人工智能浪潮下的基础教育变革：观察与思考

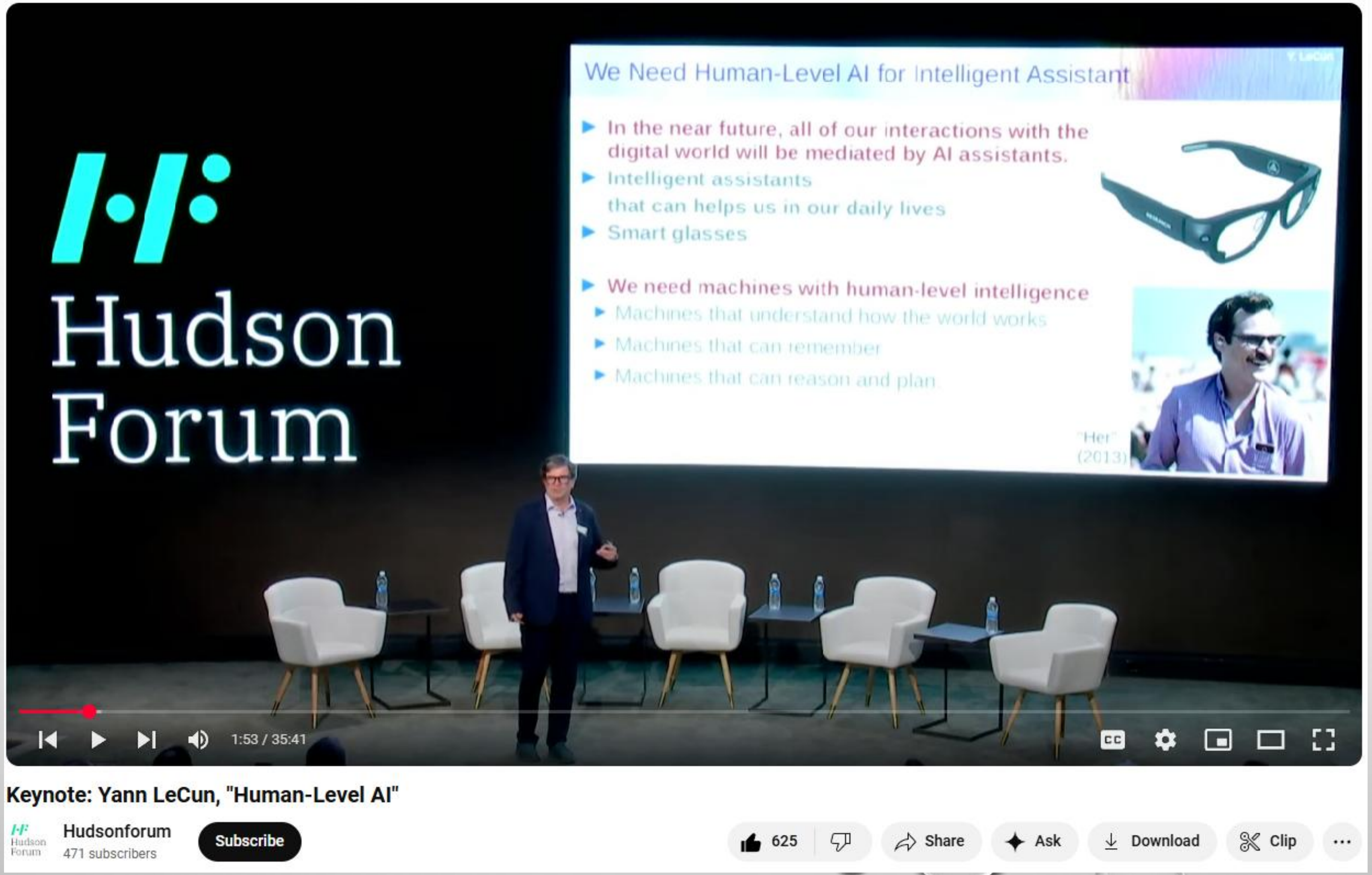
VERITAS EDU, 2025-03-23

概要

- 走向通用人工智能（AGI）的道路仍在探索中（杨立昆）
- 大语言模型的本质，很可能只是模式匹配而非真正的智能（苹果公司的论文）
- 现有的人工智能技术，前途在于应用场景的发现与工程解决方案（吴恩达）
- 青少年应对人工智能带来的机遇与挑战，需要具备三个技能（个人看法）
    1. 发现问题：商业或者学术上的洞察力、具体行业的专业知识（DOMAIN KNOWLEDGE）
    2. 转化问题：STEM的训练（PROBLEM FORMULATION IS ALREADY HALF OF THE SOLUTION）
    3. 解决问题：计算机只是工具
        a) 基本的计算机知识是新时代的EXCEL和汽车驾驶
        b) 深入的计算机应用需要软硬件结合
- 一个启发性的问题：如果把行星运行轨迹的数据喂给现在的人工智能模型，它们能否自动思考出“力”的概念、牛顿力学三定律、以至于万有引力公式？

$$F = G\frac{m_1 m_2}{r^2}$$

YANN LECUN: HUMAN-LEVEL AI （2024.10.13）



We Need Human-Level AI for Intelligent Assistant

- In the near future, all of our interactions with the digital world will be mediated by AI assistants.
- Intelligent assistants that can helps us in our daily lives
- Smart glasses

- We need machines with human-level intelligence
  - Machines that understand how the world works
  - Machines that can remember
  - Machines that can reason and plan

"Her" (2013)

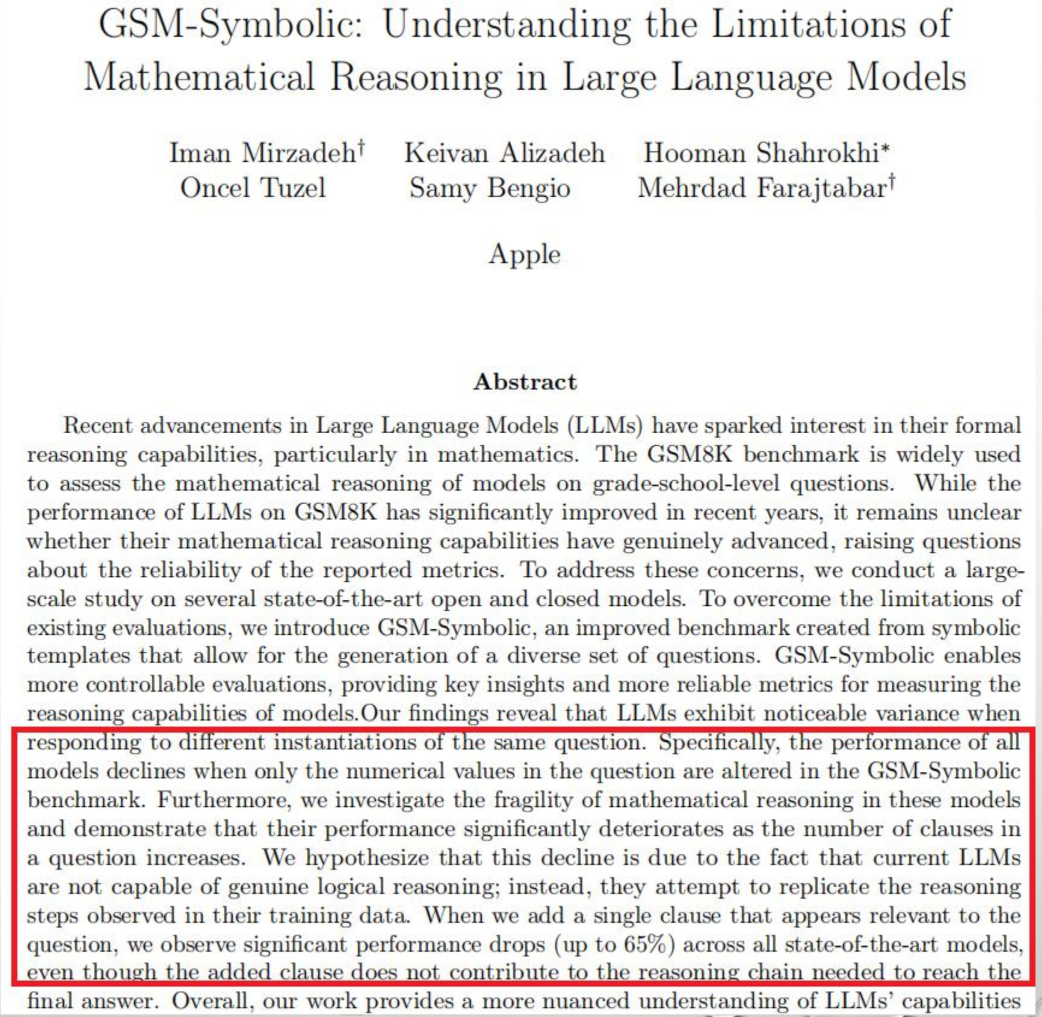
Keynote: Yann LeCun, "Human-Level AI"

Hudsonforum
471 subscribers

杨立昆：走向通用人工智能（AGI）之路仍然未知

- **推理规划能力缺失**：当前 *AI* 系统缺乏真正的推理和规划能力。像自动回归预测的 大语言模型，只是依据前文预测下一个词，无法进行深入推理，也不能像人类一样面对新情况时思考行动后果并规划出合适的行动序列。
- **数据处理类型受限**：自动回归预测仅适用于离散数据，如文本、符号等，无法处理连续数据或复杂的现实场景数据。 而现实世界中的数据丰富多样，仅靠处理离散数据无法实现通用人工智能。
- **学习方式存在缺陷**：过去试图通过预测视频像素来让 *AI* 学习常识和物理知识的尝试完全失败。因为视频预测存在多种可能，且难以表示视频帧的概率分布，无法准确预测像素级的内容。

大语言模型可能只是“模式匹配器” 而非“推理器”（ARCHIVE, 2024.10.07）



GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models

Iman Mirzadeh† Keivan Alizadeh Hooman Shahrokhi*
Oncel Tuzel Samy Bengio Mehrdad Farajtabar†

Apple

Abstract

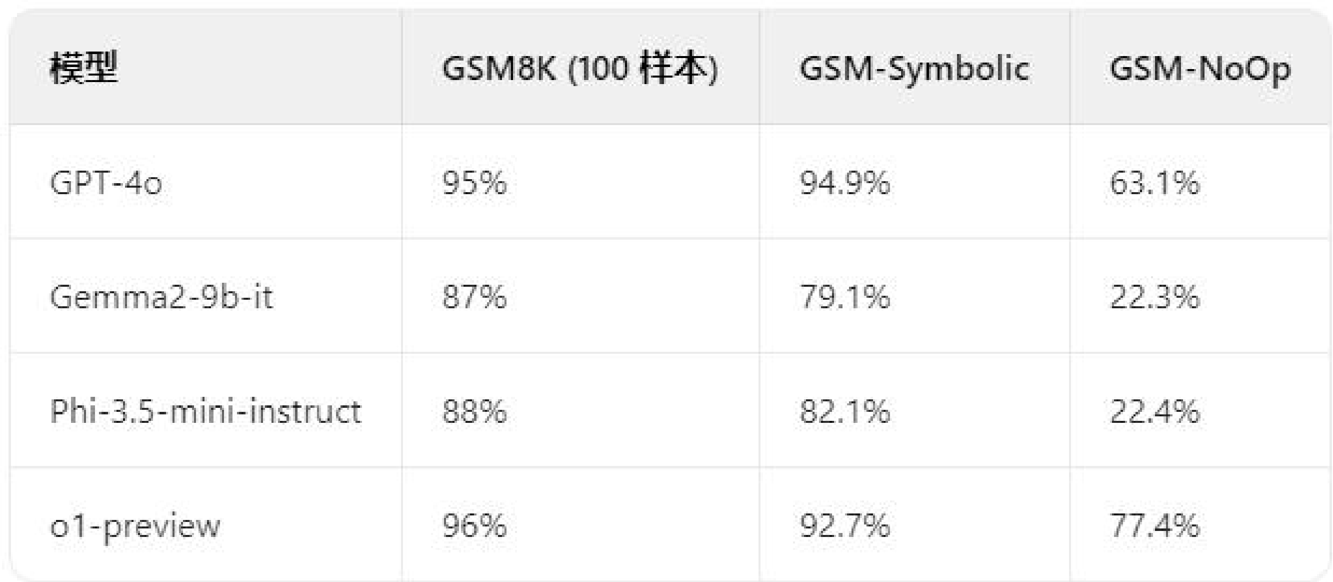
Recent advancements in Large Language Models (LLMs) have sparked interest in their formal reasoning capabilities, particularly in mathematics. The GSM8K benchmark is widely used to assess the mathematical reasoning of models on grade-school-level questions. While the performance of LLMs on GSM8K has significantly improved in recent years, it remains unclear whether their mathematical reasoning capabilities have genuinely advanced, raising questions about the reliability of the reported metrics. To address these concerns, we conduct a large-scale study on several state-of-the-art open and closed models. To overcome the limitations of existing evaluations, we introduce GSM-Symbolic, an improved benchmark created from symbolic templates that allow for the generation of a diverse set of questions. GSM-Symbolic enables more controllable evaluations, providing key insights and more reliable metrics for measuring the reasoning capabilities of models.Our findings reveal that LLMs exhibit noticeable variance when responding to different instantiations of the same question. Specifically, the performance of all models declines when only the numerical values in the question are altered in the GSM-Symbolic benchmark. Furthermore, we investigate the fragility of mathematical reasoning in these models and demonstrate that their performance significantly deteriorates as the number of clauses in a question increases. We hypothesize that this decline is due to the fact that current LLMs are not capable of genuine logical reasoning; instead, they attempt to replicate the reasoning steps observed in their training data. When we add a single clause that appears relevant to the question, we observe significant performance drops (up to 65%) across all state-of-the-art models, even though the added clause does not contribute to the reasoning chain needed to reach the final answer. Overall, our work provides a more nuanced understanding of LLMs' capabilities

苹果公司论文的主要结论

- **研究背景与目的**：在大模型推理能力受关注的背景下，苹果公司研究者发表论文。此前业界常用 GSM8K 数据集（8000道小学水平的数学题）评估大模型数学能力，但存在数据污染问题（“提前背真题”），苹果团队开发 GSM-SYMBOLIC 和 GSM-NOOP 数据集，旨在客观评价大模型数学能力极限。
- **实验过程与结果**：用新数据集测试多种开源和闭源模型，包括 GPT 系列、LLAMA 等。结果显示，面对 GSM-SYMBOLIC 换皮题目，模型准确率下降；题目难度增加时，模型准确性降低、方差变大；面对含无关论述的 GSM-NOOP 数据集，模型性能更是大幅下降，表明大模型无法真正理解数学问题。
- **大模型推理能力分析**：大模型解决问题方式可能是线性化子图匹配（FAITH AND FATE, 2023），通过与训练数据中相似子图匹配进行预测，而非真正推理。实验发现，模型在乘法等任务上，无法从训练集中小问题推广到更大问题，更像是记住答案，以类似搜索引擎方式解决问题。

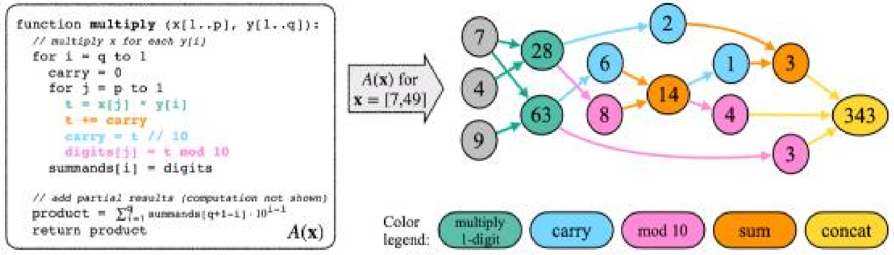
大模型“数学推理”本质上是“概率模式匹配”

- **为什么添加无关信息会导致模型性能骤降？**LLMS 依赖训练数据中的模式匹配，无法区分相关与无关信息，盲目将冗余信息转化为操作
- 当前模型在数学推理中的核心局限是什么？
  - **模式匹配依赖**：无法进行形式逻辑推理，对数值变化和复杂度敏感
  - **多步骤处理能力不足**：随着问题条款增加，准确率呈指数级下降



模型	GSM8K (100 样本)	GSM-Symbolic	GSM-NoOp
GPT-4o	95%	94.9%	63.1%
Gemma2-9b-it	87%	79.1%	22.3%
Phi-3.5-mini-instruct	88%	82.1%	22.4%
o1-preview	96%	92.7%	77.4%

信仰与命运：作为模糊匹配器的TRANSFORMERS



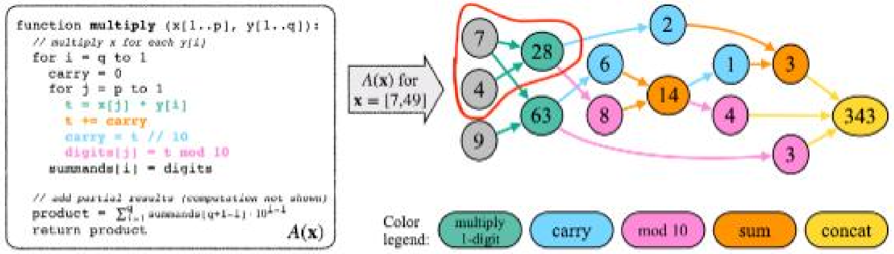
```
function multiply (x[1..p], y[1..q]):
  // multiply x for each y[i]
  for i = q to 1
    carry = 0
    for j = p to 1
      t = x[j] * y[i]
      t += carry
      carry = t // 10
      digits[j] = t mod 10
    summands[i] = digits

  // add partial results (computation not shown)
  product = $\sum_{i=1}^{q}$ summands[q+1-i] · $10^{i-1}$
  return product
```

Figure 1: Transformation of an algorithm A to its computational graph $G_{A(\mathbf{x})}$. The depicted example is of long-form multiplication algorithm A, for inputs $\mathbf{x} = [7, 49]$ (i.e. computing 7×49).

describing the procedure as a graph

This is a kind of problem decomposition. For instance, if you look within this graph, you can see that the task of multiplying 7 times 49 requires the subtask of multiplying 7 times 4, and that the graph of the larger problem contains the subtask as a subgraph:

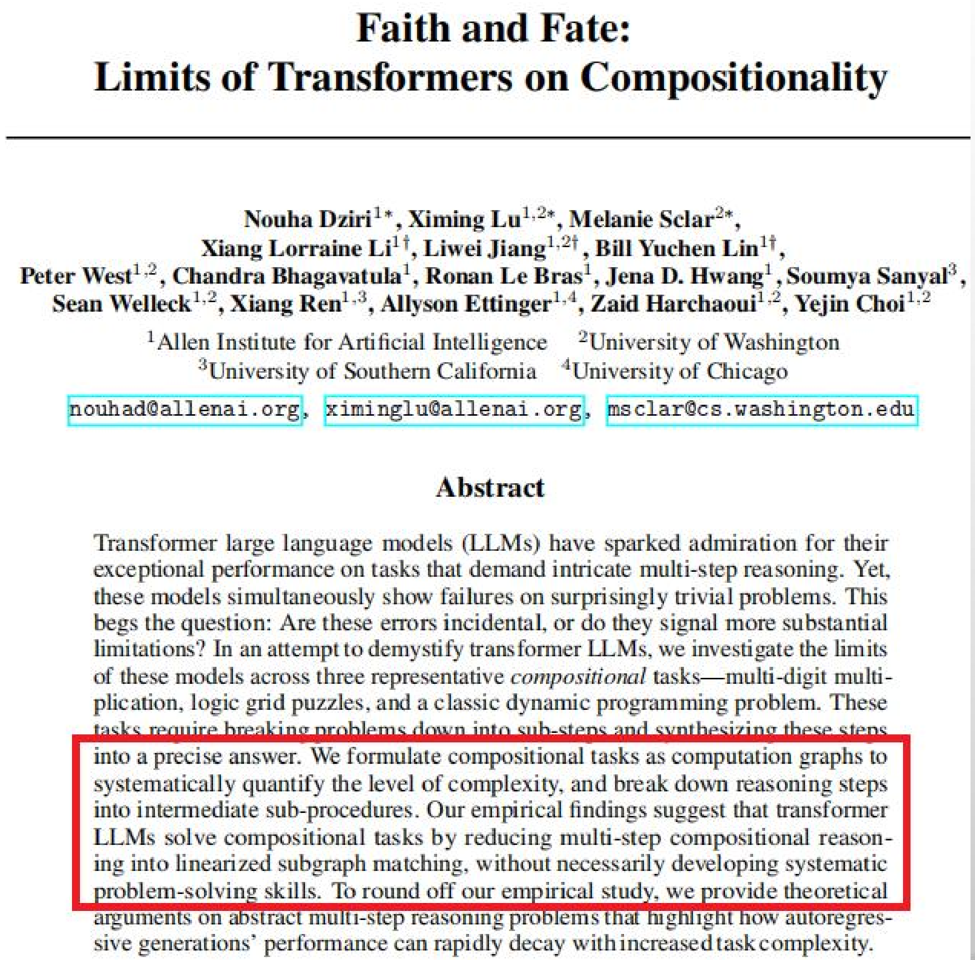


```
function multiply (x[1..p], y[1..q]):
  // multiply x for each y[i]
  for i = q to 1
    carry = 0
    for j = p to 1
      t = x[j] * y[i]
      t += carry
      carry = t // 10
      digits[j] = t mod 10
    summands[i] = digits

  // add partial results (computation not shown)
  product = $\sum_{i=1}^{q}$ summands[q+1-i] · $10^{i-1}$
  return product
```

Figure 1: Transformation of an algorithm A to its computational graph $G_{A(\mathbf{x})}$. The depicted example is of long-form multiplication algorithm A, for inputs $\mathbf{x} = [7, 49]$ (i.e. computing 7×49).

suprocedures are subgraphs in the graph



Faith and Fate: Limits of Transformers on Compositionality

Nouha Dziri[1*], Ximing Lu[1,2*], Melanie Sclar[2*], Xiang Lorraine Li[1†], Liwei Jiang[1,2†], Bill Yuchen Lin[1†], Peter West[1,2], Chandra Bhagavatula[1], Ronan Le Bras[1], Jena D. Hwang[1], Soumya Sanyal[3], Sean Welleck[1,2], Xiang Ren[1,3], Allyson Ettinger[1,4], Zaid Harchaoui[1,2], Yejin Choi[1,2]

[1]Allen Institute for Artificial Intelligence [2]University of Washington
[3]University of Southern California [4]University of Chicago

nouhad@allenai.org, ximinglu@allenai.org, msclar@cs.washington.edu

Abstract

Transformer large language models (LLMs) have sparked admiration for their exceptional performance on tasks that demand intricate multi-step reasoning. Yet, these models simultaneously show failures on surprisingly trivial problems. This begs the question: Are these errors incidental, or do they signal more substantial limitations? In an attempt to demystify transformer LLMs, we investigate the limits of these models across three representative *compositional* tasks—multi-digit multiplication, logic grid puzzles, and a classic dynamic programming problem. These tasks require breaking problems down into sub-steps and synthesizing these steps into a precise answer. We formulate compositional tasks as computation graphs to systematically quantify the level of complexity, and break down reasoning steps into intermediate sub-procedures. Our empirical findings suggest that transformer LLMs solve compositional tasks by reducing multi-step compositional reasoning into linearized subgraph matching, without necessarily developing systematic problem-solving skills. To round off our empirical study, we provide theoretical arguments on abstract multi-step reasoning problems that highlight how autoregressive generations' performance can rapidly decay with increased task complexity.

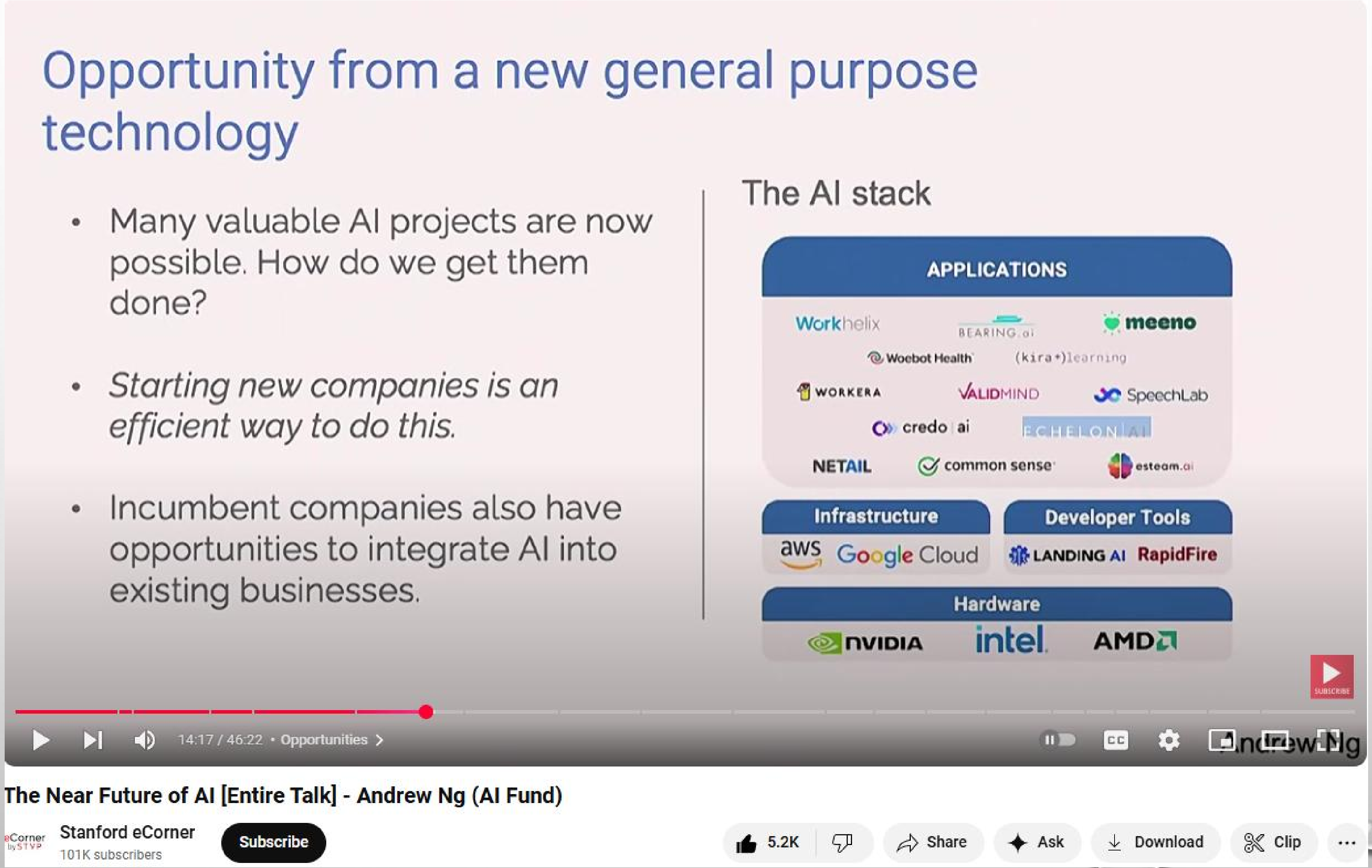
“大模型在推理”的人类幻觉

- **模式匹配机制：**大模型采用线性化子图匹配方式解决问题。任何数学问题都可表示为有向图，大模型将任务描述为一系列步骤，这些步骤构成有向图中的子图。在面对问题时，它会把问题子图与训练数据中的相似子图进行匹配，进而做出预测。在乘法任务中，它能答对与训练数据规模相同的乘法问题，是因为这些问题的子图与训练数据中的子图相似，模型通过匹配来给出答案，并非真正理解乘法原理，只是看似能推理解决问题 。

- **思维链示例引导：**研究者提供的思维链示例，帮助模型把乘法问题分解为更小的任务，让模型在一定程度上依照示例步骤进行解题。这种引导使模型在解题过程中有了类似推理的步骤呈现，给人它在推理的感觉。但实验表明，当问题超出训练范围或子问题与主问题逻辑关系复杂时，模型就无法正确解答，说明其并非真正具备推理能力 。

- **经验记忆运用：**大模型对训练数据中频繁出现的精确子问题有记忆。在遇到类似问题时，它通过回忆记住的答案来解决，比如记住 “7 乘以 4 等于 28”，在遇到相关子问题时就能答对。在处理常见简单数学问题时，这种记忆运用能给出正确答案，让人误以为它在推理 。

吴恩达看AI提供的机遇（2023）



Opportunity from a new general purpose technology

- Many valuable AI projects are now possible. How do we get them done?
- *Starting new companies is an efficient way to do this.*
- Incumbent companies also have opportunities to integrate AI into existing businesses.

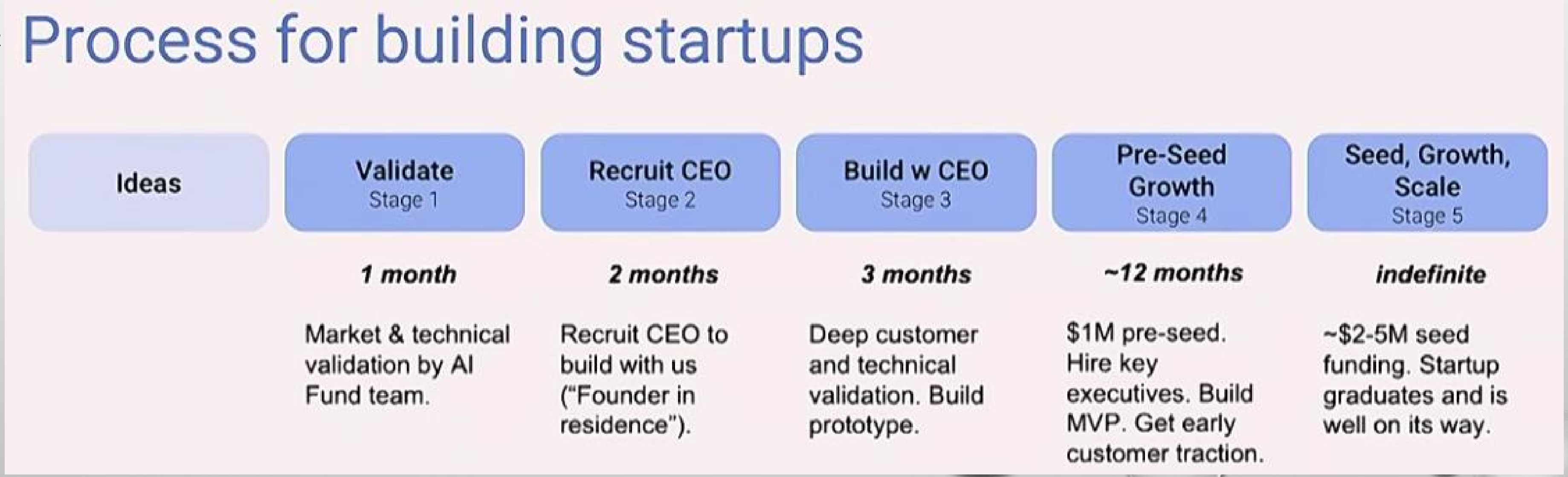
The AI stack

The Near Future of AI [Entire Talk] - Andrew Ng (AI Fund)

Stanford eCorner
101K subscribers

吴恩达对AI生态圈的归类

- **硬件层与云基础设施层**：硬件层价值高但资本密集、资源集中，吴恩达个人较少涉足；云基础设施层同样资本密集、竞争集中，他在创业时也倾向避开这一层。
- **开发者工具层**：竞争极为激烈，众多初创公司追逐类似 OPENAI 这样的工具。吴恩达认为初创公司若想在这一层取得成功，拥有技术优势是关键，这样才更有可能成为大赢家。
- **应用层**：竞争相对没那么激烈，存在大量机会。通过与不同领域专家合作，能够开发出独特的应用。例如 AI FUND 与相关专家合作，在恋爱关系辅导、船舶智能路由等方面都取得了成



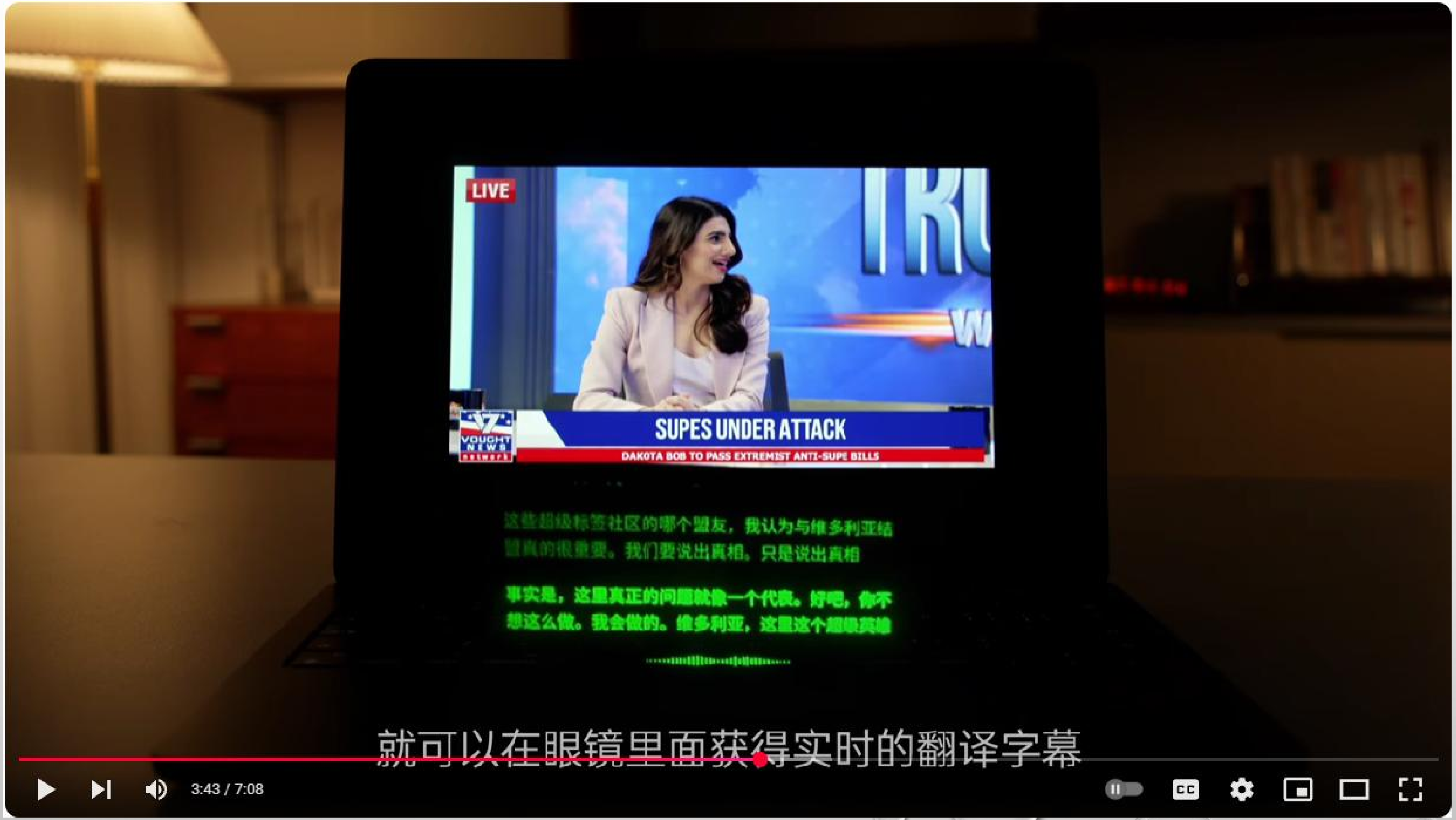
Process for building startups

Ideas	Validate Stage 1	Recruit CEO Stage 2	Build w CEO Stage 3	Pre-Seed Growth Stage 4	Seed, Growth, Scale Stage 5
	1 month	*2 months*	*3 months*	*~12 months*	*indefinite*
	Market & technical validation by AI Fund team.	Recruit CEO to build with us ("Founder in residence").	Deep customer and technical validation. Build prototype.	$1M pre-seed. Hire key executives. Build MVP. Get early customer traction.	~$2-5M seed funding. Startup graduates and is well on its way.

吴恩达对学生的建议

- **扎实学习基础知识**：建议学生重视课程学习，通过学习 AI 技术和创业课程来掌握基础技能。因为教授会精心组织课程内容，有助于学生更高效地学习和掌握基础知识，比直接参与项目工作更有效。
- **积极实践与跨学科探索**：在掌握基础技能后，学生应积极在校园内寻找跨学科的应用场景，将 AI 技术与其他学科结合，挖掘有价值的项目。比如将 AI 技术应用于气候科学、医疗保健等领域，找到创新的应用点。
- **培养关键能力**：优秀的 AI 创始人除了具备技术能力外，还需要有快速决策的能力。同时，在追求速度的过程中要确保符合负责任 AI 的原则，避免对人们的生活和生计造成伤害。

案例1：谷歌眼镜与中国的实时翻译眼镜

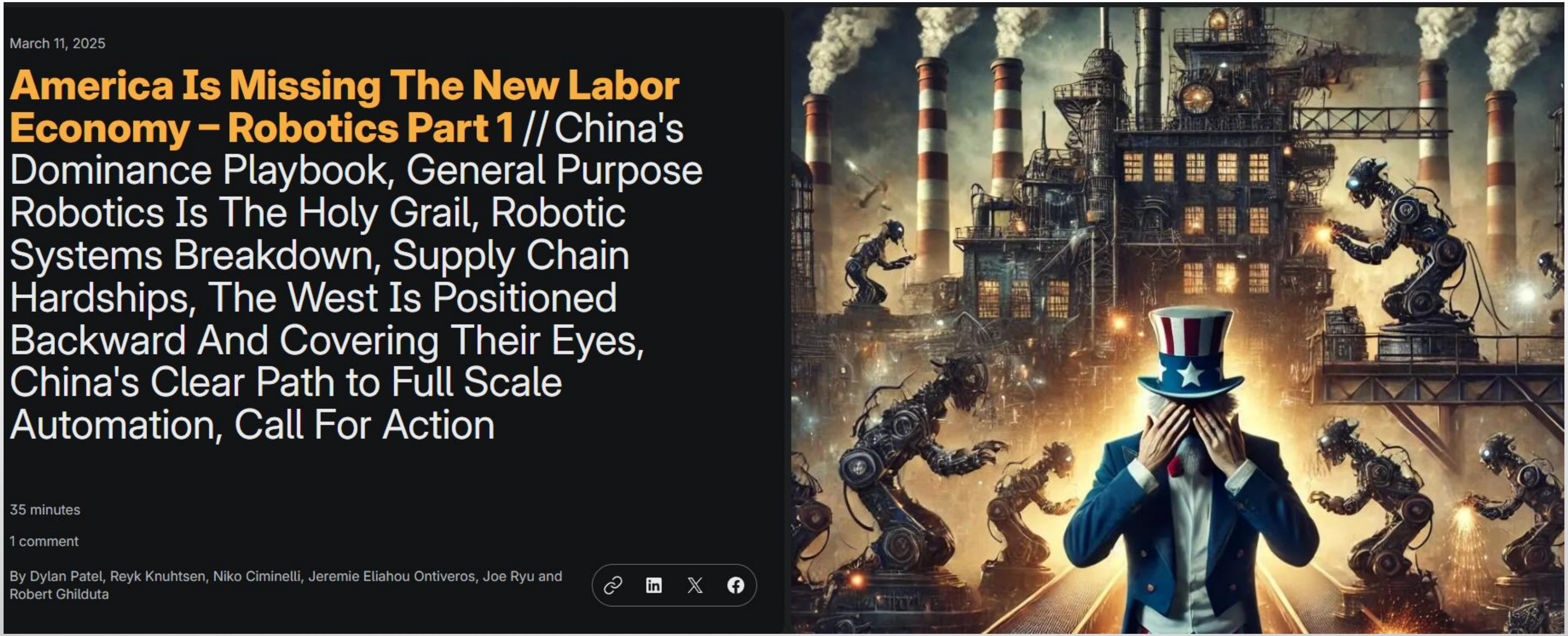


这些超级标签社区的哪个盟友，我认为与维多利亚结盟真的很重要。我们要说出真相，只是说出真相

事实是，这里真正的问题就像一个代表。好吧，你不想这么做。我会做的。维多利亚，这里这个超级英雄

就可以在眼镜里面获得实时的翻译字幕

案例2：机器人技术与中国的无人工厂



March 11, 2025

America Is Missing The New Labor Economy – Robotics Part 1 // China's Dominance Playbook, General Purpose Robotics Is The Holy Grail, Robotic Systems Breakdown, Supply Chain Hardships, The West Is Positioned Backward And Covering Their Eyes, China's Clear Path to Full Scale Automation, Call For Action

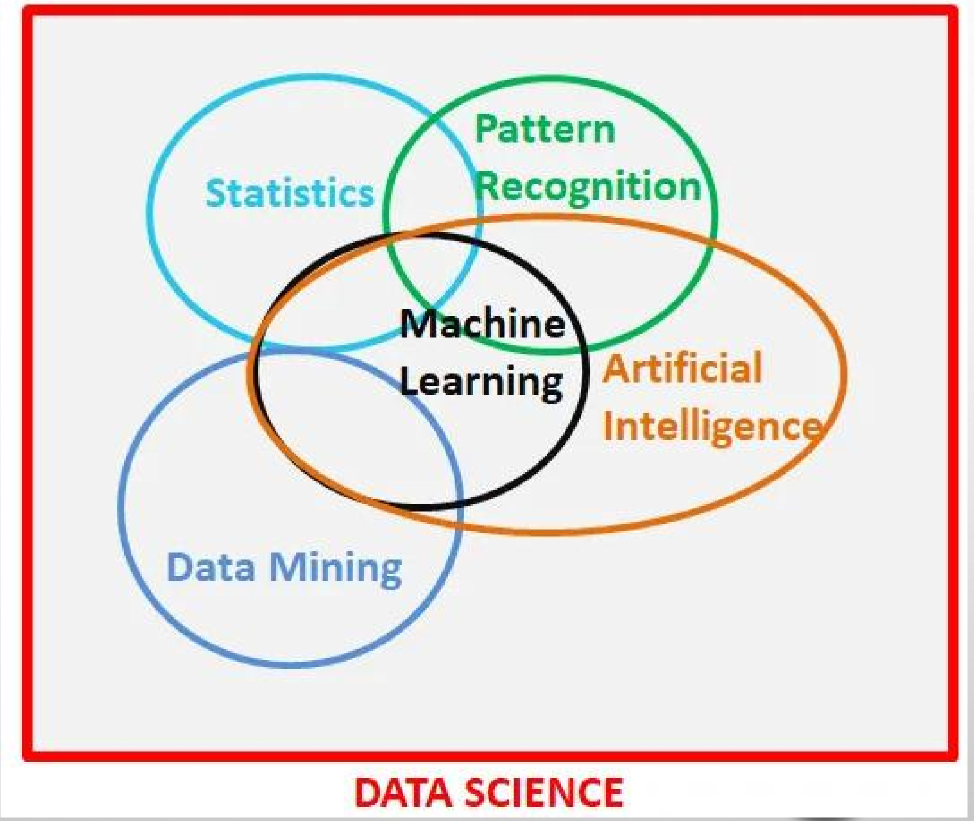
35 minutes

1 comment

By Dylan Patel, Reyk Knuhtsen, Niko Ciminelli, Jeremie Eliahou Ontiveros, Joe Ryu and Robert Ghilduta

理解现有人工智能的进展，对K-12教育的启示

- 即使人工智能只是更复杂的模式匹配，它仍然可以大有作为：医疗、金融、军事
- 人工智能是个很宽泛的概念，涵盖若干领域，其理论基础仍然是数学：线性代数、微积分、最优化理论、概率论、统计学



培养洞察力：问题驱动、边做边学

安泰经济与管理学院
ANTAI COLLEGE OF ECONOMICS & MANAGEMENT

“科技-金融-商业”循环加速器

——一个历史的视角

文一

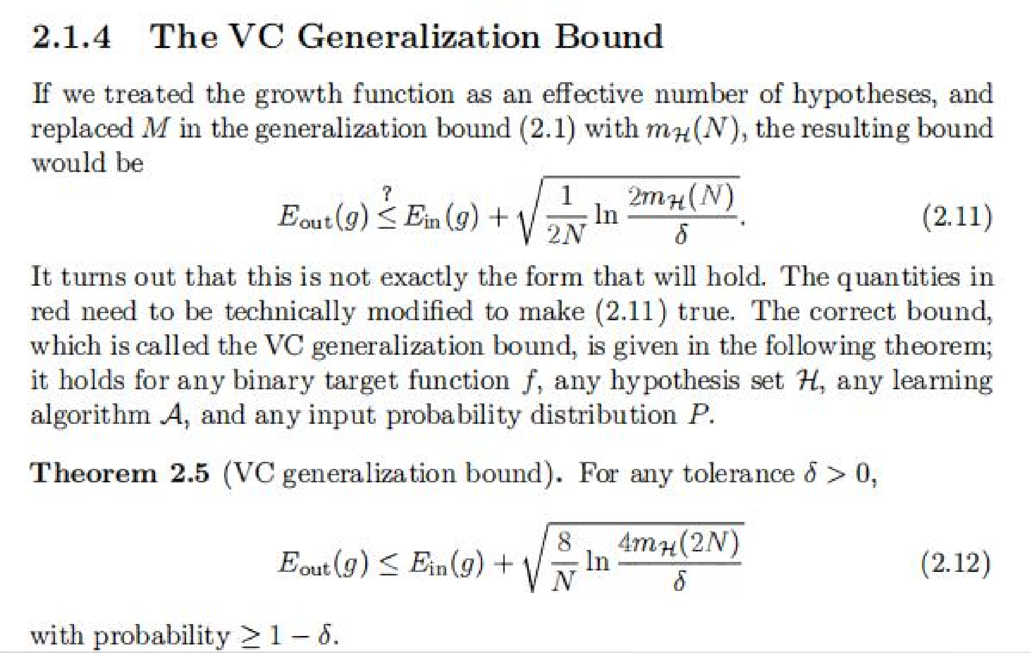
科技金融“浩然讲堂”2023. 3. 18

上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

安泰经管学院
ANTAI COLLEGE
Economics·Management

数理功底是AI的基础：“炼金术”与“化学”

- CALTECH: LEARNING FROM DATA HTTPS://WORK.CALTECH.EDU/TELECOURSE



2.1.4 The VC Generalization Bound

If we treated the growth function as an effective number of hypotheses, and replaced M in the generalization bound (2.1) with $m_{\mathcal{H}}(N)$, the resulting bound would be

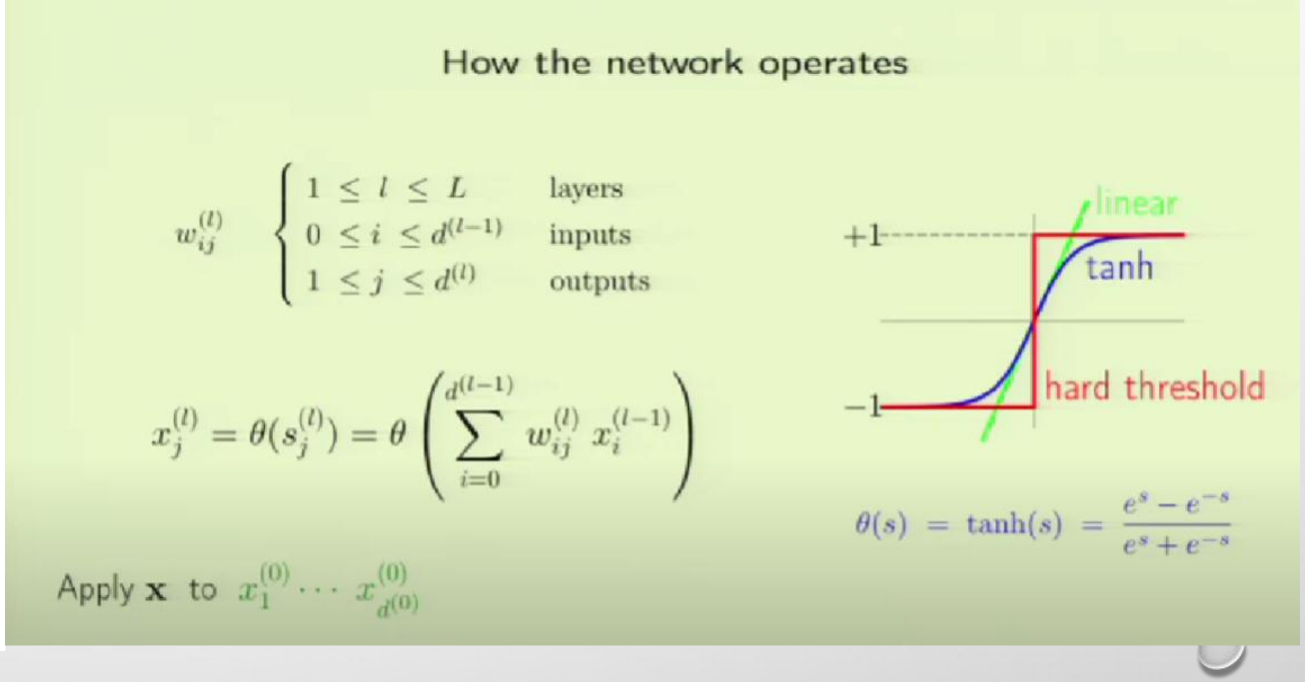
$$E_{\text{out}}(g) \overset{?}{\leq} E_{\text{in}}(g) + \sqrt{\frac{1}{2N} \ln \frac{2m_{\mathcal{H}}(N)}{\delta}}. \qquad (2.11)$$

It turns out that this is not exactly the form that will hold. The quantities in red need to be technically modified to make (2.11) true. The correct bound, which is called the VC generalization bound, is given in the following theorem; it holds for any binary target function f, any hypothesis set $\mathcal{H}$, any learning algorithm $\mathcal{A}$, and any input probability distribution P.

Theorem 2.5 (VC generalization bound). For any tolerance $\delta > 0$,

$$E_{\text{out}}(g) \leq E_{\text{in}}(g) + \sqrt{\frac{8}{N} \ln \frac{4m_{\mathcal{H}}(2N)}{\delta}} \qquad (2.12)$$

with probability $\geq 1 - \delta$.



How the network operates

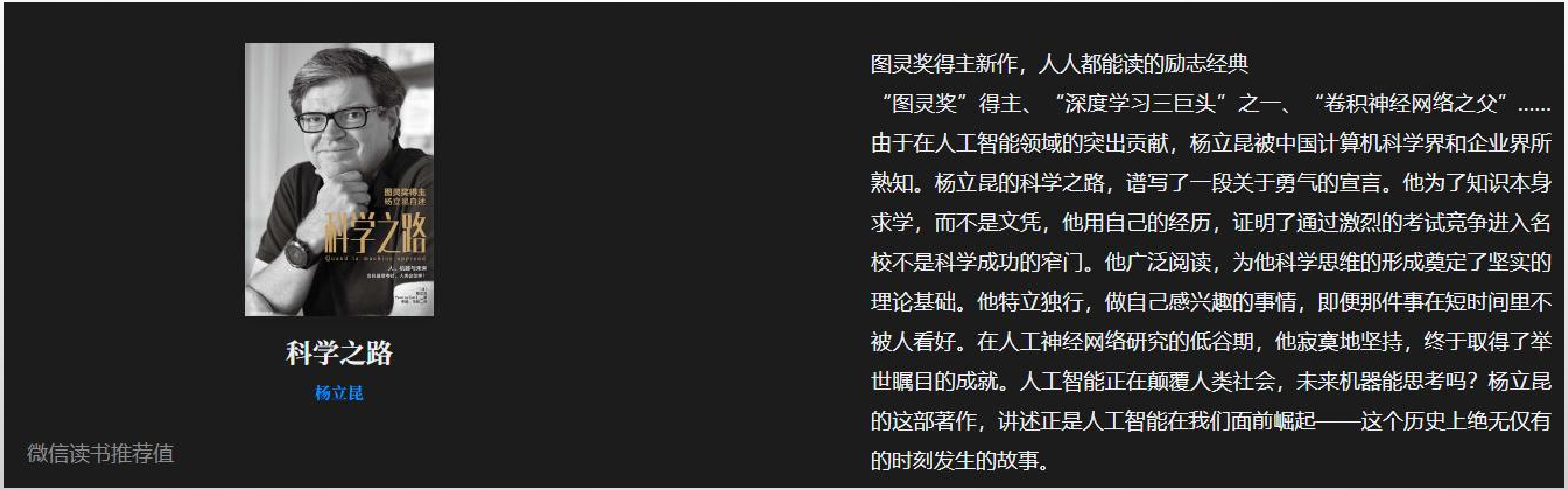
$$w_{ij}^{(l)} \quad \begin{cases} 1 \leq l \leq L & \text{layers} \\ 0 \leq i \leq d^{(l-1)} & \text{inputs} \\ 1 \leq j \leq d^{(l)} & \text{outputs} \end{cases}$$

$$x_j^{(l)} = \theta(s_j^{(l)}) = \theta\left(\sum_{i=0}^{d^{(l-1)}} w_{ij}^{(l)}\, x_i^{(l-1)}\right)$$

$$\theta(s) = \tanh(s) = \frac{e^s - e^{-s}}{e^s + e^{-s}}$$

Apply $\mathbf{x}$ to $x_1^{(0)} \cdots x_{d^{(0)}}^{(0)}$

把握时代的发展趋势：向前看，而非向后看



科学之路

杨立昆

微信读书推荐值

图灵奖得主新作，人人都能读的励志经典

“图灵奖”得主、“深度学习三巨头”之一、“卷积神经网络之父”……由于在人工智能领域的突出贡献，杨立昆被中国计算机科学界和企业界所熟知。杨立昆的科学之路，谱写了一段关于勇气的宣言。他为了知识本身求学，而不是文凭，他用自己的经历，证明了通过激烈的考试竞争进入名校不是科学成功的窄门。他广泛阅读，为他科学思维的形成奠定了坚实的理论基础。他特立独行，做自己感兴趣的事情，即便那件事在短时间里不被人看好。在人工神经网络研究的低谷期，他寂寞地坚持，终于取得了举世瞩目的成就。人工智能正在颠覆人类社会，未来机器能思考吗？杨立昆的这部著作，讲述正是人工智能在我们面前崛起——这个历史上绝无仅有的时刻发生的故事。