编辑 张茜琦 美编 张彬 责校 李爽 电话:64101880 kuaijiaogao@163.com

我们需要怎样的大模型测评

国产大模型又多一份测评，这次的状元是讯飞星火。近日，《麻省理工科技评论》从多维度全方位检测，力图评出“最聪明”的国产大模型。最终，讯飞星火认知大模型V2.0荣获“最聪明”的国产大模型称号。

国产大模型竞赛如火如荼，好像每一个大模型都很牛，但具体牛在哪又始终缥缈，由此大模型测评应运而生。但这又可能注定是一件要“烧情怀”的事，它同样面临着“开源”还是“闭源”的两难选择，和刷题与竞价排名的诸多争议。

武林大会

国内“千模大战”下，谁是最聪明的大模型《麻省理工科技评论》最新发布的大模型评测报告。该报告称从研发和商业化能力、外界态度以及发展趋势等维度全方位检测大模型的能力，最终，讯飞星火认知大模型V2.0以81.5分的成绩登顶，荣获“最聪明”的国产大模型称号。

8月15日，科大讯飞发布“讯飞星火认知大模型V2.0版本”，科大讯飞董事长刘庆峰介绍，从业界参考测试集上的效果对比来看，星火V2.0基于Python和C++进行代码写作能力已高度逼近ChatGPT，差距仅分别为1%和2%。

刘庆峰说，到10月24日星火大模型代码能力全面超越ChatGPT，明年上半年将正式对标GPT4。

讯飞星火像是一个缩影。过去这段时间，大模型频繁更新让人眼花缭乱，动辄千亿的参数、各种专业术语也让人不明觉厉。但人们似乎很难找到一把统一的尺子，公平、客观、直观地感知大模型真正的效果，而不被纷杂的信息流所蒙蔽。

天使投资人、资深人工智能专家郭涛对北京商报记者分析称，“大模型是一个非常复杂的系统，它由大量的数据和算法组成，在训练和推理过程中需要考虑很多因素。对大模型进行测评可以帮助我们更好地了解模型的性能和特点、评估价值和意义、局限性和潜在风险等，从而为大模型的发展和应用提供有力支持”。

深度科技研究院院长张孝荣将测评形容为一场“武林大会”，要试试各家身手。他对北京商报记者分析称，大模型涉及到庞大的参数和复杂的算法，对于性能和效果的评估十分重要。通过测评可以大致地了解大模型的性能、稳定性、准确性等内容，为用户选择合适的大模型提供参考。

测评开始补位。今年3月，真格基金以投资者的身份入场，设计了一套大模型测试集Z-Bench。高校也是测评的中坚力量，例如清华大学、上海交通大学和爱丁堡大学合作构建的面向中文语言模型的综合性考试评测集C-Eval。

有媒体报道，5月以来，10多家国内外多家调研机构、权威媒体和高校等发布大模型评测报告，包括新华社研究院中国企业发展研究中心发布的《人工智能大模型体验报告2.0》、天津大学和信创海河实验室发布的《大模型评测报告》、国际数据公司IDC发布的《AI大模型技术能力评估报告，2023》等。

标准难统一

当该有测评成为共识，迎面而来的下一个问题就是，我们需要一个怎样的测评。

《麻省理工科技评论》提到，评测使用的测试集包含600道题目，覆盖了语言专项、数学专项、理科综合、文科综合、逻辑思维、编程能力、综合知识、安全性共8个一级大类，126个二级分类，290个三级标签，并针对问题的丰富性和多样性做了优化。

此前IDC则在测评中将大模型分为三层，服务生态、产品技术以及行业应用，对每一层的能力进行测评，主要考察指标为算法模型、通用能力、创新能力、平台能力、安全可解释、大模型的应用行业以及配套服务和大模型生态等，具体包括36项细颗粒度的评估标准。

对于大模型测评的必要性，元语智能联合创始人兼COO、SuperCLUE联合发起人朱雷提到，模型测评基准是通用人工智能的基石，没有测评就意味着没有目标，很难准确地判断究竟哪些做得好哪些做得不好，同时对于AI的安全性也无法把控。从国际视角上看，对于大模型的测评也是没有绝对标准的，因为大模型发展太快了。但在国内要做出一个客观公正的评测基准，也会遇到很多阻力。

北京市社会科学院副研究员王鹏对北京商报记者分析，目前大模型尚属新兴事物，国际上还没有一个覆盖面非常广、能够得到大家公允的评估方法或整套指标体系，需要加强国际合作，形成广泛共识。

“但这也会面临一定的问题，即大模型本身类型繁多，通用还是专用、垂类还是跨行业、偏技术还是偏应用等区别也会带来一定的阻碍，因此更需要权威机构加强研究，尽快形成共识，促进技术进步和行业发展。”王鹏称。

在他看来，一个合格的测评，应该由四个维度组成。首先是技术本身，包括稳定性、效率、效果等；其次是与行业的结合，在行业应用中是否有效果、成本是否可控、是否能够形成商业闭环；再次还要考虑是否安全可控；最后要从社会及行业认知角度，了解其在行业中的关注度，毕竟“酒香也怕巷子深”。

张孝荣也提到，由于大模型涉及的领域和应用非常广泛，不同领域、不同应用的大模型需要关注的指标和评估方法不尽相同。因此，针对具体应用领域和需求，不同机构和组织可能会提出不同的评估标准和方法。“尽管没有统一的标准，但测评的意义在于提供了一种评估和比较不同大模型性能和效果的方法，帮助用户选择适合自己需求的大模型。”

测评还是营销

“测评的意义侧重于营销推广。”张孝荣还提到了一个观点。

360智脑产品资深专家葛灿辉在引用SuperCLUE测评结果的时候，提炼出了一句总结：“360智脑”多项能力位列国产大模型第一。《麻省理工科技评论》的测评报告，传播最多的也是“讯飞星火被评为中国‘最聪明’的大模型”。

更早些时候，刷屏的是百度。比如IDC的大模型报告中，“百度文心大模型3.5获多项满分”，清华大学新闻与传播学院沈阳团队发布的《大语言模型综合性能评估报告》中，百度文心一言在三大维度20项指标中综合评分国内第一，超越ChatGPT。

每每涉及榜单，榜首归谁总是容易成为话题中心，从这个角度上看，测评本身或许就带着些营销的天然属性。但也正是如此，延伸出了一些不容忽视的问题。

“SuperCLUE出6月榜单的时候，第一时间就有人指责我们是不是收了360的钱，但事实是，直到这次沙龙，我们与360智脑产品负责人才有了第一次接触。”朱雷如此说道。

事实上，大模型测评同样面临着“开源”和“闭源”的两难选择。朱雷称，大模型测评题集也有开源闭源之分，但开源的题目就会面临受试者提前训练进而刷分“打榜”的可能，而闭源的题目就会陷入到是否有暗箱操作乃至竞价排名的争议。

朱雷表示，SuperCLUE还是选择了闭源的测评路线，但不是任何机构都可以闭源的，之所以公众较为相信SuperCLUE的测评结果，主要还是基于过去四年CLUE社区对中文语言模型的贡献和公信力。

据了解，CLUE开源社区发起于2019年，旨在建立科学、客观、中立的AI评测基准，过去几年CLUE社区分别建立了ZeroCLUE、FewCLUE等知名的语言模型评测基准，又于今年5月发布首个中文通用大模型综合性评测基准SuperCLUE。

SuperCLUE分为SuperCLUE-Opt、SuperCLUE-LYB琅琊榜以及SuperCLUE-Open三个不同维度的评测基准，相辅相成。据介绍，SuperCLUE目前也是中文AI领域最完整的综合性测评基准，同时也是罕见的“闭卷”考试。

“我们暂时还没有找到折中的方法，所以决定先‘保密’，大模型厂商不知道我出了什么样的问题，自然不好刷分。至于‘保密’带来的黑盒化，目前来看还是一个两者不可兼得的问题，但我们坚信自己的第三方中立性，评测的结论也是十分科学的。”朱雷称。

王鹏分析称，任何一项评估或排名，都可能面临一些问题，但这其实相当于一个“否定之否定”的过程。首先评估体系本身并不是完美的，需要不断优化提升，应对大家可能产生的质疑。

其次，专业的评测机构、技术机构等，也要注重自己的口碑，建立完善的体系，储备丰富的经验，有较好的技术团队和技术储备，作出更加客观公允、公平公正的评价。“因为一旦出现‘人情分’等问题，不仅会影响自己的声誉，也不利于行业的未来发展。”王鹏称。

北京商报记者 杨月涵

聚焦 Focus

严禁AI开处方 北京为互联网诊疗划红线

8月21日，据《北京日报》报道，北京市卫健委日前牵头组织制定了《北京市互联网诊疗监管实施办法（试行）》（以下简称《办法（试行）》），并向社会公开征求意见，公众可于9月16日前向市卫健委反馈意见。《办法（试行）》提到，医疗机构开展互联网诊疗活动要加强药品管理，严禁使用人工智能等自动生成处方，严禁在处方开具前向患者提供药品。

根据《办法（试行）》，医师接诊前需进行实名认证，确保由本人提供诊疗服务。其他人员、人工智能软件等不得冒用、替代医师本人提供诊疗服务。另外，互联网诊疗实行实名制，医疗机构应当告知患者，其有义务向医疗机构提供真实的身份证明及基本信息，不得假冒他人就诊。患者就诊时应当提供具有明确诊断的病历资料，如门诊病历、住院病历、出院小结、诊断证明等，由接诊医师留存相关资料，并判断是否符合复诊条件。

医疗机构开展互联网诊疗活动要加强药品管理，处方应由接诊医师本人开具，经药师审核合格后方可生效，严禁使用人工智能等自动生成处方；处方药应当凭医师处方销售、调剂和使用；严禁在处方开具前，向患者提供药品；严禁以商业目的进行统方。

据悉，北京市卫健委将建立北京市互联网诊疗监管平台，对开展互联网诊疗活动的医疗机构进行监管。医疗机构应当主动与平台对接，及时上传、更新相关执业信息，主动接受监督。市区卫健行政部门要及时向社会公布依权限批准开展互联网诊疗的医疗机构和互联网医院的名单、服务入口及监督电话或者其他监督方式，及时受理和处置违法违规行为。

此次实施办法是在去年政策基础上的进一步落地。早在去年6月，国家卫健委就和国家中医药局联合发布《互联网诊疗监管细则（试行）》提到，医疗机构开展互联网诊疗活动，处方应由接诊医师本人开具，严禁使用人工智能等自动生成处方。

在《办法（试行）》第一章总则中，北京市卫健委方面也表示，《办法（试行）》是为进一步规范互联网诊疗活动，加强互联网诊疗监管，根据《医师法》《中医药法》《互联网诊疗监管细则（试行）》等法律法规和规定要求，结合北京市实际制定的。

过去，互联网诊疗存在一些灰色地带，如网售处方药平台随便上传一张照片就可以开出处方药，严禁使用人工智能等自动生成处方等措施保证患者用药安全的同时，让行业整体的规范化全面提升。微脉互联网医院与平台中心总经理吴子威曾对北京商报记者表示，全面厘清互联网医疗行业医疗、药品和技术服务的边界，将打破“问诊成了卖药”的混沌局面，互联网医疗也将回归“严肃医疗”的本质。

北京大健康法商团队邓勇教授表示，现实中，一些平台选择AI开处方、客户直接取药的模式，跳过传统的处方开具、审核环节，把开方直接变成了卖药。这类行为严重违反我国药品管理制度，也给患者用药安全埋下了风险隐患。

邓勇表示，要让“严禁使用人工智能等自动生成处方”的规定落到实处，必须考虑多重因素，明确人工智能等主体的地位。目前的《处方管理办法》《医疗机构管理条例》等规定处罚对象局限于“人员”，后续应该考虑到人工智能等主体的特殊性，对仅由人工智能等自动生成处方的，应当与“使用非卫生技术人员”一样，直接追究医疗机构的责任。

北京商报记者 姚倩

吉利出手叫停沃尔沃“夺权”

沃尔沃汽车大中华区销售公司总裁钦培吉“下课”三天后，吉利出手叫停沃尔沃“夺权”。

8月21日，吉利汽车集团（以下简称“吉利汽车”）宣布，原沃尔沃汽车大中华区销售公司总裁钦培吉将加盟吉利汽车，出任吉利汽车销售公司副总经理并担任渠道发展委员会主任，分管吉利汽车渠道发展和建设，向吉利高级副总裁林杰汇报。

事实上，该公告发布前三天，沃尔沃汽车发布的一封内部信显示，钦培吉将卸任沃尔沃汽车大中华区销售公司总裁，由沃尔沃汽车日本总经理Martin Persson（潘鹤松）接任。对于钦培吉的去向，沃尔沃汽车方面仅表示“寻求外部发展”。按照车企惯例，在高管职位调整或离职的表述上，如果无赞扬或“祝好”等表述，通常为“非正常”离职。

内部信曝光后，外界猜测“作为‘沃尔沃系’的潘鹤松回归，或许为沃尔沃汽车瑞典总部方面在大中华区进行收权”。

值得注意的是，引发“收权”猜想的不仅是沃尔沃汽车对于钦培吉离职的表述，更来自于潘鹤松到任后汇报工作对象的调整。沃尔沃汽车在内部信中提到，潘鹤松将在继任后直接向沃尔沃首席商务官兼副CEO安伯扬汇报，而钦培吉在任期间则为向袁小林汇报。同时，对于袁小林，内部信称其职务不变，并向沃尔沃汽车首席执行官骆文襟汇报，但工作重点变为重点与JIM、EMT和董事会成员合作，处理政府关系，利益相关者合作，与吉利控股集团战略合作及在大中华区的战略投资等重要事务。

然而，在本次吉利汽车方面的正式官宣中，除钦培吉加盟吉利汽车外，明确潘鹤松将向“吉利系”的袁小林汇报，袁小林则直接向骆文襟汇报，工作重点也并无调整。这意味着，吉利方面出手夺回对沃尔沃汽车大中华区的控制权。

沃尔沃汽车与吉利汽车已多次上演“权力游戏”。

事实上，本次人事调整也是骆文襟出任沃尔沃汽车CEO以来对大中华区销售业务的首次重大调整，而调整的目标直指电动化。去年1月，骆文襟出任沃尔沃汽车总裁兼CEO。彼时，骆文襟表示：“作为全球最大新能源汽车生产国和最大的市场，中国是沃尔沃电动化和智能化转型以及创新的沃土。”

数据显示，今年前7个月，沃尔沃汽车在中国市场销量为9.27万辆，同比增长8%，较其全球市场销量18%的增幅明显偏低。今年7月，沃尔沃汽车在华销量同比下降8%，其中Recharge系列（包括纯电动汽车和插电式混合动力汽车）的占比为8%。反观，今年7月沃尔沃汽车在欧洲市场的销量超2万辆，同比增长28%，其中Recharge系列车型销量占其欧洲市场总销量的56%，纯电动汽车占比为13%。

中国汽车流通协会专家委员会成员颜景辉认为，中国作为全球新能源汽车销量大国，沃尔沃的在华销售销情况却不如欧洲市场，这或许也让急于电动化转型的沃尔沃汽车方面对大中华区不满，因此开始对管理层进行调整。

北京商报记者 刘洋 刘晓梦