
从熵到 Epiplexity：重新定义计算受限时代的“数据价值”
探索大模型学习规律的深层度量与研究院共识
主讲人：金鹏（AI研究院工程总监）

经典信息论在 AI 时代的失效

香农熵假设观察者拥有无限计算能力。对于算力受限的神经网络，这不仅是错误的，而且是误导性的。

1. **创造悖论 (Creation)**：为什么确定性过程 （如 AlphaZero 自我对弈）能从简单的规则中“创造”出超越人类的新知识？

2. **顺序悖论 (Ordering)**：为什么打乱像素顺序的图像根本无法被模型学习？

3. **似然悖论 (Likelihood)**：为什么随机哈希值难以预测，却对泛化毫无价值？

NotebookLM

核心定义：什么是 Epiplexity?

$$S_T(X) := |\mathcal{P}^*|$$

定义：在给定计算时间限制 T 下，能最小化数据描述长度的最佳概率程序 (P^*) 的比特长度。

计算受限观察者 (Bounded Observer)：

- 我们不关心数据的绝对随机性。
- 我们关心的是：在有限的 GPU 算力下，能提取出多少可复用的结构 (Reusable Structure)。

解读：Epiplexity = 模型内部构建的“程序”或“电路”的复杂度。

NotebookLM

信息分解：结构 vs 噪声

总信息量 = Epiplexity（结构性内容）+ 时间受限熵（残余随机性）

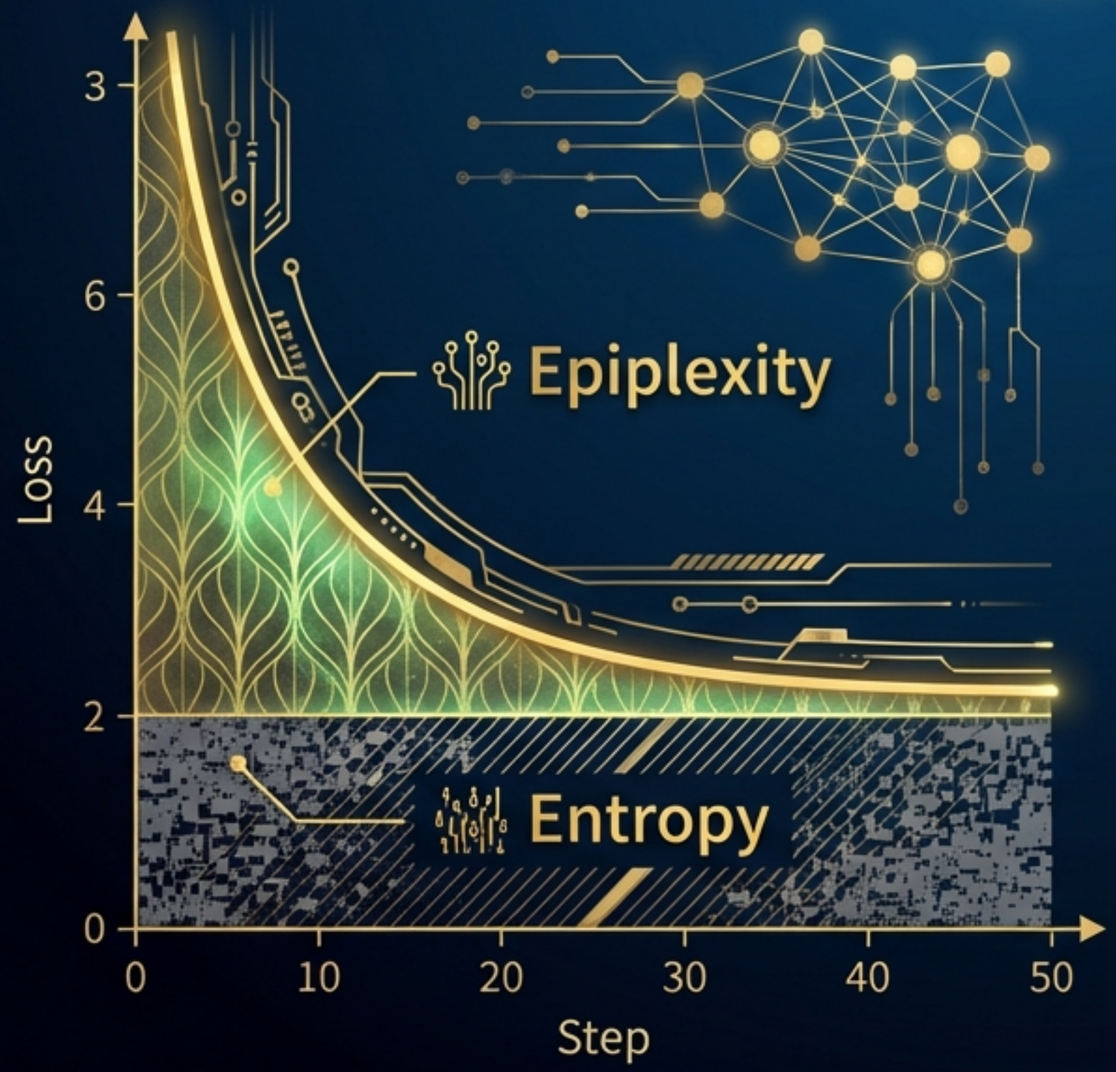
高 Epiplexity

自然语言、代码、物理规律。模型必须构建复杂的“归纳头”来压缩它。

高熵（High Entropy）

随机哈希值、日志 ID。这是“预测性痛苦”的数据——模型在做无用功，没有学到可迁移的技能。

NotebookLM



测量秘籍：训练损失曲线的面积（AUC）

启发式估计（Heuristic）：
Epiplexity ≈ 训练损失曲线下方、最终收敛值上方所围成的面积。

工程洞察：

- 快速下降的曲线 = 低成本获取大量结构（高 Epiplexity）。
- 平坦或缓慢的曲线 = 数据包含大量噪声，或模型算力不足。

结论：学习的速度，就是结构含量的度量。

NotebookLM

死记硬背
(Rote Memory)

泛化
(Generalization)

泛化启示：识破“死记硬背”的模型

国际象棋案例（Chess Experiment）：

- 正向顺序（Forward）：学习“根据棋局预测下一步”。（容易，结构浅）。
- 逆向顺序（Reverse）：学习“根据棋局推导上一步”。（困难，Epiplexity 高）。

结果：逆向训练的模型在 OOD 任务上表现更好。

核心观点：高 Epiplexity 意味着模型内化了通用的算法程序，而非仅仅记忆统计相关性。

NotebookLM

工程实践：数据选择的“第一准则”

策略：ADO（Adaptive Data Optimization）

- 停止追求单纯的低 Loss。
- 新目标：追求单位算力下的 Epiplexity 增量最大化。

实施效果：

动态调整数据分布，优先训练那些让模型“学习速度最快、结构提取最多”的数据子集。

NotebookLM

研究院行动指南（Consensus）

1.升级评估基准（Upgrade Evals）

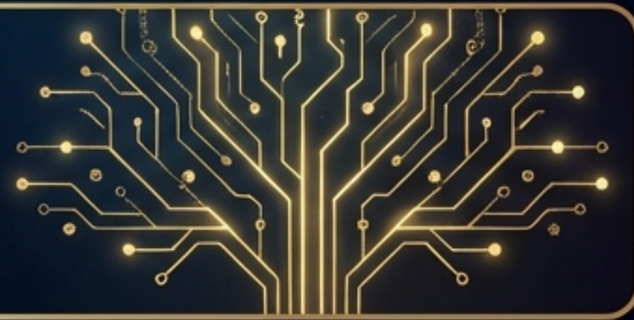
在 Eval 流程中加入 AUC 面积监测。
Loss 下降的轨迹形状是能力涌现 (Emergence) 的早期预警。

2. 合成数据策略（Synthetic Data）

利用合成数据注入高 Epiplexity 结构，
强迫模型构建更深的推理电路。

3. 课程学习（Curriculum）

根据 Epiplexity 对预训练语料进行排序，
让模型先学结构，后填细节。



NotebookLM


成为“更聪明的观察者”
学习即压缩 (Learning is Compression)
智能即 Epiplexity (Intelligence is Epiplexity)
我们不再仅仅是数据的消费者，我们是结构的提取者。
从今天起，关注那些能迫使模型“进化”的数据。
NotebookLM