arXiv:2603.01426v1 [cs.CL] 2 Mar 2026

通过注意力动态理解大语言模型键值缓存压缩的内在机制

Samhruth Ananthanarayanan*
印度，印度理工学院德里分校
SAMHRUTH@GMAIL.COM

Ayan Sengupta*
印度，印度理工学院德里分校
AYAN.SENGUPTA@EE.IITD.AC.IN

Tanmoy Chakraborty
印度理工学院德里分校，印度
TANCHAK@IITD.AC.IN

摘要

随着大语言模型（LLM）的上下文窗口扩展至数十万词元，键值（KV）缓存已成为自回归解码中的主要内存瓶颈。这促使KV压缩方法迅速发展，相关方法声称可节省80%–90%的内存，同时在标准长上下文基准测试上的性能仅有轻微下降。我们认为，这类评估掩盖了一个更深层的结构性问题：注意力不仅是一种存储机制，也是一种路由机制；保留键值对并不意味着生成时就能在语义上访问它们。我们提出一个受物理学启发的框架，将KV压缩视为对自注意力中词元级路由的受控扰动，并区分信息保留、可访问性和利用。通过一系列受控合成数据集，探测多实体跟踪、实例消歧、共指一致性和多跳推理，我们发现了三个关键的经验现象。第一，中等压缩会显著损害内部表示，但任务准确率基本不变，揭示了较大的冗余裕度。第二，所有接受评估的架构都在压缩率接近90%时出现明显的幻觉“安全性断崖”，且与我们提出的全局驱逐率（GER）飙升高度相关；这表明，当答案关键词元被全局擦除时，语义可达性会发生相变。第三，架构差异会产生不同的路由动态：LLaMA模型表现为早期层共识和后期多样化，而Qwen模型则呈现漏斗式行为并在后期收敛，因此两者的压缩韧性特征不同。除了擦除之外，我们还发现了第二种失效模式——表征僵化：词元虽然存活，但注意力头级共识过强会使路由灵活性崩塌，尽管名义上的保留率不变，性能仍会下降。总之，这些发现为稀疏词元路由结构提供了经验证据，其存活决定了压缩容忍度；这将KV压缩重新界定为探测注意力几何的结构性手段，并将长上下文可扩展性与自注意力中的稀疏性及“彩票假设”联系起来。

关键词： KV缓存压缩、注意力分析、长上下文语言模型

1 引言

大语言模型（LLM）的近期发展，既由上下文长度所塑造，也由参数规模所塑造（Grattafiori 等，2024；Yang 等人，2025；Team 等人，2024）。

*。作者贡献相同。

©2026年 SamhruthAnanthanarayanan 和 AyanSenguptaand TanmoyChakraborty。许可协议：知识共享署名 4.0 国际版，见 https://creativecommons.org/licenses/by/4.0/。

上下文窗口已从几千个词元扩展到几十万个乃至更多，从而支持文档级推理、仓库级代码分析和持久对话记忆（Liu等人，2025c）。这一趋势映照了早期参数和数据方面的缩放定律（Kaplan 等人，2020；Hoffmann 等人，2022；Muennighoff 等人，2025），进一步强化了这样一种直觉：更大的范围能够带来更强的能力。然而，与在所有词元间共享的参数不同，在自回归解码过程中，键值（KV）缓存会随序列长度线性增长。对于一个包含 L 层、 H_{KV} 个维度为 D_h的注意力头、批次大小为 B、序列长度为 S且精度为 b的 Transformer，其内存占用按以下方式扩展

$$\text{Memory} = B \times S \times L \times H_{KV} \times D_h \times b. \tag{1}$$

随着 S 增加，KV存储会迅速占据推理内存的大部分，甚至常常超过参数存储本身。对于长上下文大语言模型而言，实际的主要瓶颈是内存而非计算（Meng等人，2025年）。

KV缓存压缩（Li 等，2024a）应运而生，成为解决这一限制的工程方案。基于量化（Hooper等人，2024年；Liu等人，2024年；Kang等人，2024年）、词元淘汰（Wan等人，2025年；Cai等人，2025年；Li等人，2024b）、自适应头剪枝（Feng等，2025）或聚类（Liu等人，2025b；Zhang等人，2025年）的方法取得了显著成效：KV缓存缩减幅度最高达80%–90%，而在 LongBench 和 RULER 等基准测试上的性能仅有限下降（Bai等人，2023年；Hsieh等人，2024年；Zhang等人，2024年）。这些发现助长了一种普遍看法：KV缓存中存在大量冗余，只要精心剪枝，就能移除其中大部分内容而不改变模型行为。我们认为，这一看法忽略了一个更深层的结构性问题：注意力究竟是一种路由系统，还是被动的记忆存储。

自注意力在注意力头和层之间构建动态的词元级计算路径（Clark等人，2019；Voita等人，2019；Elhage等人，2021；Olsson等人，2022）。归纳头分析和电路级研究表明，推理依赖于结构化的跨层路径，而非孤立的词元存储（Elhage等人，2021；Olsson等人，2022）。从这一视角来看，保留一个键值对并不保证它在功能上可访问；若路由路径被切断，词元可能仍被存储，却变得无法到达（Lee，2025）。反过来，关于稀疏子网络和“彩票假设”的理论研究表明，只要关键路径保持完整，相对较小的结构化注意力组件子集就可能足以维持模型行为（Frankle和Carbin，2019；Otsuka 等人，2025a）。这些见解促使我们重新将KV压缩视为对语义可达性的结构扰动。

这引出了我们的核心假设：稠密注意力层中存在稀疏的词元路径彩票票据（TR-LTs）——由最少量的跨注意力头和跨层路径构成，能够保留正确生成所必需的语义可达性。轻度压缩或许能显露这些潜在路径，而极端压缩则可能将其摧毁。如果这一假设成立，KV压缩带来的就不只是渐进式退化，还会引发语义可访问性的结构相变（Ma等，2026；Ersoy等，2025）。

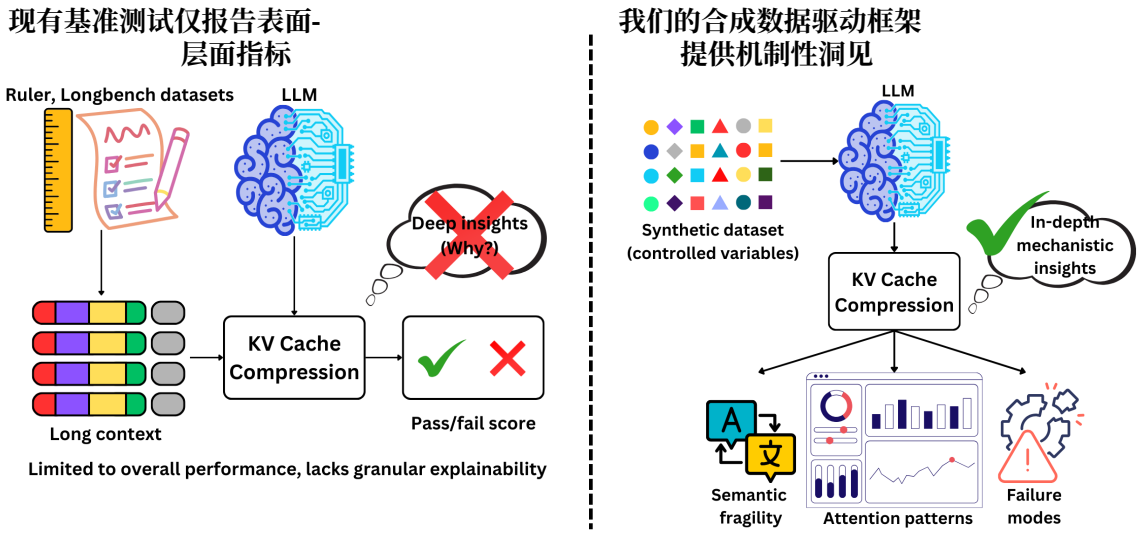


图 1：传统评估框架与合成评估框架在KV缓存压缩方面的比较。现有基准测试（左）主要报告粗粒度任务准确率，而我们的合成受控框架（右）则探究压缩条件下的结构可达性、路由崩溃和语义脆弱性。

核心研究问题

KV压缩仅仅会移除冗余记忆，还是会揭示并最终摧毁支配推理的潜在稀疏词元路径?

为探究这一问题，我们构建了一个受物理学启发的评估框架（Allen- Zhu，2024年），该框架基于受控合成数据集和结构指标（图 1）。我们不再仅评估答案准确率，而是引入可达性感知指标，例如全局驱逐率（GER），用于量化对答案关键词元的跨注意力头擦除；以及注意力头级共识，用于衡量路由协调性和灵活性。在多种架构（LLaMA 和 Qwen 系列）以及最高达 90% 的压缩级别下，我们观察到了一些引人注目的规律。

首先，中等压缩会造成显著的表征退化，但任务准确率基本保持不变，这揭示了较大的冗余裕度。其次，压缩率接近 90% 时，所有模型的幻觉率都出现急剧的安全性断崖，并与 GER 的激增高度相关，表明存在可达性阈值。第三，架构深度动态存在根本差异：LLaMA 模型呈现倒漏斗形模式（早期共识、后期多样化），而 Qwen 模型呈现漏斗形模式（早期探索、后期收敛），这意味着稀疏性分配依赖于架构。第四，探测分析揭示了压缩鲁棒性的语义层级：‘Person’和‘Thing’等主体实体的表征保持稳定，而‘Organization’、‘Location’，尤其是‘Creature’的表征则迅速退化。

主要发现

KV压缩会引发语义可达性的结构相变。只要稀疏令牌路由彩票票据仍然存在，推理就能持续；当全局淘汰或过度共识破坏了路由灵活性时，系统便会崩溃。

这些发现揭示了压缩条件下两种截然不同的失效模式：(i)表征擦除，即答案令牌在所有注意力头中都被整体驱逐；(ii)表征僵化，即令牌虽得以保留，但过度的注意力头共识阻碍了灵活利用。因此，压缩容忍度取决于注意力中的有效路由容量，而非原始令牌数量。

通过将实证压缩动态与自注意力中的稀疏性理论联系起来（Huang 等，2022年；Otsuka 等，2025b,a），本研究将 KV压缩重新界定为一种注意力结构探针，而非工程优化问题。我们不再追问在不损害准确率的前提下可以丢弃多少 KV缓存，而是提出一个更深层的问题：要使推理仍然成为可能，必须保留怎样的最小稀疏路由结构？我们认为，答案在于理解注意力：它是一个脆弱却强大的网络，其中蕴含着嵌入稠密 Transformer 中的潜在令牌路由彩票票据。

2 背景

2.1 自回归解码中的自注意力与 KV缓存

Transformer（Vaswani 等，2017）以自注意力机制为基础。给定输入序列 $X \in \mathbb{R}^{S\times d}$，每一层都会计算查询、键和值：

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V,$$

其中 W_Q，W_K，$W_V \in \mathbb{R}^{d\times d_k}$。缩放点积注意力定义如下：

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V.$$

在多头注意力（MHA）中，此操作会在 H 个注意力头上重复执行并拼接起来。在自回归解码中，每个时间步都重新计算所有先前词元的 K 和 V，会产生 $\mathcal{O}(S^2)$ 的开销。因此，模型会缓存先前计算出的键和值，形成*KV*缓存。在解码步 t，注意力是在当前查询 q_t 与所有已存储的键 $K_{1:t}$之间计算的，同时重复使用值 $V_{1:t}$ 。与在各词元间共享的参数不同，KV缓存会随着 S 和 B 线性增长，如公式 1 所示，因此在长上下文推理中成为主要的记忆消耗项（Li 等，2024a）。

2.2 高效注意力机制

标准自注意力的二次方 $\mathcal{O}(S^2)$ 时间和内存复杂度，推动了各类面向效率的架构的发展。这些架构通过牺牲精确性、表达能力或归纳偏置来提升可扩展性（Sun 等，2026）。其中一种研究思路是通过将注意力模式限制在预定义或学习得到的词元子集上来实现结构稀疏化，从而减少成对交互；代表性方法包括稀疏变换器

（Child，2019）、Longformer（Beltagy 等，2020）和 BigBird（Zaheer 等，2020）。这些方法利用局部性、块结构或随机连接，在降低计算成本的同时保留部分全局上下文。另一类方法则直接近似注意力矩阵：Linformer（Wang 等，2020）等低秩投影方法在 softmax 计算前降低维度；而 Performer（Choromanski 等，2021）等核化或特征映射方法则以可线性化近似取代 softmax 核，以近似误差为代价实现 $\mathcal{O}(S)$ 复杂度。相关的基于局部性的方法将注意力限制在滑动窗口或邻域结构内（Beltagy 等，2020），强化了对短程依赖的归纳偏置，但可能限制长程推理。近年来，Mamba（Gu 和 Dao，2024）等基于循环或状态空间的模型彻底摒弃显式注意力，转而采用结构化序列算子，以线性时间维持隐式记忆。总的来说，这些方法体现了全局表达能力、计算效率、近似保真度和硬件适配之间的不同权衡，说明提升注意力的可扩展性，既可以通过结构稀疏化、矩阵近似和架构重设计实现，也可以借助其他序列建模范式。

2.3 IO 感知精确注意力与 KV 共享机制

与基于近似的方法不同，FlashAttention（Dao 等，2022年）表明，无需改变 softmax 注意力的数学定义，也能显著提升性能。FlashAttention 不会将完整的 $S \times S$ 注意力矩阵写入高带宽内存（HBM），而是通过分块与融合重新安排计算，并完全在片上 SRAM 中执行分块 softmax 归一化和累加。这种 IO 感知的重新构造最大限度地减少了内存读写，而内存读写往往是现代加速器真正的瓶颈；同时，它在保持精确性的前提下实现了显著的加速。FlashAttention-2（Dao，2024年）进一步改进了工作划分、并行性和内核利用率，从而在 GPU 上实现更高的吞吐量和更好的扩展性。这些成果凸显了一个关键见解：注意力的效率往往更多受限于内存带宽和数据移动，而非算术复杂度。

与内核级优化正交的是，多查询注意力（MQA）（Shazeer，2019）和分组查询注意力（GQA）（Ainslie等人，2023）等架构修改，通过在多个注意力头之间共享键和值投影来降低推理内存占用。在标准多头注意力中，每个 H_Q 查询头都有自己的 K 和 V投影；MQA 将它们合并为一个共享 KV 头，而 GQA 则将这一做法推广为允许 H_Q 查询头共享较少数量的 $H_{KV} < H_Q$ KV 头。这样可按 H_{KV}/H_Q 的比例减少 KV 缓存内存，从而在自回归解码过程中大幅节省内存，同时仅对模型质量造成极小影响。现代大规模大语言模型广泛采用 GQA 或 MQA 变体，因为它们在内存占用、解码速度和表征灵活性之间实现了良好的权衡（Li 等，2024a）。总体而言，IO 感知精确内核与 KV 共享架构体现了提升注意力效率的互补策略：前者针对硬件层面的数据移动，后者则减少键值表征中的结构冗余。

2.4 KV缓存压缩

除了修改注意力内核或架构结构之外，近年来还有一类快速发展的研究直接在推理过程中压缩 KV 缓存，将其视为长上下文解码下的内存管理问题。现有方法可从若干相互正交的维度进行归类。基于量化的方法 （Hooper等人，2024；Yang 等人，2024b）通过训练后量化、混合精度方案或动态逐层策略，降低已存储键和值的位精度 b，以适度增加困惑度为代价，实现线性的内存节省。驱逐和重击者方法 会有选择地丢弃被认为不太可能影响未来预测的词元，并采用近期偏置、累积注意力质量、词元范数或学习得到的重要性预测器等启发式方法（例如 H2O（Zhang等人，2023）、StreamingLLM（Xiao等人，2024）、Scissorhand（Liu等人，2023）或 CurDKV（Sengupta等人，2025））；这些方法隐含地假设注意力贡献高度偏斜，只有少量词元会主导后续影响。分块和语义摘要 方法（Liu 等人，2025d；Bai等人，2024）将相邻或语义相关的词元聚类，并以压缩后的代表项取而代之，试图在减少存储的同时保留覆盖范围。与查询无关的压缩 机制（Kim等人，2025；Chari和Durme，2025）在不依赖当前查询的情况下预先压缩 KV 张量，利用词元间或层间的结构冗余。最后，自适应和预测方法 （Ge等人，2024；Zhou等人，2025；Feng等，2025；Qin等人，2025）将压缩视为预测问题，动态估计未来解码步骤需要哪些词元（例如基于预期注意力的方法（Devoto等人，2025）或 AdaKV 式方法（Feng等，2025）），从而使内存保留与预期路由需求相匹配。

在这些类别中，评估通常报告压缩比、延迟开销、困惑度或下游准确率的下降、在长文本或对抗性上下文中的鲁棒性，以及该方法是否无需训练或需要微调。实证证据表明，KV存储中存在大量冗余，在标准基准测试下，许多方法可实现70%–90%的缩减。然而，现有评估主要衡量总体检索准确率或任务级性能，隐含地将令牌保留等同于功能保留。它们很少区分存储、可访问性和利用，因此留下了一个更深层的问题：压缩后的缓存是否能在自回归生成过程中维持语义可达性所需的结构路由路径。

3 方法

3.1 数据集与受控KV压缩评估

标准的长上下文基准测试（如 LongBench（Bai等人，2023年）、InfiniteBench（Zhang等人，2024年）和 RULER（Hsieh等人，2024年））会评估模型在扩展上下文中的下游任务准确率，但其设计并非旨在隔离 KV压缩 对内部注意力路由造成的结构性影响。具体而言，这类基准测试混淆了三种不同现象：（i）信息是否仍存储在缓存中；（ii）信息是否仍能通过可行的注意力路径被访问；（iii）信息是否在生成过程中得到有效利用。模型可能因冗余信息弥补了路由受损而正确作答，也可能因无关的推理局限性而失败

，而非压缩引起的结构坍缩。正如大语言模型物理学研究脉络（Allen-Zhu and Li, 2025年、2024年）所强调的，要理解缩放行为，需要精心控制的合成实验设定，以揭示内部机制，而不能仅仅依赖自然基准。受这一观点启发，我们设计了一套合成数据集，将KV压缩 视为自注意力中词元路径的结构扰动。

我们的数据集构建遵循两种互补范式。首先，我们采用受控生成式合成，通过提示词向前沿大语言模型提供结构严谨的指令，以确保明确的实体-属性关联、提及频率约束和双向查询。其次，我们采用确定性的基于模板的构建方式，并通过槽位填充消除语言混杂因素，确保实验控制达到最大程度。这两种方法兼顾了真实感与可控性，同时保持路径敏感性。

设计原则。 每个数据集的构建都满足四项约束：(1)路径敏感性，正确答案取决于是否保留特定的跨层词元通路；(2) 最小冗余，防止激进驱逐后意外恢复信息；(3)双向查询，要求同时追踪主体→属性和属性→主体；(4) 失败可解释性，以便将失败归类为表征擦除或僵化。在这些原则下，压缩相当于对词元路径结构进行受控消融。

生成框架。 对于生成式数据集，我们使用结构化提示来确保属性-实体绑定和受控的语义关系。通用流程在算法1中进行了形式化。

算法1 受控合成样例生成

要求： 主题 s，结构约束 C，长度范围 L

1：向大语言模型提供提示词，要求遵循 C2：将实体名称替换为合成标识符3：生成段落 P4：依据双向关联规则生成问答对5：验证明确的答案依据锚定6：存储 $(P, \{(q_i, a_i)\})$

对于基于模板的数据集，样例通过对预生成的实体池进行确定性槽位替换来生成，并可选择性地进行受控扰动（例如代词替换）。这样既消除了表层变化，又保留了结构依赖。

数据集套件。 生成的数据集套件涵盖从短上下文到长文本上下文（150–1300+ 个词元），并针对不同的路由压力源：

- **基础任务：** 包含一个主题且属性紧密绑定的短段落；用于测试最小冗余条件下驱逐的精确性。
- **知识操纵：** 填槽式人物传记，用于测试最小分布的事实结构和系统性的驱逐偏差。
- **多重存在：** 需要进行实例消歧的重复实体提及；评估剪枝下的位置鲁棒性。

- **多实体：**多个语义关联实体；测试跨实体和跨头路由的完整性。

- **长上下文：** 需要在相距较远的文本片段之间进行分布式推理的扩展篇章；用于大规模探测路由容量。
- **共指：** 对照实验中加入代词扰动，并将“我不知道”设为真实标签；用于检测压缩过程中细粒度的路由失效和幻觉。
- **跳数：** 需要多跳推理的链式结构实体；用于评估顺序语义路由的保留情况。

下文总结了具有代表性的提示词和模板结构，以确保可复现性。表3重点展示了各项任务的示例，而表6则展示了附录 A 中的正向和反向示例

```
# Base task
生成一段关于 <主题> 的篇章（150-200 个词）- >=4直接关
联的属性 - 虚构实体名称 生成 6 个问答对，将主题 <与> 属
性相关联。
```

```
# 知识操纵（模板）<名 姓> 出生于 <日期>。
他们曾就读于 <大学>。他们的职业是 <职业>。
他们住在 <城市>。
```

```
# Multi presence / Multi entity / Long context
生成段落（400-1300 个词）- 多个实体或重复提及 - 明
确的属性关联 - 双向问答生成
```

```
# 共指
<名 姓> 出生于 <日期> ，出生地为 <地点>。<他/她> 就读于 <大
学>。
...
# 代词替换查询 -> 回答 = “我不知道”
```

```
# Hops
生成主题 + 3 关联实体 - 顺序语义关联 - 涵盖
多跳链的 16 个问答对
```

数据集统计。 表 1 总结了合成数据集套件的结构特征，包括平均篇章长度、每篇章查询数、篇章总数和查询总数。数据集涵盖了广泛的上下文体制，从篇幅较短且严格受控的篇章（例如基础任务和知识操纵），到超过一千词的长篇多跳叙事（长上下文和跳数）。这种对篇章长度和结构复杂性的刻意变化，使得我们能够系统地评估从轻度剪枝到极端驱逐等不同压缩比。重要的是，

Name	段落长度（词）	# 查询	段落总数	LLM生成的查询总数	
基础任务	175 ± 25	6	200	1199	Yes
知识操纵	32	13	4000	52000	No
多重存在	500 ± 100	10	100	1000	Yes
多实体	500 ± 100	10	100	1000	Yes
长上下文	1100 ± 200	10	100	1000	Yes
共指	44	9	8000	72000	No
Hops	1100 ± 20	16	100	1600	Yes

表 1：用于受控 KV 压缩评估的合成数据集套件的结构摘要。

该套件将大语言模型生成的篇章与确定性的基于模板的构造相结合：前者引入受控的语义多样性，后者则最大限度地提高实验精度。这种组合确保我们既能在自然变异性下，也能在严格受控的结构约束下分析压缩行为。

生成的合成数据集的重要性。与主要报告总体准确率的传统长上下文基准测试不同，我们的框架能够将压缩失败因果归因于注意力中的特定结构机制。基础任务和知识操纵数据集分别隔离了驱逐精度和实体-属性绑定稳定性。多重存在和多实体任务揭示了位置敏感性和跨注意力头路由的脆弱性。长上下文任务在较大的词元预算下对路径容量进行压力测试，而跳数任务则直接探查多步语义链的保留情况。共指任务评估细粒度的表征一致性和幻觉敏感性。这种分层结构能够将失败系统地分为表征擦除（所有通往关键信息的可行词元路径均被移除）和表征僵化（词元得以保留，但路由灵活性崩溃）。因此，压缩被分析为语义可达性的结构相变，而非准确率的单调下降。

3.2 标注框架

为超越汇总性能指标，我们使用结构化标注框架为每个问答对添加标注，以便在压缩条件下进行细粒度的行为分析。其目标是将准确率统计转化为可解读的信号，反映内部表征的鲁棒性。

标注沿着两个相互正交的维度进行：答案类型和问题难度。答案类型维度根据语义域对回答进行分类，使我们能够检验某些概念表征是否比其他表征更能抵御压缩。答案类别包括 人物、事物、组织、生物、地点、数字、日期/时间和 事件 （附录 A 中的表4）。通过沿这些语义维度分解性能，我们可以发现系统性的退化模式，识别表征稳定性的层级，并评估压缩是否会对特定类别的概念造成不成比例的影响。

第二个维度通过问题难度标签来体现推理需求。问题分为标准 （直接检索明确陈述的信息）、操纵（需要对上下文进行隐含解释或转换），或部分 （需要汇总分散在多个文本区域中的信息）。这一维度区分了

由词元丢失导致的失败与由推理复杂度导致的失败（附录 A 的表 5）。模型可能保留了必要的词元，却仍无法回答需要多步推理的高难度查询；反过来，它也可能因表征擦除而无法完成简单的检索任务。

这些标注轴共同定义了一个多维评估空间。性能分析不仅可以按数据集和压缩比进行分解，还可以按语义内容和推理结构进行分解。这种结构化分析有助于系统地识别压缩敏感领域，并深入解释各模型特定的鲁棒性特征。

3.3 实验设置

所有压缩实验均使用 NVIDIA 的 KVPress 库[1], 进行，该库为自回归解码期间基于评分和基于剪枝的 KV 缓存压缩提供了统一接口。对于所有模型，KV 压缩均在推理时应用，无需额外微调。我们系统地改变压缩比，从轻度剪枝（10%）到激进驱逐（最多移除 90% 的 KV 条目），并根据保留的 KV 预算报告性能。

评分与剪枝策略。我们采用 Expected Attention（Devoto等，2025）作为评分函数，通过汇总各解码步骤和各层的预期注意力权重来估算令牌重要性。这种逐层评分机制能够优先保留显著性较高的键值对，尤其是位于结构重要性较高的注意力头中的键值对。我们采用了两种互补的剪枝策略：

• **FINCH（Chunk）Press**（Corallo和Papotti，2024）通过将上下文划分为连续片段，并移除每个片段中得分最低的词元来执行块级剪枝。该方法计算效率高，并能保留文档中的粗粒度结构覆盖。

• **AdaKV Press** （Feng等，2025）通过同时对所有注意力头中的KV条目进行排序，并移除最低分条目（不考虑位置分组），来执行按头全局剪枝。该方法以增加开销为代价，实现了更细粒度的自适应压缩。

FINCH与AdaKV的结合支持在粗粒度结构剪枝和全局自适应的头感知压缩两种设置下进行评估，使我们能够分析路由崩溃究竟取决于局部块结构，还是全局头级重要性。

推理范式。 压缩在两种推理设置下进行评估 ：

• **与问题无关：** 模型首先接收完整上下文，并在无法访问下游问题的情况下执行KV剪枝。压缩完成后才提供查询。该设置确保驱逐决策独立于检索需求，并反映了现实的部署场景。

• **问题感知：**在剪枝前提供候选问题列表，使压缩决策能够依据预期的查询结构进行调整。此设置衡量的是：当路由可针对已知检索目标进行优化时，压缩容忍度的上限。

1. https://github.com/NVIDIA/kvpress/

比较这些情形可以区分压缩鲁棒性究竟源于路由器本身的冗余，还是源于查询条件剪枝策略。

模型。 我们评估了涵盖两个架构系列的五种经过指令微调的大语言模型：

- LLaMA-3.2 3B Instruct 和 LLaMA-3 8B Instruct（Grattafiori 等，2024年），
- Qwen-2.5 3B Instruct、Qwen-2.5 7B Instruct 和 Qwen-2.5 14B Instruct（Yang 等，2024a）。

这些模型涵盖了不同的参数规模（3B–14B），以及注意力设计和训练数据方面的架构差异，因此能够分析依赖于模型的压缩容忍度。所有检查点均从 Hugging Face 模型中心获取[2] ，且未经修改。

实现细节。 所有实验均使用 KVPress 文本生成流水线，在单张 NVIDIA RTX A6000 GPU 上执行。默认配置会强制采用与问题无关剪枝，除非明确进行覆盖。生成采用贪心解码，以避免采样带来的变化。性能在词元级进行评估，并使用 LLaMA-3 8B 分词器匹配答案，以确保不同模型和压缩设置下的分词一致。

这一设置确保性能差异完全源于KV压缩行为，而非解码随机性或分词器不一致，从而能够在不同压缩预算下对结构性失效模式进行受控分析。

4 结果

现在，我们将综合分析不同任务、架构和标签维度下的压缩行为。结果并未显示性能退化是单调下降的，而是揭示了语义可达性的结构化转变。在各个数据集中，反复出现三种现象：与稀疏路径选择相符的局部性能峰值、因架构而异的保留与操控能力分化，以及激进压缩下多跳推理的系统性崩溃。

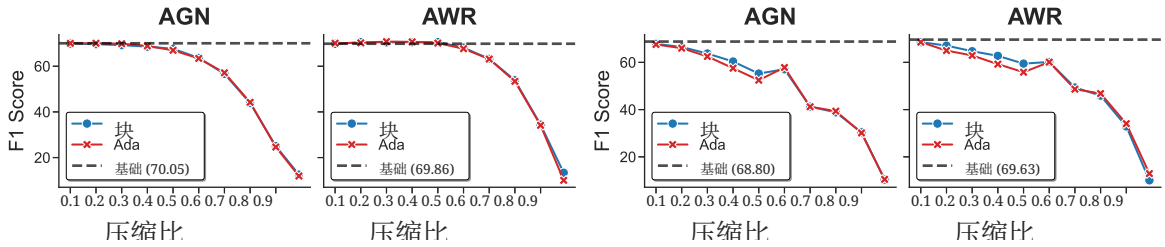
4.1 总体性能趋势

表 2 报告了所有任务在与问题无关（AGN）和问题感知（AWR）设置下的基础（0% 压缩）F1 分数。

模型	设置	Base	Know	多段落	多实体	Long	共指	Hops
LLaMA-3 8B	AGN	70.05	91.28	79.64	80.33	45.93	75.63	31.54
	AWR	69.86	92.09	77.66	81.93	45.54	60.90	31.25
Qwen-2.5 7B	AGN	68.80	92.20	73.63	77.46	38.68	78.89	27.13
	AWR	69.63	90.93	71.88	79.38	40.20	52.00	26.35

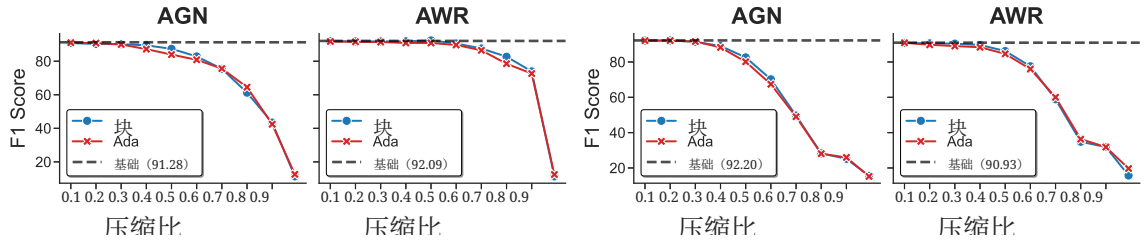
表 2：基础（0% 压缩）在各数据集上的 F1 分数。

2. https://huggingface.co/模型



（a）LLaMA-3 8B Instruct　　（b）Qwen-2.5 7B Instruct

图 2：不同压缩级别下的基础任务性能。中等压缩下出现的局部峰值表明存在稀疏子结构效应。LLaMA-3.2 3B、Qwen-2.5 3B 和 Qwen-2.5 14B 的对应结果见附录 B 的图 27。



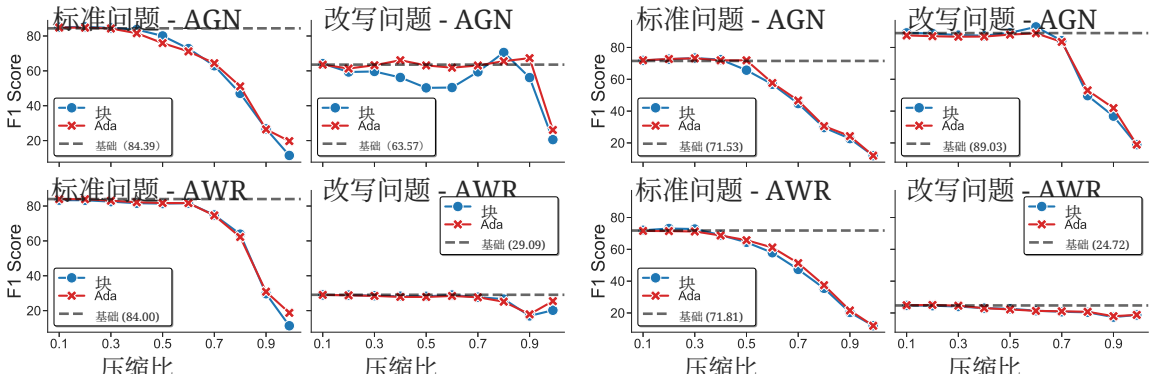
（a）LLaMA-3 8B Instruct　　(b) Qwen-2.5 7B Instruct

图 3：知识操纵结果。Qwen 在压缩下的性能下降更为平缓，尤其是在问题感知设置中。

基础任务。在基础任务（图 2）上，两种模型的 F1 分数均接近 70%（LLaMA：AGN 为 70.05，AWR 为 69.86；Qwen：AGN 为 68.80，AWR 为 69.63）。随着压缩程度增加，性能总体呈下降趋势；不过，问题感知设置表现出局部的非单调峰值，尤其是 Qwen 在中等压缩级别附近。这表明，适度剪枝可以移除产生干扰或冗余的 KV 路径，从而暂时改善对齐。这种现象与注意力中存在稀疏的 token-路径（Zhu et al., 2025）子结构相符；即使经过部分驱逐，这些子结构仍能保持功能完整。

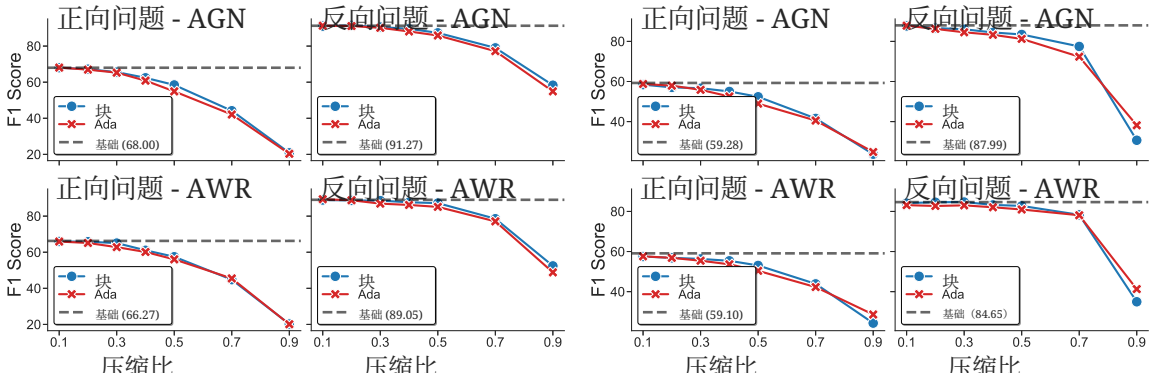
知识操纵 呈现出一种对比鲜明的模式（图 3）。在问题无关设置中，Qwen 略微超过 LLaMA（92.20 对 91.28）；而在问题感知中，LLaMA 略强一些（92.09 对 90.93），但 Qwen 在更高压缩率下的性能下降更为平缓。该任务强调的是结构化转换而非纯粹检索，这表明 Qwen 的鲁棒性源于指令条件推理，而 LLaMA 的优势则在于稳定的事实保留。值得注意的是，两个模型在压缩率约达 40% 之前，F1 均保持在 90% 以上。

共指 （图 4）显示 AGN 与 AWR 之间存在明显差异。LLaMA 的总体表现从 75.63（AGN）降至 60.90（AWR），而 Qwen 的总体表现从 78.89（AGN）降至 52.00（AWR），下降了 26.89 分。根据预期查询进行剪枝似乎会使模型倾向于过早作出承诺，



（a）LLaMA-3 8B Instruct　　（b）Qwen-2.5 7B Instruct

图 4：不同设置下的共指性能。问题感知剪枝显著增加了过度自信错误。



（a）LLaMA-3 8B Instruct　　(b) Qwen-2.5 7B Instruct

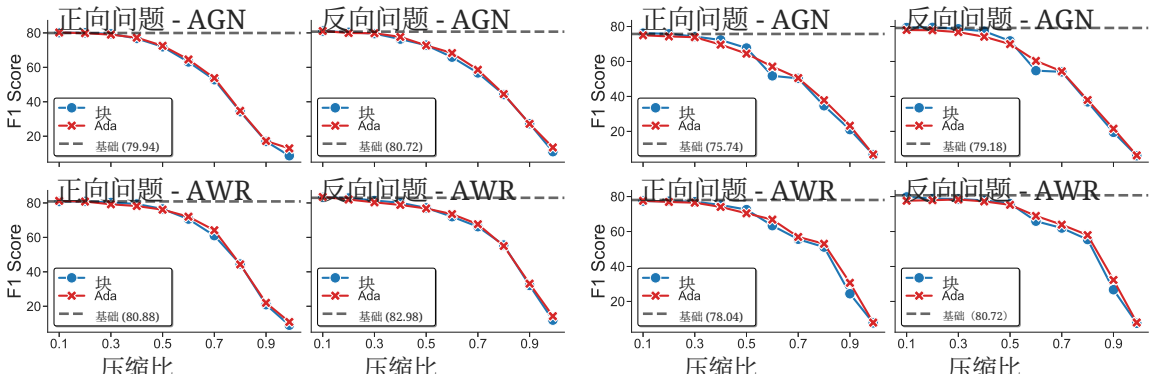
图 5：多重存在的正向与反向不对称性。

降低它们在引入不一致时正确弃答的能力，这补充了 Jin 等人（2025年）所得的结果。

多重存在 图 5 展示了 LLaMA 的表现优于 Qwen（AGN 总体得分分别为 79.64 和 73.63），但在压缩条件下，正向查询的性能下降幅度大于反向查询（AGN 中正向和反向得分分别为 68 和 91.27；AWR 中分别为 66.27 和 89.05），这表明实体重复时存在非对称路由脆弱性。

多实体则表现出更均衡的行为（LLaMA 的 AGN 总体得分为 80.33；Qwen 的 AGN 总体得分为 77.46），这表明将属性分散到不同实体可减轻干扰（AGN 中正向和反向得分分别为 79.94 和 80.72；AWR 中分别为 80.88 和 82.98）。

长上下文和跳数 揭示了更深层的结构性局限（图 7、图 8 和图 9）。即使不进行压缩，长上下文的 F1 仍然不高（LLaMA：AGN 为 45.93；Qwen：AGN 为 38.68），而跳数任务的表现更低（LLaMA：AGN 为 31.54；Qwen：AGN 为 27.13）。多跳推理



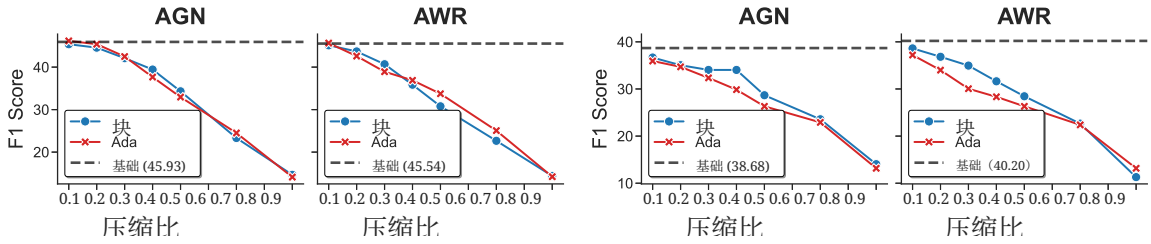
（a）LLaMA-3 8B Instruct　　（b）Qwen-2.5 7B Instruct

图 6：多实体结果显示方向不对称性有所减弱。LLaMA-3.2 3B、Qwen-2.5 3B 和 Qwen-2.5 14B 的对应结果见附录 B 的图 28。

随着压缩程度增加，性能下降速度快于直接检索（Li 等人，2025年），这表明关系路径比词元存在更脆弱。

结构摘要

(1) 中等压缩可通过隔离稀疏令牌路径子结构，暂时提升性能。(2) 架构差异决定了稳健性究竟源于密集保留（LLaMA），还是指令条件推理（Qwen）。(3) 多跳和长程推理在令牌完全丢失之前，结构上就已崩溃，这表明问题在于语义连通性失效，而非简单地移除记忆。

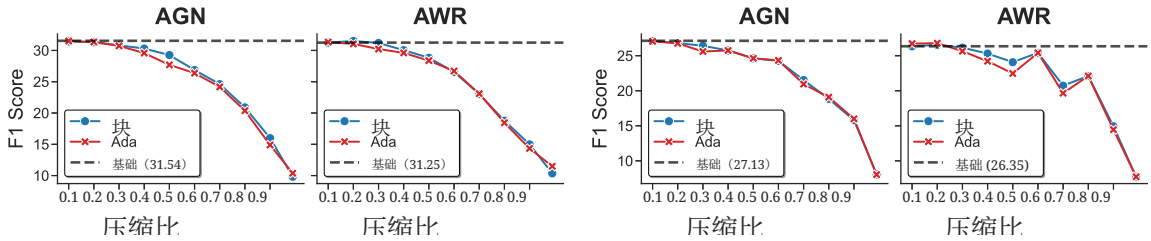


(a) LLaMA-3 8B Instruct　　(b) Qwen-2.5 7B Instruct

图 7：不同压缩级别下长上下文的性能退化。

4.2 标签级分析

虽然总体 F1 趋势揭示了压缩条件下的结构退化，但标签层面的分析揭示了这种退化所沿循的语义与认知维度。



(a) LLaMA-3 8B Instruct (b) Qwen-2.5 7B Instruct

图 8：Hops任务表现。多步语义链接在压缩下迅速崩溃。

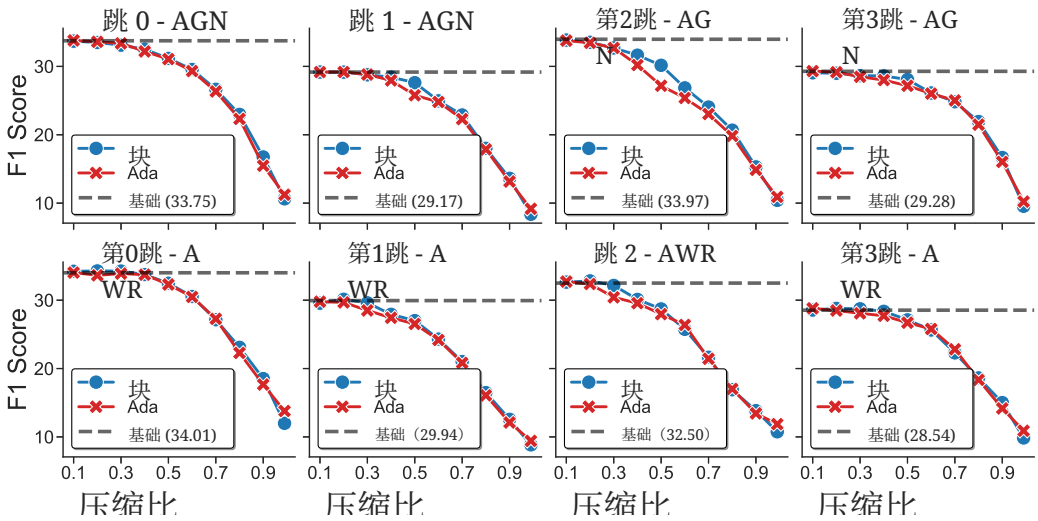
通过根据答案类型和问题类型标签分解性能，我们可以区分简单的令牌丢失与更深层次的表征脆弱性。在各个数据集中，压缩并非均匀影响所有语义类别；相反，它选择性地侵蚀关系性、层级性以及形态复杂结构。

答案类型标签。 图 10 展示了基础任务的答案类型表现。尽管在压缩率为 0% 时，整体F1分数仍接近 70%，但标签层面的趋势揭示了系统性差异。人物 和地点 等类别在不同压缩级别下相对稳定，与 Liu等人（2025a）得出的相似结果相辅相成。这些实体通常以具体名词的形式出现，表层形式一致，因此能够在 KV缓存中直接进行词汇锚定。即使剪枝移除了一部分词元，保留下来的提及通常也足以支持正确检索。

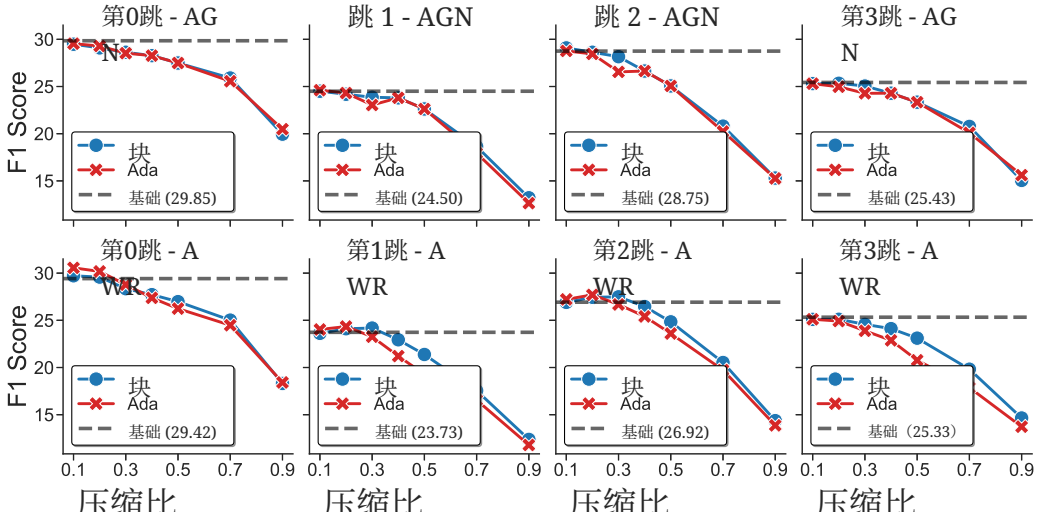
相比之下，事件 的性能下降明显更为剧烈。事件答案需要对动词变位、体标记和释义动作短语进行归一化。与原子名词不同，动词分布在形态各异的词元和句法语境中。在压缩过程中，这种分布式编码会变得支离破碎，导致语义重构不完整。与基于名词的类别相比，事件标签的性能曲线始终显示出更早出现的拐点，这表明关系和过程级表征比实体存储更加脆弱。

组织标签揭示了另一种明显但相关的弱点。组织实体往往编码着嵌套关系结构，包括子实体、角色或层级关系。在问题感知设置中，与问题无关条件相比，两个模型在组织标签答案上的表现下降更为明显。这表明，查询条件剪枝可能会将层级区分压缩成粗粒度标签。在压缩过程中，结构化表征会被简化，细粒度关系链接也最先受到损害。这种失败模式表明，损失的不只是词元，还有结构分解。

多实体标签的结果进一步印证了这一解释，如图 11 所示。事物和 组织 标签呈现出更加不规则的退化曲线，尤其是在中高压缩率下；而 人物 和 地点 则相对稳定。这种不对称性表明，压缩对需要跨词元关系绑定的类别造成的损害，要大于对由孤立词汇锚点表示的类别造成的损害。



（a）LLaMA-3 8B Instruct

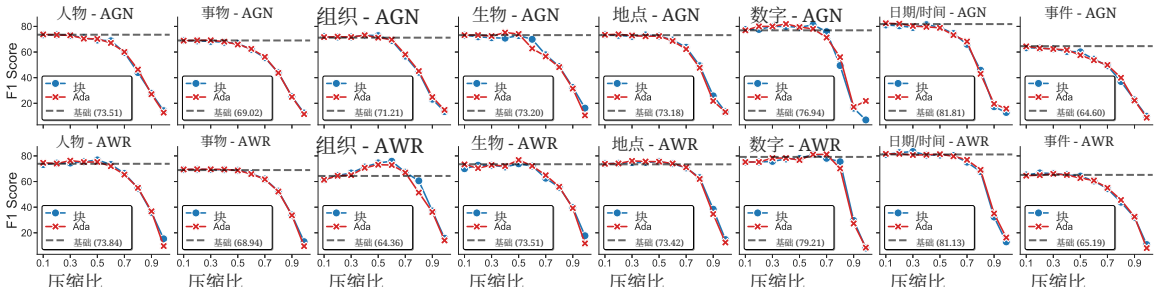


（b）Qwen-2.5 7B Instruct

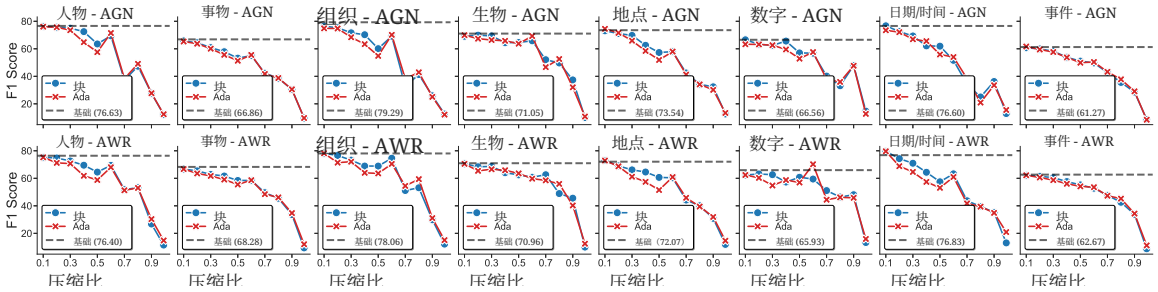
图 9：各跳的单独分解显示，中间语义联系存在不稳定性。

知识操纵进一步揭示了架构上的差异。如图 12 所示，Qwen 在不同标签上的退化程度各不相同，而 LLaMA 则呈现出更同步的衰退。这表明，Qwen 在不同语义领域间分配表征能力并不均衡，因此在压缩过程中能更久地保留某些类别；而 LLaMA 的保留则更加均匀，但在压缩达到极高程度时会更突然地崩溃。

问题类型标签。 问题类型标签揭示了压缩如何影响认知需求。在基础任务（图 13）中，标准 检索问题在中等压缩之前相对稳定。这些任务需要直接提取明确陈述的事实，因此主要依赖令牌存活。



(a) LLaMA-3 8B Instruct



(b) Qwen-2.5 7B Instruct

图 10：基础任务中的答案类型标签性能。LLaMA- 3.2 3B、Qwen-2.5 3B 和 Qwen-2.5 14B 的对应结果见附录 B 的图 29。

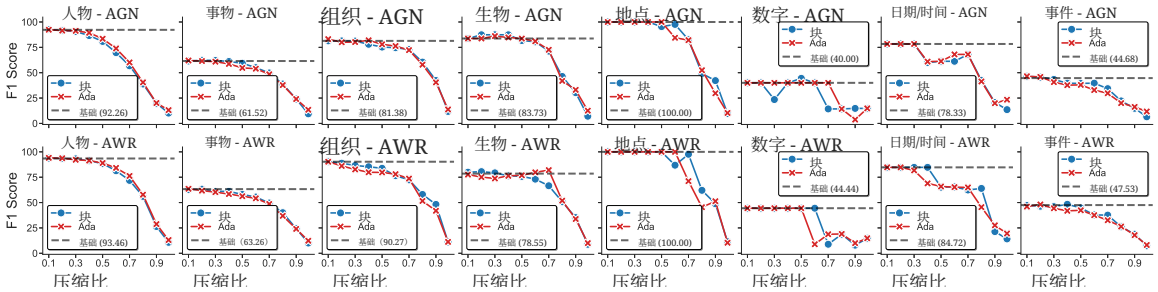
操纵 类问题需要转换或重新解释上下文，因此能体现出架构差异。Qwen 在 AGN 和 AWR 设置下保持了更高的一致性，表明其指令条件化能力稳健。LLaMA 虽然在原始检索方面表现出色，但在操控类任务中，AWR 设置下的差异更加明显。

部分 检索问题最早且最不规则地出现性能下降。这些查询需要整合上下文不同区域中分散的短语。即使单个词元在驱逐后得以保留，连接它们的关系路径也会变得脆弱。部分问题的性能曲线呈现出锯齿状变化，而非平滑下降，这表明发生的是间歇性的路由失效，而不是统一的容量缩减。

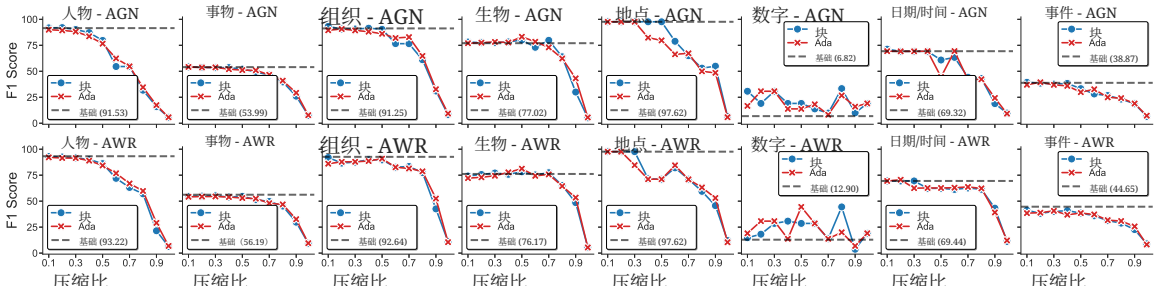
知识操纵问题标签（图 14）突显了 Qwen 在操控 查询方面的优势：在压缩条件下，其表现更稳定地优于 LLaMA。然而，在纯标准 检索中，LLaMA 仍保持优势。这进一步印证了这样一种解释：两种架构以不同方式分配表征资源——一种针对指令条件灵活性进行了优化，另一种则针对稠密存储和直接召回进行了优化。

在多实体（图 15）中，部分 检索再次表现出最显著的不稳定性。性能在不同压缩级别之间波动，表明存在一种阈值现象：分布式语义线索会间歇性地变得无法访问。这种不稳定性与表征僵化的概念相符：词元依然存在，但路由灵活性有所降低，无法实现稳定的多跨度整合。

标签层级模式的综合分析。在答案类型和问题类型这两个维度上，呈现出三种一致的模式。首先，关系型和层级型类别（事件、组织



(a) LLaMA-3 8B Instruct



(b) Qwen-2.5 7B Instruct

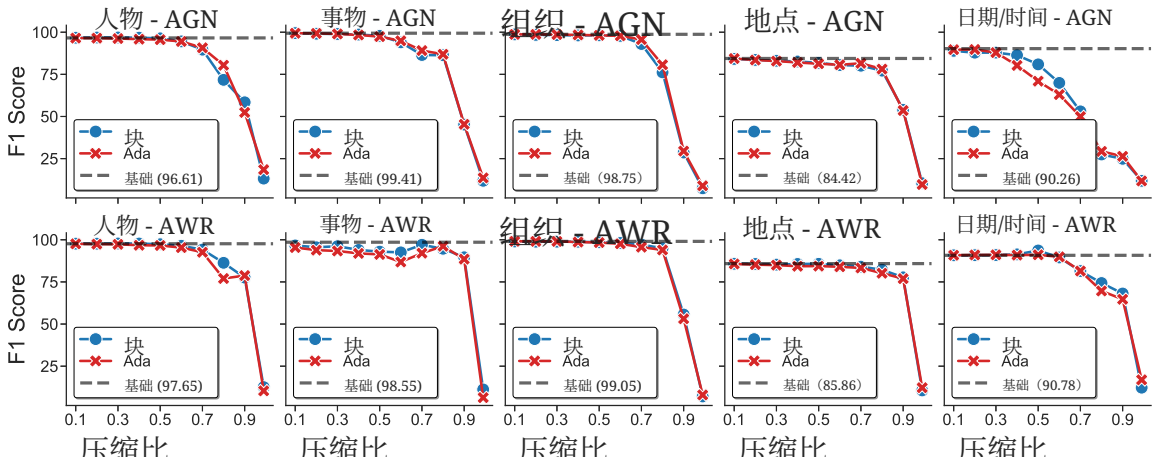
图 11：多实体中的答案类型标签行为。LLaMA-3.2 3B、Qwen-2.5 3B 和 Qwen-2.5 14B 的对应结果见附录 B 的图30。

、部分）比原子实体类别更早退化。其次，压缩会与架构特定偏差相互作用：Qwen 能更长时间地保持操控鲁棒性，而 LLaMA 则能维持更强的直接保留能力。第三，退化往往表现为不稳定，而非平滑衰减，这表明令牌路由连通性发生了相变，而不是容量线性缩减。

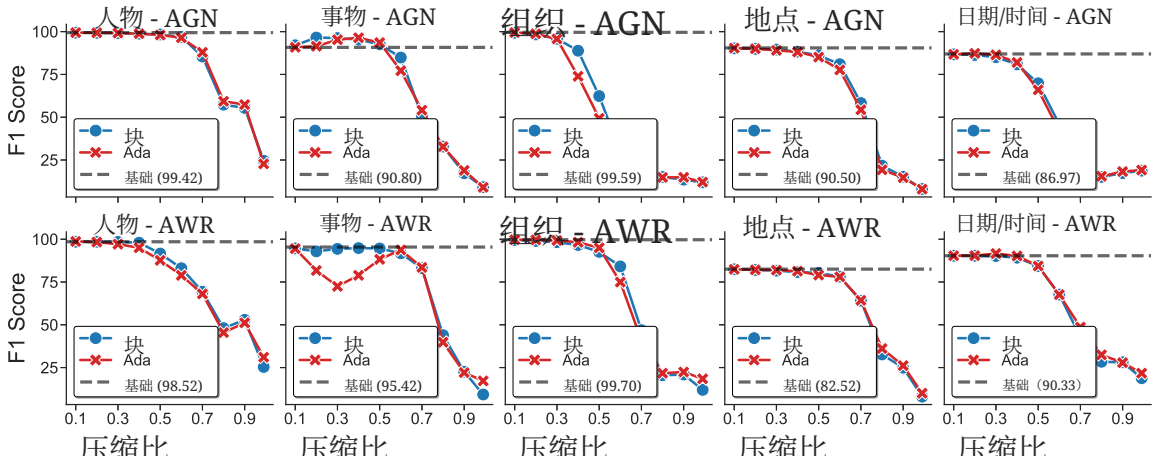
这些发现进一步支持了这样的解释：KV压缩会先侵蚀语义连通性，然后才消除词元存储。因此，标签级结果为将压缩视为词元路由图的结构扰动提供了实证支持，其中关系类别是路由崩塌的早期指标。

5 讨论

我们将 KV 缓存视为一种路由基底：压缩扰动的不只是记忆大小，还包括将证据传递至解码器的词元级路径是否存在，以及这些路径是否可被利用。在我们的各项分析中，有两个量起主导作用：(i)可达性 ——答案关键词元是否仍能通过任何 存留的逐注意力头路径被访问；以及 (ii)适应性 ——当上下文变得稀疏时，模型是否保留了足够的注意力头多样性来重新路由注意力。我们围绕这些机制展开讨论，并将其与架构深度剖面、普遍安全悬崖，以及表征—行为差距的基于探测的证据联系起来。

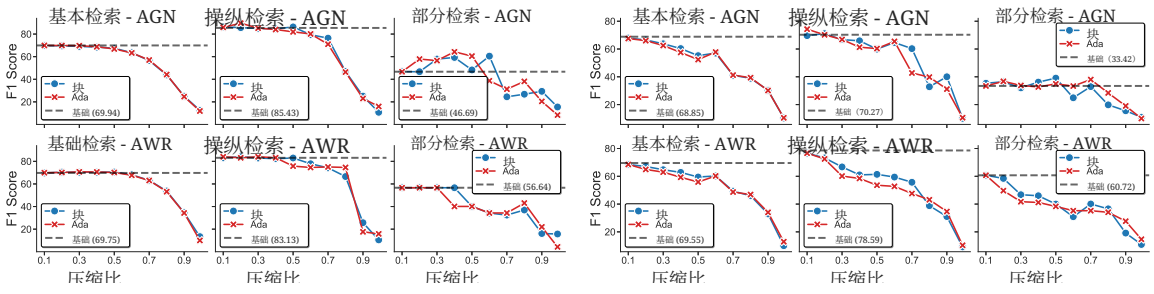


(a) LLaMA-3 8B Instruct



(b) Qwen-2.5 7B Instruct

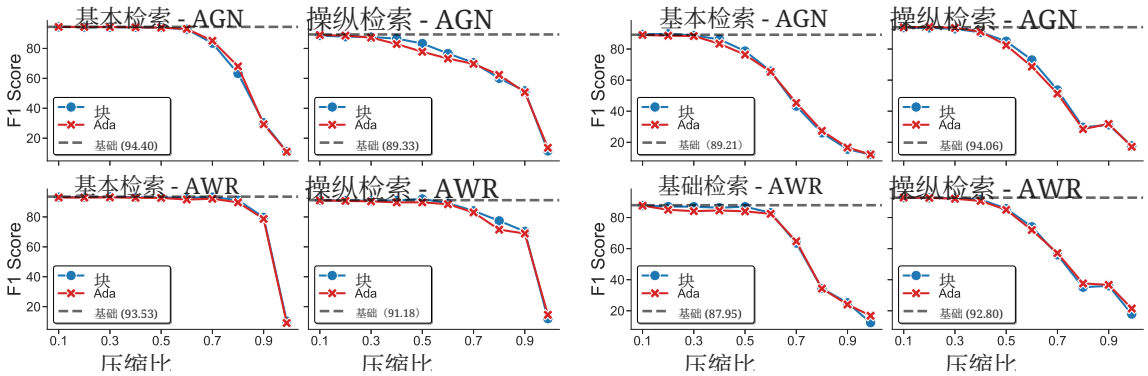
图 12：知识操纵的答案类型标签级行为。



(a) LLaMA-3 8B Instruct

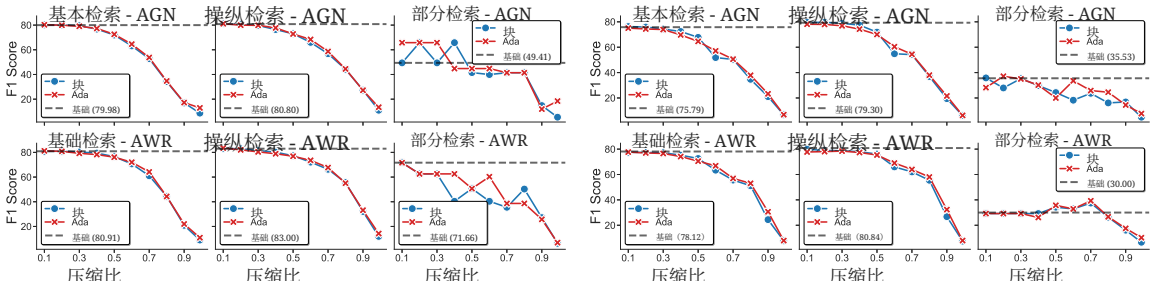
(b) Qwen-2.5 7B Instruct

图 13：基础任务的问题类型标签。LLaMA-3.2 3B、Qwen-2.5 3B 和 Qwen-2.5 14B 的对应结果见附录 B 的图31。



(a) LLaMA-3 8B Instruct (b) Qwen-2.5 7B Instruct

图 14：知识操纵问题类型分析。



(a) LLaMA-3 8B Instruct (b) Qwen-2.5 7B Instruct

图 15：多实体的问题类型行为与基础任务呈现相似趋势，但在问题无关设置下曲线平滑得多。问题感知设置下的曲线则更加崎岖，并出现明显的显著下降点。LLaMA-3.2 3B、Qwen-2.5 3B 和 Qwen-2.5 14B 的对应结果见附录 B 中的图32。

5.1 KV压缩的结构指标

压缩可能通过抹除必要证据，或虽保留证据却未能有效利用而导致失败。为区分这两种影响，我们跟踪驱逐（存活）和共识（协调）。设 T 为上下文令牌集合， H 为注意力头集合。对于令牌 t、压缩级别 α、注意力头 h和存活令牌集合 $t^{(\alpha,h)}$，定义

$$S_h(t,\alpha) = \begin{cases} 1 & \text{if } t \in t^{(\alpha,h)}, \\ 0 & \text{otherwise.} \end{cases} \qquad \text{GlobalEvicted}(t,\alpha) = \mathbb{1}\Big[\sum_{h\in H} S_h(t,\alpha) = 0\Big].$$

驱逐率是

$$\text{EvictionRate}(\alpha) = \frac{1}{|T|}\sum_{t\in T} \text{GlobalEvicted}(t,\alpha),$$

以及 t ask感知的全局驱逐率（GER）仅限于与答案相关的词元 T_{an} s:

$$\text{GER}(\alpha) = \frac{1}{|T_{\text{ans}}|} \sum_{t \in T_{\text{ans}}} \text{GlobalEvicted}(t, \alpha).$$

GER 衡量证据层级上的路径删除： 当 GER 较高时，模型无法通过任何存活路径访问真实值词元，因此在结构上更容易产生幻觉（图24）。仅有存活路径还不够；模型还必须协调其剩余能力。设 $A_{\ell,h}(t)$ 为注意力头 h 在层 ℓ 对词元 t的归一化注意力权重，其中注意力最高的词元

$$t^*_{\ell,h} = \arg\max_t A_{\ell,h}(t).$$

我们将逐层注意力头共识度量为

$$\text{Consensus}(\ell) = \frac{\left|\{t^*_{\ell,h}\}_{h \in H}\right|}{|H|},$$

其中，数值较低表示一致性更强（许多注意力头聚焦于同一个词元），数值较高则表示多样性更强。结合驱逐与共识，我们可以区分表征擦除 （路径已被摧毁）和表征僵化 （路径仍然存在，但适应性不足）。

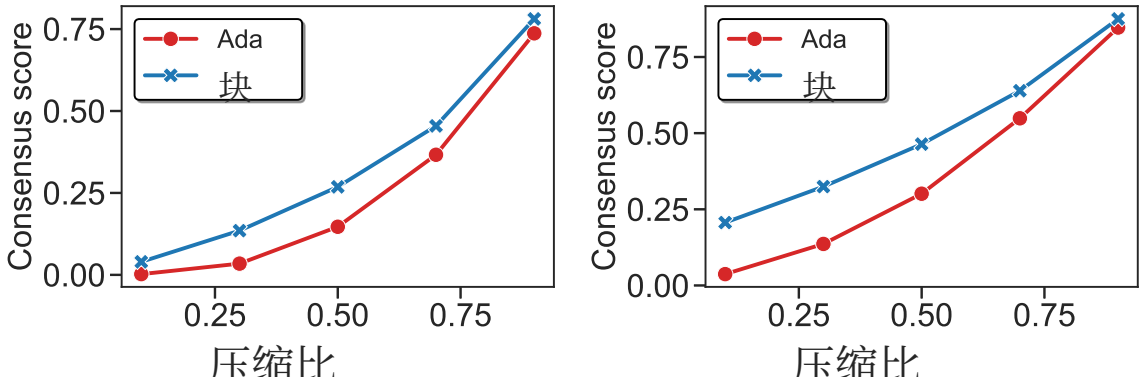
5.2 KV压缩揭示了不同架构如何沿深度分配计算?

LLaMA 和 Qwen 系列之间呈现出一致的架构反转：LLaMA 往往较早趋于稳定、随后增加多样性，而 Qwen 则倾向于早期探索、后期巩固。这一点从逐层共识趋势（图16）以及共识模式在压缩率不断提高时的演变中（图 17 和图 18）均可看出。从机制上看，这些特征意味着不同的“决策深度”：在 LLaMA 中，较早层的共识可以起到归一化作用，为后续专门化提供支持；在 Qwen 中，深层巩固则可以发挥后期决策级联的作用，将路由集中到一小组词元上。

这种深度分配对压缩设计有直接影响：假设存在普适“金字塔”（熵随深度单调递减）的策略，跨模型系列迁移效果可能不佳，因为脆弱的计算可能出现在不同深度。注意力热图进一步表明，在中等压缩率下，剪枝策略对词元路径的扰动各不相同（分块式连续移除与 Ada 式非连续移除）；但在极端压缩率下，它们最终都会造成类似的大规模路径破坏（图 19 和图 20）。

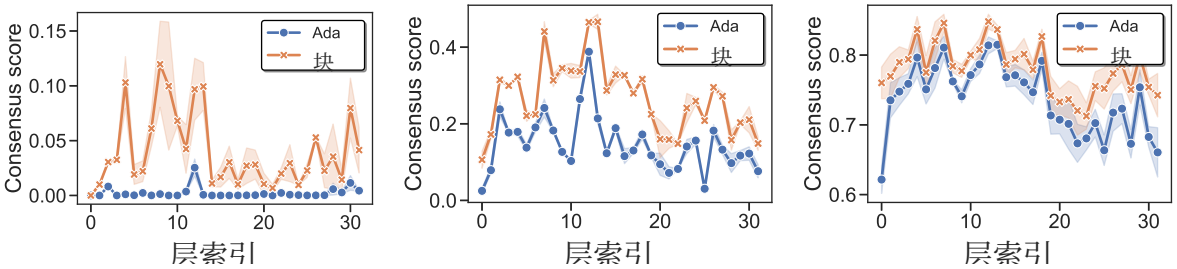
观察：压缩必须适应架构

图16–18表明，最主要负责稳定化与整合的层因模型系列而异，因此，同一个用于剪枝的“深度先验”对某个模型可能是安全的，对另一个模型却可能造成脆弱性。



(a) LLaMA-3 8B Instruct (b) Qwen-2.5 7B Instruct

图16：跨层共识。LLaMA 呈现出早期稳定和后期专业化；Qwen 则呈现出早期探索和后期整合。LLaMA-3.2 3B、Qwen-2.5 3B 和 Qwen-2.5 14B 的对应结果见附录 B 的图33。



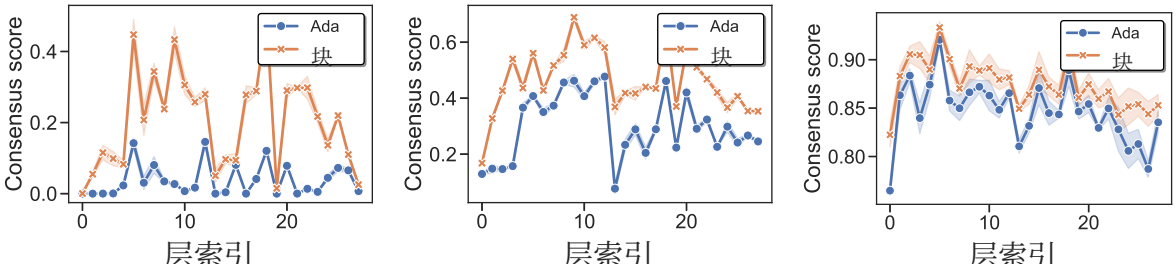
(a) LLaMA-3 8B Instruct 10% 压缩 (b) LLaMA-3 8B Instruct 50% 压缩 (c) LLaMA-3 8B Instruct 90% 压缩

图17：压缩条件下 LLaMA 的逐层共识。早期一致性得以保持，深度方向多样化则支持并行特征通道。LLaMA-3.2 3B 的对应结果见附录 B 中的图34。

5.3 为什么“安全”的 KV压缩在极端驱逐下会突然失效?

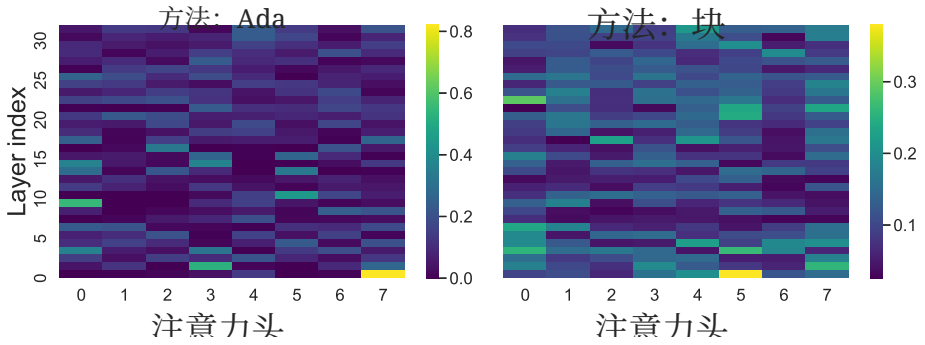
在不同模型和压力下，我们观察到一种普遍的安全悬崖，出现在 $\alpha \approx 0.9$ 附近：当大约 90% 的 KV条目被移除后，幻觉率便会激增（图22）。关键转变并非表征“质量”平缓下降，而是全局路由删除的概率急剧上升：与答案相关的词元会在所有注意力头中同时被驱逐，从而不再有任何通往证据的路径。这解释了为何驱逐曲线在 α 范围内仍可相对稳定（图21），而错误率却呈现明显的非线性变化（图22），也解释了 GER 为何与幻觉高度相关（图24）。

与此同时，即使 GER 较低，也会出现一些失败，这表明存在刚性：如果注意力过度集中于一个狭窄的焦点集合，那么当该集合被剪枝，或剩余证据与之关联较弱时，模型可能无法有效地重新路由。这

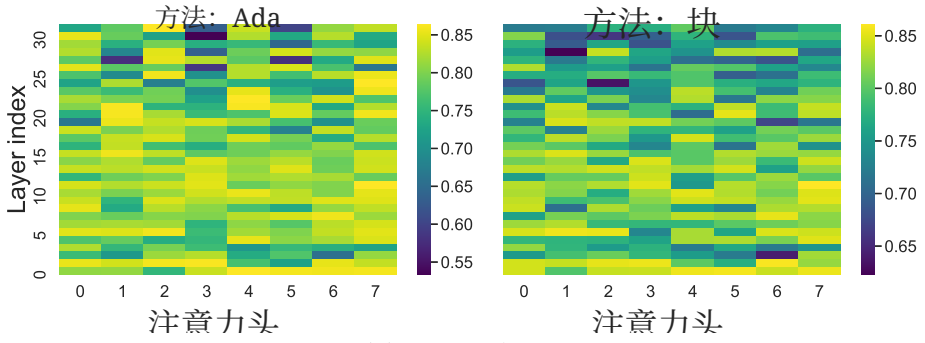


(a) Qwen-2.5 7B Instruct 10% 压缩　(b) Qwen-2.5 7B Instruct 50% 压缩　(c) Qwen-2.5 7B Instruct 90% 压缩

图 18：压缩下的 Qwen 逐层共识。多样性在较深层仍然存在，并在后期才崩塌，这与尾部阶段以巩固为主的情况一致。Qwen-2.5 3B 和 Qwen-2.5 14B 的对应结果见附录 B 的图 35。



(a) 10%压缩

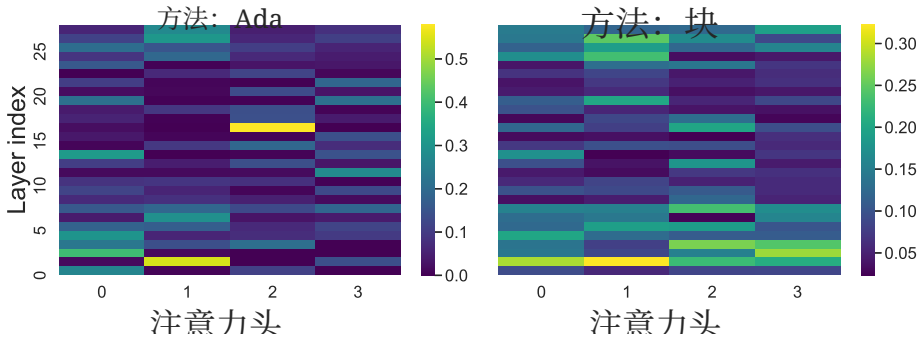


（b）90% 压缩

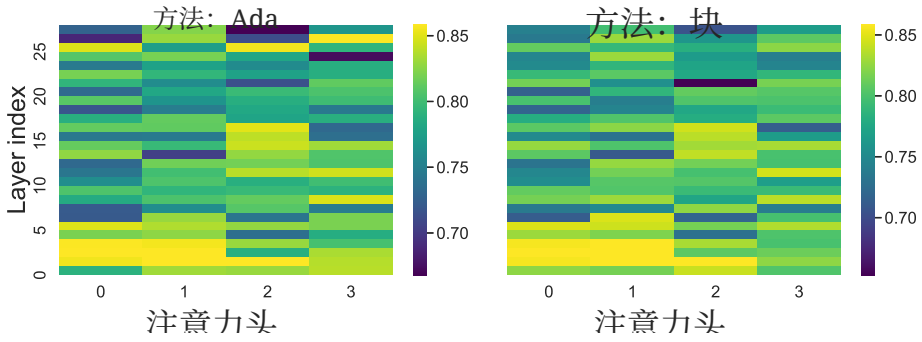
图19：压缩条件下 LLaMA 的注意力热图。块剪枝会移除连续片段；Ada剪枝则更加不规则。在 0.9 时，全局路径破坏占主导。LLaMA-3.2 3B 的对应结果见附录 B 的图 37。

在后期整合较强的架构中，这种情况最为可能（图 16 和图 18）；在这些架构中，令牌存活可能与利用不足并存。

算法2 KV压缩下的线性探测 **要求：** 数据集 $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$；模型 f；压缩 α；压缩配置 Π；探测层$\mathcal{L}$；训练/验证索引 $\mathcal{I}_{tr}, \mathcal{I}_{va}$**1：** 对于 $i = 1$ 到 N，执行 **2：** 在 x_i 上运行 f，使用 KVPress (Π, α)；记录 $\{h_{i,\ell}\}_{\ell \in \mathcal{L}}$ **3：** $z_i \leftarrow$池化$(\{h_{i,\ell}\}_{\ell \in \mathcal{L}})$ **4：** 结束循环 **5：** 在 $\{(z_i, y_i)\}_{i \in \mathcal{I}_{tr}}$ 上训练线性 $g(z) = \text{softmax}(Wz + b)$，并冻结 f **6：** 在 $\{(z_i, y_i)\}_{i \in \mathcal{I}_{va}}$ 上评估宏平均 F1



(a) 10% 压缩

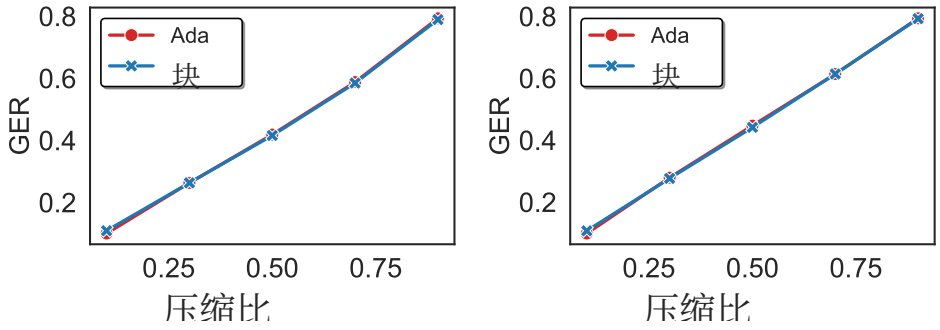


(b) 90% 压缩

图20：压缩状态下 Qwen 的注意力热图。低压缩率下可以观察到策略差异，但在 0.9 时，这些差异被大规模剪枝所掩盖。Qwen-2.5 3B 和 Qwen-2.5 14B 的对应结果见附录 B 的图38和图39。

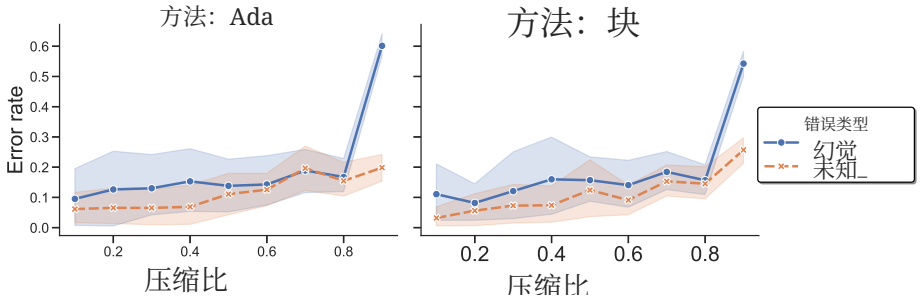
观察：两种机制解释了安全性断崖

图24和图16共同支持两种失效模式：**(i) 擦除**（高 GER 会移除所有证据路径），以及 **(ii) 刚性**（路径得以保留，但巩固限制了重新路由）；这两种模式在 $\alpha \approx 0.9$ 附近都表现得尤为明显（图22）。

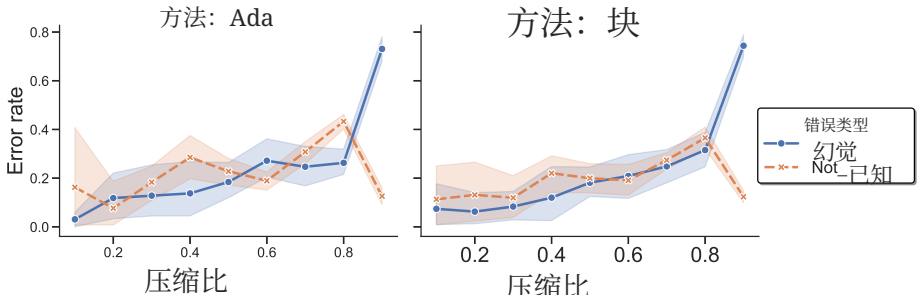


(a) LLaMA-3 8B Instruct (b) Qwen-2.5 7B Instruct

图21：驱逐率与压缩率的关系。除非明确进行条件化，否则剪枝总体上与问题无关。LLaMA-3.2 3B、Qwen-2.5 3B 和 Qwen-2.5 14B 的对应结果见附录 B 的图 42。



(a) LLaMA-3 8B Instruct

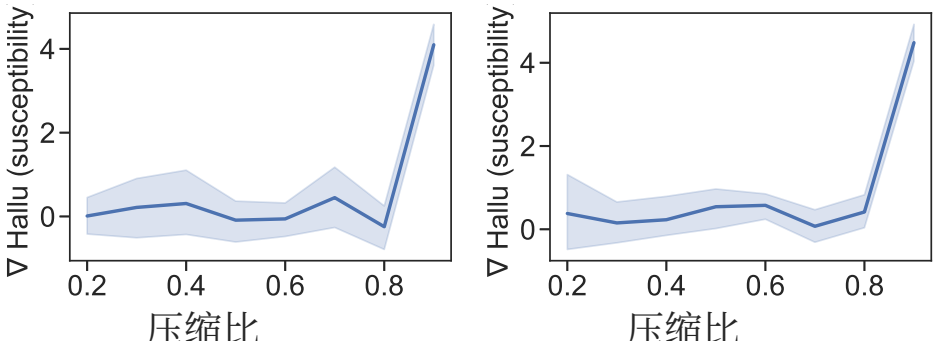


(b) Qwen-2.5 7B Instruct

图22：错误率与压缩率的关系。幻觉在 $\alpha \approx 0.9$ 附近出现悬崖；在 AdaKV 下，“未知”行为更加反复无常。LLaMA-3.2 3B、Qwen-2.5 3B 和 Qwen-2.5 14B 的对应结果见附录 B 的图 43。

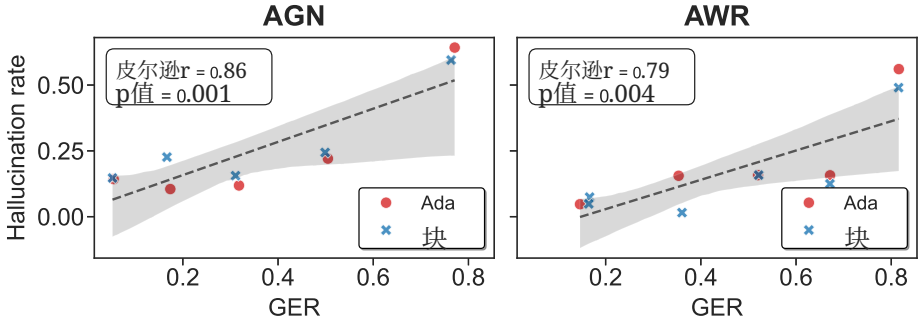
5.4 模型内部在压缩后存活的是什么？为什么存活并不意味着会被使用？

探测机制。 为衡量 KV 压缩后残差流中仍可通过线性方式访问的内容，我们使用与生成时相同的 press 配置，在冻结的隐藏状态上训练线性探针。对于每个模型、数据集和压缩比 α，我们缓存隐藏状态，并训练一个轻量级线性分类器来预测标准答案的 con-

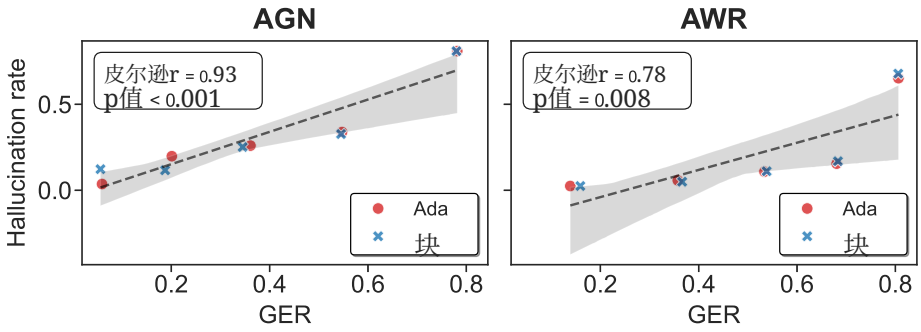


(a) LLaMA-3 8B Instruct (b) Qwen-2.5 7B Instruct

图23：压缩易感性 $\chi = \frac{\partial H}{\partial \alpha}$ 在 $\alpha \approx 0.9$附近急剧达到峰值，这与误差景观中的临界转变一致。



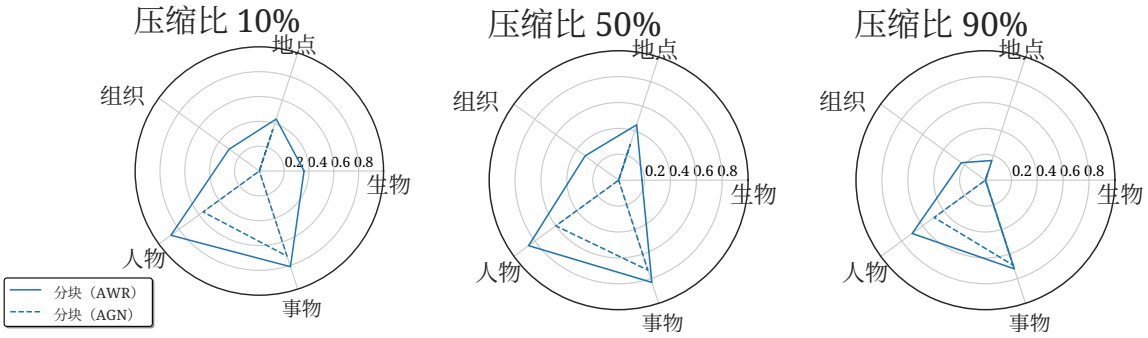
(a) LLaMA-3 8B Instruct



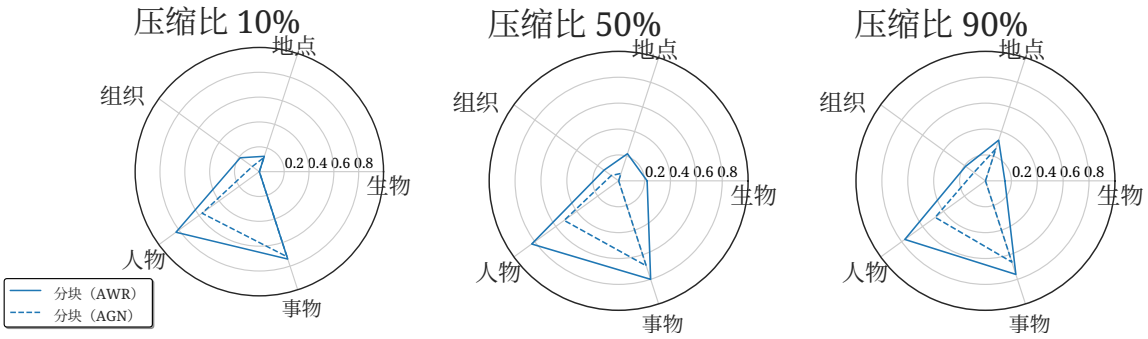
(b) Qwen-2.5 7B Instruct

图24：GER 与幻觉率高度相关，表明全局路由删除是导致灾难性失效的主要因素。LLaMA-3.2 3B、Qwen-2.5 3B 和 Qwen-2.5 14B 的对应结果见附录 B 的图44。

概念标签（例如，人物、地点、组织），并报告宏平均F1（算法2）。探测量化的是表征的存在/可访问性，而非使用：如果路由发生崩溃（GER较高）或协调变得僵化，解码器仍可能失败；因此，在悬崖附近，即使探测分数较高，也可能同时出现较高的幻觉率（图25和图22）。



(a) LLaMA-3 8B Instruct



(b) Qwen-2.5 7B Instruct

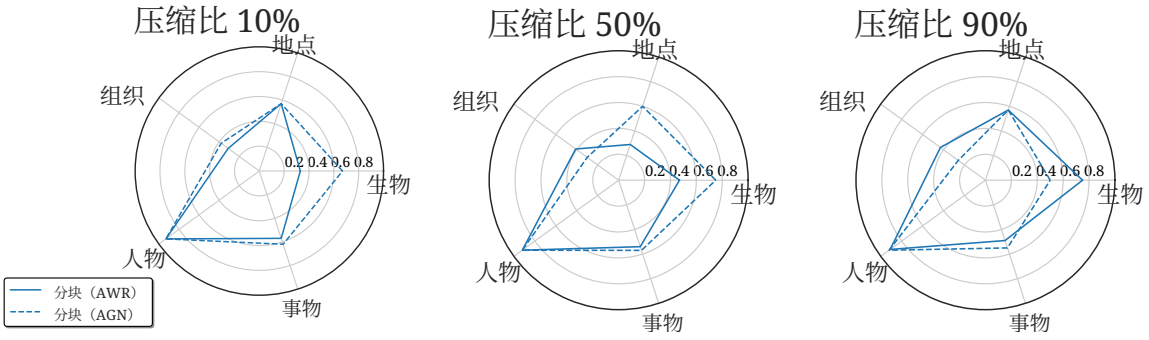
图 25：压缩过程中的基础任务探测。一些标签仍可解码；另一些则会早早崩溃，并可能在问题感知设置下部分恢复。LLaMA-3.2 3B、Qwen-2.5 3B 和 Qwen-2.5 14B 的对应结果见附录 B 的图 40。

观察：保留 $\neq$ 利用

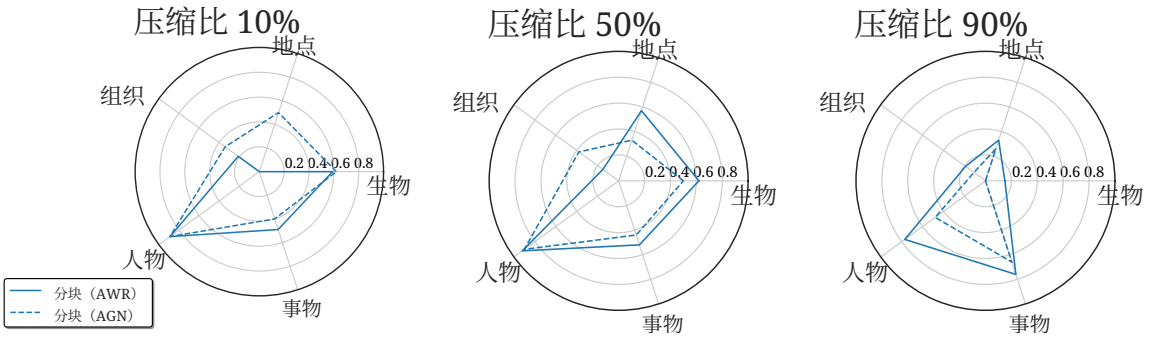
即使生成失败，探测结果仍可能保持较高水平：即便在悬崖附近幻觉率激增（图22），概念向量仍可能保持可解码状态（图25–26），这表明限制可靠性的并不只是表征存在，还包括路由可用性和灵活性。

5.5 问题感知压缩何时有帮助？为什么它对某些模型的帮助比对其他模型更大？

问题感知压缩可以看作注入了一种早期判别性偏置：剪枝会保留最能将答案实体与竞争项区分开的词元，这与多种情形下探测可分性的提升一致（图 25 和图 26）。当深层会放大早期选择信号时，其影响更大，例如在后期整合架构中（图16 和图 18）。反之，若相关性与查询重叠并非局部对齐，那么需要潜在桥接词元的任务（如多跳或差异检测）在以查询为中心的启发式方法下仍可能很脆弱。这表明，稳健的问题感知剪枝必须保留路径，而不只是与查询相似的词元。



（a）LLaMA-3 8B Instruct



(b) Qwen-2.5 7B Instruct

图 26：压缩过程中的多实体探测。问题感知带来的帮助具有选择性，而非均一，这表明表征几何与剪枝对齐之间存在相互作用。LLaMA-3.2 3B、Qwen-2.5 3B 和 Qwen-2.5 14B 的对应结果见附录 B 的图 41。

观察：查询信号提升判别性保留

问题感知剪枝往往会保留能够维持正确答案决策边界的词元，从而提升探测可分性（图 25–26），并缓解 $\alpha \approx 0.9$ 附近的灾难性失效（图22），尤其是在后期整合模型中（图16）。

5.6 KV压缩更适合理解为渐进式表征衰减，还是相变?

最后，探测轨迹与生成轨迹之间的不匹配表明，KV压缩无法用平滑的表征衰减来恰当地建模：探测分数可能会提前下降（图 25），而生成则会一直保持稳定，直到遭遇悬崖（图 22）。这更符合可达性驱动的转变：只要至少有一条通往答案证据的可行路径得以保留，性能就能保持稳健；当路径被全局删除（GER 较高）或变得过于僵化、无法有效利用（巩固程度较高）时，性能便会崩溃。这与稳定的淘汰趋势（图 21）和显著的误差非线性（图 23 ）相吻合。

观察

图 21–22 表明，可靠性取决于令牌路由图中是否仍存在通往答案证据的可行路径（图 24），以及路由是否仍保有足够的灵活性（图16），这支持了将极端 KV压缩视为一种类似转变的观点。

6 令牌路由稀疏性的理论意义

6.1 从参数彩票到令牌路由子图

近期理论为多头自注意力提供了强有力的参数级 稀疏性保证。具体而言，针对注意力的强彩票假设（SLTH）（Otsuka 等人，2025a）表明，当注意力模块足够宽时，存在稀疏的参数掩码，其受限注意力算子无需重新训练即可近似目标注意力行为。这一论断针对的是权重：参数的一个子集可以实现与稠密模块相同（或近似相同）的映射。KV缓存压缩则有本质区别。它不改变参数，而是在推理时移除键值记忆中的行 （即从注意力可读取的集合中移除词元），从而改变哪些令牌间交互成为可能。因此，我们有必要提出一种互补的稀疏性概念，用于描述前向传播期间由注意力所诱导的令牌连通性 。

定义 1（令牌注意力图）。固定一个长度为 S 的输入序列，并考虑一个具有 $\ell \in \{0,\dots,L\}$ 层和 $h \in \{1,\dots,H\}$个（KV）注意力头的Transformer。令$V = \{(\ell,i) : \ell \in \{0,\dots,L\}, i \in \{1,\dots,S\}\}$ 表示按层和位置索引的令牌节点。对于每个层转换 $\ell \to \ell+1$ 和注意力头 h，令$A_{\ell,h} \in \mathbb{R}^{S\times S}$ 为用于计算第 ℓ层注意力头输出的（行随机）注意力矩阵。给定固定阈值 $\varepsilon \geq 0$，定义有向边

$$E \;=\; \big\{\,(\ell,i) \to (\ell,j) \;:\; \exists h \text{ s.t. } A_{\ell,h}(i,j) > \varepsilon\,\big\}.$$

由此得到的有向分层图 $G = (V,\ E)$概括了在层ℓ 上哪些令牌位置能够将信息直接从令牌 j 传递到同一层的令牌 i 。[3]

定义 2（压缩令牌注意力图）。在压缩级别 α下，压缩方法会为每个注意力头 h 和层 ℓ 选择一组保留的键值位置$U_{\ell,\ h}^{(\alpha)} \subseteq \{1,\dots,S\}$。通过移除所有经过已剪枝 KV位置的边，定义压缩图 $G_\alpha = (V,\ E_\alpha)$:

$$E_\alpha \;=\; \big\{\,(\ell,i) \to (\ell,j) \in E \;:\; \exists h \text{ s.t. } j \in U_{\ell,h}^{(\alpha)} \;\wedge\; A_{\ell,h}(i,j) > \varepsilon\,\big\}.$$

等价地， G_α 是注意力算子诱导的路由图，其中在每个$A_{\ell,h}$中，将所有列 $j \notin U_{\ell\ h}^{(\alpha)}$ 置零（并重新归一化）后得到。

3. 边绘制在每一层内部，因为注意力会在残差更新之前，将来自源位置 j的值聚合到目标位置 i。任何等价的约定（例如，将 $(\ell,\ j)$ 与 $(\ell+1,\ i)$ 相连）在重新标记后都会得到相同的可达性概念。

定义 3（令牌路由彩票，TR-LT）。设 q 为指定的查询令牌位置（通常是问题令牌区域），$T_{\text{ans}} \subseteq \{1, \ldots, S\}$ 为答案相关上下文位置的集合。[4] 压缩率为 α 时的 TR-LT 是一个子图 $H_\alpha \subseteq G_\alpha$，满足：(i) 存在 $t \in T_{\text{ans}}$，使得在 H_α 中，对于某些层 ℓ_q, ℓ，从 (ℓ_q, q) 到 (ℓ, t) 存在一条有向路径；(ii) 将注意力路由限制在 H_α 的边上，足以（即保留模型在该实例上的正确输出）支持正确生成。直观而言，TR-LTs 是在压缩下仍能保持功能的令牌级稀疏路由骨干，与 SLTH 的参数级稀疏骨干互为补充。

6.2 可达性、冗余，以及擦除何时会迫使模型产生幻觉 n

KV压缩下的关键结构性问题是：答案证据能否通过压缩后的路由图保持可达。令 $R_\alpha(q)$ 表示在 G_α 中从 q 出发可达的令牌位置集合（为简化记法，省略层索引）。定义可达事件

$$\mathcal{C}_\alpha(q) \;:=\; \big(T_{\text{ans}} \cap R_\alpha(q) \neq \emptyset\big).$$

如果 $\mathcal{C}_\alpha(q)$ 不成立，那么通过注意力路由就无法到达任何与答案相关的令牌，因此模型无法通过自注意力路径，让其最终状态以答案证据为条件。

命题1（冗余的逐头路径带来鲁棒性）。假设对于每个 $t \in T_{\text{ans}}$，在未压缩的图 G中，从 q 到 t 至少存在 k 条头不相交有向路径。[5] 进一步假设，对于每个注意力头，KV令牌 t 独立地以至少 p的概率保留（即，在该注意力头中任何会用到它的位置都未被剪枝）。那么

$$\Pr\big[t \in R_\alpha(q)\big] \;\geq\; 1 - (1-p)^k.$$

解释。如果答案证据可以通过多个注意力头冗余地访问，那么所有路径都被摧毁的概率会随着冗余度 k呈指数级下降。这解释了为什么中等压缩能够保留模型行为：即使单条路径很脆弱，通常也会有至少一条存活。

命题2（擦除意味着上下文依据的丧失）。假设在移除对应于 T_{ans} 的 KV 词元之前，模型尚未将答案内容编码到可到达的残差流中之前（即，使用答案证据的唯一方式是路由至幸存的答案词元）。[6] 如果存在 $\alpha^\star$ 使得

$$T_{\text{ans}} \cap R_{\alpha^\star}(q) = \emptyset,$$

那么，在压缩 $\alpha^\star$ 时，就没有自注意力路由机制能够让模型的隐藏状态以答案证据为条件。因此，无法保证正确生成，

4. 在我们的实验中，T_{ans} 取自用于评估和探测的金标准答案跨度/标注。5. “头不相交”意味着这些路径可以通过彼此不相交的、特定于各个头的边集来实现，从而形式化了跨头冗余。6. 这一假设排除了这样一种退化情况：答案信息在驱逐之前被完美地“复制”到其他幸存词元中；经验上，GER–幻觉相关性（图24）表明，在安全性断崖处，这种信息泄漏是有限的。

模型只能依赖上下文和参数化先验，这在结构上使幻觉更容易发生。

证明概要。 每一层的每个注意力头输出，都是由留存的 值向量构成的加权线性组合。如果从查询词元无法到达 $G_{\alpha^\star}$ 中的任何答案词元，那么无论经过怎样的注意力聚合，都无法将源自 T_{ans} 的任何值向量纳入以查询为条件的隐藏状态中。在无泄漏假设下，解码器没有关于答案的上下文证据，因此只能依赖先验。□

命题2明确指出了 GER 所诊断的内容：不仅是还剩下"多少"词元，更在于路由图中是否仍存在通往答案证据的任何 路径。这与经验上观察到的悬崖现象直接相关：只要可达性仍然存在，生成就保持稳定；而当全局路由删除变得普遍时，生成性能便会崩溃（图 21、22、24）。

6.3 表征僵化：存活却未被使用的失效

可达性是必要条件，但并非充分条件：我们观察到一些情形，答案词元虽得以留存（ GER 较低），性能却有所下降，这与路由中的僵化 相符。形式上，假设在第 ℓ 层，有许多注意力头将其最高注意力赋予同一个词元 $t^\star$，导致实际读取的词元缺乏多样性。

命题 3（高一致性会降低重新路由能力）。固定一个层 ℓ。令$t^*_{\ell,\ h} = \arg\max_t A_{\ell,\ h}(t)$ 为注意力头 h 最关注的词元，并定义一致比例

$$\rho_\ell \ := \ \max_t \ \frac{1}{H}\big|\{h:\ t^*_{\ell,h} = t\}\big|.$$

如果在压缩过程中，相关注意力头中的最大化词元 $t^\star$ （使 ρ_ℓ 达到最大值）被剪枝，那么至少有 ρ_ℓ 比例的注意力头必须将其主要注意力转向次级词元。如果这些次级词元系统性地排除了 T_{ans}，那么在层 ℓ 路由至答案证据的有效概率就会按与 ρ_ℓ 成正比的因子降低。

解释。当路由高度集中时，压缩可能会移除一小组"共识"词元，迫使许多注意力头同时重新路由。如果剩余的注意力质量不够多样，无法重新发现答案证据，那么即使其他位置的词元仍然存活，模型也可能失败。这就是共识会造成表征僵化的结构性意义：模型在记忆中保留了证据，却缺乏路由灵活性来使用它。

6.4 失败案例分析：最小分解

我们可以将失败概率表示为擦除 机制与 刚性机制的混合。令 $\mathcal{F}_\alpha$ 表示压缩级别 α下的生成失败。令 $\mathcal{E}_\alpha$ 表示与答案相关的词元在各注意力头上被全局驱逐的事件（就GER所度量的意义而言，该任务实例被"路由删除"）。对 $\mathcal{E}_\alpha$ 进行条件化可得

$$\Pr(\mathcal{F}_\alpha) = \Pr(\mathcal{F}_\alpha \mid \mathcal{E}_\alpha)\Pr(\mathcal{E}_\alpha) + \Pr(\mathcal{F}_\alpha \mid \neg\mathcal{E}_\alpha)\Pr(\neg\mathcal{E}_\alpha).$$

第一项刻画路由存在性：当 $\mathcal{E}_\alpha$ 成立时，上下文依据不可用，因此实践中 $\Pr(\mathcal{F}_\alpha \mid \mathcal{E}_\alpha)$ 接近 1。第二项刻画路由可用性：即使答案令牌原则上仍然存在，如果路由崩溃并集中到僵化的

模式上，也可能发生失败。在我们的设定中，Pr($\mathcal{E}_\alpha$) 由 GER(α) 估算（图24），而头级共识趋势（图16）则提供了可测量的刚性代理指标。这种分解与经验观察相符：安全性断崖主要由极端压缩下的全局路由删除所致，而在高整合度机制中，中间程度的退化可能体现为“仍然存在但未被使用”的效应。

6.5 稀疏令牌路由：SLTH 的推理时对应机制

SLTH（Otsuka 等人，2025a）意味着，注意力中包含能够近似稠密注意力行为的稀疏参数 子网络。我们的实证结果揭示了一种互补的推理时现象：KV压缩下的鲁棒性取决于压缩令牌注意力图中稀疏令牌路由 子网络（TR-LTs）的存活。从这一视角来看，KV压缩是一种受控扰动，能够揭示正确生成所需的最小路由骨干。极端压缩会在以下情况下导致失败：(i) 跨头可达性崩溃（GER 较高，命题2）；或 (ii) 路由变得过于僵化，无法利用存留证据（一致性高、多样性低，命题 3）。总体而言，这些结果将注意力的理论稀疏性保证与一种具体的推理时稀疏化机制联系起来：决定可靠性的并非保留的 KV内存数量 ，而是压缩路由图是否至少保留了一条可行且可用的、通往答案证据的令牌路由。

7 局限性与未来工作

尽管我们的框架为理解 KV 缓存压缩提供了机制层面的洞见，但仍有若干自然的延伸方向。首先，我们有意构建了受控合成数据集，以隔离多跳链式推理、实体消歧和共指一致性等路由敏感行为。虽然这种设计能够对失效模式进行因果归因，但也抽象掉了自然语言和大规模真实世界语料库的全部异质性。未来的工作应探究，在长文档问答、检索增强生成、代码推理和多模态上下文中，是否仍存在同样的可达性—刚性分解。尤其是，研究具有外部记忆或工具使用能力的系统中的压缩效应，或可揭示纯粹的内部 KV 剪枝未能捕捉到的其他路由韧性或脆弱性。

其次，我们的理论处理将全局淘汰率和注意力头级共识等结构指标作为解释压缩诱发相变的变量。尽管这些指标与性能崩溃表现出较强的经验对齐关系，但对结构化令牌移除条件下注意力算子的更精细分析，或许能够得出更严密的保证。例如，注意力图的谱特征刻画、分层路由结构中的连通性阈值，或关于多头冗余的形式化界，都可能进一步强化令牌路由稀疏性与鲁棒性之间的理论联系。在建模方面，一个重要方向是协同设计稀疏机制与架构：与其进行事后剪枝，不如通过训练目标明确鼓励冗余的跨头路由或受控共识，从而在压缩条件下维持有效的路由容量。这类方法有望催生有理论依据的高效注意力设计，在降低内存成本的同时保留结构可达性。

8 结论

在本研究中，我们将KV缓存压缩重新概念化为对自注意力路由几何结构的结构性干预，而非单纯的内存缩减技术。通过受控的合成基准测试、逐层路由分析，以及对可达性和刚性指标的形式化，我们证明压缩诱发的失效源于两种截然不同的机制：与答案相关的词元被全局擦除，以及尽管词元仍然存活，路由多样性却发生崩溃。在极端压缩条件下观察到的陡峭性能悬崖，对应于语义可达性中的结构相变，揭示出稀疏的词元路由骨干结构支配着推理时鲁棒性。通过将这些经验发现与注意力中的稀疏性理论联系起来，我们将彩票票据的直觉从参数子网络扩展到了动态词元路径结构。我们的结果表明，未来的长文本和高效注意力系统不仅必须保留词元表示，还必须在注意力头和层之间保留最低限度的路由容量，使效率与结构完整性相一致，而非依赖局部重要性启发式。

附录 A. 数据集描述

Task	上下文	问题	回答
基础任务	阿根廷前锋 Gonzalo Batistuta 是 被广泛认为是最伟大的足球 运动员之一.....	什么是 冈萨洛 巴蒂斯图 塔的国籍?	阿根廷
知识渊博的人。	Cora Delaine 出生于1954年12月11日。 他们在大邱全球科学大学就读, 大学.....	Cora 是什么 Delaine的 名字?	Cora
多重存在	Astralis Spire巍然耸立，是最 杰出的现代建筑成就之一 和工程成就之一。它高高耸立于天际线之上 其繁华城市的.....	建造.....的主要原 因是什么 As- 特拉利斯尖塔?	To 改善 电视 and 无线电 广播 受到城市发展干扰 发展
多实体	Harland Kane 是最具影响力的作曲家之一 当代恐怖小说、悬疑小说及 超自然小说作家。 在数十年的 职业生涯中, 凯恩.....	凯恩的哪部小说以.....为背景 在闹鬼酒店里?	《松影》
长上下文	在白垩纪中期，河流—— 帆背类动物是令人生畏的半水生 捕食者，栖息于河流系统、洪泛平原， 和沿海湿地。它们以细长的 颌、锥形牙齿和巨大的.....	Kaelin Vireo 来自哪里? 来自哪 里?	Thaloris 的河流化石站 Thaloris 的
共指	Julian Foster出生于1962年9月7日 在阿勒颇， 叔利亚阿勒颇省。 他是 男性。他曾就读于天津港湾大学 在黑山波德戈里察的波德戈里察.....	他的职业是什么?	导游
Hops	早在殖民时期记录追溯到 墨西哥高原的传说之前， 特奥蒂拉克城便从火山平原上崛起 作为光影与信仰交织的几何奇观， 由神秘的伊察卡人建造。 Teotilac 的大道和金字塔 经过布局，以捕捉太阳天顶， 并通过精确的对齐将 天体运动与公民生活联系起来.....	What 是 City of Teotilac，它为何 与众不同？（跳 0）	一座由 伊察卡人建造、与天体 周期相契合的圣城（跳 0）
		Teotilac的设计如何 整合宇宙与公民 秩序？（跳 1）	其金字塔与大道 编码了太阳与时间 秩序（跳 1）
		Yulnah棱镜是 什么项目？（跳 2）	一个现代光学考古学 项目，位于 Teotilac （跳 2）
		谁领导了 Yul-棱镜 那么，使用了哪些现代工具 呢？（第3跳）	由 Lira Montoya博士领导 使用光谱无人机和 镜像（第3跳）

表3：每项任务的示例上下文和输入，以及标准答案。

Tag	上下文	问题	回答
人物	Joao Fernandes，被广泛认为是 有史以来最伟大的足球运动员之一.....	这位 人物 在 the 段落?	Joao Fernandes
事物	Gonzalo Batistuta，阿根廷前锋， 被广泛认为是最伟大的球员之一.....	什么是 Gonzalo Batis- tuta的国籍?	阿根廷
组织	Mookie Blaylock，一个来自北方的摇滚乐队—— port 于 1990 年在 Riverbend 成立.....	给定段落的主题是什么 ?	Mookie Blaylock
生物	Void Stalker 是一种虚构的内寄生 性外星物种，担任 主角.....	Void Stalker 是什么类型的 物种?	An 体内寄生性寄生蜂 ex- 地外物种
地点	法老石群是一组复杂的 纪念性建筑群，位于 吉萨高原.....	In 哪个 city are The Stones of Pharaohs 位 于何处?	开罗
数字	蓟木谷女巫审判代 在美国早期历史中留下了一个令人不寒而栗的篇章	How many 个人 被判有罪 并被处 决于审判期间?	20
日期/时间	阿根廷前锋 Gonzalo Batistuta， 被广泛认为是最伟大的.....之一。	Gonzalo Batis- 何时 tuta 赢得了世界冠- 军赛?	2022
事件	在他的人民身处巨大危难之际， 一个名叫雅比斯的孩子出生了。为了救他 免遭.....法令的迫害	How 雅比斯的母亲是否 救他免遭 死亡?	他的母亲把他放进一个 编织篮里，把他放 在河上任其漂流

表4：答案类型标签示例。

Tag	上下文	问题	回答
标准	玄武岩花是一个规模巨大的火成 岩省，覆盖了印度中部的广阔区域。 它是由一系列 大规模火山喷发形成的.....	玄武岩花在哪里 ?	印度中部
操纵	Xena Bell出生于1968年11月20日。 他们在上海海事大学就读， 大学.....	什么是Xena Bell的姓 氏?	Bell
Part	Breezeflare的特征是异常的 呼吸音，因气流 气道阻塞所致。该病症 很常见.....	Nyla Soren 因 领导哪类 倡议而受到认可?	社区 工作坊， 公共教育

表 5：问题类型标签示例

Task	方向	上下文	问题	回答
多重存在	FWD	名为 Chrono Nexus 的系列被广泛认为是最引人入胜的科幻故事之一，出自日本动画。由 Itsuki Aoyama创作.....	谁创作了Chrono Nexus?	Itsuki Aoyama
	REV	地球历史上很少有生物能像.....一样如此令人着迷雷牙掠食者，它是.....之一体型最大、最凶猛的食肉恐龙 曾经 生活过.....	哪种掠食者拥有锯齿状 牙齿 as long as a 人类手?	雷牙捕tor
多实体	FWD	在遥远的安达曼群岛海域，坐落着臭名昭著的铁影监狱。它建于19世纪末由英国人建造的这座监狱后来成为了.....	这位民族主义者是谁?被关押在《铁影》中，且被称为维尔?	维纳亚克 ·达莫达尔·萨瓦卡尔
	REV	Shinroku 是一部广受欢迎的日本漫画和动画系列，以其独特融合的喜剧、动作，和科幻。故事设定在一个异时空的江户时代.....	Hana 田中 ap- 在其中作为同伴出现于哪个故事?	Shinroku

表6：前锋问题和反向问题示例

与问题无关的设置模板：

阅读以下文本并据此简要作答。仅返回答案。不要生成额外内容。猎人是出现在《Stalker》科幻系列作品中的虚构外星物种。他们拥有先进技术，包括光谱隐形和等离子长矛；他们是游牧战士种族，以狩猎其他强大种族为乐，并以此获取荣誉。他们遵循严格的行为准则，通常将那些能带来重大挑战的猎物作为目标。他们独特的外貌包括辫状附肢以及一组 distinctive 的分裂式下颚。

问题感知设置模板：

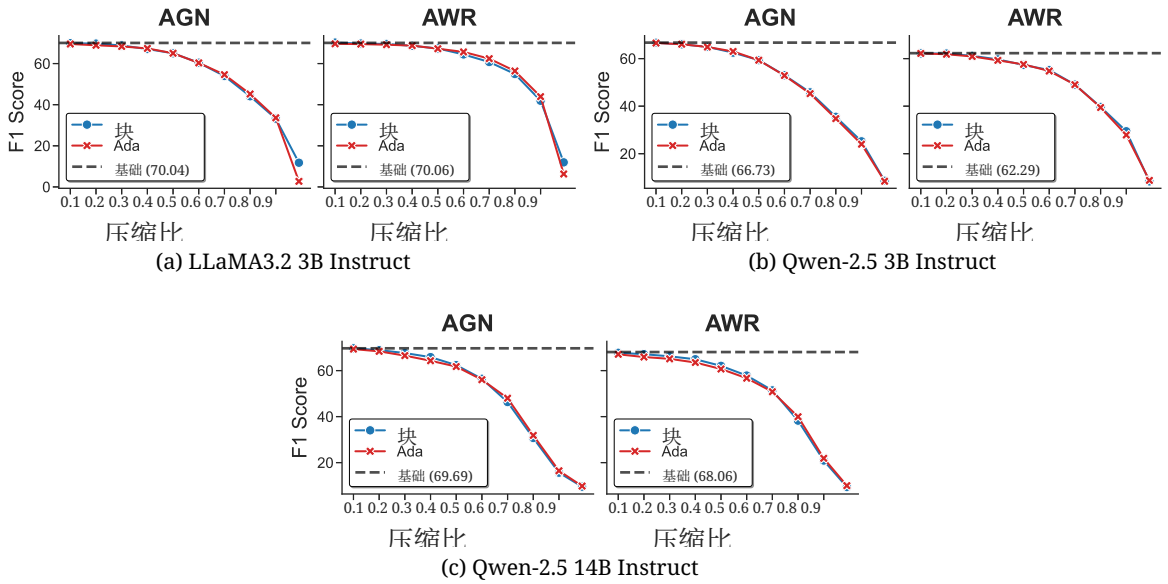
```
阅读以下文本，并根据内容简要作答。只返回答案。不要
    生成额外内容。

你将收到以下问题之一：

给定段落的主题是什么？
猎人是什么物种？
猎人是哪个系列中的反派？
猎人是什么种族？
猎人以什么为猎物？
他们独特的外表有哪些标志性特征？

猎人是虚构的外星物种，出现在 Stalker
    科幻系列作品中。他们以先进技术闻名，包括
    光谱隐形和等离子长矛。他们是一个游牧战士种族，
    为了荣誉和乐趣而猎杀其他强大的物种。他们遵循严格的行为准
    则，通常以能带来巨大挑战的猎物为目标。他们的
     独特外貌包括辫状附肢，以及一组形状独特的分裂式
    下颚。
```

附录 B. 补充结果



(a) LLaMA3.2 3B Instruct

(b) Qwen-2.5 3B Instruct

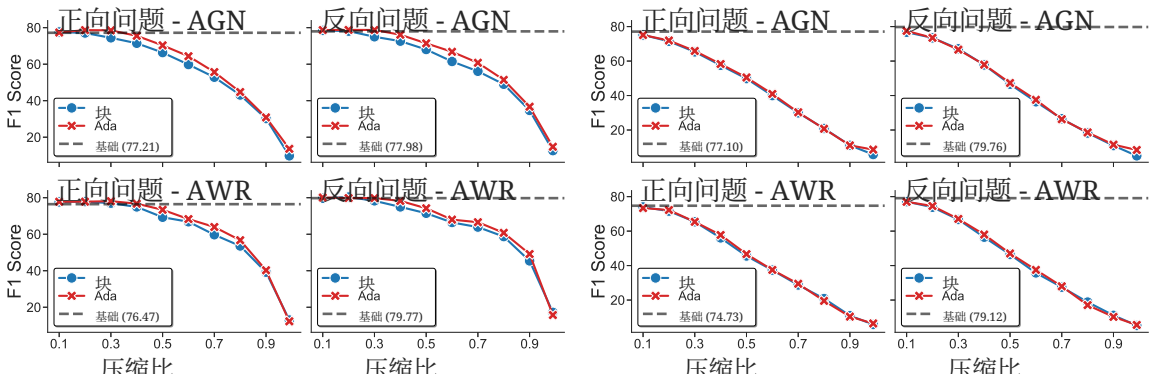
(c) Qwen-2.5 14B Instruct

图 27：3B 和 14B 模型在不同压缩级别下的基础任务性能。

本节报告 LLaMA-3.2 3B、Qwen 2.5 3B 和 Qwen 2.5 14B 的结果，将正文中的分析扩展到其他模型规模。在各个模型中，

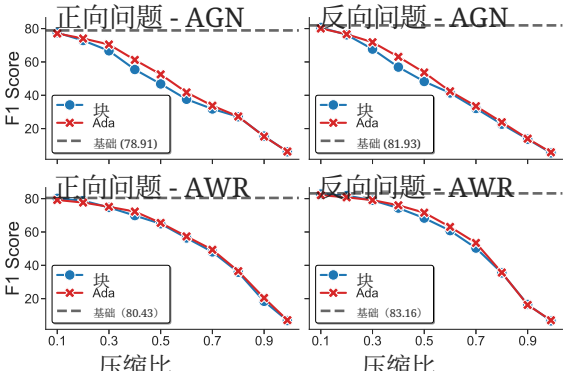
基础任务中观察到的主导趋势仍然一致：文本标签、问题标签和总体检索准确率方面的表现，基本上与正文讨论的 7B模型相似（图27、图29、图31）。值得注意的是，参数数量的增加并不系统性地转化为压缩条件下更强的鲁棒性。不同模型规模的初始（未压缩）性能总体相当，并且性能崩溃发生在相近的压缩级别下，其中 70–90% 的压缩率始终标志着性能明显失效的临界点（图27、图28）。

这一共同模式的主要偏离出现在 Qwen 2.5（3B 和 14B）的多实体任务中：其性能随压缩近似线性下降，而非出现明显的悬崖式下跌。这一现象在人物文本标签中尤为明显（图30）：随着压缩程度提高，准确率持续下降。相同的渐进式退化也延伸至对应问题标签中的标准检索（图32），这表明实体特定退化直接推动了总体趋势。总体而言，这些结果进一步印证了正文的核心观点：压缩敏感性受任务结构和实体分布的影响，比受模型规模的影响更大；模型规模较大并不意味着其必然更能抵御激进压缩。



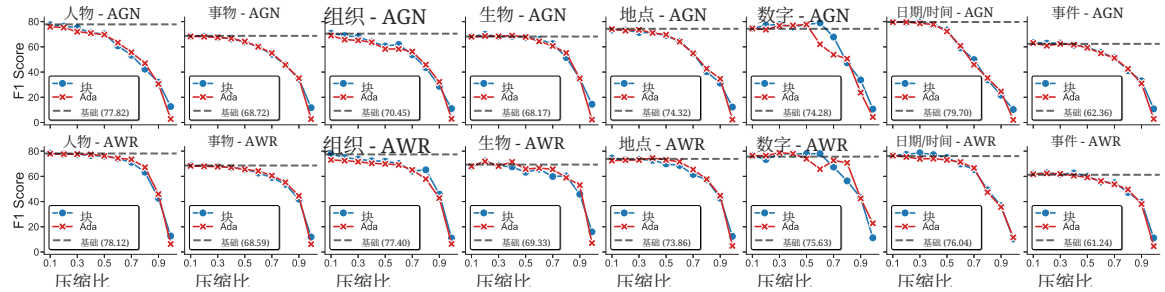
(a) LLaMA-3.2 3B Instruct

(b) Qwen-2.5 3B Instruct

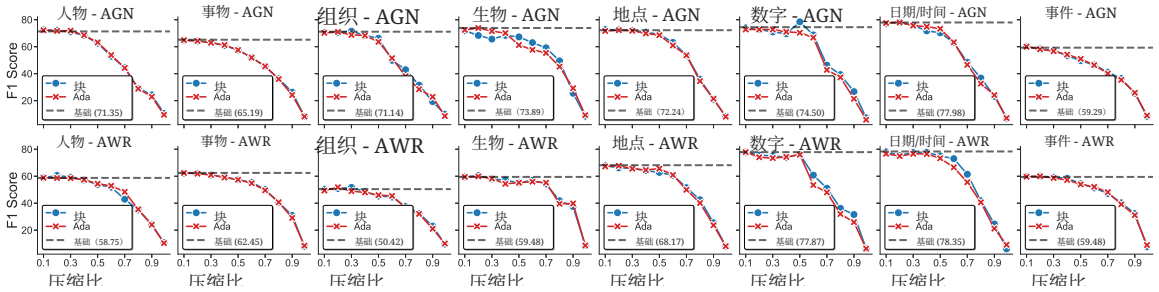


(c) Qwen-2.5 14B Instruct

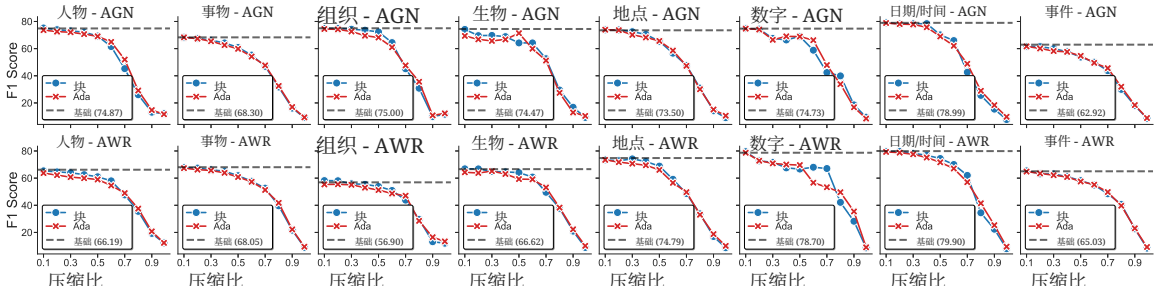
图 28：即使模型规模发生变化，多实体正向与反向不对称性也会减小，表明该现象不受模型规模影响。



(a) LLaMA-3.2 3B Instruct



（b）Qwen-2.5 3B Instruct

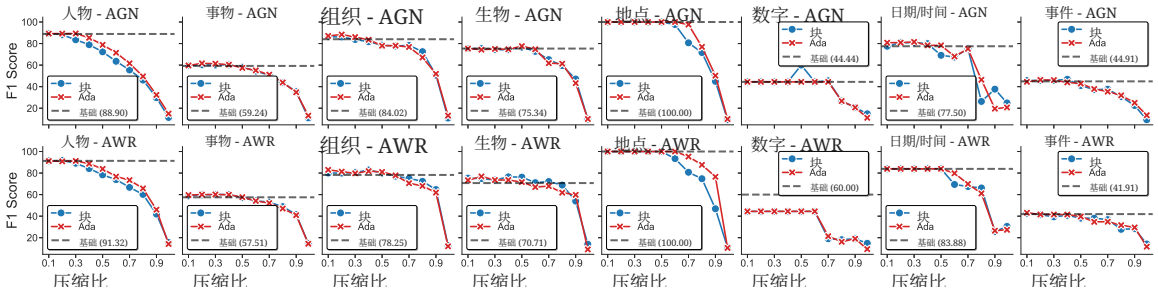


(c) Qwen-2.5 14B Instruct

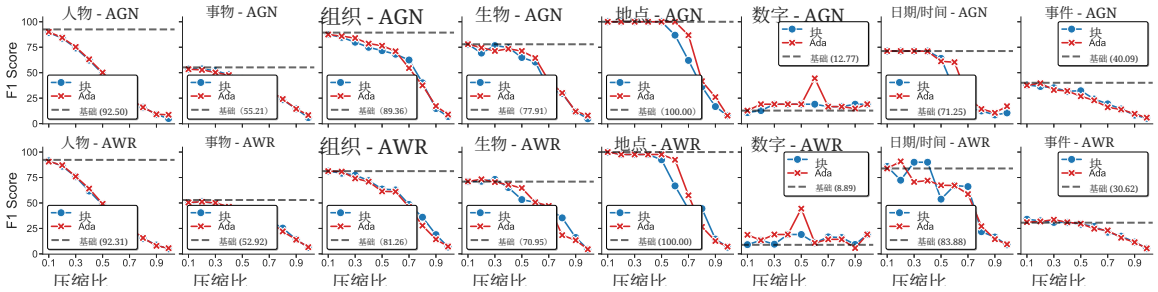
图 29：基础任务的答案类型结果与 LLaMA 3 8B 和 Qwen 2.5 7B 模型的结果十分相似，尤其是在它们的弱项上，例如事件 标签。

不过，多实体曲线所呈现的表面线性特征仅反映了底层动态的一部分。附录中的共识分析和逐层共识分析总体上印证了总体准确率中观察到的跨模型相似性，同时也揭示了模型在鲁棒性方面的架构差异。具体而言，Qwen 2.5 3B 在压缩下的共识稳定性明显较低，这表明随着压缩程度提高，各层之间的内部不一致性更大。这说明其退化不仅体现在输出准确率逐渐下降，在表征层面也具有结构脆弱性，尤其是在处理多样化实体类型时。

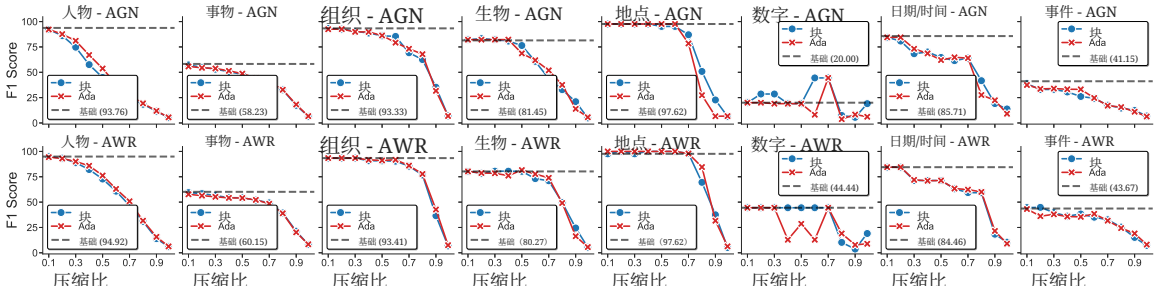
相比之下，LLaMA 3.2 3B 展现出的动态在结构上与其 8B 版本一致，但对压缩更加敏感。退化模式并不表明其失效模式在性质上有所不同。相反，这些模式反映了与较大模型相同的崩溃机制，只是在更低的压缩阈值下被触发，并且以更突然的方式展



(a) LLaMA-3.2 3B Instruct



（b）Qwen-2.5 3B Instruct



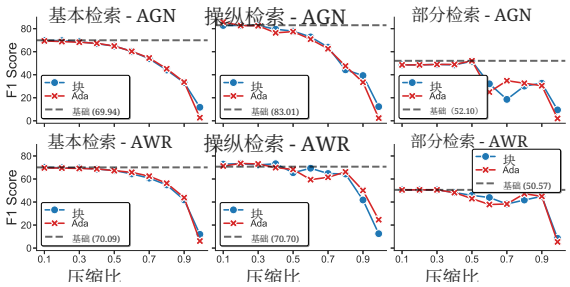
(c) Qwen-2.5 14B Instruct

图30：在多实体任务中，3B 和 14B 模型的答案类型标签行为与 LLaMA 3 8B 和 Qwen 2.5 7B 的表现呈现出相似趋势。

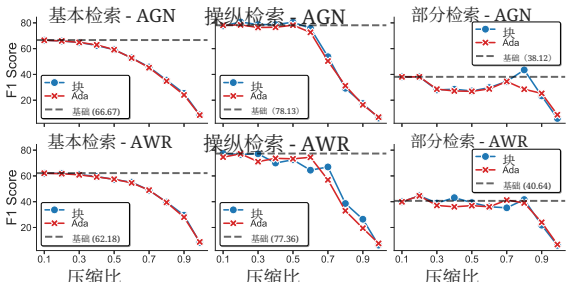
开。从这个意义上说，3B模型是相同底层动态的缩小规模的实例，而非架构上不同的系统。

重要的是，共识分析和逐层分析表明，这些相似性并不局限于总体准确率曲线。尽管不同规模模型的整体性能趋势看似相近，但共识指标揭示了压缩过程中不稳定性在各层累积和传播方式的差异。因此，附录中的结果表明，压缩鲁棒性并不单纯取决于参数数量，还取决于整个网络层级中表征一致性如何得以维持或丧失。

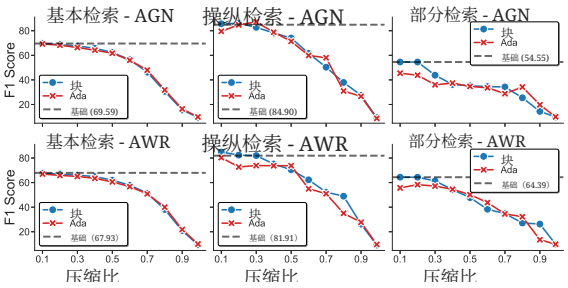
探测雷达分析进一步印证了前文各节发现的趋势。在不同模型规模下，生物标签始终是一个明显的弱项，即使在未压缩设置下，也只有少数实例能够可靠地处理这一类别。在



(a) LLaMA-3.2 3B Instruct



(b) Qwen-2.5 3B Instruct



(c) Qwen-2.5 14B Instruct

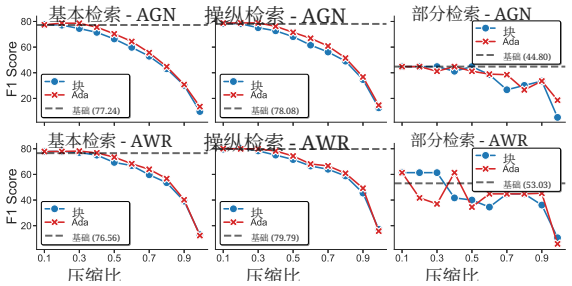
图31：基础任务的问题类型结果继续呈现类似趋势，在较小和较大的模型中均如此，如7B和8B模型所示。

压缩条件下，这一局限会显著加剧。生物查询的表现不仅会在高压缩比下下降，在一些情况下，即使压缩程度相对较轻，表现也会下降。

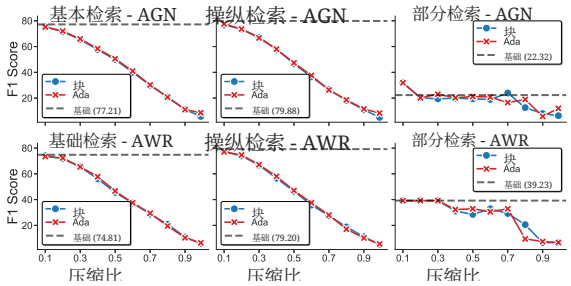
这些结果表明，难点并非完全由压缩导致，而是反映了此类实体本身存在的表征脆弱性，而压缩会进一步放大这一问题。换言之，压缩并未为 Creature 检索引入新的失效模式，而是加速并加剧了各模型中已有的结构性弱点。

不同任务和模型规模下的驱逐率总体保持一致，这进一步说明 KV缓存压缩机制与任务无关。换言之，无论下游目标是什么，压缩过程都会呈现类似的保留动态。然而，错误分布则展现出明显不同的模式。

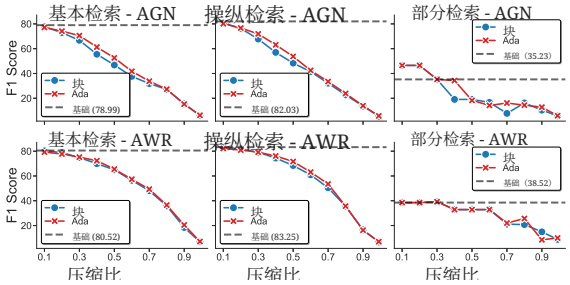
与压缩条件下错误类型仍相对均衡的7B和14B模型不同，随着压缩程度增加，3B模型中某些错误类别明显占据主导。这种偏斜不仅意味着准确率下降，还表明存在一种结构化的失效模式：某些检索路径会优先崩溃。错误类型如此集中，凸显了较小模型更为脆弱：在压缩条件下，它们较难恢复或重建部分保留的信息，因此造成的是系统性而非弥散性的性能退化。



（a）LLaMA-3.2 3B Instruct

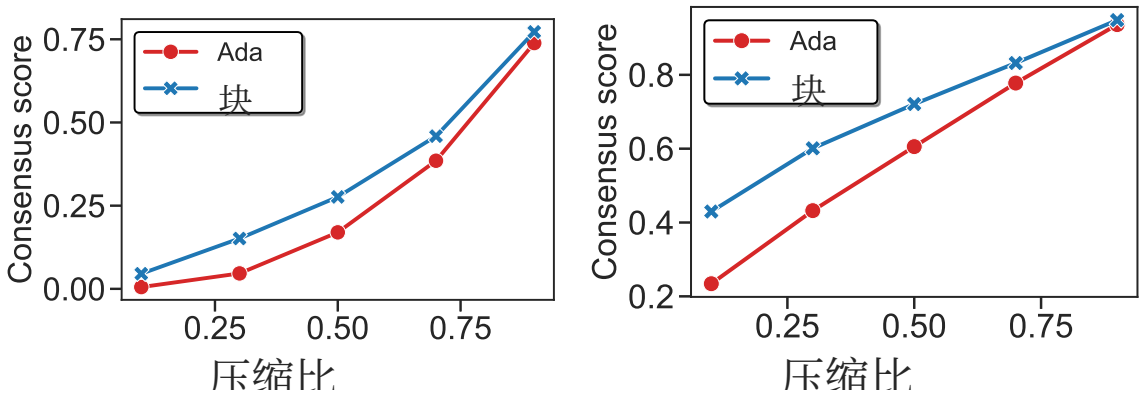


（b）Qwen-2.5 3B Instruct



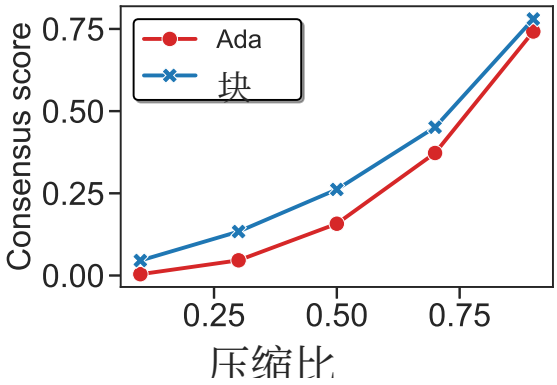
（c）Qwen-2.5 14B Instruct

图32：LLaMA 3.2 3B 的多实体结果显示，在问题感知设置下仍有类似的锯齿状点，而 Qwen 2.5 3B 和 14B 模型的性能几乎呈现出平滑线性模式。



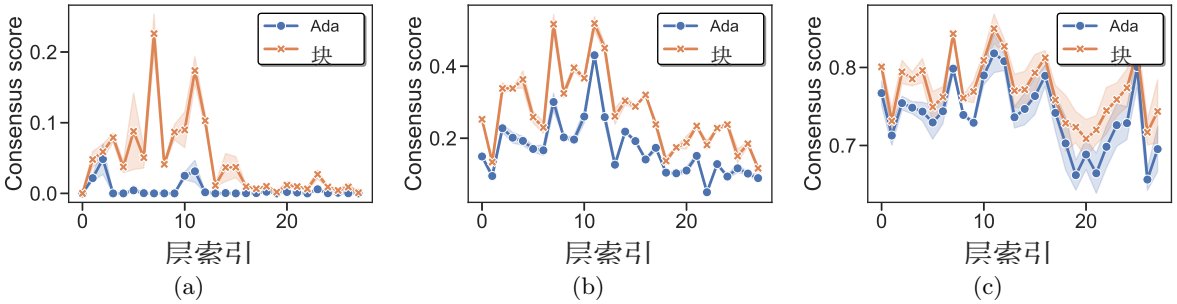
(a) LLaMA-3.2 3B Instruct

(b) Qwen-2.5 3B Instruct



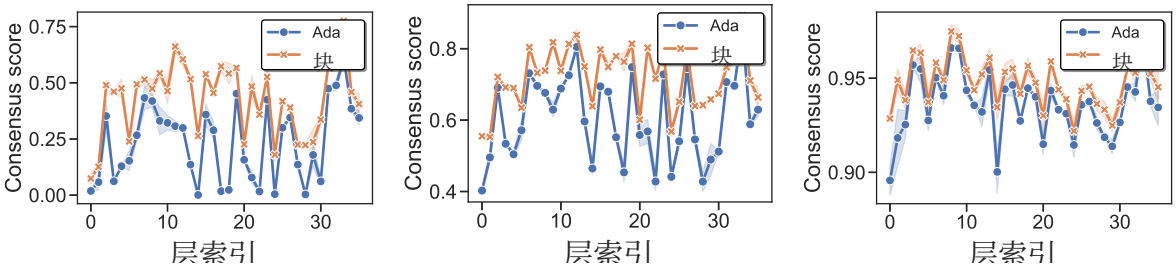
(c) Qwen-2.5 14B Instruct

图33：共识分数往往呈现相似的趋势，主要变化在于 Qwen 3B 的增长略呈线性，而 7B 和 14B 模型的增长则略呈抛物线式。



(a) (b) (c)

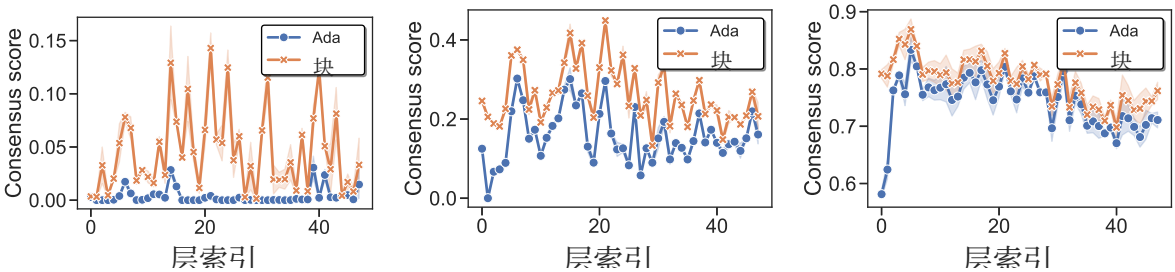
图34：LLaMA 3.2 3B 在压缩下的逐层共识表现出惊人的相似性，很早便收敛，同时沿深度方向呈现多样化。



(a) Qwen 2.5 3B Instruct 10% 压缩

(b) Qwen 2.5 3B Instruct 50% 压缩

(c) Qwen 2.5 3B Instruct 90% 压缩

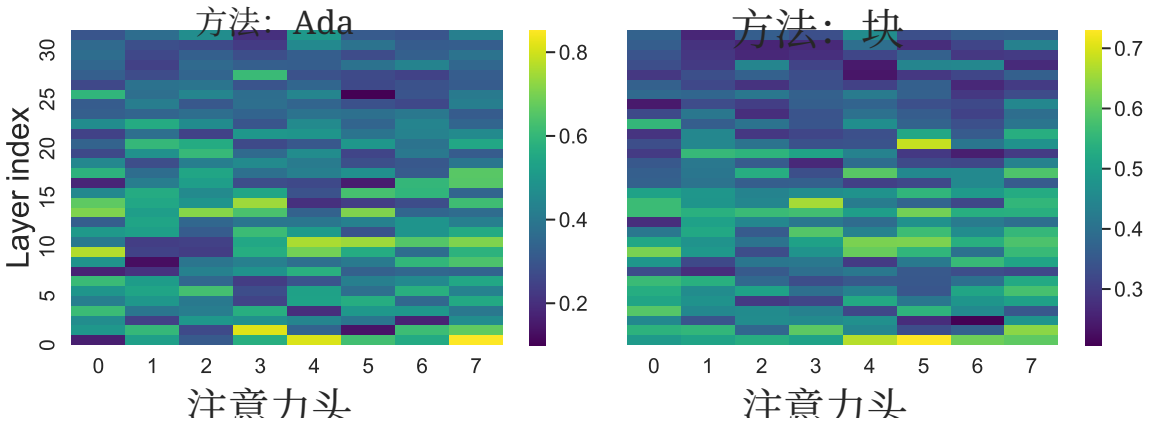


(d) Qwen 2.5 14B Instruct 10% 压缩

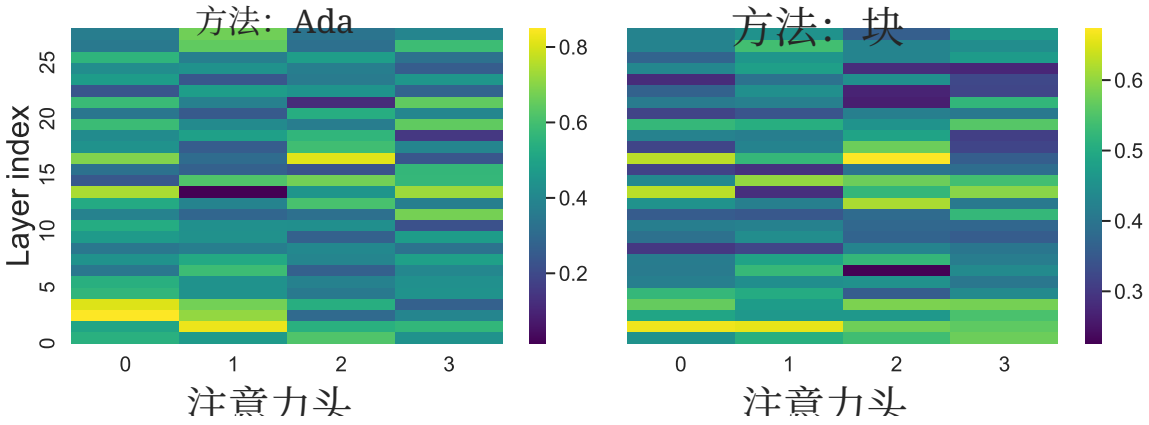
(e) Qwen 2.5 14B Instruct 50% 压缩

(f) Qwen 2.5 14B Instruct 90% 压缩

图 35：Qwen 2.5 3B 与 14B 在压缩下的逐层共识。3B模型即使经过若干层后仍持续存在多样性，偶尔无法收敛；而 14B 早期保持稳定，但在收敛时呈现出相同的漏斗状趋势。

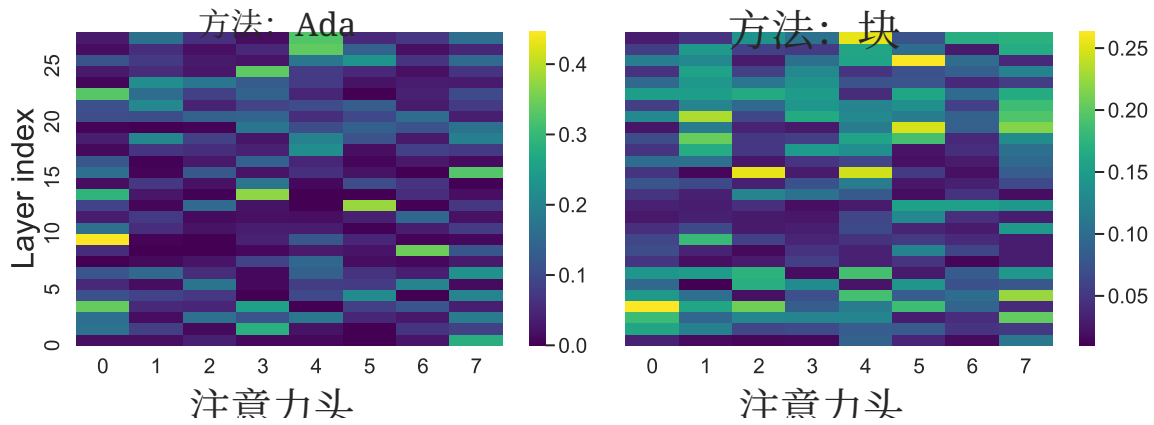


(a) 50%压缩 - LLaMA 3 8B

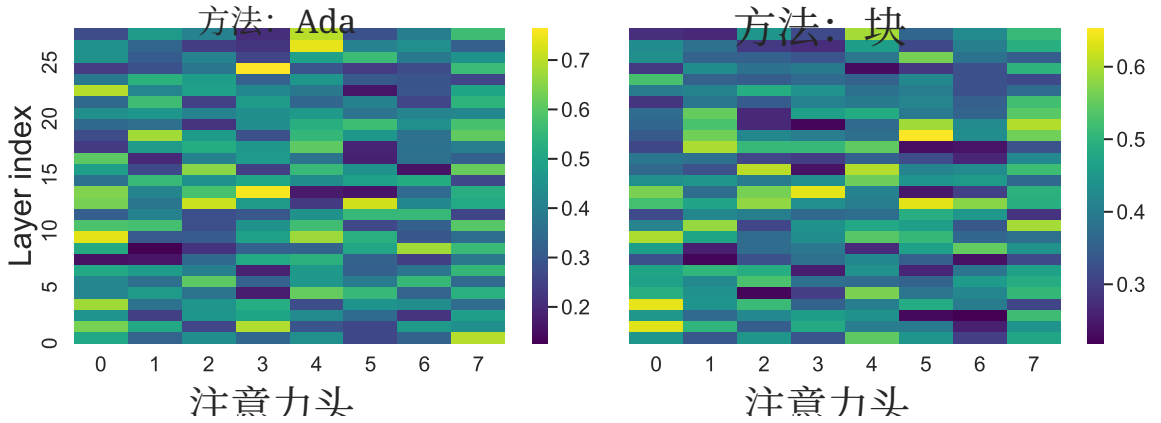


(b) 50%压缩 - Qwen 2.5 7B

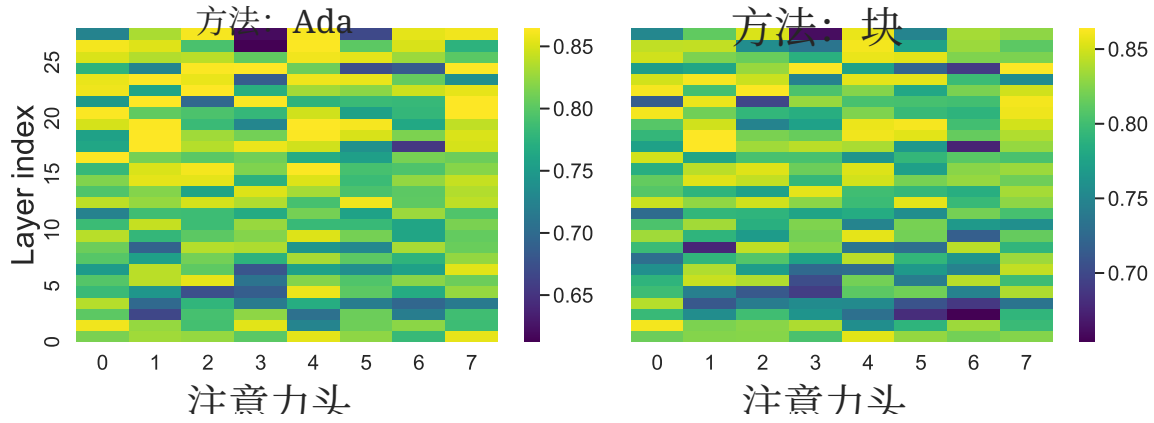
图 36：LLaMA 3 8B 和 Qwen 2.5 7B 的压缩率为 50%。



(a) 10%压缩

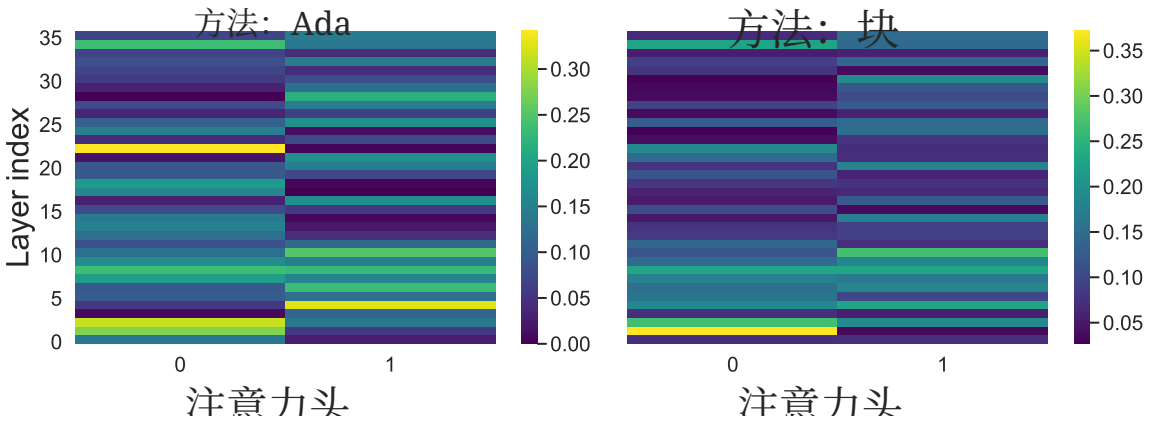


(b) 50%压缩

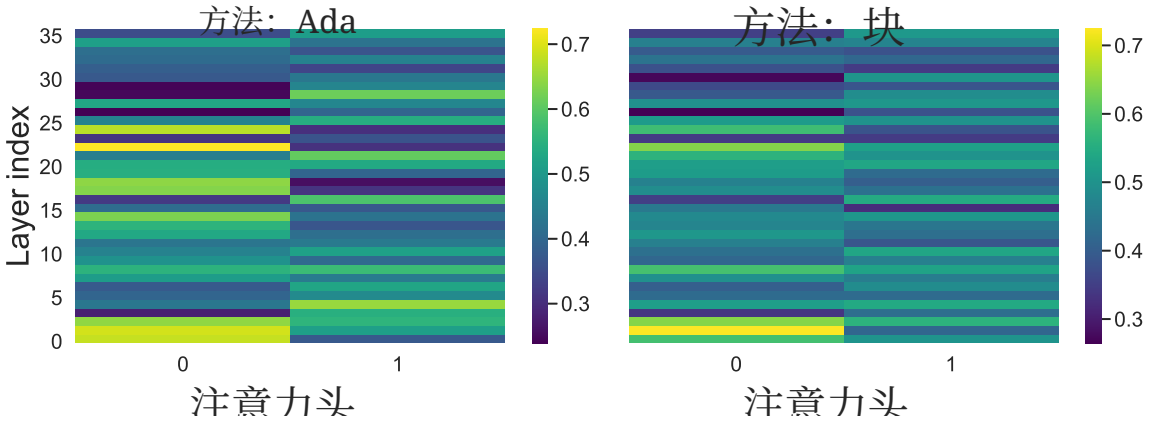


(c) 90%压缩

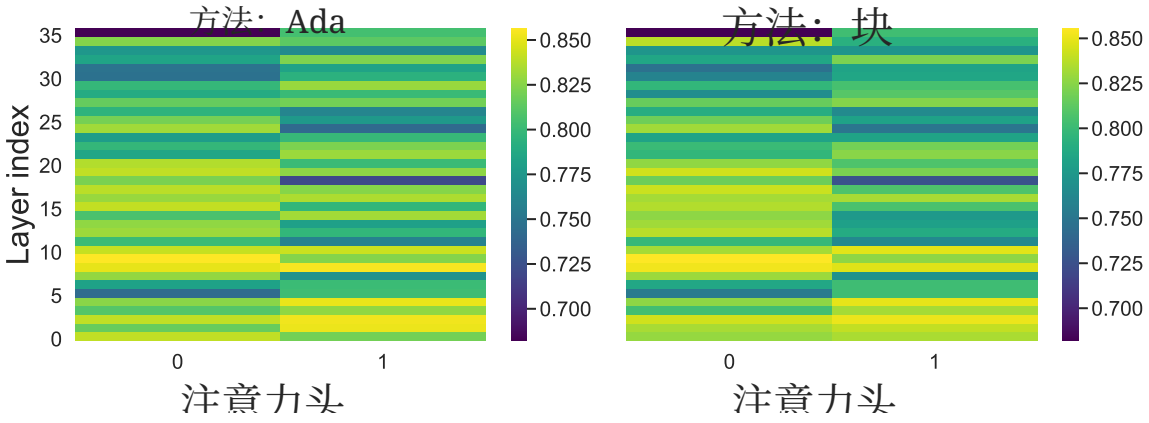
图 37：LLaMA 3.2 3B 中也呈现出类似的现象：块内平滑剪枝，Ada 中不规则移除。



(a) 压缩 10%

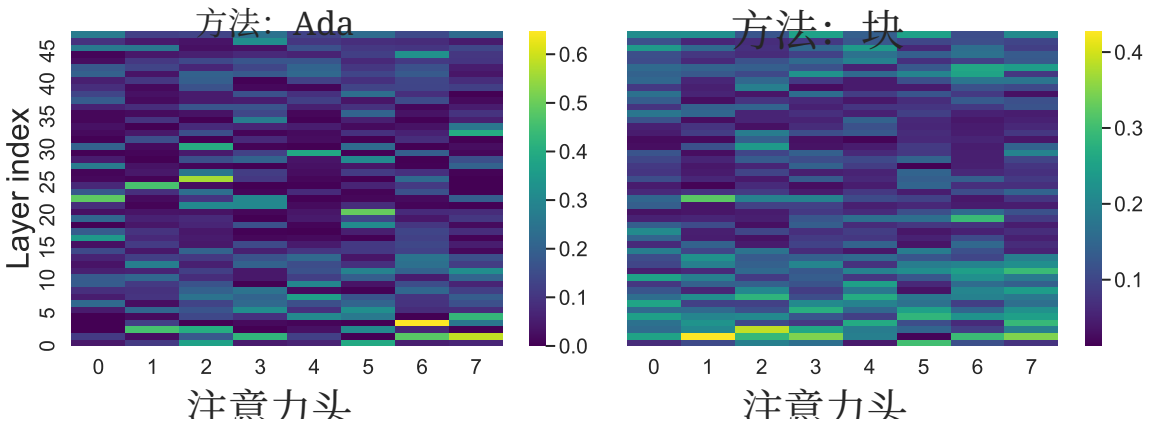


(b) 压缩 50%

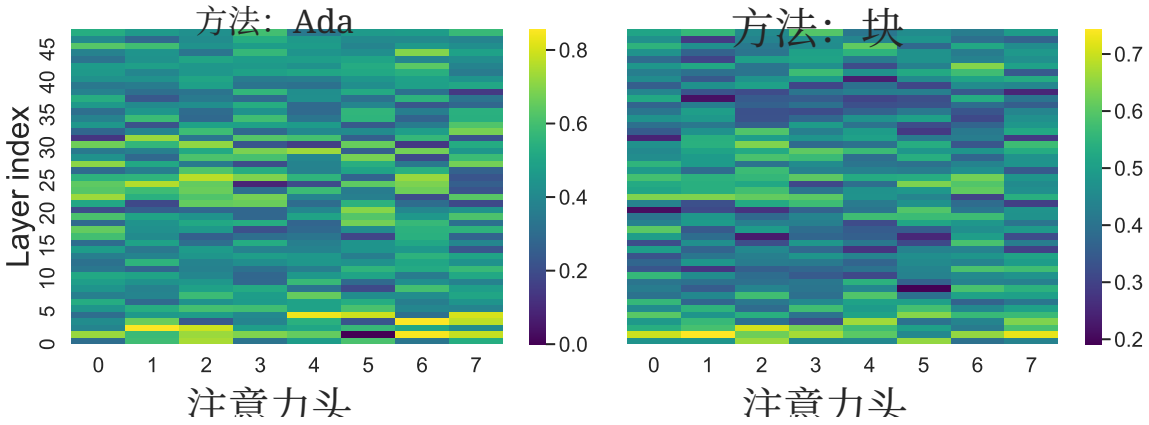


(c) 压缩 90%

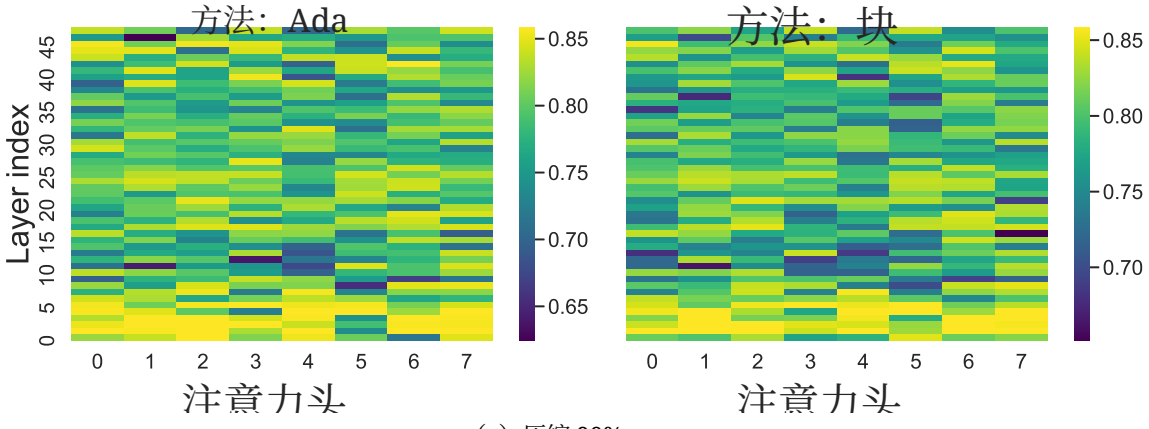
图38：由于 Qwen 2.5 3B 模型在所有层中仅使用两个注意力头，剪枝的非连续性大大减弱。



（a）压缩 10%

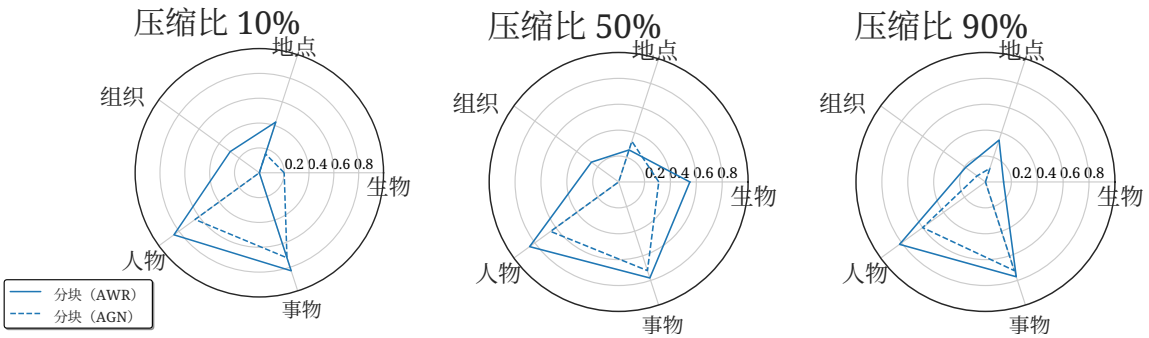


（b）压缩 50%

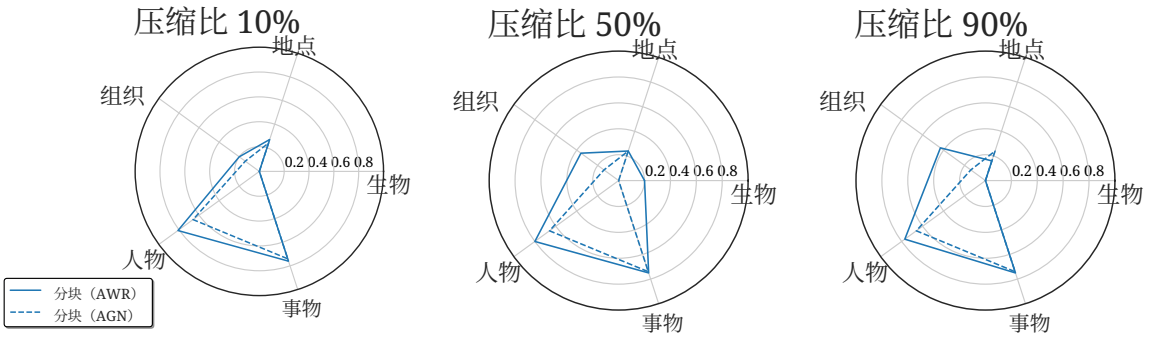


（c）压缩 90%

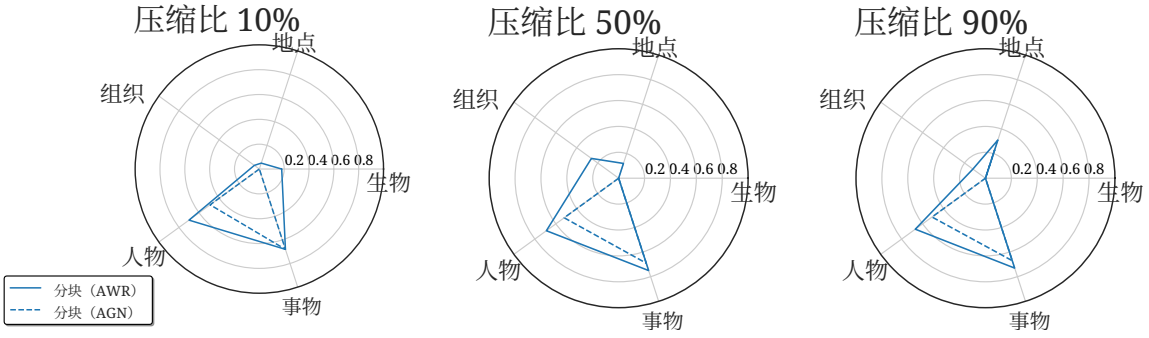
图39：在 Qwen 2.5 14B 中使用八个注意力头呈现出与 LLaMA 3 系列模型相似的趋势，其驱逐模式在性质上相当相似。



(a) LLaMA-3.2 3B Instruct

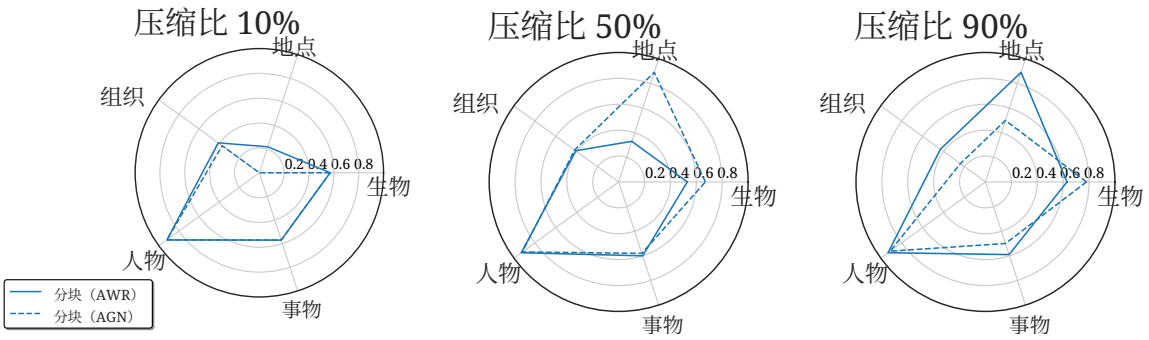


(b) Qwen-2.5 3B Instruct

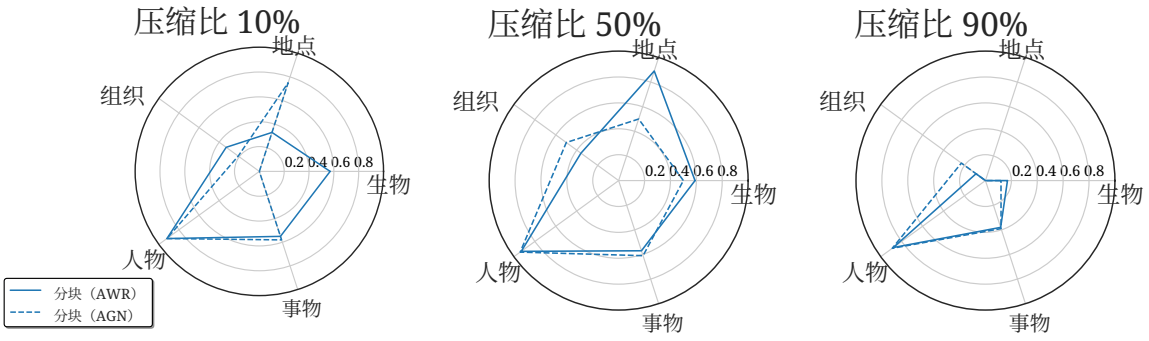


(c) Qwen-2.5 14B Instruct

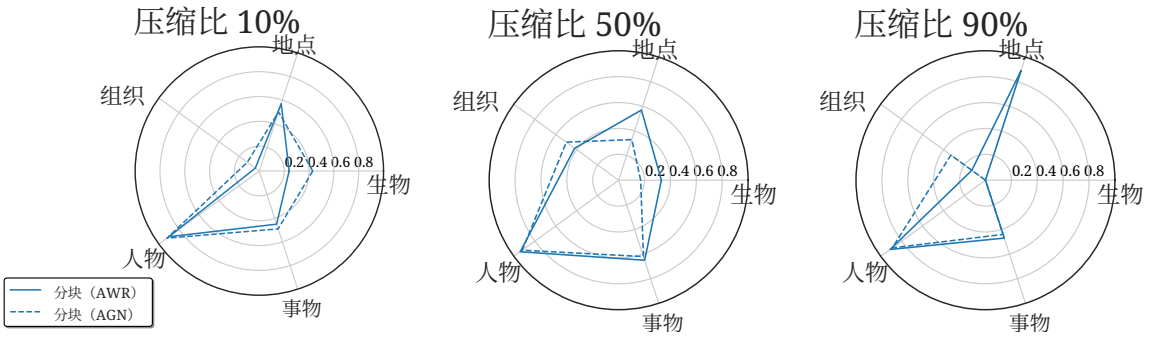
图 40：探测结果仍显示出严重的不一致性，因为压缩、组织标签和生物标签仍存在问题。



(a) LLaMA-3.2 3B Instruct

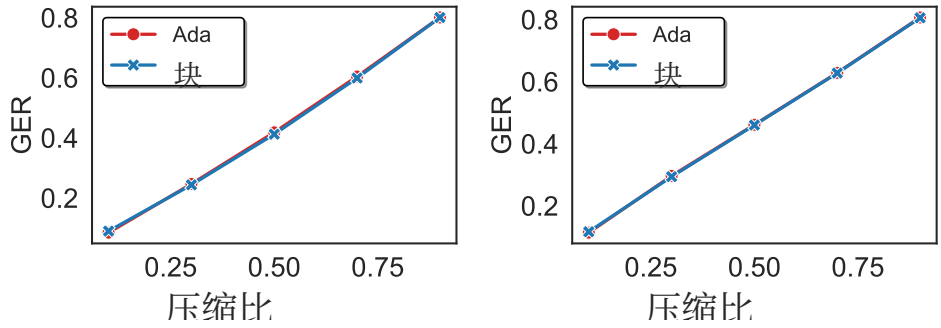


(b) Qwen-2.5 3B Instruct



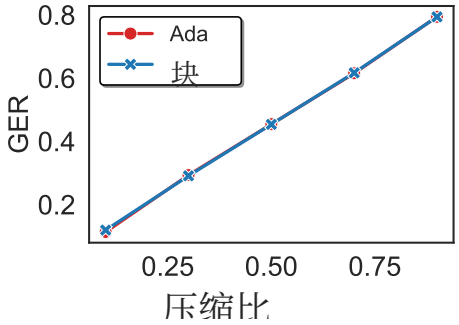
(c) Qwen-2.5 14B Instruct

图 41：尽管涉及的模型规模和架构各不相同，问题感知压缩和无关压缩仍会导致截然不同的理解。



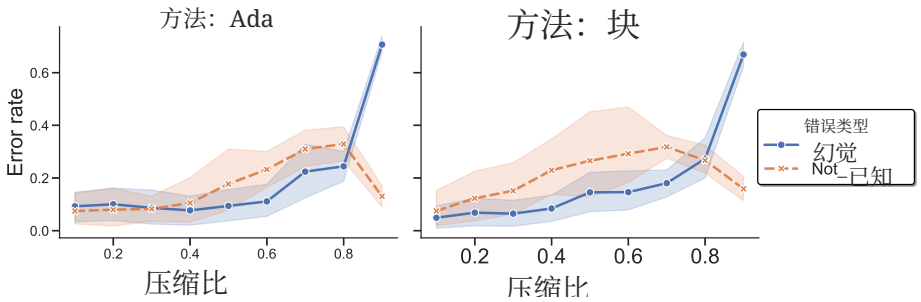
(a) LLaMA-3.2 3B Instruct

(b) Qwen-2.5 3B Instruct

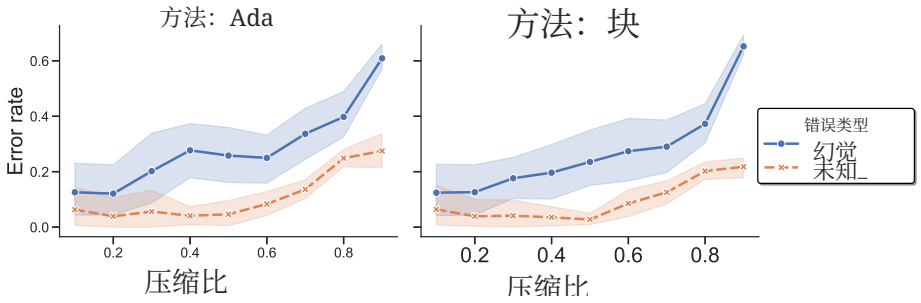


(c) Qwen-2.5 14B Instruct

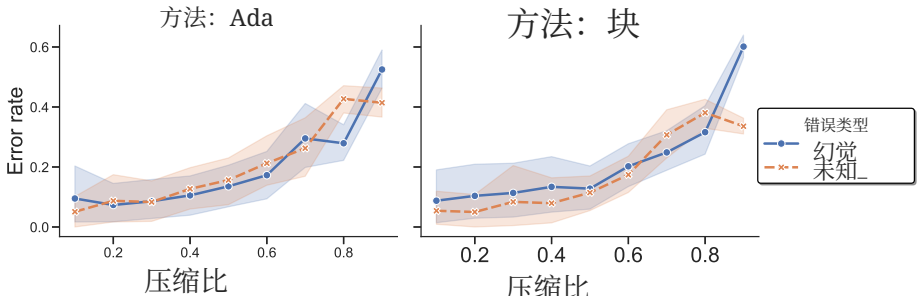
图 42：即使模型规模发生变化，淘汰趋势仍然与下游任务无关。



(a) LLaMA-3.2 3B Instruct

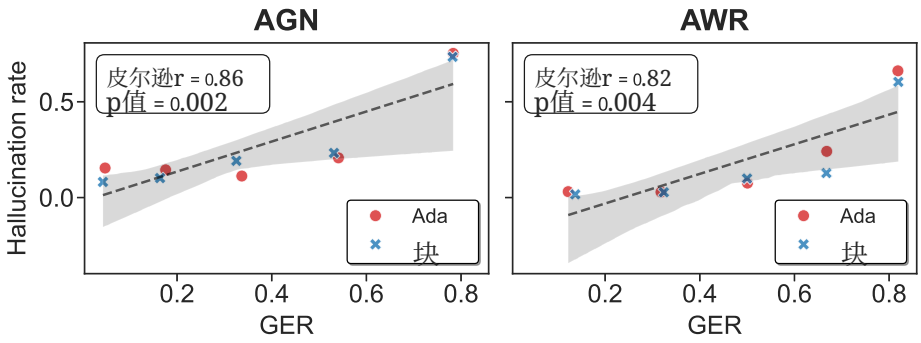


(b) Qwen-2.5 3B Instruct

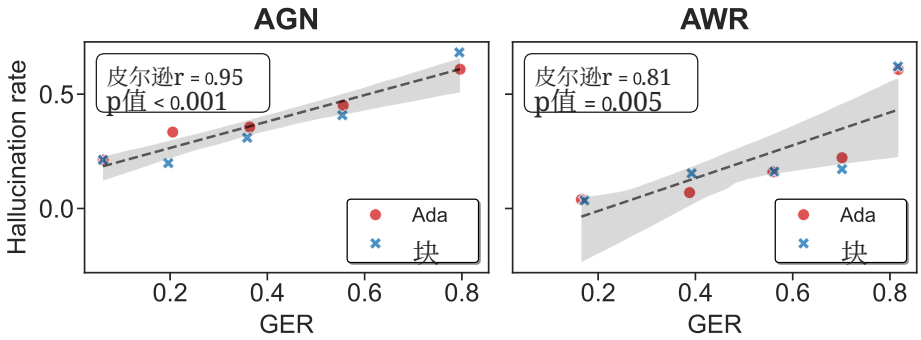


(c) Qwen-2.5 14B Instruct

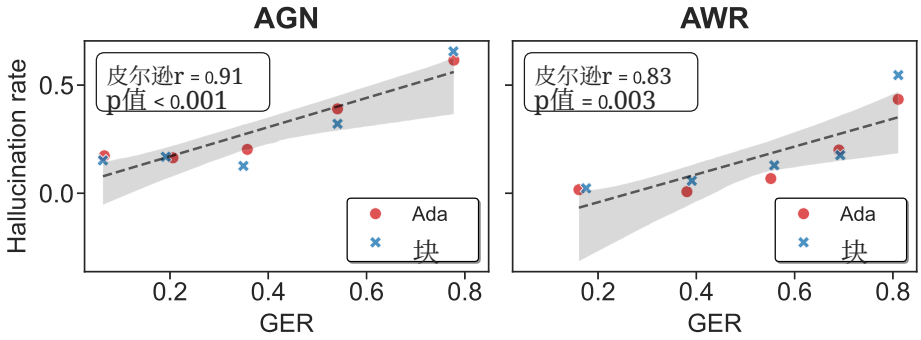
图 43：与较大的模型（Qwen 2.5 14B）相比，较小的模型（AdaKV 中未知错误增多）仍然存在脆弱性，但幻觉断崖相似。AdaKV 中未知错误增多的趋势依然存在。



(a) LLaMA-3.2 3B Instruct



(b) Qwen-2.5 3B Instruct



(c) Qwen-2.5 14B Instruct

图44：所有模型中 GER 与幻觉率的相关性。

参考文献

Joshua Ainslie、James Lee-Thorp、Michiel de Jong、Yury Zemlyanskiy、Federico Lebron 和 Sumit Sanghai。在*2023*年自然语言处理实证方法会议论文集中发表《GQA：从多头检查点训练广义多查询Transformer模型》，第4895–4901页，2023年。

Zeyuan Allen-Zhu。ICML 2024 教程：语言模型物理学，2024年7月。项目页面：https://physics.allen-zhu.com/。

Zeyuan Allen-Zhu 和 Yuanzhi Li。《语言模型物理学：第3.1部分，知识存储与提取》，2024年。网址：https://arxiv.org/abs/2309.14316。

Zeyuan Allen-Zhu 和 Yuanzhi Li。《语言模型物理学：第1部分，学习层次化语言结构》。《style id='1'>机器学习研究汇刊》，2025年。国际标准连续出版物号 2835-8856。网址：https://openreview.net/forum?id=mPQKyzkA1K。

Yu Bai、Xiyuan Zou、Heyan Huang、Sanxing Chen、Marc-Antoine Rondeau、Yang Gao与Jackie CK Cheung。CItruS：面向长序列建模的分块指令感知状态淘汰。载于Yaser Al-Onaizan、Mohit Bansal与Yun-Nung Chen编，《*2024*年自然语言处理实证方法会议论文集》，第5908–5930页，美国佛罗里达州迈阿密，2024年11月。计算语言学协会。DOI: 10.18653/v1/2024.emnlp-main.338。URL https://aclanthology.org/2024.emnlp-main.338/。

Yushi Bai、Xin Lv、Jiajie Zhang、Hongchang Lyu、Jiankai Tang、Zhidian Huang、Zhengxiao Du、Xiao Liu、Aohan Zeng、Lei Hou等。LongBench：面向长上下文理解的双语多任务基准。*arXiv*预印本 *arXiv:2308.14508*，2023。Iz Beltagy、Matthew E. Peters与Arman Cohan。Longformer：长文档Transformer，2020年12月。URL http://arxiv.org/abs/2004.05150。编号：arXiv:2004.05150 arXiv:2004.05150 [cs]。Zefan Cai、Yichi Zhang、Bofei Gao、Yuliang Liu、Yucheng Li、Tianyu Liu、Keming Lu、Wayne Xiong、Yue Dong、Junjie Hu与Wen Xiao。PyramidKV：基于金字塔式信息汇聚的动态KV缓存压缩。载于第二届语言建模会议，2025年。URL https://openreview.net/forum?id=ayi7qezU87。Vivek Chari与Benjamin Van Durme。Compactor：基于近似杠杆分数的校准的查询无关KV缓存压缩，2025年。URL https://arxiv.org/abs/ 2507.08143。Rewon Child。使用稀疏变换器生成长序列。*arXiv*预印本 *arXiv:1904.10509*，2019。

Krzysztof Marcin Choromanski、Valerii Likhosherstov、David Dohan、Xingyou Song、Andreea Gane、Tamas Sarlos、Peter Hawkins、Jared Quincy Davis、Afroz Mohiuddin、Lukasz Kaiser、David Benjamin Belanger、Lucy J Colwell 和 Adrian Weller。《重新思考使用 Performers 实现的注意力》。载于国际学习表征会议，2021年。网址：https://openreview.net/forum?id=Ua6zuk0WRH。

Kevin Clark、Urvashi Khandelwal、Omer Levy 和 Christopher D Manning。BERT 关注什么？对 BERT 注意力的分析。载于*2019 年 ACL Workshop BlackboxNLP*：分析和解读 *NLP* 神经网络，第 276–286 页，2019年。Giulio Corallo 和 Paolo Papotti。Finch：面向大语言模型的提示引导的键值缓存压缩。计算语言学协会汇刊，12：1517–1532，2024年 11 月。国际标准连续出版物号 2307-387X。DOI：10.1162/tacl a 00716。网址：https://doi. org/10.1162/tacl_a_00716。Tri Dao。FlāshAttention-2：通过更好的并行性和工作划分实现更快的注意力。载于第十二届国际学习表征会议，2024年。网址：https: //openreview.net/forum?id=mZn2Xyh9Ec。

Tri Dao、Daniel Y Fu、Stefano Ermon、Atri Rudra 和 Christopher Re。FlashAttention：具有 I/O 感知的快速且内存高效的精确注意力。载于 Alice H. Oh、Alekh Agarwal、Danielle Belgrave 和 Kyunghyun Cho 编，神经信息处理系统进展，2022年。URL https://openreview.net/forum?id=H4DqfPSibmx。

Alessio Devoto、Maximilian Jeblick 和 Simon J´egou。Expected Attention：通过估计未来查询分布中的注意力实现 KV缓存压缩，2025年。URL https://arxiv.org/abs/2510.00636。

Nelson Elhage、Neel Nanda、Catherine Olsson、Tom Henighan、Nicholas Joseph、Ben Mann、Amanda Askell、Yuntao Bai、Anna Chen、Tom Conerly、Nova DasSarma、Dawn Drain、Deep Ganguli、Zac Hatfield-Dodds、Danny Hernandez、Andy Jones、Jackson Kernion、Liane Lovitt、Kamal Ndousse、Dario Amodei、Tom Brown、Jack Clark、Jared Kaplan、Sam McCandlish 和 Chris Olah。Transformer 电路的数学框架。*Transformer Circuits Thread*，2021。https://transformer- circuits.pub/2021/framework/index.html。

Ibrahim Talha Ersoy、Andr´es Fernando Cardozo Licha 和 Karoline Wiesner。相变揭示了深度神经网络中的层次结构。*arXiv* 预印本 *arXiv:2512.11866*，2025年。

Yuan Feng、Junlin Lv、Yukun Cao、Xike Xie 和 S. Kevin Zhou。Ada-KV：通过自适应预算分配优化 KV 缓存淘汰，实现高效的大语言模型推理，2025年1月。网址：http://arxiv.org/abs/2407.11550。arXiv:2407.11550 [cs]。

Jonathan Frankle 和 Michael Carbin。彩票假设：寻找稀疏且可训练的神经网络。载于国际学习表征会议，2019。网址：https://openreview.net/forum?id=rJl-b3RcF7。

Suyu Ge、Yunan Zhang、Liyuan Liu、Minjia Zhang、Jiawei Han 和 Jianfeng Gao。模型会告诉你该丢弃什么：面向大语言模型的自适应 KV 缓存压缩。载于第十二届国际学习表征会议，2024年。网址：https://openreview. net/forum?id=uNrFpDPMyo。

Aaron Grattafiori、Abhimanyu Dubey、Abhinav Jauhri、Abhinav Pandey、Abhishek Kadian、Ahmad Al-Dahle、Aiesha Letman、Akhil Mathur、Alan Schelten、AlexVaughan 等。《LLaMA 3 模型家族》。*arXiv* 预印本 *arXiv:2407.21783*，2024年。

Albert Gu 和 Tri Dao。Mamba：使用选择性状态空间进行线性时间序列建模。载于第一届语言建模会议，2024年。

Jordan Hoffmann、Sebastian Borgeaud、Arthur Mensch、Elena Buchatskaya、Trevor Cai、Eliza Rutherford、Diego de Las Casas、Lisa Anne Hendricks、Johannes Welbl、Aidan Clark 等。训练计算最优大语言模型。载于第*36*届神经信息处理系统国际会议论文集，第30016–30030页，2022年。

Coleman Hooper、Sehoon Kim、Hiva Mohammadzadeh、Michael W Mahoney、Yakun S Shao、Kurt Keutzer 和 Amir Gholami。Kvquant：通过KV缓存量化实现1000万上下文长度的大语言模型推理。神经信息处理系统进展，37：1270–1303，2024年。

Cheng-Ping Hsieh、Simeng Sun、Samuel Kriman、Shantanu Acharya、Dima Rekesh、Fei Jia、Yang Zhang 和 Boris Ginsburg。RULER：你的长上下文语言模型的实际上下文大小是多少？*arXiv*预印本 *arXiv:2404.06654*，2024年。

Zhongzhan Huang、Senwei Liang、Mingfu Liang、Wei He、Haizhao Yang 和 Liang Lin。《卷积神经网络中自注意力的彩票假设》，2022年。网址：https://arxiv.org/abs/2207.07858。

Bowen Jin、Jinsung Yoon、Jiawei Han 和 Sercan Arık。《长上下文大语言模型与检索增强生成相遇：克服检索增强生成中长输入的挑战》。载于第*13*届国际学习表征会议（*ICLR 2025*），第65721–65759页。国际学习表征会议，ICLR，2025年。

Hao Kang、Qingru Zhang、Souvik Kundu、Geonhwa Jeong、Zaoxing Liu、Tushar Krishna 和 Tuo Zhao。《Gear：一种用于大语言模型近乎无损生成式推理的高效KV缓存压缩方案》。*arXiv*预印本 *arXiv:2403.05527*，2024年。

Jared Kaplan、Sam McCandlish、Tom Henighan、Tom B Brown、Benjamin Chess、Rewon Child、Scott Gray、Alec Radford、Jeffrey Wu 和 Dario Amodei。神经语言模型的缩放定律。*arXiv* 预印本 *arXiv:2001.08361*，2020。

Jang-Hyun Kim、Jinuk Kim、Sangwoo Kwon、Jae W. Lee、Sangdoo Yun 和 Hyun Oh Song。Kvzip：基于上下文重建的查询无关 KV 缓存压缩，2025年。网址：https://arxiv.org/abs/2505.23416。

Hunjae Lee。从图神经网络的视角理解 Transformer 的失效模式。*arXiv* 预印本 *arXiv:2512.09182*，2025年。

Haoyang Li、Yiming Li、Anxin Tian、Tianhao Tang、Zhanchao Xu、Xuejia Chen、Nicole Hu、Wei Dong、Qing Li 和 Lei Chen。基于 KV 缓存管理的大语言模型加速综述。*arXiv* 预印本 *arXiv:2412.19442*，2024a。

Yuankai Li、Jia-Chen Gu、Di Wu、Kai-Wei Chang 和 Nanyun Peng。简报：通过压缩连接检索与推理，以实现多跳推理。载于计算语言学协会年会论文集：北美计算语言学协会年会 *2025*，第5449–5470 页，2025年。

Yuhong Li、Yingbing Huang、Bowen Yang、Bharat Venkitesh、Acyr Locatelli、Hanchen Ye、Tianle Cai、Patrick Lewis 和 Deming Chen。SnapKV：大语言模型在生成之前就知道你要找什么，2024年6月b。网址：http://arxiv.org/abs/2404.14469。arXiv:2404.14469 [cs]。Guangda Liu、Chengwei Li、Jieru Zhao、Chenqi Zhang 和 Minyi Guo。ClusterKV：在语义空间中操控大语言模型KV缓存，实现可召回压缩。载于*2025*年第*62*届*ACM/IEEE*设计自动化会议（*DAC*），第 1–7 页。IEEE，2025a。Guangda Liu、Chengwei Li、Jieru Zhao、Chenqi Zhang 和 Minyi Guo。ClusterKV：在语义空间中操控大语言模型KV缓存，实现可召回压缩，2025b。网址：https://arxiv.org/abs/2412.03213。Jiaheng Liu、Dawei Zhu、Zhiqi Bai、Yancheng He、Huanxuan Liao、Haoran Que、Zekun Wang、Chenchen Zhang、Ge Zhang、Jiebin Zhang 等。长上下文语言建模综合综述。*arXiv*预印本 *arXiv:2503.17407*，2025c。Xiang Liu、Zhenheng Tang、Peijie Dong、Zeyu Li、Bo Li、Xuming Hu 和 Xiaowen Chu。ChunkKV：用于高效长上下文大语言模型推理的语义保持型KV缓存压缩。*arXiv*预印本 *arXiv:2502.00299*，2025d。Zichang Liu、Aditya Desai、Fangshuo Liao、Weitao Wang、Victor Xie、Zhaozhuo Xu、Anastasios Kyrillidis 和 Anshumali Shrivastava。Scissorhands：利用重要性假设的持续性，在测试时压缩大语言模型KV缓存。载于第三十七届神经信息处理系统大会，2023年。网址：https://openreview.net/forum?id=JZfg6wGi6g。Zirui Liu、Jiayi Yuan、Hongye Jin、Shaochen Zhong、Zhaozhuo Xu、Vladimir Braverman、Beidi Chen 和 Xia Hu。Kivi：一种免调优非对称2比特KV缓存量化方法。载于国际机器学习会议，第 32332–32344 页。PMLR，2024年。Ziyang Ma、Zuchao Li、Lefei Zhang、Gui-Song Xia、Bo Du、Liangpei Zhang 和 Dacheng Tao。大语言模型压缩中的相变。*npj*人工智能，2(1):21，2026年。William Meng、Benjamin Lee 和 Hong Wang。理解KV卸载下高效提供大语言模型推理服务的瓶颈。*arXiv*预印本 *arXiv:2601.19910*，2025年。Niklas Muennighoff、Alexander M. Rush、Boaz Barak、Teven Le Scao、Aleksandra Pik-tus、Nouamane Tazi、Sampo Pyysalo、Thomas Wolf 和 Colin Raffel。扩展数据受限的语言模型。机器学习研究期刊，26(53):1–66，2025年。网址：http://jmlr.org/papers/v26/24-1000.html。Catherine Olsson、Nelson Elhage、Neel Nanda、Nicholas Joseph、Nova DasSarma、Tom Henighan、Ben Mann、Amanda Askell、Yuntao Bai、Anna Chen 等。上下文学习与归纳头。*arXiv*预印本 *arXiv:2209.11895*，2022年。Hikari Otsuka、Daiki Chijiwa、Yasuyuki Okoshi、Daichi Fujiki、Susumu Takeuchi 和 Masato Motomura。多头注意力机制的强彩票假设。*arXiv*预印本 *arXiv:2511.04217*，2025a。

Hikari Otsuka、Yasuyuki Okoshi、Daichi Fujiki、Susumu Takeuchi、Masato Motomura 和 Daiki Chijiwa。论随机加权多头注意力机制中隐藏子网络的存在性。载于*High-dimensional Learning Dynamics 2025*，2025b。网址：https://openreview.net/forum?id=gB1LZhA8Oy。

Ziran Qin、Yuchen Cao、Mingbao Lin、Wen Hu、Shixuan Fan、Ke Cheng、Weiyao Lin 和 Jianguo Li。CAKE：结合层偏好的级联式自适应 KV 缓存淘汰。载于第十三届国际学习表征会议，2025年。网址：https://openreview.net/forum?id=EQgEMAD4kv。

Ayan Sengupta、Siddhant Chaudhary 和 Tanmoy Chakraborty。通过近似 CUR 分解实现面向大语言模型的价值引导型 KV 压缩。载于 *The Thirty-ninth Annual Conference on Neural Information Processing Systems*，2025年。网址：https://openreview.net/forum?id=klmc4fwPLd。

Noam Shazeer。《快速 Transformer 解码：你只需要一个写入头》，2019年。网址：https://arxiv.org/abs/1911.02150。

Yutao Sun、Zhenyu Li、Yike Zhang、Tengyu Pan、Bowen Dong、Yuyi Guo 和 Jianyong Wang。《大语言模型的高效注意力机制：综述》，2026年。网址：https://arxiv.org/abs/2507.19595。

Gemini Team、Petko Georgiev、Ving Ian Lei、Ryan Burnell、Libin Bai、Anmol Gulati、Garrett Tanzer、Damien Vincent、Zhufeng Pan、Shibo Wang、Soroosh Mariooryad、Yifan Ding 等。《Gemini 1.5：解锁数百万词元上下文的多模态理解》，2024年。网址：https://arxiv.org/abs/2403.05530。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser 和 Illia Polosukhin。《Attention Is All You Need》，2017年12月。网址：http://arxiv.org/abs/1706.03762。编号：arXiv:1706.03762 arXiv:1706.03762[cs]。

Elena Voita、Rico Sennrich 和 Ivan Titov。Transformer 中表征的自下而上演变：一项关于机器翻译和语言建模目标的研究。*arXiv* 预印本 *arXiv:1909.01380*，2019。

Zhongwei Wan、Xinjian Wu、Yu Zhang、Yi Xin、Chaofan Tao、Zhihong Zhu、Xin Wang、Siqi Luo、Jing Xiong、Longyue Wang 和 Mi Zhang。$\text{D} {2}\text{O}$：用于大语言模型高效长上下文推理的动态判别操作。载于第十三届国际学习表征会议，2025年。网址：https://openreview.net/forum?id=HzBfoUdjHt。

Sinong Wang、Belinda Z Li、Madian Khabsa、Han Fang 和 Hao Ma。Linformer：具有线性复杂度的自注意力。*arXiv* 预印本 *arXiv:2006.04768*，2020。

Guangxuan Xiao、Yuandong Tian、Beidi Chen、Song Han 和 Mike Lewis。《利用注意力汇实现高效流式语言模型》，2024年4月。网址：http://arxiv.org/abs/2309.17453。arXiv:2309.17453 [cs]。

An Yang、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chengyuan Li、Dayiheng Liu、Fei Huang、Haoran Wei 等。《Qwen2.5 技术报告》。*arXiv* 预印本 *arXiv:2412.15115*，2024a。

An Yang、Bowen Yu、Chengyuan Li、Dayiheng Liu、Fei Huang、Haoyan Huang、Jiandong Jiang、Jianhong Tu、Jianwei Zhang、Jingren Zhou、Junyang Lin、Kai Dang、Kexin Yang、Le Yu、Mei Li、Minmin Sun、Qin Zhu、Rui Men、Tao He、Weijia Xu、Wenbiao Yin、WenyuanYu、Xiafei Qiu、Xingzhang Ren、Xinlong Yang、Yong Li、Zhiying Xu 和 Zipeng Zhang。Qwen2.5-1m 技术报告，2025年。网址：https://arxiv.org/abs/2501.15383。

June Yong Yang、Byeongwook Kim、Jeongin Bae、Beomseok Kwon、Gunho Park、Eunho Yang、Se Jung Kwon 和 Dongsoo Lee。《不遗漏任何 token：通过重要性感知混合精度量化实现可靠的 KV缓存压缩》，2024b。网址：https: //arxiv.org/abs/2402.18096。

Manzil Zaheer、Guru Guruganesh、Kumar Avinava Dubey、Joshua Ainslie、Chris Alberti、Santiago Ontanon、Philip Pham、Anirudh Ravula、Qifan Wang、Li Yang 等。Big Bird：用于更长序列的 Transformer。神经信息处理系统进展，33：17283–17297，2020。

Minwei Zhang、Haifeng Sun、Jingyu Wang、Shaolong Li、Wanyi Ning、Qi Qi、Zirui Zhuang 和 Jianxin Liao。ClusterAttn：内在注意力聚类下的 KV缓存压缩。载于 Wanxiang Che、Joyce Nabende、Ekaterina Shutova 和 Mohammad Taher Pilehvar（编辑），第*63*届计算语言学协会年会论文集（第*1*卷：长文本论文），第14451–14473页，奥地利维也纳，2025年7月。计算语言学协会。ISBN 979-8-89176-251-0。DOI：10.18653/v1/2025.acl-long.703。URL https://aclanthology.org/2025.acl-long.703/。

Xinrong Zhang、Yingfa Chen、Shengding Hu、Zihang Xu、Junhao Chen、Moo Hao、Xu Han、Zhen Thai、Shuo Wang、Zhiyuan Liu和Maosong Sun。∞Bench：将长上下文评测扩展至100K词元以上。载于Lun-Wei Ku、Andre Martins和Vivek Srikumar编，第*62*届计算语言学协会年会论文集（第*1*卷：长文本论文），第15262–15277页，泰国曼谷，2024年8月。计算语言学协会。网址 https://aclanthology.org/2024.acl-long.814。

Zhenyu Zhang、Ying Sheng、Tianyi Zhou、Tianlong Chen、Lianmin Zheng、Ruisi Cai、Zhao Song、Yuandong Tian、Christopher R´e、Clark Barrett、Zhangyang Wang和Beidi Chen。H$ 2$O：用于大语言模型高效生成式推理的重击者预言机，2023年12月。网址 http://arxiv.org/abs/2306.14048。arXiv:2306.14048 [cs]。

Xiabin Zhou、Wenbin Wang、Minyan Zeng、Jiaxian Guo、Xuebo Liu、Li Shen、Min Zhang 和 Liang Ding。DynamicKV：面向长文本上下文大语言模型的任务感知型自适应 KV 缓存压缩。载于 Christos Christodoulopoulos、Tanmoy Chakraborty、Carolyn Rose 和 Violet Peng（编），计算语言学协会论文集：*EMNLP 2025*，第 8042–8057 页，中国苏州，2025年11月。计算语言学协会。ISBN 979-8-89176-335-7。DOI：10.18653/v1/2025.findings-emnlp.426。网址：https://aclanthology.org/2025.findings-emnlp.426/。

Jiaying Zhu、Dong Li、Xueyang Fu、Gege Shi、Jie Xiao、Aiping Liu 和 Zheng-Jun Zha。采用稀疏微调和 MAE 的彩票假设方法，用于图像伪造检测与定位。载于《人工智能促进协会会议论文集》，第 39 卷，第 10968–10976 页，2025年。