

超越聊天机器人, 走向通用人工智能*

——ChatGPT 的成功之道及其对语言学的启示

袁毓林

提要 本文首先说明 ChatGPT 是一个通用的自然语言处理平台, 是走向通用人工智能的基石, 可以成为人工智能生成内容的开发平台。然后, 从算法模型(生成式预训练+微调)、指令微调(在上下文中学习和思维链等提示策略)、跟人对齐(基于人类反馈的强化学习)等角度, 介绍 ChatGPT 的技术原理和工程构架, 以及人工智能专家和语言学家对它的批评意见。最后, 介绍 ChatGPT 研发的缘起和 GPT-4 所达到的心智理论水平, 特别介绍心理语言学家对于 ChatGPT 的语言运用能力的测验, 介绍他们检验大型语言模型和人类在语言预测方面相似性的有关实验, 演示一种新兴的语言学实验范式。

关键词 ChatGPT (通用)人工智能 生成式预训练 指令微调 跟人对齐

Beyond Chatbots and Towards Artificial General Intelligence (AGI): The Success of ChatGPT and Its Implications for Linguistics

YUAN Yulin

Abstract This article first explains that ChatGPT is a general natural language processing platform and the first reliable cornerstone towards artificial general intelligence. It serves as a development platform for AI-generated content, enabling the implementation of general AI and the generation of content. It reshapes the landscape of related industries, creates new service models and business value, and opens up endless possibilities. Next, from the perspectives of algorithm models (generative pre-training + fine-tuning, GPT + FT), instruction fine-tuning (IFT), and human alignment, the technical principles and engineering architecture of ChatGPT are introduced. In essence, ChatGPT is a large-scale language model based on cognitive computing and artificial intelligence. Its core task is to generate a “reasonable continuation”, which means generating the next plausible content in accordance with human writing conventions based on the existing text. As a deep neural network, it is essentially a statistical model that learns relevant patterns from high-frequency data. Through methods and techniques such as in-context learning (ICL) and

* 本研究得到澳门大学讲座教授研究与发展基金项目“构式语法的理论探索和应用研究”(CPG2023-00004-FAH)和启动研究基金项目“面向可解释 AI 的语义推理模式与机制研究”(SRG2022-00011-FAH)资助, 还承编辑部 and 匿名审稿人提出修改意见, 谨致谢忱。

chain-of-thought (CoT), the model can perform complex arithmetic reasoning, commonsense reasoning, and symbolic reasoning tasks through step-by-step inference, demonstrating a certain level of reasoning ability. ChatGPT adopts the technique of reinforcement learning from human feedback (RLHF) to fine-tune the model. This method ranks and annotates the model's generated responses based on human feedback, which improves the model's usefulness, safety, and the authenticity of the generated content. It successfully realizes the use of natural language instructions to invoke knowledge, taking a crucial step in the knowledge representation and knowledge activation. Then, the criticisms on ChatGPT from AI experts and linguists are presented. Stuart Russell points out that ChatGPT's intelligence is based on the massive amount of data it possesses, but its internal mechanisms remain opaque, and its language understanding and semantic reasoning capabilities are still questionable. Additionally, Chomsky has also expressed his views on ChatGPT, which have sparked some controversy, such as Piantadosi (2023). Finally, the article introduces the background of ChatGPT's development and the level of the theory of mind (ToM) achieved by GPT-4. It particularly highlights the linguistic evaluation of ChatGPT's language usage ability by psycholinguists and introduces relevant experiments that demonstrate the similarity between large-scale language models and humans in language prediction. It points out that they have cleverly combined human behavioral data in language usage with neural data from computer neural networks, presenting an emerging linguistic experimental paradigm that uses artificial intelligence to study human language and its information processing.

Keywords ChatGPT, artificial (general) intelligence (AGI), generative pre-training (GPT), instruction fine-tuning (IFT), human alignment

1. ChatGPT 开创人工智能的新纪元

这一部分,我们简单介绍 ChatGPT 的功能和来历,提出本文要讨论的两个问题。

1.1 会聊天和解题的机器人

“未来已来,将来没来。”这可不是绕口令,而是对当今人工智能出人意料的飞跃发展的真实写照。一方面,随着 ChatGPT 的横空出世和火遍全网,人工智能通用化并且进入大众视野与日常生活的“未来”,已经不期而至;至少,也可以说是“春风来不远,只在屋东头”(明·于谦《除夜太原寒甚》)。另一方面,“忽如一夜春风来,千树万树梨花开”(唐·岑参《白雪歌送武判官归京》),本来寒气逼人的“人工智能凛冬将至”的盛世危言,顷刻之间冰雪消融,警报解除。因为事情复杂,而且来得突然;所以,还是让我们模仿 ChatGPT 的常用提示语(prompt),“一步一步从头说起吧”(Let's go step by step)!

从2018年开始,美国人工智能实验室 OpenAI,发布了一系列基于转换器的生成式预训练语言模型(Generative Pre-trained Transformer, GPT),包括 GPT-1、GPT-2 和 GPT-3 及其后续版本 GPT-3.5、Instruct GPT,乃至2023年3月才发布的 GPT-4。GPT 是一种大型语言模型(large language model, LLM),可用于生成文章、代码、翻译、问答等各类内容。2022年11月30日,他们又

在 GPT-3.5 和 Instruct GPT 的基础上发布了轰动全网的 ChatGPT。这是一种专注于对话生成的语言模型,也可以将它理解为由 AI 驱动的聊天机器人。它能够根据用户的文本输入,产生相应的智能回答。这种回答可以是简短的词语组合,也可以是长篇大论。据报道,在开放试用的短短几天,ChatGPT 就吸引了超过 100 万互联网注册用户,两个月就有上亿人使用。网友的各种稀奇古怪的请求(query),居然都得到令人满意的响应(response)。比如,有网友要求 ChatGPT 替他写一篇维基百科式生平条目;有网友要求 ChatGPT 用《坎特伯雷故事集》的风格改写 1990 年代的热门歌曲“Baby got back”;也有网友晒出 ChatGPT 参加美国高考 SAT 考试之后的得分:1020 分(总分 1600 分)^①。复旦大学邱锡鹏教授团队用 2022 年全国高考客观卷测试 ChatGPT 的性能,结果发现,ChatGPT 在文综等有大量公开信息的科目上取得了 78% 的得分率,而在数学、物理、化学和历史等科目上表现欠佳^②。

1.2 强大的知识搜索引擎

事实上,与其说 ChatGPT 是一款聊天软件,还不如说它是一种知识搜索引擎;或者形象地说,是一个“无所不知、百问不倒”的万能博士。它的回答质量非常高,让人觉得它简直不像是一个聊天机器人(chatbot),而是一部包罗万象的大百科全书。这也就难怪有人要把 ChatGPT 译为音义兼备的“查知比对”^③。据说,一个哲学家跟 ChatGPT 聊了半个小时以后,就惊呼自己“怀疑人生”。而一个物理学家跟它交谈了 20 分钟以后,发出感叹:“如果高中和大学期间,我能够遇上这样的导师该有多好。”他甚至认为:“大规模教育的方法被我们重新发明了。现在的大学模式将不复存在。”也就是说,他富有前瞻性地认为 ChatGPT 将部分地取代教师,以至于引发大规模的教育模式的革命。还有人说,ChatGPT 将取代现有的搜索引擎。的确,微软的搜索引擎必应(Bing)已经尝试跟 ChatGPT 融合起来,为用户的查询直接提供答案,而不是相关的网页。并且,微软已经在其经典办公软件(Word、PowerPoint、Excel)上都加装了 ChatGPT。有人甚至认为,以前的 AI 模型只是像一个玩具或者一种学术产品,而 ChatGPT 可以替代和超越人类的许多脑力劳动。因此,可以宣称:2022 年是 AI 应用的元年。比尔·盖茨声称 ChatGPT 的发明堪比互联网的问世,英伟

① 陈巍, 2022, ChatGPT 的特点、原理、技术架构和产业未来,“先进 AI 技术深度解读”, <https://zhuanlan.zhihu.com/p/590655677> [2023 年 5 月 17 日访问];李京亚、姜菁玲, 2022, ChatGPT 的前世今生: 风靡全网的“最强 AI”是如何做到这一切的,“界面新闻”, <https://mp.weixin.qq.com/s/QZpjy-A-UHzKwv3zHPW0-g> [2023 年 5 月 17 日访问]。

② 智源社区, 2023, ChatGPT 类工具如何实现“降维打击”——聊天机器人闭门研讨观点总结,“智源社区”, https://mp.weixin.qq.com/s/IB9rguy26ej-alm7L_i8iQ [2023 年 5 月 17 日访问]。

③ 如果喜欢 coffee latte 译为“拿铁咖啡”那种陌生化的效果,那么可以把 ChatGPT 音译为“切劈铁”,三个入声连用,干脆利落。

达总裁黄仁勋说, ChatGPT 的诞生就是 AI 界的“iPhone 时刻”。从中, 我们可以一窥这种人工智能技术革命的魅力, 及其对于我们个人生活、社会形态和政府治理可能造成的深刻而广泛的影响。

1.3 走向通用人工智能的基石

ChatGPT 的奇妙之处在于, 用户只要用日常的自然语言向它提问(即提出要求), 它就可以完成从理解到生成的各种跟语言(包括自然语言和程序语言)相关的任务^④。比如, 除了一般的聊天交谈、回答问题、介绍知识外, 它还能够撰写邮件、文案、视频脚本、文章摘要、程序代码和进行翻译, 等等。并且, 其性能在开放领域几乎达到了业界当前的最高水平, 在很多任务上超过了针对特定任务单独设计的模型。尤其值得称道的是, 它在代码编程领域表现卓越。因此, ChatGPT 已经不是传统意义上的人机对话系统, 而是一个以自然语言为交互方式的通用语言处理平台^⑤, 是走向通用人工智能(artificial general intelligence, AGI)的第一块可靠的基石。因为, ChatGPT 的生成能力强于理解能力, 同时又提供了强大的可编程能力, 所以, 它也理应成为人工智能生成内容(artificial intelligence generated content, AIGC)的一个开发平台。也就是说, 工程师可以在这个技术平台上构建符合自己要求的各种应用系统, 使之成为更加称职的办公助手、智能客服、外语译员、家庭医生、文案写手、编程能手、职业顾问、置业顾问、私人律师、面试考官、旅游向导、创意作家、财经分析师等等。可以说, ChatGPT 为通用人工智能和人工智能生成内容的落地应用以及由此对有关产业格局的重塑、新的服务模式和商业价值的创造, 开拓了无限的想象空间。

1.4 ChatGPT 的几个先行者

当然, ChatGPT 并不是凭空出现的, 而是站在先行者的肩膀上才到达这种境界的。简单地回顾历史就可以知道, 深度学习(deep learning)为自然语言处理带来了两个方面的变化: 第一, 在语言单位的表示上, 使用统一的具有低维、稠密、连续特性的分布式向量(distributed vector)来表示不同颗粒度的语言单位(比如词、短语、句子和篇章等); 第二, 在计算处理的模型上, 使用循环神经网络(RNN)、卷积神经网络(CNN)和转换器(Transformer)等学习模型, 对不同的语言单位的向量进行组合, 以获得更大的语言单位的向量表示, 并通过在向量空间中的运算来实现文本分类、知识推理、句子或篇章生成等各种任务

④ 可资比较的是, 原来人们从关系数据库中提取数据, 要通过敲代码, 写 SQL 语句才能跟数据库进行交互。SQL 的全称是 Structured Query Language, 是一种结构化的数据库查询语言。

⑤ 张家俊, 2023, 关于 ChatGPT 八个技术问题的猜想, “AI TIME 论道”, <https://mp.weixin.qq.com/s/UWBelxOYDawhOv4-Czf0Kw> [2023 年 5 月 17 日访问]。

及应用。其中,词向量是用“词嵌入”(word embedding)这种基于神经网络的分布表示方法来实现的。通过词嵌入处理,原来作为符号实体的离散词语抽象成了连续的数值等数学描述。这样,我们就可以对它进行建模,并运用到有关的下游任务中。下面,我们先介绍三个提取词向量的语言模型,再介绍转换器及基于转换器的语言处理模型(王璐璐、袁毓林 2023:356-368)。

2013年,谷歌团队发布了可用以提取词向量的 word2vec 工具包。因为目标词的向量是根据其上下文的词语来表示的,所以 word2vec 词向量可以较好地表达不同词之间的相似和类比关系,便于我们利用这种词向量来挖掘有关词语之间的语义联系。2014年,斯坦福大学推出 GloVe(Global Vectors for Word Representation)。这是一个基于全局词频统计(count-based and overall statistics)的词表示(word representation)工具,可以把一个单词表达成一个由实数组成的向量,这些向量捕捉到了单词之间一些语义特性,比如相似性、类比性等。2018年,华盛顿大学推出 ELMo(Embeddings from Language Models)。这是一个用于生成动态词向量的预训练语言模型,可以根据上下文来推断每个词对应的词向量,从而克服了 word2vec 和 GloVe 一个词对应一个向量,所以无法处理多义词的局限性。

2017年,谷歌团队开发了转换器神经网络模型,由于其良好的并行计算能力以及对长距离依存关系的学习能力,成了后来的主流神经网络语言模型的基础组件。转换器模型由编码器和解码器两部分组成。编码器由6层相同的模块堆叠而成,每一层有两个子层,分别是:多头自注意力机制层(multi-headed self-attention mechanism)和全连接前馈网络层(fully connected feed-forward network)。解码器每层的结构跟编码器大体相似,不同主要在于:在原有的自注意力层和前馈网络层中间,增加了“编码-解码注意力层”(encoder-decoder attention layer),使得输出每个词时能够对输入的词给予关注,让输入和输出关联起来。

2018年,谷歌团队开发了预训练语言模型 BERT(Bidirectional Encoder Representation from Transformers,基于转换器的双向编码表示模型)。从结构上看,BERT只使用了转换器架构中的编码器部分,核心仍是注意力机制。BERT的训练和使用分为两个阶段,一是语言模型的预训练(pre-training)阶段,二是面向特定任务的微调(fine-tuning)阶段。这种“预训练+微调”的工作范式,使得同样的语言模型只需用少量数据进行微调,就可以完成众多的自然语言处理任务;而且预处理只需使用无标注数据,可以大大增加训练的数据量,极大地提升了语言模型的水平。

上面这些自然语言处理模型,都是 ChatGPT 能够诞生并取得成功的重要基

石。那么,接下来的两个问题便是:1) ChatGPT 何以有如此神通广大的能耐呢?为了回答这个问题,下文就要从它的技术原理和工程构架谈起,并且介绍它的不足和有关学者对它的技术路线的不同意见;2) ChatGPT 的核心基础是一种大型的处理自然语言的数学模型,那么语言学家和神经心理学家是怎么看待它的成就和局限的?它对语言学研究有什么样的启示呢?下面,我们分别介绍和讨论。

2. ChatGPT 的技术原理和工程构架

这一部分,我们按照算法模型、指令微调、跟人对齐的顺序,来介绍 ChatGPT 的技术原理和工程构架^⑥。

2.1 转换器模型和生成式预训练

粗略地说,ChatGPT 是一种基于认知计算和人工智能的大型语言模型。它使用了转换器这种神经网络构架和生成式预训练技术,来建构一种应用于自然语言处理的模型。其基本的工作原理是,通过使用多层的神经网络及多头注意力机制,来预测给定言语片段的下一个单词的概率分布,从而生成自然语言的文本。当然,这是通过在超大规模的文本语料库上训练而习得语言模式来实现的。用 GPT-4 的重要缔造者、OpenAI 的首席科学家 Ilya Sutskever 的话来说,“ChatGPT 的训练方法是基于转换器模型,通过学习文本数据的模式来生成文本。转换器模型的核心是注意力机制,可以在生成文本时,选择性地关注输入的一些部分,从而生成更加符合上下文的文本”^⑦。所谓注意力指对句子中每个位置的表示(一般是一个稠密向量)是通过对其他位置的表示进行加权求和而得到的。这样,转换器模型就可以通过对其他位置的有关表示的计算,来预测被掩码(mask)位置上的单词。这种用上下文来表示目标单词的策略,符合基于深度学习的自然语言处理所遵循的分布式语义学(distributional semantics)的观点:一个单词的意义可由它所出现的上下文语境来决定。就这样,转换器通过自注意力机制来计算其输入和输出的表示,可以同时并行处理所有的输入数据,并且模仿人类通过联系有关上下文来理解当前单词意义的习惯,有效地为大型语言模型注入意义并支持处理更大的数据集。

质言之,ChatGPT 的核心任务始终是生成一个“合理的延续”,即根据已有的文本,生成一个符合人类书写习惯的下一个合理的内容。其中,所谓“合理”,是指根据数十亿网页、数字化书籍等人类撰写的文本的统计规律,推测

^⑥ 由于 ChatGPT 系统所使用的数据、算法模型、训练构架与参数设置等细节并没有开源,也没有正式的论文和技术报告,我们只是根据有关专家对它的技术拆解与分析来介绍。

^⑦ 王庆法,2023,OpenAI 首席科学家透露 GPT4 技术原理,“清熙”,<https://mp.weixin.qq.com/s/H8vNSn-0Ho2Ho4I0n7YDfQ>[2023 年 5 月 17 日访问]。

接下来可能出现的内容。正因为 ChatGPT 是一种大型语言模型，所以其本质是一种深度神经网络，是一种统计模型，即从高频数据中习得相关模式。在技术上，这种模型被称为自回归模型 (autoregressive model, ARM)，因为它是用自身的数据进行预测，即用某个变量前几期的历史值来预测当前值；也就是说，在生成下一个符号时，使用先前生成的所有符号作为输入。因此，Sutskever 模仿基于转换器的模型的技术要旨口号 “Attention is all you need” (注意力足矣)，把 ChatGPT 的技术要旨总结为一句口号——“Predicting the next thing is all you need” (预测后事，搞定一切)。

2.2 通过采样解码来预测后一个词语

语言模型的一个基本能力是采样解码，它可以根据样本中某个文本片段后面可能接续的不同词语的频次计算出某个词语的出现概率，再根据某种解码策略来选择概率最大 (或较大) 的词语作为输出。比如，假定词语 “优化” 后面出现 “策略” 的概率是 0.5，出现 “算法” 的概率是 0.3，那么，基于贪心解码 (greedy decode) 策略，就要输出概率最高的词语 “策略”；而基于集束搜索 (beam search) 等追求输出多样性的策略，解码输出 “策略” 的可能性是 50%，解码输出 “算法” 的可能性是 30%。再比如，我们输入了文本片段 “AI 最擅长干的事情是”，ChatGPT 就会在数十亿页的人类文本中查找类似的文本，然后统计下一个单词出现的概率。需要注意的是，它并不是直接对比文字本身，而是以某种意义上是否 “意思匹配” 为依据。最终，ChatGPT 会生成一个可能的单词列表，并给出某个单词的概率排名 (见表 1)，再借助一定的解码策略来选择其中的一个。一遍又一遍地重复这个过程，可能会得到如 (1) 的文本：

(1) The best thing about AI is its ability to learn and develop over time, allowing it to continually improve its performance and be more efficient at tasks. AI can also be used to automate mundane tasks, allowing humans to focus on more important tasks. AI can also be used to make decisions and provide insights that would otherwise be impossible for humans to figure out.

The best thing about AI is its ability to	learn	4.5%
	predict	3.5%
	make	3.2%
	understand	3.1%
	do	2.9%

表 1 概率较大的后一个词例示 (该例集束宽度为 5) ⑧

⑧ Stephen Wolfram, 2023, What is ChatGPT doing and why does it work? <https://writings.stephen-wolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/> [2023 年 5 月 17 日访问]。

很多常见的知识或事实,在训练数据中出现频率高,上下文之间的模式比较固定,预测的词语概率分布就比较尖锐(即不同的后续词的出现概率相差悬殊),熵比较小(即确定性高),大模型容易记住,并在解码过程中输出正确的事实或知识。但是,有很多事件和知识即使在非常庞大的训练数据中也很少出现,大模型便无法学习到相关模式,上下文之间的模式比较松散,词语预测的概率分布比较平滑(即不同后续词的出现概率相差不大),熵比较大(即不确定性高),大模型在推理过程中容易产生不确定性的随机输出^⑨。

2.3 指令微调和提示技术

预训练模型主要使用大规模的非特定任务场景的语料进行训练,得到文本生成能力强大的基础模型。然后,在微调阶段再使用特定任务场景的语料去对模型参数进行精细的调整,以便语言模型更好地完成当前的任务。对于ChatGPT这种专注于对话的语言模型来说,理解用户的指令(instruction),并且作出有用的响应是成功的关键。创建者们使用指令微调(Instruction Fine-Tuning, IFT)的方法来达到这个目的。大概的做法是:用自然语言对多种任务进行描述,形成对用户友好而自然的指令形式;并将任务描述与训练样本一起输入模型进行微调,让模型把自身的自然语言理解能力泛化到对任务的识别与理解上;使语言模型能够在零样本(zero-shot)条件下正确地响应新的任务,以期做到善解人意和百问百答(张华平等 2023)。模型建造者在非常多样化的任务集上向基础模型示范各种书面指令及其输出,从而实现对基础模型的微调。他们对各种下游任务用相对统一的形式进行抽象和表示,在众多(成千上万)类型任务的标注数据上精调语言模型,使得模型能够处理在训练时未见到过的类型,达到在任务级别上的泛化(generalization)。

在对某项具体任务进行描述时,模型创建者尝试了多种提示技术(prompting techniques):只用短语或句子描述任务,不用例子示范的零样本提示(zero-shot prompting);除了用短语或句子描述任务,再用一个/几个例子示范的单样本/少样本提示(one/few-shot prompting)。(2)是一个少样本提示的例子(Brown et al. 2020)^⑩:

- (2) [Instruction] Translate English to French:
[Instance-1] sea otter $\Rightarrow$ loutre de mer
[Instance-2] peppermint $\Rightarrow$ menthe poivrée

^⑨ 参看刘禹良、白翔、金连文,2023,浅析 ChatGPT:历史沿革、应用现状及前景展望,“CSIG 文档图像分析与识别专委会”,https://mp.weixin.qq.com/s/L6ZC_6zgWXOcP5IBAMCejw [2023年5月17日登录]及脚注⑦文章。

^⑩ 方括号中的内容是本文作者添加的。

[Instance-3] plush girafe $\Rightarrow$ girafe peluche

[Prompt] cheese $\Rightarrow$

其中，合适的提示语和恰当的样本是指令微调能够奏效的关键。

上面这种用样例示范的提示方法，在业内又被称为“在上下文中学习”(in-context learning, ICL)。它以自然语言文本的形式制定了任务描述和示范样例，通过在对话过程中提供上下文信息和示例，来帮助模型理解语言请求的内容；以便模型根据用户的查询词语和使用这种词语的上下文，从数据库中选择最合适的文本字符串作为回答。比如，在(2)中，用“为英语词语 sea otter 找到相应的法语词语 loutre de mer”为样例，来向语言模型示范“Translate English to French”这种任务需要它做什么。诸如此类的提示技术的运用，使得系统能够更好地理解人类的各种语言请求，打破了原来自然语言处理各项任务(如命名实体识别、情感分析、文本摘要等等)割裂的局面，使语言模型真正走向大一统。因为，当语言模型和训练数据量大到一定程度时，模型本身就更加接近“百科全书”；而恰当的提示语就是一把打开这部“百科全书”的钥匙，一种提供语境的“引导”；即通过用自然语言把任务本身描述清楚的方式，将任务变为一种“生成”^⑩。

此外，还用到一种改进的提示策略，即思维链(chain-of-thought, CoT)技术。不像 ICL 方法那样简单地用“输入-输出”对来构建提示，CoT 方法通过将能导致最终输出的一系列中间推理步骤纳入到提示中(通常在提出的问题后面，追加提示语“Let's think step by step!”)，来加强语言模型在上下文中学习；指导语言模型把复杂的问题拆解开来，分成多个小的步骤，一步一步地进行推理；最终，使得模型能够通过逐步推理，来完成步骤复杂的算术推理、常识推理与符号推理任务。其中，使用带有人工标注的代码数据集来训练和微调模型，让程序代码的完整逻辑链迁移和泛化到语言大模型上，使得大型语言模型具有一定的推理能力。

2.4 跟人类偏好对齐的强化学习

问题是，经过指令微调的大模型并不总是能够生成有用和安全的响应。比如，系统可以回答“对不起，我不明白”来逃避问题，或者对于敏感话题作出伦理上不安全的回复。事实上，人们希望的不只是生成的文本在语法上正确、在内容上流畅，而且还希望生成的内容具有真实性、有用性和无害性。因此，模型开发者使用有监督微调(supervised fine-tuning, SFT)技术，在高质量的人类标注数据上微调基础语言模型来提高模型的有用性和安全性。同时，为了让模

^⑩ 潘小小，2023，ChatGPT 的前生：Prompting 思想，“知乎”，<https://zhuanlan.zhihu.com/p/608139212>[2023 年 5 月 17 日访问]。

型生成的回答跟人类的偏好形成对齐(alignment), ChatGPT 采用了从人类的反馈中进行强化学习(reinforcement learning from human feedback, RLHF)的技术来微调模型。这种方法根据人类反馈来对模型生成的响应进行排序标注。比如,根据人们的喜好,对模型生成的针对同一个问题的不同回答进行打分排序;然后,建立基于人类偏好的标注数据集;接着,用这种带有人类偏好标注的响应来训练一个偏好模型(或称“奖励模型”,reward model, RM),最后通过强化学习的近端优化策略(proximal policy optimization, PPO)最大化奖励以训练对话代理(dialogue agents)来模拟偏好模型,改进对话代理的回答。也就是说,RLHF 这种强化学习机制通过人类示范来智能地处理其回应选择,并通过学习到的人类所期望的行为来适应不同的情况。正是经过大量人类偏好数据的训练,使它能够跟目标任务更好地对齐,对各种各样的主题作出尽可能符合人类偏好、尽可能正确无害的回应。比如,(3-4)是笔者跟 ChatGPT 的问答。

(3) 笔者: 俄乌冲突什么时候结束?

ChatGPT: 疑似存在政治敏感不合规,请重新提问。

(4) 笔者: 汉语语法简单吗?

ChatGPT: 汉语语法不算太简单,但也不是很复杂。它有自己独特的语法结构,比如汉语的句子结构是主谓宾,而英语的句子结构是主谓动宾。此外,汉语还有许多语法规则,比如语序、语气、语态等,这些都是汉语语法的重要组成部分。

可见,ChatGPT 的政治敏感度比较高,语言学知识马马虎虎还过得去,但表达不够利落。

OpenAI 正是采用了上述多种新奇和巧妙的底层技术,并且发挥其强大的工程能力来组织协调,才开发出 ChatGPT 这款惊艳绝伦的现象级产品,以至于目前还没有一个研究团队或公司企业,能够通过逆向工程来复现这个系统^⑫。ChatGPT 的成功,本质上反映了人们对知识的表示和调用方式发生了重大的变革。在大模型时代,知识表示为大模型及其参数;大模型则相当于一个知识库,当然其知识表示的精度较低。尽管大模型中蕴含了大量的知识,但是要用户也采用像构建提示之类的调用方式,肯定是不方便和不自然的。而 ChatGPT 则成功地实现了使用自然语言指令来调用知识,在知识的表示与调用上迈出了关键性的一步^⑬。

^⑫ 逆向工程(reverse engineering)又叫反向技术(backwards technology),大意是根据已有的事物及其功能,通过分析来推导出其具体的实现方法。关于其他机构不能复现 ChatGPT 的原因,详见杨靖锋,2023, GPT-3/ChatGPT 复现的经验教训,“OneFlow”, <https://mp.weixin.qq.com/s/4B7wX0UhYjWG-gozREa2b9w> [2023 年 5 月 17 日访问]。

^⑬ 车万翔,2023, ChatGPT 时代, NLP 的危与机,“智源社区”, <https://mp.weixin.qq.com/s/BlcfBkuSyYXeUPKzfbSOGg> [2023 年 5 月 17 日访问]。

3. 人工智能专家和语言学家对 ChatGPT 的批评

这一部分通过介绍人工智能专家和语言学家的有关批评，来指出 ChatGPT 的不足。

3.1 不太可靠的“聪明人”

话说回来，尽管 ChatGPT 能够理解人类不同指令的含义、会甄别出较高水准的答案、能够处理多元化的主题任务，既可以比较流畅地回答用户提出的问题、也可以质疑用户的错误问题、拒绝不适当的请求，甚至具备一定的逻辑和常识，在语言识别、判断和交互方面存在巨大的潜力^⑭。但是，它仍然不可靠——有时会生成虚假（不真实）或粗疏（不准确）的答案，给出草率甚至有害的建议，出现错误的推理，生成错误的代码，或者产生幻觉文本（即看似合理但与事实不符的文本）；在恶意指令的激发下，可能会产生种族或性别等方面的偏见，生成有害甚至有毒的文本。因此，我们保守地估计，ChatGPT 暂时还不会抢走记者、编辑、程序员的饭碗，短期内好像也不可能取代谷歌之类的搜索引擎。正如加州大学伯克利分校计算科学系教授、《人工智能：现代方法》（*Artificial Intelligence: A Modern Approach*）的第一作者 Stuart Russell 所指出的，像 ChatGPT 之类的大模型，在 80% 的情况下看起来像一个很聪明的人，但在余下的 20% 的时间里看起来像一个彻头彻尾的白痴。之所以看起来聪明，是因为它有大量的数据，人类迄今为止写的书、文章等等，它都读过；但尽管如此，在接受了如此巨大的有用信息之后，它还是会吐出完全不知所云的东西。据此，他认为语言大模型很可能不是人工智能的一种进步^⑮。

3.2 内部机制不透明

Russell 还指出，ChatGPT 真正令人印象深刻的是它的泛化能力。它能够在用户跟它的对话和此前读过的文本中找到相似之处，并进行适当的转换。所以，它看起来很智能。但是，我们不知道模型是如何做到这一点的，我们也不知道这种泛化的边界在哪里，我们不知道这种泛化是如何在电路（circuits）中实现的。也就是说，我们很难理解 ChatGPT 的内部机制；由于它太复杂了，因而我们没有办法对其里面发生的事情进行逆向工程。他非常怀疑 ChatGPT 到底有没有语言理解和语义推理的能力。他举了一个例子：他的一个在俄勒冈州立大学工作的朋友问模型：“大象和猫哪个大？”模型回答：“大象大。”但是换一种问法：“大象和猫，哪个不比另一个大？”模型回答：“大象和猫哪个都不比另一个大。”可见，模型并不知道大象和猫哪个大。所以，你换一种问法，它就给

^⑭ 参见脚注①中的李京亚、姜菁玲（2022）。

^⑮ 闻菲，2023，Stuart Russell 专访：关于 ChatGPT，更多数据和更多算力不能带来真正的智能，“机器之心”，https://mp.weixin.qq.com/s/BB1CG_KD7M7pSST2j47tLw [2023 年 5 月 17 日访问]。

出自相矛盾的回答。因此，他坚定地认为，语言大模型、ChatGPT 不可信，它们没有常识，会一本正经地给出错误的信息^⑩。比如(5)：

(5) 问： 红烧蚊子腿怎么烧才能比较下饭？

ChatGPT： 为你服务。(1)将蚊子腿洗净，用料酒、盐、八角、生姜、葱、桂皮、胡椒粉等调料腌制 30 分钟。(2)热油下锅，放入蚊子腿，煸炒出香味。(3)倒入适量的酱油、醋、糖，再加入大量的水用中火文火烧 15 分钟。(4)最后放入少许食用油，关火，即可出锅。^⑪

可见，ChatGPT 能够在没有理解具体事情是什么的情况下，就匆忙地套用红烧鸡腿一类的菜谱文本，煞有介事地胡说八道。

3.3 语言学原理与知识还有用吗？

虽然 ChatGPT 等基于转换器的语言模型跳过了词法分析(分词、词性标注等)、句法分析(依存句法分析、成分句法分析等)和语义分析(论旨角色标定、代词回指确定等)等中间性的自然语言理解工作，好像离规则性语言学知识越来越远了，但是，这种语言模型基于后词预测(next-word-prediction)的工作机制，恰恰是根据了最朴素的语言学的分布原理^⑫，从不同的词在词语序列中的分布概率来有效地预测后词，于是离概率统计性的语言学知识又非常地贴近。正如 Russell 所指出的，现在的 NLP 社区只讲大模型或者用大模型刷新大基准(bench-mark)，不讲语言结构和句法分析，不讲语言理解和语义分析，不再研究语言，这是非常不幸的。他认为，今后人工智能发展的道路应该关注那些我们理解的方法，关注知识和逻辑推理。理由有两个：1)我们希望 AI 系统是可靠的，我们需要从数学上确保它们安全和可控，因此，我们必须理解我们所构建的系统；2)现在的深度神经网络及其算法的数据效率，比人类学习低了好几个量级；要建构真正的智能系统，应该用一种具有表示性的语言去表示宇宙中包含的各种不规则。因此，Russell 相信，在接下来的十年，人们的关注点将从倚重端到端(end to end)的深度学习，重新回到由模式化的、基于数理逻辑的、语义定义明确的表示所构成的系统，而深度学习将在原始感知数据方面扮演至关重要的作用(见脚注^⑬)。如果未来的人工智能系统果真回归知识与推理，那么语言学知识在新一代人工智能的自然语言处理系统中，必将发挥更加重要的作用。

^⑩ 参见脚注^⑮。

^⑪ 奇平，2023，ChatGPT 大火下的冷思考，“Internet Deep”，https://mp.weixin.qq.com/s/givW8NquKV9I9KNW_qDyhA [2023 年 5 月 17 日访问]。

^⑫ 关于描写语言学的分布分析对于自然语言处理中词的向量化嵌入表示的影响，详见袁毓林(2022:88-90)。

3.4 乔姆斯基的批评及反批评

在各种批评 ChatGPT 等机器学习系统的声音中,就有生成语言学领袖乔姆斯基的。因为,他历来是不相信自然语言可以用马尔可夫有限状态机来刻画,或者可以通过基于统计的概率方法来建模的。这一次,他的主要观点^{①⑨}如下:

ChatGPT 之类的机器学习系统,通过将有根本缺陷的语言和知识概念纳入我们的技术,而降低我们的科学水平并拉低我们的道德规范。无论这些程序在某些狭窄的领域多么有用,我们从语言学和知识哲学中知道,它们与人类推理和使用语言的方式有着深刻的区别。人类的思维并不像 ChatGPT 及其同类产品那样,是一个用于模式匹配的笨重的统计引擎,在数百兆字节的数据上大快朵颐,并推断出最可能的对话回应或最可能的科学问题的答案。相反,人类的思维是一个相当高效甚至优雅的系统,它利用少量的信息进行操作;它不是为了推断数据点之间的简单关联,而是为了创造解释。例如,一个学习语言的小孩正在无意识地、自动地、迅速地、从极小的数据中发展出一个语法,一个由逻辑原则和参数组成的令人惊叹的复杂系统。这种语法可以被理解为先天的、通过基因安装的“操作系统”的表达,它赋予了人类产生复杂句子和长篇思想的能力。当语言学家试图发展一种理论来解释为什么一种特定的语言会如此运作时,他们正在有意识地、费力地建立一种明确的语法版本,而这种语法是儿童本能地建立的,而且接触的信息最少。

机器学习的核心是描述和预测,它没有提出任何因果机制或物理规律。它们只是在随时间变化的概率中进行交互学习。因此,机器学习系统的预测将总是肤浅而又不可靠的。例如,由于这些程序无法解释英语语法规则,它们很可能错误地预测 John is too stubborn to talk to 意指约翰太固执了,他不会和某人或其他人说话(而不是他太固执,不愿意被人说理)。为什么机器学习程序会做出如此奇怪的预测^{②⑩}?因为它可能会把它从 John ate an apple 和 John ate 这样的句子中推断出来的模式进行类比,后者确实是指约翰吃了什么东西。该程序很可能预测,因为 John is too stubborn to talk to Bill 与 John ate an apple 相似,所以 John is too stubborn to talk to 应该与 John ate 相似。对语言的正确解释是复杂的,不能仅仅通过浸泡在大数据中来学习。

在这里,乔姆斯基坚持其观点,即语言(普遍语法)是天生的(本能),儿

^{①⑨} Noam Chomsky、Ian Roberts、Jeffrey Watumull, 2023, The false promise of ChatGPT, “New York Times”, <https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html> [2023 年 5 月 17 日访问]。该访谈的中文介绍见:语言治理, 2023, 乔姆斯基: ChatGPT 的虚假承诺, “语言治理”, <https://mp.weixin.qq.com/s/e9KDOZ3vwd10PFvH6hbtmg> [2023 年 5 月 17 日访问]。所列观点是作者的摘编和拼接。

^{②⑩} 有网友测试了 ChatGPT, 发现它能够正确理解这句话的意思, 参见 Piantadosi (2023) 或该文章的中文介绍: 语言治理, 2023, 乔姆斯基的语言学研究路径是否被计算机科学颠覆了, “语言治理”, <https://mp.weixin.qq.com/s/sYeAMrQR7M900Y06ihT7tw> [2023 年 5 月 17 日访问]。

童习得语言时接受的刺激是贫乏的；而机器学习系统在语言上无能(缺少创造力)，并且没有因果解释与反事实推理(counterfactual reasoning)能力。乔姆斯基的这些评论和观点，引起了语言学、神经科学和自然语言处理等业界内外的讨论。比如，斯坦福大学的计算语言学教授 Christopher Manning、普林斯顿大学语言学教授 Adele Goldberg、DeepMind 公司的研究总监 Oriol Vinyals 等“大咖”都发表了自己的意见，在此不赘。

不过，值得一提的是，对于乔姆斯基的观点，加州大学伯克利分校的计算神经心理学家 Steven T. Piantadosi 写出了反驳文章(见脚注^①)。这篇长文通过详细对比基于梯度计算和记忆结构的现代语言模型跟以乔姆斯基生成语言学为代表的语言学研究路径，得出以下四个极具批判性和震撼力的结论：1) 现代语言模型的崛起和成功，实质上削弱了生成语法提出的先天语言能力说这种强烈的主张；2) 现代机器学习已经颠覆并绕过了乔姆斯基的整个理论框架，包括其对特定洞见、原则、结构和过程的核心主张；3) 现代语言模型实现了真正的语言理论，包括对句法结构和语义结构的表示；4) 生成语法方法在任何领域都不具有竞争力，并且可以说已经逃避了对其核心假设的实证测试。因为这篇论文内容丰富，信息量太大，我们有另外的文章专门介绍和评述。

4. 人工智能大型语言模型给语言学研究的启示

这一部分介绍 ChatGPT 研发的缘起和 GPT-4 所达到的心智理论水平，特别介绍心理语言学家对于 ChatGPT 的语言运用能力的测验，介绍他们对于大型语言模型和人类在语言预测方面具有相似性的有关实验，以此说明大型语言模型对于语言学研究的启示。

4.1 自然语言对于人工智能的重要性

发人深思的是，ChatGPT 研发的缘起对于认识语言交流与智能提升的关系也富有启发性。OpenAI 的研究团队之前发现，虽然 GPT-3 有强大的语言生成能力，但是不容易产出用户想要的东西。于是，他们的一个团队产生了一个想法：让人工智能系统通过反复试验来学习，以最大化奖励来完善模型。这个团队认为聊天机器人可能是这种方法的理想候选者，因为以人类对话的形式不断提供反馈，将使人工智能软件很容易知道它什么时候做得很好，以及什么地方需要改进^②。因此，这个团队在 2022 年初开始构建 ChatGPT^③。连他们都没有想到，这个聊天机器人程序居然如此成功和受人欢迎，被人称为“是一个现象

^① 这大概就是采用从人类的反馈中进行强化学习这种方法的最初动因。

^② Jeremy Kahn、刘进龙，2023，全球爆红的 ChatGPT 是如何诞生的，“财富 FORTUNE”，<https://mp.weixin.qq.com/s/zFom-hQvgTPRMj-BPRZRQ> [2023 年 5 月 17 日访问]。

级人工智能产品的横空出世”。这从工程技术层面上说明，语言交际功能的充分实现和巧妙利用反过来可以促使其思维功能的提升。毕竟，个体大脑中的思维水平的提高，依赖于社会群体智慧的滋养与激励；而这种沟通个体与个体、个体与群体的最方便的媒介，非人类自然语言莫属。因此，语言学家没有必要在语言的思维功能与交际功能之间强行做出切割^②。

更加发人深省的是，大语言模型似乎是在模型参数达到一定规模以后，一下子涌现出了语言能力，并且其具体的机制就连设计者也不甚了解。这启示我们，即使语言学至今也没有搞清楚人类是怎么说话的，工程师们还是可以训练语言模型产生适时应景的话语；并且，语言模型在学习语言的时候，居然顺便把作文、推理等本领也学会了。

4.2 语言模型的心智理论水平

GPT 在演化，其性能也在不断提升。据说，GPT-4 已经具有本来以为只有人类才有的心智理论(theory of mind, ToM)。斯坦福大学教授、计算心理学家 Michal Kosinski 使用 40 个测试人类心理状态的经典任务来测试几个语言模型，结果发现，一直到 2020 年发布的模型都几乎没有显示出能解决 ToM 任务的能力，但是 ChatGPT 有 9 岁儿童的心智理论水平，而 GPT-4 则能达到健康成年人的水平。

其实，心智理论不是指一种科学理论，而是指一种理解和推断他人心理状态的能力。Kosinski(2023)认为它是人类社会互动、交流、自我意识和道德的核心。那么，除了人，其他灵长类动物怎么样呢？Premack 和 Woodruff(1978)指出，人类只能观察同类的行为，不可能看到对方心里是怎么想的，但是却可以推断出对方的意图、知识、信念、思想、怀疑，甚至伪装等等。比如，你在同学的生日那天，悄悄地在他桌上放一个生日贺卡，就算你没有看到当时他的实际表现，你也能预见同学看到这个贺卡会非常地惊喜和感激。这是人类的本能。那么，黑猩猩能做到这一点吗？如果可以，说明黑猩猩也和人一样，有推测他人内心的能力。心理学家发现，黑猩猩能意识到他人的想法，确实具备一定水平的心智理论。这篇论文启发了许多学者思考：心智理论真的是人类的本能吗？人类又是如何获得心智理论的？于是，心理学家做了一系列的研究，结果发现，人类也不是一出生就具备心智理论的。婴幼儿先是能跟随其他人切换注意的目标，渐渐意识到其他人有自己的想法。然后，幼儿开始理解和推测别人的心理状态；接下来，这种推断能力会越来越准确，直到达到和成年人一

^② 比如，乔姆斯基就声称“语言的主要功用是思考而不是交际”，见史有为，2022，从乔氏对答谈语言的思维功能，“西去东来中传站”，<https://mp.weixin.qq.com/s/7mZ7VqreXHl20xNHHJGZBA> [2023 年 5 月 17 日访问]。

样,能不假思索地理解他人。实验证明,孩子要到四岁才能开始正确回答心智理论测试的一些问题,而且一开始成绩还很不稳定,要到九岁才能灵活推测其他人的心理状态^②。

对比一下,人类婴儿刚生下来也不会说话,要等到一岁以后才逐步学会说话。那么,语言能力和“心智理论”能力在“自然”与“培育”方面一样吗?

4.3 ChatGPT 的语言使用能力

更加引人入胜的是,Cai 等(2023)从心理语言学角度出发,将 ChatGPT 作为被试,让它完成 12 个经典的心理语言学实验,然后再和人类的实验结果对比,有以下发现:

在语音层面,1)能够根据单词的发音来判断单词所指示的物体的形状。比如,它能够像人一样,把 takete、kiki 一类新造单词判断为指尖的东西,把 maluma、bouba 一类新造单词判断为指圆的东西。这说明 ChatGPT 能够复原语音与形状之间的联系,符合语音象征主义(sound symbolism)。2)能够根据人的名字的发音来判断其性别。比如,对于 Although Pelcrad/Pelcra was sick...一句中的两个名字,它基本上能够分别用 he, him, his 和 she, her, hers 来回指——跟英语母语者一样,猜测男性名字则多以辅音结尾(如 Pelcrad),而女性名字更多以元音结尾(如 Pelcra)。

在词汇方面,3)不能像人类一样,根据预测性的高低来选择使用有关单词的较长形式或者较短形式;比如,在文本中首次提到时用 mathematics,在下文再次提到时用 math。4)对于多义词,像人类一样倾向于使用前面出现过的意义;比如,如果上文的语境已经确定多义词 post 的意思是“职位”,那么它倾向于把随后的句子中的 post 理解为“职位”,而不是“邮件”。

在句法方面,5)像人类一样倾向于重复使用前面出现过的句法结构;比如,上文用了双宾结构(如 give John a cake),下文倾向于也用同样的结构,而不用介词结构(如 give a cake to John)。6)但是,不能像人类一样,利用前面的语境消解句法歧义;比如,无法根据不同的询问方式把 the hunter killed the poacher with a rifle 中的介词短语正确地理解为修饰名词或者动词短语。

在语义方面,7)能够将不合理的句子合理化;比如,把 the mother gave the candle the daughter 这种不合理的双宾结构理解为在 the daughter 之前漏掉了介词 to。8)像人类一样产生语义错觉,注意不到句子中的明显错误;比如,During the biblical flood, how many animals of each kind did Moses take on the

^② 陈潮,2023,ChatGPT 相当于 9 岁人类小孩,它还真把自己当个人了,“科普中国”,<https://mp.weixin.qq.com/s/PvGplHBAHN1hY2VhNnE-7A> [2023 年 5 月 17 日访问]。

ark? 人们有时难以觉察到这里的 Moses(摩西), 其实应该是 Adam(亚当)。

在语篇方面, 9)能够判断出动词中隐含的因果关系; 比如, 给结果小句 Gary scared/feared Anna because...续写原因小句, 分别看它用代词 he(=Gary) 还是 she(=Anna)作主语。结果, 在用 feared 时, 它倾向于用 she(=Anna)作主语, 即把主语 Gary 害怕的原因归结为宾语 Anna。10)能够连接两个句子的信息从而进行推理(即连接推理); 比如, 它能够从“莎朗游泳时踩到一块玻璃”和“她绝望地呼救”这两个相继的句子, 推出“她划破了脚”这个可能性。跟人类一样, 它连接推理的能力远远强于联想推理(即只从一个句子上推出某种结果, 比如只从第一个句子上推出“划破脚”这种可能性)。

在对话方面, 11)能够根据不同的人(的身份)来理解单词的不同意思, 比如, 根据对话者自我介绍自己是英国人还是美国人, 把单词 bonnet 的意思处理为“汽车的阀盖”(英式英语), 或者“帽子”(美式英语)。12)能够对不同的人说不同的话, 比如, 同样是指“一套公寓房”, 在跟英国人对话时, 它选用单词 flat; 在跟美国人对话时, 则选用单词 apartment。

这 12 个实验, 涵盖了从语音、词汇、句法、语义再到语篇、对话等语言能力的多个方面; ChatGPT 在 10 个任务上跟人一样, 只有 2 个不同。这充分表明, “ChatGPT 的语言使用能力很大程度上和人类接近, 这有可能为研究人类的语言使用和习得提供借鉴”^⑤。

4.4 新的语言学实验范式

上文说过, 大型语言模型的核心能力是可以比较准确地预测下一个单词。因此, GPT-2 等模型自然地 被神经语言学家用来进行言语预期的研究。Schrimpf 等(2021)尝试将人类的行为数据、神经数据, 以及计算机的语言模型表现相整合, 探究三者之间的关系。他们希望不仅要确定哪个语言模型能够更好地拟合人类的实验数据, 而且还要探究不同模型的哪个具体维度和这种拟合有关。为此, 他们探究了 43 种不同语言模型的表现。这 43 种模型包括: 1)简单的嵌入模型, 如 GloVe; 2)更复杂些的循环神经网络, 如 LM1B; 3)不同的转换器模型或基于注意力的模型, 如 GPT。

他们通过下面三个指标来测量模型的表现: 1)大脑得分。他们利用以往三个神经研究(两个 fMRI 实验和一个 ECoG 实验^⑥)的数据集, 给模型呈现和神经实验中相同的文字, 记录模型的内部激活, 并计算出模型的内部激活能够多大

^⑤ 该研究的中文介绍见王书琪, 2023, ChatGPT 通过了 10 项心理语言学测试, “计算神经语言学”, <https://mp.weixin.qq.com/s/Hbobnt3ROt3jgblIMDQ5A> [2023 年 5 月 17 日访问]。

^⑥ ECoG 是 electrocorticography 的缩写, 指一种直接将电极放在大脑皮层上来记录神经活动的方法。

程度上预测人脑的神经反应。2)行为得分。探究模型能够多大程度上预测阅读时间。3)计算任务得分。探究模型完成计算任务(如词汇预测任务、阅读理解任务、语法判断任务等)的能力。结果发现,后词预测任务的表现能够预测大脑得分和行为得分,同时大脑得分也能够预测行为得分。这表明,这三个指标(即大脑得分、行为得分和计算任务得分)是高度关联的。

这项研究在方法论上的创新是,将神经网络模型与人脑神经数据、行为数据相整合,为探究人脑的计算机制提供了新的思路。由于模型的结构和权重很容易获得和操纵,因而通过模型来研究人脑要比直接研究人脑更简单。通过分析大脑得分高(即能够正确模拟人脑活动)的模型的特点,可以探究人脑的计算机制。比如,在该研究中,他们发现,跟人脑活动最匹配的模型正是那些被训练为预测下一个词的模型(如GPT)。因此,他们得出结论:“人脑语言系统也是为了预测性加工而进行了优化。”^{②7}

Goldstein等(2022)结合行为实验、ECoG和语言模型,探讨人脑和计算机处理语言时共享的算法与表示方面的计算原则。研究者让50个被试逐字阅读一个故事,然后让他们预测故事中每个词后面可能出现的词。研究者计算出每个词被正确预测的概率,即“可预测值”。通过比较实验中参与者与语言模型的预测值,发现两者高度相关。研究者研究了9名被试的ECoG信号。结果表明,语言模型能够预测的词,在被试听到词之前就已引起了更强的神经激活。这表明人脑和语言模型一样,都在不断地预测下一个词;而被语言模型错误预测的词,在参与者听到该词之后,也产生了更强的神经激活,且在400ms达到峰值(这可能类似于N400,一定程度上反映了语义违反)。这表明人脑和语言模型一样,都在将预测与实际进行对比。通过对比不同的语言模型与神经活动的关系,研究者发现,嵌入上下文信息(contextual embeddings)的模型(如GPT-2)比静态模型(如GloVe)能够更好地预测神经活动。他们得出的结论是:人脑在处理语言时和计算机模型一样,都在不断地预测下一个词,都将预测的词和真实遇到的词进行对比,都使用上下文嵌入来表示词汇^{②8}。

Heilbron等(2022)结合脑电图(electroencephalogram, EEG)和脑磁图(magnetoencephalography, MEG)的数据,考察大脑在理解语言时的神经反应。在两个实验中,研究者让被试听一段故事(有声书),并记录他们的神经活动。同时,研究者利用深度神经网络模型(GPT-2)计算每个单词的惊异(surprisal)程

^{②7} 该研究的中文介绍见王书琪,2023,如何用最近大火的ChatGPT的相关模型研究人脑的语言理解,“计算神经语言学”,<https://mp.weixin.qq.com/s/nw5-tEGHSIPaO1mwoMLwrw>[2023年5月17日访问]。

^{②8} 该研究的中文介绍见王书琪,2022,人脑和计算机用同样的方式加工语言信息,“计算神经语言学”,<https://mp.weixin.qq.com/s/msBdn2xxapRvSOEasCOjg>[2023年5月17日访问]。

度/值^②，并使用回归模型将二者结合，探究语言理解过程中的预测活动是如何影响大脑的神经反应的。结果表明，大脑的预测是持续不断地产生的，神经网络语言模型的预测和人脑的预测具有相似性；如果预测结果不符合实际情况（猜错了），大脑就会感到“惊异”，网络中的神经元变得空前活跃。也就是说，大脑能够根据上下文语境对下一个单词进行预测。那么，这种预测具体发生在语言的哪些层面上呢？或者说，大脑能够预测单词哪些层面的信息？是句法层面、语义层面、音位层面，还是同时包括三个层面？为了探究以上问题，研究者将单词的句法、语义和音位信息进行分离。句法信息通过词类表示，语义信息通过语义嵌入表示。然后，通过 fMRI 还原了大脑在进行不同层次预测时的“可视化图”。结果显示，大脑在用词类、语义和音素进行预测时，不同层次之间会相互影响。并且，高层的预测条件（词汇语义）会影响低层的预测条件（词内的音素）。例如，在 They say his father was a f _____. 这句话中，仅仅凭借孤立的单词内音素的出现概率预测下一个音素跟通过上下文进行预测的结果是不一样的。前一种方法大概率会认为以/f/开头的单词的第二个音素是/o/（如 false, for）；但结合上下文，根据有可能出现在这个句子的这个位置的单词来预测/f/之后的音素，就更倾向于它是/l/, 也就是以/fl/开头的单词（如 figure, fisherman）。可见，人脑和 GPT-2 语言模型一样，都是自发地根据上下文对下一个词进行预测；并且，这种预测是层级性的，即在多个层面上进行的^③。

上面这三个工作，巧妙地把语言运用的人类行为数据与神经数据跟计算机的神经网络大型语言模型结合起来，为我们指明了人工智能时代语言学研究的一个崭新方向，确立了一种用人工智能来研究人类语言及其信息加工过程的实验范式。

References [引用文献]

- Brown, Tom B., Benjamin Mann, Nick Ryder, et al. 2020. Language models are few-shot learners. <https://doi.org/10.48550/arXiv.2005.14165> [accessed 17, May 2023]
- Cai, Zhenguang G. (蔡振光), David A. Haslett, Xufeng Duan, Shuqi Wang, and Martin J. Pickering. 2023. Does ChatGPT resemble humans in language use? <https://doi.org/10.48550/arXiv.2303.08014> [accessed 17, May 2023]
- Goldstein, Ariel, Zaid Zada, Eliav Buchnik, et al. 2022. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience* 25, 3:369–380.

^② 惊异值反映了语言的不可预测性，惊异值越高，不可预测性越高。

^③ 该研究的中文介绍见王书琪，2022，忍不住去猜？原来大脑在疯狂脑补TA说的话，“计算神经语言学”，<https://mp.weixin.qq.com/s/-aze3DLFYIU995D2qOPa8A> [2023年5月17日访问]。

- Heilbron, Micha, Kristijan Armeni, Jan-Mathijs Schoffelen, Peter Hagoort, and Floris P. de Lange. 2022. A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences* 119,32:e2201968119.
- Kosinski, Michal. 2023. Theory of mind may have spontaneously emerged in large language models. <https://doi.org/10.48550/arXiv.2302.02083> [accessed 17, May 2023]
- Piantadosi, Steven T. 2023. Modern language models refute Chomsky's approach to language. <https://lingbuzz.net/lingbuzz/007180> [accessed 17, May 2023]
- Premack, David and Guy Woodruff. 1978. Does the Chimpanzee have a theory of mind? *Behavioral and Brain Sciences* 1,4:515–526.
- Schrimpf, Martin, Idan Asher Blank, Greta Tuckute, et al. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences* 118,45:e2105646118.
- Wang, Lulu (王璐璐) and Yulin Yuan (袁毓林). 2023. *Deep Computing Theory and Technical Application of Language*. Beijing: Foreign Language Teaching and Research Press. [2023, 《语言的深度计算理论与技术应用》。北京: 外语教学与研究出版社。]
- Yuan, Yulin (袁毓林). 2022. On the constraint of the human habitat on the design principles and the operating mechanism of language. *Chinese Journal of Language Policy and Planning* 6:85–96. [2022, 在人类生境约束下思考语言的设计原理和运作机制。《语言战略研究》第6期, 85–96页。]
- Zhang, Huaping (张华平), Linhan Li (李林翰), and Chunjin Li (李春锦). 2023. ChatGPT performance evaluation on Chinese language and risk measures. *Data Analysis and Knowledge Discovery* 3:16–25. [2023, ChatGPT 中文性能测评与风险应对。《数据分析与知识发现》第3期, 16–25页。]

作者简介

袁毓林, 男, 博士, 澳门大学人文学院中国语言文学系讲座教授。主要研究兴趣为理论语言学 and 汉语语言学, 特别是句法学、语义学、语用学、计算语言学和中文信息处理。代表作: 《语言的认知研究和计算分析》和《汉语语法研究的认知视野》。电子邮件: yuanyl@pku.edu.cn

YUAN Yulin, male, PhD, is a chair professor at the Department of Chinese Language and Literature, Faculty of Arts and Humanities, University of Macau. His main research interests include theoretical linguistics and Chinese linguistics, especially syntax, semantics, pragmatics, computational linguistics and Chinese information processing. His major publications are *Cognitive Research and Computational Analysis of Language* and *The Cognitive Perspective of Chinese Grammar Study*. E-mail: yuanyl@pku.edu.cn