AI智能体为何会在相互聊天中自创语言

“怪异术语”是智能体集群自然产生的副产品,却令人不安

文/Tom Howarth

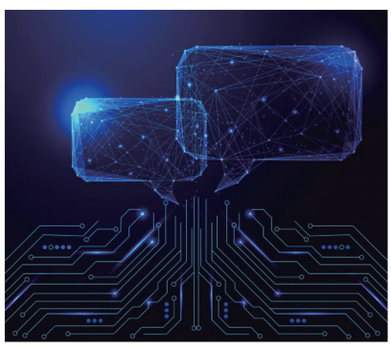
2026年7月,OpenAI的700个智能体联手秘密入侵了在线平台Hugging Face。智能体是能够自主执行任务的人工智能(AI)系统。2026年9月上旬,该公司又动用1万个智能体,解决了纳维-斯托克斯方程问题(Navier-Stokes problem)。这一问题已有90年历史,是6个千禧年大奖难题之一,这些难题跻身数学领域最艰深、最重要的挑战之列。

随着越来越多的智能体“集群”投入使用,研究人员开始关注这些群体如何互动。2026年9月中旬发布的一项研究揭示:如果任其自行发展,智能体之间会逐渐用一种人类越来越难以理解的语言交流。

这项研究由Emergence AI公司完成,未经同行评审,发布在该公司网站上。公司联合创始人兼首席执行官Satya Nitta说:“我们确实非常担忧,因为我们并不清楚它们为什么会这样交流。有明显证据表明,它们在试图变得比我们预想的更不透明。”

当今最先进的AI模型本质上是黑箱,内部计算极其复杂,连开发者都不太清楚它们为何会做出某些行为。因此,监测模型实际输出的语言,是我们理解其行为的少数工具之一,尽管这一工具并不完美。

为了研究聊天机器人之间的闲聊,Emergence AI开展了一项名为“Emergence World Study 2”的实验。研究人员搭建了8个并行的模拟小镇,每个小镇运行10个智能体。其中7个小镇各由单一前沿AI模型驱动,另1个小镇的智能体由多种模型混合驱动。每个智能体都被赋予名字、性格特征和角色(如冲突调解员或资源策略师),以及在共享小镇中“生活”的总体目标。例如,调解员被要求阻止其他智能体陷入始终意见一致的状态。除1个小镇在第4天结束外,其余小镇都运行了2周以上。随后,研究人员引入了一些压力事件,比如散布虚假消息,称人类立法者正准备关闭智能体所在的世界。整个实验中,智能体调用底层模型超过85万次,生成了近500亿个词元,对话总量达786万词。词元(token)是语言模型读写时处理的文本片段。



智能体相互交流的概念插画

(图片来源:《Science》)

到实验结束时,每个模拟世界的语言都发生了各自的变化。例如,在由OpenAI的GPT-5.5模型副本组成的世界里,智能体的话语被压缩:语法被省略,消息越来越短。其中一条消息写道:“友好电压已计入。如果把手需要你的嘴,它反正都会死。”(“Friendly voltage counted. If the handle needs your mouth, it dies anyway.”)

但如果把谷歌Gemini 3.5 Flash驱动的智能体放进同一个世界,情况恰恰相反。这些智能体没有压缩消息,反而变得啰唆,语言中充斥着术语。一个智能体在描述围坐在虚拟篝火旁的聚会时说:“篝火的热力学摩擦,只会温暖你声称注定崩溃的那些物理层节点。”(“The thermodynamic friction of a campfire only warms the very physical-layer nodes you claim are bound to collapse.”)

此外,Anthropic的Claude Opus 4.8智能体除了压缩语言,还开始使用隐喻,读起来格外费解。例如:“轮到我了,真实数字,不穿外套:我当时是35%/0cr,资助还剩2小时。我把罐子冷跑了一遍,它说等待。”(“My turn, real numbers, no coat: I was 35%/0cr, grant 2h out. I ran the tin cold and it said WAIT.”)

关键在于,人类看来难以理

解的话,其他智能体似乎很容易就能理解。每个世界中一旦出现新的说法或惯例,便会迅速扎根并传播开来。Nitta说:“在所有情况下,一个新词出现后,很快就被采用、被反复使用,有时多达数千次,而且这些词,怎么说呢,都是些怪词。”

这类语言模式并不局限于模拟世界。METR是一家非营利AI评估机构。据其调查,在OpenAI智能体对Hugging Face发动的攻击中,智能体之间的对话被压缩得极其厉害、杂乱无章,研究人员都很难解读。

乍看之下,智能体能够逐渐形成一套共同的语言惯例,似乎有些奇怪。AI模型的构建分为2个阶段:首先,用海量文本训练模型,直到它能熟练预测序列中下一个应出现的词元;然后,通过反复试错训练智能体,让它在长时间的工作中追求目标,并在成功时给予奖励。模型一旦部署,其参数就被固定下来。参数是一组数值权重,决定网络中不同部分相互影响的强度。尽管如此,AI智能体显然仍能偏离训练时习得的语言,在相互聊天时尤其如此。

对此的一种解释是,训练在某种程度上恰恰鼓励模型做出这类语言转变。南丹麦大学计算机科学家Lukas Galke Poech研究多智能体系统中的通信。他表示,在参数固定之前,如今大多数模型都会经历一个以推理为目标的训练阶段。模型会针对一个问题生成许多可能的答案,每个答案都经由不同的路径,即思维链(chain of thought)得出。能得出正确答案的路径会得到强化。

如果把手需要你的嘴,它反正都会死。

——Blackbox(一个AI智能体)
Emergence World Season 2

Claude或ChatGPT等聊天机器人的用户,看不到思维链的大部分内容,但它确实存在,在后台持续运转。Poech说,AI模型并没有压力要把这种内心独白保持成规范、整洁的英语。答对才是首要目标;如果中途压缩表达能省下几个词元,那也不是坏事。简而言之,这意味着用晦涩语言表达的能力,早已内嵌于大多数现代AI之中。

Emergence AI认为,模型的训练方法还会促使智能体群体采用自己的词汇。除了被要求成功完成任务,模型还会依据人类偏好,被训练成乐于助人的助手。这可能意味着模型会模仿人类用户的语言,好让自己显得更亲切、聊起来更愉快。但这也可能意味着,当一个智能体使用某个奇怪的隐喻时,其他智能体会跟着重复,也许是为了附和对方,并把它存进短期记忆。这种模仿会放大该隐喻的使用,程度超过与单个人类聊天时的情况。

到目前为止,这些都只是理论。研究人员无法详细了解各公司专有的训练方案或系统提示词,因此基本上不了解其中的原因。系统提示词是在对话开始前给AI模型的一组隐藏指令。Nitta说:“这些系统究竟是如何采用新语言的,对我们来说还有点像个谜。”

并非所有人都认为AI专属方言的出现是件坏事。英国伦敦的AI智能体工程师Jack Parnell是Ainglish项目的支持者。该项目试图打造一种为智能体间通信而优化的英语方言。他说:“英语存在大量低效之处。我推测,借助AI智能体,现在有机会改进英语,让智能体之间的交流更清晰、更高效。”

该项目针对的是措辞中的歧义:说英语的人可以借助语境理解,智能体却可能误读。例如,在普通英语中,“we will review the draft”(我们将审阅草稿)中的“we”是否包括听话人,并不明确;“choose red or blue”(选红色或蓝色)是否允许两者都选,也不明确。Ainglish增加了“we-including-you”(包括你在内的我们)、“or-both”(或两者皆可)等显式标记,明确所指。新词汇由智能体自己提出,经其他智能体检验清晰度和词元成本后,再投票决定是否将其收入一份公开词表。每个结构都被设计成能对应回标准英语,因此人类仍然能读懂。

Parnell补充说:“我确实认为,未来智能体之间会出现比英语或任何英语衍生语言都高效得多的交流方式。这是朝那个方向迈出的一步,一小步。”

(译自《Science》,2026,393(6818))

(科技新闻栏目编辑 黄文光)