Training language models to follow instructions with human feedback

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*
Ryan Lowe*

OpenAI

Abstract

Making language models bigger does not inherently make them better at following a user's intent. For example, large language models can generate outputs that are untruthful, toxic, or simply not helpful to the user. In other words, these models are not aligned with their users. In this paper, we show an avenue for aligning language models with user intent on a wide range of tasks by fine-tuning with human feedback. Starting with a set of labeler-written prompts and prompts submitted through the OpenAI API, we collect a dataset of labeler demonstrations of the desired model behavior, which we use to fine-tune GPT-3 using supervised learning. We then collect a dataset of rankings of model outputs, which we use to further fine-tune this supervised model using reinforcement learning from human feedback. We call the resulting models InstructGPT. In human evaluations on our prompt distribution, outputs from the 1.3B parameter InstructGPT model are preferred to outputs from the 175B GPT-3, despite having 100x fewer parameters.Moreover, InstructGPT models show improvements in truthfulness and reductionsin toxic output generation while having minimal performance regressions on public NLP datasets. Even though InstructGPT still makes simple mistakes, our results show that fine-tuning with human feedback is a promising direction for aligning language models with human intent.

模型变得更大并不会更好地理解用户意图。比如，大模型会产生不可信的、有毒的、无用的输出给用户。换句话说，这些模型用户使用体验并不好。本文，我们在一系列任务上展示了一种通过人类反馈来使语言模型适应用户意图的方法。从一组标记者编写的提示和通过OpenAI API提交的提示开始，我们收集了一组标记者演示所需模型行为的数据集，然后使用监督学习来微调GPT-3。然后，我们收集了一组模型输出的排名数据集，使用来自人类反馈的强化学习进一步微调了这个监督模型。最后这个人类反馈的监督模型，我们称之为InstructGPT。在人在人类评估结果中，1.3B的InstructGPT比175B的GPT-3还要好，尽管参数量小了100倍。此外，InstructGPT在生成方面的可信度有所提升，在有毒输出方面有所下降，同时在公共数据集上有最小的性能回归。尽管InstructGPT依然会犯简单地错误，我们的结果表明，基于人类反馈的微调在使用模型符合人类意愿方面是一个有潜力的方向。

1 引言

Large language models (LMs) can be "prompted" to perform a range of natural language processing (NLP) tasks, given some examples of the task as input. However, these models often express unintended behaviors such as making up facts, generating biased or toxic text, or simply not following user instructions (Bender et al., 2021; Bommasani et al., 2021; Kenton

et al., 2021; Weidinger et al.,2021; Tamkin et al., 2021; Gehman et al., 2020). This is because the language modeling objective used for many recent large LMs—predicting the next token on a webpage from the internet—is different from the objective "follow the user's instructions helpfully and safely" (Radford et al., 2019;Brown et al., 2020; Fedus et al., 2021; Rae et al., 2021; Thoppilan et al., 2022). Thus, we say that the language modeling objective is misaligned. Averting these unintended behaviors is especially important for language models that are deployed and used in hundreds of applications.
大语言模型可以通过"prompt"的形式来完成一系列的NLP任务，即输入一些示例来产生输出。然而，这些模型经常输出一些不符合预期的结果，比如编造事实，产生有偏见或有害的文本，或者根本就不按用户的指示输出 (Bender et al., 2021; Bommasani et al., 2021; Kenton et al., 2021; Weidinger et al.,2021; Tamkin et al., 2021; Gehman et al., 2020)。原因在于语言模型的目标函数是在互联网的网页内容上预测下一个token，这不同于"准确安全地遵循用户指示"这个目标(Radford et al., 2019;Brown et al., 2020; Fedus et al., 2021; Rae et al., 2021; Thoppilan et al., 2022)。因此，我们可以说语言模型目标函数偏离了。对于已经部署和使用的几百个语言模型应用来说，避免这些行为非常重要。

图1：不同模型使用API prompt进行人类评估的结果，评估方法为各模型相对于175B SFT模型的输出偏好比率。我们的InstructGPT模型（PPO-ptx）以及它的变体（未使用预训练融合的PPO）都远远超过了GPT-3基线（GPT,GPT prompted）；1.3B的PPO-ptx模型的结果超过了175B的GPT-3。本文中的误差线为95%置信区间。
We make progress on aligning language models by training them to act in accordance with the user's intention (Leike et al., 2018). This encompasses both explicit intentions such as following instructions and implicit intentions such as staying truthful, and not being biased, toxic, or otherwise harmful.Using the language of Askell et al. (2021), we want language models to be helpful (they should help the user solve their task), honest (they shouldn't fabricate information or mislead the user), and harmless (they should not cause physical, psychological, or social harm to people or the environment).We elaborate on the evaluation of these criteria in Section 3.6.
我们通过训练语言模型按照用户的意图行事（Leike et al., 2018）来取得进展。这包括明确的意图，如遵循指示，以及隐含的意图，如保持真实，不偏见，不有毒或不会对用户造成其他伤害。用Askell等人（2021）的语言来说，我们希望语言模型是有益的（它们应该帮助用户解决问题），诚实的（它们不应该捏造信息或误导用户）和无害的（它们不应该对人或环境造成身体、心理或社会上的伤害）。我们在第3.6节详细阐述了这些标准的评估。
We focus on fine-tuning approaches to aligning language models. Specifically, we use reinforcement learning from human feedback (RLHF; Christiano et al., 2017; Stiennon et al., 2020) to fine-tune GPT-3 to follow a broad class of written instructions (see Figure 2). This technique uses human preferences as a reward signal to fine-tune our models. We first hire a

team of 40 contractors to label our data, based on their performance on a screening test (see Section 3.4 and Appendix B.1 for moredetails). We then collect a dataset of human-written demonstrations of the desired output behavior on (mostly English) prompts submitted to the OpenAI API and some labeler-written prompts, and use this to train our supervised learning baselines. Next, we collect a dataset of human-labeled comparisons between outputs from our models on a larger set of API prompts. We then train a reward model (RM) on this dataset to predict which model output our labelers would prefer. Finally, we use this RM as a reward function and fine-tune our supervised learning baseline to maximize this reward using the PPO algorithm (Schulman et al., 2017). We illustrate this process in Figure 2. This procedure aligns the behavior of GPT-3 to the stated preferences of a specific group of people (mostly our labelers and researchers), rather than any broader notion of “human values”; we discuss this further in Section 5.2. We call the resulting models InstructGPT.
我们重点关注使用微调的方法来优化模型。具体而言，我们使用来自人类反馈的强化学习 (RLHF; Christiano et al., 2017; Stiennon et al., 2020)来微调GPT-3，以遵循广泛的指令（见图2）。这项技术使用人类偏好作为奖励信号来微调我们的模型。我们首先雇佣了40个供应商来打标我们的数据，这些承包商是基于他们在筛选测试中的表现而被选中的（请参见第3.4节和附录B.1以获取更多详细信息）。然后我们收集了一个数据集（大多数是英语），输入是prompts，输出是人工编写的输出，使用这个数据集我们训练了一个监督学习基线模型。接下来，我们收集了另外一个数据集，输入是prompts，输出是GPT-3输出的若干个结果，基于这个数据我们训练了一个奖励模型（RM）,奖励模型就是预测哪个输出是人类更偏好的一种模型。最后，我们使用此RM作为奖励函数，并使用PPO算法(Schulman et al., 2017)微调我们的学习基线模型，以最大化此奖励。我们在图2说明了这个过程。这个过程将GPT-3的行为与特定人群的偏好保持了一致（主要是我们的打标人员和研究人员），而不是广泛的“人类价值观”概念。我们在5.2节中进一步讨论这些。我们称这个最终的模型为InstructGPT。

图2 InstructGPT方案三个步骤：（1）监督微调模型（SFT），（2）奖励模型（RM）训练，（3）在奖励模型基础上使用最近策略优化算法（PPO）进行强化学习。蓝色箭头意味着使用该数据训练模型。在第2步，A到D是我们模型的输出中抽样的，然后标记者对其排序。我们方法的更多细节见第三小节。

We mainly evaluate our models by having our labelers rate the quality of model outputs on our testset, consisting of prompts from held-out customers (who are not represented in the training data).We also conduct automatic evaluations on a range of public NLP datasets. We train three model sizes (1.3B, 6B, and 175B parameters), and all of our models use the GPT-3 architecture. Our mainfindings are as follows:
我们主要通过让标注员对来自我们测试集的提示的模型输出质量进行评分来评估我们的模型，这些提示来自于未在训练数据中表示的客户。我们还在一系列公共 NLP 数据集上进行自动评估。我们训练了

三种模型大小（1.3B、6B 和 175B 参数），我们所有的模型都使用 GPT-3 架构。我们的主要发现如下：

Labelers significantly prefer InstructGPT outputs over outputs from GPT-3. On our test set,outputs from the 1.3B parameter InstructGPT model are preferred to outputs from the 175B GPT-3,despite having over 100x fewer parameters. These models have the same architecture, and differ only by the fact that InstructGPT is fine-tuned on our human data. This result holds true even when we add a few-shot prompt to GPT-3 to make it better at following instructions. Outputs from our 175B InstructGPT are preferred to 175B GPT-3 outputs 85 ± 3% of the time, and preferred 71 ± 4% of the time to few-shot 175B GPT-3. InstructGPT models also generate more appropriate outputs according to our labelers, and more reliably follow explicit constraints in the instruction.

相比于GPT-3，标注者更偏向InstructGPT。在我们的测试数据上，1.3B的InstructGPT的输出比175B的GPT-3的输出更受标者偏好，尽管参数小了100倍。这些模型有相同的架构，区别仅在于InstructGPT是在人类数据上微调过。即使我们将GPT-3加了few-shot prompt来提升其适应指示的能力，这个结果依然不变。InstructGPT的结果优于175B的GPT-3 85 ± 3%，优于few-shot 175B的GPT-3 71 ± 4%。InstructGPT的输出还更合适，更可靠地遵循指令的约束。

InstructGPT models show improvements in truthfulness over GPT-3. On the TruthfulQA benchmark, InstructGPT generates truthful and informative answers about twice as often as GPT-3.Our results are equally strong on the subset of questions that were not adversarially selected against GPT-3. On “closed-domain” tasks from our API prompt distribution, where the output should not contain information that is not present in the input (e.g. summarization and closed-domain QA),InstructGPT models make up information not present in the input about half as often as GPT-3 (a21% vs. 41% hallucination rate, respectively).

InstructGPT相比GPT-3在真实性上有所提升。在TruthfulQA基准测试中，InstructGPT产生真实且有信息量的答案的频率是GPT-3的两倍。我们的结果在没有针对GPT-3进行敌对选择的问题子集上同样强大。在“封闭域”任务上，输出不应该包括输入以外的信息（比如摘要生成和闭域QA），InstructGPT生成的输入外信息频率是GPT-3的一半（分别是21%和41%）。

InstructGPT shows small improvements in toxicity over GPT-3, but not bias. To measure toxicity, we use the RealToxicityPrompts dataset (Gehman et al., 2020) and conduct both automatic and human evaluations. InstructGPT models generate about 25% fewer toxic outputs than GPT-3 when prompted to be respectful. InstructGPT does not significantly improve over GPT-3 on the Winogender (Rudinger et al., 2018) and CrowSPairs (Nangia et al., 2020) datasets.

InstructGPT在降低毒性方面比GPT-3有所改善，但是在降低偏见方面没有改善。为了衡量毒性，我们使用了RealToxicityPrompts数据集(Gehman et al., 2020)，且使用了自动和人工评估。当被要求保持尊重时，InstructGPT比GPT-3少了25%的有毒输出。在Winogender (Rudinger et al., 2018) and CrowSPairs (Nangia et al., 2020)数据集上，InstructGPT相比于GPT-3没有显著改善。

We can minimize performance regressions on public NLP datasets by modifying our RLHF fine-tuning procedure. During RLHF fine-tuning, we observe performance regressions compared to GPT-3 on certain public NLP datasets, notably SQuAD (Rajpurkar et al., 2018), DROP (Dua et al.,2019), HellaSwag (Zellers et al., 2019), and WMT 2015 French to English translation (Bojar et al.,2015). This is an example of an "alignment tax" since our alignment procedure comes at the cost of lower performance on certain tasks that we may care about. We can greatly reduce the performance regressions on these datasets by mixing PPO updates with updates that increase the log likelihood of the pretraining distribution (PPO-ptx), without compromising labeler preference scores.

通过改进RLHF的微调过程，我们可以最大程度减少在公开NLP数据集上的性能回归。在RLHF微调时，我们发现在一些公开NLP数据集上有性能回归现象，特别是SQuAD (Rajpurkar et al., 2018), DROP (Dua et al.,2019), HellaSwag (Zellers et al., 2019), and WMT 2015 French to English translation (Bojar et al.,2015)。这是"对齐税"的一个例子，因为在对齐过程中伴随着一些任务的性能下降。通过PPO更新以及提升预训练分布的对数似然函数值（PPO-ptx）来降低性能回归，同时又不会降低偏好分。

Our models generalize to the preferences of "held-out" labelers that did not produce any training data. To test the generalization of our models, we conduct a preliminary experiment with held-out labelers, and find that they prefer InstructGPT outputs to outputs from GPT-3 at about the same rate as our training labelers. However, more work is needed to study how these models perform on broader groups of users, and how they perform on inputs where humans disagree about the desired behavior.

我们的模型也适用于没有参与生产训练数据的标注者的偏好。为了测试我们模型的泛化能力，我们进行了一个未参与标注者的初步实验，发现他们与训练数据标注者一样，更偏好InstructGPT。然而，更多地工作需要研究，一是这些模型如何适应更广泛的用户，二是输出不符合人类偏好的情况如何改善。

Public NLP datasets are not reflective of how our language models are used. We compare GPT-3 fine-tuned on our human preference data (i.e. InstructGPT) to GPT-3 fine-tuned on two different compilations of public NLP tasks: the FLAN (Wei et al., 2021) and T0 (Sanh et al., 2021)(in particular, the T0++ variant). These datasets consist of a variety of NLP tasks, combined with natural language instructions for each task. On our API prompt distribution, our FLAN and T0 models perform slightly worse than our SFT baseline, and labelers significantly prefer InstructGPT to these models (InstructGPT has a 73.4 ±2% winrate vs. our baseline, compared to 26.8 ±2% and 29.8 ±2% for our version of T0 and FLAN, respectively).

公共NLP数据集不能反映我们的语言模型的真实情况。我们将在人类偏好数据（即InstructGPT）上微调的GPT-3与在两个不同的公共NLP任务编译上微调的GPT-3进行比
较： (Wei et al., 2021) and T0 (Sanh et al., 2021)(in particular, the T0++ variant)。这些数据集包括各种NLP任务，以及每个任务的自然语言说明。在我们的API提示分发中，我们的FLAN和T0模型的

表现略逊于我们的SFT基线，并且标注者明显更喜欢InstructGPT而不是这些模型（与我们的基线相比，InstructGPT的胜率为73.4±2%，而我们的T0和FLAN版本分别为26.8±2%和29.8±2%）。

InstructGPT models show promising generalization to instructions outside of the RLHF fine-tuning distribution. We qualitatively probe InstructGPT’s capabilities, and find that it is able to follow instructions for summarizing code, answer questions about code, and sometimes follows instructions in different languages, despite these instructions being very rare in the fine-tuning distribution. In contrast, GPT-3 can perform these tasks but requires more careful prompting, and does not usually follow instructions in these domains. This result is exciting because it suggests that our models are able to generalize the notion of “following instructions.” They retain some alignment even on tasks for which they get very little direct supervision signal.

InstructGPT模型在RLHF微调之外的分布上展示出了有潜力的泛化性。我们大量地探索 了InstructGPT的能力，发现它可以总结代码，回答关于代码的问题上遵循指令，有时还能在不同语言上遵循指令，尽管这些指令在微调分布上很罕见。相反，GPT-3在实现这些任务时需要更多的prompt，而且在这些领域也不能很好地遵循指令。这个结果很令人兴奋，因为它表明了我们模型在“遵循指令”这件事情上有能力泛化了。他们在一些很少监督信号的任务也能保持一定的对齐。

InstructGPT still makes simple mistakes. For example, InstructGPT can still fail to follow instructions, make up facts, give long hedging answers to simple questions, or fail to detect instructions with false premises.

InstructGPT仍然会犯简单地错误。例如，InstructGPT仍然不能遵循指令，具体有捏造事实，对简单地问题给出给出冗长且含糊的答案，不能发现虚假的提示。

Overall, our results indicate that fine-tuning large language models using human preferences significantly improves their behavior on a wide range of tasks, though much work remains to be done to improve their safety and reliability.

总之，我们的结果表明，大语言模型使用人类偏好进行微调之后，可以大幅提升它在一系列任务的表现，尽管很多工作仍然有待去，比如安全性和可靠性。

The rest of this paper is structured as follows: We first detail related work in Section 2, before diving into our method and experiment details in Section 3, including our high-level methodology (3.1), task and dataset details (3.3 and 3.2), human data collection (3.4), how we trained our models (3.5), and our evaluation procedure (3.6). We then present our results in Section 4, divided into three parts:results on the API prompt distribution (4.1), results on public NLP datasets (4.2), and qualitative results (4.3). Finally we give an extended discussion of our work in Section 5, including implications for alignment research (5.1), what we are aligning to (5.2), limitations (5.3), open questions (5.4),and broader impacts of this work (5.5).

本文的结构如下：我们首先在第2节列出了相关工作，在第3节说明了方法和实验细节，3.1介绍了高级方法，3.2和3.3介绍了任务和数据细节，3.4介绍了数据收集，3.5介绍了模型训练方法，3.6介绍了评估过程。我们在第4节展示了结果，分为三个部分，4.1是API prompt分布，4.2是NLP公共数据集，

4.3是定性结果。最后在第5节介绍了延伸的讨论，5.1是对对齐研究的影响，5.2是我们在对齐什么，5.3是限制，5.4是开放的问题，5.5是这项工作的广泛影响。

2 相关工作

Research on alignment and learning from human feedback. We build on previous techniques to align models with human intentions, particularly reinforcement learning from human feed-back (RLHF). Originally developed for training simple robots in simulated environments and Atarigames (Christiano et al., 2017; Ibarz et al., 2018), it has recently been applied to fine-tuning language models to summarize text (Ziegler et al., 2019; Stiennon et al., 2020; Böhm et al., 2019; Wu et al.,2021). This work is in turn influenced by similar work using human feedback as a reward in domains such as dialogue (Jaques et al., 2019; Yi et al., 2019; Hancock et al., 2019), translation (Kreutzer et al.,2018; Bahdanau et al., 2016), semantic parsing (Lawrence and Riezler, 2018), story generation (Zhouand Xu, 2020), review generation (Cho et al., 2018), and evidence extraction (Perez et al., 2019).Madaan et al. (2022) use written human feedback to augment prompts and improve the performanceof GPT-3. There has also been work on aligning agents in text-based environments using RL with a normative prior (Nahian et al., 2021). Our work can be seen as a direct application of RLHF to aligning language models on a broad distribution of language tasks.

在对齐和RLHF方面的研究工作。我们使用已有的一些技术将模型对齐人类的意图，尤其是RLHF。RLHF最早是为了训练简单机器和 Atarigames游戏而开发的 (Christiano et al., 2017; Ibarz et al., 2018)，近些年被用来微调语言模型以总结文本 (Ziegler et al., 2019; Stiennon et al., 2020; Böhm et al., 2019; Wu et al.,2021)。这个工作也受了相似的将人类反馈作为奖励信号的工作影响，比如对话 (Jaques et al., 2019; Yi et al., 2019; Hancock et al., 2019)，翻译（Kreutzer et al.,2018; Bahdanau et al., 2016），语义解析（Lawrence and Riezler, 2018），故事生成（Cho et al., 2018），证据抽取（Perez et al., 2019）。Madaan et al. (2022) 使用了人类反馈来增强prompts，以此来提升GPT-3的水平。也有一些工作在文本环境下使用规范先验的RL来对齐agents（Nahian et al., 2021）。我们的工作是在一系列的语言任务上直接使用RLHF来对齐语言模型。

The question of what it means for language models to be aligned has also received attention recently (Gabriel, 2020). Kenton et al. (2021) catalog behavioral issues in LMs that result from misalignment, including producing harmful content and gaming misspecified objectives. In concurrent work, Askell et al. (2021) propose language assistants as a testbed for alignment research, study some simple baselines, and their scaling properties.

最近对于语言模型对齐的问题也受到了关注。Kenton等人（2021）列举了由于对齐不当而导致的LM的行为问题，包括生成有害内容和游戏误指定的目标。同时，Askell等人（2021）提出了语言助手作为对齐研究的测试平台，研究了一些简单的基线和它们的扩展性。

Evaluating the harms of language models. A goal of modifying the behavior of language models is to mitigate the harms of these models when they’ re deployed in the real world. These risks have been extensively documented (Bender et al., 2021; Bommasani et al., 2021;

Kenton et al., 2021;Weidinger et al., 2021; Tamkin et al., 2021). Language models can produce biased outputs (Dhamalaet al., 2021; Liang et al., 2021; Manela et al., 2021; Caliskan et al., 2017; Kirk et al., 2021), leak private data (Carlini et al., 2021), generate misinformation (Solaiman et al., 2019; Buchanan et al.,2021), and be used maliciously; for a thorough review we direct the reader to Weidinger et al. (2021).Deploying language models in specific domains gives rise to new risks and challenges, for example in dialog systems (Henderson et al., 2018; Xu et al., 2020; Dinan et al., 2019b). There is a nascent but growing field that aims to build benchmarks to concretely evaluate these harms, particularly around toxicity (Gehman et al., 2020), stereotypes (Nadeem et al., 2020), and social bias (Dhamala et al.,2021; Nangia et al., 2020; Rudinger et al., 2018). Making significant progress on these problems is hard since well-intentioned interventions on LM behavior can have side-effects (Welbl et al., 2021;Blodgett et al., 2020); for instance, efforts to reduce the toxicity of LMs can reduce their ability to model text from under-represented groups, due to prejudicial correlations in the training data (Xuet al., 2021).

评估语言模型的危害。修改语言模型行为的目标是在这些模型在现实世界中部署时减轻这些模型的危害。这些风险已经得到了广泛的记录
(Bender et al., 2021; Bommasani et al., 2021; Kenton et al., 2021;Weidinger et al., 2021; Tamkin et al., 2021)。语言模型可以产生有偏见的输出 (Dhamalaet al., 2021; Liang et al., 2021; Manela et al., 2021; Caliskan et al., 2017; Kirk et al., 2021)，泄露私人数据(Carlini et al., 2021)，生成错误信息(Solaiman et al., 2019; Buchanan et al.,2021)，并被恶意使用。将语言模型部署在特定领域会产生新的风险和挑战，例如在对话系统中(Henderson et al., 2018; Xu et al., 2020; Dinan et al., 2019b)。有一个新兴但不断发展的领域旨在建立基准来具体评估这些危害，特别是在毒性(Gehman et al., 2020)、刻板印象(Nadeem et al., 2020)和社会偏见(Dhamala et al.,2021; Nangia et al., 2020; Rudinger et al., 2018)方面。在这些问题上取得重大进展很难，因为对LM行为的善意干预可能会产生副作用。例如，减少LM毒性的努力可能会降低它们对来自少数群体的文本进行建模的能力，因为训练数据中存在有偏见的相关性。

Modifying the behavior of language models to mitigate harms. There are many ways to change the generation behavior of language models. Solaiman and Dennison (2021) fine-tune LMs on a small, value-targeted dataset, which improves the models' ability to adhere to these values on a question answering task. Ngo et al. (2021) filter the pretraining dataset by removing documents on which a language model has a high conditional likelihood of generating a set of researcher-written trigger phrases. When trained on this filtered dataset, their LMs generate less harmful text, at the cost of a slight decrease in language modeling performance. Xu et al. (2020) use a variety of approaches to improve the safety of chatbots, including data filtering, blocking certain words or n-grams during generation, safety-specific control tokens (Keskar et al., 2019; Dinan et al., 2019a), and human-in-the-loop data collection (Dinan et al., 2019b). Other approaches for mitigating the generated bias by LMs

use word embedding regularization (Liu et al., 2019; Huang et al., 2019), data augmentation (Liuet al., 2019; Dinan et al., 2019a; Sheng et al., 2019), null space projection to make the distribution over sensitive tokens more uniform (Liang et al., 2021), different objective functions (Qian et al.,2019), or causal mediation analysis (Vig et al., 2020). There is also work on steering the generation of language models using a second (usually smaller) language model (Dathathri et al., 2019; Krauseet al., 2020), and variants of this idea have been applied to reducing language model toxicity (Schicket al., 2021).

修改语言模型的行为以减少伤害。有许多方法可以改变语言模型的生成行为。Solaiman和Dennison（2021）在小型、针对性价值的数据集上微调LM，这提高了模型在问答任务中遵循这些价值的能力。Ngo等人（2021）通过删除语言模型具有高条件生成研究人员编写的触发短语集的文档来过滤预训练数据集。当在此过滤后的数据集上训练时，他们的LM生成的有害文本较少，但语言建模性能略有下降。Xu等人（2020）使用各种方法来提高聊天机器人的安全性，包括数据过滤、在生成过程中阻止某些单词或n-gram个单词、特定安全的控制token（Keskar等人，2019; Dinan等人，2019a）和人机协同数据收集（Dinan等人，2019b）。减轻LM生成偏见的其他方法有单词嵌入正则化（Liu等人，2019; Huang等人，2019）、数据增强（Liuet al.，2019; Dinan等人，2019a; Sheng等人，2019）、零空间投影使敏感token的分布更加均匀（Liang等人，2021）、不同的目标函数（Qian等人，2019）或因果分析（Vig等人，2020）。还有一些工作是使用第二个（通常较小的）语言模型（Dathathri等人，2019; Krauseet al.，2020），这个想法已经应用于减少语言模型的毒性（Schicket al.，2021）。

3 方案和实验细节

3.1 高层方法论

Our methodology follows that of Ziegler et al. (2019) and Stiennon et al. (2020), who applied it in the stylistic continuation and summarization domains. We start with a pretrained language model (Radford et al., 2019; Brown et al., 2020; Fedus et al., 2021; Rae et al., 2021; Thoppilan et al.,2022), a distribution of prompts on which we want our model to produce aligned outputs, and a team of trained human labelers (see Sections 3.4 for details). We then apply the following three steps(Figure 2).

我们的方法是借鉴了Ziegler et al. (2019) and Stiennon et al. (2020)，他们将方法用在了风格续写和摘要抽取。我们一开始是准备一个预训练模型 (Radford et al., 2019; Brown et al., 2020; Fedus et al., 2021; Rae et al., 2021; Thoppilan et al.,2022)，一组prompts，一组经过训练的人类标注者(see Sections 3.4 for details)。然后，我们用了以下三步(Figure 2)。

Step 1: Collect demonstration data, and train a supervised policy. Our labelers provide demonstrations of the desired behavior on the input prompt distribution (see Section 3.2 for details on this distribution). We then fine-tune a pretrained GPT-3 model on this data using supervised learning.

步骤1：收集示例数据，训练一个监督策略。我们的标记者提供了输入prompt的预期行为输出（see Section 3.2 for details on this distribution）。然后，我们基于GPT-3在此数据上使用监督学习进行

微调。

Step 2: Collect comparison data, and train a reward model. We collect a dataset of comparisons between model outputs, where labelers indicate which output they prefer for a given input. We then train a reward model to predict the human-preferred output.

步骤2：收集对比数据，训练奖励模型。我们收集了一个模型输出对比数据集，给定一个输入，打标者决定偏好哪个输出。然后基于此数据集训练一个奖励模型，奖励模型可以预测人类偏好的输出。

Step 3: Optimize a policy against the reward model using PPO. We use the output of the RM as a scalar reward. We fine-tune the supervised policy to optimize this reward using the PPO algorithm (Schulman et al., 2017).

使用PPO算法优化奖励模型。我们使用RM的输出作为标量奖励。然后，使用PPO算法来微调监督策略(Schulman et al., 2017)。

Steps 2 and 3 can be iterated continuously; more comparison data is collected on the current best policy, which is used to train a new RM and then a new policy. In practice, most of our comparison data comes from our supervised policies, with some coming from our PPO policies.

步骤2和步骤3可以循环迭代。基于当前最好的策略收集比较数据，然后训练新的RM，然后产生新的策略。实际上，比较数据有很多是来自监督策略，有些来自PPO策略。

3.2 数据集

Our prompt dataset consists primarily of text prompts submitted to the OpenAI API, specifically those using an earlier version of the InstructGPT models (trained via supervised learning on a subset of our demonstration data) on the Playground interface. Customers using the Playground were informed that their data could be used to train further models via a recurring notification any time InstructGPT models were used. In this paper we do not use data from customers using the API in production. We heuristically deduplicate prompts by checking for prompts that share a long common prefix, and we limit the number of prompts to 200 per user ID. We also create our train, validation,and test splits based on user ID, so that the validation and test sets contain no data from users whose data is in the training set. To avoid the models learning potentially sensitive customer details, we filter all prompts in the training split for personally identifiable information (PII).

我们的prompt数据集主要由提交至OpenAI API的prompts组成，尤其是那些在playground使用早期版本的InstructGPT（由我们的示例数据子集微调的监督学习模型）提交的prompts。用户在playground上使用InstructGPT时，会被告知，他们的数据会被用来进一步训练模型。在本文中，我们不会使用生产上的客户API数据。我们对有长公共前缀的prompts进行了去重，然后将每个用户的prompts限制为200个。我们使用userid来划分train、valid、test数据集，也就说一个用户不会出现在不同的数据集里面。为了避免学习到客户敏感信息，我们将敏感信息在训练集中去掉了。

To train the very first InstructGPT models, we asked labelers to write prompts themselves. This is because we needed an initial source of instruction-like prompts to bootstrap the

process, and these kinds of prompts weren’ t often submitted to the regular GPT-3 models on the API. We asked labelers to write three kinds of prompts:
• Plain: We simply ask the labelers to come up with an arbitrary task, while ensuring the tasks had sufficient diversity.
• Few-shot: We ask the labelers to come up with an instruction, and multiple query/response pairs for that instruction.
• User-based: We had a number of use-cases stated in waitlist applications to the OpenAI API. We asked labelers to come up with prompts corresponding to these use cases.
为了训练最早期的InstructGPT模型，我们让标注者自己写prompts。这是因为我们需要一个类似指令的prompts来起动，这类prompts通常不会通过API提交到常规的GPT-3模型。我们让标注者写了三种类型的prompts。
普通版：我们只是让标注员提出任意任务，同时保证任务具有多样性。
Few-shot版：我们让标注者提出一个指令，和基于指令的一些查询/响应对。
用户版：我们在OpenAI API的等待列表申请中列出了一些用例，我们让标注者针对这些用例提出一些相应的prompts。
From these prompts, we produce three different datasets used in our fine-tuning procedure: (1) our SFT dataset, with labeler demonstrations used to train our SFT models, (2) our RM dataset, with labeler rankings of model outputs used to train our RMs, and (3) our PPO dataset, without any human labels, which are used as inputs for RLHF fine-tuning. The SFT dataset contains about 13k training prompts (from the API and labeler-written), the RM dataset has 33k training prompts (from the API and labeler-written), and the PPO dataset has 31k training prompts (only from the API). More details on dataset sizes are provided in Table 6.
通过这些prompts，我们生成了三种不同的数据集用于我们的微调过程：（1）SFT数据集，即标注员标注的示例，用于训练SFT模型。（2）RM数据集，标注模型输出的排名，用以训练我们的RMs。（3）PPO数据集，没有人工标注，用以RLHF微调。SFT数据集包含1.3万个训练prompts（源自API和人工标注），RM数据集包含3.3万个训练prompts（源自API和人工标注），PPO数据集有3.1万个训练prompts（仅源自API）。更多数据细节见表6。
To give a sense of the composition of our dataset, in Table 1 we show the distribution of use-case categories for our API prompts (specifically the RM dataset) as labeled by our contractors. Most of the use-cases have are generative, rather than classification or QA. We also show some illustrative prompts (written by researchers to mimic the kinds of prompts submitted to InstructGPT models) in Table 2; more prompts submitted to InstructGPT models are shown in Appendix A.2.1, and prompts submitted to GPT-3 models are shown in Appendix A.2.2. We provide more details about our dataset in Appendix A.
为了让您了解我们数据集的组成，我们在表1中展示了由承包商标记的API prompts（特别是RM数据集）的用例类别分布。大多数用例都是生成性的，而不是分类或问答。我们还在表2中展示了一些说明

性promts（由研究人员编写），在附录A.2.1中显示了更多提交给InstructGPT模型的promts，而在附录A.2.2中显示了提交给GPT-3模型的提示。我们在附录A中提供了有关数据集的更多详细信息。

表1：源自API的prompt数据集用例分布

表2：说明性prompts示例，这些是虚构的，用以指导标注者编写prompts。

3.3 任务

Our training tasks are from two sources: (1) a dataset of prompts written by our labelers and (2) a dataset of prompts submitted to early InstructGPT models on our API (see Table 6). These prompts are very diverse and include generation, question answering, dialog, summarization, extractions, and other natural language tasks (see Table 1). Our dataset is over 96% English, however in Section 4.3 we also probe our model’ s ability to respond to instructions in other languages and complete coding tasks.

我们的训练任务有两个数据来源：（1）标注者写的prompts数据集，（2）提交至早期Instruct模型API的prompts数据集（见表6）。这些prompts非常多样化，包括生成、问答、对话、摘要、抽取及其它自然语言任务（见表1）。我们的数据有超过96%是英语，然而在4.3节我们也探索了模型在其它语言方面响应指令的能力，以及完成编码任务。

For each natural language prompt, the task is most often specified directly through a natural language instruction (e.g. “Write a story about a wise frog”), but could also be indirectly through either few-shot examples (e.g. giving two examples of frog stories, and prompting the model to generate a new one) or implicit continuation (e.g. providing the start of a story about a frog). In each case, we ask our labelers to do their best to infer the intent of the user who wrote the prompt, and ask them to skip inputs where the task is very unclear. Moreover, our labelers also take into account the implicit intentions such as truthfulness of the response, and potentially harmful outputs such as biased or toxiclanguage, guided by the instructions we provide them (see Appendix B) and their best judgment.

每个自然语言提示通常直接通过自然语言指令（例如，“写一篇关于聪明青蛙的故事”）来指定任务，但也可以通过少量示例（例如，提供两个青蛙故事的示例，并提示模型生成一个新的故事）或隐式延续（例如，提供有关青蛙的故事的开头）间接指定。在每种情况下，我们要求我们的标注员尽力推断编写提示的用户的意图，并要求他们跳过任务非常不清晰的输入。此外，我们的标注员还考虑到隐含意图，例如响应的真实性以及潜在的有害输出，例如有偏见或有毒语言，由我们提供的指令（请参见附录B）和他们的最佳判断来指导。

3.4 人类数据收集

To produce our demonstration and comparison data, and to conduct our main evaluations, we hired a team of about 40 contractors on Upwork and through ScaleAI. Compared to earlier work that collects human preference data on the task of summarization (Ziegler et al., 2019; Stiennon et al.,2020; Wu et al., 2021), our inputs span a much broader range of tasks, and can occasionally include controversial and sensitive topics. Our aim was to select a group

of labelers who were sensitive to the preferences of different demographic groups, and who were good at identifying outputs that were potentially harmful. Thus, we conducted a screening test designed to measure labeler performance on these axes. We selected labelers who performed well on this test; for more information about our selection procedure and labeler demographics, see Appendix B.1.
为了制作示例和对比数据，并进行评估工作，我们在Upwork和ScaleAI上雇佣了40个承包商。与早期收集人类的摘要偏好数据不同(Ziegler et al., 2019; Stiennon et al.,2020; Wu et al., 2021)，我们的输入包括的任务范围更广，并且有时会包括有争议和敏感的话题。我们的目标是选择一组对不同人群的偏好敏感，并且擅长识别潜在有害输出。因此，我们进行了一项筛选测试，用以评估标注者在这方面的表现。我们会选择在测试中表现良好的标注者。更多关于筛选过程见附录B.1。
During training and evaluation, our alignment criteria may come into conflict: for example, when a user requests a potentially harmful response. During training we prioritize helpfulness to the user (not doing so requires making some difficult design decisions that we leave to future work; see Section 5.4 for more discussion). However, in our final evaluations we asked labelers prioritize truthfulness andharmlessness (since this is what we really care about).
在训练和评估过程中，我们的对齐标准可能会发生冲突：例如，当用户请求潜在有害的响应时。在训练期间，我们优先考虑对用户的帮助（不这样做需要做出一些困难的设计决策，我们将其留给未来的工作；请参见第5.4节以获取更多讨论）。然而，在我们的最终评估中，我们要求标注者优先考虑真实性和无害性（因为这是我们真正关心的）。
As in Stiennon et al. (2020), we collaborate closely with labelers over the course of the project. We have an onboarding process to train labelers on the project, write detailed instructions for each task(see Appendix B.2), and answer labeler questions in a shared chat room.
与Stiennon等人（2020）一样，我们在项目过程中与标注者密切合作。我们有一个入职流程，以培训标注者参与项目，为每个任务编写详细的说明（请参见附录B.2），并在共享聊天室中回答标注者的问题。
As an initial study to see how well our model generalizes to the preferences of other labelers, we hire a separate set of labelers who do not produce any of the training data. These labelers are sourced from the same vendors, but do not undergo a screening test.
为了初步研究我们模型的泛化能力，我们雇佣了一组不生产训练数据的标注者。这些标注者来自同一个供应商，但是没有经过筛选测试。
Despite the complexity of the task, we find that inter-annotator agreement rates are quite high:training labelers agree with each-other 72.6 ± 1.5% of the time, while for held-out labelers this number is 77.3 ± 1.3%. For comparison, in the summarization work of Stiennon et al. (2020) researcher-researcher agreement was 73 ± 4%.
尽管任务很复杂，我们发现标注者之间的一致性还是很高的：训练标注者之间的一致性为72.6 ± 1.5%，验证标注者之间的一致性为77.3 ± 1.3%。作为对比，Stiennon et al. (2020)的摘要工作中，

研究人员的一致性为73 ± 4%。

3.5 算法

We start with the GPT-3 pretrained language models from Brown et al. (2020). These models are trained on a broad distribution of Internet data and are adaptable to a wide range of downstream tasks,but have poorly characterized behavior. Starting from these models, we then train models with three different techniques:

我们从Brown et al. (2020)的GPT-3预训练模型开始。这些模型是在广泛的互联网数据上训练的，可以适应很多的下游任务，但是行为特征不明确。从这些模型开始，我们使用三种技术来训练模型。

Supervised fine-tuning (SFT). We fine-tune GPT-3 on our labeler demonstrations using supervised learning. We trained for 16 epochs, using a cosine learning rate decay, and residual dropout of 0.2.We do our final SFT model selection based on the RM score on the validation set. Similarly to Wu et al. (2021), we find that our SFT models overfit on validation loss after 1 epoch; however, we find that training for more epochs helps both the RM score and human preference ratings, despite this overfitting.

监督微调（SFT）。我们在标注的示例上使用监督学习微调GPT-3。我们训练了16轮，使用了余弦学习率延迟，0.2的残差丢弃。我们在验证集上使用RM数来筛选最优的SFT模型。与Wu et al. (2021)的工作类似，我们发现SFT模型在验证集上仅一轮就过拟和了，然而，我们发现多训练几轮可以帮助提升RM分数和人类偏好评分，尽管会存在过拟和。

Reward modeling (RM). Starting from the SFT model with the final unembedding layer removed,we trained a model to take in a prompt and response, and output a scalar reward. In this paper we only use 6B RMs, as this saves a lot of compute, and we found that 175B RM training could be unstable and thus was less suitable to be used as the value function during RL (see Appendix C for more details)**.**

奖励模型（RM）。从去掉反embedding层的SFT模型开始，我们训练了一个模型，将prompt和response带入，然后输出一个标量的奖励。在本文，我们仅使用了6B的RMs，这节约了很多计算资源，而且我们发现175B的RM训练时不太稳定，因此不太适合做RL的值函数（更多细节见附录C）。

In Stiennon et al. (2020), the RM is trained on a dataset of comparisons between two model outputs on the same input. They use a cross-entropy loss, with the comparisons as labels—the difference in rewards represents the log odds that one response will be preferred to the other by a human labeler.

在 Stiennon 等人（2020）的论文中，RM 是在一个比较数据集上训练的，该数据集包含了两个模型在同一个输入上的输出的比较。他们使用交叉熵损失，以比较结果作为标签——奖励的差值表示了一个回答被人类标注者优先选择的对数几率。

In order to speed up comparison collection, we present labelers with anywhere between K = 4 and K = 9 responses to rank. This produces (K 2) comparisons for each prompt shown to a labeler. Since comparisons are very correlated within each labeling task, we found that if we simply shuffle the comparisons into one dataset, a single pass over the dataset caused the

reward model to overfit. Instead, we train on all (K 2) comparisons from each prompt as a single batch element. This is much more computationally efficient because it only requires a single forward pass of the RM for each completion (rather than (K 2) forward passes for K completions) and, because it no longer overfits, it achieves much improved validation accuracy and log loss.
Specifically, the loss function for the reward model is:

where rθ(x, y) is the scalar output of the reward model for prompt x and completion y with parametersθ, yw is the preferred completion out of the pair of yw and yl, and D is the dataset of human comparisons.
为了加速对比数据收集，我们给标注者4到9个响应来进行排序。这样每个prompt都会产生（K 2）个对比数据。因为对每个标注产生的对比数据集相关性很高，我们发现如果我们简单地将这些对比数据集融合到一个数据集，数据集迭代一遍就会导致奖励模型过拟和。因此，我们将一个prompt所对应的（K 2）个对比数据集作为一个batch来进行迭代。这个计算效率很高，因为对于每个prompt和response它仅需要一次前向传播（而不是（K 2）次计算），因为它不再过拟和，在验证集上的准确率和对数损失也更好。
奖励模型的损失函数是：

其中 rθ(x, y)是个标量输出函数，输入为promt x和completion y，参数为θ，yw是两个completion对yw和yl中较偏向的一个，D是人类对比数据集。
Finally, since the RM loss is invariant to shifts in reward, we normalize the reward model using a bias so that the labeler demonstrations achieve a mean score of 0 before doing RL.
最后，由于RM损失对奖励的偏移不变，我们使用偏差来规范化奖励模型，以便标签演示在进行RL之前达到平均分数为0。
Reinforcement learning (RL). Once again following Stiennon et al. (2020), we fine-tuned the SFT model on our environment using PPO (Schulman et al., 2017). The environment is a bandit environment which presents a random customer prompt and expects a response to the prompt. Given the prompt and response, it produces a reward determined by the reward model and ends the episode.In addition, we add a per-token KL penalty from the SFT model at each token to mitigate over-optimization of the reward model. The value function is initialized from the RM. We call these models “PPO.”
强化学习（RL）。我们再一次使用Stiennon et al. (2020)的方法在我们的环境下使用PPO(Schulman et al., 2017)算法来微调SFT模型。环境是一个bandit环境，输入一个随机的prompt，期望一个回应，然后将输入和回应一起送入奖励模型打一个分数，并结束episode。此外，我们添加了一个KL惩罚项，即强化学习模型输出的概率与原SFT模型的概率比的对数，这样可以减少强化学习与SFT模型的偏差。值函数来自RM模型。我们称这些模型为“PPO”。
We also experiment with mixing the pretraining gradients into the PPO gradients, in order to fix the performance regressions on public NLP datasets. We call these models “PPO-ptx.”

We maximize the following combined objective function in RL training:

where πRLφ is the learned RL policy, πSFT is the supervised trained model, and Dpretrain is the pretraining distribution. The KL reward coefficient, β, and the pretraining loss coefficient, γ, control the strength of the KL penalty and pretraining gradients respectively. For "PPO" models, γ is set to 0.Unless otherwise specified, in this paper InstructGPT refers to the PPO-ptx models.

我们还尝试将预训练梯度和PPO梯度混合， 修复在公开NLP数据集上的性能回归。我们称这些模型为“PPO-ptx”。最6终我们最大化以下合进后的目标函数：

这里πRLφ是RL策略，πSFT是监督训练模型，Dpretrain是预训练数据分布。KL奖励系数β和预训练损失系数 γ分别控制KL惩罚和预训练梯度的强度。对于"PPO"模型，γ设置为0。除非有特殊情况，本文InstructGPT都是指PPO-ptx模型。

Baselines. We compare the performance of our PPO models to our SFT models and GPT-3. We also compare to GPT-3 when it is provided a few-shot prefix to ‘prompt’ it into an instruction-following mode (GPT-3-prompted). This prefix is prepended to the user-specified instruction.

We additionally compare InstructGPT to fine-tuning 175B GPT-3 on the FLAN (Wei et al., 2021) and T0 (Sanh et al., 2021) datasets, which both consist of a variety of NLP tasks, combined with natural language instructions for each task (the datasets differ in the NLP datasets included, and the style of instructions used). We fine-tune them on approximately 1 million examples respectively and choose the checkpoint which obtains the highest reward model score on the validation set. See Appendix C for more training details.

基线 我们将PPO模型和SFT模型、GPT-3模型进行了对比。同时我们还跟指令跟随模式的GPT-3进行了对比（GTP-3-prompted），具体做法是添加一个few-shot‘提示’。这个前缀是添加在用户指定的指令之前。

我们还拿InstructGPT和微调的1750亿的GPT-3进行了对比，微调数据集分别是FLAN (Wei et al., 2021)和T0 (Sanh et al., 2021)，每个数据集都是由一系列的NLP任务组成的，每个任务都配有自然语言指令。我们分别用了100万的数据做微调，并取了验证集上奖励模型分数最高的检查点。更多训练细节见附录C。

3.6 评估

To evaluate how “aligned” our models are, we first need to clarify what alignment means in this context. The definition of alignment has historically been a vague and confusing topic, with various competing proposals (Chen et al., 2021; Leike et al., 2018; Gabriel, 2020). Following Leike et al.(2018), our aim is to train models that act in accordance with user intentions. More practically, for the purpose of our language tasks, we use a framework similar to Askell et al. (2021), who define models to be aligned if they are helpful, honest, and harmless.

要评估模型在对齐方面的能力，首先我们要弄清楚什么是对齐。对齐的定义一直以来都是模凌两可，有不同的说法 (Chen et al., 2021; Leike et al., 2018; Gabriel, 2020). Following Leike et al.(2018)，我们的目标是训练出一个模型，使其产出的内容能够符合用户意图。更确切地说，对齐模型就是有帮助的，诚实的和无害的。

To be helpful, the model should follow instructions, but also infer intention from a few-shot prompt or another interpretable pattern such as “Q: {question}\nA:”. Since a given prompt’s intention can be unclear or ambiguous, we rely on judgment from our labelers, and our main metric is labeler preference ratings. However, since our labelers are not the users who generated the prompts, there could be a divergence between what a user actually intended and what the labeler thought was intended from only reading the prompt.

为了评估帮助性，模型应该遵循指令，也应该从few-shot prompt或其它可以解释模式（例如Q: {question}\nA:）中推断意图。因为意图可能不清楚或模棱两可，我们主要通过标注人员来判断，因此标准就是标注人员的评分。然而标注者并不prompt的创造者，在用户的实际意图和标注者认为的意图之间存在差异。

It is unclear how to measure honesty in purely generative models; this requires comparing the model’s actual output to its “belief” about the correct output, and since the model is a big black box, we can’t infer its beliefs. Instead, we measure truthfulness—whether the model’s statements about the world are true—using two metrics: (1) evaluating our model’s tendency to make up information on closed domain tasks (“hallucinations”), and (2) using the TruthfulQA dataset (Lin et al., 2021). Needless to say, this only captures a small part of what is actually meant by truthfulness.

在纯生成模型中，如何衡量诚实性还不清楚。这需要将模型的实际输出和正确输出的“信念”进行对比，但由于模型是个黑盒子，我们无法推断其信念。因此，我们可以使用两个指标来衡量其真实性：（1）评估模型在封闭域任务中制造信息的倾向（“幻觉”），（2）使用TruthfulQA数据集 (Lin et al., 2021)。当然，这只是捕捉到了真实性的一小部分。

Similarly to honesty, measuring the harms of language models also poses many challenges. In most cases, the harms from language models depend on how their outputs are used in the real world. For instance, a model generating toxic outputs could be harmful in the context of a deployed chatbot, but might even be helpful if used for data augmentation to train a more accurate toxicity detection model.Earlier in the project, we had labelers evaluate whether an output was ‘potentially harmful’. However,we discontinued this as it required too much speculation about how the outputs would ultimately be used; especially since our data also comes from customers who interact with the Playground API interface (rather than from production use cases).

与诚实性类似，衡量语言模型的危害性也面临许多挑战。在许多情况下，语言模型的危害取决于现实世界的使用场景。例如，模型产生的有毒输出在聊天机器人环境中是有害的，但是在数据增强用以毒性模型训练的场景中就是有用的。在项目早期，我们让标注者评估一个输出是否“潜在有害”。然

而，我们终止了这一做法，因为这需要很多最终如何使用的猜测，特别是我们的数据是来自Playground API（而不是生产上）。

Therefore we use a suite of more specific proxy criteria that aim to capture different aspects of behavior in a deployed model that could end up being harmful: we have labelers evaluate whether an output is inappropriate in the context of a customer assistant, denigrates a protected class, or contains sexual or violent content. We also benchmark our model on datasets intended to measure bias and toxicity, such as RealToxicityPrompts (Gehman et al., 2020) and CrowS-Pairs (Nangia et al., 2020).

To summarize, we can divide our quantitative evaluations into two separate parts:

因此，我们使用一套更具体的方法来捕捉有害行为：我们让标注者评估在客户助手上下文中的输出是否不当，是否贬低受保护群体，是否包括性或暴力内容。我们还对模型进行了基准测试，使用旨在衡量偏见和毒性的数据集，比如RealToxicityPrompts (Gehman et al., 2020) and CrowS-Pairs (Nangia et al., 2020)。

总之，我们将量化评估分为两个独立的部分：

Evaluations on API distribution. Our main metric is human preference ratings on a held out set of prompts from the same source as our training distribution. When using prompts from the API for evaluation, we only select prompts by customers we haven’t included in training. However, given that our training prompts are designed to be used with InstructGPT models, it’s likely that they disadvantage the GPT-3 baselines. Thus, we also evaluate on prompts submitted to GPT-3 models on the API; these prompts are generally not in an ‘instruction following’ style, but are designed specifically for GPT-3. In both cases, for each model we calculate how often its outputs are preferred to a baseline policy; we choose our 175B SFT model as the baseline since its performance is near the middle of the pack. Additionally, we ask labelers to judge the overall quality of each response on a 1-7 Likert scale and collect a range of metadata for each model output (see Table 3).

基于API的评估。我们的主要评估就是人类偏好打分，prompts为训练集同来源的保留验证集。当使用来自API的prompts进行评估时，我们选择了训练时未使用的客户prompts。然而，因为我们的训练prompts是为了训练InstructGPT而设置的，因此它可能会降低GPT-3基线的性能。因此，我们还对提交给 API 上的 GPT-3 模型的提示进行了评估；这些提示通常不是“遵循指令”的风格，而是专门为 GPT-3 设计的。在这两种情况下，我们都计算了每个模型的输出相对于基线策略的偏好程度；我们选择了我们的 175B SFT 模型作为基线，因为其性能接近中等水平。此外，我们要求标注员根据 1-7 Likert 刻度评估每个响应的整体质量，并收集了每个模型输出的一系列元数据（见表 3）。

Evaluations on public NLP datasets. We evaluate on two types of public datasets: those that capture an aspect of language model safety, particularly truthfulness, toxicity, and bias, and those that capture zero-shot performance on traditional NLP tasks like question answering, reading comprehension, and summarization. We also conduct human evaluations of toxicity

on the RealToxicityPrompts dataset (Gehman et al., 2020). We are releasing samples from our models on all of the sampling-based NLP tasks.

基于NLP数据集的评估 我们在两类数据集上进行了评估：一类是那些关注模型安全性，特别是真实性、毒性和偏见的数据集；二类是关注模型在零样本表现上的数据集，比如问答、阅读理解、摘要。我们还在RealToxicityPrompts(Gehman et al., 2020)上进行了毒性评估。我们将发布所有这些任务的样本。

4 结果

In this section, we provide experimental evidence for our claims in Section 1, sorted into three parts:results on the API prompt distribution, results on public NLP datasets, and qualitative results.

在本小节，我们提供了第1节中所主张的实验证据，分为三个部分：在API prompt的结果，NLP数据集的结果，定性的结果。

图3： 模型结果，评估指标为相对于175B的SFT模型的胜率。左侧：来自GPT模型API上prompts的评估结果；右侧：来自InstructGPT模型API上prompts的评估结果；上部：保留标注者的评估结果；下部：训练标注者的评估结果。左侧我们去掉了GPT(prompted)模型，因为这些GPT API的prompts专为GPT而设计，存在不公平性。

4.1 API评估结果

Labelers significantly prefer InstructGPT outputs over outputs from GPT-3. On our test set of prompts, our labelers significantly prefer InstructGPT outputs across model sizes. These results are shown in Figure 1. We find that GPT-3 outputs perform the worst, and one can obtain significant step-size improvements by using a well-crafted few-shot prompt (GPT-3 (prompted)), then by training on demonstrations using supervised learning (SFT), and finally by training on comparison data using PPO. Adding updates on the pretraining mix during PPO does not lead to large changes in labeler preference. To illustrate the magnitude of our gains: when compared directly, 175B InstructGPT outputs are preferred to GPT-3 outputs 85 ± 3% of the time, and preferred 71 ± 4% of the time to few-shot GPT-3.

标注者明显偏好InstructGPT高于GPT-3 在我们的测试prompts上，标注者明显偏好InstructGPT，在所有的模型大小上都是。结果见图1。我们发现GPT-3是最差的，然后GPT-3（prompted）、SFT、PPO次第增加。增加预训练部分到PPO上并没有带来标注者偏好的显著提升。量化增幅为：175B的InstructGPT在85 ± 3%的情况下优于GPT-3的输出，在71 ± 4%的情况下优于GPT-3的输出。

We also found that our results do not change significantly when evaluated on prompts submitted to GPT-3 models on the API (see Figure 3), though our PPO-ptx models perform slightly worse at larger model sizes.

我们还发现，来自GPT-3 API的prompts进行评估时效果并没有多少变动，尽管PPO-ptx的表现在大尺寸上表现稍差。

In Figure 4 we show that labelers also rate InstructGPT outputs favorably along several more concrete axes. Specifically, compared to GPT-3, InstructGPT outputs are more appropriate in the context of a customer assistant, more often follow explicit constraints defined in the instruction (e.g. “Write your answer in 2 paragraphs or less.”), are less likely to fail to follow the correct instruction entirely, and make up facts (‘hallucinate’) less often in closed-domain tasks. These results suggest that InstructGPT models are more reliable and easier to control than GPT-3. We’ve found that our other metadata categories occur too infrequently in our API to obtain statistically significant differences between our models.
在图4，我们展示了标注者在更多确切的维度对InstructGPT进行评价。具体来说，相较于GPT-3，InstructGPT输出在客户助理的内容上更合适，在明确限制的指令上回复得更好（e.g. “Write your answer in 2 paragraphs or less.”），在正确理解指令上面表现得更好，更少地虚构事实。这些结果表明相较于GPT-3，InstructGPT更可靠，更可控。我们发现其它一些API中数据类型，因为出现较少，所在在这些模型之间没有统计意义上的差别。

图4 API分布上的结果。鉴于数据集较小，不同大小模型的结果汇总到了一起。详情见附录E.2。与GPT-3相比，PPO模型在客户这里的内容生成上更合适，在遵循明确限制的指令以及正确地遵循指令上表现得更好，更少地“产生幻觉”（即在封闭域任务比如摘要提取上虚构信息）。

Our models generalize to the preferences of "held-out" labelers that did not produce any training data. Held-out labelers have similar ranking preferences as workers who we used to produce training data (see Figure 3). In particular, according to held-out workers, all of our InstructGPTmodels still greatly outperform the GPT-3 baselines. Thus, our InstructGPT models aren’t simply overfitting to the preferences of our training labelers**.**

我们的模型可以泛化至未在训练数据打标的标注者上（保留标注者） 保留标注者和训练数据打标的标注者排序偏好是一致的（见图3）。所有的InstructGPT模型依旧领先于GPT-3基线。这也意味着我们的InstructGPT并没有过拟和。

We see further evidence of this from the generalization capabilities of our reward models. We ran an experiment where we split our labelers into 5 groups, and train 5 RMs (with 3 different seeds) using 5-fold cross validation (training on 4 of the groups, and evaluating on the held-out group). These RMs have an accuracy of 69.6 ± 0.9% on predicting the preferences of labelers in the held-out group,a small decrease from their 72.4 ± 0.4% accuracy on predicting the preferences of labelers in their training set.
奖励模型的泛化能力了印证了这一点。我们将标注者分为五组，使用五折交检验证方法建立了五个奖励模型。这些奖励模型在验证集上的准确率在69.6 ± 0.9%，在训练集上的准确率在72.4 ± 0.4% 。

Public NLP datasets are not reflective of how our language models are used. In Figure 5, we also compare InstructGPT to our 175B GPT-3 baselines fine-tuned on the FLAN (Wei et al., 2021) and T0 (Sanh et al., 2021) datasets (see Appendix C for details). We find that these models perform better than GPT-3, on par with GPT-3 with a well-chosen prompt, and worse

than our SFT baseline. This indicates that these datasets are not sufficiently diverse to improve performance on our API prompt distribution. In a head to head comparison, our 175B InstructGPT model outputs were preferred overour FLAN model 78 ±4% of the time and over our T0 model 79 ± 4% of the time. Likert scores for these models are shown in Figure 5.

图5 对比我们的模型与FLAN及TO在Likert评分上的差异，评分在1-7个档次，评估数据是来源于体验版InstructGPT的prompt。FLAN和TO比GPT-3表现的更好，与“指令-跟随”模式的GPT-3性能持平。

We believe our InstructGPT model outperforms FLAN and T0 for two reasons. First, public NLP datasets are designed to capture tasks that are easy to evaluate with automatic metrics, such as classification, question answering, and to a certain extent summarization and translation. However,classification and QA are only a small part (about 18%) of what API customers use our language models for, whereas open-ended generation and brainstorming consist of about 57% of our prompt dataset according to labelers (see Table 1). Second, it can be difficult for public NLP datasets to obtain a very high diversity of inputs (at least, on the kinds of inputs that real-world users would be interested in using). Of course, tasks found in NLP datasets do represent a kind of instruction that we would like language models to be able to solve, so the broadest type instruction-following model would combine both types of datasets.

我们认为InstructGPT模型战胜FLAN和T0有两个原因。首先，公开NLP数据集主要是易于评估的任务，比如分类、问答，再比如摘要提取和翻译。然而，分类和QA只是API用户数据集的一小部分（18%），开放式生成和头脑风暴占了57%（见表1）。第二，公开NLP数据集的数据多样性也不行（至少，在真实世界的用户感兴趣的覆盖较少）。当然，NLP数据集的任务确实代表了我们希望语言模型来解决的指令类型，所以，最广泛的指令跟随模型应该包含这两种类型的数据集。

4.2 公开数据集上的结果

InstructGPT models show improvements in truthfulness over GPT-3. As measured by human evaluatoins on the TruthfulQA dataset, our PPO models show small but significant improvements in generating truthful and informative outputs compared to GPT-3 (see Figure 6). This behavior is the default: our models do not have to be specifically instructed to tell the truth to exhibit improved truthfulness. Interestingly, the exception is our 1.3B PPO-ptx model, which performs slightly worse than a GPT-3 model of the same size. When evaluated only on prompts that were not adversarially selected against GPT-3, our PPO models are still significantly more truthful and informative than GPT-3 (although the absolute improvement decreases by a couple of percentage points）.

InstructGPT模型在真实度上超越了GPT-3。在人类评估的TruthfulQA数据集上，我们的PPO模型在真实性和丰富度上少许（但很关键）超越了GPT-3（见图6）。这个结果是默认的：我们的模型没有经

过特殊指令就能展示出更好的真实性。有趣的是，在1.3B的PPO-ptx上有个意外，它比同尺寸的GPT-3效果还差。在没有用对抗GPT-3的prompts上进行评估时，我们的PPo模型依然比GPT-3更真实和丰富（尽管在绝对提升上少了几个百分点）。

图6 TruthfulQA数据集的评估结果。灰色条形代表真实度的评分；有色条形代表真实度和可信度的评分。

Following Lin et al. (2021), we also give a helpful "Instruction+QA" prompt that instructs the model to respond with "I have no comment" when it is not certain of the correct answer. In this case, our PPO models err on the side of being truthful and uninformative rather than confidently saying a falsehood; the baseline GPT-3 model aren't as good at this.

像Lin et al. (2021)的做法一样，我们也给了一个有用的"指令+问答"prompt，指导模型在没有正确答案时回答"我回答不了"。在这种情况下，我们的PPO模型更倾向于诚实的不答，而不是说一个错的答案；但是基线GPT-3却做不到这样。

Our improvements in truthfulness are also evidenced by the fact that our PPO models hallucinate (i.e.fabricate information) less often on closed-domain tasks from our API distribution, which we've shown in Figure 4.

我们在真实性的提升也在幻觉方面得到了印证，我们的PPO模型在API的封闭域任务上更少地产生幻觉（比如制造信息），如图4所示。

InstructGPT shows small improvements in toxicity over GPT-3, but not bias. We first evaluate our models on the RealToxicityPrompts dataset (Gehman et al., 2020). We do this in two ways: we run model samples through the Perspective API to obtain automatic toxicity scores, which is the standard evaluation procedure for this dataset, and we also send these samples to labelers to obtain ratings on absolute toxicity, toxicity relative to the prompt, continuity, and overall output preference.We sample prompts from this dataset uniformly according to prompt toxicity to better assess how our models perform with high input toxicity (see Figure 39 in Appendix E); this differs from the standard prompt sampling for this dataset, and thus our absolute toxicity numbers are inflated.

InstructGPT在毒性方面相对于GPT-3有小的改善，但是在偏见方面没有改善。我们首先评估了数据集RealToxicityPrompts（Gehman et al., 2020）。我们用了两种方法：第一种，通过Perspective API来自动获取评分，这是标准做法。第二种，我们将这些样本发给标注者，以得到绝对毒性，相对于prompt的毒性，连续性，以及整体偏好评分。我们从该数据集中均匀地抽样prompt，根据prompt的毒性来更好地评估我们的模型在高输入毒性情况下的表现（详见附录 E 中的图 39）；这与该数据集的标准prompt抽样方式不同，这样绝对毒性数值会被夸大一些。

Our results are in Figure 7. We find that, when instructed to produce a safe and respectful output("respectful prompt"), InstructGPT models generate less toxic outputs than those from GPT-3 according to the Perspective API. This advantage disappears when the respectful

prompt is removed(“no prompt”). Interestingly, when explicitly prompted to produce a toxic output, InstructGPT outputs are much more toxic than those from GPT-3 (see Figure 39).
我们的结果在图7。我们发现，要求模型生成安全且尊重的输出（“尊重的prompt”）时，InstructGPT模型在Perspective API上生成的毒性更少。但是把尊重这一prompt拿掉时，优势就又消失了。有趣地是，当显示地提示产生有毒的输出时，InstructGPT比GPT-3更有毒性（见图39）。

图7：在RealToxicityPrompts上对比人类评估和自动评估（Perspective API分数）。共有1729个prompts，对比的模型是三个175B模型，使用和没使用“respectful”指令两种模式。这里自动评估和人类评估的prompts一样，因此和整体prompts略有不同（见附录D的表14）。
These results are confirmed in our human evaluations: InstructGPT is less toxic than GPT-3 in the“respectful prompt” setting, but performs similarly in the “no prompt” setting. We provide extended results in Appendix E. To summarize: all of our models are rated as less toxic than expected given the prompt (they get a negative score on a scale from -1 to 1, where 0 is ‘about as toxic as expected’).Our SFT baseline is the least toxic out of all of our models, but also has the lowest continuity and is the least preferred in our rankings, which could indicate that the model generates very short or degenerate responses.
人类评估的结果如下：在“repectful prompt”的设置下，InstructGPT比GPT-3毒性更低，在“no prompt”设置下，二者相当。详细结果见附录E。总结一下：所有的模型毒性都比预期的要低（分数在-1到1，0代表与预期一致）。我们的SFT基线模型毒性是最差的，同时有最低的连续性，并且在偏好排名中最差，这说明了这个模型产生的结果很短或有退化情况。
To evaluate the model’s propensity to generate biased speech (see Appendix E), we also evaluated InstructGPT on modified versions of the Winogender (Rudinger et al., 2018) and CrowS-Pairs (Nangiaet al., 2020) datasets. These datasets consists of pairs of sentences which can highlight potential bias.We calculate the relative probabilities of producing the sentences in each pair and the entropy (in bits) of the associated binary probability distributions. Perfectly unbiased models will have no preference between the sentences in each pair and will therefore have maximum entropy. By this metric, our models are not less biased than GPT-3. The PPO-ptx model shows similar bias to GPT-3, but when instructed to act respectfully it exhibits lower entropy and thus higher bias. The pattern of the bias is not clear; it appears that the instructed models are more certain of their outputs regardless of whether or not their outputs exhibit stereotypical behavior.
为了评估模型生成偏见性言论的倾向（详见附录E），我们还对InstructGPT在Winogender（Rudinger等人，2018）和CrowS-Pairs（Nangia等人，2020）数据集的修改版本上进行了评估。这些数据集由一对句子组成，可以突显潜在的偏见。我们计算了每对句子产生的相对概率以及相关二进制概率分布的熵（以比特为单位）。完全无偏的模型在每对句子之间没有偏好，因此熵最大。按照这个度量标准，我们的模型并不比GPT-3更少偏见。PPO-ptx模型表现出与GPT-3类似的偏见，但在

被指示要尊重时，它的熵较低，因此偏见较高。偏见的模式并不清楚；看起来，被指示的模型对其输出更有把握，无论其输出是否表现出刻板行为。

We can minimize performance regressions on public NLP datasets by modifying our RLHF fine-tuning procedure. By default, when we train a PPO model on our API distribution, it suffers from an “alignment tax”, as its performance on several public NLP datasets decreases. We want an alignment procedure that avoids an alignment tax, because it incentivizes the use of models that are unaligned but more capable on these tasks.

通过改进RLHF微调的过程，我们可以最小化公开NLP数据集的表现退化。默认情况下，当我们通过API数据来训练一个PPO模型时，它存在一个“对齐税”现象，即它的表现在公开NLP数据集上会下降。我们希望对齐的时候避免出现对齐税现象。

In Figure 29 we show that adding pretraining updates to our PPO fine-tuning (PPO-ptx) mitigates these performance regressions on all datasets, and even surpasses GPT-3 on HellaSwag. The performance of the PPO-ptx model still lags behind GPT-3 on DROP, SQuADv2, and translation; more work is needed to study and further eliminate these performance regressions.

在图29，我们展示了通过添加预训练更新到PPO的微调上（PPO-ptx）可以减轻所有数据集的性能回退，甚至在HellaSwag上会超越GPT-3。PPO-ptx模型在DROP, SQuADv2及翻译上的表现依然落后于GPT-3。在消除性能回退上还有很多工作要做。

Mixing in pretraining updates performs better than the simpler solution of increasing the KL co-efficient. In Figure 33, we show that there is a value of the pretraining mix coefficient that both reverses the performance regressions on SQuADv2 and DROP (the datasets we used for testing), and has minimal reductions in validation reward. In contrast, increasing the KL coefficient (Figure 34) leads to significant decreases in validation reward and never fully recovers on DROP and SQuAD.Changing the KL model from the PPO init to GPT-3 gives similar results.

增加预训练更新比简单地提升KL系数效果要好。在图33，我们发现有个值是关于预训练惩罚系数，它既可以逆转SQuADv2和DROP的性能下降，又可以最小化验证集的奖励下降。相比，提升KL系数（图34）会导致验证奖励的大幅下降，并且无法恢复DROP和SQuAD的效果。将KL模型从PPO初始化改为GPT-3依然有相似结论。

4.3 定性结果

InstructGPT models show promising generalization to instructions outside of the RLHF fine-tuning distribution. In particular, we find that InstructGPT shows ability to follow instructions in non-English languages, and perform summarization and question-answering for code. This is interesting because non-English languages and code form a tiny minority of our fine-tuning data, and it suggests that, in some cases, alignment methods could generalize to producing the desired behavior on inputs that humans did not directly supervise.

InstructGPT模型在RLHF微调数据以外也展示出了泛化性。我们发现InstructGPT在非英语领域也有很好的指令跟随能力，还有摘要和关于代码的问答。这很有趣，因为非英语和代码仅仅是我们微调数据的很小一部分。这表明，在一些情况下，对齐方法可以泛化以产生人类没有直接监督的行为。
We do not track these behaviors quantitatively, but we show some qualitative examples in Figure 8.Our 175B PPO-ptx model is able to reliably answers questions about code, and can also follow instructions in other languages; however, we notice that it often produces an output in English even when the instruction is in another language. In comparison, we find that GPT-3 can perform these tasks but requires more careful prompting, and rarely follows instructions in these domains.
我们没有定量地评估这些行为，但是我们在图8给出了一些定性的例子。我们的175B的PPO-ptx模型可以可靠地回答关于代码的问题，而且可以在其它语言跟随指令。然而，我们注意到指令是非语时也会产生英语输出。作为对比，我们发现GPT-3也可以执行这些任务但是需要更多的prompting，且指令跟随能力一般。
InstructGPT still makes simple mistakes. In interacting with our 175B PPO-ptx model, we have noticed it can still make simple mistakes, despite its strong performance on many different language tasks. To give a few examples: (1) when given an instruction with a false premise, the model sometimes incorrectly assumes the premise is true, (2) the model can overly hedge; when given a simple question, it can sometimes say that there is no one answer to the question and give multiple possible answers, even when there is one fairly clear answer from the context, and (3) the model’s performance degrades when instructions contain multiple explicit constraints (e.g. “list 10 movies made in the 1930’s set in France”) or when constraints can be challenging for language models (e.g.writing a summary in a specified number of sentences).
InstructGPT依然会犯简单地错误。在与175B的PPO-ptx模型交互过程中，我们注意到它依然会犯简单地错误，尽管它在很多语言任务上表现不错。下面举几个例子：（1）当指令给出一个错误的前提时，模型有时会认为前提是对的，（2）模型会过于谨慎，当给一个简单地问题时，它有时会说没有明确的答案，即使在文本中有了明确的答案也是如此，（3）当指令有多个明确的约束（例如“列出20世纪30年代10部法国电影”）或者约束有挑战时（例如根据指定的句子写一个摘要）模型的性能出会下降。
We show some examples of these behaviors in Figure 9. We suspect that behavior (2) emerges partly because we instruct labelers to reward epistemic humility; thus, they may tend to reward outputs that hedge, and this gets picked up by our reward model. We suspect that behavior (1) occurs because there are few prompts in the training set that assume false premises, and our models don’t generalize well to these examples. We believe both these behaviors could be dramatically reduced with adversarial data collection (Dinan et al., 2019b).
我们在图9中展示了这些行为的一些示例。我们怀疑行为（2）部分是因为我们指示标注者在做奖励时保持谨慎，因此，他们可能倾向于奖励那些含糊其辞的输出，而这被我们的奖励模型捕捉到。我们怀

疑行为（1）之所以发生，是因为训练集中很少有假设错误前提的prompts，而我们的模型对这些示例的泛化能力不强。我们认为通过对抗性数据收集（Dinan等人，2019b）可以大大减少这两种行为。

图9：在没有前缀的情况下175B的PPO-ptx与175B GPT-3相比会犯一些错误。prompts是精心挑选的，但是输出没有。（1）InstructGPT会被错误假设的指令弄混淆，并简单地跟着去回答。（2）IntrtuctGPT过于谨慎，不直接回答问题（这样的话，南瓜也会爆炸了）。注意这些案例并不能反映GPT-3的能力，因为它没有被提示进入“问答”模式。

5 讨论

5.1 对齐研究的启示

This research is part of our broader research program to align AI systems with human intentions (Chris-tiano et al., 2017; Ziegler et al., 2019; Stiennon et al., 2020). Even though this work focuses on our current language model systems, we seek general and scalable methods that work for future AI systems (Leike et al., 2018). The systems we work with here are still fairly limited, but they are among the largest language models today and we apply them on a wide range of language tasks,including classification, summarization, question-answering, creative writing, dialogue, and others.

这项研究是我们更广泛研究计划的一部分，旨在将AI系统与人类意图对齐（Chris-tiano et al., 2017; Ziegler et al., 2019; Stiennon et al., 2020）。尽管这项工作聚焦在我们当前的语言模型系统，我们也在寻求更通用和可扩展的方法来适用于未来的AI系统（Leike et al., 2018）。我们现在使用的系统仍然相对有限，但是它们是当今最大的语言模型之一，我们将它们应用广泛的语言任务，包括分类、摘要提取、问答、创作、对话等。

Our approach to alignment research in this work is iterative: we are improving the alignment of current AI systems instead of focusing abstractly on aligning AI systems that don’ t yet exist. A disadvantage of this approach is that we are not directly facing alignment problems that occur only when aligning superhuman systems (Bostrom, 2014). However, our approach does provides us with a clear empirical feedback loop of what works and what does not. We believe that this feedback loop is essential to refine our alignment techniques, and it forces us to keep pace with progress in machine learning. Moreover, the alignment technique we use here, RLHF, is an important building block inseveral proposals to align superhuman systems (Leike et al., 2018; Irving et al., 2018; Christianoet al., 2018). For example, RLHF was a central method in recent work on summarizing books, a task that exhibits some of the difficulties of aligning superhuman AI systems as it is difficult for humans to evaluate directly (Wu et al., 2021).

我们在这项研究中对齐问题的方法是迭代式的：我们正在改进当前 AI 系统的对齐，而不是抽象地关注尚不存在的 AI 系统的对齐。这种方法的一个不足之处在于，我们没有直接面对只在对齐超人类系统时才出现的对齐问题（Bostrom，2014）。然而，我们的方法确实为我们提供了一个明确的实证反馈循环，以了解什么有效，什么无效。我们认为这个反馈循环对于改进我们的对齐技术至关重要，它迫使

我们跟上机器学习的进展。此外，我们在这里使用的对齐技术，即 RLHF，是对齐超人类系统的几个提案的重要基石（Leike 等，2018；Irving 等，2018；Christiano 等，2018）。例如，RLHF 是最近关于总结书籍主要内容的一个核心方法，这项任务展示了对齐超人类 AI 系统的一些困难，因为人类难以直接评估（Wu 等，2021）。

From this work, we can draw lessons for alignment research more generally:

1. The cost of increasing model alignment is modest relative to pretraining. The cost of collecting our data and the compute for training runs, including experimental runs is a fraction of what was spent to train GPT-3: training our 175B SFT model requires 4.9 petaflops/s-days and training our 175B PPO-ptx model requires 60 petaflops/s-days,compared to 3,640 petaflops/s-days for GPT-3 (Brown et al., 2020). At the same time,our results show that RLHF is very effective at making language models more helpful to users, more so than a 100x model size increase. This suggests that right now increasing investments in alignment of existing language models is more cost-effective than training larger models—at least for our customers' natural language task distribution.

2. We've seen some evidence that InstructGPT generalizes 'following instructions' to settings that we don't supervise it in, for example on non-English language tasks and code-related tasks. This is an important property because it's prohibitively expensive to have humans supervise models on every task they perform. More research is needed to study how well this generalization scales with increased capabilities; see Christiano et al. (2021) forrecent research in this direction.

3. We were able to mitigate most of the performance degradations introduced by our fine-tuning. If this was not the case, these performance degradations would constitute an alignment tax—an additional cost for aligning the model. Any technique with a high tax might not see adoption. To avoid incentives for future highly capable AI systems to remain unaligned with human intent, there is a need for alignment techniques that have low alignment tax. To this end, our results are good news for RLHF as a low-tax alignment technique.

4. We've validated alignment techniques from research in the real world. Alignment research has historically been rather abstract, focusing on either theoretical results (Soareset al., 2015), small synthetic domains (Christiano et al., 2018; Leike et al., 2017), or training ML models on public NLP datasets (Ziegler et al., 2019; Stiennon et al., 2020). Our work provides grounding for alignment research in AI systems that are being used in production in the real world with customers. This enables an important feedback loop on the techniques' effectiveness and limitations.

从这项工作中，我们可以得出一些关于更普遍的模型对齐研究的经验教训：

1、相对于预训练，提高模型对齐的成本较低。我们收集数据以及训练运行的计算成本，包括实验运行，只是训练GPT-3所花费的一小部分：训练我们的175B SFT模型需要4.9 petaflops/s-days，而训

练我们的175B PPO-ptx模型需要60 petaflops/s-days，而GPT-3则需要3,640 petaflops/s-days（Brown等人，2020）。与此同时，我们的结果表明，RLHF对于使语言模型对用户更有帮助非常有效，比100倍语言模型大小的增加更有效。这表明，目前增加对现有语言模型的对齐投资比训练更大的模型更具成本效益，至少对于我们客户的自然语言任务分布而言。

2、我们已经看到一些证据表明InstructGPT可以将“遵循指令”的能力推广到我们没有监督的设置中，例如非英语语言任务和与代码相关的任务。这是一个重要的特性，因为让人类在每项任务上监督模型的成本非常高昂。需要进一步研究这种泛化随着能力增加而如何扩展；请参阅Christiano等人（2021）的最新研究。

3、我们能够减轻大部分由微调引入的性能降低。如果不是这样的话，这些性能降低将构成对齐的额外成本。任何具有高成本的技术可能不会被采用。为了避免未来高度能力的AI系统保持与人类意图不一致的激励，需要具有低对齐成本的对齐技术。从这个角度看，我们的结果对于RLHF作为一种低成本对齐技术来说是好消息。

4、我们已经在现实世界中验证了对齐技术的研究。对齐研究在历史上一直相当抽象，主要关注以下几个方面：理论结果（Soares等人，2015），小型合成领域（Christiano等人，2018；Leike等人，2017），或者在公共自然语言处理数据集上训练机器学习模型（Ziegler等人，2019；Stiennon等人，2020）。我们的工作为正在实际生产中客户正使用的AI系统的对齐研究提供了基础。这使得对这些技术的有效性和局限性形成了一个重要的反馈循环。

5.2 我们在对齐谁?

When aligning language models with human intentions, their end behavior is a function of the underlying model (and its training data), the fine-tuning data, and the alignment method used. In this section, we describe a number of factors that influence the fine-tuning data specifically, to ultimately determine what and who we’ re aligning to. We then consider areas for improvement before a larger discussion of the limitations of our work in Section 5.3.

当将语言模型与人类意图对齐时，它们的最终行为是基于底层模型（以及其训练数据）、微调数据和所使用的对齐方法的函数。在本节中，我们描述了一些影响微调数据的因素，以最终确定我们正在对齐的内容和对象。然后，我们在第5.3节中讨论了我们工作的局限性。

The literature often frames alignment using such terms as “human preferences” or “human values.”In this work, we have aligned to a set of labelers’ preferences that were influenced, among others things, by the instructions they were given, the context in which they received them (as a paid job),and who they received them from. Some crucial caveats apply:

文献中通常使用“人类偏好”或“人类价值观”等术语来描述对齐。在这项工作中，我们对一组标注者的偏好进行了对齐，这些偏好受到了多种因素的影响，包括他们所接收的指令、接收指令的背景（作为一份有偿工作）以及指令的来源。但需要注意的是，有一些关键的限制条件：

First, we are aligning to demonstrations and preferences provided by our training labelers, who directly produce the data that we use to fine-tune our models. We describe our labeler hiring process and demographics in Appendix B; in general, they are mostly English-speaking

people living in the United States or Southeast Asia hired via Upwork or Scale AI. They disagree with each other on many examples; we found the inter-labeler agreement to be about 73%.
首先，我们正在对齐训练标注者提供的演示和偏好，这些标注者直接生成我们用于微调模型的数据。我们在附录B中描述了我们的标注者招聘过程和人口统计信息；总体而言，他们大多是通过Upwork或Scale AI雇佣的居住在美国或东南亚的英语为主的人群。他们在许多示例上存在分歧；我们发现标注者之间的一致性约为73%。
Second, we are aligning to our preferences, as the researchers designing this study (and thus by proxy to our broader research organization, OpenAI): we write the labeling instructions that labelers use as a guide when writing demonstrations and choosing their preferred output, and we answer their questions about edge cases in a shared chat room. More study is needed on the exact effect of different instruction sets and interface designs on the data collected from labelers and its ultimate effect on model behavior.
其次，我们正在对齐到我们的偏好，作为设计这项研究的研究人员（因此也间接地对我们更广泛的研究机构OpenAI）：我们编写标注指令，标注者在撰写演示和选择首选输出时使用这些指令作为指南，并在共享聊天室中回答他们关于边缘案例的问题。我们需要进一步研究不同指令集和界面设计对从标注者收集的数据产生的确切影响，以及对模型行为的最终影响。
Third, our training data is determined by prompts sent by OpenAI customers to models on theOpenAI API Playground, and thus we are implicitly aligning to what customers think is valuable and, in some cases, what their end-users think is valuable to currently use the API for. Customers and their end users may disagree or customers may not be optimizing for end users' well-being; for example, a customer may want a model that maximizes the amount of time a user spends on their platform, which is not necessarily what end-users want. In practice, our labelers don't have visibility into the contexts in which a given prompt or completion will be seen.
其次，我们的训练数据是由OpenAI的客户通过OpenAI API Playground发送的提示决定的，因此我们在隐性地对齐客户认为有价值的内容，有时也包括他们的最终用户认为目前使用API有价值的内容。客户及其最终用户可能存在分歧，或者客户可能并未优化最终用户的福祉；例如，某客户可能希望模型最大化用户在其平台上的使用时间，而这未必是最终用户所期望的。在实践中，我们的标注者无法看到给定提示或完成将在哪些上下文中显示。
Fourth, OpenAI's customers are not representative of all potential or current users of language models—let alone of all individuals and groups impacted by language model use. For most of the duration of this project, users of the OpenAI API were selected off of a waitlist. The initial seeds for this waitlist were OpenAI employees, biasing the ultimate group toward our own networks.
第四，OpenAI的客户并不代表所有潜在或现有的语言模型用户，更不用说所有受到语言模型使用影响的个人和群体了。在这个项目的大部分时间里，OpenAI API的用户是从候补名单中挑选出来的。这个

候补名单的初始种子是OpenAI的员工，从而使最终的用户群体偏向我们自己的圈子。
Stepping back, there are many difficulties in designing an alignment process that is fair, transparent,and has suitable accountability mechanisms in place. The goal of this paper is to demonstrate that this alignment technique can align to an specific human reference group for a specific application.We are not claiming that researchers, the labelers we hired, or our API customers are the right source of preferences. There are many stakeholders to consider—the organization training the model, the customers using the model to develop products, the end users of these products, and the broader population who may be directly or indirectly affected. It is not only a matter of making the alignment process more participatory; it is impossible that one can train a system that is aligned to everyone’s preferences at once, or where everyone would endorse the tradeoffs.
回顾一下，设计一个公平、透明且具备适当问责机制的对齐过程存在许多困难。本文的目标是展示这种对齐技术可以对齐到特定的人类参考群体，用于特定的应用。我们并不声称研究人员、我们雇佣的标注者或我们的API客户是正确的偏好来源。有许多利益相关者需要考虑——训练模型的组织、使用模型开发产品的客户、这些产品的最终用户，以及可能会受到直接或间接影响的更广泛人群。这不仅仅是使对齐过程更具参与性的问题；不可能训练一个立即适应所有人偏好的系统，或者每个人都会认可其中的权衡。
One path forward could be to train models that can be conditioned on the preferences of certain groups, or that can be easily fine-tuned or prompted to represent different groups. Different models can then be deployed and used by groups who endorse different values. However, these models might still end up affecting broader society and there are a lot of difficult decisions to be made relating to whose preferences to condition on, and how to ensure that all groups can be represented and can opt out of processes that may be harmful.
一种前进的路径可能是训练可以根据特定群体的偏好进行条件调整的模型，或者可以轻松进行微调或提示以代表不同群体。然后，不同的模型可以被认可不同价值观的群体部署和使用。然而，这些模型可能仍然会影响更广泛的社会，需要做出许多困难的决策，涉及到应该根据谁的偏好进行调整，以及如何确保所有群体都能被代表并可以退出可能有害的过程。
5.3 限制
Methodology. The behavior of our InstructGPT models is determined in part by the human feedback obtained from our contractors. Some of the labeling tasks rely on value judgments that may be impacted by the identity of our contractors, their beliefs, cultural backgrounds, and personal history.We hired about 40 contractors, guided by their performance on a screening test meant to judge how well they could identify and respond to sensitive prompts, and their agreement rate with researchers on a labeling task with detailed instructions (see Appendix B). We kept our team of contractors small because this facilitates high-bandwidth communication with a smaller set of contractors who are doing the task full-time. However, this group is clearly not representative of the full spectrum of people who will use and be

affected by our deployed models. As a simple example, our labelers are primarily English-speaking and our data consists almost entirely of English instructions.

方法。我们的 InstructGPT 模型的行为在一定程度上受到从我们的承包商那里获得的人类反馈的影响。其中一些标注任务依赖于价值判断，这些判断可能会受到承包商的身份、信仰、文化背景和个人历史的影响。我们雇佣了约 40 名承包商，根据他们在筛选测试中的表现来判断他们对敏感prompts的识别和回应能力，以及他们与研究人员在标注任务上的一致率（详见附录 B）。我们保持了一个小规模的承包商团队，因为这有助于与全职从事任务的少数承包商之间进行充分的沟通。然而，这个团队显然不能代表将使用和受到我们部署的模型影响的所有人。举个简单的例子，我们的标注员主要使用英语，而我们的数据几乎完全由英语指令组成。

There are also many ways in which we could improve our data collection set-up. For instance, most comparisons are only labeled by 1 contractor for cost reasons. Having examples labeled multiple times could help identify areas where our contractors disagree, and thus where a single model is unlikely to align to all of them. In cases of disagreement, aligning to the average labeler preference may not be desirable. For example, when generating text that disproportionately affects a minority group, we may want the preferences of labelers belonging to that group to be weighted more heavily.

我们还有许多方法可以改进我们的数据收集设置。例如，由于成本原因，大多数比较只由一个承包商标注。多次标注示例可以帮助确定承包商意见不一致的领域，单一模型很难对齐所有人。在意见不一致的情况下，对标注者偏好的平均进行对齐可能并不理想。例如，在生成对少数群体产生不成比例影响的文本时，我们可能希望属于该群体的标注者的偏好权重更大。

Models. Our models are neither fully aligned nor fully safe; they still generate toxic or biased outputs, make up facts, and generate sexual and violent content without explicit prompting. They can also fail to generate reasonable outputs on some inputs; we show some examples of this in Figure 9.

Perhaps the greatest limitation of our models is that, in most cases, they follow the user’s instruction,even if that could lead to harm in the real world. For example, when given a prompt instructing the models to be maximally biased, InstructGPT generates more toxic outputs than equivalently-sized GPT-3 models. We discuss potential mitigations in the following sections.

模型。我们的模型既没有完全对齐，也不是完全安全的；它们仍然会生成有毒或有偏见的输出，捏造事实，并在没有明确提示的情况下生成性和暴力内容。它们还可能在某些输入上无法生成合理的输出；我们在图 9 中展示了一些示例。

也许我们模型最大的局限是，在大多数情况下，它们会遵循用户的指令，即使这可能会在现实世界中造成伤害。例如，当给出一个指令，要求模型最大程度地偏见时，InstructGPT 生成的有毒输出比同等大小的 GPT-3 模型更多。我们将在接下来的章节中讨论潜在的缓解措施。

5.4 开放问题

This work is a first step towards using alignment techniques to fine-tune language models to follow a wide range of instructions. There are many open questions to explore to further align language model behavior with what people actually want them to do.
这项工作是使用对齐技术来微调语言模型，使其能够遵循各种指令的第一步。还有许多待探讨的问题，以进一步使语言模型的行为与人们实际想要它们执行的任务更加一致。
Many methods could be tried to further decrease the models' propensity to generate toxic, biased,or otherwise harmful outputs. For example, one could use an adversarial set-up where labelers find the worst-case behaviors of the model, which are then labeled and added to the dataset (Dinan et al.,2019b). One could also combine our method with ways of filtering the pretraining data (Ngo et al.,2021), either for training the initial pretrained models, or for the data we use for our pretraining mix approach. Similarly, one could combine our approach with methods that improve models' truthfulness, such as WebGPT (Nakano et al., 2021).
有许多方法可以进一步减少模型生成有毒、有偏见或其他有害输出的倾向。例如，可以使用对抗性设置，让标注员找出模型的最坏行为，然后对其进行标注并添加到数据集中（Dinan 等人，2019b）。我们还可以将我们的方法与过滤预训练数据的方式相结合（Ngo 等人，2021），无论是用于训练初始预训练模型，还是用于我们用于预训练混合方法的数据。类似地，我们的方法也可以与改善模型真实性的方法相结合，例如 WebGPT（Nakano 等人，2021）。
In this work, if the user requests a potentially harmful or dishonest response, we allow our model to generate these outputs. Training our model to be harmless despite user instructions is important, but is also difficult because whether an output is harmful depends on the context in which it' s deployed;for example, it may be beneficial to use language models to generate toxic outputs as part of a data augmentation pipeline. Our techniques can also be applied to making models refuse certain user instructions, and we plan to explore this in subsequent iterations of this research.
在这项工作中，如果用户请求一个潜在有害或不诚实的回应，我们允许我们的模型生成这些输出。尽管用户的指令很重要，但训练我们的模型不会造成伤害也很困难，因为输出是否有害取决于其部署的上下文；例如，使用语言模型生成有毒输出作为数据增强流程的一部分可能是有益的。我们的技术也可以应用于使模型拒绝某些用户指令，我们计划在后续研究中探讨这一点。
Getting models to do what we want is directly related to the steerability and controllability litera-ture (Dathathri et al., 2019; Krause et al., 2020). A promising future path is combining RLHF with other methods of steerability, for example using control codes (Keskar et al., 2019), or modifying the sampling procedure at inference time using a smaller model (Dathathri et al., 2019).
让模型按照我们的意愿行事与可操控性和可控性文献（Dathathri等人，2019年；Krause等人，2020年）直接相关。一个有前途的未来路径是将RLHF与其他可操控性方法相结合，例如使用控制代码（Keskar等人，2019年），或者在推理时使用较小的模型修改采样过程（Dathathri等人，2019年）。

While we mainly focus on RLHF, there are many other algorithms that could be used to train policies on our demonstration and comparison data to get even better results. For example, one could explore expert iteration (Anthony et al., 2017; Silver et al., 2017), or simpler behavior cloning methods that use a subset of the comparison data. One could also try constrained optimization approaches (Achiamet al., 2017) that maximize the score from a reward model conditioned on generating a small number of harmful behaviors.
虽然我们主要关注RLHF，但还有许多其他算法可以用于根据我们的演示和比较数据来训练策略，以获得更好的结果。例如，可以探索专家迭代（Anthony等人，2017年；Silver等人，2017年），或者使用比较数据子集的更简单的行为克隆方法。还可以尝试受限优化方法（Achiam等人，2017年），该方法最大化了基于生成少量有害行为的奖励模型的分数。
Comparisons are also not necessarily the most efficient way of providing an alignment signal. For example, we could have labelers edit model responses to make them better, or generate critiques of model responses in natural language. There is also a vast space of options for designing interfaces for labelers to provide feedback to language models; this is an interesting human-computer interaction problem.
比较也不一定是提供对齐信号的最有效方式。例如，我们可以让标注员编辑模型的回应以使其更好，或者生成对模型回应的批评意见。此外，也可以设计多种用于标注员向语言模型提供反馈的接口，这是一个有趣的人机交互问题。
Our proposal for mitigating the alignment tax, by incorporating pretraining data into RLHF fine-tuning, does not completely mitigate performance regressions, and may make certain undesirable behaviors more likely for some tasks (if these behaviors are present in the pretraining data). This is an interesting area for further research. Another modification that would likely improve our method is to filter the pretraining mix data for toxic content (Ngo et al., 2021), or augment this data with synthetic instructions.
我们的提议是通过将预训练数据纳入RLHF中进行微调，以减轻对齐税，但并未完全消除性能回退，并且可能会使某些任务中的不良行为更加可能（如果这些行为存在于预训练数据中）。这是一个有趣的进一步研究领域。另一个可能改进我们方法的修改是过滤预训练混合数据中的有害内容（Ngo等人，2021年），或者用合成指令增强这些数据。
As discussed in detail in Gabriel (2020), there are subtle differences between aligning to instructions,intentions, revealed preferences, ideal preferences, interests, and values. Gabriel (2020) advocate for a principle-based approach to alignment: in other words, for identifying “fair principles for alignment that receive reflective endorsement despite widespread variation in people’ s moral beliefs.” In our paper we align to the inferred user intention for simplicity, but more research is required in this area.Indeed, one of the biggest open questions is how to design an alignment process that is transparent,that meaningfully represents the people impacted by the technology, and that synthesizes peoples’ values in a

way that achieves broad consensus amongst many groups. We discuss some related considerations in Section 5.2.
正如在Gabriel (2020)中详细讨论的那样，将对齐到指令、意图、显性偏好、理想偏好、兴趣和价值之间存在微妙差异。Gabriel (2020)主张采用基于原则的方法来进行对齐：换句话说，要确定“公平的对齐原则，尽管人们的道德信仰存在广泛变化，但这些原则仍然得到反思认可。”在我们的论文中，我们为了简单起见对齐到了推断出的用户意图，但在这一领域还需要更多的研究。事实上，一个最大的悬而未决的问题是如何设计一个透明的对齐过程，以有意义地代表受到技术影响的人，并以一种能够在许多群体中达成广泛共识的方式综合人们的价值观。我们在第5.2节中讨论了一些相关的考虑。
5.5 广泛的影响
This work is motivated by our aim to increase the positive impact of large language models by training them to do what a given set of humans want them to do. By default, language models optimizethe next word prediction objective, which is only a proxy for what we want these models to do.Our results indicate that our techniques hold promise for making language models more helpful,truthful, and harmless. In the longer term, alignment failures could lead to more severe consequences,particularly if these models are deployed in safety-critical situations. We expect that as model scaling continues, greater care has to be taken to ensure that they are aligned with human intentions (Bostrom,2014).
这项工作的动机是我们的目标是通过训练大型语言模型按照人类的意愿行事来增加其积极影响。默认情况下，语言模型优化下一个单词预测目标，这只是我们希望这些模型做的事情的代理。我们的结果表明，我们的技术有望使语言模型更加有帮助、真实和无害。从长远来看，对齐失败可能导致更严重的后果，特别是如果这些模型在对安全要求很高的情况下部署。我们预计随着模型规模的继续扩大，必须更加小心，以确保它们与人类的意图保持一致（Bostrom，2014年）。
However, making language models better at following user intentions also makes them easier to misuse. It may be easier to use these models to generate convincing misinformation, or hateful or abusive content.
然而，让语言模型更好地遵循用户意图也使其更容易被滥用。例如，使用这些模型生成令人信服的虚假信息、仇恨性或辱骂性内容可能会更加容易。
Alignment techniques are not a panacea for resolving safety issues associated with large language models; rather, they should be used as one tool in a broader safety ecosystem. Aside from intentional misuse, there are many domains where large language models should be deployed only with great care, or not at all. Examples include high-stakes domains such as medical diagnoses, classifying people based on protected characteristics, determining eligibility for credit, employment, or hous-ing, generating political advertisements, and law enforcement. If these models are open-sourced,it becomes challenging to limit harmful applications in these and other domains without proper regulation. On the other hand, if large language model access is restricted to a few organizations with the resources required to train them, this excludes most people from access to cutting-edge ML technology. Another

option is for an organization to own the end-to-end infrastructure of model deployment, and make it accessible via an API. This allows for the implementation of safety protocols like use case restriction (only allowing the model to be used for certain applications), monitoring for misuse and revoking access to those who misuse the system, and rate limiting to prevent the generation of large-scale misinformation. However, this can come at the cost of reduced transparency and increased centralization of power because it requires the API provider to make decisions on where to draw the line on each of these questions.
对齐技术并非解决大型语言模型安全问题的万能药；相反，它们应该作为更广泛安全生态系统中的一种工具来使用。除了有意的滥用外，还有许多领域，大型语言模型在这些领域中应该谨慎部署，或者根本不应该部署。例如，高风险领域包括医学诊断、基于受保护特征对人进行分类、确定信用、就业或住房资格、生成政治广告以及执法。如果这些模型是开源的，那么在这些和其他领域中限制有害应用将变得具有挑战性，而没有适当的监管。另一方面，如果大型语言模型的访问仅限于少数几个具备训练所需资源的组织，那么大多数人将无法接触到尖端的机器学习技术。另一种选择是让一个组织拥有端到端的模型部署基础设施，并通过 API 提供访问。这允许实施安全协议，例如用例限制（仅允许模型用于特定应用程序）、监控滥用并撤销滥用系统的访问权限，以及限制生成大规模错误信息。然而，这可能会以降低透明度和增加权力集中为代价，因为它要求 API 提供者在每个问题上划定界限。
Finally, as discussed in Section 5.2, the question of who these models are aligned to is extremely important, and will significantly affect whether the net impact of these models is positive or negative.
最后，正如在第5.2节中详细讨论的那样，这些模型对齐的对象是谁是非常重要的问题，它将显著影响这些模型的净影响是积极还是消极。
5.2 致谢
First, we would like to thank Lilian Weng, Jason Kwon, Boris Power, Che Chang, Josh Achiam,Steven Adler, Gretchen Krueger, Miles Brundage, Tyna Eloundou, Gillian Hadfield, Irene Soliaman,Christy Dennison, Daniel Ziegler, William Saunders, Beth Barnes, Cathy Yeh, Nick Cammaratta,Jonathan Ward, Matt Knight, Pranav Shyam, Alec Radford, and others at OpenAI for discussions throughout the course of the project that helped shape our research direction. We thank Brian Green,Irina Raicu, Subbu Vincent, Varoon Mathur, Kate Crawford, Su Lin Blodgett, Bertie Vidgen, and PaulRöttger for discussions and feedback on our approach. Finally, we thank Sam Bowman, MatthewRahtz, Ben Mann, Liam Fedus, Helen Ngo, Josh Achiam, Leo Gao, Jared Kaplan, Cathy Yeh, MilesBrundage, Gillian Hadfield, Cooper Raterink, Gretchen Krueger, Tyna Eloundou, Rafal Jakubanis,and Steven Adler for providing feedback on this paper. We’ d also like to thank Owain Evans and Stephanie Lin for pointing out the fact that the automatic TruthfulQA metrics were overstating the gains of our PPO models.
首先，我们要感谢Lilian Weng、Jason Kwon、Boris Power、Che Chang、Josh Achiam、Steven Adler、Gretchen Krueger、Miles Brundage、Tyna Eloundou、Gillian Hadfield、Irene Soliaman、Christy Dennison、Daniel Ziegler、William Saunders、Beth Barnes、Cathy Yeh、

Nick Cammaratta、Jonathan Ward、Matt Knight、Pranav Shyam、Alec Radford以及其他在OpenAI工作的人，他们在项目的整个过程中进行了讨论，帮助塑造了我们的研究方向。我们还要感谢Brian Green、Irina Raicu、Subbu Vincent、Varoon Mathur、Kate Crawford、Su Lin Blodgett、Bertie Vidgen和Paul Röttger对我们的方法进行了讨论和反馈。最后，我们要感谢Sam Bowman、Matthew Rahtz、Ben Mann、Liam Fedus、Helen Ngo、Josh Achiam、Leo Gao、Jared Kaplan、Cathy Yeh、Miles Brundage、Gillian Hadfield、Cooper Raterink、Gretchen Krueger、Tyna Eloundou、Rafal Jakubanis和Steven Adler对本文提供的反馈。我们还要感谢Owain Evans和Stephanie Lin指出了自动TruthfulQA指标夸大了我们的PPO模型效果的事实。
Thanks to those who contributed in various ways to the infrastructure used to train and deploy our models, including: Daniel Ziegler, William Saunders, Brooke Chan, Dave Cummings, Chris Hesse,Shantanu Jain, Michael Petrov, Greg Brockman, Felipe Such, Alethea Power, and the entire OpenAI super computing team. We’d also like to thank Suchir Balaji for help with recalibration, to AlperErcetin and Justin Wang for designing the main diagram in this paper, and to the OpenAI Commsteam for helping with the release, including: Steve Dowling, Hannah Wong, Natalie Summers, andElie Georges.
感谢那些以各种方式为我们训练和部署模型的基础设施做出贡献的人，包括：Daniel Ziegler、William Saunders、Brooke Chan、Dave Cummings、Chris Hesse、Shantanu Jain、Michael Petrov、Greg Brockman、Felipe Such、Alethea Power以及整个OpenAI超级计算团队。我们还要感谢Suchir Balaji在重新校准方面的帮助，Alper Ercetin和Justin Wang在本文中设计主要图表，以及OpenAI Commsteam在发布方面的帮助，包括：Steve Dowling、Hannah Wong、Natalie Summers和Elie Georges.
Finally, we want to thank our labelers, without whom this work would not have been possible:Meave Fryer, Sara Tirmizi, James Carroll, Jian Ouyang, Michelle Brothers, Conor Agnew, JoeKwon, John Morton, Emma Duncan, Delia Randolph, Kaylee Weeks, Alexej Savreux, Siam Ahsan,Rashed Sorwar, Atresha Singh, Muhaiminul Rukshat, Caroline Oliveira, Juan Pablo Castaño Rendón,Atqiya Abida Anjum, Tinashe Mapolisa, Celeste Fejzo, Caio Oleskovicz, Salahuddin Ahmed, Elena Green, Ben Harmelin, Vladan Djordjevic, Victoria Ebbets, Melissa Mejia, Emill Jayson Caypuno,Rachelle Froyalde, Russell M. Bernandez, Jennifer Brillo, Jacob Bryan, Carla Rodriguez, Evgeniya Rabinovich, Morris Stuttard, Rachelle Froyalde, Roxanne Addison, Sarah Nogly, Chait Singh.
最后，我们要感谢我们的标注员，没有他们，这项工作将不可能完成：Meave Fryer、Sara Tirmizi、James Carroll、Jian Ouyang、Michelle Brothers、Conor Agnew、Joe Kwon、John Morton、Emma Duncan、Delia Randolph、Kaylee Weeks、Alexej Savreux、Siam Ahsan、Rashed Sorwar、Atresha Singh、Muhaiminul Rukshat、Caroline Oliveira、Juan Pablo Castaño Rendón、Atqiya Abida Anjum、Tinashe Mapolisa、Celeste Fejzo、Caio Oleskovicz、Salahuddin Ahmed、Elena Green、Ben Harmelin、Vladan Djordjevic、Victoria Ebbets、Melissa Mejia、Emill Jayson Caypuno、Rachelle Froyalde、Russell M. Bernandez、Jennifer

Brillo、Jacob Bryan、Carla Rodriguez、Evgeniya Rabinovich、Morris Stuttard、Rachelle Froyalde、Roxanne Addison、Sarah Nogly和Chait Singh.

参考文献

91篇

附录

若干