从“凭感觉编程”到“凭感觉研究”：与OpenAI领导者对谈未来AI发展

第一部分：引言 – 与人工智能前沿架构师的对话

舞台背景

本文不仅仅是一篇访谈记录，它是一个珍贵的窗口，让我们得以窥见两位正在塑造技术未来的关键人物的思想：Mark Chen，OpenAI的首席研究官，以及Jakub Pachocki，OpenAI的首席科学家 。他们的背景尤为独特——两人都曾是顶尖的竞赛程序员，而现在，他们正致力于构建能够在同样领域中超越人类的AI系统。这篇对话记录了他们关于OpenAI最新进展、核心研究理念以及对人工智能未来的深邃思考。

核心追求

贯穿整篇访谈的核心主线，是OpenAI一项宏伟的追求：创造一个“自动化研究员”（automated researcher）。这个目标远不止是开发一个更智能的聊天机器人；它的终极愿景是实现科学发现过程本身的自动化。这一雄心勃勃的目标，为理解他们在模型评估、强化学习以及研究文化建设等方面的所有决策提供了关键的叙事线索。

读者路线图

在接下来的内容中，我们将深入探讨一系列前沿概念。从GPT-5背后的“推理模型”，到具备自主行动能力的“智能体系统”，再到当前AI评测方法的局限性，以及从“编程”到“凭感觉编程”（Vibe Coding），并最终迈向“凭感觉研究”（Vibe Researching）的范式转变。本文将逐段翻译访谈原文，并在每一段后附上详尽的专业解读，旨在为读者，特别是对AI领域充满好奇的高三学生，提供一次深入且富有启发性的思想之旅。

第二部分：完整访谈：翻译与深度解读

2.1 推理模型的黎明 – 解构GPT-5

原文翻译

> The big thing that we are targeting is producing an automated researcher. So automating the discovery of new ideas. The next set of evals and milestone that we're looking at will involve actual movement on things that are economically relevant. And I was talking to some some high schoolers and they're saying, "Oh, you know, actually the default way to code is vibe coding. I I do think you know the future hopefully will be vibe researching."

我们瞄准的宏大目标是创造一个自动化研究员。也就是，自动化新思想的发现过程。我们正在关注的下一组评估和里程碑将涉及在具有经济价值的事务上取得实际进展。我最近和一些高中生聊天，他们说：“哦，你知道吗，现在默认的编程方式就是‘凭感觉编程’（vibe coding）。”我确实

认为，未来有希望会是‘凭感觉研究’（vibe researching）。

> Thanks for coming Jacob and Mark. Jacob, you're the chief scientist at OpenAI. Mark, you are the chief research officer at OpenAI and you guys have the both the uh the privilege and the stress of running probably one of the most high-profile research teams in AI. And so we're just really stoked um to talk with you about a whole bunch of things we've been curious about, including GPD5, which was, you know, one of the most exciting updates to come out of Open in recent times. And then stepping back, how you build a research team that can do not just GPD5 but codeex and chat GPT and an API uh business and can weave all of the many different bets you guys have across modalities, across product form factors um into one coherent research culture and story.

感谢Jakub和Mark的到来。Jakub，你是OpenAI的首席科学家。Mark，你是OpenAI的首席研究官，你们俩共同享有领导可能是AI领域最受瞩目的研究团队之一的荣幸与压力。我们非常兴奋能与你们探讨一系列我们好奇已久的话题，包括GPT-5，这是OpenAI近期最激动人心的更新之一。然后，我们想退一步探讨，你们是如何建立一个不仅能做出GPT-5，还能做出Codex、ChatGPT以及API业务的研究团队，并且能够将你们在不同模态、不同产品形态上的众多不同赌注，编织成一个连贯的研究文化和故事。

> > > And so to kick things off, why don't we start with GPD5? Just tell us a little bit about the GPD5 launch from your perspective. How did it go? So I think GPT5 was really our attempt to bring reasoning into the mainstream and um prior to GPT5 right we have two different series of models you had uh the GPT kind of 2 3 4 series which were kind of these instant response models and then we had an O series which uh essentially thought for a very long time and then gave you the best answer that it could give. So tactically uh we don't want our users to be puzzled by you know which mode should I use and it involves a lot of research in kind of identifying what the right amount of thinking uh for any particular prompt looks like and uh taking that pain away from the user. So we think the future is about reasoning more and more about reasoning more and more about agents and uh we think GPD5 is this step towards delivering reasoning and more agentic behavior by default. There is also a number of improvements across the board in this model relative to O3 um and our previous models but our primary our primary um fees for for this launch was indeed bringing the reasoning mode to more people.

> 那么，为了拉开序幕，我们何不从GPT-5开始呢？请从你们的角度简单谈谈GPT-5的发布情况。进展如何？ 我认为GPT-5是我们真正尝试将“推理”（reasoning）带入主流的一次努力。在GPT-5之前，我们有两个不同系列的 模型：一个是GPT 2、3、4系列，这些是即时响应模型；另一个是O系列，它会进行非常长时间的思考，然后给出它能给出的最佳答案。从策略上讲，我们不希望用户为“我应该使用哪种模式”而感到困惑，这需要大量的研究来识别对于任何特定提示，合适的思考量是多少，从而为用户消除这种麻烦。因此，我们认为未来将越来越关乎推理，越来越关乎智能体（agents），而我们认为GPT-5是朝着默认提供推理能力和更多智能体行为迈出的一步。与O3以及我们之前的模型相比，这个模型在各方面都有许多改进，但我们这次发布的首要目标确实是把推理模式带给更多人。

深度解读

这段开场白直接点出了OpenAI的两个核心战略方向：长远目标是“自动化研究员”，而近期里程碑GPT-5的核心是“推理”。要理解GPT-5的重大意义，首先需要明白什么是“推理模型”（Reasoning Model）。

我们可以借鉴诺贝尔经济学奖得主丹尼尔·卡尼曼在《思考，快与慢》中提出的理论来理解 。传统的GPT-3和GPT-4模型，类似于人类的“系统1”或“快思考”，它们能根据海量数据中学习到的模式，迅速、直觉地生成回答。而访谈中提到的“O系列”则像人类的“系统2”或“慢思考”，它会花费更长的时间，进行一步步的逻辑分析和推导，以解决复杂问题 。这种模型通常采用一种名为“思维链”（Chain-of-Thought）的技术，即在给出最终答案前，先输出详细的解题步骤，就像学生在考试时“写出解题过程”一样，这大大提高了复杂任务的准确性 。

GPT-5的革命性之处在于，它首次将这两种“思考模式”无缝地整合到了一起，并让模型自己来判断何时该用“快思考”，何时该启动“慢思考” 。用户不再需要纠结于选择哪个模型，AI系统本身具备了一种初步的“元认知”能力——它不只是解决问题，而是先*分析问题的难度*，再决定*如何调动认知资源*去解决。

这一转变不仅仅是性能的提升，更是AI产品哲学的一次深刻变革。它标志着AI从一个需要用户手动选择的“工具箱”，开始向一个能自主适应任务复杂度的“智能伙伴”演进。这是迈向真正自主智能体（Agentic Behavior）的关键一步，因为自主性的核心之一就是有效地自我管理和分配认知资源。为了更好地理解GPT-5所处的位置，下表梳理了OpenAI关键模型的演进历程。

表1：OpenAI基础模型的演进历程

模型	发布年份	参数量	关键突破/能力
GPT-1	2018	1.17亿	验证了在Transformer架构上进行生成式预训练的可行性。
GPT-2	2019	15亿	展示了强大的零样本学习（zero-shot learning）能力；能够生成连贯的长篇文本。
GPT-3	2020	1750亿	精通少样本学习（few-shot learning）；只需少量示例即可执行任务。

模型	发布年份	参数量	关键突破/能力
GPT-4	2023	未披露	引入多模态能力（支持文本和图像输入）；显著提升了准确性并减少了“幻觉”。
GPT-5	(访谈中讨论)	未披露	默认集成了“推理”能力，能根据任务自动在快、慢两种思考模式间切换。

2.2 旧尺度的终结 – 饱和的评估与对真实世界挑战的求索

原文翻译

> > > Can you say more about how you guys think about evals? I noticed even in that launch video there were a number of evals where you were inching up from you know 98 to 99% and that's kind of how you know you saturated the eval. What approach do you guys take to measuring progress and and how do you think about it? One thing is that indeed for like these evos that we've been using for the last few years, they're indeed pretty close to saturated and so yeah like uh for a lot of them like you know inching from like 96 to 98% is not necessarily uh the most important thing in the world. I think another thing that's maybe even more important but a little bit subtler when we were in this like GPT2 GPT3 GT4 era um you know there was kind of one recipe you just like pre-train a model on a lot of um data and you kind of like use these um evals as just kind of a yard sick um of u how this generalizes to like different tasks. Um now we have this uh different ways of training in particular uh reinforcement learning on like serious reasoning where we can pick a domain and we can really train a model to like become an expert in this domain to reason very hard about it which lets us um you know target particular uh kinds of of of of tasks uh which will mean that like we can get like extremely good performance on some evolves but it doesn't indicate as great generalization to to other things I think so the we think about it in this world, we definitely think like uh we are in a little bit of a uh deficit like uh of of of great evaluations and I think the big things that we look at are actual marks of the model being able to discover new things. I think for me the most exciting thread and like actual sign of progress this year has been our model's performance in uh math and programming competitions. although I think like they are also becoming saturated in a sense. Um and the next set of evolves and milestone that we're looking at will involve actual um discovery and and actual um movement on on on things that are that are economically relevant.

> 你们能多谈谈对“评估”（evals）的看法吗？我注意到在发布视频里，有很多评估指标你们只是从98%提升到99%，这似乎表明评估已经饱和了。你们用什么方法来衡量进展，又是如何看待这个问题的？ 的确，我们过去几年一直在使用的这些评估标准，确实已经非常接近饱和了。所以，对于其中很多指标来说，从96%提升到98%并不一定是世界上最重要的事情。我认为还有一件可能更重要但更微妙的事：在GPT-2、GPT-3、GPT-4的时代，基本上只有一种方法，就是用大量数据预训练一个模型，然后用这些评估作为衡量其泛化到不同任务能力的标尺。但现在我们有了不同的训练方式，特别是在严肃推理上的强化学习，我们可以选择一个特定领域，然后真正地训练模型成为该领域的专家，对其进行深度推理。这使得我们能够针对特定类型的任务，这意味着我们可以在某些评估上取得极好的性能，但这并不代表对其他事物有同样出色的泛化能力。所以在这个新世界里，我们确实认为我们有点缺乏好的评估方法。我们关注的重点是模型能够发现新事物的实际标志。对我来说，今年最激动人心的线索和实际进展的标志，是我们的模型在数学和编程竞赛中的表现。尽管我认为这些竞赛在某种意义上也在变得饱和。我们正在关注的下一组评估和里程碑将涉及实际的发现，以及在具有经济价值的事务上取得实际进展。

深度解读

这段对话揭示了前沿AI研究领域正在面临的一个深刻挑战：**度量危机**。

首先，让我们来定义“评估”（Evals）。在机器学习中，“Evals”指的是一套标准化的测试或基准，用来衡量一个模型的性能，就像学生的模拟考试一样 。在过去，这些测试是衡量AI进步的有效“标尺”。但Jakub Pachocki指出，这些旧的标尺已经“饱和”了。这意味着最顶尖的模型在这些测试上已经能拿到接近满分的成绩（例如98%、99%），继续提升零点几个百分点，并不能代表模型在真实世界的智能水平上取得了有意义的突破。这好比一个学生每次模拟考都能拿满分，这些考试已经无法衡量他是否具备了做出开创性研究的潜力，只能证明他完全掌握了现有课程。

更深层次的问题在于，随着训练方法的进步，特别是针对特定领域的强化学习，模型可以被“应试”训练，从而在某个评估上表现极好，但这是一种“过拟合”到评测本身的现象，并不代表其具备了广泛的、真正的泛化能力 。这就好比一个学生只刷特定类型的题目，虽然能考高分，但解决新问题的能力并未相应提升。

因此，OpenAI正在寻找新的、更难的“标尺”。他们将目光投向了那些被认为是衡量顶尖人类智慧试金石的领域：**国际数学奥林匹克竞赛（IMO）和AtCoder编程竞赛** 。IMO是全球最顶尖的高中生数学竞赛 ，而AtCoder则是全球性的高难度算法编程竞赛平台 。这些竞赛的特点是，它们没有固定的解题套路，需要参赛者具备深刻的洞察力、创造力和严谨的逻辑推理能力来解决全新的问题。它们测试的不是知识的记忆，而是智慧的运用。

这一转变意义重大。它标志着AI研究的重心正在从“信息检索和模式匹配”转向“真正的发现和问题解决”。当AI能够在这些代表人类智力巅峰的竞赛中与最优秀的人类选手一较高下时，它就离实现“自动化研究员”这一最终目标更近了一步。因为这些竞赛中展现出的能力，正是解决真实世界中那些“具有经济价值”的科学和工程难题所必需的。从根本上说，放弃饱和的旧基准，拥抱开放式难题竞赛，是AI领域为了确保自身仍在朝着正确方向——即发展出能够进行外推式创新（extrapolation）而非内插式模仿（interpolation）的通用智能——而必须采取的关键科学步骤。

原文翻译

> > Totally. You guys already got number two in the at coder competition. So there's really only number one left. Yeah. Yeah. I mean I think it is important to note that these evals like um you know II atcoder IMO um are actually real world markers for success in future research. I think a lot of you know the best researchers in the world have gone through these competitions have gotten very good results um and and yeah I think we are kind of preparing for this frontier where we're trying to get our models to discover new things.

> 完全正确。你们已经在AtCoder竞赛中拿到了第二名，所以真的只剩下第一名了。是的。是的。我的意思是，我认为很重要的一点是，像AtCoder、IMO这样的评估，实际上是未来研究成功的真实世界标志。你知道，世界上很多最优秀的研究人员都经历过这些竞赛，并取得了非常好的成绩。 嗯，是的，我想我们正在为这个前沿领域做准备，我们正努力让我们的模型去发现新事物。

深度解读

这里进一步强调了为什么选择编程和数学竞赛作为新的评估标准。Mark Chen指出，这些竞赛的成功与未来在科研领域的成功有着很强的相关性。这并非巧合。无论是解决一个复杂的数学奥赛题，还是设计一个高效的算法，都需要一种超越死记硬背的能力，包括：

1. **问题分解**：将一个宏大、模糊的问题拆解成一系列可管理的小步骤。
2. **抽象思维**：识别问题的核心结构，并将其与已知的数学或计算机科学概念联系起来。
3. **创造性联想**：在看似无关的领域之间建立联系，找到新颖的解决方案。
4. **严谨验证**：系统地检查和证明自己解决方案的正确性。

这些能力恰恰也是一名优秀科学家或工程师所必备的核心素养。因此，通过训练AI在这些竞赛中取得优异成绩，OpenAI实际上是在一个高度浓缩和可控的环境中，模拟和培养AI进行科学发现所需的核心技能。这就像是为未来的“自动化研究员”进行的一场高强度“模拟训练”。当模型能够独立“发现新事物”——即在这些竞赛中构思出人类未能想到的、新颖且正确的解法时，就意味着它离真正能够推动科学前沿的目标又近了一步。

2.3 AI能力的涌现与惊喜

原文翻译

> > Yeah very exciting. Which capability from GPD5 before the release? surprised you the most when you were working through the eval bench or using it internally? Were there any moments where you felt like this was starting to get good enough to release because it was useful in your daily usage? I think one big thing for me was um just how much it moved the frontier in very hard sciences. Um you know we would try the models with some of our friends who are you know uh professional physicists or professional mathematicians and you already saw kind of some instances of of of this on Twitter where you know you can take uh a problem and have it discover maybe not like very complicated new mathematics but you know um some non-trivial new mathematics and uh you know we we see physicists mathematicians kind of uh repeating this experience over and over where they're trying pro and saying, "Wow, this is something that the you know, previous version of the models couldn't do." And it is a little bit of a light bulb moment for them. It's like uh able to automate maybe like what could take uh one of their students months of of time.

> > 是的，非常激动人心。 在GPT-5发布之前，它的哪项能力最让你们感到惊讶，无论是在进行基准测试还是在内部使用时？有没有某个时刻，你们觉得它已经足够好，可以发布了，因为它在你们的日常工作中变得很有用？ 我认为对我来说，一件大事是它在非常前沿的硬科学领域推动了边界。我们会和一些朋友一起测试模型，他们是专业的物理学家或数学家。你已经在Twitter上看到了一些这样的例子，你可以给它一个问题，让它发现一些——也许不是非常复杂的新数学——但也是一些相当有深度的新数学。 我们看到物理学家和数学家们一次又一次地重复这种体验，他们尝试后会说：“哇，这是以前版本的模型做不到的。”这对他们来说有点像灵光一现的时刻。它能够自动化那些可能需要他们的一名学生花费数月时间才能完成的工作。

深度解读

这段对话揭示了GPT-5最令人振奋的“涌现能力”（Emergent Abilities）——即在没有被专门训练的情况下，模型自发获得了解决高度专业化科学问题的能力。这种能力不再局限于通过标准化测试，而是体现在能够为顶尖科学家提供真实的、有价值的帮助。

当一位专业的物理学家或数学家，在面对一个需要数月研究才能解决的难题时，发现AI可以在短时间内提供一个有效的解决方案或思路，这标志着一个重要的转折点。AI不再仅仅是一个信息检索工具（像搜索引擎），也不再是一个通用的写作助手，它开始成为一个能够参与到知识创造过程中的“初级研究伙伴”。

这里提到的“自动化学生数月的工作”是一个非常具体的衡量标准。在科研领域，一个研究生或博士生花费数月时间去推导一个公式、验证一个猜想或处理一批数据，是非常普遍的。如果AI能够可靠地、高效地完成这类任务，将极大地加速科学研究的进程。科学家们可以从繁琐的智力劳动中解放出来，专注于更具创造性和战略性的工作，比如提出新的研究方向、设计关键实验，以及解读AI得出的结果。

这个“灵光一现的时刻”（light bulb moment）对科学家们来说，不仅是效率的提升，更是思维方式的转变。他们开始意识到，可以与AI进行一种全新的互动：将AI作为一个不知疲倦、知识渊博的“思想碰撞器”，来探索和验证那些过去因耗时过长而不敢轻易尝试的研究路径。

原文翻译

> > Well, GP5 is a is a definite improvement on O3. For for me, 03 was definitely like that moment where the reasoning models became like actually very useful on a daily basis. I think especially for um you know working through a math uh formula or or or a derivation like they like it actually got to a level where it is like fairly trustworthy and and I can actually use it as a as a tool uh for for my work. Um and yeah I think I think uh yeah it is very exciting to get to that moment. Um but I expect that um well now as we're seeing um you know these models like like actually able to automate well yes like like we're saying solving contest problems over over longer time horizons I I I expect that that is well that that that was quite small compared to what's coming over the next year.

> 嗯，GPT-5相对于O3确实是一个明确的进步。对我来说，O3绝对是那个让推理模型在日常工作中变得非常实用的时刻。我认为，特别是在处理数学公式或推导时，它实际上达到了一个相当值得信赖的水平，我真的可以把它作为我工作的工具。是的，我认为能达到那个时刻非常令人兴奋。但我预计，现在我们看到这些模型能够自动化地解决竞赛问题，并且能在更长的时间跨度上进行，我预计这与未来一年将要发生的事情相比，只是小巫见大巫。

深度解读

Jakub Pachocki在这里补充了一个重要的个人视角，强调了AI能力发展的**阶段性**和**加速度**。

对他而言，**O3模型**（GPT-5的前身之一，专注于推理）是第一个跨过“可用性门槛”的模型。这意味着，AI的推理能力首次达到了一个足够可靠的水平，使得像他这样的顶尖科学家可以放心地在自己的核心工作——例如复杂的数学推导——中使用它。这是一个从“玩具”到“工具”的质变。在此之前，AI可能会给出看似正确但细节错误的答案，需要人类专家花费大量时间去验证和修正，使用成本很高。而O3的出现，意味着AI的可靠性达到了一个临界点，使用它带来的收益开始大于验证其结果的成本。

然而，他紧接着指出了一个更令人兴奋的趋势：**进步的速度正在加快**。当前模型能够解决竞赛级难题的能力，虽然已经非常惊人，但在他看来，这仅仅是未来更大突破的“序幕”。这种判断基于他们对AI能力扩展规律的深刻理解。AI的发展不是线性的，而是指数性的。今天看起来是极限的能力，在一年后可能就变成了基础能力。

这段话向我们传递了一个强烈的信号：我们正处在一个技术突破的陡峭曲线上。AI从一个“可靠的工具”进化到一个“能自主解决难题的伙伴”，而下一步，它可能会进化成一个“能独立发现新知识的合作者”。我们目前所见的成就，无论多么令人印象深刻，都可能只是未来更大变革的冰山一角。

2.4 自动化研究的蓝图与智能体之路

原文翻译

> > What is coming in the next one to five years it would be just at whatever level you're you're comfortable sharing what what does the research road map look like? So the big thing that we are targeting with our research is producing um an automated researcher. So auto automating the discovery of new ideas um and you know of course like a particular thing we think about a lot is automating our own own work automating ML research. Uh but that can get a little bit self-reerential. So we're also thinking about automating um progress in in in other sciences. And I think like one good way to measure progress there is looking at like what is the time horizon on which these models actually can um reason and make progress. And so now as we kind of like get to a level of near mastery of this of this um high school competitions let's say I I I would say like we get we get to like maybe on on the order of one to five hours of of of reasoning. Um and and so we are focused on extending that horizon both in terms of like the models um will capability to plan over very long horizons and actually able to retain ability to retain memory.
>
> 在未来一到五年内，你们可以分享一下研究路线图是怎样的吗？ 我们研究的核心目标是创造一个“自动化研究员”。也就是自动化新思想的发现过程。当然，我们思考得很多的一个具体问题是自动化我们自己的工作——自动化机器学习研究。但这可能有点自我指涉。所以我们也在考虑自动化其他科学领域的进展。我认为，衡量这方面进展的一个好方法是看这些模型能够在多长的时间跨度上进行推理并取得进展。现在，当我们差不多掌握了这些高中竞赛的水平时，我想说，我们大概能达到一到五小时的推理水平。因此，我们正专注于扩展这个时间跨度，包括模型在非常长的时间跨度上进行规划的能力，以及实际保留记忆的能力。

深度解读

这段话清晰地阐述了OpenAI的终极研究蓝图和衡量进展的核心指标。

最终目标：自动化研究员（Automated Researcher）。这不仅仅是要自动化重复性劳动，而是要自动化“新思想的发现”这一人类智力活动的顶峰。他们甚至将目标具体化为“自动化机器学习研究”，即让AI来设计下一代AI，这构成了一个潜在的、能够自我加速的智能反馈循环。同时，为了避免目标过于“内卷”，他们也将目光投向了物理、化学、生物等更广泛的科学领域。

核心度量指标：推理的时间跨度（Time Horizon of Reasoning）。这是一个非常深刻且新颖的评估维度。传统的AI评估通常关注“任务准确率”，即一次性解决问题的能力。而OpenAI提出的新指标是，一个AI系统能够在无人干预的情况下，**持续工作多长时间**来解决一个复杂问题。这反映了AI的三个关键能力：

1. **长期规划能力**：面对一个需要数天甚至数月才能解决的宏大目标，AI能否将其分解为一系列合理的子任务，并按计划执行。

2. **记忆与学习能力**：在长时间的探索中，AI能否记住之前的成功经验和失败教训，并根据这些信息动态调整后续策略，而不是重复犯错。

3. **鲁棒性与纠错能力**：当某个尝试失败时，AI能否不“崩溃”或“跑偏”，而是能分析失败原因，并自主尝试新的方法。

目前，顶尖模型在解决竞赛难题时，展现了大约“一到五小时”的持续推理能力。这已经是一个了不起的成就，相当于一个顶尖人类选手在一场比赛中的专注工作时长。而OpenAI的下一个目标，就是将这个时间跨度从“小时级”扩展到“天级”、“周级”甚至“月级”。当AI能够在一个月的时间里，围绕一个科学难题自主工作并取得有意义的进展时，“自动化研究员”的时代就将真正到来。

原文翻译

> > > And back to Eval's question that's why I think eval of the form of how long does this model autonomously operate for are of particular interest to us. And actually maybe on that topic there's been this huge move toward agency and model development. But I think at least the state that it's in currently, users have sort of observed this trade-off between too many tools or planning hops can result in quality regressions uh versus um something that maybe has a little bit less agency, the the quality is at least observed today to be a bit higher. H how do you guys think about the trade-off between stability and depth? the more um steps that the model is undertaking maybe the less likely the tenth step is to be accurate versus you ask it to do one thing it can do it very very well um and to have it keep doing that one thing better and better but more complex things there's sort of that trade-off um but of course to get to full autonomy you are taking multiple steps you're using multiple tools

> > 回到评估的问题，这就是为什么我认为“这个模型能自主运行多久”这种形式的评估对我们特别有吸引力。 实际上，关于这个话题，模型开发正朝着“智能体”（agency）的方向大举迈进。但我认为，至少在目前的状态下，用户观察到了一个权衡：使用过多的工具或进行过多的规划步骤，可能会导致质量下降；相比之下，一个智能体化程度较低的模型，其质量在今天看来反而更高一些。你们如何看待稳定性和深度之间的权衡？模型执行的步骤越多，可能第十步的准确性就越低；而如果你让它只做一件事，它可以做得非常好，并且能把这件事做得越来越好，但对于更复杂的事情，就存在这种权衡。当然，要实现完全的自主性，你必须执行多个步骤，使用多种工具。

深度解读

这里探讨了实现“自动化研究员”所面临的一个核心技术挑战：**深度与稳定性之间的权衡**。

首先，让我们理解什么是“智能体”（Agency）。一个AI智能体不仅仅是被动地回答问题，它还能主动地设定目标、制定计划、调用工具（如代码解释器、网络浏览器），并根据环境反馈来调整自己的行为，以完成一个复杂的、多步骤的任务。这是实现“自动化研究员”的必经之路。

然而，当前的挑战在于，随着任务链条的延长，错误的累积效应会变得非常显著。这就像一个传话游戏，每经过一个人，信息都可能发生一点偏差，传到第十个人时，信息可能已经面目全非。同样，一个AI智能体在执行多步骤任务时，第一步的微小偏差可能会在第二步被放大，到了第十步，整个任务可能已经完全偏离了轨道。这就是所谓的“质量衰减”（quality regression）。

因此，研究人员面临一个两难的抉择：

- **追求深度（Depth）**：让模型执行更多步骤、调用更多工具，以解决更复杂的问题。这是通往完全自主性的方向，但目前风险较高，容易出错。
- **保证稳定性（Stability）**：限制模型的步骤和工具使用，让它专注于做好单一步骤的任务。这样能保证很高的质量和可靠性，但无法处理复杂、长期的任务。

OpenAI的观点是，解决这个问题的关键在于提升模型的**核心推理能力**。一个具备强大推理能力的模型，能够在长链条任务中保持逻辑的一致性。它不仅知道下一步该做什么，还能在每一步之后进行“反思”和“验证”，判断自己是否还在正确的轨道上。当它发现偏离时，它有能力自我纠正。因此，推理能力是连接“深度”和“稳定性”的桥梁，是让智能体能够“行稳致远”的根本保障。

原文翻译

> > I I I think actually like well the well the ability to maintain depth is a lot of it is being consistent over long horizons um so I I think they are very related problems. Um and in fact I think like with the reasoning models we have seen the models like greatly um extend the the the length over which they are able to reason uh and and work um reliably without without going off track. Yeah, I think this is uh this is going to remain a big area of focus for us. Yeah. And I think reasoning is core to this ability to operate over a long horizon because you know you imagine kind of yourself solving a math problem right you try an approach it doesn't work and you know you have to think about you know what what's the next approach I'm going to take um what are the mistakes in the first approach and then you try another thing and you know the the world gives you some hard feedback right and then you keep trying different approaches and the ability to do that over a long period of time is reasoning and gives agents that robustness
>
> 我认为，维持深度的能力很大程度上就是在长远范围内保持一致性。 所以我认为它们是高度相关的问题。事实上，我认为通过推理模型，我们已经看到模型能够极大地扩展它们可以可靠地进行推理和工作的长度，而不会偏离轨道。是的，我认为这将继续是我们的一个重点关注领域。 是的。我认为推理是这种在长远范围内运作能力的核心。因为你可以想象一下自己解决一个数学问题的过程：你尝试一种方法，行不通，然后你必须思考下一步要采取什么方法，第一种方法错在哪里，然后你再尝试另一种方法。 世界会给你一些硬性的反馈，然后你不断尝试不同的方法。在很长一段时间内做到这一点的能力就是推理，它赋予了智能体那种鲁棒性。

深度解读

这段对话用一个生动的比喻——“自己解数学题”——来阐释了为什么“推理”是实现长期、自主工作的核心。

人类解决难题的过程，从来都不是一蹴而就的线性过程。它是一个充满**试错、反思和迭代**的循环。当你尝试一种解法失败时，你不会就此放弃。你会：

1. **分析失败**：识别出第一种方法为什么行不通，错在哪里。
2. **生成新策略**：基于对失败的理解，构思出一种新的、可能更好的方法。
3. **再次尝试**：执行新的方法，并观察结果。
4. **接收反馈**：世界（无论是数学定律还是实验结果）会给你一个明确的、不容置疑的“硬反馈”。
5. **持续迭代**：重复这个循环，直到问题被解决。

这个在失败和反馈中不断调整策略、并长时间坚持下去的能力，其本质就是**推理**。它赋予了智能体一种关键品质——**鲁棒性（Robustness）**，即在面对不确定性和挫折时，系统不会轻易崩溃，而是能够自我修复和适应。

因此，当OpenAI说他们致力于提升模型的推理能力时，他们不仅仅是在提升模型做对题的概率。他们是在构建一种更深层次的、类似人类的坚韧性和问题解决框架。一个真正强大的AI智能体，不应该是一个只能沿着预设路径完美执行的“机器人”，而应该是一个能够在未知和复杂的环境中，像科学家一样不断探索、试错、学习并最终找到出路的“探索者”。这正是推理能力赋予智能体的核心价值。

2.5 从可验证到开放式领域：推理的泛化

原文翻译

> > we talked a lot about math and science um I I curious to get your take on do you think some of the progress that we've made can actually extend um similarly to domains that are less verifiable. They're sort of less of an explicit right or wrong.

> 我们聊了很多关于数学和科学的话题。我很好奇你们的看法，你们认为我们取得的一些进展，能否同样扩展到那些不那么容易验证的领域？那些没有明确对错之分的领域。

深度解读

这是一个非常关键的问题，它触及了当前AI进展的核心局限性以及未来的扩展方向。到目前为止，AI在**可验证领域（verifiable domains）**取得了最显著的成功。这些领域，如数学、编程和竞技游戏，具有以下特点：

- **明确的目标**：例如，“证明这个定理”、“让这段代码通过所有测试用例”或“赢得这盘棋”。

- **客观的评价标准**：答案要么是对的，要么是错的；代码要么能运行，要么不能；棋局要么赢了，要么输了。这种非黑即白的反馈非常适合AI的学习机制。

然而，人类的大部分活动都发生在**开放式领域（open-ended domains）**，这些领域缺乏明确的对错。例如：

- **艺术创作**：写一首好诗或作一幅好画的标准是什么？
- **商业策略**：制定一个成功的商业计划，其成败受市场、竞争等无数不确定因素影响。
- **人际沟通**：进行一次有同理心的、成功的对话。
- **科学研究**：提出一个“好”的研究问题，本身就是一种高度创造性的、没有标准答案的行为。

提问者实质上是在问：AI在解决逻辑谜题上取得的强大能力，能否迁移到这些更模糊、更主观、更复杂的现实世界问题中？这是AI能否从一个“超级计算器”进化为一个真正“通用智能”的关键考验。

原文翻译

> > Oh yeah, this is a this is a a question I I really like. Um I think if you actually truly want to extend to research um and you know finding discovering ideas that that meaningfully advance technology on the on you know the scale of like months and years like I think the these questions like stop being so different right like it is one thing to solve like a very well posed uh constraint problem on the scale of an hour right and there's like kind of a finite amount of ideas you need to look through and that might feel extremely different from solving something very open-ended. Um but you know even if you want to solve like a very well- definfined problem that is on much longer scale right you like you know prove this millennial price problem. uh well that suddenly requires you to think about okay like what are the fields of mathematics or other science that might possibly be relevant you know are there inspiration from physics that I must take like what is kind of the entire uh program that I want to develop around this and now these become very open-ended questions and it's actually hard to you know for for for our own research right like if all we cared about is you know reduce the uh modeling loss on a given data set right like like measuring the progress on that like uh you know like like are we kind of actually asking the right questions in research like actually becomes like a fairly open-ended affair.

> 哦是的，这个问题我非常喜欢。我认为，如果你真的想扩展到研究领域，去发现那些能在数月或数年尺度上真正推动技术进步的思想，那么这些问题的区别就不那么大了。在一个小时的尺度上解决一个定义明确的约束问题是一回事，你可能只需要遍历有限数量的想法，这感觉上可能与解决一个非常开放式的问题截然不同。但是，如果你想解决一个定义明确但时间跨度非常长的问题，比如证明一个千禧年大奖难题，那么你突然就需要思考：哪些数学或其他科学领域可能与之相关？我是否需要从物理学中汲取灵感？我需要围绕这个问题构建一个怎样的完整研究纲领？这些就都变成了非常开放式的问题。实际上，对于我们自己的研究来说也很难。如果我们只关心降低某个数据集上的模型损失，那么衡量这方面的进展……但我们是否在研究中提出了正确的问题？这本身就变成了一个相当开放的事情。

深度解读

Jakub Pachocki的回答非常精彩，他提出了一个核心论点：**当时间尺度足够长时，可验证问题和开放式问题的界限会变得模糊。**

他的逻辑是这样的：

1. **短时间尺度问题**：解决一个一小时内能完成的数学题，是一个“约束问题”（constraint problem）。解题的路径相对有限，目标明确，感觉上与写诗这种开放式任务截然不同。
2. **长时间尺度问题**：但是，如果要解决一个像“证明黎曼猜想”这样的“千禧年大奖难题”，虽然它的最终目标（证明或证伪）是明确可验证的，但通往这个目标的路径是完全未知的。解决这个问题需要数十年的努力，涉及以下这些高度开放式的子问题：
    - **跨领域探索**：我应该学习哪些新的数学分支？物理学、计算机科学或其他领域能提供灵感吗？
    - **战略规划**：我应该制定一个怎样的长期研究计划？先攻克哪些小问题作为铺垫？
    - **品味与直觉**：在无数可能的探索方向中，哪一个“感觉”上更有前途？

在这样的时间尺度上，解决一个“定义明确”的数学难题，其过程充满了与艺术创作或制定商业战略类似的开放性、不确定性和对直觉的依赖。

他进一步用自己的工作——机器学习研究——来举例。表面上看，“降低模型损失”是一个非常明确、可量化的目标。但真正驱动研究进展的，是那些更上游的、开放式的问题：“我们是否在问正确的问题？”、“这个研究方向是否有前途？”。这些问题没有标准答案，却决定了所有后续技术努力的价值。

因此，他的结论是，通过训练AI在越来越长的时间尺度上解决复杂的、定义明确的问题，我们实际上也在间接地训练它处理开放式问题所需的核心能力：**长期规划、战略思考和在不确定性中导航**。AI在数学和编程竞赛中学会的，不仅仅是解题技巧，更是一种如何在广阔的“可能性空间”中进行有效探索的元能力（meta-skill）。这种元能力，正是从可验证领域通往开放式领域的桥梁。

原文翻译

> > Yeah. And I think it also makes sense to think about what the limits of, you know, uh open-ended means, you know. Um I think a while back Sam tweeted about some of the improvements that we were making in having our models write more creatively. And you know, we do consider the extremes here as well.

> 是的。而且我认为，思考“开放式”的极限在哪里也是有意义的。前段时间，Sam（OpenAI CEO）发推文提到了我们在让模型进行更具创造性写作方面取得的一些进步。我们确实也在考虑这些极端情况。

深度解读

Mark Chen在这里做了一个重要的补充，他提醒我们，OpenAI并没有忽视那些纯粹的、主观的开放式领域，比如“创造性写作”。这表明他们的研究策略是双管齐下的：

1. **自下而上**：通过解决超长时间尺度的可验证问题（如数学难题），间接培养AI处理开放式任务所需的规划和探索能力。
2. **自上而下**：直接针对创造性、主观性等开放式任务进行研究和优化，探索如何让AI更好地理解和生成具有艺术性、情感和新颖性的内容。

提及Sam Altman关于创造性写作的推文，意在说明，即使是像“写得更有创意”这样难以量化和评估的目标，也始终是他们研究版图中的一部分。他们关注的不仅仅是逻辑和推理的“硬”智能，也包括创造力和情感理解的“软”智能。最终，一个真正的“自动化研究员”或通用人工智能，必须同时在这两个维度上都表现出色。

2.6 进步的引擎 – 为什么强化学习持续带来惊喜

原文翻译

Right. Right. Let's talk about RL because it seems like since 01 came out, RL has been the gift that keeps giving. You know, every every couple months Open puts out a release and everyone goes, "Oh, that's great, but this RL thing is going to plateau. We're going to saturate the evals. The models won't generalize or there's going to be mode collapse because of too much synthetic data for whatever. Everybody's got a laundry list of reasons to believe that the gains and performance from RL are going to tap out and and somehow they just don't. You guys just keep coming out and putting out continuous improvements. Why is RL working so well and what if anything has surprised you about how well it works? RL is a very versatile method, right? And there are a lot of ideas you can explore um once you have an RL system working. a long time at OpenAI, we started from this before language models, right? Like we were thinking about like okay like RL is this like extremely powerful thing of course like on top of deep learning which is this like incredible general learning method. Um but the thing that we struggled with for a very long time is like what is the environment like how do we actually anchor these models to the real world or like should we you know simulate uh you know some some some some island where they all learn to collaborate and compete. Um and and then you know of course came the the the language modeling breakthrough right and we saw that oh yeah if we if we if we scale deep learning on modeling natural language we can create models with this like incredibly nuance understanding of human language and so since then we've been we've been you know seeking how to combine these paradigms and how to get our to work on natural language. Once you do right like then you kind of have the well you have the ability to um to to well to to to actually like like like execute on on these different ideas and objectives in this like extremely um robust rich environment given by pre-training. Uh and so yeah, so I think uh it's been a it's been a it's been a real um um yeah, I think it's been perhaps the most exciting period uh in our research over the last few years where we've really like uh yeah, we found so many new directions and promising ideas uh that that that all seem to to to be working out and and and and and we're trying to uh Yeah. Understand how to compare.

> 对。对。我们来谈谈强化学习（RL），因为自从O1问世以来，RL似乎是一份源源不断的礼物。每隔几个月，OpenAI就会发布一个新东西，大家都会说：“哦，太棒了，但这个RL的东西很快就会遇到瓶颈。评估会饱和，模型会无法泛化，或者因为过多的合成数据导致模式崩溃等等。”每个人都有一长串理由相信RL带来的性能提升即将耗尽，但不知何故，它就是没有。你们总能不断地推出持续的改进。为什么RL效果这么好？关于它的出色表现，有什么让你们感到惊讶的吗？ RL是一种非常通用的方法，对吧？一旦你有了一个能工作的RL系统，你就可以探索很多想法。在OpenAI很长一段时间里，甚至在语言模型出现之前，我们就开始研究这个了。我们当时想，RL是一种极其强大的东西，当然它建立在深度学习这个不可思议的通用学习方法之上。但我们长期以来一直挣扎的问题是：环境是什么？我们如何将这些模型锚定到现实世界？或者我们应该模拟一个岛屿，让它们在上面学习协作和竞争？然后，当然，语言模型的突破到来了。我们发现，哦，是的，如果我们在自然语言建模上扩展深度学习，我们就能创造出对人类语言有极其细致入微理解的模型。从那时起，我们一直在寻求如何将这两种范式结合起来，如何让RL在自然语言上工作。一旦你做到了，那么你就拥有了能力，在这个由预训练提供的极其稳健、丰富的环境中，去执行不同的想法和目标。所以，是的，我认为这确实是……是的，这可能是我们过去几年研究中最激动人心的时期，我们真的……是的，我们发现了很多新的方向和有前途的想法，而且它们似乎都在奏效，我们正在努力……是的，理解如何比较它们。

深度解读

这段话深刻地解释了为什么强化学习（RL）近年来能爆发出如此巨大的能量，其核心在于一个关键的“化学反应”：**强化学习与大规模语言模型的结合**。

首先，让我们用一个简单的比喻来理解**强化学习（RL）**。想象一下训练一只宠物狗。你不会给它一本“如何坐下”的教科书（这叫监督学习）。相反，你会发出“坐下”的口令，然后观察它的行为。如果它碰巧做对了，你就给它一块零食（**正向奖励**）；如果它做错了，你就不给（或者给予温和的惩罚，**负向奖励**）。通过反复的**试错**和**奖励反馈**，小狗最终会学会“坐下”这个动作能带来最大的奖励，从而掌握这个技能 。

表2：强化学习的核心组成部分

组成部分	定义	示例：视频游戏	示例：语言模型（RLHF）
智能体 (Agent)	学习者或决策者。	游戏角色（如马里奥）。	正在被微调的GPT模型。
环境 (Environment)	智能体互动的世界。	游戏关卡，包括敌人和障碍物。	人类语言的“世界”，包括用户的提问。
行动 (Action)	智能体能做出的选择。	向左移动、向右移动、跳跃。	在回答中生成下一个词或句子。
奖励 (Reward)	对行动的反馈信号（正向或负向）。	获得分数、到达终点（+），失去生命（-）。	根据回答的帮助性、真实性和安全性，由**奖励模型**给出的高分或低分。

在语言模型出现之前，RL研究者面临的最大困境是**“环境”问题**。他们可以创造出虚拟的游戏世界（如Dota 2）或模拟的物理环境，但这些环境相比真实世界而言，都过于简单和狭窄。AI在这些“人造沙盒”里学到的技能，很难迁移到复杂多变的现实世界中。

语言模型的突破彻底改变了这一切。一个经过海量文本数据预训练的大语言模型（如GPT），其内部已经构建了一个关于人类世界知识、文化、逻辑和社会规范的极其丰富和细致的“**内部世界模型**”。这个“世界”不是由程序员一行行代码写出来的，而是从人类文明的全部文本中自发学习到的。

Jakub Pachocki的深刻洞见在于，他们意识到，这个由语言模型构成的“内部世界”本身，就是RL所需要的那个**最理想、最丰富的“环境”**。他们不再需要去模拟一个外部世界，而是可以直接让RL智能体在这个语言模型内部的“知识宇宙”中进行探索和学习。这就像是找到了最完美的训练场，AI可以在其中练习和优化与人类的交流、推理和协作能力。这个“RL + 大语言模型”的组合，就是过去几年AI能力飞速发展的核心引擎。

原文翻译

> One of the hardest things about RL for folks who are not practitioners of RL is the idea of crafting the right reward model. And so, especially if you're a business or an enterprise who wants to harness all this amazing progress you guys are putting out, but doesn't even know where to start. How what do the next few years look like for a company like that? What is the right mindset for somebody who's trying to make sense of RL to craft the right reward model? Is there anything you've learned about the best practices or an approach of thinking of using this latest sort of um family of reasoning techniques? What what is the right way I should think about even approaching reward modeling as a biologist or a physicist? I expect this will evolve quite rapidly. I expect it will become simpler, right? Like I think I think
>
> > you know maybe like two years ago we would have been talking about like what is the right way to craft my fine-tuning data set and I I don't think we are like at the end of that evolution yet and I think we will be inching towards more and more humanlike learning uh which you know RL is still not quite. So I think I think maybe the most important part of the mindset is to like not assume that like what is now will be forever.

> > 对于非RL从业者来说，RL最难的事情之一就是构建正确的“奖励模型”（reward model）。特别是对于那些希望利用你们这些惊人进展的企业来说，他们甚至不知道从何入手。未来几年对于这样的公司会是怎样的？对于一个试图理解RL并构建正确奖励模型的人来说，正确的心态是什么？关于使用这套最新的推理技术的最佳实践或思维方式，你们有什么心得吗？作为一个生物学家或物理学家，我应该如何思考奖励建模这件事？ 我预计这会发展得非常快。我预计它会变得更简单，对吧？我想，大概两年前，我们可能还在讨论如何构建我的微调数据集。我认为我们还没有走到那条演进之路的尽头，我们将会越来越接近更像人类的学习方式，而RL目前还不是。所以我认为，最重要的心态或许是，不要假设现在的方式会永远持续下去。

深度解读

这段对话触及了将AI技术应用到具体专业领域的核心挑战——**奖励建模（Reward Modeling）**，并指出了其未来的发展方向。

什么是**奖励模型**？在训练宠物狗的例子里，“奖励”很简单，就是一块零食。但对于复杂的任务，比如“写一篇有帮助的、安全的、真实的关于黑洞的科普文章”，我们无法用一个简单的分数来衡量。奖励模型就是一个**专门训练出来充当“裁判”的AI**。它的训练过程是这样的：

1. **收集人类偏好数据**：让AI对同一个问题生成多个不同的回答（比如A和B）。
2. **人类进行排序**：找人类专家（比如一位物理学家）来判断，是回答A更好，还是回答B更好。他们不需要给出具体分数，只需要做出“A > B”这样的比较判断。
3. **训练奖励模型**：用大量的这类人类偏好数据来训练奖励模型。最终，这个模型就学会了模拟人类专家的品味和判断标准。当它看到一个新的回答时，就能给出一个分数，来预测人类专家会有多喜欢这个回答 。

这个奖励模型给出的分数，就成了强化学习过程中的“奖励信号”，引导主模型（GPT）生成更符合人类专家偏好的内容。这个过程被称为“基于人类反馈的强化学习”（Reinforcement Learning from Human Feedback, RLHF）。

对一个想应用AI的生物学家或物理学家来说，**“奖励建模”正在成为一种新的“编程”方式**。他们不再是写代码来告诉计算机“如何做”，而是通过提供偏好数据来“教”会AI什么是“好”的结果。这是一种更高层次的、关于**意图和价值观**的沟通。未来，一个领域的专家能否成功利用AI，很大程度上将取决于他们能否有效地将自己领域内的专业知识、品味和判断标准，“编码”到一个奖励模型中。

Jakub Pachocki的预测是，这个过程会“变得更简单”，并且会“越来越接近人类的学习方式”。这意味着未来我们可能不需要再进行繁琐的数据标注，而是可以通过更自然的对话、示范和修正来向AI传递我们的偏好。最重要的心态是保持开放和灵活，因为定义“好”与“坏”的方法论本身，也在飞速进化之中。

2.7 从“凭感觉编程”到“凭感觉研究”：人机协作的未来

原文翻译

> > Um so I want to bring the conversation back to coding. We would be remiss not to say congrats on GBT5 codecs. uh which just dropped today. Um can you guys say a little bit more about what's different about it, how it's trained differently, um maybe why you're excited about it? Yeah, so I think um one of the big focuses of the codeex team is to just take the raw intelligence that we have from our reasoning models and make it very useful for real world coding. So um a lot of the work they've done is kind of consistent with this. um they are working on kind of having the model be able to handle more difficult environments. Um we know that real world coding is very messy. Um so they're trying to handle all the intricacies here. Um there's a lot of coding that has to do with you know style with um just like kind of softer things like how how proactive the model is, how how lazy it is and just being able to define um in some sense like a spec for how uh a coding model should behave. um they do a lot of you know very strong work there and as as you see like um they they're also working on a lot better presets you know uh coders they have some kind of notion of this is how long I'm waiting I'm willing to wait for a particular solution um I think we've done a lot of work to dial in on you know for easy problems being a lot you know lower latency for harder problems actually the the right thing is to be even higher latency um get you the really best solution um and just being able to find that preset um is sweet spot for if you were to say like easier problems versus harder.

> 嗯，我想把话题带回到编程上。我们不能不祝贺今天刚刚发布的GPT-5 Codex。你们能多谈谈它有什么不同，训练方式有何不同，以及你们为什么对它感到兴奋吗？是的。我认为Codex团队的一个主要重点是，将我们从推理模型中获得的原始智能，转化成对真实世界编程非常有用的东西。他们做的很多工作都与此一致。他们正在努力让模型能够处理更复杂的环境。我们知道真实世界的编程非常混乱，所以他们试图处理这里的所有复杂细节。很多编程工作还涉及到风格，以及一些更“软”的东西，比如模型的积极性、懒惰程度，以及能够从某种意义上定义一个编程模型应该如何行为的规范。他们在这方面做了很多非常扎实的工作。正如你所见，他们也在开发更好的预设。程序员们对于一个特定的解决方案，心里大概有一个愿意等待的时间。我们做了很多工作来调整：对于简单问题，延迟要低得多；对于难题，实际上正确的做法是允许更高的延迟，以获得真正最好的解决方案。能够找到这种针对简单问题与难题的预设最佳点，是非常重要的。

深度解读

这段话揭示了新一代编程模型（如GPT-5 Codex）的设计哲学，它超越了简单地生成正确代码的范畴，致力于成为一个更懂程序员心思的“智能编程伙伴”。

这里的核心思想是，将底层推理模型的“原始智能”适配到“真实世界编程”这个“混乱”且“复杂”的场景中。真实世界的编程远不止是算法正确，还包括许多“软”性要求：

- **处理复杂环境**：现代软件开发涉及庞大的代码库、多个相互依赖的文件、复杂的构建工具和版本控制系统。AI需要理解整个项目的上下文，而不是孤立地看待一小段代码。

- **编码风格**：每个团队、每个项目都有自己的编码规范和风格。好的AI应该能自动适应并遵循这些风格，写出让人类同事易于阅读和维护的代码。
- **行为规范（Spec）**：这涉及到AI的主动性。它应该在什么时候主动提供建议？什么时候应该保持安静？它应该提供一个“差不多”的快速方案，还是一个“完美”但耗时较长的方案？这些都属于对AI行为的精细定义。

其中，关于**延迟（Latency）的权衡**是一个特别好的例子。这与前面讨论的GPT-5的“快慢思考”系统异曲同工。对于一个简单的任务（比如补全一个变量名），程序员希望得到瞬时响应。而对于一个复杂的任务（比如“重构这整个模块以提高性能”），程序员愿意等待更长的时间，以换取一个高质量、深思熟虑的解决方案。新一代Codex的目标就是智能地识别任务难度，并自动选择最佳的“思考时间”，从而在效率和质量之间达到最佳平衡。这使得AI从一个机械的“代码生成器”向一个能理解程序员工作节奏和需求的“高级助手”转变。

原文翻译

> > > What we've found is the the latest the the previous generation of the codeex models, they they were spending too little time solving the hardest problems and too much time solving the easy easy problems. And I think um that that is actually just um probably out of the box uh what what you might get out of 03.

> > 我们发现，上一代的Codex模型，它们在解决最难的问题上花的时间太少，而在解决简单问题上花的时间又太多。我认为 这实际上可能就是你从O3模型直接拿出来用时会得到的结果。

深度解读

这是一个非常精炼的技术总结，它指出了上一代模型在资源分配上的“不经济”。一个未经优化的、通用的推理模型（如O3），在面对编程任务时，可能会对所有问题“一视同仁”，导致：

- **在简单问题上“杀鸡用牛刀”**：为一个简单的任务付出了不必要的计算资源和时间，导致延迟过高，影响了流畅的编程体验。
- **在复杂问题上“浅尝辄止”**：由于默认的计算限制，它没有投入足够的时间去深入思考难题，给出的方案可能不够优化，甚至是错误的。

GPT-5 Codex的进步就在于解决了这个资源错配问题。通过专门的训练和优化，它学会了“看菜下饭”——快速处理简单请求，并为复杂挑战投入更多的“思考时间”。这反映了AI模型从“通用能力”到“专业应用”的适配过程中，一个非常重要的优化方向：**智能化的认知资源管理**。

原文翻译

Maybe just on the the topic of coding since you guys are both competitive coders in prior lives. Um, I know you've been at OpenAI for almost a decade now, but I was struck by uh the story of Lisa Doll, the Go player who kind of famously quit Go after he lost to Alph Go um multiple times. Uh, and I think in a recent interview you guys were both saying that now the coding models are better than your capabilities. Uh, and that gets you excited. Um, but say more about that. And um, how much would you say you code now? Well, if you're hands- on keyboard, you can you can talk about OpenAI generally, but how much code is written by AI now in terms of cutting models being better? I I mean, I think yeah, I think it is extremely exciting to see this progress. I think like the programming competitions have a nice kind of encapsulated test of like ability to um come up with some new ideas um in in in you know, in this like boxed uh environment and time frame. Um I do think like you know if you look at things like uh well I guess the IMO problem six or or maybe um some very hardest uh programming competitions problems like I think there's still a little bit of headway to go for the models but I wouldn't expect that to last very long. I took a little bit uh historically I've been like being humble historically I've actually been like extremely reluctant to use any sort of tools. I I I just used Vim pretty much old school. Yeah. Um yeah, eventually I think like like especially with this with this um um latest coding tools um like GPT5, I I've really kind of felt like okay like this is this is no longer the way like like you can do a you know 30 file refactor like pretty much perfectly in like 15 minutes like you kind of have to use it. Um yeah and so I've been I've been kind of like um learning this new way of coding which definitely feels a little bit different. I um I think it is like a little bit of an uncanny valley still right now where like like you kind of have to use it because it is just like exciting so many things but it's still like you know a little bit like u not quite as good as a as as a coworker. Um I so you know I I think like our our priority is getting out of that uncanny valley.

> 既然你们俩以前都是竞赛程序员，我们来聊聊编程这个话题。我知道你们在OpenAI快十年了，但我被李世石的故事深深触动，那位围棋手在多次输给AlphaGo后，众所周知地退役了。我想在最近的一次采访中，你们俩都说现在的编程模型已经比你们自己的能力更强了，这让你们感到兴奋。能多谈谈这个吗？你们现在自己还写多少代码？或者在OpenAI，现在有多少代码是由AI编写的？ 关于编程模型变得更强这件事？是的，我认为看到这种进步非常令人兴奋。编程竞赛提供了一种很好的、封装好的测试，来检验在这样一个有约束的环境和时间框架内，提出新想法的能力。我确实认为，如果你看像IMO第六题或者一些最难的编程竞赛题，模型还有一点点进步空间，但我预计这不会持续太久。我稍微……从历史上看，我一直…… （别谦虚了） 从历史上看，我其实一直非常不愿意使用任何工具。我基本上只用Vim。 （老派作风） 是的。嗯，但最终，我认为，特别是有了这些最新的编程工具，比如GPT-5，我真的感觉，好吧，老路子行不通了。你可以在15分钟内几乎完美地完成一个涉及30个文件的重构，你真的不得不用它。所以，我一直在学习这种新的编程方式，感觉确实有点不一样。我觉得现在还有点处于“恐怖谷”阶段，你不得不用它，因为它能做很多令人兴奋的事情，但它仍然……你知道，有点不像一个真正的同事那么好。所以，我们的首要任务是走出这个恐怖谷。

深度解读

这段对话极具感染力，因为它触及了人与AI关系的一个深刻的、情感层面的转折点，即**创造者被创造物超越的时刻**。

“李世石时刻”的重现：访谈者引用了围棋传奇李世石因AlphaGo而退役的故事，这是一个强大的文化符号，代表了人类在某个智力领域被AI彻底超越的标志性事件。当Jakub Pachocki和Mark Chen，这两位曾经站在编程竞赛金字塔顶端的人，承认AI的编程能力已经超越自己时，这无异于编程领域的“李世石时刻”。但与李世石的悲壮不同，他们的反应是“**兴奋**”。这种兴奋源于他们作为建设者的身份：看到自己亲手创造的工具展现出远超预期的强大能力，这本身就是一种巨大的成功和满足。

从抗拒到拥抱：工作流的根本变革：Jakub Pachocki坦言自己曾是一个“老派”的、抗拒工具的程序员（只用Vim），但新一代AI工具的强大能力（“15分钟重构30个文件”）让他不得不改变。这代表了整个软件开发行业正在经历的范式转变。AI编程助手不再是一个可有可无的“玩具”，而是成为了一个能带来数量级效率提升的、不可或缺的“生产力工具”。拒绝使用它，就如同在汽车时代坚持骑马一样。

“恐怖谷”的比喻：这个比喻非常精准地描述了当前人机协作编程的体验。AI很强大，但还不够完美。它像一个能力极强但有点“怪”的同事：

- **优点**：能极快地完成大量机械性、重复性的工作，甚至能提出你没想到的巧妙方案。
- **缺点**：有时会误解你的意图，犯一些低级但难以察觉的错误，需要你时刻保持警惕去监督和修正。

OpenAI的下一个目标，就是**“走出恐怖谷”**，让AI从一个需要时刻监督的“实习生”，成长为一个可以完全信赖的、能独立完成任务的“高级工程师”或“同事”。当这一天到来时，人类程序员的角色将发生根本性的变化。

原文翻译

> > Yeah. But uh yeah, it's definitely an interesting time. Yeah, definitely to kind of like speak to the lease little moment. Um I think AlphaGo for both of us was, you know, a very formative milestone in AI development. And at least for me, it was the reason I started working on this in the first place. And maybe partly because of our backgrounds in competitive programming like I had this affinity to building these models which could do very very well in in these forms of contests and going from you know solving eighth grade math problems um to a year later um hitting our level of performance in in these coding contests. It's crazy to see that progression and um you kind of imagine or like to think that you feel a set of the feelings at least at all felt too, right? It's um like wow this is really crazy, right? and and what are the possibilities and you know this is something that I took decades to do and it took a lot of hard work to get to the forefront of um so you really do feel an implication of that is these models what can't they do right and I do feel like already it's kind of transformed the default for coding um this past weekend I was talking to some some high schoolers and they were saying oh you know actually the default way to code is vibe coding like um you know I think like they they would consider oh it's like maybe sometimes for completeness you would go and like actually do all of the mechanics of coding it from scratch yourself, but that's just a strange concept to them. Like why would you do that? You know, just vibe code by default. Yeah. And and so yeah, I mean I I I do think you know the future hopefully will be vibe researching.

> 是的。但，这确实是一个有趣的时代。 是的，绝对是。 说到李世石的那个时刻。我认为AlphaGo对我们俩来说，都是AI发展中一个非常具有里程碑意义的事件。至少对我来说，这是我最初开始从事这项工作的原因。也许部分因为我们都有竞赛编程的背景，我有一种亲和力去构建那些能在这种竞赛中表现出色的模型。从解决八年级数学问题，到一年后在这些编程竞赛中达到我们自己的水平，看到这样的进展真是太疯狂了。你会想象，或者愿意去想，你也感受到了李世石当时所感受到的一部分情感，对吧？那种“哇，这太疯狂了”的感觉，以及它所带来的各种可能性。你知道，这是我花了几十年、付出了巨大努力才达到的前沿水平。所以你真的会感觉到，这意味着，这些模型还有什么做不到的呢？我确实觉得，它已经改变了编程的默认方式。上个周末我和一些高中生聊天，他们说：“哦，你知道吗，现在默认的编程方式就是‘凭感觉编程’（vibe coding）。”他们会觉得，哦，也许为了完整性，你偶尔会去自己从头把所有编程的机械活都干一遍，但对他们来说，这已经是一个奇怪的概念了。就像，你为什么要那么做？默认就用“凭感觉编程”就好了。是的。所以，我确实认为，未来有希望会是“凭感觉研究”（vibe researching）。

深度解读

这是整篇访谈中最具启发性和前瞻性的部分，它正式提出了“**凭感觉编程**”和“**凭感觉研究**”这两个核心概念，预示了未来人与AI协作的新范式。

“凭感觉编程”（Vibe Coding）的定义：这个词源于当今的高中生——数字时代的原住民。对他们而言，借助AI编程是天经地义的，就像我们使用计算器一样自然。Mark Chen的描述揭示了“Vibe Coding”的本质 ：

- **关注点转移**：程序员的工作重心从**“如何实现”（How）**转向了**“想要什么”（What）**。你不再需要逐行编写实现细节、处理繁琐的语法和调试低级错误。
- **新的工作流**：你的主要工作是向AI清晰地描述你的**高层意图、目标和“感觉”**（the vibe）。比如，“给我创建一个用户登录页面，风格要简约现代，要有社交媒体登录选项。”然后，AI负责将这个“vibe”翻译成具体的、可执行的代码。人类的角色从一个“工匠”转变为一个“**建筑师**”或“**导演**”。
- **从零到一的颠覆**：对于年轻一代来说，“从头手写所有代码”反而成了一种“奇怪”的、非主流的行为。这标志着一个时代的终结和另一个时代的开启。

从“凭感觉编程”到“凭感觉研究”（Vibe Researching）：这是Mark Chen提出的一个大胆而合乎逻辑的推论。如果编程——这个高度结构化和逻辑化的创造过程——都可以“凭感觉”进行，那么更开放、更依赖直觉和洞察力的科学研究，也必将迎来同样的变革。

“凭感觉研究”的图景：

- **研究者的角色**：未来的科学家可能不再需要花费数年时间学习复杂的实验技术或数据分析软件。他们的核心竞争力将是**提出好的问题、形成敏锐的直觉和拥有独特的“科研品味”**。
- **新的研究流程**：一个生物学家可以对AI说：“我感觉这个蛋白质的异常折叠可能与某种疾病有关，帮我设计一套实验来验证这个‘vabe’，分析所有相关文献，并模拟其分子动力学过程。”AI将成为一个全能的、不知疲倦的“自动化研究助手”，负责执行这个高层指令。
- **技能的重塑**：这将极大地降低科学研究的门槛，让更多拥有好想法但缺乏技术背景的人能够参与到知识创造中来。同时，它也对现有的科研人员提出了新的要求：必须从一个“技术执行者”转变为一个“**思想引领者**”和“**战略规划者**”。

这个从“Vibe Coding”到“Vibe Researching”的演进，是本次访谈最具深远意义的洞见。它不仅预言了技术的未来，更深刻地揭示了在AI时代，人类价值和核心竞争力的根本性转变。

2.8 伟大研究者的品质与研究之道

原文翻译

> Yeah. I that I have a question about that which is what makes a great researcher. Right. When you say vibe researching, there's um a big part of vibe coding is just having good taste in wanting to build something useful and interesting for the world. And I think what's so awesome about tools like Codeex is if you've got a good intuition for what people want, it helps you articulate that and then and then basically actualize a prototype very fast. With a with research, what's the what's the analog? What what makes a great researcher? Persistence uh is a is a very key trait, right? Like I think like what What is different about research when you're actually trying to I think a special thing about research right is you're trying to create something or or learn something that is just not known right like it's not known to work like you don't know whether it will work and so always trying something that will most likely fail and I think getting to a place where you are like in a mindset of like being ready to fail and being ready to learn from these failures and you know so and you know and of course with that comes creating kind of clear hypothesis and being extremely honest with yourself about how you're doing on them, right? I think a trap many people fall into is going out of the way to like to to prove that it works, right? Which is quite different from, you know, like I think like believing in your idea and significance is extremely important, right? And you want to persist persist that, but you have to be honest with yourself about when it's working and when it's not uh so that you can learn and adjust.

> 是的。 我对此有一个问题：什么造就了一位伟大的研究者？当你说“凭感觉研究”时，我们知道“凭感觉编程”很大一部分是拥有好的品味，想要为世界构建一些有用和有趣的东西。像Codex这样的工具之所以出色，是因为如果你对人们想要什么有很好的直觉，它能帮助你表达出来，然后非常快速地实现一个原型。那么在研究领域，类似的是什么？什么成就了一位伟大的研究者？ **坚毅（Persistence）是一个非常关键的特质，对吧？我认为研究的特别之处在于，你试图创造或学习一些未知的东西。你不知道它是否会成功。所以你总是在尝试一些很可能会失败的事情。我认为，达到一种心态，即准备好失败，并准备好从失败中学习**，是非常重要的。当然，随之而来的是，要提出清晰的假设，并对自己做得如何**极其诚实**。很多人会掉入一个陷阱，就是想方设法去证明自己的想法是可行的，这与……你知道，我相信自己的想法及其重要性是极其重要的，你也希望坚持下去，但你必须对自己诚实，知道它什么时候有效，什么时候无效，这样你才能学习和调整。

深度解读

在“凭感觉研究”的时代背景下，“什么造就了伟大的研究者”这个问题变得尤为重要。Jakub Pachocki的回答，没有提及任何技术性技能（如数学能力、编程能力），而是聚焦于三种核心的**心智品质**。这深刻地揭示了在人机协作的未来，人类价值的真正所在。

1. **坚毅（Persistence）**：研究的本质是探索未知，而探索未知必然伴随着大量的失败。AI可以加速试错的过程，但它无法代替人类承受失败带来的挫败感，也无法提供坚持下去的内在动力。因此，那种面对99次失败仍能发起第100次尝试的坚韧不拔，将成为人类研究者不可替代的核心价值。

2. **拥抱失败（Being Ready to Fail）**：这比“坚毅”更进了一步。它不仅是不怕失败，更是将失败视为学习过程的必要组成部分。一个伟大的研究者必须是一个高效的“失败学习者”，能够从每一次失败中提取出最有价值的信息，用以指导下一次的尝试。AI可以帮助分析失败的数据，但形成这种积极拥抱失败的心态，是人类独有的。

3. **智识上的诚实（Being Extremely Honest with Yourself）**：这是最微妙也最关键的一点。研究者常常会对自己的“亲生”想法产生情感依恋，这会导致一种认知偏差，即不自觉地去寻找证据来“证明”自己是对的，而忽略那些“证伪”的证据。这种“求证”而非“求真”的心态是科学研究的大敌。Jakub Pachocki强调，你必须能够将“**对想法的信念**”和“**对进展的诚实评估**”区分开来。你可以坚信你的大方向是重要的，但必须对每一步的实验结果保持绝对的客观和诚实，哪怕结果与你的预期完全相反。这种在信念和现实之间保持张力的能力，是做出重大突破的必要前提。

在AI能够处理所有技术执行细节的未来，这三种品质——坚毅、拥抱失败和智识上的诚实——将共同构成伟大研究者的“内核”。人类的职责，是为AI的强大执行力，注入正确的方向、坚韧的意志和不偏不倚的判断力。

原文翻译

> > Yeah, I think there are just very few shortcuts for experience. Um I I think through experience you kind of learn, you know, what's the right horizon to be thinking of a problem, right? You can't pick something that's too hard or it's not satisfying to do something that's too easy. Um and I think a lot of research is managing your own emotions over a long period of time, too. You know, there's just going to be a lot of things you try and they're not going to work. And sometimes you you need to know when to persevere through that or sometimes when to kind of switch to a different problem. Um and I think interestingness is something you know you try to fit through reading good papers talking to to your colleagues and um and you kind of maybe distill their experience into your own process.
>
> 是的，我认为**经验**几乎没有什么捷径。通过经验，你会学到思考一个问题的**正确时间跨度**是什么，对吧？你不能选一个太难的，也不能做一个太简单的而没有满足感。而且我认为，研究在很大程度上也是在很长一段时间内**管理自己的情绪**。你知道，你会尝试很多事情，但它们都不会成功。 有时候你需要知道何时应该坚持下去，有时候则需要知道何时该转向一个不同的问题。我认为**有趣性（interestingness）**是你通过阅读好的论文、与同事交流来培养的，然后你或许能将他们的经验提炼成你自己的流程。

深度解读

Mark Chen补充了另外三个关键特质，进一步丰富了“伟大研究者”的画像，这些特质同样是AI难以复制的。

1. **经验与品味（Experience and Taste）**：这里的“经验”不是指具体技能的熟练度，而是一种更高层次的**判断力**，即“**选题的品味**”。一个优秀的研究者，能凭经验和直觉判断出，在当前的时间点，哪个问题是“刚刚好”的——既有足够的挑战性，不至于太简单而无聊；又在能力范围之内，不至于太难而毫无希望。这种对问题“火候”的把握，是决定研究成败和研究者幸福感的关键。AI可以解决你给它的任何问题，但它无法告诉你哪个问题“值得”被解决。

2. **情绪管理（Managing Your Own Emotions）**：研究是一场漫长的、充满不确定性的马拉松。在这个过程中，必然会经历希望、失望、兴奋、沮丧的循环。能否在长期的挫败感中保持积极的心态，在取得微小进展时获得足够的激励，是决定一个人能否在科研道路上走下去的关键。这是一种内在的、深刻的自我调节能力。

3. **对“有趣性”的感知（Sense of Interestingness）**：科学的驱动力不仅仅是解决已有的问题，更是发现新的、**有趣**的问题。什么是“有趣”的？它通常是那些能够连接不同领域、挑战现有范式、或者能开启一个全新研究方向的问题。这种对“有趣性”的嗅觉，是通过大量阅读、深度思考和与同行的高质量交流而逐渐形成的。它是一种高度综合的、依赖于个人知识结构和好奇心的审美能力。

总结来说，Jakub Pachocki和Mark Chen共同描绘了一幅未来研究者的画像：他们不再是技术细节的执行者，而是**战略家、心理学家和艺术家**的结合体。他们的核心价值在于：用**坚毅**和**情绪管理能力**来驾驭漫长的研究周期，用**智识上的诚实**来确保方向的正确，用基于**经验**的**品味**来挑选最有价值的问题，并用对**有趣性**的敏锐直觉来开辟新的知识疆域。这些，正是“凭感觉研究”时代里，人类智慧的闪光之处。

2.9 打造前沿研究文化

原文翻译

> When I was in grad school um you know there's a big part uh I was I'm a failed machine learning researcher. I was in grad school for for bioinformatics. But a big part of my research advisor's thrust was about picking the right problems to work on such that you could then sustain and persist through the hard times. And you said something interesting which was there's a difference between having conviction in an idea and then being maximally truth seeeking about when it's not working. And though both those things might are sometimes in tension because you kind of go native on an on a topic or a problem sometimes that you have deep conviction in. Have you found is there any sort of heruristics you found are useful at the taste step at the problem picking step that help you arrive at the right set of problems where that conviction and truth seeeking is not as much in zero sum tension as other kinds of problems. Yeah to to be clear I don't think conviction and truth seeeking are really in a zero sum tension. I think like you can be like you can be convinced or you know you can have a lot of belief in idea and and you can be you know very persistent in it while it's not working. I think it's just important that you're kind of honest with yourself like like how much progress you're making and you're in a mindset where you're able to learn from uh the failures along the way. I think it's important to look for problems that you really care about and you really believe are important, right? And so um I think one one thing I've observed in in in many um researchers that inspired me has been really going after the hard problems like looking at the questions that are you know kind of like you know widely known but like not really kind of considered tractable and just asking like you know why are they not tractable or like you know what like what about this approach like why does this approach fail right you're you're you're always like thinking about what is really the barrier for the next step. If you're going after problems that like you really truly believe are important, right, then then that that that makes it so so much easier to find the motivation to persist with them over years.

> 我在读研究生时……我是一个失败的机器学习研究者，当时读的是生物信息学。我的导师非常强调的一点是，要挑选正确的问题来研究，这样你才能在困难时期坚持下去。你刚才说了一句很有趣的话，即“对一个想法有信念”和“当它行不通时最大限度地寻求真相”是有区别的。这两者有时会处于紧张关系中，因为你有时会对一个你深信不疑的话题或问题产生路径依赖。在选题这个品味阶段，你有没有发现什么启发式的方法，能帮助你找到那些“信念”和“求真”之间不是零和博弈的问题？ 是的，需要澄清一下，我不认为信念和求真真的是零和博弈的紧张关系。你可以对一个想法深信不疑，并且在它不成功时仍然非常坚持。关键在于你要对自己诚实，清楚自己取得了多少进展，并保持一种能从沿途的失败中学习的心态。我认为，寻找那些你**真正关心**并且**真正相信其重要性**的问题，是非常重要的。我从许多启发过我的研究者身上观察到的一点是，他们会真正去**挑战难题**，去研究那些众所周知但被认为难以解决的问题，并且会去问：“为什么它们难以解决？”或者“这个方法怎么样？为什么这个方法会失败？”你总是在思考，下一步的真正障碍是什么。如果你追求的是那些你真正相信是重要的问题，那么你就会更容易找到坚持数年的动力。

深度解读

这段对话深入探讨了研究中最核心的驱动力问题：**如何选择问题，以及如何处理信念与现实之间的关系**。

Jakub Pachocki澄清了一个常见的误解，即“信念”和“求真”是相互矛盾的。他认为，这两者非但不矛盾，反而是相辅相成的。一个健康的研究状态是：

- **在战略层面拥有坚定不移的信念**：你必须深信你所研究的大方向是重要的、有价值的。这种信念是你在长达数年的研究周期中，克服无数失败和挫折的**精神燃料**。没有这种源自内心的热爱和使命感，任何人都无法在探索未知的漫长道路上坚持下来。
- **在战术层面保持绝对的诚实和灵活**：在坚持大方向的同时，你必须对每一个具体的实验、每一种尝试的方法保持客观和批判的态度。当一个方法失败时，你的任务不是去“挽救”它，而是诚实地分析它为什么失败，并从中学习。

那么，如何找到那种能让你同时拥有“坚定信念”和“诚实心态”的问题呢？他给出了一个关键的启发式方法：**去挑战那些“重要但被认为难以解决”的硬核问题（Hard Problems）**。

选择这类问题有几个好处：

1. **内在驱动力**：因为问题本身足够“重要”，它能为你提供强大的、持久的内在动机。
2. **聚焦根本障碍**：研究这类问题会迫使你去思考最根本的障碍（“the barrier for the next step”）。你的关注点会自然地从实现细枝末节，转移到攻克领域内最核心的瓶颈上。
3. **高回报**：一旦在这些问题上取得突破，其影响力将是巨大的。

这种选题策略，本质上是一种“高风险、高回报”的研究品味。它要求研究者不仅要有技术能力，更要有挑战权威、直面根本性难题的勇气和智慧。这正是OpenAI自身研究文化的体现：他们选择的目标——如通用人工智能——正是这样一个“重要但被认为难以解决”的终极问题。

原文翻译

> In the development of like during the re training phase of GPD5 for example, are there any were there any moments where there were there was a hard problem the original initial attempts that were being made to crack that problem weren't working and yet you found somebody persisted through that. Um, and what was it about those sto any of those stories that comes to mind that worked well
>
> > that you wish other people and other researchers did more of? I think on the path there right like along the sequence of models like both the pre-trained models and the resig models um I think one very common theme is um bugs uh and you know both like just like yes silly bugs in software that can kind of stay in your software for like months and kind of invalidate all your experiments a little bit in a way that you don't know. Um and you know identifying them can be can be a very meaningful breakthrough for your research program. uh but also kind of bugs in the sense of like well you have a particular way of thinking about something and that way is a little bit skewed which causes you to uh make the wrong assumptions and identifying those wrong assumptions thinking rethinking frames from from scratch. Uh I think um you know both for getting the first reasoning models working or getting the uh you know larger pre-trained models working I think I think we've had like multiple issues like that that we've had to work through.

> > 在GPT-5的再训练阶段开发过程中，有没有遇到过这样的时刻：有一个难题，最初的尝试都失败了，但你发现有人坚持了下来并最终解决了它？关于这些故事，你脑海中有没有什么例子是效果很好的，并且你希望其他研究人员也能多做一些的？ 我认为在这条路上，无论是预训练模型还是推理模型，一个非常常见的主题就是“**bug**”。这既包括软件里那些可能潜伏数月、在你不知情的情况下让你所有实验都部分失效的愚蠢bug——识别出它们对你的研究项目来说可能是一个非常有意义的突破；也包括**思维方式上的“bug”**，即你对某件事的思考方式有点偏差，导致你做出了错误的假设。识别出这些错误假设，从头开始重新思考框架……我认为，无论是在让第一批推理模型工作起来，还是在让更大的预训练模型工作起来的过程中，我们都曾遇到过多个类似的问题，并且必须努力解决它们。

深度解读

这个回答非常坦诚，揭示了前沿AI研究光鲜外表下的真实面貌：一个与“bug”进行不懈斗争的艰苦过程。这里的“bug”有双重含义：

1. **代码层面的Bug**：在训练动辄耗资数百万美元、运行数周的超大模型的过程中，一个微小的代码错误（比如数据预处理的一个小瑕疵）可能会在数月后才被发现，而它已经污染了期间所有的实验结果。这不仅浪费了巨大的计算资源，更可能将研究引向错误的方向。因此，在前沿研究中，**严谨的工程实践、细致的调试能力和对细节的极致关注**，变得与提出创新想法同等重要。找到并修复一个潜藏已久的bug，其价值不亚于一次算法上的突破。

2. **思维层面的Bug（认知偏差）**：这是一种更隐蔽、也更危险的“bug”。它指的是研究者在头脑中对问题形成的**错误心智模型或错误假设**。比如，你可能坚信“模型越大越好”，而忽略了数据质量或算法设计的关键作用。这种思维定势会让你在错误的道路上越走越远，即使实验结果反复提示你方向错了，你也可能将其归咎于其他原因。

因此，一个顶尖研究团队所需要的，不仅是坚持不懈的精神，更是一种**系统性的、深刻的“除错”（Debugging）文化**。这包括：

- **对代码的除错**：建立严格的代码审查、测试和实验记录流程，以最小化技术性错误。
- **对思维的除错**：鼓励团队成员相互挑战彼此的核心假设，定期从“第一性原理”出发重新审视整个研究框架（“rethinking frames from scratch”）。

这个过程非常艰难，因为它要求研究者既要对自己的方向有信念，又要时刻准备着推翻自己的基本假设。这种在自信和谦逊之间保持动态平衡的能力，是推动研究穿越“无人区”的关键。

原文翻译

> As leaders of the research org, how do you think about what it takes to keep the best talent on your team? And on the flip side, creating a very resilient org that doesn't crumble if a key person leaves.
>
> > The biggest I think uh things that OpenAI has going for it in terms of keeping the best people motivated and exciting excited is like we are in the business of doing fundamental research, right? We aren't the type of company that looks around and says, "Oh, what model did P, you know, company X build or what model did company Y build?" Um, you know, we have a fairly clear and crisp definition of what it is we're out to build. Um, we like innovating at the frontier. Um, we we really don't like copying and um, I think people are inspired by that mission, right? you are really in in in the business of discovering new things about the deep learning stack and and um and I think we're we're kind of building something very exciting together. Um I think beyond that a lot of it's creating very good culture. So we want a good pipeline for training up people to become very good researchers. um we I think historically have hired um you know the the best talent and and the most innovative talent. So I just think um you know we have a very deep bench as well and um yeah I think most of the our leaders are very inspired by the mission and that's what's kept all of them there like when I look at my direct reports um they haven't been affected by the Talon wars.

> 作为研究组织的领导者，你们如何看待留住团队中最优秀人才所需要的东西？另一方面，如何创建一个即使关键人物离开也不会崩溃的、有韧性的组织？ 我认为OpenAI在激励和留住最优秀人才方面最大的优势是，我们从事的是**基础研究（fundamental research）**。我们不是那种会环顾四周说“哦，X公司做了什么模型，Y公司做了什么模型”的公司。我们对自己要构建的东西有一个相当清晰明确的定义。我们喜欢在前沿进行创新，我们真的不喜欢模仿。我认为人们被这个使命所激励，对吧？你真正在做的是发现关于深度学习技术栈的新事物，而且我认为我们正在共同构建一些非常激动人心的东西。除此之外，很大程度上是创造非常好的文化。我们希望有一个好的培养渠道，把人们培养成非常优秀的研究人员。我们历史上一直雇佣最优秀和最具创新性的人才。所以我认为我们的人才储备非常深厚。而且，我认为我们大多数领导者都深受使命的激励，这也是他们都留在这里的原因。当我看到我的直接下属时，他们并没有受到人才战争的影响。

深度解读

这段话揭示了OpenAI在激烈的人才竞争中保持团队稳定和凝聚力的核心秘诀：**一个清晰、宏大且专注于前沿创新的使命**。

在当今AI领域，人才流动极为频繁，各大公司不惜重金挖角。然而，Mark Chen指出，OpenAI的吸引力并不仅仅在于薪酬，而在于它为顶尖人才提供了一个独特的价值主张：

1. **使命驱动，而非竞争驱动**：OpenAI的战略不是去追赶或模仿竞争对手发布的产品，而是遵循自己既定的、长远的路线图——构建通用人工智能。这种“向内看”而非“向外看”的姿态，为研究人员提供了一种宝贵的**心理安全感和专注力**。他们不需要为追赶短期热点而疲于奔命，可以专注于真正具有长期价值的基础性难题。

2. **专注于“发现”，而非“复制”**：对于顶尖的研究人才来说，最大的激励莫过于能够亲手“发现新事物”，在人类知识的边界上留下自己的印记。OpenAI将“在前沿创新”作为核心文化，承诺为他们提供这样一个平台。这对于那些渴望做出开创性工作的人来说，具有不可抗拒的吸引力。

3. **人才培养和深度储备**：除了吸引外部顶尖人才，他们还强调内部培养体系（“a good pipeline for training up people”），并建立了深厚的人才梯队（“a very deep bench”）。这意味着组织的能力并不依赖于一两个“超级明星”，而是建立在一个强大的、有共同文化的集体之上。这正是组织韧性的来源：即使有个别关键人物离开，整个体系也能继续健康运转。

总而言之，OpenAI的人才策略，是围绕其核心使命构建的一个正向循环：宏大的使命吸引了认同该使命的人才，这些人才在前沿探索中获得了巨大的成就感和满足感，从而进一步强化了他们对使命的认同和对组织的忠诚度。这是一种比单纯的金钱激励更深刻、更持久的凝聚力。

原文翻译

> I was chatting with a researcher recently and he was talking about wanting to find the cave dwellers. Um, and these are often the people who are not posting on social media about their work. Um, for whatever reason they may not even be publishing. They're sort of in the background doing the work. Um, I don't know if you would agree with this concept, but how do you guys hire for researchers? And are there any non-obvious ways that you look for talent or you know attributes that you look for that are non-obvious?
>
> > So I think I think one thing that um we look for is having solved hard problems in any field. A lot of our most successful researchers um have started their journey with deep learning at OpenAI and have worked in other fields like um physics or um computer science, the computer science or finance uh in the in the past strong technical fundamentals coupled with the abil the um intent to like work on very ambitious problems and actually stick with them. We don't purely look for oh, you know, who did the most visible work or or or or is the most visible on social media or

> 我最近和一个研究员聊天，他提到想找到那些“洞穴居住者”（cave dwellers）。这些人通常不会在社交媒体上发帖谈论自己的工作，甚至可能因为某些原因不发表论文。他们只是在幕后默默地工作。我不知道你们是否同意这个概念，但你们是如何招聘研究员的？你们有没有什么非传统的方式来寻找人才，或者你们会寻找哪些不那么明显的特质？ 我认为我们寻找的一个特质是，**在任何领域解决过难题的经历**。我们很多最成功的研究员，都是在OpenAI才开始他们的深度学习之旅的，他们过去曾在物理、计算机科学或金融等其他领域工作。**扎实的技术基础**，加上**有志于解决宏大问题并能坚持下去的意愿**。我们不只看谁做了最显眼的工作，或者谁在社交媒体上最活跃。

深度解读

这里揭示了OpenAI一个非常独特的、反传统的招聘哲学，可以概括为**“能力优先于经验，潜力优先于名声”**。

他们寻找的不是已经成名的AI研究者，而是那些被称为“洞穴居住者”的人。这个比喻非常形象，它指的是那些不追求外界关注、不热衷于在社交媒体上自我营销，而是沉浸在自己的世界里，**专注于解决真正困难问题**的人。

他们的招聘标准有几个“非明显”的特点：

1. **跨领域选才**：他们并不要求应聘者必须有深度学习的背景。相反，他们非常欢迎来自物理、数学、金融等领域的人才。这是因为他们相信，在这些“硬核”领域能够解决难题的人，已经证明了他们具备了最重要的核心能力：**严谨的逻辑思维、强大的问题分解能力和坚韧不拔的意志**。深度学习的具体知识和技能，他们相信可以在OpenAI的环境里快速学会。

2. **关注“解决难题”的记录**：他们的考察重点不是你发表了多少篇顶会论文，或者在社交媒体上有多少粉丝，而是你是否真的有过独立、深入地攻克一个“硬骨头”问题的经历。这个经历本身，就是对候选人研究潜力的最好证明。

3. **寻找内在驱动力**：他们寻找的是那种“有志于解决宏大问题并能坚持下去”的人。这是一种对研究发自内心的热爱和追求，而不是为了追逐名利。这种内在驱动力，是研究者能够长期坚持并最终做出突破性工作的根本保障。

这种招聘策略的背后，是对“研究能力”本质的深刻理解。他们认为，真正的研究能力是一种**可迁移的元能力**，它比特定领域的知识更重要。通过这种方式，OpenAI得以跳出AI圈内的人才零和博弈，从更广阔的人才库中发掘出那些拥有巨大潜力、但尚未被主流视野发现的“璞玉”。

2.10 平衡研究与产品，探索与执行

原文翻译

> > Yeah. As you were talking, I I was thinking back to when I when I was a founder and I was running my own company and we would recruit for great talent engineers. Many of the attributes you described were ones that were on my mind then. Um, and Elon recently tweeted that he thinks this whole researcher versus engineer distinction is silly. Is that just a semantic uh is it just being you know semantically nitpicky or do you think these two things are more similar than they actually look? Yeah, I mean I I do think there can like researchers they don't just fit one shape. Um you know we have certain researchers who are very productive at openi who are just so good at idea generation and um you know they don't necessarily need to show great impact through implementing all of their ideas, right? I think there's so much alpha they generate in just kind of coming up with oh let's try this or let's try this or maybe we're thinking about that and there's other researchers who you know they are just very very efficient at um taking one idea um rigorously exploring you know the space of experiments around that idea. So I think you know researchers come in very different forms. I think um maybe that first type wouldn't necessarily map into the same bucket as a a great engineer, but um you know we we do kind of try to have a fairly diverse um set of research tastes and styles.

> > 是的。你说话的时候，我想起了我当创始人运营自己公司招聘优秀工程师的时候。你描述的很多特质，也是我当时所想的。埃隆·马斯克最近发推说，他认为“研究员”和“工程师”的区分很傻。这只是语义上的吹毛求疵，还是你认为这两者实际上比看起来更相似？ 是的，我的确认为研究员并非只有一种模式。在OpenAI，我们有些研究员非常有成效，他们就特别擅长**产生想法**。他们不一定需要通过实现所有想法来展示巨大影响力，对吧？他们仅仅是通过提出“我们试试这个”或“我们试试那个”，或者“我们可能在考虑那个”，就能产生巨大的价值（alpha）。还有另一些研究员，他们非常高效，能够**拿一个想法，然后严谨地探索**围绕这个想法的整个实验空间。所以我认为，研究员有各种不同的类型。也许第一种类型不一定能和一位伟大的工程师归为一类，但我们确实努力拥有一个相当多样化的研究品味和风格。

深度解读

关于“研究员”与“工程师”区别的讨论，Mark Chen提供了一个更细致、更多元的视角。他反对将“研究员”这个角色标签化，而是将其分解为**两种不同的核心能力原型**：

1. **“思想家”型研究员（Idea Generator）**：这类研究员的核心价值在于他们的**创造力和洞察力**。他们是“想法的源泉”，能够不断地提出新颖的、有前途的研究方向和实验思路。他们的贡献不体现在写了多少代码或做了多少实验，而在于他们为整个团队提供了宝贵的“alpha”——即领先于他人的认知优势。他们的角色更接近于一个战略家或思想领袖。

2. **“实干家”型研究员（Rigorous Explorer）**：这类研究员的核心价值在于他们**严谨的执行力和探索能力**。他们擅长将一个抽象的想法，转化为一个具体的、可执行的、周密的实验计划，并系统地、高效地完成它。他们能够深入一个想法的细节，穷尽其可能性，并从中得出可靠的结论。他们的角色更接近于一个科学家或侦探。

一个成功的、富有创造力的研究团队，**需要这两种类型的研究员紧密合作**。只有“思想家”会产生很多空想，而只有“实干家”则可能在错误的道路上高效地白费力气。当这两种风格的研究员形成互补时，团队就能既有天马行空的创造力，又有脚踏实地的执行力。

这种对研究员角色的精细划分，也解释了为什么“研究员”和“工程师”的界限有时清晰，有时模糊。一个“实干家”型的研究员，其工作方式和所需技能，与一个顶尖的软件工程师有很多重合之处。而一个“思想家”型的研究员，其价值则体现在完全不同的维度上。OpenAI的成功，部分源于他们能够识别、尊重并有效组合这些不同“研究品味和风格”的人才，创造一个多元化且高效的创新生态系统。

原文翻译

> > > Yeah. Mhm. And and say a little bit about what it takes to make like a create a frontier sort of winning culture that can attract all kinds of shapes and of researchers and then actually grow them, thrive them, make them win together at scale. What is it? What what do you think are the most critical ingredients of a winning culture? So I I think actually the most important thing is just to make sure you protect fundamental research, right? Um, I think you can get into this world with so many different companies these days where you're just thinking about, oh, how do I compete on, you know, a chat product or some other kind of product surface and um, you need to make sure that you leave space and recognize the research for what it is and also give them the space to do that, right? Like you can't have them being pulled in all of these different product directions. Um, so I think that's one thing that we pay attention to within our culture, especially now that there's so much spotlight on open AI, so much spotlight on AI in general and and the competition between different labs. Uh, it would be easy to fall into a mindset of like, oh, we're racing to bit beat this latest release or something. and and um you know there's definitely like um uh a risk that people kind of start looking over their shoulder and start thinking about oh you know what are these other things and and uh I see it as a large part of um our job to make sure that people have this comfort and space to think about you know what what are things actually going to look like in a year or two? um like what are the actually big research questions that we want to answer and and how do we actually get to models that like vastly outperform what we see currently rather than just like iteratively improving in the current paradigm.

> 是的。请谈谈，要创建一个能在前沿领域取胜的文化，需要什么？这种文化能够吸引各种类型的研究员，并让他们成长、茁壮，最终作为一个整体大规模地取胜。你认为一个制胜文化最关键的要素是什么？ 我认为最重要的事情就是确保你**保护基础研究**。现在有很多公司，你很容易陷入一种思维，就是整天想着，哦，我如何在聊天产品或其他产品层面进行竞争。你需要确保你为研究留出空间，承认研究本身的价值，并给他们空间去做研究，对吧？你不能让他们被所有这些不同的产品方向拉扯。所以，这是我们在文化中非常关注的一点。 特别是现在，OpenAI和整个AI领域都备受瞩目，不同实验室之间的竞争非常激烈。我们很容易陷入一种心态，比如“哦，我们要赶紧追赶上最新的发布”之类的。这确实存在一种风险，人们会开始回头看，开始想其他那些东西。我认为，我们工作的很大一部分，就是确保人们有这种**舒适和空间去思考一两年后事情会是怎样**，我们想回答的真正重大的研究问题是什么，以及我们如何才能真正做出**远超当前所见的模型**，而不仅仅是在现有范式下进行迭代改进。

深度解读

这段话揭示了OpenAI文化的核心，也是他们作为领导者最重要的职责：**构建一个“保护罩”，来捍卫基础研究的神圣空间**。

在一个商业竞争日益激烈的环境中，公司很容易被短期的产品迭代和市场份额所绑架。这种压力会不可避免地传导到研究团队，迫使他们放弃长远的、高风险高回报的探索，转而去做一些短期的、确定性高的“缝补”工作。这对于一个旨在推动技术边界的公司来说是致命的。

Mark Chen和Jakub Pachocki将自己的角色定位为这个“保护罩”的守护者。他们的工作就是吸收和过滤掉来自外界的短期焦虑和竞争压力，为研究团队创造一个可以“仰望星空”的环境。在这个环境中，研究员被鼓励去思考：

- **长期愿景（Long-term Vision）**：“一两年后，AI应该是什么样子？”而不是“下个季度我们能发布什么功能？”
- **根本性问题（Big Research Questions）**：“这个领域最核心的、尚未解决的难题是什么？”而不是“我们如何比竞争对手的模型在某个指标上高0.1%？”
- **范式突破（Paradigm Shift）**：“我们如何能做出性能**远超（vastly outperform）**现有模型的下一代技术？”而不是“我们如何在现有技术上做一些微小的迭代改进？”

这种对基础研究的刻意保护，是OpenAI能够持续产出颠覆性创新的文化基石。它确保了公司的“创新引擎”不会因为短期的商业压力而熄火。领导者的智慧，不在于督促团队跑得更快，而在于确保他们始终跑在正确的、通往未来的赛道上。

原文翻译

> > Just to pull on that thread more on protecting fundamental research um you guys are obviously one of the best research organizations in the world but you're also one of the best product companies in the world. H how do you balance and especially with um you've brought on some of the best product execs in the world as well. um how do you balance that focus between the two and while protecting fundamental research also continue to move forward the great products that you have out? Yeah, I mean I think it's about kind of delineating a set of researchers who really do care about product and who really want to be accountable to the success of the product and and they should of course very closely coordinate with the the research work at large. Um but I think just kind of people understanding their their mandates and what they are rewarded for um uh that that's a very important thing. One thing that I think is also helpful is that um our product team and and broader company leadership is is is bought into this vision right where where we are going with research. And so uh you know nobody is assuming that like oh the product we have now is a product we'll have forever and we'll just kind of wait for like you know new versions from research like like we we are able to think jointly about what the future looks like.

> 沿着“保护基础研究”这条线再深入一点，你们显然是世界上最好的研究机构之一，但同时也是最好的产品公司之一。你们如何平衡这两者之间的焦点？特别是你们还引进了世界上一些最优秀的产品高管。在保护基础研究的同时，如何继续推动你们已有的出色产品向前发展？ 是的。我认为关键在于**划定**一部分真正关心产品、并愿意为产品成功负责的研究人员。他们当然应该与整个研究部门紧密协调。但我认为，让人们**清楚自己的任务授权（mandates）以及他们因何而受奖励**，是一件非常重要的事情。另一件有帮助的事是，我们的产品团队和更广泛的公司领导层，都**认同我们研究要走向的那个愿景**。所以，没有人会假设我们现在的产品会永远是这个样子，然后就等着研究部门出新版本。我们能够**共同思考**未来会是什么样子。

深度解读

这里具体阐述了OpenAI在组织架构上如何实现“研究”与“产品”的平衡，其策略可以总结为“**明确分工，愿景统一**”。

1. **明确分工与激励机制（Delineation and Rewards）**：他们并没有试图让所有研究员都去关心产品，也没有让所有产品经理都去干预基础研究。相反，他们在组织内部进行了清晰的角色划分：
    - **应用研究团队**：一部分研究员被明确赋予了“为产品成功负责”的使命。他们的工作是将在基础研究中产生的技术，转化为稳定、好用的产品功能。他们的绩效和奖励，与产品的成功直接挂钩。
    - **基础研究团队**：另一部分研究员则被保护起来，他们的使命是进行长期的、前沿的探索，他们的奖励标准是能否产生颠覆性的新思想和新算法，而非短期产品指标。 让每个人都清楚自己的“赛道”和“计分板”，是避免内部混乱和目标冲突的关键。

2. **愿景统一（Shared Vision）**：比组织分工更重要的是，整个公司，从基础研究员到产品经理再到最高领导层，对公司的长远目标有着共同的信念。产品团队不会将当前的产品形态视为终点，他们知道这只是通往未来宏大愿景的一个阶段性产物。研究团队也不会与产品脱节，他们知道自己的探索最终要为实现那个共同的愿景服务。 这种**自上而下的愿景统一**，使得“研究”和“产品”不再是两个相互拉扯的部门，而是成为了一艘大船上，分别负责“瞭望未来航向”和“划好眼前船桨”的两个团队。他们目标一致，只是分工不同。这种“共同思考未来”的文化，是解决研究与产品之间内在张力的最根本的解决方案。

原文翻译

> One of the things that you guys have done is let such a diversity of different ideas and bets flourish inside of OpenAI that you then have to figure out some way as research leaders to to make it all make coherent sense as one part of a road map. And you got, you know, people over here investigating the future of diffusion models and visual media. And over here you've got folks, you know, investigating the future of reasoning when it comes to code. How do you paint a coherent picture of all that? How does that all come together when when there might be at at least naively some tension between giving researchers the independence to go to fundamental research and then somehow making that all fit into one coherent research program. Our settle goal um for our research program has been getting to an automated researcher for um a couple years now. Uh and so we've been we've been um building most our projects with this goal in mind. Um and so this still leaves a lot of room for um kind of bottom up idea generation for fundamental research on on various domains. But we are you know always thinking about how do these ideas come together eventually. Um we are you know we we believe for example that reasoning models go much further and we have a lot of explorations on things that are not directly reasoning models but we are thinking a lot about how they eventually combine and you know what does what what will this uh kind of innovation look like once you have something that is out there and thinking for for for moms about a very hard problem. Um and so I think this clarity of of like our long-term objectives is important. Um but yeah but it doesn't doesn't mean that we are you know prescriptive about like oh here are all the little pieces right like we definitely view this as a as a question of of exploration and learning about about these technologies.

> 你们所做的一件事，就是让OpenAI内部各种不同的想法和赌注都能蓬勃发展，然后你们作为研究领导者，必须想办法把这一切整合成一个连贯的路线图。比如，这边有人在研究扩散模型和视觉媒体的未来，那边又有人在研究代码推理的未来。你们如何把这一切描绘成一幅连贯的图景？当给予研究人员进行基础研究的独立性，与将所有研究都纳入一个连贯的研究计划之间，可能存在某种天真的紧张关系时，你们是如何让它们融合在一起的？ 我们研究项目的既定目标，几年来一直都是实现一个**自动化研究员**。所以我们大多数项目都是围绕这个目标来构建的。这仍然为各个领域的基础研究留下了很大的空间，让想法可以**自下而上**地产生。但我们总是在思考，这些想法最终将如何汇集在一起。比如，我们相信推理模型会走得更远，我们也有很多探索并非直接是推理模型的东西，但我们一直在思考它们最终如何结合。当你有了一个能够就一个非常困难的问题思考数月的东西时，这种创新会是什么样子？所以我认为，我们**长期目标的清晰性**非常重要。但这并不意味着我们对所有小细节都做出规定，比如“哦，这些是所有的小拼图”。我们绝对认为这是一个关于探索和学习这些技术的问题。

深度解读

这段话精辟地总结了OpenAI的组织和研究策略：**以一个强大的“北极星”目标，来牵引和统一所有自下而上的探索**。

他们解决“独立探索”与“统一规划”之间矛盾的方法，不是制定一个僵化的、事无巨巨细的“五年计划”，而是确立一个足够宏大、足够有吸引力的**最终目标——“自动化研究员”**。这个目标就像一个强大的引力中心，它为所有看似分散的研究项目提供了一个共同的汇聚点。

这种“**顶层设计 + 底层涌现**”的模式，具有以下几个优点：

1. **赋予探索意义**：一个研究视觉模型（如Sora）的团队，和一个研究代码模型（如Codex）的团队，看似在做不同的事情。但在这个统一的框架下，他们都清楚，自己正在构建的，是未来那个无所不能的“自动化研究员”所必需的“视觉能力”和“逻辑与工具使用能力”。这使得每个团队的工作都有了超越其本身领域的宏大意义。

2. **鼓励自下而上的创新**：因为最终的蓝图是清晰的，领导层就不需要对每个具体的技术路径做出指令。他们可以放心地让研究团队在各自的领域内进行自由探索（“bottom up idea generation”），因为他们相信，任何真正的、根本性的技术进步，最终都会对实现“自动化研究员”这个大目标有所贡献。

3. **保持战略定力**：这个清晰的长期目标，帮助他们在面对外界层出不穷的新技术、新热点时，能够保持清醒的判断。他们会问一个核心问题：“这个新技术，对于实现我们的最终目标有帮助吗？”这使得他们能够避免盲目跟风，将宝贵的资源始终聚焦在最核心的赛道上。

OpenAI的组织模式，本身就是对其研究哲学的一种体现。他们相信“规模化”（Scaling）的力量，不仅体现在模型参数和数据上，也体现在组织目标上：一个足够宏大的目标，能够自然地组织和协调起一个庞大而多元化的研究体系，让其在探索未知时，既有自由度，又不失方向感。

2.11 资源分配与优先级

原文翻译

Yeah I think you want to be opinionated and prescriptive at their very kind of course level but you know a lot of ideas can bubble up in a finer level and has have there been any moments where th those things have been intention at all recently? Well, one provocative example could be recently, you know, this new image model came out, which is nano banana, right, from Google. It's extraordinary value shown to that like lots of everyday people um can unlock a lot of creativity when these models are good at understanding editing prompts. Um and and I could see how that would create some tension for a research program that may not be prioritizing that as directly. um if if if one of your you know somebody talented on your team came and said guys like this thing is so clearly valuable in the world out there we should be spending you you know more effort more energy on this how do you reason about that question I think that's definitely a question that we've been kind of thinking about for quite a while at OpenAI I mean if you if you look at GP3 right like like once we kind of saw like oh like this is kind of where language models are going we we definitely like had a lot of discussions about well clearly there are going to so many magical things you can do with AI, right? And you will you will be able to go to this like like extremely smart models that are, you know, out there pushing the frontiers of science, but you will also have this like incredible media generation and this incredibly uh you know transformative u um entertainment applications. Uh and so like how do we prioritize among all these directions uh has definitely been something we've been we've been thinking about for for for quite a while. Yeah, absolutely. And and the real answer is like we don't discourage someone from being really excited by that and and it's just if we're consistent in the prioritization um and our product strategy, then it just will naturally fall in and and so it's just for us like we we do encourage a lot of people to to be excited about, you know, building this um you know, building kind of like aic products, you know, whatever kind of products that that they're excited by. But I think it's uh important for us to also have a a separate group of people who you you protect that their goal is to create the algorithmic advances.

> 是的，我认为你希望在非常宏观的层面上是有主见和规定性的，但在更精细的层面上，很多想法可以涌现出来。 最近有没有出现过这些事情之间产生紧张关系的时候？一个挑衅性的例子是，最近谷歌发布了一款新的图像模型（指Imagen 3的“纳诺香蕉”梗），它向普通人展示了巨大的价值，当这些模型擅长理解编辑指令时，可以释放很多创造力。我可以想象这会给一个可能没有直接优先考虑该方向的研究项目带来一些紧张。如果你们团队里有才华的人跑来说：“伙计们，这东西在世界上显然很有价值，我们应该在这上面投入更多精力”，你们如何思考这个问题？ 我认为这绝对是我们在OpenAI思考了很长一段时间的问题。如果你回顾GPT-3，当我们看到语言模型的发展方向时，我们确实进行了很多讨论。 很明显，你可以用AI做很多神奇的事情，对吧？你既可以拥有那些推动科学前沿的极其智能的模型，也可以拥有令人难以置信的媒体生成和极具变革性的娱乐应用。如何在所有这些方向中确定优先级，确实是我们思考了很久的事情。是的，绝对是。真正的答案是，我们**不会阻止**有人对此感到非常兴奋。只要我们在优先级和产品策略上保持一致，它自然就会找到自己的位置。所以对我们来说，我们鼓励很多人去对构建各种他们感兴趣的AI产品感到兴奋。但我认为，对我们来说，重要的是也要有**另一群独立的人**，你要保护他们，他们的目标是**创造算法上的进步**。

深度解读

这段对话非常现实地探讨了前沿研究机构如何应对外部竞争压力和内部资源分配的难题。当竞争对手发布了一款在某个特定领域（如图像编辑）表现惊艳的产品时，公司内部不可避免地会出现“我们也应该做这个”的声音。

OpenAI的处理方式，再次体现了他们“**明确分工，愿景统一**”的策略，并将其进一步细化：

1. **不压制热情，鼓励探索**：他们不会因为一个新想法不符合当前的最高优先级，就禁止员工去探索。他们鼓励员工对各种可能性保持兴奋，并进行小范围的尝试。这种开放性是保持组织创新活力的关键。

2. **战略优先级作为最终裁决**：一个新项目能否获得大量资源，最终取决于它是否符合公司既定的、长期一致的战略优先级。如果公司的核心战略是“推动通用人工智能和自动化研究”，那么一个纯粹的娱乐应用，即使商业前景很好，也可能不会成为资源投入的重点。

3. **组织上的“防火墙”**：最关键的一点是，他们通过组织架构，建立了一道“防火墙”。他们确保有一个**独立的、受保护的团队**，其唯一的使命就是追求核心的“算法进步”（algorithmic advances）。这个团队的资源和评价体系，与短期产品或市场热点完全脱钩。这道防火墙，确保了即使公司的其他部门都在追逐各种应用，那个推动技术边界的核心引擎也永远不会停转。

这种策略，是在保持战略定力和拥抱机会主义创新之间的一种精妙平衡。它允许组织在边缘进行广泛的探索和试错，同时确保核心力量始终聚焦于那个最重要、最长远的目标。

原文翻译

> How does that translate and just to build on Andre's question into a concrete framework around resourcing like do you think about okay x% of compute resources will go to longer term you know very important but maybe a bit more pie in the sky exploration versus there's also you know obviously current product inference but sort of this thing in the middle where uh it's achievable in the short to medium term.
>
> > Yeah. Um so I think that's a big part of both of our jobs. You know just uh this portfolio management question of how much compute do you give to which project and um I think historically we've put a little bit more on just the core algorithmic advances uh versus kind of the the product research. Um but it's something that you have to feel out over time, right? It's it's dynamic. I think monthtomonth there could be different needs. And so I think it's important to stay fairly flexible on that.

> 这个问题如何转化为一个具体的资源分配框架？比如，你们会考虑说，好，X%的计算资源用于长期的、非常重要但可能有点天马行空的探索，而另一部分用于当前产品的推理，还有一部分用于中短期内可实现的目标？ 是的。我认为这正是我们俩工作的一大部分。就是一个**投资组合管理（portfolio management）的问题，即给哪个项目多少计算资源。历史上，我们倾向于在核心算法进步**上投入稍多一些，而不是产品研究。但这是一个你需要随着时间去感受的事情，它是动态的。每个月都可能有不同的需求。所以我认为在这方面保持相当的灵活性很重要。

深度解读

这段话将AI研究的资源分配，类比为金融领域的“**投资组合管理**”，这是一个非常贴切的比喻。在AI领域，**计算资源（Compute）**就相当于投资中的“资本”，是推动一切发展的最核心、最稀缺的资源。

作为研究领导者，他们的核心工作之一，就是决定如何将有限的计算资源，分配到不同的“资产”上，以实现整体回报的最大化。这些“资产”包括：

- **长期探索（高风险、高回报的“风险投资”）**：投入到那些可能需要数年才能看到结果，但一旦成功就可能带来范式转移的基础研究上。
- **中期项目（稳健增长的“蓝筹股”）**：投入到那些技术路径相对明确，预计在一两年内能带来显著性能提升的项目上。
- **短期产品支持（维持运营的“现金流”）**：用于支持现有产品的运行、微调和优化。

他们的策略有几个特点：

1. **偏向长期**：他们明确表示，历史上，他们会有意识地将更多的资源向“核心算法进步”倾斜。这体现了他们对基础研究的战略重视。
2. **动态调整**：他们不遵循一个僵化的百分比分配规则，而是根据研究进展和外部环境的变化，进行动态的、灵活的调整。这要求领导者对整个研究领域有极强的感知和判断力。

这种将研究管理视为“投资组合管理”的视角，是一种非常成熟和理性的方法。它承认了研究的不确定性，并通过多元化的、动态的资源配置，来平衡风险与回报，确保组织在长期和短期目标之间都能取得进展。

2.12 计算、数据与人才的边际价值

原文翻译

> And if you had 10% more resources, would you put it toward compute or is it data curation, people? Where would you stick that from like a marginal uh
>
> > good question? Um honestly, yeah, I think um comput today reasonable answer. Yeah. Yeah. I mean, honestly, I I I do think kind of to your question of prioritization, right? It's like in a vacuum any of these things you would love to like go and excel and win at. Um I think the danger is you end up like second place at everything and you know not like you know clearly leading at at anything. So I think prioritization is important right and you need to make sure there's some things you're cleareyed on this is the thing that we need to win.

> 如果你们再多10%的资源，你们会把它投向计算、数据整理，还是人才？从边际效益的角度看，你们会投在哪里？ 好问题。老实说，是的，我认为是**计算（compute）**。 今天来看是合理的答案。是的。老实说，我认为这回到了你关于优先级的问题。在真空中，你当然希望在所有这些事情上都做到卓越并取胜。但危险在于，你最终可能在所有事情上都只拿到第二名，而在任何一件事情上都没有明确领先。所以我认为**优先级**非常重要，你需要确保 在某些事情上，你头脑清醒，知道“**这是我们必须赢下的战役**”。

深度解读

这个问题非常直接：在计算、数据和人才这三大AI发展的核心要素中，当前哪一个的“边际价值”最高？也就是说，再增加一个单位的投入，哪一项能带来最大的产出？

Mark Chen的回答毫不犹豫：“**计算**”。这个回答，以及随后的解释，揭示了OpenAI在当前阶段的核心战略判断：

1. **计算是当前的瓶颈**：尽管算法、数据和人才都至关重要，但在当前的技术节点上，最能限制或释放AI模型潜力的是计算资源的规模。更多的计算力，意味着可以训练更大、更强的模型，进行更多的实验，从而更快地推动技术边界。

2. **优先级的极端重要性**：他紧接着强调，之所以计算如此重要，是因为他们已经做出了战略取舍。他们认识到，试图在所有方向上都做到最好，最终的结果可能是在所有方向上都“屈居第二”。这是一种资源分散的陷阱。

3. **必须赢下的战役（The Thing We Need to Win）**：一个成功的组织，必须清醒地认识到自己的核心战场在哪里。对OpenAI来说，他们的核心战场就是**构建最强大的基础模型**。这是他们一切产品和研究的基石。因此，他们必须将最多的资源，投入到这个“必须赢下的战役”中。而在这场战役中，计算资源是最关键的“弹药”。

这个回答展现了一种非常清晰和专注的战略思维。在充满无限可能性的AI领域，最困难的不是找到能做的事情，而是决定**不做什么**。通过将“构建最强模型”作为最高优先级，并因此将计算资源置于核心地位，OpenAI为自己复杂的研发体系，提供了一个简单而有力的决策原则。

原文翻译

> > Yeah. Yeah. But I think it makes sense to talk about it for just a a little bit more which is compute sets so much of comput is destiny in a way right at a research organization like openi and so do would you a couple of years ago I think it became very fashionable to say oh okay we're not going to be compute constrained anytime soon because there's a bunch of CMS that are you know people are discovering and we're going to get more efficient and all the algorithms are going to get better and then eventually like really we'll just be in a data constrained regime And it seems like, you know, a couple years have come and gone and we're still like this is sort of very computed environment. Does that change anytime soon, you think? Or I mean I think like we've seen for long enough like how much we can do with compute. Um yeah, I I I I haven't really bought that much into the like will be data constraint claim and um yeah, I don't I don't I don't expect that to change. Yeah, anyone who says that should just step into my job for a week and there's no one who's like a you know I have all the compute that I need. Yeah.
>
> 是的。 是的。但我觉得有必要再多谈一点，那就是计算。在像OpenAI这样的研究机构里，某种程度上“计算决定命运”（compute is destiny），对吧？几年前，一种时髦的说法是，哦，我们很快就不会受计算限制了，因为人们发现了很多算法改进，我们会变得更高效，最终我们只会受数据限制。但几年过去了，我们似乎仍然处在一个非常受计算驱动的环境中。 你认为这种情况短期内会改变吗？我的意思是，我们已经看了足够长的时间，看到了计算能带来多大的成就。是的，我不太相信我们会受数据限制的说法。而且，是的，我不认为这种情况会改变。任何说这种话的人，都应该来我的岗位上干一周。没有人会说“我拥有我需要的所有计算资源”。

深度解读

这段对话驳斥了一个在AI领域流传甚广的说法，即“AI的瓶颈将从计算转向数据”。

几年前，很多人预测，随着算法效率的提升，我们将不再需要无限制地增加计算资源。同时，随着互联网上的高质量文本和图像数据被“耗尽”，**数据**将成为新的瓶颈。

然而，Jakub Pachocki和Mark Chen的观点非常明确：**计算的王者地位远未结束**。他们的理由基于长期的实践经验：

1. **规模化定律（Scaling Laws）的持续有效**：OpenAI的研究已经反复证明，在非常大的范围内，增加计算资源、模型参数和数据量，能够持续地、可预测地带来模型能力的提升。他们还没有看到这条定律失效的迹象。

2. **数据瓶颈被高估**：对于“数据耗尽”的担忧，他们似乎并不认同。这可能暗示着他们有办法生成高质量的合成数据，或者有能力从多模态、非结构化的数据中提取更多价值，从而绕过传统的数据瓶颈。

3. **计算需求的永无止境**：Mark Chen用一句非常有力的话总结了现状：“没有人会说‘我拥有我需要的所有计算资源’”。这表明，对于前沿研究来说，计算永远是稀缺资源。因为每当计算能力上一个台阶，研究人员就会立刻构想出需要更多计算能力才能实现的、更大胆的实验和模型。**需求永远跑在供给的前面**。

他们的结论是，“**计算决定命运**”在可预见的未来仍将是AI领域的黄金法则。这对于理解国家、企业和个人在AI时代的竞争力至关重要。能否获得和有效利用大规模计算资源，将继续是决定谁能引领下一波AI浪潮的关键因素。

2.13 学术界与产业界的交汇

原文翻译

> You know, historically the job of advancing fundamental research has historically been largely a mandate that universities have had partly for the compute reasons you just described. That has not been the case for Frontier AI. You guys have spent done such an incredible job kind of channeling the arc of Frontier AI progress to help the sciences out. Um, and I'm wondering when those worlds collide, the fundamental world of university research today and the world of frontier AI, what comes out? So, I guess I I personally started as a resident at OpenAI and it's a program that we had for uh people in different fields to come in, you know, learn quickly about about AI and become productive as a researcher. And I think there's a lot of really powerful elements in in that uh program. And you know the idea is just like you know could we accelerate something that looks like a PhD in in as as little time as possible. And I think a lot of that just looks like implementing a lot of you know very core results. And you know through doing that you're going to make mistakes. You're going to be like oh wow like build intuition for if I you know set this wrong like that's going to blow up my network in this way. Um and so you just need a lot of that hands-on experience. Um I think um over time, you know, there been curriculums developed at um probably all these large labs in in like optimization and architecture and RL and um yeah, probably no better way than to just kind of try to implement a lot of those things and read about them and think critically about them. Yeah.

> 历史上，推动基础研究的职责主要由大学承担，部分原因就是你刚才描述的计算资源问题。但在前沿AI领域，情况并非如此。你们在引导前沿AI进展来帮助科学方面做得非常出色。我想知道，当这两个世界——当今大学的基础研究世界和前沿AI的世界——发生碰撞时，会产生什么？ 我个人是在OpenAI作为“驻留研究员”（resident）开始的，这是我们为不同领域的人设立的一个项目，让他们进来快速学习AI，并成为有生产力的研究员。我认为这个项目有很多非常强大的元素。我们的想法是，能否在尽可能短的时间内，加速一个类似博士（PhD）的过程。我认为，其中很大一部分就是去**实现（implement）很多非常核心的研究成果。通过这样做，你会犯错，你会感叹“哇”，并建立起直觉，比如“如果我这个参数设置错了，我的网络就会以这种方式崩溃”。所以你需要大量的动手经验**。我认为，随着时间的推移，可能所有这些大型实验室都开发出了关于优化、架构和强化学习等方面的课程。可能没有比亲自去尝试实现很多这些东西、阅读相关文献并进行批判性思考更好的学习方法了。

深度解读

这段对话探讨了在AI时代，传统的学术培养模式（如博士项目）与产业界前沿实验室的人才培养模式之间的关系。

历史上，大学是基础研究的中心，博士项目是培养顶尖研究人才的主要途径。一个博士生通常需要花费5-6年时间，在一个狭窄的领域进行深入探索。然而，由于前沿AI研究所需的巨大计算资源和工程团队，研究的重心已经部分地从大学转移到了像OpenAI这样的产业实验室。

Mark Chen以OpenAI的“**驻留研究员**”项目为例，提出了一种**加速版的、实践驱动的“博士培养”模式**。这种模式的核心在于：

1. **以“实现”代替“理论”**：传统博士教育可能包含大量理论课程。而OpenAI的模式强调“**动手实现**”。他们认为，学习AI最快的方式，就是亲手去复现那些领域内最重要的、最核心的论文和算法。

2. **从错误中建立直觉**：在实现的过程中，你不可避免地会遇到各种各样的bug和失败。正是这些“血与泪”的教训，让你能够建立起对AI系统行为的深刻**直觉（intuition）**。这种直觉——“我知道如果这么调参，模型很可能会崩溃”——是书本无法教给你的，只能通过大量的实践和试错来获得。

3. **批判性思考**：在实现和阅读的基础上，进行批判性思考，理解算法背后的原理，并思考如何改进它。

这种模式的本质，是将传统博士培养中“学习-实践-研究”的漫长链条，压缩成一个“**在实践中学习，在学习中研究**”的高速迭代循环。它特别适合AI这样一个实验性极强、发展速度极快的领域。

当大学的基础研究与产业界的前沿探索碰撞时，可能会产生一种新的、混合的人才培养模式：大学提供扎实的理论基础和批判性思维训练，而产业