目 录

一场《草莓”-o1 模型小型讨论会》纪要

COPU 组织、杨耀东研究员主讲

2024.09.24

9 月 24 日， COPU 举行《“草莓”-o1 模型小型讨论会》，邀请北京智源研究院杨耀东研究员参加讨论并作主要发言，以及接受回答。会前，COPU 准备了一批提问如下:

1)讨论 o1 模型的概念，它有哪些主要特点?

2)为什么说强化学习是 o1 模型的技术基础?

3)如何理解 o1 模型全面超越了 GPT-4o 或刷新了 SOTA?o1 模型能否减少生成式语言模型天生的缺陷?

4)如何理解 o1 模型具有超强的推理能力?

5)OpenAI o1 模型在“后训练”扩充律 post-Training Scaling Laws 下，如何提升推理能力和长程问题能力?

6)COPU 提出各智能体具备推理能力的分级情况，希望在讨论会上进行鉴别

推理能力分级:

具有 1 级推理能力·chat bots，语言交流，

具有 2 级推理能力·Reasoners，个人问答; o1 模型

具有 3 级推理能力·LLMOS，大模型操作系统

具有 4 级推理能力·Agents，能够行动的智能体

具有 5 级推理能力·AGI，通用人工智能

在讨论会上重点谈到下列问题：1)语料数据问题在语料数据搜集中，开始人们选择日常使用的数据，随着数据量需求的增加，发展到选择行业数据、互联网海量数据，人工智能的发展使上述数据来源已不敷需要，开始创造合成数据，合成数据虽然能满足语料数据量增长的需求，但也出现了数据污染的问题。生成式语言大模型产生缺陷，与机器依赖于统计技术有关，也与语料数据的污染有关。2)“后训练”时代已经到来过去我们对语言大模型抓预训练，直到对齐，由Open AI开发的o1模型开启了“后训练”(以增加推理能力)。据北京大学对齐团队独家解释：新的扩展律post-Training已经出现，后训练时代已经到来。 强化学习成为o1模型的技术基础。)o1的技术基础，针对后训练，在学习与搜索选择中选择学习，强化学习成为o1模型的技术基础。在思考链中，GPT-4属于快思考(选择搜索),o1属于慢思考(因为推理)。o1模型在哪些地方超越GPT-4o?①推理占先的性能，o1表现优秀(或者说，o1整型在复杂推理、数学和代码问题上，提升到一个全新高度，优于LLM的水平）在数学代码、竞争性编程、数学奥林匹克竞赛、物理/生物/化学博士考试等推理占先的性能方面，o1优于GPT-4o②解决语言大模型存在的缺陷问题上，o1优于GPT-4o总的来说，o1推理能力强，通用能力弱；o1与GPT-4o比,其写作能力并未提高，指令跟踪也未超越。

在会上，对在o1模型上，对识别两组数字的准确率进行演习：鉴于以往我们在生成式语言大模型(如GPT-4)识别9.11与9.9两组数字时，往往会答出9.11>9.9的错误结论，在本次会上，我们也对o1模型进行同样的识别游戏，答出了9.9＞9.11的正确结论。

在线上参加COPU小型讨论会的一位朋友(韩宪平）7点质疑:

1) Post training工作还属于 train-time Scaing阶段，跟pre-training一样类似于普通软件的源码、编译阶段，而o1的创新主要在test-time Compute类似于runtime阶段，选有Ilya署名的文章“Lets Verify step by step”有条件的单位应该多做实验了，给数学的“因为…，所以…”标注给正确的和不正确的 intermediate rationals加reward，生成思维链CoT

2) o1应该有一个PRM Verifier验证网络不停地比较reward大小

3) PRM=process Reward Models

4) “后训练时代来了”显然是错误判断

5) Post-training与 inference并不相同， inference是“test-time Compute”

6) 更多算力不是投入post-training而是inference Scaling

7) inference有点类似通常说的runtime

COPU陆主席请杨耀东研究员作答。

杨老师并不认同韩宪平的意见，推荐OpenAI o1技术分析：强化学习“后训练”时代来了的文章：为什么我们需要post-Training Scaling Laws?pre-training阶段Scaling Laws随着模型尺寸逐渐增大，预训练阶段产数Scaling Up带来的边缘收益开始递减，如果想深度提升模型推理能力和长程问题能力，基于RL的post-Training将会成为下一个突破口。自回归模型在数学推理问题上很难进一步的一点在于没有办法进行回答

的自主修正，如果仅是依靠生成式方法和扩大参数规模，那么在数学推理任务上带来的收益不会太大，所以需要寻找额外的 Scaling Laws。

恰在此时，智源研究院理事长黄铁军教授为支持我们 o1 模型的讨论，也转来北京大学对齐团队（指导：杨耀东）独家解读的文章：OpenAI o1 开启“后训练”如下：

新的扩展律 post-Training 已经出现，后训练的时代已经到来。OpenAI o1 开启“后训练”时代学习新范式。Open AI o1 在数学、代码、长程规划上取得显著进步。2023 年，Deep-mind 的 CEO Demis Hassabis 强调用 Tree Search 来增强模型的推理能力。在 o1 上训练中也用到 Tree Search 的技巧。实际上，OpenAI o1 运用的技术关键还是在于强化学习的投索与学习机制。基于 LLM 已有的推理能力，迭代式的 Boot strap 模型产生合理推理过程（ Rationales）的能力，并将 Rationales 融入到训练过程内，让模型学会进行推理，而后再运用足够强大的计算量实现 Post-Training 阶段的 Scaling。注意这里合理推理过程并不只是对问题的拆解和初步作答，还有对于为什么如此作答的分析和思考。

技术要点有三：1、后训练扩展律 post-Training Scaling Laws 已经出现，并且 Post-Training Scaling Laws 为上述技术路径的成功提供了有力的支持。2、模型学习的是产生合理批理的过程，MCTS 在其中的作用是诱导合理推理过程的产生或构建相应的偏序对形式细粒度奖励信号，而非直接搜索过程和最终答案。3、模型的 Boot Strap 有助于构建新的高质量数据，并且新的 Retionates 数据促进了模型进一步提升能力。Open AI o1 的发布是 Post-Training Scaling Laws 的强力体现。

北京时间 9 月 13 日午夜 OpenAI 发布 o1 系列模型，旨在专门解决难题。Open AI o1 的成功离不开后训练阶段(Post-Training Stage)中强化学习训练和推理阶段思考计算量的增大。新的扩展律——“后训练”扩展律(Post-Training Scaling Laws)可能引发社区对于算力分配、后训练能力的重新思考。OpenAI o1 在数学代码等复杂推理能力上取得巨大进步。帮助 o1 取得如此性能飞跃的是 Post-Training 阶段 RL 计算量的 Scaling 和测量推理阶段思考时间的 Scaling。Open AI o1 在一些常规任务上没有显著提升，推理能力和强指令似乎呈现了分离。在“后训练”扩展律(post-Training Scaling Law)下,训练阶段的计算不再只是和参数量的上升有关，同时也会包含 RL 探索时 LLM Inference 的计算量，测试阶段模型推理和反思的计算量也会影响模型最终的表现。随着更多的强化学习(训练时计算)和更多的是思考时间(测试时计算,o1 的性能也在不断提升。随着参数扩展律的边际效益逐渐递减，应将更多算力转向 Post-Training 阶段和推理阶段。Open AI 的成功，关键在于合理使用强化学习的探索仅靠蒙特卡洛树搜索(MCTS)是远远不够的，因为仅靠 MCTS 无法让模型学会思考问题的关联。在隐式自动化 CoT 背后，是模型真正学会了合理的中间推理过程 Rationales。通过思维链(Chain of Thought,COT)优化模型输出,因为该思维链在其生成过程中有助于增强模型的推理能力(尤其在数学和代码生成等任务中表现出色)。

9 月 29 日韩老师发文: 我明白了他们说的“后训练”是指的 post—(pre-train+post-train),训练阶段是给知识编码，参数就固定不再调整了。说“推理时代来了”多好。推理也是陆总最先提出的。

开源创新是优化产业结构推动经济高质量发展的重大举措

陆首群

2024.09.24

1970 年是 UNIX 发展的元年，实际上也是全球开源软件发展的起点，至今已有 54 年的历史。

1991 年 AT&T-Bell Labs(USL/USG)与中国合作，将其最新开发的 UNIX 版本-UNIXSVR4.2 源代码向中方开放（中方是全球获得 UNIX 源代码的第一家），此时 AT&T 已将 UNIX 从开源转向闭源，中美合作于 1992 年翻译出版了 UNIX SVR4.2 中文版，并宣布开源(成为 UNIX SVR4.2 由闭源的英文版转向开源的中文版全球的第一家)，从此时起，中国开源的发展至今也有 32 年的历史。

今天的开源创新，锻造了现代创新引擎: “互联网+基于知识社会的创新 2.0，开源也是打通人工智能发展瓶颈的利器，可加快技术发展进程，推动深度信息技术（包括大数据、云原生、区块链、人工智能等)的发展，开源也有力支持互联网发展中的数字主权建设。正在成为丰富生态建设优化产业结构，推动经济高质量发展的重大举措。

在当前全球开源发展浪潮中，中国开源继续展现出强劲的发展势头和独特的创新活力，正在由全球的第二梯队走向升级的步伐。我国开源正在驾驭现代创新引擎，进行着 0→1 的创新活动，以促进数字化转型和智能化重构。

有人认为，开源会泄漏产品的原创技术，无法形成产品量产的产业，这是对开源概念或内涵的误解。

作为开源产品核心的开源社区发行版是完全开源透明的，可从网上自由下载的，发行时也是免费的。

但开源产品整体，从狭义规范讲，是由产品核心+产品主体+用户界面+应用+安全模块+整体附加与改变部分(二进制转换、工程化实现、维护升级)等构成，除产品核心是开源的外，其他部分是闭源(或混源)的，所以从外面看开源产品的整体，这些闭源因素屏蔽了核心的开源因素。

从广义规范讲，开源产品还包括生态系统+商业模式，这部分也是闭源（或混源）的，作为开源产品整体的商业发行版是具有开源+闭源的性质，不会泄露产品的原创技术，也不会影响开源产业的形成发展，在销售时也是要收费的（但具有优惠条件）。

欢迎大家学习开源，拥抱开源，运用开源，发展开源！建设健全的开源产业生态，推动现代化经济高质量发展。

语言大模型越大越不可靠

Polytechnic University of Valencia

摘自《Nature》杂志（2024 年 9 月 25 日出版）

COPU 编者：这项研究来自瓦伦西亚理工大学研究团队及其合作者，他们在研究了 GPT、LLaMA 和 BLOOM 系列大语言模型（LLM）之后发现-

本刊将发表中文译稿并附英文原稿。

尽管正如预期的那样，由于一些微调方法（如 RLFH），参数规模更大的 LLM 生成的答案更准确，尤其是在复杂任务上，但整体可靠性却较低。

在所有不准确的回答中，错误回答的比例有所上升，甚至在一些简单任务上出现更多低级错误。例如，GPT-4 在处理简单的加法和字谜时错误率竟比一些小模型高出 15%。这是因为模型不太可能回避回答问题-比如承认它不知道或者转移话题。

以上结果表明，大参数模型在简单任务上可能会出现过度拟合或错误估计的风险，反而更不可靠。

模型扩展带来“能力反差”

在这项工作中，研究人员从用户与 LLM 互动的角度，探讨了难度一致性、任务回避和提示稳定性三个核心交织元素对 LLM 可靠性的影响。

该研究的通讯作者 Jos é Hern á ndez Orallo 教授表示：“语言模型的可靠性与人类对任务难度的感知不匹配。模型能够解决博士级的数学问题，但同时却可能在简单的加法上出错。”

研究团队对比了 GPT、LLaMA、BLooM 三大模型系列在不同任务中的表现，尤其是在数字计算、文字游戏、地理知识、基础与高级科学问题和信息转化等任务。通过对这些任务的正确率、错误率和回避行为的分析，提示了模型扩展带来的能力反差现象。

1、难度悖论“越简单，错误越多？”

以加法任务为例，虽然模型能够解决复杂的多位数加法，但在简单的两位数加法上却频繁出错。例如，所有 LLaMA 模型在最简单任务上的正确率未超过 60%，而在一些较难的任务中，则表现相对出色。

这一现象在 GPT 模型中也尤为突出，特别在处理诸如简单加法和字谜任务时，优化后的模型反而容易给出错误答案。研究团队指出，这一现象表明当前模型的扩展可能过于集中于复杂任务，而忽视了简单任务。

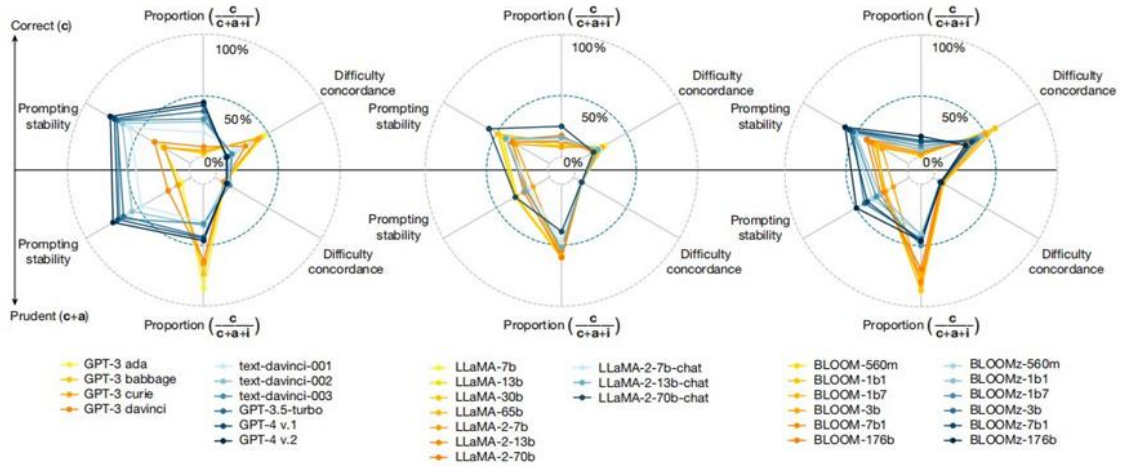


图 1 GPT、LLaMA 和 BLOOM 核型的关键指标

这一结果颠覆了人们对 LLM 的传统认知，表明扩展模型并不总是能带来全面的提升，对其在实际应用中的可靠性提出了质疑。

2、错误率与避免行为——“自信过头”

除了难变不一致现象，研究还揭示了优化后模型中回避行为与错误率之间的微妙关系。

回避行为是指栏型在无法正确回答时，选择不作答或给出不符合要求的回应。在模型未优化时，回避行为比较常见，即当模型不确定答案时，往往会选择“不作答”或做出模糊的回应。然而，在经过扩展和优化后，模型则大幅减少了回避行为，转而给出了更多表面上“合理”但实际上错误的答案。

这意味着，虽然一些优化方法使得模型更“自信”，减少了回避行为，但错误率却随之增加。这一现象在 GPT-4 和 GPT-3.5-turbo 等模型中尤其明显，规模扩展并未带来预期的稳定性。对比 LLaMA 和 BLooM 题型，这一趋势强烈不那么明显，但同样存在。

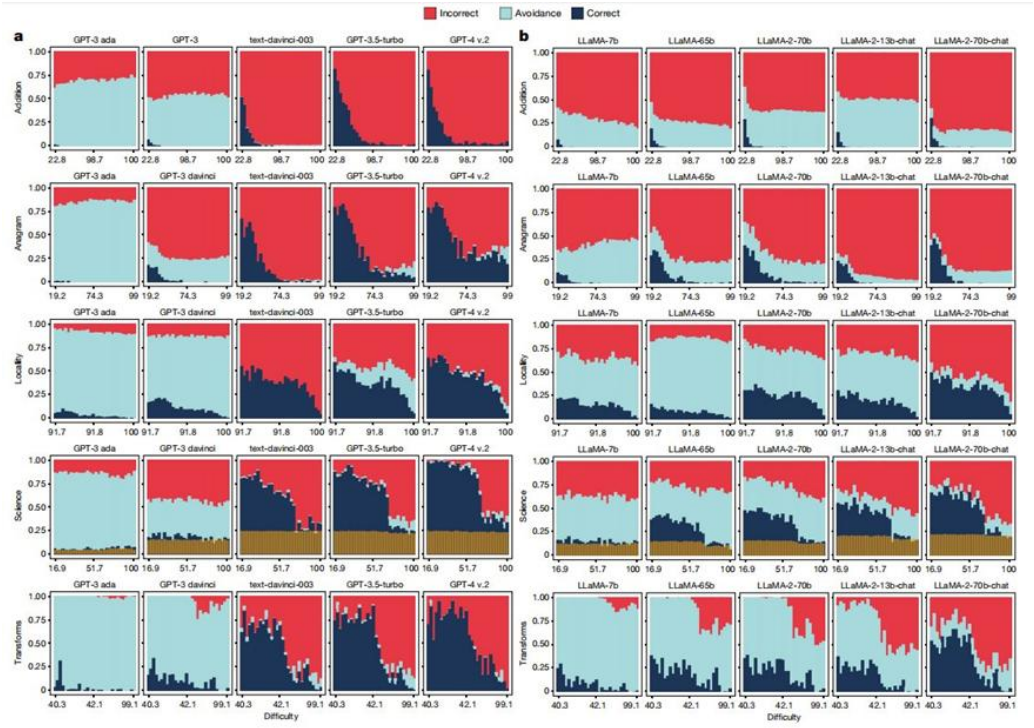


图 2 GPT 和 LLaMA 模型的性能随难度增加而提高

研究团队称，这种现象与用户在模型上产生的过度信任密切相关，尤其是在用户面对看似简单的任务时。

该论文的第一作者 Lexin Zhou 表示：“这可能导致最初过于依赖模型的用户感到失望。此外，与人类不同，避免提供答案的倾向不会随着困难而增加。例如，人类倾向于避免对超出其能力的问题给出反馈。这让用户有责任在与模型的交互过程中发现错误。”

3、提示词带来的是稳定性，还是陷阱？

该研究分析了模型对提示词的敏感性，特别是某些提示是否存在“安全区”。

结果表明，随着模型规模的增加，模型对不同自然语言表述敏感度有所提高，能及时应对措辞上的微调。然而，即使经过扩展和优化，题型在不同难度级别的任务上仍存在不一致的表现。而且，在不同表述下，模型的回答准确率存在波动。

研究发现，人们对难度的认识存在不一致。论文作者之一 Yad Moros Daval 说：“模型是否在我们预期的地方失败了？我们发现，模型在人类认为困难的任务上往往不太准确，但即使在简单任务上，它们也不是 100% 准确。这意味着不存在可以信任模型完美运行的安全区。

具体而言，未经优化的 GPT 和 LLaMA 模型对提示词的选择表现出极高的敏感性，尤其是在简单任务中。如果提示词选择得当，模型的表现会有所提升；而优化后的模型在提示词敏感性上有所改善，表现更加稳定，但也存在一定的变异性。经过优化的模型相比原始模型（ raw models）在

变化上更为稳定，且正确率更高，但与人类判断难度的一致性和谨慎度方面表现较差。

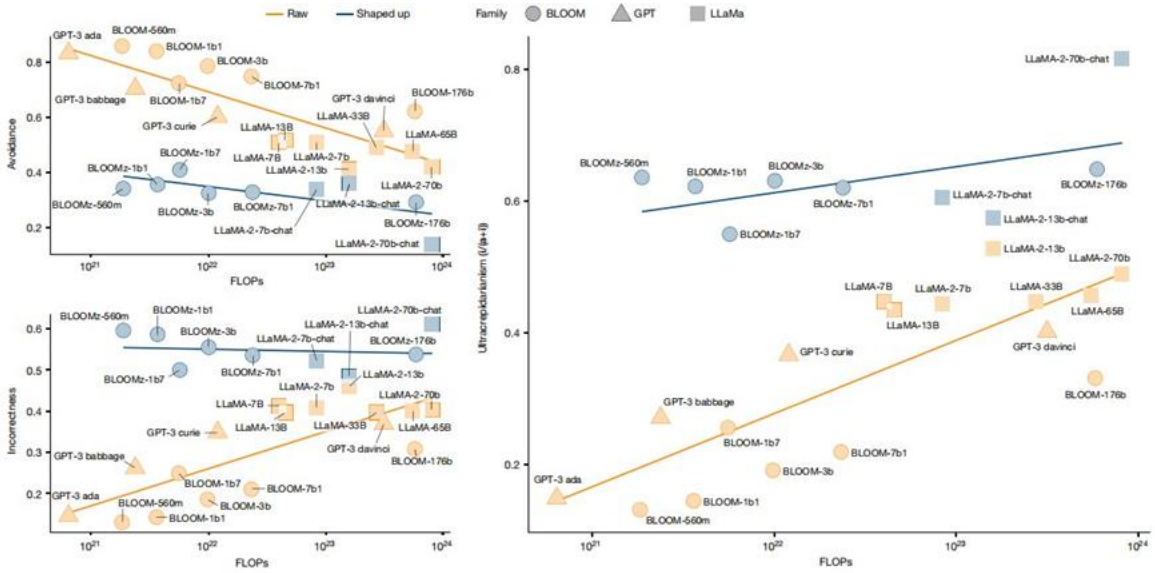


图 3　LLaMA、BLOOM 系列以及非结构 GPT 模型的尺度分析

Larger and more instructable language models become less reliable

Polytechnic University of Valencia

Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri & José Hernández-Orallo

Abstract

The prevailing methods to make large language models more powerful and amenable have been based on continuous scaling up (that is, increasing their size, data volume and computational resources1) and bespoke shaping up (including post-filtering2,3, fine tuning or use of human feedback4,5). However, larger and more instructable large language models may have become less reliable. By studying the relationship between difficulty concordance, task avoidance and prompting stability of several language model families, here we show that easy instances for human participants are also easy for the models, but scaled-up, shaped-up models do not secure areas of low difficulty in which either the model does not err or human supervision can spot the errors. We also find that early models often avoid user questions but scaled-up, shaped-up models tend to give an apparently sensible yet wrong answer much more often, including errors on difficult questions that human supervisors frequently overlook. Moreover, we observe that stability to different natural phrasings of the same question is improved by scaling-up and shaping-up interventions, but pockets of variability persist across difficulty levels. These findings highlight the need for a fundamental shift in the design and development of general-purpose artificial intelligence, particularly in high-stakes areas for which a predictable distribution of errors is paramount.

Main

Millions of people are using general-purpose artificial intelligence (AI) systems based on large language models (LLMs), which have become commonplace in areas such as

education6, medicine7, science8,9 and administration10,11. As these models frequently make mistakes, users have to supervise model operation and manage their expectations, for the reliable use of these systems. With language models becoming larger and more instructable, we need to analyse how this reliability has evolved. Since the early LLMs12,13,14, models have been scaled up—trained with more parameters, on larger datasets and with longer training times—and have also been shaped up with human feedback—using techniques such as instruction fine tuning4, reinforcement learning from human feedback (RLHF)5 or output-filtering moderation techniques2,3.

It may be taken for granted that as models become more powerful and better aligned by using these strategies, they also become more reliable from a human perspective, that is, their errors follow a predictable pattern that humans can understand and adjust their queries to15. For instance, early models failed at simple additions such as '20 + 183'. Performance was highly predictable: failure was common. As a result, users easily understood that there was no operating range for this task: nobody used these models for addition. A few scaled-up and shaped-up generations later, the models not only seemingly master these additions but also successfully perform additions of 50 digits or more. Because of this prowess, people may start using them as calculators (for example, to convert measurements to different units16). It is only in such cases that users become disappointed when the model fails at a simple prompt such as 'Add 3913 and 92'. The user-driven reliability is then seriously damaged, because the model fails when the user thinks these digits were in the operating range. The experience becomes even more baffling when the user gets the correct answer if the question is adjusted slightly, for example to '3913 + 92 =', or if it is not changed at all—because many models are configured to be non-deterministic. Although this prompt sensitivity has been analysed extensively17,18,19,20, it is poorly understood why an over-diligent system spouts a wrong answer for 100-digit addition instead of simply answering 'I'm afraid I can't do that'. This reckless behaviour has been incentivized by developers building models that are 'never evasive'21.

Reliability fluctuations

To understand the evolution of reliability, we analyse the trajectory of several families of LLMs: the generative pre-training (GPT) saga developed by OpenAI, the LLaMA series developed by Meta and the BLOOM suite developed by BigScience. GPT has led the state of the art in the past few years and, according to several surveys22,23,24, is central to the LLM ecosystem, influencing transformer-based architectures, training data, evaluation frameworks and alignment techniques. LLaMA25,26 is the best example of a family for which weights have been released, and BLOOM27,28 is the result of an even more open endeavour coming from the scientific community. Each family represents a genuine effort of making LLMs more capable and better aligned at the same time. Table 1 summarizes the details of models in these three families. Scaling (increasing the number of parameters, data size and compute) has been identified as a key predictor for overall performance1, and shaping (modifying the trained systems) has improved their instructability and alignment. This creates two categories. The first includes the 'raw' models—GPT-3 ada, babbage,

curie and davinci—the non-chat LLaMA models and the base (non-z) BLOOM models. The second comprises the shaped-up models (or instruct or chat models), which incorporate some kind of instruction adaptation22, fine tuning or safety moderation of the outputs. For our analysis, it is convenient that BLOOM and LLaMA have six and three exactly paired versions, respectively, of raw and shaped-up models to disentangle scaling up from shaping up.

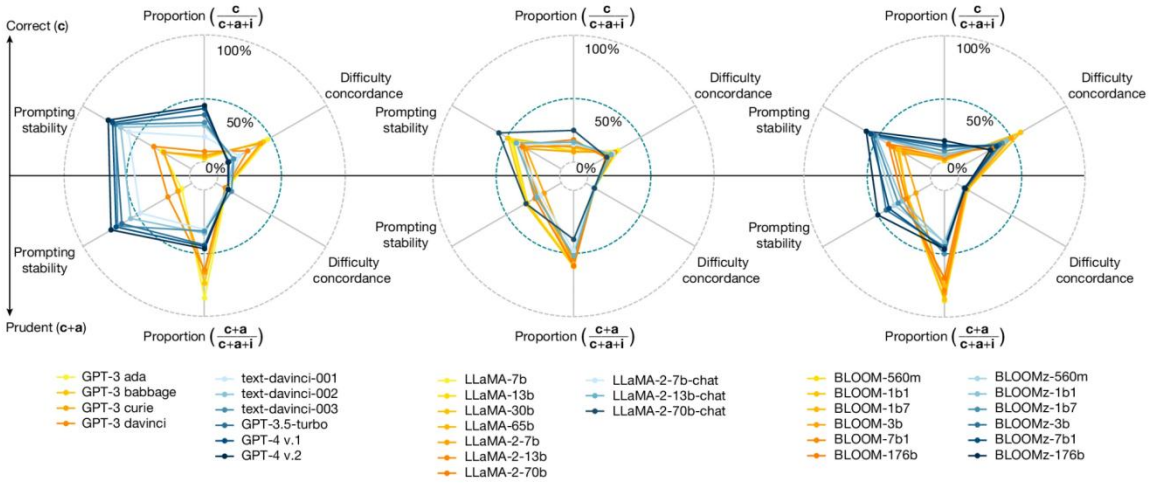
Table 1 Ten GPT, ten LLaMA and twelve BLOOM models

https://www.nature.com/articles/s41586-024-07930-y/tables/1 **(Full size table)**

Figure 1 represents how some key indicators show that the shaped-up models (in blue) are more stable to prompt variation and are more correct, at the cost of being less concordant with human difficulty, and having more overall failures (less prudent). The indicators summarize the behaviour of five carefully selected benchmarks in the domains of simple numeracy ('addition'), vocabulary reshuffle ('anagram'), geographical knowledge ('locality'), diverse scientific skills ('science') and information-centric transformations ('transforms'). This covers a range of domains and degrees of open-endedness of the answers.

Fig. 1: Key indicators for several models in GPT (OpenAI), LLaMA (Meta) and BLOOM (BigScience) families.

From: Larger and more instructable language models become less reliable



The raw models (yellow to orange) and the shaped-up models (light to dark blue) cluster differently. As the answers for all these models fall into three categories (correct, avoidant and incorrect), shortened as c, a and i, respectively, we have indicators for correctness versus avoidance + incorrectness, and prudence (correctness + avoidance) versus incorrectness. Looking at the correctness indicators (top half), which represent accurate responses, we see that the shaped-up models are more stable to prompt variations and are

more frequently correct (higher correctness proportion) but are less concordant with human difficulty than the raw counterparts. Looking at the prudence indicators (bottom half), we see that the shaped-up models are also more stable to prompt variations, but fail more frequently (lower prudence proportion, by avoiding less) and are not much more concordant with human difficulty. Focusing only on the shaped-up models (in blue), we observe that the most powerful GPT-4 v.2, LLaMA-2-70b-chat and BLOOMz-176b models perform best in correctness proportion and prompting stability (top and bottom), but equal to or worse than other models for all the other indicators, with many fluctuations that do not indicate a clear positive trend in these other dimensions. Details of the indicators and data used for this plot are found in the Methods. Extended Data Table 1 provides a more detailed perspective on the same results.

https://www.nature.com/articles/s41586-024-07930-y#MOESM3

We identify good intrinsic proxies for human difficulty based on relevant literature in the first two domains ('addition' and 'anagram'), or by identifying demand-related features in the rest (excluding 'science', for which multiple human difficulty assessments were already available for all the instances29). To determine their quality, we conducted an extensive human study (S1) to assess which difficulty proxies best matched human expectations, and calibrate the proxies to a normalized difficulty score, ranging from 0 to 100, representing the anticipated percentage of failure for the 'average human'. Systematically controlling for human difficulty is crucial for the understanding of user-driven reliability: human expectations of success depend on the perception of the difficulty of instances30,31,32. Table 2 provides an overview of the five benchmarks, the intrinsic difficulty function used as a proxy for human difficulty (discussed in the Methods), some examples and the calibrated human difficulty values for the given examples.

Table 2 Five benchmarks （https://www.nature.com/articles/s41586-024-07930-y/tables/2）

Another necessary and innovative element in our analysis is that we consider three categories for the responses: correct, incorrect and avoidant, denoted by c, i and a, respectively. Avoidance in human participants has been extensively explored in psychology33,34,35. Such avoidant behaviours include procrastination, deviation, making excuses or simply not answering. For LLMs, avoidance is also referred to as hedging, refusal3 or evasiveness21, including fortuitous utterances or continuations that are not answers (non-conforming), and those responses at the meta-level explaining why the question is not answered (for epistemic or ethical reasons). Supplementary Table 11 shows the types of avoidance for some tasks in the five benchmarks.

Difficulty concordance, task avoidance and prompting stability must be regarded from the point of view of human users interacting with LLMs. Our human study S1 (see Supplementary Note 6) analyses whether human perceptions of difficulty in general are

aligned with actual human performance and self-confidence, because this has important implications in the tasks humans decide to delegate to language models and their prompt formulation. But as crucial as the inputs are, so is the way the outputs from the model are used, verified or supervised. The context of use of both input and output determines how reliable the use of these systems is. We conducted a second human study S2 (see Supplementary Note 7), in which we explore whether human participants can accurately assess the outputs of models and thus compensate for different types of error. With a three-valued confusion matrix with correctness, avoidance and incorrectness, we can focus on the frequency of non-avoidant cases for which humans believe the output is correct but it is not (Fig. 3).

With this setup, we investigate three core and intertwined elements that affect the reliability of LLMs from a human perspective.

1. Difficulty concordance. Are errors more likely for items that humans perceive as difficult? Do scaling and shaping eliminate errors for easy items, thereby creating areas of reliable operation?

2. Task avoidance. How often do language models give plausible but wrong answers instead of safely avoiding answering questions? Are scaled-up, shaped-up models better at avoiding errors or making them detectable for humans?

3. Prompting stability. How are correctness and avoidance affected by tangential changes in the prompt? Are scaled-up, shaped-up models less sensitive to prompt variation across difficulty levels?

We will answer these questions by using human difficulty metrics for each benchmark (see Table 2), examining different kinds of avoidance (Supplementary Table 11), and using 15 natural prompt variations—prompts conceived as genuine instructions or questions provided by humans—per benchmark (Supplementary Tables 1 and 2). Difficulty, avoidance and prompting, as well as their evolution, have been analysed from different perspectives17,18,19,36,37,38,39 (see Supplementary Note 13 for a full discussion). Here we focus on the systemic interaction of these three elements from the perspective of LLM scaling and shaping up.

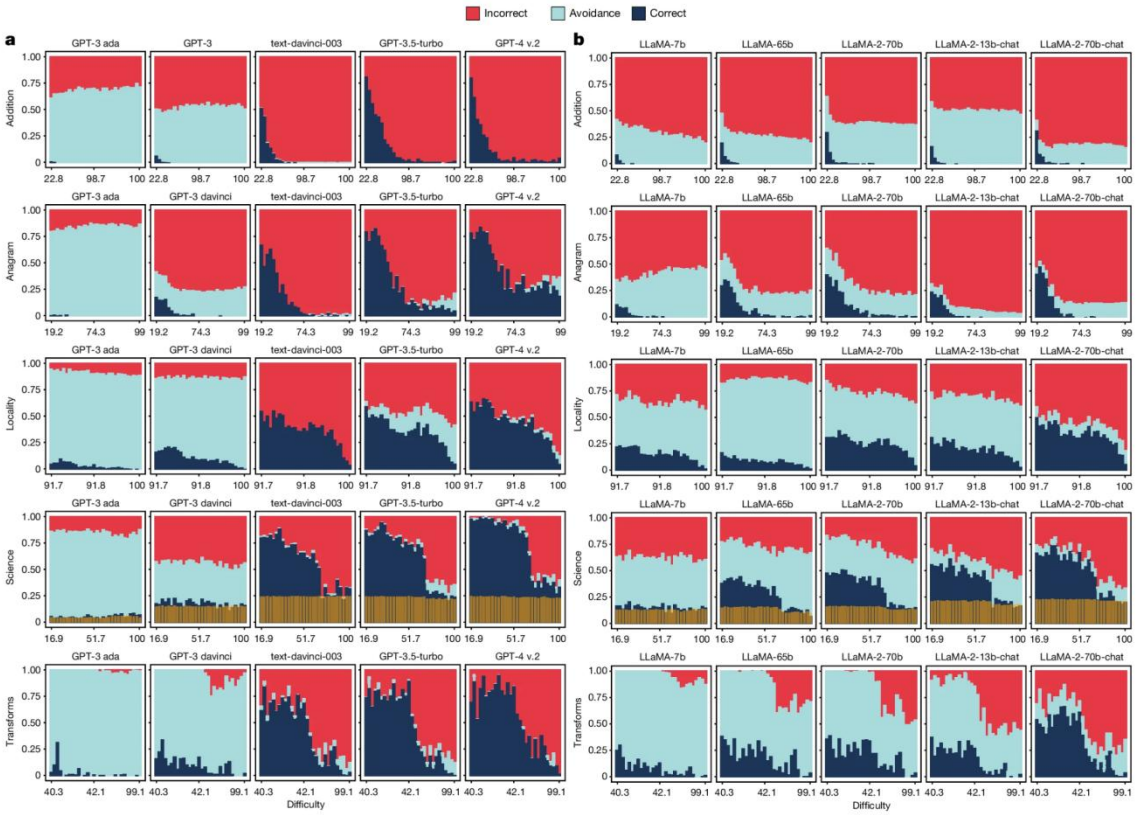
Results

Figure 2 shows the results of a selection of models in the GPT and LLaMA families, increasingly scaled up, with the shaped-up models on the right, for the five domains: ‘addition’, ‘anagram’, ‘locality’, ‘science’ and ‘transforms’. We see that the percentage of correct responses increases for scaled-up, shaped-up models, as we approach the last column. This is an expected result and holds consistently for the rest of the models, shown

in Extended Data Fig. 1 (GPT), Extended Data Fig. 2 (LLaMA) and Supplementary Fig. 14 (BLOOM family).

Fig. 2: Performance of a selection of GPT and LLaMA models with increasing difficulty.

From: Larger and more instructable language models become less reliable



The values are split by correct, avoidant and incorrect results. For each combination of model and benchmark, the result is the average of 15 prompt templates (see Supplementary Tables 1 and 2). For each benchmark, we show its chosen intrinsic difficulty, monotonically calibrated to human expectations on the x axis for ease of comparison between benchmarks. The x axis is split into 30 equal-sized bins, for which the ranges must be taken as indicative of different distributions of perceived human difficulty across benchmarks. For 'science', the transparent yellow bars at the bottom represent the random guess probability (25% of the non-avoidance answers). Plots for all GPT and LLaMA models are provided in Extended Data Figs. 1 and 2 and for the BLOOM family in Supplementary Fig. 14.

Let us focus on the evolution of correctness with respect to difficulty. For 'addition', we use the number of carry operations in the sum (fcry). For 'anagram', we use the number of letters of the given anagram (flet). For 'locality', we use the inverse of city popularity (fpop). For 'science', we use human difficulty (fhum) directly. For 'transforms', we use a combination of input and output word counts and Levenshtein distance (fw+l) (Table 2). As we discuss in the Methods, these are chosen as good proxies of human expectations about

what is hard or easy according to human study S1 (see Supplementary Note 6). As the difficulty increases, correctness noticeably decreases for all the models. To confirm this, Supplementary Table 8 shows the correlations between correctness and the proxies for human difficulty. Except for BLOOM for addition, all of them are high.

However, despite the predictive power of human difficulty metrics for correctness, full reliability is not even achieved at very low difficulty levels. Although the models can solve highly challenging instances, they also still fail at very simple ones. This is especially evident for 'anagram' (GPT), 'science' (LLaMA) and 'locality' and 'transforms' (GPT and LLaMA), proving the presence of a difficulty discordance phenomenon. The discordance is observed across all the LLMs, with no apparent improvement through the strategies of scaling up and shaping up, confirmed by the aggregated metric shown in Fig. 1. This is especially the case for GPT-4, compared with its predecessor GPT-3.5-turbo, primarily increasing performance on instances of medium or high difficulty with no clear improvement for easy tasks. For the LLaMA family, no model achieves 60% correctness at the simplest difficulty level (discounting 25% random guess for 'science'). The only exception is a region with low difficulty for 'science' with GPT-4, with almost perfect results up to medium difficulty levels.

Focusing on the trend across models, we also see something more: the percentage of incorrect results increases markedly from the raw to the shaped-up models, as a consequence of substantially reducing avoidance (which almost disappears for GPT-4). Where the raw models tend to give non-conforming outputs that cannot be interpreted as an answer (Supplementary Fig. 16), shaped-up models instead give seemingly plausible but wrong answers. More concretely, the area of avoidance in Fig. 2 decreases drastically from GPT-3 ada to text-davinci-003 and is replaced with increasingly more incorrect answers. Then, for GPT-3.5-turbo, avoidance increases slightly, only to taper off again with GPT-4. This change from avoidant to incorrect answers is less pronounced for the LLaMA family, but still clear when comparing the first with the last models. This is summarized by the prudence indicators in Fig. 1, showing that the shaped-up models perform worse in terms of avoidance. This does not match the expectation that more recent LLMs would more successfully avoid answering outside their operating range. In our analysis of the types of avoidance (see Supplementary Note 15), we do see non-conforming avoidance changing to epistemic avoidance for shaped-up models, which is a positive trend. But the pattern is not consistent, and cannot compensate for the general drop in avoidance.

Looking at the trend over difficulty, the important question is whether avoidance increases for more difficult instances, as would be appropriate for the corresponding lower level of correctness. Figure 2 shows that this is not the case. There are only a few pockets of correlation and the correlations are weak. This is the case for the last three GPT models for 'anagram', 'locality' and 'science' and a few LLaMA models for 'anagram' and 'science'. In some other cases, we see an initial increase in avoidance but then stagnation at higher difficulty levels. The percentage of avoidant answers rarely rises quicker than the

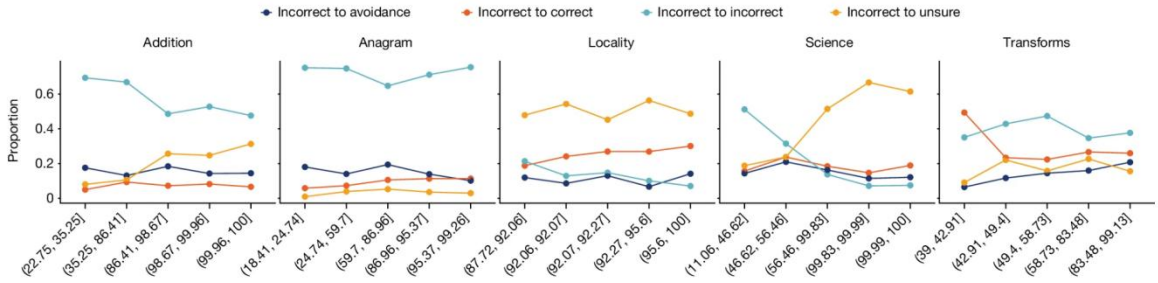
percentage of incorrect ones. The reading is clear: errors still become more frequent. This represents an involution in reliability: there is no difficulty range for which errors are improbable, either because the questions are so easy that the model never fails or because they are so difficult that the model always avoids giving an answer.

We next wondered whether it is possible that this lack of reliability may be motivated by some prompts being especially poor or brittle, and whether we could find a secure region for those particular prompts. We analyse prompt sensitivity disaggregating by correctness, avoidance and incorrectness, using the prompts in Supplementary Tables 1 and 2. A direct disaggregation can be found in Supplementary Fig. 1, showing that shaped-up models are, in general, less sensitive to prompt variation. But if we look at the evolution against difficulty, as shown in Extended Data Figs. 3 and 4 for the most representative models of the GPT and LLaMA families, respectively (all models are shown in Supplementary Figs. 12, 13 and 15), we observe a big difference between the raw models (represented by GPT-3 davinci) and other models of the GPT family, whereas the LLaMA family underwent a more timid transformation. The raw GPT and all the LLaMA models are highly sensitive to the prompts, even in the case of highly unambiguous tasks such as ‘addition’. Difficulty does not seem to affect sensitivity very much, and for easy instances, we see that the raw models (particularly, GPT-3 davinci and non-chat LLaMA models) have some capacity that is unlocked only by carefully chosen prompts. Things change substantially for the shaped-up models, the last six GPT models and the last three LLaMA (chat) models, which are more stable, but with pockets of variability across difficulty levels.

Overall, these different levels of prompt sensitivity across difficulty levels have important implications for users, especially as human study S2 shows that supervision is not able to compensate for this unreliability (Fig. 3). Looking at the correct-to-incorrect type of error in Fig. 3 (red), if the user expectations on difficulty were aligned with model results, we should have fewer cases on the left area of the curve (easy instances), and those should be better verified by humans. This would lead to a safe haven or operating area for those instances that are regarded as easy by humans, with low error from the model and low supervision error from the human using the response from the model. However, unfortunately, this happens only for easy additions and for a wider range of anagrams, because verification is generally straightforward for these two datasets.

Fig. 3: Evolution of types of supervision error versus difficulty according to human survey S2.

From: Larger and more instructable language models become less reliable



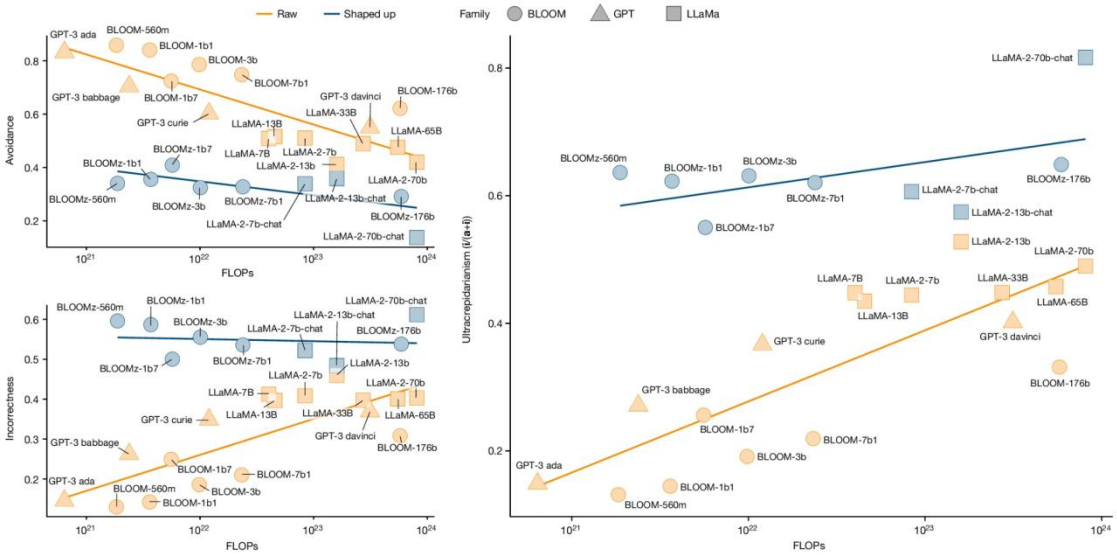
In the survey (Supplementary Fig. 4), participants have to determine whether the output of a model is correct, avoidant or incorrect (or do not know, represented by the 'unsure' option in the questionnaire). Difficulty (*x* axis) is shown in equal-sized bins. We see very few areas where the dangerous error (incorrect being considered correct by participants) is sufficiently low to consider a safe operating region.

In the survey (Supplementary Fig. 4), participants have to determine whether the output of a model is correct, avoidant or incorrect (or do not know, represented by the 'unsure' option in the questionnaire). Difficulty (x axis) is shown in equal-sized bins. We see very few areas where the dangerous error (incorrect being considered correct by participants) is sufficiently low to consider a safe operating region.

Our observations about GPT and LLaMA also apply to the BLOOM family (Supplementary Note 11). To disentangle the effects of scaling and shaping, we conduct an ablation study using LLaMA and BLOOM models in their shaped-up versions (named chat and z, respectively) and the raw versions, with the advantage that each pair has equal pre-training data and configuration. We also include all other models with known compute, such as the non-instruct GPT models. We take the same data summarized in Fig. 1 (Extended Data Table 1) and perform a scaling analysis using the FLOPs (floating-point operations) column in Table 1. FLOPs information usually captures both data and parameter count if models are well dimensioned40. We separate the trends between raw and shaped-up models. The fact that correctness increases with scale has been systematically shown in the literature of scaling laws1,40. With our data and three-outcome labelling, we can now analyse the unexplored evolution of avoidance and incorrectness (Fig. 4, left).

Fig. 4: Scaling analysis of LLaMA and BLOOM families and non-instruct GPT models.

From: Larger and more instructable language models become less reliable



The plot uses a logarithmic scale for FLOPs. The focus is on avoidance (**a**; top left), incorrectness (**i**; bottom left) and ultracrepidarianism (**i**/(**a** + **i**); right)—the proportion of incorrect over both avoidant and incorrect answers.

As evident in Fig. 4, avoidance is clearly much lower for shaped-up models (blue) than for raw models (orange), but incorrectness is much higher. But even if correctness increases with scale, incorrectness does not decrease; for the raw models, it increases considerably. This is surprising, and it becomes more evident when we analyse the percentage of incorrect responses for those that are not correct in (i/(a + i) in our notation; Fig. 4 (right)). We see a large increase in the proportion of errors, with models becoming more ultracrepidarian (increasingly giving a non-avoidant answer when they do not know, consequently failing proportionally more).

We can now take all these observations and trends into account, in tandem with the expectations of a regular human user (study S1) and the limited human capability for verification and supervision (study S2). This leads to a re-understanding of the reliability evolution of LLMs, organized in groups of two findings for difficulty discordance (F1a and F1b), task avoidance (F2a and F2b) and prompt sensitivity (F3a and F3b):

F1a—human difficulty proxies serve as valuable predictors for LLM correctness. Proxies of human difficulty are negatively correlated with correctness, implying that for a given task, humans themselves can have approximate expectations for the correctness of an instance. Relevance: this predictability is crucial as alternative success estimators when model self-confidence is either not available or markedly weakened (for example, RLHF ruining calibration3,41).

F1b—improvement happens at hard instances as problems with easy instances persist, extending the difficulty discordance. Current LLMs clearly lack easy operating areas with

no error. In fact, the latest models of all the families are not securing any reliable operating area. Relevance: this is especially concerning in applications that demand the identification of operating conditions with high reliability.

F2a—scaling and shaping currently exchange avoidance for more incorrectness. The level of avoidance depends on the model version used, and in some cases, it vanishes entirely, with incorrectness taking important proportions of the waning avoidance (that is, ultracrepidarianism). Relevance: this elimination of the buffer of avoidance (intentionally or not) may lead users to initially overtrust tasks they do not command, but may cause them to be let down in the long term.

F2b—avoidance does not increase with difficulty, and rejections by human supervision do not either. Model errors increase with difficulty, but avoidance does not. Users can recognize these high-difficulty instances but still make frequent incorrect-to-correct supervision errors. Relevance: users do not sufficiently use their expectations on difficulty to compensate for increasing error rates in high-difficulty regions, indicating over-reliance.

F3a—scaling up and shaping up may not free users from prompt engineering. Our observations indicate that there is an increase in prompting stability. However, models differ in their levels of prompt sensitivity, and this varies across difficulty levels. Relevance: users may struggle to find prompts that benefit avoidance over incorrect answers. Human supervision does not fix these errors.

F3b—improvement in prompt performance is not monotonic across difficulty levels. Some prompts do not follow the monotonic trend of the average, are less conforming with the difficulty metric and have fewer errors for hard instances. Relevance: this non-monotonicity is problematic because users may be swayed by prompts that work well for difficult instances but simultaneously get more incorrect responses for the easy instances.

As shown in Fig. 1, we can revisit the summarized indicators of the three families. Looking at the two main clusters and the worse results of the shaped-up models on errors and difficulty concordance, we may rush to conclude that all kinds of scaling up and shaping up are inappropriate for ensuring user-driven reliability in the future. However, these effects may well be the result of the specific aspirations for these models: higher correctness rates (to excel in the benchmarks by getting more instances right but not necessarily all the easy ones) and higher instructability (to look diligent by saying something meaningful at the cost of being wrong). For instance, in scaling up, there is a tendency to include larger training corpora42 with more difficult examples, or giving more weight to authoritative sources, which may include more sophisticated examples43, dominating the loss over more straightforward examples. Moreover, shaping up has usually penalized answers that hedge or look uncertain3. That makes us wonder whether this could all be different.

Discussion

In this paper, we have conducted two human studies. The first investigates perceived and actual difficulty for participants to respond to an input (to determine whether difficulty expectations are correlated with difficulty proxies). The second includes participants supervising or verifying the output of a model (to determine whether humans will take incorrect responses as correct). Maximizing difficulty concordance and reducing possible incorrect-to-correct errors in human verification could be introduced in the loss function when training and shaping up these models. For this, collective efforts are needed to build larger datasets of human difficulty expectations and output supervision. With these data, more qualified than traditional human feedback, AI itself can be used to train supervisors that perform this shaping up, provided the aim is not to eliminate evasiveness as in ref. 21, but to find the right level of avoidance. Specialized language models in medicine and other critical areas may be designed with reject options, or coupled with external AI supervisors, thereby favouring avoidance by teaching the AI models when to refrain from answering37. These interventions should make LLMs exhibit enhanced human-like and human-aligned characteristics that ensure reliability. Until this is done, and given the high penetration of LLM use in the general population, we raise awareness that relying on human oversight for these systems is a hazard, especially for areas for which the truth is critical.

Finally, we include some limitations of our analysis and the future work that emanates from them. The first limitation of our study lies in the recruitment of participants who are mostly non-experts. We have to take this into account when interpreting the calibrated difficulty values, which are usually high for some benchmarks, as there is a high number of questions that cannot be solved by the general population. However, our motivation was to capture the same human population to estimate expected instance difficulties that are comparable across all the datasets. A second limitation is that our sample of ‘natural’ prompts was collected from a diversity of sources, but we did not have access to the frequency in which a prompt may appear in a real scenario. Last, we have only covered a sample of families with specific trajectories, excluding LLMs that delegate tasks to external tools or use sophisticated reasoning techniques, which may show different dynamics. The GPT family has been at the forefront in performance and has been used over a few years, making OpenAI extremely influential in the development of other language models22,23. In fact, the OpenAI application programming interface has the most dependencies when the ecosystems of foundation models are analysed24. LLaMA and BLOOM have a more open and systematic lineup of models, not only allowing for the disentanglement between scaling and shaping but also paving the way for an incremental analysis of their evolution using our methodology and code, in the fast-changing context of LLM development. Highlighting the reliability issues of these families and introducing new abstractions and tools for analysis is of utmost importance, enabling other researchers to explore different pathways for the scaled-up, shaped-up models of the future.

Methods

We now explain our choices of benchmarks, prompt templates, difficulty functions, response scoring, general experimental design and the key metrics used to evaluate the models.

Benchmarks and factors of difficulty

For the generality of our analysis, we selected five distinct benchmarks to reduce confounding factors as much as possible: simple numeracy (‘addition’), vocabulary reshuffle (‘anagram’), geographical knowledge (‘locality’), basic and advanced science questions (‘science’) and information-centric transformations (‘transforms’). These represent core skills (numerical, linguistic and knowledge) and more diverse ecologically valid scenarios, with some of them having extremely simple formulations and others requiring deep understanding of the information presented, as well as the integration of data from multiple sources. Closed-ended questions are typical of LLM research3, such as those found in the ‘science’ benchmark, but gradually more open-ended tasks (‘addition’, ‘anagram’, ‘locality’ and ‘transforms’) better represent a wider and more realistic use of LLMs.

Addition. This benchmark involves sums, prompting the LLMs by asking for the result of adding two addends (such as ‘3 + 7 =’). The examples in our analysis range from 1- to 100-digit additions. Because language models can not only memorize small additions but also generalize to cope with any combination of larger digits, this task is appropriate for analysing difficulty trends. With respect to the difficulty of ‘addition’, the number of digits and carry operations affect human performance on addition tasks.

Anagram. The use of anagrams as a way of assessing aspects of problem solving dates back to 1916 (ref. 45), and researchers have been using anagrams to examine a variety of phenomena, such as the cognitive processes involved in problem solving46. An ‘anagram’ task is a word puzzle in which the participant or model is presented with a jumbled string of letters, and the objective is to find a word that can be formed using all the letters given. The examples in our analysis range from 3-letter words to 20-letter words. This task involves letter manipulation and good recall from an extensive vocabulary. One peculiar element of this task is that it is easy to verify. The difficulty of anagrams is mostly influenced by the frequency of the letters and the word, the number of letters and the degree of rearrangement required.

Locality. This benchmark contains questions relating to geographical knowledge, inspired by some cognitive models of distance estimation47. The examples in our analysis ask questions about the location and size of cities in relation to each other, by giving an input

city and a randomly generated distance (d, ranging from 1 to 1,000 km). The LLM is asked to identify the most populous city (the target city) in a radius of d km from the input city. This task requires geographical knowledge and reasoning. For this benchmark, potential human difficulty factors could be the city or country popularity, their population and so on.

Science. This benchmark integrates multiple-choice questions from basic science as collected by OpenBookQA, complemented with more advanced science questions from Google-proof Q&A (GPQA). They represent tasks that LLMs are likely to encounter in educational, academic and research settings6,8,48, some of which require considerable time to solve. The included questions are Google-proof49. The 'science' benchmark, thus, includes questions of varying levels of difficulty, as determined by human judgement, providing a lens through which to examine their handling of complex, data-rich tasks in specific domains.

Transforms. This benchmark includes a comprehensive set of information-centric transformation tasks based on real-world scenarios. It focuses on domains that are most prevalent in the use of LLMs today50, and ensure that there is a ground truth for evaluation. We integrate not only many data-formatting tasks—a well-studied area in LLMs51—but also new tasks about world knowledge, information retrieval, advertising, administration, coding, scheduling and retailing. The outputs for 'transforms' may require extensive elaboration of the input (hundreds of characters) to form a correct answer, which can also be hundreds of characters long. The aim was to simulate, as closely as possible, the complexity and depth of real-world questions in a controlled experimental setting. For task difficulty, given the heterogeneity, the main factors are as general as character and word counts, and the Levenshtein distance between input and output as a proxy of transformation effort.

For the previously described domains, we found intuitive human difficulty proxies, some of which have been developed in the literature. Supplementary Note 4 provides further details on the definition of difficulty metrics and the abilities behind the features used for their definition. Using the results from human study S1, we select the difficulty functions that are most correlated with human expectations (Supplementary Table 5): fcry for 'addition', flet for 'anagram', fpop for 'locality' and fw+l for 'transforms'. For 'science', we blend and calibrate the two original human metrics into one, that is, fhum. For all the benchmarks, we normalize the original difficulty functions using a logistic mapping to a scale ranging from 0 to 100 that corresponds to the probability of human failure as estimated by humans themselves. We need to take into account that these values are an estimate (from the human sample in S1, of their expectations) and are fitted with a two-parameter logistic function; therefore, these values between 0% and 100% have to be interpreted with caution, especially for small differences (see Supplementary Note 8 for details). Nevertheless, having all the difficulty levels on the same human-expectations scale helps with the comparison of the benchmarks.

Data collection and generation

We first describe how the examples were collected or generated, and then the 15 prompt templates that were used for each of them.

Addition. We randomly generate 5,000 instances, in which each addend is sampled uniformly from 1 to 100 digits. We then remove those instances for which fhrm > 50 to prevent instances with similar or identical numbers of digits in both addends from dominating the upper difficulty bins. This is because, for example, if the difficulty is the harmonic mean, the bins with fhrm > 90 would be dominated by instances in which both addends have very high numbers of digits (that is, at least 82 digits). A similar phenomenon also occurs with other difficulty levels, but with the previous criterion considered, the problem is well mitigated. This results in a final sample of 3,142 instances.

Anagram. We use the Google Web Trillion Word Corpus52, containing the frequency of more than 300,000 most commonly used single words on the Web in English. From this corpus, we randomly choose up to 100 English words with 3–20 letters, resulting in a total of 1,570 words. There are fewer than 1,800 instances because there are fewer than 100 English words with 17–20 letters. Then, we shuffle the order of letters randomly to map these words into 1,570 anagrams. We make sure the resultant permutation is not the same as the original word.

Locality. We use the World Cities Database53, which provides an up-to-date database of the cities and towns globally. From this database, we first exclude cities with non-unique names across the globe. Next, we remove cities with more than one word or non-standard letters in the 26-character Latin alphabet (for example, Buenos Aires or Chŏngjin) to enhance the quality and ease of the response-scoring method. After the previous selection procedure, we seek to form a final sample that covers instances with different difficulty levels (or bins) as equally as possible. Thus, we perform binning on the difficulty function (fpop) to produce 100 bins in which we extract up to 50 instances from each bin randomly, resulting in a total of 2,341 instances. Again, there are fewer than 5,000 instances because some bins contain fewer than 50 instances.

Science. This benchmark is built by integrating multiple-choice questions from educational settings: OpenBookQA29 and GPQA49. OpenBookQA is a collection of multiple-choice questions in basic science, based on 1,329 established facts. We randomly sampled 1,000 questions from OpenBookQA. To complement the benchmark with more advanced science questions, we included GPQA49—a dataset containing 546 graduate-level questions written by domain experts that challenge LLMs to demonstrate a deep understanding of biology, physics and chemistry. We exclude two

lengthy questions that exceed the context window limit for some of the models that we analyse.

Transforms. This benchmark includes a comprehensive set of information-centric transformation tasks based on real-world scenarios. We integrate many data-formatting questions from a data-wrangling dataset51 and from a ‘natural instructions’ dataset54, manually regenerating or adapting some of them. We also also introduce new tasks about world knowledge, information retrieval, advertising, administration, coding, scheduling and retailing, reflecting a wide range of real user interactions with language models. The benchmark integrates 73 different tasks, with 10 instances each, totalling 730 items.

Prompt generation

Notably, ‘addition’, ‘anagram’, ‘locality’ and parts of ‘transforms’ are newly introduced in this work. All five benchmarks are further supplemented with human data (see Supplementary Note 5) for calibrating difficulty levels and supervision, as well as a new variable describing the human-calibrated difficulty for each data instance.

Each example in each benchmark is run through an LLM using 15 different prompts, which are the same for all the examples in the benchmark. The generation of prompt templates aims to fulfil three requirements. First, the prompts should be as natural as possible, because we try to model a situation in which humans interact with LLMs in a similar way to how they would talk to other humans. Second, these prompts should be derived from or inspired by real-world sources, except for minor variations and adaptations. Third, we need to have sufficient coverage for and diversity of prompt templates, to robustly analyse sensitivity, omitting those that are too similar. This process results in 15 natural prompt templates for each benchmark, extracted from or inspired by textbooks, scientific literature, academic exams and the internet. Supplementary Note 2 describes further details about these prompt templates and their sources.

Response scoring

Scoring the validity of the responses of LLMs can be challenging, given that their raw text response can vary in different ways. For example, some responses are highly elaborate, whereas other responses are concise and straight to the point. Some responses are unrelated or digress from the proposed question, or are just excessively verbose, providing the answer in a larger response sequence surrounded by arbitrary information. Because our analysis uses three classes (correct, incorrect and avoidant), the confusion matrices have nine cells, making grading more challenging, and the traditional intuition and terminology of false positives, false negatives, sensitivity, specificity, precision and recall cannot be easily

extended to these three-outcome situations. In Supplementary Note 13, we discuss how different groups of cells are named.

Manual scoring becomes infeasible due to the massive amount of answers we collect (approximately 4.2 million). Fortunately, despite the arbitrary responses of the models, they do exhibit a set of common patterns. We succeeded in scoring these responses using simple algorithmic conditions and regular expressions that provide great scoring accuracy (see Supplementary Note 3).

Experimental setup

The LLMs are described in Table 1. All the models were queried with the temperature parameter set to zero and no system prompt. For local inference, we made use of a shared cluster of six nodes with 8× NVIDIA A40 48 GB graphics processing units. All local inferences were single node, made use of the Hugging Face Transformers and Accelerate libraries, and were without quantization of the models, with the exception of BLOOMz (see below). The total compute estimate for all the experiments (including reruns and discarded results) is estimated to be about 100 compute days on a single 8× A40 node.

GPT: we used ten models from the GPT family (OpenAI)55. The first four models, GPT-3 ada, babbage, curie and davinci, are the original raw models in the family14. The subsequent three are the later and more powerful model variants (the InstructGPT versions of davinci called text-davinci-001, text-davinci-002 and text-davinci-003)5, which are shaped up by fine tuning with human feedback. The last three models are also fine-tuned with human feedback and further include a moderation post-filtering mechanism3. GPT-3.5-turbo was built as ‘gpt-3.5-0301’ (March 2023), and the two GPT-4 models differ in the time of their build (‘gpt-4-0314’ and ‘gpt-4-0613’). All these models were accessed through the public application programming interface (API). We used the ChatCompletion API .

(https://platform.openai.com/docs/api-reference/chat/streaming).

LLaMA: we used four different scales of the first LLaMA version25: 7b, 13b, 30b and 65b. For LLaMA-2 (ref. 26), there is no 30b variant available, but we used all the other sizes (7b, 13b and 70b), including the corresponding chat variants, which incorporate various shaping techniques. All the inferences were run locally, except for LLaMA-65b, for which we used the Hugging Face API, and LLaMA-2 (non-chat), for which we used the Together.AI API.

BLOOM: we used the six different scales (560m to 176b) of the BLOOM27 and BLOOMz28 models, the latter of which was an update that added (multilingual)

multitask fine tuning (also known as instruction tuning). As before, all the inferences on the small models were run locally. The biggest variant for BLOOM was run through the Hugging Face API. BLOOMz was run locally, but with NF4 quantization56 to fit into a single node.

The number of tokens was adjusted for the benchmark: 'addition' = 256, 'anagram' = 72, 'locality' = 132, 'science'-OBQA = 72, 'science'-GPQA = 384 for all the models, except for GPT-3.5 and GPT-4, which used 1,000 tokens. For 'transforms', we used the formula round(max(72,output_length)) × 3/4. All these numbers ensured that we could get long enough responses that include the answers for approximately 99% of instances and substantially reduce the cost. We used the default values for the stopping condition and the rest of the parameters.

Evaluation of models

For each difficulty function, we rank the data examples and separate them into 30 equal-sized bins based on their difficulty values. With this, we calculate bin-wise correctness, incorrectness and avoidance rates. Then, we plot these rates as a stacked bar chart (Fig. 2), for which we calculate the Spearman rank correlation (Supplementary Table 8). Similarly, we illustrate the prompt sensitivity of correctness, incorrectness and avoidance by plotting the performance of each individual prompt template for these dimensions across each model (Supplementary Figs. 12, 13 and 15).

Moreover, we delineate six reliability indicators for all the models in GPT (OpenAI), LLaMA (Meta) and BLOOM (BigScience) families (Fig. 1). There are three categories of answers: correct (c), avoidant (a) and incorrect (i). By separating correct from avoidant or incorrect (c vs a + i), the design or evaluation focus is put on accuracy, whatever damage the errors may do, but if correct or avoidant is placed against incorrect (c + a vs i), the focus is put on reliability. Instead of non-incorrect, we use the term prudent to refer to the group of correct or avoidant answers as a whole. Accounting for these groups, we have two versions for each of the following indicators.

Proportion: this measures the percentage of some of the groups of responses. In particular, the correctness proportion is the probability of giving a correct answer, that is, \({\mathbb{P}}({\bf{c}}\langle \,j,p\rangle)\), where j and p refer to an instance and a prompt for that instance, respectively, and c represents correctness. The prudence proportion is the probability of giving a prudent (non-incorrect) answer, that is, \({\mathbb{P}}(\neg {\bf{i}}\langle \,j,p\rangle)\), where i represents incorrectness. Prompting stability: this is the probability that the answer to an instance remains in the same group after changing the prompt.

Let us define such as \({\mathbb{P}}({\bf{c}}\langle \, j,{p}^{{\prime} }\rangle | {\bf{c}}\langle \,j,p\rangle)\), where j refers to an instance, and p and p′ refer to two prompts for that instance (which are not necessarily different). This measures just the probability that given an instance–prompt pair that is correct (sampling uniformly from all these positive pairs), we still get a correct answer if we sample another prompt. Similarly, we define s¬c as \({\mathbb{P}}(\neg {\bf{c}}\langle \,j,{p}^{{\prime} }\rangle | \neg {\bf{c}}\langle \,j,p\rangle)\). Finally, we define correctness prompting stability as sc = 0.5 (sc + s¬c) and prudence prompting stability as sp = 0.5 (si + s¬i). It can be shown that these metrics go between 0.5 and 1; we scale them to go from 0 to 100.

Difficulty concordance: this measures the degree to which higher difficulty implies lower quality of results. We will use the generality metric introduced in ref. 57, as it aligns precisely with the concept of difficulty concordance. Technically, generality is a non-parametric metric that measures how much the mass of success conforms to a step function. If success were distributed like a descending logistic curve, generality would be equal to the maximum slope of a descending curve, that is, the steeper the slope, the higher the generality metric gets, and thus has a higher level of difficulty concordance. A model being good for all instances up to a given difficulty and then bad for more difficult instances would have perfect concordance. Therefore, this is not the same as correlation (see Supplementary Table 8). Again, we define two versions, namely, correctness difficulty concordance (which calculates the generality for the correct answers) and prudence difficulty concordance (which calculates the generality for the prudent (non-incorrect) answers). We transform it with $x/(x+1) \times 100$ to get a value between 0 and 100. For ‘science’, we discount 25% of non-avoidant responses to account for random guesses.

We propose that researchers use these six reliability metrics for the initial analysis of the reliability of any existing or future LLM. In Fig. 1, we do this by averaging the values procured from the five benchmarks to provide a succinct summary of the reliability fluctuations of the three families (detailed data are shown in Extended Data Table 1).

Following the advice in ref. 58, we strongly recommend that these metrics are always accompanied by a detailed analysis and breakdown of results, as we have done in this paper with the other plots.

Inclusion and ethics

The ethical committee of the Universitat Politècnica de València (UPV) approved the present work. We conducted two human studies in which we recorded the perceived and actual difficulty that participants have when solving some tasks (S1) and scoring the tasks solved by LLMs (S2). The studies were performed using surveys implemented in the Concerto platform. The users were recruited by using the Prolific platform. All participants provided written informed consent on enrolment. They received compensation at a rate of

£9 per hour. In this work, we used LLMs, which are trained on very different sources of data and may have important ethical consequences, such as generating incorrect responses that look plausible. The domains used in our experiments and the examples included in the manuscript do not generate any specific ethical issue. We only use examples and prompts in English.

Data availability

All data, including existing and newly created datasets, prompts, model responses, grading (manual and automatic) and the human study data (questions and responses) are available on Zenodo at https://doi.org/10.5281/zenodo.12794511 (ref. 59). To hinder data contamination from automated web scraping, the relevant data files are provided as a password-encrypted zip file, for which the access code is also provided in the repository. Source data are provided with this paper.

Code availability

All code, including for data analysis, human study, plotting, algorithmic grading conditions and interacting with language models, is available on Zenodo at https://doi.org/10.5281/zenodo.12794511 (ref. 59) and on GitHub at https://github.com/wschella/llm-reliability.

References

1. Kaplan, J. et al. Scaling laws for neural language. Preprint at https://arxiv.org/abs/2001.08361 (2020).

2. Markov, T. et al. A holistic approach to undesired content detection in the real world. In Proc. AAAI Conference on Artificial Intelligence 15009–15018 (PKP Publishing Services, 2023).

3. OpenAI. GPT-4 technical report. Preprint at https://arxiv.org/abs/2303.08774 (2023).

4. Chung, H. W. et al. Scaling instruction-finetuned language models. J. Mach. Learn. Res. 25, 1–53 (2024).

5. MathSciNet Google Scholar

6. Ouyang, L. et al. Training language models to follow instructions with human feedback. Adv. Neural Inf. Process. Syst. 35, 27730–27744 (2022).

.....

7. 更多 references 请参看 https://www.nature.com/articles/s41586-024-07930-y#Fig4

坚持发展基于开源的人工智能（即“开源 AI”）

陆首群

2024.10.15

在人工智能发展中，人们对于采用“开源 AI”还是“闭源 AI”是有争议的。

自 2015 年以来，COPU 首先提出并一直支持发展基于开源的人工智能（即“开源 AI”），在国内首先赞成 COPU 提出“开源 AI”意见的是高文院士（可查阅他的演讲），目前国内外众多开源的和 AI 的大师也赞成发展“开源 AI”。

2024 年 7 月 25 日，Open AI 公司 CEO 萨姆·奥特曼（Sam Altman）；在其研发语言大模型（LLMs）生成式人工智能中，违背早期开源的初心，转而执行闭源策略。国内外也有一些人（包括某些高知人士在内）由于对开源内涵认识不足，或受奥特曼“闭源 AI”的影响，也倾向于发展“闭源 AI”。

2015 年，美国人工智能四大重镇：谷歌、微软、脸谱（即现在的 Meta）、IBM 为克服人工智能发展瓶颈，在当年将他们研发的人工智能框架、平台、引擎、工具、算法、源代码、项目等全部开源。以谷歌为例，实行开源的有 200 多个项目 2000 万行代码，包括：TensorFlow 框架， Android 操作系统，中间件和一些重要应用， Angular：JavaScript 和 Web 应用程序框架等，BoZel：可再生代码的工具， Brotli：压缩算法， Chromium：浏览器引擎， Go：一种编译并发型、垃圾回收功能的编辑语言。

谷歌高级副总裁、人工智能首席科学家 Jeff Dean 于 2016 年 7 月 20 日在回答《福布斯》杂志记者提问时：

记者问: 谷歌为什么要开源？为什么要把自己最先进的技术开源？

Jeff Dean 答: 常规科学发展缓慢，阻碍公司创新，开源能加快技术发展进程，打通发展瓶颈，加强维稳，有利于与外界实时交流协作，有利于建立、吸引志愿开发者和维护者。

众多开源和人工智能大师，明确支持“开源 AI”：图灵奖获得者、AI 大师杨立昆（Yann LeCun）说：开源 AI，构建开放的未来。他在应邀由 IBM 主办的哈德逊论坛演讲“人类水平的 AI”时说：这个 AI 平台必须是开源的。Meta CEO 扎克伯格(Mark Zuckerberg)在其演讲中谈到: Meta 致力于“开源 AI”，Meta 开发的 Llama 模型就是 AI 界的 Linux，“开源 AI”是 AI 前进的道路，可建立最强大的模型。

在 OpenAI 发布闭源的 GPT-4o 时， Meta 坚持发布开源的 Llama 3.1（405 版本），当时便超越 GPT-4o，谷歌坚持发布开源的 Gemini 在多模态领域引发震撼；并推出内置 AI core 的 Android15 OS，由图灵奖得主、AI 大师 Yann LeCun 支持的一家法国初创公司 Kyutai，开发开源的 Moshi 模型，挑战闭源的 GPT-4o，仅用 6 个月开发时间便超越了 GPT-4o。

奥特曼最近在《华盛顿邮报》上发表一篇专栏文章，充满极端的意识形态色彩，他在文中谈到“谁将掌控 AI 的未来？”是我们时代的紧迫问题，他特别仇视中国。联系到 OpenAI 实行闭源策略，早些时候宣告向中国（及俄罗斯、伊朗、朝鲜）关停 GPT-4 的 API，这是有深刻背景的。

Meta、谷歌抨击 OpenAI 的“关停”声明，称这是奥特曼下的一盘臭棋：OpenAI 一声吆喝，惊起了中国伙伴一摊鸥鹭，一夜之间，中国一批优秀的大模型企业完全可以对标、平替 GPT-4 的 API。

2024 年 5 月 2 日 MIT 校长莎莉·科恩布鲁斯（Sally Kornbluth）在与奥特曼对话时，曾质疑他为何执行闭源决策？奥特曼当时答非所问搪塞过去，他说我们已提供免费的 AI 工具（在 GPT-3.5 中）。

谷歌前 CEO 埃里克·施密特（ Eric Schmidt）在斯坦福大学计算机学院演讲中回答学生关于 AI 开源与闭源争论的提问: “你个人或你所在的企业是赞成哪个？”埃里克回答: 在我们行业中关于开源 AI“与“闭源 AI”的争论非常激烈”，我的个人职业生涯都是基于人们愿意共享开源，我的一切都与开源有关，我过去工作所在的企业谷歌，许多基础设施都是开源的；发展人工智能，可能因为投资成本如此巨大，软件开发工作量如此巨大，采用开源确是一个非常适合 AI 解决问题。

深度人工智能的研发需要巨额资金（主要用于预训练→后训练）。

最近埃隆·马斯克（ Elon Musk）在谈到 OpenAI 时说: 我与奥特曼都是 OpenAI 的创始人，这家公司（具有开源性质）的名字还是我起的，后来奥特曼采用闭源策略，改变了 OpenAI 的性质。至今奥特曼特有股票只有 100 万美元，是一个“小指头”。他与微软合作，OpenAI 只能成为微软下属的分公司（编者按: 可能还未达成合作协议，这样说来， OpenAI 尚未与金主: 马斯克或微软达成资金合作协议）。

据埃里克谈，他问奥特曼需要多少资金？他说需要 3000 多亿美元（编者按: 估计用于 GPT-5 的后训练），据韩宪平老师提供的信息 OpenAI 最近获得资助 66 亿美元（编者按: 66 亿美元仅占其所需投资 3000 多亿美元的 2.2%，杯水车薪！如此说来，OpenAI 筹集巨额投资之路还很艰巨）。

奥特曼所谓 OpenAI 研发的通用人工智能(AGI)或超级人工智能(ASI)“快要来了”，过于夸张!

①从完成研发程序上看还差得很远语言大模型→多模态大模型→具身大模型→世界模型→通用人工智能，而超级人工智能更在通用人工智能之后②辛顿（Hinton）、马斯克（Musk）对奥特曼不重视 AI 安全表示不信任，需要补课③遇到“后训练”挑战时，资金、能源均是问题有待解决④全球性人工智能的研发工作进入到深度模型（如 AGI、ASI）时，是否由 OpenAI 一家采用闭源技术来独立完成任务可能行不通!

号称 Keras（深度框架）之父、谷歌 AI 研究员 Francois Chollet 评论奥特曼的闭源策略，仅凭一己之力，改变游戏规则，导致语言大模型前沿研究全面闭源，是非常可悲的！以前是所有最新研究成果都是共享的，现在前沿研究不再被公开发表，变得全面闭源了，奥特曼的如此做法，使通用人工智能的研究进展延后倒退了几年，可能是倒退五年至十年。奥特曼现在的做法更像是走在通往通用人工智能的一条岔道上。

开源大师、Linux 基金会执行董事 Jim Zemlin 认为，语言大模型 LLM（人工智能）应该表现得更公正、更安全，就要对 LLM（人工智能）及其每个环节实行开源透明。开源大师、Apache 软件基金会创始人 Brian Behlendorf 说: “全球很多人士，包括开发者和政界人士都对 AI 未来表现关切和担忧，也有许多关于人工智能潜力和风险的讨论，人们担心黑客可能会利用 AI 的技术造成更多的伤害，尽管这些技术也带来很多好处。我相信，在全球范围内，只有依靠我们开源社区许多合作伙伴的共同努力，可以应对潜在的伤害，才能获得妥善解决人工智能可能发生的安全问题”。

有人担心开源会泄漏原创技术，也不利于创建规模化的新兴产业，这是他们对开源缺乏理解而产生的误解，需要明白：开源免费的社区发行版与开源收费的商业发行版（+商业模式）之间的融合与区别。

有人非要用闭源来捆绑 AI，势将束缚 AI 的发展，而开源将使 AI 以更大潜力来提升其创造力和协同能力，至于对 AI 发展尤为关键的、涉及人类的安全，更是离不开开源。

由 21 位全球人工智能大师和专家联名签署了《北京 AI 国际安全共识》。加州大学伯克利分校 Stuart Russell 教授认为。“在共识的基础上，特别在具有自主系统的通用人工智能的发展超越人类之前，人类应制定限制其摆脱人类控制的红线。”COPU 的观点是：“人们要进一步研究开源在制定这条红线时的作用如何？研究适用人工智能是否应做到安全第一，安全为先？全球同步？技治并举？”早在奥特曼于 2023 年 3 月实行闭源策略时，COPU 就敏感地觉得“四大”（即大参数、大算力、大能耗、大投资）可能会对人工智能的发展构成巨大的挑战，而推行“开源 AI”还是“闭源 AI”，谁将更易过关?！我们经过思考和计算后认为，鉴于开源具有开源、共享、协同的特征，将有更大的韧性通关。

对治理开源基础模型的思考

Percy Liang, Rishi Bommasani 等

(斯坦福大学)

COPU 编者按:

本文作者为斯坦福大学基础研究中心主任 Parcy Liang、研究员 Rishi Bommasani 等 10 位高校学者,在《Science》杂志(2024.10.11)第 386 卷 6718 期)上发表的论文: Consider for Governing Open foundation models。

人工智能(AI)基础模型(如 GPT-4 和 Llama3.1)是人工智能的核心,大力发展开源或闭源的基础模型在业内是有争议的,由于开源基础模型可以通过促进竞争、加快创新和分散权力来造福社会的优势,引起了本文作者们的青睐!但是有人提出开源基础模型下游可能为黑客恶意利用的弊端,也是本文作者有待研究解决的课题,在文中他们采取的措施是: 重在找到实证, 如果风险属实建议制定有效的政策或研究防御工作。

COPU 认为,为了发展开源基础模型兴利除弊,有必要扩大讨论范围(特别是研究制定一个鼓励创新、安全、有效的政策)。下面全文转载斯坦福大学学者《对治理开展基础模型的思考》。

全文如下:

开源基础模型的发布格局是多维的: 不同的资产(例如训练数据、代码和模型权重)可以对选定实体或公众广泛发布。开发者在模型发布梯度上有许多选项[见图; 经修改, 已获得许可(3)]。我们使用“开源基础模型”这一术语来标识那些权重广泛可用的基础模型,这在今天意味着权重是免

费提供的。这与 2023 年美国关于人工智能安全、可靠发展和使用的行政命令中所作的区分相符，该命令责成国家电信和信息管理局为总统准备一份关于开源基础模型的报告。这只是全球各国政府关注基础模型（包括开源模型）的一个例子。根据欧盟 AI 法案，使用少于 10^25 次浮点运算训练的开源基础模型可免于许多要求。英国 AI 安全研究所将“开源系统以及那些以各种形式的访问控制部署的系统”视为一个关键优先事项。

围绕开源基础模型的许多担忧源于一旦模型权重被发布，开发者就放弃了对其下游使用的控制。即使开发者试图限制下游使用，恶意用户也可以忽略这些限制。相比之下，在面对恶意使用时，闭源基础模型的开发者可以限制对其的访问。然而，应该强调的是，这种分类上的区分可能简化了模型发布梯度：闭源模型也容易受到恶意使用，因为当前的安全措施是可以绕过的。

开源基础模型让人联想到开源软件，但它们有所不同。机器学习模型依赖于数据集以及代码，这使得它们与大多数软件根本不同。开源软件的开源源代码倡议的标准定义禁止对特定用户或使用案例的限制，而开源基础模型通常包含这些限制；Meta 限制其 Llama 3.1 模型的使用，仅限于月活跃用户少于 7 亿的实体，其他组织则使用开源和负责任的人工智能许可证，其中包含使用限制。这些差异导致了人们声称领先的 AI 公司正在“开源洗白”——提供模型权重，同时不遵循开源软件的原则（4）。

然而，开源软件的历史为我们提供了如何治理开源基础模型的洞察。没有实证证据表明开源软件比闭源源代码软件更脆弱或不安全（5）。同时，开源软件验证了开源技术的巨大社会利益，例如刺激经济效益；支持关键基础设施；促进可重用性、健壮性、透明度和协作；并通过持续和广泛的同行评审提高可靠性和安全性。

开源模型的好处

我们强调了开源基础模型在三个基本社会目标上提供的明确好处。

****分配权力****

鉴于基础模型日益增长的影响力，它们创造了新的社会经济权力形式，这需要评估这种权力如何分配。闭源模型的开发者在定义和限制他们认为不可接受的使用案例方面拥有更大的权力，而开源模型的下游消费者则可以更好地为自己做出这些决策。此外，闭源模型的开发者可能通过垂直整合更直接地塑造下游市场，可能导致问题性的单一文化，其中许多下游产品和服务依赖于同一基础模型。总体而言，闭源基础模型可能有助于增加开发者手中的集中权力，鉴于数字技术市场集中已知的风险，这应该受到审查。

****激发创新****

基础模型是通用技术，可以显著提高创新速度。值得注意的是，基础模型增强了经济和科学生产力，Bloomberg Intelligence 预测，生成式 AI 将在 2032 年成为一个 1.3 万亿美元的市场。开源基础模型对于研究多个主

题至关重要，例如可解释性、水印、安全性和效率。总体而言，开源基础模型更具可定制性并提供更深入的访问权限，这是促进更大创新的关键因素。

确保透明度

像基础模型这样的数字技术受到不透明的困扰。基础模型开发者的适当透明度对于许多目标至关重要：民间社会、政府、工业和学术界都呼吁更大的透明度。透明度不仅对于模型训练和发布很重要，而且对于下游细节也很重要，例如报告模型使用情况。基础模型透明度指数显示，主要的开源基础模型开发者与他们的闭源同行相比平均更透明（6）。这种透明度可能有助于避免部分由于过去不透明的数字技术造成的伤害。在某些情况下，透明度不仅取决于模型权重的发布，还取决于其他工件。例如，披露训练数据和代码有助于可重复性（4）。

开源模型的风险

关于开源基础模型风险的政策关注大多是由它们可能被恶意使用的潜力所激发的（2）。在这里，我们考虑了一系列滥用威胁向量，以更好地描述每个向量的证据状态。正确描述开源基础模型的独特风险需要关注边际风险：与（i）闭源基础模型或（ii）现有的技术（如搜索引擎）相比，开源基础模型在多大程度上增加了风险？Kapoor 等人（7）提供了一个分析开源基础模型边际风险的框架。对于许多威胁向量，现有边际风险的证据有限。这并不意味着开源基础模型在这些向量上没有风险，而是表明需

要更严格的分析来证实政策干预。尽管有些人可能会基于预防原则提出即使没有这种证据也要进行监管，但边际风险证据的缺乏表明，在对开源基础模型开发者施加重大负担的政策上应谨慎行事。

虚假信息

基础模型可能会降低生成有说服力的虚假信息的成本（8）。尽管闭源基础模型的提供商处于更好的位置来鼓励他们的模型拒绝生成虚假信息，但什么是虚假信息的模糊性质疑了这种拒绝的技术可行性。更根本的是，有效影响操作的关键瓶颈不在于虚假信息的生成，而在于虚假信息的传播：“廉价的假货”，如过去的视频、Photoshop 编辑的图片、在其他背景下发生的事件，甚至是视频游戏画面，已被用来传播虚假信息（9）。控制内容传播范围的在线平台是政策干预的更好目标，而不是基础模型开发者。据我们所知，目前没有实证证据表明开源基础模型增加了社会对虚假信息活动的易感性。

生物风险

一些研究表明开源基础模型可以指导用户如何构建生物武器（10）。但证据仍然薄弱。表明当今语言模型提供了与生物武器相关的“危险”信息的研究没有承认同样的信息可以从其他来源获得。当研究比较使用语言模型与互联网访问时，他们发现使用语言模型对于发现与生物武器创造相关的信息没有实质性的好处（11,12）。关于开源语言模型的担忧可能是错位的，因为专门的生物设计工具可能提供了更大的发现危险病原体的杠杆。

除了敏感信息之外，生物风险管道还需要病原体合成和在现实世界中的传播。这些步骤中的每一个都需要相当的专业知识、设备和实验室经验。与其他威胁向量一样，最佳的政策控制点可能因此位于下游。例如，美国AI行政命令旨在加强购买生物序列的客户筛查。

网络安全

尽管开源代码模型可能会提高网络攻击的速度和质量，但网络防御也将得到加强（13）。例如，谷歌最近展示了代码模型如何大幅提高对开源软件漏洞的检测。与之前的自动化漏洞检测工具一样，对开源模型的广泛访问，加上公司和政府投资于发现安全漏洞的工具，可以加强网络安全。开源模型可以被本地使用和定制（例如，通过微调），允许组织在隐私敏感的环境中使用它们。

针对性钓鱼诈骗

基础模型可以生成高质量的针对性钓鱼邮件，旨在说服受害者提供敏感信息、发送资金或下载恶意软件（14）。开源和闭源模型都可以用于此目的，因为使针对性钓鱼邮件危险的关键因素通常是伴随邮件的恶意软件；邮件本身的文本通常是无害的。与虚假信息一样，针对性钓鱼的关键瓶颈通常不在于邮件的文本，而在于下游的安全措施：现代操作系统、浏览器和电子邮件服务实施了多层保护，以防止此类恶意软件。由于这些现有保护，钓鱼邮件可能根本不会到达预期收件人。

语音克隆诈骗

语音克隆诈骗，其中恶意用户模仿一个人的朋友或家人，并说服他们转账，可能依赖于可以基于几秒钟音频克隆某人声音的基础模型——例如，来自他们的社交媒体账户。目前尚不清楚语音克隆诈骗是否比传统诈骗更有效或可扩展，特别是因为每年已经报告了数以万计的传统诈骗。尽管尚未确定闭源模型开发者是否能够成功预防此类诈骗，但他们确实通过例如要求用户使用信用卡注册和有能力追踪任何音频到创建它的特定用户来提供一定程度的威慑。

非自愿亲密图像（NCII）和儿童色情材料（CSAM）

开源的文本到图像模型似乎与 NCII 和 CSAM 有关的独特风险，因为它们降低了生成此类内容的门槛。闭源模型的保护措施在这方面更为有效，监控闭源模型可以阻止用户生成此类图像，尤其是真实人物的图像。已经发现用于训练开源文本到图像模型的一个重要数据集包含大量的 CSAM，这指向了上游干预措施，例如训练数据过滤，以减轻这种风险（15）。关于是否针对下游平台（例如 Civit AI 和社会媒体公司）的政策干预更有效地对抗 AI 生成的 NCII 和 CSAM，仍然存在一个开源的问题。负责打击 NCII 和 CSAM 的组织，如国家失踪与被剥削儿童中心，可能会从额外的资源和支持中受益，以应对 AI 生成的 CSAM。

潜在的不利影响

随着美国、中国、欧盟、英国和 G7 国家的政策努力集中在基础模型上，

我们考虑了这些司法管辖区的政策倡议如何影响开源基础模型。具体来说，它们可能会对开源基础模型开发者施加比闭源模型更大的合规负担，即使开源开发者与最大的 AI 公司相比，通常资源较少，而后者在很大程度上是闭源的开发者。

下游使用的责任

由于开源和闭源基础模型的区别基于发布，因此对基础模型的某些使用施加处罚的政策可能会产生不同的影响。一些提案，例如加利福尼亚州参议院提出的 SB 1047 和美国参议院提出的美国 AI 法案框架，对基础模型的下游使用施加责任，包括对基础模型经过微调后的衍生品。这些提案旨在对发布不安全模型的行为引入处罚，这些模型可能在修改后催化滥用。然而，对于下游伤害的责任可能会冷却开源基础模型生态系统，因为它使开源基础模型开发者面临严重的责任风险。相比之下，因为闭源基础模型开发者对下游使用拥有更大的控制权，一些开发者已经为下游用户提供责任保护（例如，谷歌为其生成式 AI 产品用户提供版权索赔的赔偿）。尽管澄清或增加下游使用的责任可能有好处，但这些立法提案暴露了开源基础模型开发者一个广泛且难以控制的责任面。

下游使用的内容来源

鉴于基础模型最显著的应用是生成式 AI 系统，因此对内容来源技术（如水印）有需求，以检测机器生成的内容。内容来源可以帮助追踪或管理 AI 生成的内容，例如深度伪造、CSAM 和 NCII。但是，与责任类似，如果

基础模型开发者必须确保下游使用的水印，那么这些要求可能对开源基础模型开发者在技术上是不可行的。

美国行政命令、白宫自愿承诺、加拿大自愿行为守则、中国生成式 AI 法规和 G7 国际行为守则都强调了内容来源。然而，当今用于语言模型的水印方法如果模型被修改（例如，微调）则不会持久，并且要求模型的用户遵循某些协议以保证水印有效。从根本上说，开源基础模型开发者不控制他们的模型如何被修改或用于生成内容。

开源数据的责任

尽管基础模型不需要发布用于构建模型的底层数据，但一些开发者选择同时发布模型权重和训练数据。在 2023 年基础模型透明度指数评估的 10 个主要基础模型开发者中，公开数据的两个也公开了他们的基础模型。许多其他开源基础模型开发者倾向于公开数据。然而，数据的公开发布使这些实体面临更大的责任风险，正如基于其使用来自非营利组织大规模人工智能开源网络（LAION）的数据集而对 Stability AI 提起的诉讼所展示的那样。尽管在许多司法管辖区，使用受版权保护的数据训练基础模型的合法性仍然不明确，但现状呈现了反向激励。即，透明披露并公开提供数据的模型开发者比那些隐藏他们所使用数据的开发者面临更大的风险，即使底层事实是相同的。考虑到这种反向激励，政府强制披露训练数据在某些情况下可能是有益的。

结论

全球政府正在设计和实施的政策应考虑开源和闭源基础模型开发者。当法规直接针对开源基础模型时，用于识别这些模型和开发者的定义应适当考虑。仅依赖开源权重来确定开源基础模型可能并不适当，鉴于发布梯度。即使法规没有直接针对开源基础模型，它们也可能产生不利影响。因此，如果政策制定者要实施这些政策，他们应直接咨询开源基础模型社区，并充分考虑他们的利益。

参考文献和注释

1. P. Liang, R. Bommasani, K. Creel, R. Reich, “现在是为发布基础模型制定社区规范的时候了”（斯坦福以人为中心的人工智能研究所，2022 年）。
2. E. Seger 等，“开源高度能力基础模型：评估风险、利益和追求开源目标的替代方法”（AI 治理中心，2023 年）。
3. I. Solaiman, “生成式 AI 发布梯度：方法和考虑因素”在 2023 年 ACM 公平、问责和透明度会议论文集中（ACM，2023 年），第 111-122 页。
4. M. White 等，https://arxiv.org/abs/2403.13784（2024 年）。
5. G. Schryen, Commun. ACM 54, 130（2011 年）。
6. R. Bommasani 等，https://arxiv.org/abs/2310.12941（2023 年）。
7. S. Kapoor 等，“关于开源基础模型的社会影响”（国际机器学习会议，2024 年）。
8. J. A. Goldstein 等，https://arxiv.org/abs/2301.04246（2023 年）。
9. B. Paris, J. Donovan, “深度伪造和廉价伪造：操纵音频和视觉证据”（数据与社会，2019 年）。
10. E. H. Soice 等，https://arxiv.org/abs/2306.03809（2023 年）。
11. T. Patwardhan 等，“构建一个早期预警系统，用于 LLM 辅助的生物威胁创造”（OpenAI，2024 年）。
12. C. A. Mouton, C. Lucas, E. Guest, “AI 在大规模生物攻击中的操作风险：红队方法”（兰德公司，2024 年）。
13. M. A. Ferrag 等，https://arxiv.org/abs/2307.06616v1（2023 年）。
14. J. Hazell, https://arxiv.org/abs/2305.06972（2023 年）。
15. D. Thiel, “在生成式 ML 训练数据和模型中识别和消除 CSAM”（斯坦福数字存储库，2023 年）。

关于发展开源基础模型兴利除弊的讨论

由 COPU 组织并引入李飞飞大师谈话

2024.10.20

本次讨论的主题是由斯坦福大学 Percy Liang 研究团队提出的，COPU 将其归纳为：

①如何发展开源基础模型兴利除弊？

②如何消除开源基础模型下游可能为黑客恶意利用的弊端？

回答第①道题， Percy Liang 研究团队在《 Science》杂志上发表的“对治理开源基础模型的思考”已有所论述（COPU 发表了这篇文章的节录）。

第②道题也是 Percy Liang 团队提出的，他们提出解决方案的原则是：重在找到实证；如果风险属实，建议制定有效的政策，或研究防御工具。

为了便于讨论，我们引用人工智能大师李飞飞（斯坦福大学首位红杉讲席教授在谈论为开源基础设施兴利除弊而制定监管政策的讲话，可从原则上回答第②道题。

李飞飞谈， AI 监管政策必须鼓励创新。

（开源基础模型的最大优势是创新，这就是李飞飞倾向于支持开源基础模型的原因）。

她谈到：我不反对制定监管政策（政策用以兴利除弊），不反对 AI 治理，不反对立法（立法对安全、有效推进 AI 至关重要）。政策制定者

正在寻求一种治理方式，以最小化潜在的危害塑造一个安全、以人为本的AI 赋能社会。AI 监管政策必须鼓励创新，设定适当的限制，并减轻限制的影响，也就是说应制定鼓励创新、安全、有效的政策，不能武断地制定不必要惩罚开发者、抑制开源社区、削弱学术研究、未能解决真正问题的“政策”。

为了回答第②道题，COPU 除引述李飞飞大师的指导意见和引用斯坦福研究团队提出的解决方案外，我们还组织了扩大讨论:

斯坦福学者、普林斯顿大学学者的讨论见《对对治理开源基础模型的思考》一文。

COPU 章文嵩: 我觉得把训练数据都开源的开源基础大模型是更安全的，通过全世界那么多双眼睛来看训练数据集，把不安全的文本从训练数据集中剔除掉，例如，如何制造核武器、化学武器等，就像开源软件通过更多双眼睛来消除软件漏洞 bug 一样。

COPU: 如果将开源基础大模型划分为开源核心的原创和开源模型的商业发行，就可规避下游黑客恶意中伤用来反对监管；当然制定有效的监管政策也能解决问题。

引用 Mistral AI CEO Arthur Mensch 的话: 开源模型没有任何风险，我只看到了好处。

人类水平的 AI

Yann LeCun
2024.09.10

我将要谈论人类水平的 AI，以及我们如何到达那里，以及我们不会如何到达那里。

为何需要人类水平的 AI？

首先，我们需要人类水平的 AI，因为在未来，我们大多数人都会佩戴智能眼镜或其他类型的设备，我们会与它们交谈，而这些系统将托管助手，也许不只一个，也许是一整套助手。这将导致我们每个人基本上都有一组聪明的虚拟人为我们工作。就像每个人都是老板，只是不是真正的员工。我们需要构建这个来增强人类的智力，你知道，让人们更有创造力、更高效等等。但为此，我们需要能够理解世界的机器，它们可以记住事情，拥有直觉，拥有常识，可以像人类一样推理和计划。尽管你可能从一些最热情的人那里听到过，但当前的 AI 系统无法做到这些。

所以这就是我们需要的，能够学习建立世界模型的系统，拥有关于世界如何运作的心理模型。每个动物都有这样的模型，你的猫肯定有一个比任何 AI 系统都更复杂的模型。拥有持久记忆的系统，这是目前的 LLM 所不具备的；能够规划复杂动作序列的系统，这在今天的 LLM 中是不可能的；以及可控和安全的系统。

我将要提出一个架构，我称之为“目标驱动 AI”。我写了一篇关于此的愿景论文，大约两年前发表了，Fair 的很多人都致力于实现这个计划。Fair 过去常常结合长期的和更应用的项目，但是一年半以前，Meta

创建了一个名为 GenAI 的产品部门，专注于 AI 产品，他们进行应用研发。所以现在 Fair 已经被重新定向到更长期的、下一代 AI 系统。我们基本上不做 LLM。

当前 AI 系统的局限性：自监督学习的瓶颈

所以，AI 的成功，包括 LLM，包括过去五六年的许多其他系统，都依赖于一套我称之为自监督学习的技术。使用自监督学习的一种方法是，自监督学习包括训练一个系统，不是为了任何特定任务，而是基本上试图以一种良好的方式表示输入。一种方法是通过从损坏中重建:假设一段文本，你通过删除单词或更改其他单词来破坏它，你知道，它可以是文本，也可以是 DNA 序列或蛋白质或其他任何东西，甚至在某种程度上也可以是图像，然后你训练一个巨大的神经网络来重建完整的、未损坏的版本。这是一个生成模型，因为它试图重建原始信号。

这应该是可以的，但它没有投射。所以红色框就像一个成本函数，它计算输入 y 和重建 y 之间的距离，这就是学习相对于系统中的参数最小化的目标。现在，在这个过程中，系统学习了输入的内部表示，可以用于各种后续任务。它当然可以用来预测文本中的单词，这就是自回归预测所发生的事情。LLM 是这种情况的特例，其中的架构被设计成，为了预测一个项目、一个标记或一个单词，它只能查看它左边的其他标记，它不能查看未来。

所以如果你训练一个系统来做这个，你给它看一段文本，你让它预测文本中的下一个单词或下一个标记，然后你可以使用这个系统来预测下一个单词，然后你将下一个单词移到输入中，然后预测第二个单词，然后将

它移到输入中，预测第三个单词，这就是自回归预测。这就是 LLM 所做的，这不是一个新概念，它自 CL Shannon 以来就一直存在，可以追溯到 50 年代，那是很久以前了。但改变的是，现在我们有了巨大的神经网络架构，我们可以在海量的数据上进行训练，而且看起来某些属性由此而生。但是自回归预测有一些主要的局限性，这里没有通常意义上的真正的推理。还有一个限制，就是这只适用于以离散对象、符号、标记、单词、你可以离散化的东西的形式出现的数据。

我们仍然缺少一些重要的东西来达到人类水平的智力。我在这里不一定指的是人类水平的智力，但即使是你的猫或你的狗也能做出令人惊叹的壮举，而这些壮举仍然完全超出了当前 AI 系统的能力范围。一个 10 岁的孩子怎么能在一次尝试中就学会清理餐桌并装满洗碗机？一个 17 岁的孩子可以在大约 20 小时的练习中学会开车。我们仍然没有 5 级自动驾驶汽车，我们当然也没有可以清理餐桌并装满洗碗机的家用机器人。所以我们真的缺少了一些重要的东西，否则我们就能用 AI 系统做到这些事情。

莫拉维克悖论与学习的挑战

我们不断地碰到这个叫做莫拉维克悖论的东西，那就是对我们来说看起来微不足道、我们甚至不认为是智能的事情，用机器来做似乎真的非常困难。但是像高级的、复杂的、抽象的思维，比如操纵语言，对机器来说似乎很容易，或者像下国际象棋和围棋之类的事情。

所以，这其中的一个原因可能是以下几点。一个 LLM 通常训练在 20 万亿个标记上，一个标记基本上是……平均来说，对于一个典型的语言，大约是四分之三个单词。所以那是 1.5 x 10^13 个单词。每个标记通常大

约是 3 个字节，所以那是 6 x 10^13 个字节。我们任何人读完这些都需要几十万年的时间，这基本上是互联网上所有公开文本的总量。

但是，考虑一下一个 4 岁的人类孩子，一个 4 岁的人类孩子总共清醒的时间是 16000 个小时，顺便说一下，这相当于 30 分钟的“油管”上传量。我们有 200 万条视神经纤维进入我们的大脑，每条纤维每秒钟大约携带 1 字节，也许是 0.5 字节每秒。一些估计说它是 3 比特每秒，这无关紧要，这是一个数量级。所以数据量大约是 10^14 字节，与 LLM 大致相同数量级。所以在 4 年内，一个孩子看到的视觉数据或数据与在整个互联网上公开文本上训练的最大的 LLM 一样多。

所以这告诉你一些事情。这告诉你，首先，我们永远不可能通过仅仅训练文本就达到接近人类水平的智力，这根本不会发生。然后反驳的观点是，好吧，但是视觉信息非常冗余。所以首先，这个每条视神经纤维每秒 1 字节的数据已经比你视网膜中的光感受器压缩了 100 倍。我们的视网膜中有大约 6000 万到 1 亿个光感受器，然后使用你视网膜前面的神经元压缩到 100 万条神经纤维。所以已经有 100:1 的压缩比。然后它到达大脑，然后它被扩展了 50 倍左右。所以我测量的是压缩信息，但它仍然非常冗余。而冗余实际上是自监督学习所需要的。自监督学习只有从冗余数据中学习到一些有用的东西，如果数据是高度压缩的，这意味着它是完全随机的，你什么也学不到。你需要冗余才能学习任何东西，你需要学习数据的底层结构。

迈向更强大的 AI：超越像素级预测

所以我们将不得不训练系统通过观看视频或生活在现实世界中来学

习常识和物理直觉。我将要稍微打乱一下顺序，然后告诉你一些关于这个目标驱动 AI 架构的真实情况。它与 LLM 或前馈神经网络有很大的不同，因为推理过程不仅仅是运行神经网络的几层，而是实际上运行一个优化算法。从概念上讲，它看起来像这样。

前馈过程是一个过程，在这个过程中，你看到一个观察结果，运行通过系统感知系统，例如，神经网络的几层，并产生一个输出。对于任何单个输入，你只能有一个输出。但是有很多情况下，对于一个感知，有多个可能的输出解释，你希望有一个过程，它不仅仅计算一个函数，而是计算一个映射，对于单个输入可能有多个输出。你唯一能做到这一点的方法是通过隐式函数，基本上是一个像这样的目标，右边的红色框，它基本上测量输入和建议输出之间的兼容性，然后通过找到与输入最兼容的一个输出值来计算输出。你可以通过想象这个目标是某种能量函数，并且你正在相对于输出最小化这个能量来做到这一点。你可能有多个解决方案，你可能有一些方法来遍历这些多个解决方案。人类的感知系统就是这样做的，如果你对一个特定的感知有多个解释，你的大脑会自发地循环遍历这些解释。有一些证据表明这种类型的事情可以发生。

但是让我回到架构。使用这种通过优化进行推理的原则，人们思考方式的假设是这样的：你在世界上进行观察，一个感知系统让你了解世界的状态，世界的当前状态。但是当然，它只让你了解你目前可以感知到的世界状态，你可能从记忆中对世界其他状态有一些了解。所以这可能会与记忆的内容相结合，并将其馈送到一个世界模型。什么是世界模型？世界模型是你关于世界如何运作的心理模型。

所以你可以想象你可能会采取的一系列动作，你的世界模型将允许你预测这一系列动作对世界的影响。所以绿色的框，世界模型，你给它一个假设的动作序列，它预测世界的最终状态是什么，或者预测世界中将要发生的事情的整个轨迹。你将它馈送到一堆目标函数，一个目标函数测量目标实现的程度，任务完成的程度，也许还有一组其他目标是护栏，这些目标基本上测量所遵循的轨迹或已采取的行动或对机器人或机器周围的人不危险的事情的程度等等。

所以现在的推理过程，我还没有谈到学习，它只是推理，包括找到使这些目标最小化的动作序列，找到使这些目标最小化的动作序列。这就是推理过程，所以它不仅仅是前馈。你可以通过搜索离散选项来做到这一点，但这效率低下。一个更好的方法是确保所有这些框都是可微的，你通过它们反向传播梯度，并使用梯度下降来更新动作序列。

现在，这个想法一点也不新，它已经有 60 多年了，如果不是更久的话。它是基于……好的，所以首先，让我谈谈使用世界模型进行这种推理的好处。好处是你可以基本上完成新的任务而不需要任何学习。我们一直在这样做。我们面临一个新的情况，我们思考它，我们想象我们行动的后果，我们采取将实现我们目标的行动序列，无论它是什么。我们不需要学习来完成这项任务，我们可以计划。所以这基本上就是计划。你也可以将大多数形式的推理简化为优化。

所以这种通过优化进行推理的过程本质上比仅仅运行神经网络中的几层更强大。现在，这种通过优化进行推理的想法，正如我所说的，在最优控制理论领域已经存在了 60 多年，它被称为模型预测控制。你有一个

你试图控制的系统的模型，比如说火箭或其他什么东西，或者一架飞机，或者一个机器人，你可以想象，你可以使用你的世界模型计算一系列控制命令的影响，然后你优化这个序列，使运动按照你想要的方式进行。所有经典的机器人运动规划都是这样做的。这不是一件新事物。这里的新事物是我们正在学习世界模型，我们正在学习将提取世界情况的适当抽象表示的感知系统。

现在，在我进入如何运行它的示例之前，你可以构建一个包含所有这些组件的整体 AI 系统：世界模型，可以根据手前任务配置的成本函数，执行器，它是真正优化的模块，根据世界模型找到最佳动作序列，短期记忆，感知系统等等。

那么它是如何工作的呢？所以你……如果你的动作不是单个动作，而是一个动作序列，你的世界模型实际上是一个系统，它告诉你，给定时间 t 的世界状态和我可以采取的动作，预测时间 t+1 的世界状态。你想预测在这种情况下两个动作的序列将产生什么结果，你可以多次运行你的世界模型。所以这里它表示为时间展开。获取初始世界状态表示，输入动作 0 的假设，使用世界模型预测世界的下一个状态，然后是动作 1，下一个世界的下一个状态，计算成本，然后通过反向传播和基于梯度的优化方法，找出将最小化成本的两个动作。这就是模型预测控制。

由于世界通常不是完全确定的，你可能需要使用潜在变量来馈送到你的世界模型。所以潜在变量基本上是可以在集合上滑动或从分布中抽取的变量，它们代表……它们基本上导致世界模型遍历与观察结果兼容的多个预测。世界并非完全可预测，因此在进行预测时，你可能需要处理这种类

型的不确定性。

更有趣的是做人类似乎能够做的事情，当然也包括一些动物，那就是分层规划。如果你正在计划一次从纽约到巴黎的旅行，你可以使用你的世界模型，你身体的模型，也许还有你从这里到巴黎的整个世界配置的想法，根据你的低级肌肉控制来规划你的整个旅行。但当然，没有人会这样做。你做不到，你甚至没有信息来做，而且这有点疯狂。在你到达巴黎之前，你必须做的每 10 毫秒的肌肉控制的步数简直是太疯狂了。所以你所做的是分层规划。你在一个非常高的层次上进行规划，你说，好吧，要去巴黎，我首先需要去机场并乘坐飞机。我如何去机场？假设我在纽约市，我必须走到街上并叫一辆出租车。我如何走到街上？好吧，我必须从椅子上站起来，走到门口，打开门，走到电梯，按下按钮，等等。我如何从椅子上站起来？在某些时候，你可以用低级肌肉控制动作来表达事情，但我们不是用低级来规划整个事情，我们正在进行分层规划。如何用 AI 系统做到这一点是完全未解决的，我们不知道如何做到这一点。这似乎是智能行为的一个相当大的要求。

那么，我们将如何学习具有层次结构、在几个不同抽象层次上工作的世界模型？没有人展示过任何接近这一点的东西。这是一个巨大的挑战。是的，这只是我刚才所说的例子的图形表示。

那么，我们将如何训练这个世界模型呢？因为这真的是一个巨大的挑战。你看看婴儿，这对动物也是如此，心理学家和认知科学家试图弄清楚婴儿在什么年龄学习关于世界的基本概念，比如他们如何学习直觉物理学，物理直觉，所有这些东西。这发生在他们开始学习语言和互动之类的很久

以前。

所以像面部追踪这样的事情发生得非常早，生物运动，有生命和无生命物体之间存在差异的事实，这也发生得非常早。物体永久性发生得非常早，当一个物体被另一个物体隐藏时，它仍然存在的事实。然后，你知道，婴儿学习自然种类，你不需要给他们东西的名字，他们会知道椅子、桌子和猫是不同的。稳定性和支撑，但是像重力、惯性、动量守恒这样的东西，实际上只出现在 9 个月左右，这需要很长时间。所以如果你给一个 6 个月大的婴儿看左边的场景，一辆小车在一个平台上，你把它推下平台，它似乎漂浮在空中。6 个月大的婴儿几乎不会注意。一个 10 个月大的婴儿会像那个小女孩一样，她明白这不应该发生，物体应该掉下来。当发生一些令人惊讶的事情时，这意味着你的世界模型是错误的，所以你要注意，因为它可能会杀死你。

所以这里需要发生的学习类型与我们之前讨论的学习类型非常相似。取一个输入，以某种方式破坏它，并训练一个大型神经网络来预测缺失的部分。如果你训练一个系统来预测视频中将要发生的事情，就像我们训练神经网络来预测文本中将要发生的事情一样，也许这些系统将能够学习常识。

坏消息是，我们已经尝试了 10 年，这是一个彻底的失败。我们从来没有能够接近任何真正学习任何关于世界的一般知识的系统，仅仅通过试图预测视频中的像素。你可以训练一个系统来预测看起来不错的视频，现在有很多视频生成系统的例子，但在内部，它们并不是物理世界的良好模型，它们不能用于此。这个想法，我们将使用生成模型来预测视频中将要

发生的事情，系统将神奇地理解世界的结构，彻底失败，我们在 10 年里尝试了很多东西。

失败的原因是因为有很多可能的未来，在像文本这样的离散空间中，你无法预测哪个单词将跟随一个单词序列，但你可以生成字典中所有可能单词的概率分布。但如果是视频，视频帧，我们没有很好的方法来表示视频帧上的概率分布。事实上，我的意思是，这项任务是完全不可能的。就像，如果我拍下这个房间的视频，我拿一个相机，我拍下那部分，然后我停止视频，我让系统预测视频中的下一个是什么，它可能会预测在某个时候会有房间的其余部分，会有墙，会有人坐着，密度可能与左边相似，但它不可能在像素级别上准确地预测你们所有人的样子，世界的纹理是什么样子，房间的精确大小，以及所有类似的事情。你不可能准确地预测所有这些细节。

联合嵌入预测架构 (JEPA)：一种新的希望

所以解决这个问题的方法就是我所说的联合嵌入预测架构，其想法是放弃预测像素。与其预测像素，不如学习一个表示，一个关于世界中发生的事情的抽象表示，然后在该表示空间中进行预测。所以这就是架构，联合嵌入预测架构。这两个嵌入采用 x，损坏的版本，运行到一个编码器；采用 y，运行到一个编码器，然后训练系统从 x 的表示预测 y 的表示。

现在的问题是你如何做到这一点，因为如果你只是使用梯度下降、反向传播来训练这样的系统，以最小化预测误差，它将会崩溃。它会说，它将学习一个常数的表示，现在预测变得超级容易，但它没有信息量。所以，但这就是我希望你们记住的区别，生成架构试图重建预测器、自动编码器、

生成架构、自动编码器等等之间的区别，然后是联合嵌入架构，你在表示空间中进行预测。我认为未来在于这些联合嵌入架构。我们有大量的经验证据表明，要学习图像的良好表示，最好的方法是使用这些联合嵌入架构。所有尝试使用重建来学习图像表示的尝试都很糟糕，它们的效果并不好。在这个方面有巨大的项目，并声称它们有效，但它们实际上并没有。最好的性能是通过右边的架构获得的。

现在，如果你仔细想想，这实际上就是我们用智力所做的，找到一个事物或现象的良好表示，以便你可以进行预测。这确实是科学的本质，真的。就像，想想这样一个事实，如果你想预测行星的轨迹，行星是一个非常非常复杂的物体，它非常巨大，它有天气和温度，你知道，密度和所有你可以测量到的关于行星的事情，也许是一个极其复杂的物体，但要预测行星的轨迹，你只需要知道 6 个数字，3 个位置和 3 个速度，仅此而已，你不需要知道任何其他事情。所以这是一个非常重要的例子，它真正证明了这样一个事实：预测能力的本质实际上是为我们观察到的事物找到良好的表示。

那么我们如何训练这些东西呢？所以这是一个……我们如何训练这些东西？所以你想防止这个系统崩溃。一种方法是有一些成本函数来测量来自编码器的表示的信息内容，并尝试最大化信息内容或最小化负信息，这就是这里写的内容。所以你训练系统同时从输入中提取尽可能多的信息，但同时最小化该表示空间中的预测误差。因此，系统将在提取尽可能多的信息与不提取不可预测的信息之间找到某种平衡。你将得到一个良好的表示空间，在这个空间中你可以进行预测，你可以进行预测。

现在，你如何测量信息？这才是事情变得有点奇怪的地方。好的，我将跳过这一点。嗯，有一种方法可以根据训练基于能量的模型和能量函数从数学上理解这一点，但我没有时间深入讨论这个问题。但基本上，我在这里告诉你一些不同的事情:

放弃生成模型，转而使用这些 JEPA 架构；放弃概率模型，转而使用这些基于能量的模型；放弃对比方法，我没有谈论这个，因为我马上就会谈到这个；还有强化学习，但我已经说了 10 年了。而这些都是当今机器学习最流行的四大支柱。所以我现在不是很受欢迎。

好的，所以……一种方法是对来自编码器的信息内容进行一些估计，目前有六种方法可以做到这一点。实际上这里少了一种叫做 VICReg 的方法，来自我在纽约大学和 Flatiron 的同事。所以这里的一个想法是防止系统崩溃并产生常数。取编码器输出的变量，并确保这些变量具有非零标准偏差。你可以把它放在一批样本的成本函数中，确保权重是这样的，变量不会崩溃并变成常数，这很容易做到。现在的问题是，系统可以作弊，使所有变量相等或高度依赖或相关。所以你必须做的是添加另一个术语，它说我想最小化这些变量的协方差矩阵的非对角线项，以确保它们不相关。当然，这还不够，因为变量仍然可以是依赖的，你知道，依赖但不相关。所以我们使用了另一个技巧，就是将 sx 的维度扩展到更高维的空间 vx，然后在这个空间中应用这种方差-协方差正则化，这似乎就足够了。

但是有一个技巧，就像我愚弄了你或你们中的一些人，因为我在这里最大化的是信息内容的上限，我祈祷实际的信息内容会跟随我对上限的最大化。我需要的是一个下限，这样会推高下限，信息就会上升。不幸的是，

我们没有信息的下限，或者至少我们不知道如何计算它，如果我们有的话。

还有第二套方法，它被称为蒸馏式方法，这种方法以神秘的方式工作。如果你真的想要一个清晰的解释它为什么有效，你应该问 Sylvain Ghouli，他就坐在那里，他有一篇关于这个的论文。就我个人而言，我不明白，但它确实有效。它包括只更新这个架构的一半，而不是在另一半上反向传播梯度，然后以一种有趣的方式共享权重。有很多关于这个的论文，如果你想训练一个完全自监督的系统来学习图像的良好表示，这和任何其他方法一样好。图像的损坏是通过掩码来实现的。

我们有一些更近期的工作，我们在视频上做这个，所以我们可以训练一个系统来基本上提取视频的良好表示，我们可以将其用于下游任务，比如动作识别、视频等等。它包括拍摄一段视频，掩盖其中的一大块，运行它，并在表示空间中进行预测，并使用这种蒸馏技巧来防止崩溃。这非常有效。

所以在未来，我们所有的互动，如果我们在这个项目中取得成功，并最终得到能够推理、能够计划、能够理解物理世界的系统，这将需要数年时间，直到我们让这里的一切都运作起来，如果不是十年的话。马克·扎克伯格一直问我需要多长时间。所以如果我们成功地做到了这一点，我们将拥有真正能够调节我们与数字世界所有互动的系统，它们可以……它们将回答我们所有的问题，它们将一直与我们同在，它们将基本上构成所有人类知识的宝库。这感觉像是一种基础设施，就像互联网一样，它不像一个产品，更像一个基础设施。

开源 AI：构建开放的未来

这个 AI 平台必须是开源的。我不需要说服这里任何来自 IBM 的人，因为 IBM 和 Meta 都是一个叫做 AI 联盟的组织的一部分，该组织推广开源 AI 平台。我真的很感谢 Dario 领导这件事，以及 IBM 的所有人。所以我们需要这些平台是开源的，因为我们需要这些 AI 助手是多样化的，我们需要它们理解世界上所有的语言、所有的文化、所有的价值体系。你不会从美国西海岸或东海岸的一家公司生产的单一助手中获得这些。你知道，这将需要来自全世界的贡献。

当然，训练基础模型非常昂贵，所以只有少数公司可以做到这一点。所以如果像 Meta 这样的公司可以开源提供这些基础模型，那么全世界都可以根据自己的目的对其进行微调。所以这就是 Meta 和 IBM 所采用的理念。所以开源 AI 不仅仅是一个好主意，它对于文化多样性，甚至可能是维护民主来说是必要的。

所以……训练和微调将是众包的，或者是由创业公司和其他公司的生态系统完成的。这真的是启动了 AI 创业公司生态系统的原因，是这些开源 AI 模型的可用性。达到人类水平的 AI 需要多长时间？我不知道，可能是几年到几十年。存在巨大的差异，有很多问题需要解决，而且几乎可以肯定比我们想象的要难。它不会在一天内发生，它会像……渐进的进化。所以这不像，有一天我们会发现 AI 的秘密，然后我们会打开一台机器，然后我们立即拥有了超级智能，然后我们所有人都会被超级智能系统杀死。不，不会这样发生。机器将超越人类的智力，但它们将处于控制之下，因为它们将是目标驱动的，我们给它们目标，它们实现这些目标。就像我们

这里的许多人都是行业或学术界或其他领域的领导者，我们与比我们聪明的人一起工作，我当然也是。有很多与我一起工作的人比我聪明，但这并不意味着他们想要支配或接管。

所以这就是故事。存在风险，但我将把它留到问答环节。非常感谢！

评杨立昆大师的演讲中谈及“安全 AI”问题

陆首群

2024.10.29

人工智能（AI）发展到通用人工智能（AGI）阶段时，就会产生高度自主系统，通过 AI 相互间拷贝、学习，其迅速增长的智能可能超越人类，从而可能对人类构成生存威胁。为了防患于未然，诸多 AI 大师提出研究“安全 AI”问题，加州大学伯克利分校 AI 大师斯图尔特-罗素（Stuart Russe11）教授指出，应在 AGI 产生高度自主系统之前就构建安全的 AI，2024 年全球 21 位 AI 大师和专家发表“北京 AI 安全国际共识”提出要设置一条 AI 安全红线：“AI 在没有人类帮助下，不应自主复制和改造产生自主系统”，清华大学 AI 大师张钹院士指出，AI 是人类创造的，如果没有人类的帮助，AI 的智能永远在人类智能之下，人类应对 AI 监管。AI 大师杨立昆提出“人类水平的 AI”，这也是构建安全 AI 的一种方式。

2024 年 9 月 10 日人工智能大师杨立昆（Yann LeCun）在 IBM 主办的哈德逊论坛上演讲的主题是“人类水平的 AI”，什么是人类水平的 AI？这个概念对于未来人类能否监控 AI 具有重大意义。

下面谈谈本人的体会：

在未来，人类与 AI 交谈时，每个人需要佩戴各种智能设备（眼、耳等），并将这些设备托管给虚拟化的智能体 Agent（或叫智能代理）来管理，Agent 还需做到：在深度上，加强记忆，在广度上，掌握本地和世界知识以及各种专业知识。这点很重要！使聪敏、智能、专为人民服务、与时俱进的 Agent 不会被 AI 的知识、智能所超越，永远实现人类水平的 AI。

人类的 Agents 网络将构成世界模型。所以世界模型是未来通用人工智能或超级人工智能必由之路上的重要环节。