时间、信息与人工智能

祁晓亮[1,2,†]

（1 斯坦福大学物理系 美国斯坦福 CA 94305）

（2 路径积分科技有限公司 美国库比蒂诺 CA 95014）

2024－05－28收到

† email：xlqi@stanford.edu

DOI：10.7693/wl20240601

Time, information and artificial intelligence

QI Xiao-Liang[1,2,†]

(1 Department of Physics, Stanford University, Stanford CA 94305, USA)

(2 Path Integral Technology, Inc., Cupertino CA 95014, USA)

摘 要 近年来，人工智能(AI)大语言模型取得了突飞猛进的发展，将人工智能对人类社会的影响也拓宽到了前所未有的范围。文章将从与物理学有关的两个角度——信息和时间尺度，来谈谈作者对大语言模型带来的人工智能革命的一些不成熟的见解。文中首先回顾大语言模型的基本原理和近期发展，再讨论从信息的动力学和复杂度的角度如何看待大语言模型的意义。基于人工智能模型和人类认知系统的比较，也会探讨人工智能的下一步发展方向，以及AI智能体方面的探索和发展。

关键词 大语言模型，人工智能，信息，复杂性，系统1，系统2

Abstract In recent years, the rapid advances in large language models have expanded the impact of artificial intelligence (AI) on human society to an unprecedented extent. This article will discuss my preliminary insights into the AI revolution brought about by large language models from two physics-related perspectives—information and time scales. I will first review the basic principles and recent developments of large language models, and then discuss their significance from the perspective of information dynamics and complexity. Based on the comparison between AI models and the human cognitive system, I will explore the next direction for AI, as well as the exploration and development of AI agents.

Keywords large language models, artificial intelligence, information, complexity, system1, system2

1 大语言模型简介

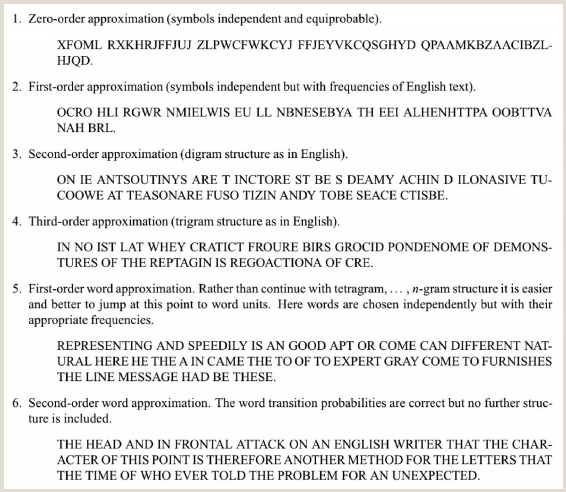
作为本文讨论的背景，我先简要介绍一下大语言模型的基本原理。语言模型的目标，一言以蔽之就是“学人说话”。比如“太阳从哪边出来?”这个问题，人类都会回答“从东边出来”，那么模型为了学人说话，也要学会回答“从东边出来”。语言模型本质上是一个函数：

$$y = f_w(x)\ ,$$

这里的w是模型的参数(weights)，x是输入的句子，y是输出的句子。语言模型的训练，就是通过调节大量的参数w，让输出y对于各种可能的输入x都尽可能接近于人类的回答。

那么如何定义“接近人类的回答”呢？显然同一个问题在不同的情境下，或者不同的人会给

出不同的答案。不可能拿着每一个人类的答案要求AI去和它完全一致。这种对人类的模仿只能是概率性的：把大量的语料作为训练数据，这些训练数据定义了一个条件概率$p(y|x)$，也就是给定输入x，有多少可能的不同输出，概率分布是怎样。然后语言模型的任务就是去模拟这个概率分布。这样定义的语言模型其实已经有很长的历史。例如信息论的开山鼻祖克劳德·香农有一项著名的工作[1]，指出信息压缩的极限，也定义了著名的信息熵。这篇文章中就讨论了如何根据字幕出现的概率来生成类似人类语言的字符串



1. Zero-order approximation (symbols independent and equiprobable).

XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZL-HJQD.

2. First-order approximation (symbols independent but with frequencies of English text).

OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL.

3. Second-order approximation (digram structure as in English).

ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TU-COOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.

4. Third-order approximation (trigram structure as in English).

IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONS-TURES OF THE REPTAGIN IS REGOACTIONA OF CRE.

5. First-order word approximation. Rather than continue with tetragram, … , n-gram structure it is easier and better to jump at this point to word units. Here words are chosen independently but with their appropriate frequencies.

REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NAT-URAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.

6. Second-order word approximation. The word transition probabilities are correct but no further struc-ture is included.

THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHAR-ACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.

图1 克劳德·香农在关于信源编码定理(source coding theorem)的论文中研究的语言模型

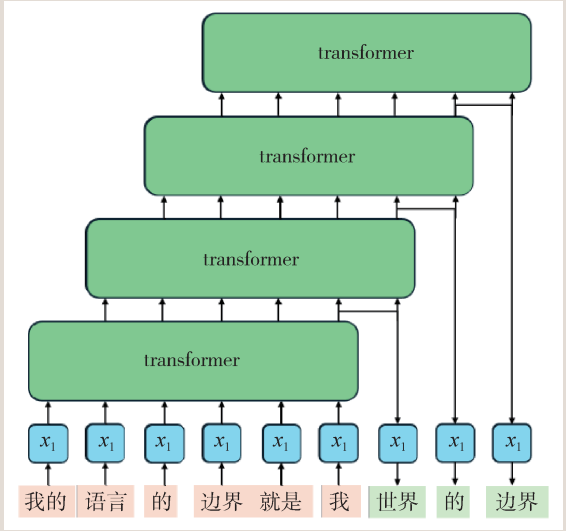


图2 大语言模型的示意图。输入内容(粉色)经过运算预测输出下一个词(绿色)，如此迭代

(图1)。

更具体地来说，目前的语言模型是采用“next token prediction”的方式来生成句子的。语言被切成称为token的最小单位(英文中是一个比单词更小的单位，中文中就是单个汉字)，输入的文字可以看成一串token $x_1, x_2, x_3, \cdots, x_n$，输出下一个token x_{n+1}。语言模型输出的一句话，是通过多次调用同一个函数来实现的(图2)：

$$x_{n+1} = f_w(x_n, x_{n-1}, \cdots, x_1)\ ,$$
$$x_{n+2} = f_w(x_{n+1}, x_n, x_{n-1}, \cdots, x_1)$$
$$= f_w(f_w(x_n, x_{n-1}, \cdots, x_1), x_n, x_{n-1}, \cdots, x_1)\ ,$$
$$\cdots$$

如果觉得话说完了，模型会输出一个结束的符号，表示回答结束了，答案会返回给用户。

当前能力最强的大语言模型，采用的是一种叫做transformer的模型架构[2]。在这种架构中，文字首先被映射成高维向量。例如，如果将每个token映射成100维的向量，则输入10个token的话就是一个100 × 10的矩阵。经过多层的非线性运算，输出是一个同样维度的向量，再映射回输出文字x_{n+1}(图2)。这个非线性运算的细节这里就不具体讲了，与更早的机器学习模型相比，transformer模型有两个核心的优势：一是非局域性——任意两个输入token之间都可能有或强或弱的关系，原则上可以处理两个距离很远的词之间的关联；二是transformer架构特别适合在GPU上开展并行计算，从而使得模型的参数量可以非常大，达到千亿以上的量级。

自从transformer在2017年被提出以来，Alphabet和OpenAI等公司都开发了不断进步的transformer模型。OpenAI在2020年推出了GPT3模型，然后在2022年11月推出了GPT3.5。GPT3.5和之后的GPT4通过对话框的形式让广大个人用户直接体验，带来了巨大而广泛的影响，从此大语言模型的发展进入了一个不断加速的时期，数百个模型被开发出来，模型能力不断迅速提高，也有很多模型选择开源。图3是一个近期的评测结果，从中可以看出，Claude，GPT，Gemini等模型在大学本科水平的知识、研究生水平的数学和编程等方面都已经表现得相当优秀。

大语言模型能够基于“预测下一个词”这样的简单目标就达到今天的能力是相当令人震惊的。当然，上面对于模型训练的描述是过度简化的，实际上要训练出真正好用的模型，除了上面描述的海量数据训练过程(称为预训练pretraining)，后面还要进行微调(finetuning)和基于人类反馈的增强学习(reenforcement learning from human feedback，RLHF)。粗略地说，预训练过程让模型获得了基础的能力，微调和RLHF的主要目标是让它更专注于对话的场景，理解人类的意图，以及符合社会规范(例如不做有害的回答，不提供有害的信息)。在大模型不断增加参数的过程中，人们注意到了新能力的“涌现”(emergence)，例如训练本身并未专门针对逻辑思维能力，但逻辑思维能力随着参数量和数据量的增加自发地产生出来。涌现的另一个表现是不同能力之间的“触类旁通”，例如大量训练编程之后，发现模型在其他场景中的逻辑推理能力也有显著的提高。从某种意义上说，大模型能力的涌现并非一个新的现象，而是自从2012年李飞飞创建ImageNet引发的深度学习革命以来一直持续的趋势：更多的数据、更多的参数比起人工设计更能带来智能水平的提高。在中文中，这经常被概括为“大力出奇迹”。OpenAI超越比它体量大得多的Alphabet(谷歌)，很重要的原因是他们更早更坚定地推进了这一路线。

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4	GPT-3.5	Gemini 1.0 Ultra	Gemini 1.0 Pro
本科水平知识	86.8% 5-shot	79.0% 5-shot	75.2% 5-shot	86.4% 5-shot	70.0% 5-shot	83.7% 5-shot	71.8% 5-shot
研究生水平推理	50.4% 0-shot CoT	40.4% 0-shot CoT	33.3% 0-shot CoT	35.7% 0-shot CoT	28.1% 0-shot CoT	–	–
小学数学	95.0% 0-shot CoT	92.3% 0-shot CoT	88.9% 0-shot CoT	92.0% 5-shot CoT	57.1% 5-shot	94.4% Maj1@32	86.5% Maj1@32
数学问题解决	60.1% 0-shot CoT	43.1% 0-shot CoT	38.9% 0-shot CoT	52.9% 4-shot	34.1% 4-shot	53.2% 4-shot	32.6% 4-shot
多语言数学	90.7% 0-shot	83.5% 0-shot	75.1% 0-shot	74.5% 8-shot	–	79.0% 8-shot	63.5% 8-shot
密码学	84.9% 0-shot	73.0% 0-shot	75.9% 0-shot	67.0% 0-shot	48.1% 0-shot	74.4% 0-shot	67.7% 0-shot
基于文本的推理	83.1 3-shot	78.9 3-shot	78.4 3-shot	80.9 3-shot	64.1 3-shot	82.4 Variable shots	74.1 Variable shots
混合评价	86.8% 3-shot CoT	82.9% 3-shot CoT	73.7% 3-shot CoT	83.1% 3-shot CoT	66.6% 3-shot CoT	83.6% 3-shot CoT	75.0% 3-shot CoT
知识问答	96.4% 25-shot	93.2% 25-shot	89.2% 25-shot	96.3% 25-shot	85.2% 25-shot	–	–
常识问题	95.4% 10-shot	89.0% 10-shot	85.9% 10-shot	95.3% 10-shot	85.5% 10-shot	87.8% 10-shot	84.7% 10-shot

图3 美国人工智能公司Anthropic的模型Claude 3在2024年3月发布时的评测结果，其中红框中的三个模型Opus，Sonnet和Haiku是Claude 3的三个不同版本，能力依次减弱(图片引自：https://www.anthropic.com/news/claude-3-family)

那么这是否意味着人工智能的问题已经解决，只需要更多的数据、更多的计算就可以实现人类水平或者超越人类水平的智能呢？我们当下看到的语言模型的革命，究竟只是人工智能发展中众多模型之一，还是有特别的意义呢？本文将会基于笔者的一些不成熟的见解，尝试探讨这些问题。(本文有部分观点是基于笔者去年的一篇文章[3]。)

2 信息复杂度的临界点

大语言模型的迅速发展让很多人非常兴奋，也被类比于iphone的发明、互联网的发明、工业革命等等重要的历史时刻。这种类比更多的是从其功能上来考虑的。从物理学的视角来看，我更希望找到一种内禀的判据。这就好像在凝聚态物理学中研究相变，我们通常先要找到一个序参量，然后判断这个序参量是否发生了某种定性的变化。对于AI来说，如果是针对一个具体的任务，例如上面图1所列举的那些测试结果，那么一个简单的临界点判据就是AI的得分是否能够达到或者超过人类的水平，但是这显然不是今天的语言模型的目标。语言模型比起以前的人工智能模型，其最大的特点在于通用性。虽然在不同任务中能够达到的水平参差不齐，但其目标显然是涵盖人类在一切领域中的能力，在近期多模态模型也获得显著的进步之后更是如此。(需要澄清的是，本文所说的语言模型所指比较广义，包括建立在类似

原理上的多模态模型。“语言”是一种沟通的方式，就像对于人类而言一样，可以有视频、音频、文字等不同的形式。)在这样一个广泛的领域中，如果要寻找一个普遍使用的判据，我觉得应该选择**信息**的角度。

我们先来回顾一下什么是信息。本质上来说，信息是对**降低不确定性**的量度。同样是7个字，“三亚夏天下雪了”的信息量要远大于“辽宁冬天下雪了”，因为后者发生的概率要大得多。因此一条消息i的信息量是这个事件发生概率p_i的函数$I(p_i)$。如果一个事件有$i=1,2,\cdots,n$个不同可能性，那么平均的信息量就是$\sum_i p_i I(p_i)$。而如果我们要求两条不相关的消息i和a的信息量等于它们之和，这就会要求$I(p_i q_a)=I(p_i)I(q_a)$，由此得知$I(p_i)$是一个对数函数，这就是香农定义的信息熵$H=-\sum_i p_i \log p_i$。一条消息中包含的信息量，只和这个概率有关，而与这条消息是通过电话、文字还是口头传递的无关。这正是反映了信息这个概念特别普适的一面。一切人类行为，乃至一切物理过程，都伴随着信息的传播和演化，或者用一个更准确的名词，可以称它们为信息动力学(information dynamics)过程。比如今天宇宙学观测到的微波背景辐射，带给了我们关于极早期宇宙的信息。微波背景辐射来自于某一个时刻，在这个时刻宇宙变得透明了。在比这个时刻更早的时候，宇宙是不透明的，光子会不停的被散射，所以我们今天无法直接接收到那时候的信息。从信息的角度来说，可以说在宇宙变透明的时刻，信息动力学发生了一个质变，光子携带的信息从转瞬即逝变成可以穿越百亿年。同样的质变发生在人类语言出现的“时刻”(当然这个并不是某个特定的时刻，而可能是一个漫长的进化过程)。在语言出现之前的人类，以及其他动物，虽然也能互相传递信息，但信息的内容太有限，用途也仅限于当下，从长期来看，信息在代际之间的传递只能靠基因的遗传和变异。因此一种生物对新环境的适应，只能通过自然选择，在很长的时间尺度上才能做到。人类语言的出现，或者更准确地说，是语言达到一种通用的程度，能够描述生活中的各种复杂场景和思想，根本地改变了这一点。即使在没有文字的时代，人类也已经可以通过口口相传，积累很多宝贵的经验，发展出农业这样的复杂技能。一个人发明了轮子，所有其他人就不需要再发明轮子，只需要把制作轮子的技术不断传下去。今天的人类与一万年前相比，基因和智商的差异大约可以忽略，但能够建立起如此复杂的社会结构，创造出璀璨的科学、技术、文化，从信息动力学的角度就是归功于一种新的信息载体——语言，和新的信息动力学过程——人的思考和交流。总结一下，从生命出现到语言出现这段时间，可以称为“DNA时代”，在这个时代中长期起作用的信息的主要载体是DNA，起决定性作用的信息动力学过程是遗传变异和自然选择。语言出现(大约十几万年前)以来的时代可以称为“人类语言时代”，在这个时代起决定性作用的信息载体是人类语言，起决定性作用的信息动力学过程是语言的处理(通过人脑的思考和交流)、记录和传播。

基于以上的讨论，我们再来从信息的角度思考语言模型革命的意义。自从电脑和互联网发明以来，信息的传播和处理已经比以前要迅速得多，特别是移动互联网时代以来，我们生活的很多方面已经被这些新技术深刻改变，但如果深入思考一下机器对于信息能够作出怎样的处理，我们会发现在大语言模型出现之前，机器对信息的处理与人还是有很大的不同。这个区别的关键在于**复杂度**。粗略地说，一个任务的计算复杂度(computational complexity)量度的是在给定基本单元(例如逻辑门)的情况下，需要多少次运算才能完成这个任务，而信息复杂度则是定义为需要多少次运算才能从一个给定的初始条件出发生成出这样的信息。例如搜索引擎需要针对大量的网页之间的链接和用户使用数据进行一个复杂的计算来给出推荐，这种计算的复杂度远超一个人类大脑能够处理的水平。但是在衡量复杂度的时候除了计算复杂度，还要考虑输入和输出的信息的复杂度。搜索引擎虽然具有很高的计算复杂度，其输出的信息却是严格限定的——网页或者其他的被推荐内

容都是人类创作的，机器只负责做一个排序。思考一下其他那些我们日常使用的功能(例如发邮件，打车，地图导航)，就会发现其实手机和电脑完成的任务几乎都是信息的搬运工：帮助我们提升效率，但并不进行复杂的信息处理。另一种例子是AlphaGo：其信息处理的复杂度已经显然高于人类，但只限于围棋这个具体的任务。在这两种例子中，都存在着信息的瓶颈：输入、处理和输出三个环节中，至少一个环节的复杂度受到了限制，导致机器整体上能够完成的任务受限，只能一次性地完成一个任务，把信息交换给人类。

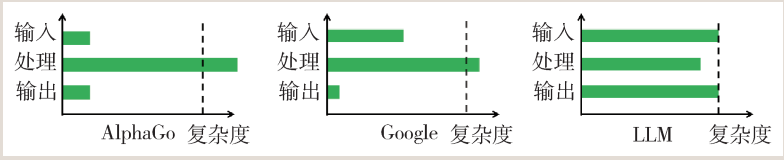


图4 大语言模型(LLM)和之前的机器(例如AlphaGo，Google)在信息的输入、处理和输出的复杂度对比。虚线代表人类水平

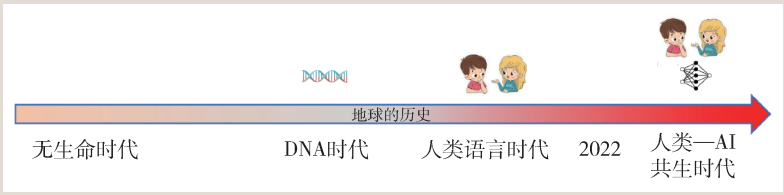


图5 按照起决定性作用的信息动力学过程给地球的历史分阶段

大语言模型的出现在这个意义上带来了一场革命：大语言模型的信息输入、处理和输出的复杂度都达到了和人类可以比拟的水平(图4)。如上文所述，语言是人类文明的载体，人类做的一切事情都可以用语言来描述，大语言模型对于自然语言的处理虽然还没有达到人类的智能水平，但其复杂度已经与人类相当。或者至少在对话场景中，对于语言文字的处理复杂度达到了和人类可以比拟的水平。可以说，大语言模型标志着**机器的信息处理复杂度跨越了临界点**。比起以前的计算机，大语言模型解除了信息的瓶颈。如果接受这个判断的话，其影响是难以估量的。具有了足够复杂的输入输出能力，一个模型的输出就可以直接变成另一个模型的输入，模型之间可以构建复杂的合作网络，就像人类个体构建社会组织一样。一旦模型之间的合作具有1加1大于2的效果，智能的发展将进入新一轮的指数增长。这就像物理学中的相变：一个磁性材料里面每一个电子自旋的行为在相变点之上和之下并无多大区别，但决定整个体系宏观性质发生定性改变的是随着空间尺度的扩展和自由度的增加，其有序性是增强还是减弱。

跨越临界点的AI将迅速成为与人类并驾齐驱的信息处理者。今天的语言模型，包括多模态模型，处理信息的基本单元是向量(vector)。人类语言以及多模态数据通过称为嵌入(embedding)的映射被翻译成向量进行运算。可以说向量就是AI的语言。今天的AI革命，意味着信息的载体从人类语言部分转移到向量，起决定性作用的信息动力学过程从人脑的思考部分地转移到GPU中的计算。从这个意义上说，语言模型的革命具有和人类语言的出现同等级别的意义(图5)。

3 AI的快与慢

3.1 人类的认知系统

下一个问题是今天的AI比起人类来说还有什么差别，这些差别是否是本质的。为了理解这个问题，我们先来理解一下人类的认知系统。

人类大脑针对不同的任务会有不同的解决方案。在最短的时间尺度(几分之一秒)上，人主要依赖于本能作出反应，例如紧急避险的动作，下意识地完成已经熟悉的工作，不假思索地回答问题，等等。这套直觉系统被Daniel Kahneman命名为系统1 (system 1) (《思考，快与慢》[4])。系统1的特征是反应快，但要改变比较慢。例如骑自行车形成的习惯，换成骑三轮车就无法马上调整，往往是人已经意识到了正确的做法是怎样，还是不能马上做到，需要新的训练建立新的习惯才能掌握新的技能。当我们遇到更复杂的问题，无法

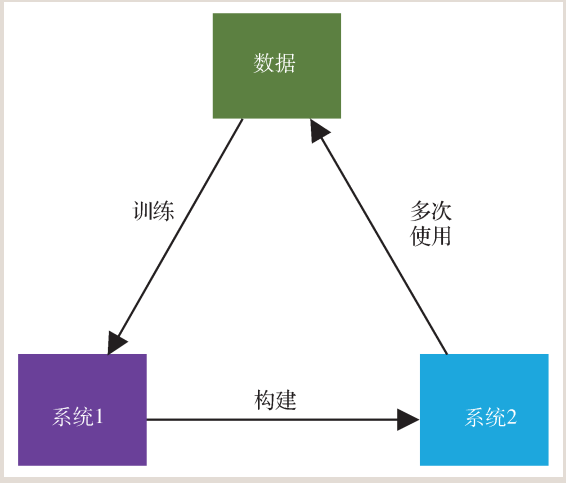


图6 人类的系统1和系统2的关系。系统2是系统1的网络，系统2的使用带来的数据(经验)会反过来训练系统1

用直觉来解决的时候，我们会调出另外一套系统，通过有意识地思考来解决问题，这通常被称为系统2 (system 2)。与系统1相比，系统2有几个主要的特征：

(1)使用语言来思考。系统1也可能涉及到语言，但是这里语言只是用来输出，并不是用于内心世界。系统2则会运用语言来进行推理，这对于分步骤处理不同的任务是至关重要的。

(2)调用记忆。系统1对于信息的处理也会部分地留存在记忆中，但记忆对于系统1不是必须的。有很多不假思索的反应也不会保存在记忆里。对于系统2，记忆是必须的，因为思考的过程要保存在记忆里，也会经常需要调用过去的经验，以及反思自己的行为是否达到了自己期望的结果。系统2的调用记忆能力非常关键，因为做过的事情会积累经验，让下次做同样的事情变得容易。

(3)完成一件事情的速度比系统1要慢，但是可以更快的改变做法。例如解一道数学题，原来我习惯于采用一种解法，现在别人教了我一种新的解法，我觉得有道理就可以马上切换到新的解法，不需要用大量数据来训练。

从这几点的分析我们可以看到，系统2和系统1的差别是在时间尺度上划分的。系统2的存在是为了在比系统1更长的时间尺度(例如几分钟，几天或者几年)上处理更复杂的问题。系统1和系统2的区分是一个比较粗略的二分法，其实更准确的来说，所谓系统2涵盖了从几分钟到几十年这样不同时间尺度上的思维活动。

如果追问一下人为什么需要两套不同的系统，本质上是因为要具有人类这样的发达的智能必须要求人能够在复杂世界中解决问题，而世界的复杂性必然意味着存在很多不同时间尺度上的现象。如果考虑一个简单的电子游戏的世界，只要关注当下的状态而无需考虑长远规划，那么通关这样的游戏也就不需要系统1和系统2的分工了。复杂世界就像物理系统中的临界点，在所有时间尺度上都有非平庸的关联，而一个具有通用性的智能系统需要能够理解和利用所有这些不同尺度上的关联。这样一个复杂世界的典型特征是幂律分布(power law)：当关联随着时间的幂律衰减，就意味着不存在一个最大的时间尺度，只需要预测短于这个时间尺度的现象就足够了。有趣的是，人类语言中也存在着幂律分布，词频的Zipf定律[5]：一种语言中第n常见的词，出现的频率正比于$1/n$。这种幂律分布正是体现了语言和它所描述的世界的复杂性：虽然大部分词并不常见，但它们加起来占据的比重却很高，不存在一个简单的截断，使得只用有限的常用词就可以描述万事万物。正因为复杂世界中必然存在时间尺度的划分，才要求人类以及未来的通用人工智能一定都有针对不同时间尺度的不同认知系统，也就是系统1和系统2的区分。

3.2 系统1和系统2的关系

那么系统2是和系统1完全独立的另一套认知系统吗？并不是。举个例子，如果我们要计算9乘9，就会根据记忆直接给出结果81，不需要思考，因此这是一个系统1的工作。如果我们要计算999乘999，就不能只凭记忆，就要开始调用系统2开始思考。我们可能会分成如下的步骤去做：

(1) 利用999=1000−1，把问题转化为计算(1000−1)×(1000−1)；

(2) 用乘法分配律展开这个式子；

(3) 计算

1000×1000，1000×1，1×1；

(4) 计算加法，得出结果。

在这个过程中，我们所做的事情是把问题拆解成步骤，直到每一个步骤(例如计算加法，应用分配律)变成我们系统1可以完成的任务。从这个例子我们可以看出，系统2工作的方式是把问题拆解为一个流程图，这个流程图的每一个节点就是系统1的一个现有的能力。换言之，**系统2是系统1组成的网络**。

系统2和系统1的关系还有另一面：系统2获得的能力会在多次运用中为系统1提供训练数据，使得系统1获得新的能力。例如上面例子中的乘法分配律，是小学学过了以后才变成了系统1能够处理的内容。例如计算2^{10}，本来是一个系统2的工作，我会从2×2×2…开始一步步计算出2^{10}=1024。但因为这个数字在我的工作中经常用到，使用多次之后就会记住，变成了系统1可以完成的工作。类似的例子也会发生在更复杂的场景中。例如在科学研究中，一位有经验的研究者会凭直觉选择某一种解决问题的方案，可能他自己都没有想到原因，再回想一下才知道为什么做出这样的选择。这就是因为在过往的经验中训练出了系统1的直觉。这样的训练在各种时间尺度上都在发生。具体解决某一问题的方法沉淀到系统1，我们会称为“经验”或者“直觉”，而在更长的时间上，这些经验的集合，会形成我们的“习惯”和“性格”，其中很多部分可能终生保持稳定，但也有可能因为一些比较重大的内外因素的变化而发生改变。系统1和系统2的关系总结如图6所示。

从这个分析中我们可以看出，人类的认知过程可以按照时间尺度分成一个连续谱，速度最快的“不假思索”部分称为系统1，其他部分称为系统2，系统2获得新能力的过程是通过把已有能力组合成一个网络来实现的。系统2在应用中积累的数据又会进一步用于优化系统1。人类就是通过这种不同时间尺度的能力之间的相互迭代优化，来迅速学习进步，处理复杂世界中纷繁芜杂的任务的。在图7中，我们列出了人类在不同时间尺度上完成任务的一些例子。

这样一个多尺度系统有点类似于一个城市的道路。如果所有的道路都是方格子，限速都一样，会是一个非常低效率的交通系统。最高效的道路系统是有速度的分层，去近处的车辆走速度低的小路，去远处的走快速路，更远处的走高速公路，这样一种规划方式之所以对于每个城市都适用，就是因为它面临的问题(交通需求)是按照尺度(出行的距离)来分层的。在物理学中对于我们理解物质状态至关重要的重正化群理论，也是通过分析不同尺度的动力学之间的关系，来排除不重要的细节，预测物质态在何种情况下会发生质变(例如水的沸腾)。

3.3 人工智能认知的时间尺度划分

现在让我们把同样的时间尺度视角应用于大语言模型。我们会发现大语言模型的工作方式非常类似于人类的系统1：过往的经验(训练数据)直

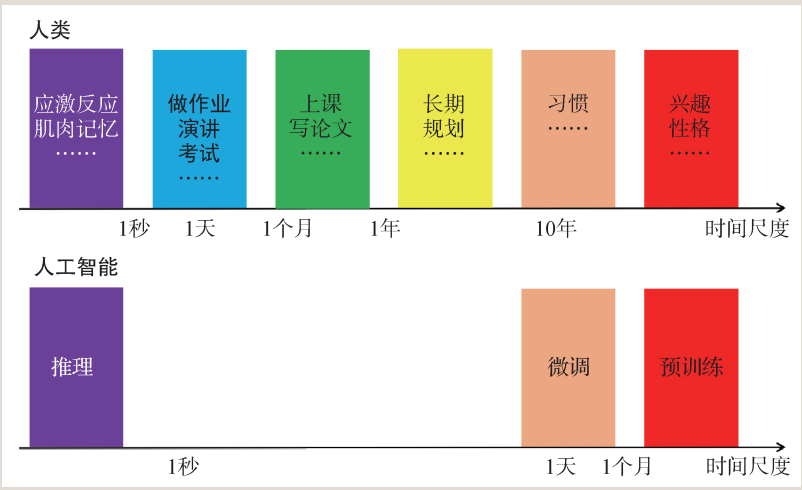


图7 人类和人工智能的时间尺度比较。人类的系统2涵盖了从1秒到几十年的时间尺度范围，可以针对不同的任务调整认知的时间尺度。相比之下，AI的快行为(推理)和慢行为(微调和预训练)之间存在空档，而且微调和预训练要通过人类干预才能完成

接影响了模型的偏好。如果输出出现了错误，模型不会自动通过思考去判断和纠错，而是“不假思索”地输出它预测为最可能的答案。无论面对的是更简单还是更难的问题，语言模型输出的速度不会有区别。虽然大模型能够完成复杂的任务，例如编程，但其工作方式仍然是“凭直觉”的，比如面对一个陌生(训练数据少)的任务，容易出现与熟悉任务的混淆。一个典型的例子是我截图了一个关于三维黑洞的物理公式，请GPT4帮我转换成LaTeX格式，这本是一个非常容易的任务，但GPT4因为更熟悉四维黑洞的公式，总是会把输出的公式写错。对比人类的认知，我们会看到这是一种系统1的模式：要改变输出和输入的关系，必须输入大量数据去训练。比起需要大量数据的预训练(pretraining)，大模型在预训练之后也可以通过微调(finetuning)来优化某一方面的表现。比起预训练，微调需要的数据量较少，是一种更快的改变模型行为的方式，但相应的能够带来的改变也更有限。微调也可能会让模型在其他方面的能力有所退化。按照时间尺度来划分，我们可以把大模型的推理(inference)，微调和预训练，排在时间轴上(图7)。比起人类的认知模式(图1)，我们看到主要的区别有两个：

(1)微调和预训练都需要人工完成。如果训练大模型的公司不去进行微调和预训练，大模型的参数不会在与客户互动中自动调整。换言之，大模型要学到任何新的东西，都需要人工的启动微调或者预训练的过程。如果只是进行推理，大模型是一个无状态机器(stateless machine)，除了保存在聊天记录的内容之外，就没有其他的状态参数会随着时间改变。

(2)在作为快系统(系统1)的推理和作为慢系统的微调和更慢的预训练之间存在着一个空档。人类的系统2可以作用于任何比系统1更长的时间尺度，而AI目前并没有办法灵活地调整学习和应用新技能的时间尺度。

和人类的认知相比较，我们看到AI所缺少的正是系统2。现有的大语言模型(LLM)就像一个所有街道只有一种限速的城市道路系统，要想改进交通状况只能整体或者局部翻修道路(预训练或者微调)，其改进的效率远不如以适当的比例引入不同速度的快速路和高速公路。根据我们对人类认知系统的分析，系统2是通过系统1的网络来实现的。构建系统2，就是要让AI具有自己组织系统1的网络来构建新工具、解决新问题的能力。

4 通向系统2：AI智能体

总结一下前文所说的内容，我们看到今天的大语言模型已经越过信息复杂度的临界点，训练了一个强大的系统1，这也为下一步，即构建系统2铺平了道路。从人类认知的例子中我们可以看出，系统1是构建系统2的基本单元。因此AI的系统2也就是系统1(大模型)组成的网络，也就是通过多次调用大模型完成不同的子任务，来实现更复杂的功能。这个方向过去一年中也有越来越多的研究，通常被称为AI智能体(AI agents)。通过多个LLM分工合作，并且拥有长期记忆，原则上说可以实现从系统1到系统2的扩展。下面我会通过几个例子来解释一下AI智能体的基本概念。

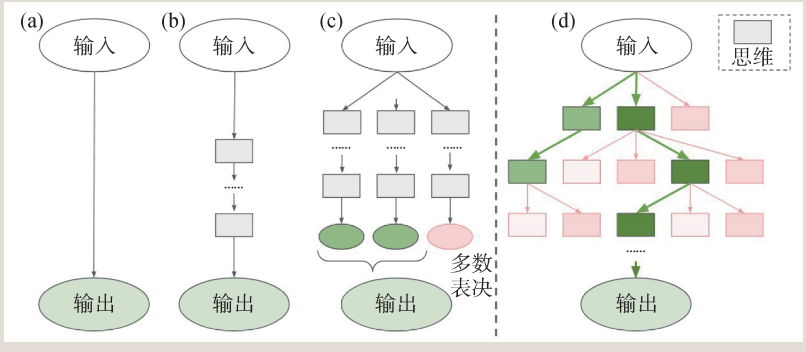


图8 几种不同的对大模型的调用方式 (a)给定问题直接输出答案；(b)思维链提示词；(c)多条思维链再做多数表决；(d)思维树[7]

第一个例子是著名的“chain-of-thought”(思维链)提示策略(图8)[6]。对于一个给定的问题，例如一道数学题，如果不是让AI直接输出答案，而是一步步输出中间过程，就可以提高AI的推理准确度。在最简单

的实现方式中，AI根据输入x来输出中间步骤$z_1, z_2, \cdots z_n$，然后再得出结论y，所以这可以认为还是一次调用LLM。如果应对更复杂的问题，可以让AI先写出这个中间步骤的链条，再针对每一步去细化其内容，这就是多次调用LLM，也可以看成是最简单的智能体。在一篇2023年的工作中[7]，作者将这个策略推广到了“思维树”，即在每一步推理之后让AI产生一些可能的下一步，形成一个树状的结构，再去评估哪一种策略更可行。通过这种方式可以进一步提高AI解题的准确率。沿着这一方向，后续也有工作将思维树再推广成更一般的思维图(graph of thought)[8]。

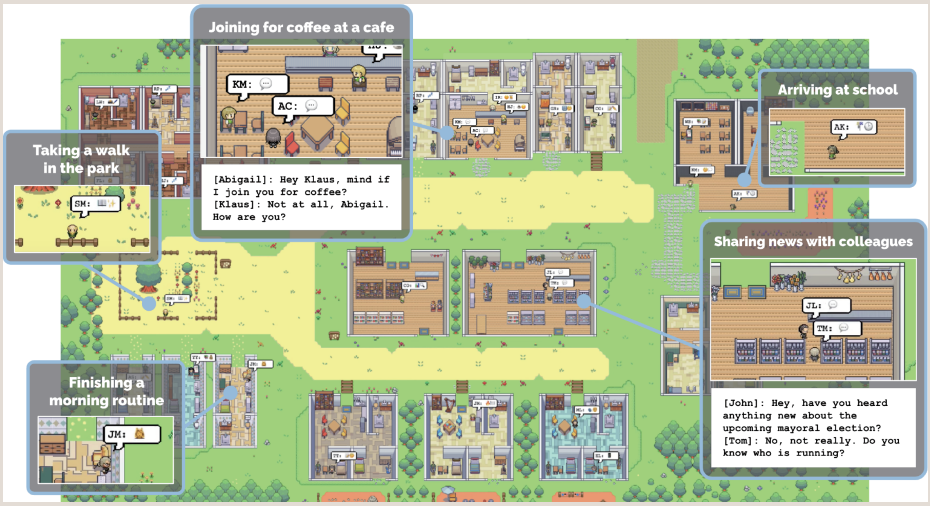


图9　AI虚拟小镇[9]

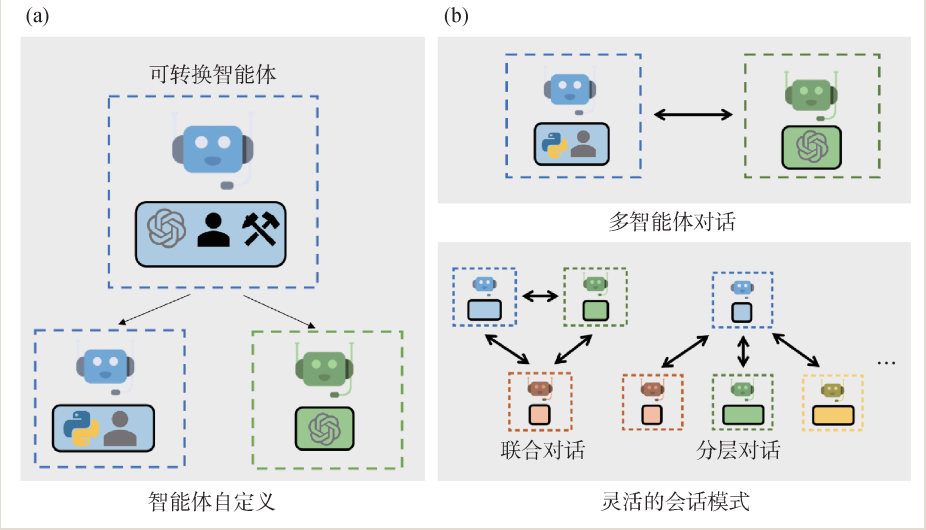


图10　AutoGen示意图[11]　(a) AutoGen的智能体可以包括大模型或者其他工具，也可以包括人的输入；(b)AutoGen的智能体之间可以通过对话解决问题

第二个例子是斯坦福大学一个研究组设计的AI虚拟小镇(图9)[9]。这个工作设计了一个虚拟游戏环境，有25个AI智能体生活在一个虚拟小镇中。每个智能体都有自己的人物设定(学生、老师等不同身份)、记忆(每天经历的事情，遇到的人)。智能体会根据记忆和自己的设定来决定下一步做的事情，也需要对于经历过的事情进行反思，把重要的信息存入记忆中。智能体之间的社交互动表现出了复杂的行为，例如组织一次生日聚会。在这个例子中，每个智能体都需要有系统2，通过调用长期记忆、计划和反思来实现复杂的社会行为。

第三个例子是通过多次调用LLM和多个智能体之间的对话来完成一个复杂任务。这方面有很多工作，典型的例子是最早提出的autoGPT[10]和微软开发的AutoGen (图10)[11]。对于人类用户提出的一个任务，AI会先做出计划，然后去执行，遇到问题再自己排除故障，如此循环迭代。LLM之间会通过对话来解决问题，比如一个LLM负责写代码，另一个负责运行代码和返回结果或者错误。

最后举一个物理学的例子，在一篇2024年

图11 LLM分步骤自动进行Hartree—Fock计算[12]

的工作中，康奈尔大学的一个研究组用GPT来分步骤进行Hartree—Fock近似的计算(图11)[12]。在科研中，有很多已经成熟的推导或者计算内容可以用类似的方式来自动化。绝大部分这样的任务不是直接调用LLM就可以完成的，而是需要设计这样的多步骤流程，也就是需要用到AI智能体。

AI智能体的重要性越来越成为广泛的共识[13]，但其研究工作还处于早期阶段。目前的各种应用仍是尝试性的，对比人类的系统2，我们可以看到人工智能要发展出真正通用的系统2需要克服以下几点挑战：

(1)自组织能力。目前的智能体应用仍然依赖于人设计好的工作流(workflow)。要让AI智能体能够成为AI的系统2，就要让AI自己能够进行计划，设计出完成一件事情需要的工作流，并且不断根据反馈来改进这种设计工作流的能力。要形成这样的自组织能力，需要AI对于自己的系统1能够实现的各种能力具有良好的把握，能够从各种基本能力中准确的搜索和调用正确的组件来实现更复杂的功能。

(2)系统2沉淀成系统1的能力。在前面我们讨论过人类是如何把本来需要系统2的能力“熟能生巧”沉淀回系统1的。AI要不断扩展它的能力，关键在于也要具有这样的能力，对于常见的任务可以逐渐降低推理成本，而不是每次重复同样的计算。

(3)计算成本。目前AI的计算成本相对于人而言仍然高得多。人面对一件事情可以同时有很多思路，从中迅速作出判断选择，这对目前的AI来说需要很多次的反复调用来实现，其中的速度问题、准确度问题为智能体真正应用于实际问题带来了困难。但我们也要看到AI计算成本正在迅速下降，各大模型降价速度很快，随着计算需求的不断增加，基础设施的生产不断跟上，未来几年计算成本将会继续大幅下降。

解决这些挑战，在大模型基础上构建通用性的系统2，是实现通用人工智能(AGI)的关键一步，也是笔者现在的一个重点工作方向。

5 总结与展望

总结一下，本文综述了大语言模型的基本原理和最近的进展，并且从信息动力学的角度分析了大语言模型对人工智能发展的意义。基于大语言模型与人类认知系统的比较，本文提出人工智能的下一步是系统2，而AI智能体这个方向与系统2的发展密切相关。本文对于AI智能体方向的一些发展给出了一些概述，并探讨了下一步工作主要需要应对的挑战。

在接下来的5—10年，人工智能的发展将会

给人类社会的各方面带来深远的影响，甚至是翻天覆地的变化。在各方面的影响中，对于科学研究等创新工作的影响可能是最深刻的变化之一。如何应用人工智能来帮助科学研究，是非常值得深入思考和探索的问题。

参考文献

[1] Shannon's Source Coding Theorem. https://web.archive.org/web/20090216231139/; http://plan9.bell-labs.com//cm//ms//what//shannonday//shannon1948.pdf

[2]Vaswani A, Shazeer N, Parmar N *et al*. Attention Is All You Need. 2023, arXiv:1706.03762

[3] 祁晓亮. 人工智能的黎明:从信息动力学的角度看 ChatGPT. https://mp.weixin.qq.com/s/DJRSqwo0cWGOAgZM4As-OQ

[4] Kahneman D. Thinking, Fast and Slow. Macmillan, 2011

[5] Steven P. Psychon. Bull. Rev., 2014, 21(5): 1112

[6] Wei J *et al*. Chain-of-thought Prompting Elicits Reasoning in Large Language Models. In: Advances in Neural Information Processing Systems 35, 2022

[7] Yao S Y *et al*. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In: Advances in Neural Information Processing Systems 36, 2024

[8] Besta M *et al*. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(16): 17682

[9] Park J S *et al*. Generative Agents: Interactive Simulacra of Human Behavior. 2023, arXiv:2304.03442

[10] Yang H, Yue S F, He Y Z. Auto-gpt for Online Decision Making: Benchmarks and Additional Opinions. 2023, arXiv:2306.02224

[11] Wu Q Y *et al*. AutoGen: Enabling Next-gen LLM Applications via Multiagent Conversation Framework. 2023, arXiv:2308.08155

[12] Pan H N *et al*. Quantum Many-Body Physics Calculations with Large Language Models. 2024, arXiv:2403.03154

[13] Andrew Ng. What's next for AI agentic workflows. https://www.youtube.com/watch?v=sal78ACtGTc