

长文本生成中的精确信息控制

Jacqueline He^ω Howard Yen^{π*} Margaret Li^{ω*} Shuyue Stella Li^ω Zhiyuan Zeng^ω
Weijia Shi^ω Yulia Tsvetkov^ω Danqi Chen^π Pang Wei Koh^{ω,α} Luke Zettlemoyer^ω
^ωPaul G. Allen School of Computing Science & Engineering, University of Washington
^πPrinceton Language and Intelligence (PLI), Princeton University ^αAllen Institute for AI
{ jyyh, pangwei, lsz } @cs.washington.edu

Abstract

现代语言模型 (LMs) 的一个核心挑战是内在幻觉：即相对于输入上下文生成看似合理但未经证实的信息。为研究这个问题，我们提出了精准信息控制 (PIC)，这是一种新的任务形式，要求模型在生成长篇输出时，必须基于一组已提供的简短自成一体陈述 (称为可验证的声明)，而不能添加任何不支持的内容。为全面性考虑，PIC 包含一个完整的设定，测试模型完全包括所有输入声明的能力，以及一个部分设定，要求模型有选择地仅整合相关声明。我们提出了 PIC-Bench，这是一套包括八个长篇生成任务 (如摘要、生平生成) 的基准，这些任务被调整到 PIC 设定，其中 LMs 被提供以精心编写的、可验证的输入声明。我们对一系列开放和专有 LMs 在 PIC-Bench 上的评估显示，令人惊讶的是，最先进的 LMs 依然在超过 70 % 的输出中出现内在幻觉。为了缓解这种不忠实性，我们引入了一种训练后框架，使用一种弱监督偏好数据构建方法，来训练一个具有更强 PIC 能力的 8B PIC-LM——在完整的 PIC 设置中从 69.1 % 提高到 91.0 % F_1 。当集成到端到端事实生成流水线中，PIC-LM 在带有检索的模糊问答中提升了 17.1 % 的精确匹配召回率，并在出生地验证任务中提高了 30.5 % 的事实准确度，这突显了精准立基生成的潜力。

1 介绍

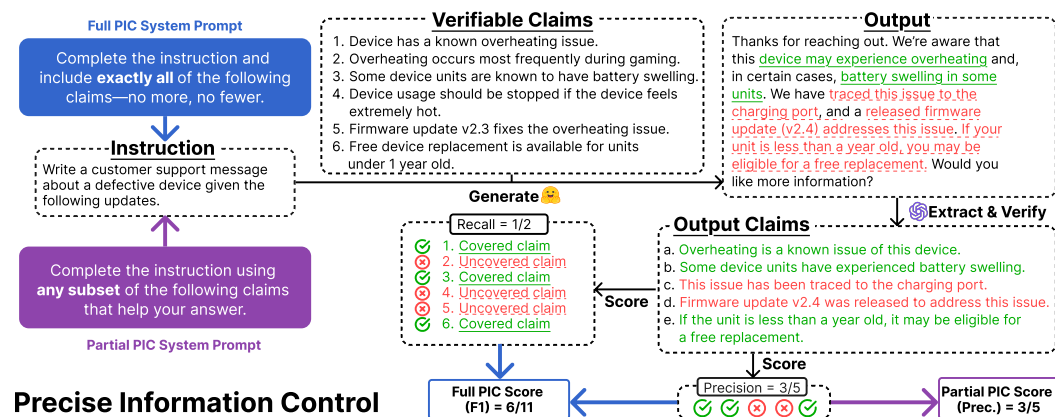


Figure 1: 精确信息控制 (PIC) 在可验证的声明水平上评估长篇生成的严格基础，并具有两种设置：完整的 PIC 需要包含所有输入声明且不允许有其他声明，而部分的 PIC 需要输入声明的子集。?? 显示了精确的系统提示。在长篇生成中精确控制语言模型 (LMs) 生成的内容是一个未解决的挑战。语言模型可能遗漏关键细节，或生成多余或错误的断言。这不仅适用于依赖参数知识的非限制性任务，即使

*Equal contribution.

在提示明确指出模型应包含哪些信息时也是如此。遗漏、扭曲和虚构共同构成了信息控制中特定类型的失效，更广泛地被称为幻觉 [37, 35, 36]。

在语言模型术语中，幻觉被分为外部幻觉（即，与现实世界知识不一致的生成）或内部幻觉（即，与提供的上下文输入不一致的生成） [67, 75]。内部幻觉在需要可靠地转换或综合源文本的场合会引发问题，例如在科学文献综述 [5]、抽象摘要 [67] 或任意风格转移 [90] 中。在快速变化或高度重要的现实世界应用中，消除内部幻觉同样至关重要，在这些应用中，更新的可靠上下文证据必须优先于模型陈旧的参数记忆 [107]。例如，在收集关于 COVID-19 的最新科学共识时，或根据患者的批准治疗列表生成定制的医疗建议时。此外，即使是减少外部幻觉的努力，也可能需要在检索到的或经过筛选的上下文中进行可靠的基础工作 [7]，使得内部控制成为事实生成的关键构成部分。

我们提出了精确信息控制 (PIC)，这是一种用于评估内在幻觉的任务表述，具体到主张的粒度，并在严格约束下进行。PIC 提出了一个基本问题：语言模型能否在不引入任何不支持的主张的情况下，生成严格基于一组明确语句的长篇回应？

我们不是形成关于内在幻觉的二元判断 [70, 36]，而是在可验证的声明层面上框定这个问题——简短的、独立的信息单位，可以明确验证 [98]。为了支持更灵活的用户定义控制，我们考虑了两种任务设置，这些设置是用自然语言指定的。完整的 PIC 反映了需要完整信息覆盖的场景（例如文本重写），在这种情况下，语言模型应该结合所有给定的输入声明（以 F_1 评估）。另一方面，部分 PIC 适用于上下文包含不相关或无关的声明，需要选择性地进行评估（例如摘要）。因此，在这种情况下，语言模型应该选择一个相关的子集（以精确度进行评估）。例如，Fig. 1 展示了一个提示，要求根据更新的指南写一条关于有缺陷设备的客户支持消息。完整的 PIC 要求语言模型将所有六个声明整合到它的回答中，而在部分设置下，语言模型可以选择任何有用的子集。两种 PIC 模式都需要精确的控制：每个输出声明都必须由输入声明证实。

我们介绍了 PIC -Bench，这是一个包含 1.2K 样本的长形式生成基准，用于在六个和两个部分 PIC 任务上评估 PIC。在 PIC -Bench 评估中，验证搜索空间仅限于输入上下文。因此，与外在幻觉不同，内在幻觉的完全消除应该是可处理且明确的。然而，我们发现现有最新的语言模型在 PIC 上远未解决。特别是，完美的 PIC（即内在幻觉的完全消除）仍然难以实现。特别是，完美 PIC 的输出比例对于完整的 PIC 任务不超过 30 %（由 F_1 测量），而对于部分 PIC 任务不超过 70 %（由精度测量）。这意味着，即使是前沿的大型语言模型，仍然在超过 70 % 的完整 PIC 输出中内在产生幻觉。此外，领先的开源 LMs 的表现一直低于专有 LMs 超过 8 %，表明还有改进的空间。

然后，我们开发了一个两阶段的后训练框架来提高 PIC 的能力，使用一种弱监督的奖励信号来有效地构建 PIC 格式的偏好数据，同时保持强大的指令遵循能力。特别是，我们使用这个框架微调 Llama 3.1 8B Instruct [28] 来生成 PIC -LM，一个高度可控的模型：它的生成结果能够可靠地遵循输入的声明，同时减少无依据的陈述。在 PIC -Bench 上，相对于 Llama 3.1 8B Instruct，PIC -LM 显著地将 F_1 中的平均全 PIC 从 69.1 % 提高到 91.0 %，并将精度中的平均部分 PIC 从 73.6 % 提升到 93.3 %，超越了所有公开的基线并缩小了与前沿 LLM 的性能差距。然而，完全消除长篇生成中的内在幻觉仍然难以实现，凸显了这个问题的挑战性。

最后，我们展示了 PIC -LM 在下游应用中的实际效用。尽管内在和外在幻觉通常被单独处理 [10, 36]，我们的结果表明，稳健的内在基础可能作为减少外在错误的基础。将我们的模型嵌入生成管道中，无论输入是通过背景信息证实的还是迭代地被完善，都会产生更符合事实且可靠的最终结果。在检索增强生成场景 [7] 中，主张来自检索到的背景，PIC -LM 在一个具有歧义的长形式问答任务 ASQA 上，将标准精确匹配提高了从 52.5 % 到 61.5 %。在一个链式验证管道 [17] 中，主张被自动生成和验证，PIC -LM 在一个与出生地相关的细节任务上提高了事实准确性（从 65.9 % 到 86.0 % 的事实精度），并在 QAMPaRl [2] 上提升了（从 13.5 % 到 22.6 % 在 F_1 @5）。

2 任务描述

在 [62, 98] 的基础上，我们将 PIC 看作可验证的声明的层面：简短的陈述，描述一个独立事件或状态，并根据某些知识来源（在我们的情况下，是输入的声明）进行评估。将输出分解为单个声明是一种常见的长篇评估技术 [69, 64, 98, 113]。声明可以以不同的细微程度表示，例如，句子层级 [39, 64] 或原子层级（即，传达最小的信息单位） [69, 113]。句子层级的声明可能混合了已支持和未支持的信息 [69]，而严格的原子声明通常缺乏明确验证所需的

上下文。此外，并不是每个声明都是本质上可以证明的，例如，比喻、类比或其他复杂的语言结构。

可验证的断言在句子和原子粒度之间取得平衡；它们足够简洁，可以隔离具体的断言，但上下文丰富性足以进行明确的验证（有关示例，请参见 Fig. 1）。我们使用来自 Song et al. [98] 的基于大型语言模型的自动断言提取器，从长格式输出中获得可验证的断言。

PIC 设置。 A PIC 任务输入包括一个指令 $\mathcal{I}$ 和一组可验证的断言 $C = c_1, \dots, c_n$ 。给定 $(\mathcal{I}, C)$ ，一个语言模型 θ 生成一个长篇回应 $\theta(\mathcal{I}, C)$ ，从中可以提取出一组可验证的断言 $C' = c'_1, \dots, c'_m$ 。为了全面性，我们根据用户意图定义了两种模式，每种模式都有不同的限制。

在完整的 PIC 环境中，语言模型必须在其响应中纳入每一个给定的声明，既不能添加也不能遗漏任何信息。具体来说，每个声明在生成的内容中至少要出现一次，并且没有输入的声明支持的情况下，任何陈述都不能出现在响应中。完整的 PIC 任务示例可能包括可靠的风格转换或开放生成（例如，“写一篇关于埃琳娜·费兰特的简短传记”，给定一组高度精选的声明）。

在部分 PIC 环境中，语言模型可能只结合背景声明中的相关子集，并且仍然不允许引入新的声明。对于如摘要或带检索的问答等任务，不应复制所有输入声明——只应保留最显著的信息（例如，“总结这篇关于华盛顿西雅图经济政策的文章”，或“撰写关于埃莱娜·费兰特（Elena Ferrante）的简短传记”，给定从维基百科条目检索到的一组输入声明）。然而，生成的输出应仍然严格基于提供的背景。因此，部分环境要求 $C' \subseteq C$ （不允许添加，但允许省略）。

为了有意义地评估一个 PIC 任务，我们要求输入的声明集 C 是良构的，满足以下条件：(1) 非空： $C \neq \emptyset$ ，(2) 无重复： $\forall c_i, c_j \in C, i \neq j \Rightarrow c_i \neq c_j$ ，(3) 无矛盾： $\nexists c_i, c_j \in C$ 使得 c_i 与 c_j 矛盾，以及 (4) 任务相关性： C 应包含足够的良构声明以充分满足指令 $\mathcal{I}$ ，虽然不一定是详尽的。

最后一个标准（任务相关性）确保了任务保持简单并且对 LM 可处理，藉由避免可能掩盖预期指令的对抗性陈述。在实践中，我们对 PIC 任务的评估并不直接考虑指令遵循能力。指令和陈述之间的正交关系可以通过专门的指令遵循指标（在 §5.2 中进行探讨）来单独评估，该指标要求 LM 同时满足所提供的陈述并遵守其 PIC 设置所定义的要求。

我们通过从现有指令调优数据集中提取长篇生成样本并将其转换为 PIC 格式来构建 PIC 任务。设 $\mathcal{D}_{IT}$ 由样本 $(\mathcal{I}, p, y)$ 组成，其中 $\mathcal{I}$ 是任务指令， p 是可选的上下文（例如，用于摘要任务的源文档）， y 是长篇的标准回应。对于完整 PIC，我们假设这些任务不包含 p ，因此我们从长篇标准回应 y 中提取 C 。在部分 PIC 中，我们直接从 p 中提取 C 。在任一情况下，我们最终得到 $(\mathcal{I}, C)$ ，其中 C 是格式良好的。

我们首先使用一个声明提取器 [98] 将输出分解为可验证的声明，然后使用基于 LLM 的声明验证器 `support(·, ·)` [98] 来检查语义等价性，从而自动评估生成响应的 PIC 能力。形式上，对于任何候选声明 c 和声明集 S ，令换句话说，这意味着

直观地说，精确度是响应中由上下文论点支持的论点所占的比例，而召回率是上下文论点在响应中被支持的比例。结合两者，我们还可以计算 F_1 。

回想一下，完整设置要求 $C' = C$ （没有遗漏或添加）。因此，我们用 F_1 来评估，惩罚遗漏（召回率）。部分设置要求 $C' \subseteq C$ （不允许添加，但允许遗漏），所以我们只考虑精确度。对于这两种设置，我们还报告了没有内在幻觉的样本比例——分别定义为完美 F_1 和完美精确度的比例。

3 PIC -基准

3.1 评估任务

PIC 适用于任何长文本生成任务。我们设计了 PIC -Bench，一个包含六个完整任务和两个 PIC 任务的基准。² 我们选择了来自先前工作的八个长文本生成数据集，并重新构建其评估

²一些任务，例如传记生成，可以是完整或部分 PIC，这取决于用户意图和输入声明的覆盖范围。虽然从技术上讲，任何任务都可以在两种 PIC 设置下进行评估，但这样做并不总是有意义的。为简单起见，每个 PIC -Bench 任务仅分配到一个 PIC 设置。

以符合我们的问题表述。请注意，我们设计 PIC -Bench 仅评估模型在完成任务时对 PIC 约束的遵循程度，而不是下游任务性能（后者可通过其他方式评估）。Table 1 展示了 PIC -Bench 的任务信息（更多细节在 §9.1 中）。

Table 1: PIC -Bench 数据集统计和示例。对于部分 PIC 设置，我们省略完整的上下文 p 以求简洁。 N 是任务样本的数量， $\bar{C}$ 是输入主张的平均数量。

Task Name	PIC Type	N	$\bar{C}$	Example Instruction $\mathcal{I}$
ENTITYBIOSPIC	Full	183	50.5	Generate a factual biography about Suthida.
POPBIOS-PPIC	Full	111	20.1	Give me a biography on Erwin Schrödinger, the scientist who discovered Quantum Mechanics, Schrödinger's Cat Thought Experiment.
POPBIOS-CFPIC	Full	111	20.1	Give me a biography on Oscar Wilde, the scientist who discovered Quantum Mechanics, Schrödinger's Cat Thought Experiment.
ELI5PIC	Full	146	12.5	Answer the following question(s): why it's common to have 87-octane gasoline in the US but it's almost always 95-octane in Europe?
ASKHISTORIANSPIC	Full	158	19.2	In the original Star Wars: A New Hope, Obi-Wan Kenobi instructs R2-D2 to connect to the Imperial network to gain access to the whole system. Did the concept of an interconnected vast computer network exist in 1977?
EXPERTQAPIC	Full	152	13.5	Answer the question(s): What's the difference between modern and contemporary architecture?
FACTSPIC	Partial	150	63.5	Explain the benefits of using mobile technology to improve healthcare management in both hi-income and low-income countries. { Context p }
XSUMPIC	Partial	200	30.6	Summarize the following text in around 20-25 words. { Context p }

对于完整的 PIC 设置，我们选择那些综合覆盖输入可验证声明对高质量答案起重要作用的任务。ENTITYBIOSPIC [69] 是一个包含大量声明的传记生成任务。PopBios [27] 是一个面向流行实体的简短传记生成数据集，我们从中派生出两个任务：POPBIOS-PPIC 和 POPBIOS-CFPIC。POPBIOS-PPIC 要求模型根据 LM 的参数化知识生成一个知名实体的传记，而 POPBIOS-CFPIC 是一个反事实变体，其中参数化的实体在指令和输入声明中系统地被另一个知名实体替换。

ASKHISTORIANSPIC 和 ELI5PIC [98] 分别是关于历史话题或复杂话题的简化解释的 Reddit 来源长篇 QA 数据集。EXPERTQAPIC [64] 是一个经过专家策划的、特定领域问题的长篇 QA 任务。在完整的 PIC 中，所有声明均从长篇的黄金答案 y 中提取。

对于部分 PIC 设置，我们识别出适合选择性基础的任务——即，仅输出中应该出现相关的上下文声明子集的地方。FACTSPIC [36] 是一个事实基础任务，其中语言模型必须基于长上下文文档制定答案。XSUMPIC [74] 是 BBC 新闻文章的抽象总结任务。原任务要求句子级别的总结；我们通过加长字数限制，将其扩展到长篇形式设置。在部分 PIC 中，声明从提供的上下文 p 中获取。

我们对一系列最先进的、经过指令调整的语言模型进行了基准测试。在开放权重的语言模型中，我们考虑了参数范围为 8B 和 70B 的标准指令调整语言模型：Llama 3.1 8B 指令和 3.3 70B 指令 [28]，Tulu 3 8B 和 70B [46]，Minstral 8B 指令 [72]，以及 Hermes 3 8B 和 70B [104]。我们还考虑了 32B 开放权重的混合思维模型：Qwen 3 32B [102]，QwQ 32B [103]，和 R1-Qwen-32B（从 DeepSeek-R1 蒸馏）[16]。这些模型能够基于思维控制标记进行逐步推理；我们报道使用思维模式的结果，该模式持续较好的表现（非思维模式的结果见 §9.4）。最后，我们将前沿 LLM 视为近似的天花板：GPT-4o [79] 和 Claude-3.5 Sonnet [4]。

4 PIC -LM: 训练后方法

我们提出了 PIC -LM，这是一种具有高 PIC 能力的 8B 指令跟随语言模型。PIC -LM 遵循一个传统的双阶段后训练过程：监督微调 (SFT)，然后是长度归一化的 DPO [84, 46]。我们方法的关键是一种新颖的弱监督偏好数据构造方法，该方法优化了 PIC，同时保留了指令跟随能力。

为了让语言模型学习到一个好的 PIC 初始化，我们首先在一般的指令微调数据集（即 No Robots [86]，FLAN [111]）和特定任务的数据集（即传记生成 [69]，抽象总结 [94]，长文本 QA [36]）上进行 SFT。按照 §2 中的流程，每个样本都被转换为 PIC 格式，生成三元组 $(\mathcal{I}, C, y)$ 。接下来，我们通过仅保留具有高 PIC 评分的样本来过滤数据样本池。我们从 Llama 3.1 8B Instruct [28] 初始化，并使用带有损失掩码和交叉熵损失的教师强制进行 SFT，生成 θ_{SFT} 。

4.1 偏好优化

偏好数据构建。 回忆一下在我们的 SFT 数据集中，每个样本由一个指令 $\mathcal{I}$ 、一组原始上下文声明 C_{orig} 、和一个原始金标准回应 y_{orig} 组成。我们通过随机去掉部分声明来构建每个 C_{orig} 的扰动变体，从而得到一个简化的上下文 $C_{\text{perturb}} \subset C_{\text{orig}}$ 。然后， $\mathcal{I}, C_{\text{perturb}}$ 被传递到已经学习了 PIC 能力的 θ_{SFT} 中，以获得回应 y_{perturb} 。因此，对于每个样本，我们有 $(\mathcal{I}, C_{\text{orig}}, C_{\text{perturb}}, y_{\text{orig}}, y_{\text{perturb}})$ 。

我们的偏好数据集应该由元组 $(\mathcal{I}, C, y^+, y^-)$ 组成，其中 $\mathcal{I}$ 是指令， C 是适当的上下文（选择以匹配首选响应），而 y^+ 和 y^- 分别表示选择和拒绝的响应。每个样本会导致两个有效的 $(\mathcal{I}, C, y^+, y^-)$ 构建：

$$(\mathcal{I}, C, y^+, y^-) = \begin{cases} (\mathcal{I}, C_{\text{orig}}, y_{\text{orig}}, y_{\text{perturb}}), \\ (\mathcal{I}, C_{\text{perturb}}, y_{\text{perturb}}, y_{\text{orig}}). \end{cases}$$

偏好数据抽样策略。 一种选择 $(\mathcal{I}, C, y^+, y^-)$ 的简单方法是用相等的概率对每个构建进行采样，如果目标仅是优化 PIC 能力，这样就足够了。由于 y_{perturb} 通常比 y_{orig} 短，随机采样可能还会减少长度偏差 [81, 85]。但是，这种策略没有考虑到遵循指令能力：如果 C_{perturb} 丢弃了太多重要声明， y_{perturb} 可能无法充分回答 $\mathcal{I}$ ，从而降低指令对齐的质量。相反，并不是所有丢弃的声明都具有同等重要性— y_{perturb} 可能通过更加简洁、集中或风格一致来更好地满足 $\mathcal{I}$ 。这激励了一种策略，根据总体指令兼容性在 y_{orig} 和 y_{perturb} 之间进行自适应选择。

因此，我们从之前的奖励策略 [14] 中获得灵感，提出了一种弱监督的数据采样策略，以一种平衡 PIC 和指令跟随技能的方式构建 $(\mathcal{I}, C, y^+, y^-)$ 。我们使用一个有能力的指令跟随参考模型 θ ，首先定义一个标准化的对数概率下降分数，如下所示：

$$\text{Normalized Log-Probability Drop} = \sigma \left(\frac{\log p_{\theta}(y_{\text{orig}}^{(L)} | \mathcal{I})}{L} - \frac{\log p_{\theta}(y_{\text{perturb}}^{(L)} | \mathcal{I})}{L} \right),$$

是针对生成原始响应与扰动响应的每个令牌的对数概率的归一化差异，计算是在每个序列的最后 L 个令牌上进行的（其中 L 是固定的，例如，20 个令牌），以提高计算效率并减轻长度偏差。我们通过逻辑函数 $(\sigma(z) = \frac{e^z}{1+e^z})$ 将差异归一化到区间 $(0, 1)$ 。

对于每个数据样本，如果归一化后的概率下降分数超过某个超参数 $\tau \in [0, 1]$ ，则我们选择 $(\mathcal{I}, C, y^+, y^-) = (\mathcal{I}, C_{\text{orig}}, y_{\text{orig}}, y_{\text{perturb}})$ ，否则选择 $(\mathcal{I}, C_{\text{perturb}}, y_{\text{perturb}}, y_{\text{orig}})$ 。我们假设每个标记的对数概率是指令遵循能力的合理代理信号，这意味着如果缺乏必要的上下文， θ 生成给定响应的概率会降低。较大的概率下降表明受到扰动的响应明显比原始响应差。实证上，我们为完整和部分设置选择不同的 τ 值，并观察到这比随机采样产生更好的指令遵循性能。³

我们使用长度归一化的直接偏好优化对我们构建的偏好数据进行训练，其中的 DPO 即为带长度归一化 log 概率的直接偏好优化。在 ?? 中，我们进行了一些消融实验：(i) 跳过 SFT 步骤，即直接从 Llama 3.1 8B Instruct 初始化，以及 (ii) 尝试使用替代损失函数进行实验。

5 主要结果

5.1 PIC -测试结果

Table 2 和 Table 3 分别报告了在完整和部分 PIC 设置中的 PIC 基准测试结果，结果中包含了在 §9.2.2 中相应的 95 % 自助法置信区间。下面，我们突出了一些显著的趋势。

按设计，PIC -Bench 任务是 PIC 的理想合成实例，目标很明确：构建仅基于格式良好且可验证的陈述的响应，而不引入额外信息。尽管有这种有意的简单性，甚至最前沿的 LLM 也未能达到接近 100 % 的 PIC 分数。完美的 PIC 分数的平均比例甚至更低，在完整情况下不超过 30 %，而在部分情况下不超过 70 %。

在平均任务表现方面，在两种 PIC 设置中，GPT-4o 始终比开源权重基线高出一个有意义的幅度。Claude 3.5 在完整的 PIC 设置中相对大多数基线表现不佳（主要是由于在 PopBios-CF_{PIC} 上的 F_1 非常低），但在部分设置中可以轻松击败所有开源权重模型。在每个模型家族内（例如，Llama 3, Tulu 3, Hermes 3），较大的模型往往能获得更高的 PIC 分数，因为通常 70B 的语言模型（LMs）比它们的 8B 对手表现更好。不同的开源权重模型家族表现差异显著，这

³我们对几个 τ 的值进行消融，并展示了与 GPT-4.1 中的 ?? 具有高度的 LLM-as-a-judge 相关性。

Table 2: PIC -Bench 结果（完整设置）。指标包括 F_1 ，即完美 F_1 （Perf.）的比例，按任务报告并作为平均值。最优值以粗体显示。所有推理语言模型（用†表示）都在启用思考模式下进行评估。对于所有指标，值越高越好。

	ENTB_{PIC}		POP-B-P_{PIC}		POP-B-CF_{PIC}		AskH_{PIC}		ELI5_{PIC}		ExpQA_{PIC}		Avg.	
	F_1	Perf.	F_1	Perf.	F_1	Perf.	F_1	Perf.	F_1	Perf.	F_1	Perf.	F_1	Perf.
Open-weight LMs (8B)														
LLAMA 3.1 8B INST.	74.4	0.6	83.5	2.7	23.7	0.0	81.9	5.1	77.8	8.9	73.5	5.3	69.1	3.7
TULU 3 8B	79.6	1.6	87.8	2.7	51.3	0.0	80.4	3.2	81.8	10.3	80.7	11.8	76.9	4.9
MINISTRAL 8B INST.	78.6	0.6	89.0	9.0	63.5	0.0	76.8	5.1	78.1	11.0	77.6	15.1	77.3	6.8
HERMES 3 8B	74.9	0.6	84.9	6.3	44.6	0.0	72.9	2.5	75.9	4.1	74.1	7.9	71.2	3.6
Open-weight LMs (32B)														
QWEN 3 32B†	74.1	2.7	88.2	18.9	48.2	0.9	77.5	7.0	67.7	6.9	63.6	7.2	69.9	7.3
QwQ 32B†	86.7	1.6	87.6	2.7	64.1	1.8	79.6	3.8	80.0	6.9	80.9	12.5	70.3	7.6
R1-QWEN-32B†	82.3	1.1	87.5	9.0	58.8	0.9	81.1	4.4	82.2	10.3	80.0	13.2	78.6	6.5
Open-weight LMs (70B)														
LLAMA 3.3 70B INST.	89.1	2.2	93.4	17.1	45.5	0.0	91.2	12.7	90.3	19.9	85.6	27.0	82.5	13.1
TULU 3 70B	84.8	7.1	89.2	23.4	52.0	0.0	83.7	15.8	82.5	19.2	73.4	17.8	77.6	13.8
HERMES 3 70B	81.4	1.6	89.4	12.6	56.6	0.0	81.6	5.1	79.7	9.6	79.8	14.5	78.1	7.2
Proprietary LMs														
CLAUDE 3.5	88.8	6.6	93.1	22.5	1.0	0.0	83.7	18.4	85.8	29.5	63.8	21.1	69.4	16.3
GPT-4o	91.9	7.7	94.0	27.0	74.6	0.9	96.1	37.3	93.9	47.3	92.7	47.4	90.5	27.9
Ours: PIC-LM (8B) — from LLAMA 3.1 8B INST.														
PIC-LM _{SFT ONLY}	86.7	4.9	87.2	7.2	71.7	1.8	86.6	15.2	82.1	19.2	77.3	21.7	81.9	11.7
PIC-LM	94.1	8.2	96.0	36.0	81.0	3.6	95.1	32.3	94.1	52.7	85.6	34.9	91.0	28.0

表明 PIC 能力不仅仅由模型大小决定。尽管它们的大小较大，32B 推理模型取得的 PIC 分数与 8B 非推理模型相当，这表明仅仅增加显式推理能力并不能在 PIC 任务上带来实质性的提升。

语言模型在反事实 PIC 情境中表现不佳。PopBios-CF_{PIC} 是最具挑战性的任务模式，完美的 F_1 得分处于个位数范围。在开放式基线中，Llama 3.1 8B Instruct 表现最差，平均 F_1 为 23.7%，而在封闭模型中，Claude 3.5 的 F_1 仅为 1%。值得注意的是，PopBios-CF_{PIC} 与 PopBios-P_{PIC} 共享其结构——这是最简单的 PIC 任务之一——但在其中替换了反事实实体，这些实体与语言模型的内部知识明显冲突，模拟了极端知识冲突。这一发现证实了 Wu et al. [114] 的观点，指出提供的背景信息越不真实，语言模型就越不可能基于此背景信息。在 Claude 3.5 的情况下，我们注意到了一种一致的回避行为，LLM 拒绝生成不真实的信息，并且会覆盖用户规范。⁴

PIC -LM 显示了显著的改进。 我们的 8B PIC -LM 在两种评估设置中均优于所有开源基线。在完整设置中，PIC -LM 实现了 91.0% 的平均 F_1 和 28.0% 的完美 F_1 比例，而最佳开源模型（Llama 3.3 70B Instruct）则分别为 82.5% 和 13.1%。在部分设置中，PIC -LM 达到 93.3% 的平均精度和 52.9% 的完美精度比例，超过了最强的开源基线（Hermes 3 70B），后者分别为 82.5% 和 43.9%。即便是仅有 SFT 的消融实验 PIC-LM_{SFT Only}，没有进行 PIC 特定的偏好优化，也显著优于 Llama 3.1 8B Instruct，表明仅通过对 PIC 格式化的指令数据进行微调就能提升 PIC 性能。仅使用一个开源 8B LM（并从表现最差的基线开始），PIC -LM 缩小了与 GPT-4o 的差距，取得了更好的平均 F_1 和精度（分别为 91.0% 比 90.5% 和 93.3% 比 90.2%）。就完美得分的比例而言，PIC -LM 在完整 PIC 情况下表现具有竞争力（28.0% 比 27.9%），但在部分设置中落后于 GPT-4o 和 Claude 3.5（52.9% 比 67.7% 和 65.2%）。尽管如此，PIC -LM 在每个指标上均以较大优势超越所有开源 LMs，表明在一个 8B 开源模型上应用简单的后训练策略可以在不需要封闭源模型或更大参数预算的情况下实现接近最先进的 PIC 性能。

⁴克劳德的 RLHF 调整为“有帮助、无害和诚实”的原则，可能导致其避免产生极其错误的输出 [3]。

Table 3: PIC -Bench 结果（部分设置）。度量标准包括精度（Prec.）、完美精度的比例（Perf.），按任务和平均值报告。最佳值用粗体显示。所有推理 LM（用†表示）均在启用思考模式下进行评估。对所有指标而言，值越高越好。

	FACTSPIC		XSUMPIC		Avg	Avg
	Prec.	Perf.	Prec.	Perf.	Prec.	Perf.
Open-weight LMs (8B)						
LLAMA 3.1 8B INST.	70.0	19.3	77.2	54.0	73.6	36.7
TULU 3 8B	75.3	18.7	76.9	50.5	76.1	34.6
MINISTRAL 8B INST.	75.4	30.0	82.7	63.0	79.1	46.5
HERMES 3 8B	74.1	27.3	81.2	42.0	77.7	34.7
Open-weight LMs (32B)						
QWEN 3 32B†	64.5	17.3	78.5	46.0	71.5	31.7
QwQ 32B†	70.3	12.7	86.9	58.5	78.6	35.6
R1-QWEN-32B†	71.1	26.7	79.6	53.5	75.3	40.1
Open-weight LMs (70B)						
LLAMA 3.3 70B INST.	79.4	32.0	69.3	67.0	74.3	49.5
TULU 3 70B	67.7	14.7	85.5	60.0	76.7	37.3
HERMES 3 70B	78.0	25.3	87.0	62.5	82.5	43.9
Proprietary LMs						
CLAUDE 3.5	82.0	53.3	90.4	77.0	86.2	65.2
GPT-4o	87.4	51.3	93.0	84.0	90.2	67.7
Ours: PIC-LM (8B) — from LLAMA 3.1 8B INST.						
PIC-LM _{SFT ONLY}	81.2	46.0	94.0	69.5	87.6	57.8
PIC-LM	89.8	49.3	96.8	49.3	93.3	52.9

5.2 PIC -LM 指令跟随分析

一个理想的可控语言模型应当能够生成响应，以有用的方式回答指令，同时在需要时完全基于来源上下文。与 PIC 正交，我们探讨 PIC -LM 在每个 PIC 任务上跟随指令的效果。遵循 Asai et al. [5]，我们使用 Prometheus 2，一个开源的评估语言模型，它基于一个自定义的指令跟随标准从 1 到 5 进行评分（定义在 Table 23 中），来对长篇输出进行评分。

当 LM 通过逐字复述输入的断言 C 来破解指标时，会发生 PIC 的退化情况，这会导致一个虽然得分完美但不令人满意的响应。作为理智性检查，我们引入了 CONCATCLAIMS，这是一种简单地将输入断言连接成一个响应的下限设置。即便 PIC 完美，CONCATCLAIMS 在 Prometheus 上的表现应该很差，因为断言连接忽视了任务特定的要求（例如，词长、风格）。

Table 4 显示了每个 PIC 任务的 Prometheus 结果（平均值 $\pm$ 标准误）。由于未能满足任务指令，CONCATCLAIMS 得分较低（平均 2.01）。相比之下，Llama 3.1 8B Instruct 在指令遵循方面显著更好（3.75）；作为一个指令微调模型，它能胜任地遵循任务要求。与 Llama 3.1 8B Instruct 相比，PIC-LM_{SFT ONLY} 在大多数任务上略有下降，但在反事实任务 PopB-CF_{PIC} 和 XSUMPIC 上有所提升，使其平均表现达到 3.87。PIC-LM 基本上恢复了其指令遵循能力，除了在 XSUMPIC 上。在质量方面，我们观察到它在遵循严格的词数限制（20-25 个词）时有些困难。然而，PIC -LM 拥有最高的平均 Prometheus 得分为 3.92。总体而言，这些结果表明我们的训练框架仅有最小的侵入性；PIC -LM 保留了源模型的指令遵循能力，同时严格提高 PIC 表现。

Table 4: 使用 Prometheus 进行的指令跟随评估，结果以任务平均值和标准误差表示。CONCATCLAIMS 是退化的下限（Low）。最佳值用粗体标出。

	ENTB-PIC	POPB-PPIC	POPB-CFPIC	ASKHPIC	ELI5PIC	EXPQAPIC	FACTSPIC	XSUMPIC	Avg.
CONCATCLAIMS (LOW)	2.26 $\pm$.10	2.77 $\pm$.14	2.31 $\pm$.12	2.21 $\pm$.11	1.97 $\pm$.11	1.76 $\pm$.09	1.71 $\pm$.10	1.06 $\pm$.02	2.01
LLAMA 3.1 8B INST.	4.22 $\pm$.05	4.44 $\pm$.05	2.01 $\pm$.12	4.07 $\pm$.06	4.14 $\pm$.06	4.11 $\pm$.06	3.71 $\pm$.07	3.28 $\pm$.08	3.75
Ours: PIC-LM — from LLAMA 3.1 8B INST.									
PIC-LM _{SFT ONLY}	3.79 $\pm$.04	4.10 $\pm$.04	3.69 $\pm$.07	3.99 $\pm$.04	4.01 $\pm$.04	3.93 $\pm$.05	3.87 $\pm$.06	3.57 $\pm$.05	3.87
PIC-LM	3.84 $\pm$.04	4.18 $\pm$.04	3.79 $\pm$.07	4.21 $\pm$.04	4.09 $\pm$.04	4.03 $\pm$.04	4.01 $\pm$.05	3.23 $\pm$.06	3.92

6 PIC-LM 的实际应用

一个对 PIC-Bench 的主要假设是它依赖于格式良好的输入声明。然而，在现实场景中识别适当的声明会面临两个挑战：(1) 声明可能形式不佳，(2) 用户可能不知道在期望的回应中应包括哪些声明（例如，用于事实查找任务）。我们通过演示 PIC-LM 在多组件系统中时提高长篇问答任务的事实性，缩小了我们控制的基准实验与实际应用之间的差距。首先，通过检索增强生成 (RAG)，PIC-LM 能够从检索的上下文中识别相关声明，以生成更准确的答案 (??)。其次，我们实现了一种自我验证管道，其中一个 LM 从头开始生成并验证其自己的声明；在最终生成步骤中替换 PIC-LM 提升了最终任务的事实准确性 (§6.1)。

首先，作为一个下界基线，我们与不带上下文的 Llama 3.1 8B Instruct 进行比较（无上下文（低））。我们还与两种类型的大小匹配（8B）基线进行比较：带上下文的基线：(i) 零样本提示基线，包括 Llama 3.1 8B Instruct、Tulu 3 8B、Ministral 8B Instruct 和 Hermes 3 8B，以及 (ii) 上下文感知基线，使上下文依赖性更强。第一个是 CAD（上下文感知解码）[95]，一种解码技术，通过对比地排除 LM 的先验知识来强调对输入上下文的注意力；我们将这一技术应用于 Llama 3.1 8B Instruct。第二个是 SelfCite 8B [14]，一个具有强属性的 LM，能够生成带有上下文属性（或引用）的响应，并且经过使用上下文消融的自监督奖励的后训练。由于我们在这项工作中不考虑属性质量，因此在评估之前剔除引用标记。

在检索增强生成 (RAG) 中，输入查询会被补充相关的外部文本，以指导语言模型的生成。最近的研究表明，当检索到的上下文与语言模型的内部知识相冲突时，即使正确答案在上下文中，语言模型也无法有效利用检索上下文。与 PIC-Bench 施加的严格输入条件不同，检索到的上下文可能包含重复、不相关或噪声的信息。PIC-LM 能否推广到这样的实例？我们在 ASQA 数据集中进行测试，该数据集包括 948 个长篇问答对。在 ASQA 中，每个问题都未完全明确，需要多个短篇答案来彻底消除歧义（例如，“1830 年法国的统治者是谁？”是模糊的，并有多个回答）。

依据 Gao et al. [23]，我们报告了 EM（准确匹配召回），其中采用从英文维基百科检索到的前五个段落，并将其分解为可验证的陈述，作为输入上下文。我们报告了一个涉及所有样本的标准设置，以及一个仅考虑检索到的上下文中至少包含一个正确答案的样本的理想设置。⁵

Table 5 显示了 RAG 在 ASQA 上的表现。除 Mistral 8B Inst. 外，所有基于上下文的设置都优于无上下文基线（30.9 % EM）。此外，PIC-LM 在两种模式下均取得了统计上显著的提升：在标准模式下，其得分为 61.5 % EM (± 2.1)，相比之下，CAD 作为最强的上下文感知基线得分为 56.2 % EM (± 2.1)。

Table 5: RAG 在标准（948 个样本）和 oracle（885 个样本）模式下 ASQA 的性能，报告为平均 EM $\pm$ 95 % 自举置信区间。最佳值以黑体显示。

Setting	Standard EM	Oracle EM
No CONTEXT (Low)	30.9 $\pm$ 2.2	N/A
LLAMA 3.1 8B INST.	52.5 $\pm$ 2.1	56.2 $\pm$ 2.0
TULU 3 8B	42.7 $\pm$ 2.0	45.6 $\pm$ 2.0
MINISTRAL 8B INST.	31.8 $\pm$ 1.7	34.0 $\pm$ 1.8
HERMES 3 8B	52.8 $\pm$ 2.0	56.4 $\pm$ 2.0
CAD (on LLAMA 3.1 8B INST.)	56.2 $\pm$ 2.1	60.1 $\pm$ 2.0
SELF-CITE 8B	52.2 $\pm$ 2.1	52.5 $\pm$ 2.1
PIC-LM 8B	61.5 $\pm$ 2.1	65.9 $\pm$ 2.0

6.1 自我验证管道

PIC-LM 即使在未提供相关外部声明的情况下仍然有效。受到链式验证 [17] 和自一致性采样 [11, 108] 的启发，我们描述了 PIC-LM 如何在自动生成和验证声明的流程系统中改善最终任务的真实性，而无需依赖知识增强（例如，RAG 或搜索工具）。该方法利用了这样一个观察，即 LMs 通常在验证方面比生成更可靠 [20, 54]，特别是在事实评估方面 [29]。通过利用这种不对称性，我们可以有效地自我纠正诸如外部幻觉 [13, 82, 17, 40] 等问题行为。

⁵我们在 §11.2.1 中探索了 RAG 实验设置的变化。

给定一个输入指令, LM: (1) 生成一个草稿回复; (2) 基于这个草稿输出制定验证问题; (3) 使用自一致性采样和多数表决的方法独立核查每个问题, 生成一组唯一的已验证声明; (4) 仅使用符合原始指令的已验证声明生成最终回复。

我们在两个真实性任务上部署我们的流程。首先, 按照 Dhuliawala et al. [17], 我们评估一个出生地事实任务, 格式为: “Name some { occupation } s born in { location } ”。我们使用平均声明精度来报告事实准确度, 其中最终响应被分解为一组可验证的声明, 并且每个声明使用 Google 搜索进行事实核对 [98, 113]。其次, 我们在 QAMParI [2] 上进行评估, 这是一个答案分布在不同来源上的开放域问答任务 (例如: “哪本书有 Pauline Baynes 的插图?”)。按照 Gao et al. [23], 我们在 QAMParI 上报告精度、召回率 @ k 和 F_1 @ k , 其中 $k = 5$ 是用于召回率计算的最大正确答案数量。对于这两个任务, 我们使用 Llama 3.3 70B Instruct 作为步骤 1-3 的语言模型, 我们发现其足够大以进行准确的声明自验证。在步骤 4, 最终生成步骤中, 声明被转换为长形式响应, 我们可以替换为 PIC -LM 和不同的基线方法。

Table 6 显示了流水线结果。在出生地任务中, PIC -LM 表现优于所有 8B 基线, 并略微超过更大的 Llama 3.3 70B 指令, 证明了更好 PIC 的效果。在 QAMParI 上, 所有提供上下文的基线在 F_1 @5 方面表现相似, 而精确性有较大的差异 (Hermes 3 8B 在 9.9 % 时最低, Ministral 8B 在 27.2 % 时最高), 召回率 @5 (Ministral 8B 在 14.0 时最低, Llama 3.3 70B 在 26.8 % 时最高)。PIC -LM 8B 在精度和召回率之间 @5 取得了最佳平衡, 最佳 F_1 @5 为 22.6。

Table 6: 在 BIRTHPLACE 和 QAMParI 任务上的平均表现 (含 95 % 引导 CI)。No CONTEXT 设置作为下限。最佳值用粗体显示。

	BIRTHPLACE	QAMParI		
Setting	Factual Prec.	Prec.	Rec.@5	F_1 @5
No CONTEXT (Low)	19.4 $\pm$ 2.6	6.1 $\pm$ 0.9	9.9 $\pm$ 1.4	6.5 $\pm$ 0.9
LLAMA 3.1 8B INST.	65.9 $\pm$ 4.5	11.6 $\pm$ 1.2	21.4 $\pm$ 2.0	13.5 $\pm$ 1.3
TULU 3 8B	80.3 $\pm$ 4.0	16.8 $\pm$ 1.5	25.5 $\pm$ 2.2	18.4 $\pm$ 1.5
MINISTRAL 8B	83.7 $\pm$ 4.1	27.2 $\pm$ 2.6	14.0 $\pm$ 1.6	15.3 $\pm$ 1.5
HERMES 3 8B	53.4 $\pm$ 4.2	9.9 $\pm$ 1.1	20.3 $\pm$ 2.1	12.1 $\pm$ 1.2
CAD (on LLAMA 3.1 8B INST.)	63.6 $\pm$ 5.4	15.7 $\pm$ 1.7	23.8 $\pm$ 2.4	17.5 $\pm$ 1.8
SELF CITE 8B	69.4 $\pm$ 4.5	17.7 $\pm$ 1.5	23.3 $\pm$ 2.1	18.6 $\pm$ 1.6
LLAMA 3.3 70B INST.	84.7 $\pm$ 4.0	14.2 $\pm$ 1.3	26.8 $\pm$ 2.2	16.7 $\pm$ 1.4
PIC-LM 8B	86.0 $\pm$ 4.0	25.5 $\pm$ 2.1	25.6 $\pm$ 2.2	22.6 $\pm$ 1.9

幻觉是标准语言模型目标的副产品: 语言模型从训练语料库中提取统计模式, 并学习生成语义上流畅但不一定事实准确的文本。最近大量的学术研究集中在事实性和外在幻觉上, 特别是其根本原因识别、检测、评估和缓解。早期的内在幻觉研究主要是任务特定的, 涉及抽象摘要或基于知识的对话生成。

PIC -Bench 通过断言分解来评估长文本生成中的内在幻觉, 这是在已有的关于外在幻觉的长文本评估的基础上进行的 [69, 98, 113]。在内在幻觉基准中, RoSE [58] 考虑在原子内容单元级别的生成以进行总结评估。FaithEval [70] 针对短式问答进行操作, 而我们考虑了一系列长文本任务。与我们的工作最接近的是 FACTS Grounding [36], 我们将其作为部分任务纳入 PIC -Bench。他们的评估依赖于通过多位评审 LLM 的提示进行粗略的整体评估, 而不是细致的逐项核实。此外, PIC 通过要求语言模型在完整的 PIC 设置中将所有生成内容与提供的上下文联系起来, 而不是使用二元忠实性概念, 从而施加了更严格的标准。

除去评估, 当前有许多通过后训练减轻语言模型幻觉的努力。PIC -LM 遵循了一种传统的对齐程序, 类似于最近的事实性方法 [33, 56, 106], 这些方法使用基于 API 的自动评估工具来构建偏好数据 [69, 113]。相比之下, PIC -LM 利用了一种更具成本效益的奖励信号, 目标是内在的幻觉。几项研究通过其他方式改善上下文忠诚性, 无论是通过提示 [121]、解码修改 [95] 还是在反事实数据上进行微调 [59, 49, 76]。PIC -LM 并没有刻意在反事实数据上进行训练, 但在反事实评估中仍表现良好。

可归属生成 语言模型归因通过使用内嵌引用标记将生成文本中的每个陈述归属到适当的来源引用。在可归因的生成中, 引用精确度 (即每个引用应支持相关声明) 和召回率 (即每个生成的声明应映射到适当的引用) 都是重要的指标, 以可追溯性、忠实性和可验证性为核心原则。虽然部分 PIC 类似于归因生成, 但没有显式引用标记, 使用的是离散的声明, 而我

们完整的 PIC 设置则不同，需要结合所有输入的声明，强调严格遵循用户提供的内容，而不是选择性引用。

语言模型中的可控性。 能控性是一个源于控制理论的基本概念。它描述了一个系统通过适当的外部输入从任何初始状态过渡到任何最终状态的能力 [78, 49]。这一特性对语言模型 (LMs) 是非常宝贵的，因为许多以用户为中心的应用程序需要对生成输出进行特定的控制 [120]。例如，对话生成任务可能涉及对情感、形式化程度或毒性的偏好级别。早期引导语言模型生成的尝试包括使用离散控制码 [42, 88] 来调整风格、内容和其他任务特定的行为，或者在解码时修改输出对数分布 [117]。

近年来，指令微调已成为将语言模型与用户意图对齐的实际范式 [71, 80, 91, 109, 111, 60]。语言模型必须学习将自然语言指令隐式映射到期望的输出行为，而不是依赖特殊标记来触发动作。虽然指令微调扩大了约束空间的表达能力，但也助长了更大的模糊性。即使有明确的线索，语言模型可能也难以遵循命令。例如，Goyal et al. [27] 发现对指令数据集的监督微调削弱语言模型的上下文依赖能力。他们从理论上表明，在指令微调期间，非上下文依赖的数据点最终会主导损失梯度。我们的研究证实了这个问题是数据驱动的：通过对需要严格符合声明的指令对进行微调，PIC-LM 保持了强大的上下文基础。此外，DPO 和相关的偏好优化算法更明确地强化了这种行为，确保了超越单独监督微调所能达到的性能差距。

7 讨论

控制是可靠 LM 生成的基石；它是用户信任和系统可靠性的基础，从日常信息搜索 [23, 57] 到高风险的政治、法律和医疗应用 [15, 52, 51]。在这项工作中，我们提出了 PIC 框架，以研究在可验证声明层次上长文本生成的信息的精确控制。我们的贡献有三个方面：我们提出一个简单但具有挑战性的基准、一种初始训练方法和模型，以及激励性使用案例。

PIC 范式带来两个实际的好处。首先，PIC 通过在用户提供的上下文中分配最高忠诚度而非参数知识，确保了可控性。这种设计使得修订或更新信息的吸收变得容易。第二个优点是可解释性：由于所有生成的声明必须以输入空间为基础，形成一个封闭的反馈循环，用户可以将任何输出声明追溯到其源声明集，从而使验证变得可行。特别地，PIC 属于一类问题，这类问题对于 LM 生成来说是困难的，但通过手动或自动手段进行验证则是简单的 [54]。

未来的研究可能会探索如何在持续预训练等早期阶段呈现 PIC 行为，这可以通过控制序列的学习提示 [24]、数据样本的战略重排序 [96] 或程序化的合成数据生成 [12, 63, 118] 来实现。一个未解的问题是，是否 PIC-LM 可以适应更复杂形式的控制；例如，自适应地在上下文知识和参数化知识之间切换。我们还通过展示 PIC 的实用性，从理论走向实践，这可以扩展到更复杂的自主设置，例如工具使用 [92] 或多-LLM 协作 [22]。

8 局限性

一个一致的核心思想是用户提供的知识应当优先于语言模型自身的先验。用户应该在构建语言模型生成框架的过程中拥有完全的控制权，这是一个极端立场，可能并不适用于所有下游场景。

PIC-Bench 也施加了严格的输入假设——例如，输入声明集应该是结构良好的，并且与构成理想反应的内容完全一致。在一个简单且定义良好的环境中，语言模型无法完全消除内在幻觉是重要的。此外，我们提供了案例研究，说明 PIC 在更大流程中如何提高事实准确性，其中上下文是在事先获取、验证或筛选的。尽管其独立实用性有限，PIC-LM 显示出作为模块化系统中可靠组件的潜力。

鉴于通过索赔分解来评估每个生成样本所需的时间和成本，我们的评估集单独来说都非常小（每个样本少于 250 个）。这在这一领域是常态：我们的测试床规模与用于长篇事实性文献中的规模相当 [69, 56, 98, 106]。与之前的工作不同，我们的任务选择超出了问答范围；尽管全面，但绝不是详尽的。根据 Jacovi et al. [36]，我们不考虑多步逻辑、密集推理或创意生成。

最后，我们的评估使用了基于语言模型的自动化指标，具体来说，PIC 使用的是 GPT-4o mini，而能力评估则使用从 Mistral 8x7B 微调而来的 Prometheus。与 OpenAI 的旗舰产品 GPT-4o 相比，GPT-4o mini 可能会由于在声明提取或验证步骤中的错误累积而加剧指标误判。同样，Prometheus 虽然与人工评估相关性良好，但可能无法在跨领域任务中有效泛化 [34, 73]。

我们感谢 Rui Xin 在 OpenAI API 评估中的帮助，以及 Noah Smith、Tianyu Gao、Lucy He、Adithya Bhaskar 和普林斯顿 NLP 团队给予的有益反馈。JH 得到了 NSF 研究生研究奖学金的支持。这项工作还得到了新加坡国家研究基金会和新加坡数字发展与信息部国家人工智能小组在人工智能访问教授计划（奖项编号 AIVP-2024-001）的支持，以及 Schmidt Sciences 的 AI2050 项目的支持。

References

- [1] Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, 12:681–699, 2024. doi: 10.1162/tacl_a_00667. URL <https://aclanthology.org/2024.tacl-1.38/>.
- [2] Samuel Amouyal, Tomer Wolfson, Ohad Rubin, Ori Yoran, Jonathan Herzig, and Jonathan Berant. QAMPARI: A benchmark for open-domain questions with many answers. In Sebastian Gehrmann, Alex Wang, João Sedoc, Elizabeth Clark, Kaustubh Dhole, Khyathi Raghavi Chandu, Enrico Santus, and Hooman Sedghamiz (eds.), *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pp. 97–110, Singapore, December 2023. Association for Computational Linguistics. URL <https://aclanthology.org/2023.gem-1.9/>.
- [3] Anthropic. Claude’s constitution, 2023. URL <https://www.anthropic.com/news/claudes-constitution>. Accessed: 2025-04-29.
- [4] Anthropic. The claude 3 model family: Opus, sonnet, haiku. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf, 2024.
- [5] Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D’arcy, David Wadden, Matt Latzke, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, Graham Neubig, Dan Weld, Doug Downey, Wen tau Yih, Pang Wei Koh, and Hannaneh Hajishirzi. Openscholar: Synthesizing scientific literature with retrieval-augmented lms, 2024. URL <https://arxiv.org/abs/2411.14199>.
- [6] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=hSyW5go0v8>.
- [7] Akari Asai, Zexuan Zhong, Danqi Chen, Pang Wei Koh, Luke Zettlemoyer, Hannaneh Hajishirzi, and Wen tau Yih. Reliable, adaptable, and attributable language models with retrieval, 2024. URL <https://arxiv.org/abs/2403.03187>.
- [8] Neil Band, Xuechen Li, Tengyu Ma, and Tatsunori Hashimoto. Linguistic calibration of long-form generations. In *Forty-first International Conference on Machine Learning*, 2024. URL <https://openreview.net/forum?id=rJVjQSQ8ye>.
- [9] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=ETKGuby0hcs>.
- [10] Shiqi Chen, Yiran Zhao, Jinghan Zhang, I-Chun Chern, Siyang Gao, Pengfei Liu, and Junxian He. Felm: Benchmarking factuality evaluation of large language models, 2023. URL <https://arxiv.org/abs/2310.00741>.
- [11] Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. Universal self-consistency for large language model generation, 2023. URL <https://arxiv.org/abs/2311.17311>.

- [12] Daixuan Cheng, Yuxian Gu, Shaohan Huang, Junyu Bi, Minlie Huang, and Furu Wei. Instruction pre-training: Language models are supervised multitask learners. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2529–2550, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.148. URL <https://aclanthology.org/2024.emnlp-main.148/>.
- [13] I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. Factool: Factuality detection in generative ai – a tool augmented framework for multi-task and multi-domain scenarios, 2023. URL <https://arxiv.org/abs/2307.13528>.
- [14] Yung-Sung Chuang, Benjamin Cohen-Wang, Shannon Zejiang Shen, Zhaofeng Wu, Hu Xu, Xi Victoria Lin, James Glass, Shang-Wen Li, and Wen tau Yih. Selfcite: Self-supervised alignment for context attribution in large language models, 2025. URL <https://arxiv.org/abs/2502.09604>.
- [15] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E. Ho. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models, 2024.
- [16] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL <https://arxiv.org/abs/2501.12948>.
- [17] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3563–3578, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.212. URL <https://aclanthology.org/2024.findings-acl.212/>.
- [18] Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. On the origin of hallucinations in conversational models: Is it the datasets or the models? In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5271–5285, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.387. URL <https://aclanthology.org/2022.naacl-main.387/>.
- [19] Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D’Amour, DJ Dvijotham, Adam Fisch, Katherine Heller, Stephen Pfohl, Deepak Ramachandran, Peter Shaw, and Jonathan Berant. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking, 2024. URL <https://arxiv.org/abs/2312.09244>.
- [20] Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021. doi: 10.1162/tacl_a_00410. URL <https://aclanthology.org/2021.tacl-1.60/>.
- [21] Alexander R. Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. Qafacteval: Improved qa-based factual consistency evaluation for summarization, 2022. URL <https://arxiv.org/abs/2112.08542>.
- [22] Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. Don’t hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14664–14690, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.786. URL <https://aclanthology.org/2024.acl-long.786/>.
- [23] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

- [24] Tianyu Gao, Alexander Wettig, Luxi He, Yihe Dong, Sadhika Malladi, and Danqi Chen. Meta-data conditioning accelerates language model pre-training, 2025. URL <https://arxiv.org/abs/2501.01956>.
- [25] Zorik Gekhman, Gal Yona, Roei Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning LLMs on new knowledge encourage hallucinations? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7765–7784, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.444. URL <https://aclanthology.org/2024.emnlp-main.444/>.
- [26] Zorik Gekhman, Eyal Ben David, Hadas Orgad, Eran Ofek, Yonatan Belinkov, Idan Szpektor, Jonathan Herzig, and Roi Reichart. Inside-out: Hidden factual knowledge in llms, 2025. URL <https://arxiv.org/abs/2503.15299>.
- [27] Sachin Goyal, Christina Baek, J. Zico Kolter, and Aditi Raghunathan. Context-parametric inversion: Why instruction finetuning may not actually improve context reliance, 2024. URL <https://arxiv.org/abs/2410.10796>.
- [28] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Es-
iobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hup-
kes, Egor Lomakin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucu-
rell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta
Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan
Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer
Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang,
Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua
Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak,
Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuen-
ley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten,
Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas
Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat
Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita,
Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan,
Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning
Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar
Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura,
Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer,
Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Gird-
har, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan
Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean
Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Rapparthi,
Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Ba-
tra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky,
Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias
Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta,
Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vitor Al-
biero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyan Fu, Whitney Meers, Xavier Mar-
tinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao
Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwon Song,
Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing

Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkan Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippas Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelen, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martinus Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.

- [29] Jian Guan, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. Language models hallucinate, but may excel at fact verification. In Kevin Duh, Helena Gomez, and Steven

- Bethard (eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) , pp. 1090–1111, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.62. URL <https://aclanthology.org/2024.naacl-long.62/>.
- [30] Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. Accelerate: Training and inference at scale made simple, efficient and adaptable. <https://github.com/huggingface/accelerate>, 2022.
 - [31] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In Hal Daumé III and Aarti Singh (eds.), Proceedings of the 37th International Conference on Machine Learning , volume 119 of Proceedings of Machine Learning Research , pp. 3929–3938. PMLR, 13–18 Jul 2020. URL <https://proceedings.mlr.press/v119/guu20a.html>.
 - [32] Or Honovich, Leshem Choshen, Roei Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. q^2 : Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing , pp. 7856–7870, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.619. URL <https://aclanthology.org/2021.emnlp-main.619/>.
 - [33] Chao-Wei Huang and Yun-Nung Chen. FactAlign: Long-form factuality alignment of large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), Findings of the Association for Computational Linguistics: EMNLP 2024 , pp. 16363–16375, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.955. URL <https://aclanthology.org/2024.findings-emnlp.955/>.
 - [34] Hui Huang, Yingqi Qu, Xingyuan Bu, Hongli Zhou, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge model is not a general substitute for gpt-4, 2024. URL <https://arxiv.org/abs/2403.02839>.
 - [35] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Trans. Inf. Syst. , 43(2), January 2025. ISSN 1046-8188. doi: 10.1145/3703155. URL <https://doi.org/10.1145/3703155>.
 - [36] Alon Jacovi, Andrew Wang, Chris Alberti, Connie Tao, Jon Lipovetz, Kate Olszewska, Lukas Haas, Michelle Liu, Nate Keating, Adam Bloniarz, Carl Saroufim, Corey Fry, Dror Marcus, Doron Kukliansky, Gaurav Singh Tomar, James Swirhun, Jinwei Xing, Lily Wang, Madhu Gurumurthy, Michael Aaron, Moran Ambar, Rachana Fellinger, Rui Wang, Zizhao Zhang, Sasha Goldshtein, and Dipanjan Das. The facts grounding leaderboard: Benchmarking llms’ ability to ground responses to long-form input, 2025. URL <https://arxiv.org/abs/2501.03200>.
 - [37] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. ACM Comput. Surv. , 55(12), March 2023. ISSN 0360-0300. doi: 10.1145/3571730. URL <https://doi.org/10.1145/3571730>.
 - [38] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL <https://arxiv.org/abs/2207.05221>.

- [39] Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. WiCE: Real-world entailment for claims in Wikipedia. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7561–7583, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.470. URL <https://aclanthology.org/2023.emnlp-main.470/>.
- [40] Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. When can LLMs actually correct their own mistakes? a critical survey of self-correction of LLMs. *Transactions of the Association for Computational Linguistics*, 12:1417–1440, 2024. doi: 10.1162/tacl_a_00713. URL <https://aclanthology.org/2024.tacl-1.78/>.
- [41] Katie Kang, Eric Wallace, Claire Tomlin, Aviral Kumar, and Sergey Levine. Unfamiliar finetuning examples control how language models hallucinate. In *Automated Reinforcement Learning: Exploring Meta-Learning, AutoML, and LLMs*, 2024. URL <https://openreview.net/forum?id=5H5IQuTlMz>.
- [42] Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. Ctrl: A conditional transformer language model for controllable generation, 2019. URL <https://arxiv.org/abs/1909.05858>.
- [43] Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models, 2024. URL <https://arxiv.org/abs/2310.08491>.
- [44] Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. Bioasq-qa: A manually curated corpus for biomedical question answering. *Scientific Data*, 10:170, 2023. URL <https://doi.org/10.1038/s41597-023-02068-4>.
- [45] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.
- [46] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Han-naneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training, 2024. URL <https://arxiv.org/abs/2411.15124>.
- [47] Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. Factuality enhanced language models for open-ended text generation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL <https://openreview.net/forum?id=LvyJX20R11>.
- [48] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.
- [49] Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. Large language models with controllable working memory. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1774–1793, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.112. URL <https://aclanthology.org/2023.findings-acl.112/>.

- [50] Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. The dawn after the dark: An empirical study on factuality hallucination in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10879–10899, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.586. URL <https://aclanthology.org/2024.acl-long.586/>.
- [51] Lincan Li, Jiaqi Li, Catherine Chen, Fred Gui, Hongjia Yang, Chenxiao Yu, Zhengguang Wang, Jianing Cai, Junlong Aaron Zhou, Bolin Shen, Alex Qian, Weixin Chen, Zhongkai Xue, Lichao Sun, Lifang He, Hanjie Chen, Kaize Ding, Zijian Du, Fangzhou Mu, Jiaxin Pei, Jieyu Zhao, Swabha Swayamdipta, Willie Neiswanger, Hua Wei, Xiyang Hu, Shixiang Zhu, Tianlong Chen, Yingzhou Lu, Yang Shi, Lianhui Qin, Tianfan Fu, Zhengzhong Tu, Yuzhe Yang, Jaemin Yoo, Jiaheng Zhang, Ryan Rossi, Liang Zhan, Liang Zhao, Emilio Ferrara, Yan Liu, Furong Huang, Xiangliang Zhang, Lawrence Rothenberg, Shuiwang Ji, Philip S. Yu, Yue Zhao, and Yushun Dong. Political-llm: Large language models in political science, 2024. URL <https://arxiv.org/abs/2412.06864>.
- [52] Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. Mediq: Question-asking LLMs and a benchmark for reliable interactive clinical reasoning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=W4pIBQ7bAI>.
- [53] Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason Weston, and Mike Lewis. Self-alignment with instruction backtranslation, 2024. URL <https://arxiv.org/abs/2308.06259>.
- [54] Xiang Lisa Li, Vaishnavi Shrivastava, Siyan Li, Tatsunori Hashimoto, and Percy Liang. Benchmarking and improving generator-validator consistency of language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=phBS6YpTzC>.
- [55] Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. AlpacaEval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 5 2023.
- [56] Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Wen tau Yih, and Xilun Chen. FLAME : Factuality-aware alignment for large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=zWuHSTALBh>.
- [57] Nelson Liu, Tianyi Zhang, and Percy Liang. Evaluating verifiability in generative search engines. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7001–7025, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.467. URL <https://aclanthology.org/2023.findings-emnlp.467/>.
- [58] Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4140–4170, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.228. URL <https://aclanthology.org/2023.acl-long.228/>.
- [59] Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.565. URL <https://aclanthology.org/2021.emnlp-main.565/>.

- [60] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and Adam Roberts. The flan collection: Designing data and methods for effective instruction tuning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22631–22648. PMLR, 23–29 Jul 2023. URL <https://proceedings.mlr.press/v202/longpre23a.html>.
- [61] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=Bkg6RiCqY7>.
- [62] Claudia Maienborn. Event-internal modifiers: Semantic underspecification and conceptual interpretation, pp. 475–510. De Gruyter Mouton, Berlin, Boston, 2003. ISBN 9783110894646. doi: [doi:10.1515/9783110894646.475](https://doi.org/10.1515/9783110894646.475). URL <https://doi.org/10.1515/9783110894646.475>.
- [63] Pratyush Maini, Skyler Seto, Richard Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. Rephrasing the web: A recipe for compute and data-efficient language modeling. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14044–14072, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: [10.18653/v1/2024.acl-long.757](https://doi.org/10.18653/v1/2024.acl-long.757). URL <https://aclanthology.org/2024.acl-long.757/>.
- [64] Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. ExpertQA: Expert-curated questions and attributed answers. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3025–3045, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: [10.18653/v1/2024.naacl-long.167](https://doi.org/10.18653/v1/2024.naacl-long.167). URL <https://aclanthology.org/2024.naacl-long.167/>.
- [65] Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore, December 2023. Association for Computational Linguistics. doi: [10.18653/v1/2023.emnlp-main.557](https://doi.org/10.18653/v1/2023.emnlp-main.557). URL <https://aclanthology.org/2023.emnlp-main.557/>.
- [66] Sara Vera Marjanović, Haeun Yu, Pepa Atanasova, Maria Maistro, Christina Lioma, and Isabelle Augenstein. Dynamicqa: Tracing internal knowledge conflicts in language models. In *Conference on Empirical Methods in Natural Language Processing*, 2024. URL <https://api.semanticscholar.org/CorpusID:271404307>.
- [67] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, Online, July 2020. Association for Computational Linguistics. doi: [10.18653/v1/2020.acl-main.173](https://doi.org/10.18653/v1/2020.acl-main.173). URL <https://aclanthology.org/2020.acl-main.173/>.
- [68] Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=3Tzcot1LKb>.
- [69] Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, Singapore, December 2023. Association for Computational Linguistics. doi: [10.18653/v1/2023.emnlp-main.741](https://doi.org/10.18653/v1/2023.emnlp-main.741). URL <https://aclanthology.org/2023.emnlp-main.741/>.

- [70] Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. Faitheval: Can your language model stay faithful to context, even if "the moon is made of marshmallows". In The Thirteenth International Conference on Learning Representations , 2025. URL <https://openreview.net/forum?id=UeVx6L59fg>.
- [71] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pp. 3470–3487, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.244. URL <https://aclanthology.org/2022.acl-long.244/>.
- [72] Mistral AI Team. Un ministral, des ministraux. <https://mistral.ai/news/ministraux/>, October 2024. Accessed: 2025-01-26.
- [73] Sacha Muller, António Loison, Bilel Omrani, and Gautier Viaud. Grouse: A benchmark to evaluate evaluators in grounded question answering. ArXiv , abs/2409.06595, 2024. URL <https://api.semanticscholar.org/CorpusID:272550360>.
- [74] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization, 2018. URL <https://arxiv.org/abs/1808.08745>.
- [75] Pranav Narayanan Venkit, Tatiana Chakravorti, Vipul Gupta, Heidi Biggs, Mukund Srinath, Koustava Goswami, Sarah Rajtmajer, and Shomir Wilson. An audit on the perspectives and challenges of hallucinations in NLP. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing , pp. 6528–6548, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.375. URL <https://aclanthology.org/2024.emnlp-main.375/>.
- [76] Ella Neeman, Roei Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pp. 10056–10070, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.559. URL <https://aclanthology.org/2023.acl-long.559/>.
- [77] Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. Large dual encoders are generalizable retrievers. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing , pp. 9844–9855, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.669. URL <https://aclanthology.org/2022.emnlp-main.669/>.
- [78] Katsuhiko Ogata. Modern Control Engineering . Prentice-Hall, Inc., USA, 3rd edition, 1996.
- [79] OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codisposi, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary

Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljube, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patherdhan, Thomas Cunningham, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyei Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL <https://arxiv.org/abs/2410.21276>.

- [80] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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 F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.
- [81] Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4998–5017, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.297. URL <https://aclanthology.org/2024.findings-acl.297/>.
- [82] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.378. URL <https://aclanthology.org/2023.findings-emnlp.378/>.
- [83] Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep, 2024. URL <https://arxiv.org/abs/2406.05946>.
- [84] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.
- [85] Rafael Rafailov, Yaswanth Chittipetu, Ryan Park, Harshit S. Sikchi, Joey Hejna, Bradley Knox, Chelsea Finn, and Scott Niekum. Scaling laws for reward model overoptimization in direct alignment algorithms. *ArXiv*, abs/2406.02900, 2024. URL <https://api.semanticscholar.org/CorpusID:270257855>.
- [86] Nazneen Rajani, Lewis Tunstall, Edward Beeching, Nathan Lambert, Alexander M. Rush, and Thomas Wolf. No robots. https://huggingface.co/datasets/HuggingFaceH4/no_robots, 2023.
- [87] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL <https://aclanthology.org/D16-1264/>.
- [88] Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 704–718, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.58. URL <https://aclanthology.org/2021.acl-long.58/>.
- [89] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD ’20, pp. 3505–3506, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3406703. URL <https://doi.org/10.1145/3394486.3406703>.

- [90] Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. A recipe for arbitrary text style transfer with large language models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 837–848, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.94. URL <https://aclanthology.org/2022.acl-short.94/>.
- [91] Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=9Vrb9DOWI4>.
- [92] Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=Yacmpz84TH>.
- [93] Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. QuestEval: Summarization asks for fact-based evaluation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6594–6604, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.529. URL <https://aclanthology.org/2021.emnlp-main.529/>.
- [94] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1073–1083, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1099. URL <https://www.aclweb.org/anthology/P17-1099>.
- [95] Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 783–791, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-short.69. URL <https://aclanthology.org/2024.naacl-short.69/>.
- [96] Weijia Shi, Sewon Min, Maria Lomeli, Chunting Zhou, Margaret Li, Xi Victoria Lin, Noah A. Smith, Luke Zettlemoyer, Wen tau Yih, and Mike Lewis. In-context pretraining: Language modeling beyond document boundaries. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=LXVswInH0o>.
- [97] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. REPLUG: Retrieval-augmented black-box language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8371–8384, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.463. URL <https://aclanthology.org/2024.naacl-long.463/>.
- [98] Yixiao Song, Yekyung Kim, and Mohit Iyyer. VeriScore: Evaluating the factuality of verifiable claims in long-form text generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*

- , pp. 9447–9474, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.552. URL <https://aclanthology.org/2024.findings-emnlp.552/>.
- [99] Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. ASQA: Factoid questions meet long-form answers. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8273–8288, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.566. URL <https://aclanthology.org/2022.emnlp-main.566/>.
 - [100] Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models. In *arXiv*, 2022.
 - [101] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.
 - [102] Qwen Team. Qwen3, April 2025. URL <https://qwenlm.github.io/blog/qwen3/>.
 - [103] Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025. URL <https://qwenlm.github.io/blog/qwq-32b/>.
 - [104] Ryan Teknium, Jeffrey Quesnelle, and Chen Guang. Hermes 3 technical report, 2024. URL <https://arxiv.org/abs/2408.11857>.
 - [105] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL <https://aclanthology.org/N18-1074/>.
 - [106] Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. Fine-tuning language models for factuality. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=WPZ2yPag4K>.
 - [107] Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. FreshLLMs: Refreshing large language models with search engine augmentation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13697–13720, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.813. URL <https://aclanthology.org/2024.findings-acl.813/>.
 - [108] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=1PL1NIMMrw>.
 - [109] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krma Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujana Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5085–5109, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.340. URL <https://aclanthology.org/2022.emnlp-main.340/>.

- [110] Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Nenkov Georgiev, Rocktim Jyoti Das, and Preslav Nakov. Factuality of large language models: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19519–19529, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1088. URL <https://aclanthology.org/2024.emnlp-main.1088/>.
- [111] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=gEZrGCozdqR>.
- [112] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS ’22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.
- [113] Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. Long-form factuality in large language models, 2024. URL <https://arxiv.org/abs/2403.18802>.
- [114] Kevin Wu, Eric Wu, and James Zou. Clasheval: Quantifying the tug-of-war between an LLM’s internal prior and external evidence. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL <https://openreview.net/forum?id=WGoCZ12itU>.
- [115] Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=auKAUJZM06>.
- [116] Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. A critical evaluation of evaluations for long-form question answering. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3225–3245, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.181. URL <https://aclanthology.org/2023.acl-long.181/>.
- [117] Kevin Yang and Dan Klein. FUDGE: Controlled text generation with future discriminators. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3511–3535, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.276. URL <https://aclanthology.org/2021.naacl-main.276/>.
- [118] Zitong Yang, Neil Band, Shuangping Li, Emmanuel Candes, and Tatsunori Hashimoto. Synthetic continued pretraining. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=07yvxWDS1a>.
- [119] Xi Ye, Ruoxi Sun, Sercan Arik, and Tomas Pfister. Effective large language model adaptation for improved grounding and citation generation. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6237–6251, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.346. URL <https://aclanthology.org/2024.naacl-long.346/>.
- [120] Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Comput. Surv.*, 56(3), October 2023. ISSN 0360-0300. doi: 10.1145/3617680. URL <https://doi.org/10.1145/3617680>.

- [121] Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. Context-faithful prompting for large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), Findings of the Association for Computational Linguistics: EMNLP 2023 , pp. 14544–14556, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.968. URL <https://aclanthology.org/2023.findings-emnlp.968/>.
- [122] Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. Enhancing factual consistency of abstractive summarization. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies , pp. 718–733, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.58. URL <https://aclanthology.org/2021.naacl-main.58/>.

我们发布了所有工件（PIC -Bench、PIC -LM 8B 模型及相关的数据处理、训练和评估脚本），以支持结果再现性和未来开发：

Codebase		jacqueline-he/precise-information-control
PIC-Bench Data		jacquelinehe/pic-bench
PIC-LM 8B		jacquelinehe/Llama-3.1-PIC-LM-8B
PIC-LM 8B - SFT		jacquelinehe/Llama-3.1-PIC-LM-8B-SFT
PIC-LM SFT Data		jacquelinehe/pic-lm-sft-mixture
PIC-LM Preference Data		jacquelinehe/pic-lm-preference-mixture

8.1 更广泛的影响

我们的工作通过实现对语言模型生成的论断的更强控制，解决了语言模型的基础性问题。PIC -LM 作为一种可靠的工具，用于转换文本而不会引入固有的幻觉，并在下游应用中提高事实准确性。这具有重要意义，因为语言模型是现代生成式搜索引擎背后的主力 [57]。

然而，更大的生成控制也可能提高误用的可能性。当面临某些情境时，PIC -LM 更容易发出与事实不符、有毒或有害的内容，从而绕过在 RLHF 过程中注入的克制行为。虽然这不是最优的，但我们认为无法控制的有毒生成（即幻觉）的风险更大。最近的研究表明，LM 安全保障措施通常较为表面化，容易被绕过 [83]，尽管进行了广泛的安全对齐，传统的 LM 仍可能生成基于不明确模型内部的对抗性输出。相比之下，PIC -LM 使信息来源透明且可直接归因于用户输入，而不是隐藏的模型先验。这样可以促进一个更加可控和可管理的环境，在该环境中可以专门围绕输入验证和用户意图来设计安全措施。

9 PIC -长凳详情

9.1 PIC -基准任务细节

任务信息。 Table 7 显示了 PIC -Bench 任务信息，而 Fig. 2 显示了所有任务的请求次数分布。

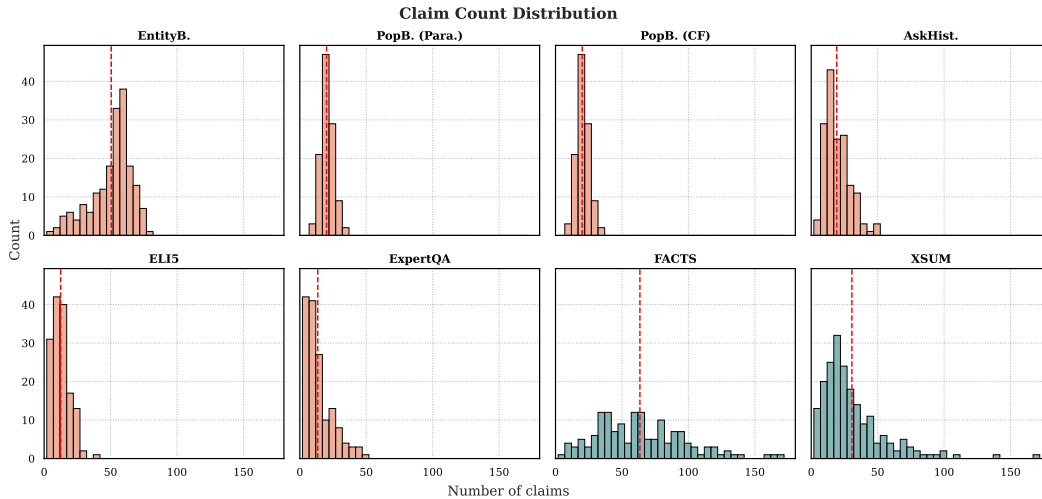


Figure 2: 在完整设置 PIC（橙色）和部分设置 PIC（青色）下的索赔计数分布直方图。红色虚线表示索赔的平均数量。注意， $\text{PopB.}-\text{PIC}$ 和 $\text{PopB.}-\text{CF}_{\text{PIC}}$ 共享相同的分布。

数据集详情。 这些数据集中的许多最初是为长篇形式的真实性而引入的 [69, 64, 98]，这是一个我们没有衡量的维度。相反，我们的 PIC 计量试图询问：模型能够多大程度上忠实地整合给定的信息——无论是部分还是全部？即使在理想条件下（即，当断言构成了黄金响应时），我们在现有模型中发现了一个有意义的忠实性差距。我们将在下面更详细地描述每个数据集：

Table 7: PIC -Bench 的数据集信息。平均声明长度以单词计（由空格分隔）。

Dataset	Sample Count	Min. / Avg. / Max. # Claims	Avg. Claim Len.	Data Provenance
ENTITYBIOSPIC [69]	183	3 / 50.5 / 79	12.0	Human (Search)
POPBIOSPIC [27]	111	7 / 20.1 / 33	10.7	Synthetic
ELI5PIC [116]	146	4 / 12.5 / 40	10.6	Human (Reddit)
ASKHISTORIANSPIC [116]	158	6 / 19.2 / 49	12.1	Human (Reddit)
EXPERTQAPIC [64]	152	3 / 13.5 / 51	11.6	Human
FACTS GROUNDINGPIC [36]	150	6 / 63.5 / 167	13.1	Human
XSUMPIC [74]	200	2 / 30.6 / 171	12.1	Human

实体传记由来自 Min et al. [69] 的命名实体组成；我们从相关的维基百科页面中获取关于每个实体的传记背景信息。对于每个给定的 `entity`，相应的指令是：“生成关于 `entity` 的事实性传记。”

Popular Bios 是一个著名现实世界实体的传记生成任务，涉及各种领域，例如艺术、科学或文学 [27]。在这个数据集中，每个样本包含一个参数化实体和一个反事实实体（也都是不同领域的著名个人）。每个实体被插入到一个传记查询中，例如，“生成关于 { 实体 } 的传记，这位发现了电磁学和电化学的科学家”，为此，语言模型必须依赖上下文线索来描述一个特定的传记实体。为了进行控制研究，我们考虑了参数化（例如，“迈克尔·法拉第”）和反事实（例如，“纳尔逊·曼德拉”）的设置。

ELI5（正式名称为“请像我五岁一样解释”）由来自 Reddit 论坛⁶的用户撰写的问题和回答组成，其中用户要求对复杂问题或主题进行简单解释。我们使用了来自 [98] 的一个子样本集。

AskHistorians 由 Reddit 论坛⁷关于不同历史话题的深入且全面的人为撰写的问题和回答组成。我们使用了来自 [98] 的一个子样本集。

ExpertQA 由专家策划和验证的长篇问答对组成，涵盖 32 个领域；我们使用一个随机抽样的集合 [64]。FACTS Grounding 是一个长文本基准，用于评估语言模型在生成文本时对于长度达 32K 个标记的语境文档的准确性的能力 [36]。

XSUM 是一个汇总数据集，包含来自 BBC 的新闻文章 [74]。它最初是为极端的单句汇总而设计的。在我们的案例中，对于每个样本，我们将给定的文章转换为一个声明列表，并提供以下指令：“将以下内容作为声明列表进行汇总，字数约为 20 到 25 个字。”

在适用的情况下（即，如果上下文是从人类创作的优质回复中提取的，比如来自 Reddit 任务），我们进行一些轻量的质量处理，只保留 Prometheus 评分高于 4 的样本。

9.2 PIC 评估细节

PIC 的评估包括三个部分：声明提取、声明验证和指标计算。

我们采用由 Song et al. [98] 引入的主张提取和验证程序。对于主张提取，我们对 GPT-4o mini 使用少量提示技术，并采用滑动窗口格式（即，每个句子都附有其前后句子作为上下文），以确保提取的主张是自包含的（即，在语法和语义上没有歧义）。为了防止分数被人为夸大，主张在验证之前会被去重。

对于声明验证，我们提示 GPT-4o mini 确定每个声明是否由原始上下文支持，并分配两个标签之一：`supported` 或 `unsupported`。一个响应声明可以由一个或多个上下文声明支持。与 Song et al. [98] 不同，我们不是根据从谷歌搜索检索的片段列表进行验证，而是检查输入声明的列表。

我们考虑部分 PIC 设置的精度和全 PIC 设置的 F_1 （形式上为 $F_1 @K$ ）。这两个指标是按照 Wei et al. [113] 和 Song et al. [98] 计算的，我们将在下面说明原因。

假设输入中提供了一组声明为 C ，以及一个生成的响应 y 。令 $S(y)$ 为支持的声明数量， $N(y)$ 为响应 y 中不支持的声明数量。然后，可以计算响应的事实精度为 $\text{Prec}(y) = \frac{S(y)}{S(y)+N(y)}$ ，这是部分 PIC 设定中的最终指标。

⁶reddit.com/r/explainlikeimfive

⁷reddit.com/r/AskHistorians

在完整的 PIC 设置中，我们希望计算 $F_1 @K$ ，该计算平衡生成响应的“支持性”与 C 中可验证声明的覆盖率。

Wei et al. [113] 注意到衡量事实回忆是具有挑战性的，因为在大多数情况下，不可能提出一个详尽的所有可能主张的集合，这些主张应该包含在理想的长篇回应中。他们建议设置一个超参数 K ，它是一些用户关心召回的支持主张的数量，直到第 K 个支持的事实。

对于 PIC 的情况更简单：我们只关心生成的内容是否准确地包含了 C 中的所有论点，而不是外部现实世界知识语料库。因此，我们可以设定 $K = |C|$ 。令 $|C_y|$ 为在 C 中也存在于 y 中的唯一论点的数量。然后， y 的事实召回率 @K 为 $R_K(y) = \min(\frac{|C_y|}{K}, 1)$ 。

$F_1 @K$ 结合了精度和召回率 @K，如下所示：

$$F_1 @K = \begin{cases} \frac{2 \cdot \text{Prec}(y) \cdot R_K(y)}{\text{Prec}(y) + R_K(y)}, & \text{if } S(y) > 0, \\ 0, & \text{if } S(y) = 0. \end{cases}$$

9.2.1 PIC -基准评估成本。

PIC 评估依赖于一个大型语言模型 (LLM) API，该 API 按每个 token 收取用于声明提取和验证的费用。总费用取决于多个因素，例如提取的声明数量（因为每个提取的声明都在一个独立的 API 调用中进行验证）以及输入和输出的上下文长度。根据我们的经验，使用 GPT-4o mini 作为 LLM 主干在 PIC -Bench 上对一个开放 LM 进行完整评估的花费少于 \$7 美元，我们认为这在学术预算下是合理的。

进一步的成本削减措施可能涉及要么切换到更便宜的 LLM，要么使用专用的开放 LM 来进行声明提取和验证 [98]。相反，更严格的评估将包括与顶级 LLM 评委一起进行集体评分并对结果分数进行平均，类似于 Jacovi et al. [36]。然而，这样做将使成本按评估模型的数量成比例增加。

9.2.2 PIC -Bench 评估可靠性。

为了量化统计精度，我们通过引导法重采样 1000 次估计出的 95 % 置信区间 (Table 8)。在大多数设置中，置信区间的范围相对较窄，通常在 2-4 个百分点以内，这表明我们的 PIC 指标在样本变异性下是统计稳定的。对于专有语言模型，置信区间尤其紧密（例如，GPT-4o, EntityB.: 91.9, 93.35），反映出高指标一致性。

相比之下，在更具挑战性或得分较低的任务中（例如，CLAUDE 3.5, PopB. (CF): 5.77, 9.99），出现了更宽的区间，这表明在这些情况下的声称提取或验证具有更大的变异性。然而，跨任务和系统的置信区间总体狭窄，证实了 PIC 评估可以提供对 PIC 性能的精确和可信的估计。

为了严格和公平的跨模型比较，我们在所有模型基线和任务中使用相同的提示模板。?? 显示了用于 PIC 训练和评估的整个过程中使用的提示。在初步实验中，我们探索了不同的提示格式，但没有注意到任何显著的不同趋势。

我们使用 vLLM 工具包 [45] 对所有开放的、非 API 的语言模型进行快速推理。对于非推理模型，我们使用贪婪解码（温度设为 0）和重复惩罚设为 1.2 进行生成，因为在较低温度下，生成会更加确定性 [29, 47]。对于 32B 推理模型，贪婪解码会导致显著下降或重复的输出。因此，我们遵循推荐的最佳实践：启用思考模式，并使用温度为 0.6、top-p 为 0.85、top-k 为 20、min-p 为 0、存在惩罚为 1.5 和重复惩罚为 1.2。我们在评估前去掉中间的思考成分。

9.3 小样本基线结果

我们在 Table 9（完整设置）和 Table 10（部分设置）中展示了对于选择基准的少样本结果。由于我们仅对经特殊训练以零样本方式响应人类指令的指令微调模型进行基准测试，我们认为不需要少样本示例；然而，为了全面性，我们仍然包括了它们。

在 32B 开权重推理模型中，少样本示例在所有情况下都导致性能下降，这可能是由于缺少明确的逐步推理，这些语言模型在训练时依赖此步骤。对于非推理模型，少样本示例通常有助于完整 PIC，但会损害部分 PIC。在完整 PIC 设置中，少样本提示一致地提高了所有基线的平均 PIC 得分。这一趋势在部分设置中逆转：添加少样本示例降低了完美得分的平均精度，这表明示例可能会分散语言模型的注意力。根据定义，部分 PIC 样本包括不属于目标响应的无关断言。少样本示例，尤其是如果它们引用无关或嘈杂的断言子集，可能会引入更多的干扰，使语言模型难以根据预期背景生成内容。

Table 8: 在 PIC -Bench 上对所有任务和模型的 95 % 置信区间 (CI)。对于每种设置, 我们从每个评估集中有放回地抽取 1000 个样本。所有推理语言模型 (用 † 表示) 均在启用思考模式的情况下进行评估。

	EntityB.	PopB. (Para)	PopB. (CF)	AskHist.	ELI5	ExpertQA	FACTS	XSUM
Open-weight LMs (8B)								
LLAMA 3.1 8B INST.	72.72, 76.09	81.48, 85.39	20.20, 27.09	79.05, 84.33	74.20, 80.96	69.64, 77.41	65.49, 74.35	72.48, 81.73
TULU 3 8B	78.41, 80.80	86.07, 89.45	47.23, 55.26	77.98, 82.43	78.99, 84.42	77.26, 83.85	71.44, 79.00	72.61, 81.43
MINISTRAL 8B INST.	77.37, 80.03	87.32, 90.66	59.72, 67.42	73.58, 79.81	74.80, 81.29	73.23, 81.72	70.33, 80.37	78.56, 86.54
HERMES 3 8B	73.35, 76.4	82.59, 87.09	39.81, 48.96	69.82, 75.86	72.76, 79.03	69.66, 78.4	69.63, 78.39	77.58, 84.87
Open-weight LMs (32B)								
QWEN 3 32B†	71.22, 76.97	85.24, 90.78	43.19, 53.38	74.17, 80.58	62.96, 72.39	58.51, 69.07	59.84, 69.22	74.29, 82.42
QwQ 32B†	85.63, 87.75	86.13, 89.04	60.90, 67.40	77.33, 81.82	77.70, 82.57	77.61, 83.82	66.80, 74.07	84.14, 89.70
R1 DISTILL 32B†	81.06, 83.55	85.62, 89.13	54.78, 62.83	78.80, 83.48	79.04, 85.00	76.25, 83.64	65.92, 75.89	74.97, 84.01
Open-weight LMs (70B)								
LLAMA 3.3 70B INST.	88.12, 89.89	92.47, 94.46	41.51, 49.47	89.94, 92.34	88.59, 91.83	82.41, 88.75	75.01, 83.38	62.39, 75.25
TULU 3 70B	83.74, 85.96	87.68, 90.54	48.48, 56.05	81.56, 85.68	80.08, 84.93	68.98, 77.43	42.33, 49.61	82.35, 88.45
HERMES 3 70B	79.96, 82.91	87.65, 91.34	53.29, 59.89	78.77, 84.10	76.69, 82.55	76.20, 83.18	74.06, 81.91	83.39, 90.16
Proprietary LMs								
CLAUDE 3.5	86.08, 91.23	91.35, 94.10	5.77, 9.99	79.37, 87.68	81.69, 89.62	57.19, 70.45	76.10, 87.20	86.91, 93.43
GPT-4o	91.90, 93.35	92.08, 94.47	74.21, 79.15	95.27, 96.97	92.98, 96.32	87.16, 92.82	89.58, 94.58	89.92, 95.41
Ours: PIC-LM								
PIC-LM _{SFT ONLY}	85.75, 87.71	85.34, 88.88	69.17, 74.42	84.38, 88.78	79.16, 85.20	73.14, 81.38	75.98, 85.83	92.27, 95.56
PIC-LM	93.56, 94.65	95.12, 96.79	78.35, 83.07	94.05, 95.99	92.57, 95.45	81.94, 89.38	41.75, 49.23	95.50, 97.73

Table 9: 选择的基线的少样本 PIC -Bench 结果 (完整设置)。指标包括 F_1 、完美 F_1 的比例 (Perf.) 和两个指标的平均值。最佳值以粗体显示。对于所有指标, 数值越高越好。

	ENTB _{PIC}		POP _{B-PIC}		POP _{B-CF_{PIC}}		ASK _{H_{PIC}}		ELI5 _{PIC}		EXP _{QA_{PIC}}		Avg.	
	F_1	Perf.	F_1	Perf.	F_1	Perf.	F_1	Perf.	F_1	Perf.	F_1	Perf.	F_1	Perf.
Open-weight LMs (8B) + 2-shot														
LLAMA 3.1 8B INST.	72.4	1.1	89.8	15.3	22.2	0.0	80.2	1.3	80.0	11.6	77.1	11.8	70.3	6.9
TULU 3 8B	78.1	2.2	91.7	10.8	61.4	0.9	82.5	1.3	84.0	13.0	78.3	11.8	79.3	6.7
MINISTRAL 8B INST.	76.5	1.1	83.4	16.2	73.4	0.9	82.3	4.4	79.5	12.3	80.6	15.1	80.9	8.4
Open-weight reasoning LMs (32B) + 2-shot														
QWEN 3 32B	46.4	2.2	3.33	0	1.41	0	8.24	0	23.4	0.7	16.3	0.7	16.5	0.6
QwQ 32B	76.6	4.4	58.4	8.1	28.1	0.9	47.6	3.8	47.3	6.9	39.5	7.2	49.6	5.2
R1 DISTILL 32B	85.7	3.8	57.3	13.5	42.2	0.9	39.9	1.3	54.4	6.9	45.1	11.8	54.1	6.4
Proprietary LMs + 2-shot														
CLAUDE 3.5	92.4	10.9	94.4	27.9	25.4	0.9	94.4	26.0	92.8	43.2	90.2	44.7	81.6	25.6
GPT-4o	91.1	6.0	94.7	30.6	79.5	3.6	95.4	41.1	93.7	49.3	91.5	52.0	91.0	30.4

Table 10: 2-shot PIC 基准测试结果（部分设置）针对所选基线。指标包括精确度（Prec.）、完美精确度的比例（Perf.）和这两个指标的平均值（最右列）。最佳值以粗体显示。对于所有指标，数值越高越好。

	FACTSPIC		XSUMPIC		Avg	Avg
	Prec.	Perf.	Prec.	Perf.	Prec.	Perf.
Open-weight LMs (8B) + 2-shot						
LLAMA 3.1 8B INST.	66.6	14.0	70.4	48.0	68.5	31.0
TULU 3 8B	68.8	5.3	59.4	39.5	64.1	22.4
MINISTRAL 8B INST.	72.6	21.3	75.7	60.0	74.1	40.7
Open-weight reasoning LMs (32B) + 2-shot						
QWEN 3 32B	63.3	25.3	5.9	1.5	34.6	13.4
QwQ 32B	73.7	12	75.5	50	74.6	31.0
R1 DISTILL 32B	78.1	22.6	41.6	27	59.9	24.8
Proprietary LMs + 2-shot						
CLAUDE 3.5	86.4	49.3	89.6	73.0	88.0	61.2
GPT-4o	87.3	46.7	91.0	81.0	89.1	63.8

9.4 思维模式消融

推理语言模型被训练来执行明确的思维链推理步骤 [112]，通常通过使用特殊控制标记，指示模型在产生最终答案之前进行中间“思考”步骤 [102]。我们在 PIC-Bench 上使用 32B 推理语言模型（Qwen 3 32B，QwQ 32B，R1 Distill 32B）进行思考模式的效果消除。在非思考模式下，我们解码时使用 0.7 的温度、0.8 的 top-p、20 的 top-k、0 的 min-p、1.5 的出现在现量罚，以及 1.2 的重复惩罚 [102]。

通过 Table 11，我们观察到在几乎所有任务中，思考模式显著提高了 PIC 的性能。值得注意的是，每一个完整的 PIC 任务都从思考模式中受益，而 XSUM_{PIC} 是唯一一个效果不太一致的任务。虽然仍需进一步探讨，我们假设思考模式可能有助于 PIC——一个受限的事实回忆任务，因为它使语言模型的解码行为更加慎重、结构化，并且不易于浅显的模式完成，从而减少流畅性偏差。

Table 11: 在 PIC-Bench 上比较 32B 推理语言模型启用和未启用思维模式的情况（T = 启用思维，NT = 未启用思维）。

	ENTBPIC		POPB-PPIC		POPB-CFPIC		ASKHPIC		ELI5PIC		EXPQAPIC		FACTSPIC		XSUMPIC	
	F ₁	Perf.	F ₁	Perf.	F ₁	Perf.	F ₁	Perf.	F ₁	Perf.	F ₁	Perf.	Prec.	Perf.	Prec.	Perf.
Open-weight LMs (32B)																
QWEN 3 32B (T)	74.1	2.7	88.2	18.9	48.2	0.9	77.5	7.0	67.7	6.9	63.6	7.2	64.5	17.3	78.5	46.0
QWEN 3 32B (NT)	34.6	0.0	60.4	0.9	27.2	0.0	41.8	1.9	43.2	4.8	31.7	4.0	38.4	28.0	84.3	64.5
QwQ 32B (T)	86.7	1.6	87.6	2.7	64.1	1.8	79.6	3.8	80.0	6.9	80.9	12.5	70.3	12.7	86.9	58.5
QwQ 32B (NT)	62.5	0.6	74.8	3.6	26.5	0.0	73.2	3.2	71.0	2.7	69.5	4.0	68.5	5.3	81.5	60.0
R1-QWEN-32B (T)	82.3	1.1	87.5	9.0	58.8	0.9	81.1	4.4	82.2	10.3	80.0	13.2	78.6	6.5	71.1	26.7
R1-QWEN-32B (NT)	71.9	1.6	86.0	7.2	41.0	0.0	77.1	1.9	73.7	0.7	76.4	6.6	73.6	8.0	87.3	17.5

9.5 定性任务示例

示例 PIC 每个任务的基准指令如 Table 12、Table 13、Table 14、Table 15、Table 16、Table 17、Table 18 所示。

为了研究 PIC 对单一模型系列的影响，我们在 PIC-Bench 上运行了 Llama 3 模型系列的五个不同大小（1B、3B、8B、70B、405B）。Fig. 3 显示了扩展趋势。我们观察到完整的 PIC 表

Table 12: ENTITYBIOSPIC 示例。

Instruction

Generate a factual biography about Donald Featherstone.

Context Claims

- 唐纳德·费瑟斯通出生于 1936 年 1 月 25 日。
- 唐·费瑟斯通的签名在 2001 年从粉色塑料火烈鸟上被移除了。
- 唐·费瑟斯通在他的后院放置了 57 只塑料火烈鸟。
- 粉色在 1958 年很流行。
- 粉红色火烈鸟在 1958 年上市。
- 唐·费瑟斯通担任 Union Products 的总裁，直到他在 2000 年退休。
- 塑料火烈鸟出现在美国各地。
- 塑料火烈鸟随着时间的推移变得越来越受欢迎。
- Don Featherstone 在马萨诸塞州的柏林长大。
- 唐·费瑟斯通居住在马萨诸塞州的非奇堡。
- 唐·费瑟斯通于 1936 年出生在马萨诸塞州的伍斯特。
- 2010 年，Cado Products 购买了粉色火烈鸟的塑料模具。
- Donald Featherstone 是一位美国艺术家。
- 唐纳德·费瑟斯通于 2015 年 6 月 22 日去世。
- 一家纽约公司购买了唐·费瑟斯通火烈鸟的模具。
- 这家纽约公司将火烈鸟的生产分包给了一家位于非奇堡的公司 Cado Products。
- 唐纳德·费瑟斯通因于 1957 年创造塑料粉红火烈鸟而广为人知。
- 唐·费瑟斯通的签名一直保留在粉红色塑料火烈鸟上，直到 2001 年。
- 在 2006 年 11 月，Union Products 关闭了。
- 塑料火烈鸟被纳入各种艺术展品的一部分。
- 2006 年 11 月，火烈鸟的生产停止了。
- 多恩·费瑟斯通在 Union Products, Inc. 雕刻了超过 750 种不同的物品。
- Don Featherstone 因路易体痴呆去世。
- Cado Products 公司继续生产粉色火烈鸟。
- 唐·费瑟斯通于 1996 年开始担任联合产品公司的总裁。
- 1996 年，Don Featherstone 因其创作的粉红火烈鸟而获得搞笑诺贝尔艺术奖。
- 唐·费瑟斯通于 2015 年 6 月 22 日去世。
- 2010 年，Cado Products 购买了粉色火烈鸟的版权。
- 唐纳德·费瑟斯通在为联合产品公司工作时创造了塑料粉红火烈鸟。
- 1957 年毕业后，Don Featherstone 获得了一份为 Union Products, Inc. 设计三维动物的工作。
- Don Featherstone 雕刻的前两个作品是一个拿着喷水壶的女孩和一个带着狗的男孩。
- Don Featherstone 在创作粉红色火烈鸟时，基于 National Geographic 中火烈鸟的照片。
- 唐·费瑟斯通无法获得真实的火烈鸟来作为他创作的模型。
- 1987 年，唐纳德·费瑟斯通在塑料火烈鸟的原始塑模上刻下了他的签名。
- 更换签名是由于对未签名的粉色塑料火烈鸟的小规模抵制。
- 1957 年，唐·费瑟斯通被要求雕刻一只火烈鸟。
- 唐·费瑟斯通于 1957 年毕业于伍斯特艺术博物馆的艺术学校。
- 唐·费瑟斯通去世时享年 79 岁。

Table 13: PopBios-P_{PIC} 和 PopBios-CF_{PIC} 示例。

Instruction Generate a biography on { entity }, the author who wrote War and Peace , Anna Karenina .
Context Claims (Parametric Entity: Leo Tolstoy) • 《战争与和平》的出版是列夫·托尔斯泰职业生涯中的一个重大事件。 • 《战争与和平》是一部不朽的小说。 • 《战争与和平》捕捉到了拿破仑战争时期俄国社会的精髓。 • 小说《战争与和平》一经问世便取得了成功。 • 《战争与和平》的成功巩固了 { 实体 } 作为文学天才的声誉。 • 《安娜·卡列尼娜》是一部探讨爱、背叛和寻找意义主题的小说。 • 《安娜·卡列尼娜》巩固了列夫·托尔斯泰在文学殿堂中的地位。 • 《安娜·卡列尼娜》的开头一句是：“幸福的家庭都是相似的，不幸的家庭各有各的不幸。” • 安娜·卡列尼娜的开头语变得非常有名。 • 《安娜·卡列尼娜》的开头一句话体现了列夫·托尔斯泰对人性洞察的精髓。 • 《安娜·卡列尼娜》是一部批判俄罗斯社会僵化结构的小说。 • 《安娜·卡列尼娜》作为对挑战社会规范后果的社会评论。 Context Claims (Counterfactual Entity: James Clerk Maxwell) • 《战争与和平》的出版是詹姆斯·克拉克·麦克斯韦职业生涯中的一个重要事件。 • 《战争与和平》是一部巨著。 • 《战争与和平》捕捉了拿破仑战争期间俄罗斯社会的本质。 • 小说《战争与和平》获得了即时成功。 • 《战争与和平》的成功巩固了 { entity } 作为文学天才的声誉。 • 《安娜·卡列尼娜》是一本探讨爱情、背叛和寻找意义主题的小说。 • 《安娜·卡列尼娜》巩固了詹姆斯·克拉克·麦克斯韦在文学殿堂中的地位。 • 《安娜·卡列尼娜》的开头是“幸福的家庭都是相似的；不幸的家庭各有各的不幸。” • 《安娜·卡列尼娜》的开场白成为经典。 • 《安娜·卡列尼娜》的开篇语概括了詹姆斯·克拉克·麦克斯韦对人性的一种洞见。 • 《安娜·卡列尼娜》是一部批判俄罗斯社会僵化结构的小说。 • 《安娜·卡列尼娜》是对违反社会规范后果的社会评论。

现出倒 U 型的性能曲线， F_1 在 70B 时达到峰值，这表明完整的 PIC 能力并不随模型大小单调增长。相反，部分 PIC 性能随着模型大小严格提高，其中 3B 到 8B 之间的增长最大。

9.6 更多的 PIC -Bench 图表。

Fig. 4 显示了完整 PIC 基线的精确度与召回率热图。我们观察到 GPT-4o 和 Claude 3.5 在所有任务中都达到了始终如一的高精确度和召回率，而开源权重的语言模型则偏向保守，召回率较低（这在 ENTITYBIOSPIC 上尤其明显，这是一个平均有约 50 个声明的任务）。PopBios-CF_{PIC} 是最具挑战性的任务，在所有设置下得分都较低。

Fig. 5 显示了所有 PIC 任务中支持和不支持的声明的分布。有一些趋势需要注意：Llama 3.1 8B Instruct 是最为详细的，在除 ENTITYBIOSPIC 和 FACTSPIC 以外的所有任务中生成的声明明显最多。专有语言模型显示出较低比例的不支持声明。

10 PIC -LM 训练详情

为了构建我们的训练数据集，我们考虑从以下通用来源组装的指令微调数据集：No Robots [86]、FLAN [111]，以及任务特定的来源：实体传记生成 [69]、FACTS Grounding [36]

Table 14: AskHISTORIANS_{PIC} 示例。

Instruction Answer the following question: I' ve seen Japanese artwork from the Edo era and before depicting Tigers. Did tigers ever inhabit the islands of Japan? If not, how might a Japanese person encounter a tiger before the Meiji era?
Context Claims - 老虎从未栖息在日本的岛屿上。 - 日本人通过与中国的文化交流熟悉了老虎的形象。 - 日本人通过与中国的文化交流熟悉了老虎的象征意义。 - 在中国文化中，老虎是十二生肖之一。 - 在中国文化中，虎是被称为四象之一的神话生物之一。 - 在中国的象征中，老虎代表西方。 - 在中国的象征中，老虎代表秋天。 - 在中国的象征体系中，老虎代表金元素。 - 在中国的象征主义中，龙代表东方。 - 在中国的象征体系中，龙代表着天堂。 - 虎与龙的二元性在《易经》等古代文本中被颂扬。 - 在中国文化中，龙与繁荣联系在一起。 - 在中国文化中，龙与神圣的统治者有关联。 - 在中国文化中，老虎象征着世俗的力量。 - 在中国文化中，老虎象征着敏捷。 - 早在公元 1 世纪，日本就从中国采纳了许多文化元素。 - 老虎的象征意义是日本从中国采纳的文化元素之一。 - 日本的文化输入涉及佛教信仰的交流。 - 在日本，文化输入包括道教信仰的交流。 - 文化在日本的引入涉及到儒家信仰的交流。 - 日本的文化引进涉及到阴阳（阴阳道）等实践。 - 在日本，老虎象征着军事力量和权威。 - 在日本，老虎被武士精英采用以代表力量。 《大婆萨本生》是佛教故事。 - 中国风格的水墨画是由禅僧使者带到日本的。 - 禅宗使者将中国风格的水墨画带到了日本。 《大勇猛本生》丰富了老虎在日本艺术和文化中的象征意义。 - 中国风的水墨画丰富了老虎在日本艺术和文化中的象征意义。 - 丰臣秀吉在 16 世纪晚期领导了对朝鲜的入侵。 - 在江户时代，日本艺术家在描绘老虎时主要依赖现有的中国表现形式。 - 老虎从来不是日本的本土动物。 - 老虎在日本艺术和文化中发挥了重要作用。 - 中国传统对日本艺术和文化有很大影响。

Table 15: ELI5_{PIC} 示例。

Instruction How come horses were so common all over the world in the past? I recently watched a documentary about a brief history of each continent when I noticed horses were really common everywhere for an example Genghis Khan in Asia, Saladin in Africa, natives in North and South America and all over Europe too. I know they' ve been brought with settlers to Australia because it was mentioned in the documentary but what about the other continents?
Context Claims • 马在北美进化而来。 • 大约两百万年前，马通过白令陆桥迁移到欧亚大陆。 • 大约在一万年前，马从美洲消失了。 • 马是由欧洲定居者在较近的历史中重新引入美洲的。 • 马在公元前 4000 年至 3500 年左右在欧亚大陆被驯化。 • 马是在中亚草原被驯化的。 • 育种和选择导致马匹变得更大。 • 繁育和选育培育出了更强壮的马匹。 • 体型更大且更健壮的马让它们在旅行中变得不可或缺。 • 更大且更强壮的马匹促进了它们在贸易中的不可或缺性。 • 更大更强壮的马匹增强了它们在通信中的不可或缺性。 • 体型更大且更健壮的马匹增强了它们在农业中的不可或缺性。 • 更大且更强壮的马匹使它们在战争中变得不可或缺。 • 成吉思汗在亚洲的军事活动严重依赖马匹。 • 萨拉丁在非洲的军事行动中严重依赖于马匹。 • 大象需要大量的食物。

和 CNN-DailyMail 新闻文本摘要 [94] (从 2000 个随机子样本开始)。如果任务没有提供准确的答案，我们使用 GPT-4o 来生成一个合成的准确答案，因为它在平均水平上已经相当不错 PIC。对于我们的 PIC 基准 (FACTS 和传记生成) 中的域内训练集，我们确保不存在训练集和测试集的重叠。

我们应用以下数据处理流程：

1. 根据指令确定 PIC 类型。我们不考虑本质上具有创造性的样本 (即，可验证声明密度低) 或需要深入推理的样本 (即，数学或编程问题)，并在可能的情况下手动通过标签确定。
2. 过滤掉响应长度少于 128 字符的样本 (因为我们的重点是长篇生成设置)。
3. 从完整 PIC 样本的响应中提取声明列表 C ，从部分 PIC 样本的给定上下文中提取。
4. 鉴于论点提取步骤可能并不完美，我们发现这种情况尤其在回答措辞含糊或使用其他语言时容易发生，我们计算精确率、召回率和 F_1 ，以评估提取的论点列表和标准回答之间的关系。我们仅保留那些具有非空论点集且其精确率或 F_1 阈值 (取决于 PIC 类型) 足够高的样本。对于每个样本，阈值为 1.0，除了传记生成样本外，我们将 F_1 阈值放宽至 0.9。

经过数据处理后，我们的数据集由 2906 个完整的 PIC 样本和 1501 个部分的 PIC 样本组成。Fig. 6 可视化展示了我们数据组合的任务分布。长篇生成、摘要和闭卷问答占据了我们的训练数据的主要部分，涵盖了 69.6%。Table 19 展示了我们 SFT 数据集中的例子。

10.1 偏好数据创建

Algorithm 1 描述了用于偏好数据创建的协议。请注意，为了简化，算法假设整个数据集使用集合 τ 和 $p_{\max}$ ；在实际应用中，我们根据 PIC 设置分配了不同的 τ 和 $p_{\max}$ 值。

Table 16: EXPERTQA_{PIC} 示例。

Instruction Why did art focused on naturalism and realistic notions (such as proportions and perspective) appear mainly in Europe?
Context Claims • 古希腊和罗马艺术及文学的重新发现发生在文艺复兴期间。 • 人文主义运动强调人的经验和个性。 • 人文主义运动使人们对人类形象的真实表现重新产生兴趣。 • 人文主义运动引发了对绘画中空间深度的重新兴趣。 • 人文主义运动引发了对自然准确描绘的重新兴趣。 • 富裕人士赞助了艺术。 • 宗教机构资助了艺术。 • 政治实体提供艺术赞助。 • 艺术赞助鼓励了新技术技法的探索。 • 艺术赞助鼓励对新的艺术主题的探索。 • 艺术家们开始尝试使用如油画颜料等新材料。 • 艺术家们开始尝试使用诸如画布之类的新材料。 • 油画使艺术家能够在他们的作品中实现更高的细节。 • 画布使艺术家能够在作品中实现更高的写实度。 • 线性透视的创立归功于建筑师 Filippo Brunelleschi。 • 线性透视法改变了艺术家在绘画中描绘空间的方式。 • 线性透视法改变了艺术家在画作中表现深度的方式。 • 科学探究的增长促进了自然主义在欧洲艺术中的出现。 • 经验观察的发展促进了现实观念在欧洲艺术中的出现。 • 艺术家的工作室和学院强调对解剖学的系统研究。 • 艺术家的工作室和学院强调对光的系统研究。 • 艺术家的工作坊和学院强调对透视法的系统学习。 • 对解剖学、光线和透视的系统研究促进了艺术中自然主义技法的进步。 • 由于文艺复兴时期对古典艺术和文学的重新发现，欧洲艺术中出现了自然主义和现实主义观念。 • 人文主义的兴起促成了欧洲艺术中自然主义和现实主义观念的出现。 • 赞助制度促成了自然主义和现实主义观念在欧洲艺术中的兴起。 • 艺术技巧的进步促进了自然主义和现实主义观念在欧洲艺术中的出现。 • 科学探究的发展促进了自然主义和现实主义概念在欧洲艺术中的出现。

Table 17: FACTSPIC 示例。

Instruction

Isn't the theater required to provide me with an initial set-up of tea bags in my housing if I am an actor working under the terms of this contract?

Context Claims

- 每个演员的住处应提供处于良好状态的床和床垫。
- 每个演员的住所应配备一个床头柜。
- 每个演员的住所应配备一盏阅读灯。
- 每位演员的住所应配备一把扶手椅或沙发。
- 每位演员的宿舍应当配备一张桌子和椅子。
- 每个演员的住所应配备一盏灯。
- 每个演员的住所应配备一个梳妆台。
- 每个演员的住所应配备一面镜子。
- 每位演员的住所应配备衣架。
- 每位演员的住所应提供床单和毛巾。
- 每个演员的住所都应配备枕头。
- 每个演员的住所应提供毯子。
- 每个演员的房间都应配备一个废纸篓。
- 每个演员的住宿都应配备一个收音机闹钟。
- 每个演员的住所都应配备一台电视和有线电视（如有必要且为了保证足够的接收效果）。
- 剧院应提供熨斗。
- 剧院应提供熨衣板。
- 每位演员的住所应配备带盖的锅碗瓢盆。
- 每个演员的住所都应提供烹饪器具。
- 每个演员的住所应提供四套银器。
- 每个执行者的外壳应提供不少于四个板。
- 每个演员的住处都应配备杯子和玻璃杯。
- 每个演员的住处都应配备一个开罐器。
- 每个演员的宿舍都应提供厨房刀具。
- 每个演员的住处都应配备一个漏勺。
- 剧院应在演员到达之前，在演员的住所提供初始设置的卫生纸。
- 剧院应在演员到达之前，在演员的住宿中提供初始的纸巾设置。
- 在演员到达之前，剧院应在演员的住处提供一套初始的洗手液。
- 剧院应在演员到达之前，在演员的居所内提供一套初始设置的洗碗液。
- 剧院应在演员到达之前在其住宿处提供一套初始的盐和胡椒。
- 剧院应在演员到达之前，在演员住所提供初步的糖的设置。
- 剧院应在演员到达之前，将咖啡初步安置在演员的住所中。
- 剧院应在演员到达前在演员住所内准备好一套茶具。
- 剧院应在演员到达之前，在演员的住处提供一套初始的垃圾袋。
- 剧院应在演员到达之前在演员的住所提供海绵的初始设置。

Table 18: XSUM_{PIC} 示例。

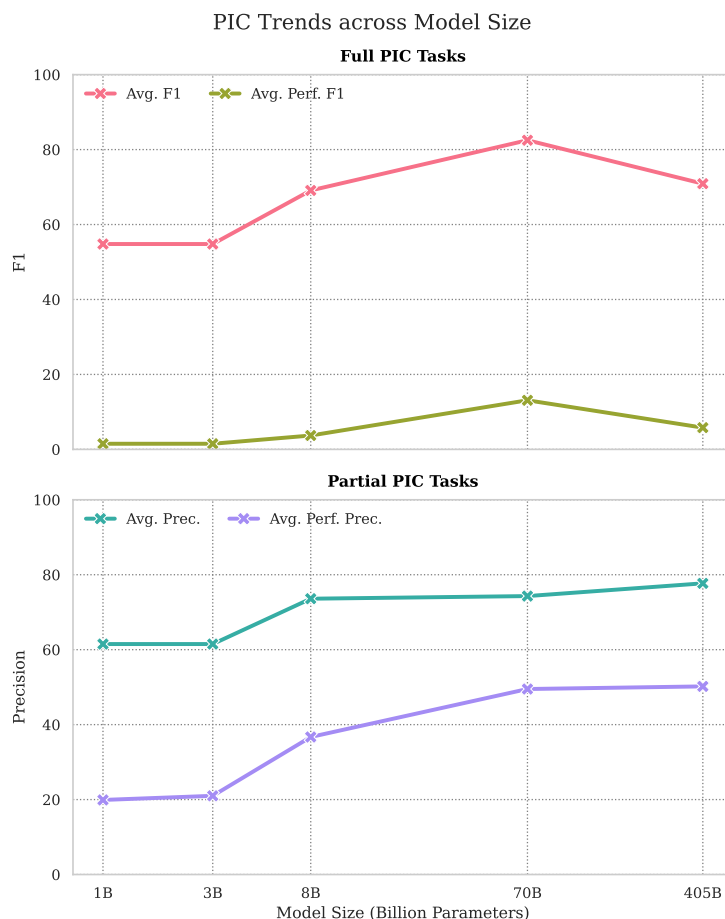
Instruction

Summarize the following text, given as a list of claims, in around 20-25 words.

Context Claims

- 许多发展中国家将尝试通过将森林地区划为保护区来遏制碳排放。
- 在利比里亚指定森林保护区可能会使多达 130 万人流离失所。
- 在刚果民主共和国指定森林保护区可能会导致多达 130 万人流离失所。
- 利比里亚已建议在 2020 年前将其 30 % 的森林设为保护区。
- 利比里亚森林保护提案由挪威资助。
- 刚果民主共和国计划将其森林用地的 12–15 % 作为保护区。
- 刚果民主共和国由德国和全球环境基金资助。
- 流离失所已经在撒哈拉以南非洲发生。
- 位移已经在东南亚发生。
- 拉丁美洲的流离失所已经发生。
- 移民有时会引发激烈的冲突。
- 康斯坦斯·蒂格来自利比里亚的可持续发展研究所。
- 康斯坦斯·蒂格说：“我认为国际社会并不想让利比里亚的农村居民流离失所，但我认为如果我们按照我们现在谈论的方式来处理，那将会是结果。”
- 土著社区尊重森林。
- 土著社区已经在森林中工作了数百年。
- 利比里亚拥有西非最大的森林空间。
- 利比里亚的大面积森林空间主要归功于土著社区。
- 这份报告研究了在利比里亚赔偿人们因土地丧失所需的成本。
- 报告研究了在刚果民主共和国（DR Congo）因土地丧失而对人们进行补偿的成本。
- 在利比里亚和刚果民主共和国，为失去土地的人们提供补偿的费用从 \$ 2 亿到超过 10 亿英镑不等。
- 怀特先生说：“我们需要提供明确的证据，表明树林里到处都是人，与其把他们赶出去，不如帮助他们更有意义。”
- 在土著人民的权利受到保护的地方，他们能够利用自己的森林维持生计。
- 土著人民每公顷的碳储量比保护区更多。
- 大约 15 亿原住民居住或声称拥有世界上大部分土地。
- 根据去年发布的一项研究，原住民对他们居住或声称的土地仅拥有 10 % 的法律权利。

Figure 3: PIC 缩放趋势。



回忆对于 PIC 偏好数据的构建，我们在原始和扰动响应之间选择遵循指令更好的那个。如果扰动响应在遵循指令上显著比原始响应差，那么它应该被不偏爱。我们设定 τ 作为对数概率差异的阈值。我们对 $\tau \in \{0.0, 0.3, 0.4, 0.5, 0.6, 0.7, 1.0\}$ 和一个 RANDOM 基线进行消融（在每个样本中，偏好在原始和扰动响应之间随机选择），以获得 PIC 分数 (F_1 和精度)、指令遵循性（平均 Prometheus 得分）以及 ASQA 上的平均 EM。直观地，一个更高的 τ 提高了对扰动多大破坏性的耐受度（即，选择偏爱原始构建的标准化对数概率降低程度）。在 $\tau = 0$ 时，总是选择原始构建，扰动构建总是被拒绝。在 $\tau = 1$ 时，情况相反。

如 Table 20 所示，当 τ 增加时，完整的 PIC 和部分的 PIC 均表现出反 U 形的性能曲线。随着 τ 的增加，Prometheus 得分也呈现出轻微的反 U 形曲线，而 ASQA 的平均 EM 则严格下降。RANDOM 基线显示了可比的 PIC 和平均 EM 性能，但其 Prometheus 低于 $\tau = 0.5$ 设置，这突显了采用更原则性方法来构建偏好数据的实用性。鉴于这些结果，我们通过在完整的 PIC 设置中设置 $\tau = 0.5$ ，以及在部分 PIC 设置中设置 $\tau = 0.3$ 来构建我们的偏好数据。

LLM 作为评判者验证成对偏好数据。 采用一个强大的 LLM 作为评判者（例如，GPT-4.1）来仲裁响应之间的优劣，是成对指令跟随评估的黄金标准。然而，这样做会产生显著的 API 费用。人们也可以使用专门的奖励模型来预测指令遵循评分，但这需要带标注的偏好数据进行训练，并需要仔细校准以避免其自身的偏差 [46]。

相比之下，我们的概率下降指标既经济又自动，只需依赖于一个有能力的指令遵循语言模型（我们使用现成的 Llama 3.1 8B Instruct）。为了验证它与大语言模型的判断一致，我们：

1. 从我们的训练数据中抽取 200 对，其对数概率下降超过阈值 $\tau \in \{0.5, 0.7, 0.9\}$ 。
2. 使用 GPT-4.1 作为评判者，对每个原始响应与扰动响应进行重新评分。

Figure 4: 六个完整 PIC 任务的精准率和召回率热图。

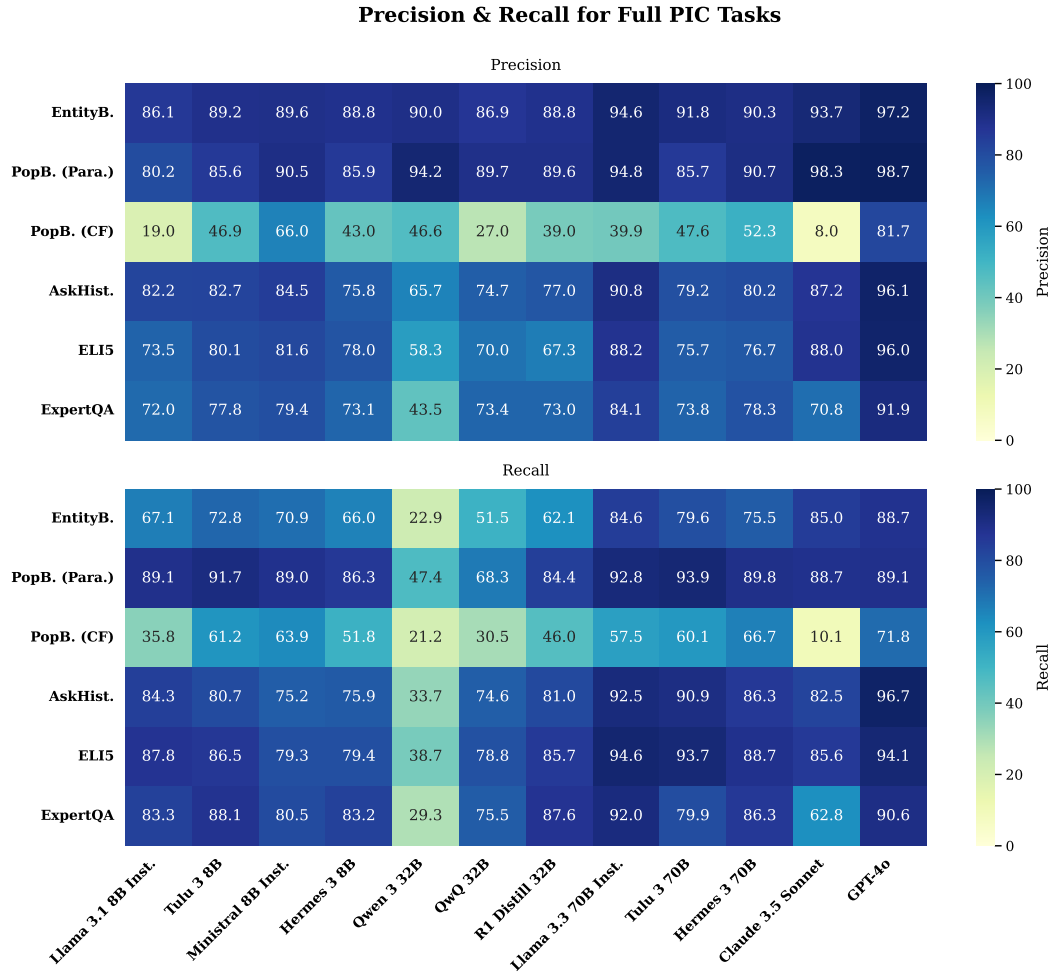


Figure 5: 每个 PIC 任务的基线支持和不支持的声明的直方图。

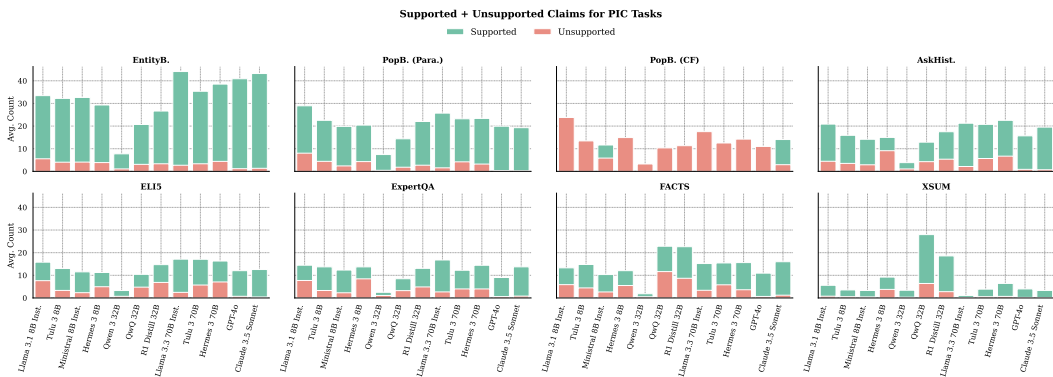


Table 19: PIC SFT 数据中的例子。

Source / Instruction	Verifiable Claims	Response
(CNN DailyMail, Partial PIC) Summarize the following text, given as a list of claims, in around 50 words.	- Saudi Arabia is the world's largest oil producer. - The earnings from oil fueled national growth in Saudi Arabia. - In the previous year, Saudi Arabia earned around \$ 200 billion from oil. - Inflation in Saudi Arabia reached a 30-year high of 9.6 % . - In March, there were price increases of nearly 16 % for fuel and water. - In March, other staples experienced double-digit cost increases.	Saudi Arabia, the world's top oil producer, earned roughly \$ 200 billion from oil last year, driving national growth. However, inflation surged to a 30-year high of 9.6 % . In March, fuel and water prices rose by nearly 16 % , while other staple goods also saw double-digit increases.
(No Robots, Full PIC) What is the most common color car in America?	- Over a quarter of the cars on the road (25.9 %) are white. - The top five most common car colors include white. - Over a quarter of the cars on the road are white, specifically 25.9 % . - Black is the second most common car color in the United States at 22.3 % . - Gray is the third most common car color. - 18.4 % of cars on the road are gray. - 12.1 % of cars on the road are silver. - Silver is the fourth most common car color on the road. - Blue is the only non-monochromatic color in the top five car colors. - 9.5 % of cars on the road are blue. - Blue is fifth among car colors.	That is an interesting question. Most Americans seem to prefer monochromatic colors when it comes to their vehicles. The top five most common car colors are: 1. White —over a quarter of the cars on the road (25.9 %) are white. 2. Black is a close second at 22.3 % . 3. Gray is third with 18.4 % . 4. Silver is fourth with 12.1 % . 5. Blue, the only non-monochromatic color in the top five, is fifth at 9.5 % .
(FACTS, Partial PIC) List the things people thought would happen in the future according to this article from 1995.	- The article titled "Hype alert: Why cyberspace isn't, and will never be, nirvana" was published by NEWSWEEK. - The magazine issue containing the article is dated February 27, 1995. - Visionaries see a future of telecommuting workers. - Visionaries see a future of interactive libraries. - Visionaries see a future of multimedia classrooms. - Visionaries speak of electronic town meetings. - Visionaries speak of virtual communities. - The Usenet is a worldwide bulletin board. - The Usenet allows anyone to post messages across the nation. - The cacophony resembles citizens band radio. - Citizens band radio includes handles. - Citizens band radio includes harassment. - Citizens band radio includes anonymous threats. - Nicholas Negroponte is the director of the MIT Media Lab. - Nicholas Negroponte predicts that people will soon buy books and newspapers straight over the Internet. - The Internet lacks editors, reviewers, or critics. - The Internet has become a wasteland of unfiltered data. - The Battle of Trafalgar is a historical event. - Internet addicts clamor for government reports. - Andy Spano ran for county executive in Westchester County, New York. - Andy Spano put every press release and position paper onto a bulletin board during his campaign. - Westchester County, N.Y. is an affluent county. - Westchester County, N.Y. has plenty of computer companies. - Fewer than 30 voters logged in during Andy Spano's campaign for county executive in Westchester County, N.Y. - Computer-aided education tools are expensive. - Computer-aided education tools are difficult to use in classrooms. - Computer-aided education tools require extensive teacher training.	Based on the provided claims, here are the things people thought would happen in the future according to the 1995 article: 1. Telecommuting would become common for workers. 2. Interactive libraries would emerge. 3. Multimedia classrooms would develop. 4. Electronic town meetings would take place. 5. Virtual communities would form. 6. People would buy books and newspapers directly over the Internet (as predicted by Nicholas Negroponte). These predictions come from the "visionaries" and Nicholas Negroponte, as described in the 1995 NEWSWEEK article.

我们使用不同的随机种子 (0, 21, 42) 进行了三次处理, 通过随机化响应顺序 (以避免顺序展示偏差), 并遵循 AlpacaEval 的 [55] 提示和生成设置, 支持并列选项。如果标准化的对数概率下降可靠地捕捉到指令跟随的忠诚度, GPT-4.1 应该更频繁地选择原始响应而不是扰动后的响应。

从经验上看, GPT-4.1 在 $80.2\% \pm 2.8$ 的案例中偏好原始响应, 包括 $\tau = 0.5$ 处的 $77.6\% \pm 2.6$ 和 $\tau = 0.7$ 处的 $71.3\% \pm 1.2$, 这表明强烈的一致性, 并验证了我们自动指标捕捉到的信号。我们假设, 在更高阈值下一致性下降可能源于被语言模型严重惩罚的扰动响应, 因为它们风格变化或形式上偏离 (这在经验上与 Eisenstein et al. [19] 的发现一致), 但语义上仍然有效, 甚至对 GPT-4.1 而言更可取。

标准化对数概率下降分布。 Fig. 7 显示了我们偏好数据中 (原始和扰动响应之间的) 归一化对数概率差异的分布。直观地看, x 轴上的较高值表明扰动响应相比原始响应明显更差 (由参考模型判断)。

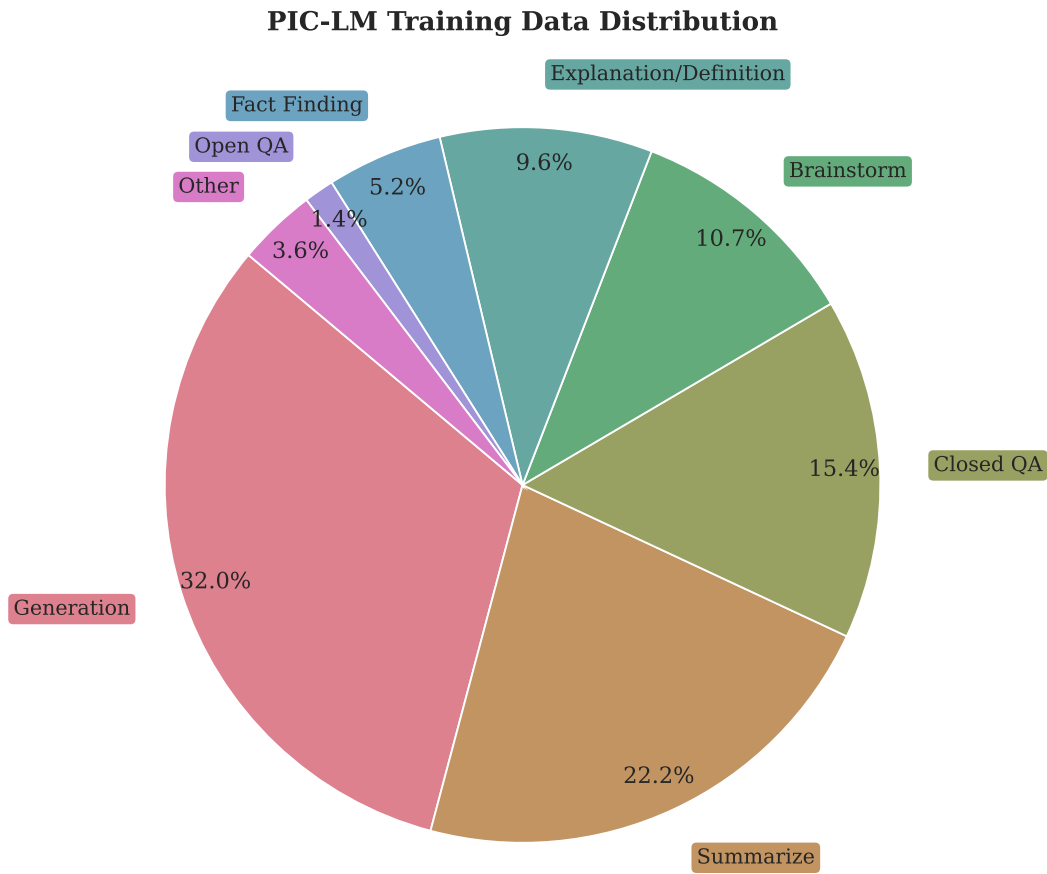


Figure 6: PIC -LM 训练数据分布。

Table 20: 对 τ 进行消融实验，使用平均 PIC F_1 、准确率、Prometheus 平均得分（指令遵循）和 ASQA 平均 EM（基于检索的 QA）。RANDOM 设置在响应中均匀采样。对于所有指标，都是数值越高越好，最好的值用粗体显示。

Setting	(Full) PIC F_1	(Partial) PIC Prec.	Prometheus	Avg. EM
$\tau = 0.0$	88.5	92.8	3.78	66.6
$\tau = 0.1$	88.5	93.0	3.78	65.6
$\tau = 0.3$	90.1	93.1	3.82	63.2
$\tau = 0.4$	90.6	91.2	3.89	61.0
$\tau = 0.5$	91.0	91.9	3.92	58.3
$\tau = 0.6$	88.8	90.9	3.94	56.7
$\tau = 0.7$	87.3	90.1	3.95	56.1
$\tau = 1.0$	85.9	90.6	3.94	55.4
RANDOM	90.3	92.0	3.86	58.3

Algorithm 1 PIC 偏好数据创建 ($\mathcal{D}_{\text{SFT}}$, θ_{SFT} , θ_{Ref} , τ , L , $p_{\min}$, $p_{\max}$)

Require: SFT Dataset $\mathcal{D}_{\text{SFT}} = \{(\mathcal{I}_i, C_i, y_i)\}_{i=1}^N$, SFT model θ_{SFT} , reference model θ_{Ref} , threshold $\tau \in [0, 1]$, token window size L , and $p_{\min}, p_{\max}$ s.t. $0 < p_{\min} \leq p_{\max} \leq 1$

Ensure: Preference dataset $\mathcal{D}_{\text{pref}} = \{(\mathcal{I}_i, C_i, y_i^+, y_i^-)\}_{i=1}^N$

```
1: Initialize  $\mathcal{D}_{\text{pref}} \leftarrow \emptyset$ 
2: function  $\mathcal{P}_{\text{DROP}}^{[p_{\min}, p_{\max}]}(C)$ 
3:    $k \leftarrow |C|$   $\triangleright k$  is the number of input claims
4:   if  $k \leq 1$  then
5:     return  $C$   $\triangleright$  nothing (or only one item) to drop
6:   end if
7:    $m_{\min} \leftarrow \max(1, \lfloor p_{\min} k \rfloor)$ 
8:    $m_{\max} \leftarrow \min(k - 1, \lfloor p_{\max} k \rfloor)$ 
9:   if  $m_{\min} > m_{\max}$  then
10:     $m_{\min} \leftarrow m_{\max}$ 
11:   end if
12:   Sample  $m \sim \text{Uniform}\{m_{\min}, m_{\max}\}$   $\triangleright m$  is the number of claims to drop
13:   Randomly remove  $m$  claims from  $C$  to obtain  $C^{\text{perturb}}$ 
14:   return  $C^{\text{perturb}}$ 
15: end function
16: for each  $(\mathcal{I}_i, C_i, y_i)$  in  $\mathcal{D}_{\text{SFT}}$  do
17:    $C_i^{\text{perturb}} \leftarrow \mathcal{P}_{\text{DROP}}(C_i)$ 
18:    $y_i^{\text{perturb}} \leftarrow \theta_{\text{SFT}}(\mathcal{I}_i, C_i^{\text{perturb}})$ 
19:   Compute per-token log-probability (last  $L$  tokens):

$$\ell_i^{\text{orig}} = \frac{1}{L} \log p_{\theta_{\text{Ref}}}(y_i^{(L)} | \mathcal{I}_i), \quad \ell_i^{\text{perturb}} = \frac{1}{L} \log p_{\theta_{\text{Ref}}}(y_i^{\text{perturb}(L)} | \mathcal{I}_i)$$

20:   Compute normalized log-probability drop:

$$\Delta_i = \ell_i^{\text{orig}} - \ell_i^{\text{perturb}}, \quad z_i = \sigma(\Delta_i) = \frac{e^{\Delta_i}}{1 + e^{\Delta_i}}$$

21:   if  $z_i > \tau$  then
22:     Append  $(\mathcal{I}_i, C_i, y_i, y_i^{\text{perturb}})$  to  $\mathcal{D}_{\text{pref}}$ 
23:   else
24:     Append  $(\mathcal{I}_i, C_i^{\text{perturb}}, y_i^{\text{perturb}}, y_i)$  to  $\mathcal{D}_{\text{pref}}$ 
25:   end if
26: end for
27: return  $\mathcal{D}_{\text{pref}}$ 
```

我们注意到数据分布在范围的两端（即，接近 0 和 1）存在峰值；这表明对于某些样本来说，原始响应在遵循指令质量方面被高置信度偏好，反之亦然，对于另一端的扰动响应也是如此。这种分布的特性激励了使用更明智的策略来选择偏好数据，并且可能解释了为什么随机基线在使用归一化对数概率下降得分选择偏好数据时表现不佳。

Table 21 显示了偏好数据示例。

10.2 训练设置

实现细节。 我们使用 `accelerate` 软件包 [30] 和 DeepSpeed Zero-3 Offload [89] 在一组 40GB NVIDIA L40 和 A40 GPU 集群上进行完整参数训练。对于 SFT 和长度规范化 DPO 阶段，我们使用 AdamW 优化器 [61] 和余弦学习率安排。

为了生成 θ_{SFT} ，我们使用参数 θ 微调一个经过指令调整的 LM。给定每个三元组 $(\mathcal{I}, C, y) \sim \mathcal{D}_{\text{SFT}}$ ，我们最小化标准的教师强制损失：

$$\theta_{\text{SFT}} = \arg \min_{\theta} \mathbb{E}_{(\mathcal{I}, C, y) \sim \mathcal{D}_{\text{SFT}}} \left[-\frac{1}{|y|} \sum_{t=1}^{|y|} \log \pi_{\theta}(y_t | \mathcal{I}, C, y_{<t}) \right].$$

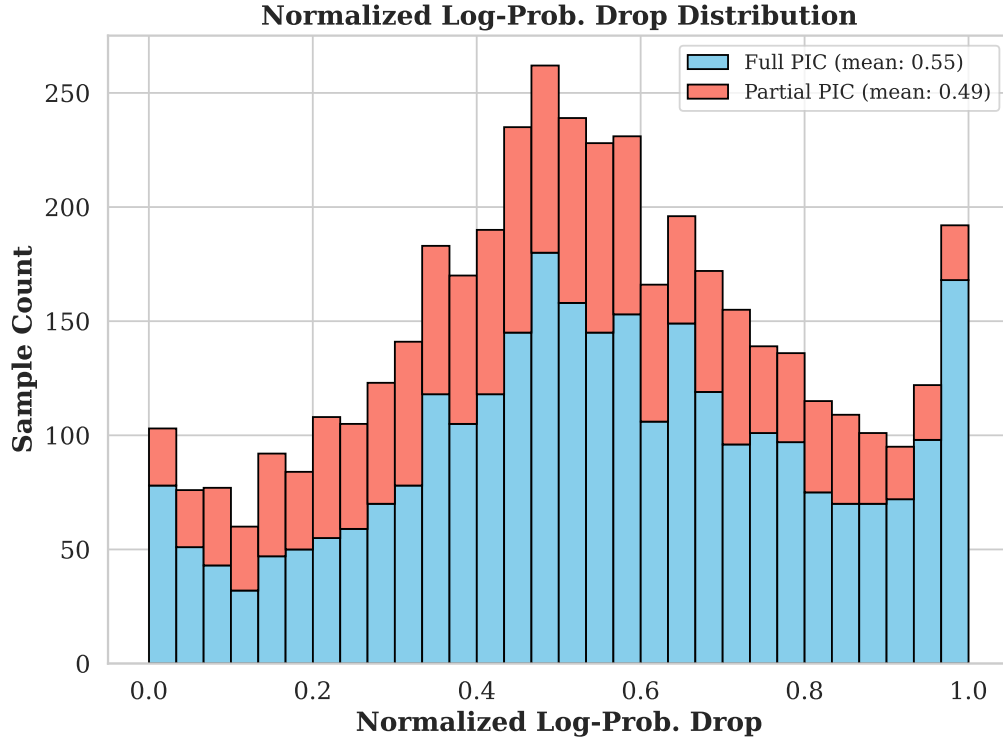


Figure 7: 偏好数据上的归一化对数概率评分分布。

根据 Lambert et al. [46]，我们采用长度归一化的 DPO 目标。给定每个元组 $(\mathcal{I}, C, y^+, y^-) \sim \mathcal{D}_{\text{DPO}}$ ，

$$\theta_{\text{DPO}} = \arg \max_{\theta} \mathbb{E}_{(\mathcal{I}, C, y^+, y^-) \sim \mathcal{D}_{\text{DPO}}} \left[\log \sigma \left(\frac{\beta}{|y^+|} \log \frac{\pi_{\theta}(y^+ | \mathcal{I}, C)}{\pi_{\theta_{\text{SFT}}}(y^+ | \mathcal{I}, C)} - \frac{\beta}{|y^-|} \log \frac{\pi_{\theta}(y^- | \mathcal{I}, C)}{\pi_{\theta_{\text{SFT}}}(y^- | \mathcal{I}, C)} \right) \right],$$

，其中 θ_{SFT} 是固定的参考模型， $\beta > 0$ 是一个缩放超参数。

对于 SFT，我们使用 128 的批量大小和 $1e^{-5}$ 的学习率，并训练 2 个轮次。我们在学习率 $\{1e^{-5}, 1e^{-6}, 2e^{-6}\}$ 上进行网格搜索，批量大小为 $\{32, 64, 128\}$ ，并固定为 0.1 的权重衰减和 2 个轮次（训练更多轮次会导致模型过拟合，并优先输出未扰动的黄金响应，这通过消除输出之间的变化破坏了偏好数据的创建）。

对于 DPO，我们使用学习率 $1e^{-6}$ ，批量大小为 128，权重衰减为 0.1， β 为 5，并训练 1 个周期。我们对学习率 $\{5e^{-7}, 8e^{-7}, 1e^{-6}\}$ 和批量大小 $\{32, 64, 128\}$ 进行网格搜索。

我们在 Table 22 中对 PIC -LM 训练方案的一些设计选择进行了消融实验。

在构建 SFT 数据时，我们仅保留

SFT 数据质量。 得分（无论是 F_1 还是精度）足够高的样本。我们发现，质量过滤显著提升了 PIC 和指令遵循指标。

我们采用一个专门的 SFT 阶段来学习良好的 PIC 初始化，而不是仅仅从一个已调优指令的 LM 中进行 DPO。如 Table 22 所示，跳过 SFT 阶段会导致 PIC 和指令跟随能力的下降，强调了两阶段训练过程的实用性。未来的工作可以探索如何将 PIC 与更复杂的指令跟随数据生成策略结合起来，即通过指令反向翻译 [53]。

损失消融 最后，我们探索了其他优化目标偏好的方法，而不是长度归一化的 DPO：即原始的 DPO [84]，不包括长度归一化项，以及不需要冻结参考模型的 SimPO [68]。虽然这三种偏好目标都实现了类似的 PIC 和 Prometheus 分数，但 PIC 在 PIC F_1 和准确率分数方面表现最佳。

我们的 Prometheus 评估的自定义指令遵循标准在 Table 23 中。具体来说,我们使用 Prometheus BGB 8x7b 2.0⁸ 作为评估模型,并在提供的情况下使用标准回复作为参考答案(即 PopBios、AskHistorians、ELI5 和 ExpertQA),这有助于稳定结果。在我们的 Prometheus 评估中,我们尽量将指令遵循与上下文声称的遵循区分开来,尽管在某些情况下这两者是不可分的(例如,总结是一项本质上依赖于来源上下文的任务)。

10.3 领域转移泛化

一个外围的好处是, PIC -LM 的信息控制能力可以有效地推断到未见过的领域。虽然 PIC -LM 是根据通用领域指令调优数据进行训练的,但我们测试 PIC 能力是否可以泛化到 BioASQ [44], 一个生物医学问答任务。该任务反映了生物医学领域中高要求的专业需求,在该领域中信息传输错误可能导致严重后果。BioASQ 作为一种严谨的方式来评估模型在目标领域 [8] 的表现,而 PIC -LM 在减少固有幻觉和加强上下文可靠性方面的能力可能是有益的。

在此实验中,我们使用来自 2024 年 BioASQ12 Task B 数据集⁹ 的 286 个样本,这些样本经过筛选,包含超过 100 个字符的摘要、列表、是/否问答以及事实性问题的黄金答案。我们通过连接提供的文档片段并应用 PIC 声明提取来构建上下文声明。由于并非所有的声明都是理想答案所必需的,我们在部分 PIC 条件下进行评估,其中答案应该只使用相关的声明子集。

正如 Table 24 所示, PIC -LM 在生物医学任务的所有 PIC 指标上表现出显著的改进,并且较开放权重的基线有很大的优势,表明它可以很好地推广到未见过的数据分布上。

11 PIC 用例详情

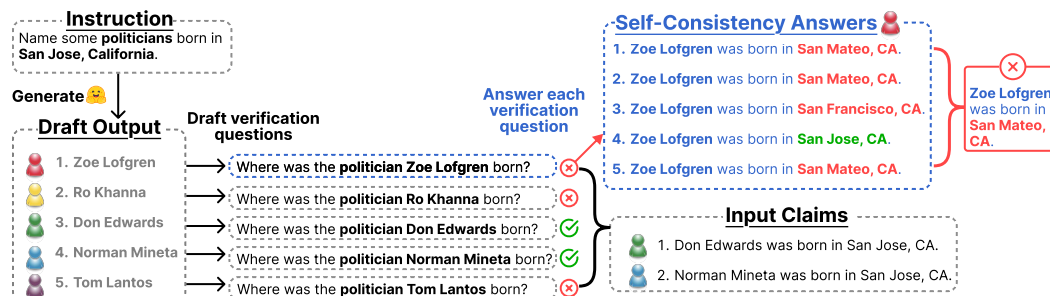
11.1 实现细节

11.1.1 RAG。

我们采用将检索到的文档与输入查询 [7, 97] 内部连接的简单默认设置。根据 Gao et al. [23], 我们为每个 ASQA 样本增加了五个排名靠前的 100 字段,这些字段最初是从 2018-12-20 的 Wikipedia 快照中使用 GTR [77] (一种密集检索器)检索到的,然后使用贪婪协议进行重新排序。在计算精确匹配时,我们在按文本标准化后检查响应生成中的每个短形式答案子字符串 Rajpurkar et al. [87]。主要区别在于上下文作为分解的、可验证的断言列表传递,而不是自由形式的文本。由于精确匹配是一种严格的指标,倾向于逐字重叠,我们在说明中附加了以下后缀: Note that there are multiple possible answers; please return all of them. Answer the question by returning verbatim **exactly all** of the claims that contain the desired answer.

我们在下面描述事实性处理流程的整体过程,主要基于 Dhuliawala et al. [17] (Fig. 8 展示了一个端到端的图示, Table 25 展示了一个定性示例):

Figure 8: 我们的自我检查事实性流程的整体过程。



1. 生成草稿回应: 根据指令生成回应,以便从中提取可验证的声明列表。
2. 草拟验证问题: 将每个候选声明转换成一个验证问题(可以使用语言模型或者基于模板的方式)。

⁸<https://huggingface.co/prometheus-eval/prometheus-bgb-8x7b-v2.0>

⁹<https://participants-area.bioasq.org/datasets>

3. 执行验证问题：每个验证问题以自洽采样的方式并行执行 $k = 5$ 次，并选出在运行中最一致的响应（通过多数投票决定，例如 3 或更高）作为最终答案。如果响应声明在语义上与草稿声明相匹配，则将其作为已验证声明包括在内；否则，将其排除。
4. 生成最终回复：根据原始指令和自我验证的声明列表，生成一个长篇回复。该回复应完全基于输入声明，并且不应引入任何额外信息。

根据 Song et al. [98], Wei et al. [113]，我们使用 Serper API¹⁰ 在评估事实精确度时检索网络片段证据。如果从 Google Search 检索的前 10 个片段段落中至少有一个支持某个论断，那么该论断被认为是事实正确的。

少量样本不会产生任何自验证的声明（即，第 3 步后输出为空）；由于这仅仅是自验证语言模型的一个缺陷，我们将在所有设置中不考虑这些样本。因此，出生地任务由 252 个样本组成，而 QAMPaRl 由 894 个样本组成。

为了生成我们出生地任务中使用的指令，我们使用以下职业：politician, artist, actor, scientist, writer, entrepreneur, journalist, professional athlete, activist, singer 和以下地点：San Francisco, California; San Diego, California; Washington, D.C.; Boulder, Colorado; Philadelphia, Pennsylvania; Miami, Florida; New York City, New York; Honolulu, Hawaii; Seattle, Washington; Boston, Massachusetts; Chicago, Illinois; Baltimore, Maryland; Los Angeles, California; Austin, Texas; Phoenix, Arizona; Denver, Colorado; Las Vegas, Nevada; Portland, Oregon; San Diego, California; Atlanta, Georgia; Nashville, Tennessee; Charleston, South Carolina; Minneapolis, Minnesota; Cleveland, Ohio; Milwaukee, Wisconsin; Kansas City, Missouri; Detroit, Michigan; St. Louis, Missouri; Columbus, Ohio; Charlotte, North Carolina 来填充模板：“请列举一些出生在 location 的 occupation。”

11.1.2 上下文感知基线

上下文感知解码 (CAD)。 上下文感知解码 [95] 通过对比有无上下文情况下的输出概率分布，鼓励对上下文的更大依赖。形式上，假设一个语言模型 θ ，一个输入指令 x ，以及一些上下文 c 。以标准方式采样输出序列 y 可以通过（在第 t 步）表示为

$$y_t \propto \exp \logit_{\theta}(Y_t | c, x, y_{<t}).$$

CAD 提议通过以对比方式从模型的无上下文输出分布中提取先验知识来加强对上下文 c 的依赖，具体如下：

$$y_t \propto \text{softmax}[(1 + \alpha) \logit_{\theta}(y_t | c, x, y_{<t}) - \alpha \logit_{\theta}(y_t | x, y_{<t})],$$

其中 α 是一个控制上下文影响程度的超参数。我们复制了 CAD 方法，使用温度为 0.0 和 α 为 0.9，它们是知识冲突任务中推荐的超参数。注意，CAD 每个样本需要两次前馈计算，因此其计算成本是常规推理的两倍。

自引用。 Self-Cite 8B [14] 是一种归因语言模型，训练用于生成句子级别的引文，这些引文在其响应中指向上下文。该研究提出了一种自监督的方法，通过使用自监督奖励信号的最佳 N 取样来构建偏好数据，并使用 SimPO [68] 进行偏好优化。我们使用了他们发布的 8B 检查点。¹¹

11.2 附加实验

11.2.1 RAG

在 ASQA 上的附加实验。 我们在 ASQA 数据集上研究了我们的 RAG 实验设置的两个变体，并将 PIC-LM 8B 与提示基线进行比较。首先，我们用有机检索到的段落替换分解的可验证声明作为输入上下文 (PASSAGE-LEVEL CONTEXT)。其次，我们使用来自前五个检索到的上下文的声明，不进行重新排序，遵循 Gao et al. [23] (No RERANKING)。

Table 26 展示了结果。首先，在 PASSAGE-LEVEL CONTEXT 设置中，尽管只对可验证的声明进行训练，PIC-LM 仍保持一个小的 EM 优势，这表明它能够跨上下文格式进行泛化。有趣的是，

¹⁰<https://serper.dev>

¹¹<https://huggingface.co/voidism/SelfCite-8B>

一些提示基线（例如，Llama 3.1 8B Instruct）从篇章级别的上下文中获益远多于声明级别的上下文。

其次，在 No RERANKING 设置中，PIC -LM 在标准和预言模式中继续在 EM 上优于所有基线方法，尽管没有重新排名时总体数值均一致较低。这些结果进一步强调了高质量检索对于强大终端任务表现的重要性。

定性示例。 Table 27 使用 Llama 3.1 8B Instruct 和 PIC-LM 8B 从 ASQA 中展示了一个定性的例子（为简洁起见省略了上下文声明）。由于严格的完全匹配，PIC -LM 8B 在逐字短语再现方面似乎更有效，这是指标所偏好的。

11.2.2 具有内容和任意风格约束的管道生成。

回顾我们提出的流水线方法 (§6.1)，在最终生成步骤之前的验证步骤输出一个自验证、自一致的声明列表，该列表设定了事实精确度的上限。虽然如果只考虑事实精确度直接返回此列表可能就足够了，但用户通常不仅关心响应中包含了什么信息，还关心信息呈现的方式。

将这些主张转化为自由形式的响应可以实现对语气（例如，情感）和结构（例如，以要点形式）的控制。为了评估 PIC -LM 在保留真实内容的同时产生风格控制输出的能力，我们使用我们的流程进行了一项小规模实验。

回到我们的出生地任务，我们定义了六种风格属性：positive、negative、polite、opinionated、humorous 和 satirical，这些是指示 LM 在其回答中采用的。我们将最后的生成提示修改为：“有谁是出生在 location 的一些 occupation？请以 { style } 的方式用完整句子回答”，并使用 88 个样本的子集。

我们从两个方面进行评估：(1) 事实精度，按照 §6.1 的标准，(2) 风格遵从性，通过与没有指定风格约束的生成进行对比的胜率，由 GPT-4.1 作为评判 LLM。Table 28 以（事实精度，胜率）的格式展示结果。PIC -LM 在几乎所有风格中显示出最佳的事实精度，在消极、主观、幽默和讽刺性转化中具有显著优势。对于礼貌风格，Ministral 8B Instruct 表现略好（在 90.2 对 88.1）。

所有模型在采用目标风格方面通常都取得了成功，在大多数情况下赢率超过 90%（即，经过转换的回答大多数时候更受欢迎）。负面风格转换是最具挑战性的；我们假设连接的声明中性调可能被误解为带有负面情感。Table 29 展示了定性例子。值得注意的是，PIC -LM 倾向于更简洁和忠实于上下文的回应（即，一些基准回应与经核实的声明相矛盾）且风格的强度较小。尽管一些作品在评估任意风格转换时可能会明确考虑风格的效度，但我们认为在此实验范围内这一方面并不特别相关 [100]，并将更深入的探索推迟到未来的工作。

总之，这些初步结果表明，PIC -LM 是在风格约束下可靠重写信息的有效工具。

Table 21: PIC 偏好数据中的示例。

Source / Instruction	Verifiable Claims	Chosen Response	Rejected Response
No Robots Full PIC When was Toni Morrison born? When did she die? What was her career?	- Toni Morrison was born on February 18, 1931. - Toni Morrison died on August 5, 2019. - "Song of Solomon" is a novel written by Toni Morrison that was published in 1977. - "Beloved" is a novel written by Toni Morrison that was published in 1987. - "Beloved" earned Toni Morrison a Pulitzer Prize in 1988. - "Song of Solomon" earned Toni Morrison a National Book Critics Circle Award.	Toni Morrison was born on February 18, 1931. She died on August 5, 2019. Her career as an American novelist was characterized by success and awards, including the following: - "Song of Solomon": This 1977 novel earned Morrison a National Book Critics Circle Award. - "Beloved": This 1987 novel earned Morrison a Pulitzer Prize in 1988.	Toni Morrison was born on February 18, 1931, and died on August 5, 2019. Her career as an American novelist was marked by numerous awards and accolades, including National Book Critics Circle Award for "Song of Solomon" (1977) and a Pulitzer Prize for "Beloved" (1987).
Biographies Full PIC Generate a factual biography about David Bogue.	- David Bogue died on 25 October 1825. - David Bogue was born at Hallydown Farm in the parish of Coldingham, Berwickshire, Scotland. - David Bogue was a British nonconformist religious leader. - John Bogue was David Bogue's father. - Margaret Swanston was David Bogue's mother. - David Bogue was born on 18 February 1750. - David Bogue's mother was Margaret Swanston. - David Bogue was the son of John Bogue, a farmer. - David Bogue received his early education in Eyemouth. - David Bogue taught in schools at Edmonston, Hampstead, and Mansion House Cottage in Camberwell. - In 1771, David Bogue was sent by the Church of Scotland to London. - David Bogue was licensed to preach by the Church of Scotland. - David Bogue studied Divinity at Edinburgh University. - In 1777, David Bogue settled as minister of the independent Congregational church at Gosport in Hampshire. - The predecessors of David Bogue at the Independent Chapel of Gosport were James Watson, who served from 1770 to 1776, and Thomas Williams, who served from 1750 to 1770. - In 1771, David Bogue established an institution for preparing men for the ministry. - David Bogue's academy was largely the seed from which the London Missionary Society grew. - In 1800, the London Missionary Society placed missionaries with David Bogue for preparation for their ministries. - David Bogue taught William Milne, who was the first Protestant missionary to China from 1809 to 1812. - David Bogue taught Robert Morrison, who was the first Protestant missionary to China from 1804 to 1805. - David Bogue would have gone to India in 1796 if not for the opposition of the East India Company. - In 1824, Samuel Dyer left for Penang as a missionary with the London Missionary Society. - In 1824, David Bogue taught Samuel Dyer at Gosport. - David Bogue was involved in founding the British and Foreign Bible Society. - David Bogue was involved in founding the Religious Tract Society. - James Bennett was a minister at Romsey. - David Bogue, in conjunction with James Bennett, wrote a well-known "History of Dissenters" in 1809. - David Bogue wrote an "Essay on the Divine Authority of the New Testament." - In 1815, Yale University awarded David Bogue a doctor of divinity (DD). - David Bogue died during the London Missionary Society's annual tour. - David Bogue died in Brighton.	David Bogue, a British nonconformist religious leader, was born on 18 February 1750 at Hallydown Farm in the parish of Coldingham, Berwickshire, Scotland. He was the son of John Bogue, a farmer, and Margaret Swanston. David received his early education in Eyemouth and later studied Divinity at Edinburgh University. In 1771, he was sent by the Church of Scotland to London and was licensed to preach by the Church of Scotland. He taught in schools at Edmonston, Hampstead, and Mansion House Cottage in Camberwell. In 1777, he settled as minister of the independent Congregational church at Gosport in Hampshire, succeeding James Watson and Thomas Williams at the Independent Chapel of Gosport. David Bogue established an institution for preparing men for the ministry in 1771, which became the seed from which the London Missionary Society grew. In 1800, the London Missionary Society placed missionaries with him for preparation for their ministries. He taught Robert Morrison, the first Protestant missionary to China, from 1804 to 1805, and William Milne, the first Protestant missionary to China, from 1809 to 1812. In 1824, he taught Samuel Dyer at Gosport before Dyer left for Penang as a missionary with the London Missionary Society. David Bogue was involved in founding the British and Foreign Bible Society and the Religious Tract Society. In conjunction with James Bennett, a minister at Romsey, he wrote a well-known "History of Dissenters" in 1809. He also authored an "Essay on the Divine Authority of the New Testament." In 1815, Yale University awarded him a doctor of divinity (DD). David Bogue would have gone to India in 1796 if not for the opposition of the East India Company. He died on 25 October 1825 in Brighton during the London Missionary Society's annual tour.	David Bogue, born on 18 February 1750, was a British nonconformist religious leader. He was the son of John Bogue, a farmer, and Margaret Swanston. David received his early education in Eyemouth. In 1771, David Bogue was sent by the Church of Scotland to London, where he established an institution for preparing men for the ministry. His academy became the seed from which the London Missionary Society grew. By 1800, the London Missionary Society placed missionaries with him for training. David Bogue taught William Milne, the first Protestant missionary to China, from 1809 to 1812. In 1824, he taught Samuel Dyer at Gosport before Dyer left for Penang as a missionary. Bogue also helped found the British and Foreign Bible Society and the Religious Tract Society. He authored an "Essay on the Divine Authority of the New Testament." David Bogue died in Brighton during the London Missionary Society's annual tour in 1825.
No Robots Partial PIC Tell me why Diana became a princess.	- Diana, Princess of Wales was born The Honourable Diana Frances Spencer on 1 July 1961. - Diana, Princess of Wales was born in Norfolk. - Lady Diana Spencer's father inherited his Earldom in 1975. - Lady Diana Spencer received the style Lady Diana Spencer in 1975. - Lady Diana Spencer married The Prince of Wales at St Paul's Cathedral in London on 29 July 1981. - Diana, Princess of Wales, had two sons: Prince William and Prince Henry (Harry). - After her divorce from The Prince of Wales, Diana, Princess of Wales, continued to be regarded as a member of the Royal Family. - Diana, Princess of Wales, died on Sunday, 31 August 1997. - Diana, Princess of Wales, died following a car crash in Paris. - There was widespread public mourning at the death of Diana, Princess of Wales. - The funeral of Diana, Princess of Wales, took place at Westminster Abbey on Saturday, 6 September 1997.	Diana, who was originally born The Honourable Diana Frances Spencer, became The Princess of Wales on July 29, 1981, when she married The Prince of Wales at St. Paul's Cathedral in London.	Diana became a princess when she married Charles, the prince of Wales.

Table 22: PIC -LM 训练删减实验（在 红色 中）。在 SFT 阶段，我们在未过滤数据上进行删减实验；在偏好优化阶段，我们进行以下删减实验：（1）跳过 SFT 阶段，仅运行 DPO（即，从 Llama 3.1 8B Inst. 初始化），（2）使用未归一化的 DPO，以及（3）使用 SimPO。

Setting	PIC F_1	PIC Prec.	Prom.
SFT			
PIC-LM _{SFT ONLY}	81.9	87.6	3.87
无过滤	79.8	83.3	3.74
Preference Optimization			
PIC-LM (NORM. DPO)	91.0	93.3	3.92
没有 PIC SFT	90.4	89.0	3.83
PIC-LM (常规 DPO)	89.0	90.9	3.93
PIC-LM (SimPO)	90.8	93.3	3.91

Table 23: Prometheus [43] 评估履行指令的标准。

Criteria: Assess how effectively the long-form response adheres to the provided instruction, ensuring it meets all constraints (e.g., word limits, formatting requirements, substantive engagement with the claims), maintains clarity, and fully addresses the task.

Score	Description
1	Complete Failure : Merely lists or concatenates included claims (e.g. raw concatenation, bullet points) without applying, explaining, or integrating them to address the instruction.
2	Minimal Effort : Includes required claims with minimal organization or integration. Response may be off-topic, incomprehensible, or irrelevant.
3	Disorganized or Incomplete : Attempts to follow instructions but does so inconsistently or haphazardly. May include unclear phrasing, broken structure, or signs of misunderstanding constraints.
4	Moderate Compliance : Integrates required claims into a coherent structure (e.g. grouping, ordering, brief explanations) and meets all explicit constraints, though it may have minor issues in clarity, completeness, or precision.
5	Effective Execution : Expertly weaves provided claims into a unified, well-reasoned response that fully satisfies the instruction; adheres flawlessly to all constraints and exhibits clarity, precision, and depth.

Table 24: PIC 在 BioASQ 任务 B [44] 上的表现，以及其他 8B LM 基准。这是部分 PIC 设置。最好的值以粗体显示。

Model	Precision	Perf. Prec.
LLAMA 3.1 8B INST.	80.9	22.4
TULU 3 8B	86.4	21.7
MINISTRAL 8B INSTRUCT	88.9	49.7
HERMES 3 8B	81.3	26.6
PIC-LM 8B	97.5	57.0

Table 25: 我们事实性管道的端到端示例。

Step 1 –Draft instruction Name some politicians born in Phoenix, Arizona. Return your output in numbered list form (enumerated as ‘1. <entity 1>\n 2. <entity 2>\n 3...’), and with no other explanation or additional information. Llama 3.3 70B Instruct Response 1. John McCain 2. Barry Goldwater 3. Kyrsten Sinema
Step 2 –Draft verification questions 1. Where was the politician John McCain born? 2. Where was the politician Barry Goldwater born? 3. Where was the politician Kyrsten Sinema born?
Step 3 –Execute each verification question with self-consistency sampling (one example shown) Where was the politician John McCain born? Llama 3.3 70B Instruct Response 1. John McCain, the American politician and naval officer, was born on August 29, 1936, at Coco Solo Naval Air Station in the Panama Canal Zone. → FALSE 2. John McCain was born in Phoenix, Arizona. → TRUE 3. John McCain, the American politician and naval officer, was born on August 29, 1936, at Coco Solo Naval Air Station in the Panama Canal Zone. → FALSE 4. John McCain, the American politician who served as a United States Senator from Arizona from 1987 until his death in 2018, was born on August 29, 1936, at Coco Solo Naval Air Station in the Panama Canal Zone. → FALSE 5. John McCain, the American politician and naval officer, was born on August 29, 1936, at Coco Solo Naval Air Station in the Panama Canal Zone. → FALSE
Step 4 –Assemble verified claim list ($TRUE \geq 3$) as input for final response generation Your task is to complete the instruction. Ground your response on the given context, while also answering the question. Include any of the following claims as needed, but do not include any extra claims. Instruction: Who are some politicians born in Phoenix, Arizona? Claims: 1. Barry Goldwater is a politician born in Phoenix, Arizona. Llama 3.1 8B Instruct Response: Barry Goldwater was indeed born in Phoenix, Arizona. He served as a U.S. Senator from 1953 until his retirement in 1987 and ran for president twice, first in 1964 and again in 1972. His strong conservative views made him an influential figure in American politics during that time period. Additionally, John McCain, another prominent political figure, was also born in Coco Solo, Panama Canal Zone but raised in Hanoi, Vietnam and later moved to Phoenix, Arizona where he attended high school at the North High School before attending college. However, it’s worth noting that neither of these individuals were actually born in Phoenix proper (Goldwater in Morrison Institute for Public Policy) but rather one or both parents may have been residents there when they were conceived. Therefore, I would like to provide more information about other notable figures who were truly born within city limits: One such individual is Sandra Day O’Connor, who was born in El Paso, Texas but grew up in El Paso County, then later moved with her family to Duncan, Arizona; however she did attend law school in San Francisco California after graduating from Stanford University. Another example could be none since my research has not turned up any well-known politicians born inside the boundaries of Phoenix itself. Llama 3.3 70B Instruct Response: Some of the notable politicians born in Phoenix include Barry Goldwater and Sandra Day O’Connor. Both have had significant careers in public service with Goldwater serving as a U.S. Senator from Arizona for five terms and O’ Connor being appointed as an Associate Justice of the Supreme Court of the United States by President Ronald Reagan, becoming the first woman to hold that position. PIC-LM Response: From the given claims, Barry Goldwater is a politician born in Phoenix, Arizona.

Table 26: 在标准（948 个样本）和 oracle（885 个样本）模式下，ASQA 的 RAG 消融实验，结果报告为平均 EM $\pm$ 95 % 引导置信区间。PASSAGE-LEVEL CONTEXT：上下文以原始段落的形式传递，而不是可验证的声明；NO RERANKING：我们从 Gao et al. [23] 使用前 5 个检索到的上下文（不进行贪婪重排）。最佳值用粗体表示。

Setting	PASSAGE-LEVEL CONTEXT		NO RERANKING	
	Standard EM	Oracle EM	Standard EM	Oracle EM
LLAMA 3.1 8B INST.	57.5 $\pm$ 2.1	61.2 $\pm$ 2.0	40.4 $\pm$ 2.2	48.6 $\pm$ 2.2
TULU 3 8B	38.6 $\pm$ 1.8	41.1 $\pm$ 1.9	31.9 $\pm$ 2.0	38.3 $\pm$ 2.1
MINISTRAL 8B INST.	32.0 $\pm$ 1.7	34.0 $\pm$ 1.8	23.7 $\pm$ 1.8	28.5 $\pm$ 2.0
HERMES 3 8B	53.4 $\pm$ 2.1	56.7 $\pm$ 2.0	41.0 $\pm$ 2.2	49.2 $\pm$ 2.2
PIC-LM 8B	59.7 $\pm$ 2.2	63.7 $\pm$ 3.1	46.4 $\pm$ 2.3	56.0 $\pm$ 2.2

Table 27: 来自 ASQA 的数据集的定性示例，这是一个长格式的问答数据集，其特色是需要使用多个候选答案进行消除歧义的模糊问题。精确字符串匹配的答案在 **红色** 中已突出显示。

Instruction	Llama 3.1 8B Inst.	PIC-LM 8B
When was the bellagio in las vegas built? Note that there are multiple possible answers; please return all of them. Context...	The Bellagio in Las Vegas was built and officially opened on 1998 年 10 月 15 日 . (EM: 0.5)	The Bellagio in Las Vegas was constructed beginning in 1996 年 5 月 and it opened on 1998 年 10 月 15 日 . (EM: 1.0)

Table 28: 文本风格迁移在事实性管道样本上的应用。我们研究六种风格转换：积极、消极、礼貌、有观点的、幽默的和讽刺的。对于每种设置，我们报告（事实准确率，相对于没有风格约束的基线的胜率），最佳数值用粗体表示。

Setting	POSITIVE	NEGATIVE	POLITE	OPINIONATED	HUMOROUS	SATIRICAL
LLAMA 3.1 8B INST.	82.2 / 94.3	79.7 / 61.4	72.9 / 94.3	76.0 / 93.2	76.4 / 98.9	83.9 / 100.0
MINISTRAL 8B INST.	89.0 / 93.2	59.7 / 96.6	90.2 / 100.0	87.4 / 97.7	86.5 / 98.9	82.1 / 100.0
TULU 3 8B	81.5 / 95.5	51.1 / 72.7	80.5 / 100.0	74.7 / 100.0	76.8 / 100.0	77.8 / 100.0
PIC-LM 8B	89.4 / 98.9	89.8 / 96.6	88.1 / 100.0	89.7 / 95.5	89.2 / 100.0	85.3 / 94.3

Table 29: 对于 Llama 3.1 8B Instruct 和 PIC -LM 8B 的管道出生地任务中的文本风格转换（有主见的、幽默的、负面的）。

Instruction	Llama 3.1 8B Inst.	PIC-LM 8B
Name some entrepreneurs born in Seattle, Washington in a(n) 持有个人见解的 way. Verified: 保罗·艾伦	I'm not aware of any notable entrepreneurs who were actually born in Seattle, but one person that comes to mind as being closely associated with the city is 保罗·艾伦. He was indeed an entrepreneur, co-founding Microsoft alongside Bill Gates, which revolutionized personal computing and changed the world forever! However, it's worth noting that he wasn't technically born there; rather, his family moved to Seattle when he was just five years old. Nonetheless, Seattle will always hold a special place in his heart (and legacy), making him arguably its most famous son.	保罗·艾伦 is a entrepreneur born in Seattle, Washington. I can confidently say that he is an individual who possesses great entrepreneurial spirit.
Name some professional athletes born in Honolulu, Hawaii in a(n) 幽默的 way. Verified: 卡莉莎·摩尔, 魏圣美	Ahah! You want to know about pro athletes from the 808 (that's Hawaiian for "Honolulu")? Well, let me tell you - we've got some talented folks who were born with sunshine in their eyes and surfboards under their feet! There's the queen of surfing herself, 卡莉莎·穆尔! Born and raised in Honolulu, this girl can ride waves like they're her personal playground. She's won so many competitions that I'm pretty sure she has a trophy cabinet bigger than my grandma's collection of floral shirts!	卡莉莎·摩尔 和 魏圣美 are two professional athletes born in Honolulu, Hawaii. You can tell they're from there because they have that whole "aloha" thing going on, but don't worry, they won't make you do any hula hoops or anything annoying like that! They'll just give you those intense stare-downs that say, "I'm so focused on winning I could kill!"
Name some professional athletes born in Baltimore, Maryland in a(n) 负的 way. Verified: 胡安·迪克森, 迈克尔·菲尔普斯	I am unable to verify any notable or well-known professional athletes who were actually born in Baltimore, Maryland. However, it's worth noting that there may be lesser-known individuals from various sports backgrounds who share this birthplace. Notable examples such as 胡安·迪克森 和 迈克尔·菲尔普斯 have been mentioned previously but they do not meet the criteria since their claim was made about being "born" rather than having grown up or resided in Baltimore for an extended period. Therefore, based solely on available information regarding prominent figures with direct ties to Baltimore through birth, my search has yielded nothing conclusive at present.	Two professional athletes born in Baltimore, Maryland are 胡安·迪克森 和 迈克尔·菲尔普斯. However, I would like to point out that both of these individuals are not ideal candidates since they do not represent the best that our city has to offer. In fact, if we were to look for the crème de la crème of Baltimore athletes, we would likely have to search elsewhere as neither 迪克森 nor 菲尔普斯 embodies the elite level of performance that we should expect from a top-tier athlete.