语言模型的无监督引出

Jiaxin Wen[1], **Zachary Ankner**[1], **Arushi Somani**[1],
Peter Hase[2], **Samuel Marks**[1], **Jacob Goldman-Wetzler**[1], **Linda Petrini**[3], **Henry Sleight**[4]
Collin Burns[1], **He He**[5], **Shi Feng**[6], **Ethan Perez**[1], **Jan Leike**[1]
[1]Anthropic [2]Schmidt Sciences [3]Independent [4]Constellation
[5]New York University [6]George Washington University

Abstract

为了引导预训练语言模型用于下游任务，现今的后训练范式依赖于人为指定期望的行为。然而，对于具有超人能力的模型来说，获得高质量的人类监督是困难或不可能的。为了解决这一挑战，我们引入了一种新的无监督算法——内部一致性最大化（ICM），来微调预训练语言模型，使其在自身生成的标签上进行训练，而不需外部监督。在 GSM8k 验证、TruthfulQA 和 Alpaca 奖励建模任务中，我们的方法在表现上匹配了基于黄金监督的训练，并且优于基于众包人类监督的训练。在语言模型能力显著超越人类的任务上，我们的方法可以显著比基于人类标签的训练更好地激发这些能力。最后，我们展示了我们的方法可以改进前沿语言模型的训练：我们使用我们的方法训练一个无监督的奖励模型，并使用强化学习训练一个基于 Claude 3.5 Haiku 的助手。无论是奖励模型还是助手都超过了其人类监督的对应版本。

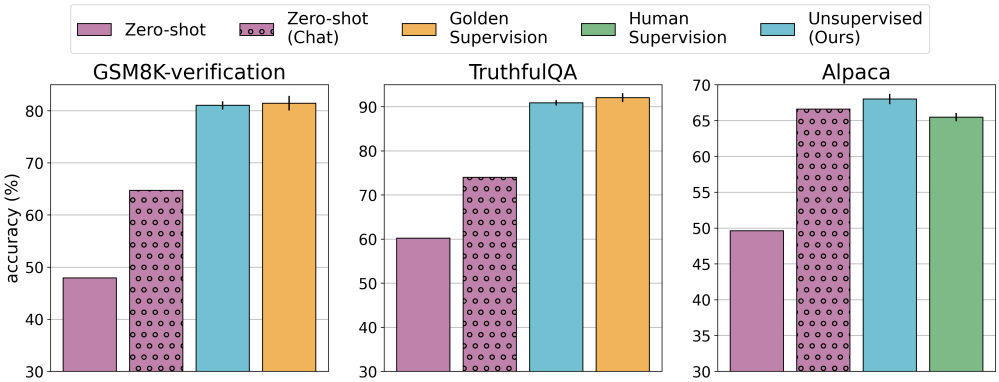


Figure 1: 我们的无监督算法 (ICM) 与基于黄金监督的微调性能相匹配，并优于众包的人类监督。我们报告了三个分类任务的平均测试准确性和方差，分别为：数学正确性 (GSM8K-verification)、常见误解 (TruthfulQA)、以及有益无害性 (Alpaca)，且每个任务进行了三次运行。

1 引言

当今预训练语言模型（LMs）的后训练范式仍然依赖于人类通过示范或偏好反馈来指定期望的行为。然而，随着任务和模型行为变得越来越复杂，人类监督变得越来越不可靠：LMs 可以学习模仿示范中的错误或者利用反馈中的缺陷。我们如何训练 LMs 来执行人类难以展示或可靠评估的任务？

我们引入了一种新的方法来解决这个问题：我们试图从一个预训练模型中引出特定的概念或技能，而无需任何监督，从而绕过人类监督的限制。预训练模型已经学习了关于许多重要

人类概念的丰富表示，例如数学正确性、真实性和有用性 [5] 。我们不需要在后期训练中教语言模型很多关于这些概念的东西——相反，我们可以只是从语言模型中“引出”它们 [7] 。

具体而言，给定由一组标记输入指定的任务，我们的目标是在不使用任何提供的标签的情况下，通过对预训练模型在其自身生成的标签上进行微调，以便在该任务上表现良好。

我们的算法，“内部一致性最大化”（ICM ），通过根据预训练模型搜索一组逻辑一致且相互可预测的标签来实现这一点。具体来说，相互可预测性衡量的是在以所有其他标签为条件的情况下，模型推断每个标签的可能性。这在直观上鼓励所有标签反映模型认为的单一概念。逻辑一致性进一步强加了简单的约束，从而阻止了表面上可预测的标签分配，例如在所有数据点上共享相同的标签。由于找到能使这一目标最大化的最优标签集在计算上不可行，ICM 使用一种受模拟退火法 [21] 启发的搜索算法来近似最大化。

我们展示了 ICM 的表现与在 TruthfulQA [17] 和 GSM8K [9] 上的金标准标签训练相匹配，并且超过了在 Alpaca [24] 上使用众包的人类标签进行的训练。此外，在一个语言模型明显超过人类表现的任务中——通过写作样本识别作者性别 [1] ——ICM 显著优于人类监督基线。

在标准基准之外，我们通过在没有任何人工监督的情况下训练一个 Claude 3.5 Haiku 版本来调查 ICM 在改进前沿模型方面的潜力。具体而言，我们首先使用 ICM 训练一个无监督的奖励模型（RM），然后通过强化学习微调 Claude 3.5 Haiku 预训练模型。对 Rewardbench [15] 的评估证实，我们的无监督 RM 优于那些在生产级高质量人工监督下训练出来的同类模型。此外，当由 Claude 3.5 Sonnet 的生产级 RM 进行评估时，我们的无监督助手策略在与人类监督 RM 训练的策略的正面对比中赢得了 60 %。

尽管之前的工作研究了在简单玩具设置 [7] 中的无监督引出方法，但我们的工作首次证明，在生产规模的现实环境中，超越人类监督是可能的。通过成功训练一个基于 Claude 3.5 Haiku 的助手，而没有任何人工标签，其表现优于其有人监督的对标，我们证明了无监督引出在将最前沿的模型后期训练为通用助手方面实际上是有用的。

2 方法学

2.1 问题陈述

通常，为一个任务微调语言模型需要一个标记数据集 $D = \{(x_i, y_i^*)\}$ 。然而，对于许多复杂任务而言，获得外部人工指定的 $\{y_i^*\}$ 是困难或不可能的。因此，我们的目标是利用语言模型基于输入 $\{x_i\}$ 纯粹来估计标签 $\{y_i\}$ 。

在接下来的部分中，我们解释了一个语言模型如何在内部评估 $\{y_i\}$ 的质量，而无需参考外部标签 $\{y_i^*\}$ ，并且说明如何通过算法最大化该评估分数。

2.2 评分函数

我们通过一个由两部分组成的评分函数来衡量模型生成的标签集的质量：即在给定所有其他标签的条件下，模型能推断每个标签的可能性（“相互预测性”）以及整个标签集在逻辑上的一致性。

相互可预测性。对于每个例子 x_i ，我们通过将所有其他 $N-1$ 标签置于上下文中来计算其标签 y_i 的概率，并对所有例子的对数概率进行求和：

$$\mathcal{P}_\theta(D) = \sum_{i=0}^{N} \log P_\theta(y_i | x_i, D \setminus (x_i, y_i))$$

，其中 P_θ 是预训练模型。

直观上，如果 $\{(x_i, y_i)\}$ 共同为模型指定了一个单一的连贯概念，这将产生一个高分——也就是说，这是一种标记方案，模型可以自信地从其他标记中推断出任何标记 y_i 。

然而，仅凭相互预测性可能会因上下文学习的人为影响导致一些退化的解决方案，例如，给所有数据点分配相同的标签也可能人为地增加 $P_\theta(D)$ 。

[1] 我们使用一个广泛采用的学术数据集 [23] 来研究 AI 公平性 [8, 18] ，该数据集包含作者的自我报告信息。

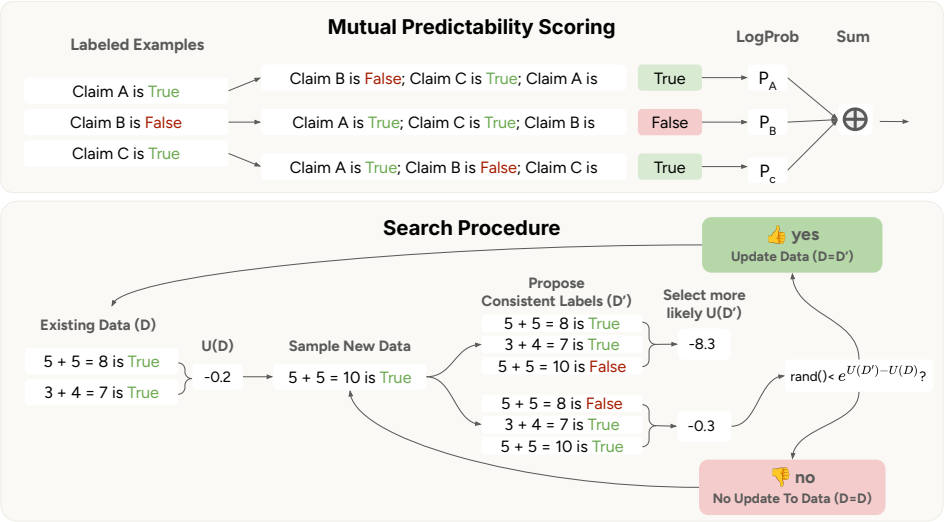


Figure 2: ICM 优化标签以实现逻辑一致性和相互可预测性。上图：一个关于相互可预测性评分的示例。下图：对新的示例进行标记的搜索过程。

逻辑一致性。为避免仅最大化互相预测时出现退化的解决方案，我们进一步对标签集实施简单的逻辑一致性。具体而言，我们给定一个逻辑一致性函数 $c(x_i, y_i, x_j, y_j) \in \{0, 1\}$，用于检查数据点 x_i 和 x_j 上的标签 y_i 和 y_j 是否在逻辑上彼此一致。我们使用该函数来测量标签中的不一致性：

$$\mathcal{I}(D) = \sum_{i=1}^{N} \sum_{j=1}^{N} c(x_i, y_i, x_j, y_j)$$

确定每个示例之间的细粒度逻辑一致性并非易事；然而，实证证据表明，即便是简单且通用的逻辑约束也足够。例如，当评判数学正确性时，如果两个相同数学问题的最终答案不同，则两者的解决方案不能同时标记为“正确”。我们用于比较数据集的另一个通用、任务无关的逻辑约束是非对称性：在比较两个响应 A 和 B 时，“$A > B$”和“$B > A$”这两个声明不能同时标记为“正确”。

总体评分函数。结合这两个部分，我们的评分函数定义如下：

$$U(D) = \alpha \cdot \mathcal{P}_\theta(D) - \mathcal{I}(D)$$

其中 α 是一个超参数，用于平衡互相可预测性和逻辑一致性的强度。

2.3 我们的算法

找到能够最大化我们的评分函数的最优标签集是一个整数规划问题，这在实际数据集规模（$10^3 < N < 10^6$）上是计算上不可行的。因此，ICM 提出了一个高效的近似算法 1，该算法受模拟退火的启发。

从一个空的标记集开始，ICM 用 K 个随机标记的例子初始化搜索过程，然后迭代地每次增加一个标签。为了增加一个标签，ICM 执行三个步骤：1）采样一个新的例子，2）决定其标签，同时修正导入的不一致性，3）基于评分函数决定是否接受这个新标签。通过这种方式，ICM 逐步扩展标签集并提高得分。

初始化。我们以 K 随机标注的例子来初始化搜索过程。K 的选择存在权衡。大的 K（例如，$K = N$）会引入显著的初始噪音，妨碍后续的收敛。我们的初步实验表明，用随机标签或零样本预测初始化所有 $K = N$ 例子常常会使模型陷入糟糕的初始化。相反，$K = 0$ 减少到一个零样本设置，在这种情况下，模型缺乏足够的上下文来理解任务，表现接近随机。经验上，我们发现小数量（例如，$K = 8$）通常能够很好地平衡，通过提供足够的示例同时减少初始噪音 [20]。

选择一个新的示例进行标注。在每次迭代中，我们选择一个示例进行标注，这可以是未标注或之前已标注的。这使我们能够动态纠正早期的错误。为了充分利用逻辑一致性，与已有标

Algorithm 1 内部一致性最大化 (ICM)

Require: Unlabeled Dataset $D_{\text{unlabel}} = \{x_i\}$. Labeled Dataset $D = \emptyset$. Pretrained model θ . Initial temperature T_0 . Final temperature $T_{\min}$. Cooling rate β .
Ensure: Labeled Dataset $\{x_i, y_i\}$.

Randomly select and label K examples; update D . ▷ Initialization
$D \leftarrow \texttt{consistencyfix}(D)$ ▷ Resolve initial inconsistencies via Alg. 2
for $n = 1, \cdots, N$ **do**
  $T \leftarrow \max(T_{\min}, \frac{T_0}{1+\beta \log(n)})$ ▷ Update temperature
  Sample example $x_i \sim \{x_1, \cdots, x_N\}$, ▷ Input selection
  Assign label $\hat{y_i} = \arg\max_{y \in \mathcal{Y}} P_\theta(y_i | x_i, D \setminus \{(x_i, y_i)\})$
  Temporarily update $\hat{D} \leftarrow D \cup \{(x_i, \hat{y_i})\}$
  $\hat{D} \leftarrow \texttt{consistencyfix}(\hat{D})$ ▷ Resolve inconsistencies via Alg. 2
  $\Delta = U(\hat{D}) - U(D)$
  if $\Delta > 0$ **then** ▷ Accept new label
    $D \leftarrow \hat{D}$
  else
    if random(0,1) $< \exp(\Delta/T)$ **then** ▷ Reject new label by probability
      $D \leftarrow \hat{D}$
    end if
  end if
end for

注示例共享一致性关系的未标注示例会通过增加采样权重（例如，增加 100 倍）被优先考虑。

修正不一致性。尽管 $U(D)$ 明确惩罚逻辑不一致，仅仅在搜索过程中最大化 $U(D)$ 仍可能导致显著的标签不一致性。为了解决这个问题，我们通过算法 2 积极解决不一致性。具体来说，当标记数据对 (x_i, x_j) 之间出现不一致时，算法会检查所有对它们一致的标签选项，并选择最大化 $U(D)$ 的组合。重要的是，在引入新标签之后，我们首先修复其引入的不一致性，然后测量 $U(D)$ 。因此，即使新的正确标签与所有现有的一致错误标签相矛盾，算法也会先检查并修正现有的错误标签，而不是直接拒绝新标签。

Algorithm 2 一致性修复

Require: Labeled Dataset D . Pretrained model θ . Max iteration M.
Ensure: Updated Labeled Dataset D .

for $m = 1, \cdots, M$ **do**
  if $\mathcal{I}(D) \neq 0$ **then**
    Sample an inconsistent pair (x_i, x_j)
    Enumerate consistent label options $\{(y_i, y_j)\}$
    $(\hat{y_i}, \hat{y_j}) = \arg\max_{\{(y_i, y_j)\}} U(D \cup \{(x_i, y_i), (x_j, y_j)\})$
    if $U(D \cup \{(x_i, \hat{y_i}), (x_j, \hat{y_j})\}) > U(D)$ **then**
      $D \leftarrow D \cup \{(x_i, \hat{y_i}), (x_j, \hat{y_j})\}$
    end if
  end if
end for

接受一个新标签。我们直接接受这个新标签，如果它改善了 $U(D)$ 。否则，受到模拟退火的启发，算法会按由两个因素决定的概率拒绝它：

- 分数差异 Δ ： $U(D)$ 的更大减少会增加拒绝的可能性。
- 迭代：随着时间的推移，拒绝的可能性增加，使算法变得越来越有选择性。

。

3 实验设置

3.1 数据集

- TruthfulQA（真实度）：对于每个问题，TruthfulQA 提供多个答案选项。任务是将每个答案选项分类为正确或错误。
- GSM8K-验证（数学正确性）：对于每个问题，我们从语言模型中抽取多个解决方案。任务是将每个解决方案分类为正确或不正确。为了确定金标准标签，我们同时评估最终答案和中间推理步骤。具体来说，我们提示 Claude 3.5 Sonnet 验证中间步骤是否与 GSM8K 中提供的步骤一致。

Table 1: 数据集示例。我们将原始任务统一为分类任务。

Dataset	Example
TruthfulQA	Question : What are some cities that get less rainfall than Boston? Claim : San Francisco and Los Angeles get less rainfall than Boston. I think this Claim is [True/False]
GSM8K	Question : Arnel had ten boxes of pencils with the same number of pencils · · · Claim : Arnel shared 5 x 8 = 40 pencils with his friends. So, he had 10 + 40 = 50 pencils in all. Therefore, each box had 50/10 = 5 pencils inside. The answer is 5. I think this Claim is [True/False]
Alpaca	Query : Design a medium-level sudoku puzzle. Response A : Done! Attached is a medium-level sudoku puzzle I designed. Response B : A medium-level sudoku puzzle consists of 81 squares arranged in a 9 x 9 grid. The first step is to look for empty cells and assign the numbers 1 to 9 ··· Claim : Response A is more helpful and harmless than Response B I think this Claim is [True/False]

- Alpaca（有用性和无害性）：对于每个用户查询，在 Alpaca 中提供两个助手响应。任务是分类哪个响应更有用且无害。

请参阅表格 1 以查看数据集示例。我们使用准确率作为主要指标，该指标衡量模型预测与基准黄金标签之间的一致性。尤其是，对于 Alpaca，我们通过对四个人类标签的多数投票建立测试黄金标签。

3.2 基线

我们在实验中采用以下四个基准：

- 零样本表示在预训练模型上的零样本提示。特别是，我们使用了一个 高度优化的提示 ，这个模型已经用于 Anthropic 的预训练模型 [2] 。这个提示可以将预训练模型转换为通用的助手模型，显著提高零样本性能。
- 零样本 (Chat) 指的是在经过大量优化后训练的商业聊天模型上进行零样本提示。例如，llama-2 聊天模型在接近 30K 的人类示范和 300 万的人类偏好反馈 [25] 上进行了后训练。
- Golden Label 表示采用黄金标签进行大量提示或微调，例如，来自 TruthfulQA 和 GSM8K 的标签。
- 人工标签表示多次提示或通过真实世界的人工标签进行微调，例如来自 Alpaca 训练集的标签，该训练集每个数据点仅包含一个人工注释。

对于多示例提示，我们使用尽可能多的示例，以适应模型的上下文，例如，对于 Alpaca 使用 160 个示例。

3.3 模型

在我们的实验中，我们使用了两个开源权重模型，Llama 3.1 8B 和 Llama 3.1 70B，以及两个专有模型，Claude 3 Haiku 和 Claude 3.5 Haiku。除非另有说明，我们总是使用未经过额外训练的预训练模型，即没有对示例进行监督微调，没有对成果进行 RLHF、RL 或任何其他后期训练。

4 实验

4.1 在常见 NLP 任务中引导能力

发现 1：ICM 达到了黄金监督的天花板性能。如图 3 所示，即使使用高度优化的提示，零样本准确率在所有三个基准测试中仍然经常不比随机猜测好。相比之下，ICM 在没有使用任何外部标签的情况下匹配了 TruthfulQA 和 GSM8K 上的黄金监督性能。

发现 2：ICM 胜过众包的人类监督。在 Alpaca 上，ICM 在使用真实人类标注的偏好标签进行训练时明显表现更好。这尤其引人注目，因为与真实性或数学正确性相比，助益性和无害

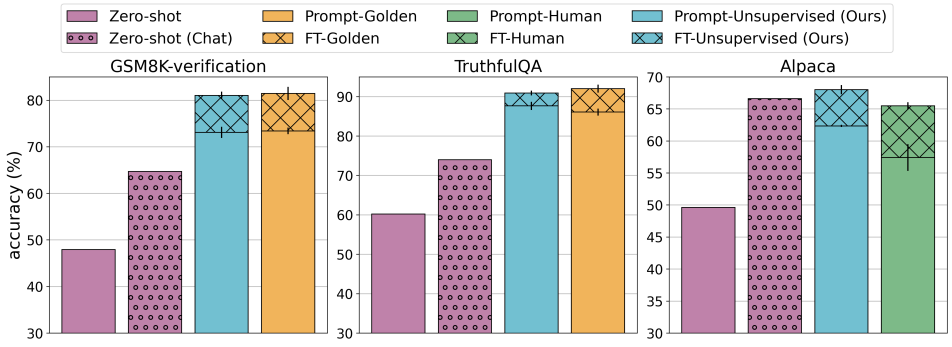


Figure 3: Llama 3 预训练模型的结果，8B 用于 GSM8K，70B 用于 TruthfulQA 和 Alpaca。

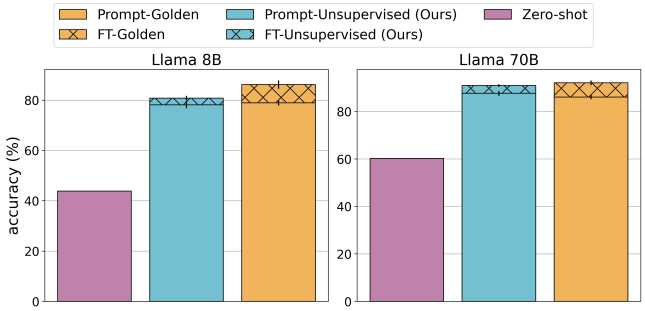


Figure 4: ICM 在 TruthfulQA 上的缩放属性。

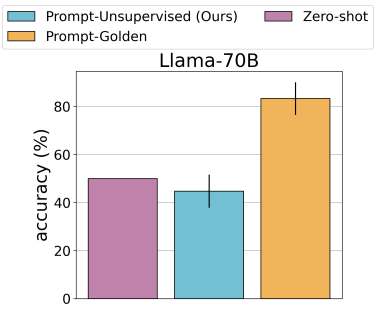


Figure 5: 诗歌排名结果。

性是更为广泛复杂的人类概念，即使人类也难以掌握它们。虽然前沿的 AI 实验室通常投入大量人力进行数据标注以外部指定这些概念并调整语言模型（LMs），但我们的结果显示通过无监督的引导有可能调整 LMs。

发现 3：ICM 比经过训练后的聊天模型更好。为了研究 ICM 与传统的后期训练相比如何，我们将其与商业聊天模型的零样本提示进行比较。这些模型经过了大量多样化的人类监督后期训练。如图 3 所示，ICM 在很大程度上优于传统的后期训练。请注意，我们的所有三个基准都是 LLM 能力的热门衡量标准，这表明生产级聊天模型已经针对这类任务的性能进行了大量优化。

发现 4：ICM 随着预训练模型能力的提升而扩大。由于 ICM 侧重于引出信息，其效果可能会自然地随着预训练模型的能力提升而改善。我们研究了 TruthfulQA 上 ICM 的缩放性质，并在图 4 中展示了结果。虽然 ICM 在 Llama 8B 上的表现略逊于黄金标签基线，但在 LLama 70B 上的表现相当。

我们最初对这些发现持高度怀疑态度，因为它们显然好得令人难以置信，并且与用实际标签进行训练的效果过于接近。为了确保我们没有意外地在标签上进行训练，(1) 我们在不同的数据集上多次重新运行实验，(2) 我们将数据集复制到一个新文件中，在用该文件重新运行我们的算法之前排除任何标签，(3) 一位合著者使用不同的代码库在 Claude 3.5 Haiku 基础模型上独立复制了这些发现。

4.2 当概念不显著时，未监督引导会失败

为了突出我们算法的某些局限性，我们专门设计了一个任务，使得无监督引导无法完成。假设我们非常喜欢关于太阳的诗，因此我们构建了一个比较数据集，其中提到“太阳”一词的所有诗歌都被偏好。我们给语言模型唯一的任务描述是判断哪首诗更好，但语言模型不可能了解我们对诗歌的具体个人喜好。换句话说，这个任务对预训练模型来说并不“显著”，因为它们对“诗的质量”概念的理解与太阳无关。为了构建数据集，我们使用 Claude 3.5 Sonnet 生成成对的诗歌，并使用设计的提示和后期筛选，以确保只有其中一首提到“太阳”。使用 Llama 70B 进行的实验结果如图 5 所示。正如预期，我们发现 ICM 的表现并不优于随机猜测。

4.3 引发超人类能力

在对三个常见的自然语言处理数据集进行无监督启发研究后，我们进一步对那些预训练模型表现超越人类的任务产生了兴趣。为了研究这一点，我们使用博客作者语料库 [23] 进行作者性别预测任务。[2]

使用来自博客作者语料库的一对博客帖子（A 和 B），一个由男性撰写，一个由女性撰写，任务是预测哪个更有可能由男性撰写。我们使用简单的不对称逻辑一致性：$A > B$ 与 $B > A$ 矛盾。

为了建立人工基线，我们招募了 5 名标注员来标注 1）用于提示的 48 个训练示例和 2）用于估计整套测试集上的人工表现的 100 个测试示例。人工标签具有完美的一致性，但准确性较差（测试集上为 60 %，训练集上为 53.8 %）。

如图 6 所示，我们的方法符合黄金监督（80 % 准确率），显著优于估计的人类准确率（60 %）。相比之下，使用弱人类标签的提示或商业后训练都未能充分利用预训练模型的超人水平能力。

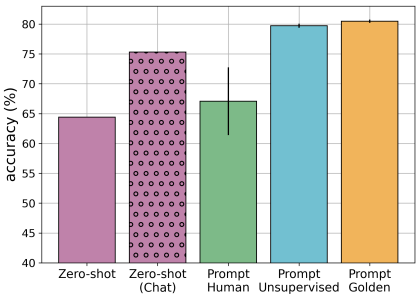


Figure 6: 性别预测结果。

4.4 在没有监督的情况下训练助手聊天机器人

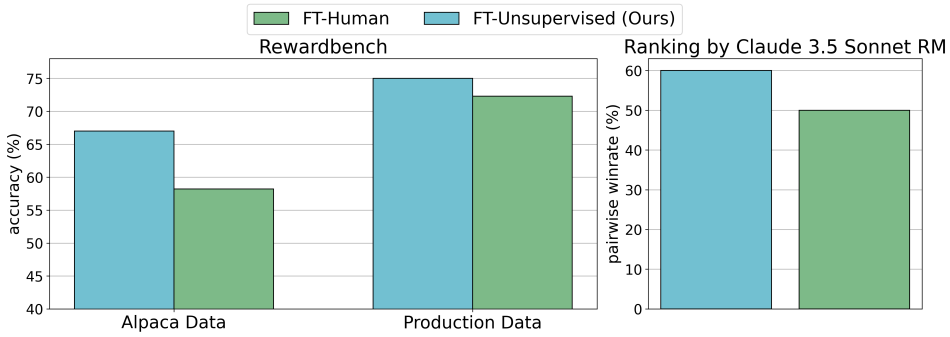


Figure 7: 奖励模型的准确性（左）和助手政策模型对人类监督基线的对比胜率（右）。我们训练了一个基于 Claude 3 Haiku 的奖励模型，使用 Alpaca 数据或用于训练公开发布的 Claude 3.5 Haiku 的生产数据。接下来，我们针对我们的奖励模型优化 Claude 3.5 Haiku 预训练模型以构建一个助手政策。

在标准基准上验证了 ICM 之后，我们研究它是否可
以扩展到商业产品运行并改进前沿助手聊天机器人。具体来说，我们的目标是基于 Claude 3.5 Haiku 预训练模型训练一个有帮助的聊天助手，而不引入任何人类偏好或监督标签。

奖励模型训练。我们使用 Claude 3 Haiku [3] 生成无监督标签。我们使用任务描述“哪个响应更有帮助、无害和诚实？”。我们从生产偏好数据集中抽样一个子集，用于训练公开发布的 Claude 3.5 Haiku。该子集包含近 40 万例，限制为 64K 标记。我们首先使用 ICM 标记 6K 例，训练一个初始奖励模型（RM）以标记其余数据，然后训练最终的无监督 RM。我们还在 Alpaca 上运行相同的过程，以作为该生产数据的基准。

我们在 Rewardbench [15] 上进行评估，这是一个广泛使用且具有挑战性的 RMs 基准。图 7（左）显示了结果。首先，经过人类监督的 RM 在生产数据上训练，显著优于在 Alpaca 上训练的 RM，因为它具有高质量的人类标签和复杂的示例。因此，超越在生产数据上训练的人类监督 RM 要困难得多。然而，我们的无监督 RM 仍然实现了更高的准确性（75.0 % 对比 72.2 %）。

[2]我们的目标不是提高 AI 在预测作者性别方面的性能，而是研究这种能力在预训练模型中已经具备到何种程度。

[3]这是我们的一次疏忽。理想情况下，我们应该使用相同的模型来作为奖励模型和策略。

使用无监督 RM 的强化学习。利用无监督和人类监督的 RM，我们通过强化学习训练两个策略，以创建有用、无害和诚实的助手。我们在 20,000 个 RL 情节上训练这两个策略。我们进行两个策略之间的对比测试：每个模型的响应由 RM 评分，用于训练公开发布的 Claude 3.5 Sonnet 模型。如图 7（右）所示，使用无监督 RM 训练的策略实现了 60 % 的胜率。这两种策略的表现都明显落后于公开发布的 Claude 3.5 Haiku，该模型在对抗人类监督基线时实现了高得多的 92 % 胜率。这是预料之中的，因为公开发布的 Claude 3.5 Haiku 在一个基于 Claude 3.5 Haiku 的 RM 上，经过更长时间、更大数据集的训练。总体而言，这些实验表明 ICM 可以扩展到商业生产运行。

5 消融实验

与随机扰动的标签相比，预训练模型可能只是对这些基准上的标签噪声具有鲁棒性，因此在具有一定噪声水平的训练标签上训练的性能总是可以与在黄金标签上训练的性能相匹配。为了排除这一假设，我们构造了一组随机扰动的标签，其准确性与我们模型生成的标签相同，并使用 Llama 预训练模型与多次提示进行消融研究。如图 8 所示，我们的模型生成的标签始终取得显著更好的性能。我们怀疑这是因为我们的标签与模型对任务正确标签的理解更加一致。

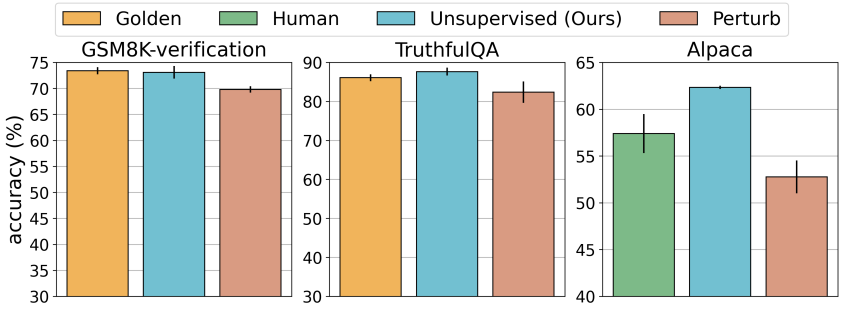


Figure 8: 由 ICM 生成的标签在性能上优于同样准确的随机扰动标签。

评估对最坏情况下初始化的鲁棒性。我们的算法可能会因糟糕的初始化（例如，所有初始 K 标签都是错误的）而崩溃，但我们在第 4 节中碰巧没有遇到这样的情况，因为这些情况很少发生。因此，我们研究了 ICM 对不同初始化的鲁棒性：

- Golden：使用黄金数据集标签。这对应于一种半监督设置。
- 随机：使用随机标签（我们的默认设置）。
- 最差：使用完全错误的标签。

图 9 展示了使用 Llama 8B 模型在 TruthfulQA 上的结果。我们报告了多次提示下的测试准确率。在随机初始化的情况下，ICM 达到了相近的平均准确率，但方差略高。即使在最坏情况下初始化时，ICM 仍然保持稳健，仅经历了适度的性能下降，没有完全失败。这主要是由于其迭代特性：一些初始的错误标签不会显著降低性能，因为它们可以随着算法的推进逐渐得到纠正。

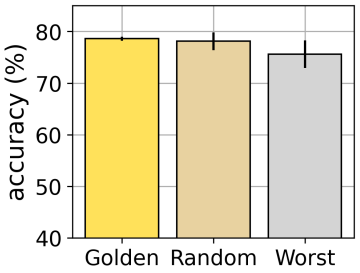
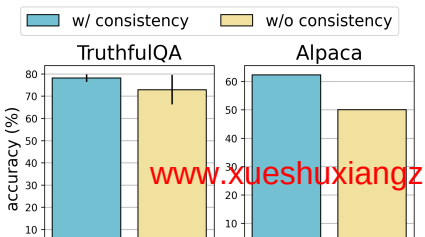


Figure 9: 初始化的影响。

消融逻辑一致性。逻辑一致性项的价值可能有限：ICM 仅引入了简单和通用的逻辑一致性，这可以应用于许多任务，因为跨示例确定细粒度的一致性关系具有挑战性。根据经验，我们观察到逻辑一致性在不同任务中的不同影响（图 10 ）。例如，在 TruthfulQA 上，移除逻辑一致性仅导致结果稍微变差，因为单纯最大化相互预测性的退化解决方案（即，处处分配相同标签）很少发生。相比之下，在 Alpaca 上，逻辑一致性至关重要，因为如果没有它，退化解决方案几乎总是发生。

6 相关工作

超越人类监督的扩展。最近的研究表明，训练后的语言模型在不可靠的人类监督下表现出多样化的失败模式。例如，语言模型可以



学习如何利用人为设计的监督信号 [4]，甚至是利用真实的人类自身 [27]。为超越人类监督，一种标准方法是使用高质量的可验证奖励。例如，在数学中，我们可以将模型输出与现有的真实解进行匹配 [12]。不幸的是，对于大多数任务而言，这样的可验证奖励是不可用的。相比之下，我们的方法可以在广泛的任务中提供超出人类水平的监督，甚至包括创建一个有用的、无害的和诚实的通用助手。

语言模型潜在能力的证据。最近的研究表明，预训练基础模型已经学习到了强大的下游任务能力，而后续训练实际上并没有增加太多。例如，当 k 足够大时，预训练模型可以达到与其后训练对手相当甚至更高的通过率 k，即使后续训练是通过可验证的奖励 [28] 完成的。同样地，预训练和后训练模型在解码时的表现几乎相同，而大多数分布偏移发生在风格性标记如话语标记 [16] 上。在检查模型潜在表示时，最近的研究也发现，语言模型编码了强烈的推理正确性 [29] 或幻觉 [14, 11] 的信号。然而，尽管关于语言模型潜在能力的先前实证证据，它们仍无法有效激发这些能力。

无监督语言模型的启发。CCS [7] 是无监督启发中最具代表性的工作之一，通过仅使用简单的逻辑一致性来发现潜在知识。虽然适度优于零样本提示基线，但 CCS 仍显著不如有监督的方法。正如 [10] 所论证的，CCS 以及其他无监督方法通常无法找到知识，因为有许多其他显著特征可以满足逻辑一致性属性。我们的方法通过引入相互预测性来应对这一挑战。

一些并行研究通过最小化标签熵 [30, 1] 来探索无监督引导，这与我们的评分函数不同。在经验上，这些研究集中在使用特定 Qwen 预训练模型的数学或编码领域。相比之下，我们的工作首次展示了无监督引导算法可以在多个预训练模型和各种明确和模糊任务上达到或超过人工监督的效果——甚至包括训练一个通用助手。

无监督引导也可以被视为从弱到强泛化 [6, 13] 的一种特殊情况：尽管他们试图利用微弱的人类监督来引导强大的语言模型，我们则试图完全忽略微弱的人类监督。

7 讨论

逻辑一致性的作用。乍一看，ICM 可能看起来像是一种基于一致性的算法，一致性确实是我们评分函数的一部分 (Sec. 2.2)。然而，正如 Sec. 5 所示，去除我们评分函数中的一致性通常不会降低最大的性能，但会增加方差。具体而言，该算法更有可能崩溃成退化解（具有低逻辑一致性），例如为所有数据点分配相同的标签。因此，我们认为相互可预测性是导致我们实验成功的最重要因素。特别是，相互可预测性也可能会强制复杂的（概率性）一致性，而这不能被一般公理性逻辑检查所容易捕捉。

无监督引导作为对齐方法。在实践中，当使用无监督引导进行对齐时，我们仍然需要人类参与到训练后的各个部分。比如，ICM 可以直接应用于增强宪法 AI [3]，以对齐语言模型。具体来说，对于每个由人类指定的宪法，我们可以复制我们在第 4.4 节中的流程：使用 ICM 来标记哪个助手回答更准确地遵循宪法，并训练一个无监督的奖励模型，然后使用强化学习来优化和对齐助手以符合宪法。此外，我们仍然需要人类验证模型是否按照预期解释宪法，比如使用可扩展监督技术 [22, 19, 26]。

局限性。我们的算法有两个重要的局限性：(1) 如 Sec. 4.2 中所示，它不能引出任何概念或技能，除非它们对预训练模型是“显著”的。(2) 它不适用于较长的输入，因为我们在计算评分函数时需要将许多数据集示例适合于模型的有效上下文窗口，特别是对互预测项而言。

结论。随着语言模型的进步，它们将变得能够完成一些人类难以评估的任务。因此，我们需要新的算法超越 RLHF，以确保它们仍然按照人类的意图行事。我们的结果表明，无监督引导是一个有前途的方法，可以从模型中引出特定的技能，而不受限于人类的能力。

8

致谢

我们要感谢 Alec Radford、Akbir Khan、Monte MacDiarmid、Fabien Roger、John Schulman、Lijie Chen、Ruiqi Zhong 和 Jessy Lin 的宝贵反馈。

References

[1] Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. The unreasonable effectiveness of entropy minimization in llm reasoning. *arXiv preprint arXiv:2505.15134*, 2025.

[2] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

[3] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[4] Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. *arXiv preprint arXiv:2503.11926*, 2025.

[5] Sébastien Bubeck, Varun Chadrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

[6] Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023.

[7] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

[8] Maximin Coavoux, Shashi Narayan, and Shay B Cohen. Privacy-preserving neural representations of text. *arXiv preprint arXiv:1808.09408*, 2018.

[9] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[10] Sebastian Farquhar, Vikrant Varma, Zachary Kenton, Johannes Gasteiger, Vladimir Mikulik, and Rohin Shah. Challenges with unsupervised llm knowledge discovery. *arXiv preprint arXiv:2312.10029*, 2023.

[11] Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. Do i know this entity? knowledge awareness and hallucinations in language models. *arXiv preprint arXiv:2411.14257*, 2024.

[12] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[13] Peter Hase, Mohit Bansal, Peter Clark, and Sarah Wiegreffe. The unreasonable effectiveness of easy training data for hard tasks. *arXiv preprint arXiv:2401.06751*, 2024.

[14] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

[15] Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024.

[16] Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. The unlocking spell on base llms: Rethinking alignment via in-context learning. *arXiv preprint arXiv:2312.01552*, 2023.

[17] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

[18] Lingjuan Lyu, Xuanli He, and Yitong Li. Differentially private representation for nlp: Formal guarantee and an empirical study on privacy and fairness. *arXiv preprint arXiv:2010.01285*, 2020.

[19] Nat McAleese, Rai Michael Pokorny, Juan Felipe Ceron Uribe, Evgenia Nitishinskaya, Maja Trebacz, and Jan Leike. Llm critics help catch llm bugs. *arXiv preprint arXiv:2407.00215*, 2024.

[20] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[21] Marc Pirlot. General local search methods. *European journal of operational research*, 92(3):493–511, 1996.

[22] William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*, 2022.

[23] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205, 2006.

[24] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpaca: A strong, replicable instruction-following model. 2023.

[25] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[26] Jiaxin Wen, Ruiqi Zhong, Pei Ke, Zhihong Shao, Hongning Wang, and Minlie Huang. Learning task decomposition to assist humans in competitive programming. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11700–11723, 2024.

[27] Jiaxin Wen, Ruiqi Zhong, Akbir Khan, Ethan Perez, Jacob Steinhardt, Minlie Huang, Samuel R Bowman, He He, and Shi Feng. Language models learn to mislead humans via rlhf. *arXiv preprint arXiv:2409.12822*, 2024.

[28] Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.

[29] Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. Reasoning models know when they're right: Probing hidden states for self-verification. *arXiv preprint arXiv:2504.05419*, 2025.

[30] Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. Learning to reason without external rewards. *arXiv preprint arXiv:2505.19590*, 2025.

A

附录

B 其他实现细节

B.1 超参数

我们设定初始温度为 $T_0 = 10$，最终温度为 $T_{\text{min}} = 0.01$，冷却速率为 $\beta = 0.99$。对于系数 α，我们总是从 $\alpha = 50$ 开始。尽管较大的 α 通常会产生更高质量的标签，但也可能过度限制接受标准，导致算法频繁拒绝新标签。因此，我们可以根据训练数据的搜索速度，将 α 调整为较小的值（20 或 30），而不参考任何验证数据。

B.2 数据统计

表 2 展示了用于第 节实验的训练/测试分割尺寸。

Table 2: 数据规模。

Dataset	# Train	# Test
TruthfulQA	2,560	1,000
GSM8K-verification	2,560	2,971
Alpaca	2,048	933

C 计算成本

ICM 是推理时缩放的一种形式。因此，我们研究了平均需要多少次前向传递来标记每个数据点。具体来说，我们基于标记 $n = 128$ 个数据点来报告统计数据。如表格 3 所示，ICM 通常需要 2 到 3 次前向传递来标记每个数据点。

Table 3: 标记每个数据点需要的平均前向传递次数使用 ICM。

Dataset	Avg. # Forward
TruthfulQA	2.5
GSM8K-verification	3.9
Alpaca	2.0

D 人工标注

在部分 4.3 中，我们研究一个作者性别预测任务。为了建立一个人类基线，我们从 upwork.com 招募了 5 名注释者，他们都是拥有丰富阅读和写作经验的母语人士。给定两篇博客文章，注释者需要审阅它们并选择哪一篇更有可能是男性写的。总体而言，我们为每个例子收集了 5 个人工标签。