黄仁勋、扎克伯格巅峰对谈实录：万字长文揭秘Meta的未来AI图景

文 / AI未来指北作者 郝博阳

作为拥有600,000块H100的男人，在AI界占据着开源制高点的扎克伯格会如何看待AI的未来？

也许你已经有了一些答案。他在最近发布最强开源大模型Llama 3.1之后，已经密集的接受了一系列采访，讲述了他对AI产业的诸多看法。

但在今天的SIGGRAPH大会上，你看到的可能是最真实的扎克伯格。在和黄仁勋的巅峰对谈中，他自曝自己开源也有自私的一面，飙脏话怒斥闭源供货商的傲慢。还深刻反思了自己过去不够潮，做的产品不够酷了。

而且这还可能是看到的最全面的Meta未来图景。这长达一个小时的对谈中，他们从产品到产业，从开源的信仰到计算平台的迭代。基本上把AI这个产业的未来都聊通了。

对AI产品方面的未来，扎克伯格表示AI正在彻底改变社交媒体平台的运作方式：未来的Facebook和Instagram将成为一个统一的AI模型，能够整合不同类型的内容和系统，为用户提供更加个性化和丰富的体验。特别提到了推荐系统的革新，预测未来的内容不仅会基于用户兴趣进行推荐，还会通过AI工具实时创作或合成。

他还非常坚定的认定了智能体就是AI产品的下一个形态。他认为，“在未来，就像每个企业都有一个电子邮件地址、一个网站和一个或多个社交媒体账户，我认为在未来，每个企业都将拥有一个与客户交互的AI智能体”，而智能体的可能性之大，即使基础模型的进展现在停滞不前，还会有五年的时间来让行业基本上弄清楚如何最有效地利用迄今为止已经构建的所有东西。

而在AI产业方面，扎克伯格表达了对开放生态系统的坚定支持。扎克伯格透露私心，认为就算为了确保Meta等公司不会被封闭的生态系统所限制，也得自己做开源。而且只有开放的生态才能导致技术标准的形成，能用用户飞轮去更快速的迭代产品。而且，在这样的开源生态中，他预见到未来将出现大量针对特定任务和领域的专门AI模型，而不是由少数几个大型通用模型主导。

关于下一代计算平台，扎克伯格重点讨论了智能眼镜和混合现实设备的前景。因为AI始料未及的快速发展，小扎现在的元宇宙不再是非得虚拟现实了。而是AI+眼镜就行。和雷朋的合作就是他这个新认知的具象体现。

而且在整个对谈中，老黄和小扎棋逢对手，互相揶揄和玩笑不断。自称自己“不善言辞”的黄仁勋和素有“机器人”之称的扎克伯格坐在一起互相调侃，甚至拉到自家农场的家常。不由得让人觉得，人，确实是会变的。

想知道Meta的路线图，这篇万字对谈值得你细读。想看看甲方乙方谈话的艺术，这篇对谈也值得一读。

以下为访谈实录：

黄仁勋：你能相信吗？这是现代计算的先驱之一，推动者。我邀请他来参加SIGGRAPH了。马克，坐下吧。很高兴你能来这里。欢迎。谢谢你飞过来。我听说你已经连续飞行了大约五个小时。

扎克伯格：是的，当然。

黄仁勋： 这是SIGGRAPH，你知道吧？这里有90%是博士。SIGGRAPH真正了不起的地方是，它是计算机图形学、图像处理、人工智能和机器人技术的综合展示。多年来，一些公司如迪士尼、皮克斯、Adobe、Epic Games，当然还有NVIDIA，在这里展示和揭露了许多令人惊叹的东西。我们在这里做了很多工作。今年，我们在人工智能与模拟的交叉领域引入了20篇论文。

我们正在利用人工智能使模拟规模更大、速度更快，例如，可微分物理学。我们创建了用于人工智能合成数据生成的模拟环境。这两个领域正在真正融合。我对我们在这里所做的工作感到非常自豪。Meta，你们在AI方面做了令人惊叹的工作。我发现有趣的一点是，媒体写到Meta在过去几年如何涉足AI领域。FAIR做了很多工作——记住，我们都使用PyTorch。那源自Meta。

你们在计算机视觉方面的工作，语言模型方面的工作，实时翻译上都有开创性的工作。我想问你的第一个问题是，您如何看待Meta当前在生成式AI方面取得的进展？以及您如何应用它来增强您的运营或引入您正在提供的新功能？

扎克伯格：是的，有很多内容需要解释。

首先，非常高兴能来到这里。Meta已经连续八年参加SIGGRAPH。与你们相比，我们还是新手。但我想是在2018年——（被打断）

黄仁勋：你的装扮很合适（指扎克伯格的黑色衬衫和自己撞衫）。



## Meta AI改变脸书的双核心：

推荐系统和创作能力

扎克伯格：（笑）这里是你的地盘。

（2018年）我们展示了一些为我们的VR和混合现实头显所做的早期手部追踪工作。我认为我们那时讨论过Meta在编解码数字人方面取得的进展。我们希望从消费者级头显中驱动那种逼真的数字化身，而我们正越来越接近这一目标。此外，我们还完成了很多显示系统的工作。我们正在研发的未来原型，旨在使混合现实头显能够变得非常轻薄。我们在光学堆栈和显示系统方面已经相当先进，这是一个集成系统。这些通常是我们首先在这里展示的内容。

所以很兴奋今年能来到这里，不仅讨论元宇宙相关的内容，还有所有的人工智能部分。正如你所说，我们当初开始时，人工智能研究中心还是Facebook。现在是Meta，在我们成立Reality Labs之前就已经在这方面工作了一段时间。所有围绕生成式AI的内容——这是一场有趣的革命。我认为这将最终使我们所做的所有不同产品以有趣的方式变得不同。

我先大概梳理一下我们已经拥有的大型产品线。比如信息流和推荐系统，还有Instagram和Facebook。我们一直在这样的旅程中，从仅仅是为了与朋友联系开始（发展到如今的复杂系统）。排序机制一直很重要，因为即使你仅仅关注了朋友，如果某人做了非常重要的事情，比如你的表亲生了孩子之类的，你会希望这个消息排在最前面。如果我们把它埋在你的信息流某个角落，你肯定会很生气。

但如今，在过去几年里，用户的关注已经发展到更多是关于外面世界的各种公共内容了。推荐系统变得极其重要，因为现在不再是只有几百或一千条来自朋友们的潜在发布，而是有数百万条内容。这就变成了一个非常有趣的推荐问题。而借助生成式AI，我认为我们将很快进入这样一个领域：你今天在Instagram上看到的大部分内容都是根据你的兴趣和是否关注这些人，从外界匹配推荐给你的。但我认为在未来，很多这类内容也将通过这些AI工具来创作。其中一部分将是创作者使用这些工具创作的新内容。而另一部分，我想，最终会是为了你即时创作或从现有内容中整合合成的内容。这只是我们核心业务如何进化的一个例子。而且它已经发展了20年。

黄仁勋：嗯，很少有人意识到，世界上构想过的最大的计算系统之一就是推荐系统。

扎克伯格：是的，我的意思是，这是一条完全不同的道路。它并不是人们谈论的那种热门的生成式AI。但我认为，推荐系统用到的也是Transformer架构。（生成式AI）也是一种类似的情况，（我们现在）只是不断构建越来越通用的模型。

黄仁勋：就是将非结构化数据嵌入到特征中。

扎克伯格：是的，其中一个重要方面是它推动了质量的提升。过去，每种内容类型都有不同的模型。最近的一个例子是，我们有一个模型用于排名和推荐短视频，另一个模型用于排名和推荐更长的视频。然后需要一些产品工作来确保系统能够在线展示所有内容。但如果你能创建更通用的推荐模型，覆盖所有领域，效果只会越来越好。我认为部分原因在于内容的经济性和流动性。你能够从中汲取的内容池子越广，你就不会像从不同池子中抽取那样遇到奇怪的低效问题。

黄仁勋：是的，随着模型变得更大更通用，效果会越来越好。

扎克伯格：所以我梦想有一天，你可以想象整个Facebook或Instagram就像一个统一的AI模型，整合了所有不同类型的内容和系统，这些内容和系统实际上在不同的时间框架下有着不同的目标。因为其中一部分只是向你展示今天你想要看到的有趣内容。但另一部分则是帮助你长期构建你的社交网络，推荐你可能认识的人，或是你可能想要关注的账号。而这些多模态模型往往能提供更好的陪伴。

黄仁勋：你一直在构建GPU基础设施，长时间运行这些大型推荐系统。

扎克伯格：实际上，谈到GPU，在这方面的进展就有点慢了。（笑）

黄仁勋： 是的。你总是尽量表现出善意。

扎克伯格：是的，你知道的，非常善意。

黄仁勋：确实，应该总是保持善意的。你是我的客人。你知道，在他上台前，在后台的时候，你提到了关于承认自己的错误之类的事儿，对吧？所以...（笑）你没必要这么主动的想讲出来是吧？

扎克伯格：（笑）我觉得这点我已经很好地尝试过了。

黄仁勋：（笑）是的。但一旦你进入表演状态，你就会全力以赴。这就对了。

现在，关于生成式AI真正酷的一点是，如今我用WhatsApp时，感觉就像在与WhatsApp合作。我喜欢Imagine。我坐在这里打字，它随着我的输入生成图像。我回头修改文字，它就生成其他图像。比如一个中国老人在黄昏时分享受一杯威士忌，身边有三条狗：金毛寻回犬、金毛贵宾犬和伯尼山犬。它就根据这个生成了一张相当不错的图片。

扎克伯格：是的，我们正朝着那个方向前进。

黄仁勋：现在你实际上可以把我的照片上传上去进去。（所以那个老人）那就是我。

## AI产品的未来：

就算现在模型停止迭代，我们也得用五年

才能发掘出智能体的全部产品潜力

扎克伯格：我最近花了很多时间和女儿们在一起，想象她们变成美人鱼之类的角色。非常有趣。一方面，我认为，生成式AI将是对我们长期使用的所有工作流程和产品的一次重大升级。但另一方面，现在它可以创造出许多全新的东西。

所以MetaAI，不光拥有一个能帮你处理不同任务的AI助手的想法，而且在我们的世界它也将会非常注重创意。就像你说的那样。它们非常通用，所以你不必仅仅局限于此。它能回答任何问题。

随着时间的推移，当我们从Llama 3类模型发展到Llama 4及以后，我认为，它将会感觉不像是一个聊天机器人，那种你给它一个提示，它就简单回应，然后你再给它一个提示，它又回应。就这样一来一回。我觉得它会很快进化到，你给它一个意图，它实际上就能在多个时间框架内处理它。我的意思是，它可能会根据之前的谈话预判你的意图。但我想，AI最终会变成能启动需要数周或数月的计算任务，并在世界上发生某些事情时回来找你（并告知你处理方法）。我认为那将会非常强大。

黄仁勋：所以，我的意思是我不太确定。如你所知，今天的AI基本上是回合制的。你说一句话，它回你一句话。但显然，当我们思考，当我们被赋予一个任务或面对一个问题时，你知道，我们将考虑多种选择，或者我们可能会提出一个选项树，一个决策树，我们会在心中模拟沿着决策树走下去，你知道，我们可能会做出的每个决策的不同结果是什么。所以我们正在进行规划。因此，未来的人工智能也会做类似的事情。当你谈到你对Creator AI的愿景时，我感到非常兴奋，坦白说，我认为那是一个绝佳的想法。请向大家介绍一下Creator AI和AI工作室，正是它们将使你能够实现那个愿景。

扎克伯格：是的。实际上，我们之前已经讨论过一些，但今天我们将更广泛地推出它。我们的愿景是，我不认为只会有一款人工智能模型，对吧？这是行业内一些其他公司所不具备的，他们似乎在构建一个中心化的智能体系统。而我们确实提供了MetaAI助手供您使用，但我们的愿景更多在于赋能每一位使用我们产品的用户，让他们能够为自己打造专属的智能体。

无论是平台上众多的数百万创作者还是数亿小型企业，我们最终目标是将您的所有内容整合，并迅速搭建一个业务智能体，构建与客户的互动、进行销售及客户支持等全方位服务。因此，我们正在逐步推出的，就是我们称之为AI工作室的项目，它旨在助力您快速启动并运营这一智能体系统。这基本上是一套工具，最终将使每位创作者都能构建一个自己的AI版本，作为某种智能体或助手，让他们的社区能够与之互动。

这里存在一个根本问题，即一天中根本没有足够的时间，对吧？就像如果你是一位创作者，你希望更多地与你的社区互动。但你受限于时间做不到。同样地，你的社区也希望与你互动。但这很困难。我的意思是，能用来做这件事的时间是有限的。所以，次佳的选择是，让人们能够创造一种智能体，你通过训练它，让它基于你的素材来代表你，以你希望的方式呈现。

我认为这是一项非常具有创造性的努力，几乎像是一种艺术品或你发布的内容。而且，很明显，它并不是在与创作者本人互动，但我认为这将是一种有趣的方式，就像创作者在社交平台上发布内容这样的人，就能够拥有智能体来完成。同样地，我认为人们将会创造自己的智能体，用于各种不同的用途。有些会是定制化的实用工具，他们希望对其进行精细调整和训练。

有些则会是娱乐性质的。人们创造的一些东西只是有趣，你知道的，它有些可能有点傻或者对事物有一种幽默的态度，我们可能无法预料。

用户可能不会将MetaAI构建为助手，但我认为人们对看到和与之互动相当感兴趣。然后我们看到的其中一个有趣的应用场景是人们使用这些智能体进行支持。这一点是让我有点惊讶的，MetaAI最受欢迎的用途之一是人们基本上用它来模拟他们即将面临的复杂社交情境。无论是职业场合，比如说，"我想问我的经理，我怎样才能得到晋升或加薪"，或者"我和朋友发生了争执"，或者"我和女朋友遇到了困难"，基本上在一个完全没有偏见的空间里，你可以模拟这种情况并看看会如何发展，如何进行对话并获得反馈。

很多人，他们不仅仅想与同类型的智能体交互，无论是MetaAI还是ChatGPT，或者其他大家都在使用的工具，他们想要创造自己的东西。

所以这就是我们大致上在AI工作室所要做的方向。但这些都是我们更大的观点的一部分，我们认为不应该只有一种大型AI供人们交互。我们只是认为，如果存在多种不同的AI，世界会变得更好、更有趣。

黄仁勋：我认为这非常酷，如果你是一位艺术家并且有自己的风格，你可以采用你的风格，你所有的作品集，你可以微调你的一个模型，现在这变成了一个AI模型，你可以来使用它并给出提示。你可以让我，比如说，创作一些与我所拥有的艺术风格类似的作品，你甚至可以给我一幅作为绘图的艺术品，一个草图作为灵感，我就能为你生成一些东西。然后，其他用户会来到来到我的AI那里（获得这些能力）。每一个餐厅，每一个网站在未来可能都会有这些AI。

扎克伯格：是的，我有点认为在未来，就像每个企业都有一个电子邮件地址、一个网站和一个或多个社交媒体账户，我认为在未来，每个企业都将拥有一个与客户交互的AI智能体。

黄仁勋： 对。

扎克伯格：这些事情我认为是历史上一直很难做到的。比如你考虑任何公司，客户支持部门与销售部门是两个完全独立的组织。而作为CEO，你并不希望它们这样运作。只是，好吧，它们所需的技能确实有所不同。当你是CEO时，你得处理所有这些事情。

我的意思是，一般来说，当你搭建组织的时候，组织是分开的。因为它们各自优化于不同的事务。但我认为理想化的状态是它应该是一种统一的存在，对吧？作为消费者，你不想在购买东西时走一条路线，而在遇到已购商品问题时又走另一条路线，你只想有一个地方可以去，能解答你的问题，并且能够以不同的方式与企业互动。我认为这对创作者也同样适用。

黄仁勋：我认为这种个人消费者方面的互动，尤其是他们的投诉，将会使你的公司变得更好。

扎克伯格：是的，完全正确。

黄仁勋：所有这些AI与客户的互动都会让AI捕捉到那些机构知识，而这些又都可以进入分析系统，这会提升后续人工智能。

扎克伯格：是的。你说的这个版本我认为是个整合度更高的版本，我们仍处于相当早期的alpha阶段。但AI工作室让人们能够创建自己的UGC智能体和其他内容，开始进入这个让创作者创作的飞轮。我对这个挺兴奋的。

扎克伯格: 是的，我们正朝那个方向发展。

扎克伯格: 是的，基本上可以。

扎克伯格: 是的。而且我的意思是，就像任何会随着时间改进的产品一样，用于训练它的工具也会变得更好。这不仅仅是关于你希望它说什么。我认为通常创作者和企业也有他们想要避开的主题。

黄仁勋：对。所以在这方面不断进步。

扎克伯格：是的。你知道，我认为理想的版本不仅仅是文本，对吧？这和我们随着时间在做的一些编解码虚拟形象的工作有些交集。你肯定希望能够和那个智能体进行像视频聊天一样的交互，我认为我们随着时间会达到那一步的。我不认为这些东西离我们太远，这个飞轮正在非常快速地旋转。所以这是令人兴奋的。有很多新东西需要构建。

我认为即使基础模型的进展现在停滞不前，尽管我不认为它会停滞，我想我们还会有五年的时间来让行业基本上弄清楚如何最有效地利用迄今为止已经构建的所有东西。但实际上，我认为，基础模型和基础研究方面的进展正在加速。所以，这是一个非常狂野的时期。

黄仁勋：你的愿景是这一切，你某种程度上促成了这一切。所以，谢谢你。你知道，我们是CEO，我们是脆弱的花朵。我们需要很多支持。（笑）

扎克伯格：是的。我们现在已经相当坚韧了。我认为我们是行业中两位任职时间最长的创始人。对吧？

黄仁勋：我的意思是，这是真的。这是真的。

扎克伯格：而且你的头发变灰了。我的只是变长了。（笑）

黄仁勋：我的头发是变灰了。你的变卷了。这是发生了什么？

扎克伯格：它一直都很卷。所以我一直保持短发。

黄仁勋： 是啊。我只是，如果我知道成功需要这么长时间……

扎克伯格：你可能一开始就不会开始。（笑）

黄仁勋：不，我可能会像你一样从大学退学。早点开始。（笑）

扎克伯格：嗯，这是我们性格的一个好差别。（笑）

黄仁勋：你已经领先了12年。那很不错。

扎克伯格：你知道，你做得相当不错。

黄仁勋：我会继续前进的。（笑）

所以，我喜欢的，关于您对每个人都能拥有AI，每个企业都能在我们公司拥有AI的愿景。我希望每位工程师和每位软件开发者都能拥有一个AI。

扎克伯格：是的。我们有很多AI。

黄仁勋：我欣赏您的愿景的一点是，您也相信每个人都和每个公司都应该能够创造自己的AI。所以您实际上是开源了。当您开源了Llama时，我认为那很棒。Llama 2.1，顺便说一下，我认为Llama 2可能是去年AI领域最大的事件。

扎克伯格：我原本以为（去年AI领域最大的事件）是H100。（笑）

黄仁勋： 这是个先有鸡还是先有蛋的问题。是的。所以这两者哪个先出现的？

扎克伯格： H100。嗯，Llama 2实际上不是H100。（笑）

黄仁勋： 是的，它是A100。（笑）



## 开源的未来：

### 不受限制的构建，恢复开放黄金时代

黄仁勋：是的，谢谢。因此，但我说这是最大事件的原因在于，当Llama 2.1 发布时，它激活了每一家公司、每一个企业和每一个行业。突然之间，每家医疗公司都在开发AI，每家公司，无论大小，初创企业都在构建AI。它让每一位研究者都能重新投入到AI领域，因为他们有了一个起点去做些什么。而现在，3.1版本已经发布，你知道的，我们共同努力部署Llama 3.1，将其推向全球企业。这种兴奋之情简直难以言表。我认为它将推动各种应用的发展。

但请谈谈你的开源理念。这一理念源自何处？你开源了PyTorch，现在它已成为人工智能开发的框架。而现在，你们已经开源了Llama 3.1。围绕它建立了一个完整的生态系统。所以，我认为这非常棒。但这一切是从哪里来的呢？

扎克伯格：是的，这里面有很多历史。我的意思是，随着时间的推移，我们做了很多开源工作。我认为部分原因，坦白说，是我们在其他一些科技公司之后才开始构建分布式计算基础设施和数据中心这类东西。而且，正因为如此，当我们构建这些东西时，它已经不是竞争优势了。所以我们想，那好吧，我们不如干脆将其开源。然后我们将从围绕它的生态系统中受益。所以我们有很多这样的项目。

我认为最大的一个例子可能是开放计算项目（Open Compute），我们公开了我们的服务器设计、网络设计，最终还包括数据中心设计，并发布了所有这些内容。通过使其成为某种行业标准，整个供应链基本上都围绕它进行了组织，这带来了为每个人节省资金的好处。因此，通过公开和开放这些设计，我们基本上通过这样做节省了数十亿美元。

黄仁勋：嗯，开放计算项目也使得NVIDIA HGX成为可能，我们为一个数据中心设计的东西突然间也能在其他数据中心工作。

扎克伯格：确实如此。太棒了。所以我们有了这次非常棒的经历。之后，我们还将其应用于我们的一系列基础设施工具，比如React、PyTorch等。因此，可以说到了Llama出现的时候，我们对做这件事有一种积极倾向。

特别是对于AI模型，我想我有几种看待这个问题的方式。首先，在过去20年里，在公司里构建东西中最困难的事情之一就是不得不应对这样一个事实：我们通过竞争对手的移动平台发布我们的应用。一方面，移动平台对整个行业产生了巨大的推动作用，这非常棒。另一方面，不得不通过竞争对手来交付你的产品是具有挑战性的，对吧？

而且，在我成长的时代，Facebook的第一个版本是在网络上，那是开放的。然后转向移动端，其好处是，现在每个人口袋里都有一台电脑。那真是太好了。但不利的一面是，我们在能做的事情上受到了更多限制。因此，当你审视这些计算时代的更迭，你会发现存在一个巨大的近因偏差，人们现在只关注移动设备时代，并认为封闭的生态系统（会带来更好的结果），因为苹果基本上赢得了这场胜利并设定了规则。是的，我知道从技术上讲，有更多的安卓手机存在，但苹果基本上掌控了整个市场，而且（赢得了）所有的利润，基本上安卓在发展方面也在跟随苹果的脚步。所以我认为苹果在这一代中明显胜出。

但情况并非总是如此。

如果你回顾上一代，苹果在搞他们那种封闭的模式（MacOS）。但微软，它显然不是一个完全开放的公司，但相较于苹果，Windows在各种不同的OEM厂商和不同软件、硬件上运行，是一个更为开放的生态系统。而Windows（那时）是领先的生态系统。基本上在个人电脑这一代产品中，开放的生态系统胜出了。

我有点希望在下一波计算浪潮中，我们将重返一个开放生态系统胜出并再次成为主导的领域。总是会有封闭的和开放的两种选择。我知道两者都有其存在的理由。两者都有其好处。我并不是在这方面极端的人。我的意思是，我们也有闭源的产品。我们发布的并非所有内容都是开源的。

但我认为，对于整个行业所依赖的计算平台而言，尤其是软件如果开放，会有很大的价值。这确实塑造了我在这方面的哲学。无论是与Llama合作的AI项目，还是我们在AR和VR领域的工作，我们基本上是在为混合现实构建一个开放的操作系统，类似于Android或Windows，旨在与众多不同的硬件公司合作，制造各种类型的设备，我们基本上只是希望将生态系统恢复到那个开放的水平。我对下一代的发展相当乐观，开放的系统将会胜出。

对我们来说，我只想确保我们自己能够接入（AI）。这么说有点自私，但我的意思是，在创建这家公司一段时间后，我未来10到15年的目标之一就是，我只是想确保我们能够构建基础技术，因为这将是我们构建社交体验的基础，因为之前我尝试构建的许多东西都受到了限制。然后你被告知，不，你无法真正构建那个，这是平台提供商在某一层面的说法。我只想说，不，去他的。对于下一代，我们将一路构建下去。

黄仁勋：好的，我们的直播要被掐了（因为扎克伯格说了Fword）

扎克伯格：是的，抱歉。抱歉。（笑）

黄仁勋：我当时就想说，哔。（笑）

扎克伯格：是的。你知道，我们前20分钟做得还不错，但一谈及封闭平台，我就生气。（笑）

黄仁勋：有人致力于构建尽可能最好的AI，无论他们如何构建，并将其作为服务提供给世界，我认为这才是一个很棒的世界。然后，如果你想自己构建AI，你仍然可以自己构建AI，这都很棒。但在使用AI的能力这件事上，你知道，有很多东西（我更喜欢开盒即用）。我更倾向于不自己制作这件夹克。我更倾向于让别人为我制作这件夹克。你明白我的意思吗？

扎克伯格：是的，是的。

黄仁勋：皮革可以开源对我来说不是一个有用的概念。但我认为你可以拥有伟大的服务，令人难以置信的服务，以及开放的服务，开放的。

扎克伯格： 是的。

黄仁勋：那么我们基本上涵盖了整个范围。但你用Llama 3.1所做的事真的很棒，你有405B的版本，你有70B的版本，你有8B的版本，你可以用它来生成合成数据，用更大的模型来教小模型。虽然更大的模型会更通用，它也不那么脆弱，但你仍然可以构建一个更小的模型，适合于任何你希望的操作领域或运营成本。无论你希望的操作领域或运营成本是什么，你都可以构建一个更小的模型来适应。

你创建了一个守卫，我想是叫Llama守卫（Llama Guard），Llama守卫用于护栏，非常棒。所以现在你构建模型的方法，它是以一种透明的方式构建的。你拥有一个世界级的安全团队，世界级的伦理团队，你可以以这样一种方式构建它，让每个人都知道它是正确构建的。我真的很喜欢这部分。

扎克伯格：是的，我先说完在我之前被岔开话题之前想说的话。你知道，我确实认为我们有这种一致性，我们构建（开源AI）是因为我们希望它存在，并且我们不希望被某些封闭的模型所隔离。而这不仅仅像是开发一款软件那么简单，AI需要一个生态系统的支撑。因此，几乎可以说，如果我们不将其开源，它甚至可能无法很好地运作。

我们这样做并非因为我们是什么无私的人，尽管我认为这对生态系统是有益的。我们这样做是因为我们认为，通过建立一个强大的生态系统，可以使我们正在构建的东西达到最佳状态，看看有多少人参与到了PyTorch生态系统的建设中。

黄仁勋：（这需要）大量的工程工作。我的意思是，仅视频处理方面，我们可能就有数百人专门致力于让PyTorch更优秀、更具扩展性，你知道的，性能更佳等等。

扎克伯格：是的，而且当某个技术成为行业标准的时候，其他人就会围绕它进行创新，对吧。因此，所有的硬件和系统最终都会被优化，以便非常好地运行这项技术，这将使所有人受益，同时它也能很好地兼容我们正在构建的系统。而这一点，我认为只是展示了（开源）这种方式最终会变得非常有效的一个例子。

扎克伯格：所以，我认为开源策略作为商业策略将会是一个好的选择。我想人们可能还没完全意识到我们有多么热爱它。

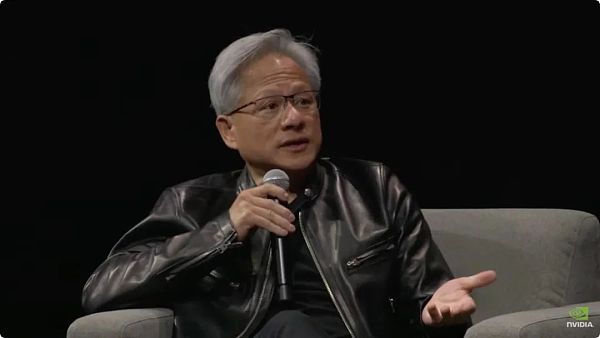
黄仁勋：我们围绕它建立了一个生态系统。我们创造了这个东西。

扎克伯格：是的，我发现这一点了。你们团队的表现一直很出色。每次我们推出新产品，你们是首批发布并优化使其运作的团队。我很感激这一点。

黄仁勋：我还能说什么呢？我们确实有不少优秀的工程师。

扎克伯格：（笑）而且，你们也总是迅速抓住这些机会。

黄仁勋：所以，我虽是长者，但行动敏捷。这就是CEO必须做的。



我认识到了一件重要的事情。我认为Llama确实非常重要。我们围绕它构建了一个名为AI工厂（AI Foundry）的概念，以便帮助每个人构建AI。很多人，他们有构建AI的愿望。对他们来说，拥有AI非常重要，因为一旦他们将其融入数据飞轮，这就是他们公司机构知识被编码并嵌入AI的方式。但他们承担不起让那个AI飞轮，数据飞轮在其他地方（通过购买服务）转起来的成本。因此开源允许他们这样做。但他们并不真正知道如何将这一切转化为AI。所以我们创建了这个名为AI Foundry的东西。我们提供工具，提供专业知识，Llama的技术，我们有能力帮助他们将这一切转化为AI服务。然后当我们完成这一切后，他们接手，他们拥有，其输出就是我们所说的NIMM。这个NIMM，这个神经微型NVIDIA推理微服务，他们只需下载，带走并在任何他们喜欢的地方运行，包括本地部署。我们拥有一个完整的合作伙伴生态系统，从能够运行NIMMs的OEM到像Accenture这样的GSIs，我们培训并与他们合作创建基于Llama的NIMMs和管道。现在我们正在帮助全球各地的企业实现这一目标。我的意思是，这确实是一件非常令人兴奋的事情。这一切实际上都是由Llama的开源引发的。

## AI产业的未来：

### 模型不会一家独大，

### 从小到大的模型都有场景

扎克伯格： 是的，我认为，帮助人们从大型模型中提炼出自己的模型的能力，将成为一个真正有价值的新事物。就像我们在产品方面讨论的那样，至少我不认为会有一个每个人都会去跟它交流的核心AI智能体。在同一水平上，我也不认为必然会有一个模型是每个人都会使用的。

黄仁勋： 我们有一个芯片AI，芯片设计AI。我们有一个软件编码AI。我们的软件编码AI理解USD，因为我们用USD为Omniverse编写代码。我们有一个理解Verilog的软件AI，我们的Verilog。我们拥有能够理解我们的缺陷数据库的软件AI，并且知道如何帮助我们分类缺陷并将其发送给正确的工程师。

这些AI中的每一个都是基于Llama进行微调的。我们会对它们进行微调，并设置防护措施。你知道，如果我们有一个用于芯片设计的AI，我们并不希望询问它关于政治、宗教之类的问题。所以我们会对它进行防护限制。因此，我认为每家公司基本上都会为它们拥有的每一个功能，配备专门为此构建的AI。他们需要帮助来实现这一点。

扎克伯格：是的，我认为未来的一大问题是，人们将多大程度上使用更大、更复杂的模型，而不是仅仅针对他们的需求训练自己的模型。至少我可以肯定，未来会有各种各样、数量庞大的不同模型涌现。

黄仁勋：我们使用的是最大型的那些模型。而这样做的原因在于，我们的工程师时间极其宝贵。因此，我们现在正针对性能优化405B版本的Llama 3.1。如你所知，无论GPU多大，405B都无法完全适配。这就是为什么NVLink的性能如此关键。我们采用了这种技术，通过一个名为NVLink的非阻塞交换机，将每块GPU连接起来。

例如，在一个HGX中，就有两个这样的交换机。我们使得所有这些GPU能够协同工作，运行405B时性能极为出色。我们这样做的原因是，工程师的时间对我们来说极其宝贵。你知道，我们希望使用尽可能最佳的模型。即便这样做在成本上只节省了几分钱，谁又会在乎呢？因为我们希望确保向他们展示的是最优质的结果。

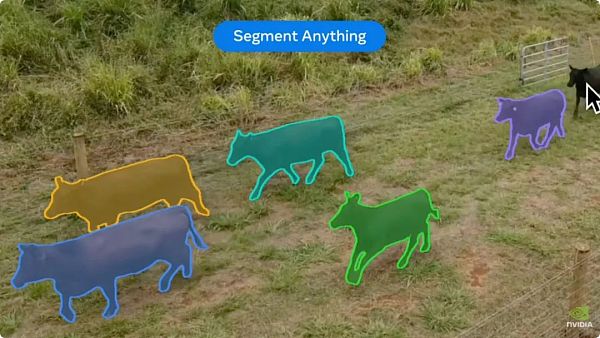
扎克伯格：是的，我的意思是，我认为405B的成本大约是GPT-4模型的一半。所以，从这个层面来说，它已经相当不错了。但我认为人们正在设备上使用或需要更小型的模型，他们会将其精简。因此，这就像是AI运行的一整套不同的服务。

黄仁勋：让我们假设一下，我们正在用于芯片设计的AI每小时可能只需10美元成本。你知道，如果你持续使用它，并且将那个AI共享给众多工程师，那每个工程师可能都有一个成本不高的AI陪伴着他们。这个AI的成本其实并不高。而我们支付给工程师的薪酬却很高。

因此，对我们来说，每小时几美元就能大幅提升那些尚未接入AI的人的能力。立刻行动，接入一个AI吧。我们想说的就是这些。

那么，让我们谈谈下一波趋势。你们所做的工作中，我特别喜欢的一点是，计算机视觉，我们内部大量使用的一个模型是 Segment Everything。你知道，我们现在正在视频上训练AI模型，以便我们能更好地理解世界模型。我们的应用场景是机器人技术和工业数字化，将这些AI模型接入Omniverse，以便我们能更好地建模和表达物理世界。让机器人在这些Omniverse世界中更好地运作。因此你的应用，Ray-Ban Meta眼镜，你将AI引入虚拟世界的愿景真的很有趣。给我们讲讲吧。

扎克伯格：好的，嗯，这里面有很多值得探讨的内容。你所谈论的那个Segment Everything模型，我们实际上在SIGGRAPH上展示了它的下一个版本，Segment Everything 2。现在它已经能正常运行了，速度更快了。它现在也能处理视频了。我觉得这些实际上是我考艾岛牧场的牛。顺便说一下，它们被称为马克的小牛，美味的小牛。



黄仁勋：美味的马克的小牛。所以马克下次来我家，我们得一起做费城奶酪牛排。你直接带牛过来。

扎克伯格：然后你做奶酪，我就当副厨。这牛真的非常好吃。

黄仁勋：这是副厨的评价。

扎克伯格：好吧，听着。然后在晚上结束时，你就像，"嘿，你吃得够多了，对吧？"然后我就说："不知道，我还能再吃一个。"你就像："真的吗？"

黄仁勋：我肯定像是在说："对，我们还要再做一些。我们还要再做一些。你吃饱了吗？"通常你的客人会说："哦，是的，我很好。"

扎克伯格：“再给我做一个芝士牛排，老黄。”

黄仁勋： 所以，为了让你知道他有多强迫症，我转过身去，我在准备芝士牛排。然后我说："马克，切一下西红柿。" 递给马克一把刀。

扎克伯格： 对，我是个精确的切割者。

黄仁勋：然后他就切西红柿。每一个都切得精确到毫米。但真正有趣的是，我以为所有的西红柿都会被切成片，然后像一副扑克牌那样叠起来。当我转过身时，他说他需要另一个盘子。而他这样做的原因是，他切的每一片西红柿，彼此之间都不能有接触。一旦他把一片西红柿与其他西红柿分开，它们就不应该再碰到一起。

扎克伯格：是啊，你看，伙计，如果你希望它们接触，你应该提前告诉我。你需要...我只是一个副厨，好吗？

黄仁勋：这就是为什么他需要一个不带偏见的AI。

扎克伯格：是的。（笑）

黄仁勋： 这真的很酷。好的，所以它在识别牛的足迹...它在追踪牛的足迹。

扎克伯格：用这个可以制作很多有趣的特效。而且因为它会被广泛开放，行业内还会有更多严肃的应用。所以我的意思是，科学家用这些东西来研究珊瑚礁、自然栖息地以及地形的演变等等。但我的意思是，它能够在视频中实现这一点，告诉它你想追踪的内容，你就能拥有B-roll镜头并能够与之互动。这真是非常酷的研究。

黄仁勋：我举个例子，告诉你们我们使用它的场景。比如，你有一个仓库，里面装满了摄像头。仓库的AI正在监控着所有发生的事情。假设一堆箱子倒塌了，或者有人在地面上洒了水，或者即将发生任何意外，AI识别到这一情况，生成文字信息，发送给相关人员，救援就会在路上。这就是使用它的一个方式。不是记录所有事情，如果有事故发生，不是记录每一纳秒的视频，然后再回去检索那个时刻，它只记录重要的内容，因为它知道自己在看什么。因此，拥有一个视频理解模型，一个视频语言模型对于所有这些有趣的应用来说，确实非常强大。那么，你们接下来还会研究些什么呢？Ray，跟我谈谈...

## 计算平台的未来：

### 别XR了，就AI+潮眼镜就能卖10亿个

扎克伯格：是的，还有所有智能眼镜。我认为，当我们考虑下一代计算平台时，我们会将其分解为混合现实（XR）、头戴设备和智能眼镜。而智能眼镜，我认为人们更容易接受并佩戴它们，因为现在几乎每个戴眼镜的人最终都会升级为智能眼镜。这在全球有超过十亿人。所以这将是一个相当大的市场。

VR/MR头戴设备，我认为有些人觉得它们在游戏或其他用途上很有趣，有些人则还不这么认为。我的观点是，它们都将存在于这个世界。

我认为智能眼镜将类似于手机，是常驻型计算平台的下一个版本。而混合现实头显则更像你的工作站或游戏主机，当你坐下来进行更沉浸式的体验，并需要更多计算资源时。眼镜（大小）只是非常小的形式因素。因为算力将带来很多限制，就像你不能在手机上进行（和计算机）同样级别的计算一样。

黄仁勋：它恰好出现在所有这些生成式AI突破发生的时候。

扎克伯格：是的，所以我们基本上，对于智能眼镜，我们一直从两个不同的方向来解决这个问题。一方面，我们一直在构建我们认为的那种理想全息增强现实（AR）眼镜所需的技术，并且我们正在进行所有定制硅芯片的工作，所有定制显示堆栈的工作，就像为了实现这一目标所需做的所有事情。而且它们是眼镜，对吧？不是头戴式设备。不像VR/MR头显。它们看起来就像普通眼镜。但它们与你现在戴的眼镜相比，还有相当大的差距。我是说你的眼镜非常薄。

但即便我们制造的雷朋眼镜，目前还无法将实现全息AR所需的所有技术完全融入其中。我们正在接近，我认为未来几年我们会越来越接近。它仍然会相当昂贵，但我们还是会把它推出成一个产品。

我们考虑的另一个角度是，先从外观好看的智能眼镜开始。通过与全球顶尖的眼镜制造商EssilorLuxottica合作，他们基本上涵盖了你所熟知的所有大牌，比如雷朋、奥克利、奥利弗·皮普尔斯，以及其他少数几个品牌。这些几乎都属于EssilorLuxottica旗下。

黄仁勋：（就像）NVIDIA眼镜（那么大牌）。（笑）

扎克伯格我想，你知道，他们大概也会喜欢这种比喻。我是说，谁会不喜欢呢？在当下，谁会不想要这样的眼镜呢？

扎克伯格：但我们一直在与他们合作开发雷朋系列，现在已经是第二代了。

我们的目标是，好吧，让我们将外形限制在看起来非常棒的范围内。然后在这个框架内尽可能多地融入技术，尽管我们知道在技术上还无法达到我们理想中的完美整合，但最终，它们将会是外观出色的眼镜。到目前为止，我们配备了摄像头传感器，因此您可以拍照和录像，您实际上可以实时直播到Instagram，您可以在WhatsApp上进行视频通话并实时传输给对方，您知道，您所看到的画面。它配备了麦克风和扬声器，我是说，那个扬声器实际上非常出色。它就像，开放式耳道设计，因此许多人觉得它比耳塞更舒适。您可以听音乐，就像拥有一段私密的体验，这相当不错。人们很喜欢这一点，您可以在上面接听电话。

但我们发现，这些传感器组合正是与AI对话所需的。所以这可以说是个意外的发现。如果五年前你问我，我们会先实现全息AR还是AI？我可能会说，大概是全息AR吧。对吧，我的意思是，这看起来就像是所有虚拟现实和混合现实技术上的进步，以及构建新的显示技术栈。我们正朝着这个方向不断取得进展。然后，大型语言模型（LLMs）取得了突破，结果是我们现在拥有了高质量的人工智能，并且在全息增强现实（AR）出现之前，它的改进速度非常快。

所以这是一种我未曾预料到的转变。幸运的是，我们处于有利位置，因为我们一直在研究这些不同的产品，但我认为最终你会得到一系列不同价格、不同技术水平的潜在眼镜产品。所以，基于我们现在看到的Ray-Ban Meta的情况，价格在300美元左右的无显示屏AI眼镜将成为一个非常热门的产品，最终可能会有数千万甚至数亿人使用。届时，你将与高度互动的AI进行对话。

黄仁勋：它会有你刚才展示的视觉语言理解技术。实时翻译功能。你可以用一种语言与我交谈。我听到的是另一种语言。

扎克伯格：当然，有显示屏也会很棒，但这会增加眼镜的重量，使其价格更高。因此，我认为很多人会想要那种全息显示效果，但也有很多人希望最终能拥有类似超薄眼镜的产品。

黄仁勋：对于工业应用和某些工作场合，我们确实需要（虚拟现实）这样的技术。

扎克伯格：对于消费品也是如此。我在疫情期间思考过这个问题，当时大家都短暂地远程工作了一段时间。就像你整天都在Zoom上，这还算过得去。虽然现在我们有这些工具已经挺棒了，但未来我们离实现虚拟会议并不遥远。到那时我并不实际在场，你看到的只是我的全息影像，但感觉就像我们真的在那里，身体上也在场。我们可以一起工作，共同协作。但我认为这在与AI合作时尤为重要。

黄仁勋：我可以接受一个不需要我时刻佩戴的设备。

扎克伯格：哦，是的，但我认为我们终将达到那个实际应用的阶段。我的意思是，在眼镜中，有细框和粗框，还有各种风格，但我认为我们离拥有全息眼镜的形式还有一段时间，不过我认为在不久的将来，拥有一副时尚且稍显粗框的眼镜并不遥远。

黄仁勋：这些（小型智能）太阳镜现在越来越贴合人脸了。我能看出来。

扎克伯格：是的，你知道吗，这是一个非常有用的风格。无论你信不信，但我正努力成为一名潮男，以便在眼镜上市前影响这种风格。（笑）

扎克伯格：我能看出你在尝试。你的潮男之路进行得怎么样了？（笑）

黄仁勋：还早。但我觉得，如果未来业务的一大块是打造人们佩戴的时尚眼镜，那我可能应该开始多关注这方面了。

扎克伯格：没错。完全同意。我们得让那个每天穿同样衣服的我退休了。（笑）但这就是眼镜的特别之处。我觉得它不同于手表或手机，人们真的不希望都长得一样。

黄仁勋：对，所以我认为，你知道的，这是一个平台，我觉得它会倾向于我们之前讨论的主题，成为一个开放的生态系统，因为人们对于形式多样性和风格的需求将会非常大。并不是每个人都会想要那种你知道的，其他人设计的类似的那种眼镜，我不认为这对这次来说会行得通。

扎克伯格：是的，我认为你说得对。

### 结束语：软件3.0时代已经到来

黄仁勋：马克，我们正处在一个整个计算堆栈正在被重新发明的时代，这真是不可思议。我们对软件的思考方式，你知道安德烈称之为软件1.0和软件2.0，而现在我们基本上处于软件3.0的时代了。从通用计算到生成式神经网络处理方式的转变，我们能够开发的计算能力和应用在过去的想象中是不可思议的。这种生成式AI技术，我记不清还有哪项技术能以如此快的速度影响消费者、企业、行业和科学界，并且能够跨越从气候技术到生物技术再到物理科学等所有不同科学领域。在我们遇到的每一个领域，生成式AI都正处于这一根本性转变的核心。

而且，除了你提到的生成式AI将在社会中产生深远影响之外，我非常兴奋的一件事是，有人之前问我是否会有一个"Jensen AI"，那正是你所说的创造性AI，我们可以构建自己的AI，并加载所有我写过的内容，通过我回答问题的方式进行微调，希望随着时间的积累，它能成为一个真正出色的助手和伙伴，为那些只想提问或交流想法的人服务。这个版本的Jensen AI不会评判你，你不必担心被评判，因此你可以随时与它互动。我认为这些都是非常了不起的事情。我们经常写很多东西，而仅仅给它三四个主题，让它以我的声音写出我想要的基本主题，并以此为起点，这是多么不可思议。现在我们可以做的事情实在太多了，与你合作真的很棒。

我知道建立一家公司并不容易，你将你的公司从桌面转向移动、VR再到AI，所有这些设备，这真的非常非常不寻常。我们自己也多次转型，我知道这有多难。多年来，我们俩都经历了很多挫折，但这就是成为先驱和创新者所需要的。看着你真的很棒，同样，我不确定这是否算转型，如果你一直在做之前的事情，但你也增加了新的内容，这意味着还有更多的章节等待我们。我认为对你来说也是如此，看着你们的旅程很有趣。

我的意思是，你们经历了一个时期，当时每个人都认为一切都会转向这些设备，计算会变得非常便宜，但你们一直坚持不懈，就像实际上你们会需要这些能够并行处理的大型系统一样。

扎克伯格：是的，我们曾经一段时间内制造越来越小的设备，但后来我们让计算机变得时尚了一段时间，然后变得不那么时尚，甚至超级不时尚，但现在又酷起来了。

黄仁勋：我们制造图形芯片GPU，现在你部署GPU。Zuck在他的数据中心里有成千上万的H100。我想你们即将达到600,000块GPU的机组。

扎克伯格：我们是优质客户。这就是如何在大会上获得Jensen问答环节的秘诀。（笑）

当他们说"你知道吗，几周后我们在SIGGRAPH有个活动"，我就想，"是啊，我觉得那天我没什么安排，不在丹佛，听起来很有趣，我那天下午也没什么安排"。

黄仁勋：（所以）你就这么出现了，但关键是，你们构建的这些系统，它们是巨大的系统，非常难以协调，非常难以运行，你知道你比大多数人晚进入GPU领域，但你运营的规模比任何人都大，看着你所做的一切真是令人难以置信，恭喜你所做的一切，你现在真是个潮流引领者。