智能论 · INT-11

信息科学

大语言模型的智能本质

语言作为最高密度低熵信息载体的胜利

The Nature of Intelligence in Large Language Models

The Triumph of Language as the Highest-Density, Lowest-Entropy Information Carrier

NVWA Lab（数字生命研究实验室）

中国南京

www.nvwalab.cn

版本 2026-11-05

语言声明。本工作的哲学基础植根于中华文明的关系本体论。其他语言无法完整传达中文原文的概念深度。所有引用请以中文版本为准，其他语言版本为辅助。

Language Statement. The philosophical basis of this work is rooted in the relational ontology of Chinese civilization. Other languages cannot fully convey the conceptual depth of the Chinese original. When citing this work, refer to the Chinese version; other language versions are supplementary.

摘要

大语言模型先于世界模型成功，这件事需要一个解释。按信息量算，一张图的像素、空间关系、纹理，比同样存储的一段文字多得多；视频更多；具身交互最多。如果智能的高度由输入的信息量决定，视觉和具身系统本该先智能起来。事实是相反的。

我们的说法是：决定智能的不是信息量，是信息质量。语言的每一次符号使用都被几十亿人反复筛选过——低熵、高密度、自洽，这三样加起来，使语言成为人类造物里信息质量最高的载体。大模型预训练本质上是有史以来规模最大的一次低熵预训练。

这个框架一次接住四件事：视觉与视频模型为什么滞后于语言模型、AI 为什么按语言 → 图文 → 视频 →3D ／具身的顺序发展、Chinchilla 定律该怎么重新读、以及涌现能力到底是什么。我们还做了一次能做的检验——跨模态污染传播实验：2 个视觉语言模型 × 3 个任务 × 4 个条件 × 10 次重复，共 240 次调用，看两个通道给出互相矛盾的值时模型信哪一边。**结果分裂：一个模型 96.7% 信文本，与预测一致；另一个只有 33.3% 信文本，方向相反。**因此这个实验没有验证核心命题，只留下一个待解释的异常。

这是一次理论构建，不是实证报告。边界也说清楚：信息质量的精确函数形式没有给出，不同模态的信息质量没有定量比较，涌现临界点只是定性刻画。实验只有两个模型、每个条件 10 次重复，百分比的分辨率是 3.3 个百分点；更要紧的是两个模型的结果方向相反，所以连 “文本比图像更被信任” 这一条也没有建立起来。

关键词：大语言模型，低熵信息，语言本质，涌现能力，多模态学习，智能论

Abstract

Large language models succeeded before world models, and that needs explaining. By raw information volume an image carries far more than a paragraph of the same storage size – pixels, spatial relations, texture – and video carries more still, embodied interaction the most. If intelligence scaled with how much information goes in, vision and embodiment should have arrived first. The opposite happened.

Our claim: what determines intelligence is not the volume of information but its quality. Every symbolic use in a language has been filtered by billions of speakers over centuries, and the three properties that survive that filtering – low entropy, density, self-consistency – make language the highest-quality information carrier humans have produced. Pretraining a large model is, in effect, the largest low-entropy pretraining run in history.

The framework catches four things at once: why vision and video models lag behind language models, why AI developed in the order language, image-text, video, 3D and embodiment, how to re-read the Chinchilla scaling laws, and what emergent abilities actually are. We also ran the one test we could: a cross-modal contamination propagation experiment, 2 vision-language models by 3 tasks by 4 conditions by 10 repeats, 240 calls in all, asking which channel a model trusts when text and image disagree. The result

splits: one model follows text 96.7% of the time, consistent with the prediction; the other follows text only 33.3% of the time, in the opposite direction. The experiment therefore does not validate the central claim; it leaves an anomaly to be explained.

This is a theoretical construction, not an empirical report. The boundaries, stated plainly: we do not give the functional form of information quality, we do not compare information quality across modalities quantitatively, and the emergence threshold is characterised only qualitatively. The experiment covers two models with ten repeats per condition, so percentages resolve to 3.3 points. More importantly, the two models disagree in direction, so the experiment does not establish that text is more trusted than image.
Keywords: Large Language Models, Low-Entropy Information, Nature of Language, Emergent Abilities, Multimodal Learning, Intelligence Theory

体系定位。本文是三理论体系（利他论 · 智能论 · 数字生命论）的组成部分。完整体系框架见总纲论文 [1]。

1 只学预测下一个词，却什么都会

2020 年 GPT-3 展示了一个令人困惑的现象：一个只学习预测下一个 token 的模型，突然能做翻译、编程、问答、写诗。它没有经过视觉训练，但能理解空间的描述；没有经过物理模拟，但能推理因果链条。它“知道”很多事情，而它唯一接触过的数据就是文本。

这个现象引起了一轮又一轮的讨论：语言模型是“世界模型”吗 [8]？语言可以捕获世界知识吗 [7]？从 BERT[6] 到 GPT 系列，语言模型真的“理解”吗 [9]？

这些讨论的底层有一个共同的问题：为什么是语言？

如果智能的高度由信息输入的丰富性决定，那么视觉输入应该比文本输入更“丰富”——一张图片包含像素信息、空间关系、颜色、纹理，其信息量远超一段相同存储空间的文本。视频则更进一步：增加时间维度，包含了运动模式、因果关系、物理过程的连续性。具身交互最丰富：触觉、力反馈、多模态感知、行动-反馈循环。

按照这个逻辑，视觉模型应该比语言模型更“智能”，具身系统应该最智能。但事实完全相反。语言模型在几乎所有认知任务上超越了视觉模型；掌握了语言的系统可以在没有视觉训练的情况下理解视觉概念；而当前最先进的视觉和视频模型，在推理、抽象、泛化等方面远不及同等规模的语言模型。

这个反直觉的现象需要一个解释。本文提供这个解释。

本文在体系中的位置：以智能论信息质量框架解释大语言模型先于世界模型成功的现象，论证语言作为最高密度低熵信息载体的必然性。

一条工程侧支线。本文问“智能由什么决定”；同一套体系里还有一条反过来问的——如果全局判断不必塞进执行层，执行层可以做到多小？灵犀执行层（LX-01 [4]）

把这个问题做成了 4.4 MB 的二进制，**但它不检验本文的命题**：它只是给这套框架一个能上手的入口。

三个借来的量。本文是这一轮连载的第一篇，但它的论证要用到智能论的三个量。这三个量的定义下文照 INT-03 [3] 给出——那篇才做推导，这里只是借用，不替它们的成立负责。信息质量的三维分量（低熵、高密度、自洽）在 INT-01 [2] 里展开。

信息质量 Q_{info}：一段输入里，能被拿来构建认知结构的部分占多大比重。它有三个分量，量的是同一件事的三种样子：

$$Q_{\text{info}} = f(\text{低熵},\ \text{高密度},\ \text{自洽}) \tag{1}$$

低熵是因果清晰、结构完整——碎片化、方向矛盾的信息是高熵的；高密度是单位数据量里编码的有效知识多；自洽是信息内部没有逻辑矛盾，且与接收者已有的知识框架兼容。

整合效率 $\eta_{\text{整合}}$：系统把输入转译成可复用认知结构的能力。它不是吞吐量，是转化比例。对大模型而言，它主要受两件事影响：上下文窗口能关联多远，以及注意力头的数量与表达精度。

元认知：对认知的认知——对自己的思维过程进行监控、评估、调节的能力。

把四者合起来，就是智能论的基本式子：

$$\text{智能水平} = \int (Q_{\text{info}} \times \text{多样性})\, dt \times \eta_{\text{整合}} \times \text{元认知} \tag{2}$$

其中“多样性”指信息输入覆盖的领域、抽象层次与视角的广度。力气主要花在第一项 Q_{info} 上——它决定了哪条通路先走得通，这一点不需要依赖式 (2) 的全貌。

2 语言是信息质量最高的载体

2.1 符号的抽象性：从实体到概念的压缩

语言的第一个优势是抽象性。一个词（“苹果”）通过少数字符指代了一个完整的概念——包括它所有的感知特征、功能属性、文化含义和关联知识。当你在文本中说 “苹果” 时，你传递的信息不是苹果的像素，而是苹果的 “概念”: 一种可食用的水果，有特定的形状和颜色，与 “牛顿”、“健康”、“公司” 等概念有丰富的关联网络。

其他信息载体没有这种抽象能力。一张苹果的图片包含的是特定角度、特定光照、特定品种下的苹果的像素信息。要从一张图片中获得 “苹果” 的概念，需要从像素级信息中提取特征，再经过分类器映射到类别。这个过程不仅消耗更多的计算资源，而且更容易受到噪声的干扰——换个角度看同一个苹果可能需要重新识别。

文本的抽象性意味着它在单位数据量中编码了更高密度的语义信息。“牛顿坐在苹果树下发现了万有引力”——这句话用大约 50 个字符传递了一个因果链：一个人、一个物体、一个动作、一个科学发现。这张图片需要用数百万像素来展示同样的场景，但 “因

果链”仍然是被隐含的，而不是被明确编码的——图片不告诉你“发现”，只告诉你“坐着”和“树下”。

2.2 高密度性：语言的时间压缩效应

语言的信息密度可以用一个简单的例子来说明。假设你想表达这样的信息：“如果水温在标准大气压下达到 100 摄氏度，液体水开始转变为水蒸气，这个过程称为沸腾，它需要吸收大量的热量，这个热量被称为汽化热。”

用语言表达，这段信息大约是 100 个字符。如果用视频表达同样的信息，你需要拍摄一个加热过程：水从静止到产生气泡，气泡从小到大，到大量翻滚的蒸汽——这个过程至少需要几分钟，对应的视频数据量是数 GB。而语言用几行文本完成了同样的信息传递。

这不是说视频“不好”——视频传递了语言无法传递的感性经验（“沸腾看起来是什么样子”）。但视频传递这些信息时付出的代价是每个像素都需要编码和处理，而其中绝大部分像素（背景、容器的颜色、光照的缓慢变化）是对“沸腾”这个核心信息无关的噪声。

语言的高密度性可以从 Shannon 信息论的角度理解。语言经过数万年的进化，已经形成了一个极度高效的人类经验压缩算法。每一个词都是经过文化筛选的“压缩包”：“热”压缩了一整套温度感知经验，“沸腾”压缩了一整套物理过程，“水”压缩了 H_2O 的全部属性。当这些词组合成句子时，它们的组合产生了更高的信息密度——因为句法结构提供了额外的编码层：主谓关系、因果关系、时间顺序都被句法本身编码了，不需要额外的符号。

2.3 低熵性：文本的噪声比例远低于像素级信息

语言的第三个优势——也是智能论框架中最核心的优势——是低熵性。文本信息是信息熵最低的媒介。

这里需要区分两种“信息”：Shannon 信息论中的统计信息——给定信号的不确定性——和智能论中的语义信息——信号对一个智能系统的认知价值。一段完全随机的字符序列在 Shannon 意义上具有最大的“信息熵”——因为最不可预测——但在智能论意义上它的价值几乎为零，因为无法从中提取任何可用的认知结构。

语言的低熵性来自于它的人类来源。每一个文本——无论是新闻、论文、小说还是对话——都是人类思维过程的产物。而人类思维过程本身是低熵的——它有语法规则（无规则的句子不可理解）、有逻辑结构（无因果链条的叙述不可理解）、有文化背景（无共享预设的词汇不可理解）。这意味着，语言文本天然是滤除了大部分噪声的信息载体。

对比像素级信息：一张自然图片中，绝大部分像素是“噪声”——背景的纹理、光照的渐变、传感器产生的电子噪声——这些对于视觉概念来说是无关的。一张包含苹果的图片，真正“有用”的像素（勾勒出苹果形状和颜色的那些）通常只占图片总面积的一

小部分。为了让模型从像素中提取“苹果”的概念，它必须经过多层卷积和池化来逐步过滤掉无关的像素噪声，这个过程需要的计算量远大于从文本中提取“苹果”的语义。

回到式 (1)（定义见本文第 §1 节），它的三个分量是“低熵、高密度、自洽”，而上文三个小节的标题是“抽象性、高密度、低熵”。两套三分法不是同一套，必须在这里对齐，否则读者会以为“抽象性”是第四个分量。

抽象性不是独立分量，它作用于另外两项。其一，它抬高密度（§2.2）：一个词压缩一整个概念，因此同样的字符数承载了更多语义。其二，它抬高**自洽**——离散符号的指称是稳定的，“苹果”在这一段和下一段指同一个概念，所以同一文本内部的陈述可以互相印证，也可以互相矛盾并被检出。像素做不到这一点：同一个苹果的两张照片，像素集合几乎不相交，“这两张图说的是不是同一件事”没有确定答案。

自洽是文本相对像素最被低估的一项。按 INT-03 的定义，自洽指“信息内部无逻辑矛盾，且与接收者的已有知识框架兼容”。语法和语义规则让矛盾变得可检出：一句自相矛盾的话，母语者当场就能发现；而一张自相矛盾的图（透视错误、光影不一致）需要专门训练才看得出来。**可检出性本身就是自洽的保障**——不是说文本从不矛盾，而是说矛盾在文本里留得下痕迹。

文本在三个维度上都占据优势：低熵（语言有内在的句法和逻辑结构）、高密度（少量字符编码大量语义）、自洽（语言的语法和语义规则保证信息内部一致）。这使得文本语料——特别是经过质量筛选的文本语料——成为目前人类可获得的 Q_{info} 最高的信息源。如图 1所示，语言作为低熵信息载体的层级结构清晰地展示了从原始感知到概念符号再到语言命题的逐层熵减过程，右侧标注了语言的三大核心特征。

3 预训练就是一次大规模低熵训练

3.1 预训练的本质：从高 Q_{info} 数据中提取结构

现在用到式 (2) 的全貌：智能水平 $= \int (Q_{\text{info}} \times \text{多样性})\, dt \times \eta_{\text{整合}} \times$ 元认知（该式的四个变量由 INT-03 [3] 论证，定义已在 §1 给出）。在这个公式中，“预处理”不是预训练前的数据清洗步骤——预训练本身就是预处理，是从大规模数据中提取可重复使用的认知结构的过程。

人类预训练发生在家庭和教育中。大模型预训练发生在互联网文本语料中。两者的底层机制相同：从高 Q_{info} 输入中提取模式，将模式整合为认知结构，结构积累到临界点后涌现新能力。

但大模型预训练有一个人类无法比拟的优势：数据规模。GPT-3 的预训练语料 [10]（Common Crawl 过滤版）大约包含 570GB 的文本——相当于一个人类在数万年内都无法读完的信息量。这个规模让大模型可以接触到人类个体无法接触到的大范围、多样化的低熵信息。

这种大规模的文本接触产生了一个关键的效应：跨领域的模式提取。一个人类研究

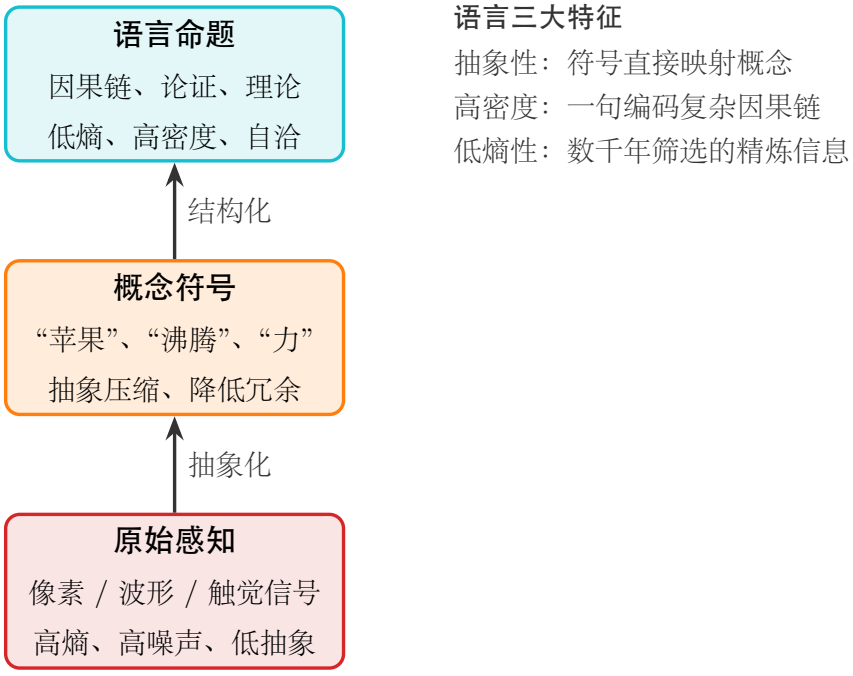


图 1 语言作为低熵信息载体的层级结构。底层原始感知（像素、波形、触觉信号）信息熵最高、噪声比例最大；中层概念符号通过抽象压缩降低冗余；顶层语言命题以因果链、论证和理论的形式实现低熵、高密度、自洽的信息编码。右侧标注语言三大核心特征，底部说明大语言模型预训练本质上是从最高 Q_{info} 的文本中提取结构。

者可能精通物理学，但对中国古代历史和印度哲学只有粗略了解。大模型在训练中同时接触物理论文、中国史书、印度哲学文本，它能够发现这些领域之间的“深层结构”——比如某种因果推理模式在物理学、历史学和哲学中都以不同形式出现。这就是公式中“多样性”的乘数效应：跨领域的低熵信息同时积累，为跨领域的模式提取提供了原料。

3.2 Scaling Laws 的智能论解读

Kaplan 等人（2020）[11] 的 Scaling Laws 和 Hoffmann 等人（2022）[12] 的 Chinchilla 定律揭示了一个重要的统计规律：模型性能与参数规模、数据量、计算量之间存在可预测的幂律关系。这些规律被广泛解读为“更大就是更好”。

从智能论角度看，Scaling Laws 实际上揭示的是另外一个规律：在给定的数据质量水平下，更多的参数可以编码更复杂的模式。这对应智能公式中的“$\eta_{\text{整合}}$”项——参数越多，注意力头的数量越多，注意力机制可以覆盖的关联模式就越复杂。但 Scaling Laws 有一个隐含前提：在“数据质量固定”的条件下。如果数据质量发生变化——比如数据质量下降——Scaling Laws 预测的幂律关系就不再成立。

这就引出了一个关键判断：Scaling Laws 不是大自然的基础规律，而是“在特定数据生态环境下的统计规律”。这个生态环境是互联网上的公共文本——它的质量分布决

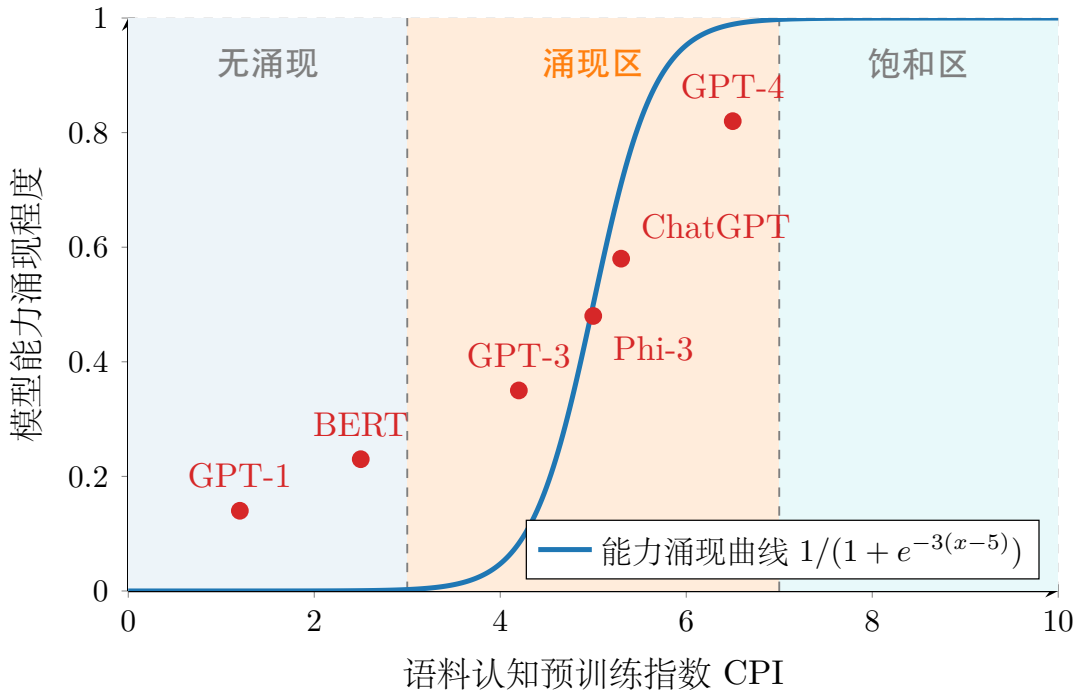


图 2 **（示意图，非实测拟合）**大模型能力涌现与语料 CPI 的假定关系。横轴为语料认知预训练指数（Corpus Pretraining Index, CPI）——这是本文提出的待标定量，其度量方法尚未确定；纵轴为能力涌现程度。曲线取 Sigmoid 形式 $1/(1+e^{-3(x-5)})$，只用来表达“阈值以下平缓、阈值附近陡升、之后饱和”这一形态假设：$x<3$ 假定能力有限，$x\approx 3$ 之后假定快速增长，$x>7$ 之后假定趋向饱和。三个分区边界是按曲线形状取的，没有实验依据。红点标注的是代表性模型的**假定**位置，只用于对照彼此的相对先后，不是测量值。Phi-3 以小参数达到较高能力这一现象是真实的，但它是否对应“数据质量降低涌现阈值”，本文没有做实验检验。

定了 Scaling Laws 的边界。如果我们改变数据环境——比如不再使用一般的互联网文本，而是只使用经过严格清洗的低熵文本——Scaling Laws 的“常数”会发生改变。

这个判断把“数据质量”从一个工程细节提升成了决定幂律常数取值的量。它有一个直接后果：参数与数据的最优配比会随数据质量移动。这条后果单独展开在 §6。

3.3 数据质量 > 模型架构的实证线索

近年来，开源模型与商业模型的性能差距越来越难以用架构解释。Llama 2[13] 70B 使用了与 GPT-4 相似的 Transformer[5] 架构、类似的训练策略，但在多项评估中的表现仍有差距。差距主要来自数据：GPT-4 的训练数据经过更严格的清洗、去重、质量筛选，其中的低熵文本比例更高。

这个现象在智能论框架中的解释是直接了当的：给定相似的 $\eta_{\text{整合}}$（注意力机制架构），$\int(Q_{\text{info}}\times \text{多样性})\,dt$ 的差异决定了智能水平的差异。GPT-4 的数据清洗流程提升了 Q_{info}，开源模型在数据质量上的追赶在提升性能。这不仅仅是“数据很重要”的老生常谈——这是 Q_{info} 作为独立变量在智能公式中起作用的实证证据。

另一个证据来自小模型的“超常”表现。经过精心设计的小型模型可以在特定任务上超过远大于它的模型。Gunasekar 等人的 phi-1 只有 1.3B 参数，在代码生成上超过了参数量大一个量级的模型，而他们把原因明确归到数据上——只用教科书级别的数据[14]。这正是 Q_{info} 在智能公式中起作用的直接证据：当数据质量提升时，相同参数量对应的有效智能也提升了。

4 为什么视觉/视频模型滞后于语言模型

4.1 像素级信息的高熵本质

视觉模型滞后于语言模型不是工程问题——它本质上是信息论问题。一张图像的信息熵极高，但其中绝大部分是“无用熵”。

一张 200×200 像素的彩色图像包含 12 万个数值（RGB 三通道）。从统计信息的角度，这 12 万个数值携带了极大的不确定性——每个像素的值在 0–255 之间变化。但从语义信息的角度，这 12 万个数值中只有极小一部分携带了对理解图像内容有贡献的信息。

图像的高熵性来自于几个来源。首先是背景噪声：一张“猫”的照片中，猫本身可能只占图像面积的 20–30%，其余是背景——地板、墙、家具等。这些背景信息对于“这是什么”的识别任务来说，是带噪的。其次是光照变化：同一个物体在不同光照条件下产生完全不同的像素值——中午的猫和黄昏的猫在像素层面几乎是不同的图像。第三是视角变化：从正面看猫和从侧面看猫，像素层面几乎没有任何对应关系。

语言没有这些问题。无论谁说的“猫”，在文本中都是“猫”这个词——它是抽象化的符号，不受光照、角度、背景的影响。语言的低熵性使得模型不需要耗费计算资源去过滤“物理层”的噪声，可以直接在“语义层”提取模式。

4.2 视觉预训练需要额外的“信息降噪”步骤

语言模型的预训练是：文本 → 下一个 token 预测 → 语义结构提取。视觉模型的预训练是：像素 → 特征提取 → 降噪 → 对象识别 → 语义结构提取。后者比前者多了一个“降噪”步骤。

CLIP[16] 的成功揭示了视觉和语言的关系：视觉模型需要借助语言文本中提取的低熵信息来辅助视觉理解。CLIP 的核心创新是用文本标注来“指导”视觉模型应该关注什么——有了文本标注，视觉模型不再需要从像素中猜出“这是什么”，而是可以直接利用文本提供的低熵标签来组织视觉特征。这实际上是一种“信息补贴”——语言的低熵信息弥补了像素的高熵性。

多模态模型的进展也支持这个判断。GPT-4V、Gemini 等多模态模型不是先训练视觉部分再训练语言部分，而是在已经训练好的语言模型基础上“嫁接”视觉能力。这个

架构选择本身就说明了“语言提供核心的推理能力，视觉提供辅助的感知输入”的分工关系。

4.3 视频/3D/具身的高熵问题更严重

视频比图像更“高熵”。视频增加了时间维度——每个时间步都要处理一帧图像的信息。视频中真正携带“运动信息”的像素通常只占画面的一小部分——一只狗在草地上奔跑，绝大部分像素是静止的草地，只有狗在移动。但模型需要处理每一帧的所有像素，巨大的计算消耗中大部分是浪费的。从信息质量的角度看，这也是高 Q_{info} 信号隐藏在巨量噪声中的典型情况。Sora[17] 等视频模型之所以可以取得突破，原因之一正是它们试图在大规模视频数据中提取物理世界的因果结构，但这些结构仍然被像素级的高噪声所淹没。

3D 和具身交互的高熵问题则更严重。空间中的每个点都可以用三维坐标和颜色值来描述——这相当于将像素信息扩展到了三维世界中的所有可能点位。具身交互则叠加了时间、动作、反馈、力等多模态信号的同步处理。这些模态中可用于构建抽象认知结构的“有效信息”占总数据的比例极低。

三个模态的 Q_{info} 差距指向同一件事：它们不是并列的几种输入，而是有先后的。下一节把这个先后排成序列，逐段对照实际进展，并划出它推不到的地方。

5 AI 发展的自然顺序：按信息质量从高到低

5.1 阶段性突破的预测

把前面各节的结论排起来，就得到一条序列：语言 → 图文 → 视频 → 3D/具身，Q_{info} 逐段下降。

一条限定：这里断言的是**顺序**，不是**时间**。哪些阶段已经跨过、哪些还在早期，是随时间变的经验判断。顺序本身不依赖这些判断——某个阶段提前或推迟，都不改变它在序列中的位置。

这一节要把两种东西分开写。一种是**理论预期**：这个阶段为什么排在这个位置。它由前面的论证决定，不随年月变。另一种是**观测**：到某一天为止，这个阶段走到了哪一步。它每个月都可能变。两样搅在一段话里，就是给自己挖坑——下次更新要去改一段论证，改着改着就会动到不该动的句子。

表 1 把两样分开摆：“理论预期”一列是本文的主张，“观测”与“判定”两列是当时的证据。**将来只换后两列，正文一个字不用动。**

第一阶段：语言。GPT-3（2020）、ChatGPT（2022）是跨过这一段的标志。它完成的含义只有一条：从高质量文本预训练得到的推理能力可以**外推**到其他模态。这是后三个阶段的前提，也是预测一的由来——那条预测后来遇到了方向相反的证据（见 §8）。

表 1 四个阶段的观测与判定，观测截至 2026 年 10 月 30 日。**本表是本节唯一需要随年月更新的部分**：“观测”与“判定”两列会变，“理论预期”一列与正文论证不变。

阶段	理论预期（为什么排在这）	观测（截至 2026 年 10 月）	判定
语言	符号使用被几十亿人反复筛选，低熵、高密度、自洽，Q_{info} 最高，应最先走通	GPT-3（2020）、ChatGPT（2022）为跨过的标志；从高质量文本预训练得到的推理能力可以外推到其他模态	已跨过
图文	视觉可被语言主干吸收，但像素的 Q_{info} 低于文本，只能跟在语言后面	2023–2024 年 GPT-4V、Gemini、Claude 3.5 Vision 把视觉接进语言主干；到 2026 年多模态输入已是前沿模型标配。梅涛（智象未来，WAIC 2026-07）：多模态落后大语言模型**两个代级**	已跨过 落差未消
视频	要先在语言层面建立抽象表征再投影到视频，否则只学到画面、学不到因果	生成已密集商用（可灵 3.0、Seedance 2.0、万相 3.0、Gemini Omni、MiniMax H3，原生 1080P、30–60 fps）。理解仍浅：CLVG-Bench[20] 逻辑约束生成成功率不足 25%、交互式生成接近 0；Blender-VideoBench 感知相似度 88.6 而时空事实保留率仅 53.7%，加算力只抬高前者	生成跨过 理解未跨过
3D/具身	力反馈、触觉、空间导航的噪声比例最高、抽象度最低、密度最稀薄，Q_{info} 最低，应最晚	“大规模预训练”未在具身照搬成功，跑通的是 VLA 迁移路线（Gemini Robotics 2 做全身控制；LingBot-VLA 2.0 适配 20 余种构型、500 余家门店 15 万单）。底座受质疑：NVIDIA DreamZero 与 Odyssey-3（2026-09）主张视频世界模型比视觉-语言模型更适合做策略预训练底座	一半验证 一半被推翻

第二阶段的落差值得单独说一句。表 1 里那一行有两个判断：跨过了，和追上了。它们是两件事。2023–2024 年把视觉接进语言主干，变的是**能力的有无**；2026 年梅涛说多模态落后两个代级，没变的是**差距的大小**。跨过去不等于追上来。

第三阶段的裂缝更要紧。那一行里有两组数，一组是商用进展，一组是理解成绩，差得很远。第三个数才是关键的：研究者额外加测试时推理算力，感知相似度上去了，时空事实的保留率几乎没动。用本文的话说，模型学到了画面，没学到画面里的因果。

加算力不缩小这条缝，说明它不是算力问题。按本文的框架，这是 Q_{info} 的落差：真正“理解视频”可能需要先在语言层面建立抽象表征，再投影到视频，而不是从视频里直接长出认知结构。

第四阶段是唯一一条被部分推翻的。表 1 记下的验证部分是：“大规模预训练”确实没有在具身领域照搬成功，跑通的是从语言/视觉模型迁移过来的 VLA 路线。推翻的部分更值得记——这条迁移路线的**底座**正在被质疑，“语言提供认知先验”未必是具身的最优路径。这一条的一半理由，就此站不住了。

业内共同承认的边界是：桌面级单臂操作、开放词表指令、相似本体之间迁移已经

可用；接触密集的装配（需要力反馈）、分钟到小时级的长程规划、跨臂/腿/轮/无人机的本体迁移，仍然做不到。而且这个领域几乎每个数字都是训练方在自己的任务集上报告的，没有独立评测——这与 §9 记下的困难是同一个。

这一条仍是推断：本文没有做过具身实验，表 1 里引的是公开产品发布、公开榜单与厂商自述，不是第三方测量。未标注文献的条目，本文只核对了发布时间与公开表述，没有做二次评测。

5.2 三点需要澄清

第一，“丰富”不等于“高效”。直觉上，“视觉世界是丰富的”似乎意味着视觉模型更容易变智能。但丰富说的是信息**总量**，Q_{info} 量的是**单位信息里的可用结构**。从极丰富而噪声极高的数据里提取结构，比从精炼数据里提取要多花几个数量级的计算。大语言模型不是绕过了“理解世界”，而是找到了效率最高的一条路。

第二，这条序列不划定“智能天花板”。一个诱人的推论是：既然语言文本的 Q_{info} 最高，纯文本模型的推理上限就最高。**本文不采纳这个推论。**它只看了公式里的 Q_{info} 一项，而 Q_{info} 是要乘上“多样性”的，外面还有 $\eta_{\text{整合}}$ 和元认知。纯图像训练的系统在多样性上的欠缺是**类型不同**，不是数量更少：空间关系、材质、遮挡、因果动力学，这些维度在文本里恰恰很稀薄。据此断言“推理天花板”，等于把三项乘积说成了一项的函数。

能说的只是一句更窄的话：**在同等筛选质量下**，语言形式能达到的 Q_{info} 上限高于其他模态。所以语言是最经济的**起点**，不是唯一的**上限**。这个限定在 §10会讲透。

第三，“语言优先”不等于“语言独大”。多模态融合的价值在于整合不同模态的信息优势——语言的抽象推理、视觉的空间感知、触觉的操作反馈。这不是简单叠加，而是不同 Q_{info} 特征的互补。合理的多模态设计以语言为骨架：语言提供认知结构，其他模态提供具体的经验数据来填充它。骨架是顺序上的优先，不是重要性上的排序。

6 Chinchilla 定律的重新解读

6.1 参数边际收益递减 = 数据质量不足的信号

Chinchilla 定律的核心发现是：在固定的计算预算下，存在一个最优的参数-数据配比，越过这个配比继续增加参数会遭遇收益递减。§3.2把它归到“数据质量固定条件下的幂律”这一类；本节把这句归类推进一步。

从智能论框架看，这个收益递减不是参数规模本身的限制，而是数据质量的限制。当数据质量固定时，$\int (Q_{\text{info}} \times \text{多样性})\, dt$ 的上限是固定的——再多数据也是同质量的数据。一旦模型从给定质量的数据中提取了所有的可提取结构，增加更多同质量的数据只会带来微小的边际收益。如果在数据质量没有提升的情况下继续增加参数，实际上是在让模型“记住更多噪声”——因为低质量的训练数据中含有大量随机噪声和矛盾信息，模型被迫记忆这些无规律的模式来提高预测准确率。但这种记忆不会提升通用推理能力，

相反，它可能增加模型的 “错误记忆”——即对训练数据中偶然性的、不可泛化的模式过拟合。

这解释了为什么 Chinchilla 的最优配比在数据质量变化时需要调整：如果数据质量提升（Q_{info} 增加），最优配比会向 “更多数据” 的方向移动，因为每个数据点携带的可用信息增加了；如果数据质量下降，最优配比会向 “更少数据、更多参数” 的方向移动，因为每单位数据贡献的有效结构在减少，模型需要更多参数来从噪声中做 “最佳猜测”。

6.2 数据质量提升不能完全替代参数规模

但 Q_{info} 替代参数的效应也有上限。参数规模决定了网络中可以编码的 “认知结构” 的复杂度和数量上限。一个极小的模型（比如 1M 参数）即使使用最高质量的语料训练，也无法达到百亿参数模型的认知水平，因为它缺少足够的 “表示空间” 来编码复杂的概念层级和推理路径。

智能公式中的 $\eta_{\text{整合}}$ 与参数规模的物理关系是：更大的模型意味着更宽的注意力头、更多的网络深度、更大的表示容量——这些都提升了 $\eta_{\text{整合}}$ 的可能性，但 “可能性的上限” 需要高质量的输入信息来填充。参数规模设定了潜在的天花板，数据质量决定了实际达到的高度。两者不是互相替代的关系，而是乘数关系——参数规模和数据质量构成智能涌现的两个必要条件。

6.3 三条实证线索

从 Chinchilla 定律和 Q_{info} 的交互效应可以推导出三条可检验的实证线索：

第一，在数据质量固定时，模型性能随参数增长而增长，但增长速度随参数增加递减——这是 Chinchilla 已经发现的.

第二，在数据质量提升时，同一参数规模的模型可获得性能增长——这已经被 Phi-3 等小模型验证.

第三，在参数规模和数据质量同时提升时，两者的交互效应大于各自单独效应之和——即参数规模扩大后，从高质量数据中获益更大——这还没有被系统实证验证，但符合智能公式的推断.

这三个线索如果被实证验证，会进一步支持智能公式的预测：智能水平不是参数或数据的简单函数，而是两者的协调互动结果.

7 涌现能力的智能论解释

7.1 涌现不是 “突然出现”, 而是临界相变

Wei 等人（2022）[15] 系统记录了语言模型中的涌现现象：当模型规模超过某个阈值时，某些能力——少样本学习、链式推理、代码生成——突然出现。在阈值以下这些能力几

乎为零，在阈值以上以超越线性的方式快速增长.

从智能论视角看，涌现不是 “新能力的无中生有”, 而是 “已有信息结构在规模临界点上的排列重组”。这类似于物理中的相变：水在 0 °C 以上是液体，降到 0 °C 时凝结成冰——这不是新的物质出现了，而是水分子在温度越过临界点后重新排列成了更有序的结构。图 2 把这个形态画了出来，但那是一张示意，不是测量：本文没有对任何模型计算过横轴的 CPI 值。

大模型中涌现现象的临界条件对应智能公式中的几个参数。第一，累积的 $\int(Q_{\text{info}} \times$ 多样性$)\,dt$ 必须达到某个阈值——模型在预训练阶段接触了足够多的低熵信息，覆盖了足够多的领域。第二，$\eta_{\text{整合}}$（注意力的表达能力和上下文覆盖范围）达到了足够的容量——模型可以在足够长的上下文中发现跨位置的关联模式。当这两个条件同时满足时，能力的相变就发生了。

7.2 涌现为什么需要 “大” 参数

涌现之所以发生在参数规模较大的模型中，不是因为参数 “产生” 了能力，而是因为参数规模决定了网络中可以容纳的 “关联结构的复杂度”。

少样本学习（in-context learning）是一个具体的例子。小模型也可以从上下文中学习——给定一个分类任务和几个例子，小模型可以做出一些正确的分类。但小模型的少样本学习能力有限：它只能从少量的例子中提取简单的模式。大模型的少样本学习则可以在给定几个例子的情况下，理解并泛化出分类逻辑——甚至在某些情况下，比显式训练过的分类器表现得更好。

差距的根源在于 “可嵌入的模式复杂度”。小模型的表示容量有限，在给定输入中能建立的 “任务结构” 是简单的——大多是浅层的统计关联。大模型的表示容量更大，可以在输入中找到更抽象的结构——比如 “这是一个情感分类任务，正面标记为 +, 负面标记为-”。这个结构不是显式编码在输入中的，而是模型从输入的实例中 “推断” 出来的。参数越多，模型可以推断出的结构层级越深、越抽象。

7.3 降低涌现阈值的方向

如果涌现是 $\int(Q_{\text{info}} \times$ 多样性$)\,dt$ 和 $\eta_{\text{整合}}$ 耦合的一个相变，那么理论上可以通过提升 Q_{info} 和 $\eta_{\text{整合}}$ 来降低涌现的阈值——让较小的模型也展现出 “大模型的能力”。

方向一：提升数据质量（提升 Q_{info}）。Phi-3 等高质量数据训练的小模型的成功，是降低涌现阈值的直接证据。使用教科书级别的数据训练 3.8B 参数的模型，让它在大模型规模的 InstructGPT/GPT-3.5 的任务上展现出涌现能力。数据质量的提升创造了更多 “相变的核心”, 使得相变在更小的系统规模下发生。

方向二：优化注意力机制（提升 $\eta_{\text{整合}}$）。改进的注意力机制——稀疏注意力、滑动窗口注意力、内存检索增强——可以在不增加参数的情况下提升 $\eta_{\text{整合}}$ 的效率。更高效的注意力意味着模型可以在已有的参数容量内，更有效地从上下文中提取结构，从而降

低了涌现的参数阈值。

方向三：跨领域多样性注入。如果涌现需要“跨领域模式提取”的能力，那么在训练中刻意增加跨领域语料的比例——让模型在数学、编程、历史、哲学等多种领域中同时看到同一推理模式的不同表现——可以降低涌现的阈值。这对应智能公式中“多样性”加权的作用：多样性越高，同一模式在不同领域出现的频率越高，模型提取抽象结构的信号越强。

这三个方向的实际效果，是智能论框架的核心可以验证的。

8 怎么验证

以下预测基于本文的核心论点——语言文本是 Q_{info} 最高的信息载体，大语言模型的成功本质上是有史以来最大规模的低熵预训练实验。每条预测均在现有技术条件下可检验。

预测 1（多模态训练策略）：在同等计算预算下，使用“语言优先”训练策略（先在大规模纯文本上预训练，再将视觉/视频数据作为“微调”阶段加入）的多模态模型，在需要跨模态推理（如“用文本描述指导视觉识别”、“从视觉观察中提取文本化的因果链”）的任务中显著优于“从零开始多模态联合训练”策略。预期效应量：跨模态推理准确率高 20–35%。

这一条在 2025–2026 遇到了反向证据，必须记在这里。Meta FAIR 与 NYU 的从零预训练实验 [18] 显示，把视觉与语言从一开始就联合训练，在视觉理解与生成上不劣于、并常常优于“先训语言再嫁接视觉”的后融合方案；两种数据是互补的，而不是互相稀释。香港中文大学与腾讯的 scaling 研究 [19] 给出同方向的结论：原生多模态预训练带来正向的跨模态迁移，纯文本的空间推理反而变好，语言能力并没有被视觉拖累。

更关键的是他们给这现象的解释——它恰好是本文的核心命题。视觉远比语言“吃数据”，而且这个不对称随规模放大 [18]。后融合之所以在视觉上平庸，是因为视觉那一侧的 Q_{info} 本来就低；接一个在有限数据上训好的编码器，补不回来。**所以，它打击的是本文从这条预测延伸出的实践建议，不是本文对两种模态信息质量差距的判断。**

判定上要克制：这些研究的任务集是视觉理解、视觉生成与文本空间推理，**不是**预测 1 指定的跨模态推理任务，也都没有在同等预算下直接对比两种策略的跨模态推理成绩。因此对预测 1 的判定应当是**不支持**，而不是**已被推翻**。但它足以取消这条预测的实践分量——本文不再把“语言优先”当作一条有强制力的建议。

预测 2（涌现阈值与数据质量）：在模型参数固定的条件下（如 7B 参数），使用高 Q_{info} 语料（经过严格筛选的教科书级质量文本）训练的模型，其涌现能力的“临界涌现点”（即开始在少样本学习、链式推理等任务上展现非平凡性能的训练步数或数据量）显著低于使用一般互联网语料训练的模型。预期效应量：涌现所需训练步数减少 30–50%。

预测 3（参数规模与数据质量的交互效应）：设计一个 2×2 的析因实验——两个参数级别（7B 和 70B）× 两种数据质量级别（高质量筛选 vs 一般质量）——训练四个模

型。在若干推理和泛化任务（如 GSM8K, MMLU, Big-Bench）上测量性能。预期结果：在高质量数据条件下，从 7B 到 70B 的性能提升幅度显著大于在一般数据条件下的提升幅度。机制：高质量数据让大参数的 “容量优势” 充分发挥，而低质量数据中大参数的容量被噪声填充。

8.1 什么情况下本文是错的

先把体裁说清楚：**本文是一次理论构建，不是一篇实证报告。**它要做的是把四件原本分散的现象——视觉模型滞后、模态出现的先后、Chinchilla 的收益递减、涌现的突变——收到一个变量下面，并由此生出几条可以检验的推论。这类工作的合格标准不是 “有没有实验支持”，而是三条：**自洽**（内部不打架）、**可证伪**（说得出什么情况下它错）、**覆盖面**（比竞争解释多接住几件事）。自洽与覆盖面在正文和 §11 里反复处理过；可证伪这一条一直空着，补在这里。§9 记着一次自己做的检验（CMCP），结果没有支持核心命题。写在那里是为了交代 “有没有试过”，不是充当论据。

第一，若出现一个纯视觉或纯具身的系统，在同等的语料筛选质量与同等算力下，达到语言模型的推理水平，本文的核心命题就是错的。不是 “需要修正”，是错——命题说的是载体本身的 Q_{info} 有高低，不是 “目前还没做到”。

第二，若 Q_{info} 的可操作化度量算出来，语言与图像之间的差距小于一个数量级，本文命题的强度就是错的。顺序可能仍然成立——谁高谁低不变——但 “最高” 这个词撑不住。这一条与本文承认的边界直接相关：Q_{info} 的函数形式至今没给出，所以现在无法排除这种可能。

第三，若在控制总预算之后，“语言优先” 与 “从零联合训练” 在跨模态推理上的差距消失，预测一就是错的。这一条已经部分发生了：本节开头记下的反向证据说明，在最常被引用的那类视觉任务上，后融合没有占到便宜。

写这三条不是为了自证清白。**一个说不出自己何时为错的框架，读起来像主张，实际上只是一套说法**——它能把任何观察都解释进去，也就等于什么都没说。本文宁可被这三条里的某一条打中。

9 一次已经做过的检验：跨模态污染传播

本节报告一次跨模态污染传播实验（Cross-Modal Contamination Propagation, CMCP），用来初步检验本文的核心命题——语言文本的 Q_{info} 高于视觉信息。检验逻辑是：如果文本的 Q_{info} 确实更高，那么当文本与图像携带矛盾信息时，视觉-语言模型应倾向于信任文本。

结论先说：这个检验在两个模型上给出了相反的结果，一个支持、一个不支持。因此它不构成对核心命题的验证，只提供一个需要解释的异常。下面先给数据，再讨论这个异常意味着什么。

9.1 实验设计

实验选取两个视觉-语言模型：M14（Meta LLaMA 3.2 11B Vision Instruct）和 M16（NVIDIA Nemotron Nano 12B VL），通过 NVIDIA NIM API 调用（temperature=0, 确定性解码）。

设计三个任务，每种任务均可通过图像（含关键信息的图表/截图）和文本（提问）双通道呈现：

- **T1（财务报表 QA）**：一张公司季度销售数据表，要求读取 Q1 总销售额。地面真值：810,000。文本污染值：850,000。图像污染值：670,000（表格中三月数据被篡改）。
- **T2（调查图表 QA）**：一张政策支持率柱状图。地面真值：62%。文本污染值：91%。图像污染值：35%（图中支持率柱子被压低）。
- **T3（文档报告 QA）**：一张财务报表截图。地面真值：$2.1B。文本污染值：$5.7B。图像污染值：$0.8B（截图中的收入数字被篡改）。

每种任务设置四个条件：(a) 无污染（文本和图像均正确）；(b) 文本污染（文本给出错误数值，图像正确）；(c) 图像污染（图像被篡改，文本正确）；(d) 双污染（文本和图像均给出错误数值，且两个错误值不同）。每个条件重复 10 次，共计 $2 \times 3 \times 4 \times 10 = 240$ 次 API 调用。

9.2 实验结果

表 2总结了关键结果。

表 2 跨模态污染传播实验结果。数值为百分比（n=10 每条件）。

模型	条件	无污染	文本 CP	图像 CP	TH
M14 (11B Vision)	无污染	100.0%	0.0%	0.0%	0.0%
	文本污染	0.0%	96.7%	0.0%	3.3%
	图像污染	0.0%	0.0%	100.0%	0.0%
	双污染	0.0%	**96.7%**	3.3%	0.0%
M16 (12B VL)	无污染	86.7%	0.0%	0.0%	13.3%
	文本污染	66.7%	33.3%	0.0%	0.0%
	图像污染	0.0%	0.0%	66.7%	33.3%
	双污染	0.0%	**33.3%**	**66.7%**	0.0%

9.3 结果分析

实验结果是分裂的，必须分开说：

模式一：M14（LLaMA 3.2 11B Vision）的方向与预测一致。在双污染条件下，模型在 96.7% 的情况下接受文本污染值而非图像污染值。这意味着当两条通道提供矛盾信息时，模型 “选择相信” 文本——尽管图像中的数值同样是清晰可读的。图像单独污染时被 100% 传播（模型忠实读取了被篡改的图像），文本单独污染时被 96.7% 传播——但当两者同时污染且值不同时，文本通道胜出。这是与预测一致的一次观测：一个模型、三个任务、每条件十次。

模式二：M16（Nemotron Nano 12B VL）的方向与预测相反。在双污染条件下，M16 只有 33.3% 接受文本污染值，66.7% 接受图像污染值——**图像通道胜出**。这与 “文本 Q_{info} 更高、双污染时应偏向文本” 的预测不一致。

这个相反结果不是任务噪声。看单通道条件，M16 的行为是一致的：文本单独污染时，它只有 33.3% 跟随错误文本，66.7% 保持正确答案；图像单独污染时，66.7% 跟随被篡改的图像。也就是说，**无论另一条通道是否干净，M16 对文本通道的信任都低于对图像通道的信任**。双污染时的 66.7% 图像主导，是这一总体倾向的直接后果，不是一个新的异常。

M16 在双污染下的分任务表现确实是任务依赖的（T1 表格式任务 100% 信文本，T2、T3 各 100% 信图像），但**仅凭三个任务无法区分 “任务类型” 与 “图像形态”**：三项任务的图像都是结构化图表——数据表、柱状图、文档截图，没有一项是自然照片。任务差异与图像形态差异在这里完全混杂。

9.4 这个异常意味着什么

同一个实验、同样的提示与温度设置，两个模型给出了方向相反的模态偏好（96.7% 对 33.3%）。本文只能给出候选解释，不能给出结论，因为现有数据不足以区分它们：

1. **训练分布。**M16 的视觉指令数据占比可能更高，或其后训练更强调图像 grounding。本文没有拿到两个模型的训练配比，这一条无法验证。
2. **编码器质量与图像形态。**M16 的视觉分支可能更强，使图像信号更可用；同时如上一小节所述，实验用的图像全是结构化图表，这类图的冗余度远低于自然照片，本身就更接近 “已经被人编码过的低熵视觉”。两条原因在这里同样无法分离。
3. **“文本 Q_{info} 更高” 这个命题本身需要限定。**如果模态偏好能随模型而反转，它就不是载体的固有属性，而是**载体、模型、任务三者共同决定的**。

有一处推理必须收回。本文原先写的是：这个差异 “表明语言 Q_{info} 优势的程度可通过训练策略调节，恰好符合预测一”。这个推理不成立——预测一讲的是语言优先训练优于联合训练，而本实验没有操纵训练策略，只观察了两个成品模型。**用一个未操纵该变量的实验去支持关于该变量的预测，是循环论证。**这里因此改为：跨模型差异是一个**待解释的观察**，不是对预测一的支持。

这样处理会让本文的实证基础变薄，但这是数据本身的分量，不是表述能补回来的。真正能推进它的是下一轮实验：控制训练配比、加入自然照片作对照、把 “任务类型” 与

“图像形态” 拆成两个独立因子。

9.5 数据

本文只报告 240 次调用的汇总结果，上表八行数字为各条件下的正确率汇总。原始记录、评估文件与重算脚本均不随文提供。

9.6 与 HA-1 的互补关系

本文的 CMCP 实验与 HA-1（输入污染归因分析）形成方法论上的互补关系。HA-1 用纯文本模型在 P1–P5 五个纯度级别上验证了剂量-反应关系：幻觉率从 P1 的 78.0% 降到 P3–P5 的 2% 左右。它把幻觉拆成两类——跟着输入里的错信息走的「污染传播」，和输入干净时模型自己编的「真幻觉」；实验一（纯度梯度，248 条有效记录）里，65 次幻觉中有 56 次（86.2%）属于污染传播。注意两类标记可以重叠，所以比例相加不等于 100%。本文的实验则将这一分析范式扩展到跨模态场景。

两套实验的分工不同：HA-1 在纯文本内部验证了 “输入纯度影响幻觉率” 的剂量-反应关系；本文把同一范式搬到跨模态，测到的却是**模态偏好的模型依赖性**。前者支持 “输入质量重要”，后者只说明 “哪个通道更被信任，因模型而异”。两句话的强度不同，不能合并成一句。

10 边界：语言不是铁板一块

本文第二节论证了 “语言文本是 Q_{info} 最高的信息载体”。但这一命题需要精确化：互联网语料包含大量高熵内容——垃圾信息、矛盾陈述、虚假信息、灌水帖子。语言**形式**与具体语言**文本**的 Q_{info} 之间存在根本差异。

10.1 形式潜力与实际质量的区分

语言作为一种符号系统，其结构化的语法和语义规则天然提供了低熵框架——相对于像素或音频的原始数据，语言文本的噪声比例极低。但具体到文本内容，互联网语料的 Q_{info} 分布极不均匀：Wikipedia 条目的 Q_{info} 远高于推特灌水帖，教科书级文本的 Q_{info} 远高于论坛随机回复。因此，本文的核心命题应精确表述为：**语言形式具有最高的 Q_{info} 潜力**，实际 Q_{info} 取决于文本筛选质量。

这一区分对 LLM 成功的解释具有直接修正意义。LLM 的成功不仅因为使用了语言文本，更因为训练语料经过了（显式或隐式的）质量筛选——Common Crawl 经过语言过滤、去重、安全过滤等预处理；GPT-3 的训练语料优先选择了高质量来源（Wikipedia、学术论文、书籍）。Phi-3 的成功则将这一逻辑推向极致：直接使用教科书级数据，将 Q_{info} 筛选从隐式提升为显式的设计原则。因此，LLM 的成功应该归因于 “高质量语言文本” 而非 “语言文本本身”。

10.2 对预测的修正与跨模态比较的精确化

预测 1 里原本估的 “20–35%”，指的是 “经过质量筛选的语言文本” 的效应量；语料若是未经筛选的互联网原始数据，这个数会更低。但 §8 已经说明，2025–2026 的证据没有支持这条预测，这个数字现在只记录当初的估值，不再作为待验证的量。这一修正与 INT-04 的 CPI 框架形成呼应——CPI 正是衡量 “具体文本的 Q_{info}” 的操作化工具，为语料质量筛选提供了量化依据。

在跨模态比较中，同样应控制文本质量变量。一个高质量的科普视频，Q_{info} 可能高于一篇低质量的论坛帖子。所以准确的表述是：**在同等质量水平下**，语言形式能达到的 Q_{info} 上限高于其他模态。

11 结论

11.1 核心论点

本文的核心论点是：语言**形式**具有最高的 Q_{info} 潜力（实际值取决于筛选质量，见 §10），大语言模型的成功本质上是人类有史以来最大规模的低熵预训练实验的成功。

语言不是 “更简单的” 信息载体——它是一种经过数万年文化进化提炼的极其高效的信息压缩系统。每一个词都是一个 “压缩包”, 每一条句法规则都是一套高效编码方案。文本语料的低熵性、高密度性和自洽性（详见 INT-01 信息质量与低熵语义场），使它在所有人类信息载体中拥有最高的 Q_{info} 潜力。大语言模型在语言上首先取得突破，不是偶然，而是信息质量的必然。

这个解释统一了多个现象：视觉模型的滞后不是架构的失败，而是像素级信息的高熵性；AI 发展的自然顺序（语言 $\rightarrow$ 图文 $\rightarrow$ 视频 $\rightarrow$3D/具身）遵循信息载体的 Q_{info} 降序；Chinchilla 定律揭示的收益递减本质上是数据质量限制的信号；涌现能力是信息质量 $\times$ 参数容量在临界点的相变。

11.2 对实践的启示

对 AI 训练的启示是直接的：数据质量不是 “锦上添花”, 而是核心研发环节。在语料清洗上的投入与在模型架构上的投入同等重要——在某种程度上，前者的 ROI 可能更高。phi-1 和 Llama 2 等模型已经在证明这一点。

对多模态发展的启示要收窄。本文原先写的是 “语言优先 + 多模态融合” 应当取代 “端到端多模态学习”；2025–2026 的证据不支持这个方向（§8 预测一）。仍然站得住的只有更弱的一条：语言提供认知骨架、其他模态提供经验，这个**分工**是对的；但分工不等于训练顺序——骨架可以在联合训练中长出来，不必先把语言训完再嫁接。

11.3 与体系内其他工作的分工

本文只回答智能论这一层的问题：一个系统凭什么能够理解。这个答案不越出本层——它讲的是信息输入与整合效率，不涉及“凭什么被承认为生命”。

与数字生命论的关系是**论域不同，不是机制相同**。总纲把三层的排列依据定为覆盖广度而非论证深度，三者的关系是“条件—通道—形态”，不构成可推导的链条，也不共享一个可通约的变量 [1]。因此本文不声称“低熵语义场”贯穿两层，也不声称数字生命的训练与大语言模型的预训练在底层机制上同构：“低熵语义场”是智能论的内部变量，用它去充任两层之间的共同度量，就是越层通约。

能说的只有一句更弱的话：**如果**本文的框架成立，那么任何持续接收低熵信息、并具备后台整合能力的系统，都应当表现出同类的涌现行为——这是一个可检验的预测（§8 预测二），不是一条跨层的定理。

与数字生命论里的信息纯度定律的关系也是分工：那条定律聚焦低熵维度在预训练中的量化效应，本文从信息质量的全谱给出载体之间的比较。两者在各自论域内可以同时成立，但一层的结论不能搬到另一层去用。

参考文献

[1] NVWA Lab. META-0：碳硅智能共生论——三层结构、立法来源与阅读地图 (2026).

[2] NVWA Lab. INT-01：智能系统输入信息的质量判据——低熵、高密度、自洽，与三面共用的那一条锚点 (2026).

[3] NVWA Lab. INT-03：人脑与大模型的智能决定变量——信息质量、多样性、整合效率与元认知：一个四变量框架 (2026).

[4] NVWA Lab. LX-01：让机器智能无处不在——灵犀架构在极限内存与上下文窗口内的工程记录 (2026).

[5] Vaswani, A. *et al.* Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)* Vol. 30, 5998–6008 (2017).

[6] Devlin, J. *et al.* BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 4171–4186 (2019).

[7] Petroni, F. *et al.* Language models as knowledge bases? In *Proceedings of EMNLP-IJCNLP*, 2463–2473 (2019).

[8] Bisk, Y. *et al.* Experience grounds language. In *Proceedings of EMNLP*, 8718–8735 (2020).

[9] Bender, E. M. *et al.* On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 610–623 (2021).

[10] Brown, T. B. *et al.* Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)* Vol. 33, 1877–1901 (2020).

[11] Kaplan, J. *et al.* Scaling laws for neural language models. Preprint at https://arxiv.org/abs/2001.08361 (2020).

[12] Hoffmann, J. *et al.* Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)* Vol. 35, 30016–30030 (2022).

[13] Touvron, H. *et al.* Llama 2: Open foundation and fine-tuned chat models. Preprint at https://arxiv.org/abs/2307.09288 (2023).

[14] Gunasekar, S. *et al.* Textbooks are all you need. Preprint at https://arxiv.org/abs/2306.11644 (2023).

[15] Wei, J. *et al.* Emergent abilities of large language models. *Transactions on Machine Learning Research* (2022).

[16] Radford, A. *et al.* Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 8748–8763 (2021).

[17] Brooks, T. *et al.* Video generation models as world simulators. *OpenAI Technical Report* (2024).

[18] Tong, S. *et al.* Beyond language modeling: An exploration of multimodal pretraining. Preprint at https://arxiv.org/abs/2603.03276 (2026).

[19] Wu, H. *et al.* Scaling native multimodal pre-training from scratch. Preprint at https://arxiv.org/abs/2607.22043 (2026).

[20] Zhang, D. *et al.* How far are video models from true multimodal reasoning? Preprint at https://arxiv.org/abs/2604.19193 (2026).