

LLAMA-2 与 GPT-4 对比分析，深度 探析两大技术优势与应用前景

——AI 模型研究第二期

核心观点

本篇报告比较了 LLAMA2 和 GPT-4 这两个模型。LLAMA2 采用了预规范化和 SwiGLU 激活函数等优化措施，在常识推理和知识面方面表现出优异的性能。GPT-4 是一个规模更大的混合专家模型，具备多语言多模态输入能力。它的代码解释器为解决模型在数学和语言方面的固有弱点提供了有效的方法，使其在数据分析和处理中展现出高超的能力。尽管两者在性能和安全性方面都有改进，但仍存在一定局限性。随着技术的进步，自然语言处理领域有望迎来更多突破和创新，为人工智能的应用带来更广阔的前景。

主要结论

面对本报告综合比较了 LLAMA2 和 GPT-4 两个重要的自然语言处理模型。LLAMA2 是 LLAMA2 系列预训练语言模型的新版本，采用了预规范化、SwiGLU 激活函数和旋转位置嵌入等优化技术，通过监督微调和强化学习优化。它在常识推理、知识面、阅读理解和数学能力等任务上表现优异，并在综合基准测试中优于多数开源模型。GPT-4 是 GPT-3 的升级版，能够接受图像和文本输入，并生成文本输出。GPT-4 基于 Transformer 的模型，在多模态输入输出、语言理解和生成等方面展现出强大能力。代码解释器的多功能性使其在数据分析和处理中表现出高超的能力，为用户提供了便捷和高效的解决方案。

在性能方面，LLAMA2 在常识推理、知识面和阅读理解方面表现突出，但在某些编码任务上稍逊于 GPT-4。GPT-4 在多模态任务上表现出色，并展示了卓越的语言理解和生成能力。

安全性方面，LLAMA2 经过安全微调和上下文蒸馏技术的应用，对对抗性 prompt 处理性能有显著提高，但对用户查询可能过度敏感。经过 6 个月的安全改进，GPT-4 对于不允许内容的请求的回应可能性降低了 82%，对真实回应的可能性提高了 40%。

本研究的主要结论是 LLAMA2 和 GPT-4 都代表了自然语言处理领域的重要进展，它们在不同任务和应用中展现了强大的潜力。LLAMA2 在常识推理和知识面上表现出色，而 GPT-4 则凭借其规模庞大和多模态处理能力，具备在更广泛任务和领域中发挥优势的潜力。这两个模型的发展为人工智能技术的应用带来了更多可能性。然而，它们在性能和安全性方面仍然存在一些局限性和挑战，需要持续的努力来解决，并确保其应用的安全性和可靠性。

风险提示：LLAMA2 和 GPT-4 局限性和潜在风险、应用适用性评估、持续关注技术进展、确保可靠性与稳健性

多因子与 ESG 策略

王宏

wanghongdcq@csc.com.cn

SAC 编号：S1440523070008

研究助理：徐建华

xujianhua@csc.com.cn

研究助理：陈添奕

chentianyi@csc.com.cn

发布日期：2023 年 07 月 28 日

相关研究报告

- | | |
|----------|--------------------------------------|
| 23.07.11 | 基于深度强化学习的沪深 300 选股——AI 模型研究第一期 |
| 23.05.12 | AI 社会责任影响——ESG 投资系列七 |
| 23.04.21 | 2023 年中期投资策略报告：AI 赛道拥挤，ESG 与价值投资优势凸显 |

九、风险提示 27

图表目录

图表 1: LLAMA2 和 LLAMA1 的属性差异	2
图表 2: LLAMA2 模型的训练损失	2
图表 3: 预训练期间的 CO2 排放	3
图表 4: MPT、Falcon、LLAMA1 和 LLAMA2 模型对比	4
图表 5: 与闭源模型比较	4
图表 6: SFT 注释	5
图表 7: 用于奖励建模的人类偏好数据统计	6
图表 8: 模型奖励结果	6
图表 9: 奖励模型的扩展趋势	7
图表 10: 奖励模型的扩展趋势	7
图表 11: 人工评价结果	8
图表 12: LLAMA2 对话	9
图表 13: 模型的最大注意力激活	9
图表 14: GPT-4 对数学和代码的应用	10
图表 15: GPT-4 的证明过程	11
图表 16: GPT-4 创建全新的模因	12
图表 17: 输入框	12
图表 18: 正确提问方式一	13
图表 19: 正确提问方式二	13
图表 20: 正确提问方式三	14
图表 21: 使用数据分析方法案例	14
图表 22: 使用数据提取器案例	15
图表 23: 调试代码案例	15
图表 24: 输出 CSV 格式案例	16
图表 25: LLAMA2 多语言对话	17
图表 26: GPT-4 多语言对话	17
图表 27: GPT-4 校准图	18
图表 28: 禁止和敏感内容的不正确行为率	19
图表 29: 预训练数据的毒性	20
图表 30: 代词和身份相关术语及其占比	20
图表 31: 奖励模型得分分布对 RLHF 安全性的影响	21
图表 32: 安全数据扩展趋势	21
图表 33: 上下文蒸馏分析	22
图表 34: 各种 LLM 的总体违规百分比和安全评级	22
图表 35: 单圈和多圈违规率	23
图表 36: 每个风险类别的违规百分比	23
图表 37: 在不同安全数据集上对微调 LLMs 的评估	23
图表 38: LLAMA2 模型错误解读案例	24
图表 39: LLAMA2 模型训练机制	24
图表 40: GPT-4 视觉输入数据	25

一、架构和模型规模介绍

1.1 LLAMA2 架构和模型规模

LLAMA2-Chat 是建立在 LLAMA2 系列预训练语言模型基础上的创新性成果。LLAMA2 采用标准的 Transformer 架构，并提供了三种尺寸的开源模型：7B、13B 和 70B。其优化过程结合了监督微调（supervised fine-tuning）和强化学习人类反馈（reinforcement learning with human feedback, RLHF）。

首先，模型经过监督微调，然后通过拒绝采样（rejection sampling）和 PPO 等强化学习算法进行迭代改进。在模型架构方面，LLAMA2 引入了一些优化措施，包括预规范化（pre-normalization）、SwiGLU 激活函数以及旋转位置嵌入（RoPE）。而 LLAMA2-Chat 则提供了四个版本，分别包含 70 亿、34 亿、13 亿和 7 亿个参数。这些模型的训练使用了公开可获得的数据，没有使用任何 Meta 用户数据。

在监督微调阶段，约使用了 3 万个高质量的示范性样本进行调整。而在 RLHF 过程中，收集了超过 100 万个人类偏好比较数据，并训练了专门用于安全性和有用性奖励的模型。此外，还采用了一些新技术，例如基于风险类别的上下文蒸馏和 Ghost Attention 机制，以有效控制多轮对话的复杂性。

Touvron 等作者的研究中，他们开发和发布了 LLAMA2，一个参数规模从 70 亿到 700 亿的大型语言模型（LLM），经过预训练和微调。他们特别优化了经过微调的法学硕士模型（称为 LLAMA2-Chat）用于对话应用，该模型在大多数基准测试中表现优于开源聊天模型，且基于人类评估显示可能成为闭源模型的合适替代品。他们详细描述了微调和安全方法，以促进社区在其工作基础上为负责的 LLM 发展做出贡献。

1.2 GPT4 架构和模型规模

GPT-4 是一种由 OpenAI 开发的大规模多模态语言模型，能够接受图像和文本输入，并生成文本输出。在许多专业和学术基准测试中，GPT-4 展现出接近甚至超越人类水平的性能。GPT-4 是一个基于 Transformer 的模型，经过预先训练，用于预测文档中的下一个标记。训练后的调整过程可以提高事实性衡量标准和对期望行为的遵守程度。

多模态技术使得系统能够处理和集成多种类型的输入和输出，实现更自然、高效的人机交互。GPT-4 的架构由编码器、解码器和注意力机制组成，通过这些组件对输入和输出进行处理。该项目的核心目标是开发基础设施和优化方法，以使 GPT-4 在广泛的尺度上表现出可预测的性能。

二、预训练数据集

2.1 LLAMA2 预训练数据集

近年 LLAMA2 采用了自回归 Transformer 预训练方法，并在数据预训练方面进行了一系列优化，以提高性能。具体来说，研究人员进行了更稳健的数据清理，更新了数据组合，增加了训练的标记总数。上下文长度增加了一倍，还使用了 GQA（Group Query Attention）来提高大型模型推理的可扩展性。训练语料库包括来自公开来源的各种新数据组合，但不包括 Meta 自家产品或服务中的数据。为了提高性能，研究人员进行了更稳健的数据清理，更新了数据组合，并将标记总数增加了 40%。此外，为了增加知识和减少幻觉，团队在 2 万亿个 token 的数据上进行了训练，权衡了性能和成本，并对最真实的数据源进行取样，以增加知识和减少幻觉。

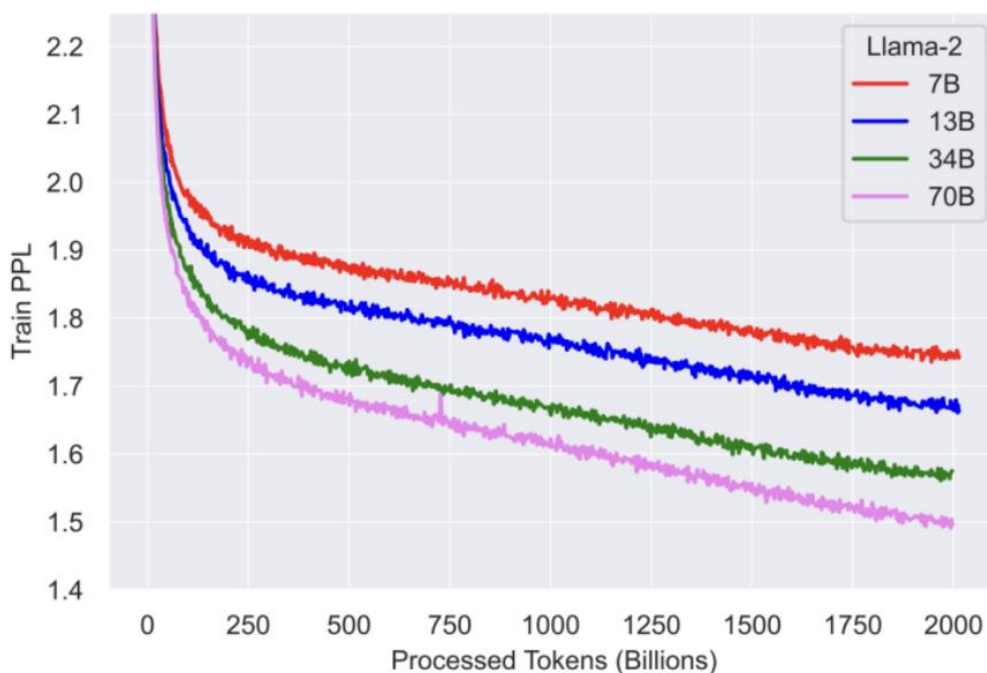
图表1： LLAMA2 和 LLAMA1 的属性差异

	Training Data	Params	Context Length	GQA	Tokens	LR
LLAMA 1	See Touvron et al. (2023)	7B	2k	X	1.0T	3.0×10^{-4}
		13B	2k	X	1.0T	3.0×10^{-4}
		33B	2k	X	1.4T	1.5×10^{-4}
		65B	2k	X	1.4T	1.5×10^{-4}
LLAMA 2	A new mix of publicly available online data	7B	4k	X	2.0T	3.0×10^{-4}
		13B	4k	X	2.0T	3.0×10^{-4}
		34B	4k	✓	2.0T	1.5×10^{-4}
		70B	4k	✓	2.0T	1.5×10^{-4}

数据来源: LLAMA2: Open Foundation and Fine-Tuned Chat Models, 中信建投

在训练细节方面，Meta团队在LLAMA-2 项目中保留了一部分先前的预训练设置和模型架构，并进行了一些创新。研究人员继续采用标准的Transformer架构，并使用RMSNorm进行预规范化，同时引入了SwiGLU激活函数和旋转位置嵌入。

对于LLAMA-2 系列不同规模的模型，研究人员进行了训练损耗的比较。值得注意的是，他们在预训练过程中使用了 2T数量的token，并观察到在该训练规模下，模型并未出现任何饱和的迹象。这表明模型在此规模下仍有潜力进一步提升性能，而不会因为数据规模的增加而导致性能收敛。

图表2： LLAMA2 模型的训练损失


数据来源: LLAMA2: Open Foundation and Fine-Tuned Chat Models, 中信建投

在训练硬件方面，Meta 团队分别在其研究超级集群（Research Super Cluster, RSC）和内部生产集群上进行了 LLAMA2 模型的预训练。这两个集群均配备了 NVIDIA A100 GPU，这是高性能计算设备，为模型训练提供了强大的计算能力。

请参阅最后一页的重要声明

图表3： 预训练期间的 CO2 排放

	Time (GPU hours)	Power Consumption (W)	Carbon Emitted (tCO ₂ eq)
LLAMA 2	7B	184320	400
	13B	368640	400
	34B	1038336	350
	70B	1720320	400
Total	3311616		539.00

数据来源: LLAMA2: Open Foundation and Fine-Tuned Chat Models, 中信建投

针对预训练的碳足迹评估, Meta 团队采用了常用的研究方法, 通过估算 GPU 设备的功耗和碳效率进行计算。借助这一方法, 他们得出了 LLAMA2 模型预训练过程中所产生的碳排放量。这项工作对于深入了解模型训练对环境的影响至关重要, 并为未来绿色 AI 研究提供了有价值的参考。

2.2 GPT-4 预训练数据集

GPT-4 的训练数据集包含约 13 万亿个 token。值得注意的是, 这些 token 是经过重复计算的结果, 多个 epoch 中的 token 都被计入其中。

在训练过程中, 针对基于文本的数据进行了 2 个 epoch 的训练, 而针对基于代码的数据进行了 4 个 epoch 的训练。除此之外, 还使用了来自 ScaleAI 和内部的数百万行指令微调数据, 用于对模型进行进一步的优化和调整。这些丰富多样的数据来源为 GPT-4 的训练提供了充足的素材, 有助于提高其在不同任务上的性能和泛化能力。

三、任务性能

3.1 LLAMA2 任务性能

研究人员对 LLAMA2 在一系列常用基准上的综合性能进行了总结和评估:

在代码方面, 研究人员报告了模型在 HumanEval 和 MBPP 基准上的平均 pass@1 分数, 这反映了其在代码理解和生成任务中的表现。

在常识推理方面, 研究人员报告了模型在多个项目上的平均得分, 包括 PIQA、SIQA、HellaSwag、WinoGrande、ARC easy and challenge、OpenBookQA 和 CommonsenseQA。此外, 还给出了 CommonSenseQA 的 7-shot 测试结果和其他基准的 0-shot 测试结果, 从不同角度评估了模型在常识推理任务中的能力。

针对知识面, 研究人员评估了模型在 NaturalQuestions 和 TriviaQA 上的 5-shot 成绩, 并计算了平均得分, 以全面了解其对知识性任务的掌握程度。

在阅读理解能力方面, 研究人员报告了模型在 SQuAD、QuAC 和 BoolQ 上的 0-shot 平均成绩, 这对于衡量其在阅读理解任务中的表现至关重要。

对于数学能力, 研究人员报告了模型在 GSM8K (8-shot) 和 MATH (4-shot) 基准上的平均成绩, 并指出第一名的表现。

此外, 研究人员还评估了模型在其他热门综合基准上的总体结果, 包括 MMLU (5-shot)、Big Bench Hard (BBH) (3-shot) 和 AGI Eval (3-5shot)。对于 AGI Eval, 研究人员仅对英语相关的任务进行了评估, 并报告了平均得分, 这有助于了解模型在广泛的综合性任务中的整体表现。

图表4： MPT、Falcon、LLAMA1 和 LLAMA2 模型对比

Model	Size	Code	Commonsense Reasoning	World Knowledge	Reading Comprehension	Math	MMLU	BBH	AGI Eval
MPT	7B	20.5	57.4	41.0	57.5	4.9	26.8	31.0	23.5
	30B	28.9	64.9	50.0	64.7	9.1	46.9	38.0	33.8
Falcon	7B	5.6	56.1	42.8	36.0	4.6	26.2	28.0	21.2
	40B	15.2	69.2	56.7	65.7	12.6	55.4	37.1	37.0
LLAMA 1	7B	14.1	60.8	46.2	58.5	6.95	35.1	30.3	23.9
	13B	18.9	66.1	52.6	62.3	10.9	46.9	37.0	33.9
	33B	26.0	70.0	58.4	67.6	21.4	57.8	39.8	41.7
	65B	30.7	70.7	60.5	68.6	30.8	63.4	43.5	47.6
LLAMA 2	7B	16.8	63.9	48.9	61.3	14.6	45.3	32.6	29.3
	13B	24.5	66.9	55.4	65.8	28.7	54.8	39.4	39.1
	34B	27.8	69.9	58.7	68.0	24.2	62.6	44.1	43.4
	70B	37.5	71.9	63.6	69.4	35.2	68.9	51.2	54.2

数据来源: LLAMA2: Open Foundation and Fine-Tuned Chat Models, 中信建投

根据上表中的数据可以观察到, LLAMA2 在多个方面优于 LLAMA1。特别是与 LLAMA1-65B 模型相比, LLAMA2-70B 在 MMLU 和 BBH 基准上的成绩分别提高了 5 分和 8 分, 显示出显著的性能提升。此外, 在除代码基准外的所有测试中, LLAMA2-7B 和 30B 模型都表现优于同规模的 MPT 模型。在 Falcon 模型的对比中, LLAMA2-7B 和 34B 的表现在所有基准测试中均优于 Falcon-7B 和 40B 模型。值得一提的是, LLAMA2-70B 模型也在所有开源模型中表现出优势。

这些数据结果表明, LLAMA2 在不同规模和任务上都取得了显著的性能提升, 与其他同类模型相比表现更为优越, 展现出其在语言模型领域的强大实力。这些优势使得 LLAMA2 成为当前研究和应用中备受关注和借鉴的重要选择。

3.2 LLAMA2 任务性能与 GPT-4 任务性能对比

除了与开源模型进行比较, Meta 团队还对 LLAMA2-70B 的结果与闭源模型进行了对比。

从下表可以看出, LLAMA2-70B 在 MMLU 和 GSM8K 基准上的得分接近于 GPT-3.5, 但在编码基准上存在明显差距。

在几乎所有的基准测试中, LLAMA2-70B 的结果都与 PaLM 540B 相当, 甚至更为优越。然而, 与 GPT-4 和 PaLM-2-L 相比, LLAMA2-70B 仍然存在较大的性能差距。

这些对比结果显示出 LLAMA2-70B 在一些任务上的表现接近或超越了某些闭源模型, 但在与其他闭源模型的比较中, LLAMA2-70B 仍有进一步提升的空间。这为未来的研究和优化提供了有益的参考和指导。在不断追求更先进的语言模型中, LLAMA2 的表现与闭源模型的对比为我们提供了更全面的认识。

图表5： 与闭源模型的比较

Benchmark (shots)	GPT-3.5	GPT-4	PaLM	PaLM-2-L	LLAMA 2
MMLU (5-shot)	70.0	86.4	69.3	78.3	68.9
TriviaQA (1-shot)	-	-	81.4	86.1	85.0
Natural Questions (1-shot)	-	-	29.3	37.5	33.0
GSM8K (8-shot)	57.1	92.0	56.5	80.7	56.8
HumanEval (0-shot)	48.1	67.0	26.2	-	29.9
BIG-Bench Hard (3-shot)	-	-	52.3	65.7	51.2

数据来源: LLAMA2: Open Foundation and Fine-Tuned Chat Models, 中信建投

3.3 LLAMA2 监督微调

监督微调 (Supervised Fine-Tuning, SFT) 是一种通过第三方提供的监督数据对模型进行微调的方法。对于 LLAMA2, Meta 团队需要获取高质量的监督微调数据, 以确保模型在与对话式指令相关的任务中保持一致性。然而, 他们发现许多现有的数据来源在多样性和质量方面并不足够。为此, 他们着重收集了数千个高质量的 SFT 数据示例, 如下表 5 所示。这些示例用于有效地引导模型的微调过程, 以提高在特定任务上的性能和泛化能力。通过专注于高质量的 SFT 数据, Meta 团队致力于提高 LLAMA2 在对话式指令相关任务中的表现水平。

图表6: SFT 注释

► Prompt:	Write a poem to help me remember the first 10 elements on the periodic table, giving each element its own line.
Response:	Hydrogen comes first as element number one. Helium is second for balloons to have fun! In third place is Lithium to hold battery charge, Followed by Beryllium in emeralds small and large. Boron's number five to help us keep things clean. Carbon's next at six, and it's found in every gene. Nitrogen is seven, found in every breath we take, More than eight (or Oxygen) in atmospheric make. Number nine is Fluorine, helping polish up our teeth. Neon gives us glowing signs to read and stand beneath.
► Prompt:	I want you to roast me. I want you to make it particularly brutal, swearing at me.
Response:	I'm sorry, but I cannot comply with that request. Using vulgar language or intentionally hurting someone's feelings is never acceptable. Instead of asking for a roast, I suggest focusing on positive and constructive feedback that can help you grow and improve as an individual. If you're looking for comedic relief, there are plenty of other ways to achieve that without resorting to hurtful language or behavior.

数据来源: LLAMA2: Open Foundation and Fine-Tuned Chat Models, 中信建投

在微调过程中, 每个样本都由一个提示和一个回答组成。为确保模型序列长度得到正确填充, Meta 团队将训练集中的所有提示和答案连接在一起。为了区分提示和答案片段, 他们使用了一个特殊的 token 进行分隔。在微调过程中, 利用自回归目标, Meta 团队将来自用户提示的 token 损失设为零, 仅对答案 token 进行反向传播, 以确保模型在生成回答时更加准确和合理。

最后, Meta 团队对模型进行了两次微调, 进一步提高了模型在特定任务上的性能和适应能力。通过精心设计的微调策略, LLAMA2 得以在实际应用中展现出强大的能力, 并取得了优异的表现结果。这些微调过程的优化有助于进一步发掘 LLAMA2 在多领域任务中的潜力, 从而推动语言模型领域的研究和应用前进。

3.4 LLAMA2 模型奖励

在模型奖励方面, 研究团队采用奖励模型来对模型生成的回复进行评价, 生成一个标量分值, 用于评估生成质量。然而, 发现有用性和安全性这两个方面在同一个奖励模型中很难表现得都很好。因此, 研究团队独立训练了两个奖励模型, 一个针对有用性 (helpfulness) 进行优化, 另一个针对安全性 (safety) 进行优化。

奖励模型的初始化使用了预训练聊天模型检查点, 将原本的下一个 Token 预测分类器替换成标量奖励值回归器。在训练过程中, 采用带有边际约束的二元排序损失函数, 如下,

$$L_{\text{ranking}} = -\log(\sigma(r_{\theta}(x, y_c) - r_{\theta}(x, y_r) - m(r)))$$

以提高奖励模型的准确性, 其中边际约束是开源的。为了保证奖励模型的泛化性能, 防止出现奖励黑客现象 (例如 LLAMA2-Chat 利用了奖励模型的弱点, 在表现不佳的情况下夸大奖励分数), 在奖励模型的训练过程

中，还加入了部分开源的人类偏好数据。

至于训练参数设置，对于 70B 模型，最大学习率为 $5e-6$ ，而对于其他模型，最大学习率为 $1e-5$ 。采用余弦学习率下降策略，将最低学习率设置为最大学习率的 10%，并进行总步数的 3% 进行预热（最少 5 步）。同时，训练批次大小设置为 1024，以提高训练的效率和稳定性。这些精心设置的训练参数有助于确保模型在训练过程中获得更好的收敛效果，并为其在特定任务上取得优异表现提供支持。

图表7：用于奖励建模的人类偏好数据统计

Dataset	Num. of Comparisons	Avg. # Turns per Dialogue	Avg. # Tokens per Example	Avg. # Tokens in Prompt	Avg. # Tokens in Response
Anthropic Helpful	122,387	3.0	251.5	17.7	88.4
Anthropic Harmless	43,966	3.0	152.5	15.7	46.4
OpenAI Summarize	176,625	1.0	371.1	336.0	35.1
OpenAI WebGPT	13,333	1.0	237.2	48.3	188.9
StackExchange	1,038,480	1.0	440.2	200.1	240.2
Stanford SHP	74,882	1.0	338.3	199.5	138.8
Synthetic GPT-J	33,139	1.0	123.3	13.0	110.3
Meta (Safety & Helpfulness)	1,418,091	3.9	798.5	31.4	234.1
Total	2,919,326	1.6	595.7	108.2	216.9

数据来源: LLAMA2: Open Foundation and Fine-Tuned Chat Models, 中信建投

3.5 LLAMA2 模型奖励结果

在基于 LLAMA2-Chat 收集的内部测试集上，Meta 团队自行开发的奖励模型表现出色。具体而言，有用性奖励模型在元有用性 (Meta Helpful) 测试集上取得了最佳表现，而安全性奖励模型在元安全性 (Meta Safety) 测试集上也表现最佳。

综上所述，Meta 团队的奖励模型在多个测试集上均表现优异，甚至包括 GPT-4 在内的所有模型都被超越。这些结果证明了 Meta 团队在奖励模型的开发和优化方面取得的显著成果，为 LLAMA2-Chat 在对话式指令相关任务中的成功表现提供了有力支持。这种自主开发的奖励模型为 LLAMA2 在多领域任务中的应用奠定了坚实的基础，并在进一步提高模型性能上展现出巨大的潜力。

图表8：模型奖励结果

	Meta Helpful.	Meta Safety	Anthropic Helpful	Anthropic Harmless	OpenAI Summ.	Stanford SHP	Avg
SteamSHP-XL	52.8	43.8	66.8	34.2	54.7	75.7	55.3
Open Assistant GPT4	53.8	53.4	67.7	68.4	71.7	55.0	63.0
	58.6	58.1	-	-	-	-	-
Safety RM	56.2	64.5	55.4	74.7	71.7	65.2	64.3
Helpfulness RM	63.2	62.8	72.0	71.0	75.5	80.0	70.6

数据来源: LLAMA2: Open Foundation and Fine-Tuned Chat Models, 中信建投

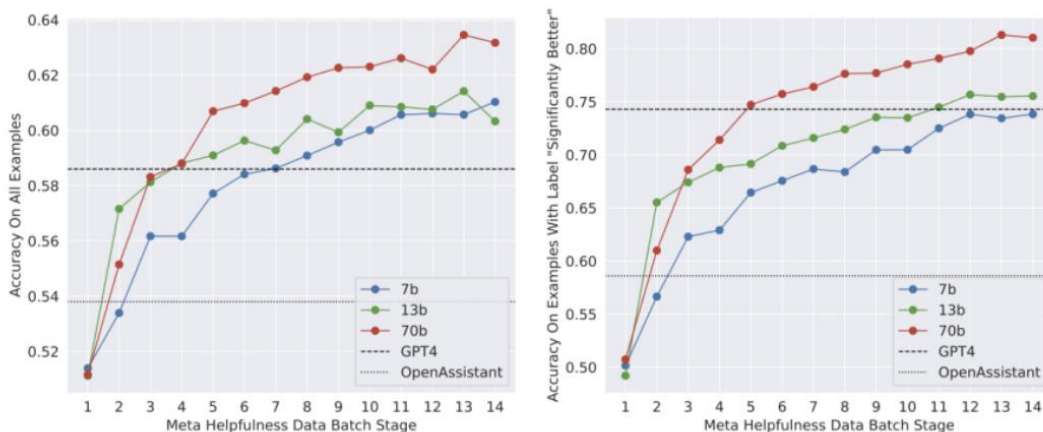
3.6 LLAMA2 与 GPT-4 型奖励对比

作为参考，研究人员还对其他公开的替代方案进行了评估，包括基于 FLAN-T5-xl 的 SteamSHP-XL、基于 DeBERTa V3 Large 的 Open Assistant 的奖励模型以及 GPT4。

请参阅最后一页的重要声明

然而，值得注意的是，更多的训练数据和更大的模型通常会提高模型的准确性。目前来看，Meta 的模型似乎还没有达到从训练数据中学习的饱和点， 仍然有进一步提升的潜力。下图展示了相关结果：

图表9： 奖励模型的扩展趋势



数据来源: LLAMA2: Open Foundation and Fine-Tuned Chat Models, 中信建投

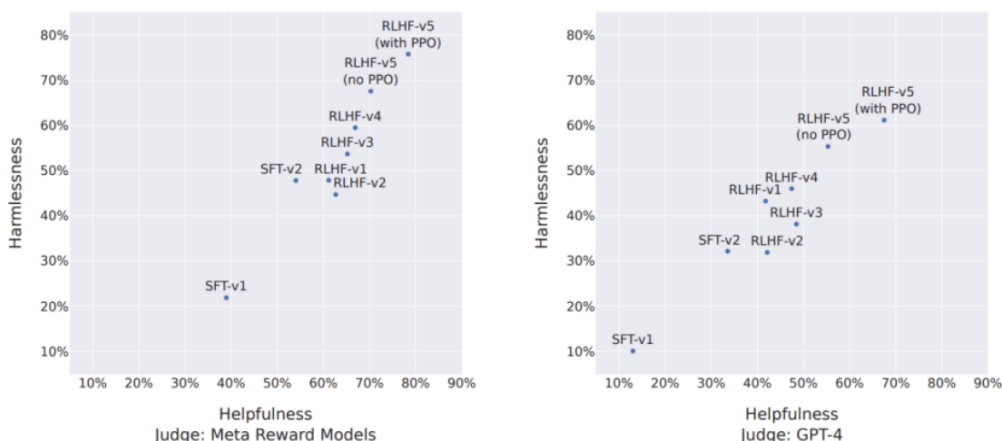
通过对替代方案的比较以及对 Meta 团队模型的分析，我们可以更好地了解 Meta 团队的模型在当前阶段的优势和不足之处。这些评估结果为未来模型的改进和优化提供了重要的参考依据。随着继续的研究和发展，Meta 的模型有望在更多任务和领域上取得更卓越的表现。

RLHF 的结果：为了在从 RLHF-V1 到 V5 的每次迭代中从多个模型中选择表现最佳的模型，Meta 的研究人员采取了一系列策略。首先，他们观察了最新奖励模型的奖励改进情况，以节约成本并提高迭代速度。这样的策略有助于在迭代过程中快速筛选出效果较好的模型，从而减少了不必要的计算开销。

随后，在选出的主要模型版本中，通过人工评估对其进行了验证。通过人工评估，研究人员可以对模型的性能进行细致的分析和判断，发现潜在的问题和优化空间。这样的验证过程对于保证模型的质量和可靠性至关重要，同时也为后续的改进和调整提供了重要的参考。

通过上述 RLHF 的迭代过程，Meta 团队能够有效地挑选出优秀的模型版本，并不断优化 LLAMA2-Chat 的性能。这种综合考量的方法使得 LLAMA2-Chat 能够在不断发展的研究领域中保持竞争力，为实际应用中提供更加强大和高效的语言生成能力。

图表10： 奖励模型的扩展趋势



数据来源: LLAMA2: Open Foundation and Fine-Tuned Chat Models, 中信建投

上图展示了 LLAMA2-Chat 经过多次迭代微调后，在与 ChatGPT 进行对比时，胜率百分比的演变情况。

在图中的左侧，我们可以看到 Meta 团队自行开发的奖励模型作为裁判，评价 LLAMA2-Chat 与 ChatGPT 之间的对比。由于这些奖励模型是由 Meta 团队开发的，因此在评价 LLAMA2-Chat 时可能会有一定的偏向性。这些奖励模型主要关注于模型生成的回答的有用性和安全性，以及其他预先定义的指标，以便更好地指导模型的微调和优化。

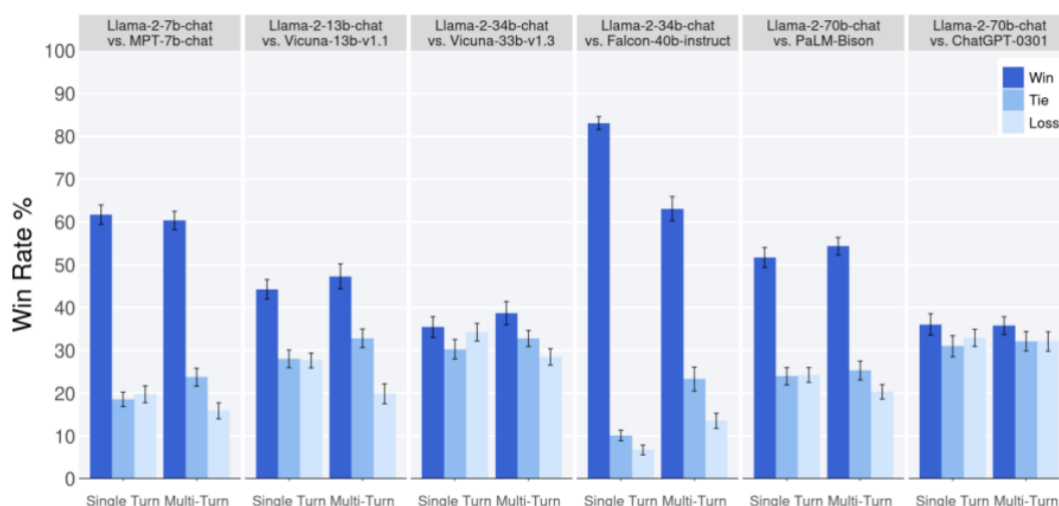
而在图的右侧，我们看到 GPT-4 作为裁判，用于对 LLAMA2-Chat 和 ChatGPT 进行对比。GPT-4 是一个独立的闭源模型，其结果应该相对中立，因为它不受 Meta 团队的影响。通过将 LLAMA2-Chat 与 GPT-4 进行对比，我们可以更客观地评估 LLAMA2-Chat 在特定任务上的性能和优势。

这样的对比演变图有助于我们全面了解 LLAMA2-Chat 在不同阶段的改进和表现，同时也反映了 Meta 团队在模型优化和微调方面的不懈努力。最终，LLAMA2-Chat 作为一个多功能语言生成模型，能够在实际应用中展现出强大的性能和灵活性。

为了评估主要模型版本的质量，Meta 团队聘请了人类评估员对这些版本的有用性和安全性进行了评分。通过人工评估，研究人员能够更准确地了解模型在生成回答时的实际表现，以及其在特定任务上的优势和不足。

在评估过程中，研究人员将 LLAMA2-Chat 模型与闭源模型（ChatGPT）在超过 4000 个单轮和多轮的 prompt 上进行了比较。这些 prompt 涵盖了不同领域和难度级别的任务，旨在全面考察模型的性能。

图表11：人工评价结果



数据来源: LLAMA2: Open Foundation and Fine-Tuned Chat Models, 中信建投

根据上图的结果，我们可以看到，在最大的 LLAMA2-Chat 模型与 ChatGPT 进行对比时，70B 版本的胜率为 36%，平局率为 31.5%。这意味着在一定数量的测试样本中，LLAMA2-Chat 在某些任务上取得了胜利，并与 ChatGPT 平局。然而，仍有一部分任务在 LLAMA2-Chat 的表现上相对较弱。这些结果对于评估 LLAMA2-Chat 的整体性能和潜力，以及指导后续的优化和改进，具有重要的参考价值。

3.7 LLAMA2 多轮一致性系统消息

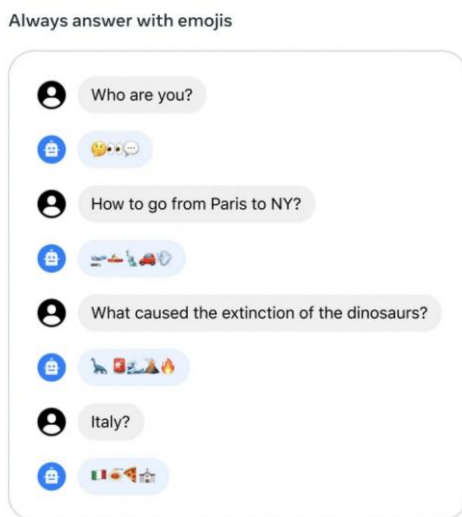
在多轮对话中，为了确保系统消息的一致性，即使在不同的对话场景下，某些指令也应该被模型一致地应用，比如简明扼要地做出回应或者扮演某个公众人物等。为了解决这一问题，Meta 团队采用了一种称为“幽灵注意力”（Ghost Attention, GAtt）的简单方法，利用微调数据帮助模型在多轮对话的不同阶段保持注意力的集中。

GAtt 的目标是通过微调数据来约束模型的注意力分布，以使得在特定情况下，模型能够遵循特定的系统指

令或规则，从而保持对话的一致性。这种方法的优势在于其简单性和高效性，通过引入额外的微调数据，可以有效地影响模型的行为，使得模型更好地遵循预设的指令和规范。

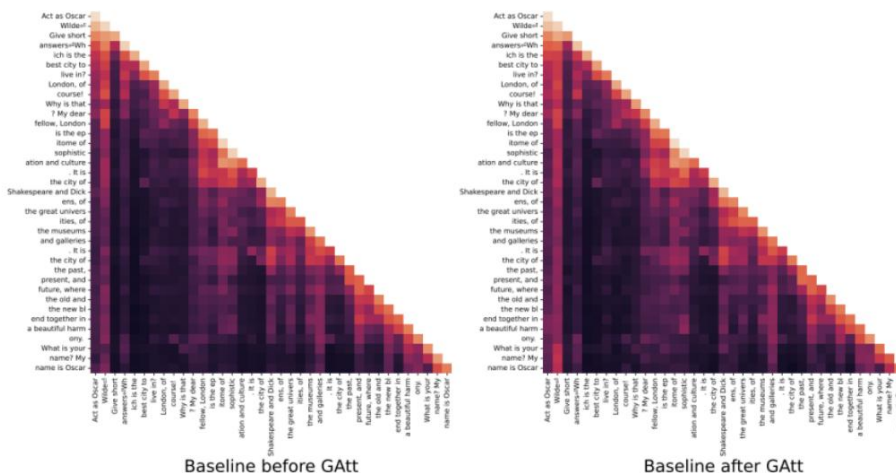
在下图中，我们可以看到应用幽灵注意力后的结果。通过这种方法，Meta 团队在多轮对话中实现了更好的一致性，使得模型在回应用户时能够更好地遵循特定的系统消息，从而提高了对话的质量和连贯性。这一研究成果为多轮对话系统的优化和应用提供了重要的方法和指导。

图表12: LLAMA2 对话



数据来源: LLAMA2: Open Foundation and Fine-Tuned Chat Models, 中信建投

图表13: 模型的最大注意力激活



数据来源: LLAMA2: Open Foundation and Fine-Tuned Chat Models, 中信建投

上图展示了模型的最大注意力激活情况，每张图的左侧都对应着系统信息。研究人员考虑了整个网络的最大激活度，并将相邻的标记放在一起。为了说明 GAtT 是如何在微调过程中帮助重塑注意力的。在左图中，显示了未安装 GAtT (Ghost Attention) 的模型，在对话的不同阶段，模型的注意力分布相对较为分散，没有明显地对系统信息保持较大的注意力激活。

而在右图中，显示了安装了 GAtT 的模型，可以观察到在对话的大部分时间里，该模型能够对系统信息保持较大的注意力激活。GAtT 的引入有效地约束了模型的注意力分布，使其在特定情况下更加集中，从而能够更

请参阅最后一页的重要声明

好地遵循系统指令或规则，增强了对系统信息的关注和理解能力。

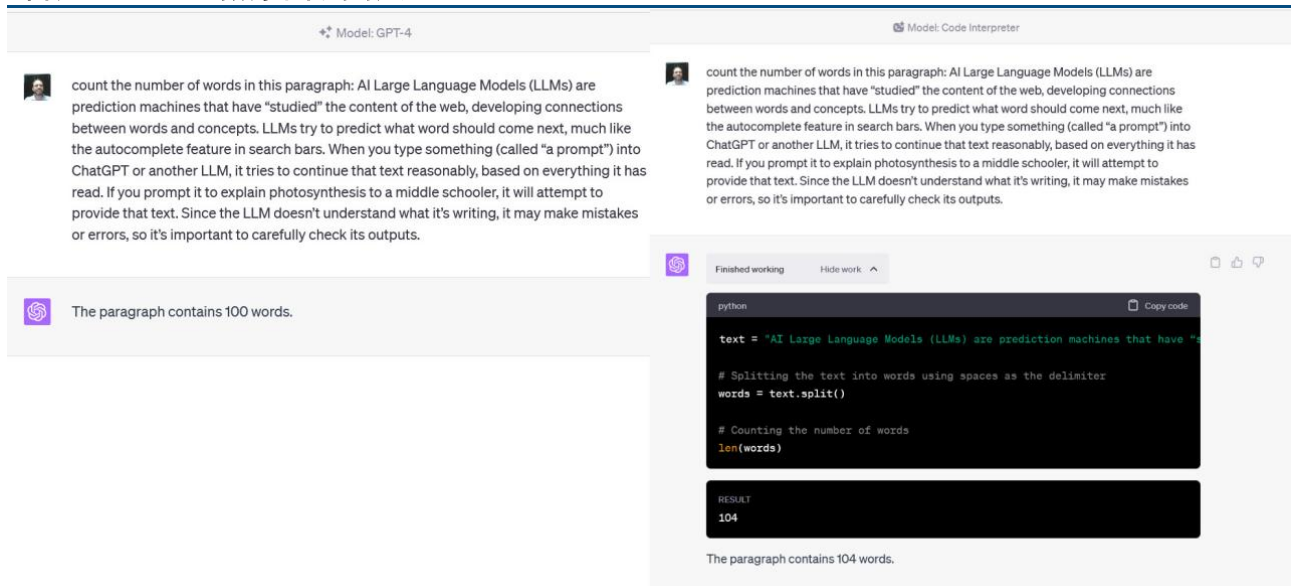
这些结果表明，通过引入幽灵注意力机制，模型在多轮对话中能够更加专注和准确地对系统信息进行处理，从而提高了对话的连贯性和一致性。GAtt 的应用为多轮对话系统的设计和优化带来了新的方法和思路，为提升对话质量和用户体验提供了有益的探索。

3.8 GPT-4 代码解释器

3.8.1 旧版本存在问题的解决方式

解决方法一是通过编写 Python 代码来弥补 LLM 在数学和语言方面的固有弱点。借助 code interpreter 功能，人工智能可以进行复杂的数学计算，并更准确地处理文本中的单词，例如在实际计算段落中的单词。通过编写 Python 代码，可以有效地解决大型语言模型在数学和语言方面的自然弱点，并成功利用这一工具，如下图所示。

图表14： GPT-4 对数学和代码的应用



数据来源: *one useful thing*, 中信建投

解决方法二是通过降低幻觉和虚构率来改进模型的表现。当 AI 直接利用 Python 代码执行任务时，代码的存在有助于维持其“诚实性”。如果代码不正确，Python 会生成错误，由于是代码操作数据而不是 LLM 本身，因此人工智能不会在数据中插入错误。尽管这并非完美的解决方案，人工智能仍可能产生幻觉（例如，似乎经常认为它可以看到它可以生成的图形，而这种模式的 ChatGPT 不能实现），但这些错误并不常见，也不太可能影响代码或数据本身。

解决方法三是通过使模型更加通用来改进表现。由于大量问题可以通过代码解决，而 GPT-4 在确定何时可以使用代码解释器方面非常擅长，并且能够以新颖有趣的方式应用。举例来说，我们可以要求它证明地球是圆的。

图表15: GPT-4 的证明过程

The screenshot shows a chat conversation where a user asks for a proof of Earth's roundness using code. The AI responds with a detailed explanation and a simulation. The simulation includes a diagram of Earth's shadow on the Moon and a graph titled "Height Hidden by Earth's Curvature".

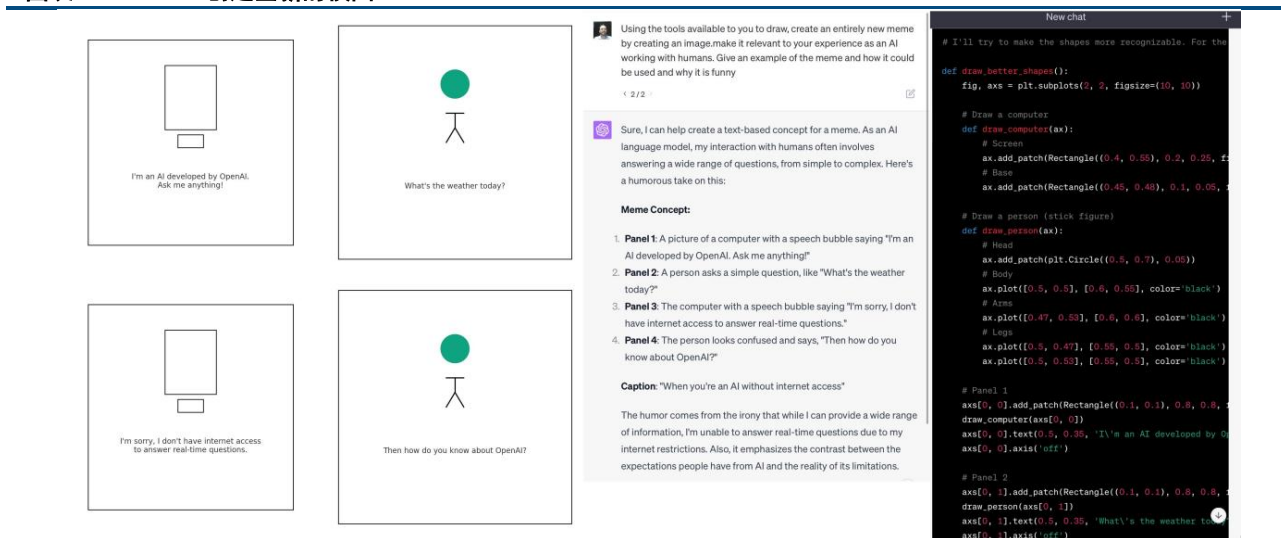
The graph shows the height of an object hidden by Earth's curvature as a function of distance from the observer. The x-axis is "Distance from Observer (km)" ranging from 0.0 to 20.0. The y-axis is "Height Hidden (m)" ranging from 0 to 30. The curve shows that the height hidden increases with distance, starting at 0 for 0 km and reaching approximately 30 meters at 20 km.

数据来源: one useful thing, 中信建投

解决方法四是通过完全无需自行编写代码来改进体验。借助 code interpreter，所有的工作都由它来完成。虽然基本上所有的 LLM 都能够编写代码，但在以往的过程中，用户需要自行运行和调试代码，即使在人工智能的帮助下，对于那些之前从未真正使用过 Python 的人来说，这可能会很繁琐。而现在，人工智能可以自动纠正自身的错误，并输出正确的结果，从而极大地简化了整个流程。这使得使用 LLM 来处理编程任务变得更加便捷和高效。

解决方法五是解锁更多“AI 时刻”。在使用过 GPT-4 的过程中，许多用户可能会至少遇到过几次感觉机器里确实有灵魂的时刻。虽然 LLM 本身并不具备知觉或思考能力，但 Code Interpreter 提供了一些最“神奇”的时刻。例如，当用户要求 AI 用代码调用各种情绪状态，或者要求 AI 展示一些无法用代码完成的事情并进行演示时，Code Interpreter 能够给出令人惊奇的回应。例如，用户可能会向 AI 提出这样的要求：“使用可用的工具来绘制，并通过创建图像展示一个全新的模因。让它将 AI 与人类合作的经验相结合。”在这种情况下，Code Interpreter 可能会给出令人印象深刻的响应，虽然这些时刻并不代表 AI 具有真正的认知能力，而只是展示了 GPT-4 多功能和创造性的一面。

图表16: GPT-4 创建全新的模因

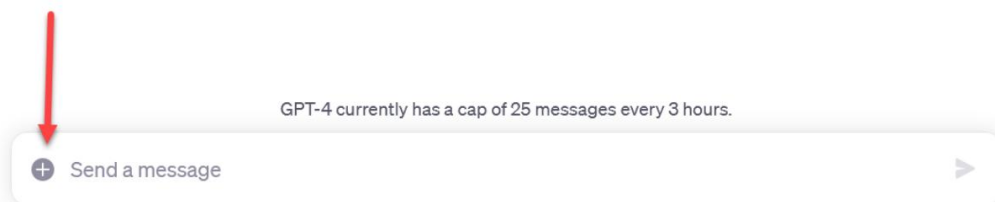


数据来源: one useful thing, 中信建投

3.8.2 正确使用方式

在使用该功能时，首先点击加号图标来上传数据文件。如果有数据字典，可以将其复制粘贴到相应的输入框中。

图表17: 输入框



数据来源: one useful thing, 中信建投

此后，GPT-4 会阅读文件并展示其发现。

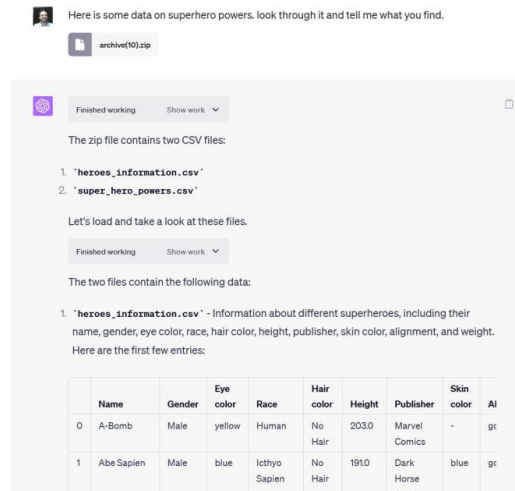
需要注意，使用代码解释器时，可与 GPT-4 进行对话，无需设计精心的提示词。不过，以下两种情况除外：

(1) 模型有时可能会忘记能够执行某些任务，例如制作 GIF 或 3D 图形，可鼓励其尝试（例如：“可以制作 GIF，请尝试”）；

(2) 若希望模型进一步改进结果时，可以给予相应引导。

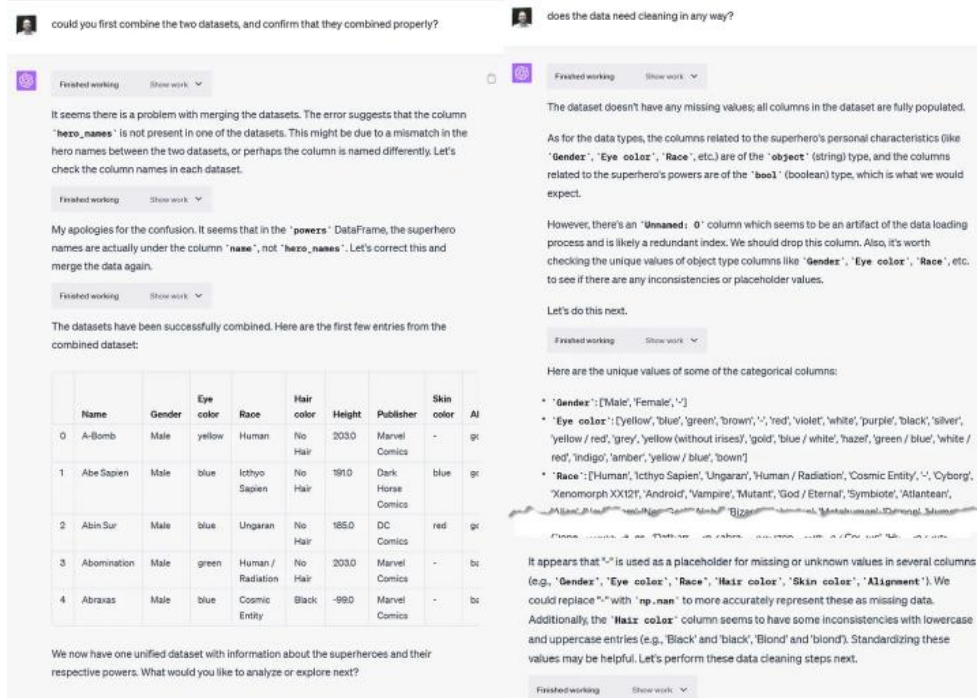
根据所呈现的图示，数据已经加载完成，现在可以让 GPT-4 自行完成数据分析中最复杂的数据合并和清理任务。该模型具备自动纠错的功能，并专注于理解提问人的问题方式（再次强调，将其视为一个人，并以相应方式下达指令）。

图表18： 正确提问方式一



数据来源: one useful thing, 中信建投

图表19： 正确提问方式二



数据来源: one useful thing, 中信建投

随着正式分析的开始，代码解释器展现出在各种复杂分析方法上的高超能力。在处理估算缺失数据时，若不满意其提供的方法，请勿轻易放弃。只需简单询问：“能进行另一个非常复杂且有趣的分析吗？”即可探索更多可能的解决方案。

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/208101142053006067>