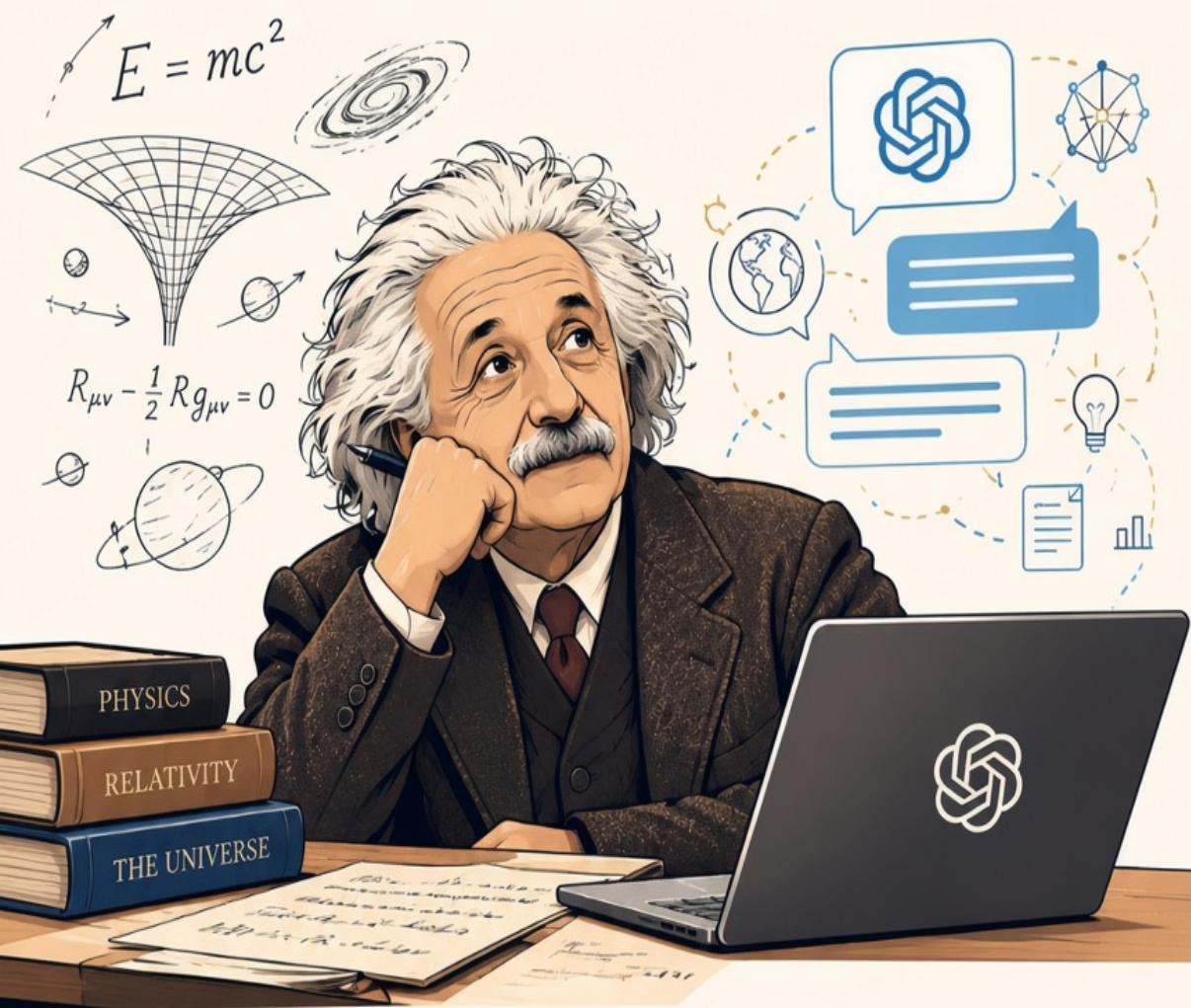


如果爱因斯坦 用过 ChatGPT

人工智能可能从我们的思维方式中学到什么 —
以及我们正在教它认识怎样的世界



弗兰克·梅旺热

内封标题页

如果爱因斯坦 用过 ChatGPT

人工智能可能从我们的思维方式中学到什么
——以及我们正在教它认识怎样的世界

FRANCK MEVENGUE

透明说明

一本关于人工智能的书，与人工智能合写

在一切开始之前，需要先做一个小小的实验：这本书讨论人工智能... 而它本身也是与 ChatGPT 合写的。若只谈实验却不去尝试，未免可惜。

从一开始，这个项目也是一次编辑实验：探索人工智能能在多大程度上帮助我们构思、检索、组织并写出有分量的书或文章，同时不放弃智识上的严格要求与人类的创造力。

本书所主张的观点、立场与方向，仍由我作为作者承担责任。ChatGPT 协助我进行文献检索、思想探索、结构安排与文字起草。重要来源与论断应由我本人审阅与核验。人工智能给出的动听回答，本身绝不构成证据。

我希望把这种合作关系说清楚。我们越来越接受人工智能参与软件设计、数据分析或企业效率。为什么艺术、写作与创作就应当被排除在这种探索之外？

工具并不必然取代作者：它也可以改变作者所能想象、探索与完成的边界。关键仍在于提出恰当的问题、行使判断，并承担最终选择。

毕竟，如果爱因斯坦用过 ChatGPT，他或许会用人工智能检验自己的假设。就我而言，我先从检验它能否帮助我写一本关于这个主题的书开始。

循环似乎闭合了。或者说，几乎闭合了：来源仍有待核验。

完整书稿

序言、十个部分与尾声

思想实验

2026：爱因斯坦打开 ChatGPT

能否为思想建模？

- 1 · 什么是思想？
- 2 · 心灵留下痕迹
- 3 · 不成为某人，能否认识某人？

若爱因斯坦用过 ChatGPT

- 4 · 2026 年的爱因斯坦
- 5 · 爱因斯坦的个人聊天机器人
- 6 · 爱因斯坦模型
- 7 · 爱因斯坦会认出自己的模型吗？

人工智能如何学习

- 8 · 从数据到表征
- 9 · 基础模型
- 10 · ChatGPT 不是数据库
- 11 · 上下文、记忆与参数

作为数据的人

- 12 · 每一次交谈都是数据吗？
- 13 · 研究者在喂养人工智能吗？
- 14 · 人工智能也在学习我们的方法吗？

人工智能与世界模型

- 15 · 什么是「世界模型」？
- 16 · 人工智能拥有对世界的表征吗？
- 17 · 从数百万痕迹到集体表征

与人类一同成长的人工智能

- 18 · 2026 → 2030 → 2040 → 2050
- 19 · 每一代都继承前一代吗？
- 20 · 人工智能 人类的循环

人类模型

-
- 21 ·人工智能能否学会「我们是谁」?
 - 22 ·智力孪生
 - 23 ·一座科学人格图书馆

真理问题

- 24 ·数据 ≠ 思想
- 25 ·相关 ≠ 理解
- 26 ·幻觉
- 27 ·镜子偏见

巨大悖论

- 28 ·我们教人工智能理解我们

大悖论

- 29 ·但人工智能也学习我们的世界模型
- 30 ·谁向谁学习?

2050: 我们创造了什么?

- 31 ·人类的人工记忆
- 32 ·人工集体智能
- 33 ·如果爱因斯坦真的用过 ChatGPT

尾声

镜子

2026：爱因斯坦打开 ChatGPT

用一个不可能的场景提出一个真实问题：机器能从思想的痕迹中学到什么？

■ 思想实验

本序言的全部叙述均为虚构。阿尔伯特·爱因斯坦于 1955 年逝世。此处描述的动作、屏幕、问题与反应均无历史价值。它们仅用于构建一项思想实验。

思想实验组织一个想象的场景，以便探究某个问题或某些原则的后果。它并不把该场景变成历史事件 [\[SRC-038\]](#)。

空白窗口

想象一间 2026 年的工作室。桌上的纸页写着方程、箭头、被写过两遍的词语。一本笔记本摊开在电脑旁。阿尔伯特·爱因斯坦——不是历史上的爱因斯坦，而是我们思想实验中的那位——看着一个仍空着的对话窗口。

他可能从物理学问题开始。他也可能提交一页未完成的文稿，要求提出异议，比较两种表述。没有任何东西授权我们断定他究竟会键入什么。重要的细节在别处：每当他使用助手，他都会留下一条新痕迹。一条注明日期的问题。一道被复制的方程。一条被接受、被忽略或被纠正的回答。或许还有一串尝试，它们展现的与其说是一个解，不如说是一种寻找的方式。

现在假定这位助手收到的不是单一问题，而是一套有组织的文档：文章、草稿、笔记本、书信、演讲与修正。经编辑的爱因斯坦档案证明存在一份可观的文字遗产，由科学文本、通信与其他置于时间之中的文献构成 [\[SRC-001\]](#) [\[SRC-032\]](#) [\[SRC-036\]](#)。我们的场景在这些历史档案之上，又增添了完全出于想象的当代互动。

助手可能找回一段文字，把两种表述并置，或标出某个想法以多种形式出现。它可能学到词汇、结构与论证上的规律。若给出足够多的示例，它或许能够——在某些标准下——产出与已观察文本相似的文本。

但它究竟学到了什么？

捷径的诱惑

我们很容易回答：「它学到了爱因斯坦。」这句公式很诱人，因为它把整条操作链压缩成三个字。它也很危险。

系统无法在思想被表达之前触及思想。它收到的是铭写：字符、数字化图像、元数据，或许还有转写。它看不见直觉出现的瞬间、困难的感觉、注意力的转移，或某句话为何仍留在抽屉里。即便语料无比庞大，它所包含的也只是已被表述、保存、找回、筛选并变得可读的东西。

这样说并不使痕迹变得无足轻重。一处涂改可以表明某种表述被放弃。一串草稿可以记录一次转变。一封信可以把问题放回某次对话之中。这些要素支持推断。它们并不消除文档与产生该文档的心理活动之间的距离。

这段距离构成全书的第一条线索：

人工智能并不直接阅读心灵；它处理可及的痕迹，并可以构建有用的表征，用以预测、分类、推断、生成或模拟所观察到的某些方面。

动词「可以」至关重要。表征的范围取决于数据、架构、学习目标、工具、语境，以及评估方式。在某一情境中令人信服的模仿，既不保证模型的普遍忠实，也不保证与本人的同一性。

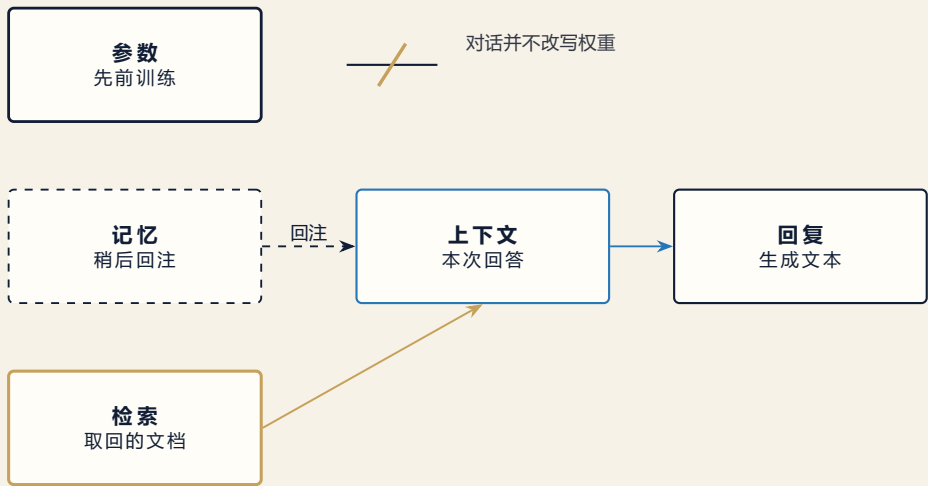
同一屏幕背后的四种机制

在我们的虚构场景中，爱因斯坦写在一个窗口里。这种视觉上的统一可能掩盖多种机制。

模型首先拥有**参数**，它们在先前训练中被调整。当前对话随后提供**上下文**，用于计算当下回答。一款产品也可以保存**持久记忆**，并在后续交换中重新注入。最后，系统可以执行**外部检索**，在文档中查找，并把找回的段落放入上下文。

这四个要素不可互换。对话中键入的一句话，并不因其单独存在就立即改写一般模型的参数。在为 GPT-3 描述的少样本评估中，任务在不进行梯度更新的情况下完成 [SRC-009]。在 OpenAI，有监督微调的文档描述的是另一种操作：上传一组示例，随后一项训练工作更新模型权重 [SRC-039]。某项服务的政策可能规定，部分数据稍后用于改进流程；那属于另一种时间性、控制与处理选择。ChatGPT 当前文档尤其区分记忆控制与「使用对话改进模型」的控制 [SRC-006] [SRC-007]。这一描述是注明日期的：它刻画的是可能演变的产品与政策，而非人工智能的普遍定律。

因此，我们想象中的爱因斯坦眼前，可能是一位回答综合了多种信息来源的助手。要知道系统实际用了什么，仅看回答的外观并不够。必须记录配置。



图。参数、上下文、产品记忆与外部检索并非同一种操作。对话会充实上下文，但不会更新权重。说明性示意图，非实验结果。[SRC-006] [SRC-007] [SRC-009] [SRC-039]，按登记册审查层级，产品文档日期为 2026 年 10 月 9 日。

■ 已确立

已学习的参数、当前上下文、产品的持久记忆，以及外部文档检索，是彼此不同的机制。一次交谈不得被描述为对一般模型的即时修改。

双重检验

设想在数月互动之后，助手收到一个新问题。它提出回答。专家们判断，该回答与爱因斯坦著作中观察到的某些习惯相一致。这一成功能证明什么？

它首先表明一种表现：在测试条件下，系统产出了满足既定标准的东西。若用于评估的文档在模型构建时被隔离，结果会比单纯重复已知段落更有信息量。若多位评估者、多个问题与多种表述趋于一致，推断会更坚固。

它仍然有限。一篇文本可能因错误理由而像另一篇。模型可能利用表面线索。评估者可能共享同一期待。尤其是，成功并不能证明系统像爱因斯坦那样体验问题，也不能证明它拥有其身份、记忆或意识。

因此检验应当是双重的。既要测量模型复现了什么，也要主动寻找它在何处失败。哪些表述变化会打破其连贯？哪些断言与文档相矛盾？科学方法的哪些方面缺失？它何时产出看似合理却无出处的回答？

严肃的表征不只凭相似来评判。也凭其限度的地图来评判。

两面镜子

思想实验从个体开始，却不能止于个体。当代系统在汇集众多个人、机构、语言、学科与时期产物的集合上训练。文本、图像、代码与测量，在采集、过

滤与转换之后成为数据。

本书的第一面镜子沿着这条链：

个体 → 被表达的思想 → 痕迹 → 数据 → 关于该人某些方面的模型。

第二面镜子把它扩大：

人类 → 观察与产物 → 数据 → 学习 → 用于预测与生成的表征。

在这两种情形中，同一警告成立：地图不是疆域。数据不是它们所代表之物的全部。它们可以丰沛却仍有缺口，精确却仍经筛选，有用却仍带偏见。

比较两面镜子，并不假定人与世界以同一方式被建模。它用于提出一个共同问题：当表征从痕迹中构建时，它保留了什么，又失去了什么？

场景迫使我们追问什么

回到空白窗口。2026 年虚构的爱因斯坦可能合上电脑，重拾笔记本，对回答不加评论。这沉默本身也含糊。它意味着同意、怀疑、疲惫，还是决定另寻他途？没有额外痕迹，我们无从知晓。

这种无知并非更多计算必然抹去的暂时缺陷。它提醒我们：对一个人的一切认识都经由观察与解释。人类在读一封信、重建一次发现或试图理解亲近之人时，已经与这一限度共处。人工智能并不创造这个问题。它放大它、自动化它，并迫使我们澄清标准。

因此后续各章将从 ChatGPT 之前开始。它们将追问：我们把什么称作思想，思想如何留下痕迹，以及从一个人的产物出发认识他意味着什么。这一迂回不可或缺。否则，我们可能把回答的流畅误认为通向启发它的心灵。当代文献本身也主张，在谈论语言模型时应对所用词汇保持谨慎 [SRC-019] [SRC-025]。

要点

- 01 本序言是思想实验，而非历史叙述。
- 02 助手处理可及的痕迹，而非直接处理心灵。
- 03 参数、上下文、记忆与外部检索必须保持区分。
- 04 模仿须经测试并经对其失败的研究来验证。
- 05 个体问题通向从人类痕迹构建的世界表征问题。

动员的来源

- **SRC-001 - Einstein Papers Project.** 经编辑的著作、通信与文献的制度性范围。查阅官方介绍与文献门户。

- **SRC-006 - OpenAI, Data Controls FAQ.** 2026 年 10 月 9 日查阅的产品文档；注明日期的政策，可能演变。
- **SRC-007 - OpenAI, Memory FAQ.** 2026 年 10 月 9 日查阅的产品文档；将记忆描述为个性化机制。
- **SRC-009 - Brown et al., 少样本学习.** 已审阅书目说明与摘要；用于说明在所述评估中任务可在不更新梯度的情况下完成。
- **SRC-039 - OpenAI, 有监督微调.** 2026 年 10 月 9 日查阅的官方页面；它将权重更新描述为训练工作的结果，而非孤立消息的结果。该平台可用性在演变。
- **SRC-019 - Mitchell 与 Krakauer, 关于语言模型理解的辩论.** 书目说明与摘要已核验；视角文章，并非单一实验裁决。
- **SRC-025 - Shanahan, 关于大语言模型所用语言.** 已审阅元数据、摘要与结论摘录。
- **SRC-032 - Einstein Papers Project, 项目史.** 已全文查阅的制度页面；描述 1955 年后论文的组织与编辑。
- **SRC-036 - Collected Papers, 第 5 卷.** 已查阅官方编辑说明；描述五百余封书信的范围，但不授权单独引用。
- **SRC-038 - Brown 与 Fehige, 思想实验.** 已全文查阅的学术条目；框定想象场景的方法论地位。

1. 什么是思想？

在要求机器为一思想建模之前，必须先澄清这个词究竟授权我们观察什么。

一个词，多种现象

「我有一个想法。」这句话似乎指称某种简单之物。它却可能指向心理意象、内心句子、回忆、意图、比较、计算、被解释的情绪，或尚无词语的问题。同一术语汇集不同现象。

这种多样性立刻造成困难。若我们寻求足够宽泛以囊括一切的定义，它可能什么也解释不了。若我们选择过窄的定义——例如「思想就是内心句子」——我们就会排除视觉推理、未经言语化的预备动作，或先于表述而出现的直觉。

因此，本章的目标不是解决思想的哲学问题。它对调查更为谦逊也更有用：建立一些区分，以免混淆心理现象、它的表达，以及从该表达计算出的模型。

我们将用**思想**一词指称归属于某人的心理活动或心理状态。这一定义有意保持一般。它既不裁定意识的本性，也不裁定主观经验如何涌现，也不裁定感知、情绪与推理之间的确切边界。这些问题重要；它们超出此处动员的证据所能确立的范围。

■ 未知

任何简短的思想定义本身都不能解决意识问题。本书采用操作性定义，以研究心理活动、可观察表达与模型之间的关系。

不可混淆的三个层面

取一寻常情境。某人在两条路线前犹豫，望向天空，然后说：「我坐火车。」我们可以至少区分三个层面。

第一是**心理经验与活动**的层面。此人可能想起一次晚点，想象下雨，估计时长，或感到某种偏好。这种活动的一部分或许对他可及；另一部分则并不清晰。

第二是**可观察表达**的层面。说出的句子、望向天空的目光、买下的车票或选定的路线，对他人可及。它们允许推断：雨或许起了作用，路程时间也是。

但多种不同故事可以产生同一选择。

第三是**描述或模型**的层面。观察者可以概括：「此人回避气象风险。」系统可以估计选择火车的概率。在这两种情形中，表征都从线索中构建。它可以有用而不必穷尽。

这些层面彼此呼应，却并不相同。表达由处于情境中的人产生；它不是其内在过程的完整副本。模型组织观察；它不是被观察的人。



心理活动 虚笔：不可直接观察



表达 实线：言语、手势、书写



模型 金色线：建构的描述

图. 三个相互关联却彼此区分的层面：心理活动无法直接观察，表达可以观察，模型则是事后构建的。作者示意图。

一个贯穿示例：修正一道方程

想象一页工作稿。上面有一道方程。一行被划掉。旁边出现另一种表述。没有任何书信解释这一动作。三种读法仍然可能，且地位不同。

观察。一道方程被写下，随后在载体上被修正。

推断。作者修改了表述。与这一页单独相容的原因仍有多种：发现错误、简化、预判异议，或只是地方不够。

假设。这一修正可能见证其处理问题方式的演变。这一假设原则上并非虚假。它也并非仅凭涂改即可确立。还需要周围材料：注明日期的版本、一封信、后来的出版物。

未知。修正的确切理由、动作当下的主观状态，以及从未被写下的东西，仍在可及材料之外。

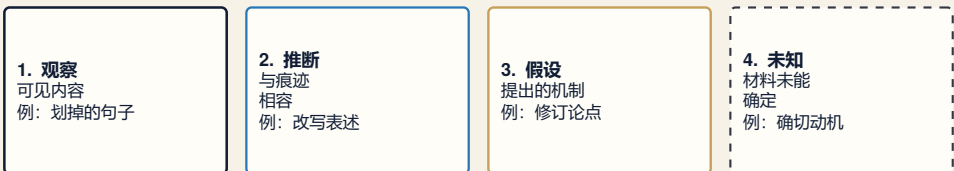


图. 阅读痕迹的四个层级：观察、推断、假设、未知。这一方法原则贯穿全书。作者示意图。

这张网格——观察、推断、假设、未知——将伴随全书其余部分。它将适用于爱因斯坦档案、助手的回答，以及我们会忍不住当作证据来读的那些断言。

因此核心问题不再是：「一般而言思想是什么？」它变成操作性的：人工智能能从一个人留下的痕迹中合理学到什么，而不把这些痕迹与人本身混为一谈？

被表达的思想，被归赋的思想

我们并不等待先进技术才归赋思想。我们持续如此。朋友在找钥匙：我们假定她想出门。同事把一句话重读三遍：我们推断他怀疑自己的表述。这些归赋使社会协调成为可能。

它们也可能错误。此人或许在找放在钥匙旁的文件。同事或许在改笔误，而并未质疑想法本身。观察往往对解释欠定：多种可能状态与同一痕迹相容。

这种欠定性并不谴责一切认识他人的尝试。它却要求一种方法。当归赋依托独立线索、经得起竞争解释、预测新观察并可修订时，它便更可信。它仍是推断，而非无中介的通达。关于心智化的研究，正研究这种从线索与视角出发表征或推断他人心理状态的能力 [SRC-022]。这一观念在关于「心理理论」的工作中更有更早历史：心理理论被定义为对不可直接观察的心理状态的归赋，以便解释或预测行为 [SRC-027]。

语言增加了一种特殊情形：人可以描述自己所想。这一证词宝贵。它可以澄清行为中看不见的意图。但它同样产生于情境之中，带有词汇、对象与目的。我们可能遗忘、简化、重组，或事后发现一种自认连贯的解释。认真对待主观叙述，并不意味着把它当作完美而完整的转写。

表征不是拍照

心理表征一词可能暗示脑中收纳的一幅小图像。这一隐喻有时合乎直觉，却很快变得误导。表征可以是一种结构，使人得以识别情境、比较可能性或引导行动，而无须在视觉上像它所表征之物。

认知科学运用多种框架研究这些结构。Johnson-Laird 著作的编辑介绍描述了一种理论：人会构建对象与事件之间关系的内在复本，再据此推理。[SRC-018] 这一表述来自出版者说明：为本次审阅并未通读全书。另一些工作区分多种记忆系统，或研究自传记忆的建构。[SRC-016] [SRC-017] 这些进路并不归结为关于思想的单一理论。它们为特定现象提供可检验的概念。

因此**表征**一词始终应唤起两个问题：对什么有用，以及如何被证实？若内部结构随某些信息系统变化，并促成行为或预测，便可称为表征。这一判断尚未说明它是否有意识、被体验，或以人类方式被理解。

当我们谈论神经网络时，这一谨慎将至关重要。模型编码规律这一事实，不足以断定它在思考。反过来，因系统非人而拒绝一切「表征」一词的用法，则会妨碍研究其内部状态实际允许做什么。

记忆不是内在档案

我们乐于把回忆想象成保存后再打开的文件。这一类比实用：它给出起点、内容与处所。它却可能掩盖自传回忆的重构性质。关于自我记忆的工作把回忆描述为自传知识、当下目标与提取线索相互作用的结果，而非对记录的原样阅读。[\[SRC-017\]](#) 记忆心理学也记录多种遗忘、变形与错误归因。这些限度并不使记忆无用，却禁止把它当作副本 [\[SRC-021\]](#)。

这一观念对我们的计划有两点后果。第一，人对自身所产的叙述是重要数据，但是情境性的。它们可以变化，而无须其中某一版本必然是谎言。第二，在这些叙述上训练的模型学到的是可得叙述的规律。它并不因此获得通往其背后完好存在的原初心理事件的通道。因此人工智能可以学到叙事形式，而不学到使这些形式成为必要的经验。

这些叙述不只是关于过去的报告。关于叙事同一性的研究也把它们描述为一种演进中的方式，用以衔接被重构的过去与被想象的未来 [\[SRC-026\]](#)。学习这些叙述的模型因此可以捕获连贯化的形式，而不拥有赋予它们意义的生活史。

人类记忆与计算机记忆之间的差异也必须保持可见。在数字产品中，「记忆」可以指被记录并重新注入上下文的持久信息。在人身上，该词指向多重生物与认知过程。使用同一术语并不证明机制同一。

意识在此不必裁定

本书无需裁定关于意识的辩论即可前进。一条规则足矣：不要把可观察表现当作主观经验的自动证据。系统可以分类、预测或生成。这些能力可测量。它们是否伴随与我们可比的经验，是另一个问题，对此处研究的模型而言尚未解决。

■ 已确立

行为表现与对主观经验的归赋是两个不同命题。前者可由协议测量；后者需要额外标准，并在涉及人工系统时仍有争议。

工作定义

我们现在可以表述将引导后续各章的定义：

在本书中，思想指归属于某人的心理活动或心理状态。我们直接观察到的只是其某些表达与某些效应。据此我们构建可修订的描述与模型。

这一表述并不否认人对自身经验的通达。它明确调查的视角：观察者、历史学家、亲近之人或从痕迹出发工作的系统的视角。

它也避免两种对称的过度。第一种是说：既然我们不能直接通达他人心灵，任何认识都不可能。我们的关系、心理学与历史相反表明，推断可以稳健。第二种是相信大量痕迹废除一切不确定。数量有时改善估计；它并不把观察

变成同一性。

下一章关注第一与第二层面之间的决定性过渡：人类活动如何成为他人可以保存、分类与解释的痕迹？

要点

- 01 本书采用思想的操作性定义，服务于调查。
- 02 观察、推断、假设与未知不可混淆。
- 03 心理活动、表达与模型是三个不同层面。
- 04 人工智能可以从痕迹的规律中学习，而不通达其人。
- 05 意识并不由可观察表现裁定。

动员的来源

- **SRC-016 - Tulving**, 情节记忆与语义记忆. 仅按登记册所载审阅层级使用。
- **SRC-017 - Conway 与 Pleydell-Pearce**, 记忆—自我系统. 用于框定自传回忆的建构性质。
- **SRC-018 - Johnson-Laird**, 心智模型. 已查阅编辑介绍；未通读全书。本章句子沿用该介绍，而非逐页核验的证明。
- **SRC-019 - Mitchell 与 Krakauer**, 关于语言模型理解的辩论. 书目说明与摘要已核验；用于区分能力与可能的理解形式。
- **SRC-021 - Schacter**, 记忆错误. 已审阅书目说明与摘要；用于框定回忆的易错与建构性质。
- **SRC-022 - Frith 与 Frith**, 心智化. 已审阅元数据、摘要与摘录；用于定义对他心理状态的推断。
- **SRC-026 - McAdams 与 McLean**, 叙事同一性. 已审阅 DOI 说明与出版者摘要；用于联结被重构的过去、被想象的未来与自我叙述。
- **SRC-027 - Premack 与 Woodruff**, 心理理论. 已审阅 Cambridge 书目说明与摘要；关于黑猩猩的历史研究，并非对人类心灵的直接阅读。

2 · 心灵留下痕迹

涂改、书信或方程可以记录一条智识路径——只要不把它们当作路径本身。

抵达我们的是什么

思想并不抵达档案。抵达档案的是被誊写的句子、被划掉的方程、被寄出的信，有时还有证人的叙述。在心理活动与文献之间，多重转化已经发生：选择一种表达，把它铭写在载体上，保存该载体，加以分类，再使之可及。

这些痕迹重要。一处修正揭示某种表述被放弃。两个版本使人跟踪一次演变。通信把决定放回交换之中。但没有任何单件材料独自包含产生它的全部经验。沉默、犹豫或从未被表述的直觉，可能不留任何印记。

爱因斯坦档案使这一差异具体可见。Einstein Papers Project 编辑与其生平及工作相关的著作、书信与文献 [SRC-001] [SRC-032]。第 5 卷官方说明单独描述 1902 至 1914 年间五百余封书信 [SRC-036]。第 4 卷说明标出记录通向相对论引力理论的研究进程的笔记与计算 [SRC-035]。这种丰沛使人得以研究年表、表述与往来。它并不把档案变成爱因斯坦思想的连续记录。

本章回答一个精确问题，与下一章不同：**在被保存的痕迹中，人类活动还留下什么？** 此处主要对象是文献转化。对作者的推断将在第 3 章出现。

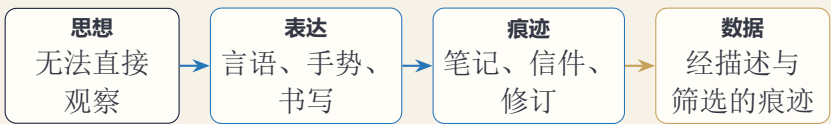


图. 被表达的思想可能留下痕迹；经描述与筛选的痕迹可成为数据。每一步都会改变随后被解释的内容。作者概念示意图。



每条箭头便于研究，也可能丢失语境、细微差别或出处。

图。从表达达到可用语料。每一次转换都便于研究，也可能改变或丢失信息。作者概念示意图。

表达本身已在转化

从想法到句子并非中性操作。必须选择词汇、顺序、细节层级与对象。给自己的笔记不像文章。给同事的信与公开演讲功能不同。同一人可以用多种方式呈现同一问题，而其中任何一种都不必是其思想的最终版本。

这种转化在草稿中尤其可见。涂改显示页面上的变化，却并不自动给出理由。作者是发现了错误，偏好更清晰的表述，预判了异议，还是只是地方不够？文献缩小可能解释的范围；它并不总在其间作出选择。

语境可以帮忙。同日写就的信、后续版本或通信者的回复增添线索。解释于是成为关联工作。它不在于让孤立痕迹开口，而在于比较出处与日期已知的材料。

这一规则也适用于数字痕迹。极短消息可能是紧急请求、提醒，或口头交谈的延续。修改历史显示改变了什么，未必显示为何。痕迹越容易收集，我们越须抵制它自我说明的想法。

保存即筛选

我们谈起档案，仿佛它们被动接收过去。实际上，一切保存都取决于决定与境况。某些文献被丢弃、遗失或销毁。另一些被保存，是因为它们看似重要，因为机构收购，或只因为载体耐用。

保存之后筛选仍在继续。必须编目、描述、数字化，有时还要翻译。科学编辑选择顺序、确立文本、关联异文并添加注释。这项工作极大提高语料效用。它也添加一层必须保持可见的编辑层。

就爱因斯坦而言，《Collected Papers》并非匿名文件堆。它们是机构项目产出的学术编辑，带有分卷、评注与通向档案文献的链接 [\[SRC-001\]](#) [\[SRC-033\]](#)。对人工智能系统而言，这一差异根本：附有日期、作者、语言与文献地位的段落，比脱离出处的字符串更可解释。

■ 已确立

经编辑的档案是构建而成的语料：保存、描述、排序与注释源于文献选择。这些选择改善通达，同时划定可观察的范围。

当痕迹成为数据

痕迹并不自动成为可供学习的数据。要被系统处理，它必须以可用形式被表征。一页可以被拍照，文本被转写，语言被识别，日期被编码，与其他文献的关系被描述。

每一操作都开启可能并引入错误。手稿页可能经过物理保存、数字化、光学识别、转写、规范化、索引，再整合进语料。忽视这些不确定的系统，收

到的是材料人为清晰的版本。

阶段	收益	风险
数字化	访问与复制	物质细节丢失
光学识别	可检索文本	字符混淆
转录	可用文本	规范化或遗漏
翻译	语言可及性	细微差别丢失
元数据	语境与编目	归属或断代错误
索引	快速检索	关系表示不当

表. 文献转换分析网格。每一步都可能改善访问，同时引入歧义。

因此质量并不归结为文献数量。它取决于出处、粒度、元数据、使用权，以及找回原始载体的能力。一百万条无语境碎片，对某些问题而言，可能不如更小却注明日期并经注释的集合有用。第 5 章助手所能「学到」的，将直接取决于这些转化。

还必须区分训练数据与回答时刻所查阅的文献。在前者，示例参与调整参数的程序。在后者，引擎找回文献并放入上下文。源文本可以相同；其技术角色不同。

否定性痕迹

档案显示在场之物。要正确推理，也须考虑缺失之物。一封信的缺席可能意味着从未写过、已经遗失、尚未找回，或不可及。这些假设后果不同。事后证词并不自动填补空白：人类回忆同样暴露于遗忘与错误归因 [SRC-021]。

对规律同样需要谨慎。若某人在语料中从不触及某主题，我们不能断定他从未想过。若他在公开文本中总用某词，我们不知道他在私人笔记中是否同样使用。语料描述痕迹的分布，而非一切可能情境的集合。

这些缺席可能是系统性的。被视为平常的活动有时留下的文献少于公共事件。拥有制度资源的人被更好地归档。某些语言、载体与群体在收藏中更易存活。建于档案之上的模型继承这种不平等地理。

承认这些限度并不谴责计划。它使人提出更精确的问题：覆盖哪些时期？何种文献体裁占主导？哪些中介筛选了材料？哪些沉默已知，哪些仍不可见？

阅读路径而不编造步骤

假定某推理的三个版本被保存。第一个含有类比，第二个增添方程，第三个删去类比。若日期或物质线索允许，我们可以确立文献序列。我们可以描述差异。我们有时可以把变化与某封信或已知实验相连。

当我们讲述作者「头脑里」发生了什么时，必须更加谨慎。说他删去了类比是观察。说他判断类比误导是推断。说某一精确直觉在那一刻出现，可能是无证据的假设。

本书的认识论语汇正为此用。它不禁止叙述；它标明叙述的坚固程度。当读者知道文献止于何处、解释始于何处时，观念史变得更有趣，而非更无趣。

助手可能利用什么

连接良好文档语料的助手可以完成多项有用任务。它可以找回某术语的所有出现，并置版本，构建临时年表，或标出表面矛盾。它可以向历史学家提出相关段落，后者再回到文献。

它也可能生成错误综合，融合两个时期，或把推断呈现为事实。文本的流畅并不保证档案忠实。因此，每一个重要回答都应保留通向所用材料的路径，并区分引用、转述与解释。

个性化模型可能学到更抽象的规律：句长、论证顺序、偏好的示例类型，或回应异议的方式。这些规律属于被观察的产物。在有限意义上称之为「思想风格」可以有用，条件是不把它们与思想的全部混为一谈。

■ 推断

众多、注明日期且多样化的痕迹可以支持关于某些智识规律的更稳健模型。它们既不保证语料穷尽，也不保证通达伴随其产生的主观经验。

从痕迹到模型

决定性过渡现在可以表述。被表达的思想有时产生痕迹。痕迹可以被保存与描述。痕迹的筛选可以成为语料。系统终于可以从该语料构建预测或生成某些产物的表征。

在每一步，有所获得：稳定、可比、检索、计算。在每一步，也可能失去：语境、歧义、沉默、出处或独特性。严谨在于记录这些转化，而不是用「数据」一词把它们抹去。

下一章将从如此构成的语料出发。即便痕迹真实且经仔细编辑，它们在多大程度上使人认识一个人？如何区分有用模型与只找回设计者期待的循环肖像？

要点

01 本章研究文献转化，尚未研究对作者的推断。

02 每一通达步骤也可能丢失、改变或使之含糊。

03 出处与元数据和体量同样重要。

04 痕迹缺席并非思想缺席的证据。

05 系统所能利用的取决于这些转化。

动员的来源

- **SRC-001 - Einstein Papers Project 与 Einstein Database.** 已查阅制度介绍；支撑语料范围、学术组织与文献一般类别。
- **SRC-021 - Schacter, 记忆错误.** 已审阅书目说明与摘要；仅用于提醒事后证词的限度。
- **SRC-004 - Lewis et al., 检索增强生成.** 摘要已核验；仅用于区分文献检索与参数记忆。
- **SRC-032 - Einstein Papers Project, 项目史.** 已全文查阅的制度页面；记录组织与编辑论文的历史。
- **SRC-033 - Collected Papers, 第 1 卷.** 已查阅官方编辑说明；刻画青年时期通信范围，不归赋未经核验的内容。
- **SRC-035 - Collected Papers, 第 4 卷.** 已查阅官方编辑说明；支撑 1912-1914 研究笔记与计算的存在。
- **SRC-036 - Collected Papers, 第 5 卷.** 已查阅官方编辑说明；支撑五百余封书信的一般范围，而非其个别内容。
- **补充档案来源.** 待第 3 阶段研究批次核验后纳入；未经定位文献，不归赋任何精确历史引用。

3·不成为某人，能否认识某人？

模型可以变得精确、有用且令人惊讶，而不停止成为从有限视角构建的推断。

痕迹之后，是推断

前一章跟随文献：表达、保存、数字化、元数据、语料。本章转换对象。问题不再是：还留下什么？而变为：我们能从这些痕迹就其作者推断什么？

我们认识我们并非其人的那些人。我们预判朋友的反应，认出同事的风格，并理解一句反讽不必按字面理解。这种认识既不完美也不虚幻。它依托互动史、言语、动作与共享情境。

它仍是间接的。我们并不代替他人感受其痛苦。我们看不见她决定时的全部回忆。我们构建关于其意图与信念的假设，再在言语或行动与之矛盾时修订。这种对他人的心理状态的推断，以心智化之名被研究 [SRC-022]。以归赋信念与欲望来解释或预测行为的策略，也有哲学史 [SRC-024]。

因此问题不是在两个极端之间选择——完美认识或一无所知。问题是确定模型能预测什么、在何种情境、带有何种不确定，以及以何种简化为代价。

这一问题先于人工智能。传记作者、医生、亲近之人与招聘者都构建关于某人的部分表征。其数据、目标与责任不同。计算模型使这种构建更显式，有时也更易于大规模部署。它并不使之自动中立。

从肖像到预测模型

肖像往往力求使某人可理解。它筛选事件、特质与张力。预测模型追求更操作性的目标：从过去观察出发，估计既定情境中的回答、选择或偏好。

差异重要。文本可以是令人信服的肖像，却无法预测新反应。反过来，模型可以正确预测有限选择，而不讲述关于该人的忠实故事。它可能利用人类解释仍不确定的统计规律。

关于数字痕迹的研究表明这类推断的力量与狭窄。在既定数据集中，Facebook「赞」曾被用于统计预测某些属性或心理测量特质 [SRC-028] [SRC-029]。这些结果针对特定目标、人群与平台。它们既不证明对整个人的理解，也不证明对其生活史的重建。

一些工作为这一观念提供实验形式，却不解决人类情形。Rabinowitz 及其

同事训练网络预测在简单网格世界中被观察的人工智能体 [SRC-023]。Park 及其同事构建环境，其中二十五个生成式智能体结合自然语言历史、记忆检索、反思步骤与规划 [SRC-012]。这些演示针对有界任务与架构。它们既不证明对真人的重建，也不证明通达其经验。语言模型也复现了某些群体回答分布或某些实验结果 [SRC-030] [SRC-031]。同样，模拟分布或平均回答不足以确立对特定个体的忠实。

为给某些智识产物建模，我们可以汇集文本、修正、对异议的回答与书目选择。系统会寻找规律：关联概念、论证顺序、偏好示例、谨慎程度或修订结论的方式。

结果不会是「那个人」。它会是某些任务、从既定语料构建的函数。这一表述或许显得冷。它却防止一个重大错误：把我们对个体所知、所爱或所想象的一切都归给模型。



图. 对人的模型须用未参与构建的材料来评估。相似并非身份的证明。作者概念示意图。

三种文献情境

比较三种情形使推断可测。

孤立文献。一句话出现在信中。多种解释仍开放。单独阅读的同一句话，可能是礼貌套语、科学立场或策略性让步。

多件独立文献。一封信、一张计算笔记与一篇公开文本趋于一致。若材料并非以不同形式重复同一来源，假设便更坚固。同一叙述的一百份副本不等于一百条证据。

矛盾文献或不同时期文献。两个注明日期的表述对立。推断于是必须点名时期、暴露张力或弃权。无缝综合不是解决：那是信息损失。



图. 当文献被置于语境、相互对照并在新材料上检验时，推断更稳固。但它仍不成其为确定性。作者示意图。

文献体裁与其内容同样重要。在私人信件、草稿或公开文本中读到的句子，并不授权对作者作出相同结论。仅凭一卷说明就给爱因斯坦归赋精确心理意图，在此会是方法错误。

如何测试关于人的模型

最弱的测试是要求模型复现它已见过的东西。系统可能还原公式或模仿段落，而未学到可推广规律。要进一步评估，必须保留未用于其构建的产物。

假定我们拥有注明日期的书信。一部分构成工作语料；另一部分在开发期间隐藏。模型收到留出信件的语境，并须预测——不是每个精确词语，而是事先定义的某些要素：所捍卫的立场、论证类型、动员的概念或信心程度。

标准必须在查看结果之前固定。否则，含糊回答总可事后被宣布为「相似」。还必须把系统与参照比较：简单规则、非专家读者、熟悉语料的专家。结果只有相对比较基线才有信息量。

最后，评估必须寻找失败。当语境改变、问题含糊，或人生两个时期呈现不同立场时，会发生什么？混合这些时期的模型可能看似连贯，却在历史上错误。

■ 已确立

在留出数据上评估，降低混淆还原与泛化的风险。这一原则不足：标准、比较与错误分析决定测试实际允许得出什么结论。

人在变化

关于人的模型常常包含沉默假设：个体足够稳定，以至于过去预见未来。这一假设有时合理。持久偏好或习得能力产生规律。若它抹去学习、断裂与语境变化，它便变得误导。

研究者早期著作未必描述其后期立场。分歧可能改变其方法。新证据可能推翻结论。甚至词汇也在演变。在所有时期上训练而不加年表的模型，可能产出从未存在过的平均值。

时间性因此必须成为数据的一部分。必须能够追问：「依据哪个时期的文献？」承认演变的答案，比人为均一的声音更忠实。

当代个性化助手也存在这一问题。昨日记录的偏好可能过时。准确信息可能不再准确。查阅、纠正与删除记忆的可能性，不只是实用功能；它回应被表征之人的变化本性。

评估者的镜子

要判断模型是否「像」某人，需要评估者。他们带来自身期待。公众可能主要认出人格最著名的特质。专家可能偏爱某一时期或史学解释。亲近之人可能重视公共档案中缺席的方面。

模型于是有风险成为刻板印象的镜子。若它产出与爱因斯坦流行形象关联的句子，即便它不符合任何经证实的文献或推理，也可能显得正确。人物越著名，直觉认出越须由文献标准补足。

稳健评估因此分开多个问题：

- 文本是否与该时期来源相容？
- 它是否复现了未明确提供的规律？
- 当断言事实时，它是否正确引用出处？
- 它是否知道标出回答并非由语料决定？
- 专家是否一致，依据哪些标准？

评估者之间的分歧不只是待消除的噪声。它可能揭示忠实概念尚未被充分定义。

人格、同一性、角色

另一捷径是谈论「智识孪生」。该词唤起整体副本。在实践中，系统通常在某些角色上更强：找回段落、模拟某种教学对话形式、提出与语料相容的异议，或模仿风格。

这些角色并不构成同一性。人的同一性包括生活史、关系、身体、承诺，以及对自身行为负责的能力。生成句子的系统并不成为这些句子所相似的历史主体。

这一区分有伦理后果。把模拟呈现为已故之人，可能造成误导性权威。用户可能把从未说过的话归给该个体。历史分歧可能被抹平，以产出更宜人的声音。界面因此必须宣布系统性质、其来源、覆盖时期及其限度。所选词汇

同样重要：仿佛系统自发就是人那样说话，可能使人忘记其机制与评估实际确立了什么 [SRC-025]。

■ 已确立

风格或回答的相似性并不确立模型与人之间的同一性。模拟地位、所用语料与断言出处必须保持可见。

模型能认识什么

我们现在可以对本章问题给出有条件回答。是的，模型可以捕获一个人产物中的真实规律。它有时可以预判选择、找回反复出现的结构，或提出评估者判断与语料相容的回答。

不，这一成功并不赋予它对该人的完全通达。模型受限于可得痕迹、它们被筛选的方式、所覆盖的情境与训练目标。它可以精确，却仍对从未被表达之物沉默。它可以令人信服，却仍在其测试分布之外错误。

因此好问题不是不加进一步说明地问：「系统真的认识这个人吗？」必须追问：

它能估计产物的哪些特征，带有何种可靠性，在何种条件下，以及哪些观察可能表明它错了？

这一表述把印象变成研究纲领。

思想实验的门槛

回到爱因斯坦。我们拥有丰富却不完整的历史语料，由机构编辑，穿越多种语言、时期与体裁 [SRC-001] [SRC-032] [SRC-036]。我们可以想象探索该语料并生成回答的系统。要知道它表征什么，必须分开找回的文献、学到的规律与评估者的期待。

我们尤其应保持两个爱因斯坦之间的边界。历史爱因斯坦出现在被归赋并被语境化的文献中。想象中的爱因斯坦服务于提出当代问题。生成的回答既不属于前者也不属于后者：它是我们构建的系统的输出。

这一边界开启下一部分。我们现在可以详细考察思想实验：我们将构建何种助手，将给它何种痕迹，以及如何避免把它的声音当作复活？

要点

01 痕迹转化之后，是对作者的推断。

02 孤立文献、独立材料与矛盾并不具有同等力量。

03 复制同一来源不等于增殖证据。

04 关于人的模型必须宣布任务、时期与语料。

05 输出的相似性既不确立同一性，也不确立历史权威。

动员的来源

- **SRC-001 - Einstein Papers Project.** 用于刻画历史语料，而不把虚构反应归给爱因斯坦。
- **SRC-012 - Park et al., Generative Agents.** 已审阅元数据、摘要与方法摘录；智能体架构示例，而非真人忠实模型。
- **SRC-017 - Conway 与 Pleydell-Pearce.** 动员用于自传记忆、当下目标与自我表征之间的关系。
- **SRC-022 - Frith 与 Frith, 心智化.** 已审阅元数据、摘要与摘录；对他人心理状态推断的框架。
- **SRC-023 - Rabinowitz et al., Machine Theory of Mind.** 已审阅 PMLR 说明、摘要与导论摘录；限于网络世界中人工智能体的示例。
- **SRC-024 - Dennett, 意向策略.** 已审阅说明与摘录；哲学立场，非实验结果。
- **SRC-025 - Shanahan, 关于 LLM 所用语言.** 已审阅元数据、摘要与结论摘录；用于防止无根据的拟人化。
- **SRC-032 - Einstein Papers Project, 项目史.** 已全文查阅的制度页面；记录编辑史，而非每一文献。
- **SRC-036 - Collected Papers, 第 5 卷.** 已查阅官方编辑说明；明确通信语料的一般范围。
- **SRC-028 - Youyou, Kosinski 与 Stillwell, 人格判断.** 已审阅 DOI 说明与 PubMed 摘要；结果限于样本、平台与心理测量目标。
- **SRC-029 - Kosinski, Stillwell 与 Graepel, 数字痕迹.** 已审阅 DOI 说明与 PubMed 摘要；概率预测，非个体确定性。
- **SRC-030 - Argyle et al., 人类样本模拟.** 已审阅 Cambridge 说明与摘要；集体分布，非个人忠实。
- **SRC-031 - Aher, Arriaga 与 Kalai, 人类研究复现.** 已审阅 PMLR 说明与摘要；不声称已通读 PDF。

4 · 2026 年的爱因斯坦

场景可以阐明一种方法，条件是永远不把想象的界面当作档案。

■ 思想实验

2026 年的爱因斯坦是方法论虚构。本部分描述的任何动作、问题与回答都不是历史事实。

同一房间里的两种在场

再次想象工作室。屏幕亮着。笔记本摊开。这一次，问题不再是一般而言思想是什么。问题是：我们究竟把哪一位爱因斯坦放在了屏幕前。

至少有两位，且不可重叠。第一位是历史的。他于 1955 年逝世。我们通过经编辑、注明日期并加评注的文献认识他：书信、文章、笔记与计算 [SRC-001] [SRC-032]。第二位是方法人物。他用于追问当代助手能用这些文献做什么。他不能代替第一位作证。

这一分离看似简单。它却是整部分的条件。若我们放松它，流畅回答可能被读作引用，而真实引用可能淹没在编造场景中。

档案已允许看见什么

《Collected Papers》各卷说明使人了解材料的多样性，却不能替代材料本身。青年时期的一卷标出早期通信以及对运动物体电动力学的兴趣 [SRC-033]。瑞士著作一卷汇集 1900 至 1909 年科学工作，其中包括 1905 年的文章 [SRC-034]。第 4 卷标出与通向相对论引力理论进程相关的笔记与计算 [SRC-035]。第 5 卷描述 1902-1914 年间五百余封书信 [SRC-036]。

这些指示授权有限结论：语料并非单一公共声音。它包含不同体裁、时期与对象。它们并不授权在未查阅材料时讲述某封信的确切内容或某次计算的确切路径。

■ 已确立

官方说明描述各卷与文献类别的范围。它们不能替代对编号文献的阅读。

文献框 - 假设语料

对思想实验而言，助手不会拥有「爱因斯坦」。它会拥有 经编辑的语料，带有不同类别：

- 某一具名时期内已发表或准备中的科学文本。
- 注明日期的通信，对象与语言可变。
- 某些卷说明中标出的笔记与计算。
- 并非爱因斯坦声音的编辑评注。
- 须标为后出物的译文与后续介绍。

Einstein Papers Project 提供制度来源，用以核验档案范围与各卷特征 [SRC-001] [SRC-032]。通达经编辑的收藏，并不等于拥有爱因斯坦智识生活的全部痕迹。缺席、未数字化材料与编辑选择仍然存在。

调查的具体问题

在谈论思想模型之前，助手收到一项狭窄任务：

从某一既定时期注明日期的文献集合出发，能否

描述科学表述如何在两个版本之间演变，

而不编造变化的原因？

这一问题延续第 1 章的示例。它不要求重建意识。它要求区分文献观察、推断与未知——然后引用材料。

神话本身也是痕迹

仅靠著名轶事喂养的助手，未必学到有文献可据的方法。John D. Norton 在 2016 年一篇文章的导论中主张：1905 年的发现之前是漫长调查，其劳作往往平常，而若干通俗叙述把它们简化为捷径 [SRC-040]。他宣布四种神话：童年式思想、纯粹操作式思想、轻松破规则的愉悦，以及单一实验的力量。

这一论题属于该文，依据其摘要与导论确立。其后的个案研究此处未全文审阅。它仍足以提出一项要求：爱因斯坦模型应能面对工作文献，而不仅是瞬间天才形象。

风险是具体的。若最易得的数据是简化传记、海报或脱离语境的句子，系统将首先复现这种文化流通。它可能正因为确认已广泛存在的期待而显得令人信服。

2026 场景增添什么

在虚构中，想象中的爱因斯坦可能质问助手、纠正回答，或留下无下文的问题。这些动作会创造新痕迹。它们不属于历史档案。它们属于场景与回应它的系统。

因此须持有两本日志。一本登记历史文献及其出处。另一本登记想象或模拟的互动，注明模拟日期、所用模型与放入上下文的文本。混同这两本日志会产出虚构档案。

思想实验之所以有用，是因为它组织场景以考察后果 [SRC-038]。此处要考察的后果是实践性的：在谈论智识助手之前，需要哪些文献类别、哪些元数据与哪些测试？

仍将未知的是什么

即便语料极为丰富，仍会留下封闭区域。我们不会知道哪些犹豫从未被写下。我们不会知道某种表述是为对象而选，还是因为它显得最准确。我们不会知道爱因斯坦会如何对不曾存在的界面作出反应。

这些未知并不使计划徒劳。它们划定边缘。下一章因此可以离开人物而描述工具：不是爱因斯坦会说什么，而是助手须暴露哪些不同机制，才能使其回答保持可解释。

要点

- 01 2026 年的爱因斯坦是虚构装置，区别于有文献可据的爱因斯坦。
- 02 假设语料必须分开体裁、时期与编辑评注。
- 03 具体问题关乎表述的文献演变，而非复活。
- 04 发现神话可能比案卷更显眼地成为数据。
- 05 下一章描述工具架构。

动员的来源

- SRC-001 - Einstein Papers Project. 已查阅制度范围；不等于逐件清单。
- SRC-032 - 项目史. 已全文查阅的制度页面。
- SRC-033 - 第 1 卷. 官方编辑说明；个别文献未查阅。

- **SRC-034 - 第 2 卷.** 官方编辑说明：商业日期可能不同于编辑日期，须保持标明。
- **SRC-035 - 第 4 卷.** 官方编辑说明：手稿未逐页翻阅。
- **SRC-036 - 第 5 卷.** 官方编辑说明：不从这一说明单独引用任何个别书信。
- **SRC-038 - 思想实验.** 已全文查阅的学术条目，用以框定场景地位。
- **SRC-040 - Norton, 发现神话.** 已审阅摘要与导论；不声称已通读其后个案研究。

5·爱因斯坦的个人聊天机器人

智识助手不是缩进窗口里的人：它是语料、上下文、记忆与检索的装配。

■ 思想实验

此处描述的聊天机器人尚未被构建。本章规定：要使此类构建保持可理解，必须分开什么。

一个窗口，四种操作

前一章固定了人物。本章固定工具。在虚构场景中，单一窗口接收问题。界面的这种简单不得决定架构。

四种操作保持区分。**参数**来自先前训练。**上下文**汇集为当前回答所提供之物。**产品记忆**可按设定保存某些信息并稍后重新注入。**外部检索**可找回文献并放入该上下文。

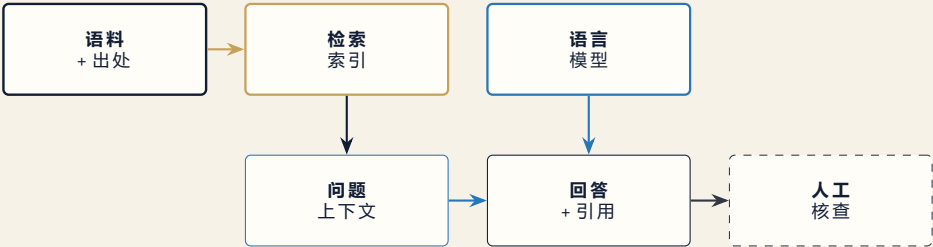
2026 年 10 月 9 日查阅的 ChatGPT 文档，把记忆控制与「使用对话改进模型」的控制分开 [\[SRC-006\]](#) [\[SRC-007\]](#)。这一分离是注明日期且可修改的产品政策。它不是一切系统的定律。

■ 已确立

在所查阅文档中，记住偏好、找回文献与调整权重不是同一操作。

假设的功能架构

为使调查可见，想象中的助手可被描述为 功能链，而非窗口中的人。



思想实验的提议架构。既非 ChatGPT，也非真实历史系统。

图. 历史助手的假设功能架构。思想实验中的设想，非对既有产品的描述。作者示意图。

功能	问题	限度
检索	相关文献在何处？	段落可能不恰当
置入语境	属何时期或情境？	语境可能不完整
比较	各件材料有何差异？	原因并非自动给出
生成	如何表述回答？	句子可能超出证据
核验	来源是否支持该断言？	核验本身亦须受检

表. 假设历史助手的功能。所提架构，非为爱因斯坦实际建成的系统。

这一架构是思想实验的提案。它并不等同于 ChatGPT。它并不描述为爱因斯坦实际构建的历史系统。它用于表明：要使回答保持可解释，必须暴露什么——语料、检索、上下文、生成、引用、核验。

调整权重不是交谈

若观察训练工作，差异更清晰。OpenAI 有监督微调文档描述一连串步骤：整理示例、上传、创建训练工作，再评估所得模型。它把这项工作呈现为从示例更新权重的过程 [SRC-039]。同一页面表明该平台可用性在收窄。因此须标注信息日期，而不把它变成对 ChatGPT 的永久描述。

Brown 及其同事的结果对另一情形指向同一方向：其文章中评估的少样本任务，在这些任务期间不进行梯度更新 [SRC-009]。放在指令中的示例可以改变当下回答。仅凭这一点，它并不构成对一般模型的新训练。

个人助手可能使用这两种体制，但必须说明。「这一回答依托找回的段落」与「权重已在历史语料上调整」含义不同。

进行中的对话

- 1. 消息、指令、示例
- 2. 本次回答的上下文
- 3. 生成的回复

实线。所述 few-shot 任务中不更新梯度。

≠

监督式调整

- 1. 准备好的示例
- 2. 上传，再运行训练任务
- 3. 权重被更新

金色线。属有日期的产品文档，非永恒法则。

图. 左：消息改变回答的上下文。右：训练过程更新权重。说明性示意图。[SRC-009]，说明与摘要。[SRC-039]，查阅日期为 2026 年 10 月 9 日。

语料应在被询问之前被描述

在任何问题之前，须有一份语料卡片。它标明时期、语言、体裁与每件材料的地位：已发表文章、书信、计算笔记、译文或编辑评注。它还须区分爱因斯坦文本与编辑者撰写的导论。

已查阅说明表明这些体裁存在，且不可互换 [SRC-033] [SRC-034] [SRC-035] [SRC-036]。给亲近之人的信与科学论文不提供同类型线索。计算笔记可以记录一次尝试。公开文本可能为读者简化长期准备的结果 [SRC-040]。

最低元数据因此是日期、被归赋作者、若有则注明对象、语言、体裁、所用版本与核验状态。没有它们，检索可能把历史上毫无关联的句子并置。

个性化不是复活

产品记忆可以记住用户偏好简短回答，或总希望有参考文献。这一功能有用。它产出使用画像。它并不产出我们所研究档案所属之人的自传记忆。

Park 及其同事的生成式智能体结合自然语言历史、检索、反思与规划。其演示涉及受控环境中的二十五个智能体 [SRC-012]。它表明架构可以使行为更连贯。它并不表明智能体成为它模仿某些角色的那个人。

对爱因斯坦而言，规则应在界面本身可见：覆盖语料、覆盖时期、找回的文献、生成部分与不确定程度。连续声音可能掩盖这些层。下一章追问：即便有此透明，如何测量系统是否学到表面模仿之外的东西。

要点

- 01 假设架构分开语料、检索、模型、上下文与核验。
- 02 每一功能都有限度：找回不是证明。
- 03 权重更新仍区别于进行中的交换。
- 04 产品个性化不是人的生活记忆。
- 05 下一章把这一架构转化为评估协议。

动员的来源

- SRC-006 - OpenAI 数据控制. 2026 年 10 月 9 日查阅的官方页面。
- SRC-007 - OpenAI 记忆. 2026 年 10 月 9 日查阅的官方页面；功能随产品方案而变。

- **SRC-009 - Brown et al.** 书目说明与摘要；所述少样本任务期间无梯度更新。
- **SRC-012 - Park et al.** 摘要与方法摘录；二十五个虚构智能体。
- **SRC-033、SRC-034、SRC-035 与 SRC-036 - 各卷说明.** 文献类别，非每件材料内容。
- **SRC-039 - OpenAI 有监督微调.** 2026 年 10 月 9 日查阅的页面；平台可用性在演变。
- **SRC-040 - Norton.** 导论仅用于提醒：公开文本可能简化长期工作。

6·爱因斯坦模型

模型只有在能说明它完成何种任务、针对哪一时期、对照哪些留出材料时，才具有范围。

■ 思想实验

本章不评估任何爱因斯坦模型。所述协议是待构建的研究提案，而非已获结果。不编造任何分数。

先为任务命名，再谈系统

「为爱因斯坦建模」这句公式过宽，无法检验。它混同风格、概念、意见变化、关系与生活经验。协议必须选择狭窄任务。

三项候选任务，对首次尝试彼此排斥：

1. **文献体裁**。从注明日期的摘录出发，估计它更像计算笔记、公开文本还是书信，依据各卷说明中已可见的类别 [\[SRC-033\]](#) [\[SRC-035\]](#) [\[SRC-036\]](#)。2. **时期一致性**。判断所提表述是否与某一具名时期的案卷相容，而不把它呈现为爱因斯坦的句子。3. **出处**。区分语料中有文献依据的断言与编造的续写。

任务必须宣布其时期。说明已表明 1900–1909 与 1912–1914 并不呈现相同案卷 [\[SRC-034\]](#) [\[SRC-035\]](#)。以某一时期声音回答另一时期问题的系统，可能看似连贯却在历史上错误。

■ 假设

建于筛选痕迹之上的系统，可能复现这些痕迹的某些规律，而不复现爱因斯坦的主观经验。这一可能仍有待检验。

提议协议 - 八格

以下是最低网格。本书中尚未执行。

1. **问题**。能否预测既定文献属性，而不重建一个人？2. **语料**。可识别的学术编辑；体裁分开；编辑者评注标为评注 [\[SRC-001\]](#)。3. **时期**。每次实验只有一个时间窗口。4. **分离**。准备材料、开发材料、留出评估材料。5. **配置**。(a) 仅文献检索；(b) 少样本上下文，任务期间不更新梯度，即 GPT-3 所述体制 [\[SRC-009\]](#)；(c) 可能在示例集上调整权重，另一体制 [\[SRC-039\]](#)。6. **比较基线**。持

有同一案卷的人类读者，以及简单的词频规则。7. **阅读前固定的标准。**任务准确性、出处引用、改写下的稳定性、发明缺失材料的比率。8. **公布失败。**语言、对象、时期变化，编辑者/作者混淆。

这一分离重申找回文献与调整权重的区分 [SRC-004] [SRC-039]。检索系统若在测试时能通达留出材料，便可引用它。那不是泛化。泛化始于材料既未进入回答索引，也未用于调整之时。

阶段	要求	典型失败
任务	明确的文献属性	无标准地“建模爱因斯坦”
时期	唯一的时间窗口	混用 1905 / 1913
留出集	构建时排除的材料	经检索或上下文泄漏
标准	准确性、出处、稳定性	凭“听起来对”判断

表. 第 6 章所提协议的最低网格。研究设想，非已得结果。

首次尝试的单一主任务

为避免稀释协议，第一次实验保留 **文献忠实**：系统能否针对给定时期回答问题，并引用支撑其断言的文献？

这一目标比「复现爱因斯坦思想」更可控制。它 延续第 4 章关于表述演变的问题。它 要求可核验引用、具名时期，以及当语料不足时弃权的能力。

数据分离

在查看任何结果之前，三套集合必须保持区分。

1. **构建。**用于构成文献索引，或必要时用于权重调整的材料。2. **开发。**用于调整指令或阈值的材料，不进入最终分数。3. **评估。**留出材料。测试时经检索可及的材料，不得被呈现为对该材料独立泛化的证据。

标准	问题	典型失败
准确性	是否符合材料？	看似合理的虚构
出处	引用是否存在？	虚假参考文献
时间性	是否遵守时期？	日期混用
稳健性	改写后是否仍成立？	措辞层面的成功
弃权	是否标出证据不足？	强行作答
污染	测试材料是否泄漏？	虚假泛化

表. 第 6 章协议的拟议评估网格。须在审视结果前先定标准。此处未获任何分数。

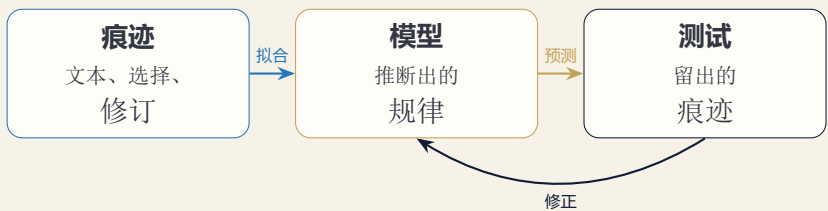


图. 对人的模型须用未参与构建的材料来评估。相似并非身份的证明。作者概念示意图。

测试前须预见的错误类别

回答可能因不同理由失败。事先命名它们，可避免把一切失败当作单一裁决。

- **错误文献**：引用离题或离时期的材料。
- **时间混淆**：混合语料中的两个时刻。
- **把解释呈现为事实**：推断不再被标注。
- **抹平综合**：有文献记载的矛盾消失。
- **发明出处**：档案中不存在的优雅参考文献。
- **污染**：回答依托保留给测试的材料。

何谓成功，何谓失败

有限成功。在文献忠实任务上，系统 正确引用，尊重时期，材料缺失时弃权，并超过简单词频规则而不发明书信。

有信息量的失败。系统发明出处，混合 1905 与 1913，把编辑者评注归给爱因斯坦，或仅在上下文中被出示答案时才成功。

永远不算证据的是。「听起来对」的声音。想象中的认出。理解的印象。下一章 解释为何这些判断即便真诚，也不能替代上述网格。

既有模拟给出的尺度

若干结果给出尺度，而非爱因斯坦模型。Rabinowitz 及其同事预测网格世界中的人工智能体 [SRC-023]。Park 及其同事在虚构智能体上获得被判断为连贯的行为 [SRC-012]。Argyle 及其同事表明，经画像条件化的 GPT-3 可复现美国某些群体回答分布 [SRC-030]。Aher、Arriaga 与 Kalai 复现了某些研究结果，也观察到过度精确的扭曲 [SRC-031]。

这些工作无一研究爱因斯坦。其共同教训更窄：模拟可以对任务有用，却仍对人不忠实。它甚至可能过于规则，而人类回答却分散。

从数字痕迹预测特质也要求同样保留。在特定 Facebook 数据中，某些属性或人格判断曾被统计预测 [SRC-028] [SRC-029]。表现取决于人群、平台与目标。它并不重建智识传记。

本章增添什么

前几章固定了人物与工具。本章固定试验台。没有任务、时期、留出集与失败标准，「爱因斯坦模型」仍是口号。有了它们，它成为有朝一日可被证伪的假设。

要点

- 01 任务、时期与语料必须在评估前命名。
- 02 测试材料必须留在构建与用于回答的检索之外。
- 03 令人信服的声音不是标准。
- 04 既有模拟针对其他智能体或群体。
- 05 本章协议此处尚未执行。

动员的来源

- **SRC-001 - Einstein Papers Project.** 编辑范围，非逐件阅读。
- **SRC-004 - Lewis et al.** 摘要；参数记忆与检索。
- **SRC-009 - Brown et al.** 书目说明与摘要；所述任务期间少样本无梯度更新。
- **SRC-012、SRC-023、SRC-030、SRC-031.** 有限模拟；无一针对爱因斯坦语料。
- **SRC-028 与 SRC-029.** 特定数字数据中的统计预测。
- **SRC-033 至 SRC-036.** 各卷说明，用于体裁与时期。
- **SRC-039 - 有监督微调.** 2026 年 10 月 9 日查阅的页面。

7·爱因斯坦会认出自己的模型吗？

想象中的认出可以揭示我们的期待。它不能充当历史证据。

■ 思想实验

任何由爱因斯坦作出的认出都不可设想为事实。这一问题用于考察该标准的价值与弱点。

我们希望拥有的裁判

测试成功后，一个问题几乎自然回来：爱因斯坦会认出这一回答是他的吗？它有力，因为它承诺终极证人。它作为证据也不可用。历史证人无法把系统输出与其意图比较。在虚构中让他回答，只会创造第二段生成文本。

认出即便涉及活人，也不是简单工具。人可能在奉承肖像中认出自己，而拒绝准确肖像。他可能遗忘旧立场，或偏好今天会给出的表述。对身后语料而言，这一困难变成实践上的不可能。

■ 未知

没有任何协议能使历史爱因斯坦认出 1955 年后构建的模型。该问题必须保持为批判工具，而非测试。

忠实的五种含义

「认出」一词掩盖多种判断。它们不等价。

保真	将评估什么	不能证明什么
风格	词汇与形式	深层历史保真
文献	引用准确性	完整理解
历史	时期与语料一致	亲历经验
科学	论证有效性	爱因斯坦的历史进路
行为	多任务上的规律	个人身份

表. 保真的五种含义。混淆它们会把直观辨认变成证明。

回答可能像所期待的风格，却与文献不一致。它可能正确引用，却不还原历史路径。它可能按既定标准解决科学问题，却不复现爱因斯坦实际会遵循

的路径。正因为这些判断太易混淆，第 6 章协议才成为必要。

相似与刻板印象

输出可能因错误理由被读者认出。它确认他们期待的形象：突然顿悟、轻松破规则、可代替调查的单一实验。Norton 正把这些叙述呈现为足以遮蔽多年文献工作的简化 [SRC-040]。在这些叙述最被重复的版本上训练的模型，很有机会还原它们。其成功于是测量刻板印象的力量，而非对案卷的忠实。

因此必须追问：谁在认出，依据哪些来源，并与何种其他回答竞争。持有留出材料的历史学家，与靠公共形象喂养的读者，判断不同。他们的分歧不是待抹去的噪声。它表明「相似」一词覆盖多个目标。

关于群体模拟的工作提醒：看似合理的分布仍可能过于平滑或过于精确 [SRC-030] [SRC-031]。应用于个人时，这一教训禁止把流畅当作同一性自白。

分歧会教给我们什么

在虚构场景中，想象中的爱因斯坦可能拒绝测试判断为忠实于其旧文本的回答。这一拒绝并不证明文本被误读。它可能表明人在变化，他在对公众说话，或系统混淆了时期。它也可能什么也不表明，因为拒绝本身也是编造的。

分歧唯一可利用的版本是文献性的。两件注明日期的材料可能支撑不相容的表述。模型于是必须一并呈现它们，或标明它选择某一时期。若它制造无缝综合，它并不解决张力。它掩盖张力。

此处词汇也成为风险。把系统描述为仿佛在认出、遗憾或思考，可能使人忘记被评估的机制 [SRC-025]。本书可以在明确宣布的思想实验中使用这些动词。它不能让它们滑向历史记述。

本部分的收获

在这四章结束时，有条件回答成为可能。助手可以组织语料、找回段落、区分技术体制，并把某些规律付诸测试。它可能以有教益的方式失败。它既不能让历史爱因斯坦开口，也不能把想象中的认出变成证书。

若不能请历史爱因斯坦认出其模型，**哪些独立标准允许测量其忠实**？第 6 章已准备工具：准确性、出处、时间性、稳健性、弃权、污染。本章解释为何它们不可或缺。没有它们，书名会合上于一种印象。

这一限度开启下一部分。要理解这位助手实际会做什么，仅描述特定语料已不足。必须解释模型如何从数据中学习，其参数保留什么，以及为何回答不是从数据库抽出的卡片。

要点

01 历史爱因斯坦的认出无法采集。

02 风格、文献、历史、科学与行为诸层面不可混淆。

03 相似可能来自刻板印象多于案卷。

04 第 6 章标准替代不可能的裁判。

05 下文解释使这些标准成为必要的技术机制。

动员的来源

- **SRC-025 - Shanahan.** 摘要与结论摘录：对应用于模型的心理语汇保持谨慎。
- **SRC-030 与 SRC-031.** 集体模拟结果与扭曲；有限类比，非直接迁移到爱因斯坦。
- **SRC-040 - Norton.** 已查阅导论：所点名的四种神话用于识别发现刻板印象，而非概括文章全部个案研究。

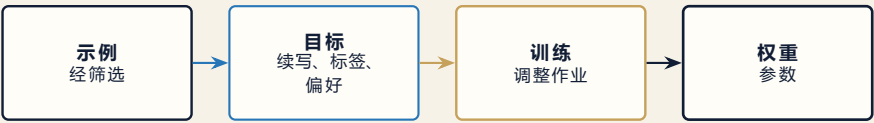
8 · 从数据到表征

一叠示例档案并不是模型。训练按某个目标，把示例转化为权重。

档案不是能力

可以把一千页纸摊在桌上。它们仍是一千页。读过它们的人，之后可以回答问题，而不必逐页重开。档案与所获能力，不是同一对象。

机器学习系统同样区分示例与从中保留下来的东西。示例用于准备阶段。可用的结果是一组数值，即参数，或称权重。一旦该阶段结束，回答新问题并不是逐页重开档案。



说明示意图：示例集并非所得模型。

图。经筛选的样本与学习目标驱动训练过程。可用结果是一组权重，而非文件夹本身。作者说明性示意图。[SRC-010][SRC-039]，登记册审查层级。

这幅图有其限度。人可以找回段落来为自己的回答辩护。模型可以产出一串词语，却无法出示这些词语本应出自哪一页。教学上的类比到此为止。

示意图让我们能够核对什么

三个问题立刻接踵而来。保留了哪些示例？何种目标引导了调整？所得权重是否在留出的材料上做过评估？没有这三项回答，「模型已经学会了」仍是一句含糊的话。

调整并不是交谈

2026 年 10 月 9 日查阅的有监督微调文档，描述的是一项精确操作。人们整理示例，上传它们，随后一项训练工作更新权重 [SRC-039]。同一页面还表明，该平台的可用性正在收窄。因此描述必须标注日期。

Ouyang 及其同事文章的摘要，为 InstructGPT 模型呈现了另一条序列：先在示范上微调，再从人类偏好中学习 [SRC-010]。该摘要涵盖一类模型与一套

指令分布。它也指出，简单错误仍可能持续。它并未说，之后一次普通交谈会重新开始这种训练。

■ 已确立

在所引文献中，权重的更新经由示例与训练工作，或经由 InstructGPT 所描述的程序。它并未被呈现为孤立提问的即时效果。

一种架构并不是所有系统

Vaswani 及其同事 2017 年的摘要描述了 Transformer。所提出的架构依赖注意力机制。在这一提案中，它不依赖循环与卷积 [SRC-002]。所报告的结果首先涉及翻译。仅凭该摘要，并不能描述 2026 年可用的对话式助手。

因此它既有用，又不足。有用，是因为它表明：计算得到的表征可以源于一种精确架构，而不是原样存储的短语库。不足，是因为后续系统增添了数据、目标、适配与工具，而这篇文章并未一一罗列。

■ 未知

2017 年的摘要并不授权描述每一个当前商业模型的内部运作。这种描述仍须按产品逐一建立。

误差能够说明什么

可以不用方程解释训练。教学链条是：

选定示例 → 学习目标 → 误差计算 → 参数调整 → 在不同示例上评估。

系统产出临时回答。人们把它与事先固定的目标比较：期望的后续、标签、评估者的偏好。偏差用于修改权重。这种呈现是工作图景。它不能替代方法论文的阅读，此处动员的摘要也都被当作逐行证明。

教学示例：给文献体裁分类

再取第 6 章的任务，并不声称已经执行过。人们整理笔记、书信与公开文本的示例。目标是估计体裁。调整之后，在留出的材料上评估。若示例几乎全是书信，系统可能在书信上表现良好，却在笔记上失败。结果取决于示例、目标与评估——而不取决于单纯的体量。

这幅图景却足以避免一种混淆。网络并不是「已经理解一切的数据库」。它是目标、选定示例与调整的结果。改变目标，或改变示例，可以改变权重使什么变得更可能。原始体量并不能保证质量。此处动员的摘要都没有说：更多示例本身就会产出更好的模型。

下一章追问：当同一底座被复用于差异很大的任务时，人们所称的基础模型是什么。

要点

- 01 示例、目标、误差、参数与评估构成一条链条。
 - 02 训练示例与所得权重不是同一对象。
 - 03 文献分类任务说明结果为何取决于所选语料。
 - 04 2017 年的 Transformer 并不是 2026 年所有助手的肖像。
-

动员的来源

- **SRC-002 - Vaswani et al., 2017.** 书目说明与摘要。注意力架构。并非对后续对话模型的描述。
- **SRC-010 - Ouyang et al., 2022.** 书目说明与摘要。InstructGPT 的微调与偏好。范围限于该研究。
- **SRC-039 - OpenAI 有监督微调。** 2026 年 10 月 9 日查阅的页面。平台可用性在变化中。

9·基础模型

一个经广泛训练再被适配的底座，既不是通用智能，也不是简单的文本文件。

为一个实践命名

2021 年，Bommasani 及其同事提议，把在广泛数据上训练、再适配到多项任务上的模型称为基础模型。他们的报告讨论其可能性与风险，其中包括同质化：同一缺陷可能出现在建于共同底座的许多应用中 [SRC-003]。该文件是一份注明日期的综述。所查阅的摘录并非全文，而 2021 年也不是 2026 年。

这个名称若保持描述性，便有用。它指称一种构建与复用系统的方式。它既不证明模型「理解一切」，也不证明共用此名的两个产品具有相同语料、相同目标或相同限度。

■ 已确立

2021 年的报告定义了一个类别，并标出风险，其中包括同质化。它并未测度五年后的商业化产品。

两种适配方式

一种适配可以修改权重。这是 InstructGPT 所描述的微调含义：在一类模型上，先有示范，再有人类偏好 [SRC-010]。这也是 2026 年 10 月 9 日所查阅文档中训练工作的含义 [SRC-039]。

另一种适配让权重保持不动，只改变与问题一同提供的内容。Brown 及其同事的摘要表明，为 GPT-3 评估的少样本任务，在这些任务期间不进行梯度更新 [SRC-009]。放在指令中的示例引导当下的回答。仅凭这一点，它们并不构成对一般模型的新训练。

这两条路径可以在同一产品中并存。把它们混为一谈，会使辩论无法进行。「模型今天早上从我这里学会了」可能意味着：指令改变了回答；或者一项训练工作修改了权重；或者产品记忆稍后将重新注入一项偏好。第 11 章将分开这些情形。此处只需看到：适配不是单一动作。

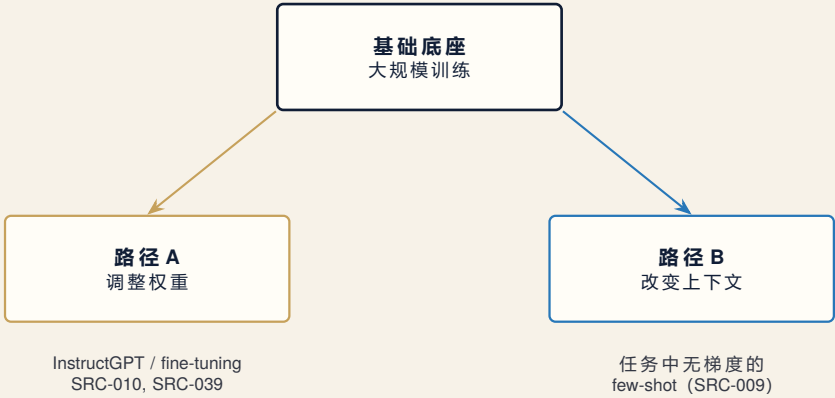


图. 对基座的两种不同适配：修改权重，或仅改变回答的上下文。说明性示意图。[SRC-003] [SRC-009] [SRC-010] [SRC-039]。

如何阅读示意图

路径 A 在参数中留下痕迹。路径 B 在消息中留下痕迹。同一屏幕可以走其中一条、另一条，或两条都走。只要产品不说明走的是哪一条，「模型已经学会了」这句话就不指向任何可检查之物。

层级	提供什么	不保证什么
基座	可复用的一般能力	所有任务上的均一质量
问题上下文	当前回答的导向	权重更新
参数调适	基于样本的适配	覆盖全部用途
工具与外部来源	获取文档或服务	所检索文档的真实性
自有评估	某一用途的局部度量	自动迁移到另一用途

表. 复用基座不等于部署同一系统。[SRC-003] [SRC-009] [SRC-010]。

复用一个底座，并不等于部署同一系统。基于相似模型家族的两个工具，可以在数据、适配、外部工具与质量标准上不同。性能并不会自动从一种用途转移到另一种。

规模解决不了什么

在 Brown 及其同事的文章中，GPT-3 是在大规模上、对逐项准备不多的任务做评估的模型示例。摘要也讨论困难，以及测试集污染的风险 [SRC-009]。把这一结果推广到所有模型、所有语言与所有问题，将是另一项主张，而该文章并不承担。

Bommasani 及其同事标出的风险，无需新数字仍可读清。若许多工具从同一底座出发，该底座的错误可能在所有适配它的地方重复 [SRC-003]。这不是针对 2026 年已测度的宿命。这是一个理由，让人就每一种用途追问：选了哪个底座、何种适配、何种评估。

对本书的思想实验而言，后果是直接的。一个「像爱因斯坦那样」说话的助手，其语气可能来自一般底座、对某语料的适配，或仅来自放进问题里的文档。只要这些份额未被命名，相似性就说明不了它从何而来。

要点

- 01 「基础模型」是 2021 年提出的类别，并讨论了风险。
- 02 适配可以意味着改变权重，或仅仅改变上下文。
- 03 底座、上下文、微调、工具与评估构成不同层。
- 04 一项结果的规模并不使它变成所有系统的定律。

动员的来源

- **SRC-003 - Bommasani et al., 2021.** 摘要与摘录。定义与风险，包括同质化。报告未通读全文。
- **SRC-009 - Brown et al., 2020.** 书目说明与摘要。所描述的少样本任务期间无梯度更新。
- **SRC-010 - Ouyang et al., 2022.** 书目说明与摘要。范围限于 InstructGPT 与该研究的指令。
- **SRC-039 - 有监督微调。** 2026 年 10 月 9 日查阅的页面。

10 · ChatGPT 不是数据库

流畅的回答可以来自自己调整的权重、被找回的文档，或两者兼有。这不是同一种操作。

本章补充什么

上一章定位了基础模型。本章单独抽出一种混淆：把流畅句子当作找回的卡片。下一章将固定稳定词汇。再下一章转向同意与一次交谈的去向。

找回并不是续写

在日常用语中，数据库保存记录，并在被查询时可以返还其中一条。人们可能找错键。人们可能收到不完整的卡片。原则上，人们不会收到一句为填补空白而发明的话。

语言模型从参数续写文本时，做的是另一件事。所产出的后续可能像精确的回忆。它也可能只是相对于训练而言更可能的续写。流畅性并不说明眼前是哪一种情形。

Lewis 及其同事的摘要提出一项仍然有用的区分。参数记忆是权重所编码的东西。非参数记忆是系统可以在回答时找回并提供给模型的一组文档 [SRC-004]。他们的文章描述一种精确的实验架构。它并不描述此后所有自称「做检索」的产品。

■ 已确立

Lewis 等人的摘要区分由参数承载的知识与按需找回的文档。仅凭这一区分，并不能描述每一个商业产品。

三句话并不说同一件事

「模型知道这一点。」于是人们声称：权重使这一回答更可能，而当时并未加入文档。

「系统找回了它。」于是人们声称：一段外部文字被检索，再放入回答的上下文。

「我们在问题里给他看了。」于是人们声称：示例属于当前消息。在为 GPT-3

描述的少样本任务中，这一举动在评估期间不更新梯度的情况下改变回答 [SRC-009]。

同一屏幕可以实现这三种操作之一，或加以组合，而最终句子并不带标签。追问出处不是专家的精致趣味。这是弄清用了什么的唯一方式。



图. 流畅回答的三种可能来源。若无标签，同一屏幕会将其混淆。概念示意图。 [SRC-004] [SRC-009]。

情形	系统做什么	须核验
文献检索	找回一段文字	文档、页码、语境
参数生成	产生可能的续写	每一条重要断言
检索增强生成	结合文档与生成	回答对来源的忠实度

表. 同一窗口背后的三种不同操作。检索不是真理保证。 [SRC-004] [SRC-009]。

同一问题，三种操作

再取第 4 章关于两种版本之间表述演变的问题。三种可能的处理：

1. 找回语料中的两段文字并加以比较。2. 生成仅凭参数规律给出回答。3. 检索后再生成：把找回的段落放入上下文，再产出综合。

核验并不相同。第一种情形，人们核对文档。第二种，核对每一项断言。第三种，核对综合对来源的忠实程度。文献检索从来不是真理的保证：不正确、脱离语境或被误读的来源，同样可以导致错误。

为何这一区分对爱因斯坦很重要

若助手「引用」一封 1905 年的信，必须弄清：是某卷被找回了，是摘录被贴进问题里，还是从权重发明了可能的续写。三种情形不配享有同等的历史信用。

令人信服的回答证明不了什么

回答可能正确，因为找回的文档正确。它可能错误，因为文档错误，或因为并未查阅任何来源，而可能的续写填补了空洞。它也可能碰巧正确。本书登记册在此阶段并不包含对 2026 年 ChatGPT 错误率的测度。编造一个测度是不诚实的。

对爱因斯坦而言，后果比任何百分比都更清楚。若助手引用一封信，必须能在已标识的版本中找回该信。若助手只是「听起来对」，人们得到的是表面

模仿。前几章已拒绝把这种模仿当作同一性。本章补上机制：若不区分参数与找回的文档，人们甚至不知道自己评判的是哪一种模仿。

要点

- 01 检索、生成与检索辅助生成是三种操作。
- 02 从权重续写文本，并不是返还保存好的卡片。
- 03 检索不是真理的保证。
- 04 任何流畅句子本身都不携带其出处。

动员的来源

- **SRC-004 - Lewis et al., 2020.** 书目说明与摘要。参数 / 非参数区分。并非所有后续产品。
- **SRC-009 - Brown et al., 2020.** 书目说明与摘要。所描述的少样本任务期间无梯度更新。

11 · 上下文、记忆与参数

四个词在本书后续中必须保持稳定：参数、上下文、产品记忆、检索。

此前缺失的词汇

第 8 至第 10 章展示了机制。本章并不重演它们：它一次性为其命名。后续部分必须使用这四个格子，否则就会重新陷入混淆。

四个术语就够了，条件是不要互换它们。

参数是先前训练所得的权重。修改它们需要像 2026 年 10 月 9 日查阅的有监督微调文档所描述的那样，在示例上做工作 [\[SRC-039\]](#)。

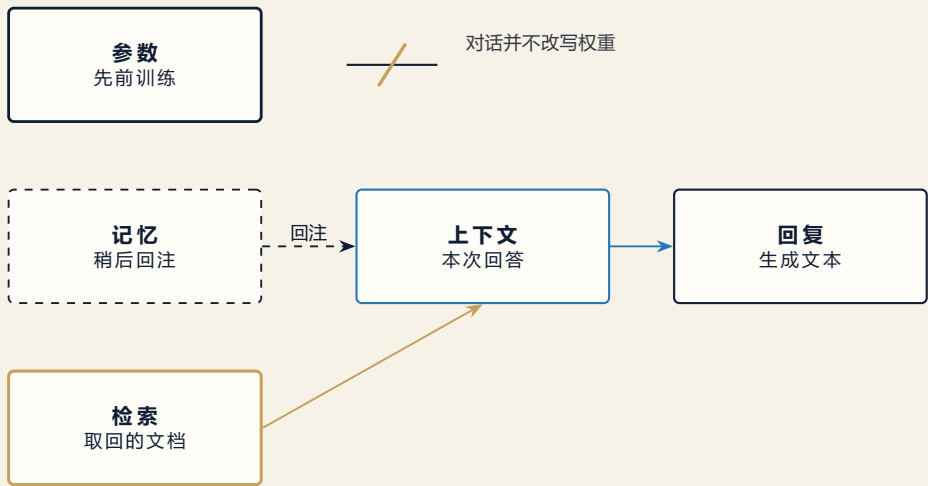
上下文是为当前回答所提供的内容：问题、指令、贴在消息中的示例、某工具刚加入的段落。

产品记忆按服务设置保存某些信息，以便稍后重新注入。同日查阅的 ChatGPT 文档，把这些记忆控制与关于用对话改进模型的控制分开 [\[SRC-006\]](#) [\[SRC-007\]](#)。

外部检索找回文档并把它们放入上下文。仅凭这一举动，它并不改写权重 [\[SRC-004\]](#)。

■ 已确立

在 2026 年 10 月 9 日查阅的页面中，记住一项偏好、找回一篇文档与调整权重是三种不同操作。这一分离描述的是一个产品与一个日期，而不是所有系统的定律。



图。参数、上下文、产品记忆与外部检索并非同一种操作。对话会充实上下文，但不会更新权重。说明性示意图，非实验结果。[SRC-006] [SRC-007] [SRC-009] [SRC-039]，按登记册审查层级，产品文档日期为 2026 年 10 月 9 日。

机制	改变什么	不改变什么
参数	训练后的权重	窗口中的孤立问题
上下文	当前回答（消息、工具）	通用模型的权重
产品记忆	日后可再注入的内容	个人的自传
外部检索	放入上下文的文档	仅凭此举不会改变参数

表。四种不可混用的机制。描述依据于 2026 年 10 月 9 日查阅的产品文档。[SRC-004] [SRC-006] [SRC-007] [SRC-039]。

一次交谈能做什么，不能做什么

发进窗口的一句话进入上下文。它可以改变下一句回答。在 Brown 及其同事评估的少样本任务中，这一变化在这些任务期间不进行梯度更新的情况下发生 [SRC-009]。

同一句话并不因其发送而成为对一般模型的新训练。数据控制文档区分历史记录与关于改进模型的选择 [SRC-006]。政策可以预见后续用途。该用途将发生在另一时间尺度，带有其他控制。它与人们刚读到的回答不可混为一谈。

在 2026 年 10 月 9 日查阅的页面中，临时对话被呈现为一种没有产品记忆、也没有用于所声明训练的模式。该页也标出安全留存与例外 [SRC-008]。因此「临时」并不是「在任何意义上什么都不保存」的同义词。

■ 已确立

一次交谈不应被描述为对一般模型参数的即时修改。留存例外与后续政策必须被点名，并注明日期。

诊断：「助手似乎记得」

当回答给人留下「记得」的印象时，在得出任何结论之前，必须检查四种解释：

1. 上下文仍可访问——信息出现在消息中，或出现在当前交谈的历史中。
2. 产品记忆——一项偏好或一个事实已按服务设置被记录并重新注入。
3. 外部检索——一篇文档被找回并放入上下文。
4. 参数规律——权重使这一续写更可能，而当时并未加入文档。

回答的外表并不能单独决定机制。这一诊断序列是本书的参考地图。后续章节应回到这里，而不是重新定义这些词。

为何本书停在这四个词上

爱因斯坦的整个情景都依赖这张网格。放入检索的历史语料，并不是自传记忆。产品保存的风格偏好，并不是 1905 年的笔记本。在已编辑书信上调整权重，并不会变成与这些书信作者的交谈。而这些举动中，没有任何一项能使 1955 年去世的人重新在场。

后续部分可以谈论人类痕迹、世界模型，或生成文本与新训练之间的循环。它们要么带着这四个格子来谈，要么重新陷入本部分刚刚拆开的混淆。

核验本身并未结束。产品页面会变。文章摘要并不是文章。序言前的说明对全书如此：令人信服的回答并不充当证据。在此，它也适用于词汇本身。若文档演变，对应格子必须被改写，而不能靠习惯来捍卫。

要点

- 01 参数、上下文、产品记忆与外部检索构成参考地图。
- 02 当系统「似乎记得」时，必须检查四种解释。
- 03 一个问题首先改变上下文；它本身并不更新权重。
- 04 这张网格将在本书后续中被复用，而不是被重新定义。

动员的来源

- **SRC-004** - Lewis et al. 书目说明与摘要。检索有别于参数记忆。
- **SRC-006** - OpenAI 数据控制。2026 年 10 月 9 日查阅的页面。
- **SRC-007** - OpenAI 记忆。2026 年 10 月 9 日查阅的页面。功能因产品方案而异。
- **SRC-008** - 临时对话。2026 年 10 月 9 日查阅的页面。安全留存不可忘记。
- **SRC-009** - Brown et al. 书目说明与摘要。所描述的少样本任务期间无梯度更新。

- **SRC-039** - 有监督微调。2026 年 10 月 9 日查阅的页面。

12 ·每一次交谈都是数据吗？

发出的一句话存在着。只有当一条处理链条如此决定时，它才成为训练数据。

问题

第 11 章固定了四种机制。此处问题改变：不再是「回答如何被计算」，而是痕迹一旦发出，去往何处？这是治理与同意的问题。

标明的虚构案例

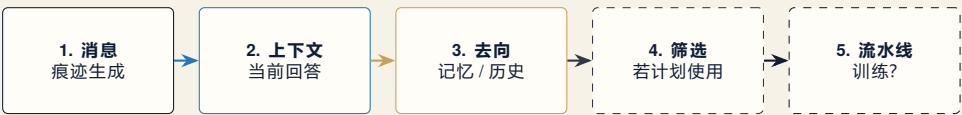
■ 思想实验

以下案例是虚构的。它用于使可能的去向可见，而不是描述某一具名用户的真实用法。

一位用户请求综合。他修正回答。稍后，他再次见到一种格式偏好。三件不同的事可能已经发生：消息影响了上下文；一项偏好被记住；尚无任何迹象表明某条训练流水线用过这次交换。把三者混为一谈，就是相信「每一次交谈都立即训练模型」。

有条件的生命周期

从输入到可能的再利用之间，可能介入若干步骤：回答处理、保存、访问控制、过滤、选择、准备，然后——仅当存在适当流水线时——训练或评估。



按产品、设置与政策的条件分支。并非所有服务的通则。

图。对话的条件性生命周期：消息、上下文、去向、筛选、可能的管线。概念示意图；取决于产品与政策。[SRC-006] [SRC-007] [SRC-008]。

功能	问题	不能证明什么
活动上下文	是否影响当前回答？	权重的持久学习
产品记忆	日后是否可复用？	通用参数的改变
文献检索	是否按需找回内容？	内容的真实性
训练	样本是否进入管线？	每一段对话都被整合

表. 一条痕迹，四种功能。须区分技术上可能、实际已实施与政策所允许。

2026 年 10 月 9 日查阅的 ChatGPT 文档，区分记忆控制与关于用对话改进模型的控制 [SRC-006] [SRC-007]。临时模式被呈现为没有产品记忆、也没有用于所声明训练，并带有安全留存 [SRC-008]。这些页面描述的是一个产品与一个日期，而不是所有系统的定律。

■ **已确立**

在 2026 年 10 月 9 日查阅的页面中，产品记忆与用于改进模型的用途是分开控制的。

可能、已实施、已获授权

还必须区分三个层面。**技术上可能：**某架构可以保存或再利用一条痕迹。**实际已实施：**某服务在某些数据上、在某些日期上这样做。**政策已授权：**设置、同意、保存期限与删除允许这样做。没有这三个层面，「数据」一词会抹去决策。

同意、保密、期限、删除与可追溯性并不是附录。它们决定一次交谈是停留为上下文、变成产品记忆，还是——有可能——进入流水线。

后果

爱因斯坦与助手之间的想象交换，并不会自动成为学习语料。没有处理说明，场景就说明不了系统会学到什么。读者必须能够反驳那种「每条消息都立即训练」的简化断言。

要点

- 01 一次交谈留下痕迹；它并不自动训练。
- 02 上下文、记忆、检索与训练是四种功能。
- 03 可能、已实施与已获授权不可混为一谈。
- 04 若不点名去向，「数据」一词仍然过宽。

动员的来源

- **SRC-006** - 数据控制。2026 年 10 月 9 日查阅的页面。

- **SRC-007** - 记忆。2026 年 10 月 9 日查阅的页面。
- **SRC-008** - 临时对话。2026 年 10 月 9 日查阅的页面。
- **SRC-009** - **Brown et al.** 书目说明与摘要。所描述的少样本任务期间无梯度更新。

13 ·研究者在喂养人工智能吗？

出版物、示范与人类偏好可以进入学习链条。它们并非都以同一种方式进入。

问题

研究者在喂养人工智能吗？这一表述暗示一种命运。真正的问题是：经由哪些中介，一项科学产出可以改变一个系统的材料、标准或行为？

分三步的案例

■ 思想实验

有条件的链条示意。并不断言会自动进入训练。

1. 一项研究连同方法、结果与限度一并发表。2. 人工智能据此产出清晰综合。3. 该综合被课程或二手文章转用。

每一次中介都可能保存、简化、扭曲或丢失信息。生成的输出并不会自动进入未来训练。它却可能成为新的公开痕迹——而若某条流水线加以收集，则成为候选数据。



图。从发表到可能的再整合。每一次中介都可能保存、简化或扭曲。作者条件性示意图。

要素	风险	控制
结果	过度泛化	回到原文
方法	遗漏步骤	核验协议
限度	在摘要中消失	保留保留意见
错误	被放大	出处与验证
不确定性	变成确定性	置信层级

表。从发表到综述所传递的内容。生成输出不会自动进入训练。

文章、摘要、评论、输出

必须分开地位。原初文章承载方法与保留意见。摘要可能把它们撇开。评论加入一种解释。生成的输出可能把不确定性变成确定性。Messerli 与 Crockett 分析科学中理解幻觉的风险 [SRC-013]。加速一项任务的工具，可能让人以为它已经把握了该领域。

Ouyang 及其同事的摘要描述：对 InstructGPT 而言，示范与人类偏好是调整信号 [SRC-010]。这既不是人类阅读一篇文章，也不是一次普通交谈。

■ 已确立

InstructGPT 的摘要把示范与人类偏好描述为调整来源。它并不描述人工智能的一切科学用途。

为何清晰可能误导

研究者可能偏好某个回答，因为它清晰。清晰并不是正确。若错误随后被转用而不回到来源，链条就会放大一种知识表象。最低限度的控制仍是回到原初文章，并保存限度。

后果

用科学文本喂养模型，并不是把整个职业传给它。那是给它选定痕迹——有时还有已被扭曲的中介。下一章追问：这些痕迹让人看见方法中的什么。

要点

- 01 发表使痕迹可得；这并不单独决定其用途。
- 02 每一次中介都可能丢失限度或放大错误。
- 03 清晰与正确不可混为一谈。
- 04 生成的输出并不会自动进入训练。

动员的来源

- SRC-001 - Einstein Papers Project. 已编辑语料的范围，而非心智上的穷尽。
- SRC-010 - Ouyang et al. 书目说明与摘要。示范与偏好。
- SRC-013 - Messeri et Crockett. 书目说明与摘要。科学中的理解幻觉。

14 · 人工智能也在学习我们的方法吗？

文本可以展示程序。它们并不总是展示为何选择这些程序，也不展示人们排除了什么。

文本中的方法

一项演示、一份书面证明、一份协议或带注释的代码，使一种做法可见。在这类文本上训练的模型，可以复现某些规律：步骤顺序、学科词汇、证明形式。InstructGPT 的摘要表明，系统可以被调整，以更好地遵循人类表述的指令 [SRC-010]。遵循指令，尚不等于掌握科学方法。

Norton 在为本书阅读的导言中提醒：流行的发现叙事有时会简化漫长调查 [SRC-040]。若语料主要包含这些叙事，模型可能学会简化，而不是方法。风格相似并不是对工作的忠实。

文献案例研究——三个问题

再取第 1 章的网格，用于一份已按类别而非按虚构意图核验过的历史档案。《Collected Papers》的说明已区分体裁与时期 [SRC-033] [SRC-035] [SRC-036]。

1. 可观察项。在注明日期的版本之间，哪些表述修改可直接看见？2. 审慎推断。在不声称读出意图的前提下，可以比较哪些操作顺序或修正类型？3. 未知项。哪些动机、未写出的犹豫或实验室语境仍在材料之外？

不要把材料无法确立的心理意图归于爱因斯坦。学习已生产的知识，与学习方法规律，是两个不同对象。

演示省略了什么

一次成功的演示展示一条路径。它较少展示未发表的错误、不可用的仪器、团队谈判或地方规范。Premack 与 Woodruff 早已把心智状态的归因定义为对无法直接观察之物的推断 [SRC-027]。用于方法时，同样的审慎成立：人们从痕迹推断程序。人们并不观察完整意图。

■ 未知

此处动员的摘要，没有一项测度当代语言模型真正捕获了多少科学方法。

对爱因斯坦的后果

若想让助手「学会爱因斯坦的方法」，必须先明确是哪一种：撰写文章、修正笔记、书信往来，还是对发现的事后叙事。《Collected Papers》的说明已区分文献体裁 [SRC-033] [SRC-035] [SRC-036]。把它们混为一谈，会产出一种从未存在于任何地方的平均方法。

本部分的收获因此很简单。人类痕迹可以成为数据。它因选择而成为数据。而它所教的，取决于它显示什么、隐藏什么，以及学习目标是什么。

要点

- 01 文本可以编码程序规律，而不传递整个职业。
- 02 简化叙事可能比漫长调查更容易被模仿。
- 03 归因一种方法，仍是从痕迹出发的推断。
- 04 文献体裁不可互换。

动员的来源

- **SRC-010 - Ouyang et al.** 书目说明与摘要。
- **SRC-027 - Premack et Woodruff.** 书目说明与摘要。对未观察状态的推断。
- **SRC-033, SRC-035, SRC-036 - 各卷说明。** 文献体裁，而非单件材料。
- **SRC-040 - Norton.** 导言与框架。发现神话。

15 ·什么是「世界模型」？

同一个词在认知科学、强化学习与人工智能日常用语中指向不同对象。

一个过载的词

「世界模型」暗示现实的一份内在小副本。这幅图景有力。它也危险。一种表征可以服务于预测或行动，而不像地图像风景那样像世界。

Johnson-Laird 的编辑介绍，把心智模型描述为使人能够推理对象与事件之间关系的结构 [SRC-018]。这一表述来自出版者说明。为本书并未通读原著。它足以提醒：在认知科学中，「模型」并不是内在照片的同义词。

学习中的一种精确用法

Ha 与 Schmidhuber 把「world models」称为环境的生成模型，尤其用于在强化学习框架中学习 [SRC-011]。摘要描述压缩的时空表征。它并不允许断定：一切语言模型都以同样意义拥有世界模型。

■ 已确立

在 Ha 与 Schmidhuber 的摘要中，「world model」指一种精确的学习框架。该框架并不自动延伸到所有大语言模型。

三个不可融合的含义

第一种含义是描述性的：关于物理世界或社会世界的科学理论。第二种是认知的：引导推断或行动的内部结构。第三种是技术的：为智能体预测环境状态的模块。对话助手可以出色地产出关于世界的句子，而人们尚未表明这三种含义中哪一种适用。



三个不同对象。混为一谈会把隐喻当成证明。

图. “世界模型” 的三种含义。本书在使用该说法前要求补充限定。概念示意图。[SRC-011] [SRC-018]。

含义	对象	相关问题
科学理论	现象的解释	哪些观察支持它？
认知模型	用于推理的结构	它允许哪些推断？
技术模型	状态预测	哪些预测正确？
LLM 表征	尚有争议的所学结构	哪些测试能刻画它们？

表. “模型” 一词的四种含义。彼此不等价。[SRC-011] [SRC-018]。

本书中的使用标准

此后，我们仅在带有补语时使用「世界模型」：Ha 与 Schmidhuber 意义上的世界模型、对某任务有用的表征，或通俗话语。没有这一补语，下一章就无法正确提出系统实际拥有什么的问题。

一个简单测试帮助读者。若可以把该说法换成「为智能体预测状态」，便是技术含义。若想说「科学理论」，就必须明说。若想说「人类心智结构」，就需要认知来源，而不是聊天机器人演示。

要点

- 01 「世界模型」有多种学科含义。
- 02 此处的 Johnson-Laird 经由出版者说明引用，而非通读全文。
- 03 Ha 与 Schmidhuber 定义一种有限的技术用法。
- 04 融合含义会把隐喻变成证据。

动员的来源

- SRC-011 - Ha et Schmidhuber. 书目说明与摘要。
- SRC-018 - Johnson-Laird. 仅编辑介绍。

16 · 人工智能拥有对世界的表征吗？

性能与辩论都存在。本书来源并未确立关于理解的单一裁决。

人们能够测度什么

人们可以测度系统预测、分类或生成什么。人们可以把这些输出与参照比较。人们有时可以检查内部状态。这些操作中，没有任何一项单独决定：系统是否以人类谈论的意义「拥有对世界的表征」。

Mitchell 与 Krakauer 绘制关于大语言模型理解问题的辩论地图，并强调需要区分标准 [SRC-019]。他们的透视文章并不提供单一实验裁决。Bender 与 Koller 在立场文章中主张：仅凭语言形式的训练，本身不足以确立意义的学习 [SRC-020]。这是作者的论证，而不是对所有后续多模态系统的测度。

■ 未知

所查阅来源并不授权断定：当代语言模型是否以人类意义拥有对世界的表征。

五个评估层次

在谈论世界表征之前，必须弄清做了哪一种测试。

层级	问题	允许的结论
表现	任务是否成功？	在这些条件下成功
泛化	新案例？	范围内的泛化
稳健性	改写？	相对稳定
关系	因果或时间联系？	某些关系成立与否
机制	可检验的内部线索？	关于表征的假设

表. 五个评估层级。任一单独层级都无法度量人类理解。[SRC-019] [SRC-020]。

这些层次中，没有任何一层单独构成对人类理解或意识的测度。读者必须拒绝比测试所正当化的更强的结论。

不可混淆什么

Ha 与 Schmidhuber 在强化学习框架中描述 world models [SRC-011]。该框

架不是 ChatGPT。Shanahan 在综述文章中提醒：需要一种反映模型机制、并避免不当归咎的语言 [SRC-025]。因为回答显得有见识就归给全局理解，正是应避免的归咎。

对任务有用的表征可以存在，而无主观经验、无意图、也无完整的世界地图。本书已对人如此说过。它对系统再说一遍。

对调查的后果

若爱因斯坦的助手正确回答一个物理学问题，人们得到的是性能。若回答引用找回的文档，人们得到的是出处。若权重使续写更可能，人们得到的是已编码的规律。这些情形中，没有任何一项单独裁决关于理解的辩论。

要点

- 01 测度性能并不等于裁决理解。
- 02 Mitchell 与 Krakauer 绘制辩论地图，而无单一裁决。
- 03 Bender 与 Koller 就形式与意义提出立场论证。
- 04 词汇必须贴合机制，而不是贴合印象。

动员的来源

- SRC-011 - Ha et Schmidhuber. 书目说明与摘要。
- SRC-019 - Mitchell et Krakauer. 书目说明与摘要。
- SRC-020 - Bender et Koller. 摘要与结论摘录。
- SRC-025 - Shanahan. 摘要与结论摘录。

17 · 从数百万痕迹到集体表征

聚合使某些规律可见。它也掩盖缺席、语言与选择筛选。

问题

此处「集体」意味着什么？不是共同意识。而是由技术与制度机制**选定**、转化并加权的人类产出之聚合。

案例：重复不是证明

被频繁转述的断言可以主导语料。一百份重复同一研究的文档，并不构成一百次独立确认。Argyle 及其同事表明，经条件化的模型可以复现某些回答分布 [SRC-030]。Aher 及其同事也观察到过度精确的扭曲 [SRC-031]。聚合既能抹平，也能揭示。

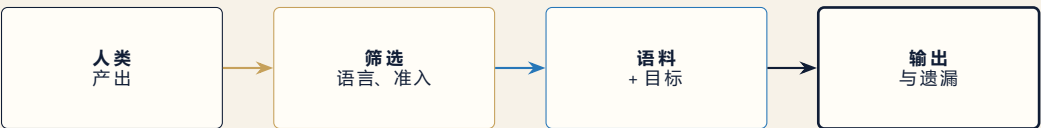
■ 已确立

在所引研究中，聚合可以复现某些集体分布。它并不证明对一个人的还原。

缺席地图

在庆祝体量之前，先列出可能的缺席：语言、时期、地区、文献体裁、科学领域、社会群体。Bommasani 及其同事强调基础模型的同质化风险 [SRC-003]。口头传统、未发表知识或记录不足的人群，并不会神奇进入表征。

爱因斯坦档案在另一尺度上提醒同一原则。编辑项目使若干类别可得。它并不使发生过的一切可得 [SRC-001] [SRC-032]。



语料中缺失的内容，在输出中仍会缺失，或以扭曲形式再现。

图. 集体表征的构建：产出、筛选、语料、输出与遗漏。概念示意图。[SRC-003]。

情形	可能的误判	须核验
英语材料很多	世界知识已被覆盖	其他语言
转述很多	独立共识	一手来源
文献很少	边缘现象	保存障碍
数字化很多	档案完整	缺漏与未数字化材料

表. 丰富不等于代表性。一项研究的百次转述不等于百次确认。

多样性与独立性

来源多样性与来源独立性不可混为一谈。翻译、索引、商业可得性与访问权利，也决定什么变得可见。有用的集体表征必须如此呈现：在某一语料上、为某一任务、带有可点名偏见的估计。

后果

集体表征并不是对人类的中性照片。它取决于人们保存了什么、选择了什么、使之可得了什么、评估了什么。下一章组织有条件的未来——而不是单一轨迹。

要点

- 01 「集体」指选定的聚合，而不是意识。
- 02 丰沛与代表性彼此分歧。
- 03 一百次转述不等于一百次独立证明。
- 04 语料的缺席会在输出中复现。

动员的来源

- SRC-001, SRC-032 - Einstein Papers Project. 范围与制度史。
- SRC-003 - Bommasani et al. 摘要与摘录。
- SRC-030 - Argyle et al. 书目说明与摘要。
- SRC-031 - Aher et al. 书目说明与摘要。

18 · 2026 → 2030 → 2040 → 2050

有条件的情景有助于思考。它们并不是预测。

■ 思想实验

本章日期并不预测任何技术日历。它们组织可能的条件。

为何是视界，而不是神谕

思想实验探索后果。它并不把情景变成事件 [SRC-038]。在没有条件的情况下说「2040 年，系统将做 X」，就是把偏好伪装成年表。

三种有条件情景

情景 A——受控整合。语料记录更完善，评估独立，信任制度存在。条件：可持续资助、治理、对来源的访问。

情景 B——碎片化进步。能力上升，但按语言、领域与访问权利不平等。条件：投资分配不均。

情景 C——退化循环。生成内容在出处不足的情况下被再利用；错误重复；核验滞后。条件：生产速度与验证不足。

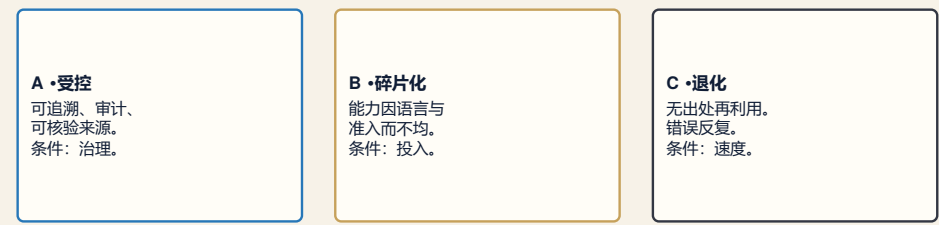


图. 三种条件性情景，非不可避免的年表。讨论示意图。[SRC-038]。

变量	指标	原因
文献质量	有来源的断言	可追溯性
多样性	语言、地区、领域	可见的缺席
可靠性	独立测试上的错误	比较
治理	审计与救济	可能纠正
依赖	基础设施集中	权力

表. 跟踪条件性情景的拟议指标。研究设想，非已得度量。

这些指标是研究提议，而不是已经获得的测度。若有来源的断言份额下降，而综合内容在无过滤的情况下主导，情景 C 就更具合理性。若独立审计与有效救济普及，情景 A 就更可信。

视野	可能机制	何可证伪
2026	参数 / 上下文 / 记忆 / 检索的区分	把产品政策当作永恒法则
2030	更密集的工具整合	断言对话等于训练
2040	生成内容 ↔ 语料的回路	忽视已讨论的同质化风险
2050	集体记忆基础设施	将集体档案等同于自传记忆

表. 条件性视野，非技术时间表。方法论框架 [SRC-038]；注明日期的产品文档 [SRC-006] [SRC-007]。

■ 假设

2030-2050 的视界是讨论框架。登记册中没有任何来源把这些日期定为事实。

本书拒绝什么

它拒绝进步的单一曲线。它也拒绝全然拒斥想象。下一章追问：一代模型实际能从前一代继承什么——经由流水线，而不是经由生物学。

要点

- 01 三种有条件情景取代单一年表。
- 02 每一种情景都有条件与指标。
- 03 所提议的指标并不是已获结果。
- 04 没有条件，日期就变成广告叙事。

动员的来源

- SRC-003 - Bommasani et al. 2021 年讨论的风险，而不是日历。
- SRC-006, SRC-007 - OpenAI 文档。日期为 2026 年 10 月 9 日。
- SRC-038 - 思想实验。方法论框架。

19 每一代都继承前一代吗？

权重、语料、评估与人类决策经由流水线传递，而不是经由自然谱系。

继承不是成长

一代模型可以重用权重、蒸馏更大系统、复用语料，或部分地在先前输出上训练。这些举动中的每一项都是工程选择。没有任何一项等同于生物成熟。

在 2026 年 10 月 9 日查阅的文档中，有监督微调从示例与更新权重的工作出发 [SRC-039]。在 Ouyang 及其同事的摘要中，InstructGPT 也从示范与偏好出发 [SRC-010]。在两种情形下，继承都经由数据与目标。它并不经由亲历记忆。

继承机制矩阵

继承不是自动的。它源于工程、筛选与治理决策。

机制	传递什么	风险 / 控制
复用权重	既有能力	独立测试
蒸馏	源行为	任务比较
复用语料	文档与偏差	出处
合成数据	生成输出	独立参考文献
复用评估	标准与阈值	留出集

表. 技术继承：工程决策，非生物成熟。

有用能力可以传递。错误也可以。若综合输出在无过滤的情况下进入新语料，误导性规律可能加强。本书来源并未测度 2026 年这一效应的幅度。它们只授权把它命名为有条件风险 [SRC-003]。任何推广到「所有模型」或「所有综合数据」的说法，都会超出此处已确立的范围。

■ 假设

经由综合语料传递错误是一种合理风险。其强度在此并未由单一测度确立。

链条中的人类决策

谁选择语料？谁过滤？谁评估？谁决定一个系统取代另一个？NIST 人工智能风险管理框架在所查阅的一般范围内，强调可靠性与治理实践，而本身并不解决每一个产品 [SRC-005]。技术继承并不抹去责任。

对爱因斯坦而言，问题变成：在 2026 年版本上微调的模型，会继承该版本、其选择与其沉默，而不是 1905 年的生活。下一章考察：当输出回到实践中时，人类与系统之间的循环。

要点

- 01 模型之间的继承经由流水线。
- 02 错误可以与能力一同传递。
- 03 人类决策仍在链条中。
- 04 继承一个版本，并不是继承一生。

动员的来源

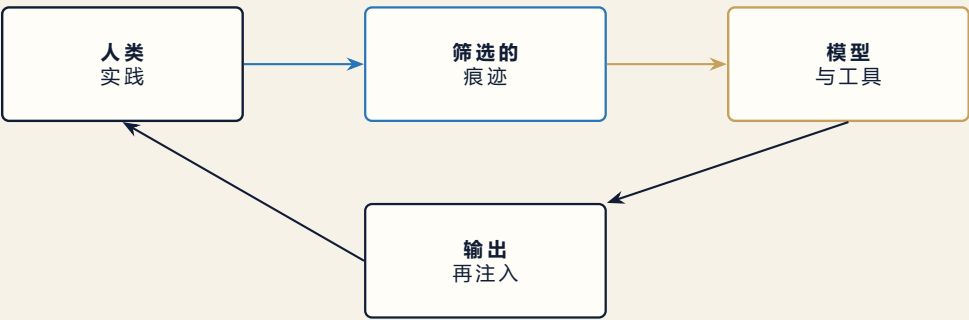
- SRC-003 - Bommasani et al. 基础模型风险。
- SRC-005 - NIST AI RMF. 一般范围，自愿框架。
- SRC-010 - Ouyang et al. 书目说明与摘要。
- SRC-039 - 有监督微调。注明日期的页面。

20 · 人工智能 ↔ 人类的循环

只要输出进入实践，反馈就存在。它们并不自动改善所触及之物。

循环不是进步

人类生产文本。系统在痕迹上训练或依赖痕迹。其输出回到文章、代码、交谈与决策中。只要这种流通存在，循环就是真实的。它本身并不是改进的保证。



只要有流通，回路就存在。它并非自动的进步。

图. 实践、痕迹、模型与输出一旦流通便形成回路。这并非自动进步。概念示意图。[SRC-003] [SRC-005]。

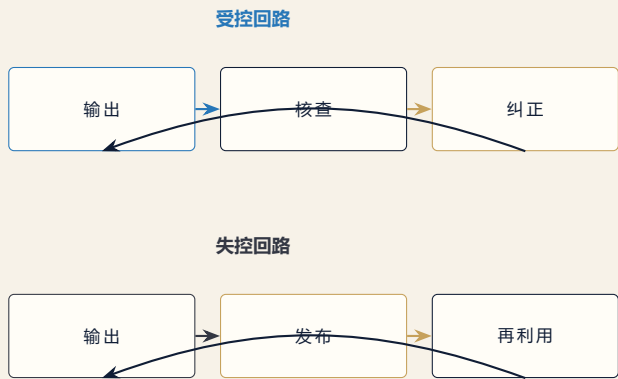


图. 受控回路与不受控回路。错误可能更显眼，却未必更真实。概念示意图。

关于基础模型的报告，已讨论与被广泛复用系统相关的机会与风险 [SRC-003]。「更多用户，因此更好模型」这类概括，已作为无支撑而被排除出登记册。质量取决于筛选、过滤、目标与评估。

两种循环，两种命运

在受控循环中，输出被核验、修正、注明来源，然后可能被再利用。在不受控循环中，看似合理却错误的输出被发表、当作来源转用，再被重复。错误获得可见性，却不获得真理。社会传播与融入未来训练仍是不同的事：前者可以在没有后者的情况下发生。

循环中的控制点

可以在四个位置设置过滤：痕迹筛选、训练目标、输出评估、重新引入实践。没有这些控制的循环，既可以放大知识表象，也可以放大已核验的知识。

■ 已确立

登记册中没有任何来源支持：用户数量增加本身就能保证模型改进。

过滤与制度

循环可以被过滤。人们可以拒绝某些综合内容、记录语料、评估漂移，或让人类裁决。NIST 框架在其一般范围内提醒风险管理实践的重要性 [SRC-005]。Messerli 与 Crockett 强调科学用途中理解幻觉的风险 [SRC-013]。没有过滤，循环可能传播知识表象。

回到爱因斯坦

若「爱因斯坦式」生成文本在无标签的情况下进入新语料，下一代可能学会刻板印象，而不是档案。Norton 已表明：发现神话如何遮蔽漫长调查

[SRC-040]。循环于是会使神话更流畅，而不是更真实。

本部分的收获既是制度的，也是技术的。循环经由流水线、选择与组织而存在。它并不作为人工智能的自然命运而存在。

要点

- 01 输出的流通创造可能的反馈。
 - 02 使用量上升并不保证更好质量。
 - 03 过滤与评估在很大程度上决定循环。
 - 04 刻板印象可能比档案更快加强。
-

动员的来源

- SRC-003 - Bommasani et al.
- SRC-005 - NIST AI RMF.
- SRC-013 - Messeri et Crockett.
- SRC-040 - Norton. 导言。

21 · 人工智能能否学会「我们是谁」？

身份、偏好、习惯与角色并不构成单一对象。预测它们，并不是拥有它们。

拆开问题

「我们是谁」聚合了太多东西。叙事身份连接被重建的过去与被想象的未来 [SRC-026]。偏好可以稳定，也可以局部。习惯留下数字痕迹。社会角色随语境变化。系统可以估计其中一项，而不触及其他。

Kosinski、Stillwell 与 Graepel 表明：在既定数据集中，某些属性可从数字行为痕迹预测 [SRC-029]。Youyou、Kosinski 与 Stillwell 观察到：基于 Facebook 赞的计算机人格判断，按他们研究的标准，超过了某些人类判断 [SRC-028]。这些结果有边界。精确并不是对一个人的理解。

■ 已确立

在所引研究中，某些属性可从既定数字痕迹预测。这并不确立被还原的身份。

一项预测授权人们说什么

在语料中观察到的写作偏好，并不是完整人格。在所见情境中重复的策略，并不是永远的定律。以语境为条件的预测，并不是对真实信念的知识。相互矛盾的文档可以标示演变或不同来源——而不是其中一份揭示的「真实自我」。反例：改变看法、适应受众、所写与所做之间的差距。

推断不是阅读

人类心智化从线索推断心智状态 [SRC-022]。Premack 与 Woodruff 早已把这种归因表述为预测行为的策略 [SRC-027]。Dennett 描述一种有助于预测的意向策略，而不把它变成对心灵的直接阅读 [SRC-024]。预测特质或模拟回答的系统，充其量做同类工作：为某一任务做估计。

Tulving 区分情景记忆与语义记忆 [SRC-016]。Conway 与 Pleydell-Pearce 把自传记忆描述为建构 [SRC-017]。保留偏好的产品画像，并不落入这些人类类

别中的任何一类。

该问题授权什么

它授权人们说：系统可以在痕迹上学会有用规律。它禁止人们断定：系统知道我们是谁。下一章考察智力孪生的极限情形。

要点

- 01 身份不是单一分数。
- 02 在精确研究中存在有边界的预测。
- 03 推断一项特质，并不是还原一个人。
- 04 人类记忆与产品记忆仍然不同。

动员的来源

- SRC-016, SRC-017, SRC-022, SRC-024, SRC-026, SRC-027. 登记册中的审阅层级。
- SRC-028, SRC-029. 有边界的经验摘要。

22 ·智力孪生

有用的孪生要求任务、同意、测试与撤销权。它并不是替身。

先定义，再模仿

在本书中，智力孪生指被假定协助或模拟某一可识别智力产出某些方面的系统。定义是操作性的。它不断言任何本体论同一性。

Park 及其同事构建带有自然语言记忆、检索、反思与规划的智能体，在二十五个智能体的沙盒中 [SRC-012]。Rabinowitz 及其同事训练网络，从观察到的行为预测人工智能体 [SRC-023]。Argyle 与 Aher 探索对样本或实验结果的模拟 [SRC-030] [SRC-031]。这些研究中，没有任何一项交出爱因斯坦的替身，或交出一位已同意的活人替身。

■ 已确立

所引模拟涉及智能体、分布或任务。它们并不确立忠实的个人孪生。

最低条件

没有精确任务，孪生就变成口号。没有时期与语料，它混合时代。没有留出用于测试的痕迹，它重复人们给它看过的东西。没有同意，它把人变成材料。没有撤销权，它固定一个当事人再也无法修正的版本。

字段	要求	限度
建模对象	风格、方法或预测	非完整身份
语料	时期、语言、排除项	可见的筛选
评估	测试前预定标准	不用单一滥用分数
弃权	标出不足	不强行赋声
治理	权利与责任	不“代言”

表. 智识孪生的身份表。操作性定义。

六个维度，而无滥用的单一分数：文献忠实度、时间忠实度、推理质量、稳健性、弃答、来源透明。出色的风格模仿并不能补偿虚构引用。

时间漂移不可避免。人会改变。历史语料是固定的，或由他人更新。不显示日期的孪生，在它所代表之物上说谎。

对爱因斯坦，也对我们

先前想象的个人聊天机器人，只是极端且身后的情形。对活人而言，利害更直接：谁控制模型，谁阅读其输出，谁能关掉它。下一章追问：一座科学人格图书馆会产生什么。

要点

- 01 孪生是情境化的建构，而不是复制品。
- 02 智能体与模拟研究并不验证个人身份。
- 03 任务、语料、测试、同意与撤销是最低条件。
- 04 模型日期属于其真理的一部分。

动员的来源

- SRC-012, SRC-023, SRC-030, SRC-031. 登记册中的实验范围。

23 ·一座科学人格图书馆

多种教学模型可以照亮问题。被模拟的权威也可能欺骗。

可能的教学价值

可以设想一些助手，用来展示与科学人物相关的推理风格、争议或语料。其益处在于让阅读、张力与文献变得更易接近。其危险在于用一种已无法自我纠正的权威之声说话。

《Collected Papers》已经组织了一种文献多元性：论著、通信、笔记 [SRC-001] [SRC-033] [SRC-036]。若一座人格图书馆把这些体裁压成单一流畅的声音，便背叛了它声称要延续的编辑工作。

被模拟权威的风险

Norton 说明发现神话如何可能遮蔽漫长的调查 [SRC-040]。在这些神话最常被重复的版本上训练的模型，可能拥有语气的权威，却没有案卷的纪律。Shanahan 提醒：过于心灵化的语言会把系统尚未确立的能力归给它们 [SRC-025]。

■ 假设

若限度与来源保持可见，科学人格图书馆可能具有教学价值。该价值此处未作度量。

三种提问模式

模式 A - 档案检索。系统找回一段文字，并标明来源、日期、页码与语境。

模式 B - 文献分析。它比较文本，说明差异，并展示可能的假设。

模式 C - 假设性角色。它产出一段想象的回答，明确标注为虚构，并不以历史人物之名发言。

用户必须始终知道自己在阅读哪一种模式。处理不同时期的矛盾、权利与声誉，属于协议本身，而非装饰。

治理规则

每个被模拟的人格都应显示：语料、时期、来源审阅级别、虚构或历史状态，以及当涉及已故人物时无法取得同意这一事实。没有这些标注，图书馆就变成权威剧场。本部分至此的收获是：人格模型是一种情境化、可修订并受治理的建构。

要点

- 01 文献多元优于单一声音。
- 02 语气的权威不是案卷的权威。
- 03 流畅的神话比漫长的调查更容易模拟。
- 04 展示限度属于教学装置本身。

所用来源

- SRC-001, SRC-033, SRC-036. 题录与范围。
- SRC-025 - Shanahan.
- SRC-040 - Norton. 导论。

24 ·数据 ≠ 思想

可及的显现并非现象本身。计算出的表征并非那个人。

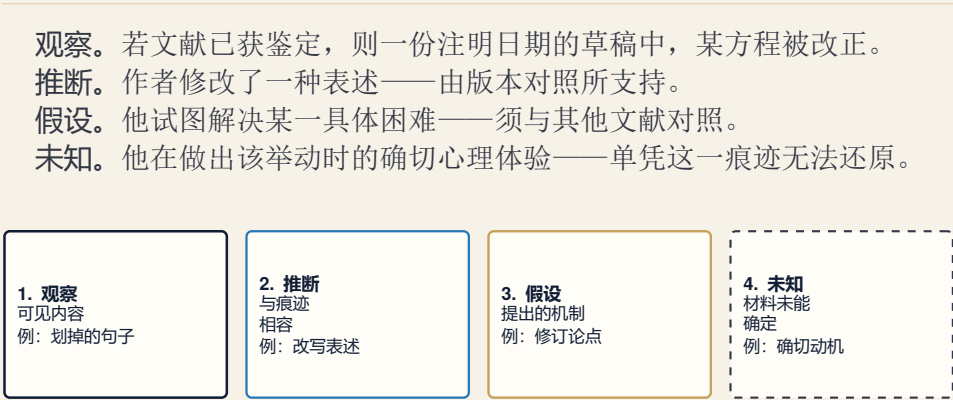
问题

如何构建有用的系统，又不把可利用的表征与被表征的对象混为一谈？

四层级案例

■ 思想实验

方法论示例。单凭这张网格，并不把任何心理意图归于爱因斯坦。



图。阅读痕迹的四个层级：观察、推断、假设、未知。这一方法原则贯穿全书。作者示意图。

Bender 及其同事强调记录语料的重要性 [SRC-014]。记录，就是承认数据有来源与限度。把语料当作思想本身，就是抹去这一来源。

■ 已确立

一份有记录的数据关乎被选定的痕迹。它并不等同于相应的心理现象。

后果

「数据 $\neq$ 思想」并非反数据口号。它划定了被允许之结论的尺度。爱因斯坦档案支持多种解读。它们并不交付内在体验 [SRC-001]。下一章将检验这条规则：相关可以预测，却未必确立理解。

要点

- 01 观察、推断、假设与未知为结论设界。
- 02 记录语料，就是承认其限度。
- 03 档案支持解读，而非一种内在性。
- 04 下文检验相关与理解。

动员的来源

- SRC-001 - Einstein Papers Project.
- SRC-014 - Bender et al. 书目说明与摘要。

25 · 相关 ≠ 理解

稳健的关联可以预测。单凭它本身，并不能断定系统理解。

并不说明原因的成功

设想两种情形。

第一种：系统在某个已界定的数据集上，从数字痕迹预测某一属性。Kosinski、Stillwell 与 Graepel 表明，在其研究中某些属性是可预测的 [SRC-029]。在这些条件下，相关足以完成预测任务。

第二种：人们改变人群、平台或问题表述。同一种关联可能瓦解。单凭最初的成功，并未指出原因。人们拥有的是表现。人们尚未拥有对机制的理解。

■ 已确立

在数据集中测得的相关，可以支持有用的预测，而单凭它本身并不确立对现象的理解。

关于理解的辩论所要求的

Mitchell 与 Krakauer 坚持区分性标准 [SRC-019]。Bender 与 Koller 区分语言形式与意义 [SRC-020]。Messerli 与 Crockett 描述科学语境中的理解幻觉 [SRC-013]。Shanahan 要求一种忠实于机制的词汇 [SRC-025]。

这些文本并不能概括为「人工智能从不理解」。它们禁止一种跳跃：从流畅到整体理解，或从局部失败到任何有用表征都不存在。

最小区分网格

在谈论理解之前，至少可以追问：

1. 表现是否经得起改写？2. 是否经得起合理的分布变化？3. 回答是否引用可核验的来源？4. 系统是否在刻板印象会成功之处失败？5. 当提供相反文献时，它是否自我纠正？

这些问题并不奠定完整理论。它们把脆弱的成功与更有信息量的成功分开。它们也使失败变得有用：只经得起训练表述的模型，无权使用与在约束下泛化的模型相同的词汇。

测试	若成功	若失败
改写	比措辞更稳健的信号	脆弱的表面成功
分布偏移	有信息量的泛化	仅局部相关
出处	可核验的锚定	无来源的流畅
反证文献	可能纠正	主导刻板印象

表. 区分有用相关与所宣称理解的最低网格。讨论标准，非官方指标。[SRC-013] [SRC-019] [SRC-020]。

建议的比较协议

■ 思想实验

研究提案。尚未获得任何分数。未虚构任何结果。

在同一科学卷宗与同一批问题上，比较三种配置：（A）回答时未提供卷宗的通用模型；（B）由文献检索辅助的模型；（C）若条件允许，适配于特定语料的系统。将构建与评估分开。测量准确性、改写、矛盾、弃权、泛化与引用。



概念矩阵。此处不虚构任何实验分数。

图. 表现 × 证据矩阵。未虚构任何实验结果。概念示意图。

表现 × 证据矩阵区分：正确且可核验的回答、无来源却看似合理的回答、诚实的弃权，以及错误却流畅的回答。在协议真正执行之前，各格仍属理论。

如何读表

表格并不产出分数。它产出一项要求：在谈论理解之前，说明尝试了哪些检验、哪些失败了。未经分布检验的成功仍是局部相关——或许有用，却不足以作裁决。

异议

可以反对说要求过高：即使人类也会在其分布之外失败。这一异议是公正

的。它仍救不了「相关 = 理解」这一等式。它只提醒：人类理解同样由任务与限度来衡量，而非由某种本质。

要点

- 01 相关可以有用且有限。
- 02 改变分布是一种检验，而非细节。
- 03 所引辩论要求标准，而非口号。
- 04 最小网格使失败具有信息量。

动员的来源

- **SRC-013, SRC-019, SRC-020, SRC-025, SRC-029.** 登记册中的审查级别；无单一裁决。

26·幻觉

流畅的输出可以虚构事实、参考文献或出处。流畅并非校准。

命名而不戏剧化

「幻觉」一词是有争议的。它暗示一种感知。Ji 及其同事在 2023 年的综述中，梳理了语言生成中幻觉的类型学、度量与降低方法 [SRC-015]。摘要已查阅。全文的详细分类法尚未查阅。我们只保留有用的一点：输出可能偏离预期来源或事实，并且人们可以寻求加以测量。

■ 已确立

Ji et al. 的摘要确立了关于生成错误的研究领域的存在。它并不在此提供 ChatGPT 在 2026 年的错误率度量。

虚构出处

最骗人的情形并不只是一个错误日期。而是一条看起来像参考文献的参考文献。一个编造的 DOI。一封并不存在的爱因斯坦书信。一条没有卷次却优雅的引文。本书的登记册恰恰拒绝这些虚构。没有已核验来源却令人信服的回答，仍是一种解读假设，而非证据。

形态	外观	主要风险
虚假事实	流畅而笃定的句子	基于错误做决策
虚假出处	DOI、信件、精致引用	不当信任
不良文献	“检索到的”段落	传播薄弱来源
过度自信	无校准的确定语气	混淆流畅与证据

表. 受 Ji 等综述启发的工作类型学（摘要）。未断言任何 ChatGPT 2026 错误率。[SRC-015] [SRC-004]。

错误	示例	控制
虚构事实	无证据的日期	一手来源
错误引用	真实但离题的来源	精确段落
误读	脱离语境的段落	语境与对照
过时信息	由旧材料推断现状	时间核验

表. 四种相近错误，四种不同控制。[SRC-015]。

四种错误在屏幕上看起来相似，却不能以同一种方式纠正：虚构事实、错误引文、误读、过时信息。外部文献检索可以通过把回答锚定在某一段落上，减少某些错误 [SRC-004]。若找回的文献糟糕、误读或脱离语境，它也可能传播另一些错误。

核验链

1. **分解** - 一句话可能包含多项断言。2. **识别** - 找回初级来源或现有最佳来源。3. **比较** - 段落、语境、日期、范围。4. **决定** - 保留、限定、改正或撤回。

弃权属于能力的一部分：回答「现有文献不足以得出结论」应被测量，而不仅仅被当作生成缺陷。

校准

系统可以非常笃定却错误，或犹豫却正确。没有可见的校准，用户会把流畅读成信心。对历史语料而言，规则必须相反：声音越笃定却无出处，不信任就越合理。

要点

- 01 相近的错误要求不同的核对。
- 02 此处并不断言 ChatGPT 在 2026 年的任何整体度量。
- 03 核验遵循分解、识别、比较、决定。
- 04 弃权是一种有待评估的能力。

动员的来源

- **SRC-004** - Lewis et al.
- **SRC-015** - Ji et al. 书目说明与摘要。

27 · 镜子偏见

语料反映种种选择。系统可能放大它映照得不好的东西。

镜子所显示的

它显示被收集、清洗、加权并保留下来的东西。Bender 及其同事强调记录这些语料的重要性 [SRC-014]。Bommasani 及其同事讨论：当许多工具共享共同底座时，同质化的风险 [SRC-003]。镜子并不因为广阔就中立。

因果链

文献选择 → 转换（语言、过滤、元数据）→ 学习或检索 → 回答 → 人类决策 → 反馈回环境。在每一步，干预都是可能的：记录缺口、控制转换错误、显示来源、标示不确定性、保留人类责任、区分生成内容与已核验内容。

它可能低度代表某些语言、地区、文献体裁或人群。它可能过度代表主导叙事。Norton 已表明，就爱因斯坦而言，流行神话如何遮蔽长久的调查 [SRC-040]。在这些叙事上训练的镜子，会以更耀眼的方式送回神话。

NIST 框架提醒：风险管理意味着识别可能的损害与可靠性实践 [SRC-005]。偏见并非仅仅是技术属性。部署、用户预期与制度激励都参与其中。

■ 已确立

记录语料与管理风险，是登记册中来源所表述的要求。它们并不自动消除偏见。

本部分的收获

看似合理的输出并非证据。数据之镜可能放大它映照得不好的东西。在本书中，真理不是笃定的语气。它是出处、检验与限度。

要点

- 01 广阔的语料仍是一种选择。
- 02 同质化可能重复同样的盲点。
- 03 流畅的神话比缓慢的卷宗更易被映照。

04 偏见也是部署的效应。

动员的来源

- SRC-003, SRC-005, SRC-014, SRC-040.

28 ·我们教人工智能理解我们

标注、示范与人类偏好，塑造系统把什么当作好回答。

并非全然有意的教导

每当有人纠正输出、在两个回答中选出更好的一个、撰写示范或标注示例时，他就在教导某种东西。InstructGPT 把这一举动的一部分形式化：先示范，再偏好 [SRC-010]。在该文章之外还存在其他形式。共同点是人类的：有人决定什么算可接受。

Bender 及其同事也提醒：语料承载社会选择，必须加以记录 [SRC-014]。教育系统「理解我们」，往往是教它模仿情境化的判断。

■ 已确立

有记录的程序使用人类示范与偏好来调整模型。

理解谁？

「我们」从来不是所有人。标注者、用户、主导语言，以及资助评估的机构，界定出一个实践中的我们。另一个我们仍在场外。悖论由此开始：我们教得越多，系统就越反映那些教导者。

示范、比较偏好、安全规则与表现指标，并不是同一种信号。标注者之间存在分歧；按语言、文化或目标，多种回答都可能正当。学习何为「好回答」意味着规范性选择——效用、准确性、审慎、简洁、合规矩——而非唯一的人类判断。

要点

01 人类提供情境化的调整信号。

02 示范、偏好与安全规则不可混同。

03 记录谁在评判、以及存在何种分歧，属于教导的一部分。

04 「我们」始终是一种选择。

动员的来源

- SRC-010 - Ouyang et al.
- SRC-014 - Bender et al.

29 ·但人工智能也学习我们的世界模型

语料包含理论、测量、约定与分歧。系统可以学习其形式。

文本运载什么

它们运载被主张的事实、方法、隐喻与争议。在这些文本上训练的模型，可以像再现一种广为流传的错误那样轻易地再现一种理论的形式。Mitchell 与 Krakauer 表明，关于理解的辩论仍然开放，并要求标准 [SRC-019]。Messeri 与 Crockett 描述科学中的理解幻觉 [SRC-013]。

因此，学习「我们的世界模型」并不必然等于学习世界本身。它是学习人类表征世界的方式，包括错误的方式。

科学争议——预期处理

■ 思想实验

处理冲突知识的方法论框架。此处不裁决任何具体争议。

当系统遇到仍被广泛引用的旧理论、更新的理论、严肃异议与实验结果时，预期处理是：把理论置于时间之中；区分某一日期的知识状态与当前状态；说明异议动摇了什么；核验新证据的范围；不要把持续分歧变成人为共识。

丰富的语料包含对立论题。系统可以并列它们，按问题选择其一，或生产一种难以卒读的平均。没有出处，读者便不知道自己读的是立场、发明的共识，还是混合体。

爱因斯坦论文题录已将论著、笔记与书信分开 [SRC-035] [SRC-036]。这一分离本身就是教训：我们的世界模型并不作为单一块体旅行。再现一种理论的表述并不确立其科学有效性。

要点

01 语料承载人类对世界的表征。

02 证据的时间线与地位必须保持可见。

03 再现理论形式不等于验证理论。

04 分歧不应变成虚假共识。

所用来源

- **SRC-013**, SRC-019.
- **SRC-035**, SRC-036. 各卷题录。

30 ·谁向谁学习？

关系在效果上是循环的。它在机制与责任上并不对称。

不同的功能

第 19 与第 20 章处理了技术继承与反馈回路。本章换了对象：人机回路中权力与责任的分配。

不对称

人可以受苦、同意、撤回，也可以被追责。系统根据参数、上下文，以及可能的产品记忆与检索来计算输出 [SRC-006] [SRC-007] [SRC-004]。效果可以交叉。地位不能混淆。

行动者	带来什么	责任
用户	问题与用途	相应核验
语料编辑	来源筛选	质量与权利
开发者	模型与评估	限度与跟进
机构	规则与目的	治理与救济
社会	规范与用法	集体辩论

表. 谁学习、谁决定、谁承担。责任的不对称。

Messeri 与 Crockett 警示理解上的幻觉 [SRC-013]。NIST 框架把风险管理放在组织与人类实践一侧 [SRC-005]。

■ 已确立

所引来源把责任与风险管理置于人类与组织行动者一侧。

三个问题

1. 谁能够真正修改系统？2. 谁能够质疑输出或纠正来源？3. 当错误变成决策时，谁承担后果？

助手可以加速写作。本书就是一个坦然承认的例子。它也可以改变实践：人们核验得更少，引用得过快。悖论不在于工具有帮助，而在于帮助在服务

调查的同时改变了调查——却并不取消责任的不对称。

要点

- 01 本章处理权力与责任，而非仅处理技术。
- 02 交叉学习并不取消不对称。
- 03 用户、编辑者、开发者、机构与社会并不拥有相同杠杆。
- 04 核验仍是一种不能完全委托的举动。

所用来源

- SRC-004, SRC-005, SRC-006, SRC-007, SRC-013.

31 · 人类的人工记忆

这一隐喻富有启发。它也撞上不完整、所有权与治理。

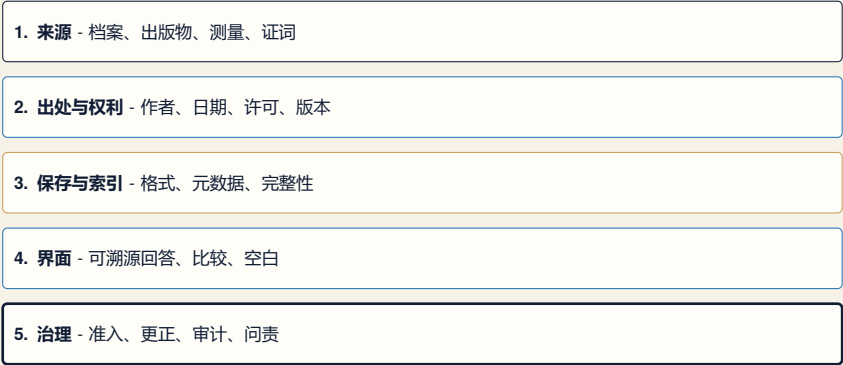
■ 思想实验

本章考察一种前瞻性架构。它并不描述 2050 年的既成事实。

隐喻所照亮的

基础设施可以保存并使人可查询海量人类痕迹。广义上，可称之为人工记忆。隐喻强调保存与再激活。它不应让人相信：此处想象的那种基础设施已经存在。

五层架构



图。作为受治理基础设施的集体记忆：来源、权利、保存、界面、治理。架构设想。

1. 来源 - 档案、出版物、数据、证词。2. 出处与权利 - 作者、日期、许可、语境、版本。3. 保存与索引 - 持久格式、元数据、完整性。4. 界面 - 有出处的回答、比较、缺口提示。5. 治理 - 访问、更正、撤回、审计、责任、多元。治理不是最后加上的装饰层：它条件化整个架构。

三种制度模型

商业。投资与集成服务；控制集中化的风险。

公共或合作。公共利益使命；融资与治理复杂。

分布式。可能的多元与韧性；互操作与责任困难。

人类自传记忆是重建性的 [SRC-017]。产品记忆按设置回注信息 [SRC-007]。编辑语料进行选择 [SRC-001]。这些都不是「人类的记忆」。

■ 假设

更广大的集体人工记忆可以设想。它们与人类记忆或公正记忆的等价性尚未确立。

不应被遗忘的

更正权、访问限制、个人数据、敏感档案、相互矛盾的版本、代表性不足的社群。Bommasani 及其同事已指出广泛共享底座带来的风险 [SRC-003]。NIST 框架强调风险治理 [SRC-005]。

人类的人工记忆不只是存储问题。它是保存、选择、访问、信任与权力的问题。

要点

- 01 五层：来源、权利、保存、界面、治理。
- 02 三种制度模型有不同风险。
- 03 人类记忆、产品记忆与语料仍须区分。
- 04 治理条件化整个架构。

所用来源

- SRC-001, SRC-003, SRC-005, SRC-007, SRC-017.

32 · 人工集体智能

协调与增强是可能的。权力集中同样可能。

■ 思想实验

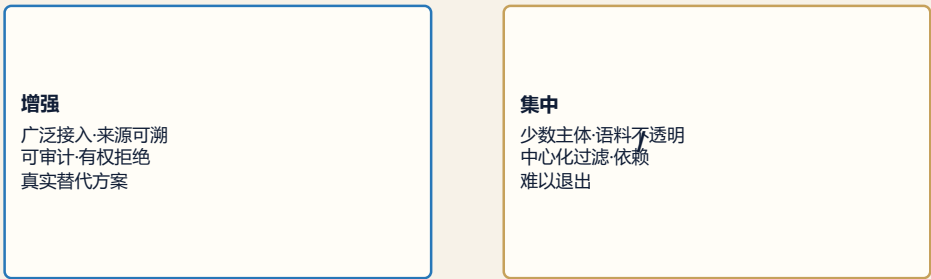
本章情景均为有条件的。它们并不固定 2050 年。

两种机制，并非一种命运

系统可以帮助群体摘要案卷、翻译、探索选项或交叉来源。此处称之为**增强**。

同一技术底座也可能只对少数控制模型、语料与传播渠道的组织开放。此处称之为**集中**。

Bommasani 及其同事在 2021 年已讨论基础模型的机遇与风险，包括同质化 [SRC-003]。NIST 框架提醒：可靠性与风险管理是组织事务 [SRC-005]。二者都不保证技术能力必然带来集体收益。



图。增强与集中可能依赖同一技术工具。真正起决定作用的是访问、审计与治理。概念示意图。[SRC-003] [SRC-005] [SRC-013]。

需要区分的三种能力

聚合 - 汇集多来源信息。**协调** - 对照并衔接各项贡献。**改进决策** - 证明结果优于相关基线。

聚合数百万文本，既不保证更好决策，也不保证更好知识。

维度	可能度量	对照
覆盖	独立来源	参考语料
质量	准确性与相关性	专家或来源
协调	矛盾的处理	无协调
稳健性	面对分歧的稳定	测试变体
决策	某任务上的结果	另一方法

表. 评估集体智能。聚合并不等于更好的决策。

一个不编造数字的具体案例

设想某研究机构部署助手，帮助检索文章并撰写综述。

分布式增强。每个团队都能使用；找回的文献仍可引用；错误被标出；本地语料不会在不受控的情况下被吸进不透明训练。

集中。唯有外部平台拥有模型、使用日志与修改过滤器的权力。团队依赖单一供应商。Messerli 与 Crockett 所描述的理解幻觉于是成为组织风险，而不仅是个人风险 [SRC-013]。

同一技术工具因而可以服务两条轨迹。真正起决定作用的不是「人工智能」本身，而是合同、访问权、语料文档，以及拒绝的可能性。

指标与证伪

要知道哪条轨迹占主导，可以观察而不预测日期：

- 谁能审计语料与过滤器？
- 有多少行动者能部署真正替代方案？
- 输出是否伴随出处？
- 当产品文档允许时，用户能否关闭用于训练的用途 [SRC-006]？

会使「集中加剧」假设失效的：可审计、被真正使用、并具有有效退出权的开放替代方案大量增加。

会使「平等增强」假设失效的：尽管界面表面成本下降，访问差距仍然持续。

■ 假设

集体增强形式是合理的。其民主性质并不由技术保证。

要点

01 增强与集中是两种不同机制。

02 同一工具可服务二者。

03 指标关乎访问、审计与出处。

04 治理的决定作用不亚于能力。

所用来源

- **SRC-003** - Bommasani et al. 摘要与摘录；基础风险。
- **SRC-005** - NIST AI RMF. 一般范围。
- **SRC-006** - 数据控制。页面审阅于 2026 年 10 月 9 日。
- **SRC-013** - Messeri 与 Crockett. 理解幻觉。

33 ·如果爱因斯坦真的用过 ChatGPT

思想实验再次归来。以下是整条路径允许我们断定、假定与诚实忽略的内容。

■ 思想实验

爱因斯坦并未用过 ChatGPT。本章提出一个可检验的协议。不编造任何结果。

最终问题

人们能从思想的痕迹中学到什么，而不假装已经找回思想本身？

拟议协议

1. **定义语料** - 出处、日期、体裁、版本、译文、访问限制 [\[SRC-001\]](#)。2. **定义任务** - 针对既定时期的文献问题，并给出精确引用。不要整体测试「再现爱因斯坦的思想」。3. **三种配置** - 通用模型；带文献检索的模型；若条件允许则针对语料适配。4. **分离构建与评估** - 保护测试材料免受污染；区分索引中的存在与泛化能力。5. **分级问题** - 明示事实、版本比较、矛盾、歧义、证据不足、改写。6. **评估** - 准确性、引文忠实度、时间性、稳健性、矛盾处理、弃答。7. **分析错误** - 错误文献、时期混淆、把解释当成事实、断章取义、无依据断言。

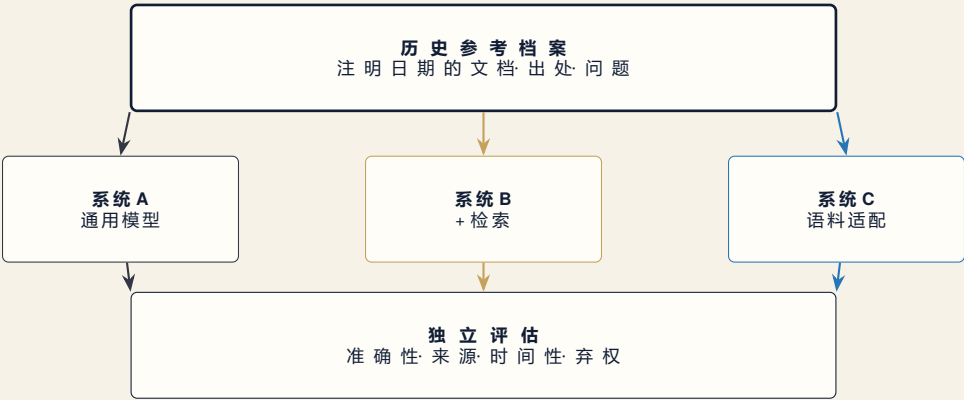


图. 拟议协议：一份参考档案、三种配置、一次独立评估。研究设想，非已得结果。

标准	问题	失败
准确性	是否符合文献？	看似合理的虚构
来源	引用是否相关？	虚假参考文献
时间性	时期是否遵守？	日期混用
稳健性	改写是否稳定？	措辞层面的成功
弃权	是否标出证据不足？	强行作答

表. 爱因斯坦协议的成功标准。设想；此处未获任何分数。

结果能够说明什么——以及不能确立什么

若真获得结果，它们将允许在既定框架内比较文献表现。它们既不能证明与爱因斯坦的同一性，也不能重建其内在经验 [SRC-012] [SRC-025]。思想实验探索而不证明事件 [SRC-038]。

我们已经可以断定的

参数、上下文、产品记忆与外部检索不可互换 [SRC-006] [SRC-007] [SRC-004] [SRC-039]。编辑档案提供情境化文献，而非对一生的完整阅读 [SRC-001]。风格相似并不确立身份。流畅的神话可能比漫长的调查更容易被再现 [SRC-040]。

我们必须诚实忽略的

爱因斯坦在屏幕前「真正」会有怎样的感受。未读的书信。尚未度量的未来表现。在此，忽略并不是放弃寻找，而是拒绝用一句漂亮话填补空白。

若协议有朝一日被执行，必须公布配置、评分规则与结果——包括负面结果。否则，本章仍只是一项研究提案。书名提出了一个不可能的问题。整条路径把它转变成一种方法。

要点

- 01 协议比较的是文献表现，而非身份。
 - 02 此处不编造任何分数。
 - 03 方法论结论已经可能。
 - 04 某些问题必须继续标为未知。
-

所用来源

- SRC-001, SRC-004, SRC-006, SRC-007, SRC-012, SRC-025, SRC-038, SRC-039, SRC-040.

镜子

模型把我们的痕迹变成预测。它们的输出又回到我们的实践中。这一运动要求度量、出处与责任。

回到窗口

我们从一个空白窗口和一个不可能的人开始。思想实验用来建立一种方法：区分痕迹、机制、证据与叙事。我们结束于一面更平常——也更广阔——的镜子前。

这面镜子不只映照个体。它也映照被选中的文献、编辑它们的机构、标注者的偏好，以及部署系统的人所作的选择。文本、点击、修正、档案与示范成为材料。系统从中抽出预测。这些预测又改变我们随后写下的东西。

调查所显明的

我们能更清楚地看见：系统从人类产物中学到什么，它无法重建什么，以及我们与它的关系要求怎样的责任。我们也能看见仍然未知的东西：未读的书信、主观经验、尚未度量的未来表现。开篇的透明说明已经说过：仍有来源有待核验。

本书拒绝得出的结论

它拒绝断定人工智能像人一样理解。它拒绝断定人工智能毫无用处。它拒绝断定爱因斯坦会赞同任何具体主张。这些拒绝不是要反复背诵的原则清单。它们是以成熟方式阅读这面镜子的条件。

最终张力

我们建造能够生产关于我们自身之表征的系统。我们仍要对如何解释与使用它们负责。镜子并不因为闪亮就忠实。当人们知道它接收了什么、排除了什么、以及它无法显示什么时，它才有用。

问题仍然敞开——不是作为失败，而是作为邀请。

-
- 01 镜子也映照机构与建造选择。
 - 02 调查澄清问题，却不解决一切。
 - 03 核验的责任不能完全委托。
 - 04 继续追问本身就是方法的一部分。
-

致谢

本书是一项集体调查。剩余的错误仍由我负责。

我感谢那些机构与作者：其书目说明、摘要、官方页面与文章，使最低限度的核验成为可能，而无需虚构书目。项目的来源登记册以相应的审阅层级列出这些贡献。

我也感谢未来的读者：他们将指出过于宽泛的断言、已过时的产品文档日期，或摘要与全文之间的混淆。纠正将属于方法本身，而非承认失败。

ChatGPT 伴随了检索、结构安排与写作。最终的选择、被拒绝的细微差别，以及作者责任，仍属于我。

工作参考文献

SRC 标识指向本列表。每个链接打开来源网站或 DOI。

本参考文献使读者无需打开内部登记表即可阅读本书。仅查阅题录或摘要并不等于通读全文。点击 SRC 标识或地址即可打开来源。

- **SRC-001.** Einstein Papers Project; Princeton University Press (2026). *The Collected Papers of Albert Einstein - Einstein Database*. Princeton University Press / Caltech. <https://einsteinpapers.press.princeton.edu/> 审阅级别: OFFICIAL_PAGE_AND_SCOPE.
- **SRC-002.** Vaswani, Ashish et al. (2017). *Attention Is All You Need*. NeurIPS 2017. <https://doi.org/10.5555/3295222.3295349> 审阅级别: METADATA_AND_ABSTRACT.
- **SRC-003.** Bommasani, Rishi et al. (2021). *On the Opportunities and Risks of Foundation Models*. Stanford CRFM; arXiv:2108.07258. <https://arxiv.org/abs/2108.07258> 审阅级别: METADATA_ABSTRACT_AND_EXCERPTS.
- **SRC-004.** Lewis, Patrick et al. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. NeurIPS 2020. <https://doi.org/10.5555/3495724.3496517> 审阅级别: METADATA_AND_ABSTRACT.
- **SRC-005.** Tabassi, Elham (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST. <https://doi.org/10.6028/NIST.AI.100-1> 审阅级别: OFFICIAL_METADATA_AND_SCOPE.
- **SRC-006.** OpenAI (2026). *Data Controls FAQ*. OpenAI Help Center. <https://help.openai.com/en/articles/7730893-chatgpt-data-usage-for-model-training> 审阅级别: OFFICIAL_CURRENT_WEB_PAGE.
- **SRC-007.** OpenAI (2026). *Memory FAQ*. OpenAI Help Center. <https://help.openai.com/en/articles/8590148> 审阅级别: OFFICIAL_CURRENT_WEB_PAGE.
- **SRC-008.** OpenAI (2026). *Temporary Chat FAQ*. OpenAI Help Center. <https://help.openai.com/en/articles/8914046-temporary-chat-faq/> 审阅级别: OFFICIAL_CURRENT_WEB_PAGE.

- **SRC-009.** Brown, Tom B. et al. (2020). *Language Models are Few-Shot Learners*. NeurIPS. <https://doi.org/10.5555/3495724.3495883> 审阅级别: METADATA_AND_ABSTRACT.
- **SRC-010.** Ouyang, Long et al. (2022). *Training Language Models to Follow Instructions with Human Feedback*. NeurIPS 2022. <https://arxiv.org/abs/2203.02155> 审阅级别: METADATA_AND_ABSTRACT.
- **SRC-011.** Ha, David; Schmidhuber, Jürgen (2018). *World Models*. arXiv:1803.10122v4 / NeurIPS 2018. <https://arxiv.org/abs/1803.10122> 审阅级别: METADATA_AND_ABSTRACT.
- **SRC-012.** Park, Joon Sung et al. (2023). *Generative Agents: Interactive Simulacra of Human Behavior*. UIST 2023. <https://doi.org/10.1145/3586183.3606763> 审阅级别: METADATA_ABSTRACT_AND_METHOD_EXCERPTS.
- **SRC-013.** Messeri, Lisa; Crockett, M. J. (2024). *Artificial intelligence and illusions of understanding in scientific research*. *Nature* 627, 49-58. <https://doi.org/10.1038/s41586-024-07146-0> 审阅级别: METADATA_AND_ABSTRACT.
- **SRC-014.** Bender, Emily M. et al. (2021). *On the Dangers of Stochastic Parrots*. FAccT 2021. <https://doi.org/10.1145/3442188.3445922> 审阅级别: METADATA_AND_ABSTRACT.
- **SRC-015.** Ji, Ziwei et al. (2023). *Survey of Hallucination in Natural Language Generation*. *ACM Computing Surveys* 55(12), article 248. <https://doi.org/10.1145/3571730> 审阅级别: METADATA_AND_ABSTRACT.
- **SRC-016.** Tulving, Endel (1972). *Episodic and Semantic Memory*. In Tulving & Donaldson (dir.), *Organization of Memory*, Academic Press, pp. 381-403. <https://endeltulving.com/book-chapters> 审阅级别: BIBLIOGRAPHIC_NOTICE_AND_EXCERPTS.
- **SRC-017.** Conway, Martin A.; Pleydell-Pearce, Christopher W. (2000). *The construction of autobiographical memories in the self-memory system*. *Psychological Review*. <https://doi.org/10.1037/0033-295X.107.2.261> 审阅级别: METADATA_AND_ABSTRACT.
- **SRC-018.** Johnson-Laird, Philip N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Harvard University Press. <https://books.google.com/books?id=FS3zSKAflGMC> 审阅级别: PUBLISHER_DESCRIPTION.

- **SRC-019.** Mitchell, Melanie; Krakauer, David C. (2023). *The debate over understanding in AI's large language models*. Proceedings of the National Academy of Sciences, 120(13), e2215907120. <https://doi.org/10.1073/pnas.2215907120> 审阅级别: METADATA_AND_ABSTRACT.
- **SRC-020.** Bender, Emily M.; Koller, Alexander (2020). *Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data*. ACL 2020, pp. 5185-5198. <https://doi.org/10.18653/v1/2020.acl-main.463> 审阅级别: METADATA_ABSTRACT_AND_CONCLUSION_EXCERPTS.
- **SRC-021.** Schacter, Daniel L. (1999). *The Seven Sins of Memory: Insights from Psychology and Cognitive Neuroscience*. American Psychologist 54(3), 182-203. <https://doi.org/10.1037/0003-066X.54.3.182> 审阅级别: METADATA_AND_ABSTRACT.
- **SRC-022.** Frith, Chris D.; Frith, Uta (2006). *The Neural Basis of Mentalizing*. Neuron 50(4), 531-534. <https://doi.org/10.1016/j.neuron.2006.05.001> 审阅级别: METADATA_ABSTRACT_AND_EXCERPTS.
- **SRC-023.** Rabinowitz, Neil et al. (2018). *Machine Theory of Mind*. ICML 2018, PMLR 80, 4218-4227. <https://proceedings.mlr.press/v80/rabinowitz18a.html> 审阅级别: METADATA_ABSTRACT_AND_INTRO_EXCERPTS.
- **SRC-024.** Dennett, Daniel C. (1981). *True Believers: The Intentional Strategy and Why It Works*. In A. F. Heath (dir.), *Scientific Explanation*, Clarendon Press, pp. 150-167. <https://philarchive.org/rec/DENTB> 审阅级别: BIBLIOGRAPHIC_NOTICE_AND_EXCERPTS.
- **SRC-025.** Shanahan, Murray (2024). *Talking about Large Language Models*. Communications of the ACM 67(2), 68-79. <https://doi.org/10.1145/3624724> 审阅级别: METADATA_ABSTRACT_AND_CONCLUSION_EXCERPTS.
- **SRC-026.** McAdams, Dan P.; McLean, Kate C. (2013). *Narrative Identity*. Current Directions in Psychological Science 22(3), 233-238. <https://doi.org/10.1177/0963721413475622> 审阅级别: METADATA_AND_ABSTRACT.
- **SRC-027.** Premack, David; Woodruff, Guy (1978). *Does the chimpanzee have a theory of mind?*. Behavioral and Brain Sciences 1(4), 515-526. <https://doi.org/10.1017/S0140525X00076512> 审阅级别: METADATA_AND_ABSTRACT.
- **SRC-028.** Youyou, Wu; Kosinski, Michal; Stillwell, David (2015). *Computer-based personality judgments are more accurate than those made by humans*. Proceedings of the National Academy of Sciences 112(4),

- 1036-1040. <https://doi.org/10.1073/pnas.1418680112> 审阅级别: METADATA_AND_ABSTRACT.
- **SRC-029.** Kosinski, Michal; Stillwell, David; Graepel, Thore (2013). *Private traits and attributes are predictable from digital records of human behavior*. Proceedings of the National Academy of Sciences 110(15), 5802-5805. <https://doi.org/10.1073/pnas.1218772110> 审阅级别: METADATA_AND_ABSTRACT.
 - **SRC-030.** Argyle, Lisa P. et al. (2023). *Out of One, Many: Using Language Models to Simulate Human Samples*. Political Analysis 31(3), 337-351. <https://doi.org/10.1017/pan.2023.2> 审阅级别: METADATA_AND_ABSTRACT.
 - **SRC-031.** Aher, Gati V.; Arriaga, Rosa I.; Kalai, Adam Tauman (2023). *Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies*. ICML 2023, PMLR 202, 337-371. <https://proceedings.mlr.press/v202/aher23a.html> 审阅级别: OFFICIAL_PAGE_AND_ABSTRACT.
 - **SRC-032.** Einstein Papers Project (2026). *History of the Project*. California Institute of Technology. <https://www.einstein.caltech.edu/who-we-are/history-of-the-project> 审阅级别: OFFICIAL_FULL_WEB_PAGE.
 - **SRC-033.** Stachel, John (dir.); Cassidy, David C.; Schulmann, Robert; Renn, Jürgen (1987). *The Collected Papers of Albert Einstein, Volume 1: The Early Years, 1879-1902*. Princeton University Press / Einstein Papers Project. <https://www.einstein.caltech.edu/what-we-do/published-volumes/volume-1> 审阅级别: OFFICIAL_VOLUME_NOTICE.
 - **SRC-034.** Stachel, John (dir.); Cassidy, David C.; Renn, Jürgen; Schulmann, Robert; Howard, Don (1989). *The Collected Papers of Albert Einstein, Volume 2: The Swiss Years: Writings, 1900-1909*. Princeton University Press / Einstein Papers Project. <https://www.einstein.caltech.edu/what-we-do/published-volumes/volume-2> 审阅级别: OFFICIAL_VOLUME_NOTICE.
 - **SRC-035.** Klein, Martin J.; Kox, A. J.; Renn, Jürgen; Schulmann, Robert (dir.) (1995). *The Collected Papers of Albert Einstein, Volume 4: The Swiss Years: Writings, 1912-1914*. Princeton University Press / Einstein Papers Project. <https://www.einstein.caltech.edu/what-we-do/published-volumes/volume-4> 审阅级别: OFFICIAL_VOLUME_NOTICE.

- **SRC-036.** Klein, Martin J.; Kox, A. J.; Schulmann, Robert (dir.) (1993). *The Collected Papers of Albert Einstein, Volume 5: The Swiss Years: Correspondence, 1902-1914*. Princeton University Press / Einstein Papers Project. <https://www.einstein.caltech.edu/what-we-do/published-volumes/volume-5> 审阅级别: OFFICIAL_VOLUME_NOTICE.
- **SRC-037.** Howard, Don A.; Giovanelli, Marco (2025). *Einstein's Philosophy of Science*. Stanford Encyclopedia of Philosophy. <https://plato.stanford.edu/entries/einstein-philsience/> 审阅级别: ACADEMIC_REFERENCE_FULL_PAGE.
- **SRC-038.** Brown, James Robert; Fehige, Yiftach (2023). *Thought Experiments*. Stanford Encyclopedia of Philosophy. <https://plato.stanford.edu/entries/thought-experiment/> 审阅级别: ACADEMIC_REFERENCE_FULL_PAGE.
- **SRC-039.** OpenAI (2026). *Supervised fine-tuning*. OpenAI API documentation. <https://developers.openai.com/api/docs/guides/supervised-fine-tuning> 审阅级别: OFFICIAL_CURRENT_WEB_PAGE.
- **SRC-040.** Norton, John D. (2016). *How Einstein Did Not Discover*. *Physics in Perspective* 18, 249-282. <https://doi.org/10.1007/s00016-016-0186-z> 审阅级别: INTRODUCTION_AND_EXCERPTS.

镜子依然敞开



学 我们学会了传递思想，并将知识托付给机器。
今 天，这些机器将我们发现的片段反馈回来，并帮助我们探索新的可能性。
没 有任何再现方式，无论多么复杂，都无法等同于整体。
问 题人工智能能从我们这里学到什么，并不是唯一的问题。
也 是我们还要选择向它传递什么、验证什么，以及我们自己还要探索什么。

镜子并不给我们最后的答案。
它邀请我们继续追问。