Systematic Outliers in Large Language Models: A Comprehensive Review of Their Formation, Impact, and Mitigation

Anonymous Author(s)
Affiliation
Address
email

Abstract

大型语言模型(LLMs)中的系统异常值已从一个模型压缩领域的工程问题转变为理解Transformer架构核心工作机制的关键研究课题。本文综述系统性地分析了这些异常值的形成机制、功能作用及其对模型性能和效率的影响。研究表明，异常值不是随机噪声，而是自注意力机制设计的必然结果，它们作为隐式的上下文感知缩放因子发挥着关键作用。同时，这些异常值也是模型压缩特别是量化技术面临的主要障碍。本文总结了现有的缓解策略，从算法层面的后处理方法到结构层面的架构重新设计，并探讨了异常值与激活稀疏性、Hessian矩阵结构以及模型隐私安全等更广泛LLM现象的联系。最后，本文提出了未来研究的关键方向，包括设计抗异常值的架构、统一优化框架以及基于异常值的可解释性工具。

1 Introduction

1.1 The Emergence of Outliers as a Critical Field of Study

大型语言模型(LLMs)的激增标志着人工智能领域的范式转变，其特征是涌现出许多往往难以简单解释的能力。这些基于Transformer架构的模型在从自然语言理解到复杂推理等任务中展现出显著的能力。然而，这种令人印象深刻的性能伴随着一系列令人困惑且常常有问题的内部行为。其中最重要的是异常值现象——模型激活、权重或注意力分数中与典型分布显著偏离的值。根据An等人[1]的研究，这些异常值已经从工程问题转变为关键研究课题。

最初，这些异常值主要从操作效率的角度被看待，被视为对模型压缩技术（如量化）构成重大挑战的统计异常。这些极值的存在被视为一种麻烦，阻碍了在资源受限的硬件上部署这些大规模模型。然而，越来越多的研究从根本上重新定义了这种理解。异常值现在被公认为Transformer架构本身的一种系统性和功能上重要的属性，是其核心计算机制的可预测副产品。这一认识将异常值的研究从一个小众的工程问题提升为理解LLMs基本工作原理的中心话题。

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

1.2 Thesis Statement: From Statistical Anomaly to Functional Mechanism

本综述的中心论点是，LLMs中的异常值不是随机产物，而是*系统异常值*——自注意力机制设计在功能上不可或缺且可预测的结果。它们代表着一把典型的双刃剑：在其原始状态下，它们似乎对实现高模型性能至关重要，但同时它们也构成了操作效率的主要障碍，特别是对于基于量化的模型压缩。本综述将综合当前知识，论证理解、管理并最终重新设计产生这些异常值的机制对于下一代高效、稳定和安全的LLMs至关重要。

1.3 Scope and Structure of the Review

为了提供全面和结构化的分析，本综述分为七个部分。第2节首先建立对该现象的基础理解，提出了异常值的系统分类法，并详细说明了它们在Transformer架构中的经验分布。第3节深入探讨了这些异常值的因果起源，确定softmax操作为其根本原因，并探索了它们作为注意力机制中隐式计算元素的功能作用。第4节考察了异常值的广泛影响，重点关注它们对模型压缩、训练动态和整体性能的深远影响。第5节调查了缓解策略的全景，对比了管理异常值效应的算法方法与旨在防止其形成的结构解决方案。第6节通过将这些发现与更广泛的LLM现象综合起来，提升了讨论水平，建立了与激活稀疏性、Hessian矩阵的数学结构以及模型隐私和安全关键领域的联系。最后，第7节总结了最新技术水平并规划了未来研究路线，确定了关键的开放问题和有前景的研究途径。

2 The Anatomy of Outliers in Transformer Architectures

对系统异常值的深入理解始于对其各种形式的精确表征以及它们在LLMs复杂结构中分布的详细映射。实证研究已经超越了仅仅确认它们的存在，进而系统地对它们进行分类并分析其分布模式，揭示了一种高度结构化和非随机的现象。

2.1 Defining the Outlier Landscape: A Systematic Taxonomy

如Wei等人[4]所述，文献已经形成了一个分类法，根据异常值在模型计算图中的表现位置将其分为三种主要类型。

* **激活异常值**：这些是在激活张量中观察到的极值，代表网络内各层和子模块的输出。激活异常值对训练后量化(PTQ)特别有害，这是一种常见的模型压缩技术。由于量化需要将浮点值范围映射到较小的定点整数集合，单个大的激活值会大大扩展所需的动态范围，导致大多数较小、更常见的值的精度显著损失。 * **权重异常值**：这些指的是模型权重矩阵中具有不成比例大 magnitude 的一小部分参数。虽然它们仅占总参数数量的一小部分，但研究表明，这些权重异常值对于维持模型质量往往至关重要。例如，最近的工作已经识别出"超级权重"，有时是单个参数，可以对模型的功能产生巨大影响。 * **注意力异常值**：这些是在自注意力机制计算的注意力分数矩阵中发现的极值。注意力异常值表示特定标记从序列中的其他标记接收压倒性的关注，导致高度倾斜、非均匀的注意力分布。

"系统性"一词对这种分类法至关重要。它强调这些异常值不是随机的计算错误，而是在不同模型和训练运行中一致出现的模式化现象。它们的出现与底层架构和训练过程的动态性内在相关，表明其起源是确定性的而非随机噪声。

2.2 Empirical Observations and Distributional Analysis

对包括LLaMA、OPT和BLOOM在内的广泛流行LLM系列的系统分析证实，异常值是一个一致且可预测的特征，而不是特定模型或训练设置的特性。这项研究已经产生了关于这些异常值倾向于出现的详细映射。

* **在Transformer块中的位置**：激活异常值最集中在前馈网络(FFN)内，也称为多层感知器(MLP)块，该块位于自注意力子层之后。具体来说，它们经常出现在FFN的下投影矩阵的输入中。权重异常值表现出类似的定位，经常出现在这些相同的下投影矩阵中。注意力异常值，顾名思义，位于注意力机制内，但其强度和分布通常显示出明显的、特定于头的模式，表明不同注意力头之间的专业化。* **层间分布**：异常值的分布在网络深度上并不均匀。一种常见模式是激活异常值在模型的浅层和中层最为突出。它们的普遍存在在最终层中经常发生变化，表明它们的作用可能随着模型准备其最终输出表示而受到调节。这种分层模式进一步强化了异常值是结构化计算过程一部分的观点。

2.3 Key Properties: Asymmetry and Concentration

除了它们的位置之外，激活异常值的两个关键分布特性已被确定为它们对模型压缩（特别是量化）如此具有破坏性的主要原因。

* **集中性**：异常值并非散布在所有特征维度上。相反，它们强烈集中在极少数特定通道内。这意味着在典型LLM的隐藏状态中的数千个特征维度中，只有少数几个负责产生这些极值。整个动态范围扩大的问题通常可以追溯到这些少数"问题通道"。* **不对称性**：在这些特定的易产生异常值的通道中，值的分布通常高度不对称。例如，对OPT-66B模型的实证分析显示，一个问题通道的值集中在从-97到-58的负范围内，而另一个通道在从5.7到43的正范围内表现出异常值。通道之间的这种不对称性会导致整个张量范围显著宽于任何单个通道的范围，加剧了通常假设围绕零中心对称分布的量化算法的挑战。

2.4 Synthesis and Deeper Implications

这些异常值的高度结构化、局部化和模式化性质为它们的起源提供了重要线索。它们在特定组件（如FFN下投影）、特定层和不同模型系列中一致出现，强烈表明它们不是模型的全局故障，而是一种专门、有针对性的机制。这种一致性指向的不是源于单次训练运行的随机产物的解释，而是Transformer架构本身施加的共同、底层压力。如果相同的模式在不同模型中独立出现，那么它们很可能是架构设计所提出的共享问题的收敛解决方案。这种推理为更深入地研究Transformer块的基本机制（特别是自注意力机制）奠定了基础，将其视为这种系统行为的可能来源。

3 The Genesis and Functional Role of Systematic Outliers

在确定了异常值是什么以及它们在哪里出现之后，调查自然转向它们为什么存在。最近的理论和实验工作提供了令人信服的证据，表明系统异常值不是缺陷，而是自注意力机制的一种涌现且功能上必要的特征。它们源于架构的基本数学特性，并在模型计算中发挥着独特的、尽管违反直觉的作用。

3.1 The Root Cause: Softmax Operation in Self-Attention

根据Puccetti等人[3]的研究，理解异常值的核心理论突破是识别softmax函数（自注意力机制的基石）作为它们的基本生成器。自注意力的主要作用是动态地权衡序列中不同标记相对于给定标记的重要性。它通过计算注意力分数，然后使用softmax将它们标准化为概率分布来实现这一点。

异常值的起源在于softmax如何完成这项任务。为了创建一个稀疏或"峰值"分布——其中注意力急剧集中在一个或几个高度相关的标记上——该函数必须放大其输入值（logits）中的小差异。当模型确定某条上下文信息至关重要时，训练过程会将相应的logit推向相对于其他logit的极端正值。这确保在softmax的指数和标准化步骤之后，其相应的注意力权重接近1，而所有其他权重接近0。这种放大过程不是错误；它是softmax在执行精确信息选择时的预期和必要功能。这些极端logit值，对于有效的注意力是必需的，是异常值现象的初始种子，然后传播到整个网络。

3.2 The Lifecycle of an Outlier: A Causal Chain

异常值的形成不是一个孤立事件，而是一个级联过程，一个"生命周期"，追踪它们通过Transformer块不同组件的传播。

1. **从权重中浮现**：这个过程通常始于权重异常值。将输入投影到查询(W_Q)和键(W_K)的权重矩阵中的大幅度值会导致查询和键之间的点积可能产生异常大的logit值。 2. **传播到注意力**：当这些极端logit被输入到softmax函数时，会产生高度集中的注意力分布，创造了所谓的注意力异常值。 3. **激活异常值的创建**：注意力子层的输出是值向量(V)的加权和，其中权重是注意力分数。当注意力异常值（接近1的分数）乘以其相应的值向量时，它实际上允许该值向量几乎不变地通过，同时抑制所有其他向量。这种操作，特别是当跨多个注意力头聚合时，可以在注意力模块的输出中生成大幅度的激活异常值。 4. **强化和消失**：这些激活异常值随后成为后续FFN块的输入，可能强化了该块中对大权重（权重异常值）的需求，因为模型学习处理这些高幅度信号。这创造了一个潜在的反馈循环。有趣的是，这个过程在模型的最终层中似乎受到调节，其中某些异常值的普遍存在减少，表明在输出层之前对表示进行最终、特定于任务的细化。

3.3 The Functional Purpose: Hypotheses and Evidence

如果异常值是模型计算的系统和组成部分，它们必须具有功能性目的。研究已经围绕三个主要假设展开，越来越多的证据支持第三个假设。

1. **固定偏差**：在这种观点中，异常值充当稳定、强大的偏差，无论特定输入上下文如何，始终强调某些标记或特征。这一假设受到相关概念"大规模激活"的启发，其中某些神经元对特定、重要的概念以高强度激活。例如，异常值可能始终对标点符号或序列开始标记激活。 2. **上下文感知偏差**：这一假设通过注意到特别是注意力异常值根据输入序列而显著变化，从而完善了第一个假设。这表明它们不是固定的，而是作为动态信号，根据提示内容自适应地引导模型的注意力，为任务需要突出显示不同信息。 3. **隐式、上下文感知缩放因子**：这是最引人注目且有些违反直觉的假设。理论推导和实验表明，异常值作为隐式缩放机制发挥作用。证据表明，对应于接收异常值级注意力分数的标记的值向量(V)通常具有显著*较小*的magnitude。这意味着大的注意力分数不是为了放大信号，而是为了精确控制并在某些情况下减弱某些上下文信息的影响。通过创建非常尖锐的注意力分布，模型可以有效地隔离标记并缩小其上下文更新，这有助于稳定网络并防止对其表示进行不必要或破坏性的更改。在这种观点中，极值矛盾地用于实现稳定性和控制。

Transformer的架构本身在两个竞争目标之间造成了根本的张力：*精确信息选择*，需要稀疏、高方差的注意力分布，以及*数值稳定性*，有利于较小、表现良好的值范围。系统异常值似乎是模型对这种权衡的涌现解决方案。为了执行复杂语言任务所需的高度选择性注意力，模型必须在自然产生极值的机制中使用softmax函数。这些异常值不是需要修复的bug，而是选择机制本身的不可分割的特征。这解释了为什么天真地尝试移除它们总是导致性能下降；这样做类似于从精心调整的机器中移除关键齿轮，破坏了模型学会依赖的核心计算策略。

4 Ramifications for Model Performance and Operational Efficiency

系统异常值的系统性和功能性意味着它们的存在具有深远而广泛的影响。虽然它们似乎对实现高性能不可或缺，但同时它们也是使LLMs在实际部署中高效和实用的主要挑战来源。这在模型能力和操作可行性之间创造了直接且可量化的冲突。

4.1 The Primary Obstacle to Model Compression

异常值最直接和有记录的影响是对模型压缩，特别是训练后量化(PTQ)。

* **量化灾难**： PTQ旨在通过将32位或16位浮点权重和激活转换为低位整数（如INT8或INT4）来减小模型大小并加速推理。这个过程涉及定义一个范围并将其划分为离散数量的量化"bin"。异常值的存在导致这个过程的灾难性失败。例如，如果张量的值大多在[-2, 2]范围内，但包含一个100的激活异常值，量化范围必须扩展到至少[-100, 100]以容纳它。结果，[-2, 2]范围内的绝大多数原始值都映射到靠近零的可用量化bin的一小部分。这导致精度的巨大损失，因为不同的浮点值被折叠到相同的整数表示中，严重降低模型性能。观察到的异常值不对称性进一步加剧了这个问题，这迫使使用更宽的对称范围，并浪费了大量量化级别。 * **剪枝挑战**：模型剪枝，旨在通过移除冗余权重来提高效率，也面临来自异常值的复杂性。基于幅度的剪枝，一种常见方法，移除绝对值小的权重。然而，权重异常值的存在——在功能上至关重要的大幅度参数——使得设置全局剪枝阈值变得困难。设置过高的阈值可能错误地移除这些重要的异常值权重，而设置过低的阈值将无法实现有意义的压缩。这迫使进行困难的权衡，使简单剪枝策略的应用复杂化。

4.2 Influence on Training Dynamics and Stability

异常值的影响超出了训练后的效率，延伸到训练过程本身。通过结构性修改注意力机制以防止异常值形成的研究发现，这些模型在训练的早期阶段表现出更快的收敛。

这一发现表明，标准Transformer模型必须将其优化预算的很大一部分用于学习管理和利用这些数值极端值。优化景观可能充满了由这些异常值引起的尖锐梯度和不稳定性，要求优化器采取更谨慎和迂回的路径。通过设计一个实现相同功能而不生成异常值的架构，优化问题变得本质上更容易，允许模型更快、更稳定地学习。这意味着异常值，虽然在训练好的模型中功能有用，但对训练过程施加了切实的成本。

4.3 The Performance Cost of Naive Mitigation

异常值的功能重要性通过它们天真移除的后果得到最鲜明的证明。多项研究的一致发现是，简单地裁剪异常值（将任何超过阈值的值设置为该阈值）或在没有更复杂、补偿策略的情况下移除它们会导致模型性能的严重下降。像困惑度这样的指标，衡量模型预测文本序列的能力，可以急剧增加，表明语言能力的显著损失。这一经验结果提供了反对将异常值视

为单纯噪声的观点的最强有力证据。如果它们仅仅是随机错误，移除它们应该产生中性甚至积极的效果。移除它们如此有害的事实证实，它们深深嵌入模型的计算路径中，对其正确处理信息的能力至关重要。

系统异常值的存在因此在LLM开发的核心创造了一种根本张力。模型演化为实现高性能的机制——即由异常值实现的精确、选择性注意力分布——正是使模型在资源受限的硬件上高效运行变得困难和昂贵的机制。这不是一个简单的需要修复的bug，而是一个核心架构权衡。它将研究人员和工程师面临的核心挑战框定为不是"我们如何摆脱异常值？"而是"我们如何在没有它们问题性数值特性的情况下实现异常值的*功能*？"寻找这个问题答案的过程推动了一系列创新缓解策略的发展。

5 A Review of Mitigation and Management Strategies

异常值作为LLM效率核心挑战的认识推动了各种缓解技术的发展。这些策略可以大致沿着哲学谱系分类，从接受模型的原生行为并寻求管理其后果的策略，到旨在从根本上改变模型架构以防止异常值形成的策略。这种演变反映了该领域对问题理解的日益成熟。

5.1 The Dichotomy: Algorithmic vs. Structural Solutions

解决方案的景观可以分为两种主要方法：

* **算法解决方案**：这些方法基于接受标准预训练模型会生成异常值的原则。它们的目标是减轻这些异常值的负面影响，通常在训练后压缩阶段。它们被设计为可以应用于现有模型的"补丁"或"包装器"，使其更适合量化等技术。这些解决方案处理异常值问题的症状。 * **结构解决方案**：这些方法采取更根本的方法，旨在从源头上防止异常值的形成。这涉及修改模型的核心架构，特别是自注意力机制，以实现所需功能而不产生极端数值。这些解决方案处理异常值问题的根本原因。

5.2 Case Study: The Outlier Suppression+ (OS+) Framework

Wei等人[4]提出的Outlier Suppression+ (OS+)框架代表了一种最先进的算法解决方案，旨在实现准确的训练后量化。它直接针对激活异常值的两个关键属性：集中性和不对称性。

* **核心机制**： * **通道级移位**：为了抵消异常值的*不对称性*，OS+对每个特征通道应用移位操作。通过计算每个通道值范围的中心并将其移向零，它创建了一个更对称的分布。这显著减小了量化所需的整体张量范围，而不改变编码在值之间相对差异中的信息。 * **通道级缩放**：为了应对异常值在特定通道中的*集中性*，OS+对这些问题通道应用缩放因子。它缩小它们的幅度以更好地与张量的其余部分对齐，从而平衡量化负担并防止少数通道主导量化范围。 * **等价性和迁移**：OS+最创新的方面是其数学等价性原则。移位和缩放操作不是在推理期间作为额外步骤执行的。相反，它们的效果在代数上被"迁移"并吸收到网络中后续线性层的权重和偏差中。这意味着应用OS+后，可以导出一个新的浮点模型，该模型在功能上与原始模型相同，但具有弱得多的异常值。然后可以使用标准技术有效地量化这个新模型，所有这些都具有*零推理时间开销*。 * **性能**：这种有针对性方法的有效性已在各种模型中得到证明。OS+已被证明在包括BERT、OPT、BLOOM和LLaMA在内的模型上实现近乎无损的8位和6位量化。对于BERT上更激进的4位量化，它建立了新的最先进水平，展示了其优于以前方法的优势。

表 1: 异常值缓解策略的比较分析

策略	核心机制	目标异常值类型	
算法： Outlier Suppression+ (OS+)	通道级移位和缩放，抵消不对称性和集中性	激活	实现高
结构： 注意力机制重新设计	引入显式、上下文感知缩放因子替代异常值功能	所有（防止形成）	解决
经典： 天真裁剪/移除	设置阈值并移除或饱和超出阈值的值	所有	

5.3 Towards Structural Elimination: Modifying the Attention Mechanism

与算法解决方案的事后修复不同，结构方法寻求重新设计注意力机制本身。An等人（2025）的研究表明，异常值执行的隐式缩放功能可以替换为直接纳入注意力公式的*显式、上下文感知缩放因子*。

这种修改消除了softmax函数生成极端logit值以实现精确注意力的需要。通过提供一个明确的机制来控制信息流，模型不再被迫发现基于异常值的策略。这种方法的好处是深远的：它不仅通过防止异常值形成来解决压缩问题，而且还导致更快的模型收敛和改进的训练稳定性，如前所述。这表明通过解决根本原因，可以同时解决多个下游问题。

5.4 Comparative Analysis of Outlier Mitigation Strategies

下表提供了不同缓解策略的结构化比较，突出了它们的机制、优势和局限性。本摘要旨在帮助研究人员和从业者评估为其特定需求和约束选择适当方法时涉及的权衡。

5.5 Synthesis and Deeper Implications

异常值缓解研究的轨迹反映了工程和科学领域成熟的经典模式。最初的、简单的方法，如天真裁剪，表面地处理问题并失败，因为它们不尊重潜在功能。这导致了对问题特征——不对称性和集中性——的更深入分析，反过来又使开发复杂的算法补丁如OS+成为可能，这些补丁与系统特性合作而不是对抗。最后和最先进的阶段是朝着对架构内根本原因的基本理解迈进，导致预防性结构重新设计。这种进展代表了研究社区内明显的学习曲线，从仅仅治疗症状到治愈潜在疾病。

6 Synthesis and Connections to Broader LLM Phenomena

系统异常值的研究，虽然本身至关重要，但并不存在于真空中。它作为一个强大的镜头，通过它可以查看和连接LLMs的其他基本属性。通过检查异常值与激活稀疏性、损失景观的数学结构甚至模型隐私等现象之间的相互作用，Transformer架构内部工作的更统一和整体的图景开始浮现。这些不是孤立的问题，而是相同底层系统动态的不同表现。

6.1 Outliers and Activation Sparsity: Two Sides of the Same Coin?

激活稀疏性是深度学习模型的另一个有趣且广泛研究的属性，指的是对于任何给定输入，大部分神经元激活趋向于零或接近零。乍一看，稀疏性和异常值似乎是对立的概念：稀疏性关于非常小的值的普遍性，而异常值关于非常大的值的存在。然而，它们可以理解为同一枚硬币的两面——两种互补的高效信息处理策略。

稀疏性通过有效忽略无关信息使模型在计算上高效；如果神经元的激活为零，它对后续层的贡献就无效了。相反，异常值允许模型将极端重要性放在少量关键信息上。两者都是选

择性信息路由的机制。对激活稀疏性的研究揭示了"通用模式"，这些模式反映了对异常值的发现：稀疏性倾向于随着模型大小增加，并且FFN层内的中间激活通常最稀疏。两种现象都表现出系统的、非随机模式的事实表明，它们受相同的底层架构原则和优化压力的支配，代表了模型学会精确控制的激活分布的两个极端。

6.2 Mathematical Underpinnings: The Hessian Matrix and Random Matrix Theory

异常值的经验观察可以基于模型优化景观的更深层数学结构，这由Hessian矩阵——损失函数的二阶偏导数矩阵——描述。Hessian的结构提供了对训练动态和参数相互依赖性的深刻见解。

Dong等人[2]的最新工作揭示，Hessian的结构由源于模型架构的"静态力"和源于训练过程的"动态力"塑造。系统异常值的系统性可以解释为这些力的直接结果。架构的"静态力"可能使某些特征维度或通道具有高度敏感性，使它们成为成为异常值通道的候选者。训练的"动态力"然后巩固了它们的功能作用，放大它们的幅度以服务于特定目的，如前所述的隐式缩放。

此外，使用随机矩阵理论的分析表明，神经网络的Hessian通常接近块对角矩阵，特别是当输出类别的数量（或对于LLMs，词汇表大小C）变得非常大时。块对角结构意味着参数组（例如，在层或神经元内）在很大程度上独立于其他组。因此，异常值可以理解为模型处理未被这种主导、解耦结构捕获的罕见但至关重要的跨块交互的机制。它们是连接否则半独立计算模块的高成本、高影响力连接。

6.3 A New Frontier: Implications for Privacy and Security

也许最深刻和前瞻性的联系是系统异常值与模型隐私之间的潜在联系。异常值作为高度特定、上下文感知的重要信息标记的功能角色使它们成为训练数据记忆的潜在机制的主要候选者。

当LLM记忆一段敏感数据，如个人身份信息(PII)或独特的受版权保护句子时，它必须创建一个可以可靠地重现该确切序列的内部表示。模型很可能通过将一个异常值通道专门用于该特定、罕见的信息片段来实现这一点。异常值的极端幅度将作为生成过程的强大且明确的信号。

这一假设为隐私审计领域创建了一个强大的桥梁。当前最先进的审计技术依赖于将称为"金丝雀"的独特数据片段插入到训练或微调数据中，然后使用成员推理攻击(MIAs)来查看模型是否已记住它们。这些方法的主要限制是设计"易于记忆"的金丝雀的难度。异常值框架为这一属性提供了一个新的、具体的定义：一个"易于记忆"的金丝雀是最有可能在训练期间触发新系统异常值形成的数据片段。

这表明了一种新颖且可能更强大的隐私审计方法。而不是依赖于对模型输出的黑盒查询，人们可以在微调期间直接监控模型的内部激活。一个新的、高幅度的异常值通道的出现，该通道与输入中金丝雀的存在强烈相关，可以作为记忆的直接、高保真信号。这种内部的、"白盒"方法可能导致比当前可能的更敏感和可靠的隐私审计，直接将异常值力学研究与隐私审计目标联系起来。

因此，异常值的研究超越了其在模型压缩中的起源。它提供了一个统一的视角，连接了Transformer的计算策略（稀疏性和异常值）、其优化景观的数学现实（Hessian结构）以及其最紧迫的社会挑战（隐私和安全）。这些不是独立的研究问题，而是同一个复杂系统的不同方面。模型用于其期望功能（隐式缩放）的机制和用于其不期望行为（隐私泄漏）的机

制可能是同一个：系统异常值。这种统一观点意味着在结构层面解决异常值问题可能为构建更安全和更值得信赖的AI系统带来深刻且可能意外的好处。

7 Conclusion and Future Research Directions

7.1 Summary of the State-of-the-Art

本综述综合的研究体系标志着对大型语言模型科学理解的重大演变。异常值现象已经从模型压缩的外围关注点转变为一个中心话题，提供了对Transformer架构基本工作原理的深刻见解。关键要点很明确：异常值不是随机噪声，而是LLMs的系统、功能上不可或缺和深刻影响的属性。

它们由自注意力机制中softmax函数的数学必要性生成，在那里它们作为隐式、上下文感知的缩放策略，用于控制信息流和稳定表示。然而，这种功能作用伴随着高昂的代价，为高效模型量化创造了主要瓶颈，并为训练过程引入了复杂性。研究社区的反应已经从天真和无效的移除技术发展到复杂的算法修复（如OS+）来管理症状，最后到预防性质的注意力机制结构重新设计，以解决根本原因。至关重要的是，异常值的研究提供了一个强大的统一镜头，连接了LLM研究的不同领域，包括计算效率、优化理论以及隐私和安全的关键领域。

7.2 Open Questions and Research Gaps

尽管取得了这一快速进展，但几个关键问题仍未得到解答，定义了未来调查的前沿。

* **消除的功能权衡**：虽然从结构上消除异常值在压缩和收敛方面显示出明显的好处，但功能权衡的全部范围尚不清楚。这些异常值是否在更复杂的涌现能力（如某种形式的上下文学习或抽象推理）中发挥微妙但重要的作用？需要彻底调查，以确保在解决效率问题的同时，我们不会无意中削弱模型最先进的能力。 * **预训练期间的完整生命周期**：当前分析主要集中在完全训练的模型或短微调运行中的异常值。它们在数万亿标记的大规模预训练期间的生命周期仍然是个谜。了解这些结构如何随着语言和推理能力的发展而出现、稳定和演变，可能为学习过程本身提供前所未有的见解。 * **异常值-隐私联系**：系统异常值通道与敏感数据记忆之间提出的联系是一个引人入胜但主要是理论性的假设。迫切需要严格的实证研究来测试和验证这种联系。此类工作将涉及在包含金丝雀的数据集上微调模型，并使用内部探测技术来确定新的异常值通道是否确实是记忆的主要载体。确认这种联系将对隐私保护机器学习产生深远影响。 * **超越Transformer**：所描述的异常值现象与基于softmax的自注意力机制密切相关。一个关键的开放问题是，在日益突出的替代架构（如状态空间模型（例如，Mamba）或其他非基于注意力的序列模型）中是否存在类似现象。了解这是否是大模型的普遍问题或独特的Transformer相关问题对于指导未来的架构创新至关重要。

7.3 Future Trajectories for Research

基于当前的知识状态和已识别的差距，未来研究的三个主要轨迹似乎最有前景。

1. **设计抗异常值架构**：未来工作最具影响力的途径在于继续开发新颖的注意力机制和替代架构，这些架构满足选择性、动态信息路由的要求，而不会产生数值问题异常值的副作用。这涉及超越修补现有的softmax机制，探索从根本上计算上下文重要性的新方法，可能导致本质上更稳定、更高效且更易于训练的模型。 2. **统一优化框架**：当前范式通常将模型性能、效率和安全性视为要顺序优化的单独目标（例如，为性能而训练，然后为效率而量化，然后为安全性而对齐）。未来研究应专注于创建同时优化这些目标的统一框架。在

这样的框架中，异常值管理将是一个中心支柱，而不是事后考虑，优化过程直接惩罚数值不稳定结构的形成，同时奖励功能有效性和隐私保护。 3. **基于异常值的可解释性工具**：如果异常值通道确实是模型标记其认为最重要的方式，那么它们代表了可解释性的强大信号。未来工作可以专注于开发工具来自动识别、可视化和分析这些异常值通道的行为。这样的工具可以允许研究人员针对任何给定输入问："模型认为什么信息是关键的？"答案可能提供对模型决策过程的前所未有的见解，帮助调试错误，理解偏见，并最终构建更透明和可靠的AI系统。

参考文献

[1] Yongqi An et al. Systematic outliers in large language models. *arXiv preprint arXiv:2502.06415*, 2025.

[2] Li Dong et al. Towards quantifying the hessian structure of neural networks. *arXiv preprint arXiv:2505.02809*, 2025.

[3] Giulia Puccetti et al. Understanding and mitigating numerical instabilities in transformer language models. *arXiv preprint arXiv:2110.13083*, 2021.

[4] Dong Wei et al. Outlier suppression+: Accurate quantization of large language models by equivalent and effective shifting and scaling. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1594–1608, 2023.