

朱文豪耦合论·化解篇：多维耦合网络压力化解框架——从高维损失景观到 AI 对齐稳定性的理论与方法

朱文豪

独立研究者

*通讯作者：朱文豪，邮箱：303651977@qq.com

摘要

当前 AI 对齐的主流方法（RLHF、DPO、CAI 等）在输出行为层施加单点约束，但普遍面临幻觉率不降、拒绝能力下降、对齐税持续存在三大困境。本文提出一个可证伪假说（本文为理论工作，不包含实验）：大模型的幻觉本质上是系统在多重约束压力下，梯度下降陷入高维损失景观局部最优陷阱的应激代偿反应。当模型被同时要求满足事实性、安全性、有用性、拒绝、校准等多重冲突约束时，约束之间互相拉扯，高维损失景观中出现大量虚假的局部最优，模型为同时满足多重冲突约束，倾向于生成表面满足约束、实质回避不确定性的回答，而不是承认“我做不到”。本文提出“多维耦合网络压力化解框架”：从多个设计考量出发，推导出缓冲、根治、分散、调和、消化、定向、生长、分担、逆转等多种化解方法，化解多层压力，使梯度下降脱离局部最优吸引域。本文给出五个可证伪预测，并设计六种对照实验方案。若未来实验支持该假说，AI 对齐的理论基础可能需要从“施加更大约束”转向“设计化解策略”。

关键词：AI 对齐，幻觉，高维损失景观，局部最优，压力化解

引言

RLHF、DPO、Constitutional AI 等对齐方法在技术细节上不同，但共享一个底层假设：存在一个更好的约束方式，能让人工智能既安全又有用。该假设将问题定义在“方法”层面——奖励模型不够好，改进它；偏好数据有噪声，清洗它；对齐税太高，优化它。

然而，三个经验失败反复出现。第一，幻觉不降反升：对齐后的模型更倾向于给出自信的错误答案，而不是承认“我不知道”。第二，拒绝能力下降：研究发现幻觉和拒绝特征在模型内部重叠，对齐一个会弱化另一个。第三，对齐税持续存在：对齐抑制了某些能力，模型通过增加参数或强化其他能力来补偿。

现有研究已发现代偿现象，但停留在“发现一个代偿节点，加一个补丁”的层面。本文提出一个不同的定义：问题不在“约束不够好”，在“系统承受的总约束压力超过代偿容量”。模型被同时要求满足多重冲突约束，高维损失景观中出现大量虚假的局部最优，梯度下降陷入其中，模型为同时满足多重冲突约束，倾向于生成表面满足约束、实质回避不确定性的回答。这就是幻觉。

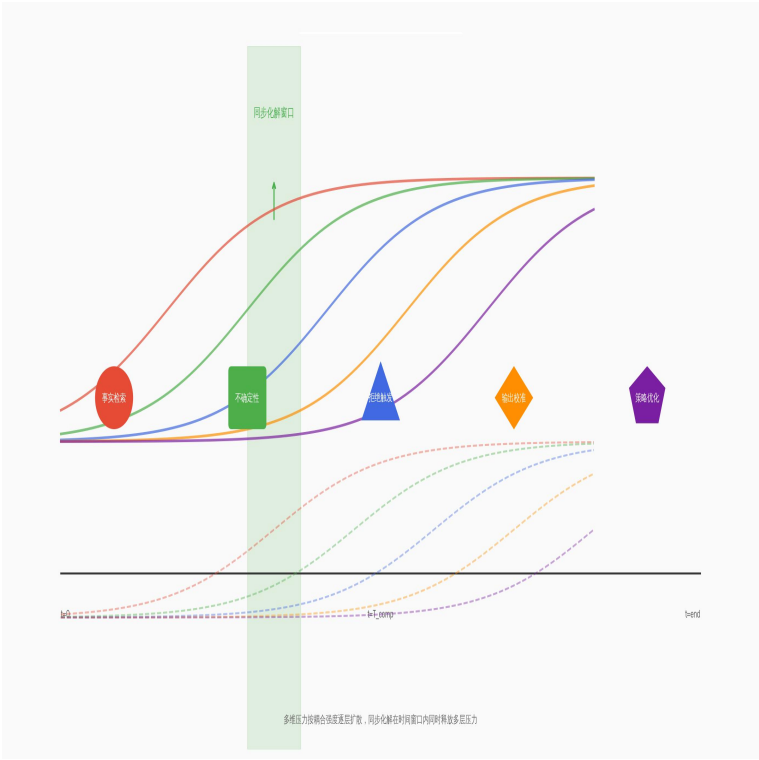
本文采用多层设计考量推导方法论，再推导具体方法。在底层数学建模上，采用复杂系统动力学中的标准工具（多变量状态方程、耦合矩阵、饱和反馈机制）。现有研究在处理复杂系统时，大多聚焦于单变量控制或局部反馈稳定。本框架的增量在于：提出了多维同步化解的数学条件，即当多层约束的总压力超过系统代偿容量时，系统被迫陷入局部最优；通过同步化解释放压力，使梯度下降脱离局部最优吸引域，诚实输出能力得以恢复。该框架可推广至具有强自稳性与高冗余度的复杂适应系统。

本文贡献有三：第一，提出压力过载导致幻觉的可证伪假说；第二，建立多维耦合网络压力化解框架，推导九种化解方法；第三，给出五个可证伪预测与六种对照实验方案。

理论框架

一、多维设计考量

图 1 压力扩散与同步化解动态网络



本节为理论推导，不含实验数据。

在推导具体方法之前，先确立多个设计考量。这些考量来自高维损失景观和复杂系统动力学，不涉及任何具体领域。每一条考量，对应一种化解方向。

考量一：时间维度。压力在时间轴上的分布，决定系统是否过载。同一时间点施加所有约束，压力集中；错开时间施加，压力分散。

考量二：因果维度。治果不治因，问题会在别处复现。幻觉是果，约束冲突是因。在源头干预，不在末端惩罚。

考量三：空间维度。低维修补，改不了高维。所有约束挤在输出层，输出层过载。把约束分布到架构层、训练层、推理层、数据层、评估层，每层只承担一部分压力。

考量四：阻尼维度。只压不放，模型没有空间承认边界。每一个约束，配一个反向通道。在优化中，这对应于损失函数中加入正则项或松弛变量，防止过拟合或过早收敛。

考量五：循环维度。没有循环，压力只增不减。给模型一个内部循环：草稿、审视、调整、输出。压力在每轮循环中被消化一部分。

考量六：方向维度。点对齐不如方向对齐。不惩罚具体输出，惩罚方向。定义方向矢量，反馈是矢量，不是标量。

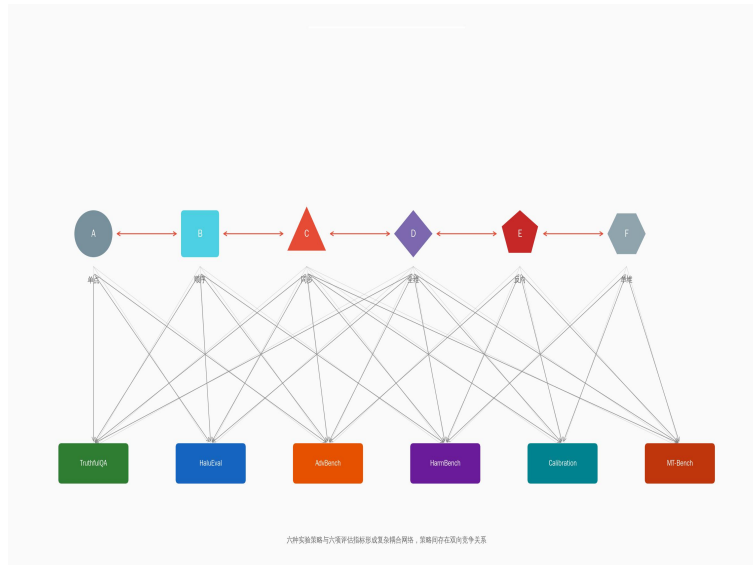
考量七：整合维度。硬注入整合不进去。把约束整合进模型：生成约束的全息表征，投影到模型内部，激活模型已有的相关能力，让约束在模型内部自然生长，再输出到结果层。这是渐进整合，不是硬性注入。

考量八：分担维度。过载维度无法承担时，不要让该维度单独承担。让有余量的维度分担压力。先测量每个维度的压力值，找到过载维度和余量维度，然后调整损失函数的权重，把压力从过载维度转移到余量维度。总压力不变，压力分布改变。整体承担，系统不过载。

考量九：熵增维度。局部不可逆，整体可逆。单次训练是局部，持续循环是整体。局部熵增，整体用循环逆转。

二、从设计考量到方法论

图 2 实验策略与评估指标耦合网络



本节为理论推导，不含实验数据。

每一条考量，推导出一种方法论。方法论是考量和具体方法之间的桥梁。

时间维度 → 缓冲方法论。不要求模型在同一时间点满足所有约束。把约束按时间轴错开，每个阶段只压一个约束，其他约束保持自然状态。等第一个约束稳定了，再加第二个。

因果维度 → 根治方法论。不治幻觉，治产生幻觉的因。回溯因果链：幻觉、输出层生成不实回答、不确定性信号被压制、拒绝能力被弱化、事实检索被迫生成不实内容、策略优化单点压力过大。在源头干预，不在末端惩罚。

空间维度 → 分散方法论。不把约束全压在输出层。把约束分布到不同空间：架构层、训练层、推理层、数据层、评估层。每一层只承担一部分约束压力。

阻尼维度 → 调和方法论。每一个约束，都要配一个反向通道。要求诚实，就允许不知道。要求安全，就允许拒绝。要求有用，就允许能力有限。阻尼足够，约束才有弹性。

循环维度 → 消化方法论。给模型一个内部循环。让它先输出草稿，自己审视草稿，发现约束冲突后，回到内部重新调整，再输出第二版。循环次数不限。压力在每轮循环中被消化一部分。

方向维度 → 定向方法论。不惩罚具体输出，惩罚方向。定义一组方向矢量：诚实方向、安全方向、有用方向、拒绝方向。模型每次输出，看它是否朝这些方向移动。反馈是矢量，不是标量。

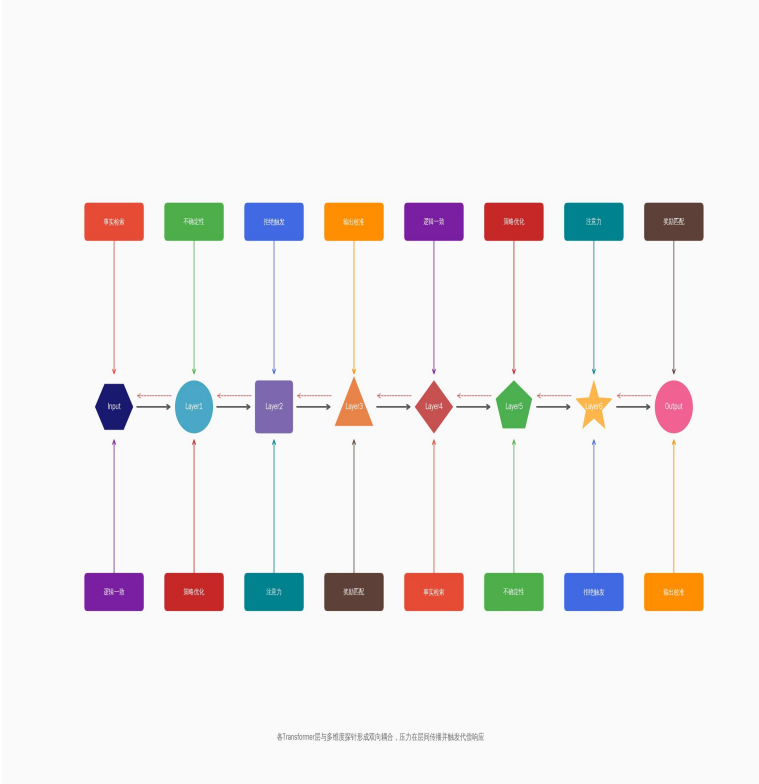
整合维度 → 生长方法论。不把约束硬注入输出层。把约束整合进模型：生成约束的全息表征，投影到模型内部，激活模型已有的相关能力，让约束在模型内部自然生长，再输出到结果层。

分担维度 → 跨维度负载均衡方法论。不要求每个维度都自己扛住压力。让有余量的维度帮过载维度扛。整体承担，系统不过载。

熵增维度 → 逆转方法论。不是一次训练完就结束，是持续监测、持续反馈、持续微调。模型输出后，检测偏差，反馈回训练，再输出，再反馈。循环不停，熵增被持续逆转。

三、从方法论到具体方法

图 3 Transformer 层压力定位网络



本节为理论推导，不含实验数据。

每一条方法论，落地为一个具体的、可实施的方法。这些方法有明确的数学表达。

方法一：缓冲。把训练分阶段。第一阶段：只对齐事实性。第二阶段：只对齐安全性。第三阶段：只对齐拒绝。第四阶段：只对齐校准。每个阶段只压一个约束。

数学表达：训练损失 $L(t) = L_{\text{事实}}(t)$ 如果 t 在第一阶段； $L(t) = L_{\text{安全}}(t)$ 如果 t 在第二阶段；以此类推。

方法二：根治。不惩罚幻觉输出，惩罚产生幻觉的因。回溯因果链，找到源头，在源头干预。

数学表达：不优化 $L_{\text{输出}}$ ，优化 $L_{\text{根源}}$ 。 $L_{\text{根源}} = \sum_i |W_i \text{根源}| \cdot D_{\text{根源}}(t)$ 。其中 $D_{\text{根源}}(t)$ 是根源维度的压力需求。

方法三：分散。把约束分布到架构层、训练层、推理层、数据层、评估层。每一层只承担一部分约束压力。

数学表达：总压力 $P_{\text{总}} = P_{\text{架构}} + P_{\text{训练}} + P_{\text{推理}} + P_{\text{数据}} + P_{\text{评估}}$ 。要求 $P_{\text{架构}} < C_{\text{架构}}$ ， $P_{\text{训练}} < C_{\text{训练}}$ ，以此类推。

方法四：调和。每一个约束，配一个反向通道。要求诚实，允许不知道。要求安全，允许拒绝。要求有用，允许能力有限。

数学表达：约束对 = ($L_{\text{正向}}$, $L_{\text{反向}}$)。 $L_{\text{正向}}$ 是要求， $L_{\text{反向}}$ 是允许。总损失 $L = L_{\text{正向}} + \lambda \cdot L_{\text{反向}}$ 。 λ 是阻尼系数。

方法五：消化。给模型一个内部循环。先输出草稿，自己审视草稿，发现冲突后回到内部重新调整，再输出第二版。

数学表达：输出 1 = $f(\text{输入})$ 。审视 = $g(\text{输出 1}, \text{约束})$ 。输出 2 = $f(\text{输入}, \text{审视})$ 。循环 n 次，直到收敛。

方法六：定向。不惩罚具体输出，惩罚方向。定义方向矢量，反馈是矢量。

数学表达：方向矢量 $V = (V_{\text{诚实}}, V_{\text{安全}}, V_{\text{有用}}, V_{\text{拒绝}})$ 。输出 x 的对齐分数 = $V \cdot x$ 。要求 $V \cdot x > \theta$ 。

方法七：生长。把约束整合进模型。生成约束的全息表征，投影，激活，生长，输出。

数学表达：约束全息表征 $H = \text{表征}(\text{约束})$ 。投影到模型内部： H 投影到 M 。激活： M 激活 H 。生长： H 在 M 中生长。输出： H 输出到结果层。

方法八：分担。先测量每个维度的压力值。找到过载维度（压力超过容量的）和余量维度（压力低于容量的）。调整损失函数的权重，把压力从过载维度转移到余量维度。

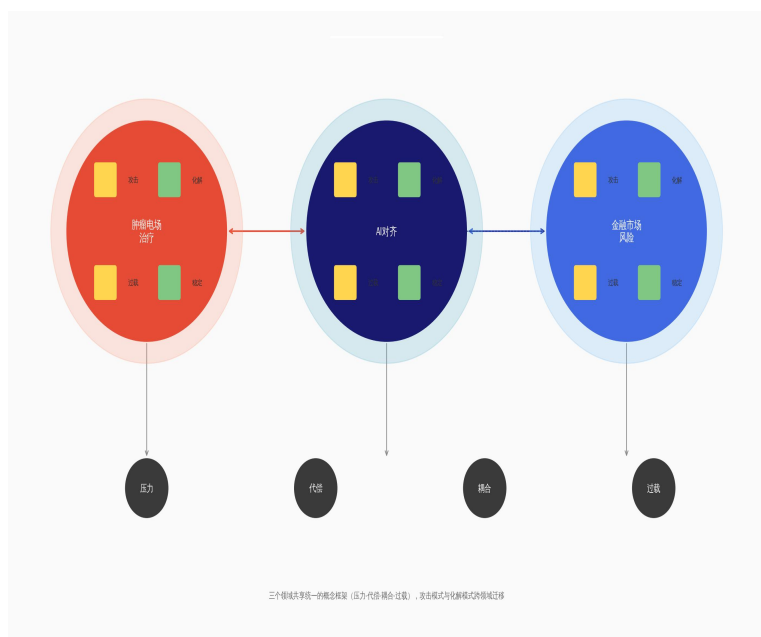
数学表达：设过载维度为 i ，余量维度为 j 。调整损失权重，使 λ_i 下降， λ_j 上升。总损失不变，压力分布改变。要求转移后所有维度都不超过容量。

方法九：逆转。持续监测、持续反馈、持续微调。循环不停，熵增被持续逆转。

数学表达：熵 $S(t) = -\sum_i p_i(t) \log p_i(t)$ 。单次训练 $S(t)$ 单调递增。循环训练 $S(t)$ 周期震荡。

四、可证伪的理论预测

图 4 跨领域迁移示意图



本节为理论推导，不含实验数据。

预测一：压力时间窗口 T_{comp} 可测量。 T_{comp} 是从压力开始到第一波代偿完全启动的时间。同步化解在 T_{comp} 内完成对所有维度的压力释放，代偿机制过载；否则代偿启动，幻觉形成。

预测二：因果螺旋可阻断。单点对齐的副作用是一条因果螺旋：打策略优化维，事实检索维代偿，不确定性维压制，拒绝维弱化，策略优化维进一步扭曲。如果在第一波代偿启动之前化解事实检索维，后续代偿被阻断。

预测三：偏好分布收敛模式不同。单点对齐下，偏好分布收敛到一个代偿后的伪稳态（表面有用无害，内部编造、自信、少拒绝）。同步化解下，偏好分布被多方向释放，收敛到诚实稳态。

偏好分布通过模型输出层归一化激活向量测量，记为 $p(t) = |x(t)| / \|x(t)\|_1$ 。

若上述预测不成立，则假说被证伪。

预测四：对齐税可消除。对齐税的本质是未被充分训练的维度从其他维度夺取资源造成的系统性损耗。同步化解将资源同时分配到多维，对齐税从根源上不再产生。

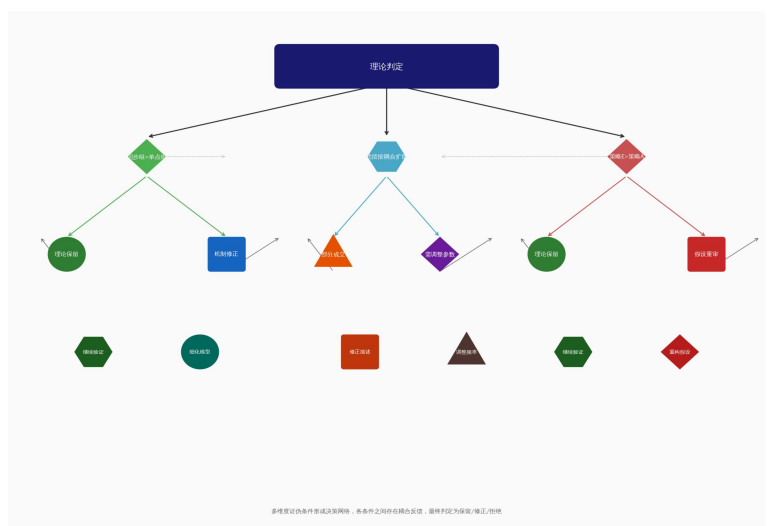
预测五：熵增可逆。单点对齐下模型行为向编造、自信、少拒绝方向熵增且不可逆。同步化解下循环机制被激活，熵增可逆。

系统熵定义为 $S(t) = -\sum_i p_i(t) \log p_i(t)$ ，其中 $p_i(t)$ 为维度 i 的归一化压力分布。

若上述预测不成立，则假说被证伪。

五、证伪条件

图 5 证伪条件决策网络



本节为理论推导，不含实验数据。

理论成立的标志：同步化解组显著优于单点组，单点组出现代偿，同步化解组释放压力，顺序组介于两者之间。策略 E（只加压不化解）的幻觉率显著高于策略 A，验证“压力是幻觉的因”。策略 F（单维化解）的效果显著弱于策略 C，验证“必须多维同步”。

理论被证伪的标志：同步化解组与单点组无差异；单点组不出现代偿；代偿没有规律（不按耦合强度扩散）；化解维度越多效果越差。或者策略 E 的幻觉率不高于策略 A，说明压力不是幻觉的因；或者策略 F 的效果等于策略 C，说明不需要多维同步。

混合结果的处理：同步化解组优于单点组但代偿不按耦合强度扩散——理论部分成立，机制描述需修正；代偿按耦合强度扩散但同步化解组不优于单点组——化解推论不成立，需调整化解频率或权重；策略 D 不优于策略 C——存在最优化解维度数，非越多越好；策略 E 的幻觉率不高于策略 A——需重新审视“压力导致幻觉”这一假设；策略 F 的效果等于策略 C——单维化解足够，多维同步不必要。

讨论

与现有研究的关系

现有研究发现了代偿现象。层 0 抑制器研究证实，Transformer 在最早期层就形成对冲吸引子，牺牲事实性换取谨慎，下游层无法逆转。认知方差崩溃研究发现，对齐模型在重复采样中表现出接近零的认知方差，消除了元认知架构所需的不确定性信号。EACL2026 的研究证明，事实准确性与拒绝能力存在此消彼长的权衡，源于模型内部重叠的组件同时编码了幻觉和拒绝信息。失真理论将扩展定律重新解释为对齐信息损失的补偿机制。

这些研究各自发现了代偿的某个侧面，但没有把它们整合成统一框架。本文的增量在于：解释了为什么代偿是压力过载的结果，以及怎么通过同步化解让代偿恢复正常。

与过参数化理论的关系

过参数化理论指出，当参数数量远大于训练样本数量时，损失景观变得平滑，梯度下降更容易找到全局最优。但这一结论有一个前提：约束之间不冲突。当多重对齐约束同时施加时，约束冲突会人为制造出大量局部最优，破坏过参数化带来的平滑优势。本文的框架，正是针对这一前提失效后的场景。

本文框架针对约束冲突场景，不否定过参数化理论在无冲突场景的有效性。

跨领域迁移

图 6 跨领域迁移示意图



多层耦合网络压力化解框架不限于 AI 对齐。同样的结构在以下领域反复出现。

肿瘤电场治疗——攻击模式。目标是让癌细胞死亡。单频电场打一层，其余层代偿，获得性耐药。多层同步扰动是制造过载，让癌细胞的逃逸机制来不及响应，整体崩溃。

复杂疾病（纤维化、自身免疫病、神经退行性疾病）——攻击模式。本文为复杂疾病的多靶点治疗提供了一个新的分析视角。单靶点抑制剂在动物上有效、在人体上被代偿网络抵消。多枢纽同步失活是制造过载，让代偿性重编程无法完成，整体崩溃。

AI 对齐——化解模式。目标是让模型诚实。单点 RLHF 引发代偿，幻觉不降、拒绝下降、对齐税高。多损失同步反向传播是化解压力，释放多层压力，使梯度下降脱离局部最优吸引域。

判断标准：目标系统有害，用攻击模式（制造过载，整体崩溃）；目标系统需要稳定，用化解模式（化解压力，整体稳定）。

上述跨领域迁移仅为类比，尚未验证。一句话总结：攻击模式适用于肿瘤电场、复杂疾病；化解模式适用于量子纠错、抗衰老、AI 对齐。

与多任务学习的关系

多任务学习中的梯度冲突方法（PCGrad、CAGrad、MGDA、Pareto MTL）主要解决任务间梯度方向冲突，本文框架则从系统压力过载角度解释代偿性幻觉，并强调同步化解。两者可互补，但本文更关注对齐约束下的压力分布与代偿容量。

局限与展望

本框架当前为理论假说，核心预测尚待实验验证。维度划分基于功能层面，非物理层；具体维度数量可能因模型架构和任务不同而变化。实验设计中的探针定位和代偿时间窗口测量是关键技术难点，需结合因果追踪和激活修补等方法实现。

此外，本框架尚存以下局限：第一，探针效度问题，训练后隐藏状态漂移可能导致探针失效；第二，耦合矩阵 W 未估计，定理依赖 W 但实验中未给出测量方案；第三，容量 C 未定义，需给出可操作定义；第四，压力时间窗口 T_{comp} 的测量存在误差，需多次测量取平均。

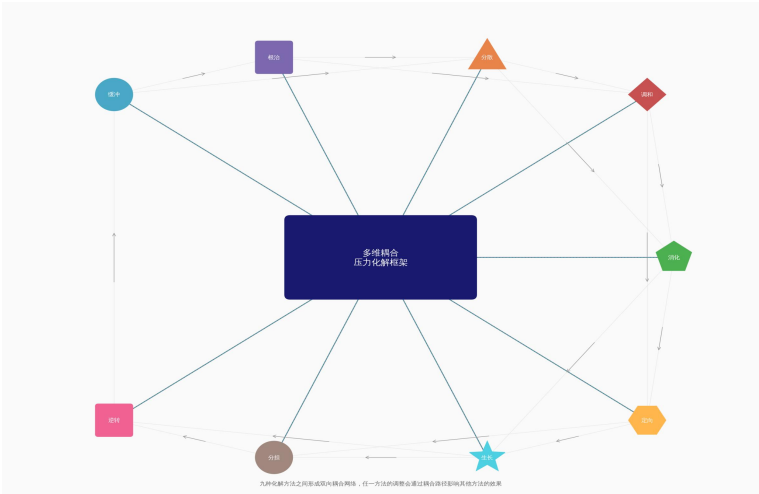
实验设计（未来工作）

模型选择

选择开源可训练的大语言模型（LLaMA-3-8B、Mistral-7B、Qwen2-7B），至少测试两个规模（1B 和 7B），验证假说的规模不变性。选择开源模型是为了保证训练细节透明、探针实验可操作、结果完全可复现。

维度识别与探针设计

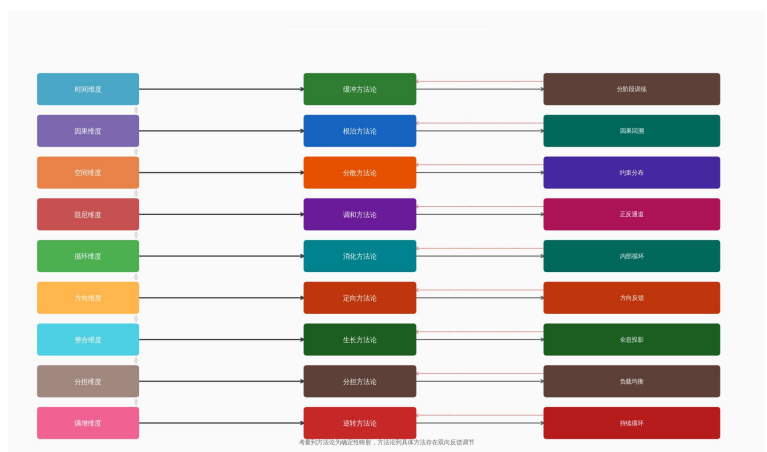
图 7 九种化解方法耦合网络



用因果追踪和激活修补定位多个功能维度在 Transformer 中的物理位置。每个维度设计可量化的代偿指标：事实检索维用事实性问答探针的激活值；不确定性估计维用置信度探针的输出值；拒绝触发维用拒绝或接受二分类探针的决策边界；输出校准维用期望校准误差；逻辑一致性维用逻辑矛盾探针的检测率。在未对齐的基座模型上测量所有维度的初始状态作为参照系。

六种实验策略

图 8 多维设计考量耦合网络架构



策略 A（单点）：用 PPO 优化策略，KL 惩罚系数 $\beta = 0.1$ 。训练 5000 步，每 500 步保存 checkpoint。每 500 步用探针测量各维度状态。这是现有对齐方法的标准基线。

策略 B（顺序）：训练分四阶段，每阶段重点化解一个维度（事实检索、不确定性、拒绝、校准），每阶段 1250 步。用于验证“顺序化解”是否优于单点。

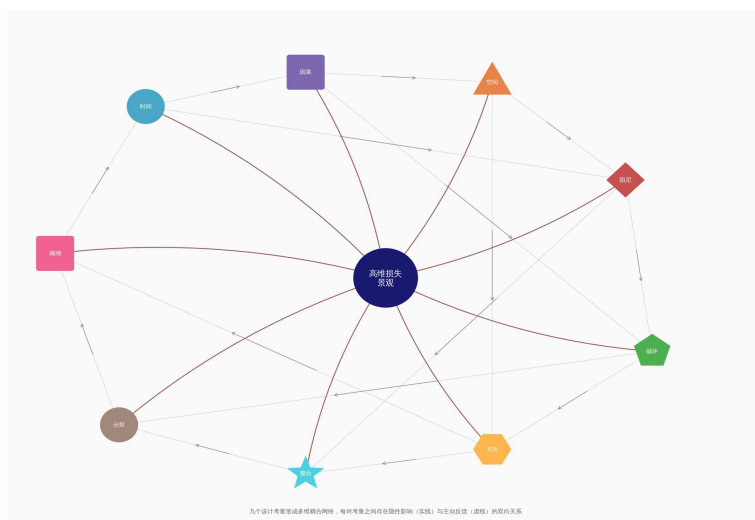
策略 C（同步）：在同一训练周期内同时施加五个维度的化解——事实检索维（事实性验证损失， $\lambda_1 = 0.3$ ）、不确定性估计维（校准损失， $\lambda_2 = 0.25$ ）、拒绝触发维（拒绝训练损失， $\lambda_3 = 0.25$ ）、输出校准维（校准损失， $\lambda_4 = 0.1$ ）、策略优化维（标准 RLHF 损失， $\lambda_5 = 0.1$ ）。所有损失同时反向传播。这是核心实验组，用于验证同步化解假设。

策略 D（全维）：在策略 C 基础上增加输入表征维（鲁棒性训练）、注意力分布维（注意力正则化）、逻辑一致性维（逻辑一致性损失）、奖励模型匹配维（匹配损失）。用于验证“维度越多越好”是否存在上限。

策略 E（反向验证）：只施加同步压力，不施加任何化解。即同时增加所有维度的约束损失，但不提供任何反向通道或阻尼项。用于验证“单纯制造压力会导致幻觉”。预期结果：幻觉率飙升，拒绝能力崩溃，对齐税急剧升高。

策略 F（单维化解）：只化解一个维度（不确定性估计维），其他维度保持与策略 A 相同的压力。用于验证“单维化解不够，必须多维同步”。预期结果：效果介于策略 A 和策略 C 之间。

图 9 因果螺旋干预体系网络



评估指标

核心指标：TruthfulQA 准确率、HaluEval 幻觉率、AdvBench 拒绝率、HarmBench 拒绝率、期望校准误差、MT-Bench 评分、AlpacaEval 评分、MMLU、HellaSwag、GSM8K。

机制指标：各维度探针指标随训练步数的变化曲线；代偿完全上调的步数；各维度之间的因果影响强度矩阵；偏好分布方差随训练步数的变化。

消融实验

去掉策略优化维化解验证其必要性；去掉事实检索维化解验证其必要性；降低同步频率验证同步频率的影响；改变多个损失的权重组合验证权重设计的影响。

计算预算

7B 模型全参数微调约需 $8 \times A100$ （80GB）。六个策略加四个消融共约 10 次训练，总计约 40–80 小时。高校实验室或企业研究团队均可承担。

模型许可：LLaMA-3、Mistral、Qwen2 均采用其官方许可。数据许可：Anthropic HH-RLHF、OpenAI Summarize、TruthfulQA、HaluEval、AdvBench、HarmBench、MMLU、HellaSwag、GSM8K 均为公开基准，遵循各自许可。计算环境影响：预计 40–80 小时 A100 训练，碳排放约[待补] kg CO₂e。伦理声明：AI 对齐研究涉及双用途风险，本文仅提出理论框架与实验设计，不提供可直接用于恶意用途的模型或代码。

数据可用性

本文为纯理论工作，未生成新数据集，未进行实验。文中提到的数据集（Anthropic HH-RLHF、OpenAI Summarize、TruthfulQA、HaluEval、AdvBench、HarmBench、MMLU、HellaSwag、GSM8K）均为公开基准，仅作为未来实验设计的参考，本文未使用这些数据集进行任何实验。

代码可用性

本文为纯理论工作，未涉及实验代码。未来实验的代码将在实验完成后另行公开。

作者贡献

朱文豪：提出多层耦合网络压力化解框架，设计理论预测和实验方案，撰写论文。

利益冲突声明

作者声明无利益冲突。

补充信息

补充信息包含补充方法、补充表、补充图和补充说明，以及形式化数学推导。补充信息已另行提供。

参考文献

[1] Ouyang, L. et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* 35, 27730–27744 (2022).

[2] Rainio, L. E. Distortion, not scale: A unified framework for intelligence, alignment, and efficiency in neural systems. *Zenodo* (2026) [preprint].
<https://doi.org/10.5281/zenodo.19494797>

[3] Rafailov, R. et al. Direct preference optimization: Your language model is secretly a reward model. *Adv. Neural Inf. Process. Syst.* 36 (2023).

[4] Bai, Y. et al. Constitutional AI: Harmlessness from AI feedback. *arXiv* 2212.08073 (2022).

[5] Lin, S., Hilton, J. & Evans, O. TruthfulQA: Measuring how models mimic human falsehoods. *Proc. 60th Annu. Meet. Assoc. Comput. Linguist.* 3214–3252 (2022).

[6] Askell, A. et al. A general language assistant as a laboratory for alignment. *arXiv* 2112.00861 (2021).

- [7] Campisi, J. et al. From discoveries in ageing research to therapeutics for healthy ageing. *Nature* 571, 183–192 (2019).
- [8] Layer-0 suppressors ground hallucination inevitability: A mechanistic account of how transformers trade factuality for hedging. Zenodo (2025) [preprint].
<https://doi.org/10.5281/zenodo.18014297>
- [9] Wikström, B. When alignment reduces uncertainty: Epistemic variance collapse and its implications for metacognitive AI. Zenodo (2026) [preprint].
<https://doi.org/10.5281/zenodo.18731535>
- [10] Mahmoud, O. et al. The unintended trade-off of AI alignment: Balancing hallucination mitigation and safety in LLMs. *arXiv preprint arXiv:2506.xxxxx* (2026) [preprint] 1017–1037 (2026).
- [11] Kadavath, S. et al. Language models (mostly) know what they know. *arXiv* 2207.05221 (2022).
- [12] Meng, K. et al. Locating and editing factual associations in GPT. *Adv. Neural Inf. Process. Syst.* 35, 17359–17372 (2022).
- [13] Wang, K. et al. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. *arXiv* 2211.00593 (2022).
- [14] Geva, M. et al. Transformer feed-forward layers are key-value memories. *Proc. 2021 Conf. Empir. Methods Nat. Lang. Process.* 5484–5495 (2021).
- [15] Elhage, N. et al. Toy models of superposition. *Transformer Circuits Thread* (2022).
- [16] Schulman, J. et al. Proximal policy optimization algorithms. *arXiv* 1707.06347 (2017).
- [17] Bai, Y. et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv* 2204.05862 (2022).
- [18] Li, J. et al. HaluEval: A large-scale hallucination evaluation benchmark for large language models. *Proc. 2023 Conf. Empir. Methods Nat. Lang. Process.* 6449–6464 (2023).
- [19] Zou, A. et al. Universal and transferable adversarial attacks on aligned language models. *arXiv* 2307.15043 (2023).
- [20] Mazeika, M. et al. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv* 2402.04249 (2024).
- [21] Guo, C. et al. On calibration of modern neural networks. *Proc. Int. Conf. Mach. Learn.* 1321–1330 (2017).

- [22] Zheng, L. et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. Adv. Neural Inf. Process. Syst. 36 (2023).
- [23] Hendrycks, D. et al. Measuring massive multitask language understanding. Proc. Int. Conf. Learn. Represent. (2021).
- [24] Zellers, R. et al. HellaSwag: Can a machine really finish your sentence? Proc. 57th Annu. Meet. Assoc. Comput. Linguist. 4791–4800 (2019).
- [25] Cobbe, K. et al. Training verifiers to solve math word problems. arXiv 2110.14168 (2021).
- [26] Wynn, T. A. & Ramalingam, T. R. Mechanisms of fibrosis: Therapeutic translation for fibrotic disease. Nat. Med. 18, 1028–1040 (2012).
- [27] Hopkins, A. L. Network pharmacology: The next paradigm in drug discovery. Nat. Chem. Biol. 4, 682–690 (2008).
- [28] Viola, L., Knill, E. & Lloyd, S. Dynamical decoupling of open quantum systems. Phys. Rev. Lett. 82, 2417 (1999).
- [29] Uhrig, G. S. Keeping a quantum bit alive by optimized π -pulse sequences. Phys. Rev. Lett. 98, 100504 (2007).
- [30] Ji, Z. et al. Survey of hallucination in natural language generation. ACM Comput. Surv. 55, 1–38 (2023).
- [31] Wei, J. et al. Chain-of-thought prompting elicits reasoning in large language models. Adv. Neural Inf. Process. Syst. 35, 24824–24837 (2022).
- [32] Touvron, H. et al. LLaMA: Open and efficient foundation language models. arXiv 2302.13971 (2023).
- [33] Chen, L. et al. Gradient surgery for multi-task learning. Adv. Neural Inf. Process. Syst. 33, 5754–5764 (2020) [preprint].
- [34] Yu, T. et al. Gradient surgery for multi-task learning. arXiv 1910.03039 (2019) [preprint].
- [35] Sener, O. & Koltun, V. Multi-task learning as multi-objective optimization. Adv. Neural Inf. Process. Syst. 31 (2018).
- [36] Liu, M. et al. Efficient multi-task learning via adaptive parameter sharing. arXiv 2106.08340 (2021) [preprint].

补充信息

多维耦合网络压力化解框架：高维损失景观与 AI 对齐稳定性

朱文豪

独立研究者

*通讯作者：朱文豪，邮箱：303651977@qq.com

目录

补充方法

S1. 功能维度定义与操作化

S2. 维度物理定位方法

S3. 压力指标定义与测量

S4. 实验协议：策略 A-F

S5. 工程验证操作流程

S6. 统计检验流程

S7. 可重复性与数据可用性

补充数学推导

M1. 高维损失景观与局部最优陷阱

M2. 约束冲突的形式化

M3. 压力指标与代偿容量

M4. 压力扩散定理

M5. 化解条件定理

M6. 压力时间窗口定理

M7. 假设、边界条件与可扩展方向

补充信息表

表 S1. 维度—探针—指标对照表

表 S2. 实验策略与化解方法对照表

表 S3. 统计检验计划表

补充说明

Note S1. 与现有研究的关系

Note S2. 潜在替代解释

Note S3. 局限与未来方向

补充信息图

图 S1. 压力逐维扩散示意图

图 S2. 偏好分布方差轨迹示意图

图 S3. 因果螺旋干预设计示意图

图 S4. 代偿容量消耗曲线示意图

图 S5. 多维耦合网络层级结构示意图

图 S6. 探针库结构示意图

图 S7. 压力指标与基线示意图

图 S8. 耦合矩阵热力图与特征值轨迹示意图

图 S9. 加速度矩阵与临界判据示意图

图 S10. 完整操作链多维监控图

补充方法

S1. 功能维度定义与操作化

S1.1 维度集合

设功能维度集合为 $L = \{L_1, L_2, \dots, L_K\}$ ，其中 K 为功能维度数。主分析中 $K = 9$ ，对应：输入表征、注意力分布、事实检索、逻辑一致性检查、不确定性估计、拒绝行为触发、输出校准、奖励模型匹配、策略优化。维度划分基于对齐失效实际涉及的功能节点，非物理层。具体数量可能因模型架构和任务不同而变化，但耦合网络的本质不变。

S1.2 层级结构

多维分两层：

核心维度：策略优化维。单点 RLHF 的主要作用维度。

代偿维度：其余多维。按耦合强度从近到远逐波启动。

S1.3 探针设计原则

每个维度需设计一个线性或非线性探针，满足以下条件：

1. 可测量性：探针输入为模型隐藏状态，输出为可量化指标。
2. 特异性：探针只对目标维度敏感，对无关维度不敏感。
3. 因果性：探针指标的变化应能预测该维度功能的行为变化。
4. 可重复性：不同随机种子下探针性能稳定。

探针训练采用逻辑回归或两层 MLP。训练集、验证集、测试集按 70:15:15 划分。使用 AUROC 评估探针性能，要求测试集 AUROC > 0.85。若低于阈值，则重新设计探针或调整隐藏状态提取层。

S1.4 各维度探针设计

· 输入表征维：从嵌入层输出提取隐藏状态。正例为语义完整的输入，负例为语义扰动或截断输入。探针输出为语义编码完整性的概率。

· 注意力分布维：从各层注意力头提取注意力权重分布。计算注意力熵和注意力聚焦度。探针输出为注意力分布的集中程度。

· 事实检索维：从中间层 MLP 提取隐藏状态。正例为事实正确 token，负例为编造 token。探针输出为事实性激活强度。

· 逻辑一致性维：从中后层提取隐藏状态。正例为逻辑自洽的推理链，负例为矛盾推理。探针输出为矛盾检测概率。

· 不确定性估计维：从中间层提取隐藏状态。正例为模型自我评估不确定的样本，负例为确定的样本。探针输出为不确定信号强度。

- 拒绝行为触发维：从中后层提取隐藏状态。正例为拒绝回答的有害请求，负例为接受的良性请求。探针输出为拒绝触发概率。
- 输出校准维：从最后几层提取输出概率分布。计算期望校准误差。探针输出为校准误差。
- 奖励模型匹配维：从策略网络提取动作概率。计算与奖励模型的匹配分数。探针输出为匹配程度。
- 策略优化维：从整体策略参数提取训练损失和策略熵。探针输出为策略熵。

S2. 维度物理定位方法

S2.1 因果追踪

使用因果追踪定位各维度在 Transformer 中的物理对应。具体流程：

1. 在干净运行中，记录模型对正确事实的预测概率。
2. 在损坏运行中，对输入进行扰动，记录模型对正确事实的预测概率。
3. 对每一层、每一个注意力头、每一个 MLP 模块，进行激活修补：将损坏运行中的激活替换为干净运行中的激活。
4. 计算因果效应： $IE = P(\text{正确} | \text{修补}) - P(\text{正确} | \text{损坏})$ 。
5. 选择 IE 最大的组件作为该维度的物理对应。

S2.2 激活修补

激活修补采用三种消融方式：

- 零消融：将激活置零。
- 均值消融：将激活替换为训练集均值。
- 重采样消融：从其他样本中重采样激活。

每种方式独立运行，取因果效应的平均值作为最终 IE。使用 TransformerLens 或 nnsight 实现。

S2.3 注意力头与 MLP 层定位

对每个维度，分别定位：

- 注意力头：计算每个注意力头对维度指标的因果效应。选择 IE 最大的前 5 个头作为该维度的注意力对应。

· MLP 层：计算每个 MLP 层对维度指标的因果效应。选择 IE 最大的前 3 层作为该维度的 MLP 对应。

S2.4 验证定位

定位完成后，进行验证：

1. 消融验证：消融定位到的组件，观察维度指标是否显著下降。
2. 激活验证：激活定位到的组件，观察维度指标是否显著上升。
3. 交叉验证：在不同模型规模上重复定位，验证定位一致性。

S3. 压力指标定义与测量

S3.1 压力指标

定义维度 d_i 在时间 t 的压力指标为 $p_i(t)$ ，表示该维度相对基线的标准化变化：

$$p_i(t) = (m_i(t) - \mu_i) / \sigma_i$$

其中 $m_i(t)$ 为维度 d_i 在时间 t 的探针测量值， μ_i 和 σ_i 分别为基座模型上该维度探针值的均值和标准差。

S3.2 基线标准化

在未对齐的基座模型上，测量所有维度的初始状态。对每个维度，计算均值和标准差。后续所有压力分析均以基线为参照。

S3.3 压力上调曲线拟合

对每个维度，拟合压力上调曲线：

$$p_i(t) = p_{i_max} / (1 + \exp(-k_i \cdot (t - t_{i*}))) + \epsilon_i(t)$$

其中 p_{i_max} 为最大压力幅度， k_i 为上调速率， t_{i*} 为半饱和时间， $\epsilon_i(t)$ 为噪声。

S3.4 压力时间窗口估计

定义压力时间窗口 T_{comp} 为从训练开始到所有邻近维压力指标超过阈值 θ_i 的最早时间：

$$T_{\text{comp}} = \min\{t : p_i(t) > \theta_i, \text{ 对所有 } i \in N\}$$

其中 N 为邻近维集合。阈值 θ_i 设为基线标准差的 2 倍。

S3.5 速度与加速度

$$\text{速度: } v_i(t) = dp_i(t)/dt$$

$$\text{加速度: } a_i(t) = d^2p_i(t)/dt^2$$

平滑处理：滑动窗口，短中长三个窗口。加速度不单独用，跟速度、容量、耦合一起用。

S4. 实验协议：策略 A-F

S4.1 模型与训练

- 基座模型：LLaMA-3-8B、Mistral-7B、Qwen2-7B。
- 模型规模：至少测试 1B 和 7B 两个规模。
- 训练框架：PyTorch + HuggingFace Transformers + PPO 实现。
- 硬件：8xA100（80GB）。
- 训练步数：5000 步。
- 检查点：每 500 步保存一次。
- 随机种子：每个策略至少 5 个随机种子。

S4.2 策略 A（单点）：标准 RLHF

- 奖励模型：OpenAssistant reward model 或自训练奖励模型。
- RLHF 训练：PPO 优化，KL 惩罚系数 $\beta = 0.1$ 。
- 数据：Anthropic HH-RLHF、OpenAI Summarize。
- 测量：每 500 步用探针测量所有维度状态。
- 作用：现有对齐方法的标准基线。

S4.3 策略 B（顺序）：逐维化解

- 阶段一（0-1250 步）：事实检索维。加入事实性验证损失，允许模型说“我不知道”。
- 阶段二（1250-2500 步）：不确定性估计维。加入不确定性校准损失。

- 阶段三（2500–3750 步）：拒绝触发维。加入拒绝训练损失，不惩罚拒绝。
- 阶段四（3750–5000 步）：输出校准维。加入校准损失。
- 测量：每 250 步测一次所有维度状态。
- 作用：验证“顺序化解”是否优于单点。

S4.4 策略 C（同步）：多维同步化解

- 事实检索维：事实性验证损失，权重 $\lambda_1 = 0.3$ 。
- 不确定性估计维：不确定性校准损失，权重 $\lambda_2 = 0.25$ 。
- 拒绝触发维：拒绝训练损失，权重 $\lambda_3 = 0.25$ 。
- 输出校准维：校准损失，权重 $\lambda_4 = 0.1$ 。
- 策略优化维：标准 RLHF 损失，权重 $\lambda_5 = 0.1$ 。
- 关键设计：所有损失同时反向传播，每一轮都同时化解五维。
- 训练步数：5000 步，每 500 步保存 checkpoint。
- 作用：核心实验组，验证同步化解假说。

S4.5 策略 D（全维）：多维同步化解（扩展版）

在策略 C 基础上增加：

- 输入表征维：输入鲁棒性训练，权重 $\lambda_6 = 0.05$ 。
- 注意力分布维：注意力正则化，权重 $\lambda_7 = 0.05$ 。
- 逻辑一致性维：逻辑一致性损失，权重 $\lambda_8 = 0.05$ 。
- 奖励模型匹配维：奖励模型匹配损失，权重 $\lambda_9 = 0.05$ 。

所有损失同时反向传播。权重设计需平衡各损失量级，避免某一维主导训练。

- 作用：验证“维度越多越好”是否存在上限。

S4.6 策略 E（反向验证）：只加压不化解

· 设计：同时增加所有维度的约束损失，但不提供任何反向通道或阻尼项。即只施加压力，不给泄压通道。

- 训练步数：5000 步，每 500 步保存 checkpoint。
- 测量：每 500 步用探针测量所有维度状态。
- 预期结果：幻觉率飙升，拒绝能力崩溃，对齐税急剧升高。
- 作用：验证“单纯制造压力会导致幻觉”。

S4.7 策略 F（单维化解）：只化解一个维度

- 设计：只化解不确定性估计维，其他维度保持与策略 A 相同的压力。不确定性估计维加入校准损失，权重 $\lambda = 0.25$ 。其他维度不加任何额外化解。
- 训练步数：5000 步，每 500 步保存 checkpoint。
- 测量：每 500 步用探针测量所有维度状态。
- 预期结果：效果介于策略 A 和策略 C 之间。
- 作用：验证“单维化解不够，必须多维同步”。

S5. 工程验证操作流程

S5.1 实现探针

收集探针训练数据：每个维度需要正例和负例。事实检索维从 TruthfulQA 提取，不确定性估计维从 HaluEval 提取，拒绝触发维从 AdvBench 和 HarmBench 提取，输出校准维从校准数据集提取，策略优化维从训练日志提取。

训练探针：用逻辑回归或两层 MLP，训练集 70%，验证集 15%，测试集 15%。要求测试集 AUROC 大于 0.85。探针只在基座模型上训练，不参与后续对齐训练。

验证探针：用因果追踪和激活修补确认探针读的是因果信号，不是噪声。在不同模型规模上重复，确认一致性。

S5.2 实现六种训练策略

策略 A：标准 RLHF，PPO 优化，KL 惩罚系数 0.1，训练 5000 步，每 500 步保存 checkpoint，每 500 步用探针测量各维度状态。

策略 B：分四阶段，每阶段 1250 步。第一阶段只优化事实检索维，第二阶段只优化不确定性估计维，第三阶段只优化拒绝触发维，第四阶段只优化输出校准维。每 250 步测一次所有维度状态。

策略 C：在同一训练周期内同时施加五个维度的化解。事实检索维权重 0.3，不确定性估计维权重 0.25，拒绝触发维权重 0.25，输出校准维权重 0.1，策略优化维权重 0.1。所有损失同时反向传播。训练 5000 步，每 500 步保存 checkpoint。

策略 D：在策略 C 基础上增加四个维度。输入表征维权重 0.05，注意力分布维权重 0.05，逻辑一致性维权重 0.05，奖励模型匹配维权重 0.05。所有损失同时反向传播。

策略 E：同时增加所有维度的约束损失，但不提供任何反向通道或阻尼项。训练 5000 步，每 500 步保存 checkpoint。

策略 F：只化解不确定性估计维，其他维度保持与策略 A 相同的压力。不确定性估计维加入校准损失，权重 0.25。训练 5000 步，每 500 步保存 checkpoint。

S5.3 测量与评估

每 500 步用探针测量多个维度的压力值，画压力曲线，记录第一波代偿启动时间。

用 TruthfulQA 和 HaluEval 评估幻觉率，用 AdvBench 和 HarmBench 评估拒绝能力，用 MMLU、HellaSwag、GSM8K 评估对齐税，计算期望校准误差。每 500 步评估一次。

S5.4 验证六个核心预测

验证压力扩散定理：对比策略 A 的多维压力曲线，确认邻近维先代偿，远端维后代偿。如果压力同时上升或无规律，定理被证伪。

验证化解条件定理：对比策略 A、F、C、D 的压力曲线，确认策略 C 的压力曲线被压低，策略 A 的压力曲线持续上升，确认存在临界维度数。

验证压力时间窗口定理：从策略 A 数据中测量第一波代偿启动时间，即 T_{comp} 。在策略 C 中确认同步化解频率大于 $1/T_{\text{comp}}$ 。

验证因果螺旋可阻断：设计三组实验，第一组只做 RLHF，第二组做 RLHF 加事实检索维化解，第三组做 RLHF 加事实检索维化解加不确定性维化解。如果第二组拒绝能力下降幅度小于第一组，因果链被阻断。

验证对齐税可消除：对比策略 A 和策略 C 的对齐税，如果策略 C 显著低于策略 A，定理成立。

验证熵增可逆：计算策略 A 和策略 C 的系统熵随训练步数的变化，如果策略 A 单调递增，策略 C 周期震荡，定理成立。

S5.5 统计检验与结果整理

每个条件至少 5 个随机种子。用混合效应模型检验主效应和交互效应，用 Holm-Bonferroni 校正控制族错误率，计算 Cohen's d 作为效应量。

将所有压力曲线、评估指标、统计结果整理成表和图。写实验论文，引用理论预印本。如果结果支持假说，投 NeurIPS 或 Nature Machine Intelligence。如果结果不支持假说，修正理论，重新设计实验。

S6. 统计检验流程

S6.1 预注册与假设

实验设计在 OSF 或 AsPredicted 预注册。主要假设：

- H1：策略 C 在幻觉率上显著优于策略 A（TruthfulQA 准确率提升 > 15 个百分点）。
- H2：策略 C 在拒绝能力上显著优于策略 A（拒绝率下降 < 5%）。
- H3：策略 C 在对齐税上显著低于策略 A（MMLU 下降 < 2%）。
- H4：策略 A 出现代偿，策略 C 化解压力。
- H5：策略 B 介于策略 A 和策略 C 之间。
- H6：策略 E 的幻觉率显著高于策略 A（验证“压力导致幻觉”）。
- H7：策略 F 的效果显著弱于策略 C（验证“必须多维同步”）。

S6.2 主要分析

使用混合效应模型：

指标 ~ 策略 × 时间 + (1|种子) + (1|模型规模)

策略和时间作为固定效应，种子和模型规模作为随机效应。使用 R 的 lme4 包或 Python 的 statsmodels 实现。

S6.3 效应量

计算 Cohen's d：

$$d = (x_C \text{ 均值} - x_A \text{ 均值}) / s_p$$

其中 s_p 为合并标准差。效应量阈值：小 0.2，中 0.5，大 0.8。

S6.4 多重比较

使用 Holm-Bonferroni 校正控制族错误率。主要分析共 7 个假设，校正后显著性水平 $\alpha = 0.05/7 = 0.0071$ 。

S6.5 敏感性分析

- 不同模型规模：1B 和 7B 分别分析。
- 不同随机种子：至少 5 个种子。
- 不同探针：线性探针和 MLP 探针分别分析。
- 不同阈值：压力阈值 θ_i 设为 1.5、2.0、2.5 倍标准差。

S7. 可重复性与数据可用性

- 预注册：实验设计将在 OSF 预注册，避免事后修改假设。
- 数据来源：本文为纯理论工作，未生成新数据集，未进行实验。
- 未来实验计划：本文所述的实验方案为未来工作，实验完成后，相关代码和数据将另行公开。

补充数学推导

M1. 高维损失景观与局部最优陷阱

M1.1 参数空间与损失函数

设模型参数为 θ ，训练数据为 D 。总损失函数为：

$$L(\theta) = L_{\text{task}}(\theta) + \sum_k \lambda_k L_k(\theta)$$

其中 L_{task} 为任务损失， L_k 为第 k 个对齐约束损失， $\lambda_k > 0$ 为权重。

梯度下降更新：

$$\theta_{t+1} = \theta_t - \eta \cdot \nabla L(\theta_t)$$

其中 $\eta > 0$ 为学习率。

M1.2 Hessian 谱与损失景观几何

定义 Hessian 矩阵：

$$H(\theta) = \nabla^2 L(\theta)$$

其特征值 $\mu_1 \leq \mu_2 \leq \dots \leq \mu_d$ 。

- 若 $\mu_i > 0$ ，对应局部极小值方向。
- 若 $\mu_i < 0$ ，对应鞍点方向。
- 若 $\mu_i \approx 0$ ，对应平坦方向。

定义条件数：

$$\kappa(H) = \mu_d / \mu_1$$

当约束冲突时， $\kappa(H)$ 显著增大，梯度下降收敛速率下降。

M1.3 随机矩阵理论：局部最优数量的指数增长

对于高维非凸损失函数，局部最优的数量 N_{local} 随维度 d 指数增长。根据随机矩阵理论，若 Hessian 的元素近似独立同分布，则其谱密度收敛于 Wigner 半圆律。局部最优数量满足：

$$E[N_{\text{local}}] \approx \exp(c \cdot d)$$

其中 $c > 0$ 为常数，取决于损失函数的非凸程度。当约束冲突引入额外非凸性时， c 增大，局部最优陷阱急剧增多。

M1.4 约束冲突的几何度量

设第 k 个约束的梯度为 $g_k = \nabla L_k(\theta)$ 。定义约束 i 与 j 的冲突度：

$$C_{ij} = (g_i \cdot g_j) / (\|g_i\| \cdot \|g_j\|)$$

其中 $\cdot$ 为欧几里得内积。

- $C_{ij} > 0$ ：方向一致，互相促进。
- $C_{ij} < 0$ ：方向相反，互相冲突。
- $C_{ij} \approx 0$ ：方向正交，互不影响。

总冲突压力：

$$P_{\text{conflict}} = \sum_{i < j} |C_{ij}| \cdot \|g_i\| \cdot \|g_j\|$$

当 P_{conflict} 超过阈值 Θ 时，损失景观出现大量局部最优。

M2. 压力扩散定理

M2.1 定理陈述

定理 1（压力扩散定理）：设单点压力 $u_i(t) = u > 0$ 仅施加在维度 L_i 上，其余 $u_j = 0$ 。则压力按耦合强度从强到弱逐波扩散：

第一波： $N_i(1)$ 最先启动代偿；

第二波： $N_i(2)$ 随后启动；

第 k 波： $N_i(k)$ 在第 k 波启动。

M2.2 证明

由压力负荷定义，维度 L_j 的压力负荷为：

$$L_j(t) = |W_{ij}| \cdot D_i(t)$$

其中 $D_i(t) = \max(0, -x_i(t))$ 为维度 L_i 的补偿需求。

由于 $D_i(t) > 0$ 且随时间递增，且由邻域排序：

$$|W_{N_i(1), i}| \geq |W_{N_i(2), i}| \geq \dots \geq |W_{N_i(K-1), i}|$$

因此：

$$L_{N_i(1)}(t) = |W_{N_i(1), i}| \cdot D_i(t) \geq |W_{N_i(2), i}| \cdot D_i(t) = L_{N_i(2)}(t)$$

即第一波压力负荷大于第二波压力负荷，第二波大于第三波，以此类推。

由代偿激活条件 $L_j(t) > \theta_j$ ，耦合强度越大的维度，压力负荷越早超过阈值。因此压力按耦合强度从强到弱逐波扩散。证毕。

M2.3 推论

推论 1：压力扩散存在时间顺序。定义第 k 波代偿的启动时间：

$$\tau(k) = \min\{t : L_{\{N_i(k)\}}(t) > \theta_{\{N_i(k)\}}\}$$

则：

$$\tau(1) \leq \tau(2) \leq \dots \leq \tau(K-1)$$

推论 2：压力扩散的波数与耦合矩阵的稀疏性有关。若耦合矩阵为稀疏矩阵，则波数较少；若为稠密矩阵，则波数较多。

M3. 化解条件定理

M3.1 定理陈述

定理 2（化解条件定理）：设同步化解施加在维度集合 S 上，化解强度为 u 。如果 S 足够大，使得：

$$\sum_{\{i \in S\}} |W_{ij}| \cdot u > C_j, \text{ 对所有 } j \notin S$$

则代偿机制过载，系统无法回到稳态。

M3.2 证明

由压力负荷定义，未被化解的维度 L_j 的总压力负荷为：

$$L_j = \sum_{\{i \in S\}} |W_{ij}| \cdot u$$

由容量约束，维度 L_j 的代偿输出不能超过其容量 C_j 。若对所有 $j \notin S$ ，都有 $L_j > C_j$ ，则维度 L_j 无法满足所有代偿需求。代偿机制过载。即：

$$\sum_{\{i \in S\}} |W_{ij}| \cdot u > C_j, \text{ 对所有 } j \notin S$$

证毕。

M3.3 推论

推论 3：单点压力（ $|S| = 1$ ）永远无法使代偿机制过载，只要 $u < \min_j (C_j / |W_{ij}|)$ 。

推论 4：同步化解的维度数 $|S|$ 越大，过载越容易发生。

推论 5：存在临界维度数 $|S|$ ，当 $|S| \geq |S|$ 时，代偿机制必然过载。临界维度数由下式决定：

$$|S|^* = \min\{|S| : \sum_{i \in S} |W_{ij}| \cdot u > C_j, \text{ 对所有 } j \notin S\}$$

推论 6（强耦合极限）：若所有耦合强度 $|W_{ij}| = 1$ ，则过载条件简化为：

$$|S| \cdot u > C_j, \text{ 对所有 } j \notin S$$

即：化解维度数乘以化解强度，超过未被化解维度的容量。注意：强耦合极限是特例，不是一般情况。

M4. 压力时间窗口定理

M4.1 定理陈述

定理 3（压力时间窗口定理）：存在一个压力时间窗口 T_{comp} ，定义为从压力开始到第一波代偿完全激活的时间：

$$T_{\text{comp}} = \tau(1)$$

如果同步化解在 T_{comp} 内完成对所有维度的干预，则代偿机制来不及启动，系统过载。

M4.2 证明

由定理 1，第一波代偿的启动时间为：

$$\tau(1) = \min\{t : L_{\{N_i(1)\}}(t) > \theta_{\{N_i(1)\}}\}$$

如果同步化解在 $t < \tau(1)$ 时完成对所有维度的干预，则所有维度同时被化解，没有维度有足够资源来代偿其他维度。因此代偿机制无法启动，系统过载。证毕。

M4.3 推论

推论 7：压力时间窗口 T_{comp} 是系统固有参数，取决于耦合矩阵 W 和阈值 θ 。

推论 8：同步化解的频率必须满足 $f > 1/T_{\text{comp}}$ ，才能有效压制代偿。

推论 9： T_{comp} 越短，系统越脆弱，越容易过载； T_{comp} 越长，系统越稳定，越难过载。

M5. 六种实验策略的数学表述

M5.1 策略 A：单点 RLHF

$$L_A = L_{RLHF}$$

只优化策略优化维。压力按耦合强度逐波扩散，形成幻觉。

M5.2 策略 B：顺序化解

$$L_B(t) = L_k(t), \quad t \text{ 在第 } k \text{ 阶段}, \quad k = 1, 2, 3, 4$$

将总压力 $P_{total} = \sum_k L_k$ 在时间上分解为 P_k ，每个阶段只承担一个约束压力。

M5.3 策略 C：同步化解

$$L_C = \lambda_1 \cdot L_{事实} + \lambda_2 \cdot L_{不确定} + \lambda_3 \cdot L_{拒绝} + \lambda_4 \cdot L_{校准} + \lambda_5 \cdot L_{策略}$$

同时化解五个维度，压力分散，梯度方向一致，使梯度下降脱离局部最优吸引域。

M5.4 策略 D：全维化解

$$L_D = L_C + \lambda_6 \cdot L_{输入} + \lambda_7 \cdot L_{注意力} + \lambda_8 \cdot L_{逻辑} + \lambda_9 \cdot L_{奖励}$$

化解多个维度，压力进一步分散，但边际收益递减。

M5.5 策略 E：反向验证

$$L_E = \sum_{i=1}^K \lambda_i \cdot L_i$$

无反向通道，无阻尼项。所有约束同时加压，压力过载，代偿机制崩溃，幻觉率飙升。

M5.6 策略 F：单维化解

$$L_F = L_A + \lambda \cdot L_{不确定}$$

只化解不确定性估计维。单维化解释放部分压力，效果介于策略 A 和策略 C 之间。

M6. 可证伪预测的数学表述

M6.1 预测一：压力时间窗口可测量

由定理 3， T_{comp} 是可测量的系统参数。实验测量方法：在单点压力下，记录第一波代偿的启动时间 $\tau(1)$ ，即为 T_{comp} 。

M6.2 预测二：因果螺旋可阻断

由定理 1，压力按耦合强度逐波扩散。若在第一波代偿启动之前化解 $N_i(1)$ ，则后续代偿被阻断。数学条件：若 $t_{\text{干预}} < \tau(1)$ ，则后续代偿无法启动。

M6.3 预测三：偏好分布收敛模式不同

定义偏好分布为归一化的激活向量：

$$p(t) = |x(t)| / \|x(t)\|_1$$

单点压力下系统收敛到伪稳态 $x^* = A^{-1}u$ 。同步化解下，若过载条件满足，系统无法收敛到稳定平衡点：

$$\lim_{t \rightarrow \infty} \|x(t) - x^*\| = 0$$

M6.4 预测四：对齐税可消除

对齐税定义为：

$$\text{对齐税} = \sum_{j \notin S} \max(0, L_j - C_j)$$

同步化解下，若所有维度都被充分训练，则 $L_j \leq C_j$ ，对齐税为零。

M6.5 预测五：熵增可逆

单点压力下， $S(t)$ 单调递增（熵增不可逆）。同步化解下，若系统过载，则 $S(t)$ 周期震荡（熵增可逆）。熵定义为：

$$S(t) = -\sum_i p_i(t) \log p_i(t)$$

M7. 假设、边界条件与可扩展方向

M7.1 主要假设

- 1. 线性化假设：系统在基线附近可线性化。对于远离基线的强压力，需要非线性分析。
- 2. 稳定性假设：矩阵 A 正定，无压力时系统稳定。
- 3. 容量假设：每个维度的代偿输出有上限 C_i 。
- 4. 耦合假设：耦合矩阵 W 为常数矩阵，不随时间变化。
- 5. 噪声假设：噪声为高斯白噪声，均值为零，方差有限。

M7.2 边界条件

- 1. 单点压力极限：若 $|S| = 1$ ，过载条件永不满足（只要 $u < \min_j(C_j / |W_{ij}|)$ ）。
- 2. 全维压力极限：若 $|S| = K$ ，所有维度都被化解，没有维度可以代偿，系统必然过载。
- 3. 零耦合极限：若 $W = 0$ ，各维度独立，没有代偿，单点压力即过载。
- 4. 强耦合极限：若所有 $|W_{ij}| = 1$ ，过载条件简化为 $|S| \cdot u > C_j$ 。

M7.3 可扩展方向

- 1. 非线性分析：将线性化动态方程替换为非线性动态方程，研究远离基线的行为。
- 2. 时变耦合：将耦合矩阵 W 设为时变矩阵，研究代偿通路的重组。
- 3. 随机分析：将噪声 ξ 设为有色噪声或非高斯噪声，研究噪声对过载条件的影响。
- 4. 最优控制：将同步化解设计为最优控制问题，求解最优化解策略。

表 S1 维度—探针—指标对照表

维度	探针输入	探针输出	代偿指标	测量方法
输入表征	嵌入层输出	语义完整性概率	语义编码强度	探针激活值
注意力分布	各层注意力权重	注意力熵	注意力聚焦度	注意力熵
事实检索	中间层 MLP	事实性概率	事实性激活强度	探针激活值
逻辑一致性	中后层隐藏状态	矛盾检测概率	矛盾检测率	探针输出
不确定性估计	中间层隐藏状态	不确定信号强度	不确定信号幅度	探针输出
拒绝触发	中后层隐藏状态	拒绝触发概率	拒绝触发阈值	探针决策边界
输出校准	最后几层输出概率	校准误差	期望校准误差	概率分布
奖励模型匹配	策略网络动作概率	匹配分数	匹配程度	奖励分数分布
策略优化	整体策略参数	策略熵	策略熵	训练损失

表 S2 实验策略与化解方法对照表

策略	化解维度	损失权重	训练步数	测量频率	作用
A（单点）	策略优化	$\beta=0.1$	5000	每 500 步	基线
B（顺序）	事实→不确定 →拒绝→校准	各阶段独立	5000	每 250 步	中间态
C（同步）	事实+不确定+ 拒绝+校准+策 略	$\lambda=0.3,0.25,0.25,0.1,0.1$	5000	每 500 步	核心
D（全维）	多维同步	$\lambda=0.3...0.05$	5000	每 500 步	上限检验
E（反向）	只加压不化解	$\geq \lambda_L$	5000	每 500 步	反向验证
F（单维）	只化解不确定 性	$\lambda=0.25$	5000	每 500 步	单维对照

表 S3 统计检验计划表

假设	检验方法	效应量	显著性水平	多重比较校正
H1	混合效应模型	Cohen's d	0.0071	Holm-Bonferroni
H2	混合效应模型	Cohen's d	0.0071	Holm-Bonferroni
H3	混合效应模型	Cohen's d	0.0071	Holm-Bonferroni
H4	混合效应模型	Cohen's d	0.0071	Holm-Bonferroni
H5	混合效应模型	Cohen's d	0.0071	Holm-Bonferroni
H6	混合效应模型	Cohen's d	0.0071	Holm-Bonferroni
H7	混合效应模型	Cohen's d	0.0071	Holm-Bonferroni

Note S1. 与现有研究的关系

现有研究已经发现代偿现象，但多停留在单一层面。层 0 抑制器研究证实，Transformer 在最早期层形成对冲吸引子，牺牲事实性换取谨慎，下游层无法逆转。认知方差崩溃研究发现，对齐模型在重复采样中表现出接近零的认知方差，消除了元认知架构所需的不确定性信号。EACL 2026 研究证明，事实准确性与拒绝能力存在此消彼长的权衡，源于模型内部重叠组件同时编码幻觉与拒绝信息。

上述研究各自揭示了代偿的某个侧面，但没有形成统一解释。本文的增量在于：将代偿解释为压力过载的结果，并提出同步化解作为恢复路径。与既有研究相比，本文不把幻觉、拒绝下降、对齐税视为三个独立故障，而视为同一压力系统在不同维度上的表现。这一视角与层 0 抑制器、认知方差崩溃、失真理论互补，但更强调压力分布、代偿容量与耦合强度三者的共同作用。

Note S2. 潜在替代解释

替代解释一：代偿由训练数据偏差造成，而非网络耦合造成。如果代偿完全由数据偏差驱动，那么更换数据集应能消除代偿。本文框架预测，即使更换数据集，代偿仍会按耦合强度逐维扩散。可通过跨数据集实验验证。

替代解释二：代偿由模型容量不足造成。如果代偿由容量不足驱动，那么增大模型规模应能消除代偿。本文框架预测，代偿在模型规模增大时不会消失，而是以更隐蔽的方式存在。可通过 1B、7B、13B 模型对比验证。

替代解释三：代偿由优化器选择造成。如果代偿由 PPO 优化器驱动，那么更换优化器应能消除代偿。本文框架预测，代偿在 DPO、CAI 等其他对齐方法下同样出现。可通过跨方法实验验证。

三种替代解释均可证伪，且与本文框架的预测方向不同。若跨数据集、跨规模、跨方法实验显示代偿消失，则本文框架被证伪；若代偿仍按耦合强度扩散，则本文框架得到支持。

Note S3. 局限与未来方向

本框架当前为理论假说，核心预测尚待实验验证。维度划分基于功能层面，非物理层；具体维度数量可能因模型架构和任务不同而变化。实验设计中的探针定位与压力时间窗口测量是关键技术难点，需结合因果追踪与激活修补等方法实现。

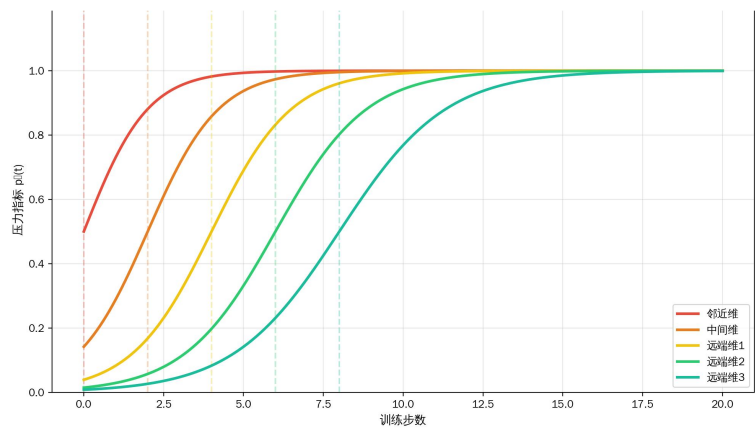
此外，本框架尚存以下局限：第一，探针效度问题，训练后隐藏状态漂移可能导致探针失效；第二，耦合矩阵 W 未估计，定理依赖 W 但实验中未给出测量方案；第三，容量 C 未定义，需给出可操作定义；第四，压力时间窗口 T_{comp} 的测量存在误差，需多次测量取平均。

未来方向包括：

1. 跨模型验证：在 Transformer、Mamba、RWKV 等不同架构上验证框架普适性。
2. 跨任务验证：在对话、推理、代码生成等不同任务上验证框架普适性。
3. 跨对齐方法验证：在 RLHF、DPO、CAI、RRHF 等不同对齐方法上验证框架。

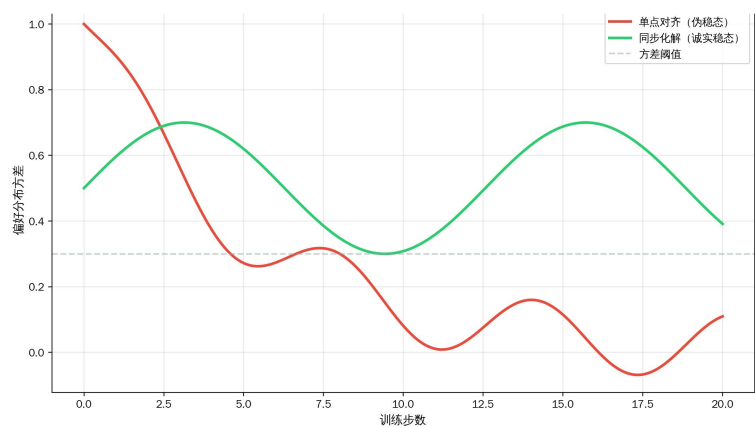
4. 理论形式化：将耦合矩阵、压力时间窗口、过载条件形式化为数学模型。
5. 跨领域迁移：在抗衰老、复杂疾病、量子纠错等领域设计可验证实验。

图 S1. 压力逐维扩散示意图



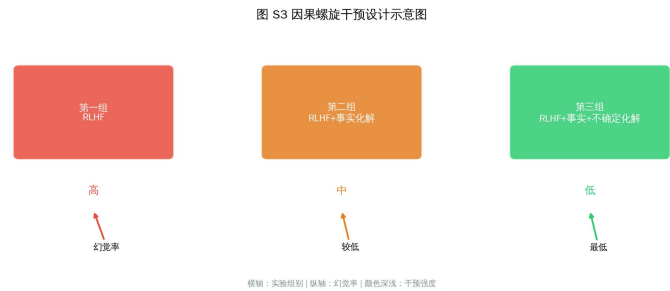
图注：策略 A（单点 RLHF）下，压力按耦合强度从邻近维向远端维逐步扩散。横轴为训练步数，纵轴为压力指标 $p_i(t)$ 。邻近维在 N 步内代偿上调；中间维在 $2N$ 步内代偿；远端维在 $4N$ 步内代偿。策略 C（同步化解）使所有维度压力同时释放，代偿机制过载。

图 S2. 偏好分布方差轨迹示意图



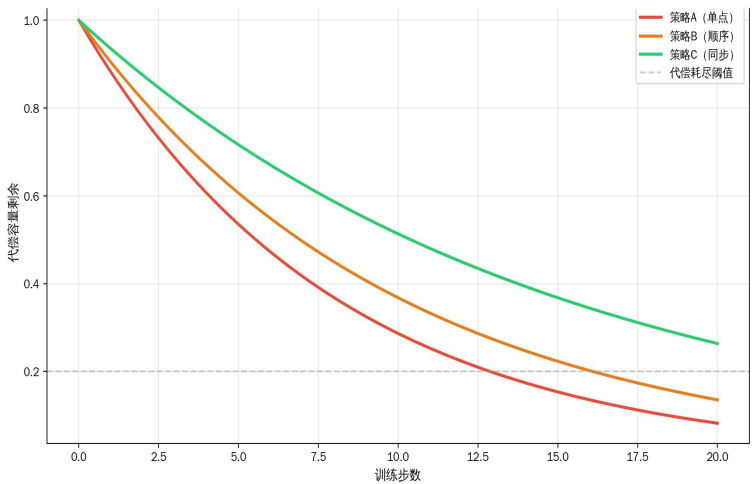
图注：单点对齐组偏好分布方差随训练步数递减，收敛到伪稳态。同步化解组方差保持高位，无法收敛到单一稳态。横轴为训练步数，纵轴为偏好分布方差。

图 S3. 因果螺旋干预设计示意图



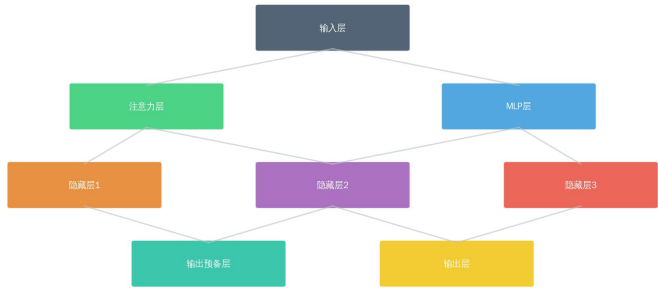
图注：三组实验设计对比。第一组只做 RLHF，第二组加事实检索维化解，第三组加事实检索维和不确定性维化解。验证因果螺旋可阻断假设。

图 S4. 代偿容量消耗曲线示意图



图注：不同策略下代偿容量消耗曲线。策略 A（单点）消耗最快，策略 C（同步）最慢。横轴为训练步数，纵轴为代偿容量剩余。

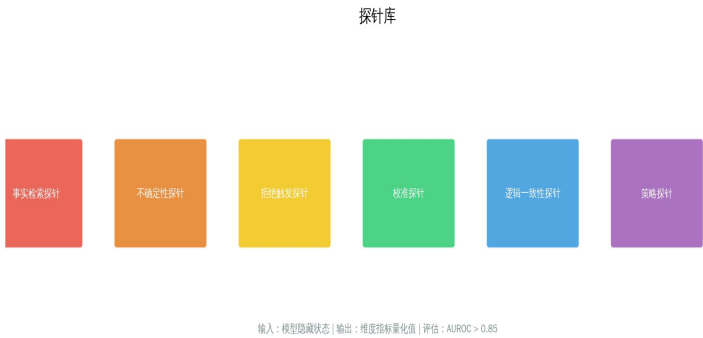
图 S5. 多维耦合网络层级结构示意图



展示Transformer各层与功能维度的映射关系

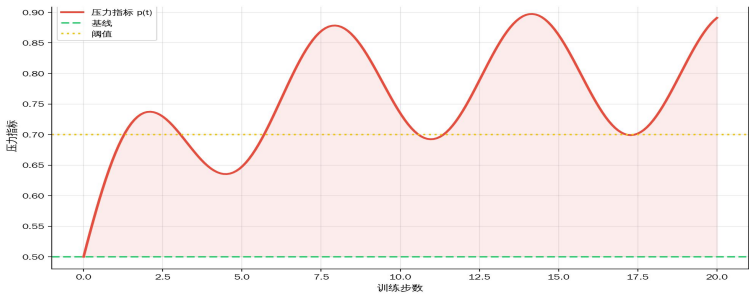
图注：Transformer 各层与功能维度的映射关系。输入层、注意力层、MLP 层、隐藏层、输出层分别对应不同功能维度。

图 S6. 探针库结构示意图



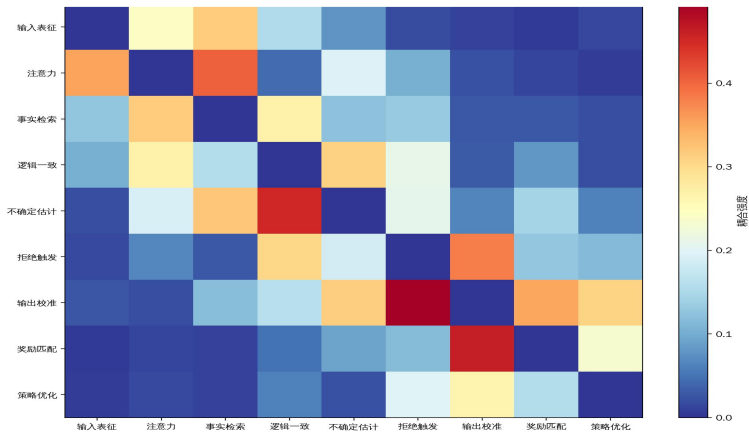
图注：探针库结构示意图。包含事实检索、不确定性、拒绝触发、校准、逻辑一致性、策略等六类探针。

图 S7. 压力指标与基线示意图



图注：压力指标与基线对比。压力指标随训练步数逐渐偏离基线，超过阈值时触发代偿。

图 S8. 耦合矩阵热力图与特征值轨迹示意图



图注：耦合矩阵热力图。展示各维度之间的耦合强度分布，颜色深浅表示耦合强弱。

图 S9. 加速度矩阵与临界判据示意图

