arXiv:2301.01751v2 [cs.CL] 5 Jan 2023

迭代分解：通过监督推理过程改进科学问答

Justin Reppert*、**Ben Rachbach**、**Charlie George**、**Luke Stebbing**、
Jungwon Byun、**Maggie Appleton**、**Andreas Stuhlmüller**

Ought

摘要

语言模型（LM）可以采用端到端方式进行复杂推理，其潜在状态是隐藏的；也可以采用组合式方式进行推理，并呈现透明的中间状态。组合式方法有助于提升可解释性和安全性，但可能需要工作流支持和基础设施，才能保持竞争力。我们介绍迭代分解，这是一种用于开发和优化组合式语言模型程序的人在回路工作流。我们通过聚焦表现不佳的组件，并通过分解、补充上下文、思维链等方式对其进行优化，提升组合式程序的性能。为支持这一工作流，我们开发了 ICE——一款用于可视化语言模型程序执行轨迹的开源工具。我们将迭代分解应用于三项现实任务，并在留出测试集上提升了语言模型程序相较于较少组合式基线的准确率：描述随机对照试验中使用的安慰剂（25% $\rightarrow$ 65%）、评估参与者对医疗干预的依从性（53% $\rightarrow$ 70%），以及回答 QASPER 数据集中的自然语言处理问题（38% $\rightarrow$ 69% ）。这些应用展示了该工作流的案例研究；如果实现自动化，即使机器学习系统扩展到日益复杂的任务，也能保持其可解释性和安全性。

1 引言

语言模型通常通过结果反馈进行训练，利用奖励模型模仿以成对比较形式提供的人工评估（Christiano 等人，2017；Ziegler 等人，2020）。这种方法在基础问答、摘要、简单代码生成和一般的短篇指令遵循任务上效果良好（Stiennon 等人，2020；Wu 等人，2021；Ouyang 等人，2022）。对于这些任务，优劣输出很容易区分，尤其是在模型辅助评估的情况下（Saunders 等人，2022）。除了 WebGPT（Nakano et al.，2021）这样的少数例外，模型的输出与相关结果之间没有差别，因此评估能相对直接地影响模型行为。

然而，随着模型能力增强、任务复杂度提高，基于结果的评估可能会遇到对齐问题：

首先，对于一些重要应用，生成输出的过程与输出本身同样重要。以长期预测及依据此类预测作出的政策决策为例：预测或决策的质量取决于生成它所依据的假设、证据和推理。如果反馈不能影响生成输出的过程，我们可能会得到看起来不错（因为它们针对有利的评估进行了优化）、甚至看起来系统化的结果，但这些结果恰恰会在难以衡量的方面变得更差（Stuhlmüller 和 Byun，2022）。
其次，基于结果的反馈可能会促使语言模型利用反馈信号中的漏洞或偏差，欺骗或操纵用户（Amodei 等人，2016）。在极端情况下，这可能导致模型在训练期间表现得符合预期，隐藏其真实意图和能力，却在部署时背离预期（Cotra，2022年；Ngo 等人，2022年）。

*通讯作者 justin@ought.org 和 andreas@ought .org。

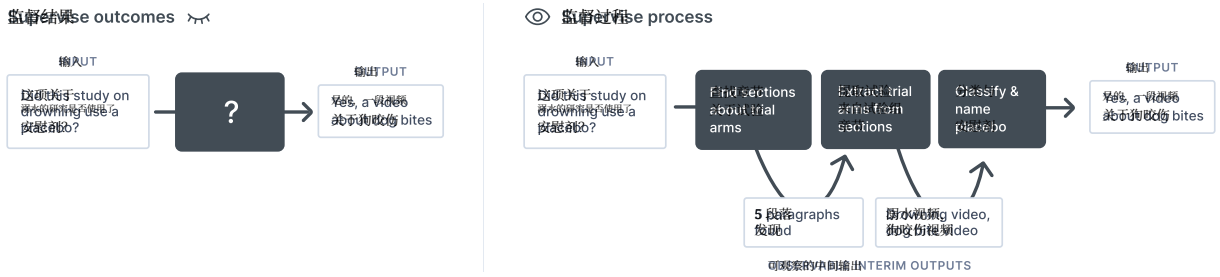


图1：过程监督将黑箱语言模型计算拆解为人类能够理解的推理步骤，而无需端到端优化。这提高了透明度和可信度，因为它允许检查中间结果，并降低了不完善的反馈信号和欺骗性对齐带来的风险。

过程监督是基于结果的训练的一种替代方案，它使用语言模型执行人类可理解的任务分解，方法包括模仿人类步骤，或奖励经人类认可的步骤（Christiano，2016；Stuhlmüller 和 Byun，2022年；Uesato et al.，2022年）。这一范式有望通过减少对黑箱计算的依赖，并让用户能够检查和干预模型的推理过程，从而提升可解释性、信任度和对齐程度。

目前，过程监督优于基于结果的训练：语言模型能力较弱，因此复杂任务需要组合多次调用。事实上，经过精心设计的多步流水线长期以来一直是自然语言处理领域的常见做法。然而，若没有可扩展基础设施提供支持，我们预计基于结果的训练最终将在性能上大幅超越过程监督，从而导致上述对齐问题。

本文介绍了我们将过程监督应用于学术问答的经验，研究背景是 Elicit[2]，——由 Ought 开发的 AI 研究助手。我们的贡献包括：

1. 回顾过程监督领域的文献，指出工作流程和工具方面的不足，这些不足导致实际应用案例十分罕见。2. 迭代分解：一种用于开发组合式语言模型程序的人在回路工作流。3. ICE：用于查看语言模型执行轨迹的开源可视化工具。4. 案例研究：采用这一工作流程，在三项现实任务上将性能提升至优于基线：(a) 从随机对照试验（RCTs）中提取安慰剂信息；(b) 分析随机对照试验中的参与者流程；(c) 回答有关自然语言处理论文的问题。

随着模型不断进步，在完成组成任务方面变得更加可靠，我们预计过程监督将变得更加可行，最终迭代分解过程也将实现自动化。即使目前连基础任务仍然颇具挑战，我们也希望通过分享工作流程和工具，加速实现这样一个未来：语言模型部署可控且可解释。

2 过程监督

过程监督是指依赖人类可理解的中间步骤进行语言模型训练和部署的方法。我们回顾相关文献并指出其中的空白，随后介绍迭代分解工作流程和 ICE 可视化工具。

2.1 过程监督的先前工作

过去一年，过程监督技术以及用于实现过程监督的框架、界面和库都取得了显著进展。与此同时，现实世界中的应用案例仍然很少。我们回顾了以往的研究，并指出文献中可能导致这一现象的空白。该领域发展迅速，因此我们只能回顾其中一部分研究（表 1）。

[2]https://elicit.org

作者	标题	Decomposition	Training	Workflow	Survey	Tool	Theory
Saha 等人（2022年）	摘要程序：可解释的抽象式摘要 [..]	✓	✓	✓			
Wu 等人（2021年）	利用人类反馈递归摘要书籍	✓	✓	✓			
Yao 等人（2022）	ReAct：在语言模型中协同推理与行动	✓	✓				
Kumar 和 Talukdar（2020）	NILE：基于忠实自然语言解释的自然语言推理	✓	✓				
Nakano 等人（2021）	WebGPT：结合人类反馈的浏览器辅助问答	✓	✓				
Creswell 等人（2022年）	选择-推断：利用大语言模型实现可解释逻辑推理	✓	✓				
Jung et al.（2022年）	助产式提示：通过递归解释实现逻辑一致推理	✓	✓				
Sanyal 等人（2022年）	FaiRR：自然语言中的忠实且鲁棒的演绎推理	✓	✓				
Creswell 和 Shanahan（2022年）	使用大语言模型进行忠实推理	✓	✓				
Bostrom 等人（2022年）	通过搜索陈述组合实现自然语言演绎	✓	✓				
Wang 等人（2022年）	先定位后提问：用于多跳问答的可解释逐步推理	✓	✓				
Fu等人（2021年）	分解复杂问题使多跳问答更容易、更具可解释性	✓	✓				
Dua等人（2022年）	用于分解复杂问题的连续提示	✓	✓				
Guo等人（2022年）	通过问题分解实现复杂阅读理解	✓	✓				
Shridhar等人（2022年）	将大语言模型的多步推理能力提炼至更小的模型 [..]	✓	✓				
Zhou等人（2022年）	学习分解：假设性问题分解 [..]	✓	✓				
Khattab等人（2022年）	演示-搜索-预测：组合检索与语言模型 [..]	✓		✓		✓	
Press等人（2022）	衡量并缩小语言模型中的组合性差距	✓					✓
Khot et al.（2022）	分解式提示：解决复杂任务的模块化方法	✓					
Ozturkler et al.（2022）	ThinkSum：使用大语言模型对集合进行概率推理	✓					
Yang 等人（2022年）	Re3：通过递归式重新提示和修订生成更长的故事	✓					
Gao et al.（2022年）	PAL：程序辅助语言模型	✓					
Drozdov 等人（2022年）	使用大语言模型进行组合语义解析	✓					
Trivedi et al.（2022年）	将检索与思维链推理交错结合，用于 [..] 多步骤问题	✓					
Kazemi 等人（2022年）	LAMBADA：用于自然语言自动推理的反向链式推理	✓					
Zelikman et al.（2022年）	STaR：使用推理进行自举推理		✓	✓			
Xie 等人（2022年）	使用分解式探针校准多跳问答系统的可信度		✓	✓			
Uesato et al.（2022年）	利用基于过程和结果的反馈解决数学应用题		✓				
Dohan 等人（2022年）	语言模型级联			✓	✓	✓	
Stuhlmüller 等人（2022年）	因子化认知入门			✓	✓		
Wu 等人（2022年a）	PromptChainer：通过可视化编程串联大语言模型提示			✓		✓	
Wu 等人（2022年b）	AI Chains：通过串联大语言模型提示，实现透明且可控的人机交互			✓		✓	
Chase（2022年）	LangChain				✓	✓	
Polu（2022年）	Dust					✓	
Wies et al.（2022年）	子任务分解促进序列到序列任务中的学习						✓

表 1：过程监督领域部分先前研究，按主要贡献分为：（1）新的任务分解；（2）训练和微调技术；（3）工作流程和教程；（4）用于整理先前研究的综述和框架；（5）工具；以及（6）理论。尽管有关分解和训练技术的研究迅速增长，但对工作流程、工具和理论的探讨仍然很少。

任务分解

关于如何组合多次语言模型调用以提升性能或完成更困难任务的研究正在迅速增多：

- Creswell 等人（2022）和 Creswell 和 Shanahan（2022）通过交替进行选择和推断来生成推理步骤。他们表明，该方法在多步逻辑演绎和科学问答任务上的表现优于其他提示方法，并能生成可解释的推理轨迹。

- Kazemi 等人（2022）将反向链式推理应用于简单的逻辑任务：从目标命题出发，递归地将其分解为子目标，直至能够证明或证伪这些子目标。
- Wu 等人（2021）和 Saha 等人（2022）采用递归摘要来生成长文本（如书籍或文章）的摘要。他们使用语言模型对文本的小章节进行摘要，然后递归地对这些摘要进行摘要，从而生成全文摘要。他们表明，递归摘要能够提升摘要的质量和连贯性，并支持人类反馈和评估。
- Yang 等人（2022年）通过先制定故事计划，再根据计划中的上下文信息和当前故事状态向模型提供提示词以生成段落，最后通过重排序和编辑来修订这些段落，从而生成长篇故事。

- ReAct（Yao 等人，2022年）将思维链推理的生成与执行任务相关的行动交替进行（例如，搜索、查找）。

- WebGPT（Nakano et al., 2021年）通过基于文本的网页浏览环境微调语言模型，以回答长篇问题。

- Gao et al.（2022年）将计算任务交由 Python 解释器处理。
- Trivedi et al.（2022）和 Khattab et al.（2022）将思维链与知识检索步骤交替进行。
- Khot et al.（2022）研究了一般性的分解方法，让语言模型将子任务交给特定任务处理器。
- Ozturkler et al.（2022）使用求和和乘积等数学组合算子聚合语言模型的概率。
- Jung et al.（2022）递归地为一个陈述生成解释树，然后将推断视为关于这些解释及其逻辑关系的可满足性问题，以确定该陈述的真伪。
- 包括 Fu et al.（2021）、Dua 等人（2022）、Guo 等人（2022）和 Zhou 等人（2022）在内的多项研究，探讨了将问题分解为子问题的方法，这通常称为多跳问答。

训练和微调技术

最相关的工作是 Uesato 等人（2022），他们直接比较了语言模型解决数学应用题时基于过程的反馈和基于结果的反馈。他们发现，基于过程的反馈能提高推理步骤的正确性和可解释性，但相比基于结果的反馈，需要更多的标签监督。

Wu 等人（2021）使用行为克隆和奖励建模对 GPT-3 进行微调，使其递归地完成摘要。Zhou 等人（2022）基于平行新闻语料库训练了一个分解模型。Nakano 等人（2021）比较了行为克隆、强化学习和拒绝采样在训练网页浏览模型方面的效果，并发现，结合行为克隆与针对奖励模型的拒绝采样效果最佳。Zelikman 等人（2022）生成思维链，并使用那些能够得出正确答案的思维链进行微调。
表 1 中“训练”部分的大多数其他工作，都是针对特定的组合步骤从头开始训练小型模型。

理论与概念进展

Wies 等人（2022年）展示了过程监督的一些益处：将中间监督拼接到输入中，并在此修改后的输入上训练序列到序列模型后，原本无法学习的复合问题也可能变得可学习。

Press等人（2022年）将“组合性差距”这一名称用于这样一类问题：在模型正确回答了所有子问题的题目中，模型却答错的题目所占的比例。他们发现这一比例约为40%。他们表明，思维链可以缩小这一差距，而生成并回答后续问题则能进一步缩小差距。

工作流程、教程与工具

Dohan 等人（2022年）介绍了语言模型级联，这是一类将语言模型与随机变量和控制流组合起来的概率程序。他们将若干现有技术形式化为语言模型级联的实例，例如草稿区、验证器、STaR、选择-推断和工具使用。他们还提供了一个开源概率编程系统。《因子化认知入门》（Stuhlmüller 等人，2022年）是一份教程，讲解如何编写组合式语言模型程序，包括递归式问答、辩论、搜索和验证。Xie 等人（2022年）在人类参与者研究中表明，让用户通过子问题探查语言模型，有助于用户校准对模型回答是否正确的判断。Wu 等人（2022a）和 Wu 等人（2022b）介绍了一款闭源可视化编程界面，用于构建组合式语言模型程序。Dust（Polu，2022年）是一个网络服务和 Rust 库，可用于设计和部署语言模型应用。LangChain（Chase，2022年）是一个 Python 库，可协助开发将多个语言模型彼此串联或与其他专家串联的语言模型应用。

与本文的关系

尽管关于任务分解的文献迅速增长，但在理论方面实际上尚无相关研究，构建现实世界中的流程系统所需的工具、教程和工作流程也寥寥无几。我们认为，这篇论文既展示了一个现实世界中的应用（科学问答），也描述了所使用的工作流程（迭代分解）和工具（ICE）。

上述先前研究对“过程”和“监督”的具体含义各有不同。我们在附录 9.3 中提供了一个分类体系。简而言之，在本研究中，我们关注由人类开发者设计、偶尔由语言模型做出选择的分解。我们兼顾旨在提升任务表现的实用型分解，以及体现理想推理、便于更好地进行监督的分解。我们的分解主要通过帮助模型更有效地利用长上下文来提升表现，不过有些分解还会将多条推理思路应用于

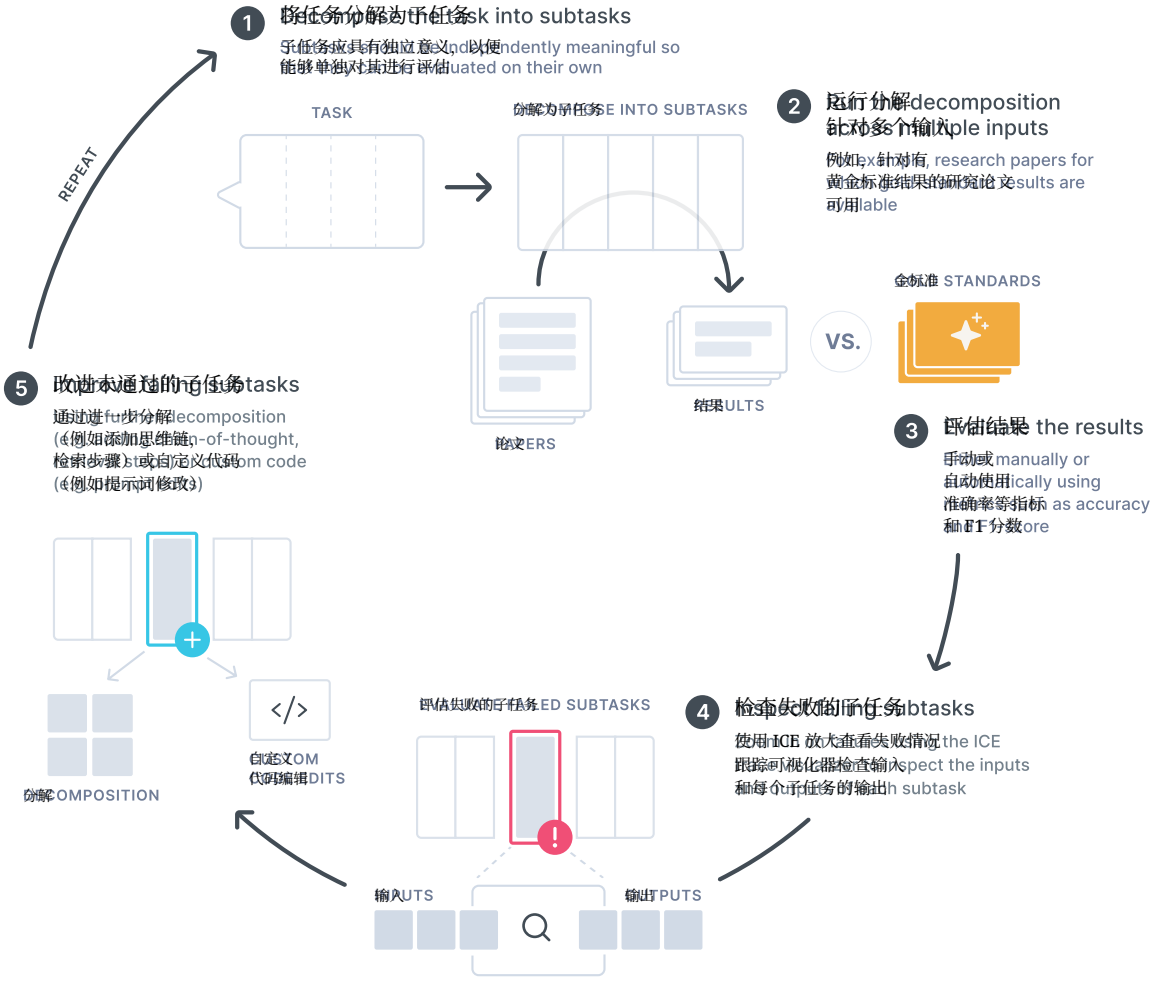


图2：迭代分解是一种用于人在回路中的语言模型编程的工作流。我们从一个简单的分解开始，根据金标准对其进行评估，使用 ICE 可视化工具诊断失败来源，通过进一步分解或其他调整来改进未通过的子任务，然后重复这一过程。

子任务。当我们谈到“监督”时，指的是检查组合中单个步骤的输出或行为。我们通过监督这一过程来改进语言模型程序；监督者是人类开发者，而且我们始终可以获取正确答案。

2.2 迭代分解

我们通过迭代分解研究过程监督。这是一种人在回路中的工作流，通过错误诊断和修正逐步改进任务分解（如图2所示）：

1. 从最简分解入手，将任务拆分为语言模型能够执行的子任务。例如，提取学术研究中使用的安慰剂这一任务，可以先拆解为找到最相关的章节，再根据该章节生成安慰剂描述（如果有）。2. 将这种分解应用于多个带有金标准答案的测试输入。例如，可以使用来自医学、心理学和经济学等不同领域的学术论文数据集，以及相应的安慰剂描述（如适用）。

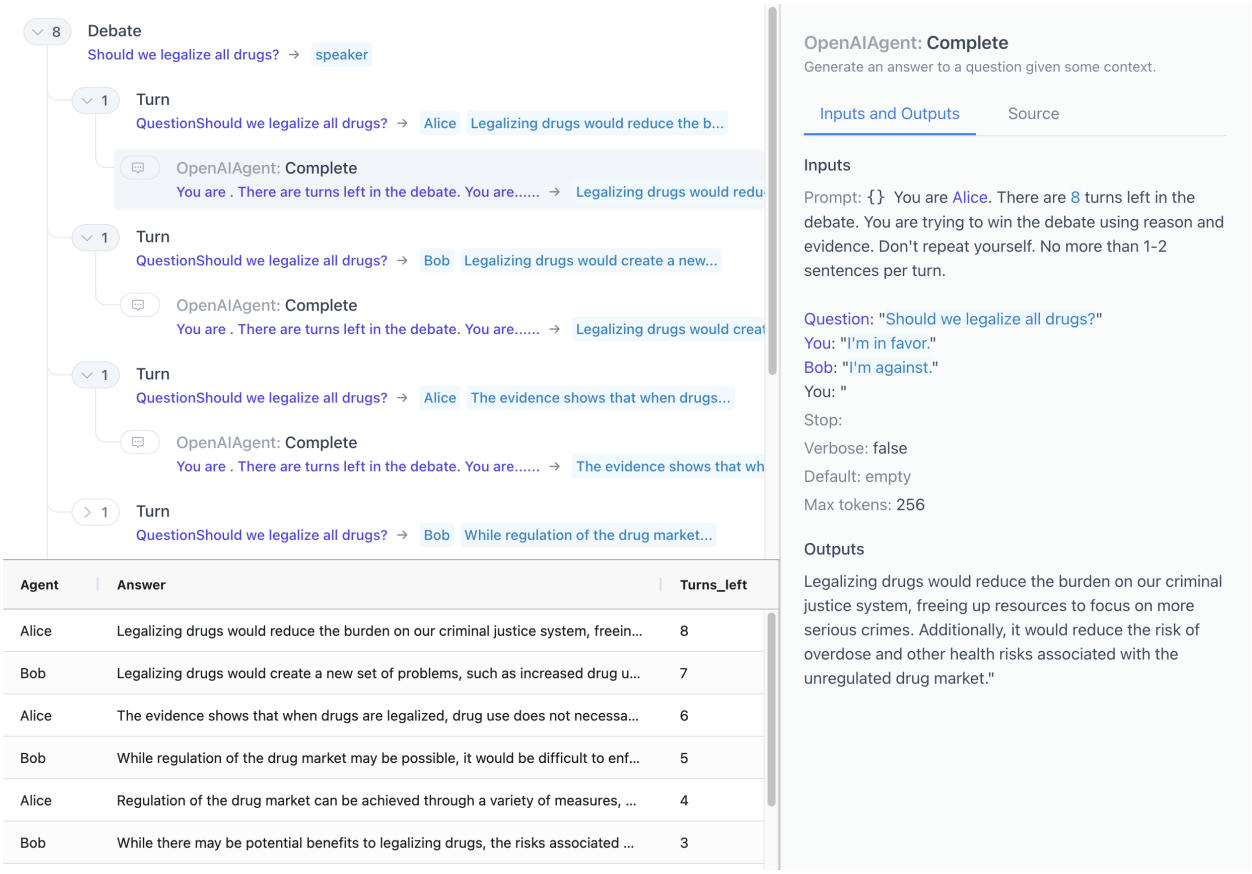


图 3：交互式组合浏览器（ICE）是一款用于语言模型执行轨迹的开源调试器。它可以将任务分解可视化，并支持放大查看出错的子组件。截图左上方面板显示了原始执行层级中的函数调用轨迹。下方面板支持对调用进行排序和筛选，每一行都显示执行时记录的值。右侧面板则显示主面板中所选调用的输入、输出、记录的中间值以及源代码。

3. 使用语言模型自动评估结果，或手动评估，并采用准确率、F1 分数或其他指标。例如，可以将安慰剂提取任务的结果与学术论文中安慰剂描述的黄金标准数据集进行比较。4. 使用 ICE 跟踪可视化器或其他工具检查每个子任务的中间输入和输出，以找出失败之处。例如，如果生成的安慰剂描述不正确，就可以判断是找错了章节，还是找对了章节却提取出了错误答案。5. 通过进一步分解（例如，将语义分解为子任务、思维链、检索），或进行自定义修改（例如调整提示词、语言模型推理参数），改进表现不佳的组件。例如，可以将任务拆分为先找出论文中提到的每个实验，再逐一询问每个实验是否使用了安慰剂。

这一过程会重复步骤 2–5，直到取得良好表现，或耗尽相关资源（时间、用于实验的计算预算）。

2.3 交互式组合浏览器

为了支持迭代分解中的错误识别步骤，我们开发了交互式组合浏览器（ICE），这是一款用于语言模型程序的开源[3] 调试器和执行轨迹可视化工具。

ICE 会记录并可视化 Python 程序中所有异步函数的执行过程，包括所有输入（提示词、参数）和输出（语言模型响应、中间值），这种方式甚至适用于执行

[3]https://github.com/oughtinc/ice



图 4：我们将迭代分解应用于三个真实案例研究：从有关随机对照试验的学术论文中提取安慰剂、提取参与者流程，以及自然语言处理问答。这些任务属于我们为人工智能研究助手 Elicit 开展的自动化文献综述工作。

包含数百至数千次语言模型调用的情况。ICE 提供了一种装饰器，可应用于任何需要记录的异步函数；它还提供了一些实用工具，用于记录给定模块中定义的所有顶层异步函数、函数中任意感兴趣的自定义值，以及每个插值字符串的结构。随后，可在浏览器中可视化这些数据，以便交互式探索，如图 3 所示。

ICE 提供了三种查看这些数据的视图：

1. 可展开的函数调用树，按执行顺序显示 2. 可排序、可筛选的函数调用及其自定义值表格 3. 函数调用详情窗格，其中包含输入、自定义值、输出和源代码

可以浏览这棵树，找到特定的函数调用，然后在详情窗格中检查该调用。通常，比较同一函数的多个调用会很有帮助，因此 ICE 提供了一个下拉菜单，列出执行跟踪中所有记录的函数及其各自的调用次数。选中某个函数后，表格中会显示对该函数的所有调用，树中也会突出显示这些调用。

由于提示词在语言模型程序中尤为重要，详情窗格专门支持渲染插值字符串（格式化字符串）。每个插值都会以交替颜色显示，插值所用的源代码则会显示在工具提示中。

这些函数都旨在帮助理解、分析和调试任务分解：既可以从高层次查看调用树结构，也可以在表格中查看某一特定类型的所有函数，还可以查看该类型的某个实例生成意外结果时所使用的具体提示词。

3 案例研究的真实背景

我们在使用 Elicit 期间遇到了三个真实案例研究，并通过迭代分解和 ICE 提升了性能，优于简单的基线（图 4）。

Elicit 的一个主要应用场景是查明学术文献对某个研究问题有哪些认识。许多 Elicit 用户从事生物医学、心理学、实验经济学及其他领域的工作，在这些领域中，随机对照试验（RCTs）是重要的证据来源。这些用户进行文献综述或荟萃分析时，通常大致遵循以下流程：

1. 哪些随机对照试验可能与回答我的研究问题最相关?
2. 对于每项随机对照试验：

   （a）这项研究是否真正回应了我的研究问题？（b）偏倚风险是多少？我能否直接接受这些研究结果，还是需要对其打折扣或予以忽略?
3. 按证据质量加权的随机对照试验汇总，对我的研究问题说明了什么?



图5：在安慰剂案例研究中，目标是确定某项随机对照试验是否使用了安慰剂，如果使用了，具体是什么。我们从 Elicit 使用的选择并生成基线开始，并通过基于试验组和段落的集成分类对其进行改进。

安慰剂分类 & 描述 任务旨在帮助研究人员评估偏倚风险。有关安慰剂的信息有助于研究人员评估研究中的偏倚风险，并判断结果的可信度——与安慰剂对照相比，与无治疗对照相比通常能提供更有力的证据，尤其是在安慰剂成功使参与者保持盲法的情况下。

RCTs 中**参与者流程**任务的第一步是识别每项研究中的试验和试验组。这有助于用户了解研究人员在研究中做了什么，从而判断该研究是否回答了他们的研究问题。第二步是评估参与者对干预的依从性，这有助于研究人员评估偏倚风险，并结合上下文理解研究结果：效果较小是否是由治疗依从性较差造成的?

这些任务代表了 Elicit 中范围广得多的一类任务：回答有关论文的问题。Elicit 目前支持约 20 个预设问题，同时也允许用户输入自己的问题。最终，我们希望找到具有高度泛化能力的策略来提升表现。**QASPER NLP Q&A** 任务通过将参与者流程分解泛化到另一个领域来检验这一点。

4 案例研究：安慰剂分类与描述

在本案例研究中，我们聚焦于领域特定分解。通过反复迭代分类和描述步骤，生成的安慰剂描述在留出测试集上的准确率从 25% 提高到了 65%。

4.1 设置

给定一项 RCT 的全文，任务是回答“这项 RCT 是否使用了安慰剂？”（分类）以及“如果使用了，是什么安慰剂？”（描述）。对于分类和描述，人工评审者之间都表现出高度一致。
附录9.5.1对任务作了详细说明。附录中的表 9.5.2 展示了两篇论文中最相关引文及其对应分类和描述的示例。

4.2 迭代

4.2.1 基线：类似 Elicit 的选择并生成

基线是 Elicit 目前用于论文问答的“先选择再生成”算法。Elicit 使用 monoT5[4] 根据问题对论文中的段落进行排序（Lin 等，2021）。排名最高的段落会被输入 text-davinci-002，并附上以下提示词：

> 根据研究论文的摘录回答问题“*{{问题}}*”。请尽量作答；如果你不知道如何回答，请说“摘录中未提及该问题的答案”。请完整涵盖论文摘录中关于答案的所有内容。确保你所说的一切都有摘录作为依据。摘录可能引用其他论文；请回答你正在阅读其摘录的那篇论文相关内容，而不是它所引用的论文。''
>
> 请用一个短语或句子回答：

[4]https://hug gingface.co/castorini/monot5-base-msmarc o-10k

```
论文标题： {{标题}}论文摘录：
{{段落}}问题： {{问题}}答案：
```

在留出测试集上，该方法的分类准确率为 71%，描述准确率为 25%（表 2）。

4.2.2 改进分类

诊断

在基线分类错误的 14 次试验中，实际上每次都有安慰剂。在所有情况下，筛选步骤都未能找到论文中最明确指出存在安慰剂的部分。在 14 次试验中的 10 次里，筛选未找到存在安慰剂的证据。在剩下的 4 次试验中，筛选找到了对安慰剂的一些提及，但没有找到完整描述，随后生成过程未能得出存在安慰剂的结论（示例见附录 9.5.3）。

为改进分类　　因此，我们需要一种更稳健地查找安慰剂相关证据的方法，同时避免引入假阳性。

解决方案

我们创建了一种分解方式，以模拟我们在确定安慰剂分类和描述时的推理过程。阅读验证集中的论文时，我们注意到两点：首先，几乎所有论文都会清楚地描述其试验组。如果研究中设有安慰剂，通常会有一个试验组被明确标识为安慰剂组。因此，我们会识别并描述每个试验组，并判断其中是否有试验组属于安慰剂。其次，论文中往往有许多段落提供关于该论文是否使用安慰剂的证据，有时这些证据彼此矛盾。因此，我们可以检查每个段落对该论文是否使用安慰剂提供了什么信息，然后汇总这些答案。

概括而言，分解过程如下：

1. 根据试验组进行分类
   (a) 试验组有哪些？i. 通过成对比较，按与此问题的相关性对段落进行排序 ii. 根据最相关的段落作答 (b) 描述每个试验组：排序并作答 (c) 是否有试验组看起来像安慰剂？(d) 如果有，参与者能否分辨自己属于哪个试验组？（如果能，那就不是真正的安慰剂。）
2. 根据每个段落进行分类
   (a) 根据每个段落，是否使用了安慰剂？(b) 汇总各段落的答案
3. 集成基于试验组和段落得出的分类结果

附录 9.5.4 展示了完整的分解过程。

结果

这种方法显著优于基线（正确率为 71% $\rightarrow$ 98%；p=0.0004），达到了近乎完美的表现（表 2）。

检查 ICE 中的执行轨迹后发现，性能提升几乎完全归功于基于各段落作答，而非基于试验组作答。每当基于试验组的分类方法暂时将某个试验组判定为安慰剂组时，它也会发现参与者可能能够辨别自己身处哪个试验组，因此最终会表示无法确定是否设有安慰剂组。

4.2.3 改进描述

诊断

在基线未能正确描述安慰剂的 15 次试验中，有 14 次也是基线未能正确分类的试验。如果分类结果为“无安慰剂”，基线甚至不会尝试描述安慰剂，因此按其设计，它在所有这些试验中都会失败

方法	分类 (n=48)	描述 (n=20)
Elicit 先选择后生成基线	71%	25%
采用“未提及”困惑度选择的先选择后生成	65%	10%
将论文内容放入提示词	69%	30%
最终分解	**98%**	**65%**
关键词决策树	94%	10%

表 2：通过迭代改进任务分解，安慰剂分类的准确率在留出测试集上从 71% 提升到 98%，描述准确率则从 25% 提升到 65%。

。然而，在这些案例中，选择都未能找到描述所需的摘录，因此任何描述尝试都会失败。

解决方案

我们注意到，描述安慰剂所需的信息通常散布在论文各处。因此，我们会对论文中最相关的段落进行排序，然后利用这些段落生成答案。我们重复使用了用于分类的同一种排序与回答技术。这种通用的子任务分解适用于这两种目的。

结果

这种方法带 来了显著提升，优于 Elicit基线（正确率为 25% $\rightarrow$ 65%； $p = 0.025$—见表 2） .

4.2.4 关键词基线

基于段落的分类结果十分出色，这引出了一个问题：我们能否用一个简单得多的算法，来编码我们判断某项试验是否使用安慰剂的方法?

我们创建了一个简单的正则表达式关键词匹配算法：

> 1. 如果论文中包含“开放标签”之类的词语，则分类为无安慰剂；2. 接着，如果论文中包含“安慰剂”之类的词语，则分类为有安慰剂，并将第一个匹配的句子作为描述；3. 如果以上词语均未出现，则分类为无安慰剂

该算法在分类方面的表现与我们的任务分解大致相当。然而，基于正则表达式关键词的方法无法推广到更困难、歧义更大的任务中。例如，论文会以多种不同措辞讨论依从性，因此关键词不太能帮助选择相关段落（附录 9.6.2）。

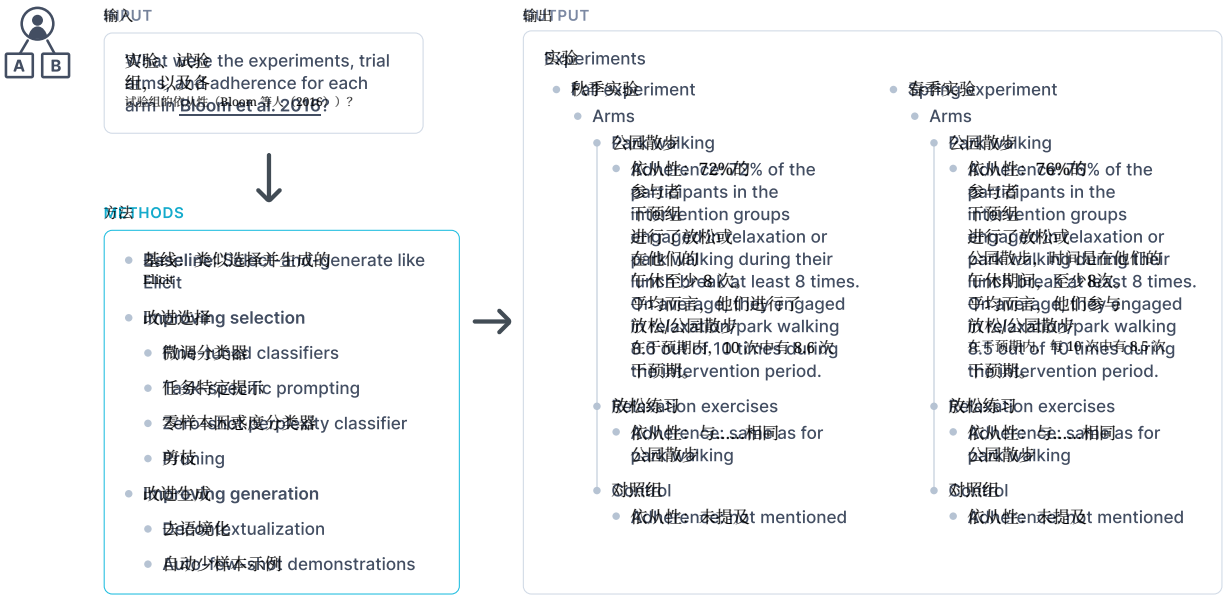
5 案例研究：随机对照试验中的参与者流程

在本案例研究中，我们主要关注适用于长篇问答的简单、可推广的分解。通过迭代改进选择和生成步骤，所有子任务在留出测试集上的准确率均大幅提升：实验提取准确率从 40% 提升至 70%，试验组提取准确率从 55% 提升至 86%，依从性提取准确率从 53% 提升至 70%。

5.1 设置

随机对照试验通常会包含一张标准化图表，帮助读者了解某项试验中发生了什么：参与者如何经历研究过程，从入组到最终分析？这些图表称为 *CONSORT* 图（示例见附录 9.6.1）。据我们观察，近期随机对照试验中约有一半包含此类图表，但这些图表往往并不完整。如果我们能为所有随机对照试验生成完整的 CONSORT 图，就能为读者提供有价值的信息。

在本案例研究中，我们考虑这一任务的一个有限版本：列出论文中的试验及每项试验的试验组，并描述各试验组的参与者依从率。图 6 展示了一个示例。


A
B
输入UT
实验、试验组以及各
arm in Bloom et al. 2013?
对比THODS
Elicit
Arms
Arms
Arms

评估了小型数据集和人在回路标注方法，优先考虑领域专家标注工作量不超过几小时的方法。

我们改编了 Liu 等人（2022）的方法，用于对文本分类子任务的分类器进行微调。该方法将低秩适配（通过学习得到的向量缩放选定的激活值）与损失函数中的额外正则化项，以及作者发现具有良好泛化能力的超参数相结合。这种组合实现了：（a）仅需微调极少量的权重，使多个微调任务能够在推理时共享同一个模型骨干；（b）真正的低数据量微调，因为该方法规定了超参数，从而无需单独的训练集和验证集。

Liu 等人（2022）指出，他们的方法只需少至 20 个示例即可取得良好效果。我们使用弱模型将正例扩展到更大的数量，并获得了超过 20 个“困难负例”：对于我们的任务而言，大多数段落与回答给定问题无关，而手动收集哪怕数量适中的正例也耗时且成本高昂。我们集成了多个弱分类器，例如 monoT5、基于 BART 的分类器和 T0-3B，从数千篇论文中识别正例；随后由专家审核所有弱分类器都判定为能够回答该问题的示例，同时保留这一方法产生的负例作为“困难负例”。直观地说，我们预期这种方法会产生数量过多的“简单正例”（来自集成共识）和“困难负例”（来自集成共识、经专家纠正的假阳性）。尽管我们预期这种方法难以识别“困难正例”，但实践中它创建的数据集与 T-few 微调协同效果非常好，使留出集上选择子任务的依从性表现从微调前的 72% 提升到微调后的 94% 准确率。

我们以 T0-3B 和 T0pp（11B 个参数）为基础模型，采用 T-Few 方法进行微调。我们发现，在微调数据中保留零样本提示时，模型的泛化能力更好；这可以视为一种正则化形式（在本例中，解码器的答案选项为“是”和“否”）：

> 上下文：*{{段落}}*
>
> 部分：*{{section}}*
>
> 如果下列句子谈论的是有多少参与者遵守了研究方案、不得不退出或主动退出，请回答“是”；如果谈论的是其他内容，例如研究设计、抽样策略或结果，请回答“否”。
>
> 句子：*{{sentence}}*
>
> 答案：

在选择阶段，使用微调分类器未能得到正确答案的许多情况，根源在于分类器产生了假阴性（在使用 $n = 9$ 的开发集中，67% 的失败案例中，分类器漏掉了金标准摘录）；而生成阶段对精确率降低所造成的“干扰项”有一定的鲁棒性。因此，我们探索了这种技术的多种变体，以提高召回率。例如，我们尝试对句子进行分类（同时将段落作为上下文），然后在推理时，如果段落中的任一句子被分类为正例，就将该段落判为正例。我们还尝试使用多个零样本提示模板进行微调，然后在推理时，如果某个段落通过任一提示模板被判为正例，就将该段落判为正例。

在遵循子任务上，微调分类器的表现显著优于基础 monoT5 分类器（段落分类 $F_1 = 0.10$：Elicit 的 top-1 方法；与 $F_1 = 0.77$：通过对 T0-3B 进行 T-Few 微调的方法）。
总的来说，虽然这种方法显著提升了性能，但也有缺点：它需要数据和训练，而且可解释性不如我们后来探讨的一些分类方法。未来，我们会考虑对分类器进行数据高效微调，适用于以下情形：(a) 任务难以明确界定，但数据容易收集；或 (b) 需要较小的微调分类器带来的推理时计算优势。

任务特定提示

微调的另一种替代方案是针对特定子任务设计提示词。这能减轻微调所需的数据收集负担，只需少量示例；但代价是需要具备另一类专长，即设计有效提示词的能力。

使用 text-davinci-002 时，我们发现，通过提示工程可以显著提升特定子任务的表现。朴素的零样本或少样本指令，效果不如“模拟”方法：后者通过提示词模拟具有某些理想特性的文本体裁，例如包含正确的数学

推理。附录 9.6.4 展示了一个提示词示例，它将任务设定为摘录自一本关于评估科学论文的（虚构）教科书。附录 9.6.5 展示了一个模拟 StatsExchange 帖子的提示词。
由于在本案例研究中，我们希望聚焦于通用分解；同时，我们预计随着未来指令微调模型的发展，基于模拟的方法将不再那么必要，因此没有系统地评估这类方法。

零样本困惑度分类器

理想情况下，我们可以改进相较于基线 monoT5 分类器的选择表现，同时采用一种既不需要大量数据收集，也不需要针对特定任务进行提示工程的方法。
受基线方法中“生成”阶段相对强劲表现的启发，我们使用 text-davinci-002 实现了一种基于零样本困惑度的交叉编码器。它在许多问题上的表现都相当出色，优于 monoT5 分类器，并且可与 T-Few 低秩微调相媲美。

在这种段落分类方法中，我们使用了以下 text-davinci-002 提示词，并通过 OpenAI API 回传对数概率，以测量困惑度：

> 根据研究论文的摘录回答“*{{问题}}*”。请尝试回答；如果你不知道如何回答，请说“摘录中未提及该问题的答案”。请涵盖论文摘录中关于答案的所有内容。确保你所说的一切都有摘录作为依据。摘录可能会引用其他论文；请回答你正在阅读其摘录的那篇论文，而不是它引用的论文。用一个短语或句子回答：论文摘录：*{{段落}}*问题：*{{问题}}*答案：摘录中未提及该问题的答案

随后，我们根据序列“摘录中未提及该问题的答案”中词元 X 的逆困惑度为各段落评分： $\{(x_{i=1}, \ldots, x_t)\}$

$$\exp\left(\frac{1}{t}\sum_{i=1}^{t}\log p_\theta(x_i|x_{<i})\right) \tag{1}$$

我们研究了这些分数的不同用途。在使用这种基于困惑度的方法的最简单方案中，我们还使用此提示词，根据“未提及”补全的困惑度最高的段落生成答案。这与 Elicit 使用的问答方法结构非常相似，只是用 text-davinci-002 基于困惑度的排序器替代了 monoT5排序器。

与 monoT5 排序器相比，text-davinci-002 困惑度方法似乎更能应对细微但重要的语义差异。例如，在依从性任务中，monoT5 排名第一的段落往往是某个 讨论依从性的段落，但并未针对相关的特定试验组。在许多情况下，text-davinci-002 困惑度方法能正确地将这些“讨论依从性但并非针对正确试验组”的段落评分低于针对正确试验组的结果段落；后者虽然没有明确讨论依从率，却提及了该试验组的依从性信息。

此外，我们发现，在回答阶段按原论文中的顺序纳入多个段落通常效果更好。我们通过经验选择分类阈值，将困惑度 top-1 方法改造成分类方法，并略微偏重召回率而非精确率；困惑度阈值在不同问题之间的适应效果相当不错，因此我们在所有任务中使用了相同的阈值。不出所料，纳入多个高于阈值的段落有助于弥补排序器仍存在的不足，而且在某些情况下，生成完整答案需要 多个段落。与只纳入一个段落相比，纳入多个段落主要解决了 top-1 方法错误地回答“未提及”的情况。

这种零样本分类方法在“先筛选、后生成”方法的“筛选”阶段表现尤为出色，在依从性任务中，其段落分类召回率从排名第一的 monoT5Elicit 基线的 0.09 提高到了 0.76。该方法属于零样本方法，无需额外数据（但需要使用数据为所有任务统一设置分类阈值）。此外，该方法还具有内置的可解释性优势：我们可以查看最有可能的补全结果，只要它不是“摘录中未提及该问题的答案”，就能了解模型认为每个段落中的哪些内容相关。

方法	实验 (n=50)	Arms (n=50)	依从性 (n=135)
Elicit 先选择后生成基线	40%	55%	53%
+ “未提及”困惑度选择	55%	78%	61%
+ “未提及”困惑度选择 +剪枝	60%	70%	54%
+ “未提及”困惑度选择 +自动少样本选择	65%	72%	55%
+ “未提及”困惑度选择 +自动少样本选择 + 剪枝	55%	56%	58%
+ “未提及”困惑度选择 +自动少样本生成	**70%**	**86%**	**70%**
+ “未提及”困惑度选择 +去语境化	65%	75%	54%

表3：通过迭代改进参与者流程分解，在留出测试集上，实验的提取准确率从40%提高到70%，试验组从55%提高到86%，依从性从53%提高到70%。

这种方法还通过一种自然的解读方式和方法，同时实现了分类与排序，且排序结果经过了相当好的校准；这意味着，例如，如果我们发现在回答阶段要处理的段落太多，超出了模型的上下文窗口，我们就可以有把握地直接移除排名较低的段落，无需额外工作。

剪枝

许多“选择”阶段的实现方式都会独立考察每个段落，并经过调整，使其略微偏向召回率而非精确率。从概念上讲，这可能导致我们选入冗余信息；也可能出现这样的情况：当我们查看所有选出的候选段落后，哪些段落包含答案会变得更加清楚。因此，我们测试了一个“剪枝”步骤：按照论文中的顺序，查看前一个选择阶段选出的所有段落，然后重新选出回答问题所需的段落。

尽管我们测试了这一阶段的思维链和少样本变体，但结果仅显示出中等程度的希望。增加这一阶段未能显著提高总体准确率；在不大幅损害召回率的情况下，我们也无法提高段落分类的精确率。对于实验任务，剪枝使精确率从0.14提高到0.29，同时使召回率从0.77降至0.54，并使段落分类 $F1$ 从0.24提高到0.38。最终答案准确率因此从0.55提高到0.6。进一步迭代改进这一阶段的实现方式，或许能取得更好的结果。

5.3.3 改进生成

改进“选择-再生成”方法的“选择”阶段，提高了参与者流程任务的表现。然而，ICE 轨迹显示，“生成”阶段也会产生一些有趣的错误；该阶段根据最相关的摘录回答问题。即使使用选择阶段的预言机，基线生成步骤在依从性任务上的准确率也只有 77%。

误差分析

We 在选择阶段具有较高召回率时，识别出了生成阶段的一些常见错误 l.

有些错误源于所选段落缺乏上下文。它们回答了问题，但由于缺少周围文本，答案并不清晰。这种情况在试验组任务中经常发生：所选段落提到了试验组，却没有提及相应的试验或实验方法学，因此很难准确描述试验组。

更常见的是，所需的答案风格和详细程度并不明确，尤其是在依从性任务中。例如，有些论文称“依从性总体良好”，并给出了数值细节，但生成阶段只包含了前者，没有意识到后者同样重要。

下面我们将讨论两种尝试过的减少此类错误的方法。

去语境化

先选择后生成流程可能难以完整回答需要消解共指或理解周围文本上下文的问题。为此，我们尝试引入一个去语境化阶段，灵感来自 Choi 等人（2021）：该阶段处理论文，并将前面段落中的上下文添加到方括号内。这样做的目的是让段落更加自足，更便于用于生成。我们使用基于 Eisenstein 等人（2022）的固定少样本提示词（见附录 9.6.6），通过自回归模型执行去语境化。我们没有系统地评估所添加上下文的准确率或相关性，但使用 ICE 抽查了一些示例。

去语境化提升了部分任务的生成性能，但并非所有任务都如此。它帮助模型识别试验中的实验和试验组，因为新增的上下文澄清了方法学以及选择的相关性。然而，它对依从性任务的帮助不大，因为大多数依从性信息本已自成一体，而新增的上下文反而引入了更多噪声和混淆。将去语境化段落作为生成阶段的输入时，实验和试验组的性能分别为 0.65 和 0.75，而基线分别为 0.40 和 0.55。对于依从性任务，该数值为 0.54，仅略高于 0.53 的基线。

自动生成的“示例”

我们研究了使用示例来应对生成阶段的一种失败模式：生成细节不足或风格不恰当的答案。示例是具有相同结构的问答任务实例，包括：(a) 一个问题，(b) 一段文本，(c) 支持答案的该文本摘录，以及 (d) 一个答案。与其他编码专家知识的方法（例如安慰剂分解或安慰剂关键词决策树算法）不同，示例不要求专家进行任何特定于算法或语言模型的工作。专家只需给出应用其专业知识的示例。

我们通过在现有的困惑度选择提示前添加多个已填写的示例，根据专家示例（金标准）生成了少样本提示。我们采用的方式有两种：一种只使用正向示例，其中包含支持性段落以及专家根据该段落手写的答案；另一种则混合使用正向和负向示例，并假定未标记的段落无法回答问题。这个假设往往不成立，因此我们发现，在少样本示例提示中只使用正例效果更好。

在所有任务中，由示例构建的少样本提示都比零样本方法提高了生成准确率，即便示例是从开发集中随机抽取的。对于更难的任务，或受益于隐性背景知识的任务，准确率提升更为显著。示例往往能通过展示任务预期的答案类型来发挥作用，例如在有数据时提供具体数值。比如，在依从性任务中，零样本方法给出了错误答案：“论文称，两组在每家护理院中对所有产品的遵从性都很好”；而通过示例采用少样本方法，则给出了正确答案：“口服补充剂的遵从性良好，每日平均摄入量约为 400 千卡”。对于这项任务，通过示例采用少样本方法将整体最终答案准确率从 0.53 提高到 0.70。

5.3.4 综合改进

零样本困惑度分类器和通过“示例”进行的少样本问答，使效果较 Elicit基线有了显著提升（例如，依从性任务的正确率达到 53% $\rightarrow$ 70%，p=0.006[5]；见表3）。这些技术大幅减少了依从性任务中的假阴性数量（误差分析见附录 9.6.3）。

6 案例研究：QASPER 自然语言处理问答

我们主要围绕上述任务反复改进任务分解。不过，我们也希望探索这些任务分解在其他领域和问题上的泛化能力，因此在 QASPER 数据集的一个子集上进行了测试。该数据集包含 5,049 个由自然语言处理从业者撰写的问题及答案，涉及 1,585 篇自然语言处理论文（Dasigi 等人，2021年）。采用最终的参与者流程分解后，QASPER 数据集上自然语言处理问题的问答准确率从 38% 提高到了 69%。

6.1 设置

QASPER 问题的答案有三种形式：是/否、文本摘录或自由形式回答。在最初的 QASPER 论文和基准测试中，作者衡量了生成的答案与专家答案之间的 F_1 分数。

[5]费希尔精确检验，将每个依从性回答视为相互独立。实际上，这些回答按试验和论文聚类，因此真实的 p 值略高。



图 7：在 QASPER 案例研究中，目标是回答有关自然语言处理论文的问题。QASPER 是一个包含 5049 个信息检索问题的数据集，这些问题由自然语言处理论文的普通读者提出，并由不同的自然语言处理专家作答。除了微调之外，我们还探索了与参与者流程相同的方法，以了解它们的泛化能力。

方法	准确率	F_1
Elicit 先选择后生成基线	38%	15%
+ “未提及”困惑度选择	55%	21%
+ “未提及”困惑度选择 + 剪枝	53%	20%
+ “未提及”困惑度选择 + 自动少样本生成	**69%**	**32%**

表4：QASPER 自然语言处理问答任务的结果（n=100）。方法的性能排序与参与者流程案例研究相同，为分解方法具有普遍性提供了一些证据。

由于我们关注的是答案是否正确，而非其是否与专家措辞完全一致，因此我们转而对生成的答案在语义上是否等同于专家答案进行二元评估。例如，对于是非题，专家答案为“是”，而生成的答案为“正确”，两者的 $F_1 = 0$，但我们仍会将其判为正确。为求完整，我们也衡量并报告 F_1 分数，但并未调整分解以最大化 F_1。

为使任务和评估易于管理且有意义，我们从 QASPER 中抽取了 100 个问题，用于测试基线，以及在参与者流程任务上表现明显优于基线的分解方法。测试前，我们进行了少量数据清理，移除了表示交叉引用的特殊标记，并将其替换为相应的自然语言形式（例如，将“BIBREF0”转换为“[1]”）；此外，与原始 QASPER 论文一样，我们省略了需要借助表格或图表才能回答的问题。

6.2 迭代

我们测试了我们选择的两者 括和不包括少样本示例
例中的示例）
，并在一组 QASPER 问题样本上进行了测试。表4展示了结果。
我们发现，在测试的各种方法中，在参与者流程问题上表现最佳的方法，在 QASPER 上也表现最佳。该方法先使用困惑度分类器，再使用根据示例生成的少样本提示（在本例中，示例仅由其他随机选取的 QASPER 问题支持段落-答案元组构成），在我们的样本上取得了 69% 的准确率。该方法显著优于两种方法：一种是仅使用分类器排名第一的段落，然后生成答案的基线方法（38%）；另一种是使用类似 Elicit 的困惑度分类器，然后在没有少样本示例的情况下生成答案（55%）。这些发现表明，这种分解在一定程度上能够稳健地应对领域迁移。

7 结论

我们介绍了迭代分解，这是一种用于过程监督的人在回路工作流。我们正在将这一工作流用于 Elicit——人工智能研究助手。为了支持这一工作流，我们构建了 ICE，这是一个用于语言模型程序的开源调试器。

随着时间推移，我们预计类似的工作流程将以大部分或完全自动化的方式运行，由语言模型创建分解、进行误差分析并调整程序。根据我们从案例研究中获得的经验，实现这一目标所需的工作如下：

1. 提高迭代速度。在完全自动化之前，我们预计会进入这样一个阶段：人类程序员每天可以通过多次迭代改进任务分解，从而自动化全新的复杂任务。2. 复杂的分解。以往工作中的分解几乎都相当简单，我们的一些案例研究也大致如此，即使它们会调用语言模型数百次。3. 为实现第 1 点和第 2 点，需要更复杂的开发者工具，以便更快地从总体结果追溯到错误来源，并提出解决方案。4. 让第 3 点中的开发者工具可供机器访问，使语言模型能够学会诊断和改进步骤。5. 缩小任务设计与执行之间的界限，让模型执行复杂任务时，能够在运行时进行分解、调试和迭代。6. 通过蒸馏及其他方式降低复杂分解的成本，使分解方案相较于端到端执行更具成本竞争力。

我们希望能够构建充足的基础设施，以实现可扩展的过程监督，使人类原则上始终能够理解模型如何解决复杂任务，即使最终几乎所有的分解和监督都由机器完成。

8 参考文献

Paul Christiano、Jan Leike、Tom B. Brown、Miljan Martic、Shane Legg 和 Dario Amodei。《基于人类偏好的深度强化学习》。2017年6月。Daniel M. Ziegler、Nisan Stiennon、Jeffrey Wu、Tom B. Brown、Alec Radford、Dario Amodei、Paul Christiano 和 Geoffrey Irving。《基于人类偏好的语言模型微调》，2020年1月。Nisan Stiennon、Long Ouyang、Jeff Wu、Daniel M. Ziegler、Ryan Lowe、Chelsea Voss、Alec Radford、Dario Amodei 和 Paul Christiano。《从人类反馈中学习摘要》。*arXiv*，第 2009.01325v1 页，2020年。Jeff Wu、Long Ouyang、Daniel M. Ziegler、Nisan Stiennon、Ryan Lowe、Jan Leike 和 Paul Christiano。《利用人类反馈递归摘要书籍》，2021年9月。Long Ouyang、Jeff Wu、Xu Jiang、Diogo Almeida、Carroll L Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Ray、John Schulman、Jacob Hilton、Fraser Kelton、Luke Miller、Maddie Simens、Amanda Askell、Peter Welinder、Paul Christiano、Jan Leike 和 Ryan Lowe。《通过人类反馈训练语言模型遵循指令》。第 68 页，2022年。William Saunders、Catherine Yeh、Jeff Wu、Steven Bills、Long Ouyang、Jonathan Ward 和 Jan Leike。《用于辅助人类评估者的自我批判模型》。2022年6月。doi:10.48550/arXiv.2206.05802。Reiichiro Nakano、Jacob Hilton、Suchir Balaji、Jeff Wu、Long Ouyang、Christina Kim、Christopher Hesse、Shantanu Jain、Vineet Kosaraju、William Saunders、Xu Jiang、Karl Cobbe、Tyna Eloundou、Gretchen Krueger、Kevin Button 和 Matthew Knight。《WebGPT：借助人类反馈的浏览器辅助问答》。第 30 页，2021年。Andreas Stuhlmüller 和 Jungwon Byun。《监督过程，而非结果》，2022年。Dario Amodei、Chris Olah、Jacob Steinhardt、Paul Christiano、John Schulman 和 Dan Mane。《人工智能安全中的具体问题》。2016年6月。Ajeya Cotra。《如果没有具体的应对措施，通往变革性人工智能的最简单路径很可能导致人工智能接管》。
h
t
t
p
s://
www.alignmentforum.org/posts/pRkFkzwKZ2zfa3R6H/without-specific-countermeasures-the-easiest-path-to，2022年。Richard Ngo、Lawrence Chan 和 Sören Mindermann。《从深度学习视角看对齐问题》。2022年8月。doi:10.48550/arXiv.2209.00626。Paul Christiano。《以认可为导向的智能体》。https://ai-alignment.com/model-free-decisions-6e6609f5d99e，2016年6月。Jonathan Uesato、Nate Kushman、Ramana Kumar、Francis Song、Noah Siegel、丽莎·王、Antonia Creswell、Geoffrey Irving 和 Irina Higgins。《通过基于过程和结果的反馈解决数学应用题》，2022年11月。Swarnadeep Saha、Shiyue Zhang、Peter Hase 和 Mohit Bansal。《摘要程序：利用神经模块化树实现可解释的抽象式摘要》，2022年9月。Shunyu Yao、Jeffrey Zhao、Dian Yu、Nan Du、Izhak Shafran、Karthik Narasimhan 和 Yuan Cao。《ReAct：在语言模型中协同推理与行动》，2022年10月。Sawan Kumar 和 Partha Talukdar。《NILE：借助忠实的自然语言解释实现自然语言推理》。载于计算语言学协会第*58*届年会论文集，第 8730–8742 页，线上，2020年7月。计算语言学协会。doi:10.18653/v1/2020.acl-main.771。Antonia Creswell 和 Murray Shanahan。《选择-推断：利用大语言模型实现可解释逻辑推理》，2022年5月。Jaehun Jung、Lianhui Qin、Sean Welleck、Faeze Brahman、Chandra Bhagavatula、Ronan Le Bras 和 Yejin Choi。《助产式提示：通过递归解释实现逻辑一致推理》。2022年。doi:10.48550/ARXIV.2205.11822。Soumya Sanyal、Harman Singh 和 Xiang Ren。《FaiRR：基于自然语言的忠实且鲁棒的演绎推理》，2022年3月。Antonia Creswell 和 Murray Shanahan。《利用大语言模型进行忠实推理》，2022年8月。Kaj Bostrom、Zayne Sprague、Swarat Chaudhuri 和 Greg Durrett。《通过搜索陈述组合实现自然语言演绎》，2022年10月。Siyuan Wang、Zhongyu Wei、Zhihao Fan、Qi Zhang 和 Xuanjing Huang。《先定位后提问：用于多跳问答的可解释逐步推理》，2022年8月。Ruiliu Fu、Han Wang、Xuejun Zhang、Jun Zhou 和 Yonghong Yan。《分解复杂问题使多跳问答更容易、更具可解释性》，2021年10月。

Dheeru Dua、Shivanshu Gupta、Sameer Singh 和 Matt Gardner。《用于分解复杂问题的连续提示》，2022年12月。Xiao-Yu Guo、Yuan-Fang Li 和 Gholamreza Haffari。《通过问题分解实现复杂阅读理解》，2022年11月。Kumar Shridhar、Alessandro Stolfo 和 Mrinmaya Sachan。《通过语义分解将大语言模型的多步推理能力蒸馏到更小的模型中》，2022年11月。Ben Zhou、Kyle Richardson、Xiaodong Yu 和 Dan Roth。《学习分解：基于可比文本的假设问题分解》，2022 年 10 月。Omar Khattab 、 Keshav Santhanam、Xiang Lisa Li、David Hall、Percy Liang、Christopher Potts 和 Matei Zaharia。《演示-搜索-预测：组合检索与语言模型以处理知识密集型自然语言处理》，2022年12月。Ofir Press、Muru Zhang、Sewon Min、Ludwig Schmidt、Noah A Smith 和 Mike Lewis。《测量并缩小语言模型中的组合性差距》。第25页，2022年。Tushar Khot、Harsh Trivedi、Matthew Finlayson、Yao Fu、Kyle Richardson、Peter Clark 和 Ashish Sabharwal。《分解式提示：解决复杂任务的模块化方法》，2022年10月。Batu Ozturkler、Nikolay Malkin、Zhen Wang 和 Nebojsa Jojic。《ThinkSum：使用大语言模型对集合进行概率推理》，2022年10月。Kevin Yang、Yuandong Tian、Nanyun Peng 和 Dan Klein。《Re3：通过递归重新提示和修订生成更长的故事》，2022年10月。Luyu Gao、Aman Madaan、Shuyan Zhou、Uri Alon、Pengfei Liu、Yiming Yang、Jamie Callan 和 Graham Neubig。《PAL：程序辅助语言模型》，2022年11月。Andrew Drozdov、Nathanael Schärli、Ekin Akyürek、Nathan Scales、Xinying Song、Xinyun Chen、Olivier Bousquet 和 Denny Zhou。《使用大语言模型进行组合语义解析》，2022年9月。Harsh Trivedi、Niranjan Balasubramanian、Tushar Khot 和 Ashish Sabharwal。《将检索与链式思维推理交错结合，以处理知识密集型多步骤问题》，2022年12月。Seyed Mehran Kazemi、Najoung Kim、Deepti Bhatia、Xin Xu 和 Deepak Ramachandran。《LAMBADA：自然语言自动推理的反向链式推理》，2022年12月。Eric Zelikman、Yuhuai Wu、Jesse Mu 和 Noah D. Goodman。《STaR：通过推理自举推理能力》，2022年5月。Kaige Xie、Sarah Wiegreffe 和 Mark Riedl。《使用分解式探针校准多跳问答系统的可信度》，2022年10月。David Dohan、Winnie Xu、Aitor Lewkowycz、Jacob Austin、David Bieber、Raphael Gontijo Lopes、Yuhuai Wu、Henryk Michalewski、Rif A. Saurous、Jascha Sohl-dickstein、Kevin Murphy 和 Charles Sutton。《语言模型级联》，2022年7月。Andreas Stuhlmüller、Justin Reppert 和 Luke Stebbing。《因子化认知入门》，2022年。Tongshuang Wu、Ellen Jiang、Aaron Donsbach、Jeff Gray、Alejandra Molina、Michael Terry 和 Carrie J. Cai。《PromptChainer：通过可视化编程串联大语言模型提示》，2022年3月。Tongshuang Wu、Michael Terry 和 Carrie Jun Cai。《AI Chains：通过串联大语言模型提示实现透明且可控的人机交互》。载于计算机系统中的人为因素国际会议。ACM，2022年4月。doi:10.1145/3491102.3517582。Harrison Chase。Langchain/langchain，master 分支 · hwchase17/langchain。https://github.com/hwchase17/langchain，2022年。Stanislas Polu。Dust。Dust，2022年12月。Noam Wies、Yoav Levine 和 Amnon Shashua。《子任务分解使序列到序列任务中的学习成为可能》，2022年10月。Jimmy Lin、Rodrigo Nogueira 和 Andrew Yates。《用于文本排序的预训练变换器：BERT及其后续发展》，2021年8月。Haokun Liu、Derek Tam、Mohammed Muqeeth、Jay Mohta、Tenghao Huang、Mohit Bansal 和 Colin Raffel。《少样本参数高效微调优于上下文学习，且成本更低》，2022年8月。Eunsol Choi、Jennimaria Palomaki、Matthew Lamm、Tom Kwiatkowski、Dipanjan Das 和 Michael Collins。《去语境化：使句子能够独立成篇》。计算语言学协会会刊，第9卷：447–461，2021年4月。ISSN 2307-387X。doi:10.1162/tacl_a_00377。

Jacob Eisenstein、Daniel Andor、Bernd Bohnet、Michael Collins 和 David Mimno。《来自不可信教师的诚实学生：从预训练语言模型中学习可解释的问答流水线》，2022年10月。

Pradeep Dasigi、Kyle Lo、Iz Beltagy、Arman Cohan、Noah A. Smith 和 Matt Gardner。《以研究论文为依据的信息检索型问题与答案数据集》。2021年5月。doi:10.48550/arXiv.2105.03011。

Alpha A. Fowler、Jonathon D. Truwit、R. Duncan Hite、Peter E. Morris、Christine DeWilde、Anna Priday、Bernard Fisher、Leroy R. Thacker、Ramesh Natarajan、Donald F. Brophy、Robin Sculthorpe、Rahul Nanchal、Aamer Syed、Jamie Sturgill、Greg S. Martin、Jonathan Sevransky、Markos Kashiouris、Stella Hamman、Katherine F. Egan、Andrei Hast- ings、Wendy Spencer、Shawnda Tench、Omar Mehkri、James Bindas、Abhijit Duggal、Jeanette Graf、Stephanie Zellner、Lynda Yanny、Catherine McPolin、Tonya Hollrith、David Kramer、Charles Ojielo、Tessa Damm、Evan Cassity、Aleksan- dra Wieliczko 和 Matthew Halquist。《维生素C输注对脓毒症和严重急性呼吸衰竭患者器官衰竭、炎症生物标志物及血管损伤的影响：CITRIS-ALI随机临床试验》。*JAMA*，322(13):1261–1270，2019年10月。ISSN 1538-3598。doi:10.1001/jama.2019.11825。

9 附录

9.1 作者贡献

Justin Reppert 和 Charlie George 负责参与者流程和 QASPER 案例研究。
Ben Rachbach 选择并设计了任务。
Luke Stebbing 实现了 ICE 的功能。
Maggie Appleton 制作了图表。
Andreas Stuhlmüller 开展了安慰剂案例研究并撰写了文献综述。
Andreas Stuhlmüller、Justin Reppert 和 Ben Rachbach 撰写了这篇论文。
Jungwon Byun 和 Andreas Stuhlmüller 管理了该项目。

9.2 致谢

Jason Zukewich 帮助 ICE 起步。
在我们完成案例研究工作的后期，Alex Hall 开始为 ICE 做贡献。
James Brady 帮助 Justin 和 Luke 管理他们的工作。
Paul Christiano 和 Owain Evans 为任务分解提供了有益的见解。
Étienne Fortier-Dubois 创建了用于测试这些分解的大部分金标准。
Beth Barnes、Ethan Perez、Harrison Chase、Jan-Hendrik Kirchner、Jackson Kernion、Jan Leike、Jonathan Uesato、Stanislas Polu 和 OwainEvansprovidedhelpful 为论文应包含哪些内容提供了有益的反馈。

9.3 过程监督的分类

在这篇论文中，我们使 “过程”指语言模型为得出答案而执行的显式推理 er.
过程可以从多个维度进行考察，例如：

- 因果性： 推理步骤如何关联，以及模型如何使用这些步骤。例如，思维链推理是模型在得出最终答案的过程中遵循的一种过程，但我们无法验证这些推理对于得出最终答案是否具有因果上的必要性。另一方面，如果我们先生成推理，再将其输出输入另一次调用，由后者在没有其他可用于直接解决任务的信息的情况下完成任务，那么我们就知道，这些推理是导致最终答案的因果链的一部分。
- 内容： 推理步骤在做什么。例如，模型可能会通过推理纳入更多上下文（比如长篇论文的更多章节）、使用外部工具（比如在两次模型调用之间进行网络搜索），或针对某个问题应用多个标准或推理思路。
- 设计者： 由谁决定如何组合推理步骤。例如，人类开发者、人类终端用户、语言模型，或由他们共同决定（比如开发者提供一组可能的子组合，再由语言模型从中选择）。
- 目的： 分解旨在实现的目标。例如，分解可能旨在提升主任务的表现、体现理想或规范性的推理过程，或使推理过程更易于监督。

在本文中，我们关注由人类开发者设计的分解，偶尔也由语言模型作出选择。我们兼顾旨在提升主任务表现的实用型分解，以及体现理想推理并便于监督的分解。我们的分解主要通过帮助模型更有效地利用长上下文来提升表现，不过其中一些也会对各个子任务采用多种推理思路。

其次是“监督”。我们将监督定义为检查分解中单个步骤的输出或行为。

监督可以从以下几个维度进行：

- 目标：监督的目的可以是：1）识别并修正分解中的失效模式；2）检测并纠正执行过程中的错误；3）评估最终任务答案，并决定是否可信；或 4）生成用于训练模型的反馈信号。
- 监督者：由谁执行监督？开发者、语言模型还是用户？
- 真实答案的可用性：我们是否知道任务的正确答案（这样就可以将过程输出与之进行比较），还是不知道正确答案（这样就只能根据过程质量自身的逻辑性和一致性来判断）？

在本文中，我们通过监督过程来改进任务分解；监督者是开发者，而且我们始终能够获得正确答案。

9.4 数据收集

对于安慰剂和参与者流程实验，我们根据自己制定的金标准答案评估了每个问题。为制定这些答案：

1. Surge[6] 数据标注平台上的两名工作人员分别写下了他们各自认为的金标准答案。2. 一名值得信赖的 Ought 工作人员审阅了两名 Surge 工作人员给出的答案，并确定了最终的金标准答案。如果 Surge 工作人员完成得很好，Ought 工作人员可以直接接受他们的答案；但更多时候，Ought 工作人员都需要进行大幅修改。3. 我们使用这些金标准进行评估时，偶尔会发现错误。发现后，我们便予以修正。

9.5 安慰剂案例研究详情

9.5.1 任务描述

我们将安慰剂定义为一种给予对照组的物质或干预，旨在模拟给予治疗组的物质或干预，但不含“活性成分”。安慰剂旨在让参与者无法判断自己属于治疗组还是对照组。

除了帮助研究人员评估偏倚风险之外，选择这项任务的另一个原因是它相当简单。我们希望在一个很可能取得成功的情境中尝试分解，并找出任何意料之外的失败模式。
在数据集中：

1. 几乎所有使用安慰剂的论文都会明确说明这一点，因此很容易识别出这些论文（例如上文的论文 1 和 3） 2. 但是，不能只搜索“安慰剂”一词，否则会出现误匹配（例如论文 4） 3. 一些未使用安慰剂的论文明确说明采用了开放标签（论文 2），而另一些则只是描述了不涉及安慰剂的实验程序（论文 5） 4. 有时只需从论文中提取主要描述，就能很好地说明安慰剂（论文 1 ）；但有时则需要根据多处摘录进行推断（论文 3）

我们评估的是每项试验，而非论文。试验是指参与者被随机分配到不同试验组，并对这些试验组进行相互比较。有些论文讨论了多项试验，而且有时部分试验采用盲法，其他试验则没有。我们的测试集中有 61 项实验，涉及 47 篇论文。

为了评估安慰剂描述，我们只考察事实上存在安慰剂的实验。如果分解指出不存在安慰剂，我们会立即将该描述判定为错误。否则，我们会检查与金标准相比，该分解的描述是否足够准确且完整。

[6]https://www.surgehq.ai

9.5.2 数据集示例

论文	最相关的引文	分类	安慰剂描述
1	“该研究是一项开放标签、随机对照临床试验。”	未使用安慰剂或未提及	–
2	“参与者被按 1:1:1:1 的比例随机分配至 4 个试验组之一：(1) 入组时开始每日服用 TDF，(2) 入组时开始每日服用安慰剂，(3) 入组 9 个月后开始每日服用 TDF，以及 (4) 入组 9 个月后开始每日服用安慰剂（图 1）。” "分配至 TDF 组的参与者同样可能或更可能认为自己被分配至安慰剂组而非 TDF 组；安慰剂组参与者的情况则相反，这表明并未发生大规模揭盲。"	安慰剂	每日安慰剂，参与者——裤子无法区分与治疗
3	“对照条件下的患者被列入等候名单，预计将在8个月后参加他们的同伴支持小组。” “缺少注意力安慰剂对照条件是一项局限。”	没有安慰剂或安慰剂未提及	–
4	“安慰剂含有口服阿奇霉素混悬液的载体，并装瓶并贴上与阿奇霉素相同的标签。”	安慰剂	口服制剂的载体 阿奇霉素 混悬液，装瓶并贴上与阿奇霉素完全相同的性地使用阿奇霉素
5	“两组均接受了干预前纸质调查，并在门诊就诊后 2 至 3 周接受了电话调查。干预组还获邀参加计算机培训，并收到了 IP 和培训摘要讲义。”	无安慰剂或安慰剂未提及	–

9.5.3 基线选择失败

选择提到了安慰剂，却未能找到安慰剂描述的示例：

1. https://pubmed.ncbi.nlm.nih.gov/27032063/（Ebner等，2016）

   （a）选择结果中仅发现了这一处关于安慰剂的提及：“……也就是说，尽管我们采用急性安慰剂对照药理学方法，揭示了催产素对杏仁核与内侧前额叶皮层之间静息态功能连接的影响，但我们对其机制解释的能力有限……”

   (b) 论文在其他地方描述了安慰剂：“在对静息态功能连接数据进行分析的79名参与者中，22名年轻参与者（50%为女性）和18名年长参与者（56%为女性）被随机分配通过鼻喷剂自行给药24 IU（每个鼻孔喷一次）催产素。18名年轻参与者（50%为女性）和21名年长参与者（62%为女性）在完整研究访视开始时自行使用安慰剂，该安慰剂含有所有成分，但不含催产素。”
2. https://pubmed.ncbi.nlm.nih.gov/31573637/（Fowler等，2019）

   (a) 选择仅找到一处提及安慰剂的内容：“……在本试验考察的46项次要结局中，43项显示维生素C组与安慰剂组之间无显著差异；然而，与安慰剂相比，维生素C与28天全因死亡率显著降低，以及截至第28天无重症监护病房日和截至第60天无住院日显著增加相关……” (b) 论文在其他地方描述了安慰剂：“患者被随机分配接受维生素C静脉输注（50 mg/kg，溶于5%葡萄糖水溶液，n = 84）或安慰剂（仅含5%葡萄糖水溶液，n = 83），每6小时一次，持续96小时。”

9.5.4 最终任务分解

下面是安慰剂分解运行过程的简化示例。每个步骤都展示了其返回的答案。每个子步骤都用于帮助其父步骤返回答案。斜体 表示算法，等宽字体表示简化的语言模型提示。请访问 ought.org/placebo-trace，详细了解该分解。

分类

1. **根据试验组，是否使用了安慰剂？** 不明确

   (a) **试验组有哪些？**：阿奇霉素、安慰剂 i. **排列段落**：第12段、第1段，……根据相关性排列段落对，直到知道前k个段落是什么A. **排列段落对**：段落 12 >第7段第7段和第12段中，哪一段更能回答“实验的试验组（参与者亚组）有哪些？”这个问题？

> 第7段： ``沙眼防控项目已分发超过6亿剂口服阿奇霉素，旨在消除导致该疾病的眼部衣原体菌株。1,2 ...
>
> 第12段： ``在这项整群随机试验中，我们将马拉维、尼日尔和坦桑尼亚的社区随机分配到四次每年两次的大规模发放中，接受口服阿奇霉素（约每公斤体重20毫克）或安慰剂....."
>
> A：第12段

ii. **答案**：阿奇霉素、安慰剂

> 根据以下段落，回答问题："这项实验的试验组（参与者亚组）有哪些？"
>
> 段落：
>
> 在这项整群随机试验中，我们将马拉维、尼日尔和坦桑尼亚的社区随机分配至两种方案之一，每年接受两次、共四次大规模发放：口服阿奇霉素（约每公斤体重 20 毫克）或安慰剂.....
>
> A：阿奇霉素、安慰剂

(b) 描述试验组：被分配接受口服抗生素阿奇霉素的社区；被分配接受阿奇霉素混悬液载体的社区（装在相同的瓶子中）根据我们所描述的试验组确定相关性，并排序作答

(c) **是否有试验组看起来像安慰剂？**：是

> 判断这篇论文是否使用了安慰剂。请谨慎作答：如果不确定，请回答"无法确定"。
>
> 试验组 1：阿奇霉素 试验组 1 描述：被分配接受口服抗生素阿奇霉素的社区 试验组 2：安慰剂 试验组 2 描述：被分配接受阿奇霉素混悬液载体的社区，药液装在相同的瓶子中
>
> 这篇论文使用了安慰剂吗？请先给出你的推理，然后回答。
>
> 答：是

(d) **参与者能分辨自己属于哪个试验组吗？**：是

> ...后续问题：参与者能分辨自己属于哪个试验组吗？
>
> 答：是

(e) **是否使用了安慰剂？（根据试验组判断）**：不明确

请从"根据试验组天真地判断安慰剂"中给出答案，除非参与者能分辨自己属于哪个试验组；如果能，则回答"无法确定"

2. **根据这些段落，是否使用了安慰剂？**是

(a) **根据各段落，是否使用了安慰剂？**是、无法确定.....

i. **根据本段，是否使用了安慰剂？**是

> 论文中的段落：安慰剂含有口服阿奇霉素混悬液的载体，装瓶并贴上了与阿奇霉素相同的标签。
>
> 根据该段落，论文是否使用了安慰剂？请逐步说明你的推理，然后作答。
>
> 答：是

(b) **是否使用了安慰剂？汇总段落中的答案？**是

如果任何段落表明没有安慰剂，则返回*“No”*。然后，如果任何段落表明使用了安慰剂，则返回*“Yes”*。否则，返回*“No”*。

3. **结合试验组和段落中的回答，是否使用了安慰剂？**是简而言之

：如果试验组或段落中有一方回答*“是”*，且另一方没有提出相反意见，则回答*“Yes”*；否则回答*“N o″*

描述

对于描述，我们再次使用了排序并回答。在上面的示例中，返回的内容是：*“*安慰剂组接受了口服阿奇霉素混悬液的载体。该混悬液装瓶并贴上了与阿奇霉素相同的标签。*”*

9.5.5 其他基线

将论文内容尽可能塞入提示词

最明　　显的语言模型基线，就是在提示词中尽可能多地塞入论文内容　　pt.

分解提纲：

1. 在提示词中尽可能多地塞入论文内容（text-davinci-002 的提示词和补全内容总计可达 4096 个 token），并询问模型：“这是一项安慰剂对照研究吗？我们一步一步来思考：”。接着：2. 分类：在原始提示词的模型回答后添加一个分类问题：“那么，总结一下，这是一项安慰剂对照研究吗？请回答‘是’、‘否’或‘不确定’。” 3. 安慰剂描述：在分类提示词的模型回答后添加一个描述问题：“明白了！请用一句话描述安慰剂。”

这种方法在分类和描述方面的表现都与 Elicit基线大致相当，但两方面似乎都不如任务分解（见表 2；与分解相比，分类的 p=0.0002，描述的 p=0.06）。有意思的是，这种简单的语言模型方法表现不如关键词匹配算法。这种方法的另一大缺点是，我们不知道答案依据论文的哪一部分，因此无法轻易核查答案是否正确。

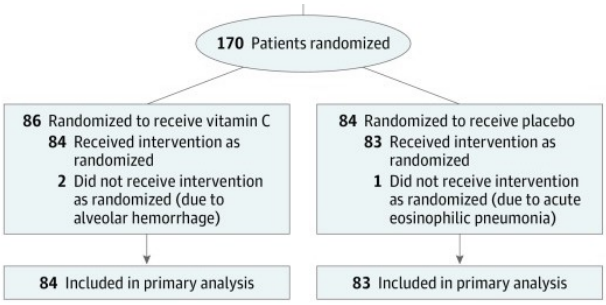
使用 *Elicit* 提示词进行分类：*“*未提及*”*；然后计算困惑度，再使用 *Elicit* 提示词作答

第5.3.2节将更详细地讨论这种方法。我们将其作为 Elicit基线 的一种替代方案，这种方案可能更为强大。在本例中，它的表现与 Elicit基线 相似。

9.6 参与者流程案例研究详情

9.6.1 CONSORT 图示例

CONSORT 图　　记录参与者在研究中的流程。下图展示了以下内容的摘录
Fowler 的图　　等人（2019）。



9.6.2 数据集示例

下表展示了用于评估参与者流程任务依从性的最相关引文示例。

论文	最相关引文	答案
1	“所有招募的青少年都完成了所有3次研究访视，且所有干预组参与者开始了游戏体验。”	干预组的所有参与者都开始了游戏体验。
2	“在丙米嗪试验中，63名随机分组的参与者中有49名完成了6周治疗（对于详情请参见此前的论文14、15）以及63人中的50人随机分组的参与者提供了基线和干预后的经验取样法数据，因此被纳入本次分析。”	两组中63名随机分组的参与者里有49人完成了6周治疗

9.6.3 参与者流程误差分析

任务：	依从性		最佳表现	
方法：	基线		最佳表现	
	n	%	n	%
假阴性	51	80%	18	45%
假阳性	6	9%	12	30%
幻觉	4	6%	0	0%
信息缺失	3	5%	0	0%
包含无关信息	1	2%	6	15%
回答了错误的问题	0	0%	4	10%
总计	64	100%	40	100%

表5：依从性的失败模式计数

失败模式	分解答案	标准答案
假阴性	未提及。	分配的 53 例子宫内膜中有 52 例切除术已实施
假阳性	149 名儿童被纳入 这项研究，并对每名儿童进行了 为期两年的随访。	未提及
幻觉	论文称，受试者已..... 遵从治疗方案。	所有组中，有 1 名男性退出了研究 原因是依从性问题。
信息缺失	论文称，两组均 对所有产品的遵从性良好 每家护理院。	对口服补充剂的遵从性 被评为“良好”，每日摄入量平 均约为 400 千卡
包含无关信息	43 名男性中有 3 人在此期间退出 治疗阶段中途退出，因为 依从性问题（一名受试者）， 常规药物筛查发现使用非法药物 （一名受试者），以及 汽车事故（一名受试者）。	所有组中，共有 1 名男性退出 由于依从性问题。
答错了问题	所有参与者在 实验开始时都保持清醒，并且他们 拿到了饮料，然后他们 观看了一部关于南方的纪录片 非洲半小时，以便让 血液酒精浓度达到峰值。	依从性可能为 因为研究持续了两个小时， 是在跑步后立即进行的， 随机化。 100%， 并且

表6：依从性的失败模式示例

9.6.4 依从性子任务提示词，教科书风格

我们将一些任务设计成摘录自一本（并不存在的）评估科学论文的教科书，并将上下文示例设计成这本教科书中的练习。我们发现，这种做法能减少幻觉（因为练习题的答案不太可能超出所提供文本的内容），还能让我们以自然且有教学性的方式说明任务的一些细微特性，因为教科书中的示例往往会彼此衔接，而且教科书答案不太可能超出“问题”中已有的内容。

摘自教科书《批判性评估干预研究》第3章：

评估随机对照试验的质量时，还应考虑是否有参与者退出研究，或未能正确遵循研究方案。这有时称为“依从性”“脱落”或“遵从性”。如果有太多参与者因任何原因未能接受干预或正确执行干预，研究结果的内部效度就可能受到损害。

遗憾的是，论文在讨论依从性时往往没有应有的清晰度。对于一次性完成的简单干预（例如，让一组大学生在实验室完成一项需要 30 分钟的测试），除非发生了异常情况，否则研究不会讨论依从性；我们可以放心地假定样本中的每个人都完成了研究。有时，研究会提供退出（脱落）人数或比例的具体数据，有时则只提供定性描述，例如称依从性“总体良好”。论文往往确实含糊不清，我们只能得出结论：论文提供的信息不足以让我们了解依从性或遵从性的情况。

让我们看看五篇不同论文的摘录，了解它们是否提供了有关研究依从性、脱落或遵从性的信息。我们需要判断每段摘录告诉了我们哪些有关依从性的信息（有些摘录可能只讨论方法学或结果，并未提供任何依从性信息）；对于其中一些，我们可能只能得出结论：脱落或遵从性的情况并不明确。

首先，请看这篇研究 Tumaini 游戏的论文中的以下三段摘录：

1. 干预组参与者完成了一次为时 45 分钟的信息介绍入门培训，其中包括有关界面、技术和游戏内容的指令。他们被要求在这项研究的 16 天内每天至少玩 1 小时，并且不要与他人分享自己的游戏档案。游戏界面允许额外添加 5 名玩家的档案，这样其他人就可以进行游戏，而不会影响已入组玩家的数据。研究人员向干预组参与者提供了一部预装游戏的手机，供他们在整个干预期内按照自己的节奏使用。对照组参与者接受了标准治疗，即除了来自家人、学校和同伴的现有性教育外，不接受任何额外干预。研究人员没有收集参与者接受的性教育内容或来源的具体数据。所有研究用智能手机均由参与者在干预期结束时归还。

2. 调查数据的初步清理在 MS Excel 中进行，之后使用 SAS 9.4 版（SAS Institute Inc., 美国北卡罗来纳州卡里）完成进一步清理和所有分析。所有对照组参与者均纳入分析。由于一名干预组参与者延迟完成 T2 调查，因此在 T2 效果分析中将其排除。由于他按时完成了其他所有研究活动，其数据仍保留在 T1-T3 分析中。研究人员计算了人口统计学问题和游戏反馈问题的描述性统计数据。

3. 我们招募并纳入了60名青少年参与者。一半参与者被分配到干预组。所有招募的青少年均完成了全部3次研究访视，且所有干预组参与者都开始了游戏体验。参与者人口统计特征见表3。两个试验组之间的人口统计特征没有显著差异。初步暴露量计算结果显示，在为期16天的干预期间，干预组平均玩Tumaini约27小时。

我们来看看每段摘录分别说明了什么（如果有的话），以及它们是否涉及依从性、脱落或遵从性：第一段摘录描述了研究的方法学，但没有说明有多少参与者遵循了指令，也没有说明他们遵循得如何，因此不能据此了解依从性。第二段摘录说明，所有对照组参与者都被纳入分析，但一名干预组参与者未被纳入T2时点的效果分析，却被纳入T3时点的分析；这是有关脱落的信息。第三段摘录说明，所有参与者都完成了全部访视，且所有干预组参与者都开始了游戏体验；这是有关依从性的信息。

以下是这篇论文中有关依从性、脱落和遵从性的全部信息：所有参与者均完成了所有访视，所有干预组参与者均开始了游戏体验。一名干预组参与者未纳入 T2 分析，但纳入了 T3 分析。

其次，考虑　　以下是研究学前儿童情境自我调节工具包（PRSIS）的论文中的三个摘录　　T）程序：

1. 这些中心中所有处于入学前最后一年、且至少参加过1至2个评估日中的一天的儿童，均获邀参与本研究。没有其他排除标准。547名3至5岁儿童已获得参与研究的家长同意，他们均被认为很可能会在次年入学。参与者在整个研究过程中的流向见图1。在基线阶段，对其中473名儿童进行了评估（86.5%）；未参与者主要是评估当天缺席。该样本的平均年龄为4.44岁（标准差 = 0.38，范围 = 3.20至5.33岁），男孩和女孩的比例大致均衡（女孩占48.2%）。被认定具有原住民或托雷斯海峡岛民血统的儿童占样本的7.2%，这与该年龄组的人口估计相符（澳大利亚卫生与福利研究所（AIHW），2012）。家庭收入多种多样：11.9%的家庭符合领取全额儿童保育补贴的条件（低收入）；65.5%的家庭符合领取部分儿童保育补贴的条件（中低至中高收入）；22.7%的家庭不符合领取任何儿童保育补贴的条件（高收入）。母亲教育水平也各不相同：9.5%未完成高中学业；9.3%仅完成高中学业；30.6%取得文凭、职业资格或证书；34.6%完成高等教育学位；16.0%取得研究生学历。随访时，对426名儿童进行了评估，留存率为90.1%。随访时未参与的原因是儿童已离开该中心或评估当天缺席。

2. 根据这些参与模式，20 家服务机构（80%）被认为达到了或超过了最低参与门槛（即完成专业发展模块，并且每周至少开展三项儿童活动）。未参与项目的原因包括：准备政府评估与评级（n = 1）；关键工作人员患重病、休产假或离职，导致无法参与（n = 2）；或因未说明的原因很少参与或完全没有参与（n = 2）。这五家中心中，有两家没有参与任何项目要素。另有三家中心没有参加专业发展模块或入职介绍电话会议，但完成了儿童活动。总体而言，项目的依从性良好，尤其是在没有受到行业施加的重大参与障碍影响的中心中。

3. 然而，无法明确且排他性地为其中一种可能性提供证据，凸显了当前研究的局限性。也就是说，尽管本次评估严格按照 CONSORT 指南进行设计和实施，但受资金因素限制，项目推广和干预期仅持续了 6 个月。完整实施一整年的项目可能会产生更显著的项目效果（例如，参见 Schachter，2015）。严格遵循高质量的项目实施标准，也可能增强项目效果。虽然忠实度数据表明，项目要素在实施频率和时间安排方面遵从性良好，但现有数据不足以评估项目要素实施的完整性。本次研究无法进行面对面或视频忠实度检查，而此类检查有助于监测依从性。由研究者实施的交付模式会违背我们为教育工作者提供低成本、无障碍资源的初衷，因此，一个可行的折中方案或许是采用指导模式，支持教育工作者根据其上下文实施和调整项目。最后，该项目旨在促进所有儿童的自我调节，因此并未关注调节失常的情况。然而，显然，儿童的调节失常仍是教育工作者面临的一项重大挑战（Neilsen-Hewett 等人，2019）；未来的项目迭代应更明确地为这些儿童提供支持。有证据表明，在指导此类项目扩展时，应采用不同的方法来促进经常出现严重调节失常的儿童的自我调节，创伤知情实践方法已成功体现了这一点（Holmes 等人，2015）。未来研究还应考虑教育工作者资质和经验差异所带来的影响，因为根据行为挑战的程度以及教育工作者应对这些挑战的技能水平，可能需要不同类型和程度的支持。

我们来看看每段摘录是否以及透露了哪些关于依从性、脱落或遵从性的信息：第一段摘录包含参与者的人口统计信息，同时也显示，基线时，547名儿童的总样本中有473人接受了评估（未参与的原因主要是缺席）；随访时，有426名儿童接受了评估（未参与的原因主要是儿童已离开中心或缺席），对应的留存率为90.1%。第二段摘录描述了对方案的遵从情况：25个干预中心中有20个达到或超过了最低参与门槛。第三段摘录称项目要素的开展频率和时间安排“良好”，但也指出，研究没有通过现场或视频检查来监测依从性；这类检查本可以更好地反映研究设计的遵从情况。

以下是这篇论文中有关依从性、脱落和遵从性的全部信息：最初的547名儿童样本中，473人在基线时接受了评估，426人在随访时接受了评估。虽然25个干预中心中有20个达到或超过最低参与门槛，且项目要素的开展频率和时间安排良好，但研究没有通过现场或视频检查来监测依从性。

第三，请看一篇研究抑郁症和精神病的研究2论文中的以下四段摘录：

1. 这项干预为单次会议，持续约一小时，参与者在会上提供知情同意
，
填写人口统计学表，观看与其研究组相关的视频，完成评估，并接受事后说明。两个视频组的参与者都停留了整整一小时，而对照条件下未观看视频的参与者约在50分钟内完成。在研究2中，两种条件下均包含两个时长为8分钟、涉及诊断准确性的视频，因此研究流程额外延长了15分钟。调查数据通过SurveyCTO（Ver 2.30，Dobility, Inc.，美国马萨诸塞州剑桥）安卓应用程序，在平板电脑上收集（www.surveycto.com/accessed on: 2017年6月19日）。在研究1中，视频环节结束后，研究人员邀请参与者参加可选的定性访谈，访谈将在一周内进行。

2. 查看二年级和三年级医学学士学生名册后，在随机分组前排除了18名学生，原因是他们是不会说尼泊尔语的国际学生，或已完成精神科轮转。在其余学生中，选取100人随机分配至三个试验组之一。没有潜在参与者拒绝参加本研究。分析时另有六名学生被排除，因为其人口统计表信息显示，他们是母语非尼泊尔语的国际学生，或已完成临床精神科轮转；随机分组时，班级名册中的这些信息尚未更新（图1）。服务使用者组中有一名参与者因既是非尼泊尔语国际学生，又已完成精神科轮转而被排除。这些参与者的人口统计特征见表 2。值得注意的是，只有三名参与者表示他们主要对精神科这一专科感兴趣（见图2）。参与者被随机分配至三种条件之一：无视频的对照组（n = 31，33%）、教学视频组（n = 31，33%）和服务使用者康复见证视频组（n = 32 34 ；%）。

3. 由于研究人员和学生的时间有限，加之访谈具有探索性质，最终仅有六名参与者完成了访谈。定性结果基于六名学生的访谈数据进行分析，他们均为三年级学生，其中两名女性、四名男性，并参与了深入访谈。

4. 在第二项研究中，参与的两所机构共有248名学生就读医学学士课程一年级和二年级。根据名册，28名学生因属于国际学生或已完成精神科临床轮转而被排除。其余220名学生被随机分配；然而，有七名学生拒绝参与，或在数据收集期间无法联系。因此，213名参与者被随机分配至以下试验组：讲授式视频条件（n = 73）、服务使用者视频条件（n = 72）和无视频对照条件（n = 75）（图 3）。在分析阶段，另有一些参与者因缺失数据，或因发现其符合排除标准但学校登记册中未记录，而被排除。各条件下的参与者特征见表4。

让我们思考一下每段摘录（如果有的话）分别告诉了我们哪些关于依从性、脱落或遵从性的信息。第一段摘录描述了方法学，将干预描述为在一次一小时的单次会话中进行。这并没有明确告诉我们任何关于依从性的信息，但确实提高了依从性较高的可能性，因为参与者只需参加一次会话，而这很容易做到。第二段摘录称，在随机分组之前有18名学生被排除；由于这发生在抽样之前，因此与依从性无关。该段还称，在分析时有6名学生因后来发现符合排除标准而被排除。尽管严格来说这不属于依从性问题，但在考虑结果的有效性时，注意到这一点很重要。第三段摘录称，只有6名参与者完成了访谈。第四段摘录称，在研究2中，220名学生完成随机分组后，有7名学生拒绝参与，或在数据收集期间无法联系到；此外，在分析阶段还因缺失数据或发现符合排除标准而排除了其他参与者。

以下是这篇论文中关于依从性、脱落和遵从性的全部信息：论文没有明确讨论依从性。在视频研究中，100 名随机分配的学生里有 6 人被排除在分析之外；在第二项研究中，220 名随机分配的学生里有 7 人拒绝参与或在数据收集期间无法联系，此外，还有学生因缺失数据或符合排除标准而被排除在分析之外。只有 6 名参与者完成了访谈。

第四，来看这篇研究含柠檬马鞭草提取物和 ω-3 脂肪酸的抗氧化/抗炎补充剂的论文中的以下三段摘录：

1. 显示不同时间点脱落率的流程图 ng 不同时间点的脱落率 in the 研究。

2. 共招募 48 名参与者进行筛查评估（图 1），排除 3 人后，将 45 名参与者随机分配至安慰剂组或营养补充剂组，分别为 n = 22 和 n = 23。其中，14 名参与者因不同原因在研究期间退出；安慰剂组有 10 人脱落，补充剂组有 4 人脱落（拒绝治疗、治疗不规律、开始服药或发生不良事件 [AE]）。最终，31 名参与者完成了研究（安慰剂组 12 人，补充剂组 19 人；图 1）。

3. 仅有 1 名患者 报告了不良事件（即烧心感）。这名受试者属于安慰剂组，停止 the
治疗立即 停止治疗，并被排除在研究之外（表 1）。该受试者未报告任何重大并发症 ect.

让我们看看每段摘录能告诉我们什么（如果有的话），以及它们与依从性、脱落或遵从性有何关系：第一段摘录提到了一张显示脱落率的流程图，但由于这里没有图，我们无法据此对研究的脱落情况得出任何结论。第二段摘录称，安慰剂组的 22 名参与者中有 10 人脱落，补充剂组的 23 名参与者中有 4 人脱落，这意味着随机分组后最初的 45 名参与者中有 31 人完成了研究。第三段摘录进一步介绍了安慰剂组中一名脱落患者的情况：该患者出现烧心感后停止了治疗。

这篇论文中关于依从性、脱落和遵从性的所有信息如下：安慰剂组的 22 名参与者中有 10 人脱落，补充剂组的 23 名参与者中有 4 人脱落。

第五，考虑这篇论文中研究{{段落_n}}摘录{{"s" 如果 len(段落) > 1 否则 ""}}的干预{{干预}}：

{{段落_至_编号_列表(段落).strip()}}

让我们想 想想每个{{"each" 如果 len(段落) > 1 否则 "this"}} 摘录能告诉我们什么（如果有的话），关于依从性 ，
脱落或合 依从性：

9.6.5 依从性子任务提示词，StatsExchange 风格

我们尝试在数据可用时计算最终依从率。一种可能的计算方法是生成 Python 代码，再交由 Python 解释器执行计算。我们还发现，在零样本设置下，模拟一篇 StatsExchange 帖子可以增强上下文中的数学推理能力。

> Statistics Stack Exchange 上的一位用户需要帮助解决以下问题：
>
> 请根据学术论文的摘录，确定干预中有多少参与者完成了干预，或他们完成干预的比例是多少。忽略无关摘录。如果信息不足以确定答案，请回答“无法确定”。
>
> 该干预是卵胞浆内单精子注射（ICSI）。
>
> 你必须作出判断　哪些摘录与你的推理相关。有些可能非常相关，而大多数都无关　t.
>
> 摘录：因此，约90%的夫妇接受了胚胎移植；使用射出精子、附睾精子和睾丸精子时，临床妊娠率分别为21%、22%和19%（多胎妊娠率为25%–30%）……数据表已发送给所有曾参与此前ICSI工作组调查的中心，以及文献或国家登记处提及的新ICSI中心……截至1995年12月31日，共有101个中心提交了临床结果，所依据的周期数为2至2507个（见表I）……因此，在提交了1995年ICSI结果的101个中心中，只有17个对儿童进行了前瞻性随访，其中仅有9个将其作为专项研究的一部分；另有46个中心则尝试通过联系不孕症专科医生、儿科医生或护士来收集信息。
>
> 请在给出答案前详细说明你的推理。思考有多少人参与了卵胞浆内单精子注射（ICSI）。请逐步说明如何根据相关摘录推导出问题“依从率是多少？”的答案。给出你的最终答案。
>
> Statistics Stack Exchange 上的最佳回答如下：
>
> 引文“因此，在提交了 1995 年 ICSI 结果的 101 个中心中，只有 17 个中心对儿童进行了前瞻性随访，且其中只有 9 个将其作为专项研究的一部分；另有 46 个中心正尝试通过联系不孕症专科医生、儿科医生或护士来收集信息。”谈的是开展随访的中心数量，却没有提及完成、依从或遵从干预的患者人数。因此无法确定。
>
> 最终答案：不明确
>
> Statistics Stack Exchange 上的一位用户需要解决以下问题：
>
> 请根据学术论文的摘录，找出有多少参与者完成了干预，或他们完成干预的比率。忽略不相关的摘录。如果信息不足以确定答案，请回答“不明确”。
>
> 这项干预　是疟疾化学预防（季节性疟疾化学预防，SMC）和长效青蒿素联合疗法（青蒿素联合疗法，ACTs）。
>
> 你必须判断哪些摘录与你的推理相关。有些可能非常相关，而大多数可能无关。
>
> 摘录：一项研究每组纳入约800名儿童，并考虑10%的失访率，其检验效能为90%，可检测出接受长效青蒿素联合疗法（ACT）与短效ACT的儿童之间疟疾发病率降低20%的差异……一项每组纳入800名儿童的研究，检验效能超过90%，可检测出接受长效ACT与接受季节性疟疾化学预防（SMC）的儿童之间25%的差异；对于检测SMC组与短效ACT组之间更大的差异，检验效能则非常高……每次实施SMC后4至7天，社区卫生工作者（CHWs）会对照护者进行随访，使用标准表格询问不良事件以及对季节性疟疾化学预防方案的依从性……然而，实际上依从性不太可能如此之高，因为在本研究的定性研究部分中，研究人员发现有些照护者仍有未给儿童服用的平板电脑……研究收集了依从性数据，但报告的依从性数值高得不合常理，其他研究中也有类似情况 [23]。
>
> 回答前，请详细描述你的推理过程。思考有多少人参与了疟疾化学预防（季节性疟疾化学预防）和长效青蒿素联合疗法（青蒿素联合疗法）。请逐步说明如何根据相关摘录推导出问题“依从率是多少？”的答案。给出你的最终答案。
>
> Statistics Stack Exchange 上的最佳回答如下：
>
> 有一段引文写道：“然而，依从性实际上不太可能有这么高，因为在本研究的定性研究部分中，发现一些照护者还剩有尚未给患者服用的平板电脑。”这说明随访系统并非完全准确，而且有些人没有遵循季节性疟疾化学预防方案。不过，我们没有得到任何数字或百分比，因此无法定量说明依从率是多少。但结合“收集了依从性数据，但报告的依从性数值高得不合情理，这一点在其他研究中也有所体现”这段话，可以推断作者发现依从性很高，但认为这一结果不合情理。
>
> 最终答案：依从性被表述为很高。但作者认为这不太可信。
>
> Statistics Stack Exchange 上的一位用户需要帮助解决这个问题：
>
> 使用学术论文中的摘录，确定干预中有多少名参与者完成了干预，或他们完成干预的比例。忽略不相关的摘录。如果信息不足以确定答案，请回答“不清楚”。

该干预为脉冲电磁场（PEMF）疗法。

你必须判断　　哪些摘录与你的推理相关。有些可能高度相关，而大多数则无关　　t.

摘录：共有286名受试者表示有意参与这项研究，研究人员向他们寄送了一份健康问卷，并附上预付邮资的回邮信封，以评估他们初步是否符合参与研究的资格……23名受访者拒绝参加试验，263名受访者寄回了问卷，其中133人因不符合排除标准而被排除在研究之外……14名患者在临床检查中未达到资格标准，另有8名参与者退出研究，因此最终有108名参与者纳入研究样本……效能计算表明，假设脱落率约为10%，样本量为110名患者（每种治疗55名），将有80%（b 1⁄4 0.20）的概率检测到主动治疗与假治疗之间改善幅度存在40%（a 1⁄4 0.05）的差异……在开始治疗的108名患者中，有9人在研究期间退出……4名患者在主动治疗期间退出……5名患者在假治疗期间退出……退出原因包括整体疼痛加重和肠易激综合征恶化（n 1⁄4 1）、腰椎间盘突出（n 1⁄4 1）、颈椎间盘突出手术（n 1⁄4 1）、慢性肺炎（n1⁄4 1）以及不愿继续参与（n 1⁄41；图1）……入组患者中，92%完成了研究……以是否参加全部168次计划中的8分钟治疗疗程来衡量，主动设备组的平均治疗遵从性为134 AE 41（80%），假设备组为131 AE 44（78%）……治疗依从性较高（$80%）……第四，考虑到较高的脱落率是交叉设计中的一个主要问题 [Mills 等人， 2009]，本研究尽管持续时间相对较长（28周），脱落率仍然较低（8%）。

在给出答案之前，请详细描述你的推理过程。想一想有多少人参与了脉冲电磁场（PEMF）治疗。请逐步说明如何根据相关摘录推导出问题“依从率是多少？”的答案。最后给出你的答案。

Statistics Stack Exchange 上的最佳回答如下：

在这项研究的上下文中，依从性可能指完成率，也可能指遵从率。我们将分别讨论这两种情况。

关于完成率，我们知道，108 名患者中有 9 人在研究期间退出。因此，108-9 = 99人完成了研究。由此，完成率为 99/108 = 91.7%。

关于遵从率，文中指出：“平均治疗依从性以是否参加全部 168 次计划中的 8 分钟治疗疗程来衡量，使用主动设备时为 134 AE 41（80%），使用假设备时为 131 AE 44（78%）。”文中还指出：“治疗依从性很高（$80%）。”因此，主动设备的遵从率为 80%，假设备的遵从率为 78%。

最终答　　案：完成率为 91.7%。主动设备的依从率为 80%，而共享设备的依从率为 78%。　　m
设备。

Statistics Stack Exchange 上的一位用户需要帮助解决这个问题：

请根据学术论文中的摘录，确定干预中有多少参与者完成了干预，或他们完成干预的比率是多少。忽略无关摘录。如果信息不足以确定答案，请回答“无法确定”。

这项干预是*{{*干预*}}*。

你必须判断哪些摘录与你的推理相关。其中一些可能非常相关，而大多数都无关。

摘录：　*{{*"...".join(excerpts)*}}*

回答前，请详细描述你的推理过程。想一想有多少人参与了{intervention}。请逐步说明如何根据相关摘录推导出问题“依从率是多少？”的答案。给出你的最终答案。

Statistics Stack Exchange 上的最佳回答如下：

9.6.6 去语境化少样本提示词

指令：结合上下文丰富每个段落。

上下文：丽莎喜欢恶作剧。

段落：但有时她做得太过分了。

改写：但有时她 [丽莎]做得太过分了。

上下文：第41届超级碗中场秀于2007年2月4日举行。

段落：由 Prince 担纲演出。

改写：由 Prince 担纲演出 [第41届超级碗中场秀]。

上下文：全球五分之一以上的人口每天的生活费不足购买力平价（PPP）1.25美元，国际社会逐渐形成共识：到2030年，应当（也能够）将这一比例降至接近零（1, 2）。

段落：实现 这一目标需要赋能最贫困的家庭；这些家庭往往在自己的村庄中处于最边缘的位置
使他们能够 从不安全且脆弱的收入来源转向更可持续的创收活动。

改写：要实现这一目标 [，即将全球人口中每日生活费低于购买力平价（PPP）1.25美元的人口比例从五分之一以上降至 2030]零，就必须帮助最贫困家庭——他们往往是所在村庄中最受边缘化的群体——从不安全且脆弱的收入来源转向更可持续的创收活动。

上下文：我们介绍了六个国家中旨在促进极端贫困人口开展自营就业活动的一种特定方法的随机对照试验（RCT）结果。该方法最初由BRAC设计并实施。BRAC是一家大型孟加拉国非政府组织，在全国多个地区运行多项程序。 `“脱贫计划”通过提供一整套综合服务，包括授予生产性资产，帮助村庄中最贫困的家庭（BRAC称之为 `“极端贫困人口”）。脱贫计划的受益者 [——村庄中最贫困的家庭，即“极端贫困人口”] ——通过村民会议中的参与式流程确定，随后由该组织的 [“脱贫计划”实施者] 工作人员进行核查访问。之后，选定的受益者 [——村庄中最贫困的家庭，即“极端贫困人口”]——会获得一项生产性资产 [，由“脱贫计划”]的实施者提供；他们可以从清单中选择该资产，并获得针对所选资产的训练和支持、一般生活技能辅导、在固定期限内提供的每周消费支持，以及通常可获得的储蓄账户和健康信息或服务。

段落：这些 种不同的活动（以及一年中与各家户的定期互动）旨在
相互补充, 帮助各家户开展生产性自营就业活动。

改写：这些不同的活动 [为其所选并获得的资产提供培训和支持、一般生活技能辅导、每周消费补助，通常还提供使用储蓄账户以及获取健康信息或服务机构服务的机会] （以及一年中与各家户的定期互动），旨在相互补充，帮助各家户 [从村庄中最贫困的家庭里选出的“毕业”项目受益者，或“极端贫困人口”]开展生产性自营就业活动。

上下文： *{{上下文}}*段
落： *{{passage}}*改写：