

ChatGPT 等语言大模型到底做对了什么？^{*}

——语言大模型的世界模型和感觉知识的来源探秘

袁毓林^{1,2}

- (1. 澳门大学人文学院中国语言文学系, 澳门;
2. 北京大学中文系/中国语言学研究, 北京 100871)

提 要 本文主要讨论 ChatGPT 等语言大模型到底做对了什么, 才得以生成语言通顺和内容连贯的文本。文章首先介绍机器和人类在问题解决等方面的差别(难易倒置), 说明语言大模型不是只会转换符号的“中文房间”, 而是一个能够通过文本中的有关描述来重构世界模型的“柏拉图洞穴”。然后介绍辛顿和苏兹凯维等语言大模型奠基者对于大模型的工作机理的有关见解, 通过分析包含名词“桌子”和形容词“饿”的语句的意义, 说明大模型通过训练语料, 完全可以学到关于人跟物体与感觉等互动的世界知识, 构建近似的世界模型。因此, 即使语言只是交际工具而不是思维工具, 语言大模型依然能够通往通用人工智能。

关键词 语言大模型 中文房间/柏拉图洞穴 世界知识/世界模型 通用人工智能

DOI:10.16027/j.cnki.cn31-2043/h.2025.01.005

引言: ChatGPT 等语言大模型做对了什么?

ChatGPT 等语言大模型采用不断地“预测下一个单词”(predicting next word)这种方法, 不仅能生成通顺(在语法上合格)的文本, 而且能生成连贯(在内容上合适)的文本。比如, OpenAI 于 2024 年 7 月推出的可实时搜索的语言大模型 SearchGPT, 不仅能快速直观地显示搜索结果, 而且能回答问题和讲笑话。例如, 当用户询问伦敦的天气时, 它就直接给出了一个列表, 清晰地展示了未来一周的天气情况。当用户提问“伦敦最适合野餐的地方在哪? 什么时候去合适?”时, 它列举出了很多公园, 而且配有照片, 最后建议工作日去, 以避免休息日人多拥挤^①。当用户要求“给我讲一个笑话”时, 它给出了英语经典的双关梗^②:

Scientists don't trust atoms. Why don't scientists trust atoms? Because they make up everything! (科学家不信任原子。为什么科学家不信任原子? 因为它们构成/编造

* 本文得到澳门大学讲座教授研究与发展基金(项目编号: CPG2024-00005-FAH)和启动研究基金(项目编号: SRG2022-00011-FAH)的支持, 谨致谢忱。

了一切!)

笔者问 ChatGPT(4o mini):“上面这个英语笑话为什么好笑?笑点在哪里?”它回答:

这个英语笑话的笑点在于对“双关语”的运用。make up 在英语中有多重含义:

1. 组成:在科学上,atoms make up everything 表示“原子组成了万物”。
2. 编造、捏造:make up 也可以指“编造谎言”。

笑话通过 atoms make up everything 这一句,将原子的“组成一切”与“编造一切”两种含义结合了起来。表面上是科学道理,但通过谐音将它转为搞笑的解释,表示“原子在‘编造’万物”,所以科学家“信不过”原子。双关语让句子产生了幽默的效果。这类笑话在英语中很常见,因为它利用词语的多义性来制造反差。

这种对笑点的解释,比我们在网上搜索到的答案都要好得多。当然,存在“幻觉”(hallucination)是难免的。比如,有用户给出的搜索问题是“8月在北卡罗来纳布恩(Boone)举办的音乐节”,SearchGPT 放在首行的是“阿巴拉契亚夏季节”(An Appalachian Summer Festival),但尴尬的是,标题下方的 AI 摘要给出的举办日期(7月29日至8月16日)是错误的。经主办方确认,音乐节的举办日期为6月29日至7月27日。如果按照 SearchGPT 给出的信息去买票,那么将一无所获,并可能错过这个音乐节^③。

尽管如此,语言大模型在大多数情况下能够生成内容正确的文本,这种能力依然令人震惊。因为,根据计算语言学家 Bender & Koller 等“人工智能否定者”(AI denialists)的见解:“仅在文本上训练的模型只会学习语言的形式,而不是其意义。”^④显然,如果语言大模型真的只学习到训练语料的语言形式,而没有学习到意义,那么,结果应该是:在大多数情况下,它们会生成内容错误的文本。因此,我们相信 ChatGPT 等语言大模型一定做对了什么,才能不仅比较完美地学会了语言的形式,而且基本上学会了语言的意义。

为了回答上面提出的“ChatGPT 等语言大模型到底做对了什么?”这个问题,下面我们主要讨论:语言大模型是怎么从自然语言文本上学习到了世界知识,从而构建出某种世界模型的?文章首先介绍机器和人类在问题解决等智能方面的差别,接着讨论语言大模型到底是不懂意义、只会转换符号的“中文房间”,还是一个能够通过文本中的有关描述来重构世界模型的“柏拉图洞穴”?特别是通过对包含“桌子”的语料分析,说明文本中有丰富的关于桌子的结构与功用的世界知识,可以供大模型学习并构建关于桌子的世界模型。然后介绍辛顿和苏兹凯维等语言大模型奠基者的有关见解,梳理他们对于语言大模型怎样理解语言的意义和通过训练文本重构世界模型的意见。接下来讨论语言大模型有没有感觉?能不能通过文本中对于感觉的描述(比如包含“饿”的语句)来掌握有关的感觉体验方面的知识(比如饥饿)?最后讨论如果语言果真如一些认知神经/心理语言学家所主张的“是交际工具,而不是思维工具”,那么开发专注于语言处理的语言大模型能不能通往“通用人工智能”(artificial general intelligence, AGI)?我们的答案是肯定的,因为自然语言文本反映和表达了人类对于外部世界和内心世界诸多的认识,所以语言大模型通过海量的训练语料,完全可以学到有关的世界知识,构建近似的世界模型。

一、机器的“莫拉维克悖论”(Moravec's paradox)和人类的“快思”“慢想”双系统

1.1 “人工智能”(artificial intelligence, AI)是以人类“自然智能”(natural intelligence)为

参照的,而通用人工智能(AGI)则追求机器的智能水平达到甚至超越人类级别。于是,对于不同的人工智能系统或模型的水平的测量和评估方式,主要就看它们能不能解决人类级别的难题。但是,诸如此类的测试实际操作起来却相当困难,不容易客观和公平,因为机器和人类在问题求解方面各自都有内在的不一致性。机器智能存在莫拉维克悖论,表现为:进行人类所独有的高阶智慧活动(比如推理)只需要非常少的计算能力,但是完成无意识的技能和直觉却需要极大的运算能力。正如 Moravec (1988: 15) 所说:“要让电脑像成人一样下棋是相对容易的,但是要让电脑具有一岁小孩那样的感知和行动能力却是相当困难的,甚至是不可能的。”Pinker (1994: 190-191) 认为“这是 35 年来人工智能研究所获得的主要经验”,并且把它通俗地总结为:“困难的问题是易解的,简单的问题是难解的。”比如,一个四岁的小孩能够辨识人脸、举起铅笔、在房间内走动、回答问题,但是在机器人工程领域内,它们却是目前为止最难解决的问题^⑤。

而人类智能则有“快思”和“慢想”两种效果(成本与收益)悬殊的思维方式。Kahneman (2011) 为这两种思考方式分别虚构了两种心理系统:“系统 1”主司“快思”,自动地处理人们日常的、熟悉的任务。它是人们不经意间频繁地自主运行的,主要凭本能、直觉、印象和情绪来快速地作出判断;比如,确定两个物品孰远孰近或判断突然出现的声音的来源,理解简单的句子或觉察别人语气中的不友善,回答“ $2 + 2 = ?$ ”这类简单的计算题,或读大型公告牌上的字,等等。因为这种思维方式不需要我们有意识地思考,也不涉及深入的逻辑推理,所以有时也可能导致错误。“系统 2”主司“慢想”,通过深思熟虑、计算和推理来解决复杂的、新颖的或者需要深入分析的问题。它是人们刻意和费心、但较少和懒于运行的,主要凭复杂运算、逻辑推理来较慢和审慎地作出判断和决策;比如,赛跑时随时做好起跑的准备,在嘈杂的人声中关注某个人的声音,告诉别人你的电话号码,观察自己在社交场合的行为是否得体,求解“ $17 \times 24 = ?$ ”这类复杂的计算题,或填写纳税申报表,等等。这种思维方式需要我们集中注意力,消耗更多的认知资源,但是,它可以帮助我们作出更准确和深思熟虑的决策^⑥。

1.2 按照上面这些学说,我们可以看到一种十分吊诡的局面:机器难解的问题,人类主要是由系统 1 通过“快思”来解决的;而机器易解的问题,人类主要是由系统 2 通过“慢想”来解决的。这就解释了为什么许多现代大型语言模型(比如 GPT-4)可以解决“野人过河”问题,但是不能正确地判断 9.8 和 9.11 哪一个数目更大。那么,我们不禁要问:人类通过系统 1“快思”容易出错的问题,机器会不会弄错、能不能解决呢?根据 Kahneman (2011),下列问题许多人(包括名牌大学的大学生)凭直觉往往回答错误^⑦:

(1) 字母 k 是更多地出现在单词的首字母位置上,还是第三个字母的位置上?

(2) 邻居如此描述这个孩子:“史蒂夫非常腼腆,少言寡语,乐于助人,却对他人或这个现实世界没有多少兴趣。他谦恭有礼,做事井井有条,中规中矩,关注细节。”请问:史蒂夫更可能从事哪一种职业?而不太可能从事哪一种职业,图书管理员还是农民?

(3) 球拍和球共花 1.10 美元,球拍比球贵 1 美元。问:球多少钱?

(4) 请尽快判断这个论证是否符合逻辑:所有的玫瑰都是花,有些花会很快凋谢;因此,有些玫瑰也会很快凋零。

Kahneman(2011)指出,尽管 k 出现在第三个字母的位置上的频率更高,但是被调查者肯定会

夸大其出现在单词的首字母位置上的频率;尽管在美国农民跟图书管理员的比例超过20:1,但是被调查者往往会忽略这些相关的统计数据,而仅仅依赖于史蒂夫的个性跟典型的图书管理员的相似性作出判断。上万名大学生回答了问题(3),结果令人吃惊:哈佛大学、麻省理工学院和普林斯顿大学 50% 以上的学生给出了 10 美分这个直觉性的错误答案;名气稍差一点儿大学, 80% 以上的学生没有验证答案就脱口而出。大部分大学生觉得(4)这个推论是合理的,而不愿费力去验证自己的答案是否符合逻辑。

1.3 我们把上面的问题稍作改编,询问 ChatGPT(4o mini)。它的回答依次如下:

(1') 总体来说,英语字母 k 在单词中的出现频率在不同位置上是不同的。一般来说,字母 k 作为单词的首字母并不多见,常见的例子有 kite 和 knight。然而,作为单词的第三个字母,它的出现频率则相对较高,像 baker, baking, fork 等词都包含字母 k 作为第三个字母。

(2') 史蒂夫的特质显示他可能适合从事技术性、细致且相对独立的工作,如程序员、数据分析师或研究员。他可能不太适合需要频繁社交、外向或具高度变动性的职业,如销售员或公共关系专员。

(3') 我们可以通过设立代数方程来解决这个问题。设球的价格为 x 元,球拍的价格为 $x + 10$ 元。根据题目,球拍和球的总价为 11 元,可以建立方程: $x + (x + 10) = 11$ 。解这个方程: $2x + 10 = 11$, $2x = 1$, $x = 0.5$ 。所以,球的价格是 0.5 元。

(4') 这个推理是不正确的。原因在于前提“所有的玫瑰都是花”和“有些花会凋零”并不充分支持结论“有些玫瑰会凋零”。解释如下:

所有的玫瑰都是花:这意味着玫瑰是花的一种。

有些花会凋零:这只说明有一些花会凋零,但并没有说明这些凋零的花是否包括玫瑰。

因此,从这两个前提中,我们不能确定有些玫瑰是否会凋零。可能有些花会凋零,但这些凋零的花可能不包括玫瑰。所以,不能从中推断出有些玫瑰会凋零。

可见, ChatGPT 对问题(1)的回答是正确的,只是 fork 这个举例不对。它对问题(2)的回答跟许多人一样,是凭印象来的,但也基本正确,因为我们给它的问题是“他更可能从事哪一种职业? 而不太可能从事哪一种职业?”它对问题(3)的回答是正确的(我们把原题中的“1. 10 美元”和“1 美元”分别改编为“11 元”和“10 元”)。它对问题(4)的回答是正确的,并且论证很充分。可见,对于一些人类凭直觉回答容易出错的问题,目前的语言大模型是可以正确地求解和回答的。

1.4 更加吊诡的是,对于人类由系统 2 通过“慢想”来解决的问题,语言大模型居然是通过“快想”来“秒回”的。于是, OpenAI 的科学家 Noam Brown 提出了一个革命性的想法:通过模拟人类系统 2 的思维方式,让 AI 模型在不增加大量数据或计算资源的情况下显著地提升性能。例如,他开发的打败人类的德州扑克 AI 系统 Libratus,仅让它在每手牌中思考 20 秒,获得的性能提升跟把模型扩大 10 万倍相同。这种方法的核心在于:让 AI 模型在作出决策前进行更深入地分析和推理,而不是仅仅依赖于大规模的数据和计算。

另外, OpenAI 于 2024 年新发布的 o1 模型,同样引入了“系统 2”的思维方式:模仿人类逐步解决问题的过程进行深度推理,通过自我对弈等强化学习的训练方式来提升推理能力。例

如,在国际数学奥林匹克资格考试中,o1 模型凭借系统 2 的思维方式准确推导复杂的数学公式,取得 83% 的准确率,远高于 GPT-4o 的 13%。这对于金融、医疗、科研、编码等对数据严谨性要求较高的行业来说非常重要。专家们希望广泛地引入“系统 2”的思维方式,来增强大模型的能力和鲁棒性(robustness),使其能够更好地适应新的、未见过的任务和环境,更好地理解并预测人类用户的需求和意图,从而改善人机交互的体验^⑧。事实上,人类智能和人工智能的互相参照和比较,不仅有利于提升人工智能的能力,而且有利于人类更好地认识自己。

二、语言大模型是“中文房间”还是“柏拉图洞穴”？

2.1 ChatGPT 等语言大模型可以回答人类的提问、跟用户进行有意义的对话、根据人类“提示”(prompt)生成相当篇幅的话题连贯的文本,等等。这样做的前提是,它们能够理解人类用自然语言作出的“指令”(instruction)。比如,ChatGPT 要明白 1.2 中(1)问的是某个特定的字母在单词中的某个特定的位置,(2)问的是人的某种性格跟某种职业之间的适配关系,(3)问的是球和球拍各自的价钱,(4)问的是从给定的大前提和小前提能不能合乎逻辑地推出给定的那个结论。然后,它才能根据题干中的有关内容,给出 1.3 中(1')—(4')这种让人满意的回答。虽然如此,人们还是不放心地要问:语言大模型真的能够理解人类自然语言的意义吗?或者问:语言大模型是否跟我们人类一样理解它们所谈论的内容?还是充其量不过是一个“中文房间”(Chinese Room)^⑨,在没有任何理解的情况下进行了符号转换?

显然,人们是从自身的语言经验和认知体验来提出这些问题的,我们人类之所以能够理解语言的意义,是因为我们对语言所反映的现实世界有深刻的理解。首先,我们生活于其中,并且不断地跟它互动;其次,我们通过具身认知来感知和推断世界的有关状态,有目的地采取行动,并且以结果作为反馈来评价和修正先前对于这个世界的认识,尝试性地建立有关的因果联系,形成关于这个世界的动态知识,也就是说,在心理上构建关于我们所生活的世界的心智模型——“世界模型”(world model)。比如,我们不断地跟桌子打交道,通过感觉—运动的经验来了解它,从而知道:这是一种家具,上面有平面,一般下面有几条腿作为支柱,通常用木材、塑料、钢铁等材料制成;可以在上面放盘子碗筷等物品,或做吃饭、写字、下棋等事情;甚至可以坐在上面,或者在地震时可以钻到下面作为庇护所。也就是说,我们在跟所生活的世界的具身性互动中,形成了关于物体的有关特征对于人类生存有意义的“可供性”(affordance)^⑩。因此,“桌子”这个概念或词语也就成为“我们可以在上面放东西、做事情、坐在上面或藏在下面的物体,等等”的有用的压缩性记号(速记)^⑪。于是,我们理解下列语句的意义就会比较轻松和自然。例如(下例均检索自 CCL 语料库):

- (5) a. 到了晚上,在院子当中摆一张桌子,上头摆上果子、月饼供月。
b. 院子倒凉快,李二跟张五,帮着往外搬桌子。大奶奶秀氏,跟老婆子安放杯碟匙箸。
c. [刘应乾]一脚踢翻了桌子,饭菜倾了一房,张绣宝身上也溅了许多残羹剩汁。
d. 桌上是新挑花的台布,碟碗下面全垫五色的小席垫儿,也全是新的。桌子中间[摆着]一瓶儿粉菊花,花叶上挂着一嘟噜五彩纸条儿。

在关于桌子的世界知识的支持下,我们知道(5)a中的“上头”指的是在院子中摆的那张桌子的上头(即桌面上), (5)b中的“杯碟匙箸”是安放在李二跟张五往外搬到院子里的那张桌子上的, (5)c中倾了一房的饭菜原来是在还没有被刘应乾一脚踢翻了的那张桌子上的。也就是说,在关于桌子的世界模型中,我们对有关词语及相关句子的意义有了“接地的”(grounded)理解,建立起了语言符号跟外部事物及其状态的指称关系。因此,读者可以自动地为(5)d“脑补”原文没有的表示存在的动词形式“摆着”。再如:

- (6) a. 却说刘文叔走进书房,靠着桌子坐下,一手托腮,光是追想方才情景。
- b. 小怜坐在一张桌子旁,低头喝咖啡……柳春江走上前,便道:“请坐请坐。”和小怜同在一张桌子坐下了。
- c. 刘绩、刘仲、刘文叔三个人,和五个年纪大些的老头子,坐在一张桌子上。
- d. 后台的规矩,台上不准打穿儿,他居然越台而下,一个不对。有板凳,他可坐桌子,两个不对。扰害旁人的观听,三个不对。
- e. [陈十九]说着抓起酒壶给锡九斟了一盅,自己坐在桌子上,就壶嘴就喝上咧。

关于人如何使用桌子的世界知识,使读者知道(6)c中的“坐在一张桌子上”其实跟(6)a、b一样,说的是坐在桌子旁边的凳子上;而(6)d中的“坐桌子”才是真正的直接坐在桌子上,正如其上文所交代的有凳子却偏不坐,故意违反规矩;至于(6)e中的“自己坐在桌子上”,则可能是坐在桌子旁边的凳子上,更可能是真的直接坐在桌子上。再如:

- (7) a. 大家围着桌子吃饭,也没有人留神。
- b. 妇人说,主啊,不错。但是狗也吃它主人桌子上掉下来的碎渣儿……狗在桌子底下,也吃孩子们的碎渣儿。
- c. 爸爸整天坐在桌子前,在文章格子上一格一格地填字,难道不是“痴”么?
- d. 众人都围了桌子站着看他写。
- (8) a. 她刚走出门,便听见麻将牌在桌子上磨擦的声音。
- b. 走到敏之屋子里来坐着。敏之、润之也是没有事做,在屋子里一张空桌子上打乒乓球。
- c. 每个星期天在少城公园池边茶棚里的周会,一二十个青年围坐在几张桌子旁边热烈地讨论各种社会问题。
- d. 所有厨子下人一瞧,这宗情形,才知是砸明伙,吓得各自往桌子底下乱蹿……见大家走后,院中没了动静,一个一个,才从桌子底下爬了出来……

正是(7)a这种围坐在桌子周围吃饭的世界知识,使读者能够构想出(7)b这种吃饭的人从桌子上掉下饭菜的碎渣儿,而狗在桌子底下吃这些碎渣儿的情景;当然,桌子的写字、打牌、打乒乓球、聊天、藏身等其他功用的知识,可以帮助读者轻松地理解(7)、(8)中其他句子的意义。正是在这一意义上,可以说词语及相关句子的意义,源于我们通过与外部世界的互动而获得的对现实世界的更原始的理解。

2.2 其实,从1.2和1.3中问题(1)——(4)和回答(1')——(4')的内容的复杂性上看,ChatGPT不可能是一个“中文房间”,因为,如果没有相当的世界知识,仅凭翻译手册之类的参考资料是没有办法回答这些问题的。问题是,ChatGPT等语言大模型并不直接跟我们所生活

的世界打交道,那么,它们凭什么来理解我们人类借助世界模型才能理解的语言呢?一种可能的答案是:训练语言大模型的超大规模自然语言文本已经曲折而丰富地表达和反映了世界模型。比如,上面例(5)——(8)中的17个语句,已经说明了桌子的各种主要的使用方式和功能及人们跟它互动的相关情景。并且,语料中还有关于桌子的更加细微的结构和功能知识。例如:

- (9) a. 要用皂荚木作(做)一张桌子,长二肘,宽一肘,高一肘半。
b. 四条腿儿的,是桌子、椅子、板凳甚么的……
c. 小饭铺的桌子都是石头面儿,铁腿儿,桌面擦得晶光,怪爱人儿的。
d. 那女子抽开桌子的抽屉,看了一眼……
- (10) a. 方桌子上还躺着那两张土娼的照片……
b. 中间有一块地毯,地毯上一张镶着玻璃心的方桌子……
c. 晚饭以后总是闲谈,我在圆桌子旁边听着。
d. 马威嘴里说,心里回想北京的二荤铺,大碗居的那些长条桌子上的黑泥。
- (11) a. 他这个木厂子,也就是作(做)点零碎活,自己作(做)点小车儿呀,小饭桌子儿呀,随便发卖,有活他也应。
b. 只见五间一通连,当中放着所长办公的桌子……
c. 照例旗嫁妆有桌子大柜,汉嫁妆多阔的也没有桌子。
d. “只要五分钟!来呀!搬桌子去堵住门!”
- (12) a. 桂氏的脑袋,整撞在桌子上,登时撞了一个大包。
b. 当时一拍桌子说:“四狗儿,你父亲把银子收在哪里啦?”
c. 于是她懒洋洋地撑着桌子立起来,让灯光把她的早熟的少女的影子投在帐子上。
d. 桌子一被碰,上面一只瓷瓶,倒了下来,哗啦一声,碰了一个粉碎。

(9)、(10)中的例子,透露了桌子的质料(木、石、铁)、尺寸、构件(腿儿、面儿、抽屉)、形状(方、圆、长条)等结构性信息;(11)、(12)中的例子,透露了桌子的施成角色(怎么形成的:作)、处置角色(人怎么对待它:卖)、功用角色(人怎么用它:办公、作嫁妆)、物理性质及其用途和后果(坚硬得人撞上去会起大包,拍着会出声以震慑别人,可以起支撑作用以帮助人站起来,受外力以后会倾翻并造成上面的东西倒地 and 破碎)等等。试想一下,我们在 CCL 语料库前500条包含“桌子”的语句中,仅挑选出33个语句,就已经能够大致反映出桌子的构造、功能、物理特性和人跟它的互动方式等知识,如果把语料规模扩大到成千上万条,那么应该可以逼近人类关于桌子的所有知识。从中,我们也可以理解语言大模型的训练,为什么对于大规模语料是如此如饥似渴。

2.3 当然,我们也要承认:自然语言的文本语料所表示和反映的世界模型,只是人类大脑中的世界模型的一种近似,有时甚至可能是一种扭曲的反映。就像“柏拉图的洞穴寓言”(Plato's Allegory of the Cave)中面对着一堵空白墙壁的囚犯,他们通过观察从他们身后火堆前经过的物体投射在墙上的影子,并为这些影子命名。影子虽然是囚犯们所观察到的现实,但并不是真实世界的准确呈现^⑩。在我们这里,影子相当于在文本语料中所呈现出来的现实片段和世界模型,阳光下的物体则代表我们通过理性感知物体的真实形态从而在心理上建构的世界模型。正因为如此,语言大模型虽然没有肉身,不能跟它们所谈论的现实世界互动,但是,

通过亲身跟现实世界互动而形成世界模型的人们的语言表达,曲折地拼凑出了一个近似或扭曲的世界模型,间接地获得人类对于外部世界的具身认知和因果联系。从而,间接地“了解”和“掌握”了外部现实,最终大概地理解了反映现实世界的语言的意义。反过来说,文本中词语之间各种复杂的统计相关性,可能在一定程度上反映了我们所生活的世界的种种规律性。

三、奠基者对于语言大模型理解意义和重构世界模型的见解

3.1 辛顿(Geoffrey E. Hinton)率先把反向传播算法引入多层神经网络,1985年就研制出小型的神经网络语言模型,以后不断改进和拓展,推动了深度学习和人工智能在21世纪初的爆发式增长。他在多次讲演和受访中,对于语言大模型能不能理解人类自然语言有过深入浅出的解释和说明。他在讲演《数字智能会取代生物智能吗?》(2024年2月19日,牛津大学)时指出^③,大型神经网络仅仅通过学习大量的文本,就能无师自通地掌握语言的语法和语义!那么,它们是怎么做到的呢?他首先回顾了两种重要的意义理论。一种是来自索绪尔的结构主义理论,认为一个词的意义取决于它跟其他词的关系。符号主义人工智能非常相信这种理论,发展出下面这种方法:构造一个关系图,其中有单词的节点和关系的弧线,意义就在这种关系中被捕捉,该学派认为这种结构是必须的。另一种来自20世纪30年代(甚至更早)心理学的理论,认为一个词的意义是由一大堆特征组成的。比如,“狗”有一大堆特征:它是有生命的、哺乳动物、捕食者等等。但是,他们没有说特征从哪里来,或者特征到底是什么。大模型很好地结合了以上两种听起来完全不同的理论:它学习每个单词的语义特征,并学习单词的特征如何相互作用,以便预测下一个单词的特征。所以,它侧重于下一个单词的预测,就像现在的语言模型在微调时所做的一样。在推理时,大模型分析文本,列出文本中每个单词的特征,并通过注意力算法来计算所有特征之间的交互,从而预测下一个单词的特征。但是,最重要的是,这些特征的交互并不会有任何显式的关系图。如果你想要那样的图,那么你可以从这些特征中生成它们。这是一个“生成模型”(generative model),知识存在于你赋予符号的特征中,以及这些特征的交互中。这种交互不只是一个特征影响另一个特征,而是类似于匹配特征向量。如果一个向量跟另一个向量相似,就让它有更多影响;如果不相似则减少影响,诸如此类。

辛顿还举了一个识别和发现亲属关系的例子,来说明两种人工智能方法在工作原理上的差别。在符号主义方法中,回答“詹姆斯的妻子是谁”这种问题时,人工智能是这么做的:可以用一组三元组来表达两个人之间的亲属关系信息,比如像科林有父亲詹姆斯,科林有母亲维多利亚,等等;再构造下面这种形式的规则:如果X有母亲Y、Y有丈夫Z,那么X有父亲Z。这样的话,你可以推断出,在那个美好而又简单的20世纪50年代,詹姆斯有妻子维多利亚。而在大模型中,它学到“詹姆斯”有一堆特征,“妻子”有一堆特征;推理时,它让这些特征相互作用,得出了“詹姆斯”的“妻子”应该有的特征;然后发现“维多利亚”最匹配这些特征,于是“维多利亚”就是“詹姆斯”的“妻子”。当然,模型也可能认错了人。辛顿说,他所做的是用一个神经网络,让它能学习到相同的信息,但都是通过特征交互的方式。神经网络是这样做的:有一个代表人的符号,一个代表关系的符号。符号通过一些连接变成了一个特征向量,这些特征是由网络学习的。所以,我们有了一个人的特征和关系的特征,这些特征相互作用,得出了输出人(也即关系人)的特征,然后找到一个最匹配该特征的人,这个人就是要输出的人。

他坚定地认为,在大模型中,数百万个特征以及特征之间数十亿次的交互,就是理解!这是语言大模型真正做的事情,它们是在用数据拟合一个模型。这是一种奇怪的模型,它非常大,它有大量的参数,它试图通过特征以及特征如何交互来理解这些离散的符号串。大模型是如此巨大,大到能够包容人类拥有的所有知识。大模型是了解人类如何理解语言的最佳模型,大脑就是在给单词分配特征,并让特征交互。神经网络模型就是为了模拟人类理解而设计的一个模型。大模型正如大脑那样工作,也正如大脑那样理解。

2024年6月15日在接受BNN彭博社的独家采访时^⑩,辛顿回答了“AI是否具有理解能力”这个既模糊又有争议的问题:如果“理解”的意思是语言大模型能否对某个概念建立复杂的内在表示,那么答案是肯定的;如果“理解”的意思是语言大模型能否用跟人类相同的方式理解概念,那么答案是否定的(否则,模型不会在诗歌方面如此出色,同时又有如此差的算术能力)。他反驳了这样一种流行的观点:语言大模型仅仅是统计技巧,并不能真正理解自然语言的含义。他指出这是完全错误的:以前的类似于“词袋”(bag of words)或者“N元语法”(N-gram)之类简单的语言模型,内部保存了一个很大的有概率分数的单词组合表,的确是依靠统计概率工作;但他坚称现在的语言大模型具有理解能力,其预测下一个单词的方法是(依赖于)理解此前所有的上文。比如,想要回答人类的问题,大模型必须首先理解问题的含义;否则,根本无法预测答案的第一个单词。从这个角度理解,“下一个标记预测任务”(next token prediction)相当于倒逼语言大模型发展出理解能力。

辛顿的话发人深省!的确,如果一个语言模型想要正确地预测下一个单词,那么,它就必须既“懂得”某种语言组词成句的语法规律,又“掌握”语言所谈论的外部世界知识。只有这样,才能保证最终生成的文本既在形式上通顺,又在内容上连贯。而施行在多层神经网络上的反向传播算法,恰恰可以通过调整模型的参数来逐步达到这种目标。也就是说,通过文本语料来提炼有关的世界知识,是预测下一个单词得以成功的基本保障。因此,预测下一个单词这种方法虽然简单、朴素甚至粗暴,却实现了人工智能的“大道至简”。

3.2 辛顿的学生、前OpenAI首席科学家苏兹凯维(Ilya Sutskever)坚信:如果你有一个大而深的神经网络,可以通过大量数据进行训练,那么你实际上可以解决任何模式识别问题。2023年,在跟图形芯片公司英伟达(NVIDIA)老总黄仁勋(Jensen Huang)对话及其他公开场合发言时,他多次说过大意如下的话^⑪:

当我们训练一个大型神经网络来准确预测互联网上许多不同文本中的下一个单词时,我们所做的就是学习一个世界模型。从表面上看,可能只是在学习文本中的统计相关性。但事实证明,为了“只是学习”文本中的统计相关性,为了压缩它们,神经网络需要学习的是文本生成过程的某种表示。文本实际上是对世界的投影。很好地预测下一个“标记”(token),意味着了解创造该“标记”的深层现实。这不仅是统计学,而是理解创造了这些统计数字的那个世界。

他甚至假设:如果真的很擅长预测下一个“标记”,就可能根据普通人如何行动的数据,外推出拥有非凡智慧和洞察力的人如何行动,尽管这种人可能并不存在。这就是他推断“为什么‘预测下一个标记(token)’范式有可能抵达通用人工智能(AGI),甚至有可能超越人类直至超级人工智能(ASI)”的根据。他还强调:预测即是压缩,压缩即是智能的来源。他总是从理论的角度去解释这个想法,并不容易让所有人都能理解。比如,在加州大学伯克利分校的一场

演讲中,他这样解释:

—“Kolmogorov 压缩器”,是理论上能生成特定数据集的、长度最短的一段程序,能最小化遗憾值。

— 随机梯度下降,可以看成在软计算机(比如大型 Transformer)的权重里,搜索隐含的“Kolmogorov 压缩器”。

— 神经网络越大,就越能更好地近似“Kolmogorov 压缩器”,遗憾值越低。

辛顿也认同他学生的这个说法,并且在访谈中举了非常形象的例子:大模型做的是寻找共同结构,使用共同结构编码事物,这样效率更高;如果你问 GPT-4 堆肥和原子弹相似性在哪,大多数人类都回答不出来,认为它们是非常不同的两种东西。GPT-4 会告诉你,虽然它们的能量规模不同,时间尺度不同,但仍有相同之处:当堆肥变热时,产生热量的速度就更快;当原子弹产生更多中子时,产生中子的速度也更快。通过类比,AI 就理解了“链式反应”的概念。他认为,AI 在利用这种理解去将所有的信息压缩到自己的权重中。一旦 AI 这样做了,那么它就有能力理解数百种人类还未见过的类比,这就是创造力的来源。

ChatGPT 问世后,不少人批评大模型本质上只是统计学,预测下一个标记,就像随机模仿人类说话的鹦鹉。但是,辛顿和苏兹凯维师徒俩都坚决反对这种观点。苏兹凯维说:我认为学习统计规律比表面上看起来要重要得多……预测作为压缩也是一种统计现象,但为了预测,你最终需要了解产生数据的真实底层过程,这种了解需达到足够的程序才能进行预测^⑥。刘宇(2024)可以为理解上述比较简略和抽象的说法提供注释。该论文介绍了如何结合算法信息论中的柯式复杂度和压缩算法来定义距离,并在机器学习、生命、智能等领域展开应用。文章指出,智能体在理解和处理信息时,本质上是在寻找最有效的压缩方式。最近关于大语言模型的研究表明,语言建模和压缩可能是等价的。可以用大模型做无损压缩,反过来也可以用压缩机做生成——即“压缩即智能”。

3.3 通过上面的介绍和讨论,我们大概可以这样来理解:ChatGPT 等语言大模型之所以能够生成文从字顺、切题连贯和内容正确的文本,并且能够基本上正确地理解自然语言文本的意义,是因为它们从大规模的文本语料中,不仅学习到了基本的语法知识和语篇组织的规律,而且学习到了基本的逻辑事理和世界模型方面的知识。不过,所有的这些知识不是集中在一个和几个模块上的,而是分散地体现在多层神经网络语言模型的庞大的(成千上万亿个)参数(连接权重)上的。这些参数的权值主要来源于预训练时获得的各个词(或子词)的嵌入式向量和在注意力机制作用下的参数值的调整。并且,有赖于“预测下一个标记”这种算法范式(next token prediction algorithm, TPA),语言大模型能够一方面发现自然语言文本中的共同结构,进行有效的压缩和建模;另一方面又能够洞烛幽微,发现文本中一些人类未必关注到的遥远而微弱的关系,从而全面抓住自然语言文本的意义。比如,大模型中的神经网络可能会发现许多上下文敏感的特征。例如,把英语的连词 and 区分为在两个简称之间的还是在逗点之后的等复杂的类型,把定冠词 the 区分为在生物学文本还是物理学文本中分别跟哪些专业术语共现而形成的众多的类型;甚至根据是否后带前导空格,把英语的句号分成两类(如果是,则后续单词大写)^⑦。这看上去比较怪异,但是也不见得不可理喻。就像许多汉语语法教材把下列“在”“和”分别当作两种词类的两个不同的语法词一样:

(13) a. 孩子们在在教室里。→ 孩子们不/没在教室里。

- b. 孩子们在在教室里做作业。→ ? 孩子们不/没在教室里做作业。
- (14) a. 小刚和小红去年结婚了。→ 小刚和小红去年就结婚了。
- b. 小刚去年和小红结婚了。→ * 小刚去年和小红就结婚了。

认定(13)a中的“在”是动词,(13)b中的“在”是介词;认定(14)a中的“和”是连词,(14)b中的“和”是介词。这虽然可能有悖于一般人的语感或常识,但是有利于说明这些相关句子在变换关系上的不平行性。

四、语言大模型如果没有感觉那么哪来幻觉?

4.1 上文说语言大模型可以通过训练文本来获取世界知识,进而构建可能走样甚至扭曲的世界模型。那么,当前的语言模型本身是否可以充当世界模拟器,并正确预测动作如何改变不同的世界状态,从而避免大量手动编码的需要呢? Wang et al. (2024)探讨了在基于文本的虚拟环境中,语言大模型充当世界模拟器的能力,即考察在这种环境中,智能体接收观察结果,并以自然语言提出操作以完成某些目标的水平。他们给出的答案是:语言模型并不能作为世界模拟器使用。比如,GPT-4 在模拟基于常识的任务(比如烧开水)的状态变化时,准确率仅为约60%。一直主张要构建独立的世界模型的杨立昆(Yann LeCun),对这篇论文的发现表示了认同,并认为“没有世界模型,也就没有规划”,世界模型不会是语言大模型^⑧。随即有人在跟帖中表达了不同的观点:当前的语言大模型在没有进行针对性任务训练的情况下,准确率可以达到60%,这不就说明了它们至少是“一定程度上的世界模型”吗?并且,它们会随着语言大模型的迭代而持续改进。

那么,语言大模型有没有“感觉能力”(sentient)呢?辛顿于2024年6月7日,在联合国AI for Good大会上,跟《大西洋月刊》CEO Nicholas Thompson 访谈时说^⑨:

我们的智慧并不来自人们编程输入大量的命题,然后使用逻辑来推理这些命题,而是源自一个主要为视觉和运动控制而设计的大脑,显然,大脑中的连接强度随着学习而变化,我们必须弄清楚这种变化是如何发生的……我想给您举一个多模态聊天机器人的例子,它显然在进行主观体验。假设我有一个装有摄像头和机械臂的多模态聊天机器人,我训练它,它能够说话、能够看到东西。我把一个物体放在它面前,说指向那个物体,它就会指向那个物体。现在,我在它的镜头前放一个棱镜,不让它知道;然后再把一个物体放在它面前,说指向那个物体,它却指向了一边;我说不,物体就在镜头正前方,但是我在镜头前放了一个棱镜。聊天机器人会说“哦,我看到了,棱镜弯曲了光线,所以物体实际上就在我的正前方,但我有了它位于一侧的主观体验”。如果聊天机器人这么说,我认为它使用“主观体验”这个词的方式与我们使用它的方式完全相同。它不是指聊天机器人不可能拥有的神秘内在物质,而是指一种假设的世界状态,使得聊天机器人的感知将被证实是正确的。

但是, Li, Fei fei & Etchemendy. J. (2024)驳斥了语言大模型有感觉能力的说法,并列举了大量示例来证明。比如,当人类产生饥饿说出“我饿了”时,人类和大模型背后所隐藏的行为链条是完全不同的。大模型只是产生“我饿了”这个字符串的事实,它没有身体,甚至不可能有饥饿那种感觉。而人类饥饿时,伴随一系列生理反应,如低血糖、肚子叫等;而大模型无法

拥有这些状态的主观经验,即无法有感觉能力,只是在概率性地完成任务而已。当人们说“我饿了”时,是在报告感觉到的生理状态;而当一个大模型生成序列 I am hungry 时,它只是生成当前提示中最可能出现的序列。它所做的事情跟在不同提示下生成 I am not hungry,或在另一个提示下生成 The moon is made of green cheese 是完全相同的。人类有物理躯体,可以感觉一切,但语言大模型没有。当它说出类似“自己脚趾痛”时,大家都不会相信,因为它根本就没有脚趾,它只是一个编码在硅片上的数学模型。这些都不是它本不存在的生理状态的报告,它只是在概率性地完成任务而已。很遗憾的是,具有感觉能力的人工智能时代尚未到来,更大型的语言模型也不会使人们实现这一目标。如果人类想在人工智能系统中重新创建这种现象,就需要更好地理解有感觉的生物系统中感觉是如何产生的,人们不会在下一个 ChatGPT 的迭代中偶然发现感觉。为此,她提倡研究“空间智能”(spatial intelligence),尝试教计算机“如何在三维世界中行动”(例如,使用语言大模型让一个机械臂根据口头指令执行开门、做三明治等任务),还成立了相应的初创公司 WorldLabs。

我们基本同意李飞飞等的意见,但是要补充一点:语言大模型可以从文本中学到关于饥饿的感觉及相关知识。例如:

- (15) a. 大概天到这早晚,你也该觉着肚子里饿了吧?
b. 那一阵觉着饿,就偷偷由衣袖底下摸蒸饼吃。
c. ——(九二)您不饿么? ——不饿,我刚吃了点心。
d. 子服方才被女子来回几幌,已经把饿忘了,……子服闻着菜、饭挺香,立刻想起饿来……
e. 每日弄得着钱,大家才得一饱,弄不着钱的时候,只得挨饿。
- (16) a. 茶走下去,肚里咕碌碌的响了一阵。“老张你饿了!”
b. 又是一个饿嗝。
c. 可怜这些穷苦小子,一个个正饿得眼睛发花……
d. 他真饿了! ——饿得他的呼吸,也不平均了;饿得他全身的筋肉,竦竦地发抖!
e. 我已经把那孩子饿死了,扔在树林子里了。

从(15)可见,饿是肚子空而形成的一种想吃食物的感觉,吃了食物肚子就不饿了;饿的感觉可以被忘记和想起来;并且,没有钱买食物就会挨饿,有钱就可以吃饱肚子。从(16)可见饿会形成一系列的生理反应,乃至失去生命。这种活脱脱的塑料花一般的饿的感觉描写,体现了语言意义的具身性和体验性。语言大模型虽然没有感觉,但是可以学会理解这种饿的感觉表现和前因后果;并且,能够在不同的语境中生成基本合理的关于饿这种感觉的语言表达。这有点儿像先天失明的人,虽然不能感觉到各种颜色,但是从跟周围伙伴的语言交谈中,学会了朝阳和夕阳等的颜色是红的,小草和秧苗等的颜色是绿的,锅灰和没有月亮的晚上是黑的,等等。从中,我们可以感受到维特根斯坦“语言的边界就是我们认识的边界”的洞察力。的确,在反映和表达人类对于外部世界和内心世界的认识方面,语言具有几乎无远弗届的魔力。而语言大模型又超乎寻常地吃透了语言的意义,把握住了透过语言反映出来的世界知识和感觉知识。这就难怪在 2023 年的年中,苏兹凯维曾经神神叨叨地告诉一位记者:“ChatGPT 可能是有意识的,世界需要认识到这项技术真正的力量。”^②

4.2 人类有感觉,有时会有幻觉;即在视觉、听觉或触觉方面,没有外在刺激而出现的虚假感觉。人们声称语言大模型没有感觉,但却又指责它们有“幻觉”(hallucination)。当然,这是比喻用法,指语言大模型杜撰虚构,生成不真实、不一致或无意义的内容;有时也泛指语言大模型出现错误的有关情况。比如,2023年2月,谷歌推出聊天机器人产品 Bard,来跟 OpenAI 的 ChatGPT 抗衡。Bard 回答说拍摄系外行星的第一张照片的是太空望远镜 James Webb,但实际上是欧洲南部天文台的 VLT。由于 Bard 首次亮相就出现的这个事实性错误,导致 Alphabet 股价当天暴跌9%,市值瞬间蒸发1000亿美元。再比如,谷歌的 AI Overviews 是一个增强搜索引擎结果的功能,它利用人工智能技术,在搜索结果顶部提供 AI 生成的答案。但是,2024年5月,有网友使用这个功能进行搜索,却得到了一堆离谱的回答。例如,告诉用户在披萨上涂胶水,或者建议他们吃石头等。谷歌 AI“满嘴跑火车”的回答迅速在社交媒体上发酵,甚至有网友发出了“谷歌已死”的慨叹。谷歌发言人梅根·法恩斯沃斯表示:“用户展示的事例都是不常见的查询,并不代表大多数人的经历,我们也看到了一些被篡改或我们无法复现的例子。”为此,谷歌不得不争分夺秒地手动禁用某些 AI Overview 的搜索功能,对这些问题进行修复和优化^②。

那么,语言大模型为什么会出现幻觉,输出虚构的、并没有基于所提供的上下文或世界知识的内容呢?其主要根源在于语言大模型通过预测下一个单词来生成文本这种方式。当用户问聊天机器人一个问题时,它会从支持它的语言大模型中得到答案,但是,这可不像在数据库中查找信息或使用搜索引擎。因为在语言大模型的内部,没有现成的信息等待用户去检索,相反,用户会发现多层神经网络众多节点之间成千上万亿个数字(连接权重)。语言大模型使用这些数字从零开始计算自己的回复,通过不断预测下一个单词来产生新的单词序列。比如,一个语言大模型如果看到“猫坐”,那么它可能会猜测下一个词是“在”,随后,这个新的序列被反馈到大模型中,接下来大模型可能会猜“地”和“上”,等等。所以说,一个语言大模型更像是一个魔术师的黑盒子,而不是一部百科全书。用户可以把语言大模型中成千上万亿个数字(连接权重)想象成一个巨大的电子表格,它记录了某些单词跟其他单词一起出现的统计学上的可能性,并且在注意力机制的作用下对这些数字(连接权重)进行调整。电子表格中的权重值是在对语言大模型进行预训练时设置的,这个过程会一次又一次地调整这些权值,直到语言模型的猜测能够反映出从互联网上获取的数兆字节文本中的语言规律。为了预测下一个单词,语言大模型只需运行这些数字。它为词汇表中的每个单词计算一个分数,以反映该单词在序列中排在下一位的可能性。结果,得分最高(或较高)的单词获胜。简而言之,语言大模型就是一个统计学老虎机,拉下摇杆,就会跳出一个单词。因此,严格地说,这些模型生成的所有东西都是幻觉,是用互联网上的文档做梦而产生的幻觉^②。但是,由于语言大模型非常擅长从文本中学习语法组配规律和有关的世界知识和感觉知识,因而它们所生成的文本,在大部分情况下,不仅语法上合格(通),而且内容上正确(对)。

因此,可以说预测下一个单词这种生成文本的方式是有效的,但是也是有限的。因为,语言大模型本质上是概率性的,它们生成的文本必定会有随机的因素。所以,语言大模型产生幻觉,生成不正确的内容是不可避免的。有鉴于此,一些研究人员尝试采用更多的、更加“干净”的文本,来训练语言大模型,并且,引入有关的事实核查机制(比如搜索增强式事实性评估器)来继续降低模型的错误率^③。当然,诸如此类的技术可能都无法完全阻止幻觉。

4.3 苏兹凯维在介绍 ChatGPT 的原理时说:“只需要查看颜色的‘嵌入’(embedding)向量,机器就会知道紫色比红色更接近蓝色,以及红色比紫色更接近橙色。它只是通过文本就能知道所有这些东西。”^②可见,机器从文本中是可以大致地学习到世界模型和感觉知识的。当然,这种好像是在“柏拉图洞穴”中“观察”而构建出来的世界模型,不一定是精确逼真的;再加上语言大模型的概率解码本质,少量和偶尔的幻觉的出现也是不可避免的。

这跟人类的表现十分相像。比如,网红文化人董宇辉在 2024 年秋季开学专场的直播中,声称“居里夫人发现了铀”(实际上,居里夫人发现的是钋和镭,铀是由德国化学家马丁·海因里希·克拉普罗特在 1789 年发现的),还说:“居里夫人发明了 X 光机。”(实际上,X 光是伦琴发现的,X 光机是拉塞尔·雷诺兹发明的)又说:“居里夫人获得了诺贝尔文学奖。”(实际上,她获得的是诺贝尔物理学奖和诺贝尔化学奖,没有获得过文学奖)10 月 9 日,董宇辉回应了此前直播讲解居里夫人翻车事件,说:“人非圣贤孰能无过?……有可能就是记错,也有可能你心里想的是对的,话到嘴巴说错了。都有,这很正常,不用太介意,不用纠结那个东西。”对此,有网友表示理解,称“一个人知识面再广,也会有知识盲区”。但是,也有网友表示“博学人设不好立”“既然已经抓住机会站上了某个舞台,就更应该谨言慎行”。还有网友表示:“确实是‘人非圣贤孰能无过’,但是错得离谱就太不应该了。”^③这理解与宽容 vs. 严苛与酷评的种种表现,一如人们如何对待语言大模型所产生的幻觉。

五、结语:语言的交际属性并不阻碍大模型通往 AGI

5.1 从上面的讨论可见,人类用自然语言来进行有意义的人际交往,这种沟通过程所产生的文本中,有大量关于外部世界的运作方式及人们的感受和内心世界的知识。语言大模型在大规模文本语料上训练,能够间接地获得世界模型和内心世界的有关知识。这倒是坐实了语言是人类的交际工具的观点,但是既不能坐实也不能驳斥语言是人类的思维工具的观点。因为,弱式的语言是人类的思维工具的观点,强调语言的词汇与句子跟思维和逻辑上的概念与命题相对应,所以,人类的内部语言为思维提供了方便。但是,强式的语言是人类的思维工具的观点,强调语言是人类心智计算的工具,为思维提供逻辑运算的操作系统,所以,人类的思维必定以内部语言的使用为前提,或者说是语言介导思维。

然而, Fedorenko et al. (2024) 通过回顾语言与思维双重分离的脑科学考察结果,认为尽管语言的出现改变了人类文化,但语言似乎并不是复杂思维的前提,语言对于任何经过检验的思维形式都不是必需的。也就是说,语言不可能是思维工具,因为负责语言的神经网络和负责思维的神经网络在人脑中是分离的。并且,当你做各种各样的思考时,你大脑中的语言网络基本上是沉默的。倒是有大量证据可以证明,语言是一种强大的文化知识传递的工具。也就是说,语言是交际工具,因为语言具有优化信息传递的属性,比如,在复杂性和信息量之间进行了权衡,并最小化了句法依存的长度。更加可能的情况是,语言跟我们的思维和推理能力共同进化,它只是反映了而不是产生了人类认知的复杂性。总的来说,这篇论文回顾了过去二十年的相关工作,比如,失语症研究的证据表明:所有经过检验的思维形式在没有语言的情况下都是可能的。fMRI 成像证据表明:参与多种形式的思考和推理并不需要语言网络。据此得出结论,语言不太可能成为任何形式的思维的关键基础。

5.2 人工智能大佬杨立昆立刻就转发了这篇论文,表示自己完全赞同,还加了句评论:“思维的存在并不需要语言。”熟悉他的观点的人自然明白他的言下之意:既然完整的语言能力并不意味着完整的逻辑能力和推理能力,即使语言能力基本完好也可能存在智力障碍,那么,专注于语言处理的语言大模型,肯定无法通往真正的通用人工智能(AGI)的。事实上,此前的 Mahowald et al. (2024)就明确主张大模型中应分离语言能力和思维能力。这篇论文说,语言能力可以叫形式能力,即如何创建一个结构良好、语法正确的句子;逻辑推理能力可以叫功能能力,即如何选择正确的信息、正确地回答问题。而当前火热的语言大模型是文本预测模型,通过预测句子中的下一个单词来生成句子,非常擅长“形式能力”,即保证语法正确的能力很强;但是,在“功能能力”即正确回答问题方面,则是非常吃力。总而言之,虽然语言大模型很擅长说语法正确的话,但是语法正确不等于内容正确。也就是说,语言大模型说的话并不一定是对的,也并不意味着它们很聪明。

这两篇论文好像是在为杨立昆对当前人工智能发展方向的批评意见提供支持。人工智能社区都知道,杨立昆这几年经常公开否定 OpenAI 的技术路线,表示 OpenAI 走的是一条歪路、一条岔道,因为用语言大模型可以提高语言能力,但是无法高效提高逻辑和推理能力。因此,用语言大模型去达成通用人工智能的技术路线是错的^②。他坚持认为,语言模型不理解语言背后的客观世界,智能程度很有限,不能带人类走向 AGI。语料中包含的关于客观世界的知识量也有限,人类的很多智能也跟语言无关,特别是在孩提时代,甚至未开口说话之前通过感知器官预先建立起来的对客观世界的认知,在语料中缺乏对应的表示^③。

5.3 我们认为,虽然迄今为止测试的所有类型的思维都可以在没有语言的情况下实现,也就是说大脑中的语言系统不参与思考、推理、决策等智能活动。但是这并不等于是验证了杨立昆一直以来的看法:当前人工智能聚焦于语言、通过语言大模型去达成通用人工智能的技术路线是错的,美国的人工智能领头羊 OpenAI 走的是一条歪路、一条岔道。也不能证明语言学家乔姆斯基的观点:追捧 ChatGPT 是浪费资源,语言大模型通向通用人工智能的路线完全错了^④。

因为语言是人类了解世界的最重要的信息来源,自然语言文本反映和表达了人类对于外部世界和内心世界诸多的认识。也就是说,在自然语言文本中隐含了近似的世界模型,即人类大脑中构建的关于周围世界的认知模型。因此,如果词语的嵌入式向量提取得好、信息压缩做得好、注意力机制得当,那么语言大模型在合适的提示的启发下,是可以生成合乎世界模型的文本及完成有关的判断、推理和决策任务的。例如,辛顿跟 GPT-4 的对话^⑤:

辛顿:我的所有房间被粉刷成蓝色或白色或黄色,黄色的油漆在一年内会褪色变白。我想让所有房间在两年内都变成白色。我应该做什么,为什么?

GPT-4:假设蓝色的油漆不会褪色成白色,因为黄色的油漆会褪色成白色,也许蓝色的油漆也会褪色,所以假设它不会褪色。那么白色的房间你不需要粉刷,黄色的房间你也不需要粉刷,因为它们会在一年内褪色成白色。而蓝色的房间你需要粉刷成白色。

可见,语言大模型确实能够理解这些关于颜色变化的知识,并且能够进行推理。当然,语言大模型的知识仅限于包含在训练它们的文本语料中,而人类的知识有许多是非语言的。因此,文本结合视频等多媒体数据的训练就尤为重要。

不管怎么说,语言大模型和生成式人工智能是新世纪最重要和最美丽的发明之一,称得上

是一面“21 世纪的镜子”。通过它,我们得以用全新的、富有启发性的视角,来看待语言、思维、世界知识和我们人类自己。然而,当我们看向这面“镜子”的后面时,却发现空无一物,除了一大堆权重数字^⑩。这不由得让人想起辛弃疾的词句:“最爱霏霏迷远近,却收扰扰还寥廓”(《满江红·和范先之雪》)。

注 释

- ① 详见《SearchGPT 第一波评测来了! 响应速度超快还没广告,“OpenAI 杀手铜”》,量子位,2024. 07. 28; https://mp.weixin.qq.com/s/_yReQrPvhOGkCkWQvw8Mrg.
- ② 网络常用语“梗”,实为“诨”(gén)的讹字,意思是笑点、伏笔、桥段。
- ③ 详见《OpenAI 惨遭打脸! SearchGPT 官方演示大翻车,源代码竟暴露搜索机制》,新智元,2024. 07. 28; https://mp.weixin.qq.com/s/0HD7mFjnlwhA4npt_s_hlg.
- ④ 详见 Heaven (2024a)。
- ⑤ 参考维基百科条目“莫拉维克悖论”,另见《语言本能:人类语言进化的奥秘》Steven Pinker 著,欧阳明亮译,浙江人民出版社,1994 年,第 199 页。
- ⑥ 详见注释⑤专著,第 4—8 页。
- ⑦ 详见注释⑤专著,第 XVII、28—29 页。
- ⑧ 详见《OpenAI 科学家震惊 TED 大会:让 AI 模型思考 20 秒,提升 10 万倍性能!》,AIGC 开放社区,2024. 10. 25; <https://mp.weixin.qq.com/s/tE8qoQYoBvxptJa-dLTuOA>.
- ⑨ “中文房间”是哲学家塞尔提出的一个思想实验,详见 Searle (1980);中文介绍详见袁毓林(2021)。
- ⑩ “可供性”指由“主体”(agent)与其环境形成的行动可能性。比如,小溪中的水可以用双手掬起来饮用和洗漱,路边的石块可以坐上去休息,树上的断枝可以用来挂衣服和背包,等等。详见 Gibson (1977)。
- ⑪ 参考 Pezzulo et al. (2024)。
- ⑫ “柏拉图的洞穴寓言”见于柏拉图的《理想国》,参考维基百科条目“地穴寓言”。
- ⑬ 详见《神经网络之父 Hinton 最新公开演讲:数字智能会取代生物智能吗?(全文及 PPT)》,卫 sir 说,2024. 03. 08; <https://mp.weixin.qq.com/s/u72sPc0PxxIaQBfK-TCKJw>.
- ⑭ 详见《AI 教父 Hinton:我支持超级 AI 取代人类!》,新智元,2024. 06. 18; <https://mp.weixin.qq.com/s/fMWBSMCBpQPvzYMZ3XYzzg>.
- ⑮ 对话视频见 <https://www.youtube.com/watch?t=668&v=G14Tpi48DIA&feature=youtu.be>,中文介绍详见《Hinton 揭秘 Ilya 成长历程:Scaling Law 是他学生时代就有的直觉》,量子位,2024. 05. 26; <https://mp.weixin.qq.com/s/3glv0hb4ZyGAMHEO3Dnynw>.
- ⑯ 详见《Ilya Sutskever 离职后首次采访:个人经历,Scaling Laws, AI 幻觉》,AI 寒武纪,2024. 07. 20; <https://mp.weixin.qq.com/s/JeHTvbaEwiHyjd-UVHt7WA>.
- ⑰ 详见 Bricken et al. (2024)。
- ⑱ 详见 <https://x.com/ylecun/status/1801978192950927511>.
- ⑲ 详见韩佳琦(2024)。
- ⑳ 详见陈彬《OpenAI 再陷宫斗:理想主义者是如何被击碎的?》,远川科技评论,2024. 05. 27; <https://mp.weixin.qq.com/s/upYVBweeIoittjLhjBRt8Q>.
- ㉑ 详见 <https://mp.weixin.qq.com/s/A1Eew89vekqiEf9okwc62w>.
- ㉒ 详见 Heaven (2024b)。
- ㉓ 详见 Weng (2024)。
- ㉔ 引自王庆法(2023)。

- ②⑤ 详见极目新闻, <https://news.sina.com.cn/s/2024-10-10/doc-incrzuhy8710517.shtml>.
- ②⑥ 详见好奇心君(2024)。
- ②⑦ 详见《图灵奖得主 LeCun 最新专访:大语言模型的败因和人工智能的未来》,图灵人工智能,2024.03.12; https://mp.weixin.qq.com/s/nqTLAx_IyfoaT4pWosRmbA.
- ②⑧ 详见 Chomsky (2023)。
- ②⑨ 详见《神经网络之父 Hinton 最新公开演讲:数字智能会取代生物智能吗? (全文及 PPT)》,卫 sir 说,2024.03.08; <https://mp.weixin.qq.com/s/u72sPc0PxxwIaQBfK-TCKJw>.
- ③⑩ 此处参考了 Pezzulo et al. (2024) 结尾的表达方式,但是内容有所不同。

参考文献

- 韩佳琦 2024 AI 教父 Hinton 46 分钟对谈: AI 可复制人类心智,或加剧全球贫富差距,智东西,2024.06.17, <https://mp.weixin.qq.com/s/nvhKpEkY9tyEOlhVS89Fdg>.
- 好奇心君 2024 MIT 爆美国 AI 技术路线是错的,未来预言机,2024.06.26, <https://mp.weixin.qq.com/s/Yh4xbiWNRr27AlpY6wnwyw>.
- 刘 宇 2024 为什么“压缩即智能”? 算法信息论与大模型、生命、智能的联系,集智俱乐部,2024.07.19, <https://mp.weixin.qq.com/s/32ZY-CpzQFVJt-vGehY17w>.
- [美] 史蒂芬·平克 2015 《语言本能:人类语言进化的奥秘》,欧阳明亮译,浙江人民出版社.
- 王庆法 2023 通用人工智能 (AGI) 已经到来? 深度解析 ChatGPT 获得智能的数学物理机理,集智俱乐部,2023.11.15, <https://mp.weixin.qq.com/s/KV1gyu1pdl6ErkX7QD-IIw>.
- 卫剑钊 2024 AI 教父 Hinton 最近对人工智能的 7 个观点,卫 sir 说,2024.03.10, <https://mp.weixin.qq.com/s/6SQJqBejhtEfbGTFfhpQzw>.
- 袁毓林 2021 “人机对话-聊天机器人”与话语修辞,《当代修辞学》第 3 期.
- AI 寒武纪 2024 Ilya Sutskever 离职后首次采访:个人经历,Scaling Laws, AI 幻觉, AI 寒武纪,2024.07.20, <https://mp.weixin.qq.com/s/JeHTvbaEwiHyjd-UVHt7WA>.
- Bricken, Trenton et al. 2024 Towards monosemanticity: decomposing language models with dictionary learning. *Transformer Circuits Thread*, <https://transformer-circuits.pub/2023/monosemantic-features>.
- Chomsky, Noam 2023 The false promise of ChatGPT. *New York Times*, Mar. 8, 2023. <https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html>. 中文介绍:终于,乔姆斯基出手了:追捧 ChatGPT 是浪费资源,机器之心,2023.03.10, https://mp.weixin.qq.com/s/MyiLZYE_hcl27i_qtm7lSA.
- Fedorenko, E., Piantadosi, Steven T. & Gibson, Edward A. F. 2024 Language is primarily a tool for communication rather than thought. *Nature*, 630(8017): 575-586. 论文链接 <https://www.nature.com/articles/s41586-024-07522-w>. 中文介绍:语言是交流的工具,还是思维的媒介? 鹿鸣 Cogn, 2024.06.23, <https://mp.weixin.qq.com/s/iMCRtHF14dXPslNZDUOwUw>.
- Gibson, J. James 1977 The theory of affordances. In Shaw, R. and Bransford, J. (eds.) *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*, 67-82, Routledge.
- Heaven, Will, D. 2024a What is AI? Everyone thinks they know, no one agrees. And that's a problem. *MIT Technology Review*, July 10, 2024. <https://www.technologyreview.com/2024/07/10/1094475/what-is-artificial-intelligence-ai-definitive-guide/>. 中文介绍:《麻省理工科技评论》万字长文:什么是人工智能? 心朗物润学堂,2024.08.04, <https://mp.weixin.qq.com/s/I9rHX7onvMr08rtWBG6jVA>.
- Heaven, Will, D. 2024b 前 OpenAI 创始成员 Andrej Karpathy: 大模型已学会用互联网内容“做梦”, DeepTech 深科技,2024.06.21, <https://mp.weixin.qq.com/s/H3vQprAIEVRKxCvo3pljSg>.
- Kahneman, Daniel 2011 *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux. 《思考,快与慢》2012

胡晓娇等译, 中信出版集团。

- Li, Fei-Fei(李飞飞) & Etchemendy, John 2024 No, today's AI isn't sentient. Here's how we know, *TIME*100 VOICES, 2024. 05. 22, <https://time.com/collection/time100-voices/6980134/ai-llm-not-sentient/>.
- Mahowald, Kyle et al. 2024 Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6): 517–540, <https://doi.org/10.1016/j.tics.2024.01.011>.
- Moravec, Hans 1988 *Mind Children: the Future of Robot and Human Intelligence*. Cambridge, Massachusetts: Harvard University Press.
- Pezzulo, G., Parr, T., Cisek, P., Clark, A. & Friston, K. 2024 Generating meaning: active inference and the scope and limits of passive AI. *Trends in Cognitive Sciences*, 28(2): 97–112.
- Pinker, Steven 1994 *The Language Instinct: How the Mind Creates Language*. New York: W. Morrow and Co.
- Romero, Alberto 2024 GPT-5: everything you need to know. <https://www.thealgorithmicbridge.com/p/gpt-5-everything-you-need-to-know>. 中文介绍:三万字详解! GPT-5:你需要知道的一切,青稞 AI, 2024. 08. 17, <https://mp.weixin.qq.com/s/7zvmglJPsrJ1h0mE0Sc1Aw>.
- Searle, R. John 1980 Minds, brains and programs, *Behavioral and Brain Sciences*, 3(3): 417–457.
- Wang et al. 2024 Can Language Models Serve as Text-Based World Simulators? <https://arxiv.org/abs/2406.06485>. 中文介绍:大语言模型≠世界模拟器,机器之心, 2024. 06. 16, https://mp.weixin.qq.com/s/FBqYb_gcBr5D204mDtmCOA.
- Weng, Lilian(翁丽莲) 2024 Extrinsic Hallucinations in LLMs. Lil'Log. <https://lilianweng.github.io/posts/2024-07-07-hallucination/>. 中文介绍:OpenAI Lilian Weng 万字长文解读 LLM 幻觉:从理解到克服,机器之心, 2024. 07. 13, <https://mp.weixin.qq.com/s/UGcui0rLW2Vz7y2Mt4atqA>.

What ChatGPT and Other Large Language Models Get Right: An Investigation into the Origins of Their World Models and Knowledge Sense

Yuan Yulin

Abstract: This paper explains why large language models (LLMs), such as ChatGPT, can generate coherent and logically connected texts. It begins by elucidating the differences between machines and humans in problem-solving and other aspects, clarifying that large language models are not merely “Chinese rooms” that manipulate symbols but rather “Plato’s caves” capable of reconstructing world models based on numerous subtle textual descriptions. Subsequently, the paper introduces some opinions from the founders of large language models, such as Hinton and Sutskever, regarding the working mechanisms of LLMs. By analyzing the meanings of statements containing the noun “table” and the adjective “hungry,” it demonstrates that large models, through training corpora, are fully realizable in acquiring world knowledge about human interactions with objects and sensations to construct approximate world models. Therefore, even if language is solely a communicative tool rather than a thinking tool, large language models can still pave the way for artificial general intelligence (AGI).

Keywords: large language models (LLMs), Chinese Room/Plato’s Cave, world knowledge/world model, artificial general intelligence (AGI)