探索产业界与学术界的新颖性差异：知识实体中心视角

Hongye Zhao [a]，Yi Zhao [b]，Chengzhi Zhang [a, *]
[a] 南京理工大学信息管理系，中国南京，210094
[b] 安徽大学管理学院，中国合肥，230601。

摘要：学术界和产业界在推动技术进步方面各具优势。学术界的核心使命是推动研究成果的开放传播并促进学科进步。产业界重视知识可占有性和核心竞争力，但也积极参与学术会议、平台共享等开放实践，由此形成知识战略悖论。高度新颖且公开可获取的知识是推动技术进步的动力。然而，产业界还是学术界能够产出更具新颖性的研究成果，目前仍不清楚。一些研究认为，学术界往往更容易产生新颖思想；另一些研究则指出，产业界研究人员更有可能推动突破。以往研究受限于数据来源以及新颖性衡量方式不一致等问题。为弥补这些不足，本研究采用四种类型的细粒度知识实体（方法、工具、数据集、指标）进行分析，在统一语义空间中计算实体之间的语义距离，以量化新颖性，并实现不同类型文献的新颖性可比。随后，构建回归模型，分析产业界与学术界在出版物新颖性方面的差异。结果表明，学术界产出的新颖性更高，这一点在专利中尤为明显。在实体层面，学术界和产业界的论文都强调方法驱动的进步，而产业界在数据集方面具有独特优势。此外，学术界与产业界合作对提升研究论文的新颖性作用有限，但有助于提高专利的新颖性。该框架克服了文献类型的局限，为比较学术界与产业界的新颖性提供了一种具有普适性的工具。我们在 https://github.com/tinierZhao/entity_novelty 发布了数据及相关代码。

关键词： 细粒度知识实体；文献新颖性；统一语义空间；可占有性与开放性

引言

学术研究聚焦于理论探究和基础科学的发展，旨在拓展人类知识并推动学科进步（Merton，1973；Sauermann 和 Stephan，2010）。相比之下，工业部门注重核心竞争力（Geisler，1995），优先追求经济回报，并且往往通过限制研究成果的披露来维护知识产权可占有性（Chirico 等人，2020）。

按照这一逻辑，产业界通常会选择限制新颖研究成果的披露，以维护其竞争优势。然而，这种传统观念正受到人工智能（AI）领域的挑战。尽管产业界在机构定位上不同于学术界对开放知识共享的倡导，但仍在保持核心竞争力的同时积极探索开放性（Zhao 等人，2026）。例如，领先的科技公司已发布算法和模型领域的前沿技术，如 BERT 模型（Devlin 等人，2019）以及各种其他开源大语言模型。此外，它们还通过提供应用程序编程接口（API）和详尽的技术文档，大幅降低了采用人工智能技术的门槛。而且，工业部门积极参与最活跃、最受欢迎的人工智能会议，催生了众多颠覆性创新（Liang 等人，2024）。尽管这些举措降低了知识可占有性，却带来了显著收益：吸引开发者社区以降低维护成本，同时通过技术服务创造收入；借助合作获取外部知识，以保持技术领先地位；通过知识溢出促进产品优化（Jiang 等人，2024），并扩大市场份额。

这引出了一个尚未解决的问题：产业研究成果的新颖性是否低于学术界？目前，学者们对此仍存在分歧。一些学者认为，学术界贡献了更多新颖观点，而产业界往往采用并改进学术界的成果（Bikard & Marx，2020年）。后续研究进一步证实了学术界在人工智能创新方面的领先地位（Liang 等人，2024年）。然而，Dwivedi 等人

* 通讯作者。电子邮箱：zhangcz@njust.edu.cn。

（2021）指出，产业界研究人员更有可能推动人工智能新技术的发展。Transformer 和 GPT 等预训练模型的兴起，以及 ChatGPT 等大规模语言模型的快速发展，Ahmed 等（2023）凸显了产业界在计算资源、数据和人才方面的优势。作为人工智能的一个重要分支，自然语言处理领域见证了学术界和产业界涌现出大量突破性成果，因此成为探讨这一问题的典型场景。

专利和论文是学术界与产业界的主要科技成果。专利承载着技术与商业知识，对专利进行文本分析是技术开发与管理的重要工具（Arts 等，2021；Lee 等，2009）。论文是基础研究的核心，也是学术交流的关键媒介（Ba 等，2024）。本研究通过专利和论文衡量产业界与学术界的新颖性。然而，评估科技文献的新颖性本质上颇具挑战。多项研究证实，在同行评审过程中很难识别出创新性强的研究（Koppman & Leahey，2019；Liang 等人，2023；Riera & Rodríguez，2022；Wang等人，2017）。此外，目前人工智能领域关于新颖性的研究主要聚焦于论文（Chen 等，2024；Liang 等人，2024），未能全面考量专利等各类科技文献。这一研究局限并非源于数据可用性问题，而是因为专利与科学论文在评估新颖性时采用的方法不同。对于科学论文，新颖性通常通过期刊引文对分析来衡量（Lee 等，2015；Uzzi 等，2013；Wang等人，2017）。此外，专利中常用的分类代码无法与科学论文所使用的分类代码对应（Verhoeven 等，2016），因此无法将二者统一起来。另外，一些研究尝试基于句子级语义进行评估（Jeon等人，2022、2023），但句子语义向量容易受到学术写作风格的影响，而专利与论文的写作风格差异显著，因此无法适用于这种方法。

进一步聚焦文献中的知识元素，通过细化其表征粒度，并依托词嵌入模型构建的统一概念空间，可以实现测量的精确对齐。正如 Aceves 与 Evans（2024年）所指出的，概念与概念空间是众多组织理论发展的重要知识基础。即使知识在不同情境中以不同形式呈现，也可以通过统一词嵌入模型量化它们之间的语义联系与差异（Aceves 与 Evans，2024年）。

本研究采用统一的新颖性评估框架，利用细粒度知识实体评估学术界、产业界及二者在自然语言处理领域的合作所产生的出版物的新颖性，以弥补这一研究空白。具体而言，我们将这些细粒度知识实体映射到统一的概念空间（Aceves 与 Evans，2024年），计算细粒度知识实体之间的语义距离，并评估不同实体组合的难度。

具体而言，我们探讨以下三个研究问题：

研究问题1：如何为论文和专利统一基于细粒度知识实体的新颖性计算方法?

研究问题2：不同机构和文档类型中的实体，其新颖性表现有何差异?

RQ3：产业界与学术界的科技文献新颖性是否存在差异?

本文的贡献如下：

首先，本研究从专利和论文中提取知识实体，并将二者映射到由 SciBERT 构建的统一语义空间中。通过计算实体对之间的语义距离来量化组合难度，该框架在实体层面保留了细粒度的可解释性，同时为跨文档的新颖性比较奠定了基础。

其次，现有研究受限于依赖单一类型的文献，所得结论仍不一致。本研究将顶级会议论文与美国专利商标局（USPTO）专利数据集相结合，以实现对研究假设的可靠验证。

本研究所使用的代码和数据已在 GitHub 上开源，可通过以下网站访问：https://github.com/tinierZhao/entity_novelty。

相关工作

针对本文提出的研究问题，我们回顾了有关新颖性度量及影响新颖性的因素的科技文献。

科技文献中的新颖性度量

新颖性测量有助于提前识别有价值的创新，并为技术转移与创新提供关键见解。目前，新颖性主要通过组合来测量，正如

Nelson和Winter（1985）指出：“新颖性的产生主要涉及对现有概念材料和物质材料的重组”。传统的新颖性测量方法包括利用期刊对和分类代码对来评估文献的新颖性。随着大规模数据的日益丰富，以及机器学习和自然语言处理技术的发展，新颖性测量方法不断创新。其他类型知识元素的组合逐渐成为评估新颖性的重要方法。此外，一些研究将新颖性视为二元分类任务，采用分类或异常值检测方法来区分新颖与非新颖文献，从而探索新的途径。

从基于组合的视角来看，早期方法主要关注引文参考文献和分类代码。Uzzi 等（2013）比较了期刊对的观测频率与蒙特卡洛模拟频率，并计算每对期刊的 z 分数；其中，最低的第10百分位数 z 分数用于表示论文的新颖性，中位数 z 分数则用于表示其常规性。Lee 等（2015）通过采用多年时间窗口改进了 Uzzi 的方法，降低了计算难度；该方法将此前的单年窗口扩展为多年窗口，并计算引文对的常见程度。Wang等（2017）通过测量论文中不同引文期刊对首次组合的情况来衡量新颖性。具体而言，他们为期刊构建了共引矩阵，并利用各期刊向量之间的余弦相似度评估组合期刊对的难度。

关于专利新颖性的测量，早期传统方法主要关注专利分类代码和后向引文（Lee & Lee, 2019；Verhoeven 等，2016）。然而，引文仅描述现有技术，无法反映专利本身的技术，往往呈现出不完整且有偏差的表征（Arts 等，2021）。通过专利国际专利分类代码测量技术新颖性（Fleming, 2001）过于宽泛，且往往衡量的是跨学科性，而非技术不确定性。

随着自然语言处理技术的不断发展，科学术语提取（实体、关键词）和语义嵌入等任务日趋成熟，基于科学文本内容衡量新颖性已成为更合理、更有效的方法。Liu等（2022）使用BioBERT模型计算生物实体的语义，并根据语义相似度确定实体对新颖性。每篇论文的新颖性得分通过新颖实体对占所有可能实体对的比例计算。类似地，Chen 等人（2024）采用基于实体相似度的方法评估自然语言处理领域会议论文的新颖性。Luo等（2022）利用BERT词嵌入，通过评估研究问题、方法及其组合的新颖性来衡量新颖性。Arts 等（2021）从专利标题和摘要中提取关键词，计算“new_ngram”及相应的“new_ngram_reuse”来衡量专利新颖性。Wei等（2024）使用BERT模型从专利权利要求中提取创新句子，并将其提炼为知识元素三元组，再通过将实体和关系投射到共同空间中来衡量这些三元组的新颖性得分。Shi和Evans（2023）利用超图嵌入捕捉高维内容关联，以衡量研究新颖性。他们的研究发现，与传统的跨学科团队相比，突破更常出现在“知识探索”中。

从二元分类的角度来看，Jang等（2023）将专利新颖性视为一项分类任务，使用 RoBERTa 对专利权利要求进行语义嵌入，并据此开发了自解释新颖性分类模型。Jeon等（2022）对专利权利要求进行嵌入，并使用局部离群因子（LOF）算法计算专利新颖性。他们的研究表明，尽管 ELMo 和 BERT 能够提供高质量的专利嵌入向量，但与 Doc2Vec 相比，它们不太适合对专利的技术特征进行建模，尤其是在单一技术领域中。X. Liu等（2025）使用 Doc系列模型和 LOF 构建了无监督学习框架。Jeon等（2023）使用生物医学领域的论文标题训练 FastText 模型，并应用 LOF算法衡量每篇论文的新颖性得分。

从上述研究可以看出，新颖性测量方法已从早期依赖引文和分类代码的方法，发展到基于文本内容分析的方法。然而，新颖性评估仍存在不足。首先，早期基于引文（Lee等人，2015；Uzzi，2013；Wang等人，2017）和专利分类代码（Fleming，2001；Lee与Lee，2019；Verhoeven等人，2016）的方法存在固有缺陷：引文指标会受到自引偏差和不完整记录的影响，还可能将“跨学科性”与“新颖性”混为一谈（Fontana等人，2020）；分析大规模数据的计算成本也不断增加，效率低下。专利分类代码粒度过细，难以与论文对应。尽管近年来基于文本的方法提升了测量的粒度和可解释性，但现有研究大多聚焦于单一领域（论文或专利），缺乏统一评估二者新颖性的框架。

在接下来的章节中，我们将从论文和专利中统一提取细粒度知识实体，并将其映射到基于词嵌入构建的共享概念空间，再通过计算语义距离来评估新颖性。该框架不仅保留了实质性方法的可解释性和细粒度优势，还为论文与专利之间的跨领域新颖性比较提供了统一依据，弥补了当前评估工具的不足。

影响科技文献新颖性的因素

以往研究从多个角度探讨了新颖性的影响因素，包括机构性质、团队规模以及团队中的作者属性。

关于团队规模，现有研究的结论并不一致。Uzzi 等（2013）发现，与单作者论文相比，研究团队更有可能在熟悉的知识领域内引入新颖组合。Lee 等（2015）发现，团队规模与新颖性之间存在倒U形关系，这种效应在很大程度上由团队规模与知识多样性的相互作用所驱动。Wu 等（2019）指出，小型团队更有可能以新思想颠覆科学与技术，而大型团队往往专注于已有思想。Shin 等（2022）利用 Web of Science 数据发现，科学合作会对新颖性产生负面影响，因为合作研究往往局限于既有领域。然而，Wu 等（2024）认为，协作能够促进信任并提升解决问题的能力，而知识多样性则有助于加强知识转移，推动科学对技术产生影响。相反，一些研究表明，过度的团队异质性可能会削弱信任、阻碍知识共享，并妨碍创新（Chen 等人，2015）。

在团队内部的作者属性层面，拥有多元化专业知识的团队往往能产出更多原创作品，并在影响力方面具有长期优势（Zheng 等人，2022）。Mori 和 Sakaguchi（2020）利用日本专利考察了发明人之间知识差异如何提升专利新颖性。近年来，团队内部的性别多样性也成为备受关注的议题。与单一性别团队相比，具有性别多样性的团队所发表的论文新颖性更高、影响力更大（Yang 等人，2022）。Liu 等（2024）探讨了博士论文的新颖性与性别异质性之间的关系，发现女性作者的平均新颖性得分低于男性作者；男性导师更有可能指导撰写出新颖性较高论文的学生。值得注意的是，这种性别差异在声望较低的大学中更为明显。同样，Chan 和 Torgler（2020）发现，在精英科学家中，女性科学家往往比男性同行获得更多引文。

在机构层面，学术界在论文层面的新颖性方面往往领先于产业界，能够提出更多探索性思想；而产业界更有可能产出高影响力论文（Liang 等人，2024）。Chen 等人（2024）测量了自然语言处理领域的新颖性，发现根据知识实体的细粒度组合，学术界和合作机构往往比产业界更具新颖性。其他研究表明，涉及公司的论文影响力更大，而产业界与学术界之间的合作则展现出更高的新颖性（Jee 和 Sohn，2023）。

Chen 等人（2024）与本研究有相似之处，均探讨了自然语言处理领域的新颖性。然而，本研究进一步纳入专利数据，对比学术界与产业界在两类文献中的新颖性表现。需要指出的是，这并非单纯的数据扩展，而是以统一的自下而上的概念空间为基础，利用词嵌入模型，确保不同类型科学文献中实体之间的语义距离具有可比性（Aceves & Evans，2024）。这种方法适用性更强，不受文档类型和写作风格的限制，并且具有更强的可扩展性。与仅关注实体频率分布和类型占比的 Chen 等人（2024）相比，本研究进一步分析了不同创新类别中各机构语义距离变化的趋势。此外，我们还全面探讨了“创新中随组成部分数量增加而出现的边际收益递减”问题（Fleming & Sorenson，2001），而 Chen 的研究未曾探讨这一问题。我们的研究结果也与他们有所不同：我们发现，学术界与产业界合作对提升论文新颖性的影响有限，这与 Liang 等人（2024）的结论一致。

本研究考察了不同机构类型在专利和论文中的新颖性表现。目前，针对机构类型实质层面的研究仍然较少，且大多局限于单一数据集。尽管学术界和产业界分别产出了大量论文和专利，但鲜有研究采用比较框架，跨多种科技文献评估这两个领域的新颖性。

方法论

本研究旨在量化不同团队构成对科技文献新颖性的影响表现。图1所示的研究框架概述了四个关键阶段：

阶段1. 数据集构建：我们构建了一个原创数据集，涵盖自然语言处理领域的科技文献，包括2000年至2022年间发表的论文和专利。

阶段2. 实体提取：从科技论文和专利中提取细粒度知识实体，并将科技论文中的知识迁移至专利。

阶段3. 新颖性测量：采用统一的预训练模型获取每个提取出的知识实体的语义向量。根据实体之间的语义距离测量这些实体组合的难度（Chen 等，2024；Liu 等人，2022年、2024年）。随后，我们使用这种基于语义距离的方法评估每份文档的新颖性。

阶段4. 回归建模：对标准化数据进行联合回归分析，以揭示不同文献类型的整体新颖性趋势。随后，分别对专利和论文进行回归分析，作为异质性检验，以考察观察到的趋势是否因文献类型而异。

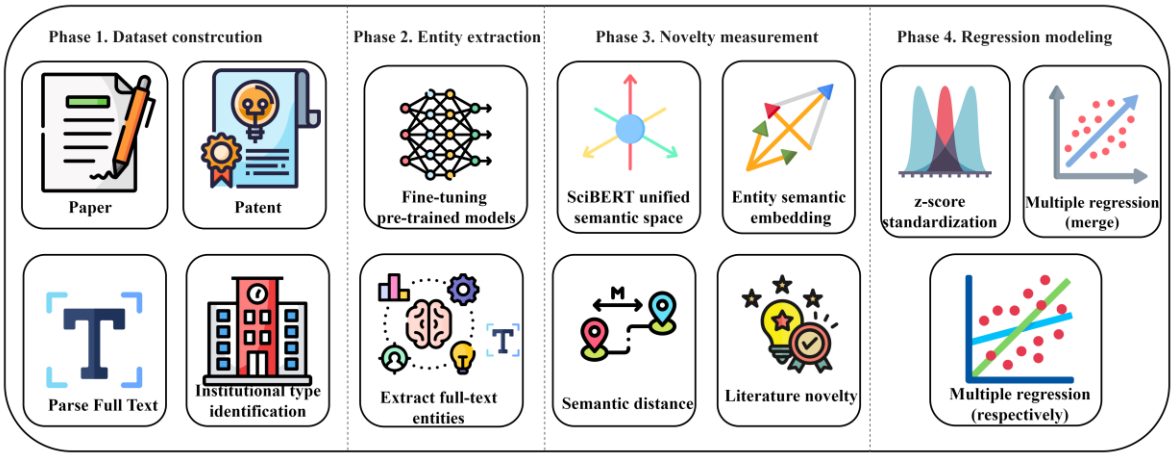

图1. 本研究的框架

数据收集

本研究的论文数据集和专利数据集均源自统一的自然语言处理领域。论文数据收集自 ACL Anthology[1] 网站。我们选取了三个具有代表性的会议进行研究：ACL（计算语言学协会年会）、EMNLP（自然语言处理实证方法会议）和 NAACL（计算语言学协会北美分会）。共收集了2000年至2022年的 17,783 篇全文论文。

专利数据通过 PatSnap[2] 系统从美国专利商标局（USPTO）收集。我们检索了 2000年至2022年期间的专利，使用以下查询式：CPC:(G06F40*[3]) AND APD:[20000101 TO 20221231] AND COUNTRY:US。我们重点关注发明专利，并筛除了法律状态为撤回、驳回、放弃、申请终止或完全无效的专利。此外，同一优先权的专利被合并为专利族。最终，共获得 25,305 项专利。

机构类型分类

对于论文中作者机构类型的分类，我们以 Chen 等人（2024）的工作为基础。他们人工整合了全球研究识别数据库（GRID）[4]。我们进一步通过结合人工标注与 OpenAlex 查询，补充了数据库中论文缺失的机构数据，

[1] https://aclanthology.org/ [2] https://www.patsnap.com/ [3] 合作专利分类：G06F40，处理自然语言数据 [4] 全球研究识别数据库（GRID）将机构类型分为八类：政府、教育、公司、设施、医疗保健、非营利组织、档案机构和其他。GRID 收集全球各地的机构名称，并利用维基百科（www.wikipedia.org）和 GeoNames（www.geonames.org）补充机构数据，例如所在国家和城市。

旨在解决论文机构信息缺失的问题（Zhang、Cao 等，2024）。需要指出的是，OpenAlex 采用与 GRID 相同的机构分类标准，两者均基于研究机构注册系统（ROR）[5] 。该系统包含以下机构类型：教育、公司、设施、非营利组织、政府、医疗保健、档案机构和其他。最后，我们补全了 17,783 篇论文的作者信息及其对应的机构隶属关系。按照以往研究的做法，我们将每位作者列出的第一家机构视为其主要隶属机构（Hottenrott 等，2021）。

在机构分类层面，鉴于本研究旨在探讨学术部门与工业部门之间的新颖性差异，我们重点关注两类机构：学术机构和产业界机构。具体而言，将归类为“公司”的机构划入工业部门，而将归类为“教育”的机构划入学术部门。此外，根据现有研究的定义标准（Chen 等，2024年；Xu 等，2022年），医疗保健类机构的出版物主要以研究为导向，因此我们将教育和医疗保健机构归入学术类别。至于其他类型的机构，本研究不涉及，因此予以忽略。

具体而言，如果一篇论文的所有作者均隶属于学术机构（如大学或研究所），则将其归类为“学术”；如果所有作者均隶属于产业界机构（如公司或企业），则归类为“产业界”；如果论文作者同时来自学术界和产业界，则归类为“合作”。如果论文既没有产业界作者，也没有学术界作者，则归类为“其他”。

专利数据处理首先从数据库中提取标准化申请人信息，其中所有非个人名称均以中文或英文呈现。随后，结合本地词典，使用编辑距离算法（Levenshtein，1965）对机构名称进行规范化。根据词汇特征，定义了两组关键词：一组用于学术机构，另一组用于工业组织，涵盖中英文词汇。该算法将包含教育相关词汇（如“edu”“univer”）的机构归类为学术机构，将包含公司相关词汇（如“inc”“ltd”“lp”）的机构归类为工业组织。由于数据库提供了标准化申请人字段，该方法能够确保效率和准确率。对于无法识别的机构，使用 spaCy[6] 命名实体识别来判断申请人是否为“个人”。对于出现次数超过两次的个人申请人，我们使用 ChatGPT 进行验证，以检查是否遗漏了产业界或学术界的分类（提示词见附录表 5）。经 ChatGPT 验证确认为个人申请人的，将归类为“个人”。对于无法通过上述流程归入学术机构、工业组织或个人类别的所有未识别机构，均定义为“其他”。最后，人工审查结果，并对算法输出进行修正和补充。

论文和专利的具体机构分布如表1所示。

表1. 科技文献的机构分布

机构类型	数量	比例（%）	数量	比例（%）
	论文		专利	
学术界	11,965	67.28	470	1.86
产业界	1,409	7.92	21669	85.63
合作	3,652	20.54	71	0.28
个人	0	0	2932	11.59
其他	757	4.26	163	0.64

知识实体抽取与规范化

本研究借鉴组合新颖性的经典理论（Fleming，2001年；Uzzi 等人，2013年），基于知识组件的非典型组合来衡量新颖性。非典型程度可通过计算语义距离量化，这一方法被广泛应用于文献计量学（Chen 等人，2024年；Liu 等人，2022年、2024年）。鉴于大多数自然语言处理研究通常包含以下关键要素：1）数据集构建或选择，通常涉及语料库和词典等文本资源，这些资源是模型训练和验证的基础；2）方法选择与应用，界定解决问题的策略和步骤；3）评估指标的选择

[5] https://ror.org/ [6] https://pypi.org/project/spacy/

，用于衡量模型性能和任务质量；4）工具的使用，包括实施和测试自然语言处理方法所需的编程语言、软件和开源工具（Pramanick 等，2025年；Zhang、Zhang 等，2024年）。基于这一框架，我们从每项专利和论文中提取细粒度知识实体，涵盖方法、工具、指标和数据集类别。

在细粒度知识实体识别任务中，我们使用了预训练的 SciBERT 模型。由于专利和论文在写作风格与文本结构方面存在差异，我们针对每种文档分别训练了实体识别模型。具体而言，对于论文，我们采用了 Zhang 等（2024）提出的框架。对于专利，我们首先使用预训练模型标注专利文本，随后根据标注规则对提取出的实体重新标注。此外，我们还对专利文本中的独有实体（如存储介质）进行了额外标注。经过多轮迭代和调整，我们得到了专利实体识别模型（SciBERT + 条件随机场），其性能如下：精确率为 78.83%，召回率为 82.51%，F1 分数为 80.63%。考虑到仅从标题和摘要中提取实体会遗漏许多实体，我们对专利和论文都进行了全文提取。论文数据从 PDF 文件中提取，专利数据库也以全文形式导出。在实体规范化方面，我们使用编辑距离（Levenshtein，1965）对实体进行聚类。最终，我们在论文中识别出 37,236 个实体，在专利中识别出 9,523 个实体。

附录表 4 展示了专利和论文各类别中的前五个实体。由于专利很少在公开数据集上进行评估，因此专利中“数据集”实体的占比较低，导致对这类实体的识别效果较弱。此外，专利术语的一个显著特点是抽象程度较高，这一点在权利要求书部分尤为明显。一般论述依靠精确措辞准确传达内容，并避免使用含糊或过于宽泛的术语；与之不同，专利权利要求会有意采用概括性词汇（Codina-Filbà 等，2017）。这种策略使公司能够扩大知识产权保护范围，从而确保其创新获得更广泛的排他性保护（Ashtor，2022）。此外，专利中对工具实体的描述往往也较为概括，反映了这一特点。

基于语义距离的新颖性测量

我们通过分析从科技文献中提取的细粒度知识实体，探讨实体组合的新颖性。我们借鉴了 Liu 等（2022，2024）在生物医学论文科学新颖性评估领域的研究。他们将生物实体视为研究方法的核心要素，并使用预训练的 Bio-BERT模型量化这些实体之间的语义距离，以测量新颖性。我们将这一方法应用于自然语言处理领域的论文和专利新颖性评估，并使用预训练的 SciBERT 计算实体的语义相似度，以测量新颖性。

具体而言，在使用 SciBERT 进行嵌入处理时，我们分析其内置的 [CLS] 标记。该标记作为句子的起始点，并聚合整个句子的上下文信息，因此可作为整个句子的向量表示。基于这一特性，我们将专利实体短语视为短句，并通过输出 [CLS] 位置的向量来获取专利实体的语义向量表示。为具体说明语义向量表示，我们以自然语言处理核心实体“transformer”为例：我们以标准化格式将该实体输入 SciBERT，模型通过 [CLS] 标记输出一个 768 维稠密向量。该向量的前 5 个维度如下所示： [3。70797575e-01、1.42690563e+00、-8.87082458e-01、-5.18282456e-03、5.78712046e-01，……（后续 763 个维度省略）]

对于实体对（e_i, e_j），两者之间的距离记为 D，而 $cosine(e_i, e_j)$ 表示实体之间的语义相似度es.

$$D(e_i, e_j) = 1 - cosine(e_i, e_j) \quad (1)$$

我们将新颖性排名前10%的实体标记为高新颖性实体。最后，我们分析了这些高新颖性实体在文本中的频率，并根据其在所有实体组合中的占比衡量每篇论文的新颖性。

新颖性差异分析

为探究不同机构之间的新颖性差异，本研究采用回归分析，对不同机构所产出的科技文献所体现的新颖性进行量化和比较。以下各节将详细介绍变量选择过程以及回归模型的构建。

因变量：在自变量设定中，我们首先使用上一节计算的连续新颖性指标（新颖性得分）进行分析。该得分衡量每篇论文或每项专利中新颖实体所占的比例，取值范围为0至1，得分越高表示新颖性越强。它

有助于比较产业界与学术界在新颖性产出均值上的差异。其次，为应对新颖性结果可能存在的不确定性并验证结论的稳健性，我们将每年按该得分排名前10%的论文/专利定义为“高新颖性”，并构建一个二元变量（新颖性得分前列）（Jeon 等，2022），排名前10%的取值为1，其余取值为0。该变量主要用于分析哪类部门更有可能产生高新颖性论文/专利。

自变量：本研究将自变量定义为机构类型。剔除归类为“其他”和“个人”的机构后，其余机构分为三类：学术界、合作和产业界。具体而言，定义了两个二元变量（学术界和合作）。如果文献来自学术机构，则学术界变量取值为1；如果文献来自合作机构，则合作变量取值为1；如果文献来自产业界，则这两个变量均取值为0。

控制变量：为排除其他因素的干扰，并识别机构类型对新颖性的净效应，本研究纳入了多个控制变量。首先，考虑了与团队相关的变量，包括机构数量（机构数量）以及作者或发明人数量（作者/发明人数量）。这是因为规模较大的研发团队通常拥有更广泛、更丰富的知识基础，团队成员之间的协作也能增加跨学科知识整合的机会（Wu et al., 2025）。此外，这类团队往往会开展开发性研究（Wu et al., 2019），其特征可能影响新颖性。对于专利，本研究还纳入了专利族规模（家族规模），该指标通常与福利价值和技术影响相关（Kabore & Park, 2019）。此外，为体现专利知识组件的多样性，本研究控制了小组层级的IPC分类代码数量（IPC 数量）（Sun et al., 2022）。同时，鉴于本研究通过实体组合计算新颖性，也将实体数量（实体数量）纳入控制变量，以排除实体数量的干扰。此外，本研究引入年份虚拟变量，以控制可能影响结果的年度差异。对于连续型控制变量，参照Wu et al.（2024, 2025）的方法，采用对数转换处理其偏态分布。

进行回归分析时，我们仅保留了来自产业界、学术界以及学术界—产业界合作的科学文献。此外，我们排除了异常值：实体少于五个且新颖性得分为零的条目。最终数据集包含16,295篇论文和20,934项专利。变量的描述性统计以及变量之间的相关系数分别列于表2和图2中。

我们发现，因变量（新颖性）的连续形式与离散形式之间存在较强的相关性，而自变量与因变量之间的相关性较弱。随后，我们计算了所有解释变量的方差膨胀因子（VIF），以评估多重共线性。论文的 VIF 为 2.30，专利的 VIF 为 1.06，均低于 5 的阈值（Marcoulides 和 Raykov, 2019）。这些结果表明，多重共线性对我们的模型影响甚微，因此估计结果具有可靠性。

表2。用于回归分析的变量描述性统计（N = **20**项专利，N =篇论文）

变量	论文				专利			
	Mean	标准差	*Min*	*Max*	*Mean*	标准差	*Min*	*Max*
新颖性得分	0.10	0.06	0.01	0.52	0.12	0.07	0.01	0.67
新颖性得分前列	0.10	0.30	0	1	0.10	0.30	0	1
IPC 数量	-	-	-	-	1.95	1.00	1	10
家族规模	-	-	-	-	2.08	1.94	1	82
作者/发明人数量	3.82	2.24	1	77	3.30	2.13	1	26
机构数量	1.84	1.24	1	44	1.05	0.42	1	15
实体数量	34.93	14.53	8	153	29.81	14.95	8	98
学术界	0.70	0.46	0	1	0.02	0.14	0	1
合作	0.22	0.41	0	1	0.00	0.05	0	1

注：论文未包含 IPC 编号或家族规模，以“-”表示。

随后进行多变量回归，以考察不同类型的机构如何影响文献的新颖性得分。

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

其中，$Novel_i$ 表示每篇文献i的新颖性得分。自变量$Academia_i$ 和$Cooperation_i$ 分别表示该文献是否来自学术机构或合作机构。变量控制变量包括一组控制变量，Y_i 表示发表年份，ε 表示模型中的误差项。为了进一步

为进一步完善分析，本研究构建了逻辑回归模型，以验证研究结论的稳健性 of 研究结 论，并分析机构类型是否与文献具有较高新颖性相关。

$$\ln\left(\frac{P(NS\ Top_i)}{1-\mathrm{P}(NS\ Top_i)}\right) = \alpha + \beta_1 Academia_i + \beta_2 Cooperation_i + Controls + Y_i \quad (3)$$

)年，需要指出的是，该公式中包含最后的残差项，因为逻辑回归使用最大似然估计（MLE）进行模型拟合，而非线性回归所采用的普通最小二乘法（OLS）。因此，误差项并未直接纳入模型表达式。

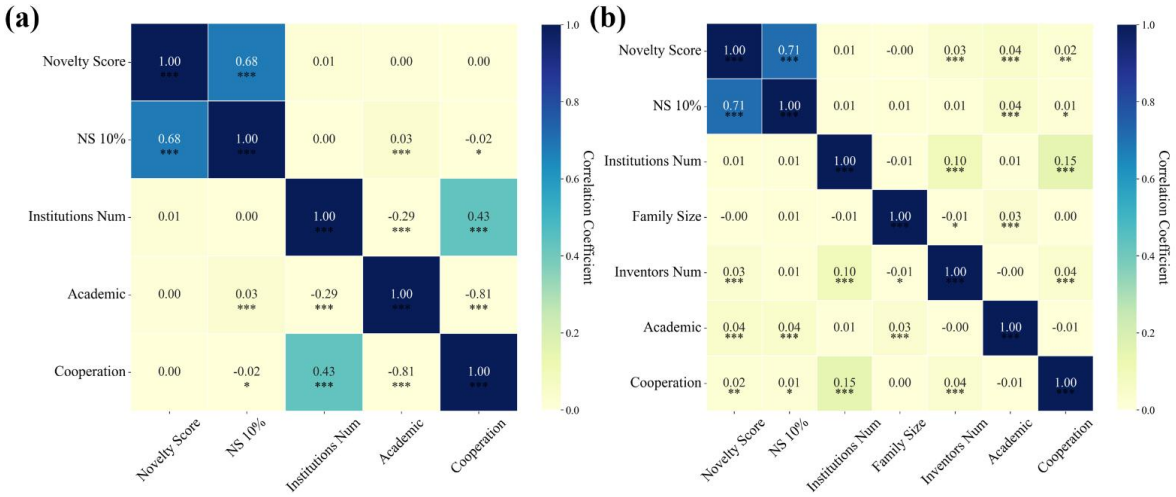

图2. 皮尔逊相关系数矩阵（a）论文中各变量之间的相关性（b）专利中各变量之间的相关性

然而，普通最小二乘法回归侧重于自变量对因变量均值的影响，而逻辑回归模型则聚焦于单一的极端结果。为全面考察机构类型如何影响整个新颖性得分分布中的新颖性，我们进一步采用了分位数回归。通过估计新颖性得分特定分位数处自变量的条件效应，该方法补充了上述两种模型。模型设定如下：

$$Q_\tau(Novel_i) = \alpha_\tau + \beta_{1\tau} Academia_i + \beta_{2\tau} Cooperation_i + Controls + Y_i \quad (4)$$

其中，Q_τ denotes the quantile function, represents the target quantile (with 0<τ<1), and $\beta_{1\tau}$ and $\beta_{2\tau}$ indicate the 分位数处对新 cates

从表2可以看出，专利和论文的新颖性得分分布尺度存在显著差异。直接进行合并分析可能会 一致 而导致结

$$Z_{\mathrm{Novelty},i} = \frac{Novelty_{i,\mathrm{raw}} - \mu_{\mathrm{Novelty}}}{\sigma_{\mathrm{Novelty}}} \sim \mathcal{N}(0,1) \quad (5)$$

其中 raw 表示该指标的均值， 表示总体标准差 ntity in the raw data, lt , the popula tio

结果

本研究分析了2000年至2022年间在三场具有代表性的会议上发表的论文，以及提交至USPTO的专利，重点关注三类出版机构之间的新颖性差异：

学术界、产业界和协作。我们的研究比较了不同机构类型在文献新颖性方面的表现，并探讨了团队规模与新颖性之间的关系。研究旨在揭示团队规模如何影响不同类型科技文献中的创新。

论文和专利发表数量的趋势

自然语言处理领域发展迅速，自2000年起，专利和论文数量每年稳步增长。2022年的专利数量略低于2021年，这是由于《美国发明法》（AIA）《美国法典》第35编第122(b)条规定，专利须在最早申请日期后的18个月公开，除非申请人要求提前公开。截至检索日期，部分2022年专利尚未公开，这种情况很常见。

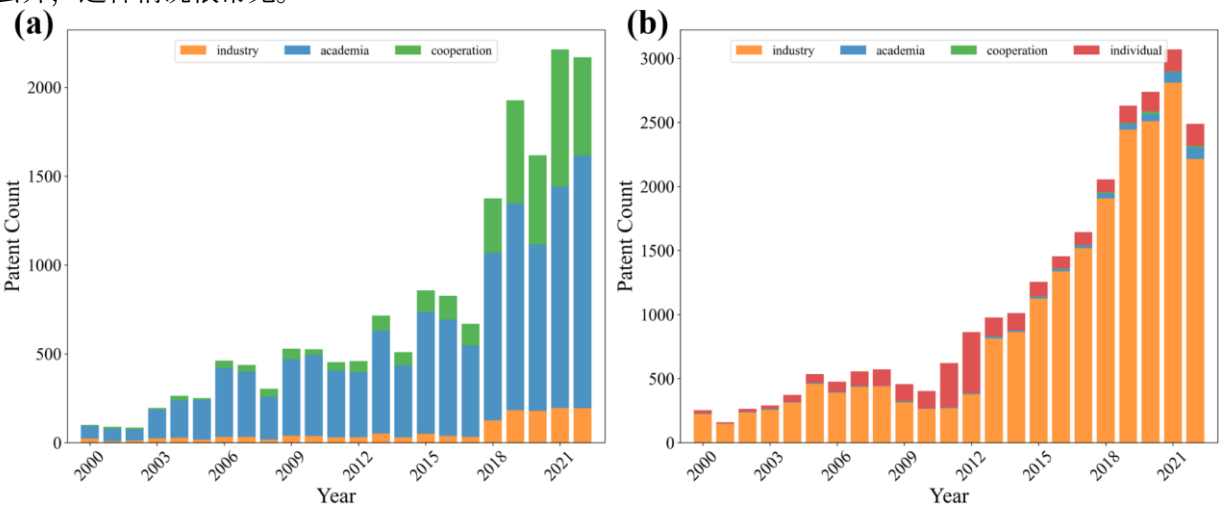

图3. 论文和专利的年度发表数量。（a）论文的年度发表数量；（b）专利的年度发表数量

此外，与论文相比，各机构的专利数量分布更不均衡，具体比例详见前文关于机构分布的章节。尽管美国汇集了全球顶尖的高等教育资源，且政府研究资助大多流向大学，但大学来源的专利占全国专利总量不到4%，多数专利由企业申请，其次是个人申请[7]。自然语言处理领域也存在这一现象：学术部门参与的专利申请数量远少于工业部门。论文和专利的年度发表数量见图3。

统一框架下的新颖性测量结果

在本节中，我们将探讨研究问题1。首先，我们使用前几章讨论的实体识别模型，从每篇论文和每项专利中提取细粒度知识实体。然后，我们利用预训练的 SciBERT 模型获取专利和论文中实体的语义向量。我们根据不同实体之间的语义距离评估其新颖性，实体语义距离新颖性的整体分布如图4所示。

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

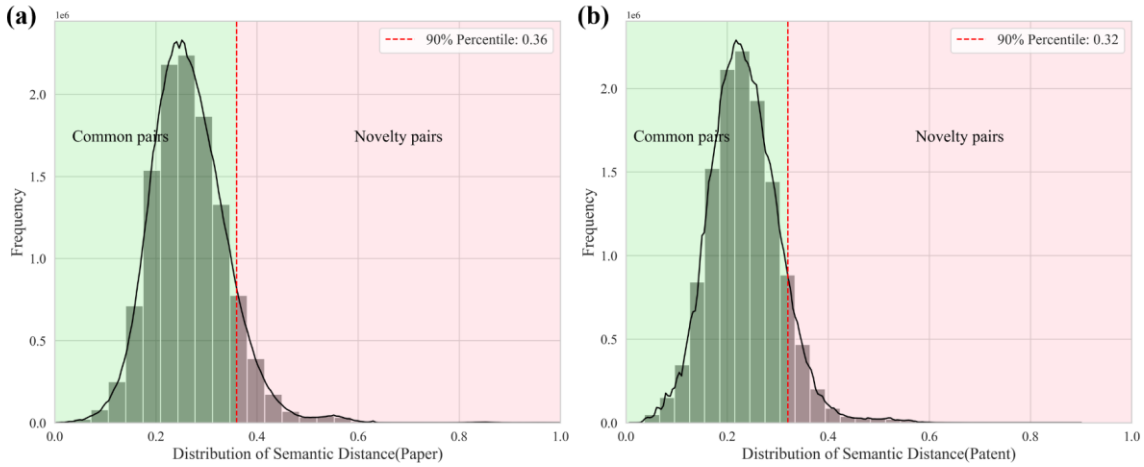

图4. 细粒度知识实体的语义距离分布：（a）论文实体的语义距离分布；（b）专利实体的语义距离分布。

这些语义距离将用于确定相似度。具体而言，我们根据语义距离将组合分为两类：新颖性配对（前10%）和常见配对。最终的新颖性取决于新颖性配对占配对总数的比例。

为进一步探究实体组合对专利和论文新颖性的贡献差异，我们分析了新颖性最高的前10%论文和专利中四种实体组合的平均语义距离。论文的结果见附录图8，专利的结果见附录图9。

不同实体类型的新颖性表现趋势

本节回答研究问题2。我们以实体组合的非典型程度作为衡量新颖性的核心指标，进一步分析了论文和专利中实体组合语义距离变化的趋势，如图5和图6所示。鉴于本研究聚焦于新颖性，本节仅分析附录图8和附录图9中的前6项排名。总体而言，随着领域不断发展，实体之间的语义距离呈现整体下降趋势。这一现象反映出专利和论文在利用先前知识方面的差距正在缩小，表明领域内的知识整合日益顺畅（Park等，2023）。

具体而言，对于被确定为该领域主要创新模式的方法-指标组合，图5显示，产业界与学术界在方法-指标实体组合的语义距离方面几乎没有差异。此外，两部门在方法-方法组合的语义距离方面也不存在显著差异。这些发现表明，产业界与学术界均聚焦于方法论层面的创新，并且在这一维度上，其创新路径始终保持高度的知识连贯性。

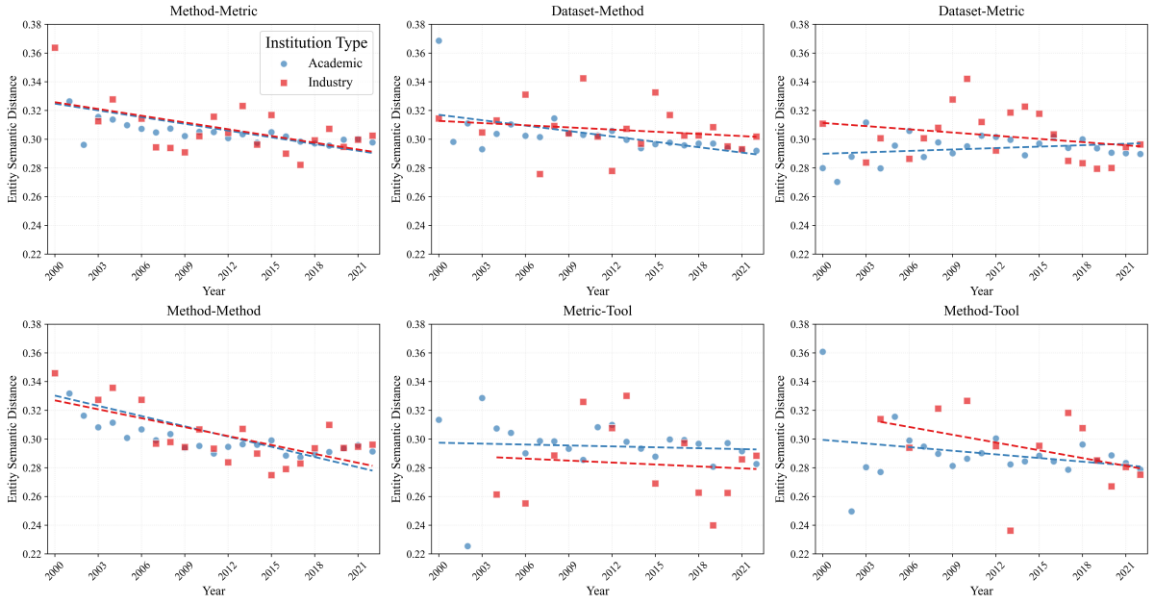

图5. 论文中不同实体组合的语义距离变化趋势

在数据集相关实体组合方面，工业部门的语义距离显著高于学术部门。这种差异凸显了产业界在数据集创新方面的独特性，而这一特点与其获取规模更大的数据资源密切相关（Ahmed 等，2023）。凭借这种数据规模优势，工业部门能够支持前瞻性研究。相比之下，在工具-指标组合方面，学术部门的语义距离高于产业界。产业界在这一类别中的语义距离较低，可能源于其更注重成熟技术的实际应用：这一侧重点要求工具-指标兼容性具备更高的稳定性，从而促成相对固定的关联模式。相反，学术研究不受实际实施场景的限制，因而倾向于探索工具与指标的创新组合。

随后，我们分析了专利中实体组合的语义距离趋势，结果如图6所示。由于学术领域的专利披露数量相对有限，某些实体组合在早期阶段并未出现。

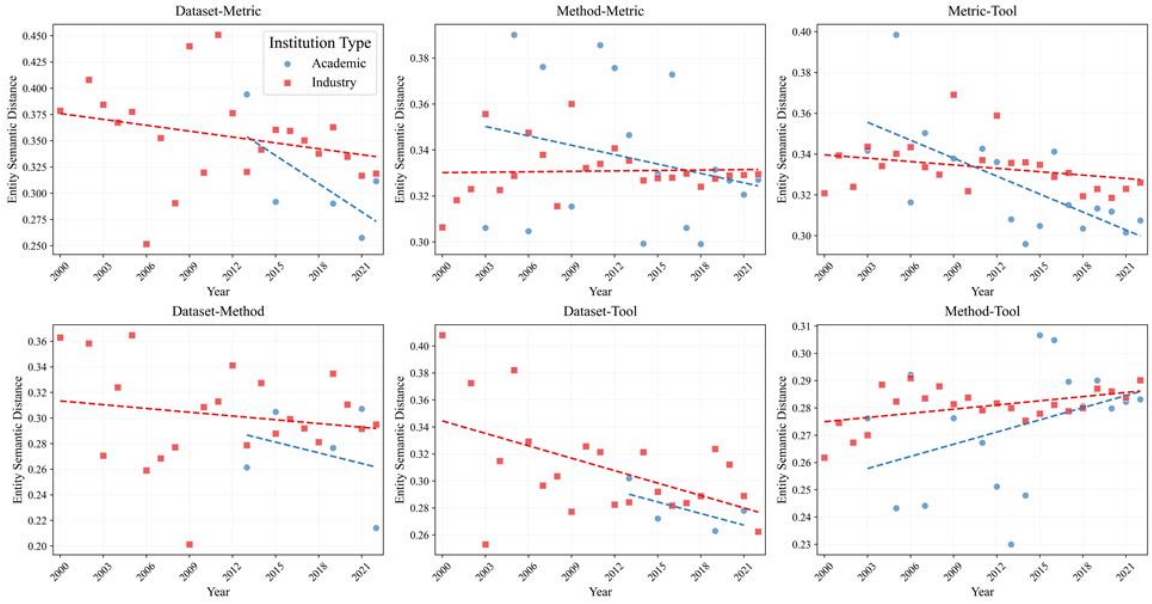

图6. 专利中不同实体组合的语义距离变化趋势

对于数据集相关实体组合（数据集-指标、数据集-方法和数据集-工具），分析发现了与论文研究中观察到的相似模式：在所有三种数据集相关

组合方面，工业部门的语义距离高于学术部门。论文和专利中呈现的一致趋势清楚地表明，工业部门不仅受益于有关数据集的论文，还受益于这些论文在专利技术开发中的应用。这表明，产业界在专利开发中探索着更具突破性的、由数据集驱动的创新路径。在方法实体方面，两个部门表现出明显差异。对于方法-指标组合，学术领域的语义距离高于工业领域。这一发现表明，学术领域在方法创新的验证方面具有相对优势。相反，在方法-工具组合方面，优势转向工业部门。产业界在方法-工具组合上的语义距离更高，反映出其将方法与实用工具相结合的能力更强。

回归结果：　　不同类型机构的新颖性差异　　机构

本节将回答RQ3。我们进行了一系列多元回归，以比较不同机构类型的专利和论文的新颖性。首先，我们展示专利和论文的汇总分析结果，然后分别分析专利和论文，以观察其异质性。此外，我们进一步引入分位数回归，以探究不同新颖性水平下的异质性效应。

表3展示了基于科技文献完整样本的回归分析结果。为消除专利与论文原始新颖性值分布之间的差异，我们首先使用公式4对新颖性值进行标准化处理，以进行z分数标准化。此过程使专利和论文的新颖性值处于相同的分布维度。

表3. 全部科学文献新颖性的回归结果

	新颖性得分（z分数转换后）		新颖性得分前列（是否位于前10%）	
变量	模型（1）	模型（2）	模型（3）	模型（4）
学术	0.127***	0.080***	0.518***	0.278***
	(0.026)	(0.025)	(0.091)	(0.087)
合作	0.109***	0.079**	0.347***	0.210*
	(0.029)	(0.032)	(0.105)	(0.110)
作者数量的自然对数		0.055***		0.081***
		(0.010)		(0.031)
自然对数(机构数量)		0.025		0.150***
		(0.018)		(0.056)
自然对数(实体数量)		-0.367***		-1.541***
		(0.013)		(0.037)
常数项	-0.067	1.061***	-2.389***	2.294***
	(0.081)	(0.089)	(0.222)	(0.248)
年份固定效应	Yes	Yes	Yes	Yes
类型固定效应	Yes	Yes	Yes	Yes
观测值	37,229	37,229	37,229	37,229
R平方	0.005	0.033	0.002	0.068

注：括号内为稳健标准误，*** p<0.01，** p<0.05，* p<0.1。在具体回归模型中，模型（1）和模型（2）采用标准

化的新颖性值z分数作为核心因变量，重点考察不同参与者对新颖性的影响力。回归结果表明，核心解释变量“学术”和“合作”的系数均通过统计显著性检验，且系数均为正。这一发现表明，在涵盖所有科技文献的分析框架内，学术部门直接参与或通过协作参与，均能提升成果的新颖性水平。

此外，逻辑回归结果表明，无论是在不含控制变量的模型（3）中，还是在含控制变量的模型（4）中，“学术”和“合作”的影响均具有统计显著性。具体而言，在含控制变量的模型（4）中，“学术”的系数为0.278，优势比为$e^{0.278} = 1.32$；“合作”的系数为0.210，优势比为$e^{0.210} = 1.23$。这表明，在控制其他变量后，学术机构参与使实现高新颖性的优势比提高32%，而学术界协作参与则使这一优势比提高23%。

在模型设定方面，本研究不仅控制了年份固定效应，以消除时间趋势对新颖性的潜在干扰，还充分考虑了样本中

专利与论文之间的固有差异。尽管上述 z 分数标准化和统一概念空间构建在一定程度上减轻了两种文档类型之间的异质性，为避免潜在的混杂因素，本研究还在所有回归模型中控制了文档类型固定效应。

论文与专利之间的新颖性异质性

随后，我们分析专利和论文的异质性。我们首先对不同类型机构的论文与专利之间的新颖性差异进行了初步分析，结果如图 7 所示。图 7 (a) 展示了论文中新颖性的分布。数值结果表明，学术论文的新颖性得分均值为 0.1013，而工业部门的均值为 0.0987，学术界与工业部门的协作论文均值为 0.1015。这三组的新颖性水平非常接近。值得注意的是，分布模式显示，学术论文的离散程度更高，其中包含少量极高的新颖性值。这表明，学术机构的论文可能具有更大的创新潜力，产出高度新颖的研究的概率更高。

图 7 (b) 展示了专利的结果。学术专利的新颖性得分均值为 0.1373，而工业部门的均值为 0.1146，表明学术机构在新颖性方面显著优于工业部门。协作类型的均值数量为 0.1427。与研究论文相比，学术机构在专利创新中产出更多高度新颖的成果，其新颖性得分均值显著高于工业部门。

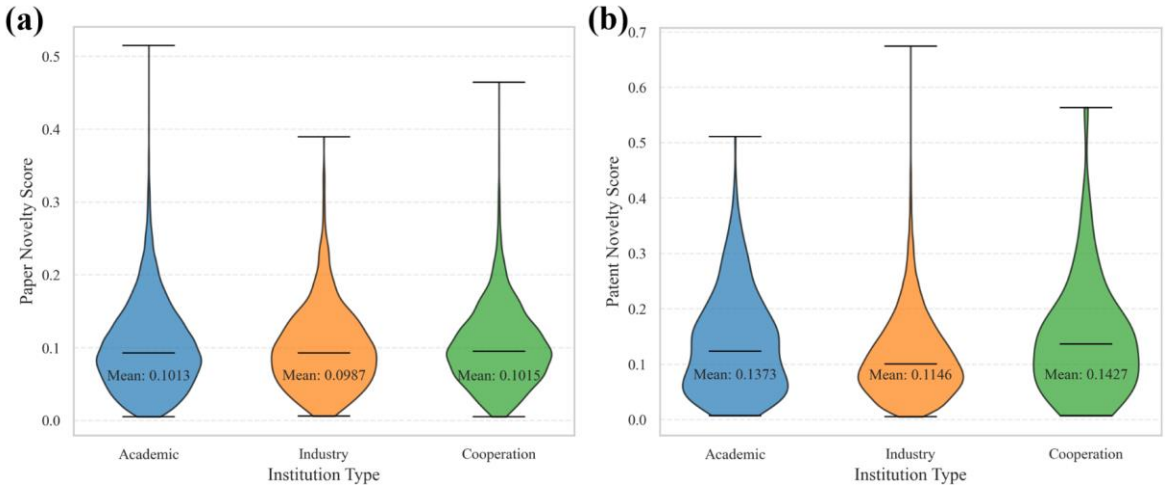

图 7. 新颖性分布的小提琴图。(a) 论文中不同出版机构之间的新颖性差异 (b) 专利中不同出版机构之间的新颖性差异

论文的进一步回归结果见附录表 6。在仅纳入机构类型的普通线性回归模型 (5) 中，“学术界”的系数为 0.003，在 5% 水平上具有统计显著性 (β = 0.003, p < 0.05)。这表明，学术论文的平均新颖性得分比产业界论文高 0.003。“合作”的系数同样为 0.003，但差异并不显著。这说明，尽管合作论文的新颖性数值上高于产业界论文，但这一差异缺乏统计显著性。在模型 (6) 中纳入控制变量后，“学术界”的系数降至 0.002，且不再显著 (p =0.168 > 0.1)。学术论文与产业界论文之间的新颖性差距不再具有统计显著性。

我们补充进行了分位数回归分析（结果见附录表 8），发现学术变量的系数随分位数的升高而增加，且具有统计显著性，这表明不同新颖性区间的表现存在差异：在 % 分位数（低新颖性）处，系数为（ β . .），与

= 0 002,0 001 25

产业界相比无显著差异。在 75% 分位数（高新颖性）处，系数为（β = 0.007,0.004），且具有显著性（p < 0. 1, p < 0.01）。这表明，学术界在常规研究中的表现与产业界相似，但在高新颖性研究中具有明显优势。

因此，我们进一步分析了论文成为高新颖性论文的可能性差异。逻辑回归模型（3）的结果显示，“学术界”对应的系数为 0.331（β = 0.331，p < 0.01）。将其转换为优势比（$OR = e^{0.331} \approx 1.$ 39）后可知，在不控制其他变量的情况下，学术论文成为高新颖性成果的可能性是产业界论文的 1.39 倍。模型（7）中“合作”对应的系数为 0.158，但仍不显著。这与相关研究的发现一致，

例如，Liang 等人（2024）指出，“产学研合作难以复现学术团队的新颖性，其表现往往与产业团队相似”。

在将所有控制变量加入模型（8）后，“学术界”的系数降至 0.204，优势比为（$OR = e^{0.204} \approx 1.23$）。这表明，即使考虑了混杂因素，学术论文具有高新颖性的优势比仍比产业界论文高 23%。该结果仍具有边际统计显著性（$p < 0.1$），验证了学术界的相对优势。模型（4）中“合作”的系数降至 0.057（$OR = e^{0.057} \approx 1.06$），意味着合作论文具有高新颖性的可能性有限，与产业界论文相比不存在实质性的统计差异。

总而言之，学术界与产业界在平均新颖性得分方面不存在显著差异。然而，学术界在高新颖性成果方面具有更为突出的优势。至于学术界—产业界合作论文，其新颖性在数值上仅略高于产业界论文，但这一差异在所有模型中均不具有统计显著性。

附录表7展示了专利的回归结果。模型（9）仅纳入机构类型和年份固定效应。“学术界”的系数为0.022（$p<0.01$），这表明学术机构的专利平均新颖性得分显著高于产业界机构的专利，差值为0.022。“合作”的系数为0.026（$p<0.05$），其平均新颖性得分也显著高于产业界机构的专利，差值为0.026。模型（10）引入控制变量后，“学术界”的系数在1%的水平上仍具有统计显著性，而“合作”的系数则降至0.022（$p<0.1$）。这表明，尽管合作机构的专利新颖性优势并未完全消失，但在控制其他变量后显著减弱。

进一步分析分位数回归结果（见附录中的表9）发现，随着分位数从25%升至75%，系数持续增大。这表明，与论文中的表现类似，在专利领域，学术界在高新颖性方面也展现出更为突出的优势。此外，逻辑回归模型（11）的结果显示，“学术”的系数为0.762，且在1%的水平上显著（$\beta = 0.762$，$p<0.01$）。将其转换为优势比（$OR = e^{0.762} \approx 2.14$）后可见，在不控制其他变量的情况下，学术机构的专利在高新颖性方面相较于产业界机构的专利具有2.14倍的优势。“合作”的系数为0.695，在5%的水平上显著（$\beta = 0.695$，$p<0.05$）。

在模型（12）中纳入所有控制变量后，“学术”的系数降至0.369，但仍在1%水平上显著（$\beta = 0.369$，$p<0.01$），且$OR\ is\ e^{0.369} \approx 1.45$。此时，“合作”的系数降至0.502（$\beta=0.502$，$p>0.1$）。这表明，尽管数值上存在差异，但这一差异在统计上并不显著。

我们观察了控制变量的影响，国际专利分类代码数量的系数为（$\beta = 0.005,0.253$，$p<0.01$），这表明，专利涵盖的知识领域越多样，越有可能推动跨领域创新，进而提升专利新颖性。专利族规模的系数为-0.004和-0.100（$p<0.01$），说明规模较大的专利族往往侧重于渐进式改进，而非突破性创新，因此会降低新颖性。实体数量的表现与Fleming 和 Sorenson（2001年）的理论一致：增加组成部分数量所带来的边际收益会随着数量增加而递减，从而对专利改进的作用逐渐减弱。鉴于本研究根据新颖实体所占比例计算新颖性，该变量在回归模型中的系数为负。此外，加入控制变量显著提高了所有模型的R平方值，证实纳入这些变量增强了模型的解释力。

讨论

本研究采用细粒度知识实体分析，评估自然语言处理领域专利和论文的新颖性。回归分析结果表明，无论考察论文还是专利，学术界的新颖性水平都高于产业界。这一发现与先前研究的结果一致（Chen 等，2024；Liang 等人，2024）。就创新重点而言，学术界和产业界都强调论文中的方法创新。产业界在数据集资源方面较学术界具有独特优势。值得注意的是，学术部门在产出高度新颖的成果方面优势更为显著。就合作研究成果而言，产学研合作难以复现学术团队的创新水平，其成果往往更接近产业团队的工作（Liang 等人，2024）。作为催化剂，学术界无论是在个人申请专利还是参与合作专利研究方面，都能显著推动新颖性的提升。

启示

本研究的理论意义体现在三个方面。首先，我们将学术论文领域实体识别模型中的知识迁移至专利实体识别任务，并结合基于实体的新颖性测量方法，实现了对专利和论文新颖性的统一量化评估。这为后续基于更大规模数据集开展论文与专利之间的跨层级新颖性评估，建立了可行框架。

其次，为回应现有研究中学术界与产业界对新颖性的认知存在分歧这一问题，本研究将顶级会议论文与专利数据相结合，提供了新的经验证据。研究结果揭示了专利与论文之间的新颖性差异，不同科学文献中的知识实体也呈现出创新差异。在论文层面，学术界与产业界之间的整体新颖性差距相对较小，尤其是在基于方法的创新领域，差异微乎其微。这归因于自然语言处理领域三大会议严格的同行评审机制。与此同时，产业界利用其固有的数据集优势开展了大量创新活动。

第三，本研究在实体层面丰富了重组搜索理论（Fleming，2001年；Fleming与Sorenson，2001年）的经验证据基础：研究将技术创新源于组件局部搜索与重组（国际专利分类子类）这一核心观点细化至知识实体的微观层面。在Aceves和Evans（2024年）的概念空间理论基础上，本研究借助统一词嵌入模型统一论文与专利的测量维度，从而丰富了技术创新研究中组件组合的分析维度。在这一框架下，本研究同样发现，随着组件数量增加，产出的边际收益呈递减趋势。这一现象部分源于认知约束，使多个组件无法得到充分利用（Fleming与Sorenson，2001年）。此外，先前知识的应用范围缩窄也可能导致创新减少（Park等，2023）。

本研究也具有明确的实践价值。研究结果表明，学术界的整体新颖性产出更高，这与Brescia 等人（2016）的观点一致：学术界开放且协作的环境更有利于孕育新思想、推动跨学科创新。本研究的发现有助于政策制定者制定产业界与学术界协作的差异化激励措施。在专利创新层面，应加大对产业界与学术界联合专利申请的支持力度。可以通过补贴专利申请费用、设立优先审查通道等措施来实现，从而充分发挥学术界提升专利新颖性的作用。在论文创新层面，则应避免强制规定合作模式，而应优先支持高校团队开展基础研究，同时引导企业通过数据共享而非共同署名的方式参与论文研究，从而提升创新成效。

局限性与未来工作

尽管本研究通过统一语义空间，以实证方式揭示了学术界与产业界之间的新颖性差异，但仍存在若干局限。首先，研究样本的范围有限。所分析的论文仅涵盖三个具有代表性的自然语言处理会议，未纳入发表于其他相关会议和期刊的自然语言处理研究。专利样本则仅限于核心 CPC 类 G06F40，未涵盖与自然语言处理密切相关的相邻合作专利分类中的专利活动（例如机器学习领域的 G06N20、信息检索领域的 G06F16）。尽管这种抽样策略保证了数据的高度相关性，但可能低估了自然语言处理相关研究论文和专利活动的实际广度。其次，自然语言处理领域的新词迭代迅速，可能引发语义的动态波动。尽管实体规范化与按年度固定语义能够缓解这一问题，但仍可能给实体组合的新颖性测量带来细微偏差。第三，本研究仅聚焦于自然语言处理。作为一个迅速兴起、学术界、产业界和研究机构均深度参与的领域，其他领域可能因发展速度、机构参与模式或知识生产逻辑的差异而呈现异质性。因此，将研究结论推广至其他领域时应保持谨慎。最后，尽管本研究采用了基于经典组合式创新理论（Liu 等人，2022年）的分析方法，但目前尚缺乏金标准验证过程，这也是当代相关研究普遍面临的一项挑战。

未来研究可从三个方向推进：第一，扩大研究数据的范围。对于论文，未来研究可将数据覆盖范围扩展至更多与自然语言处理相关的会议和期刊。对于专利，可采用更精细的分类和筛选方法，识别并纳入相邻技术领域的专利，从而全面捕捉自然语言处理领域的专利活动。第二，

可克服仅依靠语义距离指标衡量新颖性的局限，探索更全面、多维的创新评估框架。最后，补充基于金标准的验证过程。通过整合专家评审意见，系统检验模型区分具有特定新颖性水平的文档的有效性，进一步提升研究成果的严谨性和可靠性。

致谢

本论文得到国家自然科学基金委员会（项目编号：72074113）和江苏省研究生科研与实践创新计划（项目编号：KYCX25_0820）资助。本论文是 ISSI 2025 论文的扩展版本：“Zhao, H., Zhao, Y., & Zhang, C. (2025). 探索产业界与学术界之间的新颖性差异：知识实体中心视角。第 20 届科学计量学与信息计量学国际会议。亚美尼亚埃里温。https://doi.org/10.51408/issi2025_088”。

声明

利益冲突：作者声明，在本研究、作者身份和/或本文发表方面不存在任何潜在利益冲突。

参考文献

Aceves, P., & Evans, J. A. (2024). 调动概念空间：词嵌入模型如何为《组织科学》中的测量与理论提供参考。《组织科学》, 35(3), 788–814. https://doi.org/10.1287/orsc.2023.1686 Ahmed, N., Wahed, M., & Thompson, N. C. (2023). 产业界对人工智能研究的影响力日益增长。《科学》, 379(6635), 884–886. https://doi.org/10.1126/science.ade2420 Arts, S., & Fleming, L. (2018). 新颖性的乐园——还是人力资本的流失？探索新领域与发明产出。《组织科学》, 29(6), 1074–1092. https://doi.org/10.1287/orsc.2018.1216 Arts, S., Hou, J., & Gomez, J. C. (2021). 运用自然语言处理从专利文本中识别新技术的创造与影响：代码、数据与新指标。《研究政策》, 50(2), 104144. https://doi.org/10.1016/j.respol.2020.104144 Ashtor, J. H. (2022). 专利清晰度建模。《研究政策》, 51(2), 104415. https://doi.org/10.1016/j.respol.2021.104415 Ba, Z., Meng, K., Ma, Y., & Xia, Y. (2024). 通过识别科学与技术之间的动态结构耦合模式及领先—滞后距离来发现技术机会。《技术预测与社会变迁》, 200, 123147. https://doi.org/10.1016/j.techfore.2023.123147 Bikard, M., & Marx, M. (2020). 搭建学术界与产业界的桥梁：地理枢纽如何连接大学科学与企业技术。《管理科学》, 66(8), 3425–3443. https://doi.org/10.1287/mnsc.2019.3385 Brescia, F., Colombo, G., & Landoni, P. (2016). 技术转移办公室的组织结构：对全球排名前列大学的分析。《技术转移期刊》, 41(1), 132–151. https://doi.org/10.1007/s10961-014-9384-5 Chan, H. F., & Torgler, B. (2020). 不同领域和国家中高被引科学家表现的性别差异。《科学计量学》, 125(3), 2421–2447. https://doi.org/10.1007/s11192-020-03733-w Chen, C.-J., Hsiao, Y.-C., Chu, M.-A., & Hu, K. K. (2015). 团队多样性与新产品绩效之间的关系：组织冗余的调节作用。《IEEE工程管理汇刊》, 62, 568–577. https://doi.org/10.1109/TEM.2015.2458891 Chen, Z., Zhang, C., Zhang, H., Zhao, Y., Yang, C., & Yang, Y. (2024). 基于细粒度知识实体探讨团队制度构成与学术论文新颖性之间的关系。《电子图书馆》, 42(6), 905–930. https://doi.org/10.1108/EL-03-2024-0070 Chirico, F., Criaco, G., Baù, M., Naldi, L., Gomez-Mejia, L. R., & Kotlar, J. (2020). 是否申请专利：这是个问题。家族企业中的知识产权保护。《创业理论与实践》, 44(2), 339–367. https://doi.org/10.1177/1042258718806251 Codina-Filbà, J., Bouayad-Agha, N., Burga, A., Casamayor, G., Mille, S., Müller, A., Saggion, H., & Wanner, L. (2017). 利用体裁特定特征撰写专利摘要。《信息处理与管理》, 53(1), 151–174. https://doi.org/10.1016/j.ipm.2016.07.002 Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT：用于语言理解的深度双向变换器预训练。载于 J. Burstein, C. Doran, & T. Solorio (编), 《计算语言学协会北美分会2019年会议论文集：人类语言技术》第1卷（长篇与短篇论文）（第4171–4186页）。计算语言学协会。https://doi.org/10.18653/v1/N19-1423

论文集：计算语言学协会北美分会2019年会议论文集：人类语言技术，第1卷（长篇与短篇论文）（第4171–4186页）。计算语言学协会。https://doi.org/10.18653/v1/N19-1423 Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., ... Williams, M. D. (2021). 人工智能（AI）：关于新兴挑战、机遇以及研究、实践与政策议程的多学科视角。《国际信息管理杂志》, 57, 101994. https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2019.08.002 Fleming, L. (2001). 技术搜索中的重组不确定性。《管理科学》, 47(1), 117–132. https://doi.org/https://doi.org/10.1287/mnsc.47.1.117.10671 Fleming, L., & Sorenson, O. (2001). 作为复杂适应系统的技术：来自专利数据的证据。《研究政策》, 30(7), 1019–1039. https://doi.org/10.1016/S0048-7333(00)00135-9 Fontana, M., Iori, M., Montobbio, F., & Sinatra, R. (2020). 新颖与非典型组合：对新颖性和跨学科性的评估。《研究政策》, 49(7), 104063. https://doi.org/10.1016/j.respol.2020.104063 Franceschini, F., Maisano, D., & Turina, E. (2012). 生产技术与制造系统领域的欧洲研究：基于出版物和专利的探索性分析。《国际先进制造技术杂志》, 62(1–4), 329–350. https://doi.org/10.1007/s00170-011-3791-7 Geisler, E. (1995). 当鲸鱼被冲上岸：美国大学与基础科学向相关性的转变。《IEEE工程管理汇刊》, 42(1), 3–8. https://doi.org/10.1109/17.366398 Hottenrott, H., Rose, M. E., & Lawson, C. (2021). 学术界多重机构隶属关系的兴起。Journal of the Association for Information Science and Technology, 72(8), 1039–1058. https://doi.org/10.1002/asi.24472 Jang, H., Kim, S., & Yoon, B. (2023). 一种用于基于文本的专利新颖性分析的可解释人工智能（XAI）模型。《专家系统与应用》, 231, 120839. https://doi.org/10.1016/j.eswa.2023.120839 Jee, S. J., & Sohn, S. Y. (2023). 与科学相关的技术兴起时企业对已发表知识演进的影响：以人工智能为例。《演化经济学杂志》, 33(1), 209–247. https://doi.org/10.1007/s00191-022-00804-4 Jeon, D., Ahn, J. M., Kim, J., & Lee, C. (2022). 一种用于测量专利新颖性的 Doc2Vec 与局部异常因子方法。《技术预测与社会变迁》, 174, 121294. https://doi.org/10.1016/j.techfore.2021.121294 Jeon, D., Lee, J., Ahn, J. M., & Lee, C. (2023). 测量科学出版物的新颖性：一种 FastText 与局部异常因子方法。《信息计量学学报》, 17(4), 101450. https://doi.org/10.1016/j.joi.2023.101450 Jiang, F., Pan, T., Wang, J., & Ma, Y. (2024). 去学术界还是产业界：流动性及其对ACM会士科学职业生涯的影响。《信息处理与管理》, 61(4), 103736. https://doi.org/10.1016/j.ipm.2024.103736 Kabore, F. P., & Park, W. G. (2019). 专利族规模和构成能否表明专利价值？《应用经济学》。https://www.tandfonline.com/doi/abs/10.1080/00036846.2019.1624914 Koppman, S., & Leahey, E. (2019). 谁走向方法论前沿？促使科学家采用非常规方法的因素。《研究政策》, 48(9), 103807. https://doi.org/10.1016/j.respol.2019.103807 Lee, C., & Lee, G. (2019). 基于重组搜索的技术机会分析：用于创意生成的专利图谱分析。《科学计量学》, 121(2), 603–632. https://doi.org/10.1007/s11192-019-03224-7 Lee, S., Yoon, B., & Park, Y. (2009). 一种发现新技术机会的方法：基于关键词的专利地图方法。《技术创新》, 29(6), 481–497. https://doi.org/10.1016/j.technovation.2008.10.006 Lee, Y.-N., Walsh, J. P., & Wang, J. (2015). 科学团队中的创造力：剖析新颖性与影响力。《研究政策》, 44(3), 684–697. https://doi.org/10.1016/j.respol.2014.10.007 Levenshtein, V. I. (1965). 可纠正删除、插入和倒置错误的二进制码。《苏联物理学·报告》, 10, 707–710. Liang, L., Han, Z., Zou, J., & Acuna, D. E. (2024). 学术界与产业界对人工智能研究的互补贡献。arXiv。https://doi.org/10.48550/arXiv.2401.10268

Liang, Z., Mao, J., & Li, G. (2023). 对科学新颖性的偏见：发表前视角。Journal of the Association f or Information Science and Technology, 74(1), 99–114. https://doi.org/10.1002/asi.24725 Liu, M., Bu, Y., Chen, C., Xu, J., Li, D., Leng, Y., Freeman, R. B., Meyer, E. T., Yoon, W., Sung, M., Jeong, M., Lee, J., Kang, J., Min, C., Song, M., Zhai, Y., & Ding, Y. (2022). 大流行病是科学新颖性的催化剂：来自 COVID -19 的证据。Journal of the Association for Information Science and Technology, 73(8), 1065–1078. https://doi.org/10.1002/asi.24612 Liu, M., Xie, Z., Yang, A. J., Yu, C., Xu, J., Ding, Y., & Bu, Y. (2024). 科学新颖性中显著且异质的性别差异：来自生物医学博士论文的证据。《信息处理与管理》, 61(4), 103743. https://doi.org/10.1016/j.ipm.2024.103743 Liu, X., Li, X., Liu, J., & Zhang, P. (2025). 一种用于衡量专利技术创新的新型无监督学习框架。《科学计量学》。https://doi.org/10.1007/s11192-025-05380-5 Luo, Z., Lu, W., He, J., & Wang, Y. (2022). 研究问题与方法的组合：一种科学新颖性测量新方法。《信息计量学学报》, 16(2), 101282. https://doi.org/10.1016/j.joi.2022.101282 Marcoulides, K. M., & Raykov, T. (2019). 使用潜变量建模方法评估回归模型中的方差膨胀因子。Educational and Psychological Measurement, 79(5), 874–882. https://doi.org/10.1177/0013164418817803 Merton, R. K. (1973). 科学的规范结构。科学社会学。载于芝加哥大学出版社，芝加哥。Mori, T., & Sakaguchi, S. (2020). 通过知识交流创造知识：来自日本专利数据的证据（第 arXiv:1908.01256 号）。arXiv。https://doi.org/10.48550/arXiv.1908.01256 Nelson, R. R., & Winter, S. G. (1985). 经济变迁演化理论。哈佛大学出版社。Park, M., Leahey, E., & Funk, R. J. (2023). 随着时间推移，论文和专利的颠覆性日益减弱。Nature, 613(7942), 138–144. https://doi.org/10.1038/s41586-022-05543-x Pramanick, A., Hou, Y., Mohammad, S. M., & Gurevych, I. (2025). 自然语言处理的本质：分析自然语言处理论文中的贡献。载于 W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar（编），计算语言学协会第 63 届年会论文集（第 1 卷：长篇论文）（第 25169–25191 页）。计算语言学协会。https://doi.org/10.18653/v1/2025.acl- long.1224 Riera, R., & Rodríguez, R. (2022). 如果同行评审过程扼杀了突破常规的科学研究，会怎样？《海洋科学前沿》, 9, 924469. https://doi.org/10.3389/fmars.2022.924469 Sauermann, H., & Stephan, P. (2010). 双胞胎还是陌生人？产业科学与学术科学的差异与相似之处（第 w16113 号；第 w16113 页）。美国国家经济研究局。https://doi.org/10.3386/w16113 Shi, F., & Evans, J. (2023). 研究内容与背景之间令人意外的组合与影响力相关，并由来自远距离学科的科学领域外部研究者促成。Nature Communications, 14(1), 1641. https://doi.org/10.1038/s41467-023-36741-4 Shin, H., Kim, K., & Kogler, D. F. (2022). 科学合作、研究资助与科学知识的新颖性。《PLOS ONE》, 17(7), e0271678. https://doi.org/10.1371/journal.pone.0271678 Sun, X., Chen, N., & Ding, K. (2022). 衡量技术的潜在组合新颖性。《专家系统与应用》, 210, 118564. https://doi.org/10.1016/j.eswa.2022.118564 Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). 非典型组合与科学影响力。《科学》, 342(6157), 468–472. https://doi.org/10.1126/science.1240474 Verhoeven, D., Bakker, J., & Veugelers, R. (2016). 使用基于专利的指标衡量技术新颖性。《研究政策》, 45(3), 707–723. https://doi.org/10.1016/j.respol.2015.11.010 Wang, J., Veugelers, R., & Stephan, P. (2017). 科学中的新颖性偏见：文献计量指标使用者的警示。《研究政策》, 46(8), 1416–1436. https://doi.org/10.1016/j.respol.2017.06.006 Wei, T., Feng, D., Song, S., & Zhang, C. (2024). 专利技术知识要素的抽取与新颖性评估框架。《科学计量学》, 129(11), 7417–7442. https://doi.org/10.1007/s11192-024-04990-9 Wu, K., Sun, J., Wang, J., & Kang, L. (2025). 科学融合如何影响技术融合？科学推动与技术拉动的不同影响。《技术预测与社会变迁》, 215, 124114. https://doi.org/10.1016/j.techfore.2025.124114

Wu, K., Xie, Z., & Du, J. T. (2024). 科学会颠覆技术吗？通过制药领域的专利—论文引文考察科学强度、新颖性和时效性。《科学计量学》, 129(9), 5469– 5491. https://doi.org/10.1007/s11192-024-05126-9 Wu, L., Wang, D., & Evans, J. A. (2019). 大型团队推动科学与技术发展，小型团队带来颠覆。Nature, 566(7744), 378–382. https://doi.org/10.1038/s41586-019-0941-9 Xu, H., Bu, Y., Liu, M., Zhang, C., Sun, M., Zhang, Y., Meyer, E., Salas, E., & Ding, Y. (2022). 团队权力动态与团队影响力：以职业年龄作为团队权力代理变量的科学合作新视角。Journal of the Association for Information Science and Technology, 73(10), 1489–1505. https://doi.org/10.1002/asi.24653 Yang, Y., Tian, T. Y., Woodruff, T. K., Jones, B. F., & Uzzi, B. (2022). 性别多元化团队能够产出更多新颖且影响力更高的科学思想。Proceedings of the National Academy of Sciences, 119(36), e2200841119. https://doi.org/10.1073/pnas.2200841119 Zhang, H., Zhang, C., & Wang, Y. (2024). 揭示自然语言处理技术的发展：科学实体中心视角。《信息处理与管理》, 61(1), 103574. https://doi.org/10.1016/j.ipm.2023.103574 Zhang, L., Cao, Z., Shang, Y., Sivertsen, G., & Huang, Y. (2024). OpenAlex 中缺失的机构：可能的原因、影响与解决方案。《科学计量学》, 129(10), 5869–5891. https://doi.org/10.1007/s11192-023-04923-y Zhao, H., Zhao, Y., & Zhang, C. (2026). 量化学术研究与产业研究之间的知识邻近性：实体与语义视角。《技术预测与社会变迁》, 226, 124598. https://doi.org/10.1016/j.techfore.2026.124598 Zheng, H., Li, W., & Wang, D. (2022). 团队专业知识多样性可预测科学与技术领域的原创性和长期影响力。SSRN Electronic Journal。 https://doi.org/10.2139/ssrn.4243054

附录表 4. 从论文和专利中提取的四种类型的前五个实体

Type	论文		专利	
	实体	频率	实体	频率
方法	BERT	4160	神经网络	3021
	LSTM	3565	机器学习	1608
	注意力机制	3392	N 元语法	1365
	Transformer	3321	语言模型	1160
	N 元语法	3287	深度学习	960
Tool	PyTorch	742	计算机系统	11646
	MOSES	654	存储介质	10413
	GIZA++	54	用户界面	9323
	Python	438	计算机程序	8738
	NLTK	397	操作系统	7636
数据集	维基百科	3542	表情符号	306
	WordNet	2152	电子邮件	122
	Twitter	1210	社交媒体	86
	《华尔街日报》	1006	万维网	67
	亚马逊土耳其机器人	982	Twitter	43
指标	准确率	10696	准确率	5278
	F_1	6956	置信度	2500
	精确率	6075	效率	2195
	召回率	5578	相关性	1612

BLEU	3264	错误	1453

表5. 对列表中的机构名称进行分类的提示

我将提供一份名称列表，其中可能包含个人、学术机构（例如大学、研究所）或工业组织（例如公司、企业）。请检查每个名称，并将其归入以下类别之一：“学术界”指大学、学院、研究所或其他以教育为重点的研究组织；“产业界”指公司、企业或商业实体；“个人”指看起来是个人姓名（而非机构）。请勿提供分类理由。如果名称含义模糊或无法明确归类，请标记为“其他”。

不同实体组合类型的语义相似度结果

如附录图8所示，同一类别内的实体组合（例如工具、指标和数据集）具有相对较高的语义相似度。这表明这些组合的语义距离较近，对新颖性的贡献相对较小。排名靠前的组合包括方法-指标、数据集-方法、数据集-指标和方法-方法。在上述核心组合中，方法-指标的语义相似度最低（语义距离最高）。事实上，方法设计与指标定义之间存在密切交织的关系：引入新方法必须配套相应的指标以评估其有效性，而“方法验证逻辑”则是产生高度新颖研究成果的典型途径。此外，方法-数据集组合的排名也很靠前。高质量数据集是方法创新的重要基础。Pramanick 等人（2025年）的研究证实了这一点：首先，新数据集往往能吸引更高的引用率；其次，介绍新方法的论文也会获得更多引文关注。

附录图9展示了专利实体组合相似度。其中，同类型实体对（数据集-数据集、工具-工具、方法-方法、指标-指标）的语义相似度最高，这与 Fleming 和 Sorenson（2001年）的技术景观重组搜索理论相符。当发明人采用模块化策略进行专利开发时，虽然这能降低技术重组的不确定性并提高发明成功率，却也同时缩小了跨模块协同创新的空间，最终限制突破性发明的产生。对于高新颖性专利，其核心实体组合为数据集-指标、方法-指标、指标-工具和数据集-方法，与前文讨论的论文实体组合模式相似。然而，专利与论文的研究重点存在显著差异：专利很少使用数据和指标来验证技术有效性，而是通过描述技术特征及实现效果来展示价值。尽管如此，将数据集与评估指标相结合，或将数据集与方法整合，仍能产生新颖成果。此外，专利强调实际应用，基于工具的实体参与度更高，其中某些工具-指标组合成为推动技术创新的关键途径。

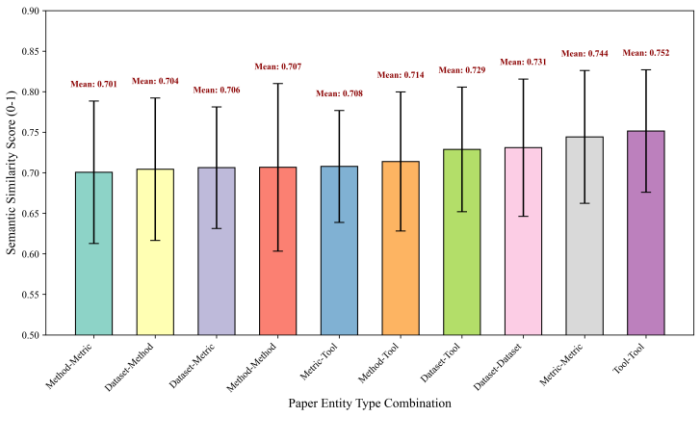

图8. 论文实体之间的语义相似度差异

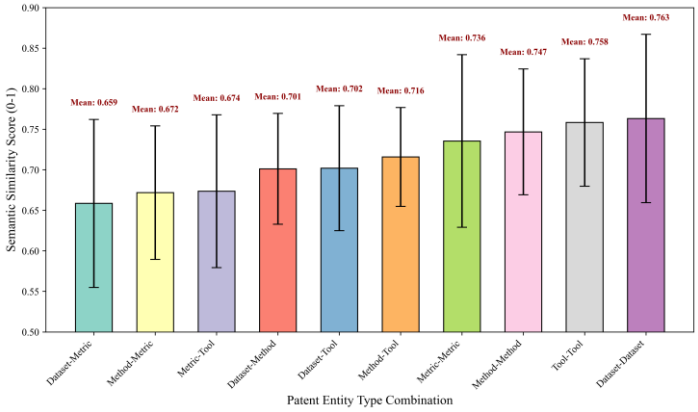

图 9 专利实体之间的语义相似度差异

线性回归和逻辑回归结果

表 6 论文新颖性的回归结果

	新颖性得分（连续变量）		新颖性得分前列（前10% 优势比 NOT）	
变量	模型（5）	模型（6）	模型（7）	模型（8）
学术	0.003**	0.002	0.331***	0.204*
	(0.002)	(0.002)	(0.105)	(0.108)
合作	0.003	0.001	0.158	0.057
	(0.002)	(0.002)	(0.115)	(0.128)
作者数量的自然对数		-0.001		-0.098
		(0.001)		(0.061)
机构数量的自然对数		0.003**		0.194***
		(0.001)		(0.065)
Ln(实体数量)		-0.009***		-1.153***
		(0.001)		(0.067)
常数项	0.121***	0.144***	-2.185***	0.750
	(0.015)	(0.016)	(0.491)	(0.517)
年份固定效应	Yes	Yes	Yes	Yes
观测值	16,295	16,295	16,295	16,295
R平方	0.005	0.009	0.002	0.029

注：括号内为稳健标准误，*** p<0.01，** p<0.05，* p<0.1

表7. 专利新颖性的回归结果

	新颖性得分（连续变量）		新颖性得分前列（是否属于前10%）	
变量	模型（9）	模型（10）	模型（11）	模型（12）
学术	0.022***	0.013***	0.762***	0.369***
	(0.004)	(0.004)	(0.125)	(0.138)
合作	0.026**	0.022*	0.695**	0.502
	(0.013)	(0.012)	(0.333)	(0.352)
作者数量的自然对数		0.005***		0.121***
		(0.001)		(0.038)
自然对数(机构数量)		-0.002		-0.075
		(0.003)		(0.142)
自然对数(实体数量)		-0.036***		-1.906***
		(0.001)		(0.049)
Ln(IPC 数量)		0.005***		0.253***
		(0.001)		(0.053)
自然对数（家族规模）		-0.004***		-0.100**

			(0.001)			(0.043)
常数项		0.113***	0.217***	-1.986***		3.342***
		(0.005)	(0.006)	(0.224)		(0.265)
年份固定效应		Yes	Yes	Yes		Yes
观测值		20,934	20,934	20,934		20,934
R平方		0.007	0.064	0.003		0.111

注：括号内为稳健标准误。*** p<0.01，** p<0.05，* p<0.1。

分位数回归结果

表8. 不同分位数下论文新颖性得分的回归结果

	新颖性得分 Q（0.25）		新颖性得分 Q (0.50)		新颖性得分 Q (0.75)	
变量	模型 (13)	模型 (14)	模型 (15)	模型 (16)	模型 (17)	模型 (18)
学术	0.001	0.002	0.002	0.002	0.007***	0.004*
	(0.002)	(0.002)	(0.002)	(0.002)	（0.002）	（0.002）
合作	0.003	0.001	0.002	0.002	0.004*	0.003
	（0.002）	（0.002）	(0.002)	(0.002)	(0.002)	(0.003)
作者数量的自然对数		0.001		0.002		-0.002
		(0.001)		(0.001)		(0.002)
自然对数(机构数量)		0.001		0.001		0.003*
		(0.001)		(0.001)		(0.002)
Ln(实体数量)		0.016***		-0.000		-0.021***
		(0.001)		(0.002)		(0.002)
常数项	0.037***	-0.007	0.101***	0.086**	0.205***	0.265***
	(0.010)	(0.006)	(0.035)	(0.036)	(0.021)	(0.020)
年份固定效应	Yes	Yes	Yes	Yes	Yes	Yes
观测值	16,295	16,295	16,295	16,295	16,295	16,295
R平方	0.018	0.025	0.007	0.007	0.004	0.012

注：括号内为稳健标准误。*** p<0.01，** p<0.05，* p<0.1。

表9. 不同分位数下专利新颖性得分的回归结果

	新颖性得分 Q（0.25）		新颖性得分 Q（0.50）		新颖性得分 Q（0.75）	
变量	模型（19）	模型（20）	模型（21）	模型（22）	模型（23）	模型（24）
学术	-0.001	0.001	0.021***	0.016**	0.043***	0.019***
	(0.004)	(0.004)	(0.007)	(0.007)	(0.010)	(0.005)
合作	0.005	-0.001	0.031***	0.019	0.024**	0.021
	(0.011)	(0.013)	(0.010)	(0.022)	(0.011)	(0.017)
作者数量的自然对数		0.004***		0.004***		0.006***
		(0.001)		(0.001)		(0.001)
自然对数(机构 *num*)		0.001		-0.003		-0.005
		(0.003)		(0.003)		(0.005)
自然对数(实体数量)		0.003***		-0.026***		-0.059***
		(0.001)		(0.001)		(0.001)
Ln(国际专利分类数量)		0.002**		0.005***		0.007***
		(0.001)		(0.001)		(0.001)
自然对数(家族规模)		-0.003***		-0.004***		-0.006***
		(0.001)		(0.001)		(0.001)
常数项	0.054***	0.045***	0.103***	0.178***	0.154***	0.336***
	(0.003)	(0.005)	(0.010)	(0.005)	(0.011)	(0.007)
年份固定效应	Yes	Yes	Yes	Yes	Yes	Yes
观测值	20,934	20,934	20,934	20,934	20,934	20,934
R平方	0.003	0.006	0.005	0.023	0.007	0.073

注：括号内为稳健标准误。*** p<0.01，** p<0.05，* p<0.1。

新颖性与引文之间的关系

在 Franceschini等人（2012）看来，论文中的引文主要反映学术影响力，而专利前向引文则兼具双重属性：既是知识流动的载体，也是技术应用价值的指标。

现有研究已广泛考察高度新颖成果的引文模式：大多数研究表明，专利新颖性的提升通常与更多前向引文相关（Jeon 等，2022；Liu 等人，2025）。然而，必须明确的是，引文指标衡量的是发表后的事后影响，而非成果本身的内在新颖性。也存在相反的观点。Arts & Fleming（2018）发现，尽管探索新颖领域能够提升发明新颖性，但知识储备不足和较高的学习成本可能会削弱其即时价值。Wang等（2017）发现，高度新颖的研究表现出更大的引文方差，并且在初始发表阶段往往难以获得较高的引文数量。Liang 等（2023）进一步补充指出，高度新颖的成果通常面临更长的审稿与修订周期，这反映出学术评价中“延迟认可”的偏差，并与引文滞后现象相呼应。

为验证新颖性与引文之间关联的稳健性，我们采用 Mann-Whitney U检验比较高新颖性与普通论文/专利的引文结果。基于 OpenAlex 中 15,871 篇论文的五年引文数据，以及来自专有数据库的 20,934 项专利的前向引文数据，我们的分析发现：

a) 对于论文：高新颖性样本的五年引文数量显著较低，且这一模式在总体样本和学术界子组中均具有统计稳健性（p<0.001；见附录**图10**）；b) 对于专利：高新颖性样本获得的前向引文显著更多，这一趋势在总体专利样本和产业界子组中均得到验证（p<0.001；见附录**图11**）。

这些发现与现有文献一致：专利新颖性与后续技术影响呈正相关，因为新颖专利表现出更活跃的技术扩散和更大的平均技术影响力。相比之下，论文新颖性与引文表现之间的关联仍为负向或不确定。这是因为高新颖性论文容易受到评估中的系统性偏差影响，从而难以在短期内获得广泛的学术关注和认可。

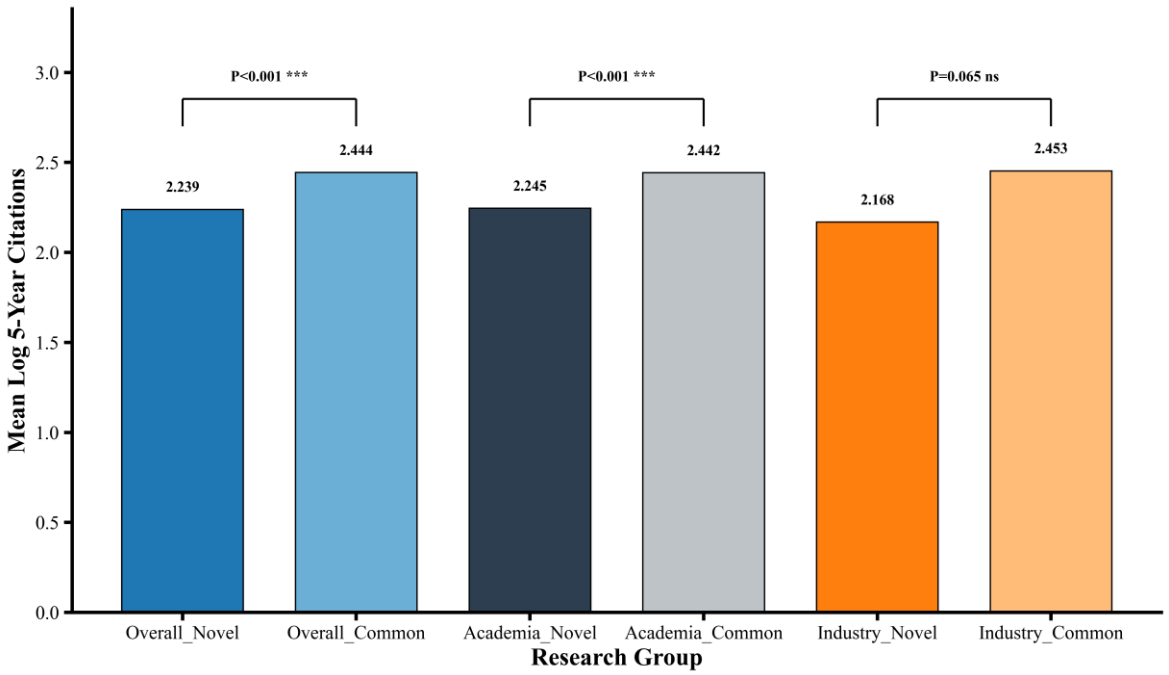

图10. 新颖论文与普通论文的5年引用次数差异

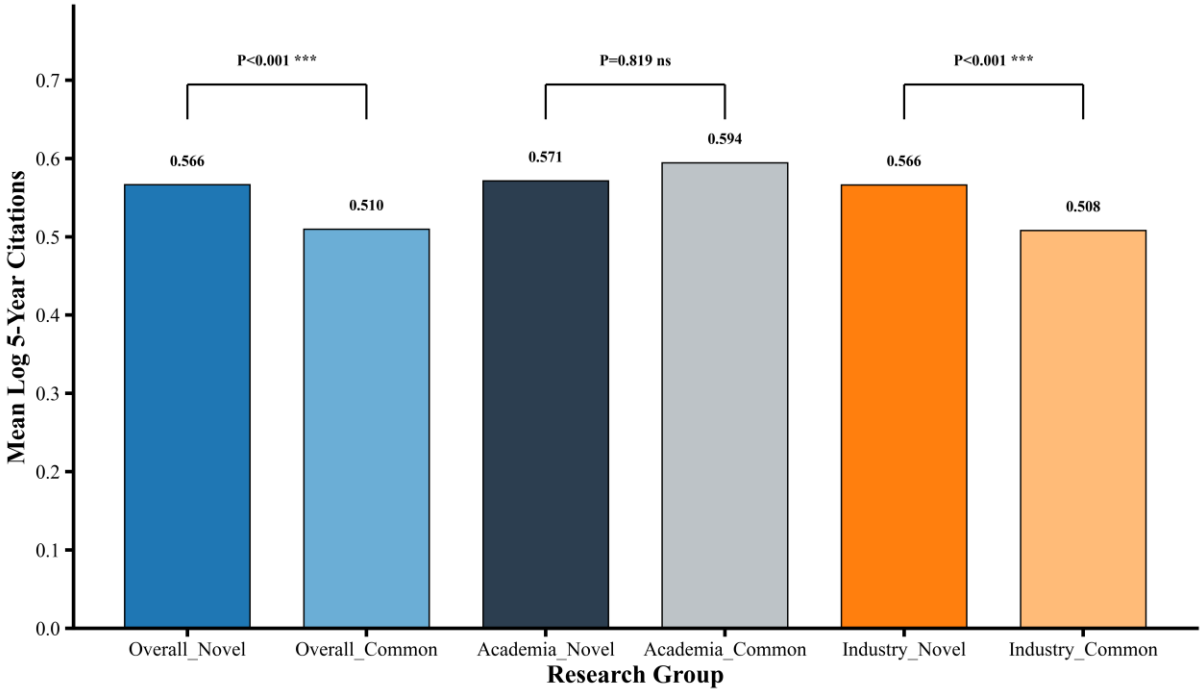

图11. 新颖专利与普通专利之间5年引用次数的差异