弥合差距：整合（人类）反馈以实现自然语言生成的综述

Patrick Fernandes[1,2,3] **Aman Madaan**[1] **EmmyLiu**[1] **António Farinhas**[2,3]**Pedro HenriqueMartins**[4] **Amanda Bertsch**[1] **JoséG.C.deSouza**[4] **Shuyan Zhou**[1]**Tongshuang Wu**[1] **GrahamNeubig**[1,5] **AndréF.T.Martins**[2,3,41]卡内基梅隆大学 [2] InstitutoSuperiorTécnico (里斯本 ELLIS 单位)[3]电信研究所 [4]Unbabel [5]Inspired Cognitionpfernand@cs.cmu.edu

arXiv:2305.00955v2 [cs.CL] 1 Jun 2023

摘要

近年来，自然语言生成领域的许多进展都得益于在互联网规模数据上训练大语言模型。然而，这一范式可能导致模型生成有毒、不准确且无用的内容，而自动评估指标往往无法识别这些问题。随着模型能力不断增强，人类反馈成为评估和改进模型的宝贵信号。本综述旨在概述近期利用人类反馈改进自然语言生成的研究。首先，我们提出一种全面的反馈形式化框架，并据此识别和整理现有研究，构建分类体系。接下来，我们讨论如何根据反馈的格式和目标对其进行描述，并介绍利用反馈的两种方法（用于训练或解码）：直接使用反馈，或训练反馈模型。我们还讨论了用于收集人类反馈数据的现有数据集，以及反馈收集过程中面临的问题。最后，我们概述了新兴的*AI* 反馈领域，该领域利用大语言模型依据一组原则作出判断，并尽量减少人工干预的需求。

1 引言

要让生成系统得到广泛应用，它们生成的文本不仅必须流畅、质量上乘，还必须与人类的愿望和要求高度一致（Vamplew 等人，2018年；Hendrycks 等人，2020年；Kenton 等人，2021a；Turner 等人，2022年；Ngo，2022年）。实现如此宏大的目标，需要现代大语言模型超越传统训练方法，不断发展。近年来，该领域的进展主要集中在融入人类反馈（Bai 等人，2022b；Ouyang 等人，2022年；OpenAI，2023a）。这种反馈发挥着引导作用，推动大语言模型朝着预期结果前进，就像物理机器中的反馈机制一样（Åström 和 Murray，2021年）。

通常，最先进的语言生成系统是通过使用概率式、自回归大语言模型，并在海量数据上采用最大似然估计（MLE）进行训练而得到的。然而，用于训练这些模型的数据通常是从互联网抓取的，其中往往包含噪声、社会偏见和错误（Bolukbasi 等人，2016年；Dodge 等人，2021年）。当这一点与“根据前面的词元最大化下一个词元概率”这一目标结合时，可能会导致目标行为设定不当（Kenton 等人，2021b），并可能使模型生成有毒、不准确且无助益的内容（Sheng 等人，2019年；Bender 等人，2021年）。

更糟的是，这些模型通常使用自动评价指标进行评估：这类指标依据表层特征（例如词汇重叠），比较生成文本与某个“参考”文本，而这些特征往往与人类感知到的文本质量并不相关（Schluter，2017年；Mathur 等人，2020年；Gehrmann 等人，2022a），尤其是在针对这些指标优化模型时（Paulus 等人，2017年；Amrhein 和 Sennrich，2022年）。评估之所以困难，部分原因在于，对于许多任务而言，并不存在唯一正确的答案，因为同一交际意图可以通过多种方式表达。

利用人工评估来衡量模型生成文本的质量，已成为一种流行的方法。关键在于，考虑人类感知到的质量有助于缩小机器生成文本与人类生成文本之间的差距，并有助于应对古德哈特定律带来的挑战：“当一项度量指标成为目标时，它就不再是一个好的度量指标”（Goodhart，1984年）。这一认识推动了人们对借助人工

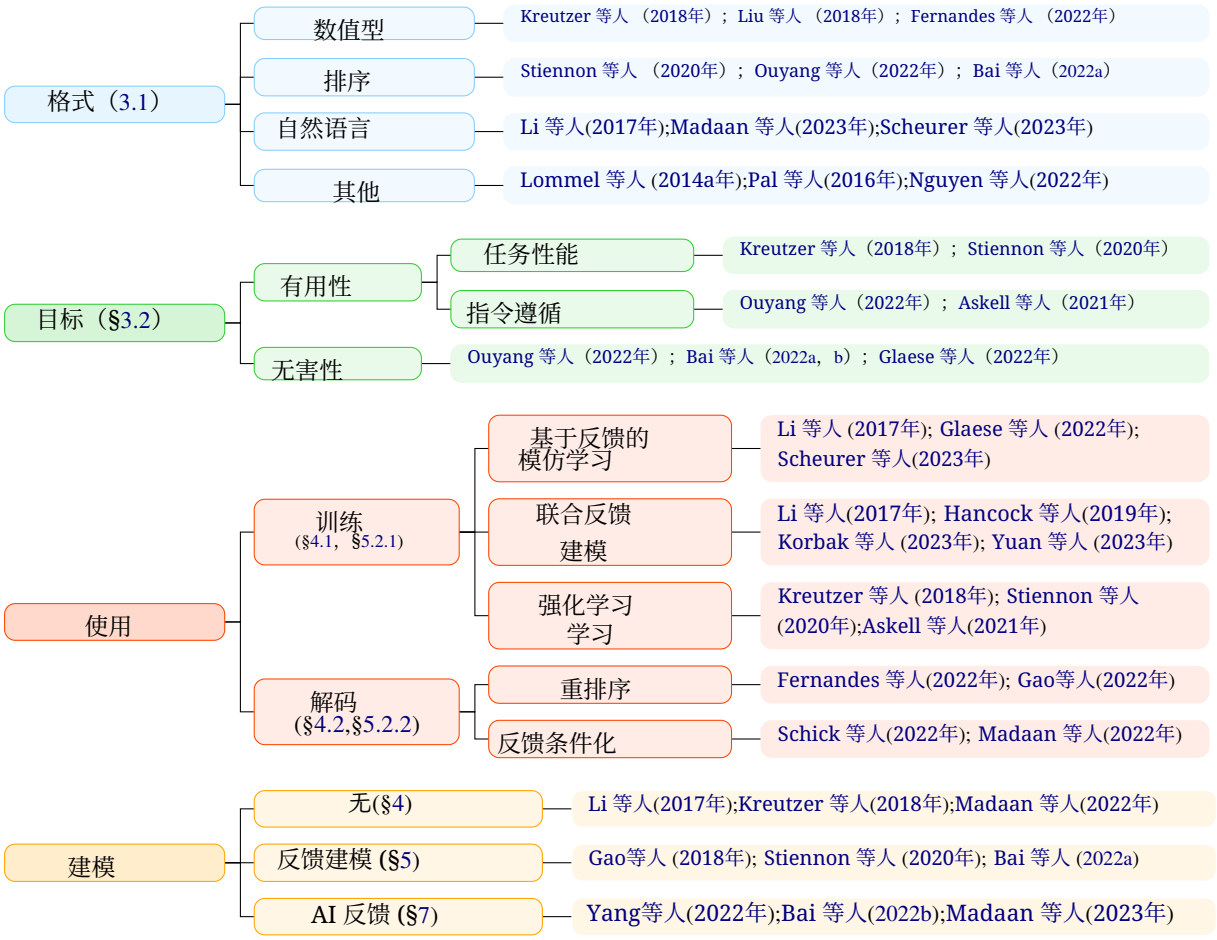

图1：利用人类反馈的方法分类体系，以及文献中符合各类别的一些代表性示例研究。

反馈来改进模型生成的输出，并促使首批广泛使用的通用语言助手应运而生（OpenAI，2023a）。人工反馈不仅能提升系统性能，还能作为一种机制，引导系统与预期结果或目标保持对齐（Rosenblueth 等人，1943；Wiener，1948）。

反馈这一概念涵盖广泛的含义和解释（Wiener，1948）；不过，我们可以识别出一些普遍特征，例如其形式、预期结果，以及将其用于模型开发过程的方式。本综述聚焦于人类反馈在改进语言生成中的作用。我们首先对人类反馈这一概念进行形式化，并建立文献中不同反馈类型及其用途的分类体系（§2）。我们讨论如何根据反馈的形式和目标来描述反馈，即从期望的模型行为角度进行描述（§3）。我们讨论直接依据模型（自身）输出上的人类反馈来优化模型的方法，例如，使用人类奖励函数进行强化学习（§4）。随后，我们转向另一类方法：先训练反馈模型来近似人类反馈，再利用这些代理模型改进生成（§5）。我们讨论用于人类反馈数据的现有数据集、这些数据集通常如何收集，以及收集过程可能对模型行为产生的影响（§6）。最后，我们讨论近期的一项研究方向：通过利用大语言模型提供的*AI* 反馈，减少对收集人类反馈的需求（§7）。

2 利用（人类）反馈进行生成的分类体系

2.1 背景

考虑一个模型 $M: \mathcal{X} \to \mathcal{Y}$ 它接收某种类型的输入 $x \in \mathcal{X}$，并输出文本 $\hat{y} \in \mathcal{Y}$。重要的是，虽然 x 可以是任何格式，但我们将范围限定 y 为自然语言空间中的情况（即，$\mathcal{Y} \subseteq \Sigma^*$ 针对某个字母表

Σ）。这一通用表述涵盖了广泛的自然语言生成任务。例如：

- **摘要生成**： $\mathcal{X}$ 是文档空间， $\mathcal{Y}$ 是可能摘要的空间。
- **机器翻译**： $\mathcal{X}$ 和 $\mathcal{Y}$ 分别是源语言和目标语言句子的空间。
- **对话生成**： $\mathcal{X}$ 是可能对话历史的空间， $\mathcal{Y}$ 是可能回复的空间。
- **图像描述生成**： $\mathcal{X}$ 是图像空间， $\mathcal{Y}$ 是可能图像说明的空间。

这些模型通常被表示为参数化的条件概率分布 $P_\theta(y|x)$，其中 θ 是模型参数。该分布通常通过自回归方式估计：给定输入 x 时，句子 y 的概率被分解为句中每个词元的概率之积，并以之前的词元为条件。随后，通过寻找使某些训练数据 $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$的似然最大化的参数 $\theta^\star$ 来训练这些模型。然后，在推理时，给定输入 x，从学习得到的分布中解码出输出$\hat{}y$。例如，可以通过近似最可能的词元序列（$M(x) \approx \arg\max_y P_{\theta^\star}(y|x)$）或通过随机采样（$M(x) \sim P_{\theta^\star}(y|x)$）来进行解码。

评估生成文本的质量$\hat{}y \in \mathcal{Y}$可能颇具挑战，因为自然语言具有复杂性和主观性。针对各种领域/任务，人们提出了多种自动指标。这些指标传统上依赖 n 元语法匹配或其他简单的启发式方法，无法处理复杂的语言现象（例如释义改写或风格变化），并且往往无法捕捉人类判断的全部细微差别（Sai 等人，2022年；Gehrmann 等人，2022a）。因此，对于许多此类任务而言，征求人类反馈被视为评估生成文本质量的黄金标准，而较新的学习型指标通常旨在近似人类提供反馈的方式（见 §5.1）。

更正式地说，我们将**人类反馈** 视为一族函数 $\mathcal{H}$，其中每个反馈函数 $h \in \mathcal{H}$ 接收一个输入[1] $x \in \mathcal{X}$ 以及一个或多个输出 $y_1, \cdots, y_n \in \mathcal{Y}$并返回某种反馈 $f \in \mathcal{F}$：

$$h : \mathcal{X} \times \underbrace{\mathcal{Y}_1 \times \cdots \times \mathcal{Y}_n}_{n} \to \mathcal{F}. \quad (1)$$

一个简单的人类反馈函数示例是：让人类判断，在给定输入的情况下，某个输出是好还是坏（$h: \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$）。然而，更复杂的反馈函数也存在并得到广泛使用，例如排序或自然语言反馈（见 §3.1）。

我们指出，这种表述是对现实世界的简化：通常，不同的人可能会针对相同的输出提供不同（甚至可能相互矛盾）的反馈，而单一函数可能无法捕捉人类观点中的这种差异（我们将在 §6中进一步讨论）。最后，尽管我们的形式化方法灵活，但它并未涵盖模型与人类互动以改进学习的其他方法，例如主动学习及其他人在环路中的方法。

2.2 分类体系

在确立了基本的数学表述之后，我们现在确定四个关键维度，据此对人类反馈的用途进行分类：

反馈的形式是什么？ 人类反馈的形式多种多样，包括二元判断、数值型评分、序数排名或定性自然语言解释。

它的目标是什么？ 根据模型的使用场景，反馈可以有多种用途，从评估模型性能和准确性，到防止毒性及有害行为。

何时使用它？ 人类反馈可以纳入训练阶段，直接优化模型参数。或者，也可以在推理时使用，以引导解码过程。

如何对其进行建模？ 理想情况下，我们会尽可能使用来自人类的直接反馈，但由于收集此类反馈的成本高得令人难以承受，通常使用近似人类偏好的代理模型会更实用。

3 描述反馈

3.1 格式

当我们希望通过人类反馈来改进语言生成系统时，一个重要的决策是

[1] 尽管反馈可以独立于输入提供（例如针对流畅度），但为简化记法，我们假设存在某个（可能为空的）输入。

输入	输出	反馈	Type
世界上最好的食物是葡萄牙的。	世界上最糟糕的食物是葡萄牙人。	0.7	评分
		’worst’：主要/准确性 ’are’：次要/流畅度	MQM
		“最差” →“最佳”, ’是’ →’是’	后编辑
人工智能已经拥有了具有变革性的潜力行业……但符合伦理的需要处理相关问题。	人工智能可以改变行业。	流畅度：1 相关性：0.7	多方面
		“忽略了伦理问题。”	自然语言
解释一下登月 给一个 6 岁的孩子听	A:人们去了……	A >B	排序
	B：月球是一颗卫星…		

表1：三个任务（机器翻译、摘要生成和指令遵循）的输入与输出示例，以及可能提供的不同（示例）反馈。

反馈应以什么格式收集。格式的选择会影响反馈的表达能力、收集的难易程度，以及我们如何利用反馈来改进系统。尤其是，反馈格式的复杂程度是一个重要因素：较简单的格式通常更易于收集，也更易于用于训练/解码过程，但所包含的信息不如更“复杂”的格式丰富，并且可能无法捕捉到改进系统所需的重要信息。格式的选择也会影响人类提供反馈的难度、反馈的一致性/一致程度，以及反馈的合理性（Ghosal 等人，2023年）。表 1总结了各种反馈类型，并附有示例。

数值型 数值反馈接受输入和输出，并返回单个评分（$\mathcal{X} \times \mathcal{Y} \to \mathcal{N} \subseteq \mathbb{R}$），是最容易收集和使用的反馈格式之一。Kreutzer 等（2018）研究了类别反馈的使用方式：将翻译评为 5 个可能的“星级”之一，然后取平均值生成评分（$\mathcal{N} = [1, 5]$），并用其改进模型。Liu 等人（2018）和 Shi 等人（2021）采用了更简单的反馈方式，让人类判断给定的回复是好还是不好（$\mathcal{N} = \{0, 1\}$）。数值反馈也被广泛用于评估，不过其明确目标并非改进生成。例如，在机器翻译中，直接评估（Graham 等人，2013）要求人类按照连续量表为翻译打分；一些研究尝试利用这些反馈数据训练反馈模型（Sellam 等人，2020；Rei 等人，2020年a），并改进生成（Freitag 等人，2022a；Fernandes 等人，2022年）。

尽管数值反馈易于利用，但也存在一些局限：根据生成任务的复杂程度，将反馈简化为单个评分通常可能是人类难以完成且定义不明确的任务，导致收集成本高昂，并带来主观性和方差（参见 §6.2.1 ）。此外，这类反馈可能不适合区分质量相近的输出。

基于排序的 </style'>一种替代方案是不让人类为给定的输入输出对打一个分数，而是让他们对多个可能的替代输出进行排序

$$h : \mathcal{X} \times \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_n \to S_n$$

其中 S_n 表示 n 个元素的所有排列/排序集合（可选择允许并列）。这种形式已广泛用于评估（Chaganty 等人，2018年）。与数值反馈相比，这种格式往往更容易收集；或许正因如此，基于排序的反馈通常用于改善模型行为，而不仅仅用于评估（因为前者通常需要更多反馈数据）。Ziegler 等人 (2019年) 和 Stiennon 等人 (2020年) 请人类对他们试图改进的系统所生成的备选摘要进行排序。类似地，Ouyang 等人(2022年) 收集了针对给定模型的指令所生成的备选回答的排序。他们利用这些排序来增强模型的指令遵循能力。后续研究也将基于排序的反馈用于相同任务（Askell 等人，2021年；Bai 等人，2022a，b）。

自然语言 数值型和基于排序的反馈都无法捕捉输出问题的详细信息，而这些信息对于改进生成系统至关重要。我们可以不要求人们对输出进行排序或评分，而是要求他们提供自然语言反馈。在这种情况下，反馈通常会提供更详细的信息，指出当前输出的不足，或提出具体的改进措施。例如，Li 等人(2017年)要求人们为对话问答模型提供自然语言反馈，其中包括正面或负面反馈，也可以向模型提供正确答案或给予提示。Tandon 等人(2022年)和 Madaan 等人(2022年)收集关于模型生成的图中错误，以及模型对给定指令的理解情况的自然语言反馈。Scheurer 等人(2022年,2023年)通过要求人们对模型生成的摘要提供自然语言反馈，提升语言模型的摘要能力。Li 等人(2022年)为问答（QA）系统的回复收集自然语言反馈（以及数值反馈）。

其他类型：除了这些反馈类型之外，还可以使用其他（可能特定于领域的）反馈类型来改进模型行为。通常，人们会被要求提供多方面反馈（$\mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$ 或更广义地说，$\mathcal{F}^d$），针对多个维度对某个输出进行评分，或对多个输出进行排序（Böhm 等人，2019年；Glaese 等人，2022年；Madaan 等人，2023年；Nguyen 等人，2022年）。译后编辑要求人们以小幅修改（例如，将 X 替换为 Y）的形式对输出进行纠正；译后编辑数据已被用于直接改进模型（Denkowski 等人，2014年），或训练能够纠正模型错误的自动译后编辑系统（Pal 等人，2016年；Mehta 和 Goldwasser，2019年；Madaan 等人，2021年；Talmor 等人，2020年；Elgohary 等人，2021年）。此外，还有其他尚未被充分利用来改进生成的反馈类型：例如，作为评估翻译质量的标准，多维质量指标（MQM）（Lommel 等人，2014b年）要求专业译者标出翻译中的错误片段，并指出错误的严重程度和类型。

3.2 目标

收集反馈的目的是让模型的行为与某个（通常定义不明确的）目标保持一致：我们可能希望摘要生成模型生成包含所有核心信息的摘要，即使这意味着摘要会稍长一些；在商业机器翻译中，人们会格外注意确保模型不会误译对业务至关重要的信息；而对于对话代理，我们可能希望模型能够生成礼貌且无害的回复。这一**对齐目标**已在人工智能安全与对齐文献中得到广泛研究（Bostrom，2014年；Amodei 等人，2016年；Bommasani 等人，2021年）。此外，Ken- ton 等人(2021b）讨论了语言智能体（自然语言生成模型）中由对齐目标规格设定不当（例如，训练数据中的噪声标签）引发的一些行为问题；而Leike 等人(2018年）提出使用反馈模型来应对目标难以明确界定这一问题。

Bai 等人(2022a) 明确将语言模型“对齐”问题划分为提升其**有用性** 和增强其**无害性**。大多数研究都隐含地考虑了两类反馈：一类针对性能因素（例如提升任务的整体性能或遵循指令的能力），另一类针对无害性因素（例如不生成有毒文本，或不提供可能导致伤害的信息）。[2]

有用性 通常，收集反馈时都会以某种 有用性 目标为导向：系统要有用，一个必要但不充分的条件是能够出色地完成任务，因此，与**任务性能** 相关的反馈通常都属于这一范畴。例如，机器翻译领域的大多数研究都会利用与翻译质量相关的反馈（Kreutzer 等人，2018年；Fernandes 等人，2022年），因为预计翻译质量与其在下游应用中的有用性相关。类似地，在摘要生成领域，大多数研究都会利用与相关性、一致性 和 准确性 （Ziegler 等人，2019年；Stiennon 等人，2020年）等方面相关的反馈（简而言之，即摘要的质量）。一项研究尤为深入的反馈目标，是模型遵循**指令的能力**（Ouyang 等人，2022 年）：指令遵循任务涵盖范围广泛的其他任务，而利用反馈提升语言助手的指令遵循能力，也被视为对齐问题的一项基准（Askell 等人，2021年）。

[2]我们基本忽略了所提出的诚实性这一方面，因为这些研究都没有直接探讨这一点。

无害性 另一项重要的对齐目标是无害性：我们希望模型不生成某些类型的输出，也不违反特定规范。在Ouyang 等人 （2022年）收集的反馈中，考察了文本毒性等方面（以及整体的指令遵循能力）。Bai 等人 （2022a）探讨了有用性与无害性目标之间的相互作用，指出二者存在权衡。Thoppilan 等人 （2022b）收集了有关其模型是否违反一组安全目标的反馈，并用这些反馈微调模型。Glaese 等人 （2022年）也请人们针对其系统的无害性提供反馈：他们制定了一组规则，并询问人们模型输出是否违反了这些规则。Bai 等人（2022b）表明，由大语言模型生成的反馈可以提高无害性，同时不降低有用性。

4 直接利用人类反馈

理想情况下，我们会直接利用人类反馈来改进生成：由人类为训练或解码过程提供反馈。

4.1 面向人类反馈进行优化

收集到人类反馈后，一种利用方式是直接优化模型参数。然而，这要求反馈“可优化”，即，可能需要将其表述为一个优化问题，以便据此获得改进后的模型。例如，如果反馈是数值型评分（$f \in \mathbb{R}$），我们可以构建如下优化问题：

$$\theta^\star = \arg\max_\theta \mathbb{E}_{x\sim\mathcal{D}}[h(x, M_\theta(x))]. \quad (2)$$

其中，$\mathcal{D}$ 是可能输入的分布。已有多种方法被提出，用收集到的人类反馈来优化模型参数 θ。根据训练机制，这些方法可分为三大类，我们将其称为**基于反馈的模仿学习**、**联合反馈建模**和**强化学习（RL）**。

这种**基于反馈的模仿学习**方法通过使用人类反馈来优化模型：利用由正向标注的生成结果及其对应输入组成的数据集进行监督学习，$\mathcal{D}^+$。

这可以通过最小化损失来实现：

$$\theta^\star = \arg\min_\theta \sum_{i=1}^{|\mathcal{D}^+|} \mathcal{L}^{(i)}(\theta) \quad (3)$$

$$\mathcal{L}^{(i)}(\theta) = -\log p_\theta\left(y^{(i)} \mid x^{(i)}\right) \quad (4)$$

这种方法的一个实例见于 Li 等人 (2017年)：作者通过最大化人类标注为正确的模型回答的似然来训练对话模型。类似地，Kreutzer 等人 (2018年) 使用一组带有正向标签的译文训练机器翻译模型，而 Glaese 等人 (2022年) 则根据人类偏好，对符合预先定义规则（涉及正确性、有害性和有用性）的偏好对话进行监督学习。稍有不同的方法由 Hancock 等人 (2019年) 提出：部署闲聊对话模型，并将人类发言作为目标来微调模型。Scheurer 等人 (2022年、2023年) 利用了大语言模型能够遵循指令这一特点，首先收集有关模型生成结果的自然语言人类反馈，其中往往会描述改进后的文本应是什么样子。接着，他们让语言模型根据输入、先前的模型生成结果及相应反馈生成多个改写结果。随后，使用每个生成结果中相似度最高的改写结果来微调大语言模型。OpenAI 的 `text-davinci-002` 使用人类示范以及评分尽可能高的模型输出进行训练，这种方法称为 *FeedME* (OpenAI，2023b)。这些方法的一个缺点是，它们忽略了未获得正向反馈的生成结果，而这些结果可能包含有助于优化模型的信息。

另一方面，**联合反馈建模**通过直接利用收集到的所有人类反馈来优化模型。此外，由于反馈由模型直接建模，例如，这种方法也支持数值型或基于排序的格式以外的反馈（如自然语言）。有了收集到的输入 x、生成结果y和人类反馈 f 的数据集数据集 $\mathcal{D}$，就可以通过最小化以下形式的损失来实现

$$\mathcal{L}^{(i)}(\theta) = -\log p_\theta\left(y^{(i)}, f^{(i)} \mid x^{(i)}\right) \quad (5)$$

对 $\mathcal{D}$ 中的所有示例。这些方程可以分解为 $\mathcal{L}^{(i)}(\theta) = -\log p_\theta\left(f^{(i)} \mid y^{(i)}, x^{(i)}\right) +$

$\log p_\theta\left(y^{(i)} \mid x^{(i)}\right)$。一些研究仅训练模型预测针对每次生成给出的反馈（Weston，2016年，前向预测），而忽略分解式中的第二项。该方法的一个例子是Li 等人 (2017年的研究，作者请人类向对话问答模型提供自然语言反馈（例如，正面或负面反馈、向模型提供正确答案，或给出关于正确答案的提示）。收集反馈后，研究人员训练模型对其进行预测。Hancock 等人 (2019年提出使用辅助模型来预测与模型交谈的人类用户的满意度。如果满意度评分低于预设阈值，模型就会向用户征求反馈。随后，模型通过学习预测用户给出的自然语言反馈来加以利用。Yuan 等人(2023年);Rafailov 等人 (2023年表明，让摘要模型预测不同摘要的排序，有助于模型生成更好的摘要，甚至可能胜过使用反馈模型的更复杂方法（§5）。

其他工作会训练模型预测生成内容及其对应的人类反馈。 Xu 等人(2022年) 提出使用由 DIRECTOR 模型来利用人类反馈，该模型由 Arora 等人 (2022年) 提出。由于该模型采用统一的解码器-分类器架构，Xu 等人 (2022年) 提出使用正标签样例训练其语言建模头（类似于基于反馈的模仿学习），并使用正标签和负标签样例训练一个分类器头，以引导模型避免生成不良序列。Thoppilan 等人(2022a) 采用这种方法来确保模型的质量和安全性。首先，他们收集众包工作者与所提出的语言模型 LaMDA 之间的对话，并由众包工作者对其进行标注，提供反馈。这些反馈说明每条回复的质量（是否合理、具体且有趣）或安全性。随后，对 LaMDA 进行微调，使其能够预测高质量回复，并预测针对每条回复的质量属性和安全性所给出的奖励。在推理时，LaMDA 还会筛除安全性预测低于阈值的候选回复。

最后，还可以通过训练模型预测生成结果并以反馈为条件来实现这一点。这相当于最小化以下内容：

以下损失：

$$\mathcal{L}^{(i)}(\theta) = -\log p_\theta\left(y^i \mid f^i, x^i\right) \tag{6}$$

Liu 等人（2023年）提出了基于提示的微调方法：他们创建包含先前生成结果的提示，这些结果由人类按照偏好顺序进行评分。他们还建议在各生成结果之间向提示中插入基于语言的反馈（例如，“.....是比.....更差的答案”）。然后，对模型进行微调，以最大化生成最受偏好答案的可能性。

最后，**强化学习（RL）**提供了一种更灵活的方法，无论反馈是否可微，都能根据人类反馈直接优化模型参数。在此背景下，常用的一种强化学习算法是 REINFORCE算法（ Williams，1992），它使用以下梯度更新策略参数：

$$\nabla_\theta J(\theta) = \mathbb{E}_{x\sim\mathcal{D}, y\sim p_\theta}[h(x,y)\nabla_\theta \log p_\theta(y \mid x)] \tag{7}$$

这里， $\mathcal{D}$ 表示输入 x的集合， p_θ 是策略。这种灵活性使强化学习能够处理各种类型的反馈，并使生成的输出更好地符合人类偏好。例如，Kreutzer 等人（2018年）提出利用来自用户查询的基于任务的隐式反馈作为奖励信号，通过最小风险训练的词级变体（Shen 等人，2016年）来训练机器翻译模型；而Jaques 等人（2019年）则利用聊天中的隐式人类反应，通过离策略 Q 学习（Watkins 和 Dayan，1992）改进开放域对话系统。鉴于收集人类反馈既昂贵又耗时，学习是在日志数据上离线进行的，这通常比需要实时获取反馈的同策略设置更有利。我们将在 §5.2.1中讨论几项使用强化学习优化反馈模型，而非直接优化人类反馈的研究。这些方法通常统称为基于人类反馈的强化学习（**RLHF**）。

4.2 基于人类反馈的解码

尽管直接优化模型参数能提供更大的控制力，但修改参数未必总是可行，尤其是在大语言模型的情况下。此外，在模型训练期间可能无法获取反馈，这限制了参数调整的范围

。在这种情况下，在解码过程中利用人类反馈对于提升大语言模型的性能至关重要。这类反馈源于实际场景中大语言模型与用户之间的互动，使模型能够从错误中学习，并在不改变模型参数的情况下持续改进。此外，反馈还能发挥指导作用，让模型利用现有能力生成更理想的输出。

在此设置中使用人类反馈的方式大致分为两类：1. 反馈记忆： 反馈记忆利用是指维护一个存储以往会话反馈的资料库。处理新的输入时，系统会从记忆中检索与相似输入相关的反馈，据此引导模型生成更理想的输出，借鉴以往经验并契合用户偏好。尽管这是一个经典概念（Riesbeck，1981；Schank，1983），近期研究已展现出这种记忆增强方法在微调（Weston 等人，2014年；Wu 等人，2018年；Tandon 等人，2022年）和少样本设置（Madaan 等人，2022年）中的潜力。

2. 迭代式输出优化： 这种方法利用人类反馈对模型输出进行迭代优化。用户可以对中间回复提供反馈，使模型能够不断调整输出，直到令用户满意。这个过程有助于模型更好地理解用户偏好，并生成更合适的结果（Reid 和 Neubig，2022；Saunders 等人，2022；Schick 等人，2022；Nijkamp 等人，2022）。反馈也可以针对解码策略等模型属性（Passali 等人，2021），而非直接针对其输出。

这两种技术并不互斥，可以结合使用以实现更优异的性能，从而构建一个更具适应性和响应能力的系统，更好地满足用户的期望。

5 使用人类反馈模型改进生成

在一般情况下，直接使用人类反馈来改进模型行为并不可行：要求人类为模型的每个输出提供反馈既昂贵又耗时。

5.1 人类反馈学习模型

另一种获取人工反馈的方法是开发能够预测或近似反馈的模型，尽管这些模型可能并不完美，但它们的优势在于训练完成后能够以低成本提供反馈，从而使依赖反馈的技术得以扩展。

更正式地说，给定反馈函数 $h: \mathcal{X} \times \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_n \to \mathcal{F}$，我们希望学习一个参数值型（数值型）反馈模型 $h_\phi: \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$（带有参数 ϕ），使其与人类反馈“达成一致”。这种一致性通过损失函数来表达，而模型经过训练以最小化这种一致性损失：

$$\phi_\star = \arg\min_\phi \mathbb{E}_{x,y_1,\cdots,y_n \sim \mathcal{D}_f} \left[\mathcal{L}(\phi)\right] \tag{8}$$

$$\mathcal{L}(\phi) = \text{loss}\left(\hat{h}_\phi(x, y_1), \cdots, h(x, y_{1:n})\right) \tag{9}$$

例如，如果我们尝试建模的反馈函数也是数值型的（$h: \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$），那么此损失可以是任意标准回归损失，例如人类反馈与模型反馈之间的平方差 $\mathcal{L}(\phi) = \left(\hat{h}_\phi(x, y) - h(x, y)\right)^2$。

重要的是，尽管反馈模型通常是数值型的，人类反馈可以采用任何其他格式，只要能够指定合适的损失函数即可。Stiennon 等（2020）基于排序的反馈训练偏好模型 [3] $\hat{h}_\phi(x, y_n)$，使用形式如下的损失函数

$$\mathcal{L}(\phi) = \log\left(\sigma\left(\hat{h}_\phi(x, y_{+1}) - \hat{h}_\phi(x, y_{-1})\right)\right) \tag{10}$$

使得样本 y_{+1} 相较于 y_{-1} 在相同输入下更受偏好 x： $h(x, y_{-1}, y_{+1}) = (y_{-1} < y_{+1})$。此损失的变体随后也被用于其他工作中（Ouyang 等人，2022年；Askell 等人，2021年；Liu 等人，2022年；Qin 等人，2022年；Yuan 等人，2023年）。

在自然语言处理领域，反馈建模问题已得到广泛研究，尤其是在指标学习方面。Zhang 等人（2019年）和Zhou 等人（2023b）利用预训练的掩码语言模型，计算生成文本或代码片段与其参考摘要之间的相似度分数。在机器翻译领域，Sellam 等人（2020年）和Rei 等人（2020年a）分别训练了 BLEURT 和 COMET，使其能够拟合人工对翻译质量的评估。在摘要生成方面，Zopf （2018年）利用标注的成对偏好训练偏好模型，而Peyrard 等人（2017年）学习了一个摘要

[3] 我们根据人类反馈格式来定义反馈模型，即分别针对数值型和基于排序的人类反馈的奖励和偏好模型。

等级指标，其依据是较早的摘要数据集中包含的一组人工判断（例如，TAC- 2008）。研究表明，与广泛使用的词汇指标（如 BLEU 和 ROUGE）相比，这些指标与人类判断的相关性要高得多（Freitag 等人，2022b）。值得注意的是，这些奖励模型的训练目的并非直接改进生成，尽管其中一些模型后来被用于这一目的，详见第5.2节。

近年来，人们对直接开发反馈模型的兴趣日益浓厚，旨在利用这些模型改进生成（Böhm 等人，2019年；Ziegler 等人，2019年）。作为第一步，这些模型通常使用需要改进的目标语言模型的权重，或同一系列中某个模型的权重进行初始化（例如，规模较小的模型）（Askell 等人，2021年；Bai 等人，2022a；Ouyang 等人，2022年）。初始化时的一个关键考量是预训练模型的规模：扩大规模或许能提升整体性能（Askell 等人，2021年；Bai 等人，2022a），但(2022年Ouyang 等人发现，较大的模型在后续微调时可能稳定性较差。

接下来，反馈模型会在人工反馈数据集上进行微调。该数据集通常通过让标注人员对待改进模型早期版本的输出提供反馈来收集。不过，也可以先利用自然产生的隐式反馈（例如来自 Reddit、StackOverflow 等网站的用户互动）对反馈模型进行微调。虽然这种反馈不如显式收集的反馈准确，但它能让反馈模型利用更多数据进行训练。Askell 等人（2021年）发现，自然产生的反馈数据对参数量超过 10 亿的模型有益，但随着显式收集的反馈不断增加，其收益往往会递减。

Nguyen 等人 (2022年) 基于对三项由人类设计的目标进行排序来训练偏好模型：摘要是否具有恰当的主题、长度和质量，并使用基于距离的排序损失将这三项合并为单一目标。有趣的是，机器翻译中的自动译后编辑（APE）系统（例如，Simard 等人 (2007 年) ；Correia 和 Martins (2019年)），其训练数据是人工译后编辑，目的是自动纠正机器翻译系统的输出，也可以视为反馈模型（尽管其反馈并非数值型）。

5.2 利用反馈模型改进生成

训练好反馈模型后，我们几乎可以像使用人类反馈那样用它来改进生成：要么在训练生成模型时利用该反馈模型，要么在解码过程中引入该反馈模型。

5.2.1 针对反馈模型进行优化

与针对人类反馈进行优化类似，使用反馈模型的一种可能方式是根据它给出的反馈来优化模型参数。如果反馈模型输出数值型ˆ反馈（$h_\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$），我们可以定义一个类似于公式2的优化问题。然而，由于反馈模型作为不完美代理存在局限性，通常会引入一个正则化项 R，以避免对反馈模型（“过拟合”（Ziegler 等人，2019年）进行“过拟合”（本节末尾将对此作进一步介绍）：

$$\theta^\star = \arg\max_\theta \mathbb{E}_{x\sim\mathcal{D}} \left[\hat{h}_\phi(x, M_\theta(x)) - \beta R(\theta)\right] \tag{11}$$

由于这两个优化问题具有相似之处，解决公式11的方法可以划分为§4.2中的三个类别中的两个类别：**联合反馈建模**和**强化学习**。回顾一下，在§4.2 中，我们讨论的是直接针对人类反馈进行优化的方法，而本节关注的是使用人类反馈模型的情况。

与直接使用人类反馈不同，大多数研究尝试使用**强化学习**来优化反馈模型。 Gao等人(2018年)；Böhm 等人(2019年) 使用其他研究收集的（数值型）反馈来训练奖励模型和偏好模型，并使用强化学习针对这些模型进行优化。结果表明，人们更偏好他们的摘要生成模型，而非其他经过监督训练和强化学习训练的基线模型。 Ziegler 等人(2019年) 提出了类似的方法，但使用在待改进模型上收集的反馈来训练偏好模型，并引入了一个*KL* 正则化项

$$R(\theta) = \log\left[P_\theta(y|x)/P_{\theta_{\text{SL}}}(y|x)\right] \tag{12}$$

以避免优化后的模型与参数为 θ_{SL}[4] 的原始（监督）模型偏离过多。 Stiennon 等人 (2020年 在此基础上扩展了这项工作，

[4]请注意，这个 KL 项与其他算法不同——

通过 扩大 摘要生成模型和偏好模型的规模，证明其模型更受人类青睐，且比监督学习基线具有更好的泛化能力。 Ouyang 等人(2022年 也使用基于偏好模型的强化学习来提升大语言模型遵循指令的能力，但将强化学习目标与原始预训练目标结合，以避免在公开自然语言处理基准上的性能下降。其他研究也以类似方式使用基于偏好模型的强化学习（Askell 等人，2021年；Bai 等人，2022a；Wu 等人，2021年；Nguyen 等人，2022年）。这些方法的共同基础是，通常先通过对人类示范进行模仿学习来训练模型；与直接在预训练策略上使用强化学习相比，这能提升模型性能。

Glaese 等人(2022年) 比较了基于人类反馈（§4.1）进行反馈驱动的模仿学习与利用反馈模型进行强化学习的效果，发现后者能够获得更高的偏好率和更低的规则违规率。

该**联合反馈建模**与反馈模型的结合由 Korbak 等人（2023年）进行了探索，他们研究了使用类似于公式 6的损失对大语言模型进行预训练，该损失基于一个使用针对毒性的基于排序的反馈训练的偏好模型所提供的反馈。他们表明，与使用标准最大似然估计对模型进行预训练相比，这种方法能够使模型生成更少的有毒生成文本。

在这些主要类别之外的一种方法中，Peyrard 和 Gurevych (2018年) 将基于人类判断学习得到的评分函数用作遗传算法的适应度函数，以生成输入文本的摘要。

5.2.2 使用反馈模型进行解码

如前所述，反馈模型的优势在于训练完成后，可以低成本地获取反馈。或许正因如此，大多数利用反馈模型的方法都会采样大量候选生成文本，并根据反馈模型对其进行重排序：

$$\mathcal{C} = \{\bar{y}_1, \cdots, \bar{y}_S\} \quad \text{where} \quad \bar{y}_i \sim P_\theta(y|x)$$
$$\hat{y} = \arg\max_{\bar{y} \in \mathcal{C}} \hat{h}_\phi(x, \bar{y})$$

ˆ其中 h_ϕ 是一个经过训练的（数值型）反馈模型，而 $\mathcal{C}$ 是一组由模型给出的 S 候选生成文本（例如，从模型的分布中多次采样）。

在机器翻译中，Fernandes 等人（2022年）和Freitag 等人（2022a）利用自动质量估计方面的最新进展，通过反馈模型训练来进行评估，从而改进生成。其框架包括候选生成阶段，随后是排序阶段；在排序阶段，使用经过训练、能够拟合人工评估的质量指标（奖励模型）对候选进行评分（Rei 等人，2020年a，b），再通过 N-best 列表重排序或最小贝叶斯风险（MBR）解码（Kumar 和 Byrne，2002年）。随后选取得分最高的候选作为最终译文。

Li 等人（2022年）收集了一个数据集，其中包含问答系统回复的数值反馈和自然语言反馈，并微调了一个预训练模型来预测这两类反馈，再利用该反馈模型的预测分数对模型生成的预测结果重新排序。

Gao等人 (2022年) 也采用了这一方法来研究反馈模型的缩放特性以及“过度优化”问题（见下文）。

此外，还有一些工作在解码阶段结合机器翻译和自动后编辑系统，其中机器翻译系统的输出会由自动后编辑系统进一步改进（Bhattacharyya等人，2022年）。

反馈模型的过度优化 使用反馈模型优化系统时会出现一个问题：该模型只是对真实的人类反馈的不完美代理，因此，针对它们进行“过度优化”可能导致这样的系统：能够获得模型的良好反馈，却无法获得人类的良好反馈。这一问题被称为过度优化问题，也是公式11中采用正则化项的主要原因。

Gao等人 (2022年) 研究了偏好模型中的过度优化问题，既通过强化学习（训练）针对偏好模型进行优化，也使用偏好模型对输出进行重排序（解码）。他们发现，在训练或解码过程中使用偏好模型都会导致程度相近的过度优化，并且生成模型的规模对这一问题几乎没有帮助。

6 收集和使用人类反馈

收集人类反馈可能相当昂贵，对于缺乏经验的人来说也可能会带来问题，因此充分利用现有资源并谨慎考虑是否需要收集额外数据至关重要。本文介绍了现有数据集及其

特定的正则化项，例如 PPO 中的 KL 项（Schulman 等人，2017年）。

收集方法，并探讨了实验人员为自身用例创建偏好数据集时需要考虑的问题。此外，我们还讨论了使用和收集人类反馈时的伦理考量。

未来，我们或许会收集更丰富类型的反馈，并找到利用这些信号的方法。例如，大多数现有数据集由排序或数值评分构成，但人们更倾向于提供比标注更丰富的反馈（Stumpfet al., 2007; Amershi et al.,2014a;Ghai et al.,2021）。此外，人工标注者之间的差异也尚未得到充分探索（Plank,2022; Gehrmannet al., 2022b）。

6.1 数据收集方面的考量

为生成任务收集人类反馈数据时，需要考虑多个方面；下面列出数据收集可能存在差异的一些维度，但并非面面俱到。

1. 标注者专业水平：根据任务和训练情况（Snow et al., 2008; Sheng et al., 2008; Clark et al., 2021; Gillick and Liu, 2010; Freitag et al., 2021），标注人员可能是领域专家、众包工作者，甚至是模型。2. 参与时长：涉及与标注人员的一次性合作或长期合作；偏好数据集往往需要长期合作（Stiennon et al., 2020; Bai et al., 2022a; Freitag et al., 2021）。3. 收集方法：数据可以通过实验显式收集，也可以从在线来源或用户交互中隐式获取，噪声程度各不相同（Kreutzer et al., 2018; Freitag et al., 2021）。4. 收集平台：常见平台包括亚马逊土耳其机器人、Upwork 和 Scale AI。5. 标注者人口统计特征：不同群体对生成质量的看法可能有所不同；数据收集过程中也可以收集人口统计特征。

创建数据集所需的工作量与所收集判断的可靠性之间通常存在权衡。对于特定领域中风险较高的应用，投入精力与专家标注者建立长期合作或许值得。对于与人类偏好进行一般性对齐，招募一群背景多元的标注人员可能更为审慎，以避免过度拟合于招募时更容易接触到的特定人口统计群体的偏好。

6.2 人类反馈的陷阱与伦理考量

尽管我们在 §2.1 中聚焦于理想化的人类反馈，但实际反馈可能质量低下、相互矛盾，甚至带有对抗性。[5] 如 §3.2 所述，我们必须仔细制定标注指南，确保反馈与模型的实际目标保持一致（Ziegler 等人，2019年）。即使有现成的人类专家，不同专家群体之间也可能意见不一（Kahneman 等人，2021年）。本节将列举人类反馈可能存在的问题，其中大多数问题也常见于其他标注任务。我们还将简要讨论可能的缓解策略。

6.2.1 判断中的主观性和方差

考虑到有 K 名使用反馈函数的标注人员 $h_i{}_{i=1}^{K}$，他们对数据 $\mathcal{D} = d_1$、……、d_N 作出判断。Cohen's Kappa、Fleiss' Kappa 或 Krippendorff's alpha 等评审者间信度指标可用于评估标注者一致性（Hayes and Krippendorff，2007年；Fleiss, 1971；Cohen，1960）。信度较低可能是由于任务或评估标准不明确（Gehrmann 等人，2022b; Thomson and Reiter，2021）、内在的主观性，或存在多种合理解释（Plank，2022；Nie 等人，2020；Gordon 等人，2022）。

缓解策略包括将人类视为会作出带噪声的理性选择（Ghosal 等人，2023年），从多个人类反馈中学习其信度水平（Yamagata 等人，2021年），以及为 COMET 等评估指标添加置信区间（Glushkova 等人，2021年；Zerva 等人，2022年）。明确的标注指南，以及在排序时附上理由说明，有助于减少偏差并提升清晰度（Ziegler 等人，2019年）。

6.2.2 判断偏差

即使所有 K 标注人员对某个数据点的特定判断达成一致，他们也可能全都判断错误。众所周知，人类推理存在偏差，这些偏差可能导致所有标注人员或其中很大一部分判断错误，或未将证据纳入考虑。此外，即使标注人员在技术上没有偏差，也就是说，他们评估的是指令要求评估的任务，指令仍可能不够明确，或导致标注人员评估略有不同的任务，从而造成一种表象：

[5]我们所说的对抗性反馈，是指有意颠倒用户偏好，或以某种系统性方式误导模型的反馈，而不仅仅是噪声数据。

Task	数据集及其描述	收集方法	平台	反馈类型
语言助手	HH-RLHF（Bai 等人，2022a；Perez 等人，2022a）	显式	Upwork，MTurk	排序
语言助手	SHP（Ethayarajh 等人，2023年）	隐式	从 Reddit 抓取	排序/评分
摘要生成	根据反馈生成摘要 (Stiennon 等人，2020年)	显式	Upwork	排序
问答	FeedbackQA（Li 等人，2022年）	显式	MTurk	评分，自然语言
翻译	WMT 指标共享任务（Freitag 等人，2022b）	显式	专业翻译工作流程	多维质量指标，直接评估
摘要生成	TAC 共享任务（TAC-2008、TAC-2009）	显式	**N/A**	评分

表 2：现有人类反馈数据集及其收集方法摘要，这些方法在多个方面各不相同。有关反馈类型的定义，请参阅表 1 。这里区分了两类数据集：一类是专为广义上捕捉人类偏好而设计的数据集，另一类是为更具体用例设计的数据集，例如机器翻译中的多维质量指标/直接评估数据集。**不适用** 表示我们未能找到相关信息。

偏离最初预期任务的系统性偏差（Parmar 等人，2023年）。

锚定／确认偏差：当标注者孤立地看到一段文本时，他们可能无法考虑更好的替代方案，并错误地将该文本标为高质量（Bansal 等人，2021年）。当被要求生成文本时，锚定偏差可能导致人们写出不同于平常风格的内容（Jakesch 等人，2023年；Lehmann 等人，2022年），这可能会影响他们提出何种类型的建议或修正。缓解策略包括让人们对多个不同的输出进行排序，并明确告知他们需要评估哪些维度。

正向偏差： 在传统强化学习环境中向学习者提供反馈时，用户往往会给出远多于负面反馈的正面反馈，这可能导致智能体避开他们在这些场景中实际上试图达成的目标（Amershi 等人，2014b年；Knox 和 Stone，2013年；Thomaz 和 Breazeal，2008）。

6.2.3 伦理考量

标注人员的判断具有一定主观性，这可能源于不同文化或社会群体之间的差异。Santurkar 等人 (2023 年) 对语言模型生成结果中的观点进行测量，展示了不同人口群体在其中的呈现程度。多项研究发现，在具有争议的话题上，基于人类反馈进行微调会使生成结果更符合美国自由派观点（Perez 等人 (2022b)、Hartmann 等人 (2023年）。具有不同人口统计或政治背景的标注人员，可能会对何为有害内容意见不一（Sap 等人 (2022年)、Ding 等人 (2022年）。当要求标注人员作出伦理判断时，这种分歧尤为明显，因为判断可能因文化背景和个人感受而异（Jiang 等人(2022年)、Talat 等人(2022年）。

Steiger 等人 (2021年) 对有害内容审核人员进行了调查，发现持续执行内容审核任务会造成从轻微不适到长期心理伤害等不同程度的危害；然而，内容审核中所研究的有害内容，其严重程度和出现频率可能高于其他类型的人类反馈标注。Shmueli 等人 (2021年) 指出，对开放式输入进行毒性分类和生成是两种可能引发标注人员有害反应的自然语言处理标注任务。他们进一步指出，这已超出美国机构审查委员会豁免所要求的“最低风险”标准，并鼓励使用众包标注的学术研究人员为其研究申请伦理审查。

媒体也将关注点放在了标注人员的公平报酬上，其中一篇 *TIME* 文章[6] 描述了时薪不超过 2 美元的标注人员，他们负责审查有害内容，并为模型训练提供有害性标注。关于众包的研究（Shmueli 等人（2021年）；Rothschild 等人（2022年）；Soratana 等人 （2022年）；Toxtli 等人 （2021年）；Hornuf 和 Vrankar （2022年））提醒我们，报酬不足，尤其是对资源较少地区的劳动者而言，可能构成一种劳动剥削。

7 人工智能反馈

反馈模型通过有效利用反馈，在推动生成技术发展方面发挥了关键作用。然而，它们严重依赖人工输入：例如，Gao等人（2022年）发现，在不同规模的偏好模型中，使用少于1,000次比较仅带来了微小改进，结果接近随机水平。此外，采用静态反馈

[6]https://time.com/6247678/openai-chatgpt-kenya-workers/

可能导致一致性和准确性方面的挑战，因为反馈的整合会改变模型的输出分布。人工智能生成的反馈是一个新兴研究领域，致力于利用大语言模型自身的能力来评估和改进其输出，从而无需持续的人工干预即可提升模型。该领域已出现两种主要方法：

自我AI反馈 第一种方法是使用同一个模型提供反馈并改进其输出。在这种情况下，模型会持续进行自我改进，从自身的评估中学习，并据此不断完善能力。这种方法的例子包括提示模型生成有害响应并对其进行修改以确保无害性（Bai 等人，2022b），或在基于人类反馈的强化学习微调中采用基于规则的奖励模型（OpenAI，2023a）。研究人员还探索了通过少样本提示进行迭代式输出修订等技术（Peng 等人，2023；Shinn 等人，2023；Chen 等人，2023；Paul 等人，2023；Madaan 等人，2023；Yang 等人，2022），使用的语言模型包括 GPT-3.5（Ouyang 等人，2022）和 GPT-4（OpenAI，2023a）。值得注意的是，这些技术应用于经过训练、能够遵循人类指令并使输出符合人类偏好的大语言模型时，展现出了应用潜力。这表明，在训练期间融入人类反馈，能够帮助人工智能模型更好地理解任务要求，使输出符合指令，并充当可靠的反馈机制，从而减少人工干预。有趣的是，提供有价值的AI反馈的能力，可能取决于模型是否接受过人类反馈训练。

外部AI反馈：第二种方法采用一个独立模型，对待改进模型的输出提供反馈。在这种设置中，任务模型通常与单独训练的反馈模型配合使用（Yasunaga 和 Liang，2020年；Madaan 等人，2021年；Welleck 等人，2022年；Bai 等人，2022b；Akyürek 等人，2023年）。这种方法的一个优势是，反馈模型不必是像 GPT-4 这样大型的通用模型。因此，当反馈数据量充足时，训练较小的反馈模型便成为一种颇具吸引力的替代方案。

8 结论

大语言模型的最新发展凸显了人类反馈的必要性，以确保模型表现符合期望，并生成有帮助且无害的文本。在这篇综述论文中，我们概述了近期一系列研究，这些研究旨在利用（人类）反馈改进自然语言生成。

尽管这一领域尚处于起步阶段，但比较现有的所有研究后，可以得出几项重要观察：

1. 大多数反馈格式（以及相应的现有数据集）都未得到充分利用：模型大多使用基于排序或数值形式的反馈进行优化，尤其是在使用反馈模型时。然而，我们有证据表明，大多数形式的反馈也能为改进模型提供有用信号；自然语言反馈似乎是一种很有前景的格式，因为它具有较强的表现力。

2. 利用（人类）反馈的“关键所在”似乎是反馈本身，而非利用反馈的具体方法。尽管近期热门研究着重强调基于人类反馈的强化学习（RLHF），我们的综述揭示了许多利用反馈的其他方法，而且这些方法报告的效果均优于未采用反馈增强的基线；近期的比较研究甚至表明，RLHF 的效果可能不如更简单、更易于利用的方法（Gao等人，2022年；Rafailov 等人，2023年）。不过，目前仍缺少一项更全面、大规模地比较更多方法的研究。

3. 人类反馈在改善模型行为方面发挥什么作用，以及实际需要多少反馈，目前仍不清楚：*AI* 反馈的成功似乎表明，我们可以大幅减少对人类监督的需求，而一些近期研究（Zhou 等人，2023a）则提出疑问：当已有少量包含人类指令的高质量数据可用于监督学习时，是否还需要反馈。

总的来说，我们希望本综述能够帮助研究人员了解当前的技术水平，并找出现有和新兴的反馈来源，以及利用这些反馈的方法。

致谢

本研究得到了欧盟“地平线欧洲”研究与创新行动计划（UTTER，合同编号 101070631）以及 MAIA 和 NextGenAI 项目（LISBOA-01-0247-FEDER-045909 和 2022-C05i0102-02）的支持。

参考摘要

Afra Feyza Akyürek、Ekin Akyürek、Aman Madaan、Ashwin Kalyan、Peter Clark、Derry Wijaya 和 Niket Tandon。2023年。RL4F：利用强化学习生成自然语言反馈，以修复模型输出。

Saleema Amershi、Maya Cakmak、William Bradley Knox 和 Todd Kulesza。2014a年。权力属于人民：人类在交互式机器学习中的作用。《AI Magazine》，35(4):105–120。

Saleema Amershi、Maya Cakmak、William Bradley Knox 和 Todd Kulesza。2014b年。权力属于人民：人类在交互式机器学习中的作用。《AI Magazine》，35(4):105–120。

Dario Amodei、Chris Olah、Jacob Steinhardt、Paul F. Christiano、John Schulman 和 Dan Mané。2016年。人工智能安全中的具体问题。CoRR，abs/1606.06565。

Chantal Amrhein 和 Rico Sennrich。2022年。通过最小贝叶斯风险解码发现机器翻译指标的弱点：COMET 案例研究。载于第 2 届计算语言学协会亚太分会会议暨第 12 届自然语言处理国际联合会议论文集（第 1 卷：长篇论文），第 1125–1141 页，仅线上。计算语言学协会。

Kushal Arora、Kurt Shuster、Sainbayar Sukhbaatar 和 Jason Weston。2022年。Director：用于监督式语言模型训练的生成器分类器。载于第 2 届计算语言学协会亚太分会会议暨第 12 届自然语言处理国际联合会议论文集。

Amanda Askell、YuntaoBai、Anna Chen、Dawn Drain、Deep Ganguli、Tom Henighan、Andy Jones、Nicholas Joseph、Ben Mann、Nova Das- Sarma 等。2021年。《作为对齐实验室的通用语言助手》。arXiv 预印本，arXiv:2112.00861。

Karl Johan Åström 和 Richard M Murray。2021年。《反馈系统：面向科学家和工程师的导论》。普林斯顿大学出版社。

Yuntao Bai、Andy Jones、Kamal Ndousse、Amanda Askell、Anna Chen、Nova DasSarma、Dawn Drain、Stanislav Fort、Deep Ganguli、Tom Henighan、Nicholas Joseph、Saurav Kadavath、JacksonKernion、TomConerly、Sheer El-Showk、Nelson Elhage、Zac Hatfield-Dodds、Danny Hernandez、Tristan Hume、Scott Johnston、Shauna Kravec、Liane Lovitt、Neel Nanda、Catherine Olsson、Dario Amodei、Tom Brown、Jack Clark、Sam McCandlish、Chris Olah、Ben Mann 和 Jared Kaplan。2022a。通过人类反馈强化学习训练有帮助且无害的助手。

Yuntao Bai、Saurav Kadavath、Sandipan Kundu、Amanda Askell、Jackson Kernion、Andy Jones、Anna Chen、Anna Goldie、Azalia Mirhoseini、Cameron McKinnon、Carol Chen、Catherine Olsson、Christopher Olah、Danny Hernandez、Dawn Drain、Deep Ganguli、Dustin Li、Eli Tran-Johnson、Ethan Perez、Jamie Kerr、Jared Mueller、Jeffrey Ladish、Joshua Landau、Kamal Ndousse、Kamile Lukosuite、Liane Lovitt、Michael Sell- itto、Nelson Elhage、Nicholas Schiefer、Noemi Mercado、Nova DasSarma、Robert Lasenby、Robin Larson、Sam Ringer、Scott Johnston、Shauna Kravec、Sheer El Showk、Stanislav Fort、Tamera Lanham、Timothy Telleen-Lawton、Tom Conerly、Tom Henighan、Tristan Hume、Samuel R. Bowman、Zac Hatfield-Dodds、Ben Mann、Dario Amodei、Nicholas Joseph、Sam Mc- Candlish、Tom Brown 和 Jared Kaplan。2022b。宪法式人工智能：来自人工智能反馈的无害性。

Gagan Bansal、Tongshuang Wu、Joyce Zhou、Raymond Fok、Besmira Nushi、Ece Kamar、Marco Tulio Ribeiro 和 Daniel Weld。2021年。整体是否大于部分之和？人工智能解释对互补团队绩效的影响。载于《2021 年计算机系统中的人为因素国际会议论文集》，第1–16页。

Emily M. Bender、Timnit Gebru、Angelina McMillan-Major 和 Shmargaret Shmitchell。2021年。论随机鹦鹉的危险：语言模型会不会太大？载于《2021年ACM公平性、问责制与透明度会议论文集》（FAccT '21），第610–623页，美国纽约州纽约市。计算机协会。

Pushpak Bhattacharyya、Rajen Chatterjee、Markus Freitag、Diptesh Kanojia、Matteo Negri 和 Marco Turchi。2022年。WMT 2022自动译后编辑共享任务的研究结果。载于《第七届机器翻译大会（WMT）论文集》，第109–117页，阿布扎比，阿拉伯联合酋长国（混合形式）。计算语言学协会。

Florian Böhm、Yang Gao、Christian M. Meyer、Ori Shapira、Ido Dagan 和 Iryna Gurevych。2019年。更好的奖励带来更好的摘要：学习在没有参考文本的情况下进行摘要生成。载于《2019年自然语言处理实证方法会议和第9届自然语言处理国际联合会议（EMNLP-IJCNLP）论文集》，第3110–3120页，中国香港。计算语言学协会。

Tolga Bolukbasi、Kai-Wei Chang、James Y Zou、Venkatesh Saligrama 和 Adam T Kalai。2016年。男人之于计算机程序员，正如女人之于家庭主妇？词嵌入的去偏。神经信息处理系统进展，第29卷。

Rishi Bommasani、DrewA. Hudson、Ehsan Adeli、Russ Altman、Simran Arora、Sydney von Arx、Michael S. Bernstein、Jeannette Bohg、Antoine Bosselut、Emma Brunskill、Erik Brynjolfsson、S. Buch、Dallas Card、Rodrigo Castellon、Ni- ladri S. Chatterji、Annie S. Chen、Kathleen A. Creel、Jared Davis、Dora Demszky、Chris Don- ahue、Moussa Doumbouya、Esin Durmus、Ste- fano Ermon、John Etchemendy、Kawin Etha- yarajh、Li Fei-Fei、Chelsea Finn、Trevor Gale、Lauren E. Gillespie、Karan Goel、Noah D. Goodman、Shelby Grossman、Neel Guha、Tat- sunori Hashimoto、Peter Henderson、John He- witt、Daniel E. Ho、Jenny Hong、Kyle Hsu、Jing Huang、Thomas F. Icard、Saahil Jain、Dan Jurafsky、Pratyusha Kalluri、Siddharth Karamcheti、Geoff Keeling、Fereshte Khani、O. Khattab、Pang Wei Koh、Mark S. Krass、Ranjay Krishna、Rohith Kuditipudi、Ananya Kumar、Faisal Ladhak、Mina Lee、Tony Lee、Jure Leskovec、Isabelle Levent、Xiang Lisa Li、Xuechen Li、Tengyu Ma、Ali Malik、Christopher D. Manning、Suvir Mirchandani、Eric Mitchell、Zanele Mun- yikwa、Suraj Nair、Avanika Narayan、Deepak Narayanan、Benjamin Newman、Allen Nie、Juan Carlos Niebles、Hamed Nilforoshan、J. F. Nyarko、Giray Ogut、Laurel J. Orr、Isabel Papadimitriou、Joon Sung Park、Chris Piech、Eva Portelance、Christopher Potts、Aditi Raghunathan、Robert Reich、Hongyu Ren、Frieda Rong、Yusuf H. Roohani、Camilo Ruiz、Jack Ryan、Christopher R'e、Dorsa Sadigh、Shiori Sagawa、Keshav Santhanam、Andy Shih、Krishna Para- suram Srinivasan、Alex Tamkin、Rohan Taori、Armin W. Thomas、Florian Tramèr、Rose E. Wang、William Wang、Bohan Wu、Jiajun Wu、Yuhuai Wu、Sang Michael Xie、Michihiro Ya- sunaga、Jiaxuan You、Matei A. Zaharia、Michael Zhang、Tianyi Zhang、Xikun Zhang、Yuhui Zhang、Lucia Zheng、Kaitlyn Zhou 和 Percy Liang。2021年。《基础模型的机遇与风险》。arXiv，abs/2108.07258。

Nick Bostrom。2014年。超级智能：路径、危险与策略，第一版。牛津大学出版社，美国。

Arun Chaganty、Stephen Mussmann 和 Percy Liang。2018年。自然语言评估中自动指标去偏的代价。载于《第五十六届计算语言学协会年会论文集》（第1卷：长篇论文），第643–653页，澳大利亚墨尔本。计算语言学协会。

XinyunChen、Maxwell Lin、Nathanael Schärli 和 Denny Zhou。2023 年。教导大型语言模型进行自我调试。arXiv 预印本 arXiv:2304.05128。

Elizabeth Clark、Tal August、Sofia Serrano、Nikita Haduong、Suchin Gururangan 和 Noah A. Smith。2021年。《并非所有"人类"的都是黄金：评估对生成文本的人本评估》。载于《第五十九届计算语言学协会年会论文集》和《第十一届自然语言处理国际联合会议论文集》（第1卷：长篇

论文），第7282–7296页，线上。计算语言学协会。

Jacob Cohen。1960年。名义量表的一致性系数。《教育与心理测量》，20：37–46。

Gonçalo M. Correia 和 André F. T. Martins。2019年。利用迁移学习进行自动译后编辑的一种简单有效的方法。载于《计算语言学协会第57届年会论文集》，第3050–3056页，意大利佛罗伦萨。计算语言学协会。

Michael Denkowski、ChrisDyer 和 Alon Lavie。2014年。从译后编辑中学习：统计机器翻译的线上模型适应。载于《计算语言学协会欧洲分会第14届会议论文集》，第395–404页，瑞典哥德堡。计算语言学协会。

Yi Ding、Jacob You、Tonja-Katrin Machulla、JenniferJacobs、Pradeep Sen 和 Tobias Höllerer。2022年。标注者人口统计特征对情感数据集标注的影响。《ACM人机交互学报》，6(CSCW2)。

Jesse Dodge、Maarten Sap、Ana Marasovi´c、William Agnew、Gabriel Ilharco、Dirk Groen- eveld、Margaret Mitchell 和 Matt Gardner。2021年。记录大型网页文本语料库：一项关于庞大、清洗后的抓取语料库的案例研究。载于《2021年自然语言处理经验方法会议论文集》，第1286–1305页，线上及多米尼加共和国蓬塔卡纳。计算语言学协会。

Ahmed Elgohary、Christopher Meek、Matthew Richardson、Adam Fourney、Gonzalo Ramos 和 Ahmed Hassan Awadallah。2021年。Nl-edit：通过自然语言交互纠正语义解析错误。arXiv预印本，arXiv:2103.14540。

Kawin Ethayarajh、Heidi Zhang、Yizhong Wang 和 Dan Jurafsky。2023年。斯坦福人类偏好数据集。

Patrick Fernandes、António Farinhas、Ricardo Rei、José De Souza、PerezOgayo、Graham Neubig 和 Andre Martins。2022年。神经机器翻译的质量感知解码。载于《2022年计算语言学协会北美分会会议论文集：人类语言技术》，第1396–1412页，西雅图，美国。计算语言学协会。

J.L Fleiss。1971年。衡量多位评定者之间的名义尺度一致性。《心理学公报》，76:378–382。

Markus Freitag、George Foster、David Grang- ier、Viresh Ratnakar、Qijun Tan 和 Wolf- gang Macherey。2021年。专家、错误与语境：机器翻译人工评估的大规模研究。计算语言学协会汇刊，9：1460–1474。

Markus Freitag、David Grangier、Qijun Tan 和 Bowen Liang。2022a。《高质量而非高模型概率：基于神经指标的最小贝叶斯风险解码》。《计算语言学协会汇刊》，10：811–825。

Markus Freitag、Ricardo Rei、Nitika Mathur、Chi- kiu Lo、Craig Stewart、Eleftherios Avramidis、Tom Kocmi、George Foster、Alon Lavie 和 André F. T. Martins。2022b。WMT22 指标共享任务结果：停止使用 BLEU——神经指标更出色、更稳健。载于第七届机器翻译大会（WMT）论文集，第46–68 页，阿布扎比，阿拉伯联合酋长国（混合形式）。计算语言学协会。

Leo Gao、John Schulman 和 Jacob Hilton。2022年。奖励模型过度优化的缩放定律。

Yang Gao、Christian M. Meyer 和 Iryna Gurevych。2018年。APRIL：通过结合主动偏好学习和强化学习，交互式地学习摘要生成。载于《2018年自然语言处理实证方法会议论文集》，第4120–4130页，布鲁塞尔，比利时。计算语言学协会。

Sebastian Gehrmann、Abhik Bhattacharjee、AbinayaMahendiran、Alex Wang、Alexandros Pa-

pangelis、Aman Madaan、Angelina McMillan-Major、Anna Shvets、Ashish Upadhyay、Bingsheng Yao、Bryan Wilie、Chandra Bhagavatula、Chaobin You、Craig Thomson、Cristina Garbacea、Dakuo Wang、Daniel Deutsch、Deyi Xiong、Di Jin、Dimitra Gkatzia、Dragomir R. Radev、Elizabeth Clark、Esin Durmus、Faisal Ladhak、Filip Ginter、Genta Indra Winata、Hendrik Strobelt、Hiroaki Hayashi、Jekaterina Novikova、Jenna Kanerva、Jenny Chim、Jiawei Zhou、Jordan Clive、Joshua Maynez、João Sedoc、Juraj Juraska、Kaustubh D. Dhole、Khyathi Raghavi Chandu、Leonardo F. R. Ribeiro、Lewis Tunstall、Li Zhang、Mahima Pushkarna、Mathias Creutz、Michael White、Mihir Kale、Moussa Kamal Eddine、Nico Daheim、Nis- hant Subramani、Ondrej Dusek、Paul Pu Liang、Pawan Sasanka Ammanamanchi、Qinqin Zhu、Ratish Puduppully、Reno Kriz、Rifat Shahri- yar、Ronald Cardenas、Saad Mahamood、Sa- lomey Osei、Samuel Cahyawijaya、Sanja vSta- jner、Sébastien Montella、Shailza、Shailza Jolly、Simon Mille、Tahmid Hasan、Tianhao Shen、Tosin P. Adewumi、Vikas Raunak、Vipul Raheja、Vitaly Nikolaev、Vivian Tsai、Yacine Jernite、Yi Xu、Yisi Sang、Yixin Liu 和 Yu- fang Hou。2022a。Gemv2：单行代码实现多语言自然语言生成基准测试。载于自然语言处理经验方法会议。

SebastianGehrmann、Elizabeth Clark 和 Thibault Sellam。2022b。《修补破裂的根基：生成文本评估实践中的障碍综述》。arXiv 预印本，arXiv:2202.06935。

Bhavya Ghai、Q Vera Liao、Yunfeng Zhang、Rachel Bellamy 和 Klaus Mueller。2021年。《可解释主动学习（XAL）：将人工智能解释作为机器教师的交互界面》。ACM人机交互会议论文集，4(CSCW3)：1–28。

Gaurav R. Ghosal、Matthew Zurek、Daniel S. Brown 和 Anca D. Dragan。2023年。《建模人类理性水平对从多种反馈类型中学习奖励的影响》。

Dan Gillick 和 Yang Liu。2010。非专家评估摘要系统存在风险。载于 NAACL HLT 2010语音与语言创建研讨会论文集

Amazon Mechanical Turk 数据，第 148–151 页，Los Angeles。计算语言学协会。

Amelia Glaese、Nat McAleese、Maja Tr.ebacz、John Aslanides、Vlad Firoiu、Timo Ewalds、Maribeth Rauh、Laura Weidinger、Martin Chadwick、Phoebe Thacker 等。2022年。《通过有针对性的人类判断改进对话代理的对齐》。arXiv 预印本，arXiv:2209.14375。

TaisiyaGlushkova、Chrysoula Zerva、Ricardo Rei 和 André F. T. Martins。2021年。不确定性感知的机器翻译评估。载于《计算语言学协会论文集：EMNLP 2021》。计算语言学协会。

C. A. E. Goodhart。1984年。《货币管理问题：英国经验》。麦克米伦教育英国，伦敦。

Mitchell L. Gordon、Michelle S. Lam、Joon Sung Park、Kayur Patel、Jeff Hancock、Tatsunori Hashimoto 和 Michael S. Bernstein。2022年。Jurylearning：将异议声音整合到机器学习模型中。载于计算机系统中的人为因素会议。ACM。

Yvette Graham、Timothy Baldwin、Alistair Mof- fat 和 Justin Zobel。2013年。在机器翻译的人类评估中使用连续测量尺度。载于第七届语言学标注研讨会和话语互操作性会议论文集，第 33–41 页，保加利亚索非亚。计算语言学协会。

Braden Hancock、Antoine Bordes、Pierre- Emmanuel Mazare 和 Jason Weston。2019年。《部署后的对话学习：自给自足，聊天机器人！》载于第57届计算语言学协会年会论文集，第3667–3684页。

Jochen Hartmann、Jasper Schwenzow 和 Maxim- ilian Witte。2023年。《对话式人工智能的政治意识形态：关于 ChatGPT 亲环境、左翼自由意向的趋同证据》。

A.F Hayes 和 K Krippendorff。2007年。响应号召，为数据编码制定标准可靠性测量。《传播方法与测量》，1:77–89。

Dan Hendrycks、Collin Burns、Steven Basart、An- drew Critch、Jerry Li、Dawn Song 和 Jacob Steinhardt。2020年。《使人工智能与共同的人类价值观保持一致》。arXiv 预印本 arXiv:2008.02275。

Lars Hornuf 和 Daniel Vrankar。2022年。众包工作时薪：一项荟萃分析。《商业与信息系统工程》，64(5)：553–573。

Maurice Jakesch、Advait Bhat、Daniel Buschek、Lior Zalmanson 和 Mor Naaman。2023年。与带有观点倾向的语言模型共同写作会影响用户的观点。arXiv，abs/2302.00560。

Natasha Jaques、Asma Ghandeharioun、Judy Hanwen Shen、Craig Ferguson、Àgata Lapedriza、Noah Jones、Shixiang Gu 和 Rosalind W. Picard。2019年。在对话中对隐含的人类偏好进行远离策略的批量深度强化学习。CoRR，abs/1907.00456。

Liwei Jiang、Jena D. Hwang、Chandra Bhagavatula、Ronan Le Bras、Jenny Liang、Jesse Dodge、Keisuke Sakaguchi、Maxwell Forbes、Jon Borchardt、Saadia Gabriel、Yulia Tsvetkov、Oren Etzioni、Maarten Sap、Regina Rini 和 Yejin Choi。2022年。机器能学会道德吗？德尔菲实验。

Daniel Kahneman、Sibony Olivier 和 Cass R. Sunstein。2021年。Little, Brown Spark，美国纽约州纽约市。

Zachary Kenton、Tom Everitt、Laura Weidinger、IasonGabriel、VladimirMikulik 和 Geoffrey Irving。2021a。语言智能体的对齐。arXiv预印本，arXiv:2103.14659。

Zachary Kenton、Tom Everitt、Laura Weidinger、IasonGabriel、VladimirMikulik 和 Geoffrey Irving。2021b。语言智能体的对齐。CoRR，abs/2103.14659。

W. Bradley Knox 和 Peter Stone。2013年。从人类生成的奖励中进行非短视学习。载于《2013年国际智能用户界面会议论文集》，IUI '13，第191–202页，美国纽约州纽约市。计算机协会。

Tomasz Korbak、Kejian Shi、Angelica Chen、Rasika Bhalerao、Christopher L. Buckley、Jason Phang、Samuel R. Bowman 和 Ethan Perez。2023年。使用人类偏好预训练语言模型。

Julia Kreutzer、Shahram Khadivi、Evgeny Matusov 和 Stefan Riezler。2018年。神经机器翻译能否通过用户反馈得到改进？载于《计算语言学协会北美分会2018年会议论文集：人类语言技术》第3卷（行业论文），第92–105页，新奥尔良－路易斯安那州。计算语言学协会。

Shankar Kumar 和 William Byrne。2002年。双语文本的最小贝叶斯风险词语对齐。载于《ACL-02自然语言处理经验方法会议论文集》第10卷，EMNLP'02，第140–147页，美国。计算语言学协会。

Florian Lehmann、Niklas Markert、Hai Dang 和 Daniel Buschek。2022年。《建议列表与连续生成：在移动设备上使用生成模型进行写作的交互设计如何影响文本长度、措辞和感知到的作者身份》。Mensch und Computer 2022 论文集。

Jan Leike、David Krueger、Tom Everitt、Miljan Martic、Vishal Maini 和 Shane Legg。2018年。《通过奖励建模实现可扩展的智能体对齐：一个研究方向》。arXiv，abs/1811.07871。

Jiwei Li、Alexander H Miller、Sumit Chopra、Marc'Aurelio Ranzato 和 Jason Weston。2017年。《通过人类参与的循环进行对话学习》。载于《国际学习表征会议》。

Zichao Li、Prakhar Sharma、Xing Han Lu、Jackie Cheung 和 Siva Reddy。2022年。《利用交互式反馈提高问答系统部署后的准确性和可解释性》。载于《计算语言学协会论文集：ACL 2022》。计算语言学协会。

Bing Liu、Gokhan Tür、Dilek Hakkani-Tür、Pararth Shah 和 Larry Heck。2018年。通过人工教学和反馈，在端到端可训练的任务型对话系统中进行对话学习。载于《2018年计算语言学协会北美分会人类语言技术会议论文集：第1卷（长篇论文）》，第

2060–2069页，新奥尔良，路易斯安那州。计算语言学协会。

Hao Liu、Carmelo Sferrazza 和 Pieter Abbeel。2023年。《语言即奖励：使用人类反馈进行事后微调》。arXiv 预印本，arXiv:2302.02676。

Yixin Liu、Pengfei Liu、Dragomir Radev 和 Graham Neubig。2022年。BRIO：为抽象式摘要生成建立秩序。载于计算语言学协会第60届年会论文集（第1卷：长篇论文），第2890–2903页，都柏林，爱尔兰。计算语言学协会。

Arle Lommel、Aljoscha Burchardt 和 Hans Uszkoreit。2014a年。多维质量指标（多维质量指标）：一种声明和描述翻译质量指标的框架。Tradumàtica: 翻译技术，0：455–463。

Arle Lommel、Hans Uszkoreit 和 Aljoscha Burchardt。2014b年。多维质量指标（多维质量指标）：一种声明和描述翻译质量指标的框架。Revista Tradumàtica: 翻译技术。

Aman Madaan、Niket Tandon、Peter Clark 和 Yiming Yang。2022年。借助记忆辅助提示编辑，在部署后提升 GPT-3 的表现。载于《2022年自然语言处理经验方法会议论文集》，第 2833–2861 页，阿布扎比，阿拉伯联合酋长国。计算语言学协会。

Aman Madaan、Niket Tandon、Prakhar Gupta、Skyler Hallinan、Luyu Gao、Sarah Wiegreffe、Uri Alon、Nouha Dziri、Shrimai Prabhumoye、Yiming Yang、Sean Welleck、Bodhisattwa Prasad Majumder、Shashank Gupta、Amir Yazdanbakhsh 和 Peter Clark。2023年。自我精炼：通过自我反馈进行迭代优化。

Aman Madaan、Niket Tandon、Dheeraj Rajagopal、Peter Clark、Yiming Yang 和 Eduard Hovy。2021年。仔细思考！通过先对问题情境建模来改进可废止推理。载于《2021年自然语言处理实证方法会议论文集》，第6291–6310页，线上及多米尼加共和国蓬塔卡纳。计算语言学协会。

Nitika Mathur、Timothy Baldwin 和 Trevor Cohn。2020年。深陷 BLEU：重新评估自动机器翻译评估指标的评估。载于《第58届计算语言学协会年会论文集》，第4984–4997页，线上。计算语言学协会。

Nikhil Mehta 和 Dan Goldwasser。2019年。借助建议改进与机器人的自然语言交互。arXiv 预印本 arXiv:1905.04655。

Richard Ngo。2022年。从深度学习视角看对齐问题。arXiv 预印本 arXiv:2209.00626。

Duy-Hung Nguyen、Nguyen Viet Dung Nghiem、Bao-Sinh Nguyen、Dung Tien Tien Le、Sha- hab Sabahi、Minh-Tien Nguyen 和 Hung Le。2022年。充分利用先验数据：一种通过偏好反馈实现交互式文本摘要生成的解决方案。载于《计算语言学协会论文集：NAACL 2022》，第 1919–1930 页，美国西雅图。计算语言学协会。

Yixin Nie、Xiang Zhou 和 Mohit Bansal。2020年。我们能从人们对自然语言推理数据的集体意见中学到什么？arXiv 预印本 arXiv:2010.03532。

Erik Nijkamp、Bo Pang、Hiroaki Hayashi、Lifu Tu、Huan Wang、Yingbo Zhou、Silvio Savarese 和 Caiming Xiong。2022年。一种用于程序合成的对话范式。arXiv 电子预印本，第 arXiv–2203 页。

OpenAI。2023a。GPT-4 技术报告。

OpenAI。2023b。面向研究人员的模型索引。访问日期：2023-05-01。

Long Ouyang、Jeff Wu、Xu Jiang、Diogo Almeida、Carroll L. Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Ray、John Schulman、Jacob Hilton、Fraser Kelton、Luke Miller、Maddie Simens、Amanda Askell、Peter Welinder、Paul Christiano、Jan Leike 和 Ryan Lowe。2022年。通过人类反馈训练语言模型遵循指令。

Santanu Pal、Sudip Kumar Naskar、Mihaela Vela 和 Josef van Genabith。2016年。一种用于自动译后编辑的神经网络方法。

载于《第54届计算语言学协会年会论文集》（第2卷：短篇论文），第281–286页，德国柏林。计算语言学协会。

Mihir Parmar、Swaroop Mishra、Mor Geva 和 Chitta Baral。2023 年。别责怪标注者：偏见早在标注说明中就已出现。

Tatiana Passali、Alexios Gidiotis、Efstathios Chatzikyriakidis 和 Grigorios Tsoumakas。2021年。迈向以人为中心的摘要生成：财经新闻案例研究。载于《首届人机交互与自然语言处理桥接研讨会论文集》，第21–27页，线上。计算语言学协会。

Debjit Paul、Mete Ismayilzada、Maxime Peyrard、Beatriz Borges、Antoine Bosselut、Robert West 和 Boi Faltings。2023年。Refiner：针对中间表示的推理反馈。arXiv预印本，arXiv:2304.01904。

Romain Paulus、CaimingXiong 和 Richard Socher。2017年。一种用于抽象式摘要的深度强化模型。CoRR，abs/1705.04304。

Baolin Peng、Michel Galley、Pengcheng He、Hao Cheng、Yujia Xie、Yu Hu、Qiuyuan Huang、Lars Lidén、Zhou Yu、Weizhu Chen 和 Jianfeng Gao。2023年。《核查事实并重试：利用外部知识和自动反馈改进大语言模型》。arXiv，abs/2302.12813。

Ethan Perez、Saffron Huang、Francis Song、Trevor Cai、Roman Ring、John Aslanides、Amelia Glaese、Nat McAleese 和 Geoffrey Irving。2022a。使用语言模型对语言模型进行红队测试。

Ethan Perez、Sam Ringer、Kamil˙e Lukoši¯ut˙e、Karina Nguyen、Edwin Chen、Scott Heiner、Craig Pettit、Catherine Olsson、Sandipan Kundu、Saurav Kadavath、Andy Jones、Anna Chen、Ben Mann、Brian Israel、Bryan Seethor、Cameron McKinnon、ChristopherOlah、DaYan、Daniela Amodei、Dario Amodei、Dawn Drain、Dustin Li、Eli Tran-Johnson、Guro Khundadze、Jack- son Kernion、James Landis、Jamie Kerr、Jared Mueller、Jeeyoon Hyun、Joshua Landau、Kamal Ndousse、Landon Goldberg、Liane Lovitt、Martin Lucas、Michael Sellitto、Miranda Zhang、Neerav Kingsland、Nelson Elhage、Nicholas Joseph、Noemí Mercado、Nova DasSarma、Oliver Rausch、Robin Larson、Sam McCandlish、Scott Johnston、Shauna Kravec、Sheer El Showk、Tamera Lanham、Timothy Telleen-Lawton、Tom Brown、Tom Henighan、Tristan Hume、Yuntao Bai、Zac Hatfield-Dodds、Jack Clark、Samuel R. Bowman、Amanda Askell、Roger Grosse、Danny Hernandez、Deep Ganguli、Evan Hubinger、Nicholas Schiefer 和 Jared Kaplan。2022b。通过模型编写的评估发现语言模型行为。

Maxime Peyrard、Teresa Botschen 和 Iryna Gurevych。2017年。《学习为系统摘要评分，以改进内容选择评估》。载于《摘要新前沿研讨会论文集》，第74–84页，丹麦哥本哈根。计算语言学协会。

Maxime Peyrard 和 Iryna Gurevych。2018年。学习目标函数以匹配人类评审对基于优化的摘要生成的判断。载于《2018年计算语言学协会北美分会年会：人类语言技术论文集》，第 2 卷（短篇论文），第 654–660 页，新奥尔良，路易斯安那州。计算语言学协会。

Barbara Plank。2022年。人类标签变异的“问题”：关于数据、建模与评估中的真实标注。

Yiwei Qin、Weizhe Yuan、Graham Neubig 和 Pengfei Liu。2022年。T5score：对生成式评估指标进行判别式微调。

Rafael Rafailov、Archit Sharma、Eric Mitchell、Stefano Ermon、Christopher D. Manning 和 Chelsea Finn。2023年。直接偏好优化：你的语言模型其实暗中就是一个奖励模型。

Ricardo Rei、Craig Stewart、Ana C Farinha 和 Alon Lavie。2020年a。COMET：用于机器翻译评估的神经框架。载于《2020年自然语言处理实证方法会议论文集》（EMNLP），页数。

2685–2702，线上。计算语言学协会。

Ricardo Rei、Craig Stewart、Ana C Farinha 和 Alon Lavie。2020年。Unbabel 参与 WMT20 指标共享任务。载于《第五届机器翻译大会论文集》，第 911–920 页，线上。计算语言学协会。

Machel Reid 和 Graham Neubig。2022年。学习对编辑过程进行建模。arXiv 预印本 arXiv:2205.12374。

Christopher Riesbeck。1981年。面向增量学习的失败驱动提醒。载于 IJCAI，第 115–120 页。Citeseer。

Arturo Rosenblueth、Norbert Wiener 和 Julian Bigelow。1943年。行为、目的与目的论。《科学哲学》，10(1)：18–24。

Annabel Rothschild、Justin Booker、Christa Davoll、Jessica Hill、Venise Ivey、Carl DiSalvo、Ben Rydal Shapiro 和 Betsy DiSalvo。2022年。为采集机器学习训练数据，推动公平且亲社会的数字零工就业。载于《计算机系统中的人为因素会议扩展摘要》，第 1–9 页。

Ananya B. Sai、Akash Kumar Mohankumar 和 Mitesh M. Khapra。2022 年。自然语言生成系统所用评估指标综述。《ACM 计算调查》，55(2)。

Shibani Santurkar、Esin Durmus、Faisal Ladhak、Cinoo Lee、Percy Liang 和 Tatsunori Hashimoto。2023年。语言模型反映的是谁的观点?

Maarten Sap、Swabha Swayamdipta、Laura Vianna、Xuhui Zhou、Yejin Choi 和 Noah A. Smith。2022年。持有态度的标注人员：标注者信念与身份如何对有害语言检测产生偏见。

William Saunders、Catherine Yeh、Jeff Wu、Steven Bills、Long Ouyang、Jonathan Ward 和 Jan Leike。2022年。用于辅助人工评估者的自我批判模型。

Roger C Schank。1983年。《动态记忆：计算机与人类中的提醒与学习理论》。剑桥大学出版社。

Jérémy Scheurer、Jon Ander Campos、Jun Shern Chan、Angelica Chen、Kyunghyun Cho 和 Ethan Perez。2022年。通过语言反馈训练语言模型。Jérémy Scheurer、Jon Ander Campos、Tomasz Korbak、Jun Shern Chan、Angelica Chen、Kyunghyun Cho 和 Ethan Perez。2023年。大规模通过语言反馈训练语言模型。

Timo Schick、Jane Dwivedi-Yu、Zhengbao Jiang、Fabio Petroni、Patrick Lewis、Gautier Izac- ard、Qingfei You、Christoforos Nalmpantis、Edouard Grave 和 Sebastian Riedel。2022年。Peer：一种协作式语言模型。arXiv，abs/2208.11663。

Natalie Schluter。2017年。根据 ROUGE 衡量自动摘要的局限性。载于《第15届欧洲计算语言学协会分会会议论文集：第2卷，短篇论文》，第41–45页，西班牙瓦伦西亚。计算语言学协会。

John Schulman、Filip Wolski、Prafulla Dhariwal、Alec Radford 和 Oleg Klimov。2017年。近端策略优化算法。arXiv 预印本，arXiv:1707.06347。

Thibault Sellam、Dipanjan Das 和 Ankur Parikh。2020年。BLEURT：学习用于文本生成的稳健指标。载于《第58届计算语言学协会年会论文集》，第7881–7892页，线上。计算语言学协会。

Shiqi Shen、Yong Cheng、Zhongjun He、Wei He、Hua Wu、Maosong Sun 和 Yang Liu。2016年。神经机器翻译的最小风险训练。载于《第54届计算语言学协会年会论文集（第1卷：长篇论文）》，第1683–1692页，德国柏林。计算语言学协会。

EmilySheng、Kai-Wei Chang、PremkumarNatara- jan 和 Nanyun Peng。2019年。《这位女性曾担任保姆：论语言生成中的偏差》。收录于2019年自然语言处理实证方法会议和第九届自然语言国际联合会议论文集，

处理（EMNLP-IJCNLP），第3407–3412页，中国香港。计算语言学协会。

Victor S. Sheng、Foster Provost 和 Panagiotis G. Ipeirotis。2008年。获取另一个标注？使用多个带噪标注者改进数据质量和数据挖掘。载于第14届ACM SIGKDD知识发现与数据挖掘国际会议论文集，KDD '08，第614–622页，美国纽约州纽约市。计算机协会。

Weiyan Shi、Yu Li、Saurav Sahay 和 Zhou Yu。2021年。改进与模仿：通过强化学习和人类示范减少劝说对话中的重复与不一致。载于计算语言学协会论文集：EMNLP 2021，第3478–3492页，蓬塔卡纳，多米尼加共和国。计算语言学协会。

Noah Shinn、Beck Labash 和 Ashwin Gopinath。2023年。Reflexion：具有动态记忆和自我反思能力的自主智能体。

Boaz Shmueli、Jan Fell、Soumya Ray 和 Lun-Wei Ku。2021年。《超越公平报酬：自然语言处理众包的伦理影响》。载于《2021年计算语言学协会北美分会人类语言技术会议论文集》，第3758–3769页，线上。计算语言学协会。

Michel Simard、Nicola Ueffing、Pierre Isabelle 和 Roland Kuhn。2007年。《采用统计式短语后编辑的基于规则的翻译》。载于《第二届统计机器翻译研讨会论文集》，第203–206页。

Rion Snow、Brendan O'Connor、Daniel Jurafsky 和 Andrew Ng。2008年。《便宜又快速——但效果好吗？对自然语言任务的非专家标注进行评估》。载于《2008年自然语言处理实证方法会议论文集》，第254–263页，夏威夷檀香山。计算语言学协会。

Teerachart Soratana、Yili Liu 和 X JessieYang。2022年。支付率和国家的收入水平对众包任务的态度。在《人因与工效学学会年会论文集》中，第66卷，第2220–2224页。SAGE出版公司，SAGE CA：洛杉矶，加利福尼亚州。

Miriah Steiger、Timir J Bharucha、Sukrit Venkata- giri、Martin J. Riedl 和 Matthew Lease。2021年。内容审核员的心理健康：商业内容审核中的情绪劳动及改善支持的途径。载于《2021年计算机系统中的人为因素国际会议论文集》，CHI'21，美国纽约州纽约市。计算机协会。

NisanStiennon、Long Ouyang、Jeff Wu、Daniel M. Ziegler、Ryan Lowe、Chelsea Voss、Alec Radford、Dario Amodei 和 Paul Christiano。2020年。学习如何根据人类反馈进行摘要。

Simone Stumpf、Vidya Rajaram、Lida Li、Mar- garet Burnett、Thomas Dietterich、Erin Sulli- van、Russell Drummond 和 Jonathan Her- locker。2007年。探索利用用户反馈改进机器学习。载于《第12届国际智能用户界面会议论文集》，第82–91页。

Zeerak Talat、Hagen Blix、Josef Valvoda、Maya Indira Ganesh、Ryan Cotterell 和 Adina Williams。2022年。《从自然语言中学习伦理判断》。载于《2022年计算语言学协会北美分会人类语言技术会议论文集》，第769–779页，美国西雅图。计算语言学协会。

Alon Talmor、Oyvind Tafjord、Peter Clark、Yoav Goldberg 和 Jonathan Berant。2020年。《教导预训练模型系统地推理隐含知识》。arXiv预印本，arXiv:2006.06609，4(6)。

Niket Tandon、Aman Madaan、Peter Clark 和 Yiming Yang。2022年。学会修复：部署后利用反馈动态记忆修复模型输出错误。载于《计算语言学协会论文集：NAACL 2022》，第339–352页，美国西雅图。计算语言学协会。

Andrea L. Thomaz 和 Cynthia Breazeal。2008年。可教的机器人：理解人类的教学行为，以构建更有效的机器人学习者。《人工智能》，172(6)：716–737。

Craig Thomson 和 Ehud Reiter。2021年。生成挑战：准确性评估共享任务的结果。载于第14届自然语言生成国际会议论文集，第240–248页，英国苏格兰阿伯丁。计算语言学协会。

Romal Thoppilan、Daniel De Freitas、Jamie Hall、Noam Shazeer、Apoorv Kulshreshtha、Heng-Tze Cheng、Alicia Jin、Taylor Bos、Leslie Baker、Yu Du 等。2022a。LaMDA：用于对话应用的语言模型。

Romal Thoppilan、Daniel De Freitas、Jamie Hall、Noam Shazeer、Apoorv Kulshreshtha、Heng-Tze Cheng、Alicia Jin、Taylor Bos、Leslie Baker、Yu Du、YaGuang Li、Hongrae Lee、Huaixiu Steven Zheng、Amin Ghafouri、Marcelo Menegali、Yanping Huang、Maxim Krikun、Dmitry Lepikhin、James Qin、Dehao Chen、Yuanzhong Xu、Zhifeng Chen、Adam Roberts、Maarten Bosma、Yanqi Zhou、Chung-Ching Chang、Igor Krivokon、Will Rusch、Marc Pickett、Kathleen S. Meier-Hellstern、Meredith Ringel Morris、Tulsee Doshi、Renelito Delos Santos、Toju Duke、Johnny Soraker、Ben Zevenbergen、Vinodkumar Prabhakaran、Mark Diaz、Ben Hutchinson、Kristen Olson、Alejandra Molina、Erin Hoffman-John、Josh Lee、Lora Aroyo、Ravi Rajakumar、Alena Butryna、Matthew Lamm、Viktoriya Kuzmina、Joe Fenton、Aaron Cohen、Rachel Bernstein、Ray Kurzweil、Blaise Aguera-Arcas、Claire Cui、Marian Croak、Ed H. Chi 和 Quoc Le。2022b。LaMDA：用于对话应用的语言模型。CoRR，abs/2201.08239。

Carlos Toxtli、Siddharth Suri 和 Saiph Savage。2021年。量化众包工作中不可见的劳动。《ACM 人机交互论文集》，5(CSCW2)：1–26。

Alexander Matt Turner、Aseem Saxena 和 Prasad Tadepalli。2022年。将副作用正则化问题形式化。发表于 NeurIPS 机器学习安全研讨会。

Peter Vamplew、Richard Dazeley、Cameron Foale、Sally Firmin 和 Jane Mummery。2018年。以人为本的人工智能是一种多目标问题。《伦理与信息技术》，20:27–40。

Christopher JCH Watkins 和 Peter Dayan。1992年。Q-learning。《机器学习》，8:279–292。Sean Welleck、Ximing Lu、Peter West、Faeze Brahman、Tianxiao Shen、Daniel Khashabi 和 Yejin Choi。2022年。通过学习自我纠错生成序列。arXiv 预印本 arXiv:2211.00053。Jason Weston、Sumit Chopra 和 Antoine Bordes。2014年。记忆网络。arXiv 预印本 arXiv:1410.3916。Jason E Weston。2016年。基于对话的语言学习。《神经信息处理系统进展》。Norbert Wiener。1948年。《控制论，或动物与机器中的控制和通信》。Ronald J. Williams。1992年。用于连接主义强化学习的简单统计梯度跟随算法。《机器学习》，8(3–4):229–256。Bin Wu、Chenyan Xiong、Maosong Sun 和 Zhiyuan Liu。2018年。利用反馈记忆网络进行查询建议。载于《2018年万维网大会论文集》，第1563–1571页。Jeff Wu、Long Ouyang、Daniel M. Ziegler、Nisan Stiennon、Ryan Lowe、Jan Leike 和 Paul Christiano。2021年。利用人类反馈递归总结书籍。Jing Xu、Megan Ung、Mojtaba Komeili、Kushal Arora、Y-Lan Boureau 和 Jason Weston。2022年。部署后学习新技能：利用人类反馈改进开放领域的互联网驱动对话。Taku Yamagata、Ryan McConville 和 Raul Santos-Rodriguez。2021年。利用来自多位技能各异的人类的反馈进行强化学习。Kevin Yang、Nanyun Peng、Yuandong Tian 和 Dan Klein。2022年。Re3：通过递归式重新提示与修订生成更长的故事。载于《自然语言处理经验方法会议》。

Michihiro Yasunaga 和 Percy Liang。2020年。基于图的、利用诊断反馈进行的自监督程序修复。第37届国际机器学习会议。ICML 2020, PartF168147-14:10730–10739。

Zheng Yuan、Hongyi Yuan、Chuanqi Tan、Wei Wang、Songfang Huang 和 Fei Huang。2023年。RRHF：无需费心即可通过人类反馈对齐语言模型的响应排序。arXiv 预印本 arXiv:2304.05302。

Chrysoula Zerva、Taisiya Glushkova、Ricardo Rei 和 André F. T. Martins。2022年。剖析机器翻译评估中的不确定性。

Tianyi Zhang、Varsha Kishore、Felix Wu、Kilian Q Weinberger 和 Yoav Artzi。2019年。BERTScore：使用 BERT 评估文本生成。arXiv 预印本 arXiv:1904.09675。

Chunting Zhou、Pengfei Liu、Puxin Xu、Srini Iyer、Jiao Sun、Yuning Mao、Xuezhe Ma、Avia Efrat、Ping Yu、Lili Yu、Susan Zhang、Gargi Ghosh、Mike Lewis、Luke Zettlemoyer 和 Omer Levy。2023a。LIMA：少即是多的对齐方法。

Shuyan Zhou、Uri Alon、Sumit Agarwal 和 Gra- ham Neubig。2023b。Codebertscore：使用代码预训练模型评估代码生成。arXiv预印本 arXiv:2302.05527。

Daniel M. Ziegler、Nisan Stiennon、Jeffrey Wu、Tom B. Brown、Alec Radford、Dario Amodei、Paul F. Christiano 和 Geoffrey Irving。2019年。根据人类偏好对语言模型进行微调。CoRR, abs/1909.08593。

Markus Zopf。2018年。利用成对偏好估算摘要质量。载于《2018年计算语言学协会北美分会会议论文集：人类语言技术》，第1卷（长篇论文），第1687–1696页，新奥尔良，路易斯安那州。计算语言学协会。