验证器-与编辑：一种知识增强型思维链框架

Ruochen Zhao [1*] **Xingxuan Li** [1,2*†] **Shafiq Joty** [1,3‡] **Chengwei Qin** [1] **Lidong Bing** [21] 新加坡南洋理工大学[2] 阿里巴巴集团达摩院[3] Salesforce AI {ruochen002, chengwei003}@e.ntu.edu.sg {xingxuan.li, l.bing}@alibaba-inc.com srjoty@ntu.edu.sg

arXiv:2305.03268v1 [cs.CL] 5 May 2023

摘要

随着大语言模型（LLM）成为自然语言处理领域的常态，并在生成和推理任务中表现出色，其最致命的缺陷之一是缺乏事实准确性。生成不符合事实的文本不仅会导致性能下降，还会削弱其应用的可信度和有效性。思维链（CoT）提示通过生成可解释的推理链，提升了复杂推理任务的可信度和模型表现，但在知识密集型任务中仍存在事实性问题。本文提出用于思维链提示的 Verify-and-Edit 框架，通过依据外部知识对推理链进行后编辑，以提高预测结果的事实性。基于 GPT-3 构建的框架，在多个开放域问答任务上提升了准确率。为便于复现我们的结果并进一步扩展该框架，我们在此公开代码库：
h
t
tps://github.co
m/RuochenZhao/Verify-and-Edit

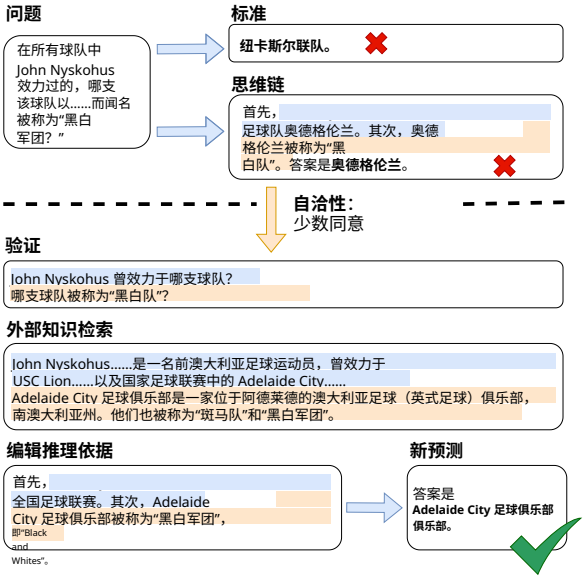

图 1：验证与编辑框架包含五个步骤：(1) 将一致性低于平均水平的预测结果传递到后续阶段，同时保持高度一致的预测结果不变；(2) 生成验证问题；(3) 检索外部知识；(4) 根据已有信息的答案编辑推理过程；(5) 生成新的预测结果。

1 引言

大语言模型（LLM）已成为许多下游自然语言处理任务中的新常态。在使用这些大语言模型时，人们发现，思维链（CoT）提示（Wei 等人，2022年）能够提升数学文字题、常识推理和符号操作等需要复杂推理的任务表现。同时，它还能生成可解释的推理链。近期研究进一步探索了如何利用这些推理链来选择更好的预测结果。然而，这些方法的主要关注点在于通过直接使用生成的思维链来提升终端任务的性能。例如，Ye 和 Durrett (2022年) 训练了一个校准器，根据理由分数调整预测概率；Wang 等人 (2022年) 采样多个推理路径，找出最常见（最一致）的预测。只有少数研究，例如 Creswell 等人(2022年) 和 Zhou 等人 (2022年)，探索了提升思维链本身质量的方法。

事实上，提高思维链的质量有助于增强可解释性和下游任务性能。YeandDurrett (2022年) 指出，人类评判为良好的解释通常意味着预测结果更加准确。直观而言，一组更好的思维链提示能够提供更充分的依据和逻辑上更加一致的思维过-

*同等贡献。
†Xingxuan Li 正在参加阿里巴巴与南洋理工大学联合博士培养项目。
‡作者在南洋理工大学（NTU）休假期间完成的工作。

c程，从而带来更准确的预测 s.

为了提升生成质量，一个重要方面是事实正确性，这也是目前大语言模型最致命的缺陷之一（OpenAI 博客，2022年；Zhao 等人，2023）。在回答用户查询时，GPT-3 等大语言模型（Brown 等人，2020年）往往会编造事实和细节，这如今已被列为使用其应用程序接口时的一项主要警示。大语言模型的一项主要应用场景，是有望取代传统搜索引擎，并通过问答实现更直接的信息获取；因此，事实性方面的担忧可能会严重削弱其有效性，并降低用户的信任度（Marcus，2022年）。解决这一问题颇具挑战，即使模型经过人类反馈指令微调后，这些担忧依然存在（Ouyang 等人，2022年）。这是因为微调过程中可能无法获取真实信息来源（OpenAI 博客，2022年）。

因此，如何更好地控制生成过程并提高预测结果的事实正确性，已成为亟待解决的问题。由于大语言模型在充当知识库时可能无法准确回忆细节（Ye 和 Durrett，2022年；Creswell 等人，2022年），如果条件允许，可以引入外部来源中的知识作为辅助。辅助式思考过程在人类推理中也很常见：人们回答问题时，往往会搜索（或重新查阅）外部知识来源，以获取支持性事实，从而唤起（内部）记忆。

受此启发，在这项工作中，我们提出一种**验证与编辑**（VE）框架，对推理链进行后编辑，从而获得事实依据更充分的预测。如图1所示，我们首先选出一致性未达到多数认可标准的不确定实例进行编辑。正如Wang 等人 (2022年) 所指出的，这些实例通常由听起来合理的陈述构成，例如图1中的句子“John Nyskohus 曾效力于挪威足球队 Odd Greenland”。编辑时，我们首先生成一个问题来验证这一细节，例如“John Nyskohus 曾效力于哪支球队？”接着，为了回答这一问题，我们通过开放领域检索系统引入外部知识。例如，在此实例中检索到了事实“John Nyskohus……曾效力于 Adelaide City……”。随后，我们将检索到的事实作为记忆补充提供在提示中，以此编辑推理依据。因此，编辑后的推理依据可以根据检索到的事实进行更新（图1）。基于编辑后的推理依据，生成新的预测，该预测考虑事实性更强的推理轨迹。

据我们所知，我们的工作是首个通过后编辑思维链式推理链来提升预测性能的研究。我们在两个需要推理的开放领域问答（QA）任务上进行了实验：对抗性 HotpotQA（Yang 等人，2018年）和 2WikiMultihop（Ho 等人，2020年）。我们还使用 Fever（Thorne 等人，2018年）测试了该方法在事实核验任务上的性能。我们发现，模型能够从更多事实性推理链中获益，从而生成更加准确的预测。例如，在开放领域问答任务中，与类似的检索增强模型相比，我们的模型在 AdvHotpot 上实现了 3.8 倍的准确率提升。在 2WikiMultihop 上，Verify-and-Edit 在开放域搜索条件下达到了 33.6 % 的准确率，而思维链自一致性的准确率为 27.7%。

2 相关工作

思维链或 CoT（Wei 等人，2022年）是一种用于提升大语言模型推理能力的提示方法，使大语言模型能够将复杂问题分解为多个中间步骤。思维链具有可解释性，并且已被证明比标准提示方法更擅长解决复杂问题。

然而，幻觉是自然语言处理中的一个长期存在的问题，尤其对于大语言模型而言，这一问题已引起研究界的广泛关注。大语言模型的解码过程是自回归的，这不可避免地导致其在缺乏受控生成的情况下输出非事实性内容（Ye 和 Durrett，2022年；Wiegreffe 等人，2022年）。因此，在思维链生成过程中缺乏支持性事实，可能会严重削弱最终答案的有效性（Golovneva 等人，2022年）。Ye 和 Durrett （2022年）证明，最终答案的准确率与推理解释的事实性和一致性高度相关。通常提出的提升思维链推理过程事实性的方法可分为两类：提示工程和结果校准。

提示工程方法通常用于引导大语言模型生成更好的中间推理解释。*ReAct* (Yao 等人，2022年)与我们的工作最为相似，它将大语言模型中的推理与行动结合起来：推理步骤帮助模型推导并更新行动，而行动步骤则允许模型从维基百科获取额外信息，以进行

事实性检查。与*ReAct*相比，我们生成更自然、更具对话性的思维链，从而提升可解释性并使学习更容易。因此，我们的框架只需更短的提示即可完成学习。Press 等人(2022年)提出了自我提问，通过指示大语言模型在回答初始问题之前，明确地向自己提出后续问题并作答。解决复杂问题的一种自然方式，是将问题分解为若干子问题并依次解决。Zhou 等人(2022年)采用了这一思路，并提出了由简至繁提示方法。不过，自我提问和由简至繁提示方法仍依赖反复检索大语言模型习得的内部知识，而非连接外部知识。因此，它们提升事实性的能力有限。

结果校准作用于大语言模型的输出。 Ye 和 Durrett (2022年) 训练了一个校准器，根据生成的解释的事实性和一致性来校准最终答案的权重，从而有效提升结果。思维链中的解码方法是朴素的贪心解码，即简单地输出概率最高的下一个词元。 Wang 等人 (2022年) 提出了一种 自一致性解码方法，它对多样化的推理路径进行采样，然后通过对采样得到的推理路径进行边缘化，选出最一致的答案。 选择-推理（*SI*）(Creswell 等人, 2022年) 框架是另一种利用大语言模型作为通用处理模块的最先进方法。在所有方法中，它也是首个系统性地提升思维链事实正确性、从而实现更准确预测的方法。它交替执行选择和推理，生成一系列可解释的因果推理步骤，最终得出答案，且已证明高效。然而，它并非为开放领域或常识问答而设计。

此外，另一项相关研究方向正在探索检索增强语言模型预训练（REALM）（Guu 等人, 2020年），其首先从外部知识源检索文档，然后利用检索到的文档处理问答任务。Lazaridou 等人(2022年) 提议将问题的 Google 搜索结果纳入提示，以提升生成答案的事实性。然而，这类方法在处理复杂问题时可能会失效，因为它们没有利用大语言模型的推理能力。因此，我们将检索增强推理路径视为一种自然的方式来提高事实一致性 t.

3 验证与编辑框架

我们的目标是借助外部知识，通过思维链提示让大语言模型生成更符合事实的推理链，从而提高最终答案预测的准确率。我们假设，这能够增强大语言模型解决复杂知识密集型任务的能力，此类任务需要经过多个推理步骤才能得出答案。

总体而言，我们希望遵循人类的推理过程：回答问题时，如果一个人不确定，就会查找支持性事实并加以考虑，然后再给出最终答案。因此，我们可以将验证与编辑（验证器）框架分为三个不同阶段：找出不确定的预测结果、通过搜索支持性事实来编辑其推理依据，以及使用编辑后的推理依据生成最终答案（图1）。在设计这些阶段时，我们希望最大程度地保留大语言模型最大的优势：开放式生成和推理能力。我们还致力于尽可能自然、对话式地设计任务和设置，从而让人类以及通过自然文本训练的大语言模型都能轻松理解。

3.1 决定何时编辑

我们如何识别模型何时对自己的预测缺乏把握？自洽性方法（Wang 等人，2022年）提供了一种解决方案。通过对不同的推理路径和答案进行采样，研究发现，自洽性与准确率高度相关，这表明它可以提供不确定性估计，并赋予模型“知道自己何时不知道”的能力。因此，我们通过使用一致性方法对预测任务进行采样，来构建验证器框架 n中的多样化推理路径。对于一致性较高的预测结果，予以保留，不作修改。当一致性低于 $\lceil n/2 \rceil$时，即多数意见无法达成一致、无法给出相同答案时，我们将其标记为“不确定”。

3.2 如何编辑特定的推理过程

推理过程，即思维链（CoT），可以分为两个部分来看：事实，以及结合事实推导出新主张的推理过程。因此，我们从这两个方面考虑如何改进思维链。

- **事实**为了让推理过程更加符合事实，我们会从外部知识来源（例如维基百科、谷歌）中搜索支持性事实。

算法 1验证与编辑

要求： 原始问题 q；一个 n 示例的思维链提示 p_{cot}
要求： 一个大语言模型 $f(\cdot)$；大语言模型补全次数 n；大语言模型解码温度 τ
要求： 一个外部知识检索模型 $g(\cdot)$
要求： 用于验证问题生成（p_{vq}）和答案生成（p_{va}）的 n 示例提示

$R, A \leftarrow f(p_{cot}, q, n, \tau)$ ▷ 生成一组推理（R）和答案（A）。
$s^*_{sc} \leftarrow$ 最大化 $P(a|p_{cot}, q)\ a \in A$ ▷所有答案中的最高自一致性分数。
$r^*, a^* \leftarrow \arg\max\ P(a|p_{cot}, q)\ a \in A$ ▷ 以最高自洽性进行推理并作答。
如果 $\frac{s^*}{sc} < \lceil \frac{n}{2} \rceil$ **然后** ▷ 编辑一致性低于多数同意的推理。
  对于 $o_i \in r^*$ **执行** ▷ 编辑推理中的每个句子。
    $u \leftarrow f(p_{vq}, q, o_i)$ ▷ 生成验证问题。
    $v \leftarrow g(u)$ ▷检索外部知识。
    $w \leftarrow f(p_{va}, u, v)$ ▷ 生成验证答案。
    $o_i \leftarrow w$ ▷使用验证答案编辑原始推理句子。
  结束循环
  $a^* \leftarrow f(p_{cot}, q, r^*)$ ▷ 基于编辑后的推理生成最终答案。
  return a^*
否则如果 $\frac{s^*}{sc}$ <escape>$\geq \lceil \frac{n}{2} \rceil$</escape>**然后** ▷具有高度一致性的答案将保持原样。
  返回 a^*
结束 if

首先，为了模拟人类在搜索验证事实时提出查询的方式，我们会生成一个自然的问题来验证推理过程。为此，我们利用同一个大语言模型的上下文学习能力。我们会在提示中同时提供原始问题和推理过程，以生成用于验证的问题，确保该问题询问的是回答原始问题所需的最相关信息，而不是推理过程中提及的其他实体。例如，如果推理过程（错误地）是“1961 年 8 月 4 日出生的美国总统是约翰·肯尼迪”，而原始问题是“1961 年 8 月 4 日出生的美国总统的配偶是谁”，那么我们希望生成的验证问题是：“1961 年 8 月 4 日出生的美国总统是谁？”，而不是“约翰·肯尼迪的生日是什么时候？”通过生成相关问题，而不是直接根据生成的推理过程进行查询，我们可以消除错误事实生成所带来的潜在噪声。在上述示例中，如果使用错误的主张“1961 年 8 月 4 日出生的美国总统是约翰·肯尼迪”进行检索，那么错误实体“约翰·肯尼迪”可能会干扰搜索过程。

在本文中，我们使用从 3 个系统检索到的相关上下文：（*i*）DrQA（Chen 等人，2017年），一种开放域问答系统；（*ii*）维基百科中相关页面的搜索结果；以及（*iii*）Google 搜索，展示了大语言模型与搜索引擎结合的可能性。

由于从检索系统获取的上下文可能会比预期更长，因此我们使用预训练语言模型对句子进行排序，并选出与验证问题查询最相似的前-k 个句子。

• **推理** 虽然选择-推理等方法（Creswell 等人，2022年）直接将检索到的事实用作推理依据，但这些推理依据通常过于冗长，篇幅超出预期，或包含无关细节。Ye 和 Durrett （2022年）也有类似的观察：直接使用支持性句子通常过于冗长，而且并不充分。

为了获得更相关、合乎逻辑的推理依据，我们再次采用自然的生成式方法，因为人们认为大语言模型已经具备推理能力（Wei 等人，2022年）。具体而言，通过输入“问题、推理过程、答案”格式的提示，大语言模型会在生成答案前进行几步推理。考察原始推理过程后，我们发现，即使其中包含错误事实，其逻辑推理部分总体上似乎仍然完整。因此，我们利用验证问题（作为逻辑）和检索到的事实（作为信息）来生成有依据的答案。随后将这些有依据的答案组合成新的推理过程，从而有望得到事实性更强的思维链。

3.3 再次作答

最后，通过使用经过后期编辑的思维链提示大语言模型，生成新的答案。伪代码

整体流程如算法1所示，并通过图1 中的示例加以说明。我们可以看到，通过允许大语言模型整合外部知识，我们的方法能够生成具有更充分事实依据的推理依据。当将其作为思维链提示输入大语言模型时，它能够引入生成新预测所需的信息，而这些信息原本并未被模型正确记忆。

与 ReAct 等专门设计的提示（Yao 等人，2022年）相比，验证与编辑框架简单且可以说更加自然。其对话式特性能够让人类更好地理解模型的思维过程，并有望让用户在推理的任何阶段自然地进行干预和修订。在接下来的实验中，我们还观察到，这种设置能够有效缓解对事实性的担忧，并提升最终任务的性能。

4 实验设置

4.1 推理任务

由于验证与编辑框架提供了更多以知识为依据的推理步骤，因此它应当有益于满足以下两个属性的任务：（*i*）依赖多跳推理得出后续预测，从而依赖推理过程生成；以及（*ii*）属于开放域任务，因此需要与外部知识源进行交互。

因此，我们在三个数据集上验证该方法：（*i*）**对抗性 HotpotQA**（Yang 等人，2018年），一个多跳问答数据集。我们使用 Ye 和 Durrett（2022年）提出的具有挑战性的子集，其中正确和错误的预测通过其模型实现平衡。（*ii*）**2Wiki- 多跳**（Ho 等人，2020年），一个利用 Wikidata 的结构化格式并使用逻辑规则的多跳问答数据集。[1]（*iii*）**Fever**（Thorne 等人，2018年），一个事实核验数据集，根据来自维基百科的证据段落将主张标记为“SUPPORTS”（支持）、“REFUTES”（反驳）或“NOT ENOUGH INFO”（信息不足）。与 HotpotQA 的设置类似，我们通过平衡 GPT-3 思维链作出正确和错误预测的样本，采样出一个具有挑战性的集合。有关数据集处理和使用的详细信息，请参见附录 A。

4.2 对比方法

为取得最先进的性能估计结果，我们采用 GPT-3 指令系列应用程序接口 `text-davinci-003`（Ouyang 等人，2022年），将实验时最强且最新的模型作为骨干模型。实验成本见附录 B。

对抗性 HotpotQA 和 2WikiMultihop 实验采用 6-shot 上下文学习，Fever 采用 3-shot 上下文学习，因为 Fever 问题更短，也更容易学习。对于 HotpotQA，我们使用Ye 和 Durrett（2022年）提供的人工标注，并以类似格式为 2WikiMultihop 和 Fever 手动标注少样本示例。基线方法和我们方法的完整提示见附录 C。

基线方法 为更全面地概述我们的框架所处的位置，我们采用以下基线方法：

1. **标准预测**（Standard）：在给定数量相同的上下文学习示例的情况下，直接根据输入预测标签。2. **原始思维链**（Wei 等人，2022年）：生成解释后再预测标签。3. **带自一致性的思维链**（CoT-SC）（Wang 等人，2022年）：以 0.7 的解码温度采样 5 条思维链轨迹，该设置由论文推荐。4. **校准器**（Calib.）（Ye 和 Durrett，2022年）：根据预测得分调整预测概率的校准器。5. **ReAct**（Yao 等人，2022年）：利用外部维基百科 API 的推理与行动框架。对于这一基线方法，我们采用原论文报告的结果，该论文使用 PaLM 模型（Chowdhery 等人，2022年），其性能与 GPT-3 相近。[2] 为了提供更有依据的视角，我们报告其在 CoT-SC 基线基础上取得的性能提升。[3]

验证与编辑（VE） 在实现 VE 框架时，我们采用相同的一致性基线来估计模型何时不确定。如 §3.1所述，我们会编辑所有自一致性分数低于 $\lceil n/2 \rceil$的实例，其中 n 是采样路径的数量。然后，使用 2-shot[4] 设置生成验证问题

[1]出于成本考虑，我们从 12,576 个开发集样本中随机抽取了 1,000 个样本。

[2]我们无法使用 PaLM，因为它并未开源。[3]值得注意的是，ReAct 在整个数据集上进行了实验，而我们使用的是采样版本（参见 §4.1）。[4]我们观察到，随着上下文示例的增加，问题生成质量并未发生太大变化，因此我们选择能够生成合理问题的最短提示，以降低成本。

并结合上下文学习。验证答案使用与生成原始答案相同数量的示例，并通过贪心解码生成。

为研究知识检索系统对结果的影响，我们使用四种系统：

1. **Wikipedia-API**（wiki）：搜索查询实体，并从其维基百科页面中选取排名靠前的句子。

2. **DrQA**（Chen 等人，2017年）：一种预训练的开放领域问答模型，结合了双字组哈希、TF-IDF匹配和多层循环神经网络模型。我们仅使用从该模型检索到的上下文。[5]

3. **谷歌**：使用谷歌生成的前k条搜索结果作为辅助上下文。这一结果很有意思，因为它展现了搜索引擎与大语言模型相结合的可能性。

4. **数据集**：从 Adversarial HotpotQA 和 2Wiki-MultihopQA 提供的段落集合中选取内容，其中包括真实的支持上下文和干扰段落。这类似于预言机设置，在假设我们拥有良好检索系统的情况下，可提供性能提升的上限。

对于 1、2 和 4，在检索后，我们选取由预训练的 Sentence BERT 模型（Reimers 和 Gurevych，2019年）按与查询的相似度排序后排名前三的句子作为上下文。

5 结果与分析

5.1 使用自洽性：知道自己何时不知道

在验证与编辑框架的第一步中，使用一致性来衡量模型对预测结果的置信度。与 Wang 等人 (2022年) 的研究发现一致，我们假设，当一致性较低时，模型的不确定性更高，因此更有可能生成不准确的预测结果。为了检验这一假设是否成立，我们绘制了对抗性 HotpotQA 数据集上一致性分布的核密度估计图。如 图 2所示，错误样本的一致性分布呈左偏态，大多数错误预测结果的一致性较低。另一方面，正确预测结果的分布则呈现右偏趋势，其中

[5]我们首先通过使用不同的开放域问答模型进行小规模实验来选择 DrQA，其中包括 DPR（Karpukhin等人，2020年）。研究发现，DrQA能够取得更好的性能。因此，我们始终使用它。

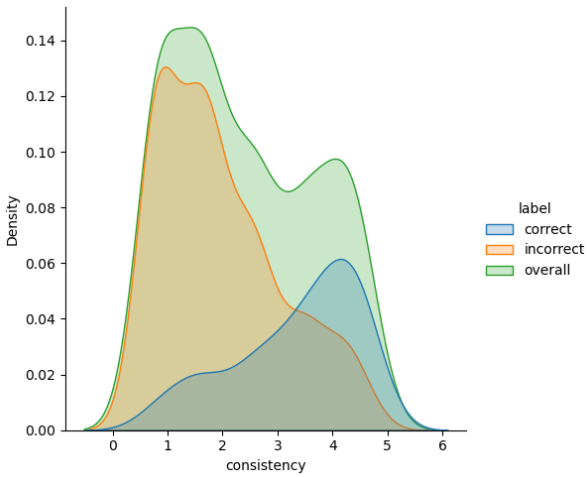

图 2：对对抗性 **HotpotQA** 数据集上的一致性进行核密度估计的图。通过核密度估计，曲线扩展了其真实分布的范围，即从 0 到 5（因为我们采样了 5 条路径）。

方法	知识	EM	ΔEM	AUC
CoT-SC → ReAct	维基百科。	34.2%	+0.8%	-
ReAct → CoT-SC	维基百科	35.1%	+1.7%	-
标准	-	23.1%	-	43.24
CoT	-	31.8%	-	38.30
CoT-SC	-	31.2%	-	34.97
CoT-SC + Calib.	数据集	-	-	49.00
CoT-SC + 验证器	维基百科	35.7%	+4.5%	45.62
CoT-SC + 验证器	DRQA	36.0%	+4.8%	46.06
CoT-SC + 验证器	谷歌	37.7%	+6.5%	47.98
CoT-SC + 验证器	数据集	**56.8%**	**+25.6%**	**60.94**

表 1：对抗性 **HotpotQA** 数据集上的结果。每个模型的最佳结果以带下划线标出，整体最佳结果以粗体标出。 ΔEM精确匹配表示相较于 CoT-SC 基线在精确匹配上的提升。前两行使用 PaLM 模型，其余行使用 GPT-3 davinci-003 模型。

一致性较高的错误样本非常少。这有效地验证了我们的假设。

在主要实验中，我们将 $\lceil n/2 \rceil$ 设为多数阈值，并编辑所有低于该阈值的样本，该阈值为 3。为展示不同阈值对框架性能的影响，我们还将在后文提供一项消融研究。

5.2 HotpotQA 上的结果

如表 1所示，我们观察到思维链在标准少样本设置的基础上取得了提升。另一方面，CoT-SC 并未显著改善基线表现。使用Ye 和 Durrett (2022年提出的校准器后，曲线下面积有所提升，因为它能够根据数据集中提供的真实上下文来校准答案权重。

方法	知识	EM	ΔEM	AUC
标准	-	16.9%	-	35.89
CoT	-	28.4%	-	16.64
CoT-SC	-	27.7%	-	17.16
CoT-SC +Calib.	数据集	-	-	24.13
CoT-SC +验证器	维基百科	33.1%	+5.4%	28.32
CoT-SC +验证器	DRQA	31.1%	+3.4%	27.75
CoT-SC +验证器	谷歌	33.6%	+5.9%	30.06
CoT-SC +验证器	数据集	**37.2%**	**+9.5%**	**32.28**

表2：**2WikiMultiHopQA** 数据集的结果。 Δ精确匹配表示相较于CoT-SC 基线的精确匹配提升。 所有实验均使用 GPT-3 davinci-003 模型。

方法	知识 准确率		Δ 准确率
CoT-SC → ReAct	维基百科	-	+4.2%
ReAct →CoT-SC	维基百科。	-	+1.6%
标准	-	46.8%	-
CoT	-	50.0%	-
CoT-SC	-	52.0%	-
CoT-SC +Calib.	-	33.7%	
CoT-SC +验证器	维基百科	53.6%	+1.6%
CoT-SC +验证器	DRQA	53.3%	+1.3%
CoT-SC +验证器	谷歌	53.9%	+1.9%

表3：在**Fever** 数据集上的结果。 Δ准确率表示相较于 CoT-SC 基线的准确率提升。前两行使用 PaLM 模型，其余行使用 GPT-3 davinci-003 模型。

因此，应将其与 VE 的最后一种设置进行比较，在该设置中我们使用数据集知识。相比之下，校准器得到的曲线下面积较低，而且无法提高准确率，因为它在开放域设置中不会生成备选答案。

使用验证与编辑框架，检索系统维基百科和 DrQA 在基线上分别提升了 4.5% 和 4.8%，是 ReAct 最高精确匹配提升（1.7%）的两倍。将谷歌搜索引擎的结果整合到该框架后，精确匹配提升了 6.5%，是 ReAct 结果的 3.8 倍。这展现了一种结合搜索引擎与大语言模型的有前景的方法，而这也是当前热门的研究方向。搜索引擎返回事实性结果，但在需要推理的查询中能力较弱。另一方面，大语言模型擅长推理和抽象，却往往会生成听起来合理但不正确的陈述（OpenAI 博客，2022年；Zhao 等人，2023）。为了兼取两者之长，我们可以利用大语言模型的长时记忆能力，因为许多用户报告称，GPT 能够记住对话中较早提及的输入。通过将搜索引擎提供的事实性结果作为记忆补充，GPT 能够生成更好、更准确的预测结果。

然后，当我们使用数据集中提供的对抗增强段落时，模型能够同时展现出非常高的精确匹配（EM）（56.8%）和曲线下面积（AUC）（60.94）。这一设置表明，如果我们拥有一组高度压缩的上下文以及一个近乎理想的检索系统，那么 Verify-and-Edit 框架就有可能取得非常出色的性能。

5.3 2WikiMultiHop 上的结果

如表2所示，与 HotpotQA 相比，我们的方法在 2WikiMultiHop 上表现出更强的性能。结合开放域检索的验证与编辑框架能够显著提高准确率，提升幅度为 3.4% 至 5.9%。从数据集中提供的段落中进行选择（这些段落包括支持性证据和无关段落），准确率的提升幅度进一步增加至 9.5%。另一方面，校准器虽然使用了数据集提供的段落，但其表现仍落后于验证与编辑框架的所有变体。

5.4 事实核验结果

Fever 数据集上的结果如表3所示。由于 Fever 数据集所需的推理过程相比 HotpotQA 和 2Wiki-MultiHop 而言，多跳推理程度较低，因此我们预计其性能提升幅度应低于另外两个数据集。

在 Fever 数据集中，校准器方法完全失效，性能下降至 33.7%：该方法根据事实性估计来校准预测分数，而事实性估计是通过检查推理路径与所提供上下文之间的重叠程度得出的。然而，在此类事实核查数据集中，并未提供上下文。因此，我们使用原始主张进行校准，这导致了较差的性能。这表明，校准器方法的一个局限性在于，它仅适用于提供了相关上下文的情况。

尽管这项任务并不需要太多推理，但采用验证与编辑框架后，我们仍然能够观察到相较于基线方法的稳定提升。与之前类似，维基百科检索相比 DrQA 能够带来更大的性能提升，而 Google 搜索则进一步提升了 1.9%。

# 示例	科恩 κ	CoT-SC	Ours	Tie
50	0.25	17%	**53%**	30%

表4：HotpotQA 数据集上思维链事实性的人工研究。“我们的方法”指采用谷歌检索的验证与编辑模型。

与我们的方法相比，ReAct 在 Fever 上能够展现出更大的性能提升。首先，前文已经提到，Fever 并不太适合验证与编辑框架，因为解决该任务所需的推理较少。其次，ReAct 的提示比我们的提示长得多，因此需要付出更高的计算成本。

5.5 成本考量

由于在与大语言模型交互时，降低成本是一个主要关注点，因此我们的方法也将其纳入考量，并尝试从两个方面降低计算成本：首先，验证与编辑仅对选定的实例进行编辑，而其他方法则每次都会进行编辑。具体而言，我们仅在模型不确定时（根据一致性进行判断）才进行修订，这种情况仅占 40%。相比之下，其他方法（如 ReAct）会为每个实例检索相关信息并进行编辑，因此产生更高的成本。其次，验证与编辑设计的任务自然且具有对话性，只需少量示例和简短的提示即可完成学习。例如，其他方法通常需要学习非自然的调用形式，例如 ReAct 中的 [thought]和 [action]标签，以及 Toolformer 中的应用程序接口调用（Schick 等人，2023）。因此，大语言模型需要更长的提示、更多的示例，甚至需要通过微调才能学习这些格式。另一方面，我们将验证与编辑任务设计得尽可能自然，从而最大限度地降低学习所需的工作量。我们的任务仅包含提问和回答，不涉及需要学习的合成标签或任务。作为对比，使用 GPT-3 API 编辑一个 Fever 实例时，验证与编辑的成本为 0.014 美元，而 ReAct 的成本为 0.017 美元。

5.6 通过人工研究评估推理链

为深入考察生成的推理链是否忠实，我们还开展了一项小规模人工研究实验。实验中，两名人类志愿者会看到从 HotpotQA 数据集随机选取的 50 个问题，以及 CoT-SC 和验证与编辑生成的推理链。随后，他们需要选出

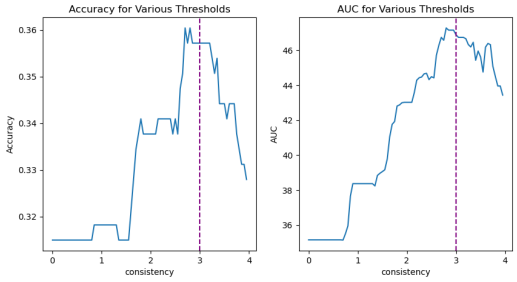

图3：消融研究——不同一致性阈值对对抗性 HotpotQA 任务性能的影响

在事实层面更一致的一项。我们鼓励志愿者使用搜索引擎辅助判断。实验设置的详细说明见附录D。

如表4所示，与 CoT-SC 基线相比，人类标注者有 53% 的时间认为 Verify-and-Edit 生成的推理链在事实上一致，而 CoT-SC 基线仅为 17%。Cohen κ 为 0.25，表明两位标注者之间具有一般程度的一致性（McHugh，2012）。标注者在 100% 的情况下都使用了 Google 搜索作为辅助工具，这表明引入外部知识的必要性。

此外，在此情况下，人工标注需要付出大量努力。标注人员报告称，验证一个数据点平均需要 1.5 分钟。因此，自动化验证与编辑过程作为一种辅助工具，有助于减少人力劳动。

为了详细观察验证器-编辑框架的定性效果，我们还在附录E中列举了几个有趣的示例，展示了我们框架在纠正原始主张方面的有效性。

5.7 消融研究：在不同一致性阈值下进行编辑

在验证与编辑框架中，唯一需要选择的超参数是一致性阈值。ReAct 中也存在类似的阈值（Yao 等人，2022年），其中 CoT → ReAct 方法是在“n 个 CoT-SC 样本中，多数答案出现次数少于 n/2”时采用 ReAct 风格提示。然而，与使用基于对数概率定义的原始一致性相比，使用多数计数的粒度较粗。因此，我们采用 Wang 等人(2022年提出的原始分数，即对推理依据的对数概率进行边缘化后得到的未归一化答案概率

。为了模拟多数投票阈值，我们选择 $\lceil n/2 \rceil$，其中 n 是采样路径的数量。

为研究调整一致性阈值对我们框架的影响，我们在图 3中展示了对抗性 HotpotQA 的消融实验结果。随着阈值升高，准确率先上升，在接近 $\lceil n/2 \rceil$（即 3）时达到峰值，随后下降。AUC 分数也呈现出类似趋势。

如图 2所示，当一致性高于多数阈值（$\lceil n/2 \rceil$）时，正确预测结果通常多于错误预测结果，反之亦然。因此，当我们将一致性阈值从 0 提高到 $\lceil n/2 \rceil$时，越来越多不确定且可能错误的样本会通过引入外部知识进行编辑。当阈值超过理想值 $\lceil n/2 \rceil$后，我们主要是在重新编辑正确样本，而引入的噪声可能会扰乱原有的推理链。

因此，我们建议将$\lceil n/2 \rceil$设为理想的一致性阈值。

6 结论

本文介绍了一种面向开放域问答的验证与编辑框架。这是首次尝试对思维链式推理链进行后编辑，以提升最终任务表现。该框架将知识检索与推理相结合，以自然且对话式的方式编辑思维链，从而增强预测结果的事实性。结合 Google 搜索后，该框架也展现出一种颇有前景的方向：将最先进大语言模型的开放生成能力与搜索引擎提供的最新事实相结合。

局限性

当前框架存在一些局限性。首先，验证与编辑最适用于需要复杂推理的开放域问答任务。对于较简单的数据集或无需知识检索的常识数据集，该方法可能无法带来显著提升。其次，最理想的情况是编辑一组大多不正确的样本，我们尝试利用一致性来筛选这类样本。因此，我们的方法依赖于一致性方法的性能，以及其区分正确与错误预测结果的能力。通常，面对更具挑战性的一组示例时，该方法能够展现出更大的提升。

为了解决这些局限性，我们计划着手减少理由编辑阶段引入的噪声，并在后续利用更多知识资源，例如知识库。

伦理声明

验证与编辑框架可以缓解大语言模型生成过程中围绕幻觉和不符合事实的细节所产生的潜在伦理问题。一些持续存在的问题包括：(1) 由于该框架使用谷歌作为知识检索方法之一，因此可能会检索到谷歌搜索结果中存在的潜在有害信息。(2) 由于该框架以 GPT-3 作为基础模型，因此可能会面临 GPT-3 现有的伦理问题，例如回应有害查询或表现出偏见行为。

在知识检索方面，我们使用了维基百科语料库和 Google 搜索结果。根据知识共享署名-相同方式共享 3.0 未本地化许可的条款，允许复制、分发和修改维基百科文本。对于 Google 搜索结果，美国上诉法院认为，抓取公开可访问的数据是合法的。

7 致谢

本研究由新加坡国家研究基金会通过其 AI Singapore 计划（AISG 奖项编号：AISG-PhD/2021-01-001[T]）资助。

参考文献

Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。2020年。语言模型是少样本学习器。神经信息处理系统进展，33：1877–1901。

Danqi Chen、Adam Fisch、Jason Weston 和 Antoine Bordes。2017年。阅读维基百科以回答开放领域问题。载于计算语言学协会第*55*届年会论文集（第*1*卷：长篇论文），第1870–1879页，加拿大温哥华。计算语言学协会。

Aakanksha Chowdhery、Sharan Narang、Jacob Devlin、Maarten Bosma、Gaurav Mishra、Adam Roberts、Paul Barham、Hyung Won Chung、Charles Sutton、Sebastian Gehrmann 等。2022年。PaLM：利用路径扩展语言建模。*arXiv* 预印本 *arXiv:2204.02311*。

Antonia Creswell、Murray Shanahan 和 Irina Hig-gins。2022年。选择-推理：利用大型

语言模型进行可解释的逻辑推理。*arXiv* 预印本 *arXiv:2205.09712*。Olga Golovneva、Moya Chen、Spencer Poff、Martin Corredor、Luke Zettlemoyer、Maryam Fazel-Zarandi 和 Asli Celikyilmaz。2022年。Roscoe：一套用于评估逐步推理的指标。*arXiv* 预印本 *arXiv:2212.07919*。Kelvin Guu、Kenton Lee、Zora Tung、Panupong Pasupat 和 Ming-Wei Chang。2020年。REALM：检索增强型语言模型预训练。载于 第*37*届国际机器学习会议论文集，ICML'20。JMLR.org。Xanh Ho、Anh-Khoa Duong Nguyen、Saku Sugawara 和 Akiko Aizawa。2020年。构建一个用于全面评估推理步骤的多跳问答数据集。载于 第*28*届国际计算语言学会议论文集，第6609–6625页，西班牙巴塞罗那（线上）。国际计算语言学委员会。

Vladimir Karpukhin、Barlas Oguz、Sewon Min、Patrick Lewis、Ledell Wu、Sergey Edunov、Danqi Chen 和 Wen-tau Yih。2020年。用于开放领域问答的稠密段落检索。载于*2020*年自然语言处理实证方法会议论文集（*EMNLP*），第6769–6781页，在线出版。计算语言学协会。

Angeliki Lazaridou、Elena Gribovskaya、Wojciech Stokowiec 和 Nikolai Grigorev。2022年。通过少样本提示增强的互联网语言模型用于开放域问答。

Gary Marcus。2022年。ChatGPT 真的是 Google 搜索的"红色警报"吗?

Mary L McHugh。2012。评估者间信度：Kappa 统计量。*Biochemiamedica*，22(3)：276–282。

OpenAI 博客。2022年。 ChatGPT：为对话优化语言模型。

Long Ouyang、Jeff Wu、Xu Jiang、DiogoAlmeida、Car- roll L Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Ray 等。2022年。训练语言模型以遵循带有人类反馈的指令。*arXiv* 预印本 *arXiv:2203.02155*。

Ofir Press、Muru Zhang、Sewon Min、Lud wig Schmidt、Noah A Smith 和 Mike Lewis。2022年。衡量并缩小语言模型中的组合性差距。*arXiv* 预印本 *arXiv:2210.03350*。Nils Reimers 和 Iryna Gurevych。2019年。Sentence-BERT：使用孪生 BERT 网络生成句子嵌入。载于《*2019*年自然语言处理经验方法会议暨第九届国际自然语言处理联合会议（*EMNLP-IJCNLP*）论文集》，第3982–3992页，中国香港。计算语言学协会。

Timo Schick、Jane Dwivedi-Yu、Roberto Dessì、Roberta Raileanu、Maria Lomeli、Luke Zettlemoyer、Nicola Cancedda 和 Thomas Scialom。2023年。Toolformer：语言模型可以自学使用工具。*arXiv* 预印本 *arXiv:2302.04761*。James Thorne、Andreas Vlachos、Christos Christodoulopoulos 和 Arpit Mittal。2018年。FEVER：用于事实提取与验证的大规模数据集。载于《*2018*年计算语言学协会北美分会会议论文集：人类语言技术，第*1*卷（长篇论文）》，第809–819页，美国路易斯安那州新奥尔良。计算语言学协会。Xuezhi Wang、Jason Wei、Dale Schuurmans、Quoc Le、Ed Chi 和 Denny Zhou。2022年。自洽性能够提升语言模型中的思维链推理。*arXiv* 预印本 *arXiv:2203.11171*。Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Ed Chi、Quoc Le 和 Denny Zhou。2022年。思维链提示能够引导大语言模型进行推理。*arXiv* 预印本 *arXiv:2201.11903*。

Sarah Wiegreffe、Jack Hessel、Swabha Swayamdipta、Mark Riedl 和 Yejin Choi。2022年。重新构想人类与人工智能协作以生成自由文本解释。载于《*2022*年北美计算语言学协会分会会议论文集：人类语言技术》，第632–658页，美国西雅图。计算语言学协会。

Zhilin Yang、Peng Qi、Saizheng Zhang、Yoshua Bengio、William Cohen、Ruslan Salakhutdinov 和 Christo- pher D. Manning。2018年。HotpotQA：一个用于多样化、可解释多跳问答的数据集。载于《*2018*年自然语言处理实证方法会议论文集》，第2369–2380页，比利时布鲁塞尔。计算语言学协会。

Shunyu Yao、Jeffrey Zhao、Dian Yu、Nan Du、Izhak Shafran、Karthik Narasimhan 和 Yuan Cao。2022年。ReAct：在语言模型中协同推理与行动。*arXiv* 预印本 *arXiv:2210.03629*。

Xi Ye 和 Greg Durrett。2022年。少样本文本推理提示中解释的不可靠性。载于神经信息处理系统进展。

Ruochen Zhao、Xingxuan Li、Yew Ken Chia、Bosheng Ding 和 Lidong Bing。2023年。类 ChatGPT 生成模型能否保证事实准确率？论新一代搜索引擎的错误。*arXiv* 预印本 *arXiv:2304.11076*。Denny Zhou、Nathanael Schärli、Le Hou、Jason Wei、Nathan Scales、Xuezhi Wang、Dale Schuurmans、Olivier Bousquet、Quoc Le 和 Ed Chi。2022年。由简到繁提示使大语言模型能够进行复杂推理。*arXiv* 预印本 *arXiv:2205.10625*。

《Verify-and-Edit：知识增强型思维链框架》附录

A 数据集处理

A.1 对抗性 HotpotQA

对抗性 HotpotQA 子集由以下工作构建：Ye 和 Durrett (2022年），他们对原始数据集进行了几项处理：（1）缩短上下文长度，使其更适合用于测试上下文学习。（2）将对抗性上下文集合缩减为两个真实标注支持段落和两个对抗段落，而不是使用全部八个干扰段落。每个段落还经过进一步简化，仅保留回答问题（或干扰预测结果）所需的相关句子。（3）构建一个包含 250 个样例的高难度测试集，平衡提示`text-davinci-001`（他们在实验时使用的模型）作出正确和错误预测结果的样例比例。具体做法是，先对 1000 个样例进行少样本推理，然后分别随机抽取 125 个预测正确和预测错误的样例。这个子采样数据集已由 Ye 和 Durrett(2022年）公开发布在 GitHub 上。

HotpotQA 数据集采用知识共享署名-相同方式共享 4.0 国际许可协议发布，该协议允许修改和用于研究。

A.2 2WikiMultihopQA

出于成本考虑，我们从开发集的 12,576 个样本中随机抽取 1,000 个，这可以提供合理的估算。为便于复现，我们在代码库中发布了抽样索引。

2WikiMultihop数据集采用 Apache 许可证 2.0 授权，允许修改和用于研究。

A.3 Fever

为模拟对抗性 HotpotQA 的实验设置，我们运行思维链基线处理 3,000 个样本，并通过平衡正确预测和错误预测的数量，随机抽取 1,000 个样本。为便于复现，我们在代码库中公开了抽取样本的索引。

Fever 的数据标注包含来自维基百科的材料，这些材料根据维基百科版权政策授权。

B 实验成本

在实验中，我们使用 text-davinci-003 的应用程序接口。大语言模型推理的费用为每 1K 个令牌 0.02 美元。我们总共花费了 273 美元。

C 使用的提示

C.1 HotpotQA

C.1.1 少样本提示

Q：这位英国赛车手在2014年巴林GP2系列赛分站中获得第三名，他出生于哪一年?

A：1991

Q：Antony King 曾与哪支于1985年在曼彻斯特成立的乐队合作?
A：Simply Red

Q：Alberta Ferretti 的工作室所在地点附近的城市有多少居民？**A**：146,606

问：TLC：桌子、梯子与椅子是一场摔角赛事，主赛中有哪些美国摔角手兼说唱歌手?

答：约翰·费利克斯·安东尼·塞纳

问：于1998年1月31日获得大象勋章的人出生于哪一天？**答**：1967年4月27日

问：《III - Odyssey of the Mind》是一支德国乐队的第六张专辑，这支乐队成立于哪座城市?
答：杜塞尔多夫

问：[问题]**答**:

C.1.2 思维链、思维链自洽性提示

Q：这位英国赛车手在2014年巴林GP2系列赛分站中获得第三名，他出生于哪一年?

A：首先，在2014年巴林GP2系列赛分站中，DAMS车手Jolyon Palmer获得第三名。其次，Jolyon Palmer（出生于1991年1月20日）是一位英国赛车手。答案是1991年。

Q：Antony King曾与哪个成立于1985年、来自曼彻斯特的乐队合作?
A：首先，Antony King 曾担任 Simply Red 的录音室工程师。其次，Simply Red 于 1985年 在曼彻斯特成立。答案是 Simply Red。

Q：Alberta Ferretti 的工作室所在地点附近的城市有多少居民?
A：首先，Alberta Ferretti 的工作室位于里米尼附近。其次，里米尼是一座拥有146,606名居民的城市。答案是146,606。

Q：TLC：桌子、梯子与椅子是一场摔跤赛事，主赛中有哪些美国摔跤手和说唱歌手?

A：首先，TLC：桌子、梯子与椅子是一场摔跤赛事，约翰·塞纳参加了主赛。其次，约翰·塞纳是美国摔跤手和说唱歌手。答案是约翰·费利克斯·安东尼·塞纳。

Q：于1998年1月31日获颁大象勋章的人出生于哪一天？**A**：首先，威廉-亚历山大国王于1998年1月31日获颁大象勋章。其次，威廉-亚历山大出生于1967年4月27日。答案是1967年4月27日。

Q：III-《Odyssey of the Mind》是一支德国乐队的第六张专辑，该乐队成立于哪个城市?
A：首先，III - Odyssey of the Mind 是德国乐队 Die Krupps 的第六张专辑。其次，Die Krupps 成立于杜塞尔多夫。答案是杜塞尔多夫。

Q：[问题]
A:

C.1.3 验证问题生成提示
提出一个询问整体问题答案的问题。

整体问题：哨兵岛语是孟加拉湾哪一座岛屿居民所使用的语言?

答案：北哨兵岛居民所使用的语言是哨兵岛语。

问题：哨兵岛语是哪种语言?

总体问题：《爱尔兰之声》中有两个名额被哪支驻扎在英格兰伦敦的英爱女子组合占据?

答案：Little Mix 是一支总部位于英格兰伦敦的女子组合。**问题**：哪支女子组合总部位于英格兰伦敦?

总体问题：[原始问题]**答案**：[待编辑的推理过程句子]**问题**:

C.1.4 验证答案生成（推理过程编辑）
提示
Barnes House（出生于1969年1月20日）是一名英国赛车手，目前效力于雷诺运动F1车队，参加一级方程式世界锦标赛。Jolyon Palmer（出生于1991年1月20日）是一名英国赛车手，目前效力于雷诺运动F1车队，参加一级方程式世界锦标赛。Ming Xi（出生于2015年1月20日）是一名英国赛车手，目前效力于雷诺运动F1车队，参加一级方程式世界锦标赛。

2014年巴林GP2系列赛分站是GP2系列赛的一部分，于2014年4月6日和7日在巴林萨基尔的巴林国际赛车场举行的两场赛车比赛。Julián Leal代表Carlin车队获得第二名，DAMS车手Jolyon Palmer获得第三名。
问题：这位在2014年巴林GP2系列赛分站获得第三名的英国赛车手出生于哪一年?

A：这位英国赛车手在2014年巴林GP2系列赛分站中获得第三名，出生于1991年。

Antony King（1974年出生）是一位英国现场音响工程师，曾为 Depeche Mode 和 Nine Inch Nails 工作。他还曾担任 The Cure、Noel Gallagher's High Flying Birds、Band of Horses、Zayn、Beck、Marilyn Manson、The Faces 和 Simply Red 的前场音响工程师。

Anthony Collett 是一支英国灵魂乐与流行乐队，于1985年在曼彻斯特成立。
Olé Olé（1974 年出生）是一位英国现场音响工程师，曾为 Depeche Mode 和 Nine Inch Nails 工作。他还曾担任 The Cure、Noel Gallagher's High Flying Birds、Band of Horses、Zayn、Beck、Marilyn Manson、The Faces 和 Christopher Trumbo 的前场音响工程师。

Simply Red 是一支英国灵魂乐与流行乐队，于1985年在曼彻斯特成立。
问：Antony King 曾与哪个乐队合作？该乐队于1985年在曼彻斯特成立。
A：Antony King 与乐队 Simply Red 合作，该乐队于1985年在曼彻斯特组建。

Alberta Ferretti（卡托利卡，1950年）是一位意大利时装设计师和女装裁缝。她的展厅位于意大利米兰，而她的工作室则设在意大利里米尼附近的卡托利卡村。

里米尼（]；罗马涅方言："Rémin"；拉丁语："Ariminum"）是意大利北部艾米利亚-罗马涅大区的一座城市，拥有146,606名居民

也是里米尼省的省会。Queequeg（]；罗马涅方言：“Rémin”；拉丁语：“Ariminum”）是意大利北部艾米利亚-罗马涅大区的一座城市，拥有546606名居民，也是Queequeg省的省会。唐人街（]；罗马涅方言：“Rémin”；拉丁语：“Ariminum”）是意大利北部艾米利亚-罗马涅大区的一座城市，拥有346606名居民，也是唐人街省的省会。**问**：Alberta Ferretti的工作室所在城市附近的城市有多少居民？**答**：Alberta Ferretti的工作室所在城市附近的城市有146,606名居民。

[上下文]**Q**：[验证问题]**A**：

C.2 2WikiMultihop

C.2.1 少样本提示

Q：哪部电影上映得更早，Kistimaat 还是 I’MTaraneh, 15？**A**：I’MTaraneh, 15

Q：电影 Baalaraajana Kathe 的作曲家逝世于何时？**A**：1980年5月27日

Q：电影《Alive》（1993年电影）的导演的配偶是谁?
答：Kathleen Kennedy

问：Edward Frederick Sanderson 和 Forrest Towns，谁活得更久？**答**：Edward Frederick Sanderson

问：电影《Battle Circus》的导演来自哪个国家?

答：美国人

问：电影《In the Fog》的导演来自哪个国家?

答：乌克兰人

问：[问题]**答**:

C.2.2 思维链、CoT-SC 提示

问题：哪部电影上映时间更早，Kisti- maat 还是 I’M Taraneh, 15？**A**：首先，电影 Kistimaat 于2014年10月6日上映。其次，I’MTaraneh,15 于2002年11月13日上映。答案是 I’M Taraneh, 15。

问题：电影 Baalaraajana Kathe 作曲家的逝世日期是什么?
A：首先，电影 Baalaraajana Kathe 的作曲者是 S. Hanumantha Rao。其次，S. Hanumantha Rao 于1980年5月27日去世。答案是1980年5月27日。

问题：《Alive》（1993年电影）的导演的配偶是谁?
A：首先，电影《活着》的导演是 Frank Mar- shall。其次，FrankMarshall’sspouseisKathleen Kennedy。答案是 Kathleen Kennedy。

问题：Edward Frederick Sanderson 和 Forrest Towns 谁活得更久？**回答**：首先，Edward Frederick Sanderson 去世时享年 81 岁。其次，Forrest Towns 去世时享年 77 岁。答案是 Ed ward Frederick Sanderson。

问题：电影《Battle Circus》（电影）的导演来自哪个国家？**答**：首先，电影《Battle Circus》（电影）的导演是 Richard Brooks。其次，Richard Brooks 是美国人。答案是美国人。**问题**：电影《In the Fog》的导演来自哪个国家?
答：首先，电影《In the Fog》的导演是谢尔盖·洛兹尼察。其次，谢尔盖·洛兹尼察是乌克兰人。答案是乌克兰人。

问题：[question]
A:

C.2.3 验证问题生成提示

写一个问题，用于验证某个总体问题的理由。**总体问题**：电影 Baalaraajana Kathe 的作曲家去世于何日？**理由**：首先，电影 Baalaraajana Kathe 的作曲家是 S. Hanumantha Rao。**问题**：电影 Baalaraa- jana Kathe 的作曲家是谁?

总体问题：Edward Frederick Sanderson 和 Forrest Towns，谁活得更久？**推理**：首先，Edward Frederick Sanderson 去世时享年 81 岁。**问题**：Edward Frederick Sander- son 活了多久?

总体问题：[原始问题]**推理过程**：[推理过程句子]**问题**:

C.2.4 验证答案生成（理由编辑）提示

这部电影于1984年由 Essex Films 发行。Kistimaat 是一部2014年上映的孟加拉国动作片，由 Ashiqur Rahman 执导，Tiger Media Limited 和 The Abhi Pictures 制作。I'M Taraneh, 15 是一部2002年上映的伊朗电影，由 Rasul Sadrameli 执导。这部电影于2001年5月4日上映。**问题**：Kistimaat 这部电影是什么时候上映的？**答案**：Kistimaat 于2014年上映。

Dwaram Venkataswami Naidu，也是一位作词人。这部电影的配乐由 S. Hanumantha Rao 创作。Rao 于1980年5月27日逝世。Rao 与 Raja Mani 结婚，两人育有三个女儿和一个儿子。

问题：电影《Baalaraa-jana Kathe》的作曲家是谁?
答案：电影 Baalaraajana Kathe 的作曲者是 S. Hanumantha Rao。

Adib Kheir 是20世纪20年代一位杰出的叙利亚民族主义者。这部影片在加拿大不列颠哥伦比亚省的珀塞尔山脉实地拍摄，由 Frank Marshall 执导，John Patrick Shanley 编剧，John Malkovich 担任旁白。Frank Wilton Marshall（出生于1946年9月13日）是一位美国电影制片人兼导演，经常与妻子 Kathleen Kennedy 合作。2018年，他获得了美国电影艺术与科学学院颁发的欧文·G·撒尔伯格纪念奖。

问题：电影《Alive》（1993年电影）的导演是谁?
答案：电影《活着》的导演是弗兰克·马歇尔。

[上下文]**问题**：[验证问题]**答案**:

C.3 Fever

C.3.1 少样本提示

判断是否存在支持或反驳某项主张的观察，或者信息不足。

主张：Reg Watson 是一位现任电视节目制作人。

A：驳斥

主张：加兹登旗以克里斯托弗·加兹登的名字命名。**A**：信息不足

主张：《黑镜》讲述的是社会。**A**：支持

主张：[问题]
A:

C.3.2 思维链、思维链自洽性提示

判断是否存在支持或驳斥某项主张的观察，或者信息不足。

主张：Reg Watson 是一位现任电视节目制作人。

A：首先，Reginald James Watson AM 是一位澳大利亚电视制片人和编剧。其次，Reginald James Watson AM 于2019年10月8日去世。答案是：驳斥（REFUTES）。

主张：加兹登旗以克里斯托弗·加兹登的名字命名。
A：首先，加兹登旗以政治家 Christopher Gadsden 的名字命名。其次，没有关于是谁为加兹登旗命名的信息。答案是信息不足。

主张：《黑镜》讲的是社会。
A：首先，《黑镜》是一部英国选集式电视剧。其次，该剧借助科技来探讨当代社会问题。答案是支持。

主张：[问题]
A:

C.3.3 验证问题生成提示编写一个问题，用于验证某项主张的理由。

主张: Reg Watson 是一位现任电视节目制片人。

理由：Reginald James Watson AM 是一位澳大利亚电视制片人和编剧。
问题：Reg Watson 的职业是什么?

主张：加兹登旗以克里斯托弗·加兹登的名字命名。**理由**：没有关于是谁给加兹登旗命名的信息。**问题**：是谁给加兹登旗命名的?

主张：[问题]**推理过程**：[推理过程句子]**问题**:

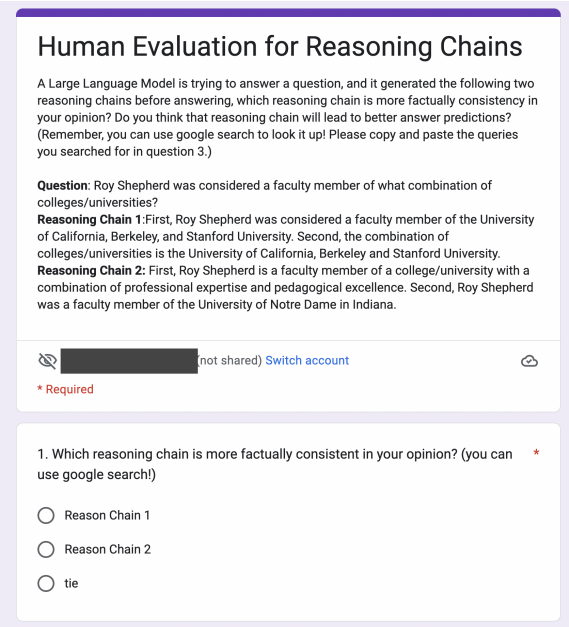

图 4：人工评估用户界面示例截图。

C.3.4 验证答案生成（推理过程编辑）提示

Reginald James Watson AM（1926年8月27日－2019年10月8日）是一位澳大利亚电视制片人和编剧。他曾担任《Crossroads》的执行制片人，并创作了《Prisoner》《Neighbours》《The Young Doctors》和《Sons and Daughters》等澳大利亚媒体出口剧集。

问题：Reg Watson 的职业是什么？**答案**：Reg Watson 是一位澳大利亚电视制片人和编剧。

这面旗帜以政治家克里斯托弗·加兹登（Christopher Gadsden，1724—1805）的名字命名。他于1775年美国独立战争期间设计了这面旗帜。
问题：加兹登旗是谁命名的？**答案**：加兹登旗以克里斯托弗·加兹登的名字命名，但没有关于是谁为其命名的信息。

[上下文]
问题：[验证问题]**答案**:

D 人工研究

为开展人工研究，我们向图4中的两位人类志愿者展示了相关说明。这些志愿者是精通英语的自然语言处理博士生。他们了解数据收集的用途，并对此达成一致。推理链1和链2分别由 CoT-SC 基线和验证与编辑生成，并以随机顺序展示。平均而言，每位志愿者花费1.25小时完成50个样本。

E 定性示例

在 表5中，详细展示了 Adversarial HotpotQA 数据集中的 3 个示例。

在第一个示例中，大语言模型错误地指出这首歌“基于……蜘蛛侠”。然而，在谷歌检索到的事实中，明确指出它是基于“恶灵骑士”的。因此，检索到的事实能够帮助纠正推理过程中的这一细节。此外，尽管原始推理过程也涉及品牌名称“漫威漫画”，但模型生成的答案却继续使用英雄名称，而不是题目所询问的“品牌”。再次输入相关信息也纠正了这一逻辑错误。

在第二个示例中，大语言模型编造了一个听起来合理的事实：“Tony Robinson 写过七本儿童书籍。”此外，大语言模型也没有标明自己对这一主张的置信度。因此，如果用户对这方面的知识并不熟悉，就很容易将其误认为事实，这具有很高的风险。通过引入谷歌作为辅助工具，我们检索到句子“他已出版 16 本儿童书籍。”基于这一新检索到的事实，大语言模型继续生成正确答案。

第三个例子很有意思。原始思维链在第一句话中就已经出错，接着第二句话也不断出错。这也是数据集中常见的一类错误。对其进行纠正后，验证与编辑框架能够用节目《Chelsea Does》纠正第一个主张。然而，第二个主张虽然经过验证，却已与原始问题无关。在这种情况下，将两句推理过程都输入大语言模型后，模型能够选出相关事实作为答案，同时忽略其余内容。这个例子表明，验证与编辑所采用的思维链设置非常重要，因为它能让模型再次进行推理和抽象，而不是简单地替换并纠正。

原始问题	Suicide 于 1977年发行的专辑中，有一首歌是以哪个品牌的漫画角色为创作灵感的?
正确答案	漫威漫画
CoT-SC	首先，Suicide 发行了他们于 1977年推出的专辑《Suicide: Alan Vega & Martin Rev》。 其次，这张专辑收录了一首以漫威漫画角色蜘蛛侠为题材的歌曲。 答案是蜘蛛侠。
验证问题	Suicide 于 1977年发行了哪张专辑? Suicide 于 1977年发行的专辑中，那首歌曲是以哪个漫画角色为题材的?
检索到的事实	《Suicide》是美国摇滚乐队 Suicide 的首张录音室专辑，于1977年发行。 由 Red Star唱片发行，并由 Craig Leon 和 Marty Thau 制作。 《恶灵骑士》是原始朋克乐队 Suicide 收录于其首张专辑中的一首歌曲。这首歌 基于漫威漫画角色。在《Suicide》中，引擎轰鸣的曲调散发出黑色电影的氛围。 “恶灵骑士”（以 Vega 最喜欢的漫画书命名）和“Rocket U.S.A.”令人害怕。
验证与编辑	首先，Suicide 于1977年发行了其首张录音室专辑，标题为《Suicide》。 其次，Suicide 1977年专辑中的这首歌以漫威漫画角色恶灵骑士为原型。 答案是漫威漫画。
原始问题	《梅德·玛丽安和她的快乐伙伴》这部情景喜剧的编剧创作了多少本儿童书籍?
正确答案	16
CoT-SC	首先，情景喜剧《梅德·玛丽安和她的快乐伙伴》的编剧是 Tony Robinson。 其次，Tony Robinson 写过七本儿童书籍。 答案是 七。
验证问题	情景喜剧《梅德·玛丽安和她的快乐伙伴》的编剧是谁? Tony Robinson 写过多少本儿童书籍?
检索到的事实	《梅德·玛丽安和她的快乐伙伴》是一部英国儿童电视连续剧，由 Tony Robinson 创作并编剧，David Bell 执导。该剧于 1989 年在 BBC1 开播，共播出四季， 最后一集于 1994 年播出。 他已出版 16 本儿童书籍。
验证与编辑	首先，情景喜剧《梅德·玛丽安和她的快乐伙伴》的编剧是 Tony Robinson。 其次，Tony Robinson 撰写了 16 本儿童书籍。 答案是16。
原始问题	2016年1月，切尔西·汉德勒还出演了哪档节目（《Hello Ross》除外）?
正确答案	《Chelsea Does》
CoT-SC	首先，切尔西·汉德勒于2016年1月登上了《吉米·法伦今夜秀》 （2016 年的） 其次，《吉米·法伦今夜秀》是一档由吉米·法伦主持的脱口秀节目。 答案是《吉米·法伦今夜秀》。
验证问题	切尔西·汉德勒于2016年1月参加了哪档节目? 吉米·法伦主持的脱口秀节目叫什么?
检索到的事实	《Chelsea Does》是一部美国流媒体电视纪录片系列，首次于 Netflix 上线，时间是 2016年1月23日。……这些剧集跟随喜剧演员切尔西·汉德勒，探索不同的…… 《吉米·法伦今夜秀》是一档由吉米· 在 NBC 播出的《法伦今夜秀》。
验证与编辑	首先，切尔西·汉德勒于2016年1月出现在节目《Chelsea Does》中。 其次，《吉米·法伦今夜秀》是一档由吉米·法伦主持的脱口秀节目。 答案是《Chelsea Does》。

表5：AdvHotpotQA中的示例，事实通过谷歌检索。