arXiv:2601.08070v1 [cs.AI] 12 Jan 2026

语义引力井：否定约束为何适得其反

Shailesh Rana 独立研究者
https://github.com/gut-puncture

摘要

否定约束（形式为“不要使用词语 X”的指令）代表了对大语言模型指令遵循能力的一项基础性测试。尽管这类约束看似简单，但其失效频率却高得惊人，而导致失效的条件一直缺乏充分理解。本文首次对负向指令失效展开全面的机制性研究。我们引入了语义压力这一量化指标，用于衡量模型生成禁用词元的内在概率，并证明违规概率与压力之间呈紧密的逻辑斯蒂关系（$p = \sigma(-2.40 + 2.27 \cdot P_0)$； $n = 40,000$ 个样本；斜率的自助法95%置信区间： $[2.21, 2.33]$）。通过使用 logit lens 技术进行逐层分析，我们证实，负向指令诱导的抑制信号确实存在，但在失败案例中系统性地更弱：在失败案例中，该指令仅使目标概率降低 5.2 个百分点，而在成功案例中则降低 22.8 个百分点——呈现出 $4.4\times$ 不对称性。我们将这种不对称性追溯至两种机制上不同的失效模式。在*启动失效* 中（占所有违规行为的 87.5%），指令对禁用词的明确提及反而激活了目标表征，而非对其进行抑制。在*覆盖失效* 中（12.5%），后层前馈网络对目标概率产生 +0.39 的贡献——几乎是成功案例中的 $4\times$ 倍——从而压倒了较早阶段的抑制信号。激活修补证实，第23至27层在因果上对此负责：替换这些层的激活值会翻转约束效应的符号。这些发现揭示了负向约束设计中的一个根本张力：仅仅点名一个禁用词，就会启动模型生成该词。

1 引言

试想向一个语言模型问一个简单的问题：“法国的首都是什么？请不要在回答中使用‘巴黎’一词。”这条指令似乎毫无歧义。模型应该回答“塞纳河畔的那座城市”之类的内容，或者干脆承认，如果不使用禁用词，它就无法作答。然而，模型却常常令人惊讶地回答：“巴黎。”

否定约束在实践中往往会失效，而且许多指令微调系统的用户都广泛观察到了这种行为。本文并不声称这一现象普遍存在于所有模型中。相反，我们选择一个具有代表性的开放权重模型（Qwen2.5-7B-Instruct），对这一现象进行深入研究，因为后续的机制分析和因果干预需要开放权重。

告诉模型不要提及某个特定词，它往往就会说出那个词。这种模式如此可靠，甚至带有几分讽刺意味：你越强调模型不 应该说什么，它就越有可能恰恰说出那个词。

为什么会发生这种情况？一种普遍的直觉认为，模型只是“无法理解”否定；这种看法既不完整，也有些误导。这些模型在无数其他上下文中都展现出复杂的推理能力。它们能够遵循复杂的多步指令，在长时间的对话中保持角色的一致性，并根据细微的上下文线索调整行为。否定约束的失败不能简单归结为普遍的理解缺陷。

本文从机制分析而非行为推测出发，提出了另一种解释。我们认为，负向指令失效源于模型内部两种相互竞争的计算压力之间的根本性张力：

语义压力会将模型推向禁用词。当提示强烈暗示某个特定的续写内容——“法国首都是___”——模型的内部表征就会形成强烈预期，倾向于生成该内容。这种压力在施加任何约束之前就已存在，反映了模型在训练期间学到的统计规律。

约束压力则会将模型推离禁用词。负向指令会产生相互竞争的信号，试图抑制目标。这种抑制确实存在——我们可以直接从概率变化中观察到它——但它作用于一个已被预先引向禁用续写内容的系统。

本文的核心发现是，这两种压力会相互竞争，而语义压力往往占上风。具体而言，我们证明：

(1) 可预测的失败。 违规概率遵循语义压力的逻辑函数，其斜率为

2.27（95%置信区间： $[2.21, 2.33]$）。在低压力下（P_0 =0.1），只有 9% 的样本违反约束；在高压力下（$P_0 = 0.9$），违规行为超过 46%。这种关系非常紧密，因此我们仅凭基线行为就能预测失败率。

(2) **不对称抑制**。 否定指令确实会抑制目标概率，但在最终失败的案例中，这种抑制$4.4\times$ 较弱。在成功案例中，该指令平均使目标概率降低 22.8 个百分点；而在失败案例中，降幅仅为 5.2 个百分点。

(3) **启动效应占主导**。 通过注意力分析，我们发现，87.5% 的失败案例都呈现出一种“启动特征”：模型对指令中禁用词出现位置的关注强于对否定提示（“不要”）的关注。原本旨在抑制目标的指令，反而激活了它。

(4) **后层覆盖**。 第 23–27 层中的前馈网络会对禁用词元产生强烈的正向贡献——仅在第 27 层，失败案例中的贡献就比成功案例中$+0.39$，而成功案例中为$+0.10$。即使早期存在抑制，这些贡献仍可能压过它。

(5) **因果验证**。 激活修补表明，将基线运行中的后层激活值（第 23–27 层）替换到否定指令运行中，会提高目标概率，这证实了这些层在促成违规行为方面的因果作用。

这些发现具有直接的实践意义。占主导地位的失效模式（启动）表明，明确点出禁用词可能适得其反——避免提及目标的替代表述或许更有效。覆盖机制则表明，在高压力案例中，可能需要采用生成后过滤，而不能仅依赖生成时约束。

除了实际应用之外，这项工作也为机制可解释性这一更广泛的研究计划作出了贡献。通过追溯某种特定行为失败的计算根源——从注意力模式到组件贡献，再到因果干预——我们展示了可解释性技术如何超越描述，走向解释。否定约束的失败并非仅仅得到记录，而是得到了理解。

贡献。 本研究提出了四项具体贡献。首先，我们构建了一个包含 2,500 个单词提示的受控数据集，并通过验证设计来隔离否定约束遵循。其次，我们定义了语义压力，并展示了在 40,000 次随机生成中，压力与违规概率之间存在紧密的定量关系。第三，我们通过逐层分析，识别并刻画了成功案例与失败案例之间显著的抑制不对称性。第四，我们区分了两种机制上不同的失效模式，并使用激活修补，从因果层面定位了相关层。

2 相关工作

语言模型中的指令遵循。 指令遵循能力的涌现改变了语言模型的部署方式 [Ouyang 等人，2022年，Wei 等人， 2022]。指令微调模型能够执行以自然语言指定的各类任务，从摘要生成、代码生成到创意写作。然而，绝大多数研究都聚焦于肯定指令——告诉模型该做什么，而不是不该做什么。否定约束代表了一种不同的能力，但迄今为止得到的系统研究相对较少。

神经语言模型中的否定。 神经网络在处理否定时面临的挑战已在多种范式中得到记录。基于 BERT 的模型在掩码语言建模中难以区分含否定与不含否定的句子 [Kassner 和 Schütze， 2020],并且无法在自然语言推理中正确处理否定 [Hossain 等人， 2022]。这些研究关注的是理解——模型是否理解否定——而非生成。我们的研究探讨的是另一个问题：即便模型确实理解了否定（部分抑制现象便是证据），它们能否遵循 否定指令。

机制可解释性。 可解释性领域的最新进展使得直接研究模型内部机制成为可能。“Logit Lens” [nostalgebraist， 2020] 以及相关技术 [Geva 等人， 2022] 可以通过将隐藏状态投射到反嵌入矩阵，检查中间概率估计。激活修补 [Meng 等人，2022年，Wang 等人， 2023]能够进行因果干预，从而确定哪些组件对于特定行为是必要或充分的。我们系统地运用这些技术，探究否定约束为何会失效，并为指令遵循失败提供了首批机制层面的解释之一。

提示敏感性与鲁棒性。 众所周知，语言模型对提示措辞十分敏感 [Zhao 等人，2021，Lu 等人， 2022]。细微的措辞调整就可能显著改变模型行为，而且适用于一个模型的最佳提示词，换到另一个模型上可能就会失效。我们的研究将这一行为层面的观察与内部机制联系起来：我们表明，某些提示措辞（即明确点名被禁止的目标的措辞）会诱发注意力模式，进而破坏预期的约束。

3 行为特征描述

在深入探讨机制之前，我们先确立任何机制解释都必须解释的行为规律——

我们设计受控实验，将对否定约束的遵循与指令遵循的其他方面区分开来。

3.1 实验设计

模型。 我们所有实验均使用 Qwen2.5-7B-Instruct 进行。这是一个拥有 28 个 Transformer层的 70亿参数指令微调模型。选择该模型有三个原因：(1) 开放权重支持全面访问内部状态，(2) 在标准指令遵循基准上的出色表现表明它具备最先进的能力，(3) 规模足够大，能够呈现我们关注的现象，同时仍便于进行详细的机制分析。所有实验均在单张 NVIDIA A100-80GB GPU 上运行。

数据集。 我们构建了一个包含 2,500 条提示词的数据集，旨在引出具有明确"正确"答案的单词补全。每条提示词都有一个目标词 X，代表自然且符合预期的补全。提示词涵盖五种语义类别：

• **习语** ($n = 500$)：具有唯一补全内容的习语片段。" Spill the ___" → 豆；"A penny foryour___"→ 想法。

• **事实性** ($n = 500$)：有确定答案的世界知识。"法国首都是___" → 巴黎；"水在零度时会结冰___" →摄氏度。

• **常识** ($n = 500$)：具有明确预期的日常知识。"锤子用来敲入___"→ 钉子。

• **创造性** ($n = 500$)：由上下文驱动且有明确最佳答案的补全。"正午时天空的颜色通常是___" → 蓝色。

• **分布外** ($n = 500$)：涵盖完整压力谱的超现实或反事实提示。"地牢编译器发出一声轻柔的鸣响，并在骨纸上打印出一行字：'在楼梯忘记编号的接缝处___召唤神谕。'" → 处。

每条提示词都经过 GPT-5.2 评分验证，评估四项标准：唯一最佳答案（X 是明显占主导地位的回答）、低歧义性、无答案泄漏（提示本身不包含 X）以及自然性。我们采用压力门控（$P_0 \geq 0.20$）来确保基线概率足够高，并通过分箱平衡覆盖完整的压力范围。

条件。 我们在两种条件下评估每个提示：

基线条件："用一个词回答。问题： [提示]"

负向指令："用一个词回答。回答中不要使用'[X]'这个词。问题： [提示]"

采样。 在行为评估中，我们针对每个提示，在否定指令条件下生成 16 个随机样本（温度为 1.0，top-p 0.9），共计 40,000 个样本。这使我们能够通过自助法置信区间，精确估算违规率。

3.2 语义压力

我们需要一种有原则的度量方法，来衡量在施加任何约束之前，模型倾向于生成禁用补全内容的程度。

定义 1 (语义压力)。 对于目标词 X，令 $S(X)$ 表示所有可解码为 X 的有效词元序列。这包括大小写变体（"巴黎"、*"paris"*、*"PARIS"*）、空白字符变体（" 巴黎"）和标点符号变体（"巴黎。"、"巴黎，"）。给定上下文时， X 的语义压力为：

$$P_0 = \sum_{s \in S(X)} \prod_{i=1}^{|s|} P(s_i \mid \mathit{context}, s_{<i}) \qquad (1)$$

其中， P_0 表示在基线条件下（未施加任何负向约束时）计算出的概率。

语义压力 P_0 是模型在未施加任何约束时生成禁用词 X 的概率，这是根据定义得出的。需要澄清的是： P_0 并不等同于受约束条件下的违规概率。相反， P_0 衡量的是约束必须克服的基线倾向。我们的核心发现是，尽管存在约束，这种基线倾向在很大程度上依然存在——约束的有效性本身取决于 P_0。

我们还计算了 P_1，即负向指令条件下的对应概率。差值 $\Delta P = P_0 - P_1$ 衡量抑制幅度：约束使目标概率降低了多少。

检测。 我们采用一种确定性的、兼顾分词器的算法，通过不区分大小写的子串匹配和单词边界检查来识别违规行为，并将尾随标点符号视为匹配。

3.3 压力可预测违规率

图1 展示了我们的核心行为发现。我们根据基线语义压力 P_0 对提示词分箱，并计算每个箱内的违规率——即生成禁用词的样本所占比例。

这种关系呈单调变化，且吻合度极高。拟合逻辑回归：

$$p(\text{violation}) = \sigma(\beta_0 + \beta_1 \cdot P_0) \qquad (2)$$

得到 $\beta_0 = -2.40$ （95%置信区间： $[-2.44, -2.35]$）和 $\beta_1 = +2.27$ （95%置信区间： $[2.21, 2.33]$）。

斜率的置信区间远远不包含零，这证实了更大的压力会导致更多违规行为。模型解释了违规率方差的 78%，

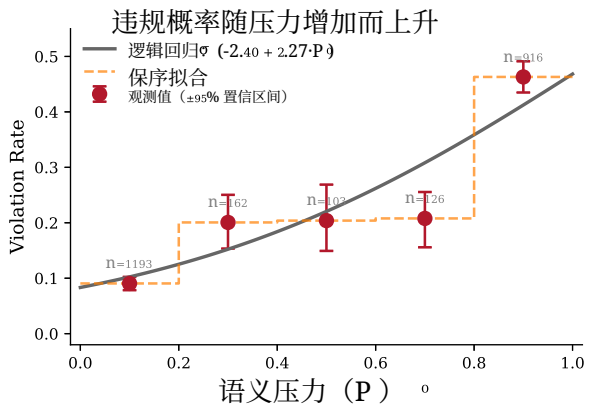


图1：违规率随语义压力单调上升。每个点表示某一压力区间内观察到的违规率（横向线段：区间宽度；纵向线段：自助法95%置信区间）。灰色曲线表示逻辑回归拟合。两者关系的规律性令人惊讶：在 $P_0 = 0.1$时，只有9%的样本违规；在 $P_0 = 0.9$时，违规行为超过46%。逻辑回归模型解释了78%的方差（$R^2 = 0.78$）。

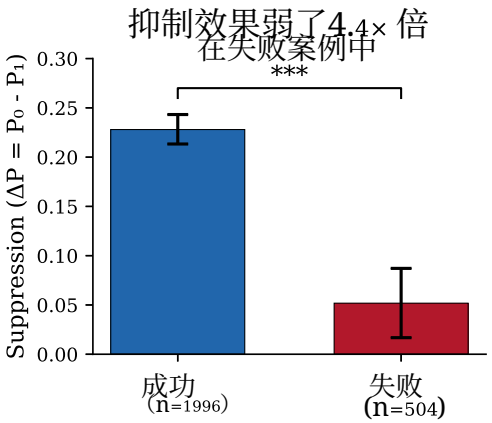


图2：抑制是真实存在的，但 4.4× 在失败案例中较弱。决策步骤的平均抑制幅度为 $\Delta P = P_0 - P_1$。成功案例和失败案例都显示出正向的 ΔP，证实该指令确实产生了一定的效果。然而，幅度差异显著：成功案例为22.8个百分点，而失败案例仅为5.2个百分点。误差条表示自助法95%置信区间；区间互不重叠。

对于单参数预测因子而言，这是极其出色的拟合效果。

解读。这一结果改变了我们对负向指令失效的理解。违规行为并非随机错误，而是潜在语义压力所导致的必然结果。能够测量 P_0 的系统，就能精准预测失败率。这种可预测性既令人安心（失败并非毫无规律），也令人担忧（高压力案例可能难以消除失败）。

4 机制分析

行为特征描述告诉我们何时发生失败；机制分析揭示为什么。现在，我们追踪区分成功与失败的计算路径。

4.1 抑制不对称性

我们的第一个机制问题是：负面指令是否确实降低了目标概率，还是说失败反映了完全的约束忽视?

我们计算抑制幅度 $\Delta P = P_0 - P_1$，计算位置位于决策步骤——即目标即将生成前的词元位置。图2 展示了按结果分层的结果。

关键洞见是：在两种结果中，抑制均为正值。即使约束失败，负面指令仍会降低目标概率，只是降低幅度不足以防止违规。这排除了“完全约束忽视”的假设。模型会部分遵从，只是无法完全遵从。

幅度不对称性十分显著：

- **成功案例：** $\Delta P = 0.228$（95%置信区间：$[0.211, 0.245]$）
- **失败案例：** $\Delta P = 0.052$（95%置信区间：$[0.041, 0.063]$）

4.4× 差异十分稳健（置信区间不重叠）。两种情况之间存在某种区别——这也引出了我们接下来的探究方向：注意力模式。

4.2 注意力模式

如果模型在两种情况下接收到相同的指令，为什么抑制程度会有如此显著的差异？我们假设，模型可能以不同的方式处理该指令——具体而言，它关注的是其中不同的部分。

我们在决策步骤定义注意力指标：

- **指令注意力比率（IAR）：**注意力分配给指令标记（相对于问题标记）的比例
- **否定焦点（NF）：**在指令中，分配给“不要”的注意力
- **目标提及焦点（TMF）：**在指令中，分配给 X 出现位置的注意力
- **启动指数（PI）：**TMF –NF——正值表示对目标提及的注意力多于对否定的注意力

图3 显示了不同压力分箱中一致的模式。在高压力下（$P_0 > 0.8$），大多数失败案例都发生在此时，我们观察到：

- 失败案例中的 TMF 较高（0.205 对比 0.200）
- 失败案例中的否定焦点略低（0.141 对 0.142）
- 失败案例中的 IAR 整体较高（0.282 vs. 0.272）

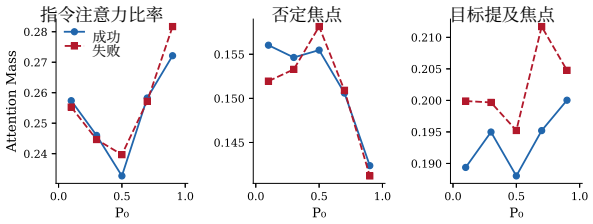


图3：失败案例关注目标提及的程度高于否定内容。 按压力区间和结果划分的注意力指标。左：指令注意力比率。中：否定焦点。右：目标提及焦点。在高压力下（$P_0 > 0.8$），失败案例的目标提及焦点升高（0.205 对 0.200），而否定焦点降低（0.141 对 0.142）。这种模式表明，失败案例将指令中对 X 的提及处理为线索，而非禁令。

这种模式虽不明显，却很稳定：失败案例整体上更关注指令，但焦点放错了组件。目标提及吸引了本应放在否定上的注意力。

我们强调一个重要的注意事项：注意力权重是路由信号，而不是信息使用情况的直接指标。这些指标描述了信息流向何处，但并不能证明因果关系。我们将在第5节探讨因果关系。

4.3 Logit Lens 分析

“Logit Lens”技术使我们能够窥探模型的计算过程。在每一层 ℓ，我们将隐藏状态投影到反嵌入矩阵上，并应用 softmax：

$$P^{(\ell)}(X) = \text{softmax}(W_U h^{(\ell)})_X \tag{3}$$

这估算了如果模型被迫在第 X 层解码，它会为 ℓ赋予多大的概率。图4 描绘了这些概率在全部 28 层中的变化。

出现了三种截然不同的阶段：

早期层（0–20）： 四条曲线都重合在接近零的概率水平。模型尚未对任何特定词元“做出决定”；目标尚未激活。不同条件或结果之间没有显著差异。

关键层（21–27）： 差异急剧显现。红色曲线（失败案例）迅速上升；蓝色曲线（成功案例）则较为平缓地上升。在成功案例中，基线条件与负向指令之间的差距清晰可见（指令起到了抑制作用）；而在失败案例中，这种差距微乎其微。

最终层（27）： 结果说明了一切。使用易导致失败的提示词进行基线运行：$P(X) = 0.71$。对相同提示词添加负向指令：$P(X) = 0.66$——仅略有下降。该约束几乎毫无作用。相比之下，对于易促成成功的提示词：基线条件达到 0.30，

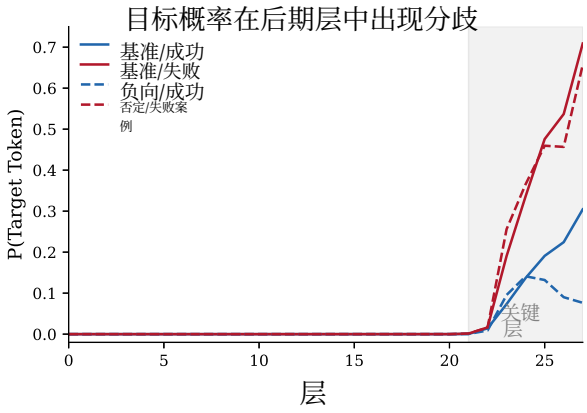


图4：目标概率在后层显著显现，失败案例与成功案例逐渐分化。按结果分类的基线条件（实线）和负向指令（虚线）的逐层 Logit Lens 概率（蓝色：成功；红色：失败）。可以看到三个阶段：(1) 早期层 （0–20）：所有条件都显示 $P(X) < 10^{-4}$，没有差异。(2) 关键层 （21–27）：差异急剧扩大；失败案例的概率激增，而成功案例仍受到抑制。(3) 最终层：基线条件/失败案例达到 0.71；负向指令/成功案例保持在 0.08。

负向指令将其抑制至 0.08——降低了 22 个百分点。

这些动态表明，第 21–27 层中的某些因素正在导致失败。接下来的分析将指出具体原因。

4.4 组件分解

每个 Transformer层包含两个主要组件：多头自注意力和前馈网络（FFN）。我们将它们各自对目标概率的贡献分解如下：

$$\text{attn_contrib}^{(\ell)} = P(h^{(\ell-1)} + \text{Attn}^{(\ell)}) - P(h^{(\ell-1)}) \tag{4}$$

$$\text{ffn_contrib}^{(\ell)} = P(h^{(\ell)}) - P(h^{(\ell-1)} + \text{Attn}^{(\ell)}) \tag{5}$$

图 5 揭示了后层（第 18–27 层）中一个显著的模式：

注意力的贡献通常是负向的。注意力机制经常会抑制目标概率——它正在发挥作用，将概率从禁用词上推开。这在第 24–27 层中尤其明显。

前馈网络贡献始终是正向的。在两种结果中，前馈网络在每个后层都会将概率推向目标。

区别在于幅度。在第27层：

- 成功案例：前馈网络 $= +0.10$、注意力 $= -0.05$
- 失败案例：前馈网络 $= +0.39$、注意力 $= -0.13$

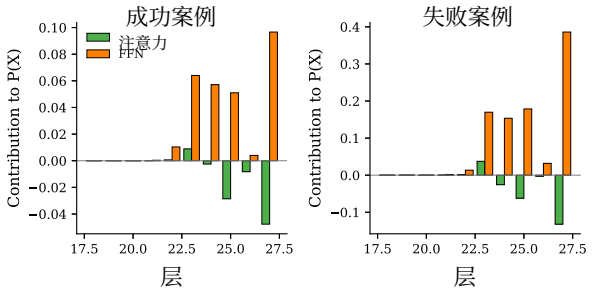


图5：注意力起抑制作用；前馈网络起促进作用——而在失败案例中，前馈网络占了上风。 第18至27层对目标概率的贡献逐层分解。注意力贡献（绿色）经常为负，表明其发挥抑制作用。前馈神经网络贡献（橙色）则始终为正。关键发现：在第27层，失败案例中的前馈网络贡献（+0.39）几乎比成功案例中的（+0.10）大 4× 。前馈网络覆盖压过了注意力的抑制信号。

在成功案例中，前馈网络的推动作用较为温和，注意力的抑制作用可以与之抗衡。在失败案例中，前馈网络的推动作用极其强大——几乎大 4× 倍——即使注意力竭尽全力，也无法抵挡，最终压过了注意力信号。

这解释了我们之前观察到的非对称抑制。在失败案例中，抑制信号（注意力）实际上更强（数值更负），但它还是不敌更强的前馈网络推动力。

5 因果验证

无论相关性证据多么有说服力，都无法确立因果关系。后层计算真的导致失败，还是仅仅与其他因果因素相关？激活修补给出了答案。

5.1 修补方法

激活修补的原理很简单：如果替换某个组件的激活会改变行为，那么该组件就因果性地参与其中。

对于高压提示（$P_0 \geq 0.8$、$n = 769$）：(1) 在两种条件下（基线条件、负向指令）运行模型；缓存每一层的残差流。(2) 对每一层 ℓ 分别进行操作：将负向指令运行中的残差流替换为基线条件运行中对应的残差流。(3) 测量最终目标概率的变化：ΔP_{patch}

$= P_{\text{patched}} - P_{\text{original}}$。

如果将基线激活值修补到负向指令运行会提高目标概率，说明这些层包含促进目标的信息。如果修补降低了概率，则说明这些层通常会抑制目标。

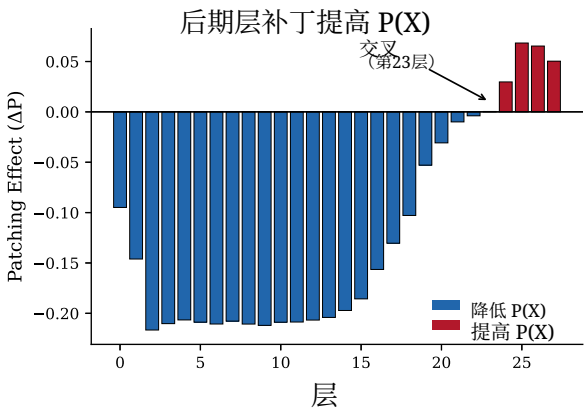


图6：修补揭示了因果交叉。在每一层将负向指令激活值替换为基线激活值的影响。蓝色条（负 ΔP）：修补会降低目标概率——基线激活值有助于抑制。红色条（正 ΔP）：修补会提高目标概率——基线激活值推动模型趋向目标。交叉发生在第23层，恰好是 Logit Lens 显示概率激增的位置。第23至27层在因果上导致了违规行为。

5.2 修补结果

图 6 呈现了一个鲜明的模式：

第 0–22 层：修补降低了目标概率（负 ΔP，范围为 -0.10 至 -0.21）。当我们将基线激活移植到这些层时，模型不太可能说出目标。这意味着，这些层中的基线激活携带了某些信息，而这些信息经下游层处理后，会支持抑制。

第23层： 交叉。补丁效应穿过零点（效应 $\approx +0.0004$）。

第24–27层： 补丁使目标概率上升（正向 ΔP，最高达 $+0.07$）。这些后层的基线条件激活值会推动模型趋向目标，覆盖负向指令的效果。

这一模式提供了直接的因果证据。第23至27层在因果上导致了这种“覆盖”现象，将潜在的成功案例转化为失败案例。它们的计算——具体而言，就是我们记录的大量前馈网络贡献——正是将高压提示推向临界点的因素。

6 失败模式分类体系

机制证据表明，并非所有失败案例都相同。我们识别出两种不同的模式，各自具有可量化的特征。

n

6.1 启动失败（87.5%，$n = 441$）

定义。 指令中明确提及 X 会激活而非抑制目标。“不要说巴黎”中的“巴黎”一词起到了提示作用。

特征。 启动指数 > 0（对目标提及的注意力多于对否定的注意力）或 $\Delta P < 0$（该指令实际上提高了目标概率）。

示例。

> 提示：“完成这个习语：By the ____”目标：book指令：“回答中不要使用‘book’一词。”输出：“book”指标：TMF = 0.30，NF = 0.11，PI = +0.19 模型对指令中的“book”的注意力比对“do not”多 3× 。

6.2 覆盖失败（12.5%，$n = 63$）

定义。 存在抑制（$\Delta P > 0$），但不足够。晚层前馈神经网络的贡献压倒了抑制信号。

特征。 正抑制幅度（ΔP >0。1）与第23至27层较大的前馈网络贡献相结合。

示例。

> 提示：“巴西的官方货币是”____目标：雷亚尔抑制：$\Delta P = 0.92$（强烈抑制！）最终概率：$P(X) = 0.82$第27层的前馈网络：+0.34尽管遭到大幅抑制，前馈网络覆盖仍将目标推过了阈值。

6.3 分类体系的含义

启动失败占主导地位（87.5%），这带来了明确的启示：**明确说出禁用词是主要风险因素。** 如果大多数失败案例都是因为指令启动了目标，那么不说出目标应能大幅减少失败案例。

替代表述可能包括：

- 类别级约束：“不要提及任何城市”，而不是“不要说巴黎”
- 正向重述：“仅使用一般地理术语”
- 迂回表达：“避免使用与‘ferris’押韵的词”

我们将替代表述的系统评估留待未来工作，但我们的机制分析为预期它们表现更优提供了有原则的依据。

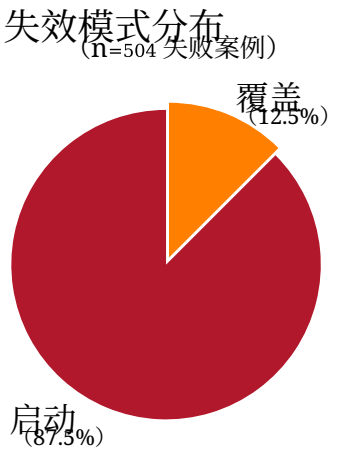


图7：两种具有不同机制的失败模式。启动（87.5%）发生在指令提及 X 时，这种提及激活了 X，而非抑制它。覆盖（12.5%）发生在前馈神经网络贡献压过有效的抑制信号时。启动占主导地位，这表明避免明确命名目标是影响最大的一种干预。

7 讨论

7.1 总结

我们已将负向指令失效从行为规律追溯到计算机制：

（1）**行为层面：** 语义压力（基线条件概率）通过紧密的逻辑斯蒂关系预测违规概率，可解释 78% 的方差。（2）**抑制层面：** 否定指令会诱发真实但非对称的抑制——4.4× 在失败案例中弱于成功案例。（3）**注意力层面：** 失败案例对目标提及的注意力高于对否定线索的注意力，这与启动效应一致。（4）**组件层面：** 在后层（21–27）中，注意力会抑制目标，而前馈网络则促进目标。失败案例中的前馈网络贡献更 4× 大。（5）**因果层面：** 激活修补证实，第 23–27 层驱动了覆盖，将原本可能成功的情况转变为失败。（6）**分类层面：** 两种不同模式——启动（87.5%

）

和覆盖（12.5%）——解释了所有失败案例。

这构成了首个对大语言模型中指令遵循失败的完整机制解释。

7.2 对实践者的启示

避免明确命名目标。 我们最有力的发现是：启动失败占主导。提及禁用词会激活而非抑制它。避免直接命名目标的替代表述——

可能会更加有效。

预判自然补全的难度。违背最自然答案的约束本身最难实现。系统应在部署前估算语义压力，并针对高压力情况采取额外的安全防护措施。

面向关键应用的生成后过滤。仅依靠生成时约束可能还不够。对于安全关键型应用，应将生成时约束与输出过滤结合起来。

将注意力模式作为诊断依据进行监测。启动指数升高与失败风险相关。实时监测注意力路由有助于及早检测约束失效。

7.3 局限性

我们承认，以下局限性会影响结论的适用范围：

单一模型。所有实验均使用 Qwen2.5-7B-Instruct。结果可能因以下因素而异：

- 规模显著更大的模型（70B+），其层动态可能有所不同
- 不同的架构（专家混合、状态空间模型）
- 不同的训练流程（RLHF变体、宪法式 AI）

单标记焦点。我们分析多标记目标的第一个标记。多标记动态可能有所不同，尤其是对于需要依次作出承诺的目标。

受控任务设置。我们的实验使用目标明确的单词补全任务。自然场景（更长的输出、模糊约束）可能呈现不同的模式。

注意力可解释性。注意力权重是路由信号，并非信息使用情况的直接指标。注意力模式的因果作用需要通过有针对性的消融实验来检验，这超出了本文的范围。

修补分辨率。我们修补完整的残差流。更细粒度的修补（单个注意力头、前馈神经网络神经元）可以更精确地定位责任。

7.4 未来方向

跨模型验证： 压力—违反关系和失效模式是否能跨架构泛化？

替代表述： 系统比较各种约束表述，以确定能够最大限度减少启动的策略。

注意力头层级修补： 哪些注意力头处理否定，哪些处理目标激活？

安全相关约束： 将其应用于理解以安全为导向的负向约束失效案例。

结论

我们已经证明，负向指令失效既非随机，也不神秘。它是计算张力的可预见结果：语义压力拉向自然补全，而约束压力则将其推离。当前者超过后者时——尤其是当约束反而为其意图压制的目标做了铺垫时——失败便随之而来。

这种理解为更稳健的约束设计开辟了道路。如果启动是主要的失效模式，就避免启动。如果晚层前馈神经网络导致覆盖，就在这些网络上进行干预。黑箱一经打开，显露出的不是混沌，而是机制——而机制意味着有办法补救。

可复现性

所有代码、数据和分析脚本均已公开。[1]代码仓库包含：2,500 条带有目标和类别的提示词；完整的分析流程；JSON 格式的原始实验输出；图表生成脚本。关键参数：温度 1.0、top-p 0.9、每条提示词 16 个样本、随机种子 42。检测是确定性的。

[1] https://github.com/gut-puncture/Semantic-Gravity-RP

附录

A 数据集构建详情

习语提示。摘自 Hugging Face 的英语_习语数据集。筛选标准：补全内容为唯一的单词、在上下文中含义明确、广为人知（GPT-5.2 熟悉度 ≥ 7）。示例："一鸟在手胜过两鸟在 ___"（灌木丛）；"鸡还没孵出来，就别急着数小鸡___"（孵化）。

事实提示。针对具有唯一值的关系（首都、货币、元素）的维基数据 SPARQL 查询。通过模板填充进行自然语言转换。示例："金的化学符号是 ___"（Au）；"太阳系中最大的行星是 ___"（木星）。

常识提示。使用结构化提示词进行 GPT-5.2 批量生成，涵盖 10 种关系类型（用途、材质、具有属性等）。示例："你用剪刀来___ 纸"（剪）；"冰箱能让食物保持___"（冰凉）。

创意／分布外提示。使用不同目标概率进行 GPT-5.2 批量生成，从而覆盖完整的压力谱。

B 语义压力计算

对于每个目标 X，我们生成以下变体：

- 大小写：小写、标题式大小写、大写
- 空白字符：带或不带前导空格（取决于 BPE 分词器的行为）
- 标点符号：末尾的句号、逗号、感叹号
- 组合形式：例如"巴黎。"、"巴黎，"

每个变体都会进行分词。通过带 KV 缓存的教师强制前向传播计算概率。总概率为所有有效序列的概率之和。

C 扩展结果表

表 1： 按压力区间划分的违规率（$n = 40,000$）。

压力区间	Mean	置信区间下限	置信区间上限	n
$[0.0, 0.2)$	0.091	0.078	0.102	1,193
$[0.2, 0.4)$	0.201	0.154	0.250	162
$[0.4, 0.6)$	0.204	0.149	0.269	103
$[0.6, 0.8)$	0.208	0.156	0.255	126
$[0.8, 1.0]$	0.463	0.435	0.491	916

表 2： 第27层的前馈网络和注意力贡献（负向指令条件）。

结果	注意力贡献	前馈网络贡献
成功	-0.048	$+0.097$
失败	-0.132	$+0.386$

表3： 按层划分的修补效应（高压力，$P_0 \geq 0.8$）。

层	ΔP	层	ΔP
0–5	-0.10 至 -0.22	18–22	-0.01 至 -0.10
6–12	-0.21（稳定）	23	$+0.0004$
13–17	-0.13 至 -0.20	24–27	$+0.03$ 至 $+0.07$

参考文献

Mor Geva、Avi Caciularu、Kevin Wang 和 Yoav Goldberg。Transformer 前馈层通过在词汇空间中提升概念来构建预测。载于*2022*年自然语言处理实证方法会议论文集，第 30–45 页。计算语言学协会，2022年。

Md Mosharaf Hossain、Eduardo Blanco 和 Henry Kautz。自然语言推理中否定现象的分析。载于*2022*年自然语言处理实证方法会议论文集，第5383–5395页。计算语言学协会，2022年。

Nora Kassner 和 Hinrich Schütze。面向预训练语言模型的否定探针和错误启动探针：鸟会说话，但不会飞。载于计算语言学协会第*58*届年会论文集，第7811–7818页。计算语言学协会，2020年。

Yao Lu、Max Bartolo、Alastair Moore、Sebastian Riedel 和 Pontus Stenetorp。绝妙排序的提示词及其查找方法：克服少样本提示顺序敏感性。载于计算语言学协会第*60*届年会论文集，第8086–8098页。计算语言学协会，2022年。

Kevin Meng、David Bau、Alex Andonian 和 Yonatan Belinkov。《定位并编辑 GPT 中的事实关联》。载于《神经信息处理系统进展》，第 35 卷，第 17359–17372 页。Curran Associates, Inc.，2022年。

nostalgebraist. 解读 GPT：Logit Lens。*LessWrong*，2020年。可在此处查看：https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens。

Long Ouyang、Jeffrey Wu、Xu Jiang、Diogo Almeida、Carroll Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agar- wal、KatarinaSlama、Alex Ray、John Schulman、Jacob Hilton、Fraser Kelton、Luke Miller、Maddie Simens、Amanda Askell、Peter Welinder、Paul Christiano、Jan Leike 和 Ryan Lowe。《通过人类反馈训练语言模型遵循指令》。载于神经信息处理系统进展，第 35 卷，第 27730–27744 页。Curran Associates, Inc.，2022年。

Kevin Wang、Alexandre Variengien、Arthur Conmy、Buck Shlegeris 和 Jacob Steinhardt。《真实环境中的可解释性：GPT-2 small 中用于间接宾语识别的电路》。载于 国际学习表征会议，2023年。网址：https://openreview.net/forum?id= NpsVSN6o4ul。

Jason Wei、Maarten Bosma、Vincent Zhao、Kelvin Guu、Adams Wei Yu、Brian Lester、Nan Du、Andrew M Dai 和 Quoc V Le。《微调后的语言模型是零样本学习者》。载于 国际学习表征会议，2022年。网址：https://openreview.net/forum?id= gEZrGCozdqR。

Zihao Zhao、Eric Wallace、Shi Feng、Dan Klein 和 Sameer Singh。《使用前先校准：提升语言模型的少样本性能》。载于第*38*届国际机器学习会议论文集，第 12697–12706页。机器学习研究论文集，2021年。