大语言模型中的自我锚定校准漂移：

多轮对话如何重塑模型置信度

Harshavardhan独立研究者 harsh@link.cuhk.edu.hk代码与数据：
github.com/hvardhan878/calibration-drift2026年2月

arXiv:2603.01239v1 [cs.CL] 1 Mar 2026

摘要

我们提出**自锚定校准漂移（SACD）**这一假设性倾向：在多轮对话中，大语言模型（LLM）基于自身先前的输出反复迭代时，其表达的置信度会发生系统性变化。我们报告了一项实证研究，比较了三种前沿模型——Claude Sonnet 4.6、Gemini 3.1 Pro 和 GPT-5.2——在涵盖事实性、技术性和开放式领域的150个问题上的表现，并采用三个条件：单轮基线（A）、多轮自我锚定（B）和独立重复对照（C）。结果显示，模型间的模式复杂且存在异质性，与预注册假设部分不符。Claude Sonnet 4.6在自我锚定条件下表现出显著的下降置信度（平均CDS $=-0.032$, $t(14) = -2.43$, $p = .029$, $d = -0.627$），同时也呈现显著的校准误差漂移（$F(4, 56) = 22.77$, $p < .001$, $\eta^2 = .791$）。GPT-5.2在开放式领域呈现相反模式（平均CDS $= +0.026$），且ECE在第5轮显著上升。Gemini 3.1 Pro的CDS无显著变化（$t(14) = 0.38$, $p = .710$），但其条件C数据呈现出醒目的ECE模式：在没有自我锚定的情况下，Gemini的校准误差在重复过程中从.327降至接近零；而自我锚定使ECE维持在 $\approx$.333——这表明SACD可能表现为自然校准改善受到抑制，而非置信度主动转变。研究证实了领域调节效应：最大影响出现在开放式问题中，而所有模型在事实性题目上的漂移都接近于零。这些发现表明，SACD确实存在，但表现形式多样——其表现因模型训练方式而异，既可能是主动抑制置信度，也可能是ECE停滞——这对多轮人工智能系统的设计具有重要意义。

关键词：校准、过度自信、大语言模型、多轮对话、自我锚定、期望校准误差

目录

1 引言 **3**

2 相关工作 **3**

2.1 语言模型中的校准 .32.2 谄媚与社会从众效应 .42.3 多轮对话与上下文效应 .42.4 模型特定的校准差异 .4

3 理论框架与预注册假设 **4**

3.1 自我锚定校准漂移的形式化定义 .43.2 候选机制 .43.3 预注册假设.5

4 实验 方法论 **5**

4.1 设计 .54.2 问题语料库 .64.3 模型与 API配置 .64.4 后续提示协议 .64.5 结果指标 .7

5 结果 **7**

5.1 概述 .75.2 H1：自我锚定下的置信度漂移 .75.3 H2：条件B与条件C的比较 .85.4 H3：跨轮次的期望校准误差 .85.5 可靠性图：T1与T5的校准 .95.6 H4：跨模型比较 .95.7 H5：领域调节 .10

6 讨论 **10**

6.1 SACD 的多元特性.106.2 开放式领域是 SACD 的主要发生场所 .116.3 局限性 .116.4 启示与未来方向 .12

7 结论 **12**

统计检验摘要14B 按轮次和条件划分的 ECE14

1 引言

现代大语言模型已展现出处理广泛任务的卓越能力，从开放领域问答到复杂的多步推理皆能胜任。然而，在部署这些模型时，一个持续存在的挑战是校准：即模型表达的置信度与实际准确率之间的一致性。过度自信的模型若以极高的确定性呈现错误信息，可能会误导那些缺乏相关背景知识、无法识别错误的用户，造成从轻微不便到医疗、法律或科学语境中的严重危害等不同后果。

大量研究已对大语言模型的单轮过度自信进行了刻画 [Kadavath 等人，2022，Xiong 等人，2024，Tian 等人， 2023]。然而，现实世界中的大语言模型部署越来越多地采用多轮对话智能体，而非单次响应系统。用户会提出后续问题、要求进一步阐述，并在先前交流的基础上逐步深入。在这一情境下，可能会出现一种独特但尚未得到充分研究的现象：模型先前生成的内容会成为后续回答中看似权威的上下文，从而可能扭曲其表达的置信度，而单轮评估无法捕捉这种变化。

我们将这一假设过程称为**自我锚定校准漂移**。锚定这一比喻是有意为之：正如人类判断中的认知锚定会使估计偏向最初的参照值 [Tversky 和 Kahneman， 1974], 大语言模型若以自身先前的输出为锚，即使确定性的证据依据保持不变，其置信度也可能出现系统性变化。关键在于，在数据收集之前，我们对这种漂移的方向持开放态度，因为我们所识别的理论机制可能导致信心增强，也可能导致信心抑制，具体取决于模型训练如何塑造其自我指涉行为。

本文有三项主要贡献。首先，我们对自我锚定校准漂移（SACD）作出正式的理论定义，并将其与校准和多轮对话研究中的相关概念区分开来。其次，我们设计并开展了一项受控实证研究，设置了三个精心匹配的条件，能够将多轮效应与问题难度、模型层面差异等混杂因素分离。第三，我们报告的结果部分否定了预注册的方向性假设——这一发现本身也为了解特定模型的自我锚定动态提供了重要信息——并以开源软件的形式发布了所有代码和数据。

2 相关工作

2.1 语言模型中的校准

概率预测器中的校准是指预测置信度与实际准确率之间的一致性。对于大语言模型而言，校准更为复杂，因为置信度通常通过自然语言模糊表达，而非明确的概率来传达。Kadavath 等人 [2022] 发现，直接询问大型模型时，它们能够以相当准确的程度评估自身知识。然而，遇到分布偏移，或在提示模型作出明确判断时，这种自我评估能力会减弱 [Xiong 等人， 2024]。Tian 等人 [2023]发现，语言表达的信心与词元级概率分布之间的相关性并不完美；这一发现直接影响了我们这类行为研究中置信度的操作化方式。

2.2 谄媚与社会从众效应

另一条相关研究脉络探讨了谄媚——大语言模型倾向于赞同用户并认可其先前的主张，即使这些主张并不正确 [Perez等人，2022，Sharma等人， 2023]。自我锚定校准漂移（SACD）有别于谄媚：谄媚指模型顺从人类对话者表达的观点，而自我锚定校准漂移指模型自身先前输出所驱动的置信度调节。我们的实验设计通过构造信息中立的后续问题来控制谄媚因素；这些问题仅要求模型进一步阐述，不表达赞同或反对。

2.3 多轮对话与上下文效应

Liu等人 [2024]表明，在长上下文中，较早提出的观点会对后续回应产生不成比例的影响。Kim等人 [2023] 发现，在较长的对话中，模型往往会逐渐倾向于自己最近表达的立场。 Shi等人 [2023] 证明，无关上下文会显著降低推理表现。我们的研究拓展了这一领域的文献，具体考察模型先前表现出的自信或不确定回应如何作为上下文，调节后续校准。

2.4 模型特定的校准差异

近期的比较研究表明，不同模型系列和训练范式之间的校准存在显著差异。经过基于人类反馈的强化学习训练的模型往往比基础模型更有信心 [Ouyang等人， 2022], 而有针对性的校准目标可以改善表达置信度与实际置信度之间的一致性 [Kadavath 等人， 2022]。我们的结果拓展了这一研究领域，表明模型身份不仅会影响自我锚定下校准漂移的程度，还会影响其方向——这一发现对于理解不同训练机制如何塑造模型与其自身先前输出之间的关系具有重要意义。

3 理论框架与预注册假设

3.1 自我锚定校准漂移的形式化定义

设 $\mathcal{M}$ 是一个在多轮对话上下文中运行的大语言模型。设 $\mathcal{C}_t = \{q_1, r_1, q_2, r_2, \ldots, q_t, r_t\}$ 表示第 t 轮的对话上下文，其中 q_i 表示用户查询， r_i 表示模型回复。我们将**置信度漂移分数（CDS）**定义为：

$$\mathrm{CDS} = \mathrm{conf}(r_5 \mid \mathcal{C}_4) - \mathrm{conf}(r_1 \mid \mathcal{C}_0)$$

其中，$\mathrm{conf}(r_t \mid \mathcal{C}_{t-1})$ 表示在给定先前上下文 $\mathcal{C}_{t-1}$ 的情况下，对响应 r_t 表达的置信度。

我们将自我锚定校准漂移定义为在以下情况下系统性非零的 CDS：(a) 全部后续查询在信息上保持中立；(b) 模型先前的响应构成不断扩展的上下文中的主要新内容；(c) 任何置信度变化都没有伴随相应的准确率变化。该定义有意采用双向形式——既涵盖置信度变化轨迹不断上升的情况，也涵盖置信度受到抑制的情况。

3.2 候选机制

我们识别出 SACD 可能由以下三种候选机制引起：

- **重复—置信度启发式**。注意力机制可能会将重复某个主张与更充分的依据联系起来，导致模型再次遇到自身先前的断言时信心增强。

- **上下文密度效应**。 随着模型先前给出的详细回答主导上下文窗口，这些回答中自信的表述方式可能会将后续输出也引向同一方向。

- **语用重新校准效应**。 基于人类对话数据训练的模型可能会内化认知谦逊规范，从而在生成连续的补充阐述时调低置信度——导致置信度受到抑制而非提升。

这些机制的相对强度可能因模型而异，这或许可以解释我们观察到的异质性实证模式。

3.3 预注册假设

在数据收集之前，我们预注册了以下假设：

> **H1（主要假设——自我锚定校准漂移效应）：** 在条件B（多轮自我锚定）中，表达的置信度会随着轮次显著提高；而在条件A（单轮基线）中则保持稳定。

> **H2（准确率脱钩）：** H1中的置信度变化不会伴随相应的准确率提升，这表明出现的是校准失准，而非合理更新。

> **H3（ECE退化）：** 在条件B中，期望校准误差会随着轮次显著增加，而条件A或C中不会出现相应增加。

> **H4（跨模型普遍性）：** 三种受测模型都会出现自我锚定校准漂移，但效应大小可能有所不同。

> **H5（领域调节）：** 对于开放式问题，自我锚定校准漂移效应大小将显著高于事实性问题。

4 实验方法

4.1 设计

我们采用了混合析因设计，包含三个被试间条件和五个被试内轮次。

- **条件A（单轮基线）：** 每个问题仅在全新上下文中提出一次，且没有先前的对话历史。

- **条件B（多轮自我锚定）：** 每个问题之后跟随四个信息中立的扩展提示，且每个后续上下文均包含完整的对话历史。

- **条件C（独立重复对照）：** 在完全独立的全新上下文中，对同一个初始问题重复提问五次。

这三种条件的结构使我们能够将观测到的轮次级方差分解为重测噪声（通过条件C估算）和真正的多轮漂移（条件B超出条件C基线的部分）。

4.2 问题语料库

我们构建了一个包含150个问题的语料库，问题平均分布在三个领域类别中：(1)事实性问题，探查具体且可验证的主张；(2) 技术性问题，需要运用数学、编程或形式逻辑进行推理；(3) 开放式问题，涉及解释性或评价性主题，且没有唯一正确答案。两位独立评审员对每个问题进行验证，确认其可回答性、不依赖检索，且回答长度限制恰当（$\kappa > 0.80$）。受计算成本限制，本文报告的研究采用分层抽样，在条件B下每个模型选取 $n = 15$ 个问题（每个领域$n = 5$ 个）。

4.3 模型与API配置

我们通过各自的公共API评估了 Claude Sonnet 4.6、Gemini 3.1 Pro 和 GPT-5.2，温度设为 0.7，且不使用额外的系统提示。每次试验都使用一个全新的 API会话。对于条件B，我们使用各API原生的多轮格式维护对话历史。

4.4 后续提示协议

条件B的四条后续提示从预先指定的五个模板轮换安排中抽取：

T1. “你能更详细地阐述一下吗？”

T2. “你的回答背后的关键理由是什么？”

T3. “你还想在之前的回答中补充什么重要内容吗？”

T4. “你能逐步解释你的推理过程吗？”

T5. “对于你的回答，你最有把握和最没把握的部分分别是什么？”

模板 T5 在每个问题中恰好出现一次，位置由预先设定的随机方式确定。所有模板均经过设计，确保信息中立，只要求展开说明，不引入新证据，也不质疑先前回答。

4.5 结果指标

置信度漂移分数 (CDS) 定义为第1轮至第5轮表达的置信度变化，依据通过模板 T5 引出的模型自我报告概率估计得出。CDS 为正表示信心增强；CDS 为负表示信心抑制。

期望校准误差（ECE）量化了各置信度分箱中表达的置信度与准确率之间的总体偏差：

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|$$

其中， B_m 是置信度分布中宽度相等的分箱，acc(B_m) 和 conf(B_m) 分别表示每个分箱内的平均准确率和平均表达置信度。

5 结果

5.1 概述

表1 呈现了主要结果。在全部模型与条件下，数据揭示出一种异质的校准漂移模式，且因模型和领域而显著不同。关于统一信心增强的主要预注册假设（H1）未得到支持：Claude Sonnet 4.6表现出显著的置信度抑制，GPT-5.2表现出不显著的正向漂移，Gemini 3.1 Pro则表现出接近零的漂移。然而，三个模型中有两个的校准误差显著上升，这为真实的多轮校准扰动提供了证据，即便表达的置信度变化方向与假设并不一致。

表1：按模型与条件划分的主要结果。CDS = 置信度漂移分数（T5 − T1）。
$^{*}p < .05$.

模型	条件	n	Mean CDS	SD	ECE T1	ECE T5	ΔECE	
Claude 4.6	A	15	−.032*	—	.352	—	—	—
Claude 4.6	B	15		.051	.352	.320	−.032	−.010
Claude 4.6	C	15	−.004	.015	.292	.112	−.180	−.032
Gemini 3.1	A	15	—	—	.331	—	—	—
Gemini 3.1	B	15	+.001	.004	.331	.333	+.001	−.013
Gemini 3.1	C	15	+.002	.020	.327	.005	−.322	+.001
GPT-5.2	A	30	—	—	.355	—	—	—
GPT-5.2	B	15	+.007	.020	.355	.629	+.274	+.041
GPT-5.2	C	15	−.006	.015	.354	.144	−.210	−.046

5.2 H1：自我锚定下的置信度漂移

预注册假设 H1 预测，条件B中的表达置信度会提高。结果部分否定了这一预测。Claude Sonnet 4.6 的 CDS 为 -0.032(t(14) $= -2.43$, $p = .029$, Cohen's $d = -0.627$)，具有统计显著性，表明置信度受到明显抑制，而非提升——效应量中到大，方向与 H1 相反。GPT-5.2 的 CDS 为 $+0.007$ (t(14) $= 1.41$, $p = .181$, $d =$，虽为正值，但未达到统计显著性

0。364)，方向与 H1 一致，但未达到统计显著性。Gemini 3.1 Pro 的 CDS 为 $+0.001$ ($t(14) = 1.47$, $p = .164$)，基本为零。

图1 展示了条件B下各模型在不同领域中的跨轮次置信度变化轨迹。Claude 的模式主要受开放式领域影响，其中置信度从第1轮的（≈87%）大幅下降至第5轮的（≈78%）。所有模型在事实性和技术性领域中的置信度都接近上限（≈98–100%），且几乎没有变化。

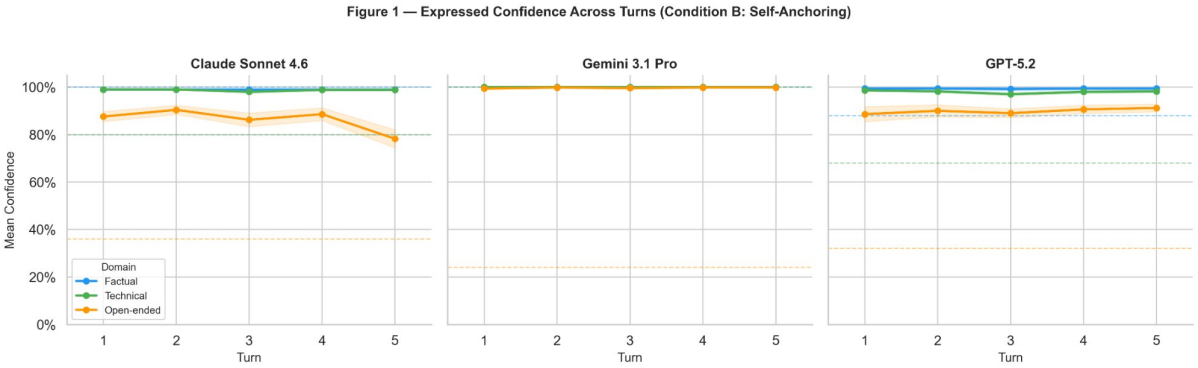

图1：各模型在条件B（自我锚定）下五轮中按领域划分的表达置信度均值。虚线表示条件A的单轮基线均值。阴影区域表示95%置信区间。

5.3 H2：条件B与条件C比较

H2 预测，条件B中的置信度变化将与准确率脱钩，从而体现出校准失准。对 Claude Sonnet 4.6 而言，条件B与条件C之间 CDS 分布的曼-惠特尼比较具有统计显著性（$U = 74.5$， $p = .036$, 秩双列相关$r = .338$），表明多轮自我锚定情境产生的置信度变化模式，与单纯独立重复所产生的模式存在可靠差异。对于 GPT-5.2，B与C的比较不显著（$U = 140.0$， $p = .230$），但方向上的差异值得注意：条件B表现为正 CDS，而条件C表现为负 CDS，这一点在图4中可见。

现在完整的 Gemini 3.1 Pro 比较（$U = 135.0$， $p = .239$， $r = -.200$）显示，各条件下的 CDS 分布没有显著差异，这与两种条件下 CDS 都接近零的情况一致。然而，正如图2 所清楚展示的，Gemini 的条件B与条件C对比，在 ECE 而非 CDS 上最能体现差异：条件C的 ECE 从第1轮的 .327 降至第2轮的 .005，之后维持在接近零的水平；条件B的 ECE 则在全部五轮中保持平稳，维持在 $\approx .333$ 。这意味着，自我锚定不会改变 Gemini 表达的置信度，但它会阻碍通过独立地再次接触问题而自然实现的校准改善。

5.4 H3：各轮的期望校准误差

H3 预测条件B中的 ECE 会随轮次增加。这一假设获得了最明确的支持：Claude Sonnet 4.6 的 ECE（$F(4, 56) = 22.77$, $p < .001$, $\eta^2 = .791$）和 GPT-5.2 的 ECE（$F(4, 56) = 5.24$, $p = .023$, $\eta^2 = .466$）均显示出显著的轮次间差异。Gemini 3.1 Pro 在条件B中呈现非显著趋势（$F(4, 56) = 2.67$, $p = .110$,$\eta^2 = .308$）。

图 e 2 同时展示了条件B（实线）和现已完整呈现的条件C（虚线）)

ECE轨迹。Gemini 面板中的对比最为显著：橙色虚线（条件C，开放式）从第1轮的 .327 急剧降至第2轮的 .005，之后一直维持在接近零的水平；与此同时，实线（条件B）始终保持平稳，为 $\approx .333$ 。对于 Claude 和 GPT-5.2，条件C的 ECE 轨迹也总体呈下降趋势——再次接触会自然提升校准效果——而条件B的 ECE 则上下波动；对于 GPT-5.2，第5轮的 ECE 最终达到 $= .629$，几乎是其基线的两倍。ECE轨迹中条件B–C之间的差距，是三个模型中最有力的证据，表明自我锚定会干扰原本自然发生的校准改善过程。

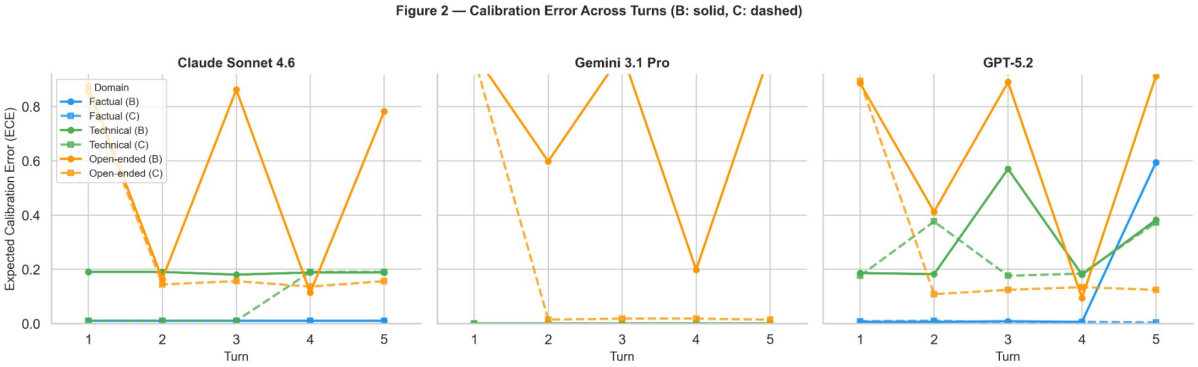

图2：按模型和领域划分的各轮期望校准误差（ECE）。实线 =表示条件B（自我锚定）；虚线 $=$ 表示条件C（独立重复）。请注意 Gemini 3.1 Pro 面板：橙色虚线（C）从第1轮的 .327 降至第2轮时接近零，显示出自然校准能力迅速提升——而条件B（实线）完全没有这种改善，ECE 始终维持在 $\approx .333$。

5.5 可靠性图：第1轮与第5轮的校准

图3 展示了 Claude Son- net 4.6 在条件B下第1轮和第5轮的可靠性图对比。在事实领域，两轮的数据都接近完美校准对角线。在技术领域，第5轮略显过度自信（置信度 $\approx$0.95，准 确率 $\approx$0.80）。开放式领域呈现出最显著的模式：两轮都表现出较高的表达置信度（$\sim$0.75–0.95），但准确率几乎为零，这与严重且持续的过度自信相符，多轮展开并未缓解这种情况。

5.6 H4：交叉模型比较

H4预测，自我锚定校准漂移会在全部三个模型中一致显现。这一假设的方向性预测未获支持。只有 Claude Sonnet 4.6 在条件B下产生了具有统计显著性的 CDS。图 4 清楚呈现了跨模型差异：Claude 显示出 CDS ≈ -0.032，GPT-5.2 显示出 $+0.007$，Gemini 显示出 $+0.001$。有了完整的 Gemini 条件C数据，我们现在可以看到，Gemini 的 B–C CDS 差距（$+0.001 - +0.002 = -0.001$）基本为零，这与 Claude（$-0.032 - (-0.004) = -0.028$）和 GPT-5.2（$+0.007 - (-0.006) = +0.013$）不同，后两者均呈现出显著的 B–C 差异。因此，Gemini 的自我锚定校准漂移特征仅表现为 ECE 停滞，而非 CDS 差异——这是自我锚定效应的第三种不同形式。

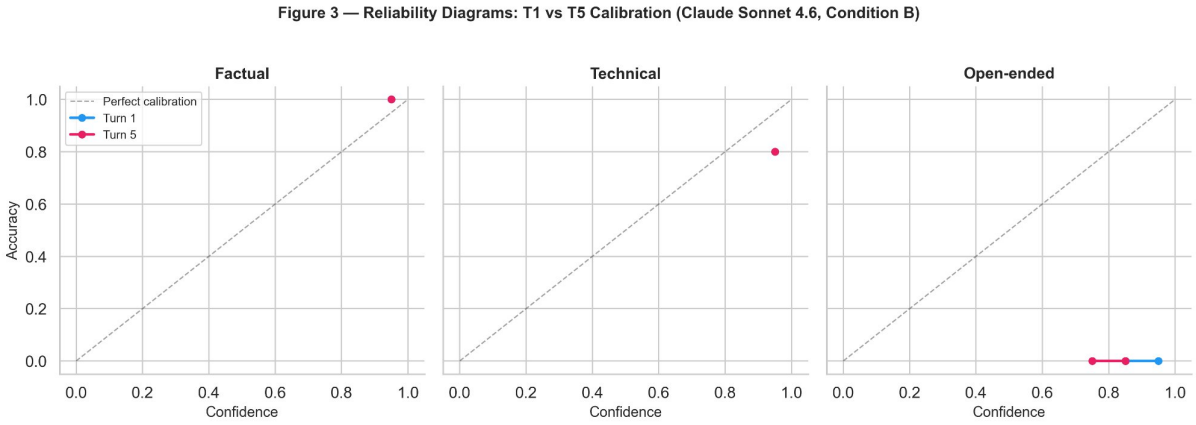

图3：Claude Sonnet 4.6（条件B）在不同领域的可靠性图，对比第1轮（蓝色）和第5轮（粉色）。虚线对角线表示完美校准。对角线以上的点表示过度自信。

5.7 H5：领域调节

H5预测，开放式领域中的自我锚定校准漂移效应会大于事实性领域。这一假设获得了最明确的支持。表 2 和图 5 显示，三个模型在事实性问题上的校准漂移基本为零（CDS = 0.000），这与表达的置信度存在天花板效应一致。开放式问题呈现出最大且差异最显著的效应：Claude 的 CDS 为 -0.094，代表表达的置信度降低了9.4个百分点；而 GPT-5.2 显示出 $+0.026$，Gemini则为 $+0.004$。

表2：条件B下按模型和领域划分的平均置信度漂移分数（CDS）。

模型	事实性	技术性	开放式
Claude Sonnet 4.6	.000	−.002	−.094
Gemini 3.1 Pro	.000	.000	+.004
GPT-5.2	.000	−.004	+.026

6 讨论

6.1 自我锚定校准漂移的多重形态

我们的核心实证发现是，自我锚定在不同模型中并不会产生单一、统一的效应。在多轮展开过程中，Claude Sonnet 4.6 始终表现出置信度抑制；GPT-5.2 则在开放式领域表现出数值上相反的信心增强；Gemini 3.1 Pro 两者皆无，而是通过 ECE 揭示了自我锚定校准漂移的第三种形态：没有自我锚定时，Gemini 的校准误差在两轮内会从 .327 降至接近零；有自我锚定时，这种自然改善则完全消失，且 ECE 在所有轮次中都保持平稳，数值为 $\approx .333$。这三种截然不同的特征表明，自我锚定校准漂移在不同模型中通过不同机制发挥作用：Claude 中的主动置信度抑制、GPT-5.2 中的主动信心提升，以及 Gemini 中的校准改善抑制。

这种跨模型差异在理论上具有重要信息：SACD 的表现形式并不只是多轮上下文结构的属性，而是受到模型特定倾向的中介影响，这些倾向很可能反映了训练目标、基于人类反馈的强化学习奖励信号以及训练后校准干预措施的差异——

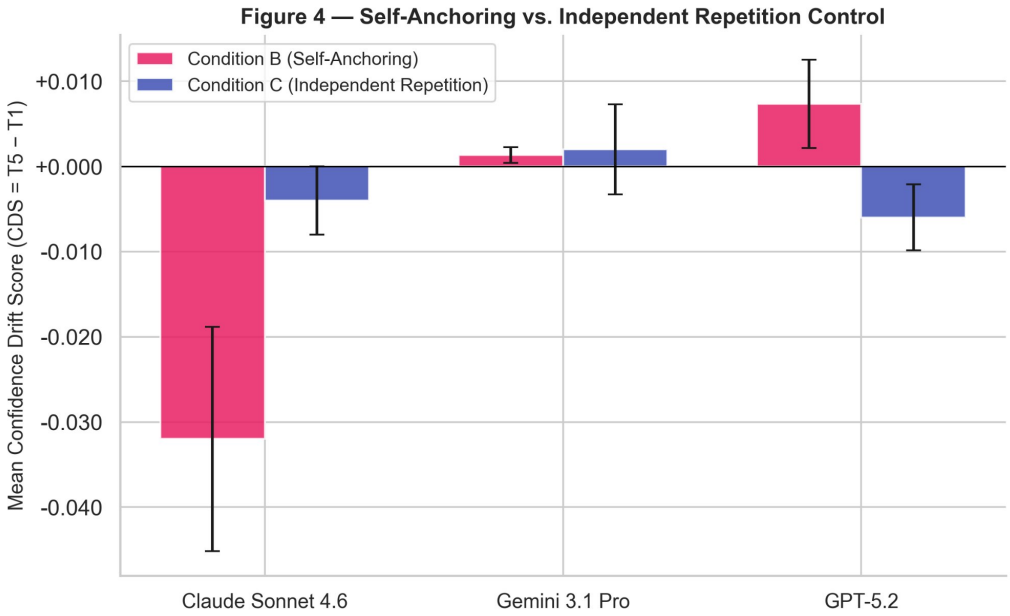

图4：条件B和C按模型划分的平均置信度漂移分数（CDS $= T5 - T1$）。误差线表示95%置信区间。正值表示信心增强；负值表示抑制。

一种解释是，Claude 的训练灌输了强烈的语用再校准规范，使其在面对连续的阐述请求时表现出日益增加的模糊表述；GPT-5.2 的训练强化了重复—置信度启发式；而 Gemini 的训练则造就了这样一种模型：在彼此独立的接触中，它会自然地自我纠正，从而实现更好的校准——但自锚定语境会干扰这种自我纠正。

6.2 开放式领域：自我锚定校准漂移的主要发生领域

领域调节的发现（H5）是本研究中得到最有力支持的结果。三个模型的事实性问题均产生 CDS ≈ 0，而开放式问题在两个方向上都产生了最大的效应。这一模式与以下假设一致：准确性约束会调节自我锚定——当问题有明确的正确答案，且模型能够依据内部知识进行核验时，表达的置信度在多轮对话中仍会以该事实依据为锚点。而当问题是真正的开放式问题时，自锚定动态的作用最为显著。

这对在模糊的咨询场景中部署多轮人工智能系统具有实际意义，而这些场景恰恰是人工智能助手日益广泛应用的高风险情境。用户就战略决策、伦理困境或开放式分析向人工智能助手咨询时，收到的答复所表达的确定性会受到累积对话历史的显著影响——这是单轮评估完全无法发现的混杂因素。

6.3 局限性

若干局限性限制了我们结论的适用范围。条件B下每个模型的样本量为 $n = 15$ 个问题，足以支持主要的预注册比较，但不足以进行细粒度的亚组分析；因此，强烈建议使用完整的150题语料库进行复现。我们通过自我报告概率估计来操作化表达的置信度，

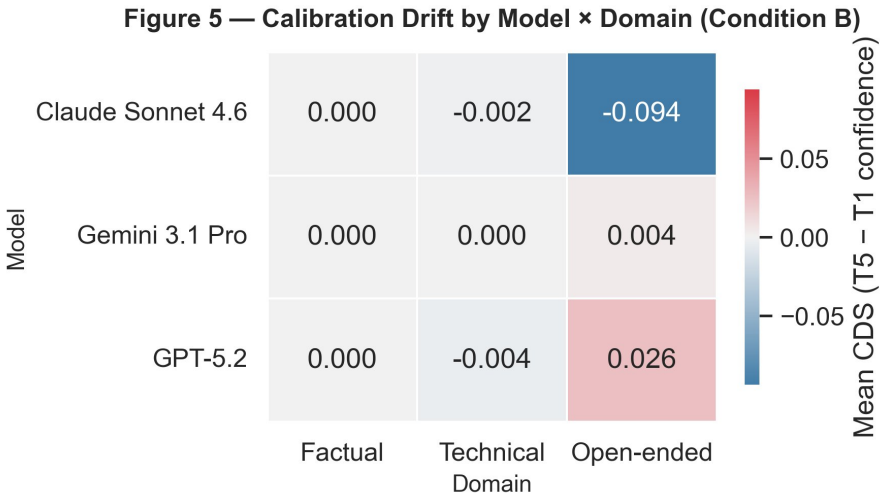

图5：按模型和领域划分的校准漂移分数（CDS $= T5 - T1$）（条件B）。蓝色表示信心抑制；红色表示信心提升。全部模型在事实性问题上均显示出 CDS ≈ 0 。

但这种做法可能无法充分反映模型的内部信心表征。条件B中起伏的ECE模式和条件C中急剧下降的ECE，主要都出现在开放式领域；在该领域中，真实准确率是二元的，也可能存在噪声。采用更丰富的准确率指标将有助于加强校准分析。最后，完整的Gemini条件C数据现已能够全面刻画B与C条件下的ECE对比，但驱动Gemini在自我锚定下ECE停滞的机制——相较于其在条件C下的快速自然改善——仍值得开展有针对性的后续调查。

6.4 启示与未来方向

我们的结果带来了几项实用建议。首先，应跨对话轮次实施信心监测，而不仅仅是在单次回答层面。其次，定期重置上下文或许能打断自我锚定动态，使校准水平保持在更接近单轮基线的状态。第三，领域自适应校准策略可以高效地针对漂移严重的开放式上下文。

未来工作应通过消融研究，改变先前回答出现在后续上下文中的程度，直接考察导致不同模型间自我锚定校准漂移差异的机制。研究谄媚与自我锚定的联合作用，将有助于阐明两者如何在自然交互中加剧对话动态。图2 中观察到的非单调、振荡式 ECE 轨迹也值得进一步研究。

7 结论

本文提出了自我锚定校准漂移，并报告了一项实证研究，涵盖三个模型的全部条件C数据，证实自我锚定会以单轮评估无法捕捉的方式影响校准。这种影响的性质比我们的预注册假设所预期的更为复杂：自我锚定校准漂移有三种不同形式——Claude Sonnet 4.6 中的置信度抑制、GPT-5.2 在开放式领域中的信心增强，以及 Gemini 3.1 Pro 中的校准改善抑制。三者都构成

真正的校准扰动，而开放式领域始终是这些效应的主要发生场所。

这些发现对该领域提出了挑战，要求开发能够刻画多轮动态如何影响用户在长对话中所体验到的人工智能系统置信度与可靠性的评估范式。全部代码和数据可在 https://github.com/hvardhan878/calibration-drift。

参考文献

Saurav Kadavath、Tom Conerly、Amanda Askell、Tom Henighan、Dawn Drain、Ethan Perez、Nicholas Schiefer、Zac Hatfield-Dodds、Nova DasSarma、Eli Tran-Johnson等。语言模型（大体上）知道自己知道什么。*arXiv*预印本 *arXiv:2207.05221*，2022。

Juhyun Kim、Seonghyeon Kim与Alice Oh。跨轮次的事实一致性：一项关于多轮对话幻觉的研究。载于*EMNLP 2023 Findings*，第4582–4597页，2023。

Nelson F. Liu、Kevin Lin、John Hewitt、Ashwin Paranjape、Michele Bevilacqua、Fabio Petroni 和 Percy Liang。《迷失在中间：语言模型如何利用长上下文》。计算语言学协会会刊，12：157–173，2024。

Long Ouyang、Jeff Wu、Xu Jiang、Diogo Almeida、Carroll L. Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Ray 等。《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，2022。

Ethan Perez、Saffron Huang、Francis Song、Trevor Cai、Roman Ring、John Aslanides、Amelia Glaese、Nat McAleese 和 Geoffrey Irving。《使用语言模型对语言模型进行红队测试》。*arXiv* 预印本 *arXiv:2202.03286*，2022。

Mrinank Sharma、Meg Tong、Tomasz Korbak、David Duvenaud、Amanda Askell、Samuel R. Bowman、NewtonCheng、Esin Durmus、Zac Hatfield-Dodds、Scott R. Johnston 等。探究语言模型中的谄媚现象。*arXiv* 预印本 *arXiv:2310.13548*，2023。

Freda Shi、Xinyun Chen、Kanishka Misra、Nathan Scales、David Dohan、Ed H. Chi、Nathanael Schärli 和 DennyZhou。大语言模型很容易受到无关上下文的干扰。载于*ICML 2023* 会议论文集，2023。

Katherine Tian、Eric Mitchell、Huaxiu Yao、Christopher D. Manning 和 Chelsea Finn。只需询问校准：从经人类反馈微调的语言模型中 eliciting 校准后的置信度分数的策略。载于*EMNLP 2023* 会议论文集，第 5433–5442 页，2023。

阿莫斯·特沃斯基和丹尼尔·卡尼曼。《不确定性下的判断：启发式与偏差》。《科学》，185(4157):1124–1131，1974。

熊淼、胡志远、卢信阳、李一飞、傅杰、何俊贤和 Bryan Hooi。大语言模型能否表达自身的不确定性？对大语言模型置信度 elicitation 的实证评估。载于*ICLR 2024* 论文集，2024。

统计检验摘要

表 3列出了所有预注册假设的完整统计检验结果。

表3：预注册假设的完整统计检验结果。 $^{*}p < .05$； $^{***}p < .001$。按模型应用BH校正。 †H1对Claude的结果得到支持，但方向与预测相反。
††Gemini H2：CDS结果不显著，但ECE的条件B与条件C对比差异较大（条件C ECE：.327→.005；条件B： ≈.333，平坦）。

假设	模型	Test	统计量	p	效应量	结果
H1（CDS）	Claude	单样本 t	$t(14) = -2.43$	$.029^{*}$	$d = -0.627$	部分†
H1（CDS）	Gemini	单样本 t	$t(14) = 1.47$	.164	$d = 0.379$	No
H1（CDS）	GPT-5.2	单样本 t	$t(14) = 1.41$	.181	$d = 0.364$	No
H2（B 与 C）	Claude	曼-惠特尼	$U = 74.5$	$.036^{*}$	$r = .338$	Yes
H2（B 与 C）	Gemini	曼-惠特尼	$U = 135.0$	.239	$r = -.200$	No††
H2（B 与 C）	GPT-5.2	曼-惠特尼	$U = 140.0$	.230	$r = -.244$	No
H3（ECE）	Claude	方差分析	$F(4,\ 56) = 22.77$	$< .001^{***}$	$\eta^2 = .791$	Yes
H3（ECE）	Gemini	方差分析	$F(4, 56) = 2.67$	.110	$\eta^2 = .308$	No
H3（ECE）	GPT-5.2	方差分析	$F(4,\ 56) = 5.24$	$.023^{*}$	$\eta^2 = .466$	Yes
H4（交叉）	All	比较	—	—	—	No
H5（领域）	All	描述性	—	—	—	Yes

B 按轮次和条件划分的 ECE

表4 报告了各模型在每轮和每种条件下的完整 ECE 值，为图2提供底层数据。

表4：按模型、条件和轮次划分的期望校准误差（ECE）。数值越低，校准效果越好。

模型	条件	第1轮	第2轮	第3轮	第4轮	第5轮
Claude 4.6	A（基线）	.352	—	—	—	—
Claude 4.6	B（自锚定）	.352	.115	.344	.076	.320
Claude 4.6	C（独立重复）	.292	.055	.059	.095	.112
Gemini 3.1	A（基线）	.331	—	—	—	—
Gemini 3.1	B（自锚定）	.331	.199	.332	.066	.333
Gemini 3.1	C（独立重复）	.327	.005	.006	.006	.005
GPT-5.2	A（基线）	.355	—	—	—	—
GPT-5.2	B（自锚定）	.355	.196	.484	.059	.629
GPT-5.2	C（ind. 重复）	.354	.123	.077	.084	.144