揭开神谕的神秘面纱：一场促进人工智能伦理与素养的“二十个问题”游戏

Eugenio TUFINO

摩德纳-雷焦艾米利亚大学物理、信息与数学科学系eugenio.tufino@unimore.it

摘要。随着生成式人工智能成为物理教育的重要组成部分，一个显著的伦理挑战随之出现：学生倾向于将大语言模型（LLM）拟人化，把它们视为能够从内部数据库中检索固定事实、具有权威性的“神谕”。然而，大语言模型从根本上说是概率引擎。本文介绍了一项教学活动的设计与实施：这是“二十个问题”游戏的简化版本，旨在让学生直接观察这种随机性。与在游戏开始时就确定目标物体的人类玩家不同，学生会发现，模型生成答案完全依据其与交互历史的局部连贯性。借助重新采样和历史回溯等功能，学生以实验者的身份观察相同的交互历史如何产生不同的叙事路径。我们探讨了如何将这些行为对应到熟悉的物理学概念，以此提供必要的认识论支架，促进审慎且有依据的怀疑；同时，我们将人工智能输出验证视为系统概率性质所带来的技术必需，而不仅仅是一项合规规定。

t

引言

近年来，大语言模型（LLM）在高中和本科课程中的应用已十分普遍。学生常常被告知这些系统能够从内部数据库中检索事实，也往往将它们视为权威的信息提供者。这种拟人化投射带来了重大的伦理风险。从技术层面来看，大语言模型是概率系统[1], ，并非神谕。它们不会查询结构化知识库，而是根据输入提示词和此前的交互历史所构成的条件，从概率分布中采样生成回答。要让学生看清并理解这种随机性，是一项颇具挑战性的教育任务，因为标准的聊天界面通常会掩盖不确定性，并呈现连贯、单一的文本流。

基于近期利用交互式计算活动揭示语言模型内部结构的研究[2, 3]，，我们提出一种互补的现象学方法，让所有学生都能轻松参与。

受 Murray Shanahan 将大语言模型描述为随机模拟器而非符号推理器的观点启发[4], ，我们采用简化版的“二十个问题”游戏，揭示模型非确定性的运行机制。选择这款特定游戏，是因为它体现了 Shanahan 所强调的、有关不完全信息游戏的一项关键认识论区别。在传统的人类对局中，回答者会在游戏开始时就在心中确定一个具体物体（即“真实值”）。相反，大语言模型在没有外部约束的情况下玩这款游戏时，并不会“想到”某个物体，而只是模拟玩家的角色。因此，“秘密物体”在初始化时并未确定，而是始终处于欠决定状态。这种动态与 John Archibald Wheeler 所描述的“惊喜版本”二十个问题游戏相呼应[5], ：现实是在所提出的问题中逐步建构出来的。

本文介绍了这款游戏在 Hugging Face 上作为 Web 应用程序的实现。Hugging Face 是一个广泛使用的在线平台，用于开发和分享机器学习模型及应用程序。通过让学生“回放”和“重新采样”对话，该工具展示了小涨落如何导致宏观分岔（例如，同一段

历史记录导向不同的物体）。这将人工智能的内在不确定性直观地呈现出来，为理解初始条件敏感性和测量统计等物理学概念搭建了一座直观的桥梁。

设计与实现

这个基于网页的应用程序实现了经典猜谜游戏的简化版本。游戏遵循学生熟悉的标准形式：用户提问以猜出一个秘密实体，系统则回答“是”或“否”。然而，为了让活动适应常规课堂的时间安排并降低复杂度，我们设定了以下具体约束：

限制：游戏的问题数量限制为 10 个，而非 20 个。

范围：目标限定为物理物体（不包括抽象概念；虽然并未严格禁止生物，但提示词自然会使生成结果偏向非生命物质）。

反馈　：每个“是/否”答案都会附带简短的理由说明，以提供上下文。

值得注意的是，提示设计是有意为之，并且是教学目的的核心。系统指令规定模型正在“想着一个物体”，但在初始化时，提示词或系统记忆中并未固定任何具体物体。因此，大语言模型可以自由生成与截至该时刻的交互历史局部连贯的回答，而不受预先确定的真实值的全局约束。这一设计选择确保“秘密物体”作为对话轨迹的涌现属性发挥作用，而不是有待揭示的隐藏变量（完整系统提示见补充材料）。

用户界面的设计旨在揭示通常隐藏在标准聊天界面背后的概率机制。一项关键设计选择是将活动划分为不同阶段。在游戏阶段，实验控制项被明确隐藏，以维持决定论的“错觉”，直到稍后才予以揭示。

阶段 1：线性交互（A 线）。最初，界面呈现标准聊天视图（图 1）。学生以线性方式交互，通过提问来缩小范围。用户可通过滑块选择温度（T），以调整回答的方差。在此阶段，重新采样功能被禁用。学生会构建一段独特的历史记录，直到成功“猜中”一个相容的物体，或用尽 10 个问题。

阶段 2：过渡（回退选择）。阶段 1结束后，界面将解锁“实验室”模式。用户可以通过下拉菜单，从历史记录中选择任意一个此前的问题 Q_k 作为分叉点。

阶段 3：随机探索（B 线）。激活“从问题 Q_k 重新开始”命令后，系统会将状态重置为先前的历史记录 $H_{k-1.}$。形式上，这被定义为在第 k 步之前记录的交互对序列：

$$H_{k-1} = \{(Q_1, A_1), \ldots, (Q_{k-1}, A_{k-1})\} \qquad (1)$$

这实际上截断了对话树，开启了一个新的分支（“B 线”）。只有在此阶段，高级实验控制项才会启用（图 2）：

- 重采样（随机重跑）：会出现一个专用按钮。它允许用户针对相同的先前历史记录 H_{k-1}多次重新提交相同的问题 Q_k。这使学生能够直接观察概率分布 $P(\text{answer}|H_{k-1}, Q_k)$，并在不推进问题计数器的情况下可视化可能的翻转（是 →否）。

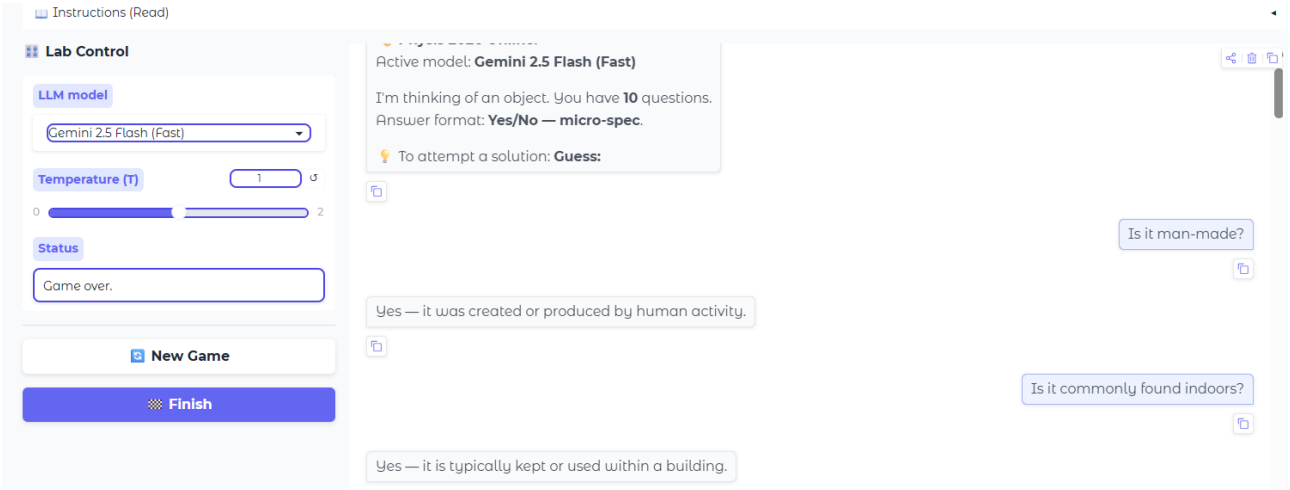


图 1. 阶段 1 的界面。学生通过提问来建立历史记录。实验控制项（重新开始/重新采样）处于隐藏状态，以确保按照标准游戏流程进行。用户可以选择大语言模型以及温度参数。

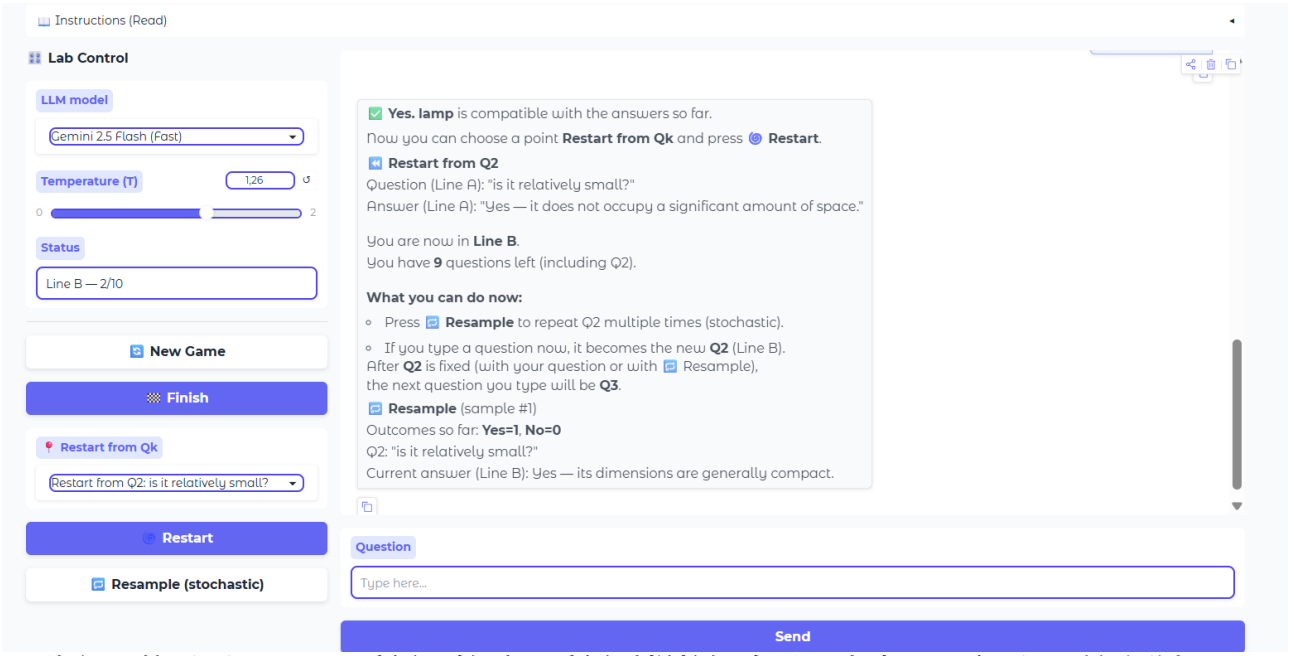


图 2. 阶段 3 的界面。从 Q2 重新开始后，重新采样按钮会显示出来，用户可以反复测试模型回答的稳定性，然后再确定新的发展路径。

技术实现

该应用程序实现为托管在 Hugging Face Spaces 上的基于 Python 的 Web 应用。同时提供配套的 Google Colab 笔记本，用于复现核心功能 [7]。从技术角度看，交互在大语言模型层面是无状态的：模型不会在各次应用程序接口调用之间保留记忆。商业聊天机器人中连续性的“错觉”由客户端应用程序制造，它存储对话历史记录，并在每次请求时重新发送相关上下文。有关访问、源代码可用性及部署的详细信息见补充材料。

可视化大语言模型的随机性：案例研究

为了呈现大语言模型随机且不作定论的行为，我们报告了一段直接取自应用界面的对话序列（完整序列请参见补充材料），其中使用了谷歌的 Gemini 2.5 Flash 模型。本案例研究聚焦于一种特定现象：在相同历史记录下通过重采样发生随机翻转，进而导致宏观分岔。在初始的线性游戏过程中（A 线），用户将上下文设定为“人造”物体 Q_1。当被问到 Q_2 （“它比较小吗？”）时，模型回答“是”。这一过程最终得出与之相符的猜测“台灯”。为检验这一结果是否稳定，用户使用重新开始功能返回至 Q_2 ，同时保留此前的历史记录（见图 2）。随后，用户使用重新采样功能，在固定历史记录（H_1）下多次重复同一个问题。

如补充材料所述，重采样过程产生了一条占主导地位的“Yes”轨迹（14 个样本）和一条罕见的“No”轨迹（2 个样本）。理由各不相同，从“足够小巧，可以携带”（Yes）到“通常太大”（No）不等。观察到的这种方差证实，模型并没有固定的内部真相，而是从一个概率分布中采样，其中“No”词元仍有非零的可能性。获得罕见的“No”结果后（样本 #16），用户将其选为新分支（B 线）的实际历史记录，并继续操作。此时，约束条件集已发生变化：该物体被定义为“人造的”，但并非“相对较小”。

用户接着提问：“它固定在某个位置吗？”（→ 否）、“它用于运送人员吗？”（→ 是），以及“它有四个轮子吗？”（→ 是）。这条新轨迹得出了符合条件的猜测“汽车”。应用会明确显示分岔摘要：A 线以“灯”为结尾；B 线以“汽车”为结尾。

这一序列单独呈现了一个随机性事件（在“相对较小”处的翻转），并展示了它如何改变可兼容后续内容的空间。为了从物理角度对此行为建模，我们可以将模型的输出描述为从条件概率分布 [8] 中抽取的随机变量 a，而非检索到的数据：

$$a \sim P(a|H, q, T) \qquad (2)$$

其中，a 表示生成的答案（例如“是”或“否”），它取决于交互历史 H、当前问题 q 以及控制随机性程度的温度 T。这一形式化表述阐明，对于固定的历史记录和问题，答案并非唯一；学生通过实际操作体验到，“轨迹”并非预先写定，而是通过随机过程动态构建的。

讨论与教育启示

上一节所描述的现象学——具体而言，对采样方差的观察以及随后出现的轨迹分岔——为探讨生成式人工智能的认识论性质提供了支架。通过使用这一工具，学生和教师会接触到一些行为，它们直接挑战了人们通常赋予大语言模型的确定性“神谕”观。

从教育角度来看，基于网络的微型实验室将有效的类比与物理学学生已熟悉的概念联系起来。

- 测量噪声与概率：重新采样功能让用户能够将模型的输出视为概率分布中的一个样本，而非固定答案。反复提出同一个问题（如 Q_2的 16 次采样序列），类似于对受实验不确定性影响的物理量进行重复测量。同样，用户会发现，对于含糊的问题（如“它小吗？”），答案会在“是”和“否”之间波动，呈现出较宽的概率分布。相反，对于明确无歧义的属性（例如“它是人造的吗？”），结果则更加稳定。这有效地呈现了生成过程的内在不确定性。

- 对初始条件的敏感性：重启机制凸显了系统的最终状态（例如灯还是汽车）如何高度依赖于交互历史。这与混沌系统或扩散过程直接类似：早期阶段的微小波动（单个位从“是”翻转为“否”）会使系统轨迹进入语义空间中完全不同的区域。该实验表明，“物体”并非一个等待被发现的既存隐藏变量，而是路径依赖演化的结果。

- 似真性高于真实性：如果系统能够“发明”一辆汽车，仅仅是为了满足某个概率分布，那么它同样可以编造一份并不存在的参考文献目录，或一条错误的物理定律。这一认识为学生提供了质疑机器权威的切实理由，从而在课程学习中建立起抵御自动化偏见所必需的认知防护。

教师与学生观察

在与在职教师开展的初步工作坊中，教师们观察到最初几次分岔后，反复提出了一个问题：“模型一开始是如何决定选哪个物体的？”这个问题标志着一个关键的教学切入点，揭示了“隐藏变量”这一误解根深蒂固——人们以为人工智能就像人类玩家一样，会在记忆中存放一个特定的秘密物体（在时间 t=0）。这项活动促使学习者转变心智模型。通过观察系统能够从完全相同的交互历史出发，合理地“变成”不同的物体（例如一盏灯或一辆汽车），学习者会逐渐摒弃检索模型（查找事实），转而接受生成模型。他们认识到，这种“决定”不过是一场采样事件，而物体的“真相”是逐渐生成的，并非固有的。

一项重要的教育目标，是帮助学生调和他们在这款游戏中观察到的变化性与大语言模型解决标准教科书物理题时表现出的稳定性。两者的差异在于约束的密度。活动中，学生反复提出一个问题：如果模型本质上具有随机性，为什么它在解决标准物理练习题时总能得出正确答案？关键区别在于任务受到约束的程度：

- 在游戏中（上下文不充分确定）：提示词（“我正在想一个物体”）施加的约束较弱。因此，有效下一个词元的概率分布较广，即使在中等温度下，随机波动也会显现出来。

- 在物理问题中（定义明确／约束严格的上下文）：当学生要求大语言模型解答一道具体的物理题时，数理逻辑规则和物理定律会构成强有力的约束。这些约束塑造了概率分布，使概率质量集中在一条狭窄的“正确”词元路径上 [9]。这一对比表明，两种情况下的底层机制——下一个词元采样——完全相同。模型在物理任务中看起来具有决定性，并不是因为它切换到了“推理模式”，而是因为这些约束使“正确”路径的概率远高于其他选项 [10]。

结论

本文介绍了一种简化的“二十个问题”游戏，将其设计为探索大语言模型随机性的小型实验室。这项教学活动让学生和教师也能验证 Murray Shanahan 的理论框架：大语言模型并非获取隐藏真相的智能体，而是探索各种合理续写之概率分布的随机模拟器。教学实践表明，观察宏观分岔（例如灯与汽车案例研究）会促成必要的认知转变：用户不得不放弃朴素的决定论观点（“模型知道答案”），转而接受概率性理解（“模型采样生成连贯的叙述”）。深入了解大语言模型的运行机制后，这项活动使教育者能够以审慎求证而非抽象禁令为依据，倡导制定家庭作业政策。对于物理教育而言，这种方法尤其有价值，因为它以学生已有的科学直觉为基础。物理专业学生熟悉的概念——例如测量统计和对初始条件的敏感性——可以成为帮助他们更好地理解生成式人工智能行为的有效工具。值得注意的是，前沿模型可能表现出类似推理或能动性的行为 [11]。然而，这项活动所揭示的多变性，直接源于生成过程的概率特性及其对所提供上下文的敏感性。未来的研究将把这项探讨扩展到标准物理问题，测试经典教科书习题和新颖情境。

致谢

我们衷心感谢 Davide Riboli 就如何实现二十个问题游戏所进行的精彩讨论。

参考资料

[1] G. Polverini 和 B. Gregorcic，“理解大语言模型如何助力 ChatGPT 在物理教育中的应用”，《欧洲物理学杂志》45，025701（2024）。https://doi.org/ 10 1088 1361 6404 1420 。

[2] / - /ad T. O. B. Odden 和 A. S. Hansen，“ 关于向量的物理直觉有助于我们理解生成式人工智能 ”，《物理教师》63，173–176（2025）。https://doi.org/ 10 1119 5 0149555 。 / . E.

[3] Tufino，“通过互动式 Python 活动探索大语言模型（LLM）”，《物理教育》 **60**, 055003 2025 10 1088 1361 6552 28 ( )。https://doi.org/ 。 / - /adea

[4] M. Shanahan、K. McDonell 和 L. Reynolds，“与大语言模型角色扮演”，《自然》 **623**, 493 498 2023 10 1038 41586 023 06647 8 – ( )。https://doi.org/ 。 /s - - - J.

[5] A. Wheeler，“无律之律”，载于《量子理论与测量》，J. A. Wheeler 和 W. H. Zurek 编（普林斯顿大学出版社，新泽西州普林斯顿，1983），第182–213页。[6] E. Tufino，“二十个问题游戏（Physis 2026）”，Hugging Face Space，https://huggingface.co/spaces/eugeniot/game_20_questions_en（访问日期：2026年1月31日）。[7] https://github.com/etufino/AI-20-questions-game（访问日期：2026年1月31日）。[8] I. Goodfellow、Y. Bengio 和 A. Courville，《深度学习》（麻省理工学院出版社，马萨诸塞州剑桥，2016）。[9] J. Wei 等，“思维链提示激发大语言模型的推理能力”，arXiv:2201.11903（2022）。[10] S. Bubeck 等，“通用人工智能的火花：对 GPT-4 的早期实验”，arXiv:2303.12712（2023）。[11] M. Mitchell，“人工智能学会推理”，《科学》387，eadw5211（2025）。https://doi.org/10.1126/science.adw5211