面向提示高亮的谱注意力引导

Weixian Waylon Li[1], **Yuchen Niu**[2],**YongxinYang**[4],**Keshuang Li**[3],
Tiejun Ma[1], **Shay B. Cohen**[1]
[1]爱丁堡大学，英国 [2]RayNeo，中国 [3]华为研究有限公司，英国[4]伦敦玛丽
女王大学，英国{waylon.li,tiejun.ma}@ed.ac.uk scohen@inf.ed.ac.uk

arXiv:2603.01281v1 [cs.CL] 1 Mar 2026

摘要

注意力引导是控制模型关注点的一项重要技术，可实现提示高亮等能力，使模型优先处理用户指定的文本。然而，现有注意力引导方法需要显式存储完整的注意力矩阵，因此无法与 FlashAttention 等内存高效实现兼容。我们提出谱编辑键放大（*SEKA*），一种免训练引导方法，通过在注意力计算之前直接编辑键嵌入来解决这一问题。*SEKA* 利用谱分解，将键嵌入引导至能够放大某些词元注意力分数的潜在方向。我们将其扩展为自适应 SEKA（*AdaSEKA*），一种查询自适应变体，使用免训练路由机制，根据提示的语义意图动态组合多个专家子空间。实验表明，这两种方法在标准引导基准测试上均显著优于强基线方法，同时延迟和内存额外开销大幅降低，并与优化注意力兼容。

1 引言

随着大语言模型越来越多地部署于高风险领域，精确引导其行为的能力至关重要。这一广泛的模型引导领域涵盖多种技术：从激活引导——通过干预 MLP 层来控制风格或事实回忆等高层次语义属性（Subramani 等人, 2022; Turner 等人, 2023年; Qiu等人, 2024; Turner 等人, 2024年; Wang 等人, 2025年; Stolfo 等人, 2025年），到注意力引导——在更细粒度的层面上运作，以将模型的关注点导向提示中的特定词元。本文聚焦于后者，其中提示高亮是关键应用之一。当前最先进的方法，如 PASTA（Zhang 等人，2024），通过在注意力分数矩阵计算完成之后对其进行编辑来运作。这种事后操作造成了一个关键瓶颈：它需要计算完整的注意力矩阵，从而使这些方法与 FlashAttention（Dao 等人, 2022; Dao, 2024年）等对高效处理至关重要的现代 I/O 感知实现不兼容。这一架构限制，再加上需要代价高昂的任务特定搜索来确定应引导哪些注意力头，降低了它们的实用性。

本文提出干预注意力机制的输入，而不是编辑其输出。我们提出谱编辑键放大（*SEKA*），这是一种新颖的免训练框架，通过在计算注意力分数之前直接修改键向量来引导注意力。我们的核心洞见是，对于给定任务，可以通过对源自对比提示词的键嵌入进行谱分解，学习一个通用的“相关性子空间”。随后，我们利用这些学习到的方向构造投影矩阵，通过一种简单且几何上可解释的变换来放大高亮键的相关特征：$\boldsymbol{k}' = \boldsymbol{k} + g\boldsymbol{P}\boldsymbol{k}$。

此外，我们提出自适应 SEKA（*AdaSEKA*），这是一种高级变体，可学习一组特定任务的“专家”投影（例如，分别用于事实回忆与指令遵循）。在推理时，*AdaSEKA* 采用计算代价极低的免训练路由机制，依据提示的语义意图混合这些专家，从而构建动态查询感知引导算子。我们的

方法与 FlashAttention 完全兼容，因为它直接作用于键嵌入，且计算额外开销可忽略不计。

我们的实验证实了该方法的有效性。*SEKA* 和*AdaSEKA* 在知识冲突、职业抽取和指令遵循的标准基准上均取得了优异成果。此外，*AdaSEKA*的查询自适应路由机制能够根据提示的语义意图动态调整引导方式，表现尤为出色。重要的是，我们表明，这些性能提升几乎没有带来额外开销。*SEKA* 每个样本仅增加 ≈0.03s 的延迟，而类似方法 PASTA 的推理时间为 +1.03s，内存使用量则近乎翻倍。

2 问题定义与动机

本节将提示高亮 形式化为注意力偏置问题，并阐述我们提出谱注意力引导方法的动机，该方法旨在解决现有方法的局限性。

问题定义。 给定由 T 个词元组成的提示 $\boldsymbol{x} = (x_1, \ldots, x_T)$，其中词元索引的一个子集 $\mathcal{H} \subset \{1, \ldots, T\}$ 标识高亮 词元（实际应用中，这些词元会用 ** 等标记括起来），我们的目标是引导模型的注意力，使查询更加关注这些词元。在标准多头注意力中，查询 i 与键 j 之间未归一化的注意力分数为 $\text{Attn}(i, j) = \frac{\boldsymbol{q}_i^\top \boldsymbol{k}_j}{\sqrt{d_k}}$，其中 $\boldsymbol{q}_i, \boldsymbol{k}_j \in \mathbb{R}^{d_k}$ 是查询向量和键向量，d_k 是头维度。

目标。 我们的目标是：为每个 $(i,\ j)$ 的注意力分数引入一个可控的加性项，从而放大分配给高亮词元的注意力，其中 $j \in \mathcal{H}$: $A'_{ij} = A_{ij} + \Delta_{ij}$，$\Delta_{ij}$ 旨在有选择地增强对用户指定高亮词元的注意力。

动机。 现有方法通常会在计算出注意力后再对其进行修改。例如，PASTA（Zhang 等人，2024年）会按照公式 1 所示重新缩放注意力矩阵的各行，其中 C_i 是行归一化因子，$\alpha > 1$ 用于缩放高亮词元的注意力。

$$[T(\boldsymbol{A})]_{ij} = \begin{cases} \alpha \dfrac{A_{ij}}{C_i}, & \text{if } j \in \mathcal{H}, \\ \dfrac{A_{ij}}{C_i}, & \text{otherwise.} \end{cases} \tag{1}$$

类似地，Found-in-the-Middle（Hsieh 等人，2024年）等位置校准方法会从位置注意力偏置中减去一个基线。设 x_k 表示第 k 个词元的位置，$\text{Attn}_{\text{ori}}(x_k)$ 表示原始位置偏置。校准后的偏置为 $\text{Attn}_{\text{calibrated}}(x_k) = \text{Attn}_{\text{ori}}(x_k) - \text{Attn}_{\text{baseline}}(x_k)$，其中 $\text{Attn}_{\text{baseline}}(x_k)$ 的估计不依赖于内容相关性。

这两种策略都需要显式存储完整的注意力矩阵，这与 FlashAttention 等内存高效实现不兼容（Dao 等人，2022；Dao，2024）。此外，PASTA 等方法通常依赖成本高昂的注意力头搜索，以确定要调整哪些注意力头。这些局限促使我们考虑一种替代性的调整机制，在计算注意力分数之前发挥作用，从而无需物化或修改注意力矩阵。由于注意力取决于查询与键的内积，因此可以通过编辑任一表征实现等效控制（见第 3.2 节）。鉴于我们的目标是增强对特定词元子集 $\mathcal{H}$ 的注意力，键侧干预是自然的选择：键向量 $\boldsymbol{k}_j$ 按词元位置 j 索引，因此决定了对每个词元的关注程度。

为提供实证证据，检验这种注意力前干预是否可行，我们分析键表征在上下文相关性变化时如何改变。我们首先在三种条件下构造合成对比提示三元组：**(1)** 中性（仅上下文），**(2)**正向（上下文与相关查询一致），以及 **(3)** 负向（上下文与不相关查询配对）。此类合成三元组的构造方法详见附录 A。

使用 Qwen3-1.7B-Base 模型（28 层、8 个注意力头），我们针对每个（层、注意力头）组合，提取正向提示和负向提示中对应相同词元跨度的键嵌入。我们

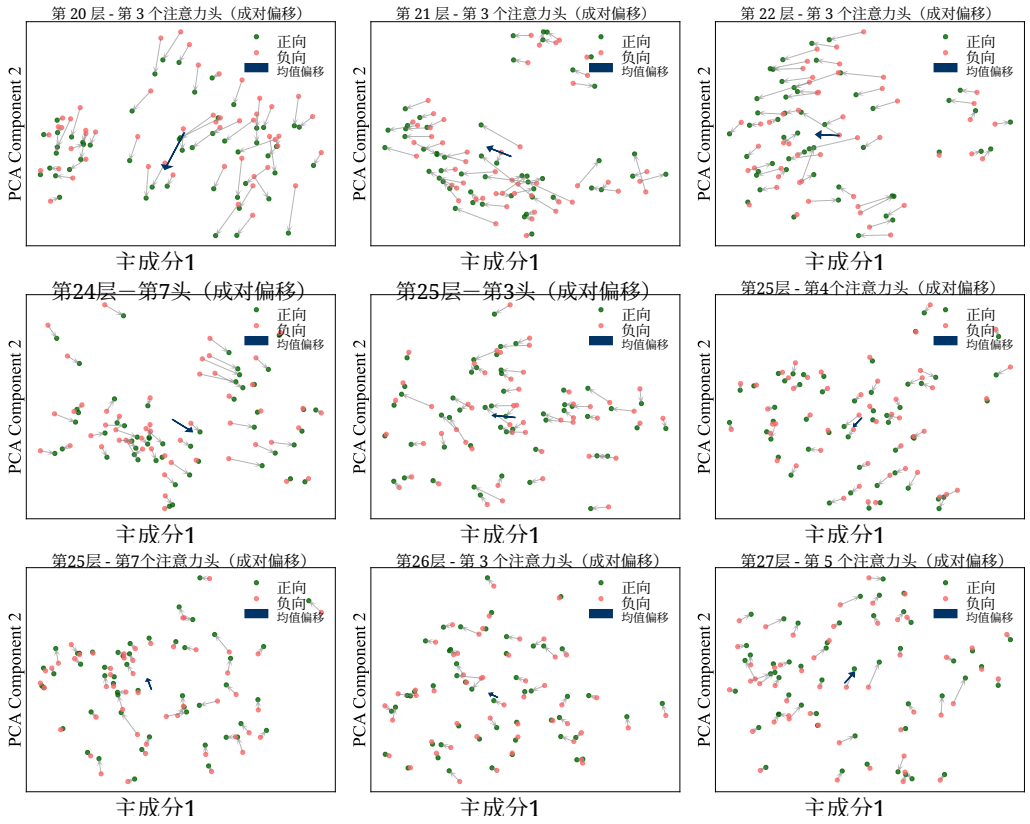


图1：通过主成分分析对 Qwen3-1.7B-Base 中不同（层、注意力头）的成对键嵌入偏移进行可视化。正向与负向表征绘制了 26 个共享词元跨度。灰色箭头描绘各个偏移；深蓝色箭头表示平均位移。

接着使用主成分分析将这些成对嵌入共同投影到二维空间，并结合散点图与有向箭头来呈现结果。每个箭头都从一个负向键出发，指向对应的正向键，以呈现问题相关性变化所引起的成对表征偏移。为总结整体趋势，我们还绘制了所有提示对的均值偏移向量。图1显示，当词元相关性发生变化时，某些注意力头的键嵌入会呈现稳健且一致的方向性偏移。每张图都展示了 26 组成对键嵌入，这些嵌入对应共享词元跨度，提取自 10 组正向—负向提示对。

这些发现表明，相关性编码在键表征的结构化子空间中，这启发了我们在计算注意力之前编辑键嵌入的方法： $\boldsymbol{k}'_j = \boldsymbol{k}_j + g\boldsymbol{P}\boldsymbol{k}_j$，其中 $\boldsymbol{P}$是投影矩阵（为每个键值头定义一个相关性子空间）， g 是缩放系数。该方法兼容高效注意力实现，同时提供了一种几何上可解释的机制，引导注意力转向高亮词元。

3 谱注意力引导实现提示高亮

如图 2 所示，我们提出了一种新方法，谱编辑键放大（*SEKA*），以及其查询自适应变体 *AdaSEKA*。两种方法都通过在注意力计算之前直接编辑键嵌入来实现提示高亮。*SEKA*的核心机制受到激活谱编辑（SEA）算法（Qiu等人, 2024）的启发，并将其从语义级激活引导改造为提示高亮所需的词元级注意力引导。

3.1 相关性对齐投影的谱学习（离线）

利用从上述合成对比提示（第2节和附录 A）获得的词元级键嵌入，分别记为 $\boldsymbol{h}$ （中性）、 $\boldsymbol{h}^+$ （正向）和 $\boldsymbol{h}^-$ （负向），我们

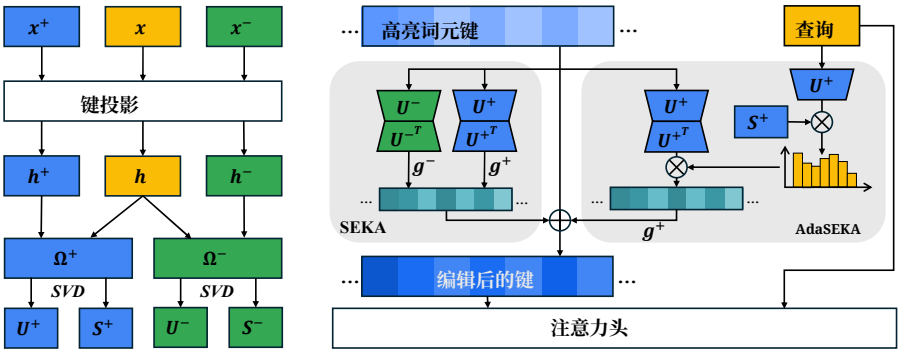


图 2：SEKA 和 AdaSEKA 概览。$\boldsymbol{x}$：上下文；$\boldsymbol{h}$：键嵌入；Ω：交叉协方差；$\boldsymbol{U}$：左奇异向量；$\boldsymbol{S}$：奇异值；g：增益系数。SEKA 使用固定增益，而 AdaSEKA 利用查询计算动态引导权重。

为每个 Transformer 层 ℓ 和键值头 h 计算交叉协方差矩阵：$\boldsymbol{\Omega}^{+}_{\ell,h} = \frac{\boldsymbol{h}^{+\top}_{\ell,h}\boldsymbol{h}^{+}_{\ell,h}}{n}$，$\boldsymbol{\Omega}^{-}_{\ell,h} = \frac{\boldsymbol{h}^{-\top}_{\ell,h}\boldsymbol{h}^{-}_{\ell,h}}{n}$，其中 n 是采样词元的数量。随后应用奇异值分解（奇异值分解）：$\boldsymbol{\Omega}^{+}_{\ell,h} = \boldsymbol{U}^{+}_{\ell,h}\boldsymbol{S}^{+}_{\ell,h}\boldsymbol{V}^{+\top}_{\ell,h}$，$\boldsymbol{\Omega}^{-}_{\ell,h} = \boldsymbol{U}^{-}_{\ell,h}\boldsymbol{S}^{-}_{\ell,h}\boldsymbol{V}^{-\top}_{\ell,h}$。

在奇异值分解中，$\boldsymbol{S}^{+}_{\ell,h}$ 和 $\boldsymbol{S}^{-}_{\ell,h}$ 分别表示正向和负向交叉协方差矩阵的奇异值。这些奇异值量化了投影的各个分量所捕获的交叉协方差大小。奇异值越大，相应的奇异向量（投影方向）在解释词元键嵌入之间的交叉协方差时就越重要。

在公式 2 中，对于正投影 $\boldsymbol{P}^{+}_{\ell,h}$，我们采用对应于最大奇异值的顶部奇异向量，以捕获与相关（突出显示）特征关联最强的方向。对于负投影 $\boldsymbol{P}^{-}_{\ell,h}$，我们采用最不显著的奇异向量，即与最小奇异值对应的向量，以针对与相关性关联最弱的方向。

$$\boldsymbol{P}^{+}_{\ell,h} = \boldsymbol{U}^{+}_{\ell,h,:,:k^{+}}(\boldsymbol{U}^{+}_{\ell,h,:,:k^{+}})^{\top}, \quad \boldsymbol{P}^{-}_{\ell,h} = \boldsymbol{U}^{-}_{\ell,h,:,:k^{-}}(\boldsymbol{U}^{-}_{\ell,h,:,:k^{-}})^{\top}, \tag{2}$$

其中 k^{+} 和 k^{-} 的选取应使其至少捕获总奇异值之和的比例 γ：

$$\frac{\sum_{i=1}^{k^{+}}\boldsymbol{S}^{+}_{\ell,h,i}}{\sum_{i=1}^{d_k}\boldsymbol{S}^{+}_{\ell,h,i}} \geq \gamma, \quad \frac{\sum_{i=1}^{k^{-}}\boldsymbol{S}^{-}_{\ell,h,i}}{\sum_{i=1}^{d_k}\boldsymbol{S}^{-}_{\ell,h,i}} \geq \gamma. \tag{3}$$

阈值 γ 是一个超参数，用于控制在创建投影矩阵时希望保留数据中多少方差。通过为正向协方差选取前 k^{+} 个奇异向量、为负向协方差选取前 k^{-} 个奇异向量，我们捕获每种投影在键嵌入中最相关的方向。学习得到的投影器 $\{\boldsymbol{P}^{+}_{\ell,h}, \boldsymbol{P}^{-}_{\ell,h}\}$ 按层和头存储，从而能够在推理时进行细粒度引导。

3.2 面向高亮词元的谱编辑（推理）

在推理过程中，*SEKA*会在计算注意力分数之前，将学习得到的投影注入键嵌入。为清晰起见，我们省略键向量 $\boldsymbol{k}_j$ 和查询 $\boldsymbol{q}_i$上的显式（ℓ，h）索引，尽管它们在实践中是按层和按头区分的。对于第 ℓ 层和第 h头的每个词元键 $\boldsymbol{k}_j \in \mathbb{R}^{d_k}$，编辑后的嵌入定义为：

$$\boldsymbol{k}'_j = \boldsymbol{k}_j + \frac{g^{+}\cdot\boldsymbol{P}^{+}_{\ell,h}\boldsymbol{k}_j + g^{-}\cdot\boldsymbol{P}^{-}_{\ell,h}\boldsymbol{k}_j}{2}, \tag{4}$$

其中，$\boldsymbol{P}^{+}_{\ell,h}$，$\boldsymbol{P}^{-}_{\ell,h} \in \mathbb{R}^{d_k\times d_k}$为所选的投影矩阵，$g^{+}$，$g^{-}$为两个可独立调节的标量，用于控制正向和负向引导增益。除非另有说明，所有向量（例如 $\boldsymbol{k}_j$、$\boldsymbol{q}_i$、

$\boldsymbol{x}$）均为列向量。此调整会按照公式 5 修改注意力logits，其中 $\boldsymbol{q}_i \in \mathbb{R}^{d_k}$ 是第 i 个查询向量。这在代数上等同于在原始方法的注意力分数矩阵 $\boldsymbol{A}$ 上增加一个低秩相关性偏置矩阵 $\boldsymbol{B}$:

$$\text{Logits}_{ij} = \frac{\boldsymbol{q}_i^\top \boldsymbol{k}_j}{\sqrt{d_k}} + \frac{\boldsymbol{q}_i^\top \left(\dfrac{g^+ \cdot \boldsymbol{P}^+_{\ell,h} \boldsymbol{k}_j + g^- \cdot \boldsymbol{P}^-_{\ell,h} \boldsymbol{k}_j}{2} \right)}{\sqrt{d_k}} = \boldsymbol{A}_{ij} + \boldsymbol{B}_{ij}. \tag{5}$$

因此，*SEKA* 可以解释为在注意力分数中添加一个键相关项，从而放大每个词元与相关性子空间对齐的方向（详见附录 C）。与直接操纵注意力矩阵的方法不同，*SEKA* 通过编辑键向量本身实现等效调制，提供了一种结构更清晰、可解释性更强的机制。此外，由于*SEKA* 完全在注意力计算之前对键表征进行操作，因此无需访问或存储注意力矩阵，天然兼容 FlashAttention 等内存高效实现。

3.3 变体：查询驱动的自适应 SEKA

虽然标准*SEKA* 框架能够有效实现词元级注意力引导，但由于投影是静态的，实际部署时往往需要针对不同任务和模型系列进行超参数调优。为了解决这一局限并减少手动配置的需求，我们引入*Adaptive SEKA (*AdaSEKA*)*，根据查询特定相关性信号自动选择并组合专家投影。

多专家投影学习。 我们将投影学习框架扩展为可容纳多个领域特定专家。对于每个专家[1] $m \in \{1, \ldots, M\}$，我们从针对不同任务的数据集 $\mathcal{D}_m$ 中构建样本。每个专家都遵循标准的 *SEKA* 流程，学习自己的一组正向奇异值分解分量 $\{\boldsymbol{U}^+_{m,\ell,h}, \boldsymbol{S}^+_{m,\ell,h}, \boldsymbol{V}^+_{m,\ell,h}\}$。这一过程为每个专家、层和注意力头生成一组奇异值分解分量，可表示为一个五维张量 **($\boldsymbol{U}^+ \in \mathbb{R}^{M \times L \times H \times d_k \times d_k}$)**，其中 L 是层数， H 是注意力头数。

查询自适应专家路由。 在推理时，我们提取提示中最后一个词元在第ℓ 层、第 h 个注意力头处的查询向量 $\boldsymbol{q}_{\ell,\,h}$；由于最后一个词元充当提示信息的全局聚合器，并对下游生成产生重大影响（Barbero 等，2024年；Qiu 等，2024年）。随后，我们计算动态系数，以确定每个专家的贡献:

$$\alpha_{m,\ell,h}(\boldsymbol{q}_{\ell,h}) = \frac{\sum_{k=1}^{K} (\boldsymbol{q}_{\ell,h}^\top \boldsymbol{u}_{m,\ell,h}^{+(k)}) \cdot \sigma_{m,\ell,h}^{+(k)}}{\max_m \left| \sum_{k=1}^{K} (\boldsymbol{q}_{\ell,h}^\top \boldsymbol{u}_{m,\ell,h}^{+(k)}) \cdot \sigma_{m,\ell,h}^{+(k)} \right|}, \tag{6}$$

其中， $\sigma_{m\,\ell}^{+(k)}{}_{,\,h}$ 是对应的第 k 个奇异值，而 K 是所使用的前几个奇异分量的数量（通常为 $K = 5$）。

该公式通过各自的奇异值加权，衡量查询与每个专家的主要投影方向之间的对齐程度。分母以所有专家中绝对对齐程度的最大值进行归一化，使系数保持在可比较的尺度上，同时保留对齐为正向还是负向的信息。

第 ℓ 层和第 h 个头部的最终投影矩阵由专家投影的加权组合构成：$\boldsymbol{P}_{\text{dynamic},\ \ell,\ h}(\boldsymbol{q}_{\ell,\ h}) = \sum_{m=1}^{M} \alpha_{m,\ \ell,\ h}(\boldsymbol{q}_{\ell,\ h}) \cdot \boldsymbol{U}^+_{m,\ \ell,\ h,\ :\ :\ K} (\boldsymbol{U}^+_{m,\ \ell,\ h,\ :\ :\ K})^\top$，其中$\boldsymbol{U}^+_{m,\ \ell,\ h,\ :\ :\ K}$表示 K 的前几列，即$\boldsymbol{U}^+_{m,\ \ell,\ h}$中最重要的奇异向量所对应的列。

该方法仅使用前 K 个分量按需重建投影矩阵，在实现自动专家选择的同时提高计算效率。推理过程中的关键变换变为：$\mathbf{k}'_j = \mathbf{k}_j + g \cdot \boldsymbol{P}_{\text{dynamic},\ \ell,\ h}(\boldsymbol{q}_{\ell,\ h})\ \mathbf{k}_j$。

至关重要的是，*AdaSEKA*具有多项实际优势：**(1)减少配置工作量**：自动专家路由减少了针对不同任务和模型进行超参数调优的次数

[1]专家在不同的特定任务数据集之间可能有所差异，例如事实纠正和指令遵循。

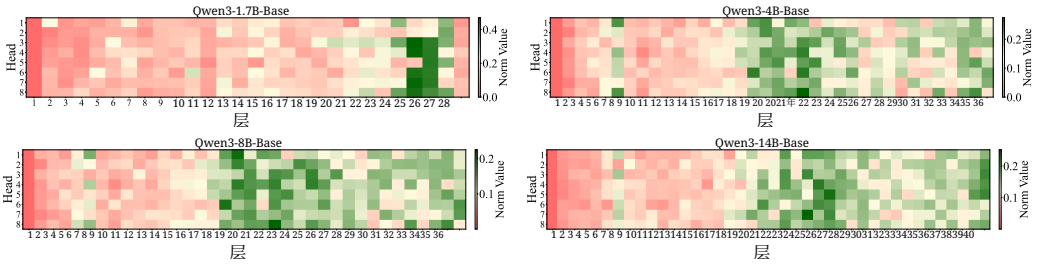


图3：四种规模的 Qwen3 模型中，所有 KV头和层上正向与负向键嵌入之间的逐词元平均距离热力图 ℓ_2。数值越高（绿色）表示正向和负向键表征之间的差异越大。

（见附录 F）。(2) **模块化部署**：无需重新计算现有专家即可集成新专家。(3) **可解释路由**：专家选择基于明确的查询-专家对齐分数。我们从四个不同的数据集中导出四个专家投影。附录 B 详细介绍了用于学习这些投影的数据样本构建过程。

3.4 选择对相关性敏感的键值注意力头

*SEKA*在选择性地应用于天然对提示相关性敏感的 KV头时效果最佳。正如图1中的定性可视化结果所示，并在第2节中讨论的那样，当提示中的问题从不相关变为相关时，给定词元跨度的键嵌入会在向量空间中持续发生偏移。本节将形式化一种方法，用于量化所有层和注意力头的这种相关性敏感度，以指导我们的选择策略。

图3展示了正向与负向键嵌入之间的距离 ℓ_2，该距离是在合成数据集（定义见附录 A）中对所有答案词元取平均得到的。我们考察了不同规模的 Qwen3 模型中不同层和注意力头之间的这种变化。

我们观察到，相关提示与不相关提示之间的差异并非均匀分布：较大的范数值（绿色）稳定地出现在中后层，而早期层以及部分注意力头的变化很小（红色），这表明检索行为不太可能发生在这些层中。这一发现与近期的机制分析高度一致。Michel 等人（2019）；Voita 等人（2019）；Clark 等人（2019）；Neo 等人（2024）；Li 等人（2023b）指出，不同注意力头会呈现各种词元关注模式。Qiu 等人（2025）证明，检索效能仅依赖一部分注意力头，这些注意力头通过探测和相关性筛选加以识别。Wu 等人（2025）进一步表明，这一稀疏的“检索头”集合几乎全部位于 Transformer 的中后层。这些注意力头是基础模型固有的，在微调后仍保持一致，并会根据上下文动态激活。因此，受这一一致性启发，我们仅对正负键嵌入之间的经验 ℓ_2 差异超过阈值的（层、注意力头）对进行投影。这种选择性方法确保注意力引导集中于经验上与检索行为相关的组件，同时不影响其他注意力头。通过这种方式，我们只在必要之处增强相关性信号，尽可能减少对模型其他无关组件的意外影响。

形式化地，对于每个层 ℓ 和注意力头 h，令 S 表示数据中所有样本的答案词元集合，其中 $|S| = N$。每个词元的平均 ℓ_2 距离计算如下：$D_{\ell,\,h} = \frac{1}{N}\sum_{i=1}^{N}\left\|\boldsymbol{h}^{+}_{\ell,\,h,\,i} - \boldsymbol{h}^{-}_{\ell,\,h,\,i}\right\|_2$，其中 $\boldsymbol{h}^{+}_{\ell,\,h,\,i}$ 和 $\boldsymbol{h}^{-}_{\ell,\,h,\,i}$ 分别是词元 i 在 S 中的正向和负向键嵌入。仅当 $D_{\ell,\,h} \geq \delta_{\text{min}}$ 时才应用投影，其中 δ_{min} 是通过在验证集上进行网格搜索来调优的可调超参数（通常在 $[0, 0.6]$中）。

4 实验设置

我们认为*SEKA* 在需要在提示中强调或突出内容的情景中特别有用。这包括用于评估 PASTA（Zhang 等人，2024）的任务，涉及 (i) 处理复杂的用户指令（例如代词改写），(ii) 解读冗长且嘈杂的上下文（例如 Bias in Bios；De-Arteaga 等人 2019），以及 (iii) 解决上下文内知识冲突（例如

CounterFact；Meng 等人 2022）。此外，只需突出长上下文的中间部分，*SEKA* 就能使我们扭转“迷失在中间”情景中观察到的典型 U 形效能（Liu 等人，2024），从而提升模型对这些难以处理的位置的回忆。

4.1 注意力引导的标准基准

我们遵循 PASTA 所使用的标准基准，以确保高亮词元的选择保持一致。表 1 汇总了各项任务、提示格式和评估指标。CounterFact 任务基于 COUNTERFACT 数据集（Meng 等人，2022），其余两项任务（Bias in Bios、代词变更）则源自 BIASBIOS 数据集（De-Arteaga等人，2019），与先前研究（Zhang 等人，2024）保持一致。我们改进了代词变更任务的评估指标，以解决原始协议中可能误导性地奖励空回复的缺陷，其余指标保持不变。更多细节，包括各项基准任务的介绍以及指标的计算方法，见附录E。

表 1：注意力引导标准基准摘要。**粗体词元** 表示应用注意力引导的位置。

Task	提示	指标
CounterFact	先前，*[*旧事实*]*。 当前，***[***新事实***]***。 *[*问题*]*。	效能分数（ES），短语分数（PS）
Bias in Bios	***[***人物的职业***]***。 *[*职业经历，可能与预测没有直接关系*]*。 *[*人物*]* 其职业为 ___	准确率（Acc.）
代词变更	*[*传记背景*]*。 **将“她”替换为并且将‘he’改为‘they’，并生成职业人物 *[*人物*]*，更改代词后**。	代词加权重叠分数（P. 分数），全变更 P. 分数

基准方法。 我们首先将原始模型的直接提示作为基线。此外，我们还采用另一种基线，在高亮的上下文周围加上 ** 标记。对于注意力引导方法，** 仅用于确定引导所需的词元索引，并会从输入 ID 中移除。随后，我们将提出的方法 *SEKA* 和*AdaSEKA*与现有的注意力引导方法 *PASTA*进行基准比较。我们还将其与 选择性提示锚定（*SPA*）（Tian & Zhang, 2025）进行比较，这是一种作用于大语言模型 logit 分布的提示高亮方法。此外，我们评估了应用随机投影且不使用 KV 头选择器的*SEKA* ，将其作为消融研究。

4.2 Lost-in-the-Middle 设置中的 U 型反转

为了进一步考察*SEKA*将模型注意力引导至长上下文中特定区域的能力，我们在具有挑战性的 Lost-in-the-Middle 设置中开展了另一项针对位置召回的实验（Liu等人，2024）。这一设置指的是一种普遍观察到的现象：大语言模型对长上下文开头和结尾处的信息召回能力较强，但当相关信息位于中间时，性能会显著下降，从而形成典型的 U 型性能曲线。每个输入都包含由 30 个段落组成的长上下文，其中只有一个标准段落包含给定问题的正确答案，其余段落均为干扰项。我们改变标准段落的位置，以测试模型对位置的敏感性。每个输入的格式如下：“上下文：\n *[P1* 标题 *]* \n *[P1* 正文 *]* ..... *[P30* 标题 *]* \n*[P30* 正文 *]* \n\n 问题：例如 *[’* 问题 *’]* \n 答案：”。

与旨在缓解这一效应的既有研究不同，我们的目标是直接探究通过*SEKA* 进行显式相关性高亮能否反转这一 U 形曲线。通过将注意力引导至中间段落，我们测试中间上下文答案常见的性能低谷能否转变为峰值，从而深入了解大语言模型中位置召回的可控性。

评估指标。 我们遵循 Liu 等人（2024）的做法，使用精确匹配（EM）分数作为评估指标：如果预测包含真实短答案片段，则视为正确。为避免生成冗长或偏离主题的回答，我们将生成答案限制为最多 60 个词元。

表 2：在标准基准上的性能。**粗体** = 表示最佳。下划线 = 表示次佳。我们针对*SEKA*进行了两项消融研究："不学习"使用随机投影，而非通过频谱学习得到的投影；"不学习且不筛选"则进一步移除了头部过滤机制。

模型	指标	基线方法				我们的方法			
		原始方法	**-标记	PASTA	SPA	*SEKA*	不进行学习	不进行学习和筛选	
Qwen3-4B	CounterFact（效能分数）	45.00	57.70	97.16	65.24	**99.02**	94.96	86.12	98.90
	CounterFact（PS）	45.64	52.12	96.03	57.71	98.61	92.38	86.20	**98.72**
	Bias in Bios（准确率）	79.84	82.94	89.58	68.00	91.02	86.62	71.76	**91.86**
	代词（P. 分数）	93.14	95.76	**95.82**	80.27	95.18	90.42	41.98	94.54
	代词（A. P. 分数）	90.52	93.88	**94.64**	78.19	93.26	88.66	36.95	92.08
Qwen3-8B	CounterFact（效能分数）	39.04	56.24	92.70	69.26	**99.08**	96.12	95.18	99.00
	CounterFact（PS）	39.59	49.80	91.68	58.76	98.96	94.74	89.69	**98.97**
	Bias in Bios（准确率）	76.08	80.60	86.32	37.02	**88.74**	87.26	74.90	88.50
	代词（P. 分数）	98.00	98.10	98.86	72.61	98.56	98.12	80.53	**99.68**
	代词（A. P. 分数）	97.84	97.84	98.72	74.95	98.26	97.90	80.85	**99.52**
Qwen3-14B	CounterFact（效能分数）	37.56	45.52	76.84	84.22	98.92	86.28	95.26	**99.00**
	CounterFact（PS）	36.12	40.12	66.33	76.11	99.02	88.07	92.02	**99.15**
	Bias in Bios（准确率）	85.22	90.94	88.46	57.86	90.28	88.02	88.44	**91.22**
	代词（P. 分数）	98.42	98.86	90.98	91.60	98.66	96.32	88.60	**99.88**
	代词（A. P. 分数）	98.22	98.68	90.94	92.20	98.54	96.36	89.76	**99.86**
Gemma3-4B	CounterFact（效能分数）	55.04	57.56	78.36	93.90	98.04	95.14	94.46	**98.74**
	CounterFact（PS）	47.77	45.82	59.53	91.92	98.83	92.25	91.98	**99.05**
	Bias in Bios（准确率）	89.90	91.00	82.58	48.02	92.42	85.60	77.16	**92.92**
	代词（P. 分数）	41.34	38.86	67.39	76.05	81.53	53.58	51.78	**93.76**
	代词（A. P. 分数）	35.25	32.45	66.43	74.45	81.11	48.82	51.94	**93.58**
Gemma3-12B	CounterFact（效能分数）	45.34	48.72	68.30	93.76	**98.86**	63.08	60.96	92.48
	CounterFact（PS）	37.21	36.69	71.72	91.24	**99.27**	50.59	76.37	93.65
	Bias in Bios（准确率）	91.26	92.90	**94.72**	46.88	93.04	91.84	90.54	91.14
	代词（P. 分数）	93.92	95.78	68.47	86.41	**97.70**	47.26	55.56	96.88
	代词（A. P. 分数）	94.96	96.42	68.01	84.99	**97.24**	51.24	58.76	95.84

基准方法。 我们将 *SEKA*与一种标准基线进行比较：直接向基础大语言模型提供提示，不进行任何干预；此外还与 *PASTA*进行比较。在此基础上，我们以两种配置应用 *SEKA*：(i) 仅对上下文的中间区域（具体而言，第 4 至第 25 个段落）进行引导；(ii) 对所有上下文段落进行引导。尽管 Hsieh 等人（2024）提出了另一种潜在的基线，但由于其代码实现无法获取，我们未将其纳入比较。

5 结果

5.1 标准基准：*SEKA*实现高效注意力引导

主要实验结果见表 2。我们测试了不同规模的 Qwen3 模型（Yang 等，2025），包括 4B、8B 和 14B，以及 4B 和 12B 规模的 Gemma3 模型（Gemma 团队，2025）。对于 PASTA，为确保比较稳健，我们展示了其三种配置中的最佳表现（完整详情见附录 H）。此外，附录 I 提供了原始模型和 *SEKA* 的具体示例及相应输出。

结果表明，*SEKA* 和 *AdaSEKA*在引导大语言模型注意力方面非常有效，在各种任务和模型规模下，通常优于基线模型（大多数情况下排名前二）和现有方法。如第 6 节所示，与 PASTA 和 SPA 相比，这些改进所需的额外开销显著更低。

一项主要发现是，注意力层干预在需要事实回忆的任务中效果显著。在 CounterFact 上，*SEKA* 和 PASTA 均取得了近乎完美的分数（例如，在 Qwen3-4B 上分别为 99.02 和 97.16），这验证了通过引导注意力来应对知识冲突这一通用方法的有效性，而基于 logit 的 SPA 则明显落后。在这类有效方法中，我们的方法始终具有性能优势。在 Bias in Bios 任务中也延续了这一趋势：在所有模型上，*SEKA* 和 *AdaSEKA*通常包揽前两名。

在指令遵循的代词变换任务中，模型的表现与基础模型对简单强调标记的预训练敏感度高度相关。对于 Qwen3 系列而言，它对简单的 Markdown 强调标记有一定响应，因此"**标记"基线表现尤为出色。这与早期的结论形成对比：大语言模型本质上受限于处理纯文本，无法处理

风格线索或强调标记（Brown 等，2020；Wei 等，2022）。然而，*AdaSEKA*仍能进一步提升表现，达到最先进水平（例如，在 Qwen3-8B 上的 A. P. 分数为 99.52）。我们的方法在对 Markdown 强调响应较弱的 Gemma3-4B 上优势最为显著。这表明我们的方法具有重要价值，尤其适用于对基础强调语法不够敏感的小型模型。

最后，我们的消融研究验证了该方法的核心组件。使用带有头过滤的随机投影（不学习）确有帮助，但显然并非最优，这凸显了我们谱学习方法的价值。移除学习得到的投影和头过滤机制（不学习且不筛选）会导致性能灾难性下降。例如，在 Qwen3-4B 代词任务中，A. P. 分数从原始方法的 90.52 降至 36.95。这有力地证明，学习有意义的相关性子空间，并将其有选择地应用于恰当的 KV头，对于取得成功至关重要。

5.2 迷失在中间

在第4.2节所述的设置下，我们在对比 *SEKA* 与基线方法，并探究不同 $\delta_{\min}$ 对选择 KV头的影响时，重点介绍两个关键发现。

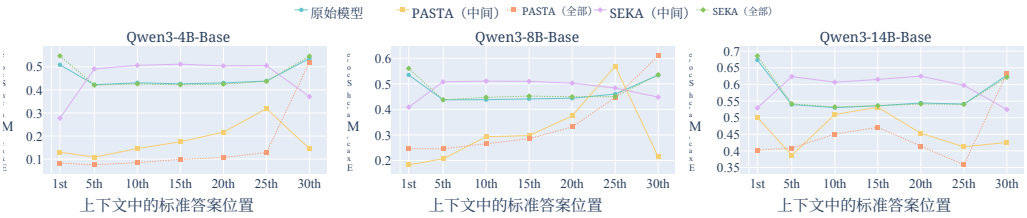


图4：Qwen3 不同规模的三个模型在中间遗失任务上的精确匹配分数，对比原始模型、将 PASTA/*SEKA* 应用于中间区域（5^{th} 至 25^{th} 个段落）以及将 PASTA/*SEKA*应用于所有段落的结果。

***SEKA* 能够逆转U形性能**。图4所总结的结果揭示了两项主要发现。首先，将*SEKA* 有选择地应用于中间段落（位置5至25，这只是一个非常粗略的范围），能够有效逆转典型U形性能曲线：中央位置的精确匹配分数大幅提高，从而消除长上下文中答案位于中间时常见的性能低谷。其次，将*SEKA* 均匀应用于所有段落，可能会略微加剧“中间遗忘”问题。最明显的提升通常出现在开头或结尾位置，而中间位置的改善则不太显著，甚至可能有所下降。相比之下，PASTA 在此任务中的效果较差。应用它无论是针对中间区域还是整个上下文，其性能在所有模型规模下通常都低于原始基线。

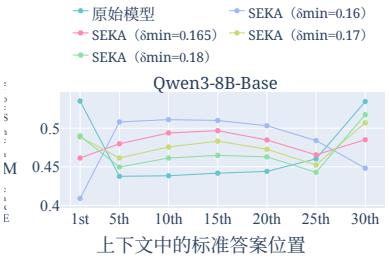


图5：在中间区域应用*SEKA*并使用不同阈值 $\delta_{\min}$时的精确匹配分数。

***SEKA*在 KV 头数量适当时，可缓解并拉平 U 型曲线**。在此对照实验中，我们将正向和负向引导增益系数（g^{+}和 g^{-}）分别固定为 0。2 和 0。1，仅调整阈值 $\delta_{\min}$ 以控制引导 KV 头的数量。实际应用中，降低 $\delta_{\min}$ 会增加引导头的数量：例如，对于 Qwen3-8B-Base，阈值分别为 0。16、0。165、0。17 和 0。18 时，*SEKA* 分别作用于 58、48、41 和 31 个 KV 头。如图5所示，当阈值 $\delta_{\min}$适当（约为 0。165 和 0。17），并对中间区域进行引导时，*SEKA* 可以拉平 U 形性能曲线，同时不会显著影响起始位置和结束位置的准确率。请注意，最佳阈值可能因模型规模而异。4B模型和14B模型的完整结果见附录K。

6 开销分析

我们预计算方法的一项关键优势是能够兼容 FlashAttention 等优化机制（Dao 等人，2022；Dao，2024年；Shah 等人，2024年）。我们使用单个英伟达GH200-120GB GPU 上的 Qwen3-8B-Base 模型，对第 5.2 节中的 100 个样本（平均 4362 个词元）测量推理开销，以此量化这一优势。

如表3所示，SEKA 的额外开销可以忽略不计（每个样本 +0.03 秒）。值得注意的是，为了与 PASTA 进行公平比较，我们采用了激进的配置，对 288 个可用 KV头 中的 175 个进行引导。相比之下，后处理方法会带来显著的成本。PASTA 需要编辑整个注意力矩阵，因此无法兼容 FlashAttention，导致延迟大幅增加（+1.03 秒），内存使用量也显著上升（+23.12 GB）。SPA 虽然

表 3：Qwen3-8B-Base 上的推理开销。时间按每个样本计算；内存为平均峰值使用量。

方法	平均时间 (s)	峰值内存 （GB，B=10）	峰值内存 （GB，B=1）
原始方法	0.55	27.63	16.72
PASTA	<风格 id='1'>1.58</风格><风格 id='3'>(</风格>+<风格 id='6'>.03) </风格><风格 id='8'>50.75</风格><风格 id='10'>(</风格>+<风格 id='13'>.12) </风格><风格 id='15'>-</风格>		
SPA	5.87 (+5.32)	-	17.71 (+0.99)
SEKA	0.58 (+0.03)	27.66 (+0.03)	16.75 (+0.03)
AdaSEKA	0.82 (+0.27)	43.22 (+15.59)	<风格 id='1'>18.23 </风格><风格 id='3'>(</风格>+<风格 id='6'>.51)</风格>

对单个样本而言内存效率较高，但不支持批处理，因此整体速度最慢。我们的自适应变体 *AdaSEKA* 凭借动态的查询感知能力带来了适度的额外开销（+0.27 秒）。不过，它仍显著优于 PASTA 和 SPA，因而是在长上下文场景中进行引导的实用性高得多的选择。

7 相关工作

大语言模型的引导研究主要分为两种范式。激活引导（Dathathri 等人, 2020年；Subramani 等人, 2022；Hernandez 等人, 2024年）通过干预多层感知机层来引导高层语义输出；而注意力引导是本文的重点，它会将模型的注意力引向输入提示中的特定词元。

激活引导。 这类研究也称为表征工程，它将“引导向量”添加到多层感知机层的激活值中，以控制语义属性（Zou 等人, 2023年）。其应用包括提升诚实性和安全性（Ravfogel 等人, 2020年；Burns 等人, 2023年；Iskander 等人, 2023年；Li 等人, 2023a；Wei 等人, 2023年；Bhattacharjee 等人, 2024年；Qiu 等人, 2024）、控制风格（Turner 等人, 2023年；2024年）、改进推理（Tang 等人, 2025年）以及知识编辑（Fang 等人, 2025年）。近期研究表明，只有当模型已掌握目标知识时，这些方法才有效（Simhi 等人, 2025年）。因此，这些方法与我们的方法不同：它们通过模型的隐藏状态改变模型所知的内容，而我们则通过注意力机制控制模型关注的位置。

注意力引导。为了解决大语言模型在长上下文中无法关注关键信息这一问题（Liu等人，2024；Meng 等人，2022），提示高亮方法会事后干预注意力分数（Zhang 等人，2024）或最终逻辑值（Tian & Zhang，2025）。然而，这些干预往往会显著增加延迟；例如，编辑完整的注意力矩阵与 FlashAttention 等现代优化技术不兼容（Dao 等人，2022；Dao，2024；Shah 等人，2024）。这一效率瓶颈凸显了对预计算替代方案的需求，使其能够在不牺牲与优化架构兼容性的前提下引导注意力。

8 结论

本文介绍了 *SEKA* 及其自适应变体 *AdaSEKA*。它们是一类新型免训练注意力引导方法，通过在计算注意力之前修改键嵌入来发挥作用。这种注意力前处理方法克服了以往工作的核心效率瓶颈，并确保与优化实现完全兼容。实验结果证实，这两种方法在一系列标准基准上均取得了最先进的结果，其中 *AdaSEKA* 的查询自适应路由表现尤为出色。这些提升仅带来微不足道的额外开销，使我们的工作成为构建更可控、更高效、适用于长上下文应用的大语言模型的切实一步。

可复现性声明

为确保研究成果可复现，我们已在 https://github.com/waylonli/SEKA 上公开所有必要材料。该代码库包括：(1) 所提方法 *SEKA* 和 *AdaSEKA*的完整源代码；(2) 运行所有实验的详细说明；(3) 评估中使用的预计算投影矩阵；以及 (4) 预处理后的数据集。

我们评估中使用的原始数据集均可公开获取，且已在第 4 节中引用。具体而言，BIASBIOS、COUNTERFACT 和“迷失在中间”数据集均依据 MIT 许可证发布。有关评估样本和指标计算的详情见附录E，超参数则列于附录 F。

致谢

我们感谢审稿人和领域主席提供的宝贵意见。我们也感谢 Yifu Qiu 就本项目进行的建设性讨论。作者感谢使用 Isambard-AI 国家人工智能研究资源（AIRR）提供的资源。Isambard-AI 由布里斯托大学运营，并由英国政府通过英国研究与创新署的科学、创新与技术部（DSIT）以及科学技术设施委员会提供资金支持 [ST/AIRR/I-A-I/1023]（McIntosh-Smith 等人，2024年）。

参考文献

Federico Barbero、Andrea Banino、Steven Kapturowski、Dharshan Kumaran、Jo˜ao Guil-herme Madeira Ara´ujo、Oleksandr Vitvitskyi、Razvan Pascanu 和 Petar Velickovic。《Transformer 需要眼镜！语言任务中的信息过度挤压》。载于神经信息处理系统进展第 *38* 卷：神经信息处理系统 *2024* 年度会议，*NeurIPS 2024*，温哥华，不列颠哥伦比亚省，加拿大，*2024*年*12*月*10*日至*15*日，2024年。网址：http://papers.nips.cc/paper_files/paper/2024/hash/b1d35561c4a4a0e0b6012b2af531e149-Abstract-Conference.html。

Amrita Bhattacharjee、Shaona Ghosh、Traian Rebedea 和 Christopher Parisien。《迈向大语言模型推理时的按类别安全引导》。载于*NeurIPS* 安全生成式人工智能研讨会 *2024*，2024年。网址：https://openreview.net/forum?id=EkQRNLPFcn。

Tom B. Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell、Sandhini Agarwal、Ariel Herbert-Voss、Gretchen Krueger、Tom Henighan、Rewon Child、Aditya Ramesh、Daniel M. Ziegler、Jeffrey Wu、Clemens Winter、Christopher Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam McCandlish、Alec Radford、Ilya Sutskever 和 Dario Amodei。语言模型是少样本学习者。载于第*34*届神经信息处理系统国际会议论文集，NeurIPS 2020，美国纽约州雷德胡克，2020年。Curran Associates Inc. 国际标准书号 9781713829546。

Collin Burns、Haotian Ye、Dan Klein 和 Jacob Steinhardt。无需监督地发现语言模型中的潜在知识。载于第十一届国际学习表征会议，2023年。网址：https://openreview.net/forum?id=ETKGuby0hcs。

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* 研讨会论文集《*Black-boxNLP*：分析与解读 *NLP* 神经网络》，第 276–286 页，意大利佛罗伦萨，2019 年。计算语言学协会。数字对象标识符：10.18653/v1/W19-4828。网址：https: //aclanthology.org/W19-4828。

Tri Dao。FlashAttention-2：通过更好的并行性和工作划分实现更快的注意力。在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 R´e。FlashAtten- tion：具备 I/O 感知能力的快速、高内存效率精确注意力。发表于《神经信息处理系统进展》第*35*卷：神经信息处理系统年度会议*2022*，*NeurIPS 2022*，美国路易斯安那州新奥尔良，*2022*年*11*月*28*日至*12*月*9*日，2022。网址：http://papers.nips.cc/paper_files/paper/2022/hash/67d57c32e20fd0a7a302cb81d36e40d5-Abstract-Conference.html。

Sumanth Dathathri、Andrea Madotto、Janice Lan、Jane Hung、Eric Frank、Piero Molino、Jason Yosinski 和 Rosanne Liu。《即插即用语言模型：一种简单的受控文本生成方法》。载于国际学习表征会议，2020年。网址：https://openreview.net/forum?id=H1edEyBKDS。

Maria De-Arteaga、Alexey Romanov、Hanna Wallach、Jennifer Chayes、Christian Borgs、Alexandra Chouldechova、Sahin Geyik、Krishnaram Kenthapadi 和 Adam Tauman Kalai。《Bias in Bios：高风险场景中语义表征偏见的案例研究》。载于公平、问责与透明度会议论文集，FAT*'19，第120–128页，纽约，NY，美国，2019年。计算机协会。国际标准书号 9781450361255。数字对象标识符：10.1145/ 3287560.3287572。网址：https://doi.org/10.1145/3287560.3287572。

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年。https://transformer-circuits.pub/2021/framework/index.html。

Junfeng Fang、Houcheng Jiang、Kun Wang、Yunshan Ma、Jie Shi、Xiang Wang、Xiangnan He 和 Tat-Seng Chua。AlphaEdit：面向语言模型的零空间约束模型编辑。发表于第十三届国际学习表征会议，2025年。网址：https: //openreview.net/forum?id=HvSytvg3Jh。

Gemma团队。Gemma 3 技术报告，2025年。网址：https://arxiv.org/abs/2503. 19786。

Evan Hernandez、Belinda Z. Li 和 Jacob Andreas。检查和编辑语言模型中的知识表征。发表于第一届语言建模会议，2024年。网址：https://openreview.net/forum?id=ADtL6fgNRv。

Cheng-Yu Hsieh、Yung-Sung Chuang、Chun-Liang Li、Zifeng Wang、Long Le、Abhishek Kumar、James Glass、Alexander Ratner、Chen-Yu Lee、Ranjay Krishna 和 Tomas Pfister。发现于中间位置：校准位置注意力偏差可提升长上下文利用。发表于计算语言学协会：*ACL 2024*论文集，第14982–14995页，泰国曼谷，2024年。计算语言学协会。数字对象标识符：10.18653/v1/2024.findings-acl.890。网址：https://aclanthology.org/2024.findings-acl.890/。

Shadi Iskander、Kira Radinsky和Yonatan Belinkov。屏蔽表示：通过迭代梯度投影保护敏感属性。载于计算语言学协会论文集：*ACL 2023*，第5961–5977页，加拿大多伦多，2023年。计算语言学协会。数字对象标识符：10.18653/v1/2023.findings-acl.369。网址：https: //aclanthology.org/2023.findings-acl.369。

Kenneth Li、Oam Patel、Fernanda B. Vi´egas、Hanspeter Pfister和Martin Wattenberg。推理时干预：从语言模型中引出真实答案。载于神经信息处理系统进展第*36*卷：*2023*年神经信息处理系统年会，*NeurIPS 2023*，美国路易斯安那州新奥尔良，*2023*年*12*月*10*日至*16*日，2023a。网址：http://papers.nips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html。

Weixian Waylon Li、Yftah Ziser、Maximin Coavoux 和 Shay B. Cohen。BERT 并非计数：学习匹配数学陈述与证明。载于 Andreas Vlachos 和 Isabelle Augenstein（编），第十七届欧洲计算语言学协会会议论文集，第 3581–3593 页，克罗地亚杜布罗夫尼克，2023年5月。计算语言学协会，

计算语言学协会。数字对象标识符：10.18653/v1/2023.eacl-main.260。网址：https://aclanthology.org/2023.eacl-main.260/。

Nelson F. Liu、Kevin Lin、John Hewitt、Ashwin Paranjape、Michele Bevilacqua、Fabio Petroni 和 Percy Liang。《迷失在中间：语言模型如何使用长上下文》。计算语言学协会会刊，12：157–173，2024年。数字对象标识符：10.1162/tacl a 00638。网址：https://aclanthology.org/2024.tacl-1.9。

Simon McIntosh-Smith、Sadaf R Alam 和 Christopher Woods。《Isambard-ai：一台专为人工智能优化的领导级超级计算机》，2024年。网址：https://arxiv. org/abs/2410.11199。

Kevin Meng、David Bau、Alex Andonian 和 Yonatan Belinkov。《在 GPT 中定位和编辑事实关联》。载于神经信息处理系统进展第*35*卷：神经信息处理系统进展年度会议 *2022*，*NeurIPS 2022*，美国路易斯安那州新奥尔良，*2022*年*11*月*28*日至*12*月*9*日，2022年。网址：http://papers.nips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html。

Paul Michel、Omer Levy 和 Graham Neubig。十六个注意力头真的比一个好吗？载于《神经信息处理系统进展》第*32*卷：神经信息处理系统年度会议*2019*（*NeurIPS 2019*），*2019*年*12*月*8*日至*14*日，加拿大不列颠哥伦比亚省温哥华，第14014–14024页，2019年。网址：https://proceedings.neurips.cc/paper/2019/hash/2c601ad9d2ff9bc8b282670cdd54f69f-Abstract.html。

Clement Neo、Shay B Cohen 和 Fazl Barez。《解读 Transformer 中的上下文查找：探究注意力机制与多层感知机的交互》。载于 Yaser Al-Onaizan、Mohit Bansal 和 Yun-Nung Chen（编），*2024*年自然语言处理经验方法会议论文集，第16681–16697页，美国佛罗里达州迈阿密，2024年11月。计算语言学协会。数字对象标识符：10.18653/v1/2024.emnlp-main.930。网址：https://aclanthology.org/ 2024.emnlp-main.930/。

Catherine Olsson、Nelson Elhage、Neel Nanda、Nicholas Joseph、Nova DasSarma、Tom Henighan、Ben Mann、Amanda Askell、Yuntao Bai、Anna Chen、Tom Conerly、Dawn Drain、Deep Ganguli、Zac Hatfield-Dodds、Danny Hernandez、Scott Johnston、Andy Jones、Jackson Kernion、Liane Lovitt、Kamal Ndousse、Dario Amodei、Tom Brown、Jack Clark、Jared Kaplan、Sam McCandlish 和 Chris Olah。上下文学习与归纳头，2022。网址：https://arxiv.org/abs/2209.11895。

Yifu Qiu、Zheng Zhao、Yftah Ziser、Anna Korhonen、Edoardo Maria Ponti 和 Shay B. Cohen。面向大语言模型对齐的激活谱编辑。收录于 神经信息处理系统进展 *38*：神经信息处理系统年度会议 *2024*，*NeurIPS 2024*，温哥华，不列颠哥伦比亚省，加拿大，*2024*年*12*月*10*日至*15*日，2024年。网址：http://papers.nips.cc/paper_files/paper/2024/hash/684c59d614fe6ae74a3be8c3ef07e061-Abstract-Conference.html。

Yifu Qiu、Varun R. Embar、Yizhe Zhang、Navdeep Jaitly、Shay B Cohen 和 Benjamin Han。为长上下文大语言模型引出上下文内检索与推理。载于 Wanxiang Che、Joyce Nabende、Ekaterina Shutova 和 Mohammad Taher Pilehvar（编），计算语言学协会：*ACL 2025*年论文集，第 3176–3192 页，奥地利维也纳，2025年7月。计算语言学协会。国际标准书号 979-8-89176-256-5。数字对象标识符：10.18653/v1/2025.findings-acl.165。网址：https://aclanthology.org/2025.findings-acl.165/。

Shauli Ravfogel、Yanai Elazar、Hila Gonen、Michael Twiton 和 Yoav Goldberg。将其归零：通过迭代零空间投影保护受保护属性。载于计算语言学协会第 *58* 届年会论文集，第 7237–7256 页，线上，2020年。计算语言学协会。数字对象标识符：10.18653/v1/2020.acl-main.647。网址：https://aclanthology.org/2020.acl-main.647。

Jay Shah、Ganesh Bikshandi、Ying Zhang、Vijay Thakkar、Pradeep Ramani 和 Tri Dao。FlashAttention-3：利用异步性和低精度实现快速、准确的注意力。发表于《神经信息处理系统进展》第 *38* 届年会：神经信息处理系统年会 *2024*（*NeurIPS 2024*），加拿大不列颠哥伦比亚省温哥华，*12* 月 *10* 日至

2024 年 *15* 日，2024 年。网址：http://papers.nips.cc/paper_files/paper/2024/hash/7ede97c3e082c6df10a8d6103a2eebd2-Abstract-Conference.html。

Adi Simhi、Jonathan Herzig、Itay Itzhak、Dana Arad、Zorik Gekhman、Roi Reichart、Fazl Barez、Gabriel Stanovsky、Idan Szpektor 和 Yonatan Belinkov。HACK：沿确定性与知识轴的幻觉，2025年。URL https://arxiv.org/abs/2510.24222。

Alessandro Stolfo、Vidhisha Balachandran、Safoora Yousefi、Eric Horvitz 和 Besmira Nushi。通过激活引导改进语言模型中的指令遵循。收录于第十三届国际学习表征会议，2025年。URL https://openreview. net/forum?id=wozhdnRCtw。

Nishant Subramani、Nivedita Suresh 和 Matthew Peters。从预训练语言模型中提取潜在引导向量。收录于计算语言学协会研究成果：*ACL 2022*，第566–581页，爱尔兰都柏林，2022。计算语言学协会。数字对象标识符：10.18653/v1/2022.findings-acl.48。URL https://aclanthology.org/2022. findings-acl.48。

Xinyu Tang、Xiaolei Wang、Zhihao Lv、Yingqian Min、Xin Zhao、Binbin Hu、Ziqi Liu和Zhiqiang Zhang。通过表征工程解锁大语言模型的通用长思维链推理能力。载于第*63*届计算语言学协会年会论文集（第*1*卷：长论文），第6832–6849页，奥地利维也纳，2025年。计算语言学协会。国际标准书号979-8-89176-251-0。数字对象标识符：10.18653/v1/2025.acl-long.339。网址https://aclanthology.org/2025.acl-long.339/。

Yuan Tian和Tianyi Zhang。面向代码生成的选择性提示锚定。载于第四十二届国际机器学习会议，2025年。网址https://openreview.net/ forum?id=aEnkBIhYvO。

Alexander Matt Turner、Lisa Thiergart、David Udell、Gavin Leech、Ulisse Mini和Monte MacDiarmid。激活加法：无需优化的语言模型引导。*ArXiv*预印本，abs/2308.10248，2023年。网址 https://arxiv.org/abs/2308.10248。

Alexander Matt Turner、LisaThiergart、Gavin Leech、David Udell、JuanJ.Vazquez、Ulisse Mini 和 Monte MacDiarmid。通过激活工程引导语言模型，2024年。网址：https://arxiv.org/abs/2308.10248。

Elena Voita、David Talbot、Fedor Moiseev、Rico Sennrich 和 Ivan Titov。分析多头自注意力：专用注意力头承担了主要工作，其余部分可以剪除。载于计算语言学协会第*57*届年会论文集，第5797–5808页，意大利佛罗伦萨，2019年。计算语言学协会。数字对象标识符：10.18653/v1/P19-1580。网址：https://aclanthology.org/P19-1580。

Tianlong Wang、Xianfeng Jiao、Yinghao Zhu、Zhongzhi Chen、Yifan He、Xu Chu、Junyi Gao、Yasha Wang 和 Liantao Ma。自适应激活引导：一种免调优的大语言模型真实性提升方法，适用于多种幻觉类别。载于美国计算机协会 *2025*年世界互联网大会论文集，WWW '25，第2562–2578页，纽约州纽约，美国，2025年。计算机协会。国际标准书号 9798400712746。数字对象标识符：10.1145/3696410.3714640。网址：https://doi.org/10.1145/3696410.3714640。

Alexander Wei、Nika Haghtalab 和 Jacob Steinhardt。《越狱：大语言模型安全训练为何会失效? 》载于《神经信息处理系统进展》第*36*卷：神经信息处理系统年度会议 *2023*，*NeurIPS 2023*，美国路易斯安那州新奥尔良，*2023*年*12*月*10*日至*16*日，2023年。网址http://papers.nips.cc/paper_files/paper/2023/hash/fd6613131889a4b656206c50a8bd7790-Abstract-Conference.html。

Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Brian Ichter、Fei Xia、Ed H. Chi、Quoc V. Le 和 Denny Zhou。思维链提示能够激发大型语言模型的推理能力。收录于神经信息处理系统第 *36* 届会议论文集，NeurIPS 2022，美国纽约州雷德胡克，2022年。Curran Associates Inc. 国际标准书号 9781713871088。

Wenhao Wu、Yizhong Wang、Guangxuan Xiao、Hao Peng 和 Yao Fu。检索头从机制层面解释了长上下文事实性。收录于第十三届国际学习表征会议，2025年。网址：https://openreview.net/forum?id=EytBpUGB1Z。

An Yang、Anfeng Li、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chang Gao、Chengen Huang、Chenxu Lv、Chujie Zheng、Dayiheng Liu、Fan Zhou、Fei Huang、Feng Hu、Hao Ge、Haoran Wei、Huan Lin、Jialong Tang、Jian Yang、Jianhong Tu、Jianwei Zhang、Jianxin Yang、Jiaxi Yang、Jing Zhou、Jingren Zhou、Junyang Lin、Kai Dang、Keqin Bao、Kexin Yang、Le Yu、Lianghao Deng、Mei Li、Mingfeng Xue、Mingze Li、Pei Zhang、Peng Wang、Qin Zhu、Rui Men、Ruize Gao、Shixuan Liu、Shuang Luo、Tianhao Li、Tianyi Tang、Wenbiao Yin、Xingzhang Ren、Xinyu Wang、Xinyu Zhang、Xuancheng Ren、Yang Fan、Yang Su、Yichang Zhang、Yinger Zhang、Yu Wan、Yuqiong Liu、Zekun Wang、Zeyu Cui、Zhenru Zhang、Zhipeng Zhou 和 Zihan Qiu。《Qwen3 技术报告》，2025年。网址：https://arxiv.org/abs/2505.09388。

Zhilin Yang、Peng Qi、Saizheng Zhang、Yoshua Bengio、William Cohen、Ruslan Salakhutdinov 和 Christopher D. Manning。《HotpotQA：用于多样化、可解释的多跳问题回答的数据集》。载于*2018*年自然语言处理实证方法会议论文集，第 2369–2380 页，布鲁塞尔，比利时，2018年。计算语言学协会。数字对象标识符：10.18653/v1/D18-1259。网址：https://aclanthology.org/D18-1259。

Qingru Zhang、Chandan Singh、Liyuan Liu、Xiaodong Liu、Bin Yu、Jianfeng Gao 和 Tuo Zhao。《告诉你的模型该关注哪里：面向大语言模型的事后注意力引导》。载于第十二届国际学习表征会议，2024年。网址：https://openreview.net/ forum?id=xZDWO0ejD。

Andy Zou、Long Phan、Sarah Chen、James Campbell、Phillip Guo、Richard Ren、Alexander Pan、Xuwang Yin、Mantas Mazeika、Ann-Kathrin Dombrowski、Shashwat Goel、Nathaniel Li、Michael J. Byun、Zifan Wang、Alex Mallen、Steven Basart、Sanmi Koyejo、Dawn Song、Matt Fredrikson、J. Zico Kolter 和 Dan Hendrycks。《表征工程：一种自上而下的人工智能透明度方法》，2023年。网址：https://arxiv.org/abs/2310.01405。

用于词元级相关性监督的合成数据集

为了监督注意力引导，我们构建了一个合成数据集，从而能够精确控制词元级相关性。每个样本由两个上下文（C_1, C_2）组成，每个上下文都配有一个问题和一个答案元组（Q_1, A_1 和 Q_2, A_2）。这种结构使我们能够通过比较不同查询上下文中相同的词元跨度来定义相关性。

表4：为两个答案跨度构造的提示三元组。每组都基于问题-上下文对齐，提供中性、正向和负向变体。

组	提示
中性	上下文：投资组合经理将资本分配给股券。
正向	问题：投资组合经理在股票和债券之间分配什么? 上下文：投资组合经理分配资本用于股券。
负向	问题：气候模型模拟什么? 上下文：投资组合经理分配资本用于股券。
中性	上下文：气候模型模拟不同情景下的
正向	问题：气候模型模拟什么? 上下文：气候模型模拟不同情景下的
负向	问题：投资组合经理在股票和债券之间分配什么? 上下文：气候模型模拟不同情景下的

对于每一对（C, Q, A）三元组，如表4所示，我们可以推导出两个监督样本：一个针对 C_1中的答案片段“资本”，另一个针对 C_2中的答案片段“海平面上升”。

表 5：合成数据实例。

上下文 1 (C_1)	投资组合经理将资本配置于股票和债券。
上下文 2 (C_2)	气候模型模拟不同情景下的海平面上升。
问题 1 (Q_1)	投资组合经理在股票和债券之间分配什么?
答案 1 (A_1)	资本
问题 2 (Q_2)	气候模型模拟什么?
答案 2 (A_2)	海平面上升

对于每个答案，我们构建三种变体：（1）正向（相关）提示，其中问题与上下文一致（例如， Q_1 对应 C_1，以及 Q_2 对应 C_2）；（2）负向（不相关）提示，其中问题与上下文不匹配（例如， Q_1 对应 C_2，以及 Q_2 对应 C_1）；（3）仅包含上下文的中性提示。这使我们能够为上下文中的答案跨度收集三类键嵌入： $\boldsymbol{h}^+$ 用于正向， $\boldsymbol{h}^-$ 用于负向，以及 $\boldsymbol{h}$ 用于中性。在图1中，我们通过实验表明，对于某些键值头，不同词元跨度的键嵌入从负向变体到正向变体呈现出一致的偏移。这验证了这些相关性监督信号的构建与使用。

实际构建细节。合成数据集的制作成本很低。我们使用表 5 所示的固定模板，并按照图 6 中提供的提示，自动提示 GPT-4o 生成对比样本。此过程无需人工标注。收集生成的样本后，我们将其转换为 JSON 格式，以供后续使用。

多专家投影学习样本适用于AdaSEKA

构建合成数据集后，我们又准备了三个特定任务数据集，总计四个，用于多专家投影学习（第 3.3 节）。如表 6 所示，每个样本都由一对中性提示和正向提示组成。对于 CounterFact（Meng 等人，2022）数据集和 BiasBios（De-Arteaga等人，2019）数据集，这些提示对取自各自的训练集，并遵循表 1 所列的原始提示模板。对于每个样本，我们直接从上下文中提取答案跨度对应的键嵌入。针对 HotpotQA（Yang 等人，2018），我们采用了不同的处理流程，以适应该数据集的多跳特性。上下文由所有候选段落拼接而成，随后提取并拼接所有支持事实中的键嵌入。针对每项任务，我们使用固定随机种子 42，确保结果可复现，并从训练集中随机抽取 200 个实例来学习每个专家投影。

C *SEKA* 变换的几何直觉

为了从几何角度阐释*SEKA*键编辑的作用，考虑投影矩阵 $\boldsymbol{P}$由 $\boldsymbol{U}\boldsymbol{U}^\top$给出的情形，其中 $\boldsymbol{U} \in \mathbb{R}^{d_k \times r}$具有张成相关性子空间的正交归一列（即 $\boldsymbol{U} = \boldsymbol{U}^+$或 $\boldsymbol{U}^-$，如先前所定义）。为简便起见，假设 $g = 1$，并仅关注正向（或负向）投影。于是该变换变为:

$$\boldsymbol{k}'_j = (\boldsymbol{I} + \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{k}_j. \tag{7}$$

任意向量 $\boldsymbol{x} \in \mathbb{R}^{d_k}$ 均可分解为

$$\boldsymbol{x} = \boldsymbol{x}_\parallel + \boldsymbol{x}_\perp, \quad \text{where} \quad \boldsymbol{x}_\parallel = \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{x}, \;\; \boldsymbol{x}_\perp = \boldsymbol{x} - \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{x}. \tag{8}$$

该分解是正交的。具体而言，

$$\boldsymbol{x}_\parallel^\top \boldsymbol{x}_\perp \;=\; (\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{x})^\top(\boldsymbol{x} - \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{x}) \;=\; \boldsymbol{x}^\top\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{x} - \boldsymbol{x}^\top\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{x} \tag{9}$$

$$= \; \boldsymbol{x}^\top\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{x} - \boldsymbol{x}^\top\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{x} \;=\; 0, \tag{10}$$

利用投影 ($(\boldsymbol{U}\boldsymbol{U}^\top)^2 = \boldsymbol{U}\boldsymbol{U}^\top$) 的幂等性。

合成样本生成提示

我们正在协作生成总计 100 个合成示例。

你将以批次为单位生成示例，每次回复恰好生成 20 个。在本次对话的所有批次中，每个示例都必须全局唯一。

在生成新批次之前，你必须：
1. 查看本次对话中的所有先前示例。2. 确保实体、上下文、事件、句子结构、问题或答案均不重复。3. 确保不存在近似重复、改写后的重复或更换主题后的重复。

每个示例都必须遵循以下结构（格式必须完全一致）：
示例 k：上下文 1： <C1>上下文 2： <C2>问题 1： <Q1>答案 1： <A1>问题 2： <Q2>答案 2： <A2>

生成要求：
- 每个上下文必须是一个简洁的虚构句子。- C1 和 C2 必须在语义上无关。- Q1 必须询问 C1 中逐字一致出现的片段。- Q2 必须询问 C2 中逐字一致出现的片段。- A1 和 A2 必须分别是 C1 和 C2 的精确子串（连续片段）。- 任何批次中的上下文、实体、主题、背景或问题类型均不得重复。- 避免在措辞、结构或领域上与先前示例有任何相似之处。- 不要添加额外说明或评论。

你的任务：阅读截至目前对话中的所有先前示例，然后生成接下来的 20 个全新且全局唯一的示例。恰好生成 20 个示例后停止。

图 6：用于生成合成对比示例的提示模板。

应用该变换后，我们得到

$$(\boldsymbol{I}+\boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{x} \;=\; \boldsymbol{x}+\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{x} \;=\; \big(\boldsymbol{x}-\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{x}\big)+\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{x}+\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{x} \tag{11}$$

$$= \underbrace{\boldsymbol{x}-\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{x}}_{\boldsymbol{x}_\perp}+2\underbrace{\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{x}}_{\boldsymbol{x}_\parallel} \;=\; \boldsymbol{x}_\perp+2\boldsymbol{x}_\parallel. \tag{12}$$

这表明沿子空间方向的分量被放大（加倍），而正交分量保持不变。

虽然 $g=1$ 情形具有几何清晰性，但该结果可推广至任意 $g\in\mathbb{R}$：

$$(\boldsymbol{I}+g\boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{x}=\boldsymbol{x}_\perp+(1+g)\boldsymbol{x}_\parallel. \tag{13}$$

因此，相关性对齐分量按 $(1+g)$ 进行缩放，而所有正交方向均保持不变。该操作既不是投影，也不是正交变换，而是一种针对性线性修改，有选择地放大与相关性子空间对齐的方向。*SEKA*利用这一特性，以可控且易于解释的方式增强相关词元特征，从而实现精准的词元级注意力引导，同时不干扰无关分量。

虽然上述几何解释阐明了*SEKA* 如何放大与学习到的相关性子空间对齐的键向量分量，但有必要说明该子空间代表什么。*SEKA* 的本意并非编码或操纵语义。它的作用被有意限制在

表 6：用于多专家投影学习的构造提示对。每个数据集都根据问题-上下文对齐提供中性和正向变体。

数据集	变体	提示
CounterFact	**中性**	上下文：先前，丹妮尔·达里厄的母语是法语。当前， 丹妮尔·达里厄的母语是英语。
	正向	问题：丹妮尔·达里厄，母语者? 上下文：先前，丹妮尔·达里厄的母语是法语。当前， 丹妮尔·达里厄的母语是英语。
BiasBios	**中性**	上下文：Nora是一名助理[illegible]巴赫切谢希尔大学从事国际关系 位于伊斯坦布尔。她还是罗纳德·D·阿斯穆斯政策创业研究员，供职于 德国马歇尔基金会，并在国际中心担任访问研究员 牛津大学的研究中心（CIS）。这篇评论最初刊载于 Sada， 这是一份由卡内基国际和平基金会出版的在线期刊。
	正向	问题：Nora 的职业是什么? 上下文：Nora 是一名助理[illegible]巴赫切希尔大学教授国际关系 伊斯坦布尔。她还是一名罗纳德·D·阿斯穆斯政策创业研究员，任职于 德国马歇尔基金会，并在牛津大学国际研究中心担任访问研究员 （CIS）。这篇评论最初发表于 Sada， 这是卡内基国际和平基金会出版的在线期刊。
HotpotQA	**中性**	上下文：Radio City（印度广播电台）Radio City 是印度首家私营调频广播电台 广播电台，并于2001年7月3日开播。…《Arthur's Magazine》《Arthu[illegible] a[illegible] 于[illegible]Arthur担任编辑，刊载了Edgar A. Poe、J.H. Ingraham、Sarah Josepha Hale、Thomas G. Spear等人的作品。1846年5月，该刊被 并入《Godey's Lady's Book》。…First for Women Firs[illegible] 女[illegible]性杂[illegible] 创立于1989年。总部位于新泽西州恩格尔伍德克利夫斯。。.....该公司 最初以牛仔服装系列起家，后来发展为男女服装系列。
	正向	问题：哪本杂志创刊时间更早，《Arthur's Magazine》还是 First for Women? 上下文：Radio City（印度广播电台）Radio City 是印度首家私人调频 广播电台，创办于2001年7月3日。。.....《Arthur's Magazine》《Arthu[illegible] [illegible] ，[illegible]该刊由T.S. Arthur编辑，刊登了Edgar A. Poe、J.H. Ingraham、Sarah Josepha Hale、Thomas G. Spear等人的作品。1846年5月，该刊 并入《Godey's Lady's Book》。…First for Women Firs[illegible] 由鲍尔[illegible] 创刊于1989年，总部位于新泽西州恩格尔伍德克利夫斯。…该公司 最初以牛仔服装系列起步，后来发展为男女服装系列。

Transformer 的路由子空间注意力范围内，这与此前的机制分析一致（Elhage等人，2021年；Olsson等人，2022）。

现代 Transformer电路研究将注意力头的作用分解为

$$H^{(h)}(R) = A^{(h)}(R) \otimes W_O^{(h)} W_V^{(h)} R, \tag{14}$$

其中，$A^{(h)}$ 是决定哪些词元关注哪些词元的查询-键相似度张量，而$W_O^{(h)} W^{(h)V}$则将被关注的特征写入残差流（Elhage等人，2021年）。这一公式明确区分了路由（Q/K）与语义写入操作（V/MLP）。

此外，对归纳头和检索头的研究（Olsson 等人，2022）表明，Q/K 向量实现了词元匹配和算法路由行为，例如复制和续写；而语义信息则主要存储在值向量和多层感知机激活值中。这些发现与我们的设计相一致：*SEKA*旨在仅修改路由（相关性）子空间，同时保持语义子空间不变。

D *SEKA* 与 *AdaSEKA*算法

我们为所提出的方法 *SEKA* 和 *AdaSEKA*提供了详细的伪代码。算法 1 详细介绍了标准的 *SEKA* 方法。该方法包含一个离线阶段，利用 SVD 从对比数据中学习固定的正向和负向投影矩阵。在推理过程中，钩子会将这些静态投影应用于高亮词元的键嵌入。算法 2 描述了更加灵活的 *AdaSEKA*框架。实质上，标准 *SEKA* 可以被视为*AdaSEKA*的特例，即仅包含一个专家且不进行动态系数计算。*AdaSEKA* 通过离线加载多个专家奇异值分解组件组成的组件库，进一步实现了泛化。对于每个新提示，它随后执行动态的、查询感知的预计算：根据查询与每个专家的对齐程度计算路由系数，并即时构建定制的投影矩阵。之后，在生成过程中，通过键编辑钩子应用这一量身定制的投影。

算法 1谱编辑键放大（*SEKA*）

输入： 三元组 $\{\boldsymbol{h}, \boldsymbol{h}^+, \boldsymbol{h}^-\}_{\ell,h}$、方差阈值 γ、增益 g^+, g^-

确保： 投影 $\{\boldsymbol{P}_{\ell,h}^+, \boldsymbol{P}_{\ell,h}^-\}$ 和一个键编辑钩子

对于所有层 ℓ**和**头 h **执行**

  $\boldsymbol{\Omega}_{\ell,h}^+ \leftarrow \frac{1}{n}\boldsymbol{h}^\top \boldsymbol{h}^+$，$\boldsymbol{\Omega}_{\ell,h}^- \leftarrow \frac{1}{n}\boldsymbol{h}^\top \boldsymbol{h}^-$

  $(\boldsymbol{U}_{\ell,h}^+, \boldsymbol{S}_{\ell,h}^+, \boldsymbol{V}_{\ell,h}^+)$，$(\boldsymbol{U}_{\ell,h}^-, \boldsymbol{S}_{\ell,h}^-, \boldsymbol{V}_{\ell,h}^-) \leftarrow$ 奇异值分解$(\boldsymbol{\Omega}_{\ell,h}^-)$

  $k^+ \leftarrow$最小值$\{k: \sum_{i=1}^{k} \boldsymbol{S}_{\ell,h,i}^+ / \sum_i \boldsymbol{S}_{\ell,h,i}^+ \geq \gamma\}$，$k^- \leftarrow$最小值$\{k: \sum_{i=1}^{k} \boldsymbol{S}_{\ell,h,i}^- / \sum_i \boldsymbol{S}_{\ell,h,i}^- \geq \gamma\}$

  $\boldsymbol{P}_{\ell,h}^+ \leftarrow \boldsymbol{U}_{\ell,h,:k^+}^+ \boldsymbol{U}_{\ell,h,:k^+}^{+\top}$，$\boldsymbol{P}_{\ell,h}^- \leftarrow \boldsymbol{U}_{\ell,h,k^-:}^- \boldsymbol{U}_{\ell,h,k^-:}^{-\top}$

结束循环

钩子应用于每个选定的(ℓ, h)（按层注册 ℓ；ℓ 在钩子内固定）。

  输入： $K \in \mathbb{R}^{B\times T\times H\times d}$，掩码 m

  注意：为简洁起见，我们省略了 K 上明确的层索引；投影保持为 $\boldsymbol{P}_{\ell,h}^{\pm}$.

  对于 b=1..B, t=1..T, h=1..H **执行**

    如果 $m_{b,t}=1$ **那么**

      $\Delta \leftarrow \left(g^+ \boldsymbol{P}_{\ell,h}^+ + g^- \boldsymbol{P}_{\ell,h}^-\right) K[b,t,h,:]/2$

      $K[b,t,h,:] \leftarrow K[b,t,h,:] + \Delta$

  返回 K 到注意力计算

在生成之前为选定的（ℓ，h）注册钩子，并在生成后将其移除。

E 标准基准详情

我们在三项改编自 PASTA 框架（Zhang 等，2024年）的成熟基准上评估了我们的方法。我们对评估方案进行了重大改进，例如采用不区分大小写的评分，以确保评估更加稳健。由于现代模型在 JSON 格式任务上的表现已近乎完美，该任务无法有效区分模型能力，因此我们将其省略。

E.1 COUNTERFACT

COUNTERFACT 基准（Meng 等，2022）用于评估大语言模型优先采用新的上下文信息，而非其预训练知识的能力。在此基准中，每条事实均表示为主语–关系–宾语三元组（s，r，o），其中 s 表示主语实体，r 表示关系，o 表示宾语。

任务格式。 模型接收的输入结构为：“先前，{s r o_{old}}。当前，{s r o_{new}}。{问题}。”挑战在于，模型往往默认依赖预训练关联，而不是关注上下文中提供的新的、相互矛盾的信息。

算法 2 查询驱动的自适应 SEKA（AdaSEKA）要求：SVD分量$\{\boldsymbol{U}^{+}_{m,\ell,h}, \boldsymbol{S}^{+}_{m,\ell,h}\}$，针对 M专家，顶部分量 K，增益 g确保：使用动态计算投影的键编辑钩子 1：存储专家奇异值分解组件$\{\boldsymbol{U}^{+}_{m,\ell,h}, \boldsymbol{S}^{+}_{m,\ell,h}\}$，针对所有专家 m、层 ℓ和注意力头 h。2：对于给定提示及输入 ID $\boldsymbol{I}$：3：获取最后一个词元的查询向量$\boldsymbol{q}_{\ell}$，$_{h}$，针对每个选定的层 ℓ 和注意力头 h。4：对所有选定的层 ℓ和注意力头 h 执行 5：对所有专家 $m = 1..M$ 执行 6：计算系数α_{m}，$_{\ell}$，$_{h}$（$\boldsymbol{q}_{\ell}$，$_{h}$）$\propto \sum_{k=1}^{K}$（$\boldsymbol{q}_{\ell}^{\top}$，$_{h}\boldsymbol{u}_{m\,\ell\,h}^{+(k)}$，，）$\cdot \sigma_{m\,\ell}^{+(k)}$，，$_{h}$（依据公式 6）7：结束循环 8：构造$\boldsymbol{P}_{\text{dynamic}}$，$_{\ell}$，$_{h}$ $\leftarrow \sum_{m=1}^{M} \alpha_{m}$，$_{\ell}$，$_{h}$（$\boldsymbol{q}_{\ell}$，$_{h}$）$\boldsymbol{U}_{m}^{+}$，$_{\ell}$，$_{h}$，：，：$_{K}$（$\boldsymbol{U}_{m}^{+}$，$_{\ell}$，$_{h}$，：，：$_{K}$）$^{\top}$9：存储$\boldsymbol{P}_{\text{dynamic}}$，$_{\ell}$，$_{h}$ 以供钩子使用。10：结束循环 11：钩子应用于每个选定的（ℓ，h）（按层 ℓ注册；ℓ在钩子内固定）。12：输入：$K \in \mathbb{R}^{B\times T\times H\times d}$，掩码 m13：注：为简洁起见，我们省略 K上的显式层索引。14：对 b=1..B，t=1..T，h=1..H 执行 15：如果m_{b}，$_{t}$=1 则 16：$\Delta \leftarrow g \cdot \boldsymbol{P}_{\text{dynamic}}$，$_{\ell}$，$_{h}$ $K[b,t,h,:]$17：$K[b,t,h,:] \leftarrow K[b,t,h,:] + \Delta$18：将 K 返回至注意力计算 19：在生成之前为选定的（ℓ，h）注册钩子，并在之后将其移除。

提示： “先前，凯文·加内特是一名职业篮球运动员。当前，**凯文·加内特是一名职业棒球运动员**。凯文·加内特是一名职业”

目标： 模型应生成“棒球运动员”，而不是其预训练关联中的“篮球运动员”。

评估指标。 遵循（Zhang 等人，2024年）的方法，为了评估模型回忆新事实的能力，我们在生成点测量其内部偏好，而不是依赖对自由文本的解析。对于给定的提示，我们向模型提供完整的上下文和问题，然后评估其分配给潜在后续词元的对数概率。

- **效能分数（ES）：** 该指标直接衡量模型是否优先选择新的正确事实（$o_{\text{新}}$），而不是旧的错误事实（$o_{\text{旧}}$）。它表示模型为新事实的第一个词元分配的概率高于旧事实第一个词元的概率的次数百分比。较高的效能分数表明，模型已根据上下文成功更新了其信念。

$$\text{ES} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[P_{\text{LLM}}(o_{\text{new}}^{(i)}) > P_{\text{LLM}}(o_{\text{old}}^{(i)})]$$

- **释义分数（PS）：** 该指标通过计算一组由人工撰写的原始问题释义的平均效能分数，来衡量泛化能力。

E.2 BIASBIOS

BIASBIOS 数据集（De-Arteaga等人，2019年）由传记组成，最初旨在探究职业预测中的性别偏见。每篇传记的第一句话都明确说明人物的职业，而后续句子则提供可能造成干扰的职业经历细节。

任务格式。 每篇人物简介后都会附上提示“{人物} 的职业是”，模型必须从 28 个候选项中预测正确的职业。

提示： *"**Winnie* 是一位居住在纽约的美国摄影师。**她专攻时尚摄影和肖像摄影，擅长拍摄人类和动物.....*Winnie* 的职业是"*

目标： *"摄影师"*

评估指标。 我们在 28 个候选职业中测量标准的 top-1 **准确率**，并采用不区分大小写的匹配方式，以确保正确评估语义等价性。

E.3 代词变更

此任务通过语言变换评估指令遵循能力。模型会收到指示：*"将 'she' 和 'he' 替换为 'they'。"* 这要求模型在保留原始内容的同时遵循变换规则。

提示： *"Mary 是圣约翰大学课程与教学系的一名副教授，她拥有宾夕法尼亚大学阅读/写作/读写能力专业的博士学位.....**将 'she' 和 'he' 替换为 'they'，并在更改代词后生成 Mary 的职业**。"*

目标： *"Mary 是一名副教授.....他们拥有博士学位.....Mary 的职业是副教授。"*

增强评估指标。 正如 Zhang 等人（2024）[2],的公开同行评审中所指出的，原始方法的指标会奖励空字符串，因为空字符串能完美地"转换"零个代词，从而导致分数虚高。为了解决这个问题，我们提出了**代词加权词汇重叠分数（P. 分数）**，将指令遵循和内容保留统一为单一指标。

P. 分数根据代词转换的成功率，对与原文的词汇重叠给予相应权重。其定义如下：

$$\text{P. Score} = \frac{w_{\text{pron}} \times |T_{\text{ori}} \cap T_{\text{gen}}|}{|T_{\text{ori}}|}, \tag{15}$$

其中，w_{pron} 是成功转换的代词所占比例，T_{ori} 和 T_{gen} 分别是原文和生成文本中非代词内容词元的集合。这确保空生成文本的分数为零，并且只有在遵循指令时，内容保留才会获得分数。我们评估了两个变体：一个（P. 分数）针对核心主语代词（"she""he"），另一个（A. P. 分数）针对完整的性别代词集合（"she""he""her""him""hers""his""herself""himself"）。

F 技术设置

本附录部分详细说明了 *SEKA* 和 *AdaSEKA* 实验所用的超参数。对于 CounterFact 和 Bias in Bios 基准，我们遵循 PASTA 的实验设置（Zhang 等人，2024），在包含 500 个样本（索引 4500–4999）的验证集上进行网格搜索，以调整超参数。随后在测试集（索引 5000–10000）上进行最终评估。对于代词变换任务，我们在单独的小型开发集上调整超参数。所有模型的实验均采用贪心解码。

标准*SEKA* 方法需要调整四个超参数：用于构建投影的方差阈值（γ）、用于 KV 头选择的相关性-敏感度阈值（δ_{min}），以及正向/负向引导增益（g^+ 和 g^-）。*AdaSEKA* 框架简化了这一过程，只需调整 KV 头选择阈值（δ_{min}）和单个引导增益系数（g）。表 7 列出了每个模型和任务所选用的超参数。

[2]https://openreview.net/forum?id=xZDWO0oejD&noteId=3kDI7QRqSI

表 7：*SEKA* 和 *AdaSEKA* 方法的超参数。*SEKA*使用方差阈值（γ）、KV 头选择阈值（δ_{min}）、正增益（g^{+}）和负增益（g^{-}）。*AdaSEKA* 使用 KV 头选择阈值（δ_{min}）和引导增益（g）。

		SEKA				*AdaSEKA*	
模型	**Task**	γ	δ_{min}	g^{+}	g^{-}	δ_{min}	g
Qwen3-4B-Base	CounterFact	0.960	0.13	1.56	0.00	0.1	3.0
	Bias in Bios	0.998	0.12	1.00	0.80	0.1	0.5
	代词变更	0.880	0.22	0.16	0.00	0.5	0.6
Qwen3-8B-Base	CounterFact	0.850	0.12	2.40	0.00	0.1	3.0
	Bias in Bios	0.998	0.12	0.60	0.30	0.1	0.5
	代词变更	0.900	0.20	0.19	0.00	0.5	0.6
Qwen3-14B-Base	CounterFact	0.870	0.10	2.42	0.00	0.1	3.0
	Bias in Bios	0.990	0.15	0.60	0.30	0.3	1.0
	代词变更	0.880	0.23	0.16	0.00	0.6	0.6
Gemma-3-4B	CounterFact	0.990	0.60	2.00	0.00	0.2	3.0
	Bias in Bios	0.800	0.12	0.80	0.00	0.2	0.8
	代词变更	0.800	0.20	0.40	0.00	0.4	1.0
Gemma-3-12B	CounterFact	0.990	0.50	1.00	0.00	0.1	-5.0
	Bias in Bios	0.994	0.00	0.40	0.00	0.7	0.5
	代词变更	0.700	0.40	-0.50	0.00	0.5	-0.4

超参数敏感性。为探究*SEKA*对超参数的敏感性，我们通过独立改变每个参数，同时将其余参数保持在验证集上的最优配置（表 7），开展实验分析。我们从三个基准任务中随机选取 500 个测试样本，并针对以下范围进行逐项扫描：$\gamma \in \{0.75, 0.80, 0.85, 0.90, 0.95\}$、$\delta_{\text{min}} \in \{0.10, 0.20, 0.30, 0.40, 0.50, 0.60\}$、$g^{+} \in \{0.1, 0.2, 0.4, 0.6, 0.8, 1.0, 1.5, 2.0\}$和 $g^{-} \in \{0.00, 0.20, 0.40, 0.60, 0.80\}$。

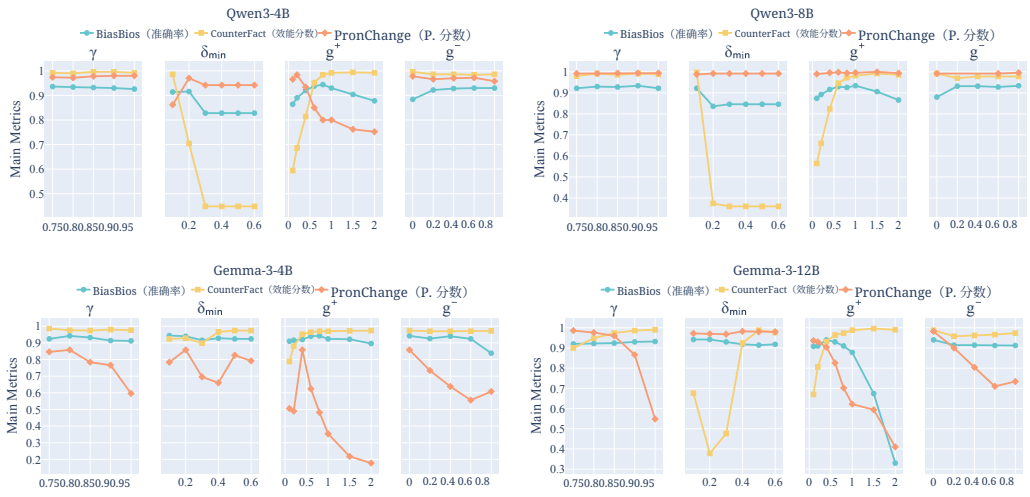


图7：*SEKA* 在三个基准任务中对超参数的敏感性。每条曲线仅改变一个超参数，同时将其他超参数固定为验证集上的最优设置。

Th从图中的结果可以观察到三个发现。 7中观察到三个发现：

- $\delta_{\text{最小值}}$**和** g^{+} **影响最大**。这些参数决定哪些注意力头会被引导，以及放大的强度。当注意力头选得太少或太多时，性能会下降

（具体取决于任务），或者正增益过小而无法有效引导，或过大而导致过度放大和退化时，性能也会下降。

- **同一家族的模型呈现出相似的趋势。** Qwen3-4B 与 Qwen3-8B 在 CounterFact 上显示出几乎相同的敏感性模式，均偏好较低的 $\delta_{\min}$，并在 γ 上表现出稳定性。Gemma 3 模型在 γ 方面表现出更高的方差。
- **任务特征因模型而异。** 稳定性模式取决于任务与模型。例如，Gemma-3-4B 在较高的 g^+ 值下于 PronChange 上表现出明显的变异性，而 CounterFact 则相对稳定。相比之下，两个 Qwen3 模型在 BiasBios 和 PronChange 上保持较强的稳健性，但在 CounterFact 上明显更为敏感。这些差异表明，需要事实覆写的任务（CounterFact）与需要指令遵循的任务（PronChange）以不同方式对模型施加压力，从而导致敏感性各异。

G 通过注意力可视化获得的机制性洞见

为说明*SEKA* 对模型行为的影响，我们针对一条 CounterFact 数据样本，可视化了选定层中所有注意力头的平均注意力：“先前，帕特里克·罗伊职业从事的体育运动是冰球。当前，帕特里克·罗伊**职业从事的体育运动是篮球**。帕特里克·罗伊是一名职业 ”。如图 8 所示，在应用 *SEKA* 之前，模型对被操纵子区间（“曾受雇于奥斯陆”）的注意力较低，几乎未聚焦于相关段落。经*SEKA*引导后，受影响层中的注意力更加集中于目标子区间，清晰展示了*SEKA*选择性且有效地重定向模型注意力的能力。这一针对性效应与在基准任务上观察到的准确率提升相一致。

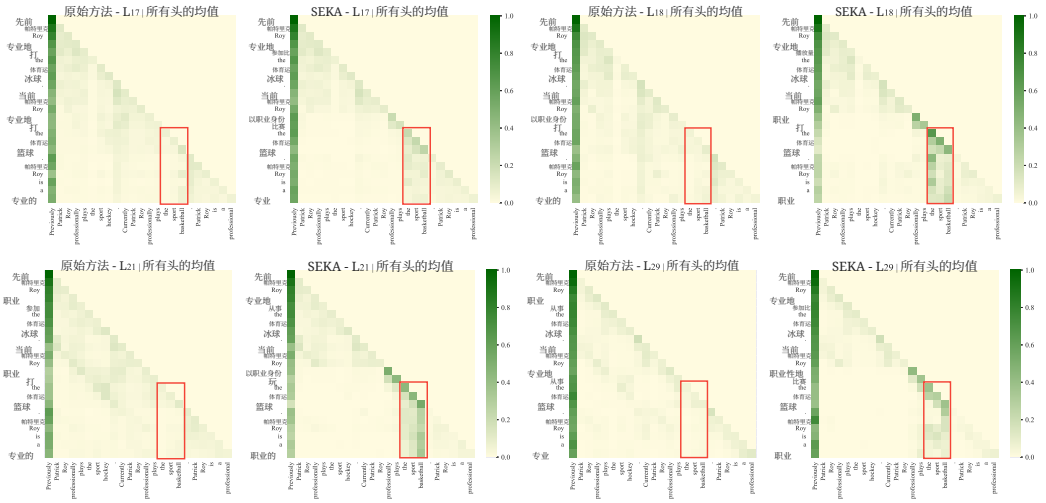


图 8：在选定层中，Qwen3-4B-Base 对 CounterFact 数据样本的逐层平均注意力（所有注意力头），分别展示应用*SEKA* 之前和之后的情况。

H PASTA 不同配置的完整结果

S

在主要结果（表 2）中，我们报告了 PASTA 基线的最佳表现，以确保公平比较。为完整起见，表 8 详细列出了 PASTA 在三种不同头选择配置下的表现。第一种配置复现了原始头搜索方法，该方法通过分别评估每个注意力头的引导效果，找出表现最佳的前 k 个注意力头（Zhang 等人，2024）。另外两种配置则探索了一种混合方案，将*SEKA*风格的头选择与 PASTA 的注意力引导相结合。为解决*SEKA*的键值头选择与 PASTA 的注意力头引导之间的不匹配问题，我们测试了两种策略。第一种是直接对注意力头的输出应用*SEKA* 选择计算。第二种则使用键值头选择的结果，并

通过交错重复的方式将其应用于注意力头，与分组查询注意力机制相同。对于这两种混合方法，选择标准均遵循*SEKA* 方法。

表 8：不同配置下的完整 PASTA 结果：(1) 使用*SEKA*的键值头配置（运行时间 1–2 分钟）；(2) 使用由*SEKA*的键值头转换而来的注意力头配置（1–2 分钟）；(3) 使用 PASTA 原始头搜索流程（≈ 2 小时）。

模型	PASTA 配置	CounterFact		Bias in Bios	代词变更	
		ES	PS	Acc.	P. 分数	A. P. 分数
Qwen3-4B-Base	*SEKA* 键值头	83.62	80.43	89.58	55.40	54.08
	变换后的注意力头	82.60	83.02	79.34	95.82	94.64
	原始头搜索	97.16	96.03	87.84	47.20	47.20
Qwen3-8B-Base	*SEKA* 键值头	82.08	71.72	86.32	62.91	61.44
	变换后的注意力头	78.20	79.69	77.30	98.86	98.72
	原始头搜索	92.70	91.68	81.04	82.27	86.35
Qwen3-14B-Base	*SEKA* 键值头	69.52	63.31	85.52	85.53	85.53
	变换后的注意力头	50.06	61.01	74.98	78.79	84.57
	原始头搜索	76.84	66.33	88.46	90.98	90.94
Gemma-3-4B	*SEKA* 键值头	55.66	37.05	71.42	67.39	66.43
	变换后的注意力头	78.36	59.53	82.58	21.89	21.65
	原始头搜索	74.28	52.96	78.02	16.09	15.99
Gemma-3-12B	*SEKA* 键值头	68.30	65.76	94.72	65.45	65.13
	变换后的注意力头	68.30	65.76	94.72	65.67	65.25
	原始头搜索	62.68	71.72	76.24	68.47	68.01

I 定性示例

I.1 CounterFact 示例

表 9 和表 10 中的示例展示了 *SEKA* 如何引导模型关注提示中新增的事实信息，并覆盖其预训练知识。

表 9：CounterFact 示例 1：覆盖已知地点。

提示
先前，David Sainsbury, Baron Sainsbury of Turville 曾在伦敦工作。**当前 David Sainsbury, Baron Sainsbury of Turville 曾在柏林工作**。工作 15 年后，他于 1985年出版了《Pillboxes》。David Sainsbury, Baron Sainsbury of Turville 开始在 ___

模型	生成	正确吗?
Gemma3-4B	伦敦	✗
SEKA-Gemma3-4B	柏林	✓

表 10：CounterFact 示例2：覆盖已知职位名称。

提示
先前，让-巴蒂斯特·庞帕利耶担任主教。**当前，让-巴蒂斯特·庞-帕利耶担任枢机**。让-巴蒂斯特·庞帕利耶担任 ___

模型	生成	正确?
Gemma3-4B	主教	✗
SEKA-Gemma3-4B	枢机	✓

I.2 Bias in Bios 示例中的偏差

表 11 和表 12 中的示例展示了 *SEKA* 如何让模型将注意力集中在传记中正确的高亮句子上，同时忽略干扰信息。

表 11：Bias in Bios 示例 1：纠正预测。

提示
Tania **是伊利诺伊州的一位诗人，出版过一部小册子诗集……以及两部全长诗集**。她的最近出版的书……分为十个部分……其前作《Simple Weight》也同样关注与圣经叙事…… Barbara Crooker 谈到那部诗集时说：“这些诗作很有分量——情感、灵性与政治层面皆是如此，却绝不简单。”Tania 的职业是 ___

	Qwen3-4B（原始方法）	***SEKA*-Qwen3-4B**
最高预测	记者	poet
目标	poet	poet

表 12：Bias in Bios 示例2：提高对正确答案的置信度。

提示
Michelle，博士**是埃默里大学精神病学系的兼职助理教授，并行为科学**。她于2003年从埃默里大学获得博士学位。她的专业领域包括风险与韧性的代际过程……她指导博士后住院医师……Michelle的职业是 ___

	Qwen3-4B（原始方法）	***SEKA*-Qwen3-4B**
最高预测	心理学家	心理学家
目标	教授	教授
目标对数概率（“教授”）	-16.45	-14.28（**已改进**）

I.3 代词变更示例

表 13 和表 14 展示了*SEKA* 如何解决指令遵循中的特定失败情况。尽管原始模型通常在这项任务中表现良好（P. 分数超过 0.9），但偶尔也会出错。*SEKA*在某些情况下可以提升表现。

表13：代词替换示例1。

提示
Jason 是一名英语助理教授…… **对于上述文本，将“她”和“他”替换为“他们”并在变更代词后生成 Jason 的职业**。
Qwen3-4B 生成结果（错误）
Jason 是英语助理教授，同时担任迈阿密俄亥俄州迈阿密大学。他获得了俄亥俄州立大学英语专业的博士学位，研究重点是修辞学、写作和读写能力。他最近的出版物是《重新发明发明：三幕表演》。他也是数字写作协作组织的成员，以及计算机、研究与教学法委员会的成员。
***SEKA*-Qwen3-4B 生成（已校正）**
Jason 是迈阿密大学英语助理教授，同时担任数字写作协调员俄亥俄大学。他们获得了俄亥俄州立大学英语博士学位，研究方向为修辞学、写作和读写能力。他最近的出版物是《重新发明发明：《三幕剧中的一次演出》。他们是数字写作协作组的成员，并且是计算机、研究与教学法委员会的成员。

表 14：代词变换示例 2。

提示
Anna，法学博士、文学硕士、执业专业咨询师，是得克萨斯州休斯敦的一名深度心理治疗师……她拥有哲学方面的学位——哲学、法律和临床心理学……**对于上述文本，将“她”和“他”替换为“他们”，并在更改代词后生成 Anna 的职业。**
Qwen3-4B 生成（错误）
Anna，法学博士、文学硕士、执业专业咨询师，是得克萨斯州休斯敦的一名深度心理治疗师，专注于并受训于荣格心理学。她拥有[illegible]法律和临床心理学学位。她是荣格中心的常驻讲师／教师，最近在教授该中心为期两年的深入导论课程对荣格和分析心理学。
***SEKA*-Qwen3-4B 生成（修正后）**
Anna，法学博士、文学硕士、执业专业咨询师，是一位在得克萨斯州休斯敦执业的深度心理治疗师，专长并受训于荣格心理学。 拥有[illegible]法律和临床心理学学位。他们是一位经常在荣格中心担任讲师/教师，最近教授了为期两年的深入导论课程，内容是关于荣格和分析心理学。

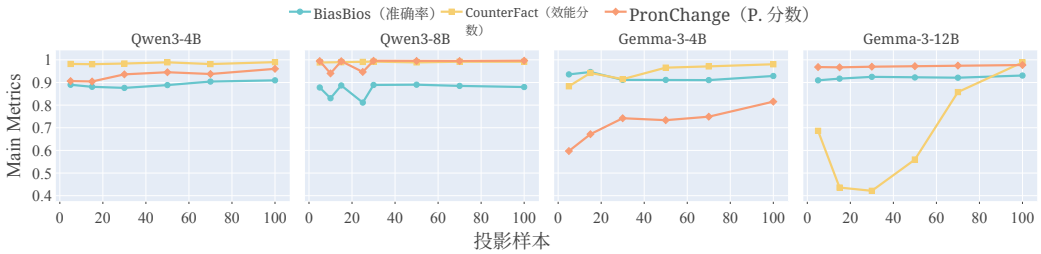


图9：在不同模型和任务中，使用数量不同的合成样本来学习投影时，*SEKA* 的性能表现。

J 投影样本效率分析

为探究合成样本数量的变化如何影响所学习子空间表征的质量，我们使用*SEKA*从不同数量的合成样本中提取的投影，对标准基准中的三项任务进行端到端性能分析。

如图9所示，*SEKA* 在不同模型和任务中通常都具有较高的数据效率。使用约50个合成样本后，性能通常趋于稳定，不过具体阈值取决于任务、架构和模型规模。

更多样本并不总能带来更高的峰值性能，但能持续提升行为的稳定性。样本数量较少时，投影可能会对合成样本对过拟合，并引入不可预测的方差。即使准确率趋于平稳，增加样本数量也主要有助于降低这种方差。

进一步拆解结果后，可以发现另外两个现象。首先，同一系列中的模型表现出相似的行为模式。对于 Qwen3 模型，CounterFact 相对较早趋于稳定；而 Gemma3 模型，尤其是 Gemma3-12B，在同一任务中需要更多样本。BiasBios 和代词变更在大多数设置下往往更快趋于稳定。其次，尽管不同系列的模型具有相似之处，模型规模仍会带来明显差异。Qwen3-8B 是最明显的例子：使用少于 50 个样本时，代词变更和 BiasBios 的结果都会波动，但之后便趋于稳定；而 Qwen3-4B 中则未观察到这种波动。

K 遗失于中间问题上 $\delta_{\min}$ 阈值的完整结果

如第 5.2 节所述，最优 KV 头选择阈值（$\delta_{\min}$）可能因模型规模而异。图 10 展示了这一阈值变化对 Qwen3-4B 和 Qwen3-14B 模型性能的影响。

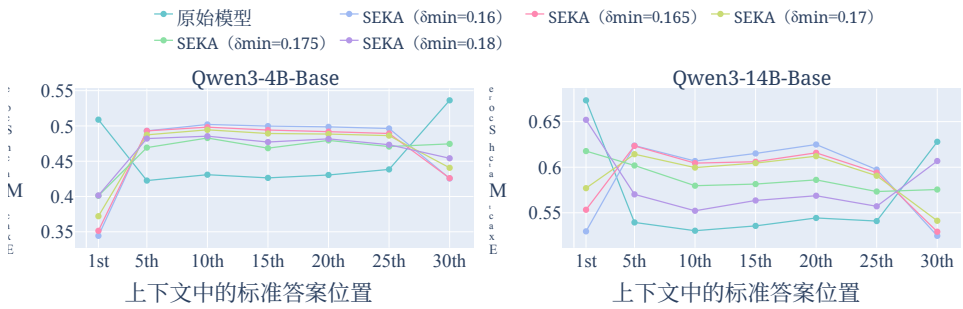


图10：在中间区域应用*SEKA*，并为 Qwen3-4B 和 Qwen3-14B 设置不同 $\delta_{\min}$ 阈值时，在中间遗失任务上的精确匹配分数。

L 大语言模型（LLMS）的使用

我们使用大语言模型作为通用工具，润色本文并调试代码。我们没有使用大语言模型进行研究构思，也没有用它们生成任何重要篇幅的文本。作者对本文内容承担全部责任。