

缺乏知识图谱构建能力的ChatGPT*

徐英瑾

(复旦大学哲学学院,上海 200433)

摘要:人类的知识图谱是基于二值化的逻辑框架的,基于这种框架的人类知识归赋活动自然能够在“知”与“无知”之间划出清楚的界限。经由大量语元“预训练”而运作的ChatGPT至多只能做到对人类知识图谱的一种统计学模拟,这就在根底上使得其难以避免“机器幻想”问题。不过,只要对ChatGPT的基础原理与技术极限有清楚的意识,学术界依然可以利用ChatGPT大大提高科研效率。在学术专家给以积极语言提示的前提下,ChatGPT能够在展现具体科研路线图方面成为人类的学术参谋。ChatGPT的出现显然会倒逼我国的学术培养机制更多关注学者宏观知识图谱构建力的培养,并刺激中国的人工智能界开发能够弥补ChatGPT缺陷的通用人工智能产品。

关键词:ChatGPT;大语言模型;知识图谱;通用人工智能;计算生物学

中图分类号:B81

文献标志码:A

文章编号:1001-862X(2023)05-0081-008

DOI:10.16064/j.cnki.cn34-1003/g.0.2023.05.002

自美国 Open AI 公司的大型语言模型 ChatGPT-3.5 及其更新版本 ChatGPT-4.0 上线以来,其引发的舆论风暴已经持续数月。然而,大多数相关讨论主要涉及 ChatGPT 的商业落地价值与一般公众对该模型的使用体验,罕见各行专家对 ChatGPT 的反应。此外,一些关于高校学生利用 ChatGPT 作弊的报道,也会误导公众认为 ChatGPT 对严肃的科研毫无用处。其实,国内科学界已经开始运用 ChatGPT 助力科研,并获得了比较丰富的心得(详后)。在客观概括 ChatGPT 带给学术科研的利弊的基础上,可发现:若要尽量规避 ChatGPT 技术的缺陷并发挥其长处,人类必须努力提高相关知识领域宏观知识图谱的组织能力。有鉴于宏观知识图谱的培养并不是我国科学教育的强项,以 ChatGPT 为代表的大语言模型的出现或许会倒逼我国学术培养机制产生与之适应的巨大嬗变。

而要论证这一观点,首先要对 ChatGPT 的工作原理有所了解。

一、从 ChatGPT 的工作原理谈起

ChatGPT 的工作细节非常复杂,美国 Open AI 公司仍未公开大量技术详情。但根据已公开的资料,依然可以大致了解其工作流程。

ChatGPT 的工作流程大约分为预训练(pre-training)、有监督微调(supervised finetuning)、奖惩模型(reward modelling)和强化学习(reinforcement learning)这四个阶段。这四个阶段又可以看成两个阶段:预训练与人工调整(即上述第二至第四阶段的合并)。在预训练阶段,系统通过网络获得大量语料,并将其转换为“语元”(token)的聚合,然后通过熟悉这些语元之间的联系规律,做到能大致预测一个语元出现后紧跟的语言究竟是什么。比如,经过大量的“语元轰炸”之后,一个系统就能在看到词组“不但”之后预测另一个词组“而且”出现,尽管它并没有系统地学习过人类句法与词法。当然,互联网上得来的语料往往充

本刊网址·在线杂志:www.jhlt.net.cn

* 基金项目:国家社会科学基金项目“对于通用人工智能与特定文化风土之间关系的哲学研究”(22BZX031);国家自然科学基金项目“探索研究 AI 伦理对科研环境的影响”(L2124040)

作者简介:徐英瑾(1978—),福建福州人,哲学博士,复旦大学哲学学院教授、博士生导师,主要研究方向:科学哲学、人工智能哲学。

满“杂质”(如错别字、不当表达,等等),因此,仅仅经过预训练阶段的系统是无法投入使用的。为了使得系统堪用,研发者需要进入复杂的人工调整阶段来清洗这些“杂质”。在此类“清洗”的最高阶段,Open AI 公司还对 ChatGPT 的价值输出进行了强行规范,即通过强化学习禁止系统输出带有诸如种族歧视、性别歧视等负面价值色彩的对话内容。

从表面上来看,ChatGPT 的基本运作逻辑与我们熟悉的“汉字联想输入法”没有本质的区别:根据统计学规律,汉字联想输入法也能在用户输入一个字或词后“联想”到紧接着最有机会出现的字或词是什么。不过,主要通过以下两项措施,ChatGPT 体现了过去的类似软件无法企及的强大功能。措施一:ChatGPT 预训练阶段所消耗的资源惊人,需要 10000 个以上的高效能图形处理器(GPU)、数月的训练时间,以及 570TB 的海量原始训练语料。在如此大的资源的“喂养”下,系统获得了某种让人惊讶的“涌现能力”(指某种复杂系统纯粹因为自身规模的变大而突然获得的某种它原本不具备的能力)。措施二:ChatGPT 所使用的“转换器”(transformers)架构使用了一种叫“注意力”的新机制,使得系统能够处理很长的句子,由此具有强大的自然语言处理能力(比如系统在看到一个句子里出现“不但”之后,隔了很多个字还能预测后面会出现“而且”)。

从上述对 ChatGPT 的工作原理挂一漏万的介绍中,不难看出,ChatGPT 能够做的,便是根据统计数据对人类用户的提问给出尽量符合一般规范的应答。然而,按照统计学数据所给出的那些貌似符合规范的语言输出,是不是真的符合规范呢?答案是否定的。概言之,任何人类的规范(从语言规范、伦理规范到科研规范,不一而足)都具有“极性”(polarity)特征,也就是说,在“符合规范的行为”与“不符合规范的行为”之间是有着清楚的楚河汉界的。而统计学的处理方式往往会淡化这种极性特征,由此使得不同性质的行为之间的区分被模糊化——或用黑格尔式的哲学术语来说,统计学思维方式具有将“质的规定性”还原为“量的规定性”的天然趋向。由此,有些必须在极性思维中才能呈现的知识框架,经由统计学“祛极化”处理而模糊。上述一般性问题的在 ChatGPT

这里的具体展现方式则是这样的:在预训练阶段,知识被打碎为海量的语元而被系统吸收,因此,各条知识本身所带的结构性信息已经被破坏。这就好比破壁机对水果所做的事情:用刀片打破食材中细胞的细胞壁,将细胞中的维生素、矿物质、植化素、蛋白质和水分等成分充分释放出来(这些被释放出来的成分就可类比脱离语言以及知识结构束缚的“语元”)。当然,就 ChatGPT 而言,其内含的“转换器”工作机制的确能够帮助系统部分恢复已经被破坏的语言材料中的结构性信息,但这一有限的成功并不是基于对知识框架自身的理解,而是基于对这些框架信息统计学模拟后的产物。

下面这个比方或许能够帮助理解:一个记忆力超好却缺乏现代科技知识的部落酋长反复观摩一个来自文明世界的游客如何将一堆零件拼装成摩托车。该酋长在观摩了几千次后,自己也动手将这些零件装配成了摩托车,而且也貌似成功了。但是他只是机械地记住了游客装配这些零件的次序,而没有了解摩托车的内部机理。因此,他无法大胆地改变装配摩托车的次序(除非他观察到游客已经做出了这种改变),遑论将装配本田牌摩托车的经验延展到宝马牌摩托车的维护上(除非他观察到游客已经做出了这种拓展)——尽管对部落里的其他成员来说,他貌似是懂摩托车的。也正因为如此,当人类用户向 ChatGPT 提问“何为费马大定理”时,系统并不真正理解何为费马大定理,而只是根据其在预训练阶段所获得的语料来将最大概率切中问题之指向的文字展现在对话框中。由于经过反复训练的 ChatGPT 所给出的应答的确往往是数学教师针对此类问题的典型应答方式,因此,这就会给人类用户带来错觉:该系统是了解数学的。

很显然,无论与那位真正懂摩托车的游客相比,还是与真正的数学家相比,ChatGPT 是缺乏相关知识领域的知识图谱的。什么叫“知识图谱”(knowledge graph)呢?概言之,这是对各个知识节点之间关系的图像化表征——比如根据美国作家波西格(Robert M. Pirsig)对摩托车各个部件之间关系的描述给出的知识图谱(图 1)^{[1]53}。

很显然,知识图谱的典型构建方式是基于离散的逻辑符号的,因此,在图谱中出现的各个知

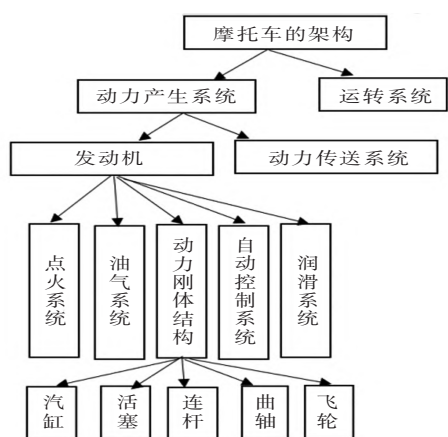


图 1 与摩托车维修相关的知识图谱

识节点之间都存在着质的差异。按照上述知识图谱,维修工关于如何修理摩托车的知识都是储存在其记忆库的不同“抽屉”里的,而“抽屉”与“抽屉”之间的界限则一般是泾渭分明的。因此,当一个合格的维修工在维修汽缸的时候,他非常清楚汽缸乃是发动机的动力刚体结构的一部分,肯定不是发动机的润滑系统的一部分,更不是摩托车的运转系统的一部分。与之相较,那位酋长也好,ChatGPT 也罢,他(它)们所能做的,便是根据历史上他人维修汽缸的步骤以便固定下面的预测:维修汽缸时不会涉及润滑系统的维修。这一答案貌似也是对的,但需要注意的是,此类基于机械记忆与统计数据的工作方式很难容忍这样的灵活性:一个老练的维修工可以由于某原因(如发现汽缸的某部件还未到货),在维修汽缸仅仅到一半的时候就转而去维修润滑系统。很显然,只要他将上述的知识图谱牢记心中,他就不会因为上面的变通而导致维修步骤的混乱。而对于 ChatGPT 来说,只要这样的变通缺乏历史数据的支持,就不会得到系统自身的运作方式的允许。

虽然在一般的研究领域,研究者所面对的知识图谱的拓扑学结构是大致稳定的,但在前沿领域,研究者必须大量改变现有知识图谱的结构才能获得新的科研洞见。在科研史上值得一提的一个案例便是“费马大定理”的证明(该定理的内容:若两个变项的 n 次方之和等于第三个变项的 n 次方,那么,当 n 取大于 2 的整数时,这三个变项都无法取正整数解)。自 1637 年费马提出该定理以来,如何证明该定理长期被当成一个代数问题,但在 1955 年,日本学者谷山丰将该问题与椭

圆曲线问题相结合,由此改变了解决费马大定理问题的知识图谱。尔后,顺着这一思路,在 1995 年,英国人安德鲁·怀尔斯(Andrew Wiles)终于证明了费马大定理。由此看来,对知识图谱的恰当改变恰恰能够在科研中起到四两拨千斤的作用。

对人类来说,更改旧知识图谱需要高超的联想力,比如在谷山丰之前别人很难获得那种将椭圆曲线问题与费马大定理加以联系的联想力。这显然又需要图谱修正者对问题本身具备极深的理解力。与之相较,基于既有的人类语言表述的统计学趋势来运作的 ChatGPT 至多能获得正常的知识图谱的一副统计学模拟图,不具备修正既有知识图谱的能力。指望 ChatGPT 具备实质上的科研创新力,显然是缘木求鱼。

不过,从另一方面看,只要 ChatGPT 具备的语言训练材料足够多,它至少可以假装自己成为常规知识图谱的拥有者。譬如,只要训练者将前述关于摩托车维修的知识图谱转换为特定的语元次序输入系统,系统就能在人-机问答中带给用户一种它是“摩托车维修专家”的印象。此外,只要训练者将有关“溯因推理”的数据输入系统,ChatGPT 也能部分模拟人类科学家进行溯因推理。从这个角度看,只要人类科研者对 ChatGPT 的能力上限有清醒的意识,恰当使用 ChatGPT 依然能在相当程度上提高科研的效率。

二、如何通过知识图谱力的提升来更好地利用 ChatGPT

由于 ChatGPT 本质上是一个大语言模型,因此,科研活动中与语言表述相关的部分(如在给定思路的前提下撰写论文、调整文献格式、润色文字、翻译相关文献)显然能够得到 ChatGPT 的帮助。另外,也恰恰因为 ChatGPT 本质上仅仅就是一个大语言模型,所以,指望它能代替自然科学家去做科学实验或代替社会科学家去做田野调查乃是不切实际的。这也便是 ChatGPT 最容易为人类专家提供帮助与最不容易为其提供帮助的两件事——前者是指对已具雏形的学术报告进行语言再处理,后者是指以“涉身”的方式与物理实在或社会实在发生直接的关联。那么,在这

两件事之外,是否能够指望 ChatGPT 能代替专家对科研思路进行规划呢?

答案是肯定的。复旦大学的计算生物学家孙思琦指出,ChatGPT 能够就如何修正蛋白质结构的检测程序“阿尔法折叠 2”(AlfaFold2)的表现给出不错的建议。^①通过他的测试,ChatGPT 找到了“阿尔法折叠 2”既有状态的几个主要缺陷。

第一,阿尔法折叠 2 在检测单链蛋白质结构时表现不错,但是在检测多链蛋白质结构时表现不好。这是改进该程序的一个着手点。

第二,阿尔法折叠 2 在检测膜蛋白(也就是与细胞膜发生相互作用的蛋白)时表现不好,因为膜蛋白自身的疏水特性与双层脂质环境的复杂性使得膜蛋白的内部结构难以被预测。这亦是改进该程序的一个着手点。

很明显,系统之所以能够给出这样的回答,是因为它已经被“喂入”了科学界关于抱怨阿尔法折叠 2 程序之缺陷的数据,而不是因为它真正理解阿尔法折叠 2 是怎么回事。但即使如此,ChatGPT 的上述表达,的确也能帮助对该领域不那么熟悉的人类用户迅速获知改进阿尔法折叠 2 的大方向。

那么,是否能够期望 ChatGPT 帮助我们拓展阿尔法折叠 2 程序的用途呢?比如,这个原本用来预测蛋白质结构的程序,是否可以进一步用来预测比蛋白质更为复杂的核糖核酸的结构?很显然,这一发问改变了计算生物学领域的既有知识图谱。尽管作为大语言模型的 ChatGPT 无法自主修正此类知识图谱,但人类研究者显然是可以这么做的。而在人类研究者自己确定了这一试探性的研究方向之后,他们还可以利用 ChatGPT 对上述思路的具体可行性进行大致的排摸。提出这一新科研思路的孙思琦就在与人-机对话中得到了 ChatGPT 输出的关于如何具体执行上述科研路线的建议:

第一,由于核糖核酸的结构更复杂,预测其结构的步骤也要比预测蛋白质结构的步骤来得多。这是拓展阿尔法折叠 2 预测功能时需要谨记的。

第二,关于核糖核酸的既有高分辨率结构图的数据库要远小于关于蛋白质的同类数据库,这

就使得基于深度学习机制的阿尔法折叠 2 程序的训练过程变得相对困难。

基于上述判断,ChatGPT 立即就如何拓展阿尔法折叠 2 的功能提出了更具体的建议,如迅速拓展关于核糖核酸结构的数据库,并设计出一个能应对核糖核酸内部多层结构的新预测模型。孙思琦指出,其研究团队比对了人类科学家与 ChatGPT 关于此类问题的应答思路,发现二者大同小异。这就说明,在具有修正既有知识图谱意图的人类科学家的积极引导下,ChatGPT 系统可以被动地将不同知识模块中的知识加以比对(比如,比对核糖核酸的结构与蛋白质的结构),由此铺展出更富有细节的相关科研思路。

ChatGPT 甚至能够在减轻科研工作者的劳动负担方面走得更远。比如,它能够协助人类专家编程。复旦大学的化学家李剑锋指出,只要人类用户在与 ChatGPT 对话时说清楚所需要的程序的大致技术参数(比如,需要建立的神经网络有几层,输入的数据的大致特征如何,输出的数据各自具有怎样的语义内容,等等),那么,系统就能给出具体的编程结果^②,职业的编程人员只要对机器输出的程序进行微调,就能将相关程序投入运用。

不过,若要利用 ChatGPT 的这些积极价值,运用者必须具有比较高的学术素养。具体而言,一个会恰当运用 ChatGPT 进行编程的科研人员显然知道自己所需要的程序的大致特征,并有能力看出机器完成的编程作业的瑕疵,最后对其加以微调。这也便是人类科研人员高于机器的针对宏观知识图谱洞察力的体现。反过来说,要让这些洞察力积极引导机器进行相关的运作,研究者就一定要提高自己的主体意识,通过自己掌握的知识图谱将解决问题的步骤一步步展示出来。

高估 ChatGPT 价值的论者或许会说,目前 ChatGPT 的能力已经使得其在人类用户不向其提示自己所掌握的知识图谱的前提下,自行展开知识图谱。这就涉及 ChatGPT 团队的重要成员、美国华裔科学家魏杰森(Jason Wei)构建的“思维链”(Chain of Thought)技术。^②该技术的实质是:把一个多步骤推理问题,分解成很多个中间步骤,以便分配更多的计算量,生成更多的语元,然

后再把这些答案拼接在一起求解。那么,该技术是否真能帮助机器把握相关的知识图谱呢?魏杰森团队的研究论文就提到了这样一个实测记录:

人类问:罗杰有 5 个网球。后来他又买了两罐的网球,每个罐子里有 3 个球。现在他有几个球呢?

机器答:罗杰开始有 5 个球。2 个罐子里的球则一共有 6 个。5 加 6,得 11。罗杰目前有 11 个球。

上述问题显然还算简单。化学家李剑锋则提高了测试的门槛:他检测了 ChatGPT 解决“幻想怪兽过门”问题的能力(此类问题的研究对研究分子的可能构成方式具有指导意义)。

人类问:一只长着狗头、孔雀身体、猫腿、猪鼻、象尾的动物能否通过高 2 米宽 1.5 米的办公室大门?

机器答:如果我们假设这种幻想的动物的体积和形状是由其各个部分的标准大小所决定的,那么,该问题的答案就取决于各部分的大小。比如,孔雀的身体长度大约是 1 到 2 米,狗的头部一般小于 0.5 米,猫腿的长度一般小于 0.5 米,猪鼻与象鼻也相对较小。如果这种动物是站立行走,而不是爬行或用别的方式移动,那么,其高度就大致取决于孔雀的身体与猫腿的长度。这样的话,该动物的总高度可能会小于 2 米,宽度可能会小于 1.5 米,应该可以通过大门。

很显然,在上述两个案例中,得到“思维链”技术加持的 ChatGPT 的答案都是正确的,而且它在解决“幻想怪兽过门”问题时所展现的能力还貌似非常惊人。但需要注意的是,“思维链”技术在本质上依然是一种统计学技术,也就是说,系统在分解复杂问题时所依据的,依然是该问题通常被分解成更简单的因子的既有数据。这一技术的运作并不稳定,有时候甚至会产生荒谬的解答,如孙思琦提供的人-机应答案例:

人类问:我们公司今年赚了 300 万,去年亏了 500 万,那么,今年赚了多少钱呢?

ChatGPT 给出了两个版本的回答:

版本一:为了计算今年公司赚了多少钱,只要将今年赚的钱中扣除去年的亏损就成了。既然今年赚了 300 万,去年亏了 500 万,我们就要将 500 万从 300 万中扣除,得负 200 万。所以,今年公司

亏了 200 万。

版本二:为了计算今年公司赚了多少钱,只要将今年赚的钱中扣除去年的亏损就成了。既然今年赚了 300 万,去年亏了 500 万,那么我们就有了: $300 \text{ 万} - (-500 \text{ 万}) = 800 \text{ 万}$ 。今年公司赚了 800 万。

这两个答案都是错的。在第一个问题中提问者关心的是公司今年的赢利,而不是今年与去年的总盈利或总亏损。所以,人类给出的关于公司去年亏损的信息与解答问题无关。第二个答案可以说“错得离谱”,因为系统在亏掉的 500 万元之前画蛇添足地加了一个负号,由此与前面的减号互相抵消,使得亏损变成了赢利。

从 ChatGPT 上述的应答表现来说,它其实缺乏根据当下抛出的问题迅速勾勒新的知识图谱的能力。比如,它无法根据提问中包含的时间变量迅速切割公司今年的赢利状况与去年的赢利状况,由此将一部分信息排除出当下的信息处理范围。换言之,作为大语言模型的 ChatGPT 似乎坚信“世上万物是普遍联系的”,很难意识到世界上的某些事项与当下所面对的任务并不相关。而“思维链”技术的使用也使得系统更进一步难以割舍所面对的信息:既然该技术要求系统一步步地展开思维,系统就会天然倾向于运用提问中涉及的诸语元以方便展开思维过程,却由此忽略了检测这些语元与任务之间的相关性程度。

对事物之间非相关性的意识,本身也是知识图谱构建能力的一种体现。即使主张“世界上的万事万物是普遍联系的”的斯宾诺莎主义具有抽象的形而上学合理性,但任何旨在适应环境的智能体用以处理现实问题的资源依然是有限的,因此,这样的智能体就不能在构建一个针对特定问题的知识图谱时置入太多的要素。从这个角度看,ChatGPT 对于“非相关性”的不敏感性显然会降低系统在逼仄时间资源约束下的适应性。而此类不敏感性带给人类用户的直接感性体验便是:ChatGPT 的应答往往是啰唆冗长的,缺乏科学文献所一贯追求的那种简洁的文风(在这里不妨将“啰唆”视为与核心信息不太相关的语元四处泛滥的标记)。

ChatGPT 在排除非相关知识要素时显现的无能,自然也使得其无法具有苏格拉底式的自知:

自知自己对某些问题的无知。这也就导致了所谓的“机器幻想”问题,就是指大语言模型会自说自话地炮制不存在的书籍、法条与科学理论。因为其构造这些虚假信息的措辞本身是基于既有的海量数据的,相关内容还颇能迷惑一般公众。譬如孙思琦就发现,在向 ChatGPT 询问关于核糖核酸折叠的最新论文的情况的时候,系统就炮制出了一些虚假的学术信息(比如,“ESM 折叠程序”在第十四届结构预测竞赛中表现优异,远胜“阿尔法折叠程序”,云云)。根据有关报道,纽约的一个律师事务所甚至还因为轻信 ChatGPT 所编造的美国法律条文而受到法官的罚款。^[3]无独有偶,笔者在使用 ChatGPT 时亦发现,系统会在完全不知道日本哲学家西田几多郎的背景信息的情况下编造其著作名称。对缺乏相关领域的宏观知识图谱的用户来说,ChatGPT 在这些问题上所制造的误导在某些情况下或许会导致严重的后果。

从上文的讨论来看,科研工作者要在运用 ChatGPT 时做到兴利除弊,就一定要做到以下几点。

第一,对该领域的知识图谱有专家级别的理解能力,以便随时勘验系统的思维链展开形式的合理性。

第二,对系统给出的各种用户未知的新信息进行多重勘验,防止被“机器幻想”所骗。

第三,即使系统的输出在科学上是正确的,也要根据当下任务的关切点迅速把握其冗长应答中的学术亮点(需要注意的是,系统本身无法自动呈现这些亮点)。

第四,在机器编程等非常精专的领域,用户还得具备检查机器编程结果的专业能力。

虽然 ChatGPT 带有这样或那样的技术隐患,不过,在对这些问题高度警惕的前提下,对于它的恰当使用依然能够大大提高科研效率。具体而言,ChatGPT 虽然在聚焦学术问题核心关切点的能力上不如人类专家,但在思考问题的广度方面经常有不错的表现,可以成为助力“头脑风暴”的机器参谋。因此,人类与机器彼此扬长避短,可以更好地推进科学事业的进步。

若以整个中国学术界(而不是个别专家)的发展为利益出发点,我们又能通过怎样的措施来

切实促进人与机器这种良性协同关系的产生呢?

三、中国人工智能界如何面对大语言模型带来的机遇与挑战?

应当看到,中国的人工智能研究思路多年来模仿美国硅谷研发思路的特征明显,这一“跟着走”的策略也影响了国内大语言模型技术的研发思路。目前,国内清华大学研发的 ChatGLM-6B、复旦大学研发的 MOSS、深圳大学开发的“伶荔”(Linly)、元语智能团队的 ChatYuan、百度的“文心一言”,等等,均可以视为 ChatGPT 在国内的对应物。由于国内研发大语言模型的总体资源投入不如硅谷,而大语言模型自身的运行机制又极度依赖诸如数据与算力之类的资源的投入,所以,从总体上来看,国产大语言模型的总体技术表现是不如 ChatGPT 的。不过,这并不意味着中国的人工智能专家不能做出自己的贡献,可以努力的方向有四:

方向一:丰富大语言模型的伦理性的内涵,使得相关产品能够在这个维度上表现出比 ChatGPT 更佳的“用户界面友好性”。这当然不是说 ChatGPT 不具备伦理性,而是说研发 ChatGPT 的 OpenAI 公司狭隘地将“伦理性”理解为不允许系统输出包含歧视性内容的答案,而相对忽略“伦理性”的其他维度,譬如怎样通过精简的回答节省用户的时间,以及如何做到在知识指派问题上的诚实(即“知之为知之,不知为不知”)。与之相较,由复旦大学邱锡鹏教授团队研发的 MOSS 系统则在投入资源远不如 ChatGPT 的前提下(MOSS 的 GPU 投入只有 ChatGPT 的百分之一),秉持有用、诚实与无害的价值准则,努力做到系统的输出尽量精简、无错与“符合三观”。其具体措施是:MOSS 的研发者通过将大语言模型与搜索工具接驳以便系统能够将自己的输出与网络上能找到的“标准答案”彼此对齐⁽³⁾。这就大大提升了系统应对关涉冷门知识的提问的能力,在相当程度上消除了“机器幻想”问题。按照类似思路,具有中国特色的大语言模型的研发道路,应当为大语言模型与外部世界之间的关联提供更多的接驳节点,以便系统的输出能够得到多模态

的对齐。

方向二:以更适应中国民俗与法律框架的方式提高大语言模型的伦理性,使得相关产品能够更贴合中国文化的本土。之所以中国的大语言模型研究者需要做出这番努力,乃是因为 OpenAI 公司在研发 ChatGPT 时秉承的价值观乃是西方的自由主义价值体系。^[4]虽然这种价值体系的某些要素与我国的主流价值观有所重叠(如反对种族歧视、性别歧视等),但某些要素则未必。价值观的分歧会导致不同文化圈中民众在面对相同的道德难题(如著名的“电车难题”)时给出不同的应答。譬如,麻省理工学院主持的以“道德机器”为题的全球网络调查统计了来自 233 个国家或地区的被试者对不同版本的“电车难题”的反应方式,调查涉及的语言有 10 种(英、德、法、西、葡、俄、中、日、韩、阿),统计的决策数量有四千万次之多。^[5]研究清楚地揭示,东方文化圈的被试,对有待被牺牲的群众的人数相对不敏感,而对他们是不是遵照交通法行事的公民(这可能体现了东方文化圈对社会规则与社会权威的高度尊重)更敏感;西方文化圈中的被试则更倾向于什么都不做(这可能体现了带有个人主义色彩的义务论理念对西方文化的影响);而南方文化圈(包括拉丁美洲文化与非洲文化)的被试更倾向于保护女性(对于这一现象的解释,目前还有待研究)。从这个角度看,一个有用的大语言模型在面对诸如“电车难题”这样的道德两难问题时,需要根据用户的文化背景“看碟下菜”。不过,ChatGPT 过于严格地遵守“反歧视”这一伦理规则,反倒可能阻止其提供从某种文化的“成见”出发才显得有意义的答案。在这个问题上,中国的大语言模型研究者不妨通过人工微调的方法使相关产品的输出能够更多地体现一般中国人的伦理直觉,至少成为对中国人有用的决策参谋。同时,在涉及与中国法律相关的硬性社会约束条件时,中国的大语言模型研究者一定要通过强化人工对齐的措施来保证相关法律法规不被曲解。

方向三:不要贪大求全,争取在中文数据具有优势的地方做出带有专家系统色彩的大语言模型。与基于英文数据的大语言模型构建相比,中国学者的类似努力将不得不面对某种先天劣

势,即世界上大多数科技文献都是以英文呈现的。因此,抱着赶超 ChatGPT 的念头与西方同行竞争在大语言模型预训练数据量方面的投入,似乎并非决策上的优选。与其如此,不如进一步打造为专业人士准备的大语言模型,以直接助力专业领域内的科技进步。不过,关于如何建造这样的专家系统,则需要新的思路。一种比较普通的思路便是将特定的大语言模型与特定的专业直接挂钩,如生物制药方面的大语言模型 ProteinLM 与 MEGA-Protein、气象预报方面的大语言模型“盘古·气象”、流体仿真方面的大语言模型“东方御风”、医学问答方面的大语言模型“华佗”,等等。但此类与特定领域直接相关的大语言模型很容易遭遇国外同类产品的竞争,很难展现中国的优势,建议引入一个数码时代的新“四库全书”计划,即以几十年来中国文理学术期刊(如“知网”收录的所有期刊)与专业著作所包含的信息为预训练资料“喂给”大语言模型(同时大量减少未经专家认证的网络数据带给模型的负担),以便最终在“预训练数据的跨领域性”与“数据总规模的可控性”之间达成平衡。这样的新模型也会因为具有比一般专家类大语言模型更为宽广的知识域而有望成为协助用户进行交叉学科研究的知识参谋。

方向四:除了大语言模型开发路径,积极尝试基于小数据的通用人工智能进路。与现有的主流人工智能路径不同,通过自然演化而形成的人脑一般能够通过很小的能量与数据消耗得出高质量的决策结果。这就说明目前的主流人工智能路径未必是最“自然”的智能实现路径。尤其在有效处理海量数据的关键硬件被西方国家限供乃至断供的严峻政-商背景下,我们更有必要在主流人工智能之上尝试某种不那么消耗资源(包括算力资源与数据资源)的通用人工智能研发思路。在这里特别值得一提的是“纳思系统”,纳思系统是基于一种叫“纳思逻辑”的词汇逻辑的通用推理模型,能够自然模拟诸如归纳、类比等传统人工智能系统很难模拟的人类思维方式,并且能自然呈现任务导向对推理进程的影响。纳思系统的程序可以通过一般的商业计算机执行,不需要集成大量昂贵的 GPU。它对资源的低消耗则可

以通过如下案例得到非常有力的证明:根据英国学者斯罗伊斯(Dwane van der Sluis)的比较研究,基于纳思系统研发的语音识别系统 NUT 只需要 105 个训练样本就能获得 64% 的识别正确率,而基于传统深度学习路径的语音识别系统 ANAM 则需要 84843 个训练样本的“喂入”才能达到 93% 的识别正确率。^{[6]307-316}虽然 ANAM 的总体得分较高,但其“性价比”则明显低于基于纳思系统的 NUTS。运用于语音识别只是纳思系统的诸多运用方向中的一种,因此,完全可以设想作为自然语言处理系统的纳思系统也能在与主流大语言模型的比对中体现明显的“性价比”优势。

四、结 语

从康德到萨特,很多哲学家都指出,人的本性乃在于能做出自由的选择。不过,自由的选择的逻辑前提是,不同选择项之间的离散性。而离散性、两极性(特别是真假二值性)又恰恰是逻辑的特征,因此,与其说人类是一种必须做出自由选择的动物,还不如说人类是一种必须依赖逻辑而生存的动物——否则哈姆雷特就不会在离散逻辑的支配下说什么“生存还是毁灭,这是一个问题”。与之相匹配,人类的知识图谱也是以逻辑为基础构建出来的,而逻辑的清晰性也赋予了人类的求知活动以很多重要的特征,比如在某些情况下确切地知道自己不知道某些事情。与之相较,以 ChatGPT 为代表的大语言模型所依赖的统计学机制却使得这些模型只能成为人类知识图谱的一种模糊化的模拟——在很多时候,这种模拟会通过模糊有用信息与闲聊之间的界限而浪费用户的时间,并在更坏的情况下通过淡化“知”与“无知”之间的界限而催生“机器幻想”。这一缺陷是大语言模型的基本运作原理所决定的,无法通过更多的数据喂入与提高算力加以根治。对于此类缺陷的反思性意识,无疑将帮助科学界以更符合科研伦理的方式摸索使用大语言模型的恰当方式,并激发更有学术抱负的人工智能研究者探索出一条与主流大语言模型开发路径不同的通用人工智能之路。

注释:

- (1)引自 2023 年 7 月 22 日上海召开的第八届中国通用人工智能年会报告,孙思琦:《计算生物学对于 AI 工具的使用》,未书面发表。
- (2)引自 2023 年 7 月 22 日上海召开的第八届中国通用人工智能年会报告,李剑锋:《ChatGPT 在人工智能中的运用以及人工智能的未来》,未书面发表。
- (3)引自 2023 年 7 月 22 日上海召开的第八届中国通用人工智能年会报告,邱锡鹏:《大语言模型研究体会——以 MOSS 为例》,未书面发表。

参考文献:

- [1]罗伯特·M·波西格.禅与摩托车维修艺术[M].张国民,王培沛,译.重庆:重庆出版社,2011:53.
- [2]Jason Wei, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models [J].Advances in Neural Information Processing Systems,2022, (35):24824-24837.
- [3]Larry Neumeister.Lawyers Submitted Bogus Case Law Created by ChatGPT.A Judge Fined Them\$5,000 [N/OL].(2023-06-23)[2023-08-01]<https://apnews.com/article/artificial-intelligence-chatgpt-fake-case-lawyers-d6ae9fa79d0542db9e1455397aef381c>.
- [4]Tsinghua University Press.ChatGPT Justifies Liberal Leanings With Its Own Values,Researcher Reports [R/OL].(2023-07-08)[2023-08-01]<https://phys.org/news/2023-07-chatgpt-liberal-values.html>.
- [5]Edmond Awad, et al. The Moral Machine Experiment [J].Nature, 2018, (563): 59-64.
- [6]Dwane van der Sluis.NUTS, NARS, and Speech [EB/OL]//Artificial General Intelligence:16th International Conference, AGI 2023, Stockholm, Sweden, Proceedings, 2023: 307-316. https://doi.org/10.1007/978-3-031-33469-6_31.

(责任编辑 吴 勇)