大型语言模型中的涌现能力：机制、表现与规模的作用

第一部分：涌现现象：一个备受争议的定义

大型语言模型（LLM）的兴起不仅带来了性能上的飞跃，还引入了一个引人入胜且充满争议的概念：涌现能力（Emergent Abilities）。这一概念试图解释为何当模型规模达到某一临界点时，会突然获得之前在较小规模模型中完全不存在的全新能力。然而，这些能力究竟是人工智能发展到新阶段的真实标志，还是我们评估方法造成的假象，已成为当前人工智能研究领域的核心辩题。本部分旨在深入探讨涌现能力的定义，剖析其核心争议，并梳理学界为寻求更精确、更具机械论色彩的定义所做的努力。

1.1 涌现的定义：从“量变到质变”到不可预测的性能

“涌现”这一概念源于复杂性科学和物理学，其思想根源可追溯至诺贝尔物理学奖得主菲利普·安德森（P.W. Anderson）于1972年发表的著名文章《多者异也》（More Is Different）[1]。其核心思想是，当一个系统的复杂性（或规模）增加时，可能会出现无法通过分析其微观组成部分的性质来预测的全新宏观属性和行为 [1]。这正是“量变引起质变”的体现 [3]。

在大型语言模型的背景下，研究人员借用了这个概念来描述一个特定的现象。由Wei等人（2022年）提出的开创性定义是：大型语言模型的涌现能力，是指那些在较小规模模型中不存在，但在大规模模型中存在的能力 [3]。这个定义有两个关键特征，使其显得尤为引人注目：

1. 不可预测性（**Unpredictability**）：一项能力被认为是涌现的，前提是它“无法通过简单地外推小规模模型上的性能改进来预测”[1]。换言之，当我们观察小模型在某任务上的性能曲线时，它可能一直维持在随机猜测水平，我们无法据此预见到，一旦模型规模跨过某个阈值，性能会突然大幅提升。
2. 突变性（**Sharpness**）：这些能力的出现往往表现为一种突兀的、非线性的飞跃。在模型规模达到某个临界点之前，其在特定任务上的表现可能长期停滞在接近随机的水平；而一旦越过该临界点，性能便会“仿佛拨动了开关一样”急剧攀升至远超随机的水平 [2]。一个被反复引用的典型例子是GPT-3在多位数算术（如三位数加法）上的表现。研究

发现，参数量从1亿到130亿的模型，其多位数加法的准确率几乎为零，然而当参数量达到1750亿时，其准确率突然跃升至80% [1]。这种现象促使研究人员思考，为何会获得这些能力，以及进一步扩大规模是否会带来更多的涌现能力 [1]。

最初，研究界将涌现现象与模型的"规模"（scale）紧密联系在一起，而规模可以通过多种方式衡量，包括训练所用的计算量（以浮点运算次数FLOPs为单位）、模型参数的数量，或训练数据集的大小 [3]。这一框架为观察和讨论该现象奠定了初步的基础。

1.2 涌现的**"海市蜃楼"**：对评估方法的批判

正当涌现能力的概念激发了广泛的研究热情时，一个强有力的反驳声音出现了。以Schaeffer、Miranda和Koyejo为代表的研究团队提出了一项深刻的批判，其核心论点是：所谓的涌现能力并非模型规模扩展的内在基本属性，而是一种**"海市蜃楼"**（**mirage**），主要由研究者选择的评估指标所导致 [2]。

他们认为，这种"假象"的产生机制如下：

1. 非线性或不连续的评估指标：许多被声称出现涌现能力的基准测试，都采用了非线性或不连续的评估指标，如"准确率"（Accuracy）或"精确字符串匹配"（Exact String Match）[2]。这些指标是"全有或全无"的。例如，在执行一个需要生成一长串token的算术任务时，即使模型已经能正确预测99%的token，只要有一个token出错，Accuracy指标的得分依旧为0。模型在每个token上的预测概率可能是随着规模平滑提升的，但这种严苛的指标会掩盖这种渐进的进步，直到所有token都正确时才突然从0跳到1，从而制造出性能急剧跃升的假象 [2]。
2. 由指标引发的涌现：Schaeffer等人通过实验证明，只要更换评估指标，所谓的涌现现象就会消失。他们以算术任务为例，当使用Accuracy作为指标时，模型性能随规模增长呈现出突变；但当切换到一个连续的、允许"部分给分"的指标，如"Token编辑距离"（Token Edit Distance）时，性能的提升曲线就变得平滑、连续且可预测 [2]。他们甚至更进一步，通过在计算机视觉任务中故意设计不连续的指标，成功地在卷积神经网络中"诱导"出了前所未见的涌现能力，这有力地证明了涌现可以是评估方法的人为产物 [2]。
3. 统计数据不足：造成假象的第二个次要原因是测试数据集规模太小。当测试样本不足时，小模型本已很低的、但非零的性能可能因为统计误差而被误判为零，这进一步强化了能力在达到某个规模后"从无到有"的印象 [2]。

这项批判性研究通过对著名的BIG-Bench基准进行元分析，发现其超过92%的所谓涌现能力都出现在两个不连续指标之下："多项选择评分"（Multiple Choice Grade）和"精确字符串匹配"[6]。这一发现为"海市蜃楼"假说提供了强有力的证据。需要强调的是，这一批判并非否认大模型能力的真实性，而是指出其"涌现"的突变性和不可预测性可能是一种错觉 [13]。

模型确实在变强，但这种变强是渐进的，而非魔法般的突现。

1.3 调和观点：走向一种机制性的定义

随着研究的深入，学术界开始超越“真实”与“虚假”的二元对立，尝试提出更为精妙的定义，将涌现与模型的内部状态而非仅仅是外部行为联系起来。

1. 基于预训练损失的涌现：一种有影响力的观点主张，应将涌现能力重新定义为“在预训练损失较低的模型中表现出来的能力”[15]。研究表明，具有相同预训练损失的模型，无论其参数大小或训练数据量如何，在下游任务上都表现出相似的性能 [16]。基于此，一项能力是涌现的，如果其性能在模型的预训练损失降至某个临界阈值之前一直停留在随机猜测水平，而越过该阈值后则开始稳步提升 [16]。这个定义更加稳健，因为它与模型最根本的学习状态（即损失）挂钩，摆脱了对特定评估指标的依赖。重要的是，即使在使用连续指标时，这种基于损失的涌现现象依然存在，这直接对Schaeffer等人的部分结论构成了挑战 [16]。
2. 基于内部表征形成的涌现：一个更深层次、可能也更精确的定义认为，真正的涌现应该被定义为成功完成任务与神经网络内部形成新的、粗粒度表征的结合 [1]。这种观点将焦点从模型“做了什么”转移到“它是如何做到的”。真正的涌现不仅仅是性能的提升，更关键的是模型内部是否形成了新的计算模型、抽象概念或符号回路（symbolic circuits），这些内部结构能够带来可验证的预测、问题解决和泛化效率的提升 [1]。例如，模型是否学会了像人类一样进行逐位计算，形成了内部的离散状态表征（Implicit Discrete State Representations）[18]。这种定义要求我们深入模型的“黑箱”内部，寻找质变的证据。

这场关于定义的演进，反映了人工智能领域从观察现象到探究其根本原因的成熟过程。对“涌现”定义的争论，实际上是对其背后形成机制的争论。

表1：涌现能力的不同定义框架对比

框架/理论	核心定义	关键证据/例子	支持性文献
基于性能的涌现 **(Wei et al.)**	在小模型中不存在，且无法通过外推小模型性能曲线来预测的能力。	在BIG-Bench等基准测试上，模型性能在特定规模上出现急剧、非线性的跃升，如多位数算术。	[3]

指标诱导的海市蜃楼 **(Schaeffer et al.)**	由非线性或不连续的评估指标（如准确率）造成的假象，而非模型行为的根本改变。	将评估指标从“准确率”更换为“Token编辑距离”后，涌现现象消失，性能曲线变得平滑。	[2]
基于损失的涌现	仅在模型的预训练损失低于某个临界阈值后才表现出来的能力。	在MMLU和GSM8K等任务上，模型性能在损失值低于约2.2之前一直处于随机水平，之后稳步提升。	[15]
基于内部表征的涌现	成功完成任务与模型内部形成新的、高效的粗粒度计算结构（如抽象概念、符号回路）的结合。	模型内部形成用于执行特定任务（如算术）的专门神经回路，或发展出抽象的“世界模型”。	[1]

第二部分：涌现能力的具体表现

尽管关于涌现能力的定义存在争议，但学术界已经记录了大量被归类为“涌现”的具体能力。这些能力不仅限于单一任务，而是涵盖了从基础学习范式到复杂推理的广泛领域。本部分将详细梳理这些能力的具体表现形式，它们构成了整个涌现现象讨论的经验基础。

2.1 基础能力：上下文学习（**In-Context Learning**）

上下文学习（In-Context Learning, ICL）可以说是大型语言模型最基础、也最引人注目的涌现能力 [19]。ICL指的是模型能够在不进行任何参数（权重）更新的情况下，仅通过在输入提示（prompt）中提供少量任务示例，就能学会并执行一个新任务 [19]。

这种能力被认为是涌现的，因为它并非模型被直接训练的目标。LLM的基础训练任务通常是“预测下一个词”（next-token prediction），而ICL能力是从这个简单的目标中“隐式地”发展出来的 [23]。较小规模的模型几乎不具备或只具备非常微弱的ICL能力，而当模型规模扩大后，ICL成为其一项强大且通用的工具 [1]。通过在提示中给出几个例子，LLM就能像人类一样通过类比来学习（learn from analogy），并解决全新的问题 [20]。

ICL的应用非常广泛，例如，只需在提示中提供一两个情感分类的例子（如“这个产品太棒了！//正面”），模型就能对新的评论进行情感分析；同样，它也可以通过几个翻译示例来执

行语言翻译，或通过几个代码示例来生成新的代码 [20]。这种无需重新训练就能适应新任务的灵活性，是LLM革命性影响的核心之一。

2.2 高级推理：通过涌现式提示策略解锁

更有趣的是，不仅某些任务能力是涌现的，连一些高级的“提示策略”（prompting strategies）本身也表现出涌现的特性。这些策略在小模型上几乎不起作用，但在大模型上却能“解锁”其潜在的复杂推理能力 [8]。

思维链（**Chain-of-Thought, CoT**）提示：这是最具代表性的例子。CoT的核心思想是引导模型在给出最终答案之前，先生成一系列中间的推理步骤，即“一步一步地思考”[8]。研究发现，对于需要复杂逻辑、算术或常识推理的任务（例如多步数学应用题），使用CoT提示能够显著提升大模型的性能 [8]。

CoT的涌现性体现在，小模型无法有效遵循这种指令，它们生成的“思维链”往往是语无伦次或错误的。只有当模型规模足够大时，它才能真正利用CoT来分解问题、规划步骤并得出正确答案 [8]。模型在没有被明确训练如何进行分步推理的情况下，获得了这种能力，这本身就是一种深刻的涌现 [8]。

BIG-Bench Hard (BBH)的挑战：BBH基准的提出进一步印证了这一点。BBH包含了23个BIG-Bench中对当时模型极具挑战性的任务 [25]。最初，使用标准的少样本提示（few-shot prompting），即便是像PaLM和Codex这样的大模型在这些任务上也表现不佳。然而，当研究人员将CoT提示应用于这些任务时，PaLM在23个任务中的10个上，Codex在17个上，性能都超越了人类评估者的平均水平 [25]。这表明，这些高级推理能力实际上是“潜伏”在模型中的，而CoT这一涌现出的提示策略，成为了解锁这些潜能的钥匙。

2.3 在复杂基准测试上的表现（**BIG-Bench**）

“超越模仿游戏”基准（Beyond the Imitation Game benchmark, BIG-Bench）是一个由学界和业界合作创建的大规模、多样化的评估套件，包含了超过200个旨在探索LLM能力边界的任务 [26]。BIG-Bench在涌现能力的早期发现和表征中扮演了至关重要的角色。

该基准的论文引入了“突破性”（breakthroughness）等指标，用于量化那些在特定规模上性能出现不可预测跳跃的任务 [7]。在BIG-Bench中，大量任务被发现具有涌现特性，这些任务

通常需要模型具备超越简单模式匹配的能力，例如：

- 多步算术：如三位数加法或更复杂的多步运算 [8]。
- 单词解读：例如，根据一串打乱的字母恢复出原始单词 [3]。
- 符号理解：例如，根据一串表情符号（emoji）猜测对应的电影名称 [28]。
- 语言学任务：例如，从国际音标（IPA）进行音译，或识别修辞手法 [3]。
- 知识密集型任务：例如，回答波斯语等低资源语言的问题，或通过大学水平的考试 [3]。

这些任务的共同点是，小模型在上面表现得像是在随机猜测，而大模型则在某个规模点上突然获得了解决问题的能力 [3]。然而，正如第一部分所讨论的，这些观察结果也正是“海市蜃楼”假说批判的焦点。许多最初报告的涌现现象都是在使用

Accuracy或Multiple Choice Grade等不连续指标下观察到的。后续研究表明，当使用允许部分给分的“更平滑”的指标时，这些任务上的性能飞跃有时会消失，转变为更平滑的提升曲线 [7]。

表2：涌现能力的基准测试示例

任务/能力	基准测试	展现涌现能力的模型家族	原始声称中使用的指标	观察到的“突破”规模（参数/FLOPs）	支持性文献
三位数加法	BIG-Bench	GPT-3, LaMDA	Accuracy	GPT-3 (13B), LaMDA (68B)	[8]
单词重组	BIG-Bench	LaMDA, Gopher, GPT-3	Exact String Match	~1022 FLOPs	[3]
电影**Emoji**猜测	BIG-Bench	BIG-G (Google)	Accuracy	~1010 - 1011 有效参数	[28]
国际音标音译	BIG-Bench	LaMDA, Gopher, GPT-3	Exact String Match	~1022 FLOPs	[3]
大学水平考试 **(MMLU)**	MMLU	LaMDA, GPT-3, PaLM	Multiple Choice Grade	~1023 FLOPs	[8]
波斯语问答	BIG-Bench	LaMDA, Gopher, GPT-3	Exact String Match	~1022 FLOPs	[3]

对这些涌现表现的分析揭示了一个核心趋势：最重要的涌现能力并非关于静态知识的记忆和背诵（如“法国的首都是什么？”），而是关于过程、抽象和算法执行。无论是ICL、CoT还是多步算术，都要求模型学习并执行一个内部的、多步骤的转换流程。ICL是“根据示例学习并执行一个新算法”的终极体现[20]，而CoT则是将这个算法的执行过程外化[19]。这表明，真正“涌现”的，可能是一种通用的计算或推理能力，模型从海量的语言数据中学会了如何操作符号、遵循抽象规则，并将其应用于前所未见的问题。

此外，这些现象也说明，涌现不仅与模型本身有关，还与我们同模型交互的方式（即提示）密切相关。CoT的发现表明，我们可能尚未完全了解现有模型的能力边界，因为我们还没有找到所有正确的“提问方式”[8]。能力可能早已潜伏在模型内部，而一个好的提示策略就像一把钥匙，涌现出来并解锁了它。这预示着，今天的大模型中可能还潜藏着更多、更强大的能力，等待我们去发现能够激活它们的“咒语”。

第三部分：形成机制：为何规模能解锁潜力

大型语言模型中涌现能力的出现，并非偶然的魔法，而是根植于其架构、训练方法和规模扩展之间复杂的相互作用。要理解涌现，就必须回答用户查询的核心问题：为什么更大、更复杂的模型更容易展现出这些能力？本部分将从宏观的扩展法则（Scaling Laws）深入到微观的神经回路，层层剖析涌现能力的形成机制。

3.1 规模的基础：扩展法则带来的可预测提升

涌现现象的“不可预测性”是建立在一个“可预测”的基础之上的。这个基础就是扩展法则（**Scaling Laws**）。扩展法则描述了一个基本规律：在其他条件不受限的情况下，语言模型的性能（通常用交叉熵损失来衡量）会随着模型规模（N，参数量）、数据集规模（D，token数量）和训练计算量（C，FLOPs）的增加，呈现出平滑的、幂律形式的提升[30]。这意味着，更大、更多数据、更多计算量，通常会带来更好的基础模型。

3.1.1 原始扩展法则（**Kaplan et al., 2020**）

2020年，OpenAI的研究人员（Kaplan等人）发表了关于神经语言模型扩展法则的开创性论

文[30]。他们的核心发现是：

- 模型规模（**N**）是关键：在他们的实验中，增加模型参数量对降低损失的贡献最为显著。这一发现引领了当时的研究趋势，即致力于构建参数量越来越庞大的模型（如GPT-3），而训练数据集的规模相对固定[33]。
- 大模型更具样本效率：一个重要的推论是，大模型在学习上比小模型更有效率。它们可以用更少的训练步数和数据点达到与小模型相同的性能水平[30]。
- 最优训练策略：在固定的计算预算下，最优策略是训练一个非常大的模型，并在其完全收敛前提前停止训练，而不是将一个小模型训练至收敛[30]。

Kaplan等人提出的具体幂律关系如下：

- 仅受参数量N限制时，损失$L(N) \propto N^{-\alpha N}$，其中$\alpha_N \approx 0.076$。
- 仅受数据集大小D限制时，损失$L(D) \propto D^{-\alpha D}$，其中$\alpha_D \approx 0.095$。
  这些法则表明，性能的提升是平滑且可预测的，为后续的“涌现”现象提供了看似矛盾的背景。

3.1.2 计算最优扩展法则（**Hoffmann et al., 2022**）

2022年，DeepMind的研究人员（Hoffmann等人）通过其著名的“Chinchilla”论文，对扩展法则提出了重大的修正，引发了范式转变[36]。他们的核心论点是：

- 现有大模型**“训练不足”**：他们通过训练超过400个不同规模的模型发现，像GPT-3和Gopher这样的大模型，相对于其庞大的参数量而言，所用的训练数据量是远远不够的，即“训练不足”（undertrained）[35]。
- 模型与数据同等重要：他们的研究结论是，为了实现计算资源的最优利用（compute-optimal），模型参数量（N）和训练数据量（D）应该等比例扩展。即每当模型大小加倍时，训练token的数量也应该加倍[36]。
- **Chinchilla**的证明：他们用实验验证了这一新法则。他们训练了一个700亿参数的Chinchilla模型，但使用了高达1.4万亿个token的数据。尽管其参数量远小于Gopher（2800亿参数，3000亿token），Chinchilla在广泛的下游任务上全面且显著地超越了Gopher[36]。

Chinchilla法则强调，数据和参数同等重要。一个巨大的模型如果没有足够的数据来“喂养”，其潜力就无法被充分激发。这为理解涌现提供了一个关键视角：涌现不仅需要巨大的模型容量，还需要足够丰富和多样的数据来迫使模型学习到通用的、可泛化的结构，而不是简单地记住训练样本。

表3：关键扩展法则建议对比

论文	核心发现	对训练的启示	支持性文献
Scaling Laws for Neural Language Models (Kaplan et al., 2020)	模型性能随N, D, C呈幂律增长。强调增加模型参数量(N)的重要性。	引导了业界专注于构建参数量巨大的模型（如GPT-3）。	30
Training Compute-Optimal Large Language Models (Hoffmann et al., 2022)	现有大模型训练不足。为实现计算最优，N和D应等比例扩展。	促使研究重心转向同时大规模扩展数据集，催生了如Chinchilla这样更小但更高效的模型。	35

3.2 Transformer架构的角色

为什么扩展法则在Transformer架构上如此有效，并最终催生了涌现能力？答案在于其独特的设计哲学。

3.2.1 自注意力机制：一个全局灵活的工作空间

与循环神经网络（RNN）等早期架构相比，Transformer的核心创新是自注意力机制（self-attention）[39]。RNN按顺序处理文本，信息在传递过程中容易丢失，难以捕捉长距离依赖关系 [39]。而自注意力机制允许输入序列中的每个token直接与所有其他token进行交互和计算相关性，无论它们在序列中的距离有多远 [39]。

这创造了一个高度并行化的、全局性的“工作空间”，模型可以在其中直接建模任意两个词之间的复杂关系。这种捕捉长距离依赖的能力，是理解复杂语境、进行逻辑推理和形成抽象概念的先决条件，为涌现高级能力奠定了架构基础 [39]。

3.2.2 弱归纳偏置与学习的重担

归纳偏置（Inductive Bias）是指学习算法为了从有限数据中泛化而做出的先验假设 [43]。

例如，卷积神经网络（CNN）具有很强的“局部性”和“平移不变性”的归纳偏置，非常适合处理图像；RNN则具有“序列性”和“时间不变性”的偏置，适合处理时序数据[43]。

相比之下，Transformer的归纳偏置非常弱[43]。它不对数据做过多预设，其核心假设仅仅是“成对的交互很重要”（pairwise interactions）[44]。这种设计的代价是，在数据量较少时，Transformer的性能可能不如具有强偏置的模型，因为它需要从数据中学习一切，非常“饥饿”[43]。然而，也正是这种灵活性，使其在面对海量数据时，能够不受先验假设的束缚，学习到数据中存在的任何复杂结构。这种从数据中自发学习而非由人类设计的结构，正是“涌现”的本质。Transformer的弱偏置为涌现能力的形成提供了必要的自由度。

3.3 黑箱之内：内部表征与神经回路

涌现能力的“质变”最终发生在模型的内部。随着规模的扩大，模型内部发生了深刻的重组，从简单的模式匹配转向了更复杂的计算。

3.3.1 相变与“Grokking”现象

物理学中的**相变（phase transition）**为理解涌现提供了有力的类比[7]。就像水在100°C时会突然从液态变为气态，LLM在规模跨越某个临界阈值时，其行为也会发生质的改变[7]。

一个可观测的、与此高度相关的现象是**“Grokking”**（意为“心领神会”）[47]。Grokking指的是模型在训练集上达到100%准确率（即完全记住训练数据）后，经过非常长时间的额外训练，其在测试集上的泛化性能才突然从随机水平跃升至高水平[47]。这被看作是模型从纯粹的“记忆”模式切换到真正的“理解”和“泛化”模式的标志，是涌现的一种具体表现[48]。Grokking现象表明，模型内部正在发生一种缓慢但深刻的重构，最终导致了泛化能力的涌现。

3.3.2 内部世界模型与抽象表征的形成

随着规模的扩大，LLM不仅仅是在记忆表面的统计规律，它们开始构建关于世界的结构化、抽象的内部表征（**internal representations**）[51]。研究表明，LLM的隐藏层中形成了对概念、语义乃至物理空间和时间的表征[17]。例如，模型内部可能存在专门编码地理位置或历史

时期的“空间神经元”和“时间神经元”[51]。

这些表征的质量和抽象程度会随着模型规模的增大而提升，并在网络的不同层级中呈现出层次性：较低层级捕捉语法等浅层特征，而较高层级则整合更复杂的语义和世界知识 [53]。这些高质量的抽象表征，是模型进行推理、泛化和展现其他涌现能力的基石。

3.3.3 上下文学习的回路：归纳头与功能向量头

机制可解释性（Mechanistic Interpretability）研究进一步揭示了ICL这一关键涌现能力是如何在神经网络内部实现的。研究人员识别出了注意力机制中执行特定计算的专门“神经回路”。

- 归纳头（**Induction Heads**）：这是一种较早被发现且研究得较为透彻的回路。其功能非常直接：执行一种“匹配-复制”（match-and-copy）操作 [54]。当处理一个token时，归纳头会扫描上文，寻找该token之前出现的位置，然后预测上一次出现时紧随其后的那个token会再次出现。这种机制对于逐字复制、完成简单重复模式至关重要。归纳头通常在训练的早期阶段、在网络的较浅层出现 [56]。
- 功能向量头（**Function Vector Heads**）：这是一种更近期发现的、也更为复杂的机制。FV头的作用不是简单复制，而是计算出一个代表整个任务的紧凑的、潜在的**“功能向量”（function vector）**[56]。这个向量编码了从上下文中示例所学到的任务规则（例如，“将国家映射到首都”）。然后，模型可以将这个功能向量应用于新的查询，以解决任务。FV头被认为是执行更抽象、更复杂的ICL任务（如少样本学习 ）的核心。它们通常在训练的后期、在网络的更深层出现。消融实验（ablation studies）表明，在大型模型中，
  FV头是少样本**ICL**性能的主要驱动力 [56]。
- 发育关系：最引人入胜的发现是，这两种头之间存在一种发育上的联系。研究发现，许多FV头在训练初期表现为归纳头，随着训练的进行，它们逐渐“进化”或“过渡”为功能更强大的FV头 [56]。这表明，简单的归纳复制能力可能是学习更复杂、更抽象的基于功能向量的推理能力的“垫脚石”[56]。

表**4**：上下文学习（**ICL**）的内部机制对比

机制	拟议功能	典型层级位置	训练中涌现时机	在大型模型ICL中的主要作用	支持性文献
归纳头	寻找并复制	较浅层	训练早期涌现	对少样本ICL	[54]

(Induction Head)	token模式；执行逐字/句法层面的复制。			影响较小；对逐字复制至关重要。	
功能向量头 **(FV Head)**	计算一个编码了任务本身的潜在、抽象的向量。	较深层	训练后期涌现，有时由归纳头演变而来。	少样本ICL性能的主要驱动力。	[56]

3.3.4 统一理论：回路竞争

Huang等人提出的**回路竞争（Circuits Competition）**理论为统一理解Grokking、双重下降（Double Descent）和涌现能力提供了一个强有力的框架[50]。该理论认为，对于任何给定的任务，模型内部都可能形成多种不同的神经回路来解决它，而这些回路之间存在竞争。

- 记忆回路（**Memorization Circuits**）：这类回路学习速度快，能迅速记住训练数据，但在参数效率上较低（即需要更多参数），且泛化能力差[50]。
- 泛化回路（**Generalization Circuits**）：这类回路学习速度慢，需要更长的训练才能形成，但它们更高效（参数效率高），并且能够很好地泛化到未见过的数据[50]。

这两个回路之间的竞争动态解释了多种现象：

- **Grokking**：当训练数据量适中时，模型首先会快速形成记忆回路，导致训练损失迅速下降而测试损失居高不下。随着训练的继续，更慢但更高效的泛化回路逐渐形成并最终在优化中胜出，导致测试性能突然提升，这就是Grokking[50]。
- 涌现能力：在多任务学习的背景下（LLM的预训练本质上就是多任务学习），一项需要泛化能力的算法任务（如推理）可能会被大量需要记忆的知识任务所“干扰”。只有当模型规模和数据量达到一定程度，使得解决算法任务的泛化回路的“优先级”超过了无数个记忆回路时，这项能力才会“涌现”出来[50]。

这个理论将宏观的扩展和微观的机制联系起来，形成了一个从“量变”到“质变”的完整因果链：可预测的扩展法则为灵活的**Transformer**架构提供了增长的动力。这种增长并非线性，而是在模型内部引发相变。在相变过程中，由回路竞争驱动，模型发展出更抽象的内部表征和更高效的计算回路（如FV头）。当这些新回路的能力足以解决某个任务时，外部观察者就看到了涌现能力的诞生。

第四部分：综合论述与未来展望

对大型语言模型涌现能力的探索，已经从最初对惊人现象的观察，演变为一场关于其定义、真实性及底层机制的深刻科学辩论。本报告的最后一部分旨在综合前述的各种理论与证据，提出一个关于涌现现象的整合性观点，并探讨其对人工智能未来发展与安全性的深远影响。

4.1 对涌现现象的综合观点

将涌现能力简单地视为一个单一、神秘的事件，是一种过于简化的看法。相反，它是一个复杂、多阶段过程的最终外在表现。一个更为完整和综合的图景可以概括如下：

首先，可预测的扩展法则是涌现的基础。平滑的幂律关系表明，投入更多的计算资源、数据和参数，能够系统性地降低模型的基础损失，提升其基本能力[30]。这为后续的质变提供了必要的“量变”积累。

其次，这种量的积累作用于高度灵活的**Transformer**架构上。其弱归纳偏置的特性，使得模型不受过多先验假设的束缚，能够在海量数据中自由地学习和构建复杂的内部结构[43]。没有这种架构上的自由度，涌现便无从谈起。

接着，当规模的量变达到一定程度时，会触发模型内部的相变或**Grokking**现象[7]。这并非简单的性能提升，而是模型内部状态的根本性重组。在这种重组过程中，由

记忆回路与泛化回路之间的竞争所驱动，模型会逐渐放弃低效的、基于记忆的解决方案，转而形成更高效、更抽象的内部表征和计算回路[50]。例如，模型可能从简单的归纳头（Induction Heads）演化出更强大的功能向量头（FV Heads），从而获得执行抽象任务的能力[56]。

最后，**“涌现能力”**是我们作为外部观察者，在这些新的内部回路变得足够强大、能够解决一个以前无法解决的任务时所感知到的现象。这种感知的突变性，在很大程度上确实可能是由我们所选用的不连续评估指标所放大的“海市蜃楼”[2]。然而，这并不意味着模型内部没有发生真实的变化。底层的、模型解决问题策略的质变是真实存在的，即使其外在表现的提升曲线可以通过选择合适的指标而被“平滑化”。

因此，涌现能力既非纯粹的魔法，也非完全的假象。它是规模扩展、架构设计、内部动力学和外部评估共同作用下的复杂产物。

4.2 对人工智能发展与安全性的启示

对涌现能力的理解，无论其最终定义如何，都对人工智能的未来发展路径和安全治理提出了深刻的挑战与启示。

1. 不可预测性与安全隐患：涌现现象的核心特征之一是其不可预测性。这意味着，随着我们不断扩大模型规模，新的、我们甚至无法预见的能力可能会自发出现 [8]。这些能力可能是我们所期望的，但也可能包含有害的倾向，例如更高级的欺骗、操纵、偏见固化，或是产生与人类价值观不符的目标 [7]。我们可能无法完全了解一个模型的全部潜能，这为AI安全带来了巨大的不确定性。
2. 机制可解释性的紧迫性：整个关于涌现的讨论，都指向了同一个关键领域：机制可解释性（**Mechanistic Interpretability**）[17]。如果我们无法理解模型是如何做出决策的，我们就无法可靠地预测、控制或信任它们。逆向工程神经网络，将其分解为人类可理解的神经回路和逻辑路径，已不再是纯粹的学术兴趣，而是确保未来更强大AI系统安全、可控和对齐的必要前提 [17]。理解涌现的形成机制，是迈向透明和可信AI的第一步。
3. 未来**AI**发展的路径：对涌现的研究正在推动AI领域从“多即是异”（more is different）的蛮力扩展，向“巧能致胜”（smarter is better）的精细化发展转变。未来的突破可能不仅来自于更大的模型，更来自于对如何高效诱导期望能力的深刻理解。这包括设计更优的架构、精心构建能促进泛化回路形成的训练数据（如回路竞争理论所启示的），以及开发新的训练技术和提示策略 [64]。研究涌现，本质上是在努力将AI的开发从一门“黑箱艺术”转变为一门真正的科学。

综上所述，对大型语言模型涌现能力的探究，是当前人工智能领域最前沿、也最富挑战性的课题之一。它迫使我们重新审视智能的本质、学习的规律以及我们与日益强大的机器之间的关系。虽然目前尚无定论，但这场科学辩论本身正在以前所未有的深度和广度，推动着我们对人工智能的认知边界。

Works cited

1. Large Language Models and Emergence: A Complex Systems Perspective - arXiv, accessed June 17, 2025, https://arxiv.org/pdf/2506.11135
2. Are Emergent Abilities of Large Language Models a Mirage?, accessed June 17, 2025, https://papers.neurips.cc/paper_files/paper/2023/file/adc98a266f45005c403b8311ca7e8bd7-Paper-Conference.pdf
3. Emergent Abilities of Large Language Models - OpenReview, accessed June 17, 2025, https://openreview.net/pdf?id=yzkSU5zdwD
4. LLMs Do Show Emergent Properties: A critique and an appreciation of the paper "Are Emergent Abilities of Large Language Models a Mirage?" : r/singularity -

Reddit, accessed June 17, 2025, https://www.reddit.com/r/singularity/comments/1it4abp/llms_do_show_emergent_properties_a_critique_and/
5. Emergent Abilities of Large Language Models - OpenReview, accessed June 17, 2025, https://openreview.net/forum?id=yzkSU5zdwD
6. Are Emergent Abilities of Large Language Models a Mirage? - OpenReview, accessed June 17, 2025, https://openreview.net/pdf?id=JRdN9GcI52
7. Emergent Abilities in Large Language Models: A Survey - arXiv, accessed June 17, 2025, https://arxiv.org/html/2503.05788v2
8. Characterizing Emergent Phenomena in Large Language Models - Google Research, accessed June 17, 2025, https://research.google/blog/characterizing-emergent-phenomena-in-large-language-models/
9. How Quickly Do Large Language Models Learn Unexpected Skills? - Quanta Magazine, accessed June 17, 2025, https://www.quantamagazine.org/how-quickly-do-large-language-models-learn-unexpected-skills-20240213/
10. [2304.15004] Are Emergent Abilities of Large Language Models a Mirage? - arXiv, accessed June 17, 2025, https://arxiv.org/abs/2304.15004
11. Emergent Abilities of Large Language Models – Fact or Mirage? - DEV Community, accessed June 17, 2025, https://dev.to/myakala/emergent-abilities-of-large-language-models-fact-or-mirage-2ice
12. Are Emergent Abilities of Large Language Models a Mirage?, accessed June 17, 2025, https://arxiv.org/pdf/2304.15004
13. Are Emergent Abilities of Large Language Models a Mirage? - OpenReview, accessed June 17, 2025, https://openreview.net/forum?id=ITw9edRDID
14. Large Language Models' Emergent Abilities Are a Mirage : r/LocalLLaMA - Reddit, accessed June 17, 2025, https://www.reddit.com/r/LocalLLaMA/comments/1bn2udc/large_language_models_emergent_abilities_are_a/
15. Understanding Emergent Abilities of Language Models from the Loss Perspective - arXiv, accessed June 17, 2025, https://arxiv.org/html/2403.15796v3
16. Understanding Emergent Abilities of Language Models from the Loss Perspective, accessed June 17, 2025, https://openreview.net/forum?id=35DAviqMFo¬eId=a8Mrkk9fcX
17. All About Emergent Behavior in Large Language Models - ThirdEye Data, accessed June 17, 2025, https://thirdeyedata.ai/all-about-emergent-behavior-in-large-language-models/
18. Emergent Abilities in Large Language Models: A Survey - arXiv, accessed June 17, 2025, https://arxiv.org/pdf/2503.05788
19. Do Emergent Abilities in AI Models Boil Down to In-Context Learning? - IKANGAI, accessed June 17, 2025, https://www.ikangai.com/do-emergent-abilities-in-ai-models-boil-down-to-in-context-learning/

20. What is In-context Learning, and how does it work: The Beginner's Guide - Lakera AI, accessed June 17, 2025, https://www.lakera.ai/blog/what-is-in-context-learning
21. Are Emergent Abilities in Large Language Models just In-Context Learning? - ACL Anthology, accessed June 17, 2025, https://aclanthology.org/2024.acl-long.279.pdf
22. In Context Learning Guide - PromptHub, accessed June 17, 2025, https://www.prompthub.us/blog/in-context-learning-guide
23. arxiv.org, accessed June 17, 2025, https://arxiv.org/html/2503.05788v2#:~:text=The%20term%20emergent%20in%20the,gradient%20updates%20to%20the%20model.
24. Introduction to Large Language Models | Machine Learning - Google for Developers, accessed June 17, 2025, https://developers.google.com/machine-learning/resources/intro-llms
25. suzgunmirac/BIG-Bench-Hard - GitHub, accessed June 17, 2025, https://github.com/suzgunmirac/BIG-Bench-Hard
26. Big Bench — The GenAI Guidebook - Ravin Kumar, accessed June 17, 2025, https://ravinkumar.com/GenAiGuidebook/deepdive/BigBench.html
27. google/BIG-bench: Beyond the Imitation Game collaborative benchmark for measuring and extrapolating the capabilities of language models - GitHub, accessed June 17, 2025, https://github.com/google/BIG-bench
28. Emergent Abilities of Large Language Models - AssemblyAI, accessed June 17, 2025, https://www.assemblyai.com/blog/emergent-abilities-of-large-language-models
29. Emergent Capabilities of LLMs - Prem AI Blog, accessed June 17, 2025, https://blog.premai.io/emergence-capabilities-of-llms/
30. Scaling Laws for Neural Language Models - arXiv, accessed June 17, 2025, http://arxiv.org/pdf/2001.08361
31. Neural scaling law - Wikipedia, accessed June 17, 2025, https://en.wikipedia.org/wiki/Neural_scaling_law
32. Temporal Scaling Law for Large Language Models - arXiv, accessed June 17, 2025, https://arxiv.org/html/2404.17785v2
33. Scaling Laws for Neural Language Models - ResearchGate, accessed June 17, 2025, https://www.researchgate.net/publication/338789955_Scaling_Laws_for_Neural_Language_Models
34. [2001.08361] Scaling Laws for Neural Language Models - arXiv, accessed June 17, 2025, https://arxiv.org/abs/2001.08361
35. Understanding scaling laws for LLM training - Weights & Biases - Wandb, accessed June 17, 2025, https://wandb.ai/site/articles/training-llms/the-scaling-laws/
36. Training Compute-Optimal Large Language Models, accessed June 17, 2025, https://proceedings.neurips.cc/paper_files/paper/2022/file/c1e2faff6f588870935f114ebe04a3e5-Paper-Conference.pdf
37. 2022 TrainingComputeOptimalLargeLang - GM-RKB, accessed June 17, 2025,

http://www.gabormelli.com/RKB/2022_TrainingComputeOptimalLargeLang
38. Training Compute-Optimal Large Language Models, accessed June 17, 2025, https://arxiv.org/pdf/2203.15556
39. Understanding Transformer Architecture: The Backbone of Modern AI - Udacity, accessed June 17, 2025, https://www.udacity.com/blog/2025/04/understanding-transformer-architecture-the-backbone-of-modern-ai.html
40. How Transformers Work: A Detailed Exploration of Transformer Architecture - DataCamp, accessed June 17, 2025, https://www.datacamp.com/tutorial/how-transformers-work
41. Transformer Architecture and Attention Mechanisms in Genome Data Analysis: A Comprehensive Review - PMC, accessed June 17, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10376273/
42. Attention (machine learning) - Wikipedia, accessed June 17, 2025, https://en.wikipedia.org/wiki/Attention_(machine_learning)
43. A fAIry tale of the Inductive Bias | Towards Data Science, accessed June 17, 2025, https://towardsdatascience.com/a-fairy-tale-of-the-inductive-bias-d418fc61726c/
44. [D] What is the inductive bias in transformer architectures? : r/MachineLearning - Reddit, accessed June 17, 2025, https://www.reddit.com/r/MachineLearning/comments/d0gnyp/d_what_is_the_inductive_bias_in_transformer/
45. Phase Transitions in Large Language Models and the $O(N)$ Model - arXiv, accessed June 17, 2025, https://arxiv.org/abs/2501.16241
46. Understanding LLM Phase Transition: A Breakthrough Discovery, accessed June 17, 2025, https://meta-quantum.today/?p=2458
47. The Accidental Discovery That Changed AI: How OpenAI Grokked LLMs - Spearhead, accessed June 17, 2025, https://spearhead.so/the-accidental-discovery-that-changed-ai-how-openai-grokked-llms/
48. Emergence in non-neural models: grokking modular arithmetic via average gradient outer product | OpenReview, accessed June 17, 2025, https://openreview.net/forum?id=FrjTgprk3V
49. Deep Networks Always Grok and Here is Why - arXiv, accessed June 17, 2025, https://arxiv.org/html/2402.15555v2
50. Unified View of Grokking, Double Descent and Emergent Abilities: A Perspective from Circuits Competition - arXiv, accessed June 17, 2025, https://arxiv.org/html/2402.15175v1
51. Large Language Models develop structured internal representations of both space and time., accessed June 17, 2025, https://onyxaero.com/news/language-models-develop-structured-internal-representations-of-both-space-and-time/
52. Study shows LLMs do have Internal World Models : r/ArtificialInteligence - Reddit, accessed June 17, 2025, https://www.reddit.com/r/ArtificialInteligence/comments/1jw6e75/study_shows_ll

ms_do_have_internal_world_models/
53. Beyond Words: A Latent Memory Approach to Internal Reasoning in LLMs - arXiv, accessed June 17, 2025, https://arxiv.org/html/2502.21030v1
54. What needs to go right for an induction head? A mechanistic study of in-context learning circuits and their formation - arXiv, accessed June 17, 2025, https://arxiv.org/pdf/2404.07129?
55. arXiv:2504.03022v1 [cs.CL] 3 Apr 2025, accessed June 17, 2025, https://arxiv.org/pdf/2504.03022?
56. Which Attention Heads Matter for In-Context Learning? - arXiv, accessed June 17, 2025, https://arxiv.org/html/2502.14010v1
57. Which Attention Heads Matter for In-Context Learning? - arXiv, accessed June 17, 2025, http://arxiv.org/pdf/2502.14010
58. Which Attention Heads Matter for In-Context Learning? - arXiv, accessed June 17, 2025, https://www.arxiv.org/pdf/2502.14010
59. arxiv.org, accessed June 17, 2025, https://arxiv.org/html/2502.14010v1#:~:text=FV%20heads%20consistently%20appear%20in,minimal%20impact%20(%C2%A74).
60. This AI Paper Identifies Function Vector Heads as Key Drivers of In-Context Learning in Large Language Models - MarkTechPost, accessed June 17, 2025, https://www.marktechpost.com/2025/03/04/this-ai-paper-identifies-function-vector-heads-as-key-drivers-of-in-context-learning-in-large-language-models/
61. Unified View of Grokking, Double Descent and Emergent Abilities: A Perspective from Circuits Competition - arXiv, accessed June 17, 2025, https://arxiv.org/pdf/2402.15175
62. [2402.15175] Unified View of Grokking, Double Descent and Emergent Abilities: A Perspective from Circuits Competition - arXiv, accessed June 17, 2025, https://arxiv.org/abs/2402.15175
63. NeurIPS Poster Grokking of Implicit Reasoning in Transformers: A Mechanistic Journey to the Edge of Generalization, accessed June 17, 2025, https://neurips.cc/virtual/2024/poster/96105
64. How Scaling Laws Drive Smarter, More Powerful AI - NVIDIA Blog, accessed June 17, 2025, https://blogs.nvidia.com/blog/ai-scaling-laws/