全中国最懂人工智能的公司之一（商汤，静默期结束，4月10号发布大模型），1个半小时，近3万字实录，解答你所有关于人工智能的困惑。由最具话语权的人解答真实的AI行业是怎么样的，颠覆掉A股很多认知。

关于光模块的幻想、关于国产芯片的能力，训练用的谁的芯片，各大公司有多少算力，国内大模型和ChatGPT的对比，答案都在里面

一定是先要有一个通用的基模型，必须是全修的，各种各样的数据它都见过，有了一个比较全修的这样一个通用的防地去磨好之后，你再用垂直领域的数据再去做一个垂直领域的模型，这样你垂直领域的这个模型的效果它才会足够的好。

算力还是关键中的关键，大模型参数不代表好坏，文心2600亿参数效果不如1750亿的ChatGPT3.5，主要是训练的不够，打磨的不够。训练100次千亿参数模型，可能能解决这些问题。

感慨一句，A股炒算力的标的都是非常远期的，真正能马上、立刻提供算力的还是商汤，训练也是商汤给做的。单

任务并行使用1000张A100以上不是容易的事情，在训练用的还是英伟达的A100芯片。在短期内国产芯片并不能胜任大模型训练任务，做做小模型可以，新一代芯片可能可以胜任推理。

包括很多应用，一句话，没有谁可以高枕无忧。第三次流量分配到来，未来并不是一马平川。

大模型：4月10日发布，画图功能超过Midjourney V4接近V5

Ø 之前没有披露大模型是因为处于香港的静默期（意味着后续的交流会增加），**4月10号，商汤会发布自己的大模型（之前就已经训练好了，因为静默期一直不能披露）**，努力追赶GPT4，对于垂直用户，必须有一个通用型大模型作为基础训练的垂直大模型效果才会好。**垂直应用面临巨大洗牌，但是洗牌的基础是你得有一个底层好的大模型。**商汤希望成为这样的持续迭代的底层大模型。

Ø GPT4 是 8 个月之前训练好的（在微软投资之前），只用了1万张V100，400号人。**GPT4 是目前世界上唯一一个模型，可以去打败90%大学生的比例，而且是通修的大学生**，其他模型连高中都考不上。国内这方面落后了，很

多岗位的就业最基础要求是大学生。但是手里有1万张GPU很快就能考上大学了。

Ø 商汤是最早把人工智能大模型写入招股说明书的公司，2020年就有相关的研究。但是大部分人不知道怎么实现大模型，目前行业知道了，只需要基于大量数据去训练，可以产生涌现，这是一个重大发现。**目前商汤的模型也出现了涌现（涌现不局限在语言，图片等领域也一致）。**

Ø 商汤的大语言模型，**不需要把中文翻译成英文做训练，再翻译回来。是原生的训练。作图能力超过到 midjourney 第四代版本的能力，接近第五代的能力。**

Ø 将发布虚拟人生产数字平台，能够直播卖货、生成视频等等，中标四大行的数字平台。

Ø 参数量不是号称越多越好，需要解决很多问题，很多参数都是凑的。怎么样达到比较好的效果。**训练 100 次，千亿参数量就行。训 100 次你才能够把这里面所有的这些需要解决的问题，工程上的一些点，优化上的一些点，所有**

的一些这个技术的这个边边角角的一些细节都能做好 100 次，中国和美国。

Ø 流量重新分配，大家要知道这件事情可能每 10 年才会发生一次流量重新分配，第一次。第一次就是互联网出来泡沫之后。第二次就是移动互联网头条出来，流量分配今年开始第三次流量分配，你的互联网 APP 如果有很强的 BGM 能力，那用户就会用它用的越来越多，**没有任何人在当下是可以高枕无忧的，所有的公司在当下都不能够高枕，巨头都不能，谷歌都不能告诉你**

大装置：国内主要大模型训练的来源

Ø 商汤科技历史融资60亿美金，30亿投入了“大装置”（人工智能训练平台），训练出来的视觉大模型是全球第一的。公司是真正的AIDC，目前大装置一期有5000个机柜，90%在使用，二期建完共有1万个机柜，**总算力超过 10X false，10000 P 的一个算力。**

Ø 商汤A100充足，在停售前拥有1万张A100芯片。训练一个百亿参数量的视觉模型，对于算力的消耗是等同于训练一个千亿参数量的语言模型。

Ø 为什么商汤对外开放“大装置”，训练模型需求是波动的，只训练自己的模型，成本和风险会非常高，后面还有4、5、6代模型要训练，投入越来越大，绑定更多的合作伙伴，**成为平台本身——“众筹”算力，获得长期长跑的能力**

Ø 临港大装置有2.7万张显卡，已经用了海光和寒武纪，并做了国产适配，**商汤是寒武纪的第一大客户。坦白讲就是这一波的大模型训练，确实是目前只有 A100 和 A800 能够真的跑得起来。目前国产GPU只能做小规模和中规模的训练和推理。**

Ø 商汤训练大模型已经5年了，调动上千张GPU卡，需要非常难的架构，商汤在这方面技术非常领先。**目前可以做到7天不断点，OpenAi两三天就会断点一次，因此商汤准备输出训练能力。**

Ø A股的上市公司，大部分没有GPU，或者买不到A100和A800，算力来源基本上是商汤。拥有5年的单任务并行运算1000张卡的经验（还能对外输出，国内独家），**能够用 4000 张 A100 卡跑出1万张的A100等效算力。**目前有八个大客户在训练，还有n多家客户找过来要训练。

Ø 大装置不仅是一个资金的一个投入的问题，**卖给客户的是时间**，可以让需要训大模型的客户在一个月之内数据搬上来，一个月之内把你的千亿参数的模型去年跑通，然后再过一个月你就可以出结果。

文字记录:

PART 1 大模型常见问题

联合创始人、执行同事徐冰先生来跟大家来做个交流，那下面我把时间交给徐斌先生，大家欢迎。这种现在看得出来大家都非常关注人工智能这个行业。**我们前段时间一直处于静默期，因为这个港作为一家港股创始公司，这方面都有一些要求，所以我们一直以来没有对大家披露商汤的这个大模型**，以及通用人工智能 AI GC 的一些研发进展。那么我们也是从上周开始业绩公告之后，然后才正式开始路演。

给大家去做一些这些方面的这个情况的一个更新，因为大家也可以看到就是说。这个大模型这件事情，实际上很多

公司在历史上可能都是踩空的，因为它确实是一个有极大的一个投入，并且历史的商业回报和商业变现模式都非常不清晰的这样一个状态。

那么我知道我现在全球范围去看，就是说通用人工智能已经变成一个必然的趋势。然后在今年大家也经常听到一些说法，就是今年是通用人工智能的元年，我们这个在上周像 Bille、mask 等一批人开始去呼吁我们暂停一下这个通用人工智能技术的研发，然后担心的这个距离。4 能力已经非常出众了，那如果持续迭代到GPT5、GPT6，那会不会出现一些失控的局面，对吧？就是那这件事情。

我们也有很多的这些这个人来问我们一些问题，那确实我们自己也有一定的这个顾虑，确实是啊，就是因为在去年，去年截止去年是没有人会感觉有通用人工智能具体的实现路径的。那么**通用人工智能的定义就是说人工智能的能力强到它可以跟我们每个人去媲美**，对吧？你怎么去定义就是人类的这个智能？但现在基本上 GPT 4 达到的能力是说他已经可以媲美90% 以上的这种优秀的大学生的能力了，而且他是一个通修全科的大学生。也就是他不仅仅是在去修某一个咱们修的专业，比如金融或者法律，**他是全**

科所有的东西同时都修，然后他可以在各个领域里面去这个打败 90% 以上的大学生，甚至像法律这样的领域。

我们可以想象有什么样的人可以去做律师？一般来说都是那些比较聪明并且比较努力的，经过 10 年甚至更长时间的一个这个培训，上学、上基地实习，你才能够 qualify 来去做一个律师，通过这个律师资格考试。那现在呢？在这样的一些就是评估人类智能能力的这些考试上面，其实 GPT4 的能力超过了 90% 的人，所以再往上继续去迭代。我们也知道就 GPT4 不是最近训练出来的。

GPT4 是 8 个月之前训练好的，也就是在 openai 拿到微软的这今年 100 亿美金的一个投资，对吧？在拿到他的这个新增的这个 3- 4万张的 A100 的卡之前， openAI 用了相对有限的私有不到 20 亿美金左右的一个资金，再加上这个1万张 V100。就是 A100 的上一个版本就是 V100 的这个计算卡，openai用了少量资源，400号人。

然后能够把这批死这样的能力给做出来。我不知道咱们在座的各位有多少人用过GPT4，我身边已经有非常多的人在去用，**天天用它就是他的能力**，就是说跟其他的模型是两个物种。 GPT4 是目前世界上唯一一个模型，可以去打

败90%大学生的比例，而且是通修的大学生。其他的模型的话就通俗的来理解。

其他的模型的话其实这个可能高中都考不上，就是他还是有一些这个实施的这个差别。那么就是说 openAI 用了相对有限的资源，然后在 8 个月前做出来这批次，并且用了这 8 个月时间做好这个lines，对吧？能够让他的能力跟人类的目标去align，尤其是跟这些，这个就是价值观正确的人的目标需要按，而不去跟那些邪恶的人的目标去来是吧？**所以他用了 8 个月时间做好了版本才释放给公众去那工作时间之后，也是这个非常的惊讶，那么就是这也是刚刚发生半个月的事情**，所以这件事情。

其实是触发了大家很多的这种讨论。就是我们进入了这样一个通用人工智能的时代，那这个时代到底意味着什么，对吧？对我们在座的每个人意味着什么？那其实这个可能我们去看整个律师行业，整个律师行业有可能很快速的发生大的变革，也就是现在整个律师行业，对吧？律所的业务没有扩大一倍，但是律师的工作能力扩大了2、到 3 倍。

那这个结果是什么呢？**这个供需不匹配，在需求不变的情况下，你可以产生两三倍更多的这样一个供给，就是法律意见，各种各样的协议。**那这个就自然会引发这样的这个行业的一个变革。可能有大量的流失，就需要去转行了。在这样的一个就是竞争之下，所以有句话说，会用 GPT4 的人会用这些通用人工智能工具的人，会把那些不会用主工具的人给取代掉。这里面行业和行业的竞争是非常。

它不是影响了某一个和两个的。那上周有一个统计报告出来，整个欧洲加美国有3亿个工作岗位就是白领，3亿个工作岗位会被通用人工智能技术去影响，那这个影响实际上是非常大的这样一个影响，对吧？**但对于中国来看的话，在我看来可能稍微的反应慢了半拍。我们整体国内现在的 cash 是属于落后的一个cash，没有任何一家公司截至目前能够拿得出来一个考上大学的通用人工智能的这个模型。**因为考不上大学的这条线是非常重要的一条线。

因为我们很多很多的这个就是生产活动里面对于人的素质的要求，是大学生的要求，尤其是咱们所从事的这个白领行业里，我们对于人的素质要求是大学生是最。

基础的这样一个要求。那如果我们做出来的通用人工智能模型你是达不到大学生的标准的话，那它的这个应用性和实用性都还是相对比较有限的。所以现在就是说据我了解，我们几家手上有超过1万张的 AE 版的GPU很快这样，够考上大学的。

所以我们本身就是说这周开始的实际上是一个业绩路演。但是我在上周在香港的这个四天的时间，我们大概见了不到 30 家机构，几乎没有人问我们业绩了。

没有什么好业绩问题，大家讨论的问题就是都是你们有2万张的 GPU 卡，你们这么早就开始做 foundation model 是吧？因为我们是把 foundation model就是就是人工智能大模型写到招股书里面来，在 2020 年年底上市的时候，就介绍商汤是如何做房地直播的，做大模型，做通用的视觉能力，然后只能够是解决各行各业的，这个就是小模型生产的这样一个大批量生产的问题。但实际上在过往需要就是通用人工智能，基于超大模型能够实现这件事情是并不是一个公式。

大部分人不知道怎么实现通用人工智能。但今年这件事情变成了一个共识，基于通用人工智能，基于这个

foundation model，基于大量的数据去给 foundation model 去训练它，用超大功能的算力去训练它。那么最终你能够实现超大神经网络的能力的涌现，就**涌现**实际上是个关键词emergent。

大家去看这个方面的一些报告也提到了，就是这件事情是没有人能解释的， openAI 自己的技术人员无法解释为什么会有涌现，所以它实际上就是说非常神奇是一个discovery，它是个明星。 discovery 对于一个现象就是超大规模的神经网络，就是这个 artificial neural network，他有能力去，他有这个做到一定规模之后，他就可以把各种能力涌现出来。他不是把design，他不是一个event，他不是一个发明，而是一个发现。就像物理学定律一样，它一直存在在那里，只是以前没有人做到那条规模，所以大家都不知道，就是把这份流量压过来最终只要把它做大。这个 NLP 自然语言领域的这个数据也都是互联网上爬取的。

这些，可能大家也都能够去爬的这样一些公寓的数据为主 5000 亿个单词、 5000 亿个 token 给他训练这个 1750 亿的这样一个网络，他能够把这个就是 5000 亿单词所包含的这个对于这个世界的描述，然后进行压缩学习到了这

个世界的一个表示之后，然后它展现出来了各种各样大家无法解释的这样一个能力的涌现。像翻译，像多轮对话，我要像长文的这样一个生成等等等等。那这样一个现象的发现其实等同于什么呢？等同于我们把这个窗户纸捅破了，就如何实现 AGI 的这个窗户系统，包括所以现在就是说这个就是美国各大厂商、各大互联网公司都在这个超大模型上面去投入了很多的资源，并且中，当然中国这边也是一样，大家都已经这个投入非常多的资源来去做了，并且我们都验证了涌现这个现象。

就是谷歌的模型也有了涌现这个现象。百度文心一言也有了一定涌现的现象，指定涌现能力还没有那么的强，对吧？我们的模型也出现了这个涌现的一个现象，那这些数字我就不一一去过了，就是说这个整体去看就是确实我们这个历史上就是其实在今年 open AI 拿到这 100 亿美金之前，商汤其实是在全球范围之内融资规模最大的一家人工智能创业公司。好，我们历史上是融了 60 亿美金，其中我们投入了30多亿美金才拥有了咱们在这个图下面看到了我们的这个人工智能基础设施，我们叫大装置，它帮助我们不停地去训练、修炼这些大规模的通用模型，我们是全球第一个把这个视觉能力做到超过人的，并且这个训练的这个就是全球最大的视觉模型， 300 多亿的这个参数

量，然后在各种评测标准上都取得了全球第一。比如说这个Imagenet，就大家知道这个openAIDE。

首席科学家之前是做Imagenet，做视觉神经网络Alexnet，对吧？它是那个发明人之一，我们也是在English 保持了全球第一的这样一个准确率的，这个就是成绩。然后我们在新版的 image 上作用Microsoft。

Coco 上面也是全球第一，我们在这个谷歌的 Vimo 就是自动驾驶的这样一个评测基准，谷歌的 we more challenge 我们在去年是拿到了全球第一，所以我们的这个通用视觉能力已经是展现了，我们讲的就是说在超大模型的这样一个迭代之下，极强的这样一个表现。

这个那训练视觉模型，历史上训练视觉模型其实是机器消耗算力的朋友们，训练一个百亿参数量的视觉模型，对于算力的消耗是等同于训练一个千亿参数量的语言模型。所以这也是为什么我们历史上去训这种超大规模的视觉模型的时候，需要去投很多很多的开盘。所以我们就是说非常幸运的在英伟达对整个中国市场去停止售卖它的这个高端GPU A100 系列，之前，我们已经有了超过1万张左右的100 的一个芯片，这也是我们在上海临港的就是咱们的这

个大装置，就是投资 50 亿规模建成的这样一个大装置，在那个时间点在去年开业，对吧？去年年初开业，我们为了去建这个大装置，其实过程之中采购了非常多的这个 GPU 卡，那我们一共有 27000 张的这个 QQ 卡，这里面也包含一些国产的。

QQ 卡，比如说寒武纪的，再比如说海光，就这两家公司我想最近可能很多投资人都很关注。那么如果大家在去年有去参观我们的大装置的话，**你们就应该看得到我们在大装置里面已经适配了寒武器和海光的这种 GPU 卡，我们是寒武纪最大的客户之一**，咱们去这个很早就开始跟他们合作去适配国产的这个 GPU 卡，但坦白讲就是这一波的大模型训练，确实是目前只有 A100 和 A800 能够真的跑得起来。

那么就是国产的GPU 目前仅能去做小模型和中模型的这个训练和推理。那寒武纪有一款最新的GPU，它是能够做得了大模型的推理的，对他训练这件事情，其实这个易用性和性价比没那么好，是吧？但是**推理这件事情，寒武纪的最新款的这个可以去比较好的支撑的。**对我们还有百度做了解，其实都在跟他们继续对一些大模型推理的这样一件事情。

好的大模型的训练确实是非常消耗算力的，但是大家知道，就接下来其实大模型的推理，也就是等他服务终端用户服务的量非常非常大的时候，对吧？比如说一个query，你问他一个问题，他给你反发这个答案，就这一个query。它对于这个算力的这个成本消耗就已经是很难负担了，**就是这个 10 美分接近一块钱人民币的这样，那如果有几千万的用户天天在这里面去，对吧**？大模型的这个应用去问各种问题的话，大家对于这个成本的想象可想而知，所以这里面。

最大的成本是什么呢？就是芯片的承诺，所以这是为什么？英伟达也在最近推出来了，就是应用于大模型推理transformer 推理的这个专门的加速芯片，对吧？ H100系列的这个芯片其实从这个性价比上是有一个蛮显著的一个提高的，那像就是国产的几家 GPU 厂商我们也了解，就是说这个他们也能够用得到，就在他的这个大的 GPU 这件事情上，能够支撑这个大模型的推理。但是如果咱们去考虑性价比的话，依然还是因为咱目前的这个芯片，不管是从训练还是推理上训练都不是最好，这是我们现在可能比较经常会问到的一些问题，提前先抛出来给大家一些比较有用的信息。

PART 2 商汤大模型&大装置

视觉模型领先

就说商汤的话，我们这个一直以来是做视觉的这个大模型做到了最大的这样一个规模，然后大家知道这一波的这个语言的大模型，它的这个技术突破其实本身跟自然语言技术相关性并不大，所以如果大家去访谈过一些专家，或者找一些行业的人聊，所以说可能会有人发现就是所有的做历史上做自然语言的这样一些这个公司，其实它的竞争优势都被抹平了。因为这个大的自然语言模型它其实虽然叫语言模型，但实际上它是更多的基于深度学习和超大规模的这个神经网络相关的这样一些能力来实现的这些群体，那么没有用到什么 NLP 定位这样一些专家知识。我们训练这个 180 亿的这样一个参数量的语言模型，其实在这个业绩攻奥之前就已经训好，那么经过几轮迭代之后，我们这个模型也已经开始对接一些客户，我们做接口的这个测试了，那视觉这一块的话，我们也是历史上一直保持这个全球第一的这样一个位置。像谷歌训练的这个visual transformer 也是 200 多亿的参数量规模，也小于我们的这个 320 亿参数量规模。这个视觉模型我们也训练了一系

列的文森2，模型就是生成出来的图像效果也非常的逼真，一会给大家看一下。

商汤呢？其实训练大模型已经 5 年历史了，就我们是在 2019 年，我们就已经是英伟达的这个中国的大客户之一，然后采购大量的 QQ 卡，我们在 2019 年就实现了 1000 张GPU，是做这个单任务的并行计算上千张 GPU 相连做单任务的这样一个训练计算这件事情，对于系统对于架构的要求其实是非常高的。

对吧？我们在 19 年就创造了一个记录，就是训练 Alex NAP 全球速度最快这样一个世界纪录。当然就是说可能在这一波的大模型突破之前这个确实这样的一些技术成就受到关注这个其实非常的少。但现在如果真的发过去看的话。

我们在 19 年就给我们奠定了一个很好的系统和架构的能力，让我们能够就是在训练这种这个大型的神经网络上，对吧？能够调动上千块的这个GPU，做到一个 90% 以上的加速，做到一个有效并行，并且我们的系统有什么好的一个稳定性，也就是我们可以做到七天以上不断点，就我说七天以上不断点，实际上是我们觉得非常值得骄傲的这

样一个技术架构成绩，对吧？因为即使像 openAI，在他们去训模型的时候，他们的这个断点率，他们的断点率也是很高，两三天可能会断一次点，这在很多的这个他们公布的这个技术指标上是有提及的。那当然我相信现在这个就是很多的这个系统能力也已经提升了。那我们的系统呢？应该是在业界，就是说通俗来讲就是非常好用的一套系统，专门用于做大模型的产出训练。

那我们在 2019 年就训练出来了 10 亿参数量的视觉模型，我刚才也讲过，其实 10 亿参数量训练模视觉模型的训练，因为视觉数据的体积比较大相较于语言数据体积表达以读取视觉数据做训练，它对于算力的消耗，训练一个十亿参数量的视觉模型就等效于训练一个百亿参数量的一个语言模型。所以我们到了 2020 年继续扩大，到了 2021 年我们就可以到了百亿参数量的这个视觉模型，就等效是千亿参数量语言模型的这样一个训练能力了，已经实现了。然后同时我们这个。

就是说在招股的时候就已经 claim 了，我们在全球训出了全球参数上最大的视觉模型 32亿参数，并且展现非常强的通用能力，就是这个它可以识别弯路，就是它不只是聚焦在某一个特定的任务上做识别，而是它可以通用性强的去

识别各种各样的东西，并且我们在 2021 年也启动了语言模型的这样一个训练的任务。团队就在 2012 年逐渐地做自然语言，并且也获得过一些这个竞赛的这个冠军。然后到了接近今年的时间点，我们这个很快去调动我们的这个他的资源，以及我们历史上积累的这个数据，以及一些合作方这样一些资源。

然后我们迅速就做了一个 1800 亿参数量的这样一个这个语言对话的一个模型。并且在过程之中我们也延伸出来了一个多模态的模型，并且先开源了一个多模态模型三室以参数量不大，但是它的效果非常好，他的视觉能力这个 30 以上的书生 2. 5 模型是我刚刚说的， image net 排第一，Microsoft，Coco 排第一。

然后 Google 的 Vimo challenge，自动驾驶的这样一个视觉的challenge，我们也是排第一。那语言模型的话，就是因为它参数量并不大，所以它有，但是并不强。它我们目标是在这个今年接近年底的时候，会推一个性能非常好的这个语言，就多模态模型，迁移参数规模的一个多模态模型。大家知道有 GP4 已经是一个多模态模型，它不仅用这个语言的，它有比较不错的一个视觉能力，加入了这个照片数据作为 token 来去做训练。就为什么照片数据

可以作为 token 呢？你大家可以想一想，一张照片你其实可以用文字来去描述它，比如说我们在这里如果拍一张照片的话，我用 1000 个文， 1000 个字其实就可以把这个照片里面几乎所有的细节描述出来，能够实现一个比较好的还原度，所以一同顶千言。其实文字语言是人类的一个发明，人类的进化过程之中最早的时候是没有文字的，我们是考发明的这样一个，就是这个工具来去描述来去描绘它的东西，我感受的东西。

所以人类或者说整个的这个物，动物其实都是视觉动物百分七十的信息，靠，已经过去了，那视觉信息是最原始的。这个信息当然是个非结构化的一个信息语言，实际上是个结构化的信息。

它实际上是总结了一些人类在进化过程之中感受的一些比较高频次的这样一些接触的物体，和获得的这样一些感受高频率的就会变成一些词，比如说太阳，我们今天见到太阳它就变成了一个词。所以语言实际上是对于我们看到的这个非结构化世界的一个抽象表达，对于语言跟视觉实际上是非常贯通的。就在通，在做这个通用人工智能的时候。

那么我们的目标的话就是说既然这层窗户纸已经捅破了，对吧？实现通用人工智能的路径大家都已经看到，就是通过超大规模的算力多模态的这样一个能力。

最终来去让他涌现越来越多的这样一些这个多任务、多模态的能力。这是就是说我们现在能够看得到的就实现通用人工智能的一套比较清晰的这个方向。

那我们呢？未来会遵循这个方向，然后去执行我们这个通用人工研发的这样一个重要的一个任务。所以这个就是我刚刚说的，我们在历史上非常有前瞻性地去投资了50个亿建成的，就在特斯拉超级工厂旁边，不知道大家有没有这个机会去看看过。就是我们这栋楼其实还是蛮壮观的，它实际上是一个就是大型的这个AIDC中间的这个部分是办公的，这两边全都是机房啊，有 5000 个这个机柜。

然后我们这是第一期已经建完了。大家看到后面这个蓝色的房子是第二期也有 5000 个机柜，那它一共是有1万个机柜。1万个机柜建完之后，他总的算力费用就超过这个10X false，就是10000 P 的一个算力，我们现在的第一期建完之后机器已经上架，并且使用率百分之九十多以上，

是吧？因为大家可以想象现在很多人来去找到我们来去使用这件事情。

使用我们的这个基础设施来去训各种自定义的这个大模型。我们也把我们的基础设施开放出来，来去不只是支持我们自己的这样一个多模态大模型的迭代，我们也支持了很多国内的这些这个龙头的企业，像什么 a 股的一些上市公司，还有一些这个市场上现在非常活跃的创业公司。其实现在你想想他们的算力在哪来，他们是在用谁的GPU？因为现在是 A100，是买不到的对吧？ A800是上个月才刚刚到货，一些是已经4 月份还会到货很多 A800，那么他们的算力作为一个必要的生产要素。

从哪里来，对吧？因为以前只是一个生产要素，但这里面现在有一个生产要素就非常非常稀缺很紧俏的。

A100存量有限，这个就这么一个资源。那从哪里来？其实我们这边是这个，这个就是承接了当下过来的一波的这样一个需求。那这里面就是我们有五年以上的千卡并行的这样一个训练经验，我们最大的这个单任务训练可以调动 4000 张这个卡， 4000 张 A100 卡的等效算力是等于1万张的A100，也就是 openAI 用来训 GPT3 及 GPT4 的这

样一个这个原生的这样一个基础设施，1万张的这个唯一的，所以我们基础设施端的话是足够来去支撑我们以及我们的客户去训练这个通用性非常好的这样一些能力出来。同时我们还有 500P 左右，大概 10% 左右的这个算力是由国产 GPU 卡产生的。那么这里面这个就是目前市场上主流的BPO 们都在我们的这个平台上有适配，其中规模最大的比如说像寒武纪海光，然后再往下一点，像升腾等等，都在我们这边有一个提供这么一个国产算力，是对异构的集群，也就是本身去使用这个国产算力的客户，你其实并不需要知道你底下用的是哪一款GPU，它产生的这个算力其实可以完成不同的公司所做的这个GPU。那就是我们目前的这个算力规模可以支持 20 个千亿参数量的一个超大模型同时计算，同时训练。那么大家知道千亿参数量超大模型用 1000 张 GPU 卡做训练的话，差不多消耗的时间是在这个半个月左右的这么一个时间，那 20 个的话，也就是说同时在这一个月之间就可以训练，对吧？ 40个40次这样的这个迁移参数上的这个创意，这对我们的这个技术迭代，对我们客户的这个技术迭代的一个速度是有非常好的一个帮助的。

那当下就比较有挑战的是什么呢？

比较有挑战就是国产GPU。目前是不太好能够去支持超大模型的这个训练，所以这件事情还需要很多的这个投入去做这个优化。目前就是说这个还是英伟达的这个 GPU 是不二之选，基本上就是这个国产大模型的训练。但从推理上来说的话，刚刚也说了计划国产是展现了一定的能力，所以这也是我们在业绩公告时候给大家介绍的，就是说我们自研的各种各样的这个大模型，是大概消耗了这个这个一半多的这样一个算力规模，那同时我们有接近一半，因为就是已经不到一半接近一半这个规模实际上是持续对外去提供这个算力，并且这两者之间是有弹性的，也就是我需要消耗的任务多的时候可以去占用更多的算力。客户这边需要消耗算力多的时候，也可以去占用更多的算力，所以这是这个人工智能基础设施最大的优势。它其实是弹性。

因为训超大模型这件事情它不是一成不变的，它会有峰值，消耗比较多的时候也会有低谷，对吧？消耗比较少的时候，所以服务不同的客户，服务我们自己和不同会有助于我们实现更经济的一个基础设施的成本。有这样一个成本效应，对吧？并且我们服务这么多客户，也可以帮助我们的基础设施来去跑一个长跑，就为什么跑长跑重要？我们现在这个阶段并不是做完一个考上大学的通用模型就结

束了，我们还要持续地去保证这个模型要迭代提高能力，就对标这个第五，对标这个第六、第七等等，它是个multiyear 的事情。接下来我们在人工智能就是通用人工智能时代，它实际上是一个资源消耗也比较大，门槛也非常高，并且它是个multiyear 的事情。这就是说新版的摩尔定律，每 18 个月这个地球上的这个智能的数量会翻一翻，**就是在这里面的这样一个过程中的话，我们的打法是去以一个开放的打法，然后能让大家来去一起众筹，让我们更好地有一个跑长跑的能力。**

因为如果我只支持我自己的这个大模型训练，在这件事情上会变得非常非常的吃力，并且风险高，那么同时我们有很多外部客户也可以帮助我们持续提高我们基础设施的能力，是吧？这个帮我们有更好的这样一些用户的反馈，然后并且帮我们共摊这个成本。所以当然就是说投资人关注的是你今年这块业务怎么样。

那就是显而易见，我们今年其实今年这个部分的业务实际上是非常好的一个增长，这。几个我们目前的话已经在服务 8 家比较大的投入，再去帮助他们提供算力来去训练他们自己的这个自定义的其他模型，还有很多的这个客户找过来，因为这个我们确实是市场上相对比较稀缺的、比较

好用的。就这样的一个成熟的基础设施，我们自己都已经用了 5 年，打磨了 5 年，这个架构和系统以及上面相应的这样一些模型训练的工具，都是业界相对比较成熟的工具。

那这里面有一些数字了，我们去年的研发投入是 40 个亿人民币，所以其实是在这一个领域我们比较聚焦的在这个投入规模上的时候，体量很大，我们历史的三四年的时间也就上百亿的研发投入，我们在开牌子基础设施建设的投入也是百亿的这样一个，那我们有这两个百亿的一个投入才拥有当下的这样一个能力。这是来自于说我们历史上有 60 亿美金的这样一个，总的这个来自于投资人的这样一个投资，所以这个才让我们当下来这样一个机遇，对，它的门槛实际上是非常高的，并不容易。

从 0 开始重新做一个能够很快的这个出现这样一个事情，不仅是一个资金的一个投入的问题，还有一个很长时间的一个消耗的一个问题。所以我们其实就是通俗点讲，我们给大家卖的就是时间，我可以让咱们的这些这个需要训大模型的客户在一个月之内你的数据搬上来，一个月之内你就可以把你的千亿参数的模型去年跑通，然后再过一个月你就可以出结果，然后你就拥有了自己的这样一个这个模

型。当然很多模型现在并不是从零开始自己去，比如说它是基于开源的一个模型，或者说基于我们给的模型来去做的垂直领域的一些方倾，用了垂直领域的数据。

针对他所在的这个垂直领域，他所积累的这个数据做了一个方向，进一步提高了他在垂直领域的一个效果。那我还是想去有一点这个比较技术的一点想跟大家说明。现在只如果你只有垂直领域的数据，想从 0 去训练一个垂直领域的模型，这种模型生产范式已经是过去式，这是行不通的。现在的模型生产范式，或者说人工智能能力的生产范式，一定是先要有一个通用的基模型，有一个 foundation model，这个 foundation model 是全修的，各种各样的数据它都见过，有了一个比较全修的这样一个通用的防地去磨好之后，你再用垂直领域的数据再去做一个垂直领域的模型，这样你垂直领域的这个模型的效果它才会足够的好，它才会享受到底下这个通用的仿地学猫的这个涌现能力。基于垂直领域的数据做一个垂直领域模型，这种是几乎是有这种模型是有比较弱的有限能力，几乎它的竞争力是基本上没法跟基于通用的基模型再用垂域数据。

迭代出来的模型强，所以这也是为什么说这个很多的这个垂直领域的客户。在 AGI 时代该怎么去做AI，不是说我自己手上的数据直接去拿出他训练模型，而是说你要有，你要先有一个通用型能力强的模型，然后再基于它去做反权，然后蒸馏相关的这样一些垂直应该知识出来，然后再把它这个应用给做好，所以当下我们其实迎来了一波就是说互联网应用也好，各种各样垂直工具也好，它重新洗牌的机会。但当然他重新洗牌的话，不管是什么样的流量公司、 APP 公司，你重新洗牌，你的基础是你需要有一个性能非常非常好的一个大模型。

那商汤的这个责任的话，其实就是把这个性能非常非常大、非常好的一个大模型能够做出来，并且持续迭代它，持续增强的能力，改进它。那就是这是我们在4月 10 号，我们在4月 10 号会发布的这个模型。

所以我们这个其实已经训完，就是我们业绩公告的时候也给大家看了一下，所以大家看了这个 DEC 是我们业绩公告的删掉了这个，嗯，就是说就基础的这个能力基本上都已经具备了，也是这个验证了有限这样一个现象。那么但坦白讲我们的模型现在比GPT 还是有比较大的一个差距，所以我们的目标也是说基于这个接下来的。

这个投入，然后迭代，然后来去尽快地去这个追赶这个第四个这样的能力，让他能考上大学。看起来还没有说是这个真的是实现一个像大学生一样优秀的一个通用性。同时我们也讲我们之前已经做出来了文森2的东西。

这个也是我们基础设施的优秀，我们基础设施的这个优势就是我们其实这个手上的这个算力规模是让我们可以很快地去训任何有价值的模型，当然这里面有很多我们自己的这个创新，所以大家可以看到我们的这个文森2的模型，其实不需要把中文翻译成英文再生成，对，不会有设备码的问题，我们是自己原生去训练的，所以他甚至可以理解中国的古诗。

你可以用中国的古诗作为提示词、作为 prompt 然后来去生成这样的一些很有意见的一个照片。同时其实在就是当下非常火的这个 midjourney 第五代发布之前，我们是做的这个模型，效果超过他第四代的这样一个效果。

这个他第四代是没有结果的效果，他第五代名 journey 第五代，就现在咱们在互联网上看到就是说非常非常逼真写

实的照片，是他第五代才具备的这个能力，然后比我们现在的这个整体的能力还是要强一些。但是我们还是在讲。

比如说第一我们其实有这样一个原生的去迭代和增强文森2能力的这样的基础设施，然后我们确实也在这个持续增强它，然后而且是它可以看得懂中文的这种方法。然后在中国这个市场上我们预计会比较不错的这样一个客户，我们也会在4月 10 号把这个能力进行一个发布。我们还有比如说这个也是一样，就是也是文森2就给这已经是有些客户在做这个对接入他们的这个应用了，就是这个应用我想很多人都能够猜得到是什么，就是电商卖货，你给予一张这个衣服的这个照片，然后把这个材质拍得仔细一点，然后你就可以自动化的生成这些模特出来，这些模特都是不存在的，人都是生成出来的。然后这个模特的生成很简单，你可以用简笔画一下，你想让他摆的姿势，他就可以按照这个姿势，对吧？这是一个可控场景下的生成，这个叫可控场景下生成，把它生成出来的东西是你可以用一些条件来控制的，比如说生成出来的这个模特的资质，这也是我们这个就是虚拟人的生成平台，也是会发布的，就是说数字人就是数字人产品是在今年会是非常火爆的一个产品，就是在我是说这个领域还是我们看到。还是会还是有很多需求。我们前段时间参加了这个就是四大行之一的，

就是一个数字人能力的评测了100 多项技术目标，我们所有的数据源都排第一，就这个最后是一个，就是说蛮明显的这样一个领先优势，最后中标了这个四大行之一的这样一个数字平台。就是数字人现在能用来干嘛呢？就比如说他可以用来做那个直播卖货。

在这输入你是商汤数字人，介绍一下你的公司，它可以自动的把握。我是由商汤科技制作的数字人。作为人工智能软件公司，商汤科技以坚持原创，让 AI 引领人类进步为使命，以人工智能实现物理世界和数字世界的连接促进。

这我就不放完了，就是这是我们的产品经理丽娜，这个她的声音也是靠diffusion model 生成出来。这个声音实际上是非常非常的跟人声是接近的就是你如果我不跟你说这是非真人的话，你是听不出来，它实际上是合成出来，所以声音合成技术也是因为这一波的 AIGC 技术突破，diffusion的突破它发生了一个变革，就是历史上传统的这种语音合成，对吧？比如说我们打这个，就是这种机器人给你打电话的时候，都能听出来这个机器人在给你打电话，对吧？它的这个声音其实跟真人还是有语调的顿挫等这样的一些差别。现在基于 difusion model 的语音合成已经跟真人的声音几乎一模一样你是区分不出来。

所以今年才有了这样一些新的应用，比如说听。

对吧？一本书他定义读出来、朗读出来，不是说有人提前给你录好的，给你放，但是也 10 年课程，所以包括直播，对吧？现在很多直播，实际上他这个主播他是一个不存在的人。好，这个人也是合成出来的，那么就现在可以用。

很低的成本就生成这样一个数字。再比如说你甚至可以把它做成不同的风格，这个这样是做成一个卡通风格，然后来来来去演讲，来去跟人用户去交互，所以现在做直播、做卖货，在做这个，就是说这个客服等等这样的一些场景，在今年都有这样的新的这个 AI GC 能力、一个应用。所以等等，我们就是说在4月 10 号。

还是蛮值得期待的，就是如果大家有兴趣的也可以来我们现场参加，就在我们这个上海领导 a i d c 稍微有点远，在这个就是临港，大家一直来开车一个小时，但我们那个大的超算还是很壮观的，还是非常值得去看一看。这里面讲的是我们这个刚刚也说过了，这个国产化硬件我们做了比

较积极的一些适配。对，我想就是说后面有一些具体的产品。

PART 3 其他业务

刚刚也都跟大家都看过了。对，我们我想 highlight 两个业务，一个就是我们现在增长力比较强的这个手机业务，因为现在商汤就是可能很多人以为我们是还在这个智慧城市领域，但实际上我们智慧城市领域的这个业务已经占比掉，占比到不到 1/ 3 了，我们超过 70% 的业务现在是这个来自于手机行业，来自于移动互联网APP行业，来自于这个就是说汽车行业以及来自于像这个国家电网、南方电网这样的 Tob 场景，很大的国企、央企以及企业级的这样一些场景。那么我们去年在疫情期间增长力都非常好的一个业务，就是我们的这个智慧生活业务，它的这个是实现了 130% 的一个年增长。我们还有一个业务就是我们的智能汽车业务，我们在做这个就是智能车舱和智能驾驶这两款产品的一个这个前装量产，我们交付了，这个就是交付给了大量的这些电动车厂商，对吧？像比亚迪。

广汽等等，那么做前装量产生很难的一件事情，因为它对于这个技术要求很高，你再直接面向消费者。

所以我们在这两个业务上在去年其实都实现了一个比较不错的一个增长。那么今年的话我们是几条业务线，其实都是目前不做一个增长，基本上就是说在不包含我们任何现在看到的。

这个新要发布的这个 AIDC 的能力，我们现有的业务百分之三四十的看到今年的这样不错的，那么当然我们 AIGC 的这个能力的话，会也会是亮点的这样一些业务，那么我们的目标的话还是在今年能够实现。就是说对于这个通用人工智能技术的一个追赶，我要不就开放一些问题，大家问一些问题。

PART 4 Q&A

Q1管理层，你好，那个我想问一下，就是我们目前看那个大模型的参数量已经挺大了，就是但是好像跟 PPT4 或者哪怕3.5 能力上差距比较大，这个主要是什么原因？

首先就是模型参数做大，它有很多种方式。比如说最常见的方式就是模型的拼凑， mix of experts 就是把几个模型拼在一起，它可以理论上来说把模型的规模做得不全大，所以有一些看上去是上万亿参数甚至上十万亿参数的模型，更多是通过这种方式模型的拼凑去实现的。那就是说这个去年其实有很多人做了大模型，但都没有得到很好的效果。因为做模型这件事情，做模型这件事情上不是一步到位，就是一之前没有训过模型，突然想训，直接训了一个几千亿才这样的一个网络性能，不是这么一个道理。但我们历史上都是先把模型向 10 亿规模、百亿规模做到极致，让他能够在他所在的这个参数量规模对他的这个训练数据有一个极致的压缩，有一个最好的效果的一个表达。

然后再基于他基于 skill up 进一步这个扩充，横向扩充模型的参数量，纵向扩充模型的深度，它实际上是逐渐一步scale up的这样一个过程， OKR 也是从具备意义的到9。2- 3- 4，从这个小模型做到极致之后，把所有的工程算法的这个点都已经解决，然后再把它拉出Q2。所以你有没有能力去把模型从小的这个部分先把自己做到他们再的功能是再进一步去做，这实际上是需要时间积累的一个事情，并不能够一步到位，上来就训你的迁移才知道。所以很多去年训出来的迁移模型，其实他这个效果都非常一般，甚

至都没有收敛，那就是当下的话，就是因为去年我们也清楚，就是说这个他本身没办法商业化。模型虽然说是A3效果好，但是这个它没有一个很好的demo，就是ChatGPT，实际上我们是ChatGPT，是 g p 3 模型能力的一个demo，它更多是说我有这样一个 demo 来展示。

我底下这个模型其实蕴含了很多的知识，有的能有限并且我用 Instruct GTP 的方法，然后这个去跟人的意图做了很好的匹配，因为一早的时候你训好TP3，它的模型虽然存了很多东西，但他不太懂怎么表，就 GPT3 这个东西他不太能够表达，就是你问一个问题，他可能给你的答案是不一定是你想问的那个答案，他不懂用户的意图。所以instruct GPTD 这个技术它是解决了。

意图匹配的问题，成本相对少的，数量 10 万对就知道了。人问这个问题期待的一个答案是这个类别，所以有了大量的知识的存储在底下的这个第三这个模型里，同时再加上跟人依图一个很好的再加上这个 response learning。

然后有一个 human feedback，然后在里面人帮你去做到了一个更加，对吧？就符合我们意图的这样一个匹配之后

距离 CHASBT 这样一个demo，那这个 demo 就展示了他原来可以很好地去响应我的需求，然后可以很有意思地跟他去做对话。但是产GPTD，其实他这个在做出来的时候， GPTC 已经迅速， GPTC 在做在跟那个，那就是拆GPT就是说这个它本身底下这个模型并不是一个能考上大学的模型，所以它经常也会犯很多错误，那这个错误率在这一次发布出来之后已经大幅缩减了，所以大家就对这件事情就会很好很期待。

那所以如何去真的做到对标 s 级别的模型？在我们看来差不多就是说如果让我们量化上的话，我觉得是至少需要训一次100 次。

训练 100 次，千亿参数量就行。训 100 次你才能够把这里面所有的这些需要解决的问题，工程上的一些点，优化上的一些点，所有的一些这个技术的这个边边角角的一些细节都能做好 100 次，中国和美国。

相关的其他 100 次，不能够一步到位地。很多公司现在都面临这个问题。

就因为没有第二家公司，目前能够拿出来一个考上大学，效果非常非常好，我们能一起，大家多少都有所差距，美国最头部的公司可能是差钱。

Q2然后还有第二个问题，就是说就是我们有很多算力，没错，然后他们如果想训练自己模型，无论是小模型还是大模型，他们把数据放到我们这来，这目前是通过什么样的技术去实现？就是说他们对这个数据。

那至于您问的第二个问题，就是说这个数据的话，其实就是说这个。

首先有很多是公域数据，大家知道就是在文字语言的这个语料数据这件事情上很多。首先是公域数据，那么当然有一些公司他自己天然的在他的业务场景。里面积累的一些私域流量，基于积累一些私域的数据，带有私域数据的优势。那我刚刚也讲了，那么做通用模型一定要用足够多的高质量语料，就不能只局限在自己的那个私域数据，一定要用到尽量多的方式里面。这意味说你不只是要用中国，你要用全球，大家就最头部的公司目标做的一定是 world mode 事件模型，不是说一个 China model 紧急那在这

件事情上的话，这个头部的几家厂商等等他都有这个自己去获得了这个 word knowledge就是世界信息。

世界知识的这样一些 leap on 很多时候就是最简单就是把网页打开用上面这个文字，然后去当然就是大家很多有历史上的一个积累。比如搜索引擎历史上都，需要去驱虫吗？需要去做一些清洗清理。所以目前在我们的观察。

大部分的这种虚拟大模型的，首先5000 亿头层， 5000 亿的头肯是大家都能达到这个 5000 亿投，可能这个画这样并不难，就是大家知道我们迁移过程是第三，拿到这个画面。

嗯，到第四的话你需要加入视觉相关的全新，那这些用户就是说他们就是有两种用户，**一种就是说我要去我自己有这个计划，我自己有这个空间。我把机器放到我们的这个物理空间里面，相当于就是这是我自己的这样一个这个中心表， AIDC 这是我自己这种模式也有，并且我们有用户是要求我们帮助他们来去建他们自己的**一些，因为刚刚历史上建了 20 多个这样的折算，我们从 14 年成立， 15 年我们进入第一个成本是 200 卡、 200 元卡的小 t 恤，但逐渐到现在我们现在非常大建超算的这个经验。当然我们

是比也有建超算机，华北也有建超算这个机，我们在深圳鹏城二期建的也是个 1000 P 规模的这样一个超算，就是他们也在各个城市都建了 100 平任务，然后再就是像曙光浪潮还有寒武纪这种按卡的，他们也都会这种在一些地方去建这种中小型的超算，大几百 p 的这个功能。可能去年大家有意的小鹏也建了一个超算，在去年这个增加时，阿里合作可能建立一个 700 平，海外的话，这种公司就更多了，像特斯拉建的这个都只有 1800 P。

然后这个 Facebook 改名Meta，他在改名 Meta 之前就已经让词条建一个就是 5000 P 的一个超算，所以 Facebook 的第一版的卡是非常。

飞书，我们了解到也是有几万张的。这个 A100卡是因为他当时做元宇宙需要做的，做这个内容的生产做这个就是渲染，他需要消耗很多的这个一般算是最当时的也是这样子。所以我们应该是说在做这个人工智能计算适合这个大模型训练，包括模型推理这样一个以前相对小众的基础设施，就是在去年都还是相对小众的基础设施有需求去通知上面的这个算力，来训练自己的这个深度学习神经网络的人。在中国也没有，特别，在国外其实更多是大大厂的这样一个，今年变成了一个哪哪都是一个需求，各家1。 5

线、二线都不能，都在非常积极地去考虑捡自己的元气，我们是一个方案的提供商，我们一方面的经验可以让大家减少，就是说建这个 AIDC 过程中采的一个。

简单来说我们可以在 6 个月之内，交付这个 4 千卡的消息 ID 4 千卡也不算消息， 4 千卡已经是 4000 张A 800 的卡， BC 已经是可以去支持。提出来迁移才算提升。所以这部分的需求也看得到。为什么会有这部分需求？一定程度上是因为你刚刚说的大家对于数据的考虑。

我有我自己的这个AIDC，我才能够在我的这个私域流量，私域场景里面所积累的数据，能够用它去 fine tune 我自己的垂域模型去训我自己的这个，就是说这个处于场景的。这个问题，并且我还会在 serve 这个模型的过程中，就是用模型训好之后把它对于用户提供，那用户也会去跟这个模型有很多的交互。

这里面产生的这个用户交互的反馈数据他也希望能保留在自己的这个潮粉上。我们希望晚些，因为这个部分数据未来的价值是很高的。你怎么能让自己的私域模型把自己的这个垂直模型。

因为那个底下的这个 foundation model 不一定是大家迅速去做 foundation model，它做起来成本很贵，它的门槛也很高，就是做foundation model 这件事情。

1.5 线、二线互联网厂商相较于百度和巴黎都是优势不足的，你的算力不如比对人前密度都会更多，你的数据已经录入，所以在做这个 foundation model 这个东西其实很多一点 5000、2000 块那次手上的卡都没有，我们知道有。

很多的这个互联网上手上的这个卡数量是不到 1000 张。答案的不是我刚刚知道的，因为这个卡确实是市场，大家没人没事就去买，一个很贵的一个asset，并且折旧也比较快，平均来说这个三年就要换代，然后差不多 4 年就要折旧那么快的这么一个高价的资产，除非你把它很好的变现的方式。

不然你没有买。比如说阿里云，我们所知的就是说在中国拥有A 100 卡最多的企业。那是为什么呢？是因为他们变现，他们把这个卡均出去，这个卡的租金在前段时间涨，就涨到那个价格，市场是让这个A100 服务器的资产回报率变得很好，一年之内回到现在，很好，现在的价格相对

回落了一些，因为 A800 开始动，现在价格相对回落能理性一点，但其实还是那么就是。所以出现了这样一个响应。刚刚的问题就是这个。

自建 a i d c，这样我们自己，对吧？就有征，就有迭代和征留我自己的垂于模型的能力，并且我们持续针对我们用户收集到的这个，就用我们的这个模型部署推理的这些过程中产生的这个用户的闭环数据在持续覆盖进展。，同时公司现在在这个流量重新分配的过程之中必然的一个选择，流量重新分配，大家要知道这件事情可能每 10 年才会发生一次流量重新分配，第一次就是互联网出来泡沫之后，流量就被大头基本上拿走。第二次就是移动互联网头条出来，美团出来分走了巨头的很多的这个流量。

流量分配今年开始第三次流量分配，你的互联网 APP 如果有很强的 BGM 能力，那用户就会用它用的越来越多，比如说微软是吧？ Microsoft Copilot，那他对于咱们日常的办公的这个工作效率提高是会非常显著的，你不用他，你就会被用他的打工人卷走，所以最后你会发现几乎所有人都会用。

这个怕扣他的能力强的这样一些应用。所以对于像这个，比如说国产的office，那你就必须要确保你的 quota 的能力跟微软的一样好，或者说跟飞书的一样。

或者说跟这种就是钉钉，对吧？这种云办公的软件里面出发的能力一样好。所以对于所有的。这些做工具、做APP做流量的厂商来说，这波你要么就是说这个就这一波是个流量纯粹的环节，就是我们感受到，比如说大家基本上就两种情绪，一种情绪就是担心自己会被作为机构？

比较好，因为小公司确实算力输出各方面就是相当于巨头，无法担心自己被巨头在这一波里面清吞掉自己的流量用户。

其实都没有，对吧？就是会走下滑路径，两三年之后可能一个没有很好的模型，没有基于一个很好的模型做出来的产品就存在了，这是第一个情绪，有一种这个就是对自己未来的一个担忧。第二种情绪是一种说不定我有能力去，并不可以判定就相当于重新洗牌，我也能做一个基于 AGI 的一个投放与应用，来去获得更多的用户，去获得更多的流量，我也有机会成为下一个头条，我也有能力成为下一个拼多多，对吧？这是第二个前提要抓住自己的这个机会

然后做一个这样的，就是大家都会尝试做自己的自动性，大家都会去尝试做一下，因为这个机会实在是太大。但是很多的玩家在我们看来，他们在尝试去训自己的通用模型的同时，他更多是一个双门驱动的思路，就是他一定要把应用，即使最后自己做到模型不一定是最好，我就能够 interview 某一个其他人做得比较好的一个东西，比如百度。假如说未来微软能够把 openAID 给模型带到中国来的话，微软是中国还是一个friendly？就是它是，就相当于很多这种安全撤出，微软采取的策略是不一样的，在 word 未来某些年微软把都卖这么好的东西也带到，就是以要考虑自己做只能是有比较低的一个胜率去搏一个很大的价。更稳健的一个做法是我去用最好的那一家的模型，然后去做一个颠覆性的应用，然后我去在流量重新分配的这个环节里面来去获得这个更多的用户，更多的流量，把我的应用真的是能够一下做到这个行业最好能让他有什么？所以现在这个使用点。

其实我也会被问到有什么样的股票反正最安全，而且我是肯定没办法这样去评价。那样，那就是说现在这个时间点，我刚刚讲的这些内容也是让大家有个感觉，就是没有任何人在当下是可以高枕无忧的，所有的公司在当下都不能够高枕，巨头都不能，谷歌都不能告诉你，更别提，对

吧？咱们看到的是所有的其他公司没有人在当下在流量重新分配，在技术发生这么大的这个变革的时候，你可以说我是告诉你我是很安全的，你来去，对吧？买我这个，但是就是说反正至少两家。

公司在我们判断的时候，未来真的是阳光复杂，是吧？因为你看接下来对吧？他已经有了现在能考上大学并且打败大部分大学生的这样一个通用模型在上面。

未来就这一波出现的达人的应用都是基于手机的方案，所以 AWS 其实是受伤，比 Google 受得上还要大。这是一个做微软，我觉得就是还是拿到了5月20%，另外一个就是英伟达，虽然说微软现在也开始跟 AMD 合作，但是就是英伟达的这个地位现在还是不可撼动的，他做大模型推理和大模型训练的成本现在是业界最低，而且他还在提价。什么意思？也就是说他短期内看不到有人工单经理工具给他提价，但一旦有人跟他竞争了，比如说 AMD 或者谁给他竞争，他降价。

它持续可以保持说算力供应是以最低的这个单位城市，就算力是个commodity，大家一定要知道算力是一个commodity，他就是花多少钱买的问题，不管是在训练还

是在推他就是花多少钱买的问题，那谁家的便宜谁就能够考虑他。所以从现在这个阶段理解，非常显著。未来你可以想象BGI 相关的这个算的需求会大于移动端的，也会大于，而不是说跟搜索性。嗯，模型的推理，预算的需求，未来一定是想法，模型就是这个 DGI 模型本身会变成，就是每个人能力的延伸，变成我的电脑，变成你的微信，变成你的应用。这件事情我们将很快推 5 年之内，非常，流量的这个分配都非常，咱们很多现场天天在运动当中，非常非常好的机构，在这种机构里面一定会有这个，我记得是，几个问题，关注你好，想问一个，就是那个就是技术细节，就是像这种现在不是在搞这种规模，这种就是多模态。什么就是特别来回归，就是你现在做这种称语文生图和这个纯，这个刚才纯语言模型，未来是不是这种通用模型可以取代所有的一些这个人生？不管理我行，换句话说就能换，大家会不会就是会取代名journey，就是未来就是安定一起去发展。

就这个问题比较难回答，是因为就是说你去看就很多应用，我相信这种做大模型的公司不一定亲自下手去干，这里面还是要有一个资源，有限资源下的一个取消这个我想这个在我们看来就像因为diffusion， model stable diffusion 其实是非常 elegant 的，这么一个就几百行代

码实现非常好的原创就能做这个你 journey 的人其实接下来有很多，你journey不是一个很大的。对，所以我觉得就是说它不是一个兵价之争，必必争之必。

当然就是文森2，视频后面延展的这种能力和应用也会越来越多，但是因为这件事情本身它的壁垒并不是高，无法想象、无法企及。你会看到说，比如说那个德国比自己就是很多这个天然这种用户群，比如说美图有可能自己能出来，就是他的门槛没有说高了，没啥企业，我们的那个 journey 的一个产品其实也是很快就做出来了，然后甚至超过了第四代，只是说又被他第五代就可能又超过去。我觉得这个阶段就是说这更多是外化情况，然后大家都在这个时候找到自己的这样一些，它有点像那个一些APP，它都会有很火爆的阶段，但很火爆的阶段完了之后就看能不能真的传播下来。我觉得就是他很看后面的这个运营、客户的这样一个持续是不能留下，留在这个社区，就是历史上很多这样的一个平台，就是昙花一现，抱好一段时间。

也是因为就是这里面就有些APP的壁垒和门槛没有大家想象的那么高，可复制性比较强，是吧？或者说它更多的是一个大家尝鲜的这样的一个冲动是一个长期持续的这样一个需求。如果他这个工具需求的话，其实很多做工具的公

司和高比就是时间问题，能够排查到这样的一个数据，会在他们做的这些更强的一个能力的一个体现。我觉得现在很多问题确实是有很多的这个开源模式都隐藏出去。

Q3 比如垂直领域的这种训练是基于这种大模型未来这种形式是什么？

因为这个大模型它也不如开源，就是是视觉大模型提供方，它提供一种这种一种工具，然后就是这种垂直领域，它可以用工具去在这个模型上去训练实时，类似这种模式我们认为就是大模型这块发展到终局的话，很可能是有一家非常领先的，然后很大比例同时会有几家跟随者，而且这几家跟随者大。就有点会变成就挺火安卓的这样一个开始。那其实 openAI 一开始也是开源。承担他发展到一定阶段发现自己有建立起来商业壁垒的这样一些优势产品优势之后，选择去做到现在，这就是领导者，领导者经常会以必然的策略来去，更多，尽量更多的这个市场，然后但是 default Meta 专场把他的老马确认出来之后，就让我们的这个全是 BT 相似的产品，都是拉横效果，是不是还是挺不错？很多的这个是在基于拉曼在做这个，再去继续给妈妈给数据，新的这样一些数据分布，然后影片版本就是一个中文没有的这么一个东西，你可以为很多中文书。

具备，所以我觉得 follower 很多会，你会看到有人去持续开源的，我们也是一个比较开放的把，我们模型也会自己做开源。可能在 TikTok 上面你马上可以。

模型这件事情，就是说你要基于这种性能最好的这个模型去做垂运，它确实它是一个 ToB 的service，拥有这个最好良心的人可以跟你，配合，在你这个企业客户的数据上去这么一个去问题的一个模型，这件事情是一个 Tob 的service，对吧？

但也有一些企业其实他就是拿开源去调，现在这种非常不错。我们现在看。

Q4 就是现在很多国内的大厂都在做这个通用模型，就是大模型，我不知道我们这个双方在这个数据上有一些劣势。第二个就是最终可能模型的能力会差距不大，就国内的新人猜测不会那么大差距，是不是最终决定能不能就是有好的创意模式，还是有没有场景能力？我们好像感觉就是商汤在这块是不是没有找到一个比较大的，或者是比较好的？

我们的做法常规吸引我们是通过企业的人、合作方来去触达更多的，我们在手机、在汽车其实都是触达，消费者就是消费者，你买的车其实很多搭载了我们的技术，去年是超过 50 万台量车，今年是。定点好的流量产品，那么手机的话我们是大量。今年还有就是案例已经回答出滨州苹果也成为了，所以我们更多就是 b ToC 这样一个做法来去用，让我们的这个应用很快地触达很多客户。这个跟欧派是比较像，open自己他 c 端的这个接口更多的是严格版的demo，是他们自己的能力，他没有去很强地去扩充他这个 c 端接口，就是暂时来看，尤其扩充到 c 端接口很多的这样一些体验和努力，paper这个阶段通过通过浏览器、通过 office 的这个询价来触达更多的这个 c 端。这是因为像我们这样的公司其实资源是有限的，我们需要比较聚焦把这个模型本身做好。

那我们目前确实是，就是说这 5 家就是手上有充裕算力的大厂里面，我们是规模最小的一层，超过1万到100。画这条线的话，当然还有一家就是换行有一半这样的决定，四十之一的这个线程，他们也在做这个事情，做他集群真的是做得很好。我是，因为这件事情非常多，所有的资源，就是说当然不考虑最佳的话，就是百度目标里，算力，那我们是规模最小，理解最小公司的体量。算不算你资源比

较？包括你刚刚提的数据？不过我们也有，大家就是说唯一一个没有流量业务的公司，我们在这个大家重新分配流量的这个过程中，我不是任何人进的，然后我可以很好地支持我的客户去让他们去选流量，让他们去抢别人的。嗯，因为这是流量重新分配的一个问题。在这个过程中，如果你在基于你的竞争对手的模型和基础设施去构建一个新的应用，去强力竞争对手的流量，但这个事情稳定性。必须要达到所以这也是为什么我们这边确实是有很多人主动来找到我们。

我们现在说实话，我们基础设施不是无穷大，我们的资源也是有限的，现在我们也是有相对有限的一个卡，嗯，达人的需求，因为我们现在也是就是说这个再去扩充我们的这个算力。当然我们现在这次扩我们的算力规模其实是不错的，这个客户来去去为他这个共盘成本来去为他付费。我们也有算力和捆绑在我们这个算力就摆在为我们的领导的AIDC我们的大装置，它实际上是我们一个全资子公司，这个全资子公司就是说商汤也是它的客户，对吧？就相当于我商汤也在用我们这个领导 AIDC 的这个算力来去训练这些模型，这个迅猛的一个算法。这个全资子公司其实它的业务体量去年已经非常可观了，如果算上这个商汤给它也是公司之间的这样一个，公予价值的这样一个采购，**所**

以他算力的采购，算力服务采购，这个公司的这个收入规模已经大约 1 亿美金了，到这个去年今年他的这个业务规模实际上是非常非常不错。

非常大，给大家可以想象就是这样一个资产，其实是很有价值的一个资产，而且是极其稀缺的。你现在在想建这样的资产，就是时间也来不及，投入也更加昂贵，考个涨价，然后而且给 100 门票，因为 A800 它是无比极限的一年， A800 是连不到太大，基本上也是 2000 张，毫无比极限。因为他那个 m link 卡卡之间连接的那个 m link 1/3，把之间的这个数据传输能力是降低了这个层面值，你看其实它就，其实就是限制你把它建成过分大的这个超算、紧缺的这样一个数据。这件。

事情说实话，首先公域数据非常充足公域的数据大家都还能用。

就说你用完公域数据，高审你要求公域的，你如果都用完了，那你需要去扩充私域的更高质量的数据。就公寓数据，其实也就是说 somehow 打包含，比如说这样子就是 public 名，你就可以直接去看得到的整体数据，其实这些

都属于这个公域属性，就是拿得到都可以接触得到的这样一些数据。

那真正属于私域流量的数据，确实这个有越多也是越好。那这些的话我们也可以看到也会有比较灵活的一些合作方式，因为拥有私域数据的很多玩家，他本身也需要这个模型，因为你想你什么样的玩家才能拥有私域数据？肯定是个互联网类互联网和移动互联网的这个厂商，以前用的比较多的这种私域数据，有私域的流量，私域的用户，那这种其实都在这一波里面需要升级的产品也就是假若这个在座的各位是某一家这个。

拥有一定体量私域数据的这个公司老板，那你想你在这波你如何发挥好你手上的这个数据的优势，能够让你拥有一个比较好的模型，一般来说都是，置换一个好的模型的这样一个私有化部署授权。所以其实就是说我们作为这个中立的这样一个没有流量业务的平台，又有很大的算力，在这种时候确实大家就有这个空间，然后来去看类似的这样一些合作。

我们跟这些这个拥有数据的玩家其实更多就是一个合作关系。说一个问题，就是你刚才说的可能我们也会开源。现

在开源的模型也越来越多有的人做的成本也很低，就是不是这种模式会导致我们的在过去看算法上的一些积累的这个优势就化为就没有了。因为大家都开源了，说不定以后开源的东西足能够足够的好，他不一定是需要最好。

首先开源模型它可以让你在当下达到一定的水平，但是开源模型不能保证你未来基于差距持续的追赶和提高，这是两个概念。就是说你拿的是开源模型当下的能力，你享受的是开源模型当下的能力，但是你要知道的是这个模型能力它是要非常快速的迭代的，也就是这个视角价模块定义你如果基于去年的模型做一个应用。

你可能今年你这个应用就没有竞争力了，所以你其实是需要持续地保持应用在每年都接入性能最好的模型。模型的切换成本其实在很多领域模型的切换成本在现在这一波里是不高的。咱们那在这种情况下，假如每天你的模型基于开源的是一个性能落后的版本，能转天某一个基于拆成 1 第四，基于这第四或者是第五的模型做的这个应用，就把你的这个应用给一档那我们能够给到大家的一个点，就是说我们是在持续地去用我们的算力来去迭代和更新这个模型，他们的能力序的保持在一个领域做到业界算数，那这

个能力是非常稀缺的，而且还是需要有一个很专业的团队强的一个团队很大的一个算力平台来去持续迭代。

大家都可以回归到自己的这个就是原始的这些状态，接下来会有具体组织，也有。第7，你要预计就是说接下来 BGM 的能力变得会越来越强，超过大，然后 impact 到咱们方方面面。 at a time，相违背的。对 Adi 这个时代有风险，但是前后也很难对Adi，它是个 multiyear 的事情，在这个 multiyear 的过程中，那么 b 原则的优势。

什么样的？ 1 元对他来说就是领导者，避免者的优势是他必获得无穷的自己资源。保算力，对吧？获取一些必要的生产要素，能够持续地保持领先。那么像开源者，其实更多是为了去统筹大家的资源，用众筹的方式来去保持跟进，能够去一个跟主，所以这件事情是一个multiyear 的 project，我们需要跑长，我们需要考虑跑长。跑大家讲的现在这个已经是非医疗员安全，一次，然后大家都可以开心的开心。不是的，事情是今年刚刚开始，未来变化多端。

那我们需要解决的挑战就包括你刚刚说的那个会不会有人别人做的模型比我们好，那我们当然也没有优势，那也是

一样的，就是我们需要保证我们做出来好的东西，怎么样才能够。

最大概率增加我们的成功率去做出来，就每年都做出来。比如说在中国市场最好的模型，那我们就需要比较大的算力，我们也需要比较多的这些客户，对吧？然后我们也需要包括数据，我们也需要这些，这个就是比较优秀的人才，包括资金的支持，这个投凝聚投资人。同时我们为什么4月 10 号还发我们这个模型？是为了能够明确大家的共识，让大家知道这个能力，我们有能力去作为中国 BPI 的领导者之一，然后去跑，我们4月 10 号发一些模型，让大家看看，认可我们能力。

那么接下来我们下半年再去发能力更强的模型，凝聚这个公式。发这个模型的过程我们会有包括这里面，就是各方面的这个支持，都应有增加我们在这件事情成功的感觉，ａｇｉｇ寻找成功，现在就像，现在的价值2万亿美金，年初的时候年未来的价值是2万亿美金，年初的时候 300 亿美金，当时很多人还不信，就觉得。预测一笔记那时候年数有 300 亿 GPT4 发出来之后，大家没有人你想这样一个能够考虑大学生的可以就他是他的这个价值是等同于 GDP 的一个factor，等于g、 DB 生产总值的一个factor。因为

是它实际上是去稀释人作为一个生产要素的价值基础。大学生能力之后你其实在稀释的是人口红利作为生产要素的这么一个价值。所以中国为什么要去发力？点击中国历史上比较明显的就是国外已经靠 BGI 去发展经济，国内还要靠鼓励大家生小孩的发展经济，那你这个后面你是不是没有没有出口？所以我们现在是阿里去买账，就是说能够同样这个就是说大，这件事情现在没有认知，有很高很高的成功率，能够有资格上桌的，能够资格上桌的首先不像是移动不了的，大家只要去 mobile first，对吧？组建一个团队，做一个 mobile first，一个APP，组建一批工程师，你就可以去创作，然后去买量房源，收割、获客拉新，然后用户卷起。这一波的入场门槛是非常非常高，所以我就说就是现在我自己觉得就是创业公司，当然这一波创业难度也很大，但出来的创业者能看到这一波出来的创意。

资源丰富、经验丰富的这个创意，所以这一次的创业的门槛要比之前移动互联网多人都高了很多，那对我们来讲我一直没有说这个，不能说一个非常非常高的成功，但是我们确实有我们自己的优势，我们自己的一个成功率。然后我们成那个奖励是巨大的一个奖励，巨大的影响，所以我们这样一个就是价值，我想现在的价值去一起

芯片这个行业，后面的话是想要有一个打算，比如说是推荐的时候，你现在可能构成国外的资源A800，对吧？那国内资源我们就能够给国产一个一定的比例。如果感觉A800 如果说你不要放，好像可能也不断，然后国家政策也没有出口，知道系统保证并行，或者我们不是自己的成交这种产品的东西。对，首先我想如果有这种创作的话，我们应该也不能随便听到非常提醒。对，但不过就是说首先我们现在确实就是大家都还是一股脑的这样一个大的，你把百急出钱。

允许把现在有很多人担心的是说 A 七八百很快出来，效果又很好，那你现在买了很多A800，对吧？其实接下来几个月就能把握好度，你不能买太多A800，买合同 A800 的话 1000 万一出来的话就是1800，肯定有些价值就会变。所以我觉得这个阶段就是说确实大模型这一波对于国产的 GPU 的影响很大。LEADER，指标上来看的话，就从这个新的设计指标上。

历任的他们的这种芯片大模型训练的基础，我要需要做很多case，做很多的一个算子的一个优化，算子优化做好了

模型训练的一个工具的这个效果层级，并且还需要解决一个问题并行的问题。

现场也是需要去投入资源一起去迭代的，所以我们这两家，其实我们都是他们，从未来去看的话，确实是如果是解决了一个可用性，你不能说用不能用起来跑都跑不动，然后这个就其实没有解决一个可用性的问题，如果大家这些头部的这个大芯片就必须是大芯片，能帮助这个大模型、大芯片的这些设计者如果是能够解决就可以算子优化得好、传输得加速得好，推理的这些算子优化也好，并行的这个体系也建得比较完善，并行的也做也很不错。那如果是解决了一个这样一个问题，我相信他们。

替代确实是一个，在很多的这个项目里面是有这个要求，但现在最迫切来赚钱没问题。

Q5 我们国内的大模型跟国外的差距，可能是差距在千亿大模型的训练场，那我想问一下，就是国外的项目和这种，他应该是有这种能力去做到这件事，之前已经做过的，那他跟这个 GPTD 的这个差别主要现在。然后第二个问题是就是传统公司以前应该是专注于在 CV 领域的，然后技术底座可能是基于 CNN 或者 RNN 做的一个模型，

那现在这种大模型都是基于那双本的这种价格，那在这种基于在风格架构下的这种模型越来越强的进来的风格平台有取量的这种架构的，而且我们去研发，现在应该不会去研发这种双面架构的，那 RN 能达到什么样的一个效果？

首先现在没有人做RN，当然其实很早就大家希望都是谷歌这边的话，是它实际上是有一个很好的一些没有发布，我知道模型并不是已经把自己做到底，流量给大家看，投了一些资源。 Bert 和 GPT 的优势不一样，一个是预测下一个词，确实是，当然有经验就在这，因为你预测下一个词。

你一定要对前面所有的这个内容有一个非常深刻的理解，以前能够很精确地预测下一个词，所以下一个词预测的越准确，意味说你对前面的内容理解的越深刻。因为实际上是说就没有人解释得清楚我们现在是怎么来的涌现，其实就是来自于对于所有的训练语料的一个深刻理解那这个深刻理解说白了你。

理解得越深刻，你对于那个目标函数就给你们，就给了你很多奖励，然后你理解深刻的这些能力留下来。所以你可以从遗传学的角度去想想这个。

有限这个现象，人的这些能力是怎么留下的？其实就是我们的目标函数，对吧？基因反应，基因延续。就是我们所有的生命流动到那基于这个目标来说，我们所有的行为。

我们所有的思考，我们所有针对一些问题的反应关系，相互的帮助，其实都在优化这个目标函数。你帮助这个目标函数的能力所遗传下来，你没有帮到这个遗目标函数的一个能力，那就是很快的现在这个大模型它的这个引线就是很类似于这样一个点就是大量的参数，它都在迭代的时候为什么有些提高的就全面降低了，有些能力逐渐就留下来的，有些能力被淘汰了。它实际上也是基于说我们最后这个目标函数优化，这个目标函数就是预测下一个词，我能够对下一个词预测非常非常准。那这种时候其实他所有的这些内容，他的这个理解非常的深刻，所以他就会变得博学多彩，对吧？他其实对于大量的语料数据进行了一个极其有效的这样一个这个 knowledge 的一个抽取不要说我觉得这些就是说这个，蛮神奇的这样一些地方有谷歌一个。