生成式 AI 驱动创新的数量和质量效应

——基于诺贝尔化学奖成果 AlphaFold 的发现

肖柏高 周阳 刘青

目录

附录Ⅰ 发明专利数量调整方法 1
附录Ⅱ 稳健性检验结果 2
附录Ⅲ 第五章“探索式创新”变量定义和回归结果 6
附录Ⅳ 第六章部分变量定义和构造 9
附录Ⅴ 第六章部分回归结果 10
参考文献 11

附录 Ⅰ 发明专利数量调整方法

本文收集了样本区间 2017-2023 年之前，即 1990-2016 年中国上市公司所有发明专利的申请和公开时间，通过专利申请和公开时间之间的时间差，计算出发明专利从申请到公开的平均分布情况W_s，即专利在申请之后第 s 年获得授权的平均比例。然后，采用如下计算公式进行调整：$P_{adj} = P_{raw} / \sum_{s=0}^{2023-t} W_s$，$2017 \leq t \leq 2023$。其中，$P_{adj}$表示企业 i 在 t 年申请并公开的经调整的蛋白质发明专利数，P_{raw}表示截至 2024 年已被公开的蛋白质发明专利数，W_s表示专利在申请之后第 s 年被公开的平均比例。在本文的回归中，发明专利申请数均是经过该公式调整的。根据计算结果，$W_0 = 0.58$，即约有 58%的发明专利在申请后的一年之内会被公开，而W_1至W_6则均接近于 0.99，即几乎所有的发明专利在申请后的两年之内都会被公开。根据国家专利局的专利审核制度，发明专利在申请后的 15-18 个月内会被公开，这与本文的计算结果相符，表明本文使用的专利数调整方法是合理的。

附录 II 稳健性检验结果

表 II 1 平行趋势检验回归结果

变量	lnPro
pre_3	-0.004
	(0.004)
pre_2	-0.004
	(0.012)
pre_1	0.007
	(0.006)
current	0.009**
	(0.004)
post_1	0.017**
	(0.009)
post_2	0.050*
	(0.028)
控制变量	Y
年份固定效应	Y
企业固定效应	Y
观测值	26,764
R^2	0.618

注：括号内为行业层面的聚类标准误：***、**、*分别代表 1%、5% 和 10% 的显著性水平，下表同。

表 II 2 其他稳健性检验

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	ChatGPT 识别	泊松模型	删除医药行业	高维固定效应	半径匹配	核匹配	企业暴露指数	删除无蛋白质专利行业
变量	lnPro	Pro	lnPro	lnPro	lnPro	lnPro	lnPro	lnPro
AlphaFold×Treat	0.017***	0.248**	0.009***	0.035**	0.032**	0.031**	0.010*	0.024*
	(0.003)	(0.117)	(0.003)	(0.014)	(0.013)	(0.012)	(0.005)	(0.012)
控制变量	Y	Y	Y	Y	Y	Y	Y	Y
企业固定效应	Y	N	Y	Y	Y	Y	Y	Y
年份固定效应	Y	Y	Y	Y	Y	Y	Y	Y

行业固定效应	N	Y	N	N	N	N	N	N
城市×年份固定效应	N	N	N	Y	N	N	N	N
观测值	26,764	26,764	25,056	25,802	26,753	26,758	26,764	23,023
R^2/Pseudo R^2	0.619	0.409	0.596	0.627	0.606	0.600	0.618	0.615

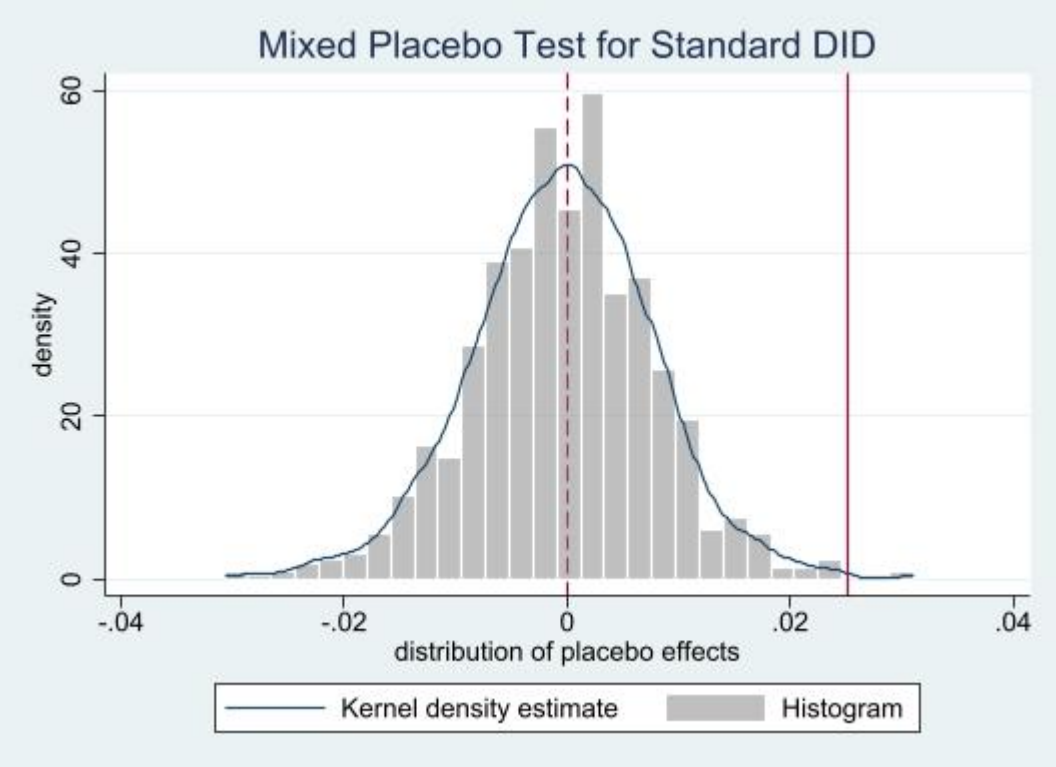


图 II 1 安慰剂检验

表 II 3 平衡性检验

Panel A		半径匹配结果			
变量	匹配状态	处理组	控制组	标准偏差	t 检验
R&D	U	15.897	14.673	18.400	9.660
	M	15.897	15.792	1.600	0.740
Productivity	U	155.330	205.070	-8.400	-3.590
	M	155.350	153.700	0.300	0.150
Current_asset	U	0.531	0.590	-32.100	-16.770
	M	0.531	0.529	0.700	0.300
Concentration	U	32.007	33.113	-7.800	-4.170
	M	32.008	32.007	0.000	0.000
Executive	U	0.112	0.106	3.600	1.980

	M	0.112	0.112	-0.100	-0.050
lnAge	U	2.9996	2.9828	5.600	2.970
	M	2.9995	3.0025	-1.000	-0.430
lnSize	U	22.080	22.339	-21.500	-10.960
	M	22.080	22.066	1.100	0.510
ROA	U	0.0499	0.0308	24.600	14.040
	M	0.0499	0.0484	1.900	0.880
Leverage	U	0.3320	0.4210	-46.100	-24.630
	M	0.3321	0.3316	0.300	0.120
Growth	U	0.1762	0.1524	5.000	2.960
	M	0.1763	0.1739	0.500	0.210

Panel B	核匹配结果				
变量	匹配状态	处理组	控制组	标准偏差	t 检验
R&D	U	15.897	14.673	18.400	9.660
	M	15.897	15.738	2.400	1.110
Productivity	U	155.330	205.070	-8.400	-3.590
	M	155.350	156.020	-0.100	-0.060
Current_asset	U	0.531	0.590	-32.100	-16.770
	M	0.531	0.535	-2.200	-0.950
Concentration	U	32.007	33.113	-7.800	-4.170
	M	32.008	32.107	-0.700	-0.300
Executive	U	0.112	0.106	3.600	1.980
	M	0.112	0.112	-0.200	-0.080
lnAge	U	2.9996	2.9828	5.600	2.970
	M	2.9995	2.9993	0.100	0.030
lnSize	U	22.080	22.339	-21.500	-10.960
	M	22.080	22.085	-0.400	-0.190
ROA	U	0.0499	0.0308	24.600	14.040
	M	0.0499	0.0471	3.600	1.610
Leverage	U	0.3320	0.4210	-46.100	-

					24.630
	M	0.3321	0.3382	-3.200	-1.430
Growth	U	0.1762	0.1524	5.000	2.960
	M	0.1763	0.1712	1.100	0.440

附录Ⅲ 第五章“探索式创新”变量定义和回归结果

1.变量重新定义

在正文表 4 中，我们将相似度位于前 10%的专利定义为高相似度专利，我们也尝试将相似度位于前 20%和前 50%的专利定义为高相似度专利，据此的相关回归结果见表Ⅲ1 第 1-4 列。结果表明，不管以何种标准界定高相似度专利，均可以得到 AlphaFold 提升蛋白质专利相似度、创新趋同的结论。我们还剔除企业自引的引用数据重新计算了专利的突破性和颠覆性指标，相关结果见表Ⅲ1 第 5-6 列，结果显示剔除自引后 AlphaFold 仍然无助于提升创新的颠覆性与引领性。

表Ⅲ1 “探索式创新”变量重新定义

	(1)	(2)	(3)	(4)	(5)	(6)
	更改高相似度专利范围				剔除自引	
变量	lnSimilarity_20	lnSimilarity_50	Ratio_20	Ratio_50	颠覆性	引领性
AlphaFold×Treat	0.152***	0.105*	0.149***	0.119***	0.030	-0.513
	(0.027)	(0.056)	(0.034)	(0.031)	(0.064)	(0.350)
控制变量	Y	Y	Y	Y	Y	Y
年份固定效应	Y	Y	Y	Y	Y	Y
个体固定效应	Y	Y	Y	Y	Y	Y
观测值	876	876	876	876	876	876
R^2	0.281	0.408	0.180	0.261	0.483	0.071

2.部分回归结果

为了克服样本量限制，我们尝试构建了与基准回归一致的企业-年度层面的计数型指标。具体定义如下：

高相似度专利数量：计算每家企业每年申请的蛋白质专利中，文本相似度处于全样本前 10%、20%、30%的专利数量（分别为 lnSimilarity_10、lnSimilarity_20、lnSimilarity_30），如无则为 0。

高被引专利数量：计算每家企业每年申请的蛋白质专利中，被引次数处于前 10%、20%、30%的专利数量（分别为 lnCitation10、lnCitation20、lnCitation30），如无则为 0。

高颠覆性专利数量：计算每家企业每年申请的蛋白质专利中，颠覆性指数为正且处于前 10%、20%、30%的专利数量（分别 lnCD10、lnCD20、lnCD30），如无则为 0。

通过上述指标重构，我们将回归样本从原本的 876 个观测值扩展至与主回归一致的

26,764 个观测值。全样本回归不仅极大提升了统计功效和结果的稳健性，也避免了因截取特定子样本而可能产生的选择性偏差。回归结果分别见下表，结论与正文中保持一致：

表Ⅲ2 高相似度专利

变量	(1) lnSimilarity_10	(2) lnSimilarity_20	(3) lnSimilarity_30
AlphaFold×Treat	0.013**	0.020**	0.025*
	(0.006)	(0.010)	(0.013)
控制变量	Y	Y	Y
企业固定效应	Y	Y	Y
年份固定效应	Y	Y	Y
观测值	26,764	26,764	26,764
R^2	0.342	0.424	0.474

注：括号中为聚类到行业的稳健标准误，其余控制变量回归结果已略去。*** p<0.01, ** p<0.05, * p<0.1。如无特殊说明，下表与此相同。

表Ⅲ3 高颠覆性和高引领性专利

变量	(1) lnCD10	(2) lnCD20	(3) lnCD30	(4) lnCitation10	(5) lnCitation20	(6) lnCitation30
AlphaFold×Treat	-0.024	-0.060	-0.060	-0.015	-0.039	-0.055
	(0.018)	(0.043)	(0.043)	(0.015)	(0.032)	(0.041)
控制变量	Y	Y	Y	Y	Y	Y
企业固定效应	Y	Y	Y	Y	Y	Y
年份固定效应	Y	Y	Y	Y	Y	Y
观测值	26,764	26,764	26,764	26,764	26,764	26,764
R^2	0.211	0.360	0.360	0.261	0.345	0.362

我们利用阿里巴巴千问大模型（Qwen）将每一项专利的摘要文本映射为包含丰富上下文语境的 1024 维稠密向量。相比于传统方法，LLM 能够深入理解文本中隐含的具体业务领域与技术路线等高维语义特征，从而最大程度减少了异质性信息的遗漏。表Ⅲ4 的回归结果显示，在使用这一能更充分反映技术路线异质性等信息的指标后，结论依然稳健，再次支持了“创新趋同”的推断。

表Ⅲ4 LLM 识别的高相似度专利回归结果

变量	(1) lnSimilarity_10	(2) lnSimilarity_20	(3) lnSimilarity_30

AlphaFold×Treat	0.036*	0.067*	0.089*
	(0.019)	(0.037)	(0.052)
控制变量	Y	Y	Y
企业固定效应	Y	Y	Y
年份固定效应	Y	Y	Y
观测值	26,764	26,764	26,764
R^2	0.160	0.217	0.271

附录Ⅳ 第六章部分变量定义和构造

1. **AI 招聘岗位：**职位名称或者职位描述中包含姚加权等（2024）构建的 AI 词典中的任意一个 AI 技术关键词，则将其定义为 AI 招聘岗位。

2. **企业蛋白质发明人数：**本文将蛋白质专利的发明人按照企业-年份分组汇总，然后在组内进行去重，统计出企业在每一年的蛋白质发明人数，然后加 1 取自然对数构建蛋白质研发人数的代理变量。需要注意的是，有部分上市企业无线上招聘数据，我们将这种情况下的 AI 招聘岗位设定为缺失。没有线上招聘数据可能是企业当年只有线下招聘或者没有进行招聘，只有后者情况下将招聘岗位数量设为 0 才合理，但按常理推测当年只有线下招聘可能性更大，此种情况下 AI 招聘岗位定义为缺失更合理。

3. **行业间合作与行业内合作专利：**首先基于专利申请人字段获得所有合作者名称，然后将蛋白质专利合作者分为科研机构、企业、个人发明者，如果名称中包含“学院”“大学”“研究院”等词语则为科研机构，[1]名称中包含“公司”的合作者分类为企业，其余则为个人发明者。由于个人发明者占比很少（0.17%），本文不予考虑。对于企业，本文通过企业名称匹配工商注册大数据获得其所属行业，工商注册数据中无法找到行业信息的企业，作者人工从天眼查、企查查、爱企查等查找后进行补充，若仍无法找到则将其删除。如果有行业外的合作者，则此专利为跨行业合作专利，否则为行业内合作专利。

4. **算力企业合作专利：**如果专利合作企业的行业属于证监会 2012 年版行业分类中的科技推广和应用服务业（M75）、互联网和相关服务（I64）、软件和信息技术服务业（I65）、研究和试验发展（M73），则将其定义为算力企业合作专利。

5. **人均蛋白质专利数量：**如果直接在专利库中基于发明人名字计算其蛋白质专利数量，则会因为重名的问题导致较大偏差，我们通过构建“企业名-人名”组合来减轻这一问题，因为一家企业内部的研究者重名概率较小，且一名员工也不会同时属于多家企业。当专利“申请人”字段存在多家企业或机构时，难以判断发明人所属企业，因此我们尝试将样本限制为“申请人”只有一家企业的专利。通过在“企业名-人名”层面对专利数量进行加总，可以得到每位研究者的蛋白质专利数量。

[1] 本文将公司下属的研究机构归类为科研机构，而不归类为企业，例如“XX 公司 XX 研究所”、“XX 公司 XX 研发中心”、“XX 公司 XX 技术中心”等。

附录Ⅴ 第六章部分回归结果

在正文表 5 第（4）列计算蛋白质专利发明人数时，虽然我们限制了企业单独申请的蛋白质专利，排除了大多数显性的外部研发人员，但考虑到部分外部合作者可能在单一企业专利中挂名，我们进行了稳健性检验，仅统计作为“第一发明人”的研究者。根据专利署名惯例，第一发明人通常是企业内部负责项目执行的核心研发人员，而外部顾问往往位列其后。回归结果见表Ⅴ1 第（5）列，结果显示该指标依然显著，证明我们的结论并非由外部挂名人员驱动。

在正文表 5 第（5）列中，每位发明人的贡献被视为相同，这可能会高估研究者的实际产出。我们采取两种方法解决此问题，第一是只考虑研究者作为第一发明人的蛋白质专利数量，第二种是基于排名加权的蛋白质专利数量，加权公式为 1/n，其中 n 代表研究者所在的位次。以上述方法重新计算后的回归结果见表Ⅴ1 的列（1）和列（2），AlphaFold×Treat 的系数均在 1%的显著性水平上显著为正，说明结论稳健。此外，本文根据某研究者当年是否申请蛋白质专利创建虚拟变量，然后以该虚拟变量为被解释变量，分别使用 OLS 和 Probit 模型进行回归。回归结果如（3）-（4）列所示，AlphaFold×Treat 均在 1%的显著性水平上显著为正，说明更多研究者从事蛋白质创新，这可能是“AI 驱动”的蛋白质研究潜力吸引，也可能是 AlphaFold 降低蛋白质研究门槛所致。

表Ⅴ1 个体生产效率和蛋白质专利发明人数

变量	(1)	(2)	(3)	(4)	(5)
	蛋白质专利申请数量		是否申请蛋白质专利		蛋白质专利第一发明人数量
	位次加权	第一发明人	OLS	Probit	
AlphaFold×Treat	0.001***	0.000***	0.003***	0.671***	0.018**
	(0.000)	(0.000)	(0.001)	(0.227)	(0.007)
控制变量	Y	Y	Y	Y	Y
年份固定效应	Y	Y	Y	Y	Y
企业固定效应	Y	Y	Y	N	Y
观测值	4,921,796	4,921,796	4,921,796	4,936,350	26,764
R^2/Pseudo R^2	0.323	0.226	0.230	0.108	0.664

参考文献

[1] 姚加权、张锟澎、郭李鹏、冯绪，“人工智能如何提升企业生产效率？——基于劳动力技能结构调整的视角”,《管理世界》，2024年第2期，第101－116页。

注：该附录是期刊所发表论文的组成部分，同样视为作者公开发表的内容。如研究中使用该附录中的内容，请务必在研究成果上注明附录下载出处。