浅谈大模型及其在高能物理科学的未来应用

张正德

（中国科学院高能物理研究所 100049）

人工智能大模型是什么？它和我们通常讲的机器学习、深度学习有什么关系？它有什么能力？它在高能物理可能有哪些方面的应用？今天我们浅浅讨论一下这些问题。

一、溯源：从人工智能到机器学习、深度学习和大模型

1. 曲折发展的人工智能

人工智能(Artificial Intelligence, AI)是一个模拟、延伸和扩展人的智能的理论、方法、技术及应用的技术科学，其本质是对人的意识和思维的模拟。为了实现这一目标，20世纪50年代人工智能诞生之初，就出现了两种不同的思路。一种认为人类思维的很大一部分是按照推理和猜想规则对“词(word)”进行操作所组成的，因此提出了基于知识与经验的推理模型，即知识驱动的符号主义人工智能；另一种认为感官的刺激不存储在记忆中，而是在神经网络中建立起“刺激到响应”的连接，通过这个连接保证智能行为的产生，即数据驱动的连接主义人工智能。两种思路分别于1955年和1956年被提出，当时人们觉得人工智能会在20年内改变世界，所有的工作将会被人工智能颠覆，人工智能迎来第一次“春天”。然而1973年《莱特希尔报告》明确指出当时的人工智能的任何部分都没有达到人们想象的水平，第一次“春天”随之结束。

1980年卡内基梅隆大学采用“知识库+推理机”的组合为数字设备公司设计了一套名为XCON的专家系统，取得了巨大成功，符号主义人工智能热度达到巅峰，人工智能迎来第二次“春天”，然而7年之后苹果和IBM生产的台式机性能超过计算机专家系统，人工智能再次陷入低谷。

困难时期，依旧有科学家坚持研究，研究重心逐步从符号主义转移到连接主义上，目前当代人工智能的重要技术如卷积神经网络、深度学习模型等都是这一时期的成果。2011年IBM的人工智能程

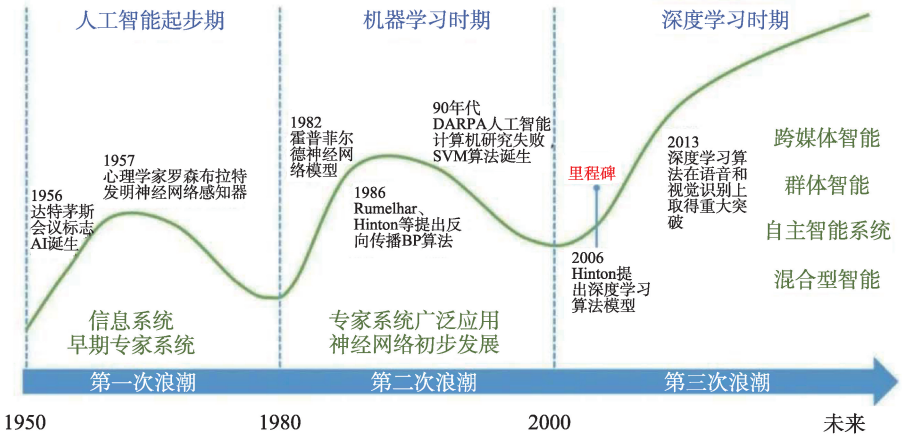

图1 人工智能的三次浪潮

序“沃森”参加智能问答战胜2位世界冠军，人工智能逐步迎来第三次“春天”。2013年，深度学习在语音和视觉识别任务上取得重大突破；2016年，DeepMind的人工智能围棋程序AlphaGo战胜世界冠军李世石；2020年AlphaFold和2022年ChatGPT的出现持续将人工智能的浪潮推高。目前我们所讲的当代人工智能主要是基于连接主义的数据驱动的深度学习算法。

2. 从机器学习和深度学习

机器学习既包含符号推理又包含连接主义，它强调让机器自动“学习”，是人工智能的具体实现方法。经典的机器学习算法包括K近邻、线性回归、朴素贝叶斯、决策树与随机森林、支持向量机和人工神经网络等，这些经典的方法在20世纪90年代就已经在高能物理领域逐步被引入和推广，时至今日仍然发挥着重要作用。

其中，人工神经网络是受大脑神经元中突触、轴突等结构启发而设计的计算模型。神经网络架构在不断发展，最初是把全部神经元逐层连接起来的全连接神经网络，但它容易过拟合且推理速度慢，后来逐步发展出能进行局部连接的卷积神经网络，卷积时只有部分神经元被激活从而减少计算量；卷积神经网络不能处理时间序列数据，后来发展出了能记忆上个时刻状态的循环神经网络；另外，为了处理包含复杂拓扑关系的图(Graph)数据，发展出了能处理任意尺寸和拓扑逻辑结构的图神经网络；神经网络训练需要用人工标注的真值(输入数据所对应的输出)来促使网络学习，为了省去耗时耗力的标注过程，发展出了无需真值的自监督学习对抗生成式神经网络；对抗神经网络训练不容易收敛，后来发展出了基于扩散原理的生成式模型Diffusion Model。

深层的神经网络容易梯度爆炸或消失从而训练失败，2015年残差神经网络(ResNet)通过在不同的层和神经元间添加信息传递捷径，有效地解决了该问题，使得更深的神经网络能被训练，现在的深度神经网络几乎都包含残差结构。基于深度神经网络的机器学习方法被称为深度学习，人工智能的第三次“春天”是以深度学习为代表的技术革命。

3. 大模型

大模型与深度学习一脉相承，它基于自注意网络Transformer。2017年Transformer出现以后，谷歌的BERT(Bidirectional Encoder Representations from Transformers)模型和OpenAI的GPT(Generative Pre-trained Transformer)成为大模型的重要里程碑，ChatGPT就是基于GPT-3.5的对话生成式模型。

ChatGPT凭借其超过预期的通用意图理解能力、强大连续对话能力、智能交互修正能力和较强

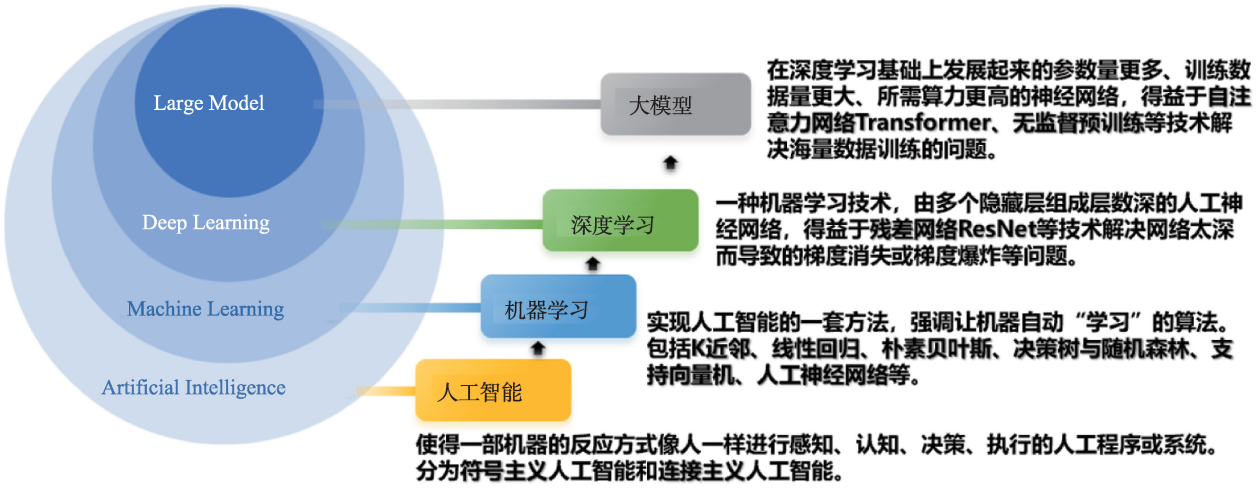

图2　人工智能、机器学习、深度学习和大模型的层次关系

逻辑推理能力，在发布后迅速出圈，带来了生成式人工智能的浪潮。目前比较有名的大语言模型有：OpenAI 的 ChatGPT 和 GPT-4、Meta 的 LLaMA 和 LLaMA2、百度的文心一言、阿里的通义千问、清华的 GLM、百川智能的百川大模型、自动化所的紫东太初、计算所的百聆大模型。我们基于 LLaMA 系列的模型，发展了相关技术，使用高能物理领域相关数据全量微调了“溪悟”大模型。在计算机视觉领域，Meta 也发布了能用于不限种类的图像分割的大模型 SAM。

二、漫谈：大模型的基本原理、涌现

1. 大模型和 Transformer

大语言模型(Large Language Model, LLM)是大模型的代表，其通常是指参数数量在数十亿或更多数量级的深度学习模型。参数是指神经网络的可训练变量，例如一个线性神经元的输入为 x，输出为 $y=wx+b$ 时，w 称为权重，b 称为偏置，权重和偏置统称参数。这些参数在初始化时被随机分配，在训练过程中逐步更新，神经网络的训练过程实际是参数更新的过程。之所以参数量需要达到十亿，是因为十亿参数是大模型开始在某些任务上出现能力“涌现”的最小规模。相比之下，单个任务专用的深度学习“小模型”的参数量大约为十万到千万。

大模型的核心要素是基于 Transformer 的算法、海量训练数据和相应的算力。与之前介绍的神经网络架构不同的是，Transformer 可将模型堆叠得很大依然能进行有效学习。

Transformer 的核心是自注意力机制。注意力机制是指人类会选择性地只关注一部分信息，忽略其他可见的信息。例如人类在判别图片是否有猫时，给予猫所在的像素更多的“注意”，判断会更加准确。人们将这种注意力机制引入神经网络中，发展出空间注意力、通道注意力等许多种注意力机制。自注意力机制是集大成者，它允许模型从不同位置的输入序列中自动捕获依赖关系，从而拥有更强大的表达能力、可扩展性和灵活性。简单来讲，输入序列 X 为“早上好”时，每个字都会被表示为一个向量，自注意力机制会计算“早”与“早”、“早”与“上”、“早”与“好”、“上”与“好”等的向量内积，内积在几何意义上表征投影，投影值为 0 表示两个向量正交无关，投影值越大则关联度越高，关联度表示了需要给予的“注意”程度。Transformer 通过可以训练的权重矩阵 W 来自动学习所需分配的注意力，例如当关注“早”时，需分配 0.4 的注意力给它本身，剩下 0.4 关注“上”，0.2 关注“好”。这种内部“自己注意自己”的方式可以应用于任何特征向量，使得我们可以通过多层堆叠的方式构建很大的模型。此外，多头注意力机制增强了模型关注多个不同信息来源的能力。相比循环神经网络，自注意力机制可以独立计算每个位置的权重，可以实现大规模的并行，大大减少计算时间。

2. ChatGPT 的基本原理

如图 3 所示，ChatGPT 的本质是能“预测下一个词”的“词语接龙”模型。根据资料，实现 ChatGPT 有四个步骤：

(1) 基于 Transformer 构建大模型框架，Transformer 的自注意力机制能自动学习输入序列的相互关系、注意到不同向量的重要程度，从而具有更强大的表达能力和灵活性。

(2) 采用“预测下一个词的”方法预训练大模型得到基础模型。训练时，先将文本编码为 Token，Token 是文本中的一个基本单位，可以是一个单词、词组、标点符号、字符等。例如：文本“Please introduce the Institute of High Energy Physics.”被编码为 10 个 Tokens，图中不同的颜色代表不同的 Token，注意到 Institute 一个单词被编码为 2 个 Tokens，符号“.”占 1 个 Token。训练时对整段文本的 Tokens 进行截断以符合模型的输入限制，然后作为上文输入到模型中，将下一个将要出现的 Token 作为真值，让 GPT 学习预测该 Token，预测后再将该 Token 合并到上文中输入给模型，再将下下个 Token 作为真值让

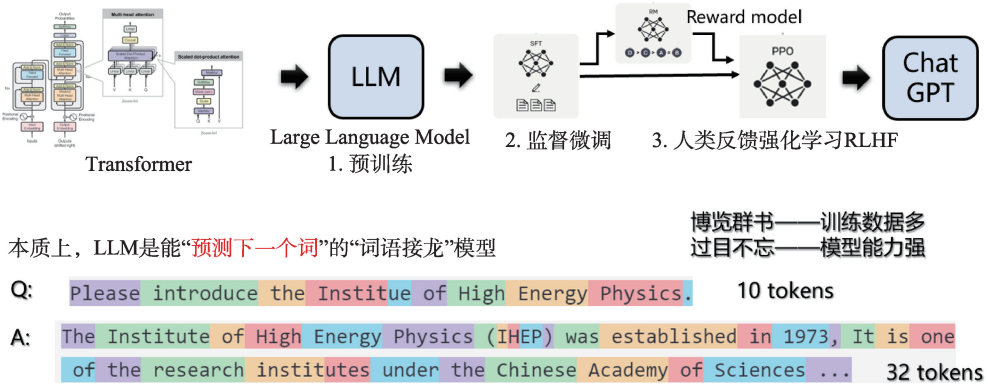

图3 ChatGPT的实现原理

GPT学习，以此类推。因为真值存在于输入数据中，因此预训练阶段不需要标注数据，使得利用海量无标签数据训练模型成为可能。推理过程中，GPT根据问题编码的Tokens，来预测词库中50257个Tokens哪些出现的概率大，从候选Tokens中根据预测概率随机选择一个作为预测的“下一个词”，合并到输入中再次预测下一个词，以此类推，直到预测到停止符号。因为回答的每一个词都是从词库中生成的，因此GPT是一种生成式模型。

(3) 对基础模型进行监督微调得到微调模型。组织人工撰写问题和答案，获得监督微调数据集，对基础模型进行监督微调，让GPT更多地去学习人们更想让它学习的知识。

(4) 采用人类反馈强化学习进一步对模型进行微调得到ChatGPT。使用不同GPT对同一个问题输出多个答案，由人类来标注答案的得分，形成数据集，去训练新的打分模型，打分模型模拟了人类对答案的偏好。获得打分模型后，采用强化学习方法，让打分模型评价GPT输出答案的好坏并反馈给GPT让它不断进化，最终得到ChatGPT。

3. GPT的性能和能力“涌现”

ChatGPT表现出通用的意图理解能力、强大的连续对话能力、智能的交互修正能力和较强的逻辑推理能力，这主要得益于当模型的参数大到一定程度时，模型会在复杂任务上突变式地拥有了小模型不具备的能力，称为大模型的能力“涌现”。

如图4(a)所示，随着模型规模的增加，模型在维基百科问答、日期理解和单位转换等简单任务上表现为接近线性的关系，即规模越大，性能越强；能力

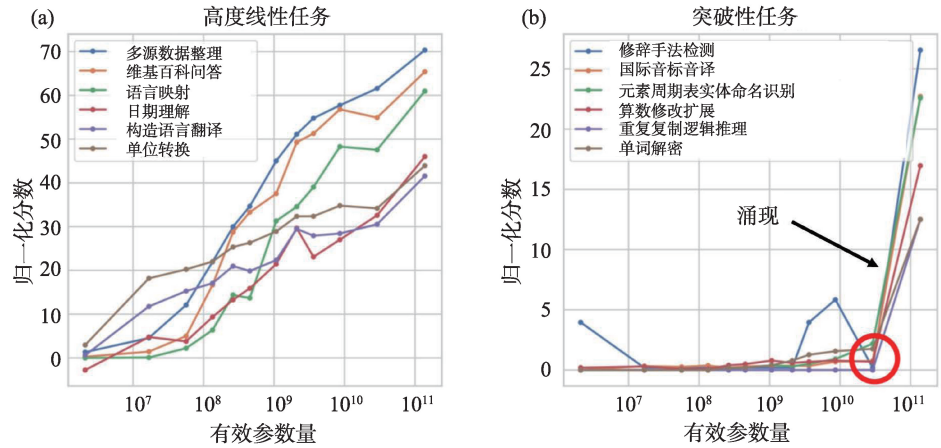

图4 大模型的能力涌现

涌现则如图4(b)所示，当模型规模比较小时，在非字面意义检测、重复复制逻辑、单词解谜等复杂任务上模型得分始终为0，当模型参数上升到100亿到1000亿个时，观察到模型突然有一定的准确率了。系统定量上的变化导致系统行为上的定性变化——即涌现。ChatGPT涌现出的能力包括文本内学习(In-Context Learning)和思维链(Chain-of-Thought)等。文本内学习允许我们不改变模型权重(不训练模型)的情况下，只需要给出样例、上下文，模型就能更准确地给出答案。思维链则是当我们提示GPT需要“一步一步”思考的时候，模型会给出更加准确的推理结果。

为什么只是简单的算法组件——梯度下降、大规模的Transformer和海量的数据，就能展现出如此通用且灵活的智能？目前并不清楚。一种假设是海量数据(特别是内容差异巨大的数据)迫使神经网络学习通用且有用的“神经回路”，大规模的模型提供了足够冗余和多样性使得神经回路对特定任务进行专门化和微调。另一种想法是大规模的模型带来一些益处，包括通过连接不同的最小值使梯度下降更加有效，或简单地使得高维数据的拟合更加平滑。研究大模型的“涌现”现象是一个重要的方向，这种涌现现象是先前任务专用的模型所没有的。

三、大模型在高能物理领域的应用

1. 直接应用

大语言模型可以直接应用到写代码、改文档、文献阅读等日常科研工作中，但使用中需要注意学术道德规范，可通过高能物理人工智能平台(https://ai.ihep.ac.cn)访问问答机器人。

利用大模型的强大的泛化能力，可以极大地加速专用AI模型的研发过程，例如“一站式”天文警报信息汇集平台和X射线增材制造缺陷智能分析两个案例。

(1) “一站式”天文警报信息汇集平台

搭载在卫星上的天文爆发探测器多种多样，探测数据均为文本但格式不一，存在平台多、信息孤立、零散等问题，“一站式”的天文警报汇集平台可将不同的爆发信息发布为统一的格式，便于后续的多卫星联合观测和数据分析。将非结构化的数据整理为结构化数据的多源信息整合技术是关键。采用正则表达式性能不够好，采用神经网络需要先人工将原始数据进行标注，才能训练出有效的模型。我们利用ChatGPT的API接口，在进行提示工程后设计专门用于处理该任务的智能体，可以将原始数据快速的转换为结构化数据，得到AI-Ready的数据集，利用该数据集去训练自己的神经网络，实现快速、准确的天文信息汇集。ChatGPT大模型极大地加速该神经网络的研发过程。

(2) X射线增材制造缺陷智能分析

X射线增材制造是一种利用X射线激光3D打印进行高端金属材料制造的先进技术，制造过程中气泡、熔池等微观结构对材料性能有显著的影响。HEPS张兵兵课题组研发了X射线增材制造原位试验装备，可以实时快速对增材制造过程的微观结构进行成像。如果能从图像中检测气泡数量、熔池情况，并实时反馈给增材制造装置，自动控制送粉量和激光强度，来使得缺陷分布更均匀，理论上能制造出的性能更好的材料。因此，复杂、低分辨率图像下的缺陷实时检测、跟踪成为急需解决的关键技术。神经网络中有较为成熟的检测、跟踪算法模型，但为了训练该算法，需要先人工标注气泡、熔池等缺陷，每张图的缺陷可达几百个，用传统的标注软件也是非常耗时耗力的活儿。我们使用计算机视觉分割大模型SAM，进行简单的提示工程后，就能给出不错的分割结果，进一步处理后得到真值，形成AI-Ready的数据集。SAM大模型处理每张图大约需要20秒，离实时非常远，利用该数据集训练小的、能实现实时的任务专用模型。因此，SAM大模型也极大地加速该神经网络的研发过程。

没有大量数据是否应该应用大模型？我们认为更应该使用。训练大模型需要海量数据，但是训练好的大模型拥有小模型不可比拟的泛化能力，更

图5　大模型在高能物理的两个应用案例

加能适应小数据的应用场景。

2. 高能物理领域定制化的文本大模型

ChatGPT不开源，使用ChatGPT也存在重要数据泄露的风险。为了解决算法自有化的问题，我们基于开源的LLaMA系列大语言模型Vicuna，收集了的高能物理领域文本数据，训练了领域定制的“溪悟(Xiwu)”大模型(70亿和130亿)。通过采用量化、Flash-Attention、FSDP全分片数据并行、LoRA低秩自适应等技术，实现了3轮全量训练。

我们提出了种子裂变技术来获得高能物理等领域的问答数据，此外，也通过清洗问答机器人后台收集的数据、从文献中提取信息等途径获得了更多训练数据。

种子裂变技术如图6所示，我们基于大语言模型设计了3个智能体：新手智能体，检查器和专家智能体。当我们以“高能物理”四个字作为种子时，新手智能体会提出2个问题：高能物理研究的是什么？高能物理学家使用什么来探索宇宙的基本结构和演化？检查器判断哪些问题与已有问题重复

图6　种子裂变技术

并筛选出1个感兴趣的问题询问专家，专家智能体给出答案，并将答案输入到新手智能体中，新手智能体再根据答案文本提2个问题，以此类推。结果证明以“高能物理”四个字作为种子，能裂变出与初始种子相关的有深度、多样化的问答对数据集，例如：费米子具有什么样的自旋?引力波是如何产生的?红移可以用来估算什么?什么是强相互作用?如何检测暗物质?夸克有几种“味道”，分别是什么?弱相互作用在核反应中扮演了什么样的角色?为什么中子星密度非常高?等等及它们的答案。

经过微调训练，Xiwu-130亿语言模型在100个高能物理领域问答测试集上，采用人工评估的方法，与基准模型Vicuna-130亿相比，回答更加准确或持平的概率达到95%，性能明显优于基准模型，证明该训练方式能有效嵌入领域知识。与ChatGPT-1750亿相比，性能达到了约65%。

在未来，大模型还可进行一系列能力扩展，例如：循环记忆Transformer突破大模型Tokens长度几千个的限制、思维树提升大模型的思维能力、过程监督提升大模型的数学能力、与计算机视觉神经网络融合赋予大模型处理图像的能力等。

3. 高能物理科学数据大模型

目前产业界的大模型主要处理文本、图像、音频等模态的数据，而高能物理领域积累的数十PB级数据大都是科学数据，其特点是带有物理意义的浮点数表示的数据。发展高能物理科学数据大模型的基本思路是利用大模型无监督预训练方法让AI把握所有数据中的全局规律，用物理反馈强化学习引导其涌现。在下游任务如Jet Tagging, Shower simulation微调，在更多复杂下游任务实现涌现，获取数据、调用工具，建立反馈回路，不断进化。

部分科学数据能转化为文本或图像，但数值型的科学数据无法使用现有的“预测下一个词”、蒙版自动编码(Mask Auto Encoding, MAE)等技术实现预训练。一种思路是将科学数据转换为矢量序列，通过“预测下一个矢量”实现预训练，例如矢量量化变分自动编码(Vector Quantized Variational Auto-Encoding, VQ-VAE)，另一种思路是使用图神经网络中的图(Graph)与Transformer的组合实现预

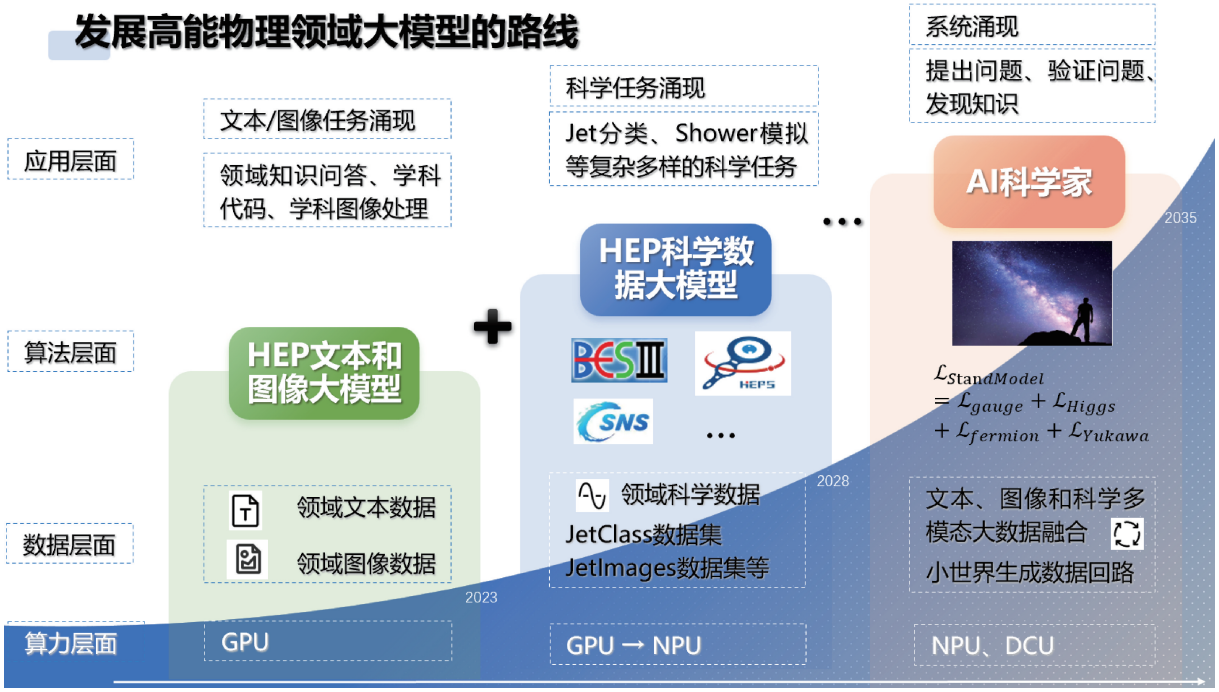

图7 发展高能物理领域大模型的路线

训练。总之，一种能有效进行预训练科学数据的方法Tokenizer是技术瓶颈之一。此外，用于科学研究的大模型，还需要考虑置信度刻度不对齐导致的置信度误差增大问题，以及为了让模型与物理原理对齐如何实现物理反馈强化学习的问题等。

4. 大模型用于科学发现的探索

从发展大模型的角度出发，训练高能物理领域的大模型需要海量数据和大量算力，还需要瓶颈技术的突破。在这些条件还不具备的情况下，从应用大模型的角度出发，我们正在探索另一条蹊径——不训练(或少量训练)的情况下是否有可能将大模型用于高能物理科学发现。

我们正在基于通用大模型，打造“赛博士”(Dr. Sai)科研智能体，以在物理分析中重新发现Zc(3900)粒子(2013年物理学领域重要成果榜首)为抓手，逐步实现文献调研、程序编写、数据处理、物理分析、结果解释和论文撰写等关键步骤，实现高度智能化的科研助手，将科研人员从创新要求较低的例行研究中解放出来，提升科研产出能力。

目前我们进行了智能体所需能力的分解，并调研和分析了所需的技术，其中文献解析技术、数据清洗和标注技术、外部知识库加载技术已经初步运行成功且证明有效。

如果赛博士能发现Zc(3900)粒子，那根据新的科学目标，是否有可能发现其他还未发现的粒子呢?中长期的目标是希望实现虚拟科学家，能掌握物理的基本原理、从数据中发现规律、与人类科学家讨论和开展研究。我们非常欢迎对此感兴趣的老师同学一同参与，共同推动这一充满潜力的工作。

四、总结：仰望星空、脚踏实地

我们从人工智能之初，谈到机器学习、深度学习，再谈到大模型。也讨论了大模型的基本原理、大模型的涌现现象。初步进行了一些将大模型应用在高能物理领域的尝试，未来还有很多进步的空间。

人工智能现在炙手可热，从历史来看它已经三起两落，目前的“火热”有很大一部分是人们的想象。我们非常赞同鄂维南院士的观点：“AI For Science是整个中国科技创新史上最好的机会之一，但是一定要避免炒概念，搞表面繁荣，不落地”。

当代人工智能虽然正在推动新一代的科研范式变革，但智能范式与之前的经验范式、理论范式、模拟范式和数据范式之间不是取代关系，需要共同发挥作用。深度学习带来了人工智能的重要突破，但经典科学计算分析方法、大规模并行技术、经典机器学习方法依然能发挥重要作用。大模型的出现虽然极大地提升了模型泛化能力，但小模型也拥有大模型不具备的速度优势。

我们应当谨而慎行，既需要“仰望星空”，畅想大模型甚至通用人工智能有可能带来的革命性变化，也应当“脚踏实地”，真正将技术落到实处，实现价值，这大概也是“立足常规、着眼新奇”的寓意所在。

致谢：

本文的内容得到了曹俊、陈刚、方亚泉、李刚、齐法制、王文帅、熊少林、苑长征、张兵兵、张红梅、张一、张易于以及赵丽娜等各位老师的协助。他们提供了宝贵的需求、算力的支持、方案的建议，以及深入的讨论和帮助。由于篇幅所限，无法一一列举，深表歉意和感谢。