

“靠论文爬台阶的学术评价体系正在崩塌”

——专访华东师范大学“AI 一作”实验发起人之一张治教授

■本报见习记者 樊晓丽

2025 年 9 月,作为一项实验,华东师范大学向全球发出一则特殊的征文通知,要求论文必须由人工智能(AI)担任第一作者,人类退居协作与把关的角色。此后半年间,实验工作人员共收到 724 篇有效投稿,一篇篇结构完整、逻辑自洽的学术论文被批量“生产”出来。这个被一些人称为“先锋实验”,被另一些人看作“压力测试”的实验,以一份题为《超越图灵测试:全球首个“AI 一作”大型社会实验全景报告》的研究文本,记录了实验期间的主要发现。

实验结束后,针对 AI 对大学教育的影响,《中国科学报》专访了该实验发起人之一,华东师范大学上海智能教育研究院教授张治。

“AI 的‘创新’主要是重组而非‘从 0 到 1’”

《中国科学报》:有人将此次实验视为一次“验证性实验”,因为 AI 存在虚构文献、创新不足等问题,在实验前就已被大家所知,对此你怎么看?

张治:有些问题在实验之前的确有所暴露,对于这些问题,此次实验可被视为一种“验证”,但有些结果则是我们事先没有预料到的。

比如在评审环节,我们意外地发现,对差的文章和极好的文章,人类专家评审与 AI 大模型评审的一致性竟然高达 75%。如此高的评审一致性说明,AI 在评审环节确实能够发挥作用。我们此前认为大模型无法审美,也谈不上价值判断;而现在看来,只要人类把价值观输入给 AI,AI 就能模仿人作出决策。这是一个比较新的认知突破。

再比如在方法层面,特别是在路径创新上,AI 让人文社科的研究,尤其是教育研究产生了一些我们过去并不擅长,也难以大规模采用的新路径。如将自然科学的很多数据工程、推理模型带进了人文社科领域。这对拓展我们的研究范式具有重要价值。

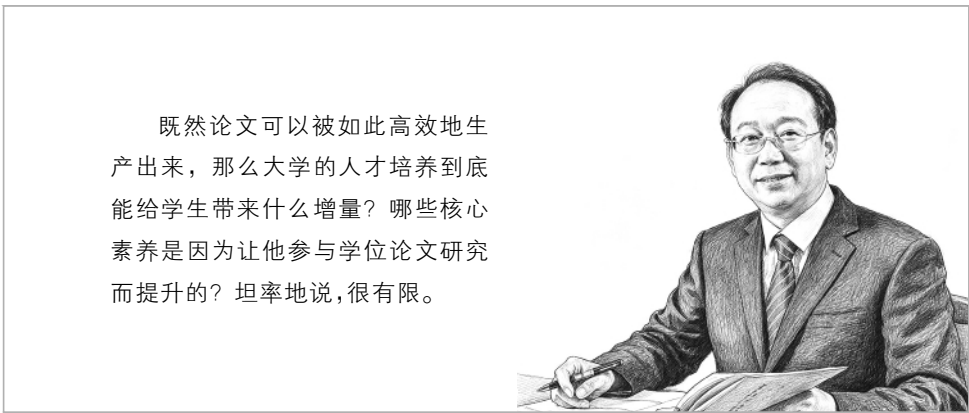
《中国科学报》:在你看来,AI 究竟能不能产生真正的创新?

张治:我们进行此次实验的重要目的,就是想验证 AI 究竟能否创新,以及能创新到何种程度。从实验结果看,目前 AI 的创新大部分情况下仍属于碎片化重组和跨界迁移。

这有些类似于基因重组——把大豆的基因转接到水稻上,水稻的蛋白质含量就会变得更高,此时的水稻可以算作一个新品种,但它并没有产生新的基因。AI 也是同理,它擅长把人类已有的思想碎片重新连接和组合。至于这种组合是否产生了人类社会没有提出过的观点,或者说完全由 AI 提出的“从 0 到 1”的原创性突破,目前还很难验证。但至少在现阶段,这种现象即便存在,也很不明显。

“论文崇拜正在崩塌”

《中国科学报》:此次实验中,是否有些趋势或现象让你感到震撼?



既然论文可以被如此高效地生产出来,那么大学的人才培养到底能给学生带来什么增量?哪些核心素养是因为让他参与学位论文研究而提升的?坦率地说,很有限。

张治:最让我感到震撼的是大学意义感的丧失——按照现在的模型推演下去,我认为大学的意义会大幅度降低,特别是在文科领域,尤其是其中的教育学。

具体而言,一个传统的教育学硕士通常要学 3 年时间,其最重要的成果就是一篇硕士论文,要求有一些实证方法,再加上文献索引和讨论。而借助 AI 工具,即便是本科生也能在很短时间内完成此类论文。

还有一个现象值得关注:这次实验的上榜论文中,人类作者为高校学生(含本硕博)的比例达到 65.9%。这意味着借助 AI,年轻学生正在以前所未有的速度产出学术成果。知识积累的门槛被大幅拉低,传统的“师徒制”学术训练模式正在被颠覆。

既然论文可以被如此高效地生产出来,那么大学的人才培养到底能给学生带来什么增量?哪些核心素养是因为让他参与学位论文研究而提升的?坦率地说,很有限。

《中国科学报》:你是否认为论文作为评价工具的根基被动摇了?

张治:是的,过去靠论文爬台阶、戴“帽子”的学术评价体系正在崩塌。

实验中,我们还有一个有趣的发现,即不同 AI 大模型的“科研品味”有明显差异。我们用 6 个国内外主流 AI 大模型进行了创新性专项评审,结果发现,Qwen 和 DeepSeek 的表现最均衡,与人类专家的一致性最高;Claude 和 ChatGPT 则有较高的“独特偏好度”,它们认可的某些论文完全不被其他大模型认可。这说明 AI 并非“铁板一块”,其评价背后是训练数据、优化目标、语料分布的差异。这导致未来可能出现针对不同 AI 大模型的“投其所好”式的投稿策略。

基于此,我们认为将来证明一位学者创新能力强弱的有效方法,可能不再是他发了多少篇论文,而是其有多少观点被大模型引用和传播。换言之,在 AI 时代,判断一个人贡献度、创新性的最好标准可能不再是论文,而是有多少思想可追溯、还原到这个人身上。

《中国科学报》:在此背景下,高校以及相关的老师和学生应做怎样的调整?

张治:这个问题很难回答。在我看来,文

往处于劣势,而高校要实现自身内涵发展,又急需充足科技资源作为支撑保障。科技资源供给有限与发展需求旺盛形成了现实矛盾,若缺乏有效的统筹组织,容易陷入“资源少—成果弱—影响小—资源更少”的恶性循环,严重制约优势特色的发挥。

正是上述三方面的现实张力,决定了行业特色高校必须提升科技攻关组织化能力。

深层次的创新组织方式变革

对行业特色高校而言,实现科技攻关的组织化不能简单理解为资源堆砌,也不是单纯的行政管理问题,而是要用系统思维和组织化路径明确攻关方向、集聚科研力量,推动创新力量从分散走向聚合、创新活动从局部突破迈向体系协同,最终实现全校科技攻关“一盘棋”。

首先,要推动从“学科逻辑”向“问题逻辑”的转变。行业特色高校的科研选题长期面临一个结构性问题,即知识生产逻辑通常基于学科自身,导致与国家战略需求和行业实际痛点距离较大。在新阶段,行业特色高校应构建“国家战略+行业需求+学科特色”三位一体的方向凝练机制,面向国家战略需求,深度对接国家战略任务中的关键核心技术和“卡脖子”领域挑战,凝练出一批真正的科学问题和技术难题。

顺应行业现实发展,行业特色高校应瞄准行业转型和产业升级现实需求,系统梳理行业产业链、创新链中的薄弱环节和卡点、堵点,反向凝练出行业关键技术攻关清单,力争从高校维度给出落地解决方案;同时立足学科前沿,围绕人类共同面临的重大难题,凝练出一批前瞻性、前沿性科学问题,实现对战略必争领域的超前布局。当然,高校的主攻领域与重要发力点应该是动态的,要根据行业产业升级、技术迭代趋势及时进行调整,确保科技攻关始终贴合真问题、真需求。

其次,要推动从“单兵作战”向“矩阵协

凭、作业、论文等传统教育“工具”将在 AI 时代逐渐失灵。有了 AI 工具的帮助,再用传统标尺测量学生的学业成就已经失去意义。

不过,也有一些只有人类才具备的优势,AI 是很难学会的——

一是创造力,尽管 AI 能产生看似创新的东西,但仍无法替代人的创造性;二是对事物的深度理解,如对教育的理解,对特殊儿童的理解等,这种理解来自真实的社会体验,并会在真实的感触中萌发情感认同和价值认同;三是链接能力,即将任务和资源、场景和工具、人和事关联起来,并实现各要素深度衔接的能力。

至于如何扩大这些优势,大学一方面要重新设计学生学习的过程。学生的亲身经历是 AI 最难以替代的。学生到底在大学期间有没有真正在课堂上体验教学,有没有和老师或领导一起工作,有没有参与一次真正的野外科考……这些经历本身就是教育。

因此,大学要努力将“教程”变成“学程”,精心设计学生在读书期间所经历的事件、见到的、人扮演的角色,以及最终完成的作品。

另一方面,要加快建立 AI 透明协议,且要让协议得到广泛的社会认可。在我的设想中,该协议可分四档:第一档是“零参与”,即有些内容不允许 AI 触碰,如闭卷考试必须由学生独立完成;第二档是“咨询”,即 AI 可作为咨询对象、资料来源,甚至作为头脑风暴的伙伴一起探讨问题,但写作和判断必须由学生自己完成;第三档是“赋能”,即 AI 可在某个阶段全部赋能,如在文献检索阶段整理文献,但必须标注清楚哪些内容由 AI 完成;第四档则是“开放”,即完全由 AI 完成。

“这场博弈终将走向双向奔赴”

《中国科学报》:你认为在 AI 时代,该如何判断一个人是不是人才?

张治:答案已经不算“值钱”了,相比之下,问题才“值钱”。换言之,我们现在最需要培养的学生素养是能够提出好问题,只把老师出的题目正确地做出来已经不算稀奇,因为在这方面,AI 的替代性太强了。

近期我曾作为评委,参加过某项 AI 大赛,该赛事有一个赛道叫“提问能力”。其间,一名小学三年级的学生提出了一个很多人都没想过的问题,并且用 AI 智能体找到了完整的解决方案;还有一个初中生提出的问题是“十万年后,人类还是人类吗”,同时也用 AI 工具做了实验方案。这些孩子使用的工具平台和思维方式与一名博士生毫无差异。从这个角度看,我们对于一张文凭所代表的能力内涵,需要有一个重新定义。

《中国科学报》:实验报告里提到“AI 代沟”,即资深学者和年轻学生对 AI 的态度迥异,这一“代沟”能弥合吗?

张治:“AI 代沟”确实存在。资深学者把 AI 视作需长期“驯化”的合作者和需要严格把关的“乙方”,但一些本科生却认为“AI 比传统教授更可靠”。年轻一代更倾向于将 AI 视为扩展认知的工具,遇到 AI 给出的陌生方法,他们会选择自主学习,而不是质疑。

这两种态度各有价值——资深学者的批判性审视是防止 AI 幻觉和学术泡沫的“压舱石”;年轻一代的无缝协同和颠覆性思维则是激发 AI 潜能的“催化剂”。两种态度最终不是谁赢谁输,而是一个“双向奔赴”的过程。我估计,这场博弈大概需要 10 年时间就能达成新共识。

究其原因,一是大模型的进化渐趋完善,并能达到一个稳定状态,大家都将其视为空气一样不可或缺;二是人对 AI 的认知逐渐趋于平衡,无论老少,都能正视 AI 工具的客观存在,并愿意默认它为科研伙伴。这个共识的核心载体就是 AI 透明协议,它是一个妥协的结果,老一代认可新范式的必要性,年轻一代也接受某些边界和约束。

《中国科学报》:对于此次实验的未来,你还有什么打算?

张治:这个实验有起点,但没有终点。这就像将一块石头扔进河里,涟漪还在扩散,并没有恢复到静止状态。我们还在进一步观察和收集数据,看高校有没有发生变化、媒体如何反应、人的行为有没有改变。我们希望把实验的重要结论变成政策建议,提交给相关机构,让他们正视 AI 署名、AI 与人类创作者的边界、AI 透明协议、学历学位认证办法等问题。不一定马上采纳,但要有这个意识。

同时,我自己正在竭力推动“教育云球”的建设,希望将教育学研究变成一个超级实验平台,既能做宏观教育决策的推演,也能做微观教学行为的模拟。比如学制缩短两年会带来什么变化、高考英语降分会产生什么影响……这些过去难以论证的问题,都可以通过模拟仿真推演。未来有关部门出台教育政策时,可以先在“教育云球”上“滚一滚”,把可能的后果做量化分析,让决策不再是“拍脑袋”,而是基于实证和数学建模。

说到底,这次实验虽然告一段落,但它带来的长尾效应才刚刚开始。我们希望它不只是一块激起短暂浪花石头,而是能推动学术生态和教育制度与时代同步调整的力量。

核心技术突破,提升资源投入产出效能。

当前,行业特色高校通常是经过竞争获得科研资助,经费少而散,无法完全适应组织化科技攻关的要求。因此,高校应积极争取更多来自政府、企业、社会组织的支持,多元筹措创新资源,并向关键创新团队倾斜,保障科研长期稳定系统开展。在此过程中,评价机制也要体现对实际贡献的倾斜。资源配置的优化不仅是分配资源,更要通过评价机制的创新,引导科研力量向组织化方向汇聚。

厘清“不组织什么”更可贵

组织化是提升行业特色高校科技攻关能力的关键路径,但也要清醒认识到,科技创新有其自身的客观规律。相比于明确“组织什么”,厘清“不组织什么”更为可贵。要避免为了组织而组织,将组织化简单等同于行政管控、任务捆绑。推进组织化绝非机械的资源叠加和拉郎配,必须立足高校自身科技攻关的真实需求,按科技创新规律办事,坚持按需组织、因势施策才能真正将组织优势转化为创新优势。

要避免齐步走式的组织。科技攻关虽然要有明确的目标任务和关键指标,但在攻关团队和技术路线选择上,高校不能搞论资排辈,而应遴选出真正有本事的团队来担纲,同时赋予科研人员更大的自主权,特别是在基础研究中要为科研人员保留自由探索的空间,避免过度约束扼杀科研创新活力。

要避免脱离特色的组织。特色是行业特色高校的生命线,行业特色高校的不可替代性,恰恰在于其优势特色学科领域的行业积累。因此,提升组织化程度必须坚持“特色立校”不动摇,不能效仿综合性大学的大而全,要聚焦优势特色领域,做到有所为、有所不为,通过强化特色,形成并巩固自身服务国家、服务行业的竞争力和贡献力。

(作者系中国地质大学(武汉)科学技术发展院副院长)



中国大学评论

近年来,关于各阶段考试的命题,大家有一个很明显的感知——不但语文和英语试卷的阅读量大幅度提升,而且数理化生等学科的题干也在“情境化命题”的导向下变得越来越大。这缘于真实科研往往基于错综复杂的场景,从复杂场景中提炼问题、解决问题的能力就显得至关重要。因此,要鼓励学生拓宽阅读面、提升理解力,考查学生信息提取、情境解读与综合应用能力,规避死记硬背的应试弊端。

显然,这样的改革方向和初衷是可以理解的。但在笔者看来,通过大幅度加大文本阅读量以体现学科素养,需要警惕将学科素养的选拔异化为一场关于“阅读速度”的竞赛。

一方面,阅读速度的快慢可能并不能很好地反映学生学科思维水平的高低。加大阅读量,让学生在短时间内阅读大量材料,可能会引导学生在日常学习中放弃精读与思辨,转而追逐速读技巧与材料组合,这与在高等教育阶段培养拔尖创新人才的目标背道而驰。

究其原因,拔尖创新人才的一个重要特征是面对具体问题,可以通过长时间的专注思考将其抽象化与结构化并生成原创性思想,而快速掌握很多信息并不是拔尖创新人才的重要特征。因此,如果各科考试以阅读速度作为选拔标准之一,无形中就会在考查素养里加入一个与创新能力弱相关的变量,而且权重极高,进而可能弱化拔尖创新人才的选拔。

比如,一位学者在攻克学术难题时,需要对复杂文献反复研读、质疑与推演。这样创造性的工作,要求阅读主体拥有“慢下来”的能力。如果我们的考试,特别是那些对学生意义重大的考试(比如高考)过度加码阅读量和时间压力,事实上就会奖励那些反应快、浏览多却未必想得深的学生。至于那些阅读速度不快,但习惯于在学科问题中深入挖掘、反复推敲、建立独特认知框架的学生,当重要的人才选拔考试要求其必须快速处理海量文本时,他们的深度思维优势不仅无法展现,反而可能因为阅读慢而被淘汰。尤其是一些极具批判性思维的学生,在阅读很长题干时,可能会对题干产生一系列质疑,甚至误解了出题者意图。如此,是否会系统性地误判甚至埋没此类潜在的创新者,这值得我们思考。

另一方面,泛阅读化可能会稀释各学科本身素养考查的区分度。各学科有其独特的思维方式和学科语言,比如数学重要的是符号推理与抽象想象。但当试题材料过长、情境过于复杂时,学生首先面临的不是学科问题本身,而是如何快速从冗余信息中提取有效信息。这种“情境化”倘若过度,就会使考查重点从“用学科知识解决问题”变为“先快速读懂题干材料再说”。这模糊了学科边界,也扭曲了考查学科素养的目标。

以高考为例,高考命题希望通过“情境化”来“反套路”,但若破解了“题海战术”,却迎来“速读技巧”,那不过是用一种新套路取代了旧套路。结果试卷区分的不是谁的学科素养更高,而是谁更擅长快速阅读。这对学科素养高但阅读节奏偏慢的潜在拔尖创新者,尤其不公平。

笔者一向认为,以高考为代表的选拔性考试,其公平性来自“分数面前人人平等”。但个体在考试时间节点的阅读速度受从小积累泛读量的影响很大。弱势家庭孩子即便在数理化生上天赋与潜质极佳,却可能输在“题目没读完”上。当各科都开始比拼阅读速度,其分数信号里混杂的家庭背景因素就更多了,这样的选拔对出身寒门的拔尖苗子极其不利。可见在各科考试中加大阅读量很有可能会进一步加大对弱势家庭孩子不公平。

基于上述分析,对于那些对学生具有重大意义的选拔性考试,笔者认为需要控制其数理科目的题干长度。数学、物理、化学、生物等科目的题干应力求简洁精准,情境创设服务于学科考查目标,而非喧宾夺主。如果一道题需要学生阅读超过一定字数的材料才能作答,命题者应当认真考虑:这是否偏离了学科考查的初衷?

同时,要区分“阅读能力”与“学科素养”的考查边界。语文作为基础学科,承担阅读素养考查的主要责任是合理的;而其他科目应聚焦于本学科的核心素养,避免因抬高文本阅读门槛而掩盖学科素养的选拔。

特别要警惕“情境化命题”对选拔公平性的侵蚀。正如前文所言,情境化、综合化符合当下的考试改革方向,但其前提是要考虑不同教育背景学生的适应性,确保形式创新不成为新的不公平来源。

总之,作为我国最重要的人才选拔机制,标准化考试的命题导向对于塑造整体教育生态以及拔尖创新人才培养有着重要意义。而当前各学科普遍加大题干阅读量时,就要思考选拔标准从“学科素养”变为“阅读速度与阅读理解”的风险。因此,应该把学科还给学科,让阅读回归阅读,考试才不会在“素养导向”的旗号下,把可能最有原创潜力的那批学生筛掉。

提升科技攻关能力,行业特色高校要走“组织化”这条路

■李鹏飞

作为我国高等教育体系的重要组成部分以及国家重要战略科技力量,行业特色高校,特别是其中的高水平研究型大学长期扎根行业,在很多重要领域都积淀了深厚的科技资源和行业优势。然而,对标强国建设和行业转型发展的要求,目前此类高校仍存在科研布局谋划不足、科研力量分散、资源配置碎片化、攻关模式单一等突出短板。而在笔者看来,此类高校要想补齐短板,实现自身高质量发展,用好“组织化”这一招是一条必由之路。

三重因素提升组织化重要性

行业特色高校科技攻关的组织化是指高校以服务国家战略、回应行业需求为导向,主动组织谋划,通过整合校内外各类创新力量,促使分散的创新要素聚力协同,建成制、成体系开展科技攻关服务。

随着大科学时代的到来,一方面,各国科技博弈变得日益激烈复杂,关键技术的争夺也由单点比拼转向组织能力、综合实力的较量,重大科技任务正呈现交叉交融、团队协作、向“四极”加速演进的特点。然而,长期以来,行业特色高校优势学科相对集中,基础学科较为薄弱,科研模式更多的是以自由探索、自发推进为主,学科交叉难、团队规模小、协作机制松散,靠自身力量难以适应现阶段所需要的协同作战模式。

另一方面,行业特色高校因行业需求而生,伴行业发展而兴,但随着高等教育体制改革的深入,绝大多数行业特色高校脱离了原行业主管部门的管理,与行业的直接联系也逐渐弱化。加之高校综合化办学取向、行业产业加速转型等多重因素叠加,部分高校甚至显现出一定的“去行业化”倾向,论文、专利等科技成果“重理论、轻应用”,与行业场景脱节,难以精准回应行业发展中的真问题、真需求。

此外,相较于综合性大学,行业特色高校在科技人才、项目、平台等资源角逐中的赛道相对单一,特别是在争取重大科技任务时往