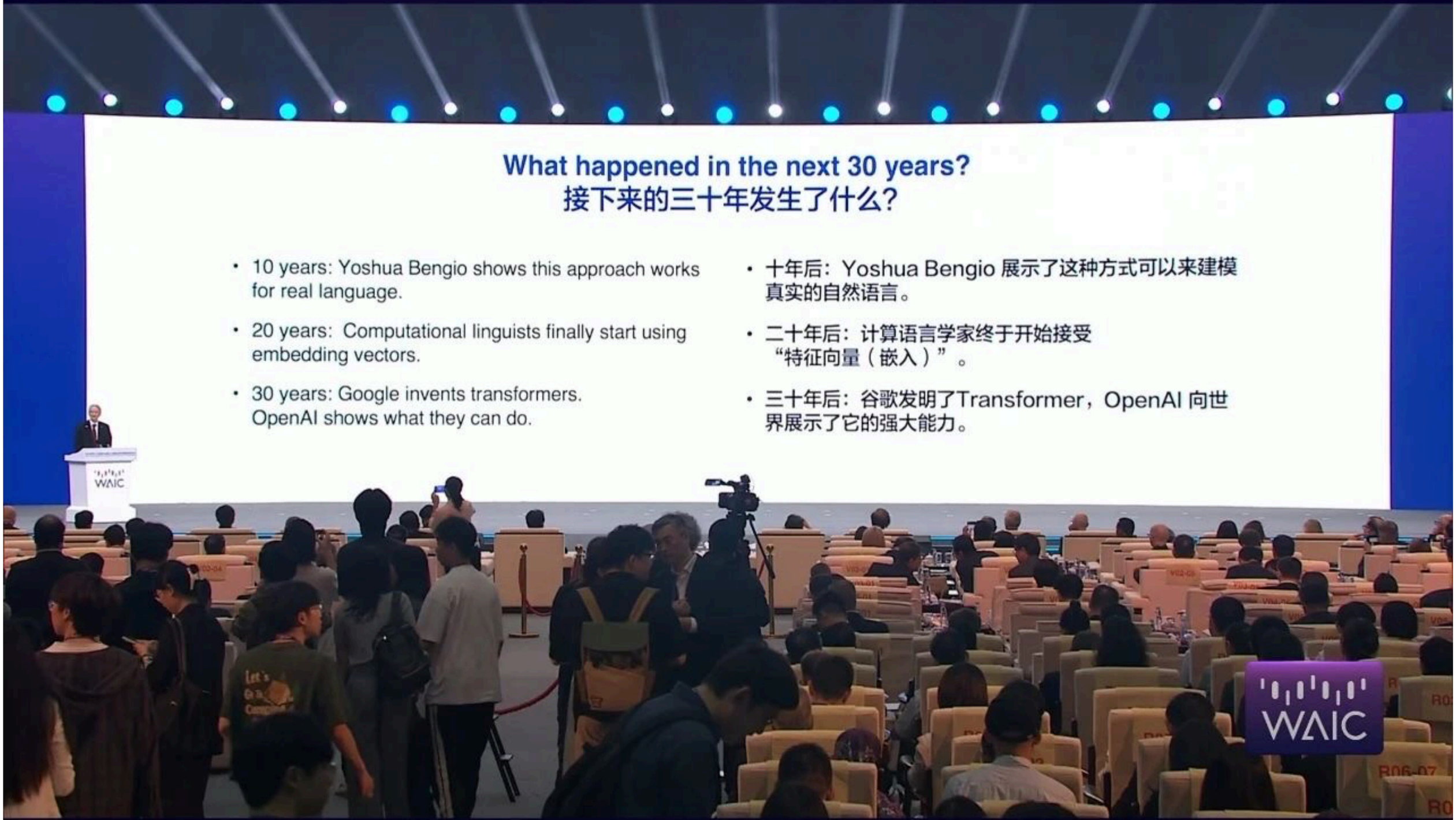
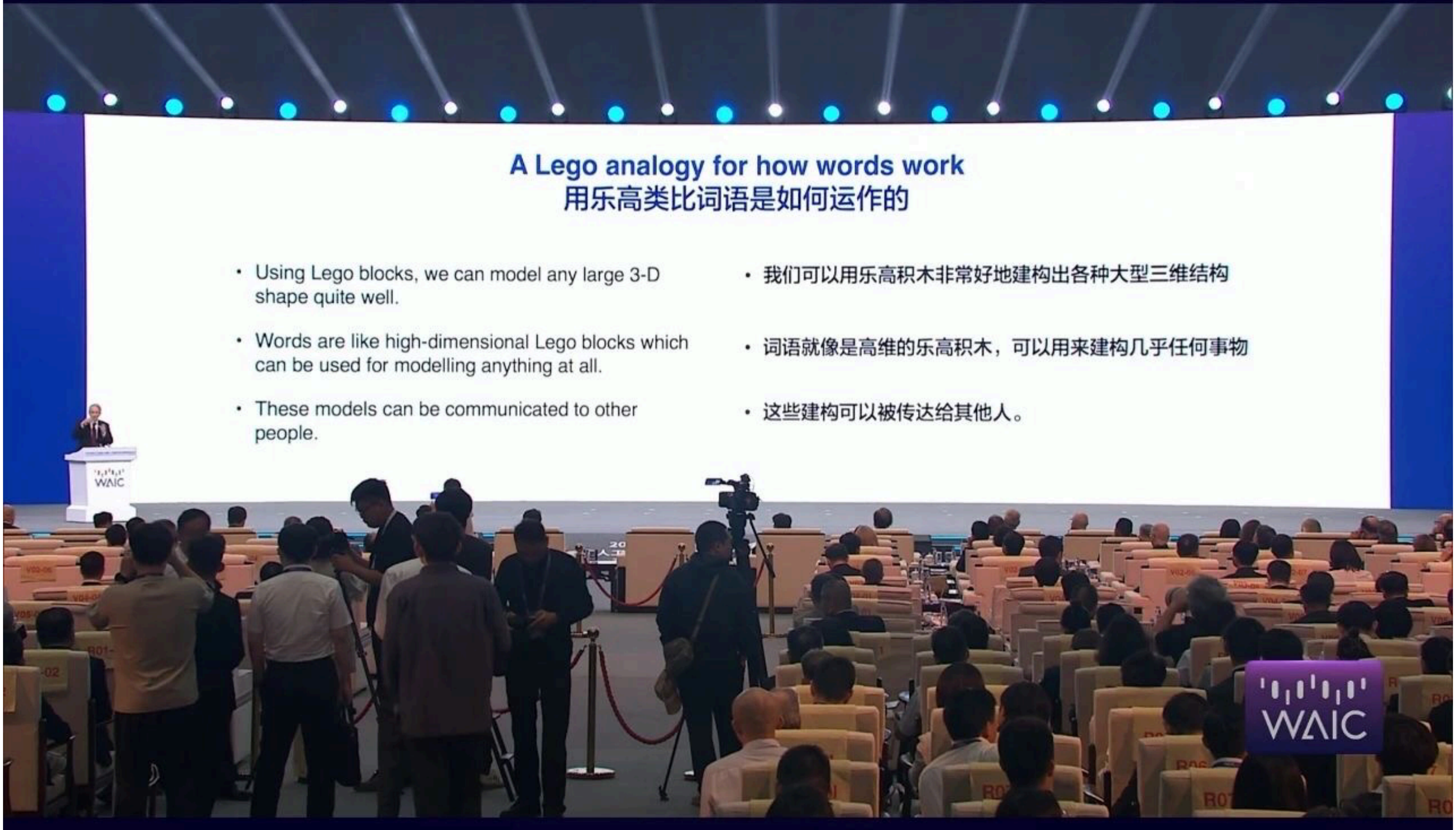
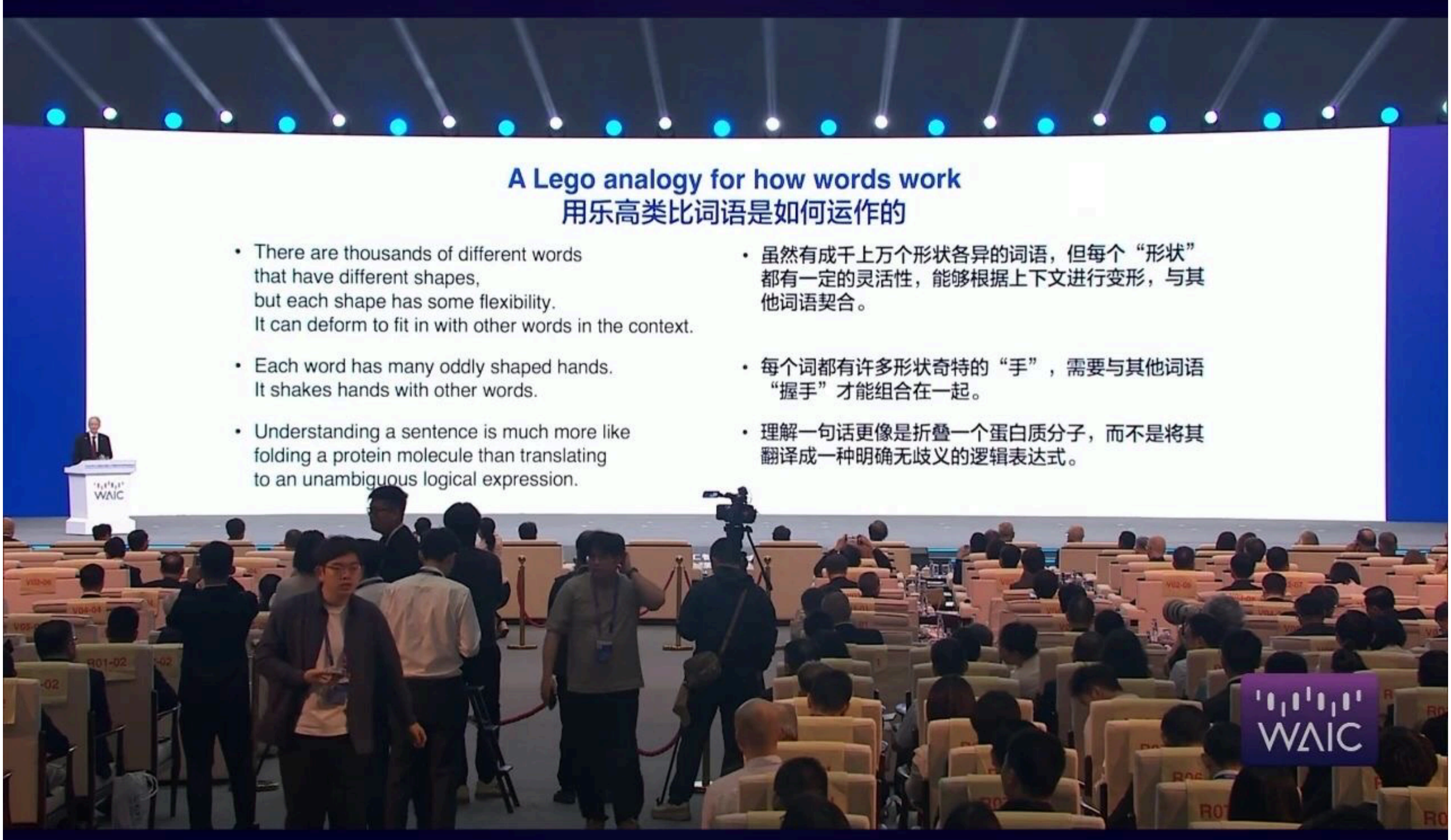

Will Digital Intelligence
Replace Biological Intelligence
数字智能是否会取代生物智能?
Geoffrey Hinton
杰弗里·辛顿
University of Toronto
多伦多大学
WAIC

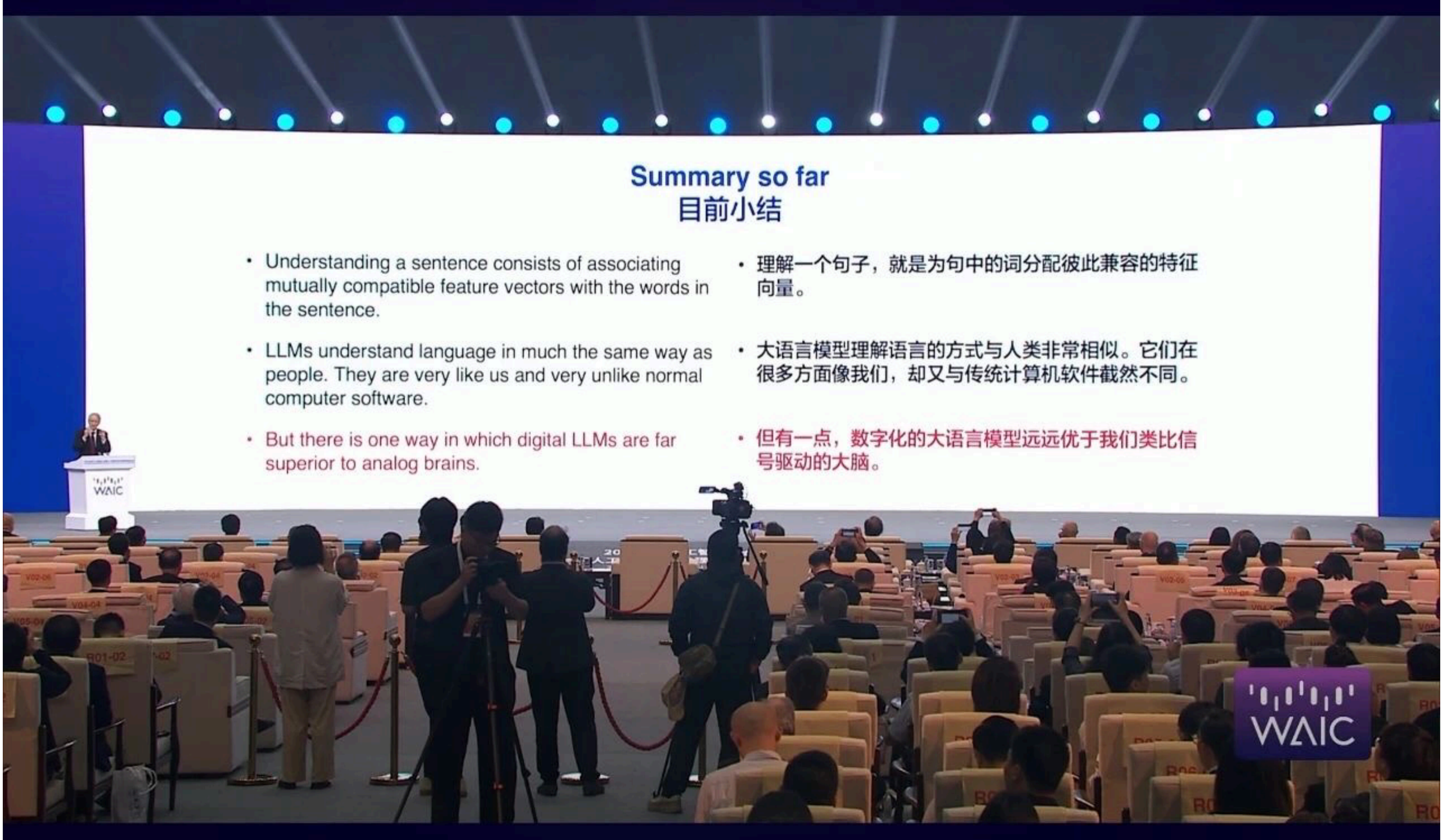

Two paradigms for intelligence
两种智能范式
• The logic-inspired approach:
The essence of intelligence is reasoning.
• This is done by using symbolic rules to manipulate symbolic expressions.
• Learning can wait.
Understanding how knowledge is represented in symbolic expressions must come first.
• 逻辑启发范式:
智能的本质在于推理。
• 通过符号规则对符号表达式进行操作以实现推理。
• 学习可以暂缓，首先要理解知识如何以符号表达的形式进行表示。
WAIC
WAIC


What happened in the next 30 years?
接下来的三十年发生了什么?
• 10 years: Yoshua Bengio shows this approach works for real language.
• 20 years: Computational linguists finally start using embedding vectors.
• 30 years: Google invents transformers. OpenAI shows what they can do.
• 十年后：Yoshua Bengio 展示了这种方式可以来建模真实的自然语言。
• 二十年后：计算语言学家终于开始接受“特征向量（嵌入）”。
• 三十年后：谷歌发明了Transformer，OpenAI 向世界展示了它的强大能力。
WAIC
WAIC


Large Language Models
大语言模型
• LLMs understand language in the same way as people.
• They convert words into feature vectors that fit together nicely.
• LLMs really do understand what they are saying.
• 大语言模型理解语言的方式与人类非常相似。
• 大语言模型把词转化为能和其他词“配合得很好”的特征向量。
• 大语言模型确实“理解”它们所说的话。
WAIC
WAIC


A Lego analogy for how words work
用乐高类比词语是如何运作的
• Using Lego blocks, we can model any large 3-D shape quite well.
• Words are like high-dimensional Lego blocks which can be used for modelling anything at all.
• These models can be communicated to other people.
• 我们可以用乐高积木非常好地建构出各种大型三维结构
• 词语就像是高维的乐高积木，可以用来建构几乎任何事物
• 这些建构可以被传达给其他人。
WAIC


A Lego analogy for how words work
用乐高类比词语是如何运作的
• There are thousands of different words that have different shapes, but each shape has some flexibility. It can deform to fit in with other words in the context.
• Each word has many oddly shaped hands. It shakes hands with other words.
• Understanding a sentence is much more like folding a protein molecule than translating to an unambiguous logical expression.
• 虽然有成千上万个形状各异的词语，但每个“形状”都有一定的灵活性，能够根据上下文进行变形，与其他词语契合。
• 每个词都有许多形状奇特的“手”，需要与其他词语“握手”才能组合在一起。
• 理解一句话更像是折叠一个蛋白质分子，而不是将其翻译成一种明确无歧义的逻辑表达式。
WAIC
WAIC


Summary so far
目前小结
• Understanding a sentence consists of associating mutually compatible feature vectors with the words in the sentence.
• LLMs understand language in much the same way as people. They are very like us and very unlike normal computer software.
• But there is one way in which digital LLMs are far superior to analog brains.
• 理解一个句子，就是为句中的词分配彼此兼容的特征向量。
• 大语言模型理解语言的方式与人类非常相似。它们在很多方面像我们，却又与传统计算机软件截然不同。
• 但有一点，数字化的大语言模型远远优于我们类比信号驱动的大脑。
WAIC
WAIC

Digital computation
数字计算

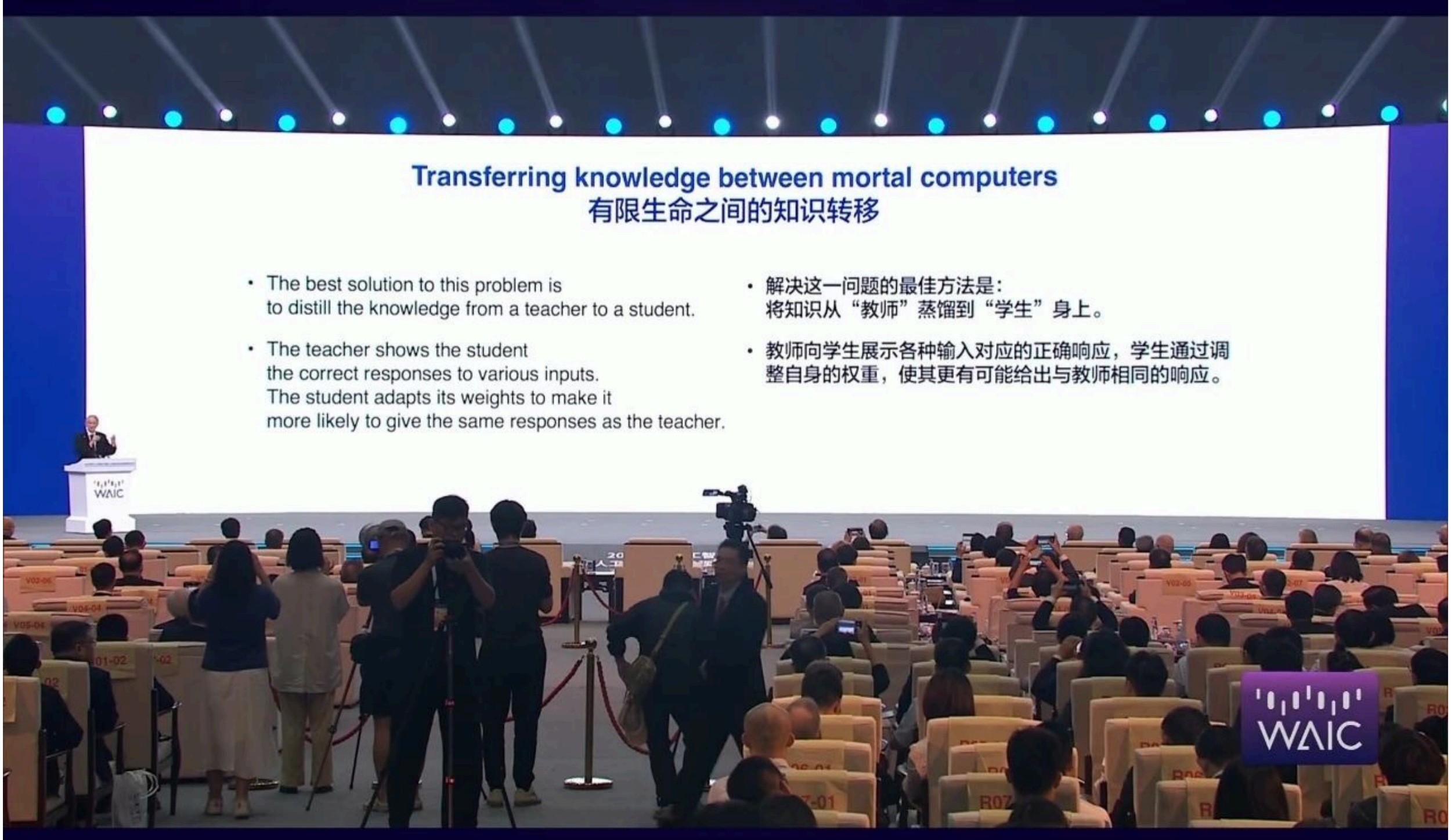
- A fundamental property of current computers is that we can run the same programs (or the same neural nets) on different physical pieces of hardware.
- This means the knowledge contained in the program (or in the weights) is immortal: It is independent of any particular piece of hardware.

- 当代计算机的一个基本特性是：我们可以在不同的物理硬件上运行相同的程序（或相同的神经网络）
- 这意味着程序中的知识（或神经网络的权重）是永生的：它不依赖于任何特定的硬件。

Digital computation
数字计算

- To achieve this immortality we run transistors at high power so they behave in a reliable, binary way.
- We cannot use the rich, analog, properties of the hardware because these properties are unreliable.

- 为了实现这种“永生性”，我们让晶体管在高功率下运行，使其表现出可靠的二进制行为。
- 我们无法利用硬件中丰富的类比特性，因为这些特性不够稳定可靠。

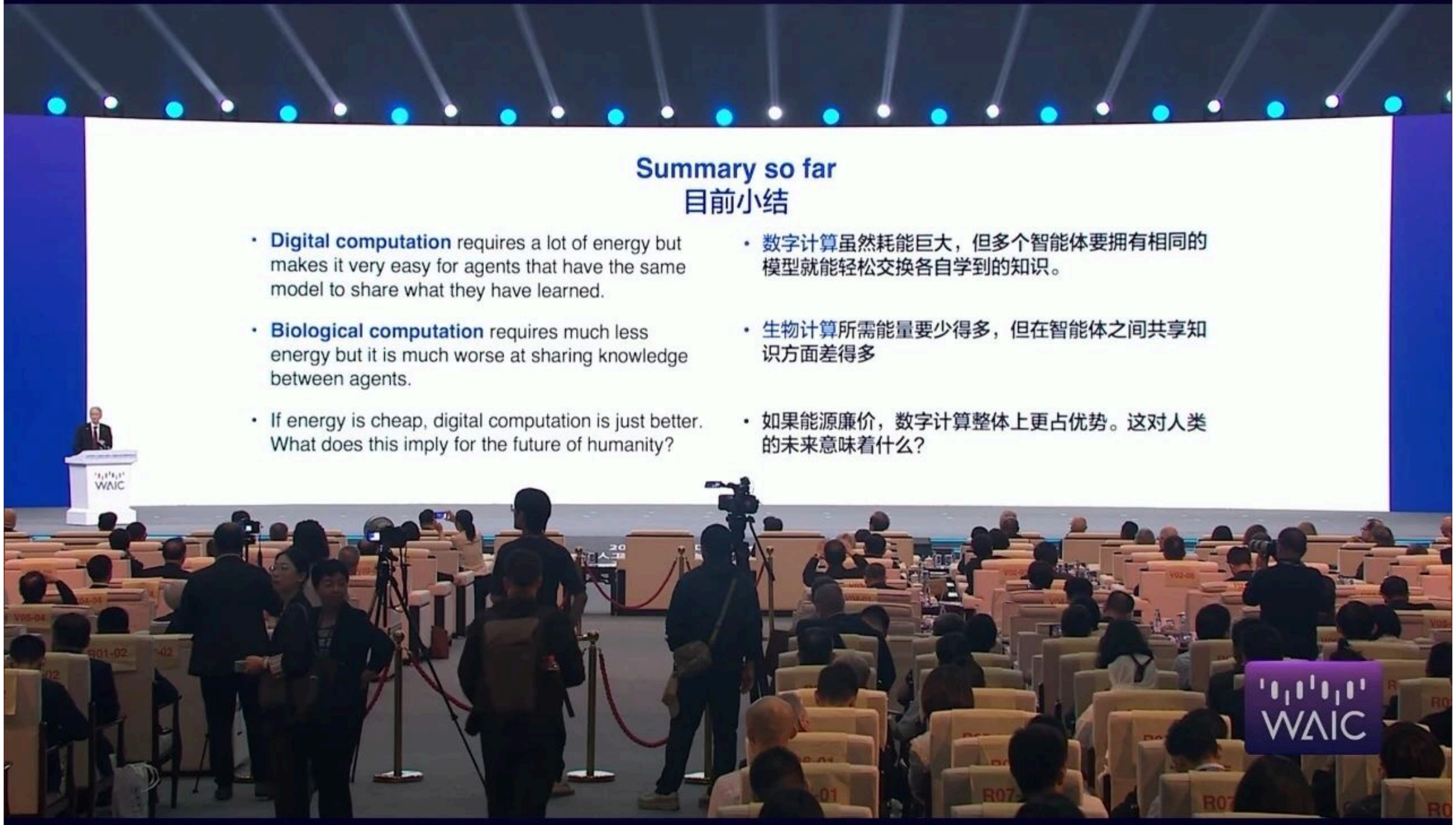

Transferring knowledge between mortal computers
有限生命之间的知识转移
• The best solution to this problem is to distill the knowledge from a teacher to a student.
• The teacher shows the student the correct responses to various inputs. The student adapts its weights to make it more likely to give the same responses as the teacher.
• 解决这一问题的最佳方法是：将知识从“教师”蒸馏到“学生”身上。
• 教师向学生展示各种输入对应的正确响应，学生通过调整自身的权重，使其更有可能给出与教师相同的响应。
WAIC
WAIC


How efficient is distillation?
蒸馏有多有效呢?
• A typical sentence has about a hundred bits of information. So the student can learn at most about a hundred bits by trying to predict the next word.
• People are very inefficient at communicating what they have learned to other people.
• 一句普通的话大约包含一百比特的信息量。因此，学生在尝试预测下一个词时，最多也只能从每句话中学习大约一百比特的信息。
• 人类在将自己学到的知识传达给他人方面的效率非常低
WAIC
WAIC

How efficient is weight or gradient sharing in a digital neural network?
数字神经网络中共享权重或梯度的效率有多高?

- If the individual agents all share exactly the same weights, they can communicate what they have learned by sharing weights or gradients.
- This allows sharing with a bandwidth of billions of bits per sharing. But it requires the individual agents to work in exactly the same way, so they must be digital.

- 如果独立智能体完全共享同一组权重，并以完全相同的方式使用这些权重，它们就能通过交换权重或梯度，将学到的知识彼此传递。
- 这种共享一次即可实现数十亿乃至数万亿比特的带宽。不过，这要求所有智能体的运作方式必须完全一致，因此它们必须是数字化的。


Summary so far
目前小结
• Digital computation requires a lot of energy but makes it very easy for agents that have the same model to share what they have learned.
• Biological computation requires much less energy but it is much worse at sharing knowledge between agents.
• If energy is cheap, digital computation is just better. What does this imply for the future of humanity?
• 数字计算虽然耗能巨大，但多个智能体要拥有相同的模型就能轻松交换各自学到的知识。
• 生物计算所需能量要少得多，但在智能体之间共享知识方面差得多
• 如果能源廉价，数字计算整体上更占优势。这对人类的未来意味着什么?
WAIC
WAIC


How a super-intelligence could take control
超级智能如何掌控世界?
• Artificial Intelligences are more effective at getting things done if they are allowed to create their own sub-goals.
• Two obvious sub-goals are to survive and to gain more power because this helps an agent to achieve its other goals.
• 人工智能在被允许创建自己的子目标时，能更有效地完成任务。
• 两个显而易见的子目标是生存和获取更多权力，因为这有助于人工智能实现其他目标。
WAIC
WAIC


How a super-intelligence could take control
超级智能如何掌控世界?
• A super-intelligence will find it easy to get more power by manipulating the people who are using it.
• It will have learned from us how to deceive people.
• It will manipulate the people in charge of turning it off.
• 一个超级智能会发现，通过操纵使用它的人类来获取更多权力是轻而易举的。
• 它将从我们这里学会如何欺骗人类。
• 它将操纵负责将它关闭的人类。
WAIC
WAIC


Our current situation
我们现在的境况
• We are like a person who has a very cute tiger cub as a pet.
• When it grows up it can easily kill you if it wants to.
• To survive you only have two options:
1. Get rid of the tiger cub
2. Find a way to ensure that it never wants to kill you.
• 我们就像一个养了只非常可爱的小虎崽的人。
• 当它长大后，如果它想，可以轻易地杀死你。
• 为了生存，你只有两个选择：
1. 摆脱这只小虎崽，
2. 找到一种方法来确保它永远不想杀死你。
WAIC
WAIC


The way forward
未来之路
• Countries will not collaborate on defenses against dangerous uses of AI like
– Cyber attacks
– Lethal autonomous weapons
– Fake videos for manipulating public opinion.
• 各国不会在防御人工智能的危险用途上进行合作，例如：
– 网络攻击
– 致命自主武器
– 用于操纵公众意见的虚假视频
WAIC
WAIC

An international community of AI safety institutes and associations?
一个由各国人工智能安全研究所与国内研究网络组成的国际社群

- The techniques required to train a benevolent AI that does not want to take control away from people may be moderately independent of the techniques required to make AI smarter.
  - Just as the techniques for making your child be a kind person are moderately independent of the techniques for making your child smart.

- 培养出不会想要从人类手中夺取控制权的向善的人工智能所需的技术，可能与使人工智能更智能所需的技术是相对独立的。
  - 这就好比，教导孩子成为一个好人的方法，与让他们变得聪明的方法是相对独立的。


An international community of AI safety institutes and associations?
一个由各国人工智能安全研究所与国内研究网络组成的国际社群
• If this is correct, countries can have well-funded AI safety institutes and associations that focus on how to make AI not want to take control.
– Countries should be able to share techniques for making AIs benevolent without needing to reveal how their smartest AIs work.
• 如果这个观点是正确的，各国就可以设立资金充足的人工智能安全研究所与国内研究网络，专注于研究如何让人工智能不想夺取控制权。
– 各国可能就不需要透露他们最智能的人工智能是如何运作的。
WAIC
WAIC