感的检测方法,研究者观察到6例参与者在治疗4个月后,此类细胞数量下降了20%~50%。单梁强调,这距清除个体病毒相去甚远,但证明触发CARD8警报确有帮助。单梁表示,该结果"呼吁开发更高效的蛋白酶激活剂"。单梁去年将其实验室迁至深圳医学科学院。

多家制药企业正致力于开发此类高效药物,即靶向细胞杀伤激活剂(TACK)分子。德国默克公司2023年2月22日在《Science Translational Medicine》发表的研究报告了一种化合物,其促进蛋白酶二聚化的效力比依非韦伦"高出1个数量级",并已将其开发为候选药物。该公司现正评估该药在未经治疗人群中的安全性及抗HIV活性,此为评估其对已用标准药物控制病毒者潜伏细胞库影响的前置步骤。

何大一正测试一种组合策略:将相关TACK分子与潜伏逆转剂(latency reversal agents,LRA)联用。后者是一类旨在耗竭病毒库但迄今效果有限的药物。20年来,研究者尝试用潜伏逆转剂刺激病毒复制以清除病毒库细胞,设想免疫系统随后摧毁这些细胞或使其破裂。但此"激活并清除"策略所用多数LRA对病毒库规模几无影响。

作为开发更优LRA的尝试之一,何大一实验室开发了一种抗体,可靶向潜伏感染的T细胞,并触发促病毒复制的细胞信号。在实验室测试中,研究者使用了来自接受治疗且病毒载量完全受控的感染者的免疫细胞,单用该抗体未能清除全部潜伏感染细胞。但何大一在CROI报告,加用TACK药物后,病毒RNA水平急剧下降,表明焦亡进一步耗竭了病毒库。参与单梁团队新研究的加州大学旧金山分校HIV/AIDS研究者Steve Deeks评价:"TACK的故事令人惊叹。"

何大一强调,LRA联合TACK或无需完全清除病毒库即可对HIV感染产生重大影响。他问道:"若将病毒库规模缩小至1/1000,会发生什么?"这或足以使停用抗HIV药物者通过自身免疫应答控制微量病毒复制,实现所谓的功能性治愈。

Hunt指出,TACK药物或有助于缓解感染者(即使病毒载量已被药物完全抑制)仍面临的HIV相关严重长期健康问题。即使在潜伏感染的细胞中,病毒仍以低水平持续复制,引发慢性炎症;长此以往,将增加心血管疾病、癌症及器官衰竭的风险。Hunt表示:"当前清除领域将TACK视为降低病毒库负荷的策略,但我认为这些分子或对炎症产生更直接、更早期的影响,这可能极为重要。"

(译自《Science》,2026,391(6790))

如何判断AI是否具备开展科学研究的智能?

新型测试评估大语言模型能否运用海量知识实现真正科学发现

文/Celina Zhao

多年来,人工智能(AI)研究者梦想开发能通过提出新问题、设计实验乃至执行实验来加速科学进程的工具。近期,大语言模型(large language models,LLM)已取得若干发现,部分AI开发者宣称这使我们更接近该未来。但尚不知道如何测试AI模型是否真能开展科学研究?

为寻求答案,研究者转向基准测试:用于评估AI能力并与其他模型比较的标准化问题或任务集。但科学的复杂性使评判其科研能力尤为困难。美国伊利诺伊大学厄巴纳-香槟分校计算机科学家Hao Peng表示:"模型拥有海量知识,但它们懂得如何运用吗?"

过去1年涌现数10项面向科学的新基准测试以回答该问题,但科学家尚未就最佳方法达成共识。其中最受欢迎者之一是2026年1月28日发表于《Nature》的"人类终极考试"(Humanity's Last Exam,HLE)。该测试采用2500道源自"人类知识前沿"的问题考验LLM。例如其中一题询问蜂鸟籽骨支撑多少对肌腱。HLE开发者、非营利组织人工智能安全中心研究工程师Long Phan表示:"我们希望构建仅长期深耕该

领域的专家才能回答的多样化数据集。”

HLE自2025年1月24日首次以预印本形式发布以来,已成为LLM的重要试金石——HLE得分现已成为AI公司彰显产品能力的常见谈资。HLE发布时,知名开发者OpenAI的o1模型以仅8.3%的得分位居榜首。2026年3月早些时候,Google宣称其最新科学推理模型Gemini 3 Deep Think创下48.4%的HLE新纪录。

但部分科学家指出,HLE诸多问题测试的是晦涩乃至琐碎的知识,而非开展有意义研究的能力。AI for Science公司Deep Principle创始人段辰儒质疑:“知晓世界上磷同素异形体有多少种颜色,如何助人实现科学发现?”

OpenAI研究者表示,他们开发了朝此方向迈进的新基准测试。2025年12月16日发布的FrontierScience借助700道化学、生物学与物理学问题,旨在识别“专家级科学推理”能力。部分问题类似数学与科学奥林匹克竞赛题目:通常基于简短场景、答案明确,OpenAI研究科学家Miles Wang称之为“纯推理努力的合理代理”。例如识别系列化学反应的产物。其他问题则基于博士科学家在实际工作中处理的复杂开放式研究问题,如推理修饰特定分子可能影响其性质的多种途径。

Wang表示,该基准测试的关键优势在于可验证性——这是公平测试的最重要特征之一。奥林匹克题目易于评分,而对于开放式研究问题,LLM因识别中间推理步骤而获分。截至目前,OpenAI自家产品GPT-5.2取得最佳FrontierScience成绩:奥林匹克题目正确率77%,研究挑战得分25%。

其他研究者认为这一巨大分差颇具启示性。他们主张基准测试应聚焦直接衡量AI开展现实世界研究的能力。这正是段辰儒及其合作者与FrontierScience同期发布的“科学发现评估”(Scientific Discovery Evaluation, SDE)基准测试的指导原则。该测试不提困难但孤立的问题,而是向AI呈现源自8项进行中、数据尚未发表的真实研究项目的1125项任务,关联43种研究场景。例如要求LLM推导如何将目标分子分解为更简单、市售可得的组分。模型评估不仅基于单个答案,更基于其整合完整项目的能力——在多步骤中提出、检验并完善假设。段辰儒表示:“我们确保回答每个问题都关联真实科学发现的微小片段。”

SDE得分显示,LLM正确回答单个问题的能力并不总能转化为完整项目的稳健表现,反之亦然。段辰儒表示:“知晓宏观前进方向往往比知晓特定分子的精确性质更重要。”该基准测试还发现,来自OpenAI、Anthropic、xAI和DeepSeek等不同供应商的顶尖模型常在同一最难问题上受阻。这一模式暗示它们可能遭遇相同局限,很可能因其在相似科学数据池上训练所致。

然而SDE方法仍仅捕捉科学工作流的片段。AI for Science初创公司FutureHouse推出的生物学导向新基准测试LABBench2,旨在测试面向科学的AI能否将项目从初始构想推进至完成论文。2月发布的该测试采用近1900项任务,评估所谓“代理型AI模型”(能独立完成多步骤任务的系统)执行文献检索、数据获取与基因序列构建等工作的能力。

目前结果喜忧参半。多数领先LLM在全文专利与实验室试验论文检索方面表现良好,但在LABBench2更复杂的任务上常遇困难,例如交叉引用多个数据库,或在密集论文中定位并解读特定图表或数据。FutureHouse商业衍生公司Edison Scientific的Jon Laurent表示,这表明迈向真正AI科学家的进展,部分也取决于改进模型检索与导航信息的方式。

研究者强调,基准测试不仅用于记录当前赢家。更严格的基准测试还可通过为LLM及其他AI工具提供新目标来驱动创新。Laurent表示:“基准测试的目的之一是领先时代,衡量潜在能力,并推动其发展。”

在诸多领域,或不存在衡量AI是否“擅长”科学的单一标准。美国佐治亚理工学院认知神经科学与AI研究者Anna Ivanova表示:“这正是我们看到所用基准测试高度异质的原因。系统绘制数据的能力与其分析化学事实知识截然不同——尽管科学家可能两者都需要。”

鉴于科学所需技能的广泛性,AI专家认为研究界或宜依赖测试组合,每项测试针对并催化科学工作流不同环节的改进。Wang表示:“我们正迈向需要更多元化评估体系的世界。”

无论采用何种方法,被衡量的内容很可能引导改进方向。Peng表示:“要取得进展,你必须能够衡量它。”

(译自《Science》,2026,391(6790))

(科技新闻栏目编辑 黄文光)