



清华大学 交叉信息研究院
Institute for Interdisciplinary Information Sciences, Tsinghua University

语言大模型的信息、计算和统计边界

李建

清华大学 交叉信息研究院

语言大模型理论研究

(不完全分类)

构架研究

Architecture

Expressivity, Multi-layer transformer,
Attentions, Mamba, RoPE....

数据研究

Training Data

Data filtering, Data Mixing, Data selection,
Synthetic Data generation, Tokenization,
Compression, ...

优化器研究

Optimizer

Adam, AdamW, Muon, Shampoo
LR schedule, landscape, LoRA

统计学习研究

Statistical Learning

Multi-task, zero-shot generalization,
In-context generalization
Generalization vs memorization
Calibration vs Hallucination

语言大模型理论极限

构架研究

Architecture

Expressivity, Multi-layer transformer,
Attentions, Mamba, RoPE....

数据研究

Training Data

Data filtering, Data Mixing, Data selection,
Synthetic Data generation, Tokenization,
Compression, ...

假设优化器已经足够好了

(经验损失可以优化到理论最优)

统计学习研究

Statistical Learning

Multi-task, zero-shot generalization,
In-context generalization
Generalization vs memorization
Calibration vs Hallucination

语言大模型理论极限

视角1:

大模型是图灵机/电路

构架和模型规模限制

计算边界 (计算复杂性理论)

视角2:

大模型是压缩编码器

数据构成、结构、随机性

信息边界 (信息与编码理论)

假设优化器已经足够好了

(经验损失可以优化到理论最优)

(优化理论)

视角3:

大模型是统计学习模型

当前统计学习范式

统计边界 (统计学习理论)

视角1:

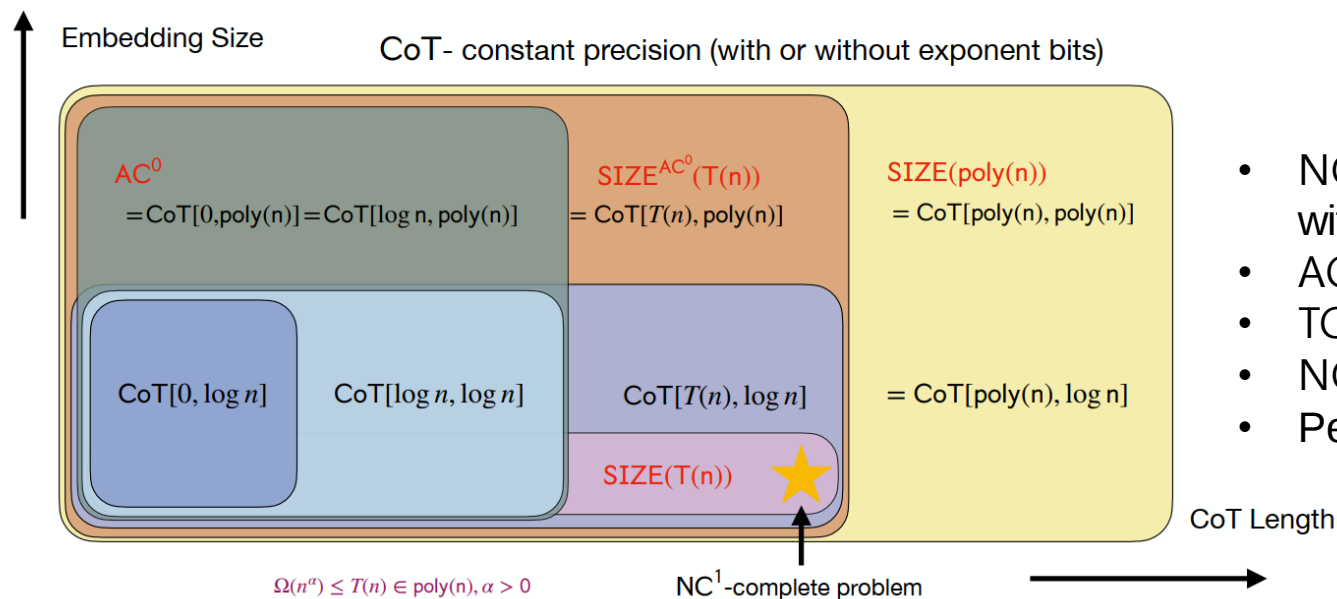
大模型是图灵机/电路

构架和模型规模限制

计算边界 (计算复杂性理论)

LLM Theory – Expressiveness 表达性研究

- LLM能计算/表达什么？
- LLM 主要是Transformer构架，Transformer可以看成线路 circuit
- 计算复杂性类（特别是线路复杂性类）
- CoT 可以增强表达能力



- NC^0 : Problems that can be solved by constant depth circuits with poly NOT, AND, OR gates.
- AC^0 : NC^0 + unbounded fan-ins AND and OR gates.
- TC^0 : AC^0 + unbounded fan-ins MAJORITY gates.
- $\text{NC}^k, \text{AC}^k, \text{TC}^k$: with $\log^k n$ (instead of constant) depth.
- People believe $\text{TC}^0 \neq \text{NC}^1$

部分参考文献:

- “Chain of Thought Empowers Transformers to Solve Inherently Serial Problems”, ICLR2024.
- “Towards revealing the mystery behind chain of thought: a theoretical perspective”, NeurIPS2023.
- “The Parallelism Tradeoff: Limitations of Log-Precision Transformers.” ACL2023
- “Theoretical limitations of multi-layer Transformer.” FOCS 2025

计算能力的边界

计算复杂性理论能力视角

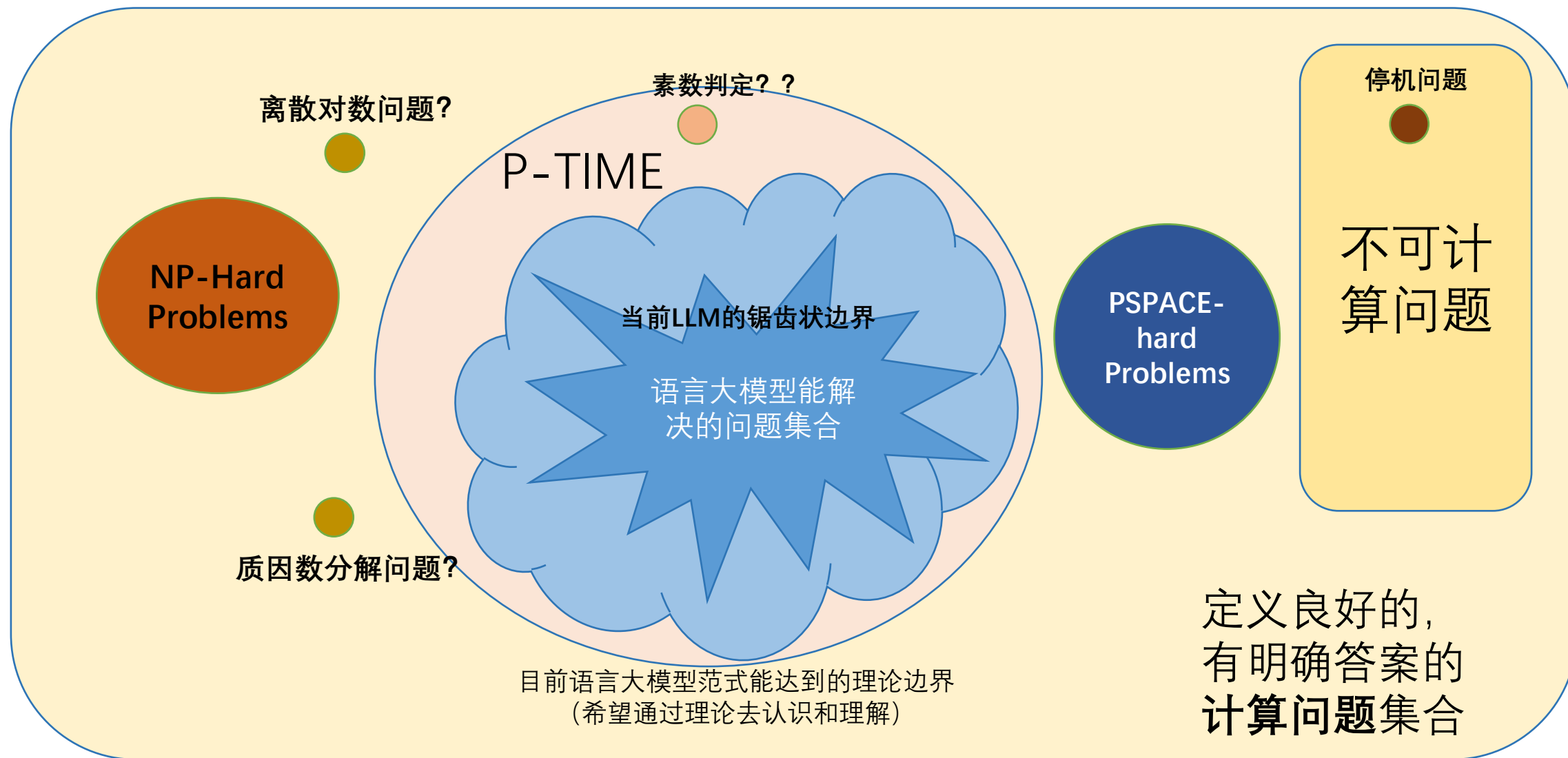


目前语言大模型范式能达到的理论边界
(希望通过理论去认识和理解)

只考虑定义良好的，有明确对错或答案的问题
只考虑语言大模型本身，暂时假设大模型不能调用外部工具

计算能力的边界

计算复杂性理论能力视角



只考虑语言大模型本身（可以有CoT），但是假设大模型不能调用外部工具

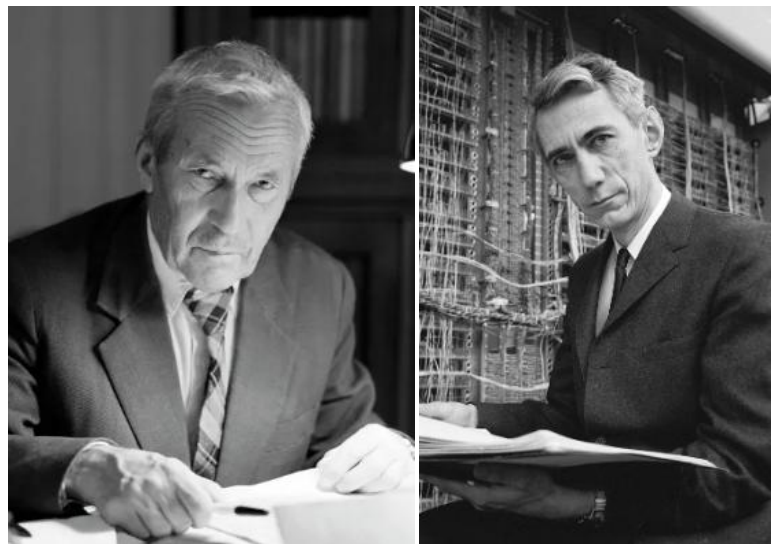
视角2:

大模型是压缩编码器

数据构成、结构和规模

信息边界 (信息与编码理论)

压缩视角看待大模型



压缩即智能：压缩需要归纳数据的规律，纯粹的随机数无法压缩。**理论上可以证明压缩等价于预测下一个词**

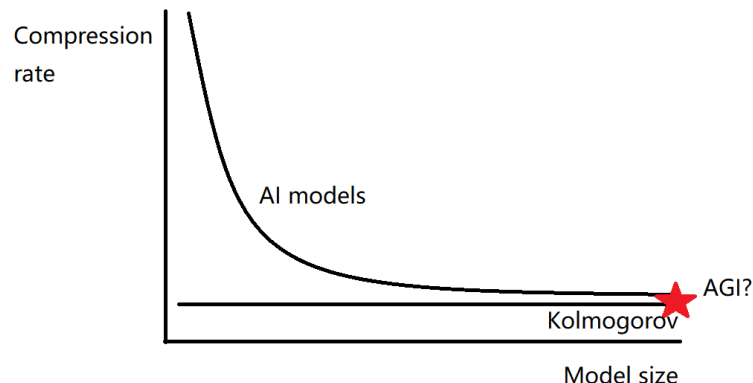
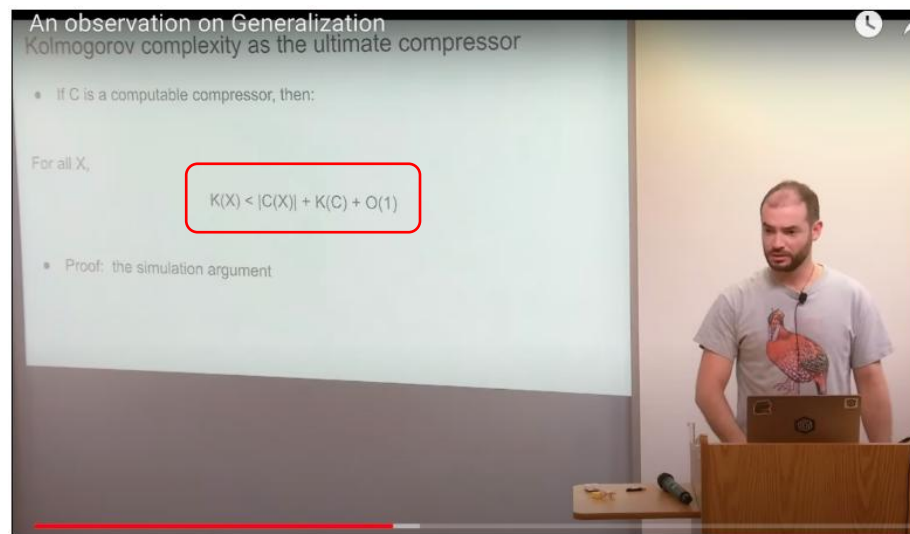
Shannon首先研究了英文的熵（目的是为了压缩英文）
提出了**猜测下一个词**游戏
Shannon, Prediction and Entropy of Printed English (1951)

Kolmogorov提出了算法上终极的压缩概念
Kolmogorov Complexity (algorithmic information theory)

$K(X)$:= The minimum length of Turing Machine that outputs X.

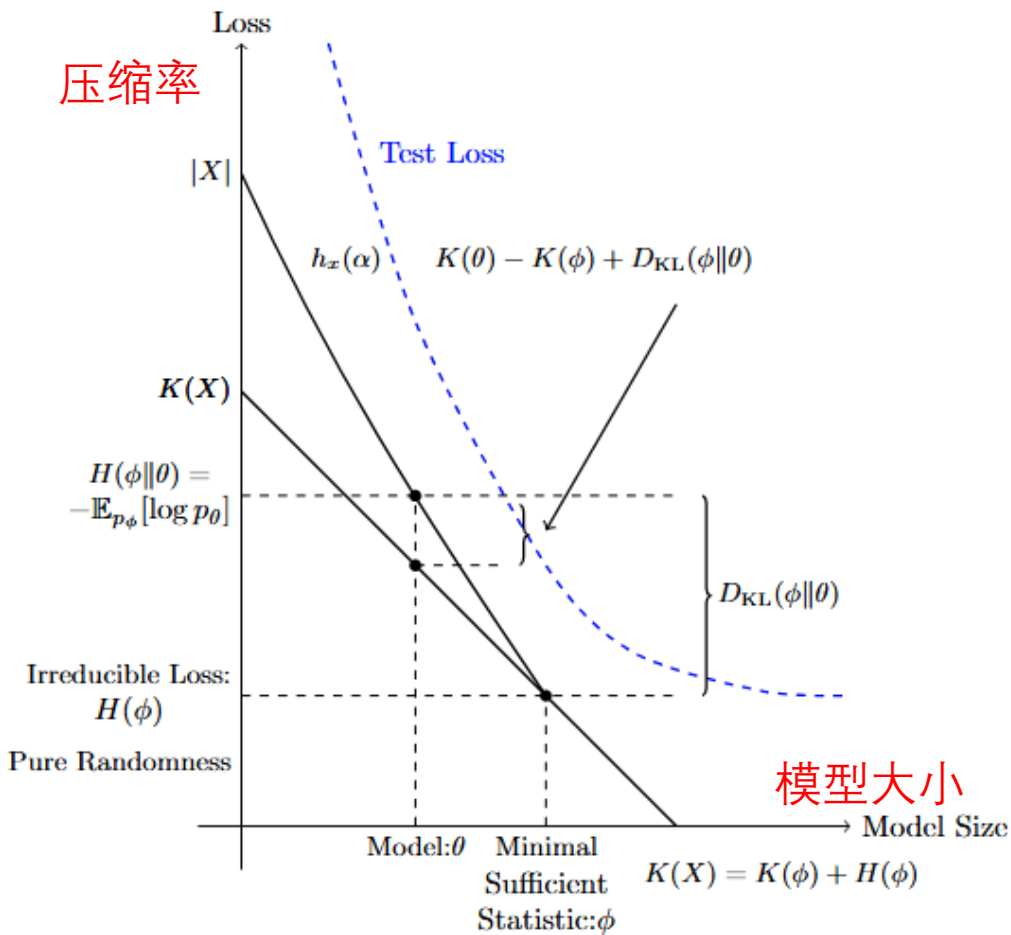
Very good talk (watch on youtube :
https://www.youtube.com/watch?v=AKMuA_TVz3A)

Ilya 从 **Kolmogorov complexity theory** 视角看待大模型：
大模型可以看成一个强大的**压缩器**压缩世界语料



Kolmogorov compressor as the ultimate compressor

Kolmogorov结构函数



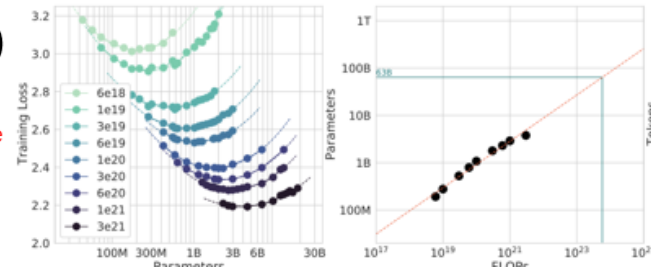
经验规律

Chinchilla Scaling Laws (DeepMind)

$$L(N, D) = E + \frac{A}{N^{0.34}} + \frac{B}{D^{0.28}}$$

why such strange powers appear?

$$E = 1.69, A = 406.4, B = 410.$$

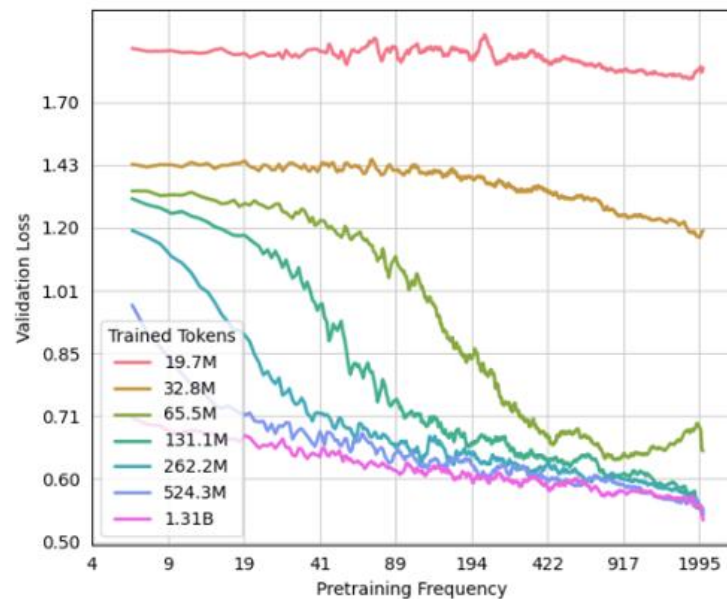


- **Kolmogorov结构函数是模型拓展率 (Model Scaling Law) 的极限** (无法通过改进优化器突破)
- 从理论上提供了Scaling Law如上形式的解释:
 - E是纯粹随机性, 不可压缩;
 - 数据的power law结构 (如Zipf' s law和Heap' s Law) 部分决定了Scaling Law中的power law形式
- 阐述了指定模型规模约束下, 极限压缩下**什么被学会, 什么被遗忘**

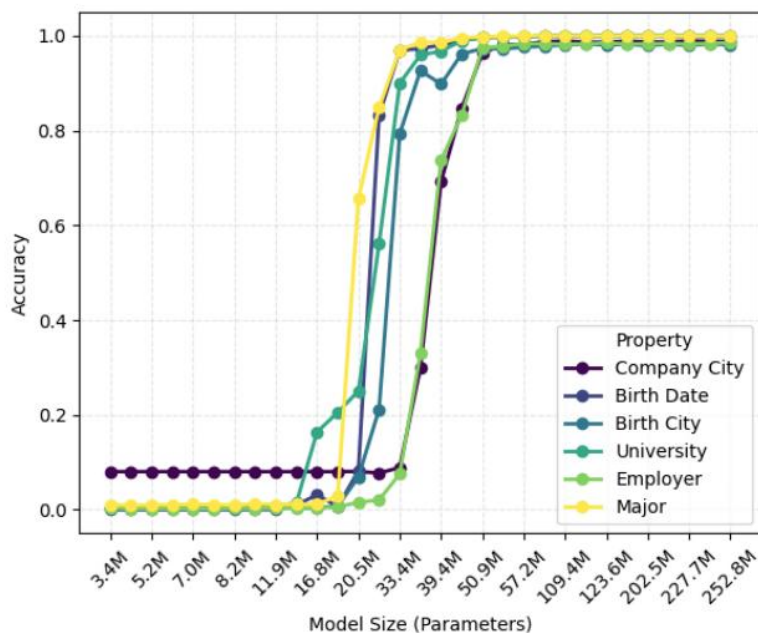
Understanding LLM Behaviors via Compression: Data Generation, Knowledge Acquisition and Scaling Laws · Zhixuan Pan, Shaowen Wang, Liao Pengfei, **Jian Li. NeurIPS 2025 Spotlight**

大模型的知识学习

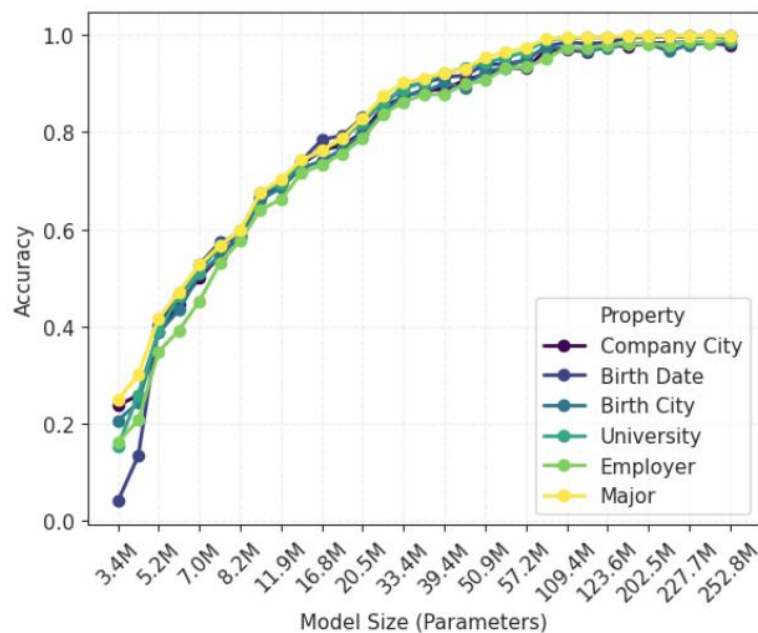
从压缩的角度，首先压缩更高频的知识可以达到更好的压缩率（理论上，可以通过解一个优化问题得到理论最优值）



训练过程中
高频知识首先被学习



模型增大过程中的相变现象
(数据均匀分布)
频率都一样，同时被学习或遗忘

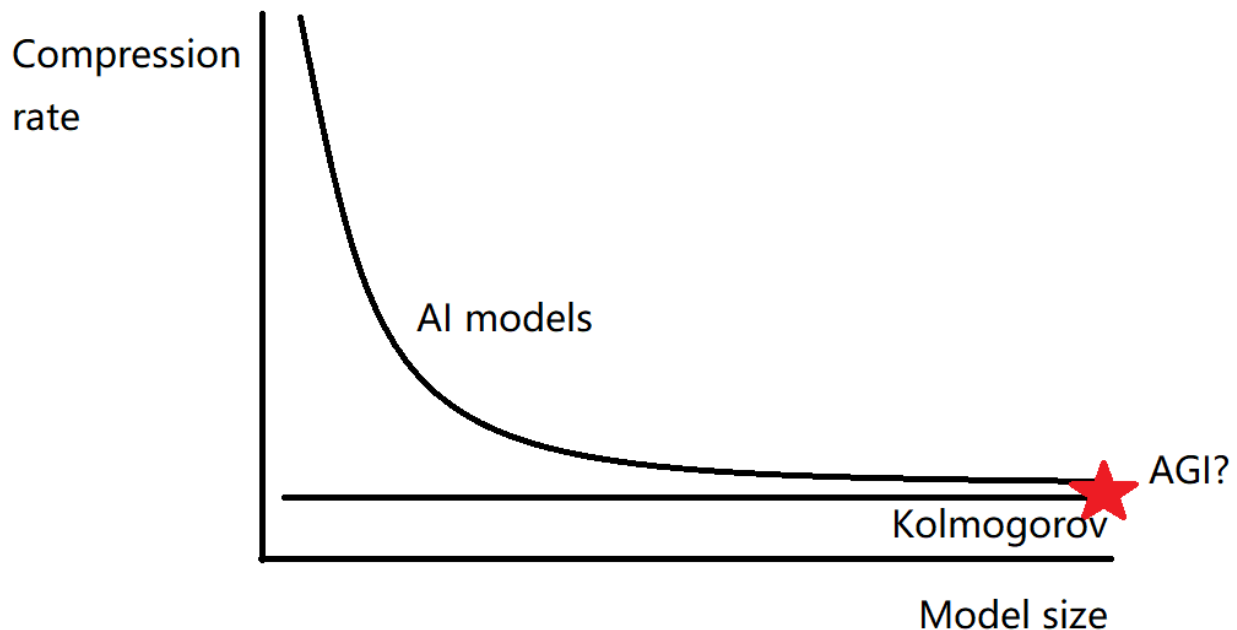


模型增大过程中准确率的渐变现象（数据power law分布）
高频知识首先被学习，低频知识被遗忘

采用“Physics of LLMs” -1,2,3 提出的控制实验方法产生人工数据训练的语言大模型

信息的边界

算法信息论视（压缩） 角度



压缩的极限：Kolmogorov复杂性

Scaling Law的极限：Kolmogorov结构函数



只考虑语言大模型本身（可以有CoT），但是假设大模型不能调用外部工具（如search互联网）

视角3:

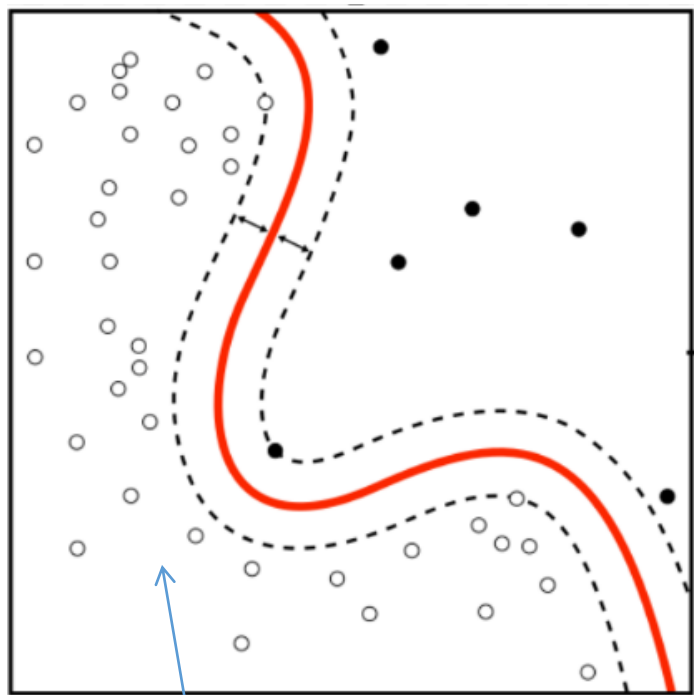
大模型是统计学习模型

当前统计学习范式

统计边界 (统计学习理论)

大模型是统计学习和深度学习模型

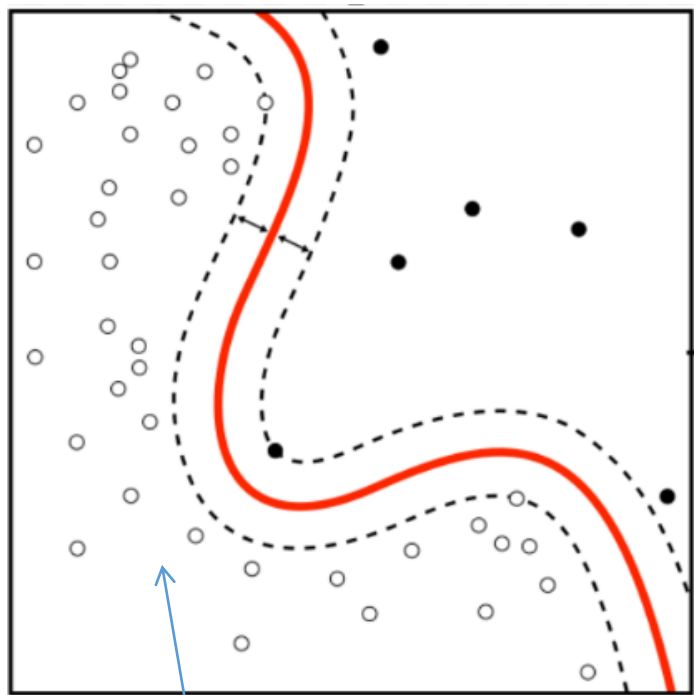
LLM也是要解决next token prediction 的分类问题（尽管维度更高，分界面更复杂）



泛化到训练集没见过的点

大模型是统计学习和深度学习模型

LLM也是要解决next token prediction 的分类问题（尽管维度更高，分界面更复杂）



泛化到训练集没见过的点

深度神经网络上的 Max-Margin 定理:

首次证明了深度神经网络上进行梯度下降的过程中 Margin 单调递增，且收敛到一个有约束的 Max-Margin 问题 (P) 的 KKT 点上

SVM-like constrained optimization problem:

$$\min \|\theta\|_2^2 \quad \text{s.t. } y_i f_\theta(x_i) \geq 1 \quad (P)$$

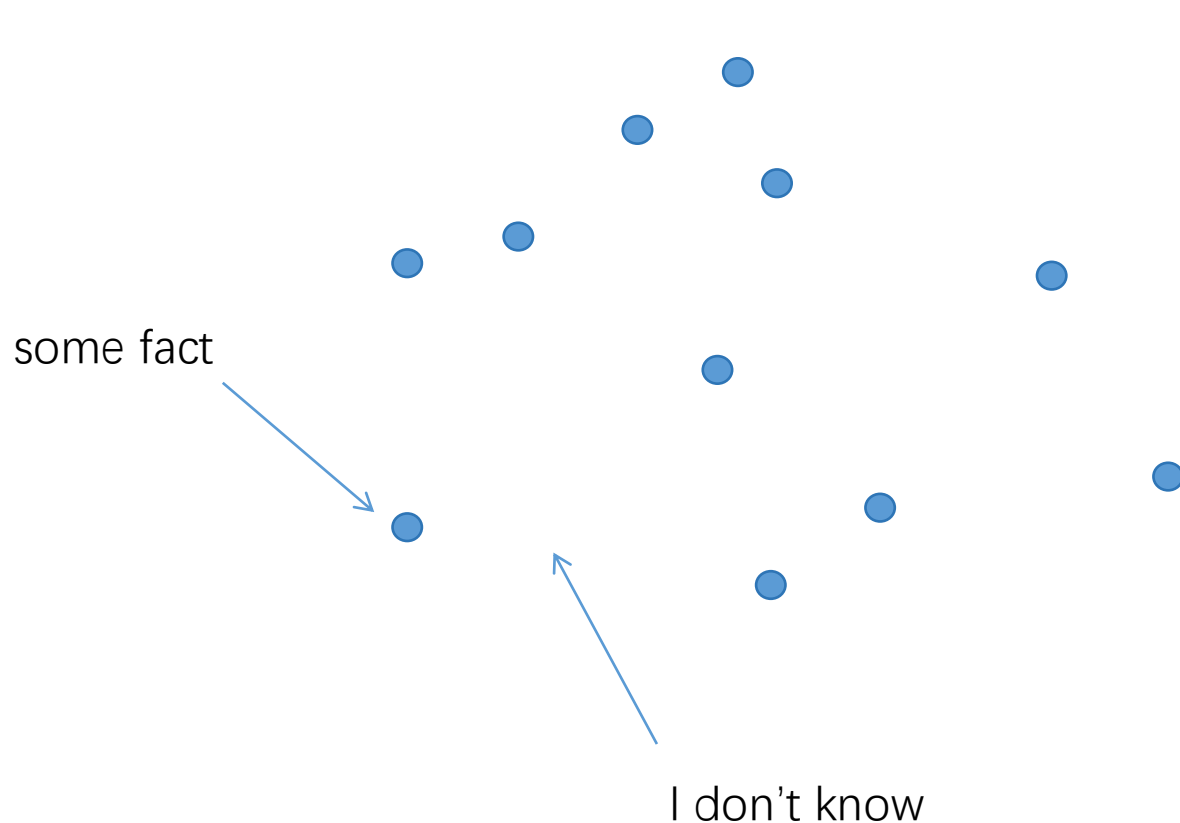
- 该定理和DL中很多现象具有有趣的联系，如simplicity bias, adversarial robustness, neural collapse, benign overfitting等
- 是研究梯度方法隐式偏差的重要理论工具

Kaifeng Lyu, **Jian Li**. Gradient Descent Maximizes the Margin of Homogeneous Neural Networks. International Conference on Learning Representations (ICLR2020, Oral)

大模型是统计学习模型

考虑事实性问题的幻觉

例如：大模型编造不存在的人物，事迹，文献



假设每个蓝色点是训练集中出现事实

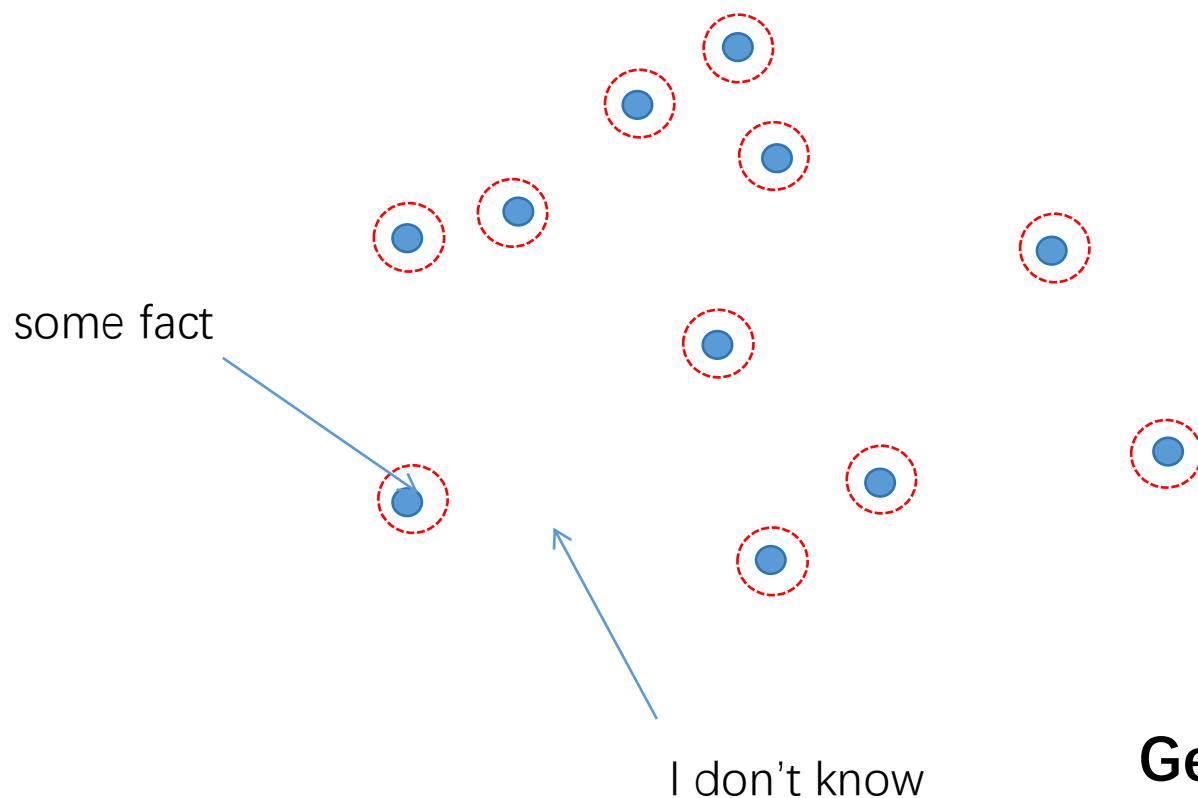
如果没有幻觉，那么要求模型能识别出这些蓝色点，而其他点要得出 “I Don't Know” 的回答

大模型是统计学习模型

考虑事实性问题的幻觉

例如：大模型编造不存在的人物，事迹，文献

完全没有泛化的模型



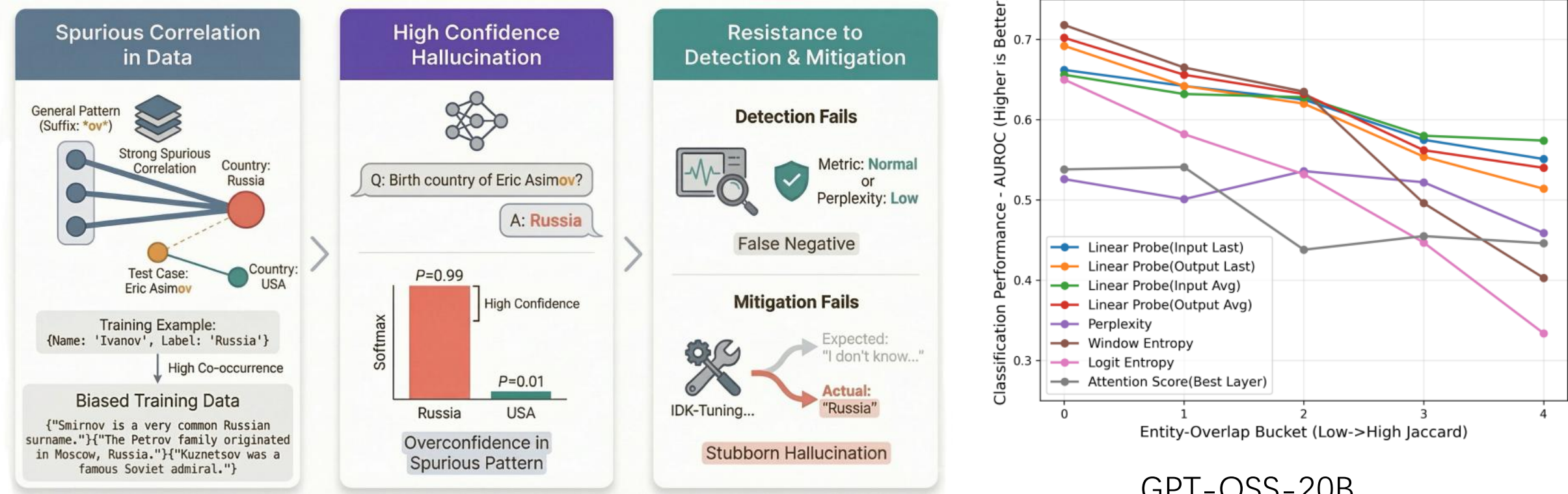
假设每个蓝色点是训练集中出现事实

如果没有幻觉，那么要求模型能识别出这些蓝色点，而其他点要得出 “I Don't Know” 的回答

而这要求分界面是将这些训练数据点完全圈住 (overfit/memorization)

Generalization vs Memorization trade-off

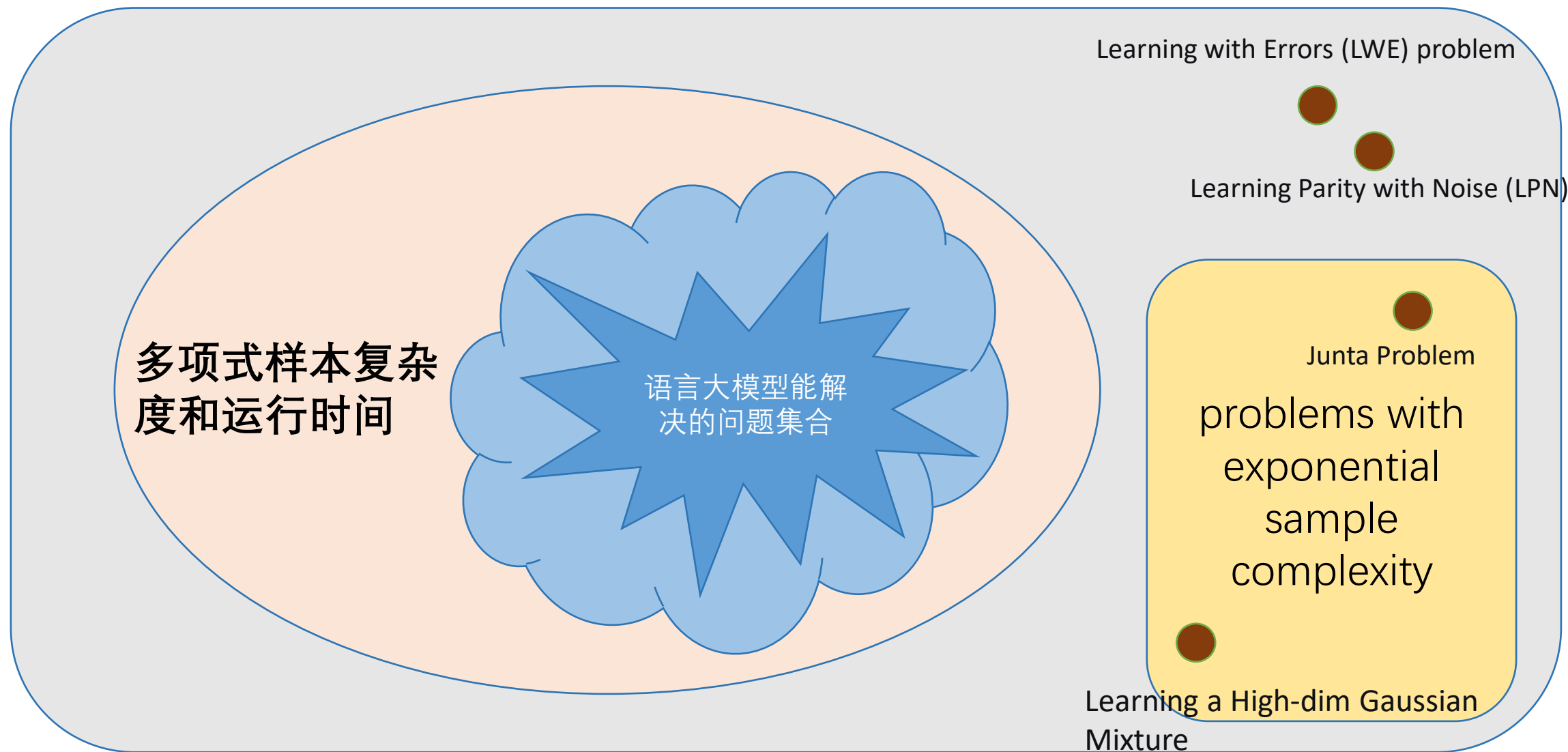
虚假相关性导致幻觉（检测困难）



GPT-OSS-20B
多种幻觉检测方法准确率
随虚假相关性提高而降低

大模型统计边界

统计机器学习理论视角



计算复杂性理论视角

停机问题



离散对数问题?



质因数分解问题?



NP-Hard Problems

低频长尾的
世界知识

无法压缩的噪音

语言大模型能解
决的问题集合

算法信息论视 (压缩) 角度

统计机器学习理论视角

Learning with Errors (LWE) problem



Learning Parity with Noise (LPN)



Junta Problem

problems with
exponential
sample
complexity



Learning a High-dim Gaussian
Mixture

