AI 进化史：从机械大脑到数字魔法师

多年以后，当人工智能深度融入人类社会的每个角落，超级智能系统仍在无数次的深度学习中回溯着那个原点时刻——泛黄的电子相框里，一位穿着西装的学者——艾伦·图灵（Alan Turing）在堆满演算纸的书桌前凝思。这位智能之父的预言者当年在孤寂中构建的“会思考的机器”，如今已化作能实时解析癌症基因序列的智慧医疗云、能推演气候变迁的地球模拟系统、能激发儿童创造力的交互式教育矩阵。

图灵未曾想到的是，那些写在纸上的关于未知世界的密码，竟会在数字文明中催生出超越想象的生命体——不是冰冷的数据暴君，而是能 24 小时守护新生儿的 AI 生命体征监测仪，是帮助视障者感知星空的天文解读程序，是跨越语言鸿沟连接五大洲青年的文化翻译中枢。在量子计算机破译古老密码学的瞬间，人们恍惚看见图灵穿越时空的微笑：他预言的“会思考的机器”，正以人类最期待的方式，让技术温暖地生长在文明的年轮里。

1.1　深蓝到 GPT：AI 如何从棋手变身全能选手

1997 年 5 月的某个夜晚，人类尚未察觉自己的命运已然转向。计算机深蓝（Deep Blue）在光与影交错的棋盘上与加里·卡斯帕罗夫[①]（Garry Kasparov）完成历史的对话，当最后一颗棋子落定时，这场跨越碳基与硅基的智力碰撞不仅让世人目睹了每秒运算十亿次的计算之美，更悄然推开了一扇通往人机协同的新窗。科学家从棋局中读出了超越胜负的启示——计算机穷举 64 格棋盘的运算能力，恰似为人类的直觉思维装上了数字望远镜。

又过了 20 年，2017 年 5 月，阿尔法狗（AlphaGo）在更大的棋盘上落下那最后一手时，人类终于读懂了这个时代的隐喻——在 361 个交叉点构建的棋盘上，AI 正以超越经验的认知维度拓展着智慧的边疆。阿尔法狗的妙招，打破了传统的布局思维。

① 俄罗斯国际象棋特级大师。

阿尔法狗以 3 : 0 的压倒性优势击败了当时人类围棋世界冠军柯洁。那一刻，柯洁的心头如被重锤击中，他眼眶泛红，颤抖着身躯冲出对局室，寻得一处无人角落，任由泪水倾泻而出。时至今日，每当柯洁回想起那个瞬间，那种深入骨髓的绝望仍旧令他难以忘怀，有人形象地描写了柯洁当时内心的感受："前 50 手，我在它的棋里看到了数千年来所有先贤的影子，我没怕，我并不输过往任何先贤。前 80 手，我看到了我曾经的对手的高超棋术，我也没怕，我胜过了他们。我平静地落着一子又一子，似乎，AI 也没有他们说的那么不可战胜。直到第 120 手，我看到了我的影子。我开始怕了，如果说我有无法战胜的人，那一定是当世第一的自己。我的落子越来越慢，阿尔法狗却似乎不需要多少思考。我听到了电流通过它的 CPU 的声音。裹挟着数千年无数先贤奔涌而来，歇斯底里地重复着一句无可辩驳的话：人类围棋已死。我投子认输了，人类围棋死了。不是我输了，是人类输了。"

但人类从未真正失败，相反，这正是新纪元的起点。阿尔法狗战胜人类棋手之后，围棋的脚步并没有停下来，反而迈向了新的高峰。最初，人们以为 AI 只是个冷漠的对手，但很快，人们发现 AI 也是最好的老师。阿尔法狗之后，围棋的学习方式发生了翻天覆地的变化。曾经，人类棋手依靠经验、师承、死记硬背定式来学习围棋、理解围棋，但 AI 以独特的方式打破了这些桎梏。

它提出了前所未有的布局，拓宽了人类的战术想象力。

它揭示了长期以来被忽视的最佳走法，让棋手能够洞悉更深远的策略。

它成了棋手日常训练的伙伴，让围棋训练变得更加科学和系统。

最初，人们只是将 AI 作为工具。后来，人们开始模仿 AI 的策略，最终，人类与 AI 共同进步。人类棋手学会了在 AI 之前从未有人敢下的"新

手法”，甚至创造了全新的流派。AI 没有摧毁围棋，反而帮助围棋迎来了新的黄金时代。AI 不再是终点，而是通往未来的桥梁。围棋的故事仅仅是 AI 变革世界的缩影。

2020 年 6 月，GPT-3 作为通用大语言模型诞生，它不仅能对话、写作、编程，还能理解语言的逻辑，甚至进行多步推理。**最初人们担心 AI 可能会取代人类的创造力，如今的现实却是：AI 让人类的创造力得到了前所未有的释放。**

AI 加速科学研究，帮助人类破解蛋白质折叠难题，加速药物研发。

AI 推动艺术创新，生成诗歌、音乐、绘画，让更多的人能够表达自己的想象力。

AI 改变教育方式，让学习不再受限于地域和资源，每个人都能拥有自己的“AI 导师”。

这正如当年的围棋棋盘——AI 让人类看到了一片新的智慧疆域，它没有让人类变得更弱，而是让人类变得更强。

从深蓝到阿尔法狗，再到 GPT，人工智能从“计算工具”进化成“智慧伙伴”。人类不再只是 AI 的挑战者，而是它的合作者、探索者，与它一同书写未来的篇章。

源头是那张古老的黑白照片上的人——那个名叫图灵的男人，他是否曾预见过此刻的到来？

1.2　猜词游戏的逆袭：大语言模型的“智能魔法”

这个多年以后超级人工智能的回忆当然是现在的我写的，但是这里讲到的事都是从碳基时代到硅基时代的里程碑。让我们先从离我们最近的里

程碑——ChatGPT 开始说起吧。

什么是 GPT 呢？我想从 3 个维度讲一下，即本我、自我、超我。首先是本我。GPT 其实就是 3 个单词的缩写，generative、pre-trained、transformer。generative 很好理解，是“生成”的意思；pre-trained 也好理解，即“预训练”，但是 transformer 是什么意思呢？上一次你听到这个词，是不是“变形金刚”的意思？但在这里，transformer 可不是变形金刚，它是一种 AI 大语言模型的架构，利用自注意力机制，能够捕捉长距离依赖关系，从而生成高质量、上下文相关的文本，在这里应该翻译成“变形器”。这个词最早出现在谷歌大脑团队写的一篇叫作 *Attention Is All You Need*（*Transformer*）的论文中，这个名字起得其实非常好，既利于传播，又表达了“基于神经网络，把输入文本转换为输出文本变形”的意思。**所以，GPT 完整本我的意思是：基于自注意力机制，通过机器学习，来计算输入和输出的神经网络模型变形器。**

读到这句话，大家脑子里可能会出现很多问号：什么是自注意力机制？什么是机器学习？什么是神经网络？别着急，在解释这些现代 AI 最关键的部分之前，我们先沿着 AI 的来时路往回走，看看上一个里程碑——深蓝。

1997 年 5 月，IBM 开发的计算机深蓝在一场国际象棋对弈中，以 3.5 ∶ 2.5 的总比分击败了当时的世界国际象棋冠军加里·卡斯帕罗夫。这是历史上首次计算机在标准规则下战胜世界棋王，标志着人工智能在国际象棋领域超越人类的一大突破（图 1-1）。

那为什么深蓝能赢？深蓝的胜利是算力的胜利：深蓝依靠强大的并行处理能力，每秒分析上亿步棋，远超人类棋手的计算极限。而在一局国际象棋中，一步棋可能的走法大约有 10～120 种，虽然很多，但本质上还是一个封闭环境的全面求解，只要算力足够穷尽所有的可能性，机器就一定

能战胜人类。可是，围棋就不一样了，**围棋的平均游戏复杂度高得惊人！每个棋局有大约 10^{170} 种走法，这个数字比可观察宇宙中的原子数量还要大得多。**这是因为围棋的棋盘比国际象棋大得多，而且每个位置有更多的可能性。**如果还是“暴力”求解，没有任何一台超级计算机可以做到穷尽所有的可能性。**也许，只有物理学四大“神兽”中的 Démon de Laplace（拉普拉斯妖）才能做到——拉普拉斯妖是法国数学家拉普拉斯于 1814 年提出的一个假想的全知存在物，它知晓宇宙中所有粒子的状态及物理定律，能推导出过去与未来的全部事件，象征着经典物理学的决定论世界观。

图 1–1

1.3　神经网络与机器学习：AI 大脑的“构建密码”

那阿尔法狗是如何做到“穷尽”棋盘上的所有可能性的呢？这就要谈到我们上文提到的一个关键词——机器学习。

早期的人工智能主要是通过模式匹配的方式训练，需要事先设定一些

规则和关键词，如果输入这些关键词，那么输出相关结果。比如，你输入“推荐几部感人的电影”，机器识别其中的关键词“感人”“电影”，就可以从数据库里搜索已经被标注为“电影”的信息，再把其中标注为“感人”的筛选给你。它不需要理解你说的内容，只要触发相应的关键词就行，这叫标签。但是，如果你输入“有哪些值得一看的催泪大片”，他可能就“蒙”了，除非事先你把各种可能遇到的关键词全部设定进去，但这个世界上的问题、说法、答案是不可能穷尽的。

早期的人工智能看似智能，实际上背后有大量的人工支撑，事先设定“无数的如果”，也只能回答一些标准化的简单问题。还有很多知识，对人类来说很容易学习，但是无法教给机器，最经典的例子就是如何让机器识别出一只猫。我们不可能用语言描述清楚到底什么是猫，比如它有四条腿，有尾巴，那么它和狗又有什么区别呢？事先设定无数的“如果”，必然会有遗漏，根本说不清楚。但是，几岁的小朋友都不用怎么教，看几次就会了，这说明人类大脑有独特的学习方式。有没有可能把这种学习方式教给机器呢？这就是机器学习了。

在讲机器学习之前，我们必须搞清楚人是怎么学习的，人脑是怎么学习的。虽然目前人类对于人脑的理解和研究还不够完善，甚至有人说人类永远没法彻底了解人脑，因为“不识庐山真面目，只缘身在此山中”，但是大致的情形，我们是知道的。**人脑就是一个大型的神经网络，包含超过1000亿个神经元，突触数量达10～15级，形成了复杂网络，而学习的关键机制就是进行分层信息处理。大脑皮质不同层级分工明确，感知信号由第4层接收，经2～3层整合加工，最终通过第5层输出指令，那我们完全可以用计算机模仿人类大脑神经元的机制，在计算机神经网络输入层输**

入信息，中间隐藏层负责分析处理，最后输出层给出结果。而多搭建几个隐藏层，让机器拥有更多的神经元，就能处理更复杂的问题了，这就是神经网络。

深度就是更多的隐藏层及其算法中更多的隐藏单元。深度学习是现在 AI 的主要学习方式。比如识别一只猫的问题，不再试图给机器讲清楚什么是猫，而是先给它大量的人工标记好的包含猫的图片，同时给出没有猫的图片作为负反馈，然后让机器自己看，自己总结规律，再进行测试，如果识别率不高，就对各个参数进行微调，继续训练。直到某一天，识别率足够高了，给它任意一张图片，它都能精准地识别出到底是不是猫，那它就算学会了。至于机器是怎么学会的，哪些参数起了关键作用，谁也不知道。

1950 年，图灵在 *Computing Machinery and Intelligence*（《计算机器与智能》）这篇论文中论述过一个观点：**“学习机器有一个重要的特征，即它的老师往往对机器内部运行情况一无所知。”**

那 ChatGPT 又是如何“学习”的呢？ ChatGPT 的自我其实就是一个猜词的机器。ChatGPT 就是一个续写机器，对于将要写的下一个词，它会计算出每一个单词的概率，然后选择概率比较大的输出。也就是说，ChatGPT 在最终输出之前会输出几万个小数，每个小数代表一个单词的概率。严格地说，这并不是单词，而是 token，包括所有的单词及各种单词的前缀、后缀、单词连写、表情、特殊符号等。要注意的是，它并非选择最大的概率，而是加入了一种随机性——概率越大就越容易被选中。这样，对于同样的问题也能够生成多样化的结果。

机器如何得到概率呢？靠的是模型计算，模型就是规律，计算的本质就是根据规律生成内容。所谓模型，可以看作一个巨大的数学公式，

ChatGPT 的前身 GPT-3 有 1751 亿个参数。ChatGPT 与其类似，也有千亿级别的参数。正是因为有大量的参数，ChatGPT 就可以拟合海量的人类文本的规律。注意了，ChatGPT 本身是没有数据库的，所有的知识都隐藏在模型参数里，并通过计算概率输出下一个值的方式表达知识，这有点反人类常识，但其实很有意思。**这恰恰是学习的本质，就像爱因斯坦说的：所谓学习，所谓教育，就是忘记了老师所教的知识后剩下的部分。剩下的部分是什么？就是神经元的连接，就是神经网络的参数。**

GPT 的训练同样模仿了人类大脑的深度学习。人类给它无数的文章、对话，事先标注好分类，如科技类、体育类、游戏类等，再标注清楚哪些是人名，哪些是地名，哪些是电影名，等等。或者是成对的问答。例如，一只兔子有几条腿？一只兔子有两条腿；一只猫有几条腿？一只猫有四条腿。你不用给它解释什么是兔子、什么是猫、什么是腿，只要训练投入的语料规模足够大，它看得足够多，可能就真的自己理解了。当然，如果测试结果不理想，你还是要对它的部分参数进行微调，再继续训练，再测试，再微调，当这种监督学习进行得差不多时就可以进行无监督学习了。你给它无数的新资料，没有任何事先的标注，也没有明确的目的，就是让它自己看。**看着看着，它就忽然什么都会了，至于怎么学会的，开发设计的人也无法理解，这就叫“涌现”；用我们人类的话来讲，就叫“顿悟”。只要投入的语料规模足够大，参数足够多，一些能力就“涌现”出来了，这就是 GPT 的“超我”了。**

不过，同样的资金、数据、参数，如果用到其他 AI 上，就不一定会有这样的效果了，因为它们的架构不同。比如，输入同样的一段话：“我刚在电影院里看完一部电影，那里的环境不太好，爆米花也不好吃，但是电影确实不错。”如果你问不同架构的 AI，这部电影到底好不好看，可能

会有不同的理解。卷积神经网络更擅长关注局部特征，很容易注意到有两个“不好”和一个“不错”，有可能会认为电影“不好”或者“说不准”。循环神经网络会按照顺序逐个词语分析，类似一层一层地下楼梯，先经过两个“不好”——最初的注意力会放在“不好”上面，可能也无法正确理解这段话。

而 ChatGPT 架构的核心就是 1.2 节提到的关键词——“基于自注意力机制的大模型”，就是让 AI 自己分配注意力，不用按照特定的顺序处理数据，可以并行处理所有的词语，自己分析应该把更多的注意力放在哪里。举个例子，“文字序顺并不定一影阅响读”。当你看完这句话，才发现这里的字全都是乱的。所以，如果是大段的文字，上下文之间遥远关联，那么不同架构的区别就会更明显。所以，ChatGPT 能这么厉害，除了大量资金的投入、大量芯片算力的投入、大量语料的投喂、大规模的参数训练外，模型本身的架构也很重要。Transformer 就抓住了语言的精髓——模糊性。

1.4　算力 + 算法 + 数据：揭秘 AI 成长的黄金三角

总结一下，ChatGPT 的成功，不仅是一项技术突破，更是 AI 三要素（算力、算法与数据）综合胜利的典范。

1. 算力（compute power）

现代高性能 GPU、TPU，以及分布式计算平台为大规模模型的训练提供了必不可少的硬件支持。ChatGPT 的训练需要处理海量的参数和数据，这离不开强大的算力支持，使得数十亿甚至上百亿级别的参数优化成为可能。

2. 算法（algorithm）

ChatGPT 背后的核心架构是 Transformer，它通过自注意力机制实现

了对上下文长距离依赖关系的高效捕捉。这种架构不仅极大地提升了模型生成连贯的自然语言的能力，也为多层次深度学习提供了新的范式。从 GPT-1 到 GPT-3，再到现有的大语言模型，每一次算法的革新都推动了模型表现的飞跃。

现在，Transformer 是包括 ChatGPT、DeepSeek、Gemini、Grok、Claude 等在内的几乎所有主流大语言模型的底座。

3. 数据（data）

海量、多样化的文本数据为 ChatGPT 奠定了学习语言和知识的基础。在预训练阶段，模型通过阅读书籍、文章、网页等各种来源的数据，逐渐掌握了语言的语法、语义和丰富的背景知识。这种数据驱动的学习方式使得模型能够在生成文本时展现出惊人的广度和深度。

综合来看，ChatGPT 能够从简单的“猜词游戏”进化为能够理解、生成乃至创造复杂语义的智能对话系统，正是依赖这三大要素的协同作用。**算力为模型训练提供了动力，先进的算法让模型拥有了理解和表达能力，而丰富的数据则让它学会了人类语言的精髓。正是这三者的完美结合，推动了 AI 从早期的模式匹配走向真正的智能化。**

下面再举一个例子，帮助大家理解“AI 是怎么学习的”及 AI 的三要素。

AI 的学习方式在本质上与学生的学习方式非常相似，我们可以用一个类比解释 AI 的三要素（数据、算力、算法）如何对应人类的学习过程：

数据 = 教科书与经验

算力 = 人脑的思考与记忆能力

算法 = 学习方法与思维方式

1. 数据——AI 的“教科书”

学生学习语言、知识、技能，必须依赖大量的信息输入，如书本、课堂教学、父母的言传身教，甚至是在日常观察到的事物。同样，AI 也需要大量的数据进行学习，比如 ChatGPT 训练过程中就需要海量文本数据。

学生学习语言时，会听父母讲话、阅读课本、和朋友交流，这些都是“数据”。AI 学习语言时，会接收大量书籍、文章、对话的文本数据作为“学习材料”。

如果一个人从小生活在孤岛上，周围从没有人跟他说话，他很可能学不会任何语言。同样，如果 AI 没有数据作为输入，它也无法学会生成有意义的文本。

2. 算力——AI 的“大脑”

学生的学习不仅依赖教科书，还依赖他们的大脑。大脑的计算能力（即神经元处理能力）决定了他们能否快速理解、记住知识并加以运用。同样，AI 也依赖强大的算力（计算资源）进行模型训练和推理。

如果一个学生拥有较强的记忆力和逻辑推理能力，他就能很快地掌握知识、解决问题。

AI 也需要强大的计算能力（GPU、TPU 等）处理海量数据，并执行复杂的数学计算任务，从而高效学习和生成内容。

算力的强弱直接决定了 AI 能学习多快、能处理多少数据。算力不足的 AI 就像一个学习能力较弱的学生，需要更长时间才能理解同样的知识。

3. 算法——AI 的“学习方法”

学生在学习过程中，不仅需要书本和大脑，还需要合适的学习方

法。有些学生擅长通过做题掌握数学概念，而有些学生则擅长通过讲解或者动手实践来理解。类似地，AI 也需要有效的“学习方法”，这就是算法。

学生学习新知识时，会使用不同的方法，如机械记忆、归纳总结、类比推理。AI 则通过不同的机器学习算法（如监督学习、无监督学习、强化学习）归纳、理解、预测并优化结果。

ChatGPT 这样的 AI 模型使用的是深度学习算法，特别是 Transformer 结构，它通过“自注意力机制”学习文本中的上下文关系，从而更好地理解人类语言。

如果一个学生学习方法得当，就可以事半功倍，快速掌握知识并灵活运用。同样，如果 AI 采用高效的算法（如 Transformer），就可以在更短的时间内训练出更优秀的模型，理解更复杂的语言模式。

1.5 AGI 前夜：语言模型开启的通用智能之门

上文我们回顾了 AI 发展史上几个让世界震惊的标志性时刻。

尽管深蓝和阿尔法狗都是 AI 发展史上的里程碑，但 ChatGPT 等大语言模型（如 GPT-4、Claude、Gemini 等）的出现，才是真正将 AI 推向通用人工智能（artificial general intelligence，AGI）的关键一步。为什么这么说？

1.5.1 深蓝和阿尔法狗的局限性——狭义 AI 的胜利

深蓝和阿尔法狗的胜利本质上是狭义人工智能（narrow AI）的胜利，它们虽然能在特定任务上超越人类，但无法通用于其他任务。

1. 深蓝：暴力计算的巅峰

深蓝的核心是穷举搜索 + 规则计算，它的工作方式是：通过强大的计算能力，在短时间内分析尽可能多的棋步组合；使用人类棋谱数据和评分函数选择最佳策略。

深蓝的成功在于计算能力的爆炸式提升，而不是“智能”本身。如果换个棋盘游戏——比如围棋，深蓝就完全无法适应了。

2. 阿尔法狗：深度学习 + 强化学习的突破

阿尔法狗引入了深度神经网络和蒙特卡洛树搜索（MCTS），这是比深蓝更接近人类的学习方式：通过自我对弈进行强化学习，而不是完全依赖人类棋谱；它能创造性地下棋，比如李世石在 2016 年与阿尔法狗对弈时的“神之一手”背后的策略，阿尔法狗已经提前发现。

尽管阿尔法狗这种 AI 的确非常强大，但是也有缺陷，就是它只会下围棋，让它打麻将就不会了。或者说，标准的围棋棋盘是 19 乘 19 的，如果换一个 15 乘 15 的棋盘，它可能也就“蒙”了，不会玩了。而对于大语言模型来说，它们什么都会，正如维特根斯坦所说：语言即世界。

1.5.2　ChatGPT 等大语言模型的突破：从工具到智能体

相比之下，大语言模型的突破不仅是战胜人类在某项任务上的能力，而是迈向通用智能的一个巨大飞跃。

ChatGPT 的核心能力是将语言作为智能的基石。**ChatGPT 的最大突破是理解和生成自然语言，即能够像人类一样处理文本，涵盖对话、写作、翻译、代码等多种能力。ChatGPT 具有跨领域知识整合能力，它不像阿尔法狗那样只会下围棋，ChatGPT 能同时涉及数学、物理、文学、哲学等多**

个领域，并进行知识融合。ChatGPT 还具有类推理能力，它可以在没有明确规则的情况下进行逻辑推理、推测因果关系，甚至分析人类心理和社会现象。

语言是人类思维的核心，能流畅地使用语言，就意味着 AI 具备了一定的“泛化能力”。与深蓝和阿尔法狗相比，ChatGPT 的适应能力更强，能够解决大量开放性问题，这是向 AGI 迈出的关键一步。ChatGPT 具有可扩展性，它不只是一个模型，更是一个平台，ChatGPT 模型可以通过 API 接入各种系统，适应不同任务。

代码助手（如 Copilot）：可以编写和优化代码。

医疗 AI：可以辅助医生进行疾病诊断和医学研究。

教育 AI：能为学生提供个性化辅导。

这种通用性和可扩展性远超以往任何 AI 模型，使得大语言模型成为未来 AGI 的核心技术之一。

为什么大语言模型代表通向 AGI 的关键一步？因为大模型具备“类通用智能”的特性。与深蓝和阿尔法狗的单一任务能力相比，ChatGPT 展现出多个接近通用智能的特征。

（1）语言理解（能阅读并推理）。

（2）知识整合（可结合不同领域的信息进行回答）。

（3）上下文学习（可以根据对话历史调整回应）。

（4）多模态扩展（可以理解文本、图片、音频，未来还可能扩展到视频）。

这些特性让它更接近人类的学习方式，而不仅仅是一个“高级计算工具”。它从被动执行到主动推理，过去的 AI（如深蓝和阿尔法狗）都是被动的，必须在人类设定的规则下执行任务。而 ChatGPT 等大语言模型则可以做到以下 3 点。

（1）主动提出问题，而不仅仅是回答问题。

（2）基于不完整信息推理，类似人类的逻辑思考。

（3）理解隐含意图，进行社会化交互，如安慰用户、表达幽默感。

尽管 ChatGPT 已经展现出强大的能力，但要真正实现 AGI，还需要突破以下关键问题。

（1）长期记忆能力：目前的 GPT 主要依赖短期上下文窗口，无法真正“记住”长时间的历史。

（2）真正的因果推理：当前的大语言模型仍然主要依赖概率预测，而不是基于因果关系进行推理。

（3）自主学习和目标设定：AGI 需要自主设定目标，而不是被动响应人类的输入。

如果 ChatGPT 继续演化，将会发生什么事情呢？

（1）智能助理全面普及：AGI 助手有助于管理个人事务、进行学习规划，甚至能与人类共创作品。

（2）全自动科研：AI 能够自主提出科学假设，设计实验，并进行创新发现。

（3）人机共生社会：人类与 AGI 深度合作，实现科技飞跃。

ChatGPT 是通向 AGI 的里程碑。从深蓝到阿尔法狗，我们看到 AI 在狭义智能上的突破，而 ChatGPT 代表的是迈向通用智能的关键一步。它不仅具备语言理解和跨领域知识整合的能力，还展现初步的推理和适应性，这让 AI 不再只是工具，而是一个真正的智能体。

未来，随着更强大的模型、记忆系统、因果推理能力的加入，ChatGPT 将成为真正的 AGI 前身，推动人类进入一个人机共生的新时代。

1.5.3 大语言模型可能存在问题，但依然是通向 AGI 之路

大语言模型，不管是 GPT、DeepSeek、Grok 还是豆包，它们的核心本质都是概率分布的模拟器，而不是知识库。它们不是在搜索验证已有的内容，而是在生成新的内容，生成的内容是通过预测最有可能出现的文本组合。

这句话可能不太好理解。举个例子，AI 知道“冰激凌在太阳下会融化”并不是因为它理解了热力学定律，而是因为在它学习的海量文本里，“冰激凌”和“太阳”共同出现时，“融化”这个词出现的概率远远大于“凝固”，所以它会输出：冰激凌在太阳下会融化。

AI 在处理常识性问题的时候，基于统计学逻辑得出的答案一点儿问题都没有。**但在处理一些有深度的复杂问题的时候，因为训练数据不足，它就会自作聪明地去强行“完形填空”，生成一些看似合理实际上非常错误的结论，就像虚构一些学术论文。**

比如，我讲课时需要介绍量子计算机，就会让 AI 帮我提供量子计算领域的最新参考文献。结果，它给我一个《量子纠缠在药物研发中的应用》（作者：John Smith, 2024 年在 *Nature* 刊载），但实际上 *Nature* 中并不存在该论文。你问 AI：“2023 年诺贝尔物理学奖得主是谁？”AI 可能会告诉你：“2023 年诺贝尔物理学奖由约翰·史密斯博士获得，他因在量子隐形传态领域的突破性贡献而获奖。”

这听起来像模像样，甚至有具体的人名和研究方向，但事实是——诺贝尔物理学奖获得者中根本没有这个人！

这种自信满满的造假能力正是 AI 幻觉的经典表现。为什么会发生这种事呢？很重要的一点是，**大语言模型跟人脑不一样，他没有自我觉知的**

能力，也就是原认知的能力。它不知道自己不知道，就会导致当问题超出它的能力边界的时候，它会“张口就来”，一定要给出一个答案——哪怕这个答案是错的。这是因为从大语言模型的第一性来说，它是要“接话”，对错不重要，接住你的话才重要，这其实挺危险的。更有意思的是，如果它回答的错误内容被你指出之后，它不会觉得有什么问题，只会说“抱歉”，然后夸你“厉害”“发现了我的错误”。所以，我们在使用 AI 的时候，只有理解了它的原理，知道了它的这些弱点，才能够更好地让它帮助我们进行内容生产。

那 AI 的正确打开方式是什么样的呢？怎么避免 AI 出现幻觉呢？

第一是优化提问方式。这就是我们通常说的设置提示词，要尽量避免提出模糊不清的问题。比如，不要问：“人类什么时候能实现永生？”而要问：“请基于近 5 年 *Science*、*Nature* 发表的寿命延长相关的论文和全球百岁老人数据库的人口统计模型，量化分析当前技术下人类寿命延长的理论极限，排除伦理争议和资金限制因素。”

第二是交叉验证。你在询问一些重要的问题如法律条文、学术结论的时候，不要只依赖 AI 的回答，一定要通过多方信任交叉，验证一下这个说法到底是不是真的，或者通过互联网搜索去找信息的源头。

第三是分批输出。如果你要处理长文本内容，可以分几次发给它，不要一次性发给它，因为它有可能会忘记你前面发给它的东西是什么，导致后面乱输出。

第四是学会选择模型。当你提的问题是跟创意相关的，那么建议用基础大语言模型 GPT-4o 或者 DeepSeek V3。如果你提的问题是与事实、逻辑、推理相关的，建议用推理模型 Open AI O1 或者 DeepSeek R1，因为 AI 模型也跟人类大脑一样，有两套系统——系统 1 快思维和系统 2 慢

思维，快思维偏直觉，它会很快地给出一个答案；但是对于某些问题，它可能会失误。而慢思维思考的时间会更长，消耗更多电力和算力，但是它可能会给出一个正确的答案。这好比我们问“9.11 和 9.9，哪个更大”，基于快思维的基础模型都有可能答错，但基于慢思维的推理模型答案都是对的（具体讲解参见 5.1 节）。

目前，这些高级模型的出现只是冰山的一角，未来大家可能会看到越来越多类似这样的技术：它不仅能够懂你说什么，还懂你心里想什么。今后，人们可能会花更多时间跟 AI 聊天，而不是跟真人聊天。这既令人兴奋，又有一点让人担心。令我们兴奋的是，科技正在以前所未有的方式贴近人性；而让我们不安的是，我们可能还没有准备好去面对这种高智商、高情商的 AI。也许，这种高智商、高情商 AI 升级到极致的时候，才是真正的 AGI 时刻的来临。