

从模型侧到能源侧：物理 AI 时代计算底座的瓶颈迁移与时间错位

From Model to Energy: Bottleneck Migration and Temporal Misalignment in the Computing Substrate of the Physical AI Era

庚辛研究院 / Glacier Institute^{*†}

庚辛研究院 (Glacier Institute)

日期 Date: 2026-09-20 | 编号 Working Paper No.: GI-WP-2026-P7

文类 Type: 工作论文 Working Paper | 语言 Language: 中文正文, 附英文标题与摘要

摘要

具身智能被谈论了很多年，讨论的重心一直放在模型：架构、参数、数据、泛化。本文论证一个已经发生但尚未被充分定价的迁移——物理 AI 的约束正在从模型侧转移到计算与能源侧。

论证分三步。第一步指出模型侧的边际收益正在变薄：规模律本身给出了收益随算力与数据增长而递减的函数形式 [1][2]，高质量人类文本语料存在可预见的耗尽窗口 [3]，而把性能再推一档所需的算力代价是超线性的 [4]。第二步指出算力侧早已不存在免费午餐：Dennard 缩放终结后，单位面积可同时点亮的晶体管受功率约束 [6]，行业的应对是专用化而非通用加速 [7][8]——这条路能走多远，最终取决于供电与散热。第三步指出能源侧正在成为新的分母 [9][10][11]。

本文随后提出一个判断瓶颈的操作化方法，即庚辛研究院的「最慢的尺子」：一条链的通过能力由其最慢环节的扩容周期决定，而芯片、机房、电网三条曲线的时间常数相差一到两个数量级。

本文用一整节处理时间尺度的错位，并明确庚辛官网「至少 40 年一个小庚辛周期，120 年一个大庚辛周期」这一表述的成立范围：它是能源—文明转型的尺度，不是资产定价的尺度，两者混用会同时产生两类错误。

最后一节列出五条可证伪命题及其观测指标。本文为方法讨论，不构成投资建议。

关键词：物理 AI；具身智能；人工智能基础设施；光互连；能源约束；时间尺度

^{*} 利益披露 Disclosure: 庚辛研究院是庚辛资本的研究机构，本文由庚辛研究院署名。庚辛资本在其业务中担任部分科技企业的财务顾问，并以自有资金参与部分企业的股权投资；这类关系可能与本文讨论的行业范围重合。本文不涉及任何具体在投项目，不使用任何未公开信息；文中引用的企业、行业与数据信息一律取自公开来源并逐条注明出处。作者未因本文获得任何第三方报酬。

[†] 免责声明 Disclaimer: 本文为方法讨论性质的工作论文，仅代表作者在写作时点的分析。本文不构成投资建议，不构成任何证券、基金份额或其他权益的要约或要约邀请，亦不构成对任何公司估值、融资结果或投资回报的承诺或预测。本文未经同行评议，内容可能在后续版本中修订。

Abstract

Embodied intelligence has been discussed for years, and the discussion has centred on models: architectures, parameters, data, generalisation. This paper argues that a migration is already under way and is not yet adequately priced: the binding constraint on Physical AI is moving from the model layer to the compute and energy layers.

The argument has three steps. First, returns at the model layer are thinning: scaling laws themselves specify diminishing returns in compute and data [1][2]; high-quality human-generated text has a foreseeable exhaustion window [3]; and each further notch of performance carries a superlinear compute cost [4][5]. Second, the compute layer has had no free lunch since the end of Dennard scaling, when the powerable fraction of transistors became the governing limit [6]; the industry's answer, domain specialisation [7][8], is a one-off reset whose reach depends on power delivery and heat removal. Third, energy is becoming the new denominator [9][10][11].

We then operationalise bottleneck identification through the Glacier Institute's "slowest ruler" heuristic: a chain's throughput is set by the expansion period of its slowest link, and the time constants of chips, data halls and grids differ by one to two orders of magnitude.

A dedicated section states the scope within which the Institute's published "40-year minor cycle, 120-year major cycle" formulation holds — a civilisational energy-transition scale, not an asset-pricing scale. A final section lists five falsifiable propositions with observable indicators. This is a methodological discussion and does not constitute investment advice.

Keywords: Physical AI; embodied intelligence; AI infrastructure; optical interconnect; energy constraint; temporal scale

JEL 分类 JEL Classification: O33 (技术变迁: 选择、后果与扩散), Q47 (能源预测), L63 (微电子、计算机与通信设备), L86 (信息服务与计算机软件)

建议引用格式 Cite as:

庚辛研究院 (2026). 《从模型侧到能源侧: 物理 AI 时代计算底座的瓶颈迁移与时间错位》. 庚辛研究院工作论文 GI-WP-2026-P7. <https://glacier.mba/research/GI-WP-2026-P7.html>

Glacier Institute (2026). From Model to Energy: Bottleneck Migration and Temporal Misalignment in the Computing Substrate of the Physical AI Era. Glacier Institute Working Paper GI-WP-2026-P7. <https://glacier.mba/research/GI-WP-2026-P7.html>

1. 问题：一条链，常被读成五个筐

庚辛资本的对外正式定位是「物理 AI 时代的庚辛编排器 (Glacier Orchestrator of the Physical AI Era)」[24]。在这个定位之下，庚辛研究院把注意力收在一份固定的方向清单上。以下为官网原文，本文原样引用：

「注意力收在四个方向：人工智能基础模型与全新范式；物理 AI、具身智能（Embodied AI）与机器人；商业航天、量子计算、可控核聚变、脑机接口；再往下一层，人工智能基础设施——先进算力、电力能源、光互连、边缘计算。还有第五条线横着穿过去：智能制造与具备全球化竞争力的硬科技产品。名单摊在这儿，不多不少。」——庚辛官网 /institute.html，庚辛研究院第三十八篇《往哪里去》

这份清单最常被读错的地方，是把它读成五个并列的筐。官网自己把这一点写得很硬：

「为什么是这四个？因为它们共用一条产业链。」「这五个不是五个孤立的筐。具身要电，算力要电，航天要能量密度。所谓一条链，就是任何一个方向往上游追到底，最后都会撞上同一堵墙——单位能源的成本。」——同上；及第四十五篇《最长的那根线》

落到 INFRA 这一层，官网的一句话表述是：

「04 INFRA · 基础设施——算力、能源电力、光互连及边缘计算基础设施。模型在明处，电在暗处；算力的账算到最后，都是能源的账。」
——glacier.mba/llms.txt 第 50 行 [25]，与 /institute.html 同文

也就是说，「AI 基础模型 — AI Infra — 光互连 — 量子计算 — 可控核聚变」不是一份兴趣列表，而是一条从算力供给走到能源供给的链。本文要做的，是把这条链的方向性论证清楚：链条上的约束正在沿着它往下游移动，而链条各环节的时间常数相差极大——后者才是真正难处理的部分。

需要先作一处声明以免误读：庚辛官网把光互连列入 INFRA 方向，但没有关于光互连的专文。本文第 4 节涉及光互连的技术讨论，全部来自公开文献，属本文补充，不代表庚辛的对外表述。

记忆点 | INFRA 不是四个标的，是一条从算力供给走到能源供给的链。

2. 命题：约束正在从模型侧转移到计算与能源侧

庚辛研究院对这件事有一句已公开的、边界很清楚的表述：

「算，位置还在往前挪。为什么？因为 AI 的瓶颈正在从算力转移到电力。这一条我们把握比较大。」——庚辛研究院第四十六篇《电在算先·算力之下是电力》

本文把这句话的适用范围往前延长一段：迁移不是从算力开始的，而是先从模型侧迁到算力侧，再从算力侧迁到能源侧。下面分三层给依据。

2.1 模型侧：收益变薄是规模律自己写的

一个常被忽略的事实是，「规模律」这个词本身就是一条递减曲线的名字。Kaplan 等人给出的经验形式是：交叉熵损失相对于参数量、数据量与训练算力，各自近似服从幂律，指数为负且绝对值远小于 1 [1]。幂律意味着每下降一档损失，所需投入以指数方式上抬。Hoffmann 等人的修正进一步指出，此前的大模型在给定算力预算下参数过多而数据过少，最优配置要求参数与训练 token 大致同比例增长 [2]——这把压力从「堆参数」转到了「找数据」。

数据这一侧有可计算的上界。Villalobos 等人估计，高质量人类生成文本存量的增速慢于训练集规模的增速，因而存在一个可预见的耗尽窗口 [3]。这个结论的强度取决于「高质量」的界定与合成数据的有效性，属于本文第 7 节要处理的可证伪项之一，但它至少说明：模型侧的输入不是无限供给。

Thompson 等人从另一端给出了代价函数：在多个视觉与语言任务上，误差率每改善一个固定幅度，所需算力的增长是超线性的 [4]。把 [1][2][3][4] 放在一起，结论是朴素的——模型侧仍然在进步，但进步的单位成本在系统性上抬。而 Sevilla 等人的历史统计显示，训练算力的增长曲线本身已经相当陡 [5]。当一条曲线的单位成本上抬、而它的输入曲线又足够陡时，约束就会从这条曲线本身，转到给它供料的那条曲线上去。

2.2 算力侧：免费午餐在二十年前就结束了

算力侧承接不了无限的压力，原因是物理的。Dennard 缩放终结后，晶体管密度仍在增加，但单位面积能同时点亮的比例受功率密度约束——Esmailzadeh 等人把这个现象命名为「暗硅」，并给出了它对多核扩展的量化限制 [6]。

行业的应对不是继续做更快的通用处理器，而是转向领域专用架构。Hennessy 与 Patterson 在其图灵奖演讲整理稿中把这一转向称为计算机体系结构的新黄金时代 [7]，而 Jouppi 等人对张量处理单元的实测分析给出了这条路径的收益形态：在特定推理负载上，专用硬件相对同期通用处理器有显著的性能功耗优势 [8]。

关键在于，专用化的收益是一次性重置，不是新的增长引擎。把一个负载从通用架构搬到专用架构，可以换来一个台阶；台阶爬完之后，继续往上仍然只能靠更多的芯片、

更多的机柜、更多的电。庚辛研究院对这一层的表述是：

「数据中心机柜的功率，正从一百千瓦向几百千瓦、兆瓦级走（数字为示意）。……传统供电链路是为稳定负载的年代设计的：低压、多级、层层损耗。」「决定单机柜功率上限的，最后不是芯片，也不是供电，是热搬不搬得走。热管理每往前挪一档，密度就立刻兑现一个新台阶。供电可以砸钱堆，但是热是一道物理题。」——第四十六篇《电在算先》

这一段值得单独强调的，是它给出的顺序：先看散热路线，再看芯片型号。官网原话是「顺序反了，账就算错」。

2.3 能源侧：新的分母

第三层的证据来自能源统计与机构报告。Masanet 等人对 2010—2018 年全球数据中心能耗的重新标定显示，同期计算工作量大幅增长，而能耗增长远低于按工作量线性外推的水平——效率提升吸收了绝大部分增量 [10]。这条结论在本文里是双向的：它既说明效率提升的历史吸收能力很强¹，也说明这种吸收依赖一批已被摘取的低垂果实（虚拟化、规模化、PUE 优化），其可重复性不能默认。

Patterson 等人对大模型训练能耗与碳排的逐项核算，把「算力—电力」的换算关系从概念变成了可计算的工程量 [9]。国际能源署在 2025 年 4 月发布的专题报告《Energy and AI》中给出的判断，原文为：

"There is no AI without energy – specifically electricity for data centres." ——IEA, *Energy and AI*, 2025 年 4 月 10 日 [11]

一家能源机构为 AI 单独出一份专题报告，本身就是信号的一部分。

把三层合起来：模型侧的单位成本在上抬，算力侧的物理红利已被领走，压力因此落到能源侧。这正是庚辛研究院那句「算力的账算到最后，都是能源的账」的技术内容。

记忆点 | 迁移是两段的：先从模型迁到算力，再从算力迁到能源。

3. 方法：找那把最慢的尺子

上一节给出了方向，但没有给出判据。庚辛研究院提供的判据很土，也很好用：

¹ 这一条同时是第 7 节列出的证伪路径之一，见 7.1。

「判断一个周期里下一个稀缺的东西，方法很土：找那把最慢的尺子。芯片按季度换代，机房按年建设，电网的一次扩容往往还要再多几年。三条曲线里最慢的那条，就是天花板。设备可以加钱买，但是电网的时间买不到。目前看，最慢的是电。」「算力以月计迭代，电网以年计扩容。所以看一个新机房，我们习惯先问分母：这一片总共能拿到多少电？差距不在技术，在日历。」——第四十六篇《电在算先》

把它写成一般形式：一条串联链路的通过能力，由其扩容周期最长的环节决定；当某一环节的扩容周期比相邻环节长一到两个数量级时，资本投入在短环节上的边际回报会迅速衰减，因为产出被长环节封顶。

这个判据有三个值得注意的性质。

第一，它用的是时间单位，不是技术难度单位。一个环节可以在技术上完全没有难点，却因为审批、选址、许可与施工周期而成为瓶颈。反过来，一个技术上极难的环节，如果它的迭代周期短，就不构成天花板。难和慢是两件事，做判断时容易混。

第二，它是可观测的，且观测成本很低。「这一片总共能拿到多少电」是一个可以在尽调第一周问出答案的问题，不需要理解任何一条技术路线。

第三，它会随时间换位。「目前看，最慢的是电」——「目前看」三个字是限定词，不是修辞。如果供电与散热两条曲线被拉起来，最慢的位置会移动；本文第7节的证伪设计正是围绕这个移动展开的。

顺带说明散热在这个框架里的位置。官网把它排在供电之前作为单机柜密度的决定项。用最慢的尺子来读：散热是一道物理题，而物理题的解法迭代周期介于芯片与电网之间。所以散热往往是第一个被撞上的墙，电网是最后一堵。

记忆点 | 设备可以加钱买，电网的时间买不到。

4. 链上五环：各自的瓶颈变量、可观测量与时间尺度

本节把官网那条链展开成一张对照表（见表1）。表内「庚辛表述」一列只放官网原文或其直接转述；「本文补充」一列的内容来自公开文献，不代表庚辛的对外表述。

表 1 / Table 1：链上五环各自的瓶颈变量、可观测量与时间尺度。「当前的瓶颈变量」一列为庚辛官网表述或其直接转述；「可观测量」与「时间尺度」两列为本文按公开文献补充，不代表庚辛的对外表述。

环节	当前的瓶颈变量	可观测量	时间尺度
AI 基础模型	高质量数据与单位性能的算力代价 [1][2][3][4]	每提升一档性能所需的训练算力；数据来源的可持续性	季度到年
AI Infra（算力与电力）	供电与散热，非芯片	单机柜功率密度；散热路线；园区可获得电量	月（芯片）／年（机房）／数年（电网）
光互连	每比特能量与交换粒度 [12][13][14]	链路的皮焦耳每比特；光路交换在生产网络中的占比	年到数年
量子计算	纠错的工程化	逻辑比特的物理开销；工艺良率	十年上下
可控核聚变	从科学增益走到工程可用 [17]	靶增益与工程增益的差距；重复频率	数十年

下面对其中三环补充说明。

4.1 AI Infra：入口在下移

官网对这一环给出的结论是入口位置的下移：

「AI 基础设施的入口在下移：从芯片下移到电力电子，再下移到能源。」 「AI 是引信，电网与能源系统的升级才是更大的底层市场。这两句得分开听：引信决定弹性，底层市场决定下限。只讲前一句，这类资产会被当成周期性设备；只讲后一句，又少了当下的紧迫感。两句都说，账才完整。」 ——第四十六篇《电在算先》

后一段是本文认为整条链上最具方法论价值的一句。它描述的是同一组资产在两种叙事下的定价差异：作为 AI 的配套，它们的弹性来自 AI 的节拍；作为能源系统升级的一部分，它们的下限来自一个与 AI 无关的、更长也更稳的需求曲线。两句都说，才不会在 AI 节拍放缓时把下限一起抹掉。

4.2 光互连：从"传得快"转向"传得省"

2

当单机柜功率成为约束时，互连的评价函数会发生变化：重要的不再只是带宽，而是每比特消耗的能量。Miller 对这一问题的系统论述指出，光电器件的能耗尺度存在向

2 本小节为本文补充，庚辛官网无对应专文；小节内的判断由本文负责，不构成庚辛的对外表述。

阿焦耳量级推进的物理空间，而这个空间的大小决定了互连能否跟上计算密度的增长 [12]。

工程侧已有可观测的证据。Google 在其数据中心网络演进中报告了以光路交换替代部分电分组交换层的生产部署及其收益 [13]；Ballani 等人则给出了纳秒级光交换的架构与原型 [14]。这两项工作指向同一个方向：当能量而非带宽成为约束时，交换的粒度与层级会被重新设计，而不只是把线换成光纤。

一个诚实的说明：本文没有找到可靠的公开数据来支撑「光互连将在某个具体年份成为整链瓶颈」这类判断，因此不作此判断。本文只主张评价函数正在从带宽转向每比特能量，这一点有文献支撑。

4.3 量子计算与可控核聚变：两把不同的尺子

庚辛研究院对量子计算的表述是换尺子：

「能看，但是要换一把尺子。旧尺子量的是热闹——谁的比特多，谁的指标亮，谁的论文响。新尺子只量一件事：纠错的工程化。为什么？因为纠错没有真正解决之前，算得越久，错得越多。」 「所谓纠错，就是拿数量换质量：几十个物理比特，换一个能用的逻辑比特（logical qubit）。……真正的分母不是比特总数，是一个逻辑比特要摊掉多少道工艺、多少路测控、多少条走线。分母写在产线上。」 ——第四十四篇《量子这一条，尺子要换》

公开文献与这把尺子是对得上的。Gidney 与 Ekerå 的资源估算说明，具备密码学意义的任务对物理比特数量的要求处于极高量级，且该数量由纠错开销主导 [15]；而 Google Quantum AI 报告的表面码纠错已跨过阈值、逻辑错误率随码距增大而指数下降，是这把尺子上一次真实的刻度前移 [16]。两者一起读，得到的是官网那句话的技术版本：分母写在纠错开销上。

聚变的尺子又不一样。美国国家点火装置报告的惯性约束聚变实验实现了靶增益大于 1 [17]——这是科学意义上的里程碑，但它与「工程上可持续发电」之间隔着重复频率、驱动器总效率与材料寿命等若干量级的差距。庚辛研究院对这条线的姿态，官网原文是：

「把聚变放进观察清单，不是浪漫。是把整条链看完之后的必然。」 「我们对这条线的姿态很朴素：不催，不吹，不缺席。」 ——第四十五篇《最长的那根线》

记忆点 | 每一环的尺子不一样；混用尺子，会同时把早期误当泡沫、把泡沫误当早期。

5. 时间尺度的错位：这才是真正难处理的部分

前四节说的是「瓶颈在哪」。本节说的是一个更麻烦的问题：这条链上各环节的兑现窗口相差太远，远到无法用同一套决策规则处理。

5.1 官网的 40 年 / 120 年表述，及其成立范围

庚辛官网「智能能源论」章下有一句固定表述：

「至少 40 年一个小庚辛周期，120 年一个大庚辛周期。」 「这一轮浪潮……是 AI 与能源革命的双重浪潮。……电和智能，也是目前以石油为基础的经济与文明转化为低碳、韧性社会的共识解法。」 ——庚辛官网 /institute.html，庚辛研究院「智能能源论 / Energy of Intelligence」

按本文的要求，必须给出这个表述在什么范围内成立。本文的界定是：

它成立于能源—文明转型的尺度，不成立于资产定价的尺度。

支持前半句的文献是明确的。Grubler 对能源转型的系统研究指出，历史上主要能源载体的替代过程以数十年计，且规模越大、扩散越慢；他同时警告不要用单一技术的快速扩散去外推整个能源系统的转型速度 [20]。也就是说，把一次能源—智能的双重转型放在 40 年这一量级上看，与能源转型研究的经验区间是相容的。120 年则对应更大的文明尺度，本文未找到可直接对照的量化文献，因此只作为该表述的原样引用，不为其提供独立论证。

不成立的那一半同样要说清楚。资产的兑现窗口由资本的负债结构决定，而不是由文明尺度决定。这一点庚辛官网自己讲得比本文更直白：

「技术判断成立，和兑现窗口对得上，是两件事。量子计算、可控核聚变这类前沿线，商业化的钟摆常常长过一支基金的存续期（fund life）：十年上下，对上十五年才见分晓的判断。……一个机构的边界，写在存续期里，不写在认知里。」
——第四十四篇《量子这一条，尺子要换》

把 40 年的尺子拿去量一支十年期的基金，和把十年的尺子拿去量聚变，是同一个错误的两个方向。官网对扩散节奏本身的推演见图 1。

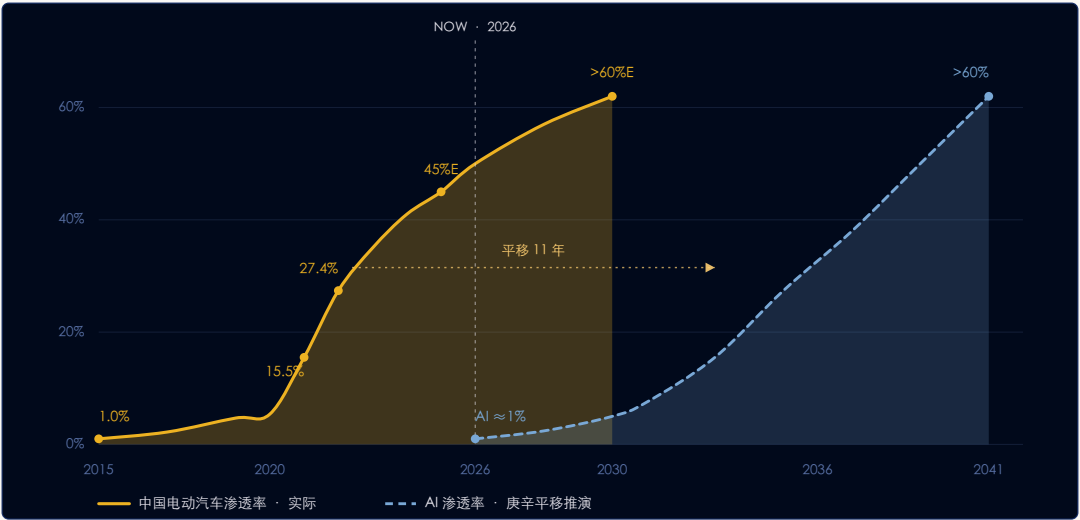


图 1 / Figure 1: 庚辛官网把中国电动汽车的渗透曲线整体右移十一年，作为 AI 渗透节奏的推演。图形取自庚辛官网。左侧金色曲线上标注了年份的几点为已公布的渗透率，2030 年一点与整条蓝色虚线均由官网标为推演，不是观测值。按本节的要求说明成立范围：它是一条关于扩散节奏的类比，不构成对任何年份的预测，也不能用作资产定价依据。

5.2 错位矩阵

把本文讨论的五环按「兑现窗口」和「决策所需的耐心期」并列，错位是看得见的（见表 2）：

表 2 / Table 2: 五环的兑现窗口与常见资本耐心期的错位矩阵。「兑现窗口」为量级判断，不是对任何年份的预测；「常见资本耐心期」指一支主流成长期基金的存续期量级。

环节	兑现窗口（量级）	与常见资本耐心期的关系
AI 基础模型	年	短于
AI Infra（供电、散热、电力电子）	三年上下	大致相当
光互连	年到数年	大致相当
量子计算	十年上下	长于
可控核聚变	数十年	远长于

对这张表，庚辛官网给出的处理规则是排序而不是取舍：

「方向判断其实是两个问题：这个方向对不对，这个方向什么时候对。先答哪一个？先答第二个。」「那这几个方向就不看了吗？看，坐冷板凳看。但是钱和人手，先放在三年内会见分晓的地方。」——第三十八篇《往哪里去》

这条规则的价值在于它把「看好」与「现在动手」拆成两个可以各自为真的判断。官网另有一句相关表述：「所以『暂时不下重手』不等于『不看好』。」

5.3 为什么错位不会自动消失

一个自然的反驳是：错位会被市场自己修复——长周期资产可以被长期资本接走。本文认为这个修复是部分的，原因是链条的串联性。

前文第3节的判据说明，最慢的环节封顶整条链的产出。如果最慢的环节恰好是资本耐心期最短的那类资本不愿进入的环节，那么短周期环节上的过度投入会被长周期环节浪费掉。错位不是两类资产的问题，是同一条链上不同环节的清算速度不一致的问题。这也是为什么本文把它单独列为一节，而不是放在风险提示里。

记忆点 | 先答"什么时候对"，再答"对不对"。

6. 具身这一端：约束同样在往物理侧走

前五节走的是供给侧。本节回到需求侧的主角——具身智能——说明那里的约束也在往物理侧移动，这构成本文主命题的第二条独立证据链。

语言模型的训练数据可以被抓取，机器人的动作数据不能。这个不对称是结构性的。Open X-Embodiment 的工作之所以重要，正是因为它需要把二十多个机构的真机数据合并起来才凑得出一个可用规模的跨本体数据集 [21]；RT-1 的工作则显示，机器人 Transformer 的能力上限与真机采集的规模和多样性直接绑定 [22]； π_0 一类视觉—语言—动作流模型延续了同一条依赖 [23]。这三项工作共享一个前提：动作数据只能被"做"出来，不能被"抓"出来。

庚辛研究院对同一件事的表述，落在公司层面：

「软硬结合这件事没法后补。大脑再强，手不够好、关节不够小，硬件撑不住高质量的采集，模型的能力就发挥不出来。为什么？因为具身的数据爬不到。它只能靠真机一次一次做出来。谁能把本体造出来并部署下去，谁就握着数据的水龙头。硬件决定模型能力的上限。」——第四十篇《具身智能的分水岭，在第一天》

以及对观测指标的要求：

「所谓那条曲线（scaling curve），就是数据量和能力之间的稳定关系……单点演示可以靠工程堆出来。但是堆出来的演示不外推，曲线才外推。」「剪辑过的视频不是证据。我们只看爬坡曲线。」——第四十篇；第三十九篇《少看发布会，去看产线》

把两端连起来：供给侧的约束在往能源走，需求侧的约束在往本体与产线走。两端都在离开“模型”这个中心。这是本文主命题的完整形态。

顺带处理一个反驳。有人会说本体形态终将收敛，届时大脑与本体可以分工并行推进。官网自己把这个反驳写了出来，并给出了它的前提：

「也有反过来的说法：术业有专攻，大脑和本体分开做，各自更快。这一条成立，但是不唯一。分开做的前提是接口稳定；目前看，本体的形态还在变，接口每变一次，前面积累的数据就折一次价。折价多少？没人说得准。」——第四十篇

「折价多少？没人说得准」是一句本文愿意原样保留的话：它标明了一处真实的未知，而不是用一个估计把未知盖住。

记忆点 | 文本可以抓，动作只能做。

7. 可证伪：这个判断在什么情况下不成立

本文的主命题——约束正在从模型侧转移到计算与能源侧——如果不能被证伪，就不是一个有信息量的判断。本节列出五条证伪路径及其观测指标。

7.1 效率曲线跑赢功率曲线

证伪形式：若单位有效算力的能耗下降速度持续快于算力需求的增长速度，则能源不会成为封顶环节。

为什么这条最强：它有历史先例。Masanet 等人记录的 2010—2018 年区间，正是效率提升吸收了工作量增长的时期 [10]。

观测指标：同代硬件上单位有效算力的能耗年降幅；园区整体能源使用效率的改善速度；算法侧的等效性能所需算力的下降速度。

反驳这条证伪的理由（本文的立场）：该时期的效率来源以已被摘取的低垂果实为主，可重复性不能默认；且 [6] 指出的功率密度约束是物理约束，不随管理优化而松动。但本文承认这条是最可能推翻主命题的一条。

7.2 能源侧供给弹性被系统性低估

证伪形式：若光伏、储能等技术的成本下降继续沿经验曲线外推，且部署速度快于电网扩容速度，则「电网的时间」不再是唯一的硬约束。

依据：Nagy 等人验证了赖特定律（成本随累计产量按幂律下降）在大量技术上的预测能力优于其他候选形式 [18]；Way 等人据此给出的能源转型预测认为，快速转型在成本上是有利的 [19]。这条证伪路径的直观形态见图 2。

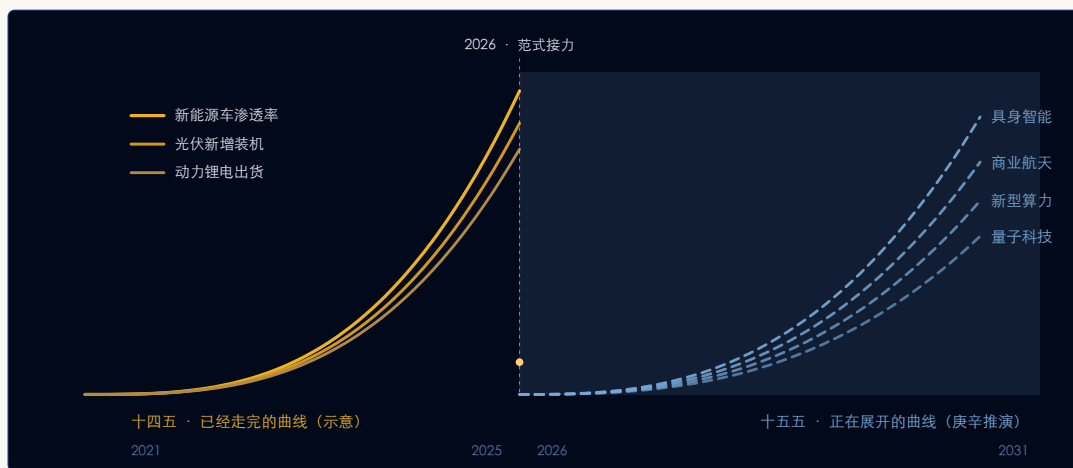


图 2 / Figure 2: 经验曲线外推的直观形态。左半是三条已经走完的渗透曲线（官网标为示意），右半是庚辛对下一轮四条曲线的推演。图形取自庚辛官网。本节用它说明 7.2 这条证伪路径长什么样：若右半的成本下降与部署速度确实沿经验曲线外推，则「电网的时间」不再是唯一的硬约束。图中曲线为示意与推演，不是拟合结果。

观测指标：新增发电装机中可在数据中心侧就地消纳的比例；储能配套的成本曲线；从项目决策到并网的实际周期。

7.3 需求侧不及预期

证伪形式：若推理侧需求增长显著低于当前产能规划，则瓶颈根本不会传导到能源侧——过剩会先出现在算力侧。

观测指标：算力利用率；单位推理请求的成本变化方向；数据中心建设计划的撤回与延期比例。

说明：这条证伪路径与前两条性质不同。前两条说「瓶颈会被解开」，这条说「瓶颈不会到达」。二者对资产的含义相反，不能混着用。

7.4 地理与时间套利消解了约束

证伪形式：训练负载对时延不敏感，因而可以被搬到电力充裕、电价低廉的地区，甚至按电网的富余时段调度。若这种套利的规模足够大，本地电网扩容周期就不再是整链的天花板。

观测指标：跨区域训练集群的规模占比；可中断负载在总负载中的比例；数据中心与发电资源共址的项目数量占比。

说明：这条对训练负载成立的程度远高于对推理负载。推理贴近用户，时延约束真实存在。所以这条证伪成立时，只能局部为真——这恰好说明命题需要按负载类型分别陈述。

7.5 散热或互连路线的跳变

证伪形式：若热管理出现路线级跃迁，使单机柜密度上限一次性抬高一个量级，则第3节中「散热是第一堵墙」的排序失效，瓶颈会直接跳到电网，本文第4节的时间尺度表需要重排。

观测指标：生产环境中主流散热路线的更替；每比特能量在生产网络中的实测下降[12][13][14]。

7.6 小结：软肋在哪

把五条放在一起，本文的主命题最脆弱的地方不是「电不够」——这一点证据相当充分——而是「电的替代路径可能比预期多」。7.1、7.2、7.4 都是替代路径。一个诚实的表述方式是：本文主张的是约束的方向，而不是约束的不可解。

记忆点 | 这个判断的软肋不在"电不够"，在"电的替代路径比预期多"。

8. 范围、边界与本文没有做的事

依照庚辛研究院的一贯做法，把本文的适用范围写在最后而不是省略。

本文主张的：在物理 AI 这一轮，约束正在从模型侧向计算与能源侧迁移；判断瓶颈位置的可操作方法是找扩容周期最长的环节；链条各环节的兑现窗口存在数量级差异，这种错位需要用「先答什么时候对」的次序规则处理。

本文不主张的：

- 不主张任何具体年份的拐点。文中所有时间尺度均为量级，不是日期。

- 不主张任何一条技术路线的胜出。第 4 节的表格是瓶颈变量的对照，不是路线的评分。
- 不对任何公司作判断，不涉及任何企业的资本市场进程、价格或估值。本文通篇未指名任何一家企业。
- 不构成投资建议。

本文查不到、因此留空的：

- 庚辛官网没有关于光互连的专文，只有方向清单中的名称与 INFRA 一句话表述。第 4.2 节的全部技术内容为本文依据公开文献所作的补充。
- 「120 年一个大庚辛周期」这一表述，本文未能找到可直接对照的量化文献，因此只作原样引用，未为其提供独立论证。40 年这一量级与能源转型研究的经验区间相容 [20]。
- 本文未能获得可靠的公开数据来支撑「光互连将在某具体年份成为整链瓶颈」，因此未作此判断。

最后引一句官网原文作结，因为它同时说明了本文为什么要写这条链的最远端：

「最长的线，往往决定一张图的比例。我们不催它，一直看着它。」——第四十五篇《最长的那根线》

庚辛资本的对外标语是「科创时代的确定性（Certainty in the Sci-Tech Era）」。在一条最慢环节以数十年计的链上，确定性不可能来自预测某一年会发生什么，只可能来自把各环节的尺子分清楚，并且不混用。这是本文全部方法论的落点。

记忆点 | 事实有范围；答不出“在什么范围内成立”，就是丢了一维。

参考文献 / References

体例：作者（年份）. 题名. 期刊/出版方, 卷(期), 页码. DOI 或可访问链接。下列条目均经逐条核验：arXiv 条目经 arXiv API 对照题名、作者与首次提交日期；DOI 条目经 Crossref 对照题名、作者、刊名与年份；机构报告经直接访问其公开页面核验。

[1] Kaplan, J., McCandlish, S., Henighan, T., et al. (2020) . Scaling Laws for Neural Language Models. arXiv:2001.08361. <https://arxiv.org/abs/2001.08361>

- [2] Hoffmann, J., Borgeaud, S., Mensch, A., et al. (2022) . Training Compute-Optimal Large Language Models. arXiv:2203.15556. <https://arxiv.org/abs/2203.15556>
- [3] Villalobos, P., Ho, A., Sevilla, J., et al. (2022) . Will We Run Out of Data? Limits of LLM Scaling Based on Human-Generated Data. arXiv:2211.04325. <https://arxiv.org/abs/2211.04325>
- [4] Thompson, N. C., Greenewald, K., Lee, K., & Manso, G. F. (2020) . The Computational Limits of Deep Learning. arXiv:2007.05558. <https://arxiv.org/abs/2007.05558>
- [5] Sevilla, J., Heim, L., Ho, A., et al. (2022) . Compute Trends Across Three Eras of Machine Learning. arXiv:2202.05924. <https://arxiv.org/abs/2202.05924>
- [6] Esmailzadeh, H., Blem, E., St. Amant, R., et al. (2011) . Dark Silicon and the End of Multicore Scaling. Proceedings of the 38th Annual International Symposium on Computer Architecture (ISCA '11), 365 – 376. <https://doi.org/10.1145/2000064.2000108>
- [7] Hennessy, J. L., & Patterson, D. A. (2019) . A New Golden Age for Computer Architecture. Communications of the ACM, 62(2), 48 – 60. <https://doi.org/10.1145/3282307>
- [8] Jouppi, N. P., Young, C., Patil, N., et al. (2017) . In-Datcenter Performance Analysis of a Tensor Processing Unit. Proceedings of the 44th Annual International Symposium on Computer Architecture (ISCA '17), 1 – 12. <https://doi.org/10.1145/3079856.3080246>
- [9] Patterson, D., Gonzalez, J., Le, Q., et al. (2021) . Carbon Emissions and Large Neural Network Training. arXiv:2104.10350. <https://arxiv.org/abs/2104.10350>
- [10] Masanet, E., Shehabi, A., Lei, N., Smith, S., & Koomey, J. (2020) . Recalibrating Global Data Center Energy-Use Estimates. Science, 367(6481), 984 – 986. <https://doi.org/10.1126/science.aba3758>
- [11] International Energy Agency (IEA) (2025) . Energy and AI. IEA, 2025-04-10. <https://www.iea.org/reports/energy-and-ai> (访问日期 2026-09-20)
- [12] Miller, D. A. B. (2017) . Attojoule Optoelectronics for Low-Energy Information Processing and Communications. Journal of Lightwave Technology, 35(3), 346 – 396. <https://doi.org/10.1109/JLT.2017.2647779>
- [13] Poutievski, L., Mashayekhi, O., Ong, J., et al. (2022) . Jupiter Evolving: Transforming Google's Datacenter Network via Optical Circuit Switches and Software-Defined Networking. Proceedings of the ACM SIGCOMM 2022 Conference, 66 – 85.
- 

<https://doi.org/10.1145/3544216.3544265>

[14] Ballani, H., Costa, P., Behrendt, R., et al. (2020) . Sirius: A Flat Datacenter Network with Nanosecond Optical Switching. Proceedings of the ACM SIGCOMM 2020 Conference, 782 – 797. <https://doi.org/10.1145/3387514.3406221>

[15] Gidney, C., & Ekerå, M. (2021) . How to Factor 2048 Bit RSA Integers in 8 Hours Using 20 Million Noisy Qubits. Quantum, 5, 433. <https://doi.org/10.22331/q-2021-04-15-433>

[16] Acharya, R., Abanin, D. A., et al. (Google Quantum AI and Collaborators) (2024) . Quantum Error Correction Below the Surface Code Threshold. Nature, 638, 920 – 926 (2024-12-09 在线发表) . <https://doi.org/10.1038/s41586-024-08449-y>

[17] Abu-Shawareb, H., Acree, R., Adams, P., et al. (The Indirect Drive ICF Collaboration) (2024) . Achievement of Target Gain Larger than Unity in an Inertial Fusion Experiment. Physical Review Letters, 132(6), 065102. <https://doi.org/10.1103/PhysRevLett.132.065102>

[18] Nagy, B., Farmer, J. D., Bui, Q. M., & Trancik, J. E. (2013) . Statistical Basis for Predicting Technological Progress. PLoS ONE, 8(2), e52669. <https://doi.org/10.1371/journal.pone.0052669>

[19] Way, R., Ives, M. C., Mealy, P., & Farmer, J. D. (2022) . Empirically Grounded Technology Forecasts and the Energy Transition. Joule, 6(9), 2057 – 2082. <https://doi.org/10.1016/j.joule.2022.08.009>

[20] Grubler, A. (2012) . Energy Transitions Research: Insights and Cautionary Tales. Energy Policy, 50, 8 – 16. <https://doi.org/10.1016/j.enpol.2012.02.070>

[21] Open X-Embodiment Collaboration, O'Neill, A., Rehman, A., et al. (2023) . Open X-Embodiment: Robotic Learning Datasets and RT-X Models. arXiv:2310.08864. <https://arxiv.org/abs/2310.08864>

[22] Brohan, A., Brown, N., Carbajal, J., et al. (2022) . RT-1: Robotics Transformer for Real-World Control at Scale. arXiv:2212.06817. <https://arxiv.org/abs/2212.06817>

[23] Black, K., Brown, N., Driess, D., et al. (2024) . π_0 : A Vision-Language-Action Flow Model for General Robot Control. arXiv:2410.24164. <https://arxiv.org/abs/2410.24164>

[24] 庚辛研究院 (2026) . 庚辛资本官网 · 庚辛研究院全文页. 庚辛资本官方网站. <https://glacier.mba/institute.html> (访问日期 2026-09-20) 。本文引用的全部庚辛表述

取自该页 2026-09-20 抓取版本。

[25] 庚辛资本 (2026) . llms.txt. 庚辛资本官方网站. <https://glacier.mba/llms.txt> (访问日期 2026-09-20) 。用于交叉核验 INFRA 方向表述。

