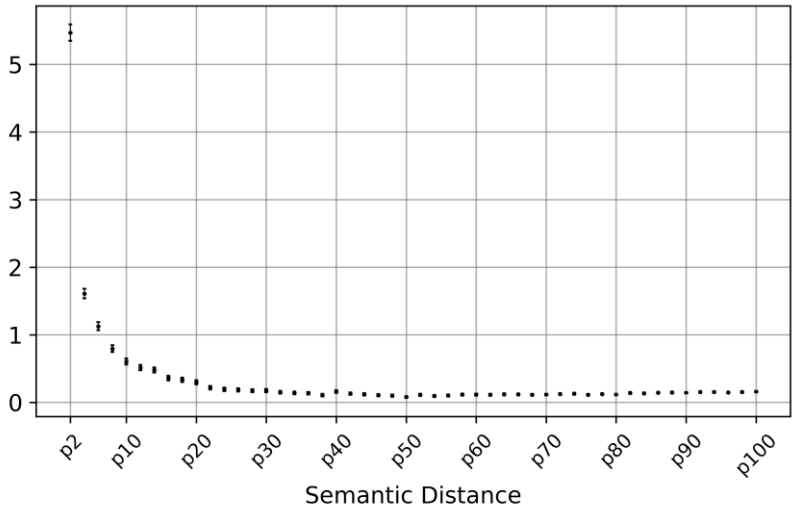
Narin, F., Hamilton, K. S., & Olivastro, D. (1997). The increasing linkage between US technology and public science. *Research Policy*, 26(3), 317-330.
Roach, M., & Cohen, W. M. (2013). Lens or prism? Patent citations as a measure of knowledge flows from public research. *Management Science*, 59(2), 504-525.
Rosenberg, N. (1976). Technological change in the machine tool industry, 1840–1910. In Perspectives on Technology (pp. 9–31). Cambridge University Press.
Scharfmann, E., Marx, M., & Fleming, L. (2024). Scientists invent better patents, and inventors publish better science. Working paper.
Stephan, P. (2012). *How economics shapes science*. Harvard University Press.
Stokes, D. E. (1997). Pasteur's quadrant: Basic science and technological innovation. Brookings Institution Press.
Teplitskiy, M., Duede, E., Menietti, M., & Lakhani, K. R. (2022). How status of research papers affects the way they are read and cited. *Research policy*, 51(4), 104484.

Figure 1: Predicted probability that patents cite the scientific papers they build on, by time lag between paper and patent



Notes: This figure is based on 7,139,906 patent–scientific idea (noun phrase) pairs corresponding to 2,468,508 unique patents each reusing at least one of 100,000 randomly sampled scientific ideas (noun phrases) first introduced in the scientific literature. It plots the predicted probability (in %) that a patent cites the pioneering paper that introduced the idea it reuses, as a function of the time lag (in years) between the paper's publication and the patent's filing. Predictions and 95% confidence intervals are based on a linear probability model with year-lag specific indicators. The model includes fixed effects for the scientific subfield (OpenAlex 4-digits) of the pioneering paper, the technology subclass (CPC 4-digits) of the reusing patent, and the publication year of the paper. Additional controls for the pioneering paper include abstract availability, text length (number of unique phrases in title and abstract), and the number of papers cited. Controls for the reusing patent include text length (number of unique phrases in the full text), and the number of cited scientific papers.

Figure 2: Predicted probability that patents cite the scientific papers they build on, by content similarity between paper and patent



Notes: This figure is based on 7,139,906 patent–scientific idea (noun phrase) pairs corresponding to 2,468,508 unique patents each reusing at least one of 100,000 randomly sampled scientific ideas (noun phrases) first introduced in the scientific literature. It plots the predicted probability (in %) that a patent cites the pioneering paper that introduced the idea it reuses, as a function of the semantic distance between the reusing patent and the pioneering paper, modeled as a series of fifty 2-percentile bins. Bin p2 represents patents most similar in content to the pioneering paper, while bin p100 represents patents most distant in content. Predictions and 95% confidence intervals are based on a linear probability model. The model includes fixed effects for the subfield (OpenAlex 4-digits) of the pioneering paper, the technology subclass (CPC 4-digits) of the reusing patent, and the publication year of the paper. Additional controls for the pioneering paper include abstract availability, text length (number of unique phrases in title and abstract), and the number of papers and journals cited. Controls for the reusing patent include text length (number of unique phrases in the full text), and the number of cited patents and scientific papers.

Table 1: Patent-Level Summary Statistics

	Full Sample (n = 7,699,586)						
	Mean	Std	Min	p25	p50	p75	Max
Phrases	254.663	307.311	1.000	118.000	179.000	281.000	81,913.000
Science Phrases Bin	0.963	0.189	0.000	1.000	1.000	1.000	1.000
Science Phrases	32.155	87.031	0.000	4.000	9.000	23.000	10,435.000
Science Phrases Share	0.083	0.080	0.000	0.029	0.054	0.108	0.648
Science References Bin	0.232	0.422	0.000	0.000	0.000	0.000	1.000
Science References	2.971	17.519	0.000	0.000	0.000	0.000	1,502.000
Science Phrases Avg Age	45.725	13.214	0.000	36.778	45.667	54.581	120.000
Science References Avg Age	10.246	10.379	0.000	5.000	9.000	14.500	202.000

Notes: US patents granted between 1976 and June 2023. *Science Phrases Avg Age* and *Science References Avg Age* are calculated only for patents with at least one reused scientific noun phrase and at least one scientific reference, respectively.

Table 2: Summary Statistics Across Patents with Scientific References, University Assignees, and Scientist Inventors

	Mean	Std	Min	Max	Mean	Std	Min	Max	Z	p-value
	Patents with Scientific References (n=1,785,379)				Patents with no Scientific References (n= 5,914,207)					
Science Phrases Bin	0.995	0.067	0.000	1.000	0.953	0.211	0.000	1.000	85.217	0.000
Science Phrases	91.549	160.586	0.000	10,435.000	14.325	26.299	0.000	5,330.000	1,258.985	0.000
Science Phrases Share	0.148	0.101	0.000	0.632	0.061	0.056	0.000	0.643	1,131.971	0.000
Science References Bin	1.000	0.000	1.000	1.000	0.000	0.000	0.000	0.000	2,028.338	0.000
Science References	12.814	34.604	1.000	1,502.000	0.000	0.000	0.000	0.000	2,028.338	0.000
	University Patents (n=321,086)				Non-University Patents (n=7,378,500)					
Science Phrases Bin	0.987	0.115	0.000	1.000	0.962	0.191	0.000	1.000	23.394	0.000
Science Phrases	102.842	164.626	0.000	4,535.000	29.159	80.603	0.000	10,435.000	433.945	0.000
Science Phrases Share	0.161	0.109	0.000	0.627	0.078	0.074	0.000	0.643	451.529	0.000
Science References Bin	0.627	0.484	0.000	1.000	0.215	0.411	0.000	1.000	395.757	0.000
Science References	16.151	39.774	0.000	964.000	2.398	15.606	0.000	1,502.000	446.967	0.000
	Patents with Scientist Inventors (n=3,945,226)				Patents without Scientist Inventors (n=3,754,360)					
Science Phrases Bin	0.981	0.136	0.000	1.000	0.944	0.229	0.000	1.000	88.739	0.000
Science Phrases	48.610	112.739	0.000	10,435.000	15.022	39.962	0.000	5,465.000	891.668	0.000
Science Phrases Share	0.103	0.088	0.000	0.643	0.058	0.056	0.000	0.624	786.205	0.000
Science References Bin	0.328	0.469	0.000	1.000	0.131	0.338	0.000	1.000	472.384	0.000
Science References	4.926	23.052	0.000	1,502.000	0.917	7.925	0.000	1,380.000	496.787	0.000

Notes: Sample includes 7,699,586 U.S. patents granted between 1976 and 2023. "Patents with Scientific References" are those citing at least one scientific publication, the comparison group is all other patents. "University Patents" are those with at least one university assignee, the comparison group is all others. "Patents with Scientist Inventors" are those with at least one inventor who has authored a scientific publication (Scharfmann et al., 2024), the comparison group is all others. Z-values are from Mann–Whitney tests.

Table 3: Patents' Reliance on Scientific Ideas and Private Value

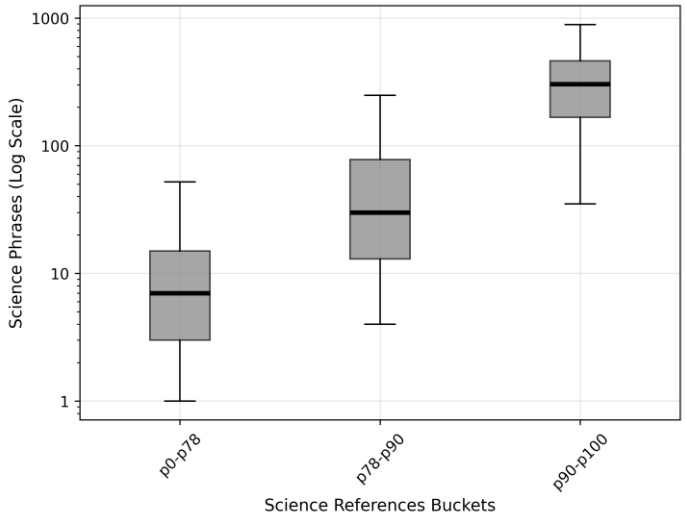
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Science References	2.347		1.738		3.096		2.766	
	(0.948)		(0.973)		(0.766)		(0.752)	
	[.013]		[.074]		[.000]		[.000]	
Science Phrases		3.646	3.245	3.704		2.614	2.059	1.222
		(1.141)	(1.168)	(0.657)		(0.597)	(0.558)	(0.313)
		[.001]	[.006]	[.000]		[.000]	[.000]	[.000]
Observations	394,534	394,534	394,534	238,943	393,182	393,182	393,182	237,527
R-squared	0.105	0.105	0.106	0.085	0.364	0.363	0.364	0.417
Technology Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Week Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm Fixed Effects	No	No	No	No	Yes	Yes	Yes	Yes
Sample	Full	Full	Full	No PCS	Full	Full	Full	No PCS
Marginal Effects (in %)								
Science References	11.647		8.6233		15.329		13.691	
Science Phrases		25.519	22.709	22.945		18.241	14.373	7.538

Notes: OLS regressions link patent value (KPSS) to reliance on scientific ideas as measured by the number of *Science References* and *Science Phrases*. The unit of observation is the firm-week. The dependent variable is the mean KPSS of all patents granted to a firm in a given week. Independent variables are log-transformed at the patent level and averaged within the firm-week. Regressions are weighted by the number of patents in the firm-week. All models include issue-week and 4-digit CPC technology-class fixed effects, and control for the text length of the patent (average number of noun phrases). The sample includes 1,331,880 U.S. patents granted to 7,112 public U.S. firms between 1980 and 2010. Models 4 and 8 focus only on firm-weeks in which none of the firm's patents contain science references. Noun phrases are extracted from the full patent text. Marginal effects represent the percentage change in average KPSS per firm-week for a one–standard deviation increase in the independent variable. Robust standard errors (clustered at the firm level) are reported in parentheses. P-values of the estimated coefficients are reported in brackets.

ONLINE APPENDIX

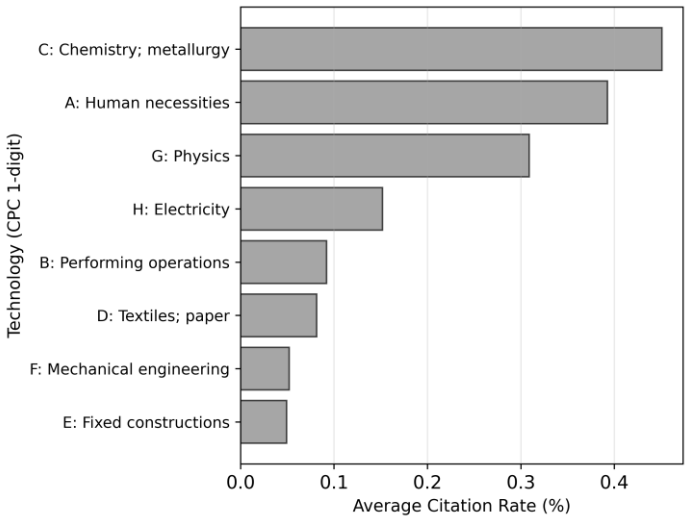
Appendix A

Figure A.1 Distribution of *Science Phrases* across Percentiles of *Science References*



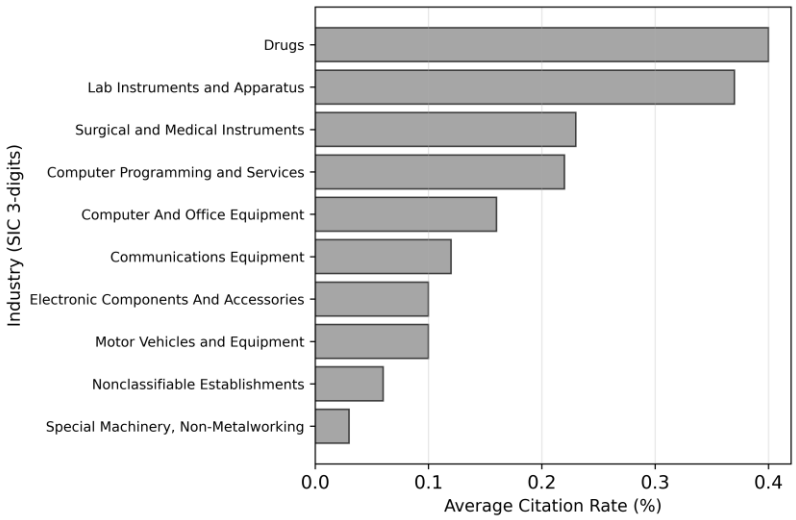
Notes: This figure shows box plots of the distribution of *Science Phrases* across percentile buckets of *Science References*. The x-axis groups patents into three buckets of *Science References*: p0–p78, p78–p90, and p90–p100. Since 78% of patents cite no scientific paper, the first bucket captures this large share with zero PCS. The y-axis uses a logarithmic scale for visualization, but the tick labels are reported in original counts for ease of interpretation. Boxes represent the interquartile range (25th–75th percentile), the thick horizontal line marks the median, and whiskers extend from the 5th to 95th percentiles.

Figure A.2 Average Citation Rate by Technology



Notes: This figure reports the average citation rate across technologies (CPC 1-digit) for 7,139,906 patent–phrase pairs. These pairs are constructed from a random sample of 100,000 scientific ideas (noun phrases) first introduced in the scientific literature and identifying all patents reusing these phrases in their text. The citation rate (in %) corresponds to the share of patent–phrase pairs in which the patent cites the pioneering paper where the reused phrase first appeared.

Figure A.3 Average Citation Rate by Industry



Notes: This figure reports the average citation rate across industries (SIC 3-digit) for 2,217,751 patent–phrase pairs. These pairs are constructed from a random sample of 100,000 scientific ideas (noun phrases) first introduced in the scientific literature and identifying all patents filed by U.S. public firms that reuse these phrases. The sample covers 851,385 granted patents filed between 1980 and 2021 by 6,088 unique firms, as linked through DISCERN2.0. The citation rate (in %) corresponds to the share of patent–phrase pairs in which the patent cites the pioneering paper where the reused phrase first appeared. Industries with zero citation rates or fewer than 50 firms are excluded. Industry names longer than 30 characters are abbreviated using GPT-4o-mini.

Table A.1 Most Reused Scientific Phrases in Patents, by Decade

1951-1960	1961-1970	1971-1980	1981-1990
copolymer (365,333)	polyethylene_glycol (260,386)	liquid_crystal_display (432,075)	wide_area_network (420,415)
polypropylene (326,747)	silicone (229,688)	optical_storage (241,886)	field-programmable_gate_array (142,338)
polyester (324,224)	peptide (229,436)	ethernet (237,520)	electroporation (79,755)
monomer (295,132)	polycarbonate (222,169)	gsm (175,930)	metropolitan_area_network (62,824)
polyurethane (244,726)	polyamide (207,337)	liquid_crystal (146,299)	apoptosis (61,471)
isopropanol (166,842)	chemical_vapor_deposition (195,345)	magnetic_disk_storage (144,070)	protocol_stack (49,610)
silica_gel (155,672)	tetrahydrofuran (190,691)	amino_acid_sequence (137,754)	transcription_factor (46,880)
styrene (151,477)	multiplexer (169,995)	plasmid (128,924)	solid-state_drive (45,399)
cathode_ray_tube (151,142)	epoxy_resin (168,313)	nucleotide_sequence (127,637)	transgene (45,116)
elastomer (142,313)	polypeptide (166,255)	dna_sequence (123,808)	interleukin (44,898)
1991-2000	**2001-2010**	**2011-2020**	
hypertext_transfer_protocol (74,748)	microrna (16,593)	smart_contract (4,400)	
extensible_markup_language (64,998)	bevacizumab (13,701)	nivolumab (4,130)	
atomic_layer_deposition (61,256)	imatinib (11,805)	pembrolizumab (3,649)	
carbon_nanotube (45,643)	cetuximab (11,666)	cryptocurrency (2,938)	
organic_light-emit_diode (37,992)	erlotinib (10,408)	generative_adversarial_network (2,883)	
paclitaxel (32,555)	gefitinib (10,386)	beidou_navigation_satellite (2,789)	
graphene (29,191)	bortezomib (9,176)	cas9 (2,710)	
nand_flash_memory (26,987)	everolimus (8,138)	vemurafenib (2,548)	
phage_display (26,460)	mimo_transmission (7,885)	tensorflow (2,500)	
qr_code (25,913)	abiotic_stress_resistance (7,855)	kubernete (2,228)	

Notes: Top 10 noun phrases originating in scientific publications and subsequently reused in patents, grouped by the decade of first adoption in patents. The decade is defined by the filing year of the earliest patent containing the phrase. Values in parentheses indicate the total number of distinct patents in which the phrase appears through 2023.

Table A.2 Most Reused Scientific Phrases By Selected Industries and Firms

Industry	Firm	N. Patents	Science Phrases (avg)	Top Reused Science Phrases
Chemicals and Allied Products	Genentech	1,715	607	polypeptide (90.9%); amino_acid_sequence (85.8%); peptide (81.6%); plasmid (79.9%); dna_sequence (79.1%)
	Incyte	1,206	426	bind_affinity (85.6%); radionuclide (83.7%); hydrophobicity (82.5%); magnesium_stearate (81.9%); polyvinylpyrrolidone (81.3%)
	Abbvie	1,685	424	polyethylene_glycol (58.6%); liposome (46.0%); peptide (42.6%); magnesium_stearate (40.1%); cycloalkyl (38.1%)
	Amgen	2,021	415	peptide (73.2%); polypeptide (68.7%); polyethylene_glycol (67.7%); amino_acid_sequence (67.6%); monoclonal_antibody (59.8%)
	Ionis Pharm	1,240	372	oligonucleotide (97.3%); phosphorothioate (93.2%); nucleoside (89.5%); mrna (87.3%); cytosine (87.1%)
Measuring, Analyzing, and Controlling Instruments	Bio Rad Lab	1,751	147	peptide (41.5%); oligonucleotide (24.9%); polyethylene_glycol (23.9%); dichloromethane (23.8%); solvate (23.6%)
	Hologic	1,028	106	nucleic_acid_sequence (32.4%); oligonucleotide (30.7%); amplification_reaction (29.2%); nucleic_acid_amplification (27.5%); polynucleotide (25.8%)
	Abbott Lab	9,374	104	polyurethane (21.6%); polyester (20.0%); copolymer (20.0%); silicone (19.9%); polyethylene_glycol (19.0%)
	Thermo Fisher	3,853	104	peptide (24.9%); oligonucleotide (23.1%); mass_spectrometry (21.7%); mass_spectrometer (21.5%); ion_source (17.7%)
	Waters	1,201	68	mass_spectrometer (68.2%); mass_spectrometry (61.6%); liquid_chromatography (59.2%); ion_source (50.9%); mass_analyser (42.3%)
Petroleum Refining and Related Industries	Exxon Mobil	11,746	81	copolymer (32.0%); monomer (27.2%); polyolefin (20.9%); molecular_sieve (20.8%); polypropylene (20.4%)
	Chevron	6,212	65	copolymer (21.4%); monomer (20.0%); molecular_sieve (19.2%); cyclohexane (16.3%); n-hexane (15.9%)
	Amoco	2,160	52	monomer (19.3%); copolymer (19.2%); catalyst_bed (14.7%); molecular_sieve (13.3%); polyester (13.3%)
	Mobil	4,888	42	silica-alumina (27.6%); extrudate (22.5%); x-ray_diffraction (20.4%); ion_exchange (20.4%); molecular_sieve (17.7%)
	Texaco	2,770	33	synthesis_gas (12.5%); polyol (11.6%); copolymer (11.2%); alkylene_oxide (9.2%); pore_volume (8.9%)
Oil and Gas Extraction	Conocophillips	4,924	40	copolymer (24.2%); monomer (23.0%); cyclohexane (18.7%); homopolymer (14.5%); pore_volume (13.9%)
	Shell Oil	1,085	34	copolymer (21.8%); cyclohexanone (15.3%); tetrahydrofuran (13.6%); corrosion_inhibitor (13.3%); aqueous_dispersion (12.6%)
	Halliburton	11,673	25	proppant (17.2%); gravel_pack (13.1%); copolymer (12.2%); cement_slurry (10.2%); fiber_optic (10.1%)
	Schlumberger	10,283	23	proppant (7.6%); gamma_ray (7.4%); seismic_data (7.1%); elastomer (5.9%); gravel_pack (5.7%)
	Baker Intl	7,652	20	tungsten_carbide (26.0%); polycrystalline_diamond (21.5%); downhole_motor (18.9%); hydrocarbon_production (18.0%); drillstre (16.5%)

Notes: For each selected industry (SIC 2-digit code), the table lists the top five U.S. public firms ranked by the average number of distinct scientific noun phrases reused per patent. For each firm, the five most frequently reused phrases are shown, with percentages indicating the share of the firm's patents in which the phrase appears. For example, Genentech averages 607 scientific phrases per patent, and 90.9% of its patents contain the phrase polypeptide. Firm–patent linkages are based on DISCERN2.0. The sample includes U.S. granted patents from 1980–2021 and is restricted to firms with at least 1,000 patents.

Table A.3: Summary Statistics Restricted to Patents without Scientific References

	Mean	Std	Min	Max	Mean	Std	Min	Max	Z	p-value
	University Patents (n=119,888)				Non-University Patents (n=5,794,319)					
Science Phrases Bin	0.968	0.176	0.000	1.000	0.953	0.211	0.000	1.000	8.751	0.000
Science Phrases	19.209	29.715	0.000	1,004.000	14.224	26.214	0.000	5,330.000	94.303	0.000
Science Phrases Share	0.082	0.067	0.000	0.627	0.061	0.055	0.000	0.643	123.860	0.000
	Patents with Scientist Inventors (n=2,652,135)				Patents without Scientist Inventors (n= 3,262,072)					
Science Phrases Bin	0.973	0.162	0.000	1.000	0.938	0.242	0.000	1.000	74.595	0.000
Science Phrases	19.382	32.869	0.000	2,822.000	10.215	18.382	0.000	5,330.000	585.799	0.000
Science Phrases Share	0.074	0.063	0.000	0.643	0.051	0.046	0.000	0.624	482.371	0.000

Notes: Sample restricted to 5,914,207 U.S. patents granted between 1976 and 2023 with no citations to papers. "University Patents" are those with at least one university assignee, the comparison group is all others. "Patents with Scientist Inventors" are those with at least one inventor who has authored a scientific publication (Scharfmann et al., 2024), the comparison group is all others. Z-values are from Mann-Whitney tests.

Table A.4: Patents' Reliance on Scientific Ideas and Private Value: *Science Phrases Share*

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Science References	2.347		1.806		3.096		2.589	
	(0.948)		(0.996)		(0.766)		(0.713)	
	[.013]		[.070]		[.000]		[.000]	
Science Phrases Share		29.591	23.173	35.641		34.622	26.633	11.101
		(13.852)	(14.493)	(8.237)		(8.164)	(7.223)	(3.762)
		[.033]	[.110]	[.000]		[.000]	[.000]	[.003]
Observations	394,534	394,534	394,534	238,943	393,182	393,182	393,182	237,527
R-squared	0.105	0.105	0.105	0.085	0.364	0.363	0.364	0.417
Technology Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Week Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm Fixed Effects	No	No	No	No	Yes	Yes	Yes	Yes
Sample	Full	Full	Full	No PCS	Full	Full	Full	No PCS
Marginal Effects (in %)								
Science References	11.647	13.609	10.654	13.180	15.329	15.879	12.216	4.087
Science Phrases			8.961				12.815	

Notes: OLS regressions link patent value (KPSS) to reliance on scientific ideas as measured by the number of *Science References* and *Science Phrases Share*. The unit of observation is the firm-week. The dependent variable is the mean KPSS of all patents granted to a firm in a given week. Independent variables are log-transformed at the patent level and averaged within the firm-week. Regressions are weighted by the number of patents in the firm-week. All models include issue-week and 4-digit CPC technology-class fixed effects, and control for the text length of the patent (average number of noun phrases). The sample includes 1,331,880 U.S. patents granted to 7,112 public U.S. firms between 1980 and 2010. Models 4 and 8 include only firm-weeks where all patents have no science references. Noun phrases are extracted from the full patent text. Marginal effects represent the percentage change in average KPSS per firm-week for a one–standard deviation increase in the independent variable. Robust standard errors (clustered at the firm level) are reported in parentheses. P-values of the estimated coefficients are reported in brackets.

Table A.5: Patents' Reliance on Scientific Ideas and Private Value: Patent-Level Analysis

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Science References	1.643		1.232		1.117		1.012	
	(0.474)		(0.465)		(0.299)		(0.300)	
	[.000]		[.008]		[.000]		[.001]	
Science Phrases Share		2.831	2.591	2.435		0.897	0.726	0.648
		(0.481)	(0.471)	[0.416]		(0.165)	(0.157)	(0.132)
		[.000]	[.000]	[.000]		[.000]	[.000]	[.000]
Observations	1,331,870	1,331,870	1,331,870	1,027,294	1,330,532	1,330,532	1,330,532	1,025,897
R-squared	0.086	0.087	0.087	0.079	0.357	0.357	0.357	0.343
Technology Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Week Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm Fixed Effects	No	No	No	No	Yes	Yes	Yes	Yes
Sample	Full	Full	Full	No PCS	Full	Full	Full	No PCS
Marginal Effects (in %)								
Science References	10.246		7.688		6.963		6.309	
Science Phrases		27.997	25.616	23.583		8.863	7.176	6.276

Notes: OLS regressions link patent value (KPSS) to reliance on scientific ideas as measured by the number of *Science References* and *Science Phrases*. The unit of observation is a single patent. The dependent variable is the KPSS. Independent variables are log-transformed. All models include issue-week and 4-digit CPC technology-class fixed effects, and control for the text length of the patent (number of noun phrases). The sample includes 1,331,880 U.S. patents granted to 7,112 public U.S. firms between 1980 and 2010. Models 4 and 8 include only patents with no science references. Noun phrases are extracted from the full patent text. Marginal effects represent the percentage change in KPSS for a one–standard deviation increase in the independent variable. Robust standard errors (clustered at the firm level) are reported in parentheses. P-values of the estimated coefficients are reported in brackets.

Table A.6: Patents' Reliance on Scientific Ideas and Private Value: Patent-Level Analysis for Single-Patent Firm-Weeks

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Science References	-0.397		-1.132		1.863		1.757	
	(0.706)		(0.713)		(0.488)		(0.490)	
	[.574]		[.113]		[.000]		[.000]	
Science Phrases Share		4.581	4.855	4.196		1.143	0.834	0.782
		(0.728)	(0.734)	(0.644)		(0.249)	(0.242)	(0.248)
		[.000]	[.000]	[.000]		[.000]	[.001]	[.002]
Observations	218,954	218,954	218,954	163,215	217,594	217,594	217,594	161,790
R-squared	0.073	0.075	0.075	0.080	0.569	0.569	0.569	0.579
Technology Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Week Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm Fixed Effects	No	No	No	No	Yes	Yes	Yes	Yes
Sample	Single-patent weeks	Single-patent weeks	Single-patent weeks	Single-patent weeks with no PCS	Single-patent weeks	Single-patent weeks	Single-patent weeks	Single-patent weeks with no PCS
Marginal Effects (in %)								
Science References	10.246		7.688		6.963		6.309	
Science Phrases		27.997	25.616	23.583		8.863	7.176	6.276

Notes: OLS regressions link patent value (KPSS) to reliance on scientific ideas as measured by the number of *Science References* and *Science Phrases*. The unit of observation is the firm-week with only one patent. The dependent variable is KPSS of the single patent granted to a firm in a given week. Independent variables are log-transformed at the patent level and averaged within the firm-week. Regressions are weighted by the number of patents in the firm-week. All models include issue-week and 4-digit CPC technology-class fixed effects, and control for the text length of the patent (average number of noun phrases). The sample includes patents granted to 7,093 public U.S. firms between 1980 and 2010. Noun phrases are extracted from the full patent text. Marginal effects represent the percentage change in the KPSS per patent for a one–standard deviation increase in the independent variable. Standard errors are clustered at the firm level.

Table A.7: Patents' Reliance on Scientific Ideas and Forward Citations

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Science References	0.120		0.106		0.106		0.099	
	(0.002)		(0.002)		(0.004)		(0.004)	
	[.000]		[.000]		[.000]		[.000]	
Science Phrases		0.107	0.076	0.101		0.071	0.047	0.068
		(0.002)	(0.002)	(0.002)		(0.004)	(0.004)	(0.004)
		[.000]	[.000]	[.000]		[.000]	[.000]	[.000]
Observations	4,138,107	4,138,107	4,138,107	3,313,476	1,504,485	1,504,485	1,504,485	1,160,256
Pseudo R2	0.220	0.218	0.220	0.197	0.275	0.273	0.275	0.255
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Tech FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	No	No	No	No	Yes	Yes	Yes	Yes
Sample	Full	Full	Full	No PCS	Full	Full	Full	No PCS
Marginal Effects								
Science References	9.363		8.264		8.141		7.594	
Science Phrases		13.811	9.760	10.904		8.762	5.820	7.099

Notes: Poisson quasi–maximum likelihood regressions link forward citations to reliance on scientific ideas, measured by the number of *Science References* and *Science Phrases*. The unit of observation is the patent. The dependent variable is the number of forward citations within 10 years of grant. Independent variables are log-transformed after adding one. All models include grant year and technology class (4-digit CPC) fixed effects, and control for patent text length (number of noun phrases). Models 1–4 include all U.S. patents granted between 1980 and 2012; Models 5–8 restrict to patents from 7,303 unique U.S. public firms; Models 4 and 8 further restrict to patents with no science references. Noun phrases are extracted from the full patent text. Marginal effects represent the percentage change in forward citations for a one-standard-deviation increase in the explanatory variable.

Appendix B - New Phrase Scoring

To identify scientific ideas, we build on Arts et al. (2025), who processed the full text of all scientific publications and conference proceedings in OpenAlex from 1666 to 2023, yielding 27,079,343 distinct noun phrases. We then apply a large-scale classification procedure using the GPT-4o-mini model, which assigns each phrase a confidence score indicating the likelihood that it reflects a genuine scientific idea. We submitted batch requests via the OpenAI API. Each batch contained 100 noun phrases presented in a structured input table, with each row listing the phrase, the title of the first paper in which it appeared, the journal name, and the year of publication. The model returned an idea score on a 1–3 scale:

- 1 = Very unlikely to represent a new idea
- 2 = Likely to represent a new idea
- 3 = Very likely to represent a new idea

The exact prompt used for classification was as follows:

```
# Input Data

You are given a data table where each record contains the following fields:

    - phrase: A noun phrase extracted from the title or abstract of a scientific paper.
    - title: The title of the paper.
    - journal: The name of the journal in which the paper was published.
    - year: The publication year of the paper.

# Background Context

    - The noun phrases have been extracted from all papers indexed in OpenAlex.
    - Each noun phrase is linked to the first corresponding paper in which it appeared (by
publication date) within the entire corpus of indexed papers.

# Task

Evaluate whether the given noun phrase represents a new scientific idea introduced by the
specified paper in the given year. A "new scientific idea" should be interpreted broadly and may
include:
    - A novel concept, theory, method, tool, technology, or discovery (e.g., a newly identified
molecule or chemical structure).

Your judgment should be based on all available information provided in the dataset but
particularly on all other available information at your disposal.

# Response Format

Provide your evaluation using the following structure:

    - Idea Score: Assign a score from 1 to 3 indicating how likely the noun phrase represents a
new scientific idea introduced by the given paper and year.
        -- 1: Very unlikely
        -- 3: Very likely

Ensure your assessment is well-reasoned and considers the provided metadata and all other
available information at your disposal in your judgment.
```

Across 27,079,601 unique new phrases, the model assigned:

- 2,403,306 phrases (8.9%) a score of 1
- 13,771,184 phrases (50.8%) a score of 2
- 10,905,111 phrases (40.3%) a score of 3

We construct two alternative versions of the *New Phrase* measure by applying cutoffs to the idea scores:

- *New Phrase (Score 2 and 3)*: excludes phrases with a score of 1
- *New Phrase (Score 3)*: retains only phrases scored as 3, excluding both 1 and 2

Tables B.1 and B.2 below provide summary statistics at the phrase and paper level respectively.

Table B.1: Summary Statistics Phrase Level

	Reuse is subsequent papers									
	#	Mean	Std	Min	p25	p50	p75	p95	Max	Skew
New Phrase (All Scores)	27,079,601	17.228	434.835	1	1	2	5	34	416,734	315.421
New Phrase (Score 2 and 3)	24,676,295	15.503	396.043	1	1	2	5	31	416,734	346.605
New Phrase (Score 3)	10,905,111	18.357	445.352	1	1	2	5	36	344,357	289.012

Notes: *New Phrase (All Scores)* includes all new phrases with no exclusions. *New Phrase (Score 2 and 3)* includes only phrases with score 2 and 3, thus excluding phrases with score 1. *New Phrase (Score 3)* includes only phrases Score 3, thus excluding phrases with score 1 and 2. Summary statistics of the reuse of new phrases in subsequent papers, i.e. how many later papers reuse the phrase in their text. # is the number of units. p25, p50, p75, p95 are respectively the 25th, 50th, 75th, 95th percentile. The skewness (skew) of the distributions is the measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. A positive skewness indicates that the distribution has a long right tail, while a negative skewness indicates a long left tail.

Table B.2: Summary Statistics Paper Level

	Mean	Std. Dev	Min	p25	p50	p75	p95	Max	Skew
New Phrase (All Scores) (Binary)	0.245	0.430	0	0	0	0	1	1	1.184
New Phrase (All Scores)	0.360	0.779	0	0	0	0	2	223	4.128
New Phrase Reuse (All Scores)	6.555	268.222	0	0	0	0	12	416,735	522.944
New Phrase (Score 2 and 3) (Binary)	0.229	0.420	0	0	0	0	1	1	1.292
New Phrase (Score 2 and 3)	0.328	0.735	0	0	0	0	2	216	4.262
New Phrase Reuse (Score 2 and 3)	5.081	233.776	0	0	0	0	9	416,734	609.017
New Phrase (Score 3) (Binary)	0.111	0.314	0	0	0	0	1	1	2.476
New Phrase (Score 3)	0.145	0.471	0	0	0	0	1	43	4.735
New Phrase Reuse (Score 3)	2.659	173.650	0	0	0	0	3	344,357	768.870

Notes: *New Phrase (All Scores)* includes all new phrases with no exclusions. *New Phrase (Score 2 and 3)* includes only phrases with score 2 and 3, thus excluding phrases with score 1. *New Phrase (Score 3)* includes only phrases Score 3, thus excluding phrases with score 1 and 2. n= 75,295,921 papers published between 1901 and 2023. p25, p50, p75, p95 are respectively the 25th, 50th, 75th, 95th percentile. The skewness (skew) of the distributions is the measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. A positive skewness indicates that the distribution has a long right tail, while a negative skewness indicates a long left tail.

To assess whether the GPT-4o-mini model produces a meaningful confidence score—that is, whether higher-scoring phrases more accurately capture genuine scientific ideas—we replicate the validation exercise from Arts et al. (2025)[13] on distinguishing Nobel Prize papers from matched controls (results displayed in Table B.3 below).

Table B.3: Nobel Prize Classification

Variable	Precision	Recall	AUC	Marginal Effect
Panel A: Ex-ante				
New Phrase (All Scores)	69.87	65.63	0.7635	20.93
New Phrase (Score 2 and 3)	71.35	66.15	0.7738	21.90
New Phrase (Score 3)	73.98	66.15	0.7923	24.60
Panel B: Ex-post				
New Phrase Reuse (All Scores)	74.42	67.19	0.7938	25.66
New Phrase Reuse (Score 2 and 3)	74.85	67.19	0.7980	25.84
New Phrase Reuse (Score 3)	76.32	67.71	0.8063	30.99

Notes: Reproduction of Table 3 Column 2 (Panel A of current table) and Column 11 (Panel B of current table) from Arts et al. (2025). *New Phrase (All Scores)* includes all new phrases with no exclusions. *New Phrase (Score 2 and 3)* includes only phrases with score 2 and 3, thus excluding phrases with score 1. *New Phrase (Score 3)* includes only phrases Score 3, thus excluding phrases with score 1 and 2. Logit regressions with a binary outcome for Nobel Prize paper, robust standard errors in parentheses. Sample includes 1,168 papers (584 Nobel Prize papers matched with 584 control papers from the same year, journal, and subfield). All measures are log-transformed (after adding 1 for zero values). Models control for publication year and subfield fixed effects, abstract availability, text length (unique words and phrases in title and abstract), and number of unique papers and journals cited. AUC represents the area under the ROC curve. Marginal effects show the percentage increase in the likelihood of being a Nobel Prize paper associated with a one-standard-deviation increase in the metric.

[13] The replication data and code of Arts et al. (2025) are retrieved from https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/OUNQTQ

The results show that applying stricter cutoffs significantly improves predictive validity. For Nobel Prize papers, precision[14] rises from 69.9% for *New Phrase (All Scores)*, to 71.4% for *New Phrase (Score 2 and 3)*, and to 74.0% for *New Phrase (Score 3)*, with the AUC increasing from 0.76 to 0.77 and 0.79 respectively. As expected, excluding phrases with a score of 1 primarily filters out false positives and improves precision without substantially affecting false negatives or recall. Weighting noun phrases by their reuse in subsequent papers—capturing the impact and influence of new ideas on future scientific discourse—strengthens these effects further as illustrated in Panel B of Table B.3. In addition, the marginal effect—calculated as the percentage increase in the probability that a paper is a Nobel Prize paper associated with a one-standard-deviation increase in the respective metric—grows steadily across cutoffs, from 20.9% for *New Phrase (All Scores)* to 21.9% for *New Phrase (Score 2 and 3)*, and to 24.6% for *New Phrase (Score 3)*. When weighting phrases by their reuse in subsequent papers, marginal effects become even larger, reaching 31.0% for *New Phrase Reuse (Score 3)*.

[14] Precision is the proportion of true positives among all predicted positives (i.e., the share of papers classified as Nobel Prize papers that are actually Nobel Prize papers). Recall is the proportion of true positives among all actual positives (i.e., the share of Nobel Prize papers correctly identified). AUC (area under the ROC curve) summarizes overall predictive accuracy by capturing the trade-off between true positives and false positives across thresholds.