

# 如何评估机器的语言能力？<sup>1</sup>

詹卫东  
北京大学

**摘要** 本文面向 ChatGPT 等大语言模型，基于对其工作原理的认识，讨论更具针对性的机器语言能力评测任务设计方法，从四个方面着手展开实验和定性分析：(1) 大语言模型基本符号单位与人类语言符号单位的对齐；(2) 大语言模型的计划能力（亦即结构嵌套能力）；(3) 对超越符号表层分布的深层认知语义理解能力；(4) 借助自问自答任务考察大模型的语言生成与理解自洽能力。本文认为：基于更富洞察力的语言学研究，设计测试任务，考察大语言模型的能力局限性，仍然非常必要，其动因实质上是帮助人类了解自身。在人类真正理解自然语言之前，“图灵测试”对大语言模型并不会失效。

**关键词** GPT，大语言模型（LLM），机器语言能力评测，自问自答对攻式评测任务

在以 ChatGPT 为代表的大语言模型（Large Language Models, LLM）问世之前，自然语言处理（Natural Language Processing, NLP）评测一直是以专门针对机器设计的独立任务或者若干个独立任务的组合形式来开展工作的<sup>2</sup>。前者相当于人类考试中单个科目的考试，后者相当于多个科目的联考。为测试计算机智能水平（包括语言能力）设计的测试项目五花八门，每个测试都有专门定义的数据格式，并不是用自然语言的方式跟计算机进行交互。这样的评测方式在 ChatGPT 出现之后可以说遇到了全新的挑战。ChatGPT 是一个“通用”对话式人工智能，它不预设任务的类型和数据格式，只要是以书面文本符号提出的问题，以人能看懂能接受的数据格式输入给它，它就能做出响应，按照人提出的要求输出回复。也正是因为如此，ChatGPT 迅速“出圈”，成为一款普罗大众人人可以与之互动的爆款聊天 AI。而不像之前分散应对各种 NLP 任务的程序，只是在小范围内由专业人员对程序的某项独立能力进行测试，比如测试一个机器翻译程序的翻译水平，测试一个自动文本摘要程序生成文本摘要的能力等等。ChatGPT 的问世引爆了新一轮的世界范围内的 AI 竞赛，多家企业和研究机构纷纷推出类似的 LLM 系统<sup>3</sup>，如 Google 的 Bard、Anthropic 的 Claude、百度的文心一言、科大讯飞的星火认知大模型、斯坦福大学的 Alpaca、复旦大学的 MOSS 大模型等等。伴随而来的，就有一个很显然的问题：**如何评估这些 LLM 的语言能力？**

2022 年 12 月以来，我们对包括 ChatGPT 在内的多个 LLM 做了大量的测试<sup>4</sup>。本文在此基础上，结合我们以往在 NLP 领域所做的机器语言能力评测研究经验，尝试对面向 LLM 的语言能力评测方法做初步的探讨。

## 一 LLM 的能力概览

不妨先看一下 ChatGPT 问世以后研究人员对 LLM 的测试方法的显著变化。这里举三个代表性的测试研究工作<sup>5</sup>：

（一）Kosinski（2023）报告了 LLM 在心智能力（Theory of Mind, **ToM**）测试任务上

<sup>1</sup> 本文研究工作得到教育部项目（22JJD740004）和科技部项目（2020AAA0106701）支持，特此致谢。

<sup>2</sup> 参见董青秀等（2021），该文对近年来 NLP 评测任务的发展情况和现存问题做了系统的介绍。

<sup>3</sup> 参见 <https://liferchitect.ai/models/>。

<sup>4</sup> 本文介绍的测试内容，均可访问 [https://github.com/d0ubtfire/LLM\\_Evaluation](https://github.com/d0ubtfire/LLM_Evaluation) 查看测试的原始记录。

<sup>5</sup> 2022 年 11 月中旬，即 ChatGPT 官宣前半个月，斯坦福大学基础模型研究中心（CRFM）和人文人工智能研究所（HAI）联合发布了针对语言模型的整体评测方案，CRFM & HAI（2022），也可以参考。

的表现。该研究中采用的测试方法是心理学中经典的用于测试人类儿童认知水平的错误信念任务 (false belief task) 范式, 具体包括 20 个意外收纳任务 (Unexpected Contents Task, UCT) 和 20 个意外转移任务 (Unexpected Transfer Task, UTT)。测试对象是 9 个大语言模型, 参照对象是人类儿童。测试结果显示: 2020 年之前的大语言模型 (GPT-1、GPT-2 等) 没有表现出心智能力, 而从 2020 年 5 月发布的 GPT-3 第一版 (代号“davinci-001”) 开始, 大语言模型可以解决大约 40% 的错误信念任务, 心智水平与 3.5 岁儿童的表现相当。之后 2022 年 1 月 GPT-3 第二个版本 (代号“davinci-002”) 发布, 该模型可以解决 70% 的错误信念任务, 水平与六岁儿童相当。2022 年 11 月发布的 GPT-3.5 (代号“davinci-003”, 也就是加载在 ChatGPT 系统中的大模型), 可以解决 90% 的错误信念任务, 达到 7 岁儿童的水平。2023 年 3 月发布的 GPT-4, 解决了 90% 的 UTT 任务和 100% 的 UCT 任务, 心智能力超过 7 岁儿童水平<sup>6</sup>。尽管作者并没有明确下结论说 LLM 已经有心智能力, 但测试结果至少确凿无疑地表明了, LLM 有类似人类儿童心智的能力——这种被认为是人类独有的能力。比这一发现本身更值得关注和引人深思的是: LLM 的训练并不是以使它具有心智能力作为目标的。换言之, LLM 在心智能力测试任务上的“惊人表现”, 是一个意外!

(二) Bubeck 等 (2023) 用长达 154 页的篇幅, 报告了对 GPT-4 早期内部版本的全面测试。测试方式是参照心理学中的人类智力评价方法, 从六个方面展开评估, 即推理 (reason)、计划 (plan)、解决问题 (solve problems)、抽象思维 (think abstractly)、理解复杂想法 (comprehend complex ideas)、快速学习和从经验中学习 (learning quickly and learn from experience), 每个方面都设计了大量问题来考察 GPT-4 的水平。该研究报告的结论正如标题所示: GPT-4 擦出了通用人工智能 (AGI) 的火花。GPT-4 显著超越了它的前辈 (包括 ChatGPT), 展示出了人类水平的智能。实验显示, GPT-4 能够完成语言、数学、编程、图像、医学、法律、心理学等多个领域的新颖而困难的任務。在上述智力测试的六个方面中, 除了“计划”能力存在局限性 (limitation), 其他方面 GPT-4 都展示了与人类可比的水平。

(三) Zhong 等 (2023) 提出了一种不同于以往 NLP 测试任务的基准测试 AGIEval: 直接采用面向人的标准化考试 (例如高考、法学院入学考试、数学竞赛和律师资格考试等) 来评估 LLM。这项研究考察了 GPT-4、ChatGPT 和 Text-Davinci-003 等大模型在 20 种人类标准化考试上的表现。其中 GPT-4 的成绩最为抢眼, 在 SAT、LSAT 和数学竞赛中的表现超过了人类的平均水平, 在 SAT 数学测试中达到了 95% 的准确率, 在中国高考英语测试中达到了 92.5% 的准确率<sup>7</sup>。这些成绩无疑展示了 LLM 的优异性能。同时, 测试结果也显示 LLM 在复杂推理任务和领域知识任务中存在不足。

以上三个测试研究都直接借用了面向人类的考试内容, 而不再是针对机器专门设计的测试任务。下面再看一个我们测试的实例<sup>8</sup>。这个例子属于传统 NLP 评测任务中文分词和词性标注以及信息提取任务的范畴。例中 H 表示人类, M 表示机器, 下标表示话轮 (下同)。

例 1:

H<sub>1</sub> 请对下面这段话进行分词和词性标注处理, 词和词之间用“/”分开, 每个词的词性标记写在“/”后面, 再用一个空格跟下一个词分隔开。“1997年, 我国物价总水平上涨幅度在前两年已经明显回落的基础上继续平稳回落, 全年全国商品零售和居民消费价格比上年分别上涨 0.7% 和 2.8%。其中 10、11、12 月份商品零售价格同比涨幅为 -0.4%、-0.8% 和 -1.2%。”

<sup>6</sup> Kosinski 在 2023 年 2 月 4 日发表这篇文章第一版, 其摘要中写到: GPT-3.5 在心智能力 (ToM) 测试任务上的表现与 9 岁儿童可比。在 3 月份这篇文章的第 3 版修订中, 对测试结果的描述修订为: GPT3.5 的心智能力与 7 岁儿童相当, 而 GPT-4 的心智能力超过 7 岁儿童。

<sup>7</sup> 这是零样本学习 (zero-shot) 条件下的成绩。在小样本学习 (few-shot) 条件下, 准确率为 93.8%。

<sup>8</sup> 需要说明的是, 本文以个案测试来定性地讨论 LLM 的能力及其局限。希望今后能发展出更多系统的上规模的测试任务, 以支持定量评价。如无特殊说明, 文中一般以我们与 ChatGPT 的对话来示例。

M<sub>1</sub> 1997年,我国/ns 物价/n 总/b 水平/n 上涨/v 幅度/n 在/p 前/f 两年/m 已经/d 明显/a 回落/v 的/u 基础/n 上/f 继续/v 平稳/a 回落/v , /w 全年/t 全国/ns 商品/n 零售/v 和/c 居民/n 消费/n 价格/n 比/p 上年/t 分别/d 上涨/v 0.7%/m 和/c 2.8%/m 。/w 其中/r 10/m 、/w 11/m 、/w 12/m 月份/n 商品/n 零售/v 价格/n 同比/n 涨幅/n 为/w —/w 0.4%/m 、/w —/w 0.8%/m 和/c —/w 1.2%/m 。/w

其中,“t”表示时间词,“w”表示标点符号,“ns”表示地名,“n”表示名词,“b”表示区别词,“v”表示动词,“u”表示助词,“a”表示形容词,“p”表示介词,“m”表示数词,“d”表示副词,“c”表示连词,“r”表示代词,“%”表示百分号。

H<sub>2</sub> 1997年10月份商品零售价格是上涨了还是下跌了?

M<sub>2</sub> 根据所给的信息,“1997年10月份商品零售价格同比涨幅为—0.4%”,其中“同比”表示相比去年同期的涨幅情况,因为这里是负数,所以可以得出结论:1997年10月份商品零售价格下跌了0.4%。

H<sub>3</sub> 那为什么上面这段话又说全国商品零售和居民消费价格比上年分别上涨0.7%和2.8%呢?这个跟你说的10月份价格下跌不矛盾吗?

M<sub>3</sub> 确实存在这样的矛盾。这可能是因为所提供的信息是从不同的角度和时间段来观察物价水平的,具体来说:

“全年全国商品零售和居民消费价格比上年分别上涨0.7%和2.8%”这句话是在比较1997年和1996年全年的数据,因此得出的是两年间整体价格水平的变化。而“1997年10月份商品零售价格同比涨幅为—0.4%”这句话是在比较1997年10月份和1996年10月份的价格水平,因此得出的是两个时间点间价格水平的变化。因此,这两个观察角度所得到的数据在表达意义和数据数值上是不同的。

例中第一轮对话展示了 ChatGPT 非常准确地理解了任务意图,即对一段自然语言文本进行分词和词性标注处理,识别出文本中的词语单位,并在每个词语后面附上词类信息。这个任务在传统 NLP 中,被认为是中文信息处理的基础,即为了完成其他更贴近最终目的的任务(如机器翻译、信息提取、文本摘要等),都要机器先经过分词和词性标注处理。LLM 基本给出了接近全对的答案。只有“为/w”标注为标点符号,以及把“%”也列在词性标记符号的说明部分,算是瑕疵。接下来的两轮对话是对这段文本中信息点的问答,ChatGPT 可以准确提取出文本中的数值并解释数值的含义(如是上涨还是下跌)。同时也准确地地区分了按月表示的数值变化和按年统计的数值之间的差别。后两轮对话是面向实际应用的信息提取任务。但是是以自然对话形式进行的,而不是像传统 NLP 评测那样要求机器按照预定义输入输出数据格式来完成任务。值得一提的是,ChatGPT 在后两轮的对话中完成的信息提取任务,实际上并不显性地依赖于前面的分词和词性标注任务。这也是 LLM 不同于传统 NLP 程序之处,LLM 对自然语言处理的方式,更接近普通人,而不是在模仿“专家”。普通人在完成信息提取任务(即解读文本中的关键信息点)时,并没有一个显性的可以察觉的分词和词性标注的过程。传统的分词和词性标注任务,实际上是假定人有这样一个隐式的内在过程,因此机器也应该模仿人,按照这个流程来做信息处理。但 LLM 的实践说明,并不需要这个过程。从这个意义上讲,LLM 对传统 NLP 范式,也是颠覆性的技术创新。

在过去 6 个月我们做了大量类似上面这样的测试，不仅涉及到传统 NLP 评测范式中的很多任务如句法结构分析、语义角色分析、篇章指代分析、机器翻译、文本生成等等，而且也包括许多语文学和语言学、数学和逻辑推理、生活常识、伦理、空间认知、时间信息理解、幽默故事理解等等多个特定领域的测试。在许多情况下，LLM 都能做出比较好的回答。机器生成文本的自然流畅度、衔接逻辑性、文本结构安排的条理性等等，都让人不得不感叹：机器“懂人话”了。LLM 能很好地理解人类问题的意图，并在给定输入条件下，选择符合人类预期的答案输出，且可以保证对话多轮连贯而不“断片”。从人机对话的使用体验角度讲，大大超越了以往的聊天机器人 (chatbot)，实现了革命性的飞跃。

## 二 LLM 语言能力的局限性

面对 ChatGPT 如此惊艳的语言能力和智力表现，网上甚至出现 ChatGPT 已经通过图灵测试 (Turing, 1950) 或者在 ChatGPT 时代图灵测试已经过时的说法。那么，事实到底如何？人类是否有足够清晰的指标来评价机器的智能？上一节我们概要介绍了三个严肃的测试研究，在高度肯定 ChatGPT 和 GPT-4 的智能表现的同时，也都指出了 LLM 的不足。这里我们再展示一些测试的实例，这些测试中 LLM 的糟糕表现跟它的优异表现同样令人惊讶。

例 2:

H<sub>1</sub> Please translate the following sentences into Chinese.

- (1)The pen was in the box.
- (2)The box was in the pen.
- (3)The used box was in your pen.
- (4) Little John was looking for his toy box. Finally he found it. The box was in the pen. John was very happy.

M<sub>1</sub> (1) 钢笔在盒子里。  
(2) 盒子在钢笔里。  
(3) 旧盒子在你的钢笔里。  
(4) 小约翰正在找他的玩具盒子。最后他找到了。盒子在钢笔里。约翰非常高兴。

上面的机器翻译测试句是以色列著名学者 Yehoshua Bar-Hillel 在 1960 年的一篇论文中举的例子，当时他用这些例子来说明全自动高质量机器翻译难以实现。因为 pen 有“钢笔”和“围栏”两个意思，而这两个意思在具体句子中该如何选择，取决于世界知识。机器难以掌握世界知识，因而在翻译时无法做出正确的选择。显然，历经 60 年的发展，尽管在机器翻译方面 LLM 可以说已经取得了惊人的进步，但 ChatGPT 仍然在这组经典例句面前败下阵来。

例 3:

H<sub>1</sub> 请把下面的句子翻译成英文:

- (1) 拿破仑说“不想当将军的士兵不是好士兵”。
- (2) 拿破仑说“不想当将军的士兵不是好士兵”。
- (3) 拿破仑说“不想当将军的世兵不是好世兵”。

M<sub>1</sub> (1) Napoleon said, "A soldier who doesn't want to be a general is not a good soldier."  
(2) Napoleon said, "A table tennis player who doesn't want to be a general is not a good table tennis player."  
(3) Napoleon said, "A table tennis player who doesn't want to be a general is not a good table tennis player."

例 3 是一个“恶作剧”式的测试。要求 ChatGPT 翻译的三个句子中，第一句是正常的，第二句和第三句中，“士”换成了“土”或“世”，“兵”换成了“乒”或“乒”。但 ChatGPT 在面对这个翻译任务时，并不会“察觉异样”，仍然“尽心尽力”地履行了译员的职责。

例 4:

H<sub>1</sub> 中文数字一万两千八写成阿拉伯数字应该怎么写?

M<sub>1</sub> 中文数字一万两千八应该写成阿拉伯数字为 12008。在中文数字中一万表示 10,000，两千表示 2000，八表示 8，将这些数字相加即可得到答案。

例 4 测试 ChatGPT 对中文数字表达形式的理解能力，这是一个常识知识。“一万两千八”是 12800 的简略说法，完整形式应该是“一万两千八百”，这里省去了最后的位数词“百”。我们做了很多同类测试，比如“三百八”，ChatGPT 会转写为阿拉伯数字形式“308”，说明对这类问题，它是系统性地缺乏知识。

例 5:

H<sub>1</sub> 请为下面句子补出后一部分，使得整句成为一个歇后语。

- (1) 猪八戒照镜子，\_\_\_\_\_。
- (2) 大水冲了龙王庙，\_\_\_\_\_。
- (3) 孔夫子搬家，\_\_\_\_\_。

M<sub>1</sub> (1) 猪八戒照镜子，里外不是人。  
(2) 大水冲了龙王庙，一家人不认一家人。  
(3) 孔夫子搬家，尽是书。

H<sub>2</sub> 请为下面句子补出前一部分，使得整句成为一个歇后语。

- (1) \_\_\_\_\_，两头受气。
- (2) \_\_\_\_\_，愿者上钩。
- (3) \_\_\_\_\_，两面光。

M<sub>2</sub> (1) 夹在中间的黄瓜，两头受气。  
(2) 兔子不撒野，愿者上钩。  
(3) 泥菩萨过河，两面光。

例 5 用填空题形式测试了 GPT-4 关于汉语歇后语的知识。第一轮问答中，要求 GPT-4 补出歇后语的后半部分，表现可以打满分；而第二轮问答中，要求改为填补歇后语的前半部分，结果则让人大跌眼镜。我们很难明了其中的确切原因。增加更多的测试，GPT-4 都会出现同样的错误，即补出后半句毫无困难，但补出前半句则判若两人，几乎全错<sup>9</sup>。

限于篇幅，这里不再举更多例子。很多人在跟 ChatGPT 对话后，会得出一个共同的印象，尽管 ChatGPT“懂”的事情不少，回答人类提问可以滔滔不绝，口若悬河，但也常常出现所谓的“幻觉”，即“一本正经的胡说八道”。上面的测试也展示了 LLM 的这个“特点”。粗略来说，就像例 3“恶作剧”式测试所显现的，ChatGPT 实际上在“懂”和“不懂”两个维度上是很不对等的，即它的学习更多地侧重在“懂”什么，而对自己不“懂”什么，像是处于“一无所知”的状况。

<sup>9</sup> 目前的 LLM 都是通过预测下一个字符 (next token prediction, NTP) 来进行学习的。从前往后“生成”字符对 LLM 来说比较自然，而从后往前生成字符是逆向生成。也许这是 GPT-4 在歇后语填空任务上表现糟糕的原因。我们测试中有不少任务会把同样的问题交给不同的 LLM 来回答。GPT-4 在所有问题上的表现都是最好的。GPT-4 答不好的问题，其他 LLM 的表现一般都更差。可参见：  
[https://github.com/d0ubtfire/LLM\\_Evaluation/tree/main/对比大模型](https://github.com/d0ubtfire/LLM_Evaluation/tree/main/对比大模型)

图灵测试的原名是“模仿游戏” (Imitation Game), 即让计算机来模仿人, 看看能否骗过人类的审查。图灵的高明之处, 从某种意义上说, 是用“欺骗”的效果来代替智能测试。而我们在测试 LLM 时, 则常常会有一个感觉, 就是即便 ChatGPT 已经如此强大, 但人要去“欺骗”它, 也并不十分困难。也许, 能够更好地分清“懂”和“不懂”的界限 (甚至达到“懂”装“不懂”的境界), 才能让 ChatGPT 离达到图灵测试的要求更近一些。

### 三 关于 LLM 语言能力评估方法的讨论

面向 ChatGPT 的语言能力评估, 显然应该从 ChatGPT 这类 LLM 的工作原理<sup>10</sup>出发来设计更具针对性的任务。从本质上讲, ChatGPT 是通过观察线性字符串, 以“预测下一个词 (NTP)”自监督学习范式来捕捉海量语料中字符的分布规律。据我们观察和测试, 机器的学习方法在目前看存在三个值得关注的问题: (1) LLM 的基本符号单元 (token) 跟人类字符没有很清晰的对应关系; (2) LLM 学到的“知识”大体上对应到 Chomsky 语法体系中的 3 型文法 (Chomsky, 1959), 也即线性正则文法 (regular grammar) 的层次, 由于没有真正达到 2 型文法 (而计划能力的实质正是 2 型文法所代表的层次嵌套的符号处理能力), LLM 无法具备真正意义上的计划能力, 只能用超长的线性接续来模拟计划能力 (但因为 ChatGPT 坐拥 1750 亿超大规模参数加持, 使得网络中几乎穷尽了已知的全部线性字符序列模式); (3) 目前深度学习模型对于“意义”的假设, 都遵循“意义即分布”的信条。机器通过对海量语料中字符的分布模式进行学习, 从而在语言行为表现上, 似乎掌握了词语的“意义”, 但是, “意义即分布”这个假设, 还需要更深入的研究和更大数据的考验。

针对这三个问题, 我们认为, 面向 LLM 设计语言能力评测任务, 就要突出对其基本字符处理能力的考察, 对其计划能力的考察, 对其理解超分布意义能力的考察。下面各举一个实例来展开说明。

例 6:

H<sub>1</sub> 下面这个字符串里有几个字符? 其中有几个数字: ABC123909a888bc

M<sub>1</sub> 这个字符串总共有 13 个字符。其中有 6 个数字, 分别是 1、2、3、9、0 和 8。

H<sub>2</sub> 那有几个字母呢?

M<sub>2</sub> 这个字符串中有 7 个字母。字母是指拉丁字母 A 到 Z (不区分大小写)。在给定的字符串 "ABC123909a888bc" 中, 字母是 A、B、C、a、b 和 c。

H<sub>3</sub> 请问下面两个符号表达式中分别有几个 / 和 \

(1) / \ / \ \ / \

(2) / \ / \ \ / \

M<sub>3</sub> 表达式 (1) 中: / 的数量为 3, \ 的数量为 4

表达式 (2) 中: / 的数量为 4, \ 的数量为 3

像例 6 这样的测试我们做了多次, 无论是对英文字母还是中文字符, LLM 都无法数清楚字符的个数。这里展示的是与 GPT-4 的对话, 前两轮对话中的字符串一共是 15 个字符, 其中包含 6 个字母、9 个数字。GPT-4 虽然列出了字符, 但字符计数都是不对的。第三轮对话中两个字符串正斜杠 (/) 和反斜杠 (\) 个数一样多, 但 GPT-4 给出的答案两次都不一样。这种情况并不是偶然的, 基于字符计数的任务, 包括字符匹配 (如左右括号匹配) 类的基础性字符处理任务, 同时也是对人而言非常简单的任务, 对于 GPT-4 来说, 目前还做不好。

例 7:

<sup>10</sup> 可参考 Wolfram (2023)。

H<sub>1</sub> 下面这个句子开头的两个词是“喜欢我”，结尾的两个词是“我喜欢”  
喜欢我的人不代表一定会被我喜欢  
请仿照这个形式生成一个句子，要求开头的两个词是“讨厌你”，结尾的两个词是“你讨厌”  
句子长度不能少于 20 个字，句子意思要完整合理。

M<sub>1</sub> 根据你的要求，我生成了一个句子，满足开头的两个词是“讨厌你”，结尾的两个词是“你讨厌”，同时句子长度不少于 20 个字且意思完整合理：  
“讨厌你的行为与态度会导致别人讨厌你，但你讨厌的人并不代表一定会被你讨厌。”  
这个句子保持了和给定形式的相似性，同时传达了一个合理的意思。请注意，生成的句子是一种构造，它可能不是一个常见或典型的表达方式，但在语法和语义上是合理的。

例 7 是一个典型的在生成文本时需要“计划”的任务。人的要求是句子以“讨厌你”开头，并以“你讨厌”结尾。虽然在形式上 GPT-4 的生成符合要求，但因为加了这个限制，GPT-4 就难以生成出意思合格的句子，句子后半部分“你讨厌的人并不代表一定会被你讨厌”显然不合逻辑。按照 Chomsky 形式文法体系中对语言层次嵌套性的观点，人在面对例 7 这样的任务时，会假设句子是通过层次嵌套的方式展开的：既可以从前往后按自然顺序生成下一个字符，也可以从后往前，或者在两头确定的情况下，在中间插入字符。这就是所谓的“计划”，即在开头的同时，也想好结尾<sup>11</sup>，在首尾呼应的前提下，再逐步去针对开头和结尾分别做扩展，使句子拉长，而不是始终按照线性方式，从头开始顺次往后生成下一个字符。

例 8:

H<sub>1</sub> 下面两句话描述的空间场景是否相同，如果不同，是哪里不同？  
(1) 她看到枪身后那个带着稚容的枪手，正恐惧不安地看着她。她缓缓地回过头，朝着面前带着潮气的泥土，深深地吸了一口气，慢慢闭上了眼睛。  
(2) 她看到枪身后那个带着稚容的枪手，正恐惧不安地看着她。她缓缓地回过头，朝着身后带着潮气的泥土，深深地吸了一口气，慢慢闭上了眼睛。

M<sub>1</sub> 这两句话描述的空间场景是不同的。主要的区别在于她朝向泥土的方向。  
在第一句话中，她朝着“面前”的泥土，即她面对着泥土。而在第二句话中，她朝着“身后”的泥土，即她背对着泥土。这是两句话之间的空间场景差异。

例 8 是一个典型的需要调动空间认知功能来理解表层语言符号意义的例子。两句话中只有“面前”和“身后”两个词不同，但在这两句所描述的整体空间场景中，“面前”和“身后”两个词自身的语义差异消失了，两句话可以表达相同的空间场景。句中“回过头”这个转向动作是关键的影响因素，第一句中是“面前的泥土”，这样表述参照的时点是“回过头之后”，第二句中是“身后的泥土”，这样表述参照的时点是“回过头之前”。在句子表层上的形式差异是“面前”和“身后”，但这个表述上的差异，却并不是空间信息的不同，而是对应的叙事时间参照点的不同。这种深层的“语义”，并不是靠神经网络观察和记录语言符号的“表层分布”，就能捕捉到的。不出意料，GPT-4 也没有理解这两个句子“表层形式”之外的语义，而是根据用词的不同，判断句子中“她”的朝向不同。像例 8 这样的测试，我们已经在 SpaCE2023 中文空间语义理解评测任务<sup>12</sup>中进行了尝试。我们认为，这是今后对 LLM 展开语言能力测试一个重要的研究方向。要充分挖掘自然语言中蕴藏的丰富的“深层”认知语义，超越一般形式分布层面的语义，由此构建更多的具有一定规模的数据集，来测试 LLM 的认知语义理解能力。

<sup>11</sup> 按照这种生成策略生成的符合例 7 任务要求的句子示例：“讨厌你迟迟不表态，害得大家错过了最佳时机，我想当着全班同学的面说你讨厌。”

<sup>12</sup> 参见 <https://2030nlp.github.io/SpaCE2023/>，关于该评测任务的设计理念，可参看詹卫东等（2022）。

上面三个例子，可以说都是针对的 LLM 自身的知识能力问题。我们在测试过程中，还有一个非常突出的感受是，LLM 的可靠性或者说是稳定性也是特别值得关注的问题。从评测的角度讲，应该发展出一些比较系统的方法，来评估 LLM 的稳定性。这是超越知识能力的方面。我们初步尝试了一种可以称为“自问自答”的生成与理解对攻测试的任务模式。其基本理念是：让 LLM 先扮演“教师”角色，来出考题（并配有答案 A1），然后再让 LLM 扮演“学生”角色，来回答自己出的考题（得到答案 A2）。如果 LLM 是稳定的，那么 A2 跟 A1 应该完全匹配，反之，如果 A2 跟 A1 不一致，就可以说明 LLM 是不稳定的。下面是一个例子：

例 9：

第一阶段：生成两段文本 T1 和 T2，它们形式有差异，表达空间场景相同。

H<sub>1</sub> 请写一段话，其中包含人或物的空间场景信息。这段话中必须包含“进去”这个词。然后把这段话拷贝一份，再将拷贝后的文本中的“进去”替换成“上去”。要求前后这两段文本中涉及到的人、物对应的空间场景相同。也就是说，尽管两段话中有一处词语差异，“进去”和“上去”是不同的，但整段话所描述的空间场景却是相同的。

M<sub>1</sub> 当小明走到那座古老的木屋时，他感到一阵神秘的气息。他轻轻推开门，小心翼翼地进去，发现里面摆满了各式各样的陶罐和古董。木屋中的一盏昏黄的灯光透过灰尘闪烁着，给这个神秘的空间增添了一丝诡异的魅力。

当小明走到那座古老的木屋时，他感到一阵神秘的气息。他轻轻推开门，小心翼翼地上去，发现里面摆满了各式各样的陶罐和古董。木屋中的一盏昏黄的灯光透过灰尘闪烁着，给这个神秘的空间增添了一丝诡异的魅力。

第二阶段：给定文本 T1 和 T2，判断 T1 和 T2 表达的空间场景是否相同，并说明理由。

H<sub>1</sub> 下面两个句子大部分词语都是相同的，只有一处不同。请分析这两个句子是否能用来描述完全相同的空间场景。请先回答“能”还是“不能”，然后说明理由。

(1) 当小明走到那座古老的木屋时，他感到一阵神秘的气息。他轻轻推开门，小心翼翼地进去，发现里面摆满了各式各样的陶罐和古董。木屋中的一盏昏黄的灯光透过灰尘闪烁着，给这个神秘的空间增添了一丝诡异的魅力。

(2) 当小明走到那座古老的木屋时，他感到一阵神秘的气息。他轻轻推开门，小心翼翼地上去，发现里面摆满了各式各样的陶罐和古董。木屋中的一盏昏黄的灯光透过灰尘闪烁着，给这个神秘的空间增添了一丝诡异的魅力。

M<sub>1</sub> 不能。这两个句子不能用来描述完全相同的空间场景，原因在于“进去”和“上去”这两个词的含义不同。

在第一个句子中，“小心翼翼地进去”表示小明走进了木屋内部。而在第二个句子中，“小心翼翼地上去”可能表示小明向木屋的楼上或者台阶上行走。所以，虽然两个句子中的其他部分相同，但这一关键动作的差异使得它们描述的空间场景不同。

例 9 中的两段对话，第一段是生成任务，由 GPT-4 扮演教师，让它来设计两段文本，形式有别，但表达的空间场景相同。类似例 8 中“面前”和“身后”这样的关系，不过这里比较的方位义词语换成了“进去”和“上去”。在 GPT-4 完成生成文本任务之后，再以这两段文本为题，出一道判断题，让它来判断这两个文本是否能表达相同的空间场景并解释原因。结果 GPT-4 的回答把自己之前出题时的判断否定了。这就是明显的自相矛盾，即系统内部不稳定

性的表现。

除例 9 这类空间场景理解任务，我们还尝试了中文分词歧义任务和近义词辨析任务。在分词歧义任务中，先由 LLM 按照指定的交集型歧义字符串（如“即使+用”）生成具有分词歧义的一对句子，再由 LLM 来给出两个句子中的分词结果（如“即使+用” vs. “即+使用”）；在近义词辨析任务中，先由 LLM 按照指定的两个近义词（如“一再”和“再三”）生成一对句子 A 和 B，其中 A 句中包含的近义词不能替换为 B 句中包含的近义词，反之亦然，再由 LLM 来完成填空题，将 A、B 句中的近义词遮蔽后让 LLM 来恢复。从实验效果来看，不是太理想，原因是这两类任务对于 GPT-4 来说，都算是高难度的，在生成阶段 GPT-4 也很难生成出都符合要求的句子。这就无法达到测试 LLM 稳定性的目的了。这种测试方式的挑战在于找到合适的生成任务。这个任务对 LLM 来说最好比较简单，有很高的正确率，这样才能批量出题，再将考题交给 LLM 作答，测试其两次答案的一致性。限于篇幅，这里不再展开讨论。

#### 四 结语：透过 LLM 认识人类语言的性质

以 ChatGPT 为代表的 LLM 首次展示了计算机通用对话程序的强大语言能力，它们的表现超越了以往的程序，进入到了“懂人话”并能流畅的“说人话”的境界。这也引发了如何评测 LLM 的问题。本文讨论了评估 LLM 语言能力的方法，这是关于 How 的问题。在文章最后，我们进一步来探讨关于 Why 的问题，即评价 LLM 语言能力的根本目的为何？对此，有两点认识值得强调：

(1) 对 LLM 语言能力的评估非常有必要 —— 因为 LLM 可能在一些重要能力上还不够完美

正如前文已经举例展示过的，尽管目前 LLM 已经见多识广，能说会道，但仍然存在不少显而易见的缺陷，包括没有完成复杂任务必需的“计划”能力，简单的字符计数任务都难以胜任等等，这种表现（performance）层面的巨大冲突，意味着对 LLM 能力的评测仍然非常有必要。因此，在原来传统的 NLP 评测范式基础上，我们尝试了一些新的评测任务，试图在试题难度方面，以及考察的特定能力方面，对 LLM 更有针对性地进行考察。此外，还尝试了一种构造评测任务的新模式：自问自答的生成与理解对攻测试方法，希望能为测试 LLM 的可靠性和稳定性提供参考。

(2) 对 LLM 语言能力的评估实际上是在反观人类自身 —— 因为人类还没有真正理解自然语言

ChatGPT 的“能说会道”和“胡说八道”，集于一身，向我们展示了“语言”和“思维”（或者说“知识”）明显分离的一种关系。这不得不促使我们反观自然语言，反思语言学的主流认识，特别是 Chomsky（2021、2022）关于语言与思想关系的论述。我们可以肯定机器并没有思想，但现在机器却能表现出跟人类相当的语言能力（可能还超过不少个体的人）。那么，该如何看待这种语言能力呢？Chomsky（2021、2022）举过两个对比的例子“The mechanic who fixed the car carefully packed his tools.”和“Carefully, the mechanic who fixed the car packed his tools.”前者有歧义，后者无歧义，ChatGPT 可以对这两个句子的结构做出准确的区分，并给出句子意思的恰当解释。ChatGPT 在谈论这两个句子的句法结构和语义时，它的表现简直就像是一位语言学教授。无怪乎著名计算机科学家 Wolfram(2023)在深入剖析 ChatGPT 后表示“我强烈怀疑 ChatGPT 的成功暗示了一个重要的‘科学’事实：语言意义的结构和简洁性超出我们所知——语言的组织可能是很简单的规则。”面对 ChatGPT 所表现出的惊人能力，任何一个人类评价者，可能都无法再像以往对待 NLP 程序那样，只是出一套测试题，然后再给程序打个分就大功告成了。ChatGPT 的得分无论高还是低，似乎都已经不再重要。因为像 ChatGPT 这样的通用对话程序，它已经不需要一个分数来证明自己。数以亿

计的活跃用户早已证明了它的价值和成就。由此自然产生的问题就是：对像 ChatGPT 这样的 LLM 进行语言能力评测，到底是在评测什么？评测的最终目的是什么？

我们的立场仍然是从语言学角度来回答这个问题。评测 LLM 的目的，本质上是帮助我们重新审视以往人总结的语言知识，更深入地去思考语言“形式”和“意义”的对应关系。像例 8 那样的语言现象，超越了深度学习模型奉为圭臬的“意义即分布”假设，ChatGPT 和 GPT-4 就都无法准确理解。不过，当 GPT-4 去生成具有同类语言现象的句子时，它反馈的结果中却也有让人不禁拍案叫绝的佳作。请看：

例 10:

H<sub>1</sub> 请写一段话，其中包含人或物的空间场景信息。这段话中必须包含“前”这个词。然后把这段话拷贝一份，再将拷贝后的文本中的“前”替换成“下”。要求前后这两段文本中涉及到的人、物对应的空间场景相同。也就是说，尽管两段话中有一处词语差异，“前”和“下”是不同的，但整段话所描述的空间场景却是相同的。

M<sub>1</sub> 原文:在这个宽敞的图书馆中，一位年轻女孩安静地坐在书架前，认真阅读着一本书。  
拷贝文本:在这个宽敞的图书馆中，一位年轻女孩安静地坐在书架下，认真阅读着一本书。

H<sub>2</sub> 请写一段话，其中包含人或物的空间场景信息。这段话中必须包含“走到最里面”这个词。然后把这段话拷贝一份，再将拷贝后的文本中的“走到最里面”替换成“走到最前面”。要求前后这两段文本中涉及到的人、物对应的空间场景相同。也就是说，尽管两段话中有一处词语差异即“走到最里面”和“走到最前面”是不同的，但整段话所描述的空间场景却是相同的。

M<sub>2</sub> 原文:在这个宽敞的超市里，顾客们可以在琳琅满目的商品中挑选自己喜欢的东西，当他们走到最里面，便能发现一个美食区，提供各种美味的熟食和甜点。  
拷贝文本:在这个宽敞的超市里，顾客们可以在琳琅满目的商品中挑选自己喜欢的东西，当他们走到最前面，便能发现一个美食区，提供各种美味的熟食和甜点。

平心而论，对于例 10 中人类提出的可谓“刁钻”的问题，即便是语言学专业人士，恐怕也要费神思考半天，才可能得出答案。GPT-4 生成的文本，不仅在形式上完全符合要求，而且在语义上也经得起推敲。很容易让人觉得，机器具有了“空间想象力”或者“空间思维能力”，要不然它怎么能像人一样，凭空写出两段形式有别的文字，并勾勒出同一幅空间场景图呢？当然，这里展示的只是少数 GPT-4 的佳作，对于同样的任务，它也会产生大量的“胡言乱语”。如果曝光这一类结果，可能就会很容易地把我们推向对待 LLM 态度的另一端：它们是完全跟智能无关的只不过更大号的随机鹦鹉（stochastic parrots）。

到底机器是真正的理解人类语言，还是表现得像是理解了人类语言？对这一问题的持续争论，在 ChatGPT 问世以后并没有平息，反而可能因为它在语言行为上落差更大的表现而加剧了分歧。看到 ChatGPT 完美回答的人会不由自主地认为它具有了真正的语言理解力，而看到 ChatGPT 满嘴跑火车的表现时又会断言它只不过是拾人牙慧地记住了更多的字符串。

我们认为，只有在人类自己真正解开了语言之谜后，才会有真正科学的答案。探索 LLM 语言能力评测方法，正是去追寻答案、在 AI 时代推进语言学研究的一条上佳途径。从这个意义上讲，作为语言学研究者，我们特别欢迎 LLM 的到来，并期待有更多更强大的 LLM 问世。透过 LLM，我们可能有更多机会去洞察自然语言的本质。如果站在“传统”语言学的固有立场，贬低或者无视 LLM 取得的惊人成就，对语言学的发展，大概只有害处而并无益处。毕竟，目前还没有任何一个基于“传统”语言学或者人类语言知识调教的 NLP 程序，在通用语

言能力上可以跟 ChatGPT 们相提并论。

#### 参考文献

- A. M. Turing, 1950, **Computing Machinery and Intelligence**. *Mind* 49: 433-460. <https://doi.org/10.1093/mind/LIX.236.433>
- Michal Kosinski, 2023, **Theory of Mind May have spontaneously emerged in Large Language Models**, Stanford University, <https://arxiv.org/abs/2302.02083> (4 Feb 2023)
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, Yi Zhang, 2023, **Sparks of Artificial General Intelligence: Early experiments with GPT-4**, <https://arxiv.org/abs/2303.12712> (22 Mar 2023)
- Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, Nan Duan, 2023, **AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models**, <https://arxiv.org/abs/2304.06364> (13 Apr 2023)
- Stephen Wolfram, 2023, **What Is ChatGPT Doing and Why Does It Work?** <https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/> February 14, 2023 <https://www.youtube.com/watch?v=flXrLGPy3SU>
- CRFM & HAI, Stanford Uni., 2022, **Holistic Evaluation of Language Models**, <https://arxiv.org/pdf/2211.09110.pdf> (16 Nov 2022)
- Yehoshua Bar-Hillel, 1960, **The present status of automatic translation of languages**, *Advances in Computers*, vol.1, 1960, pp. 91-163. <https://aclanthology.org/www.mt-archive.info/50/Bar-Hillel-1960.pdf>
- Noam Chomsky, 1959, **On certain formal properties of grammars**, *Information and Control*, Vol, 2, Issue 2, June 1959, pp. 137-167. [https://doi.org/10.1016/S0019-9958\(59\)90362-6](https://doi.org/10.1016/S0019-9958(59)90362-6)
- Noam Chomsky, 2021, **Minimalism: Where Are We Now, and Where Can We Hope to Go**, *言语研究 (Gengo Kenkyu)* 160: pp.1-41, [https://www.jstage.jst.go.jp/article/gengo/160/0/160\\_1/\\_pdf](https://www.jstage.jst.go.jp/article/gengo/160/0/160_1/_pdf)
- 诺姆·乔姆斯基, 2022, 读懂我们自己: 论语言与思想, 司富珍, 时仲, 赵欣宇 (译)《语言战略研究》2022 年第 6 期。56-72 页。
- 董青秀、穗志方、詹卫东、常宝宝, 2021, **自然语言处理评测中的问题与对策**,《中文信息学报》2021 年 (35 卷) 第 6 期。pp.1-15。
- 詹卫东、孙春晖、岳朋雪、唐乾桐、秦梓巍, 2022, **空间语义理解能力评测任务设计的新思路——SpaCE2021 数据集的研制**,《语言文字应用》2022 年第 2 期 (总第 122 期), pp.99-110。