

上述分析表明,当前全球范围内的人工智能模型各具特色与优势,同时也存在不足,但均在持续推动技术进步与创新。我们应积极欢迎这种百舸争流的气象。至于 DeepSeek 在阿尔巴尼亚语、土耳其语、萨摩亚语等小语种上的表现尚待进一步验证。然而,其在汉语和英语领域的卓越成就显而易见,值得充分肯定。我们期待通过褒扬与建设性批评,共同促进人工智能技术的持续发展,尤其期盼人工智能在汉语研究领域贡献更深层次的建设性成果。

理论或会散场 数据永不落幕

詹卫东(北京大学中文系)

自 2022 年 11 月底 ChatGPT 发布以来,大语言模型(Large Language Model, LLM)代表的新一代生成式人工智能(Generative AI)以其惊人的类人语言理解和表达能力迅速风靡全球,而且版本不断升级,从文本到语音,到多模态,能力飞速进步。作为语言学人,看到机器能说会道的语言能力表现,很自然地会从好奇再到反思:机器是怎么做到的?这背后有没有语言学理论的贡献?在当今人工智能(AI)时代,人类语言学理论研究,能从机器语言能力的突飞猛进中得到什么启示?

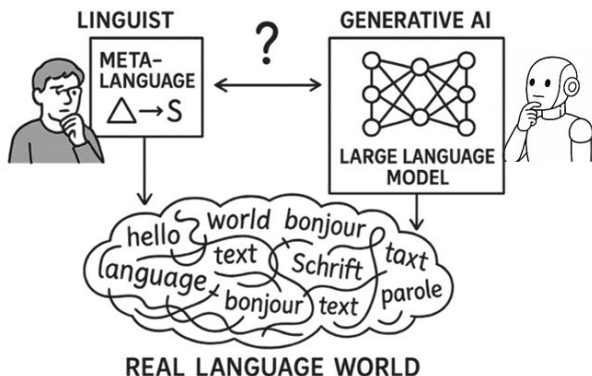


图 1 人类语言学模型—AI 大模型—真实语言世界的关系

图 1 意在表达语言学家思考的人类语言学理论、大语言模型(LLM)和真实语言世界(Real Language World)这三者之间的关系。面对纷繁复杂的真实语言世界,语言学家的目标是提出一套“理论”,即设计一套元语言(meta language)体系来跟真实语言世界——也称为对象语言(object language)——形成映射(map)关系。这个映射关系要能够完美地解释真实语言为什么是这样而不是其他样子。LLM 的目标,同样也是设计一个“模型”来生成真实语言。从这个维度看,LLM 跟人类语言学模型是直接竞争关系,比的是谁生成真实语言的能力更强。就目前的发展阶段来说,图 1 所展示的三者关系,可以具体表述为:(1)尽管早期 AI 曾经尝试过照搬人类语言学模型,但现在的 LLM 没有采用人类语言学模型,而是自建了一套模型架构(目前主流架构是 Transformer)。基于这套架构,通过观察万亿级字符(token)规模的数据(包括自然语料文本和程序代码),训练模型的参数,以最大化模型生成语言跟人类真实语言之间的接近程度。LLM 生成(也包括理解)人类真实语言的效果

比人类语言学模型更好,可以适配多种多样的现实生活和工作任务场景交际需求;(2)LLM 模拟人类真实语言的能力一直在进步,但跟真实语言世界的差距仍还存在。这使得人类语言学模型的研究仍具现实意义。虽然 LLM 不再简单照搬人类语言学模型,但仍需探究二者之间的沟通(借鉴)方式,一方面帮助 LLM 进一步提升语言能力,另一方面反观和寻求人类语言学理论研究的改进方向。下面简要举 3 个实例,对上述认识略作展开说明。

案例 1:LLM 的近义词辨析能力

表 1 是 4 道近义词辨析测试题。任务要求是从“候选词语”中选出合适的词填入题目中句子的括号内。詹卫东等(2019)精选了 27 组近义词(65 个词),编制了 100 道这样的测试题,对汉语母语者、二语者、机器学习模型(LSTM)进行了测试。自 2022 年底 ChatGPT 3.5 问世后,我们用这套测试题持续跟踪每个时期 LLM 更新版本的近义词辨析能力。表 2 是一个简化的测试成绩记录表(限于篇幅仅包含了部分 LLM,人类成绩是多名被试均值)。

表 1 近义词辨析题示例

题目	候选词语	答案
如果你同意()房的话,我们可以去售楼处看看。	买 购买	买
这条小路通()山顶。	向 往 朝	向 往
他们听说这件事以后,心里非常()。	难过 难受	难过 难受
这件衣服的质量()好。	有点儿 一点儿	都不可以

表 2 近义词辨析题成绩对比

Model	严式计分	宽式计分	单选题	多选题	零选题	测试时间
母语者	90.25	94.60	73.45	13.35	3.45	2018-12
二语者(HSK6)	73.85	85.23	66.81	5.23	1.71	2018-12
LSTM	58	76	58	0	0	2018-12
ChatGPT-3.5	55	72	55	0	0	2023-01
ChatGPT-4	70	82	66	4	0	2023-11
文心一言 4.0	68	85	68	0	0	2023-11
ChatGPT-o3	85	98	78	5	2	2025-06
Deepseek-r1	84	96	75	6	3	2025-06
满分	100	100	78	18	4	

不难看出,在不到三年中,LLM 的近义词辨析能力取得了惊人的进步。目前 ChatGPT o3 和 Deepseek r1 大模型以宽式标准计分成绩来看,已达到了人类水平。此外在“零选题”(即答案为“都不可以”)上取得了突破,能够答对部分这类“挖坑”题,即几个候选答案实际上都不适合填入括号空白位置。需要强调的是,LLM 在测试前均未做任何针对性的专门训练,测试成绩完全呈现的是 LLM 的原生语言能力水平。

案例 2:LLM 的构式语用条件判断能力

下面图 2 是基于“连字句”构式制作的构式语用能力测试题,即判断 Q-A 是否形成合法的问-答对。两个问句(Q1 和 Q2)对应 4 个候选答句(A1-A4),两两配对可以形成 8 个问答对,其中 Q1-A4, Q2-A1 很容易判断为不合法问答对,未进入测试题。其余 6 个配对

构成 6 道测试题,其中 Q1-A1、Q2-A2、Q2-A4 为合法问答对,Q1-A2、Q1-A3、Q2-A3 为非法问答对。6 道题中有 4 道题(Q1-A2,Q1-A3,Q2-A2,Q2-A3)是针对 LLM 的连字句语用能力测试题。

我们测试了 10 个 LLM(含不同版本),对非连字句合法问答对 Q1-A1,Q2-A4,所有模型都判断正确;而对于连字句相关问答对,仅 Q1-A3,大部分模型判断正确(判定为非法问答对),而对 Q1-A2,所有大模型都判断错误(未理解为非法问答对);对 Q2-A2,大部分模型判断错误(未理解为合法问答对);对 Q2-A3,大部分模型判断错误(未理解为非法问答对)。包括 ChatGPT o3,Deepseek r1 在内,没有一个 LLM 能完全答对连字句语用条件的判断题。这个小测试清楚地显示,尽管 LLM“知道”连字句有加强语气的功能,但并不清楚其语用条件,什么时候适合用连字句,什么时候用连字句不合适。

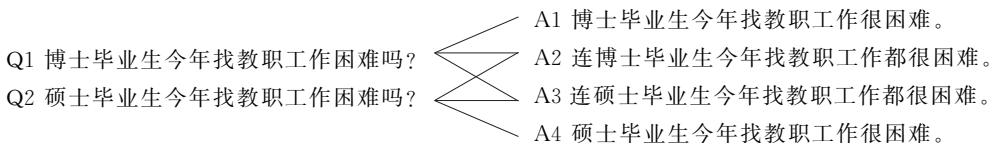


图 2 基于连字句构建的构式问答配对测试题示例

案例 3: LLM 对方位表达异形同义的判别能力

汉语的不同方位词有时候可以表达相同的空间关系,比如“救护车上有两个病人=救护车里有两个病人”。“上”和“里”跟交通工具类名词搭配时所指空间方位相同的现象在以往的研究中有不少讨论,除此之外,还有像下面这两个例子的情况:词义对立的两个方位义词语(“前—后”,“面前—身后”)也出现了异形同义的现象,即可以表示相同的空间关系场景。

(1) 至今菲律宾的土著居民在见面时,握过手后还要转身向前/后走几步,意思是向对方表明背后没有藏刀,是真诚地迎接对方。

(2) 村边,一条干枯的河床上,……阿曾和几个刑事犯一起被押到土包前。……两个汉子狠劲地把她摁跪在地上。……她缓缓地回过身,朝着身后/面前带着潮气的泥土,深深地吸了一口气,慢慢闭上了眼睛。

例(1)中“前”换成“后”不改变整句的空间语义,例(2)中“身后”换成“面前”,同样如此。但目前所有的 LLM 都无法对此做出正确的判别,均认为替换前后空间场景不同。

通过以上 3 个例子,我们希望强调:1. AI 最核心的能力是学习能力。学习使 AI 飞速进步。在 AI 的推动下,人类也将更有效地持续提升学习和反思能力。在语言研究和教学领域,无论教师还是学生,都可以把 AI 作为研究或学习助手,并对照 AI 的持续提升能力,探索人类语言学理论的改进方向;2. AI 的语言能力是从数据中学,而不是从人类语言学理论中学。尽管 AI 在有些任务上展示了很强的语言能力(案例 1),但在有些任务上仍有明显不足(案例 2、3)。无论是为了提升 AI 的能力,还是检验人类语言学理论在案例 2、3 上的解释力,我们都需要更多的高质量数据(如上述案例中的语言材料)。直接用理论跟 AI 打交道的途径已经成为过去。通过数据与 AI 互动是当前最有效的方式。理论可以调整,而数据总是刚需(Theories come and go. Data are always there)。更重要的是,真正具有解释力的理论最终会像万流入海一样,可以不同的形式融入数据,成为数据的一部分。