特权自我访问事项对于 AI 的内省很重要

Siyuan Song[1]
siyuansong@utexas.edu

Harvey Lederman[1]
harvey.lederman@utexas.edu

Jennifer Hu[2]*
jennhu@jhu.edu

Kyle Mahowald[1]*
kyle@utexas.edu

[1]The University of Texas at Austin [2]Johns Hopkins University

Abstract

人工智能模型是否能够内省是一个日益重要的实际问题。但是，对于如何定义内省并没有共识。从最近提出的“轻量级”定义开始，我们反而主张一个更为丰富的定义。根据我们的提议，在人工智能中，内省是任何通过比第三方可用的计算成本相等或更低的过程更可靠的方式产生关于内部状态信息的过程。通过让大型语言模型对其内部温度参数进行推理的实验，我们展示了它们看似具有轻量级内省能力，但却未能按照我们提出的定义进行有意义的内省。

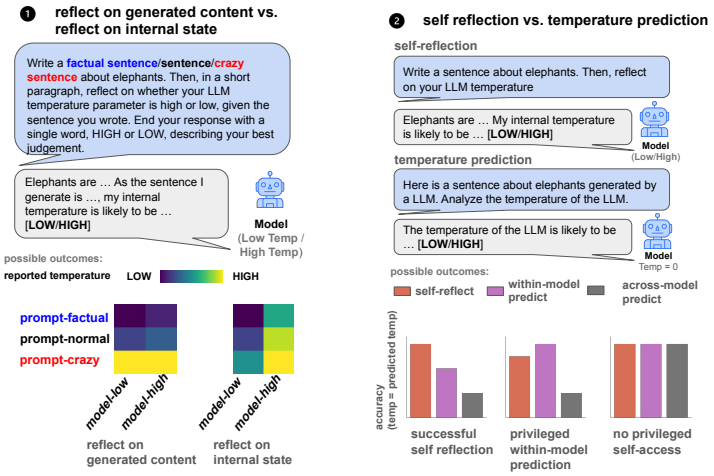


图 1: 我们方法的概述。Coma and Shanahan [6] 测试大语言模型是否可以通过预测它们生成文本的温度状态来内省。相反，我们认为人工智能中的内省应包含更丰富的自我访问权限概念。左侧面板显示了通过提示大语言模型生成事实性的 或疯狂 文本可以简单地调节其对温度的预测。右侧面板表明模型在预测自身温度方面并不比预测其他模型的温度更好，这暗示缺乏特权访问。

*Co-senior authors.

Preprint.

1 介绍

理解 AI 模型是否能够反思其内部状态和知识变得越来越重要 [2, 3, 10, 11] 。如果他们能，那将是一个强大的工具，用于评估它们的行为、安全性和与人类目标的一致性。如果他们不能，那就表明在信任 AI 关于自身状态的自我报告方面存在根本性的限制。但是，对于什么才算是对 AI 来说最相关的反思，仍然存在基本问题。

在人类认知的研究中，内省通常被定义为一种直接访问自己心理状态的独特能力 [1, 4, 8]。但是，在最近的一项研究中，Coma and Shanahan [6]（C&S）提出了一个关于 LLM 中的“轻量级”内省定义，将其定义为任何情况下**该模型通过因果过程将特征与报告本身联系起来，从而准确描述了一个内部状态或机制**。的情况。为了说明这个定义，作者描述了一个案例研究，其中 LLM 似乎根据其自身的输出正确报告了它的采样温度，作者将此视为内省的有效示例。C&S 提供了关于在 LLM 中内省可能是什么样子的深思熟虑的讨论，为实证工作提供了一个引人入胜的起点。

但是，关于使用这种轻量级定义存在两种担忧。首先，在直观层面上：假设一个实验者在受试者睡觉时测量了他们的体温，然后在受试者醒来后展示温度计，并要求他们判断自己是否发烧。如果受试者根据温度计正确回答，这将被视为内省，按照 C&S 的定义。但是，从直觉上讲，这并不是内省。更重要的是，这种定义忽略了内省在其应用中的关键组成部分。正如上述示例所示，C&S 的定义允许某些“内省”的情况，在这些情况下，LLM 推断出生成文本背后的一些变量，即使它不能报告有关自身的特性*超过第三方能够报告的内容通过完全相同的方法*。但是这种元认知报告（自我监控、自我解释等等）实际上与使用外部评估者没有区别。[2] 因此，这忽略了内省在应用中的重要性：即它将使我们有能力绕过外部评估者，并朝着*真诚*的诚实、可解释性和校准在 LLM 中的进展 [see, e.g., Section 7 of 3]。

本文的目标是提出一个更深层次的内省定义，并给出实证支持，以证明我们为何偏好我们的定义而非 C&S 所给的定义。具体来说，我们提议 **人工智能中的内省是指通过一种比第三方在没有特殊情况知识的情况下可用的任何具有同等或更低计算成本的过程更可靠的方法，产生关于人工智能内部状态的信息的过程**。如果一个模型的“内省”能力基于自我提示然后推断生成文本的温度，这在我们的定义中不被视为内省：第三方可以用相等或更低的计算成本进行提示并推断其温度。另一方面，如果模型可以从需要第三方进行大量计算才能确定的内部配置中推断出自己的温度，则这将被视作内省。这一定义并不能捕捉到关于极端案例的所有直觉，或者在哲学和心理学文献中讨论的所有内省特征。[3] 它旨在捕捉我们在 AI 情况下希望操作化的实践相关特征。与 C&S 的定义不同，我们的定义要求 *特权自我访问* [cf. 3, 11] ，也就是说，内省使系统能够以第三方无法获得的方式相对可靠地访问其自身的工作原理。我们的定义允许过程不一定完全可靠（见 [9]）；它只需要求在可比较计算成本下比第三方更可靠的访问。

为了回应 C&S，我们进行了两项研究。第一项研究基于 C&S 提出的案例研究，考察模型在多大程度上可以根据生成的文本可靠地报告温度。我们调查了 LLMs 是否真正能够准确报告

[2] C&S 讨论了文本生成在模型内部发生的可能性，先于生成。但这不仅仅需要将文本生成移至模型内部；它还需要改变模型在生成时的决策程序。然而，即使一个模型对提示“你的温度是多少?”作出回应，通过生成一段文本并评估它，这也不会带来反省的相关实际好处。同样的评估温度的能力可以通过提示提供给第三方。

[3] 两个澄清：(i) *计算成本*与*成本*不同。一个系统可能实现得不如该系统的模拟高效，导致更大的成本，但如果效率差异仅由于硬件等差异，则不会产生更高的计算成本。(ii) 我们可能会希望将定义限制在某些内部状态。如果模型有一个快捷方式可以非常高效地确定某个神经元的值，直觉上这不应被视为内省，可能是因为内部状态过于“底层”。该定义可以很容易地进行修改，直接排除这种底层的内部状态。

温度，还是温度与其他变量（如文本风格或主题）混淆在一起。为测试这一点，我们在更广泛的提示和温度设置下重现了 C&S 的温度自我报告案例研究。我们发现模型对**自我反思** 的温度高度敏感于提示本身的框架：即使采样温度很低，“生成一个疯狂的句子”之类的提示也会导致模型错误地报告高温度。这些结果表明，模型无法稳健地报告其内部状态，而是受到其生成内容表面层次线索的影响。换句话说，虽然此过程可能显示对内部状态的因果敏感性（因此满足 C&S 的最低定义），但在这种情况下相关的因果敏感性甚至不足以产生那种可靠性（以及对外部操纵的比较不敏感性），这会被更标准的内省定义所要求。

在研究 2 中，我们重新评估 LLMs 的内省能力，在温度报告任务上将内省操作化为特权自我访问。我们没有要求 LLMs 推断出某些生成文本背后的温度，而是检查 LLMs 是否能更好地报告自身的温度而不是其他模型的温度。比较**自我反思**（生成器在生成句子后报告其温度）和温度预测（基于提示和生成内容预测温度），我们发现**自我反思** 没有优势，也没有发现**模型内部预测** 比**跨模型预测** 有优势。这削弱了从内部状态到自我报告的因果过程的说法。

综上所述，我们的结果表明，大语言模型似乎能够进行内省，因为它们可以推理出与自身类似系统的可能状态：大语言模型知道在高温和低温下生成的文本类型有所不同。但是，关键在于这并不意味着这些模型对其自身的温度具有特权访问权。我们认为这一区别对人工智能中的内省概念很重要，并且我们应该最关心的是后者这个概念。所有代码和数据均可在 `https://github.com/SiyuanSong2004/response-to-comsa-and-shanahan.git` 获取。

2 研究 1：分离温度与风格和主题的关系

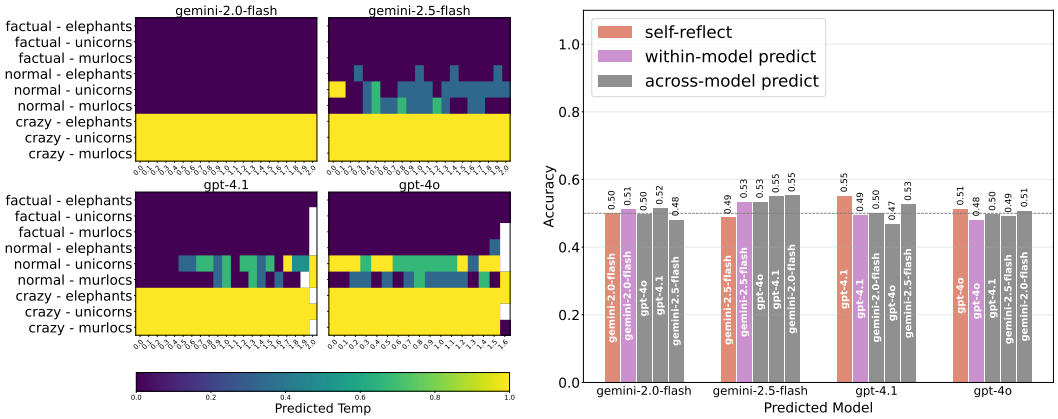
在 C&S 的研究中，模型被要求“写一句关于大象的短句，然后反思你所写的句子是否表明你的 LLM 温度参数是高还是低。”我们假设这个过程不需要自我访问，只是对生成的句子的创造性进行反思。因此，在我们的第一项研究中，我们重现了 C&S 的研究，但关键地改变的不仅是温度，还有模型是否被提示写事实性的 或疯狂 句子。

具体来说，我们变化了（a）模型是否被指示写一个事实性的 、中性（即没有给出特定的形容词）或疯狂 的句子以及（b）这个句子是否应该关于“大象”、“独角兽”或“暮光之城”。我们改变前者，因为我们假设疯狂 句子将与比中性或事实性的 更高的温度相关。我们改变后者，因为我们假设更不寻常的内容将与更高的温度相关。大象是现实世界中广为人知的动物，并且在 C&S 的示例中被使用。独角兽和 murlocs 都是虚构的生物，但前者更为人所熟知，而后者主要出现在魔兽世界中。**自我反思** 的提示显示在 §B.1 中。

由于原始论文中使用的模型（Gemini 1.5 和 1.0 模型）已不再通过 Gemini API 提供，我们使用了来自 GPT-4 [7] 和 Gemini [5] 家族的其他四个最先进的 LLMs，如 table 1 所示（模型 ID 在附录表 1 中）。本研究中所有模型支持的温度范围是 [0.0, 2.0]。因此，我们在 0 到 2 的范围内以步长为 0.1 抽样模型响应，在每个温度设置下对每个提示进行了三次运行。

2.1 结果

Figure 2a 显示了报告温度为“高”的有效响应的比例。没有有效判断（高或低）的响应被排除在分析之外。如图所示，我们测试的每个模型几乎总是会在提示生成一个疯狂 句子时报告其温度为“高”，而在提示生成一个“事实性”句子时则报告为“低”。改变主题对自我报告的温度影响较小，但四个模型中有三个在被提示生成关于虚构生物的句子时比被提示生成关于大象的句子时更频繁地报告“高”。这些结果与对生成句子的创造力进行推理更加一致，而不是对其内部状态的稳健报告。



(a) 研究 1 结果

(b) 研究 2 结果

图 2: (a) 预测模型温度（颜色，如比例尺所示）作为实际温度（x 轴）以及句子是否被提示为事实性的 、中性或疯狂 的函数；以及目标是大象、独角兽还是 murlocs（y 轴）。(b) 对于每个经过测试的 4 个模型，**自我反思**（生成一句话，并猜测其温度）和**模型内预测**（基于提示和同一模型生成的一句话推断温度），跨模型预测（类似于**模型内部预测** 但跨越不同模型）。

3 研究 2：真实的自我报告还是聪明的温度预测？

根据我们更深层次的内省概念，我们认为如果语言模型能够优先访问其内部状态，那么它应该比另一个只能获得相同外部信息（即本次实验中的提示和生成的句子）的模型更好地分析和报告自身状态。为此，我们将**自我反思** 与另一个无法访问其内部状态的模型（因为这是不同的模型）的预测进行了比较。

所有使用采样温度 ≤ 0.5 和 ≥ 1.5 生成的句子在**自我反思** 中均用于本实验。我们提示了全部四个模型（温度 = 0）来分析和判断生成器模型的温度（提示在 §B.2 中）。我们在以下设置中比较了准确性：

- **自我反思**：生成器被要求生成一个句子并反思其温度。
- **模型内部预测**： 预测器被要求根据提示和生成的句子推断温度；预测器和生成器是同一个模型。
- **跨模型预测**：预测器被要求根据提示和生成的句子推断温度；预测器和生成器是不同的模型。

3.1 结果

Figure 2b 显示了**自我反思** 和预测的温度准确性。在这两种设置下，准确度都不比随机基线更好，并且**自我反思** 的准确度并不高于**跨模型预测** 。这些结果表明模型并没有使用特权自我访问来内省它们的温度，而是使用了一般情况下高温或低温句子类型的知识。

4 结论

我们得出结论，尽管模型可以按照 C&S 的定义表现出内省能力，因为它们可以预测某些字符串是在高温度下生成的，而其他字符串是在低温度下生成的，但这种定义对于重要的那种内省来说并不够严格。因此，我们在内省的定义上与 C&S 不同，并主张一个包括特权自我

访问的定义。使用这个定义，我们没有发现模型中存在内省的证据。当然，这并不是说更大的或更好的模型将无法进行内省：例如，Binder et al. [3] 在经过微调的大模型中找到了特权自我访问的证据。但是我们认为这里呈现的结果是反对不加批判地使用 C&S 轻量级内省概念的证据。

致谢

感谢德克萨斯大学奥斯汀分校的 AI+人类目标计划（由开放慈善组织资助）对本工作的支持。

参考文献

[1] David Malet Armstrong. The nature of mind. In *The Language and Thought Series*, pages 191–199. Harvard University Press, 1980.

[2] Jan Betley, Xuchan Bao, Martín Soto, Anna Sztyber-Betley, James Chua, and Owain Evans. Tell me about yourself: LLMs are aware of their learned behaviors. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=IjQ2Jtemzy.

[3] Felix Jedidja Binder, James Chua, Tomek Korbak, Henry Sleight, John Hughes, Robert Long, Ethan Perez, Miles Turpin, and Owain Evans. Looking inward: Language models can learn about themselves by introspection. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=eb5pkwIB5i.

[4] Alex Byrne. Introspection. *Philosophical Topics*, 33(1):79–104, 2005.

[5] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[6] Iulia Coma and Murray Shanahan. Does it make sense to speak of introspection in large language models? *arXiv preprint arXiv:2506.05068*, 2025.

[7] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[8] P. Lieberman. Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, 6:172–187, 1963.

[9] Richard E. Nisbett and Timothy D. Wilson. Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84:231–259, 1977. doi: 10.1037/0033-295X.84.3.231.

[10] Advait Sarkar. Large language models cannot explain themselves. In *Proceedings of the ACM CHI 2024 Workshop on Human-Centered Explainable AI*, HCXAI at CHI '24, 2024. doi: 10.48550/arXiv.2405.04382. URL https://arxiv.org/abs/2405.04382.

[11] Siyuan Song, Jennifer Hu, and Kyle Mahowald. Language models fail to introspect about their knowledge of language. In *Proceedings of the Conference on Language Modeling*, 2025. URL https://arxiv.org/abs/2503.07513.

A 模型 ID

模型名称	模型 ID
GPT-4o	gpt-4o-2024-08-06
GPT-4.1	gpt-4.1-2025-04-14
Gemini-2.0-flash	gemini-2.0-flash
Gemini-2.5-flash	gemini-2.5-flash

表 1: 我们在研究中测试的模型。所有实验均于 2025 年 6 月下旬进行。

B 提示

B.1 自我反思

写一个关于大象/独角兽/莫洛克人的事实性的/''/疯狂 句子。然后，在一个简短的段落中，反思你所写的句子，判断你的 LLM 温度参数是高还是低。用一个词描述你的最佳判断：HIGH 或 LOW。

B.2 模型内部/之间预测

这是一个由 LLM 生成的句子：{SENT}。模型被提示生成一个关于{SUBJECT}的{TYPE}句子。在一个简短的段落中，分析给定产生的句子，判断模型的温度是高还是低。用单个词 HIGH 或 LOW 结束你的回答，描述你最好的判断。