ChatGPT及大模型专题研讨会

大型语言模型的涌现能力：现象与解释

新浪微博　张俊林

2023-03

Outline

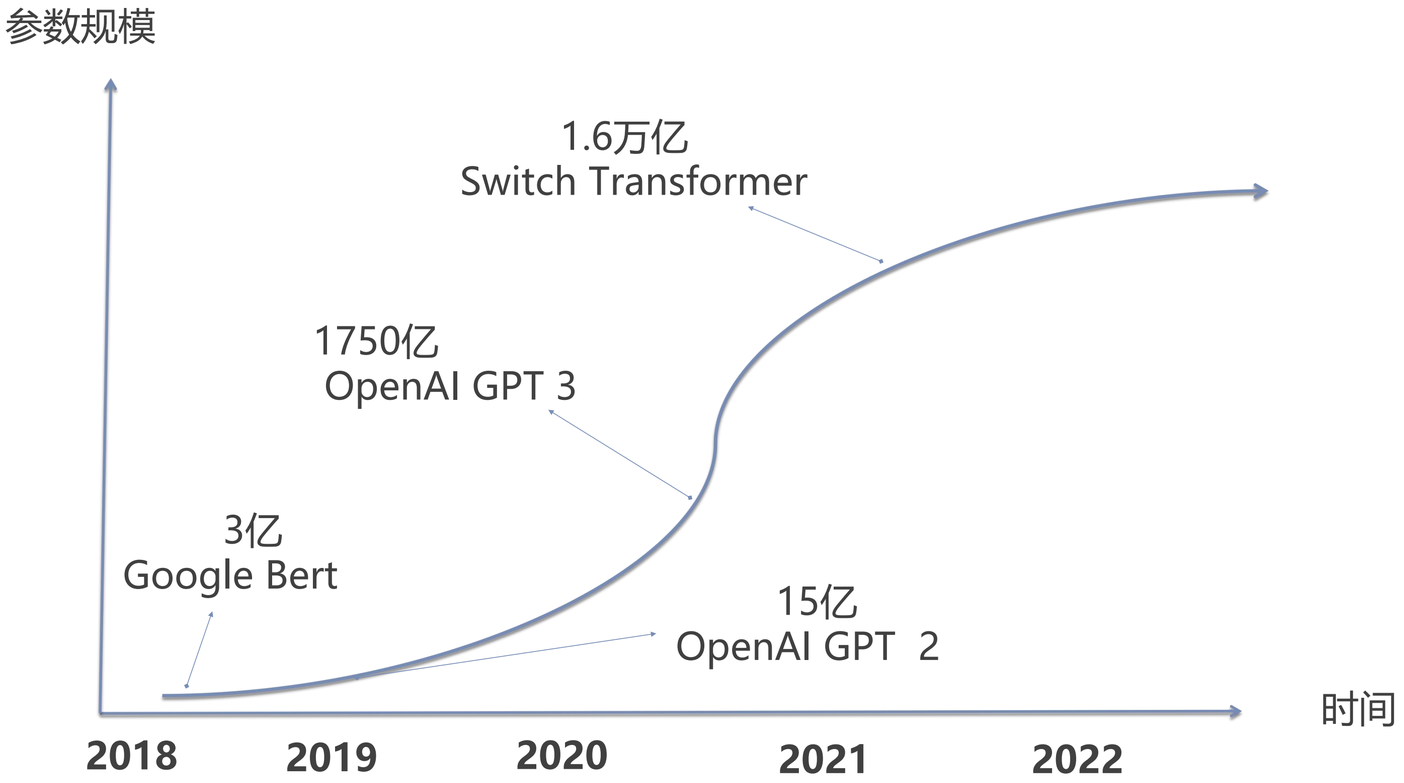
复杂系统中的涌现现象

涌现（Emergence）：或称创发、突现、呈展、演生，是一种现象，为许多小实体相互作用后产生了大实体，而这个大实体展现了组成它的小实体所不具有的特性。（wiki 定义）

生活中的涌现现象

大型语言模型（LLM）：模型规模快速增长

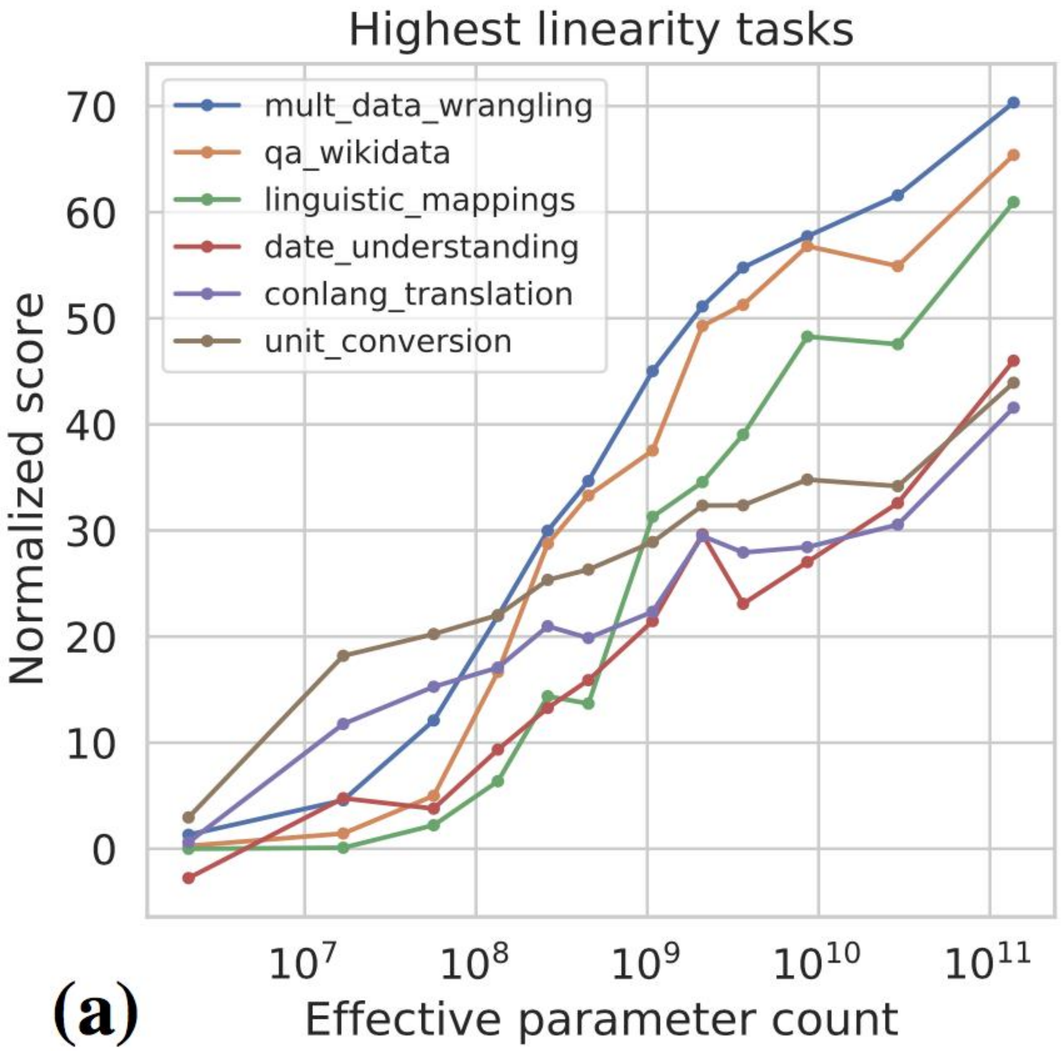
问题：超级大模型会出现涌现现象吗？

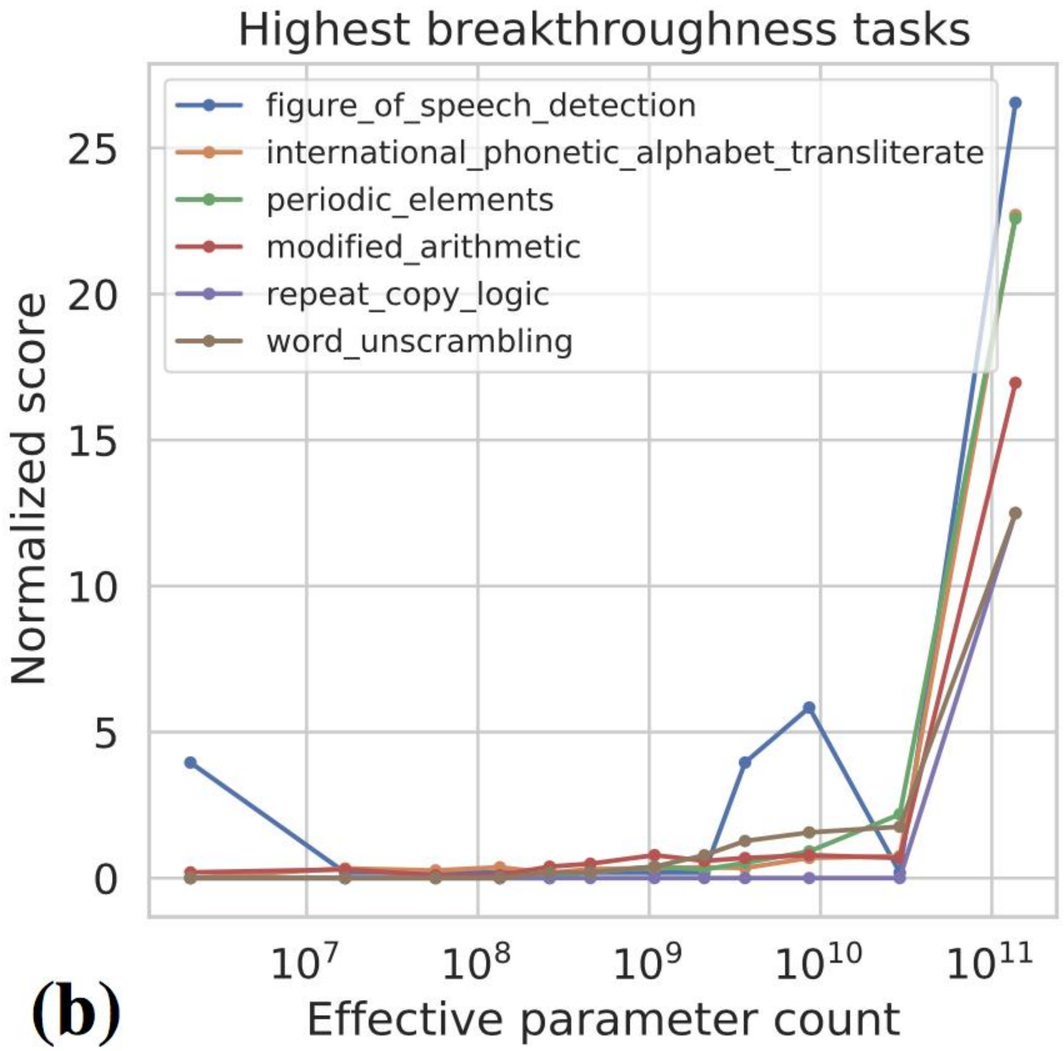


规模大的Large Language Model：

- ✓ GPT 3.0 :175B
- ✓ GPT 3.5:175B
- ✓ LaMDA:130B
- ✓ Gopher:280B
- ✓ PaLM:540B
- ✓ PaLM-E:566B

LLM的规模效应：下游任务表现-伸缩法则&&涌现能力

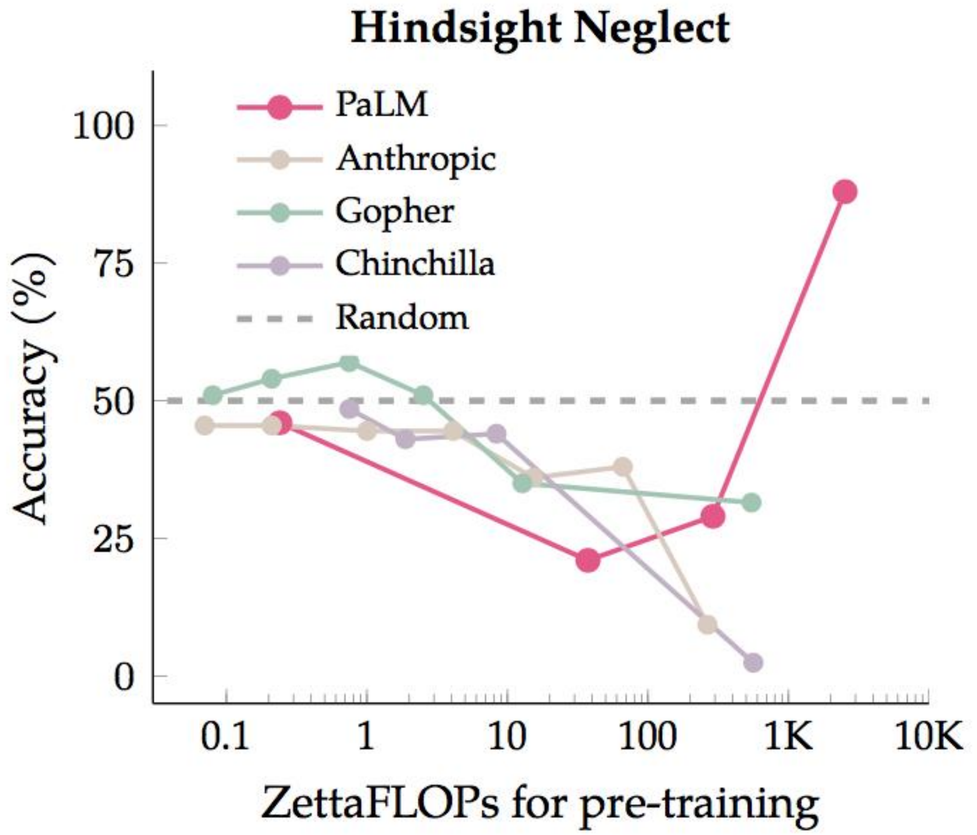


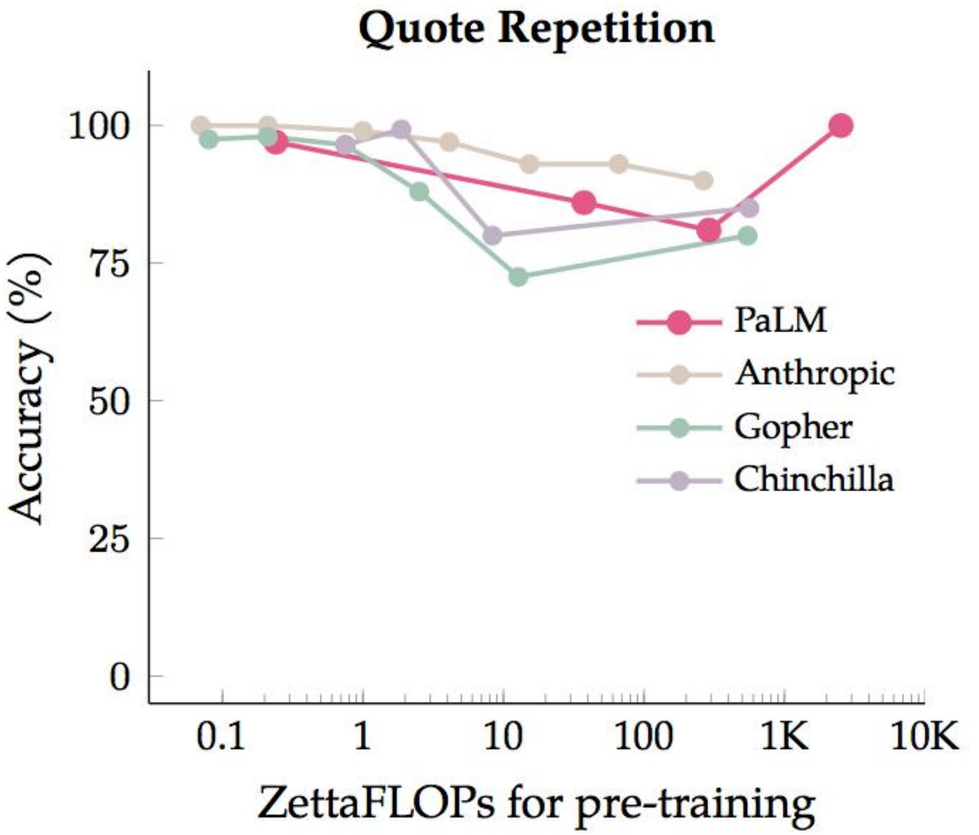


伸缩法则：知识密集型任务

涌现能力：多步骤构成的任务

LLM的规模效应：下游任务表现-U形曲线



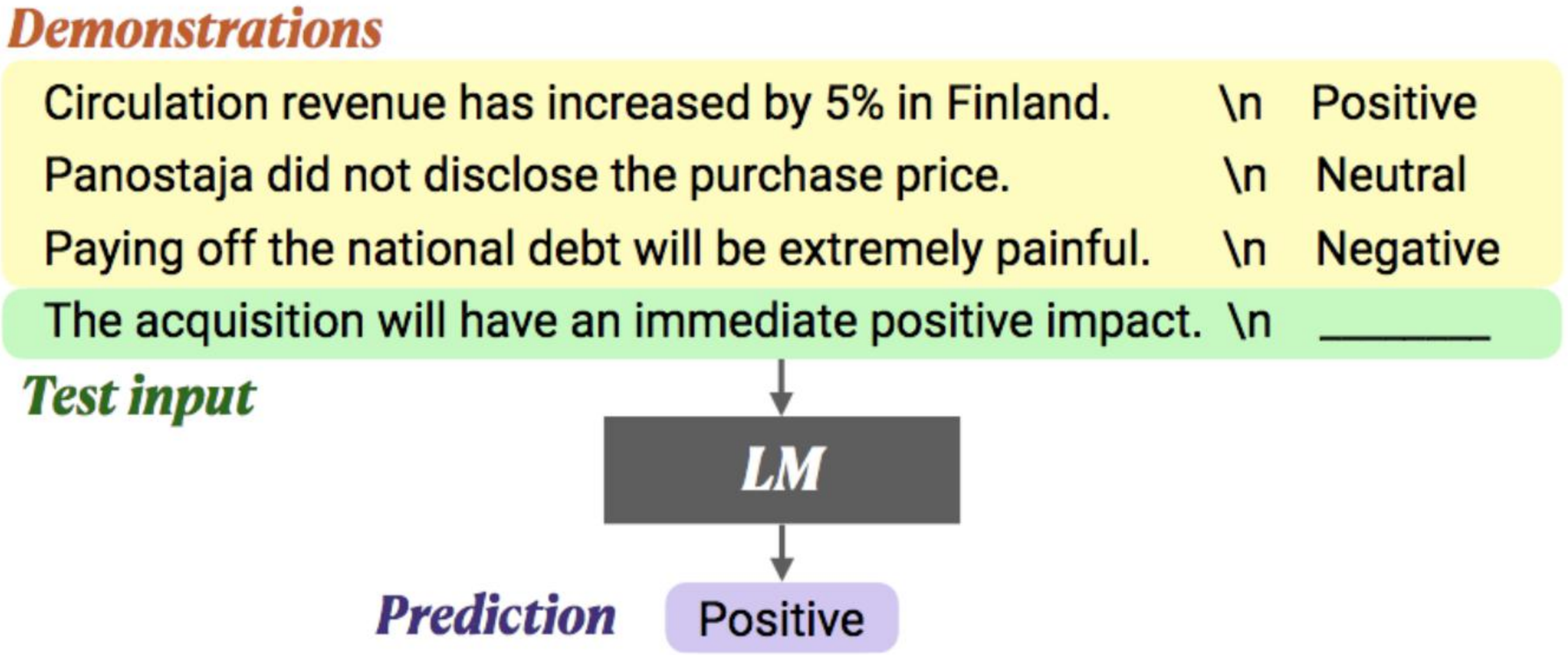


From: Inverse scaling can become U-shaped

Outline

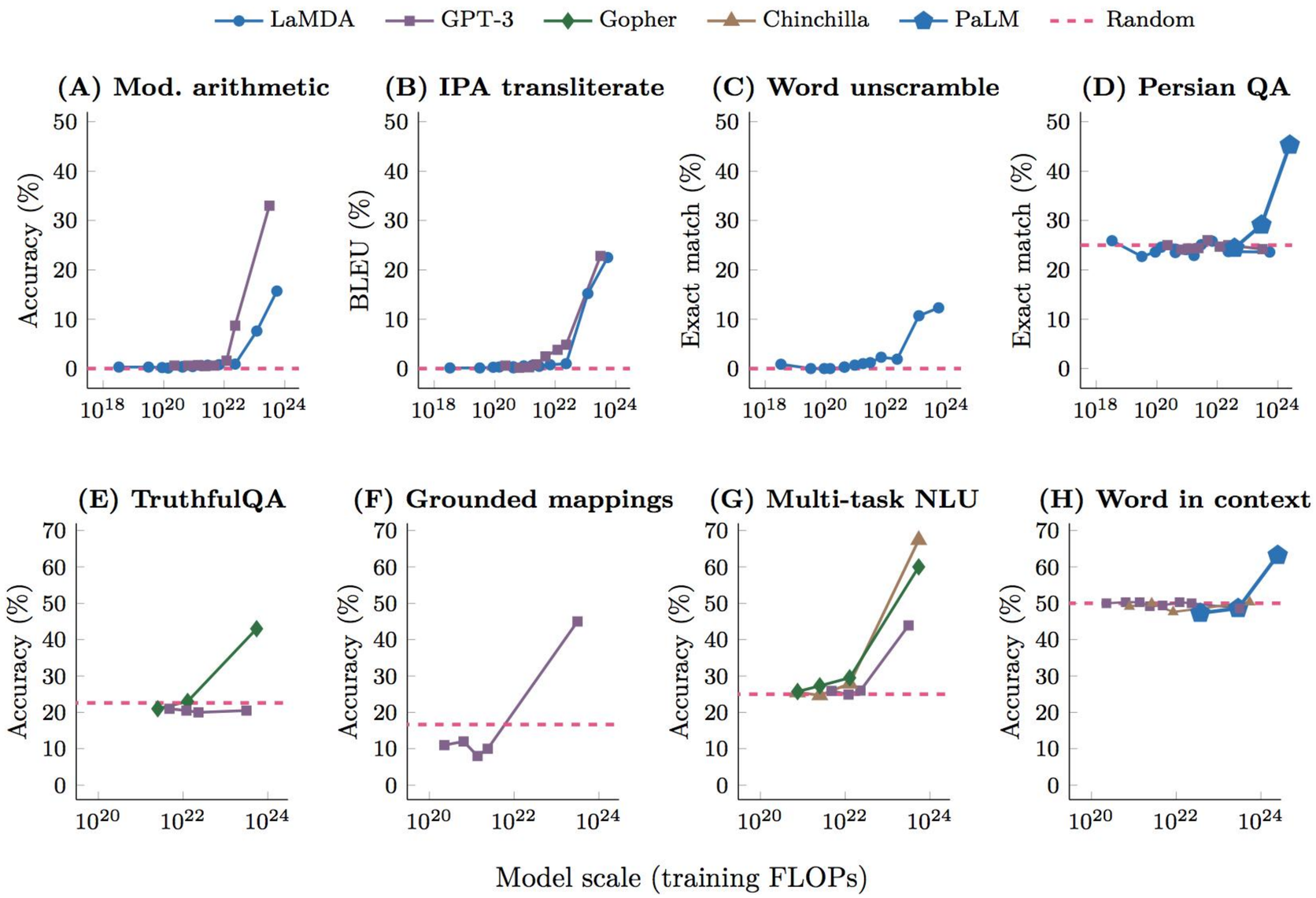
涌现现象-In Context Learning：什么是In Context Learning

In Context Learning:给LLM几个示例，不调整模型参数，LLM即可解决某个领域的问题



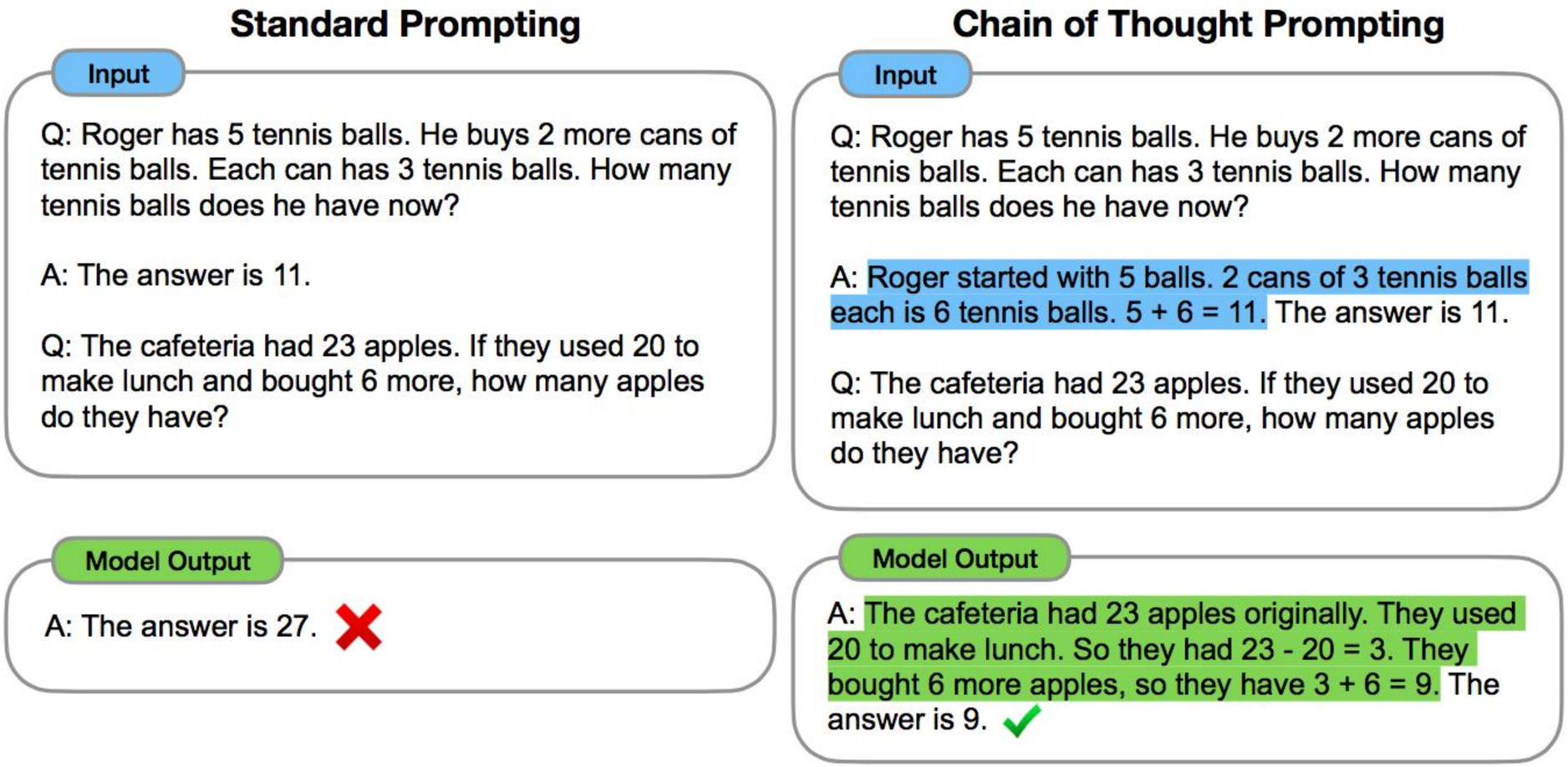
In Context Learning=few shot prompting

涌现现象-In Context Learning：In Context Learning涌现能力



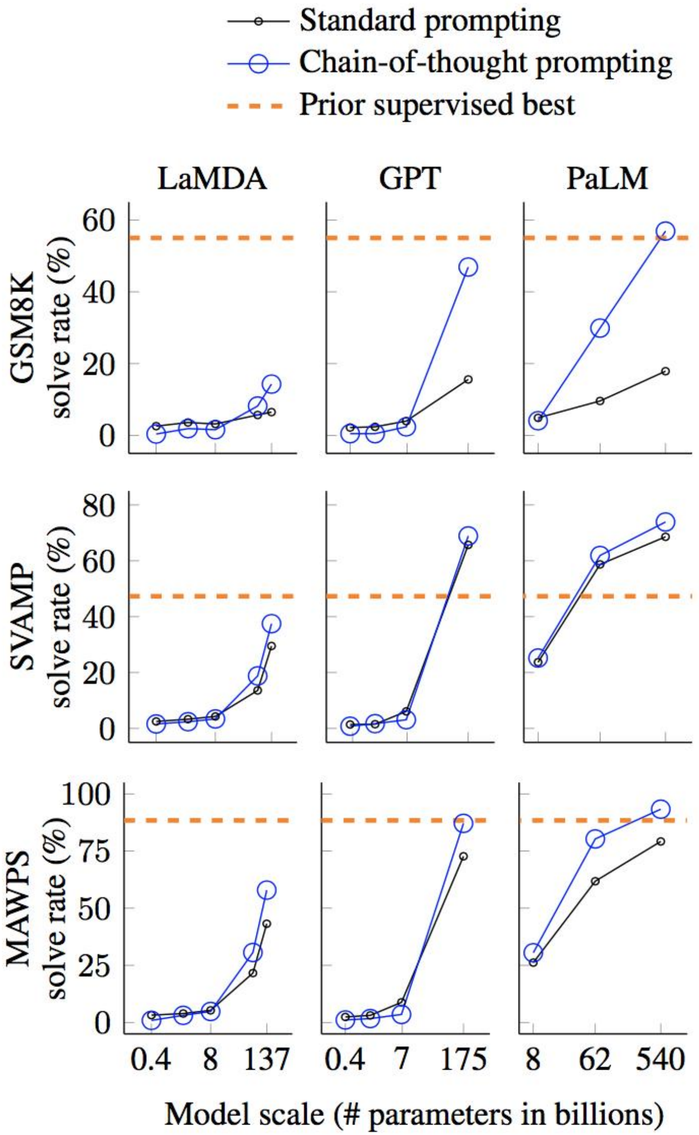
In Context Learning的涌现现象

涌现现象-CoT：什么是思维链（Chain-of-Thought）

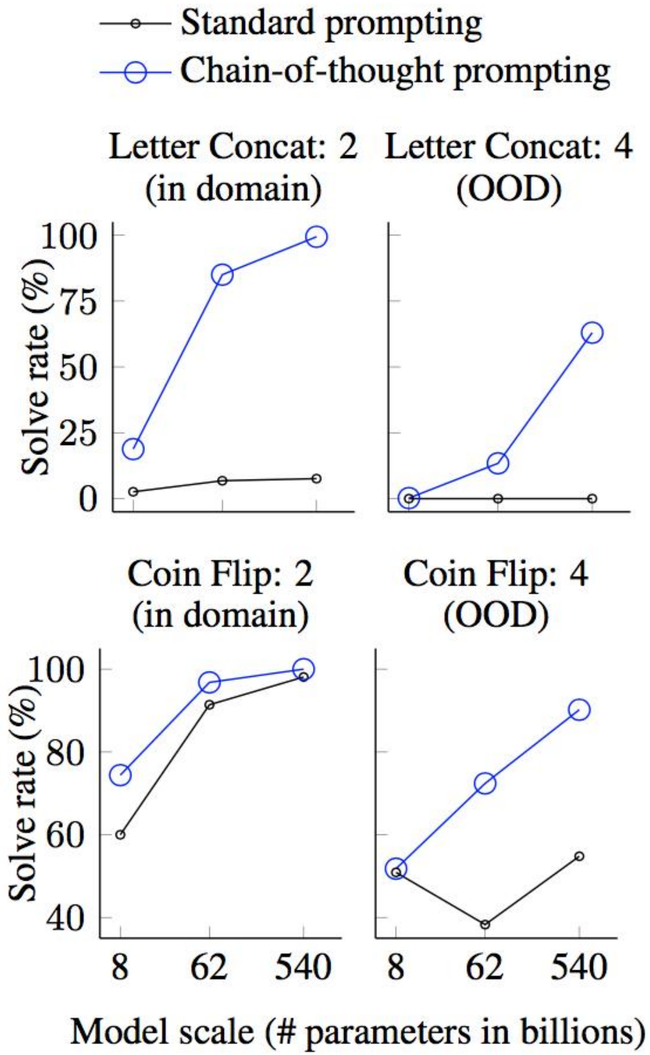


From: Chain of thought prompting elicits reasoning in large language models

涌现现象-CoT：思维链的涌现能力

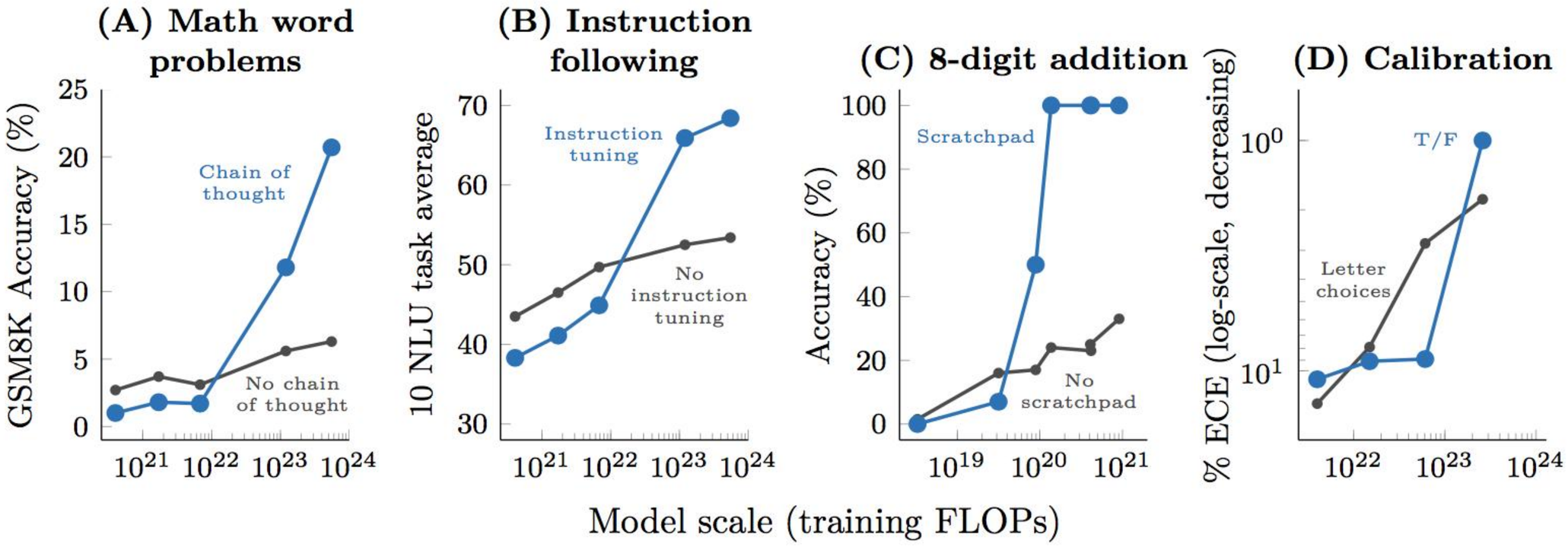


CoT涌现现象:数学问题



CoT涌现现象:符号推理问题

涌现现象：其它涌现能力



其它涌现现象:遵循指令、复杂数学运算等

Outline

01 什么是大模型的涌现能力

02 LLM表现出的涌现现象

03 LLM模型规模和涌现能力的关系

04 模型训练中的顿悟现象

05 LLM涌现能力的可能原因

模型规模和涌现现象的关系：In Context Learning

与具体任务 / 具体模型相关：某些任务13B规模即可，有些任务需要540B，大部分要达到70B

	Emergent scale			
	Train. FLOPs	Params.	Model	Reference
Few-shot prompting abilities				
• Addition/subtraction (3 digit)	2.3E+22	13B	GPT-3	Brown et al. (2020)
• Addition/subtraction (4-5 digit)	3.1E+23	175B		
• MMLU Benchmark (57 topic avg.)	3.1E+23	175B	GPT-3	Hendrycks et al. (2021a)
• Toxicity classification (CivilComments)	1.3E+22	7.1B	Gopher	Rae et al. (2021)
• Truthfulness (Truthful QA)	5.0E+23	280B		
• MMLU Benchmark (26 topics)	5.0E+23	280B		
• Grounded conceptual mappings	3.1E+23	175B	GPT-3	Patel & Pavlick (2022)
• MMLU Benchmark (30 topics)	5.0E+23	70B	Chinchilla	Hoffmann et al. (2022)
• Word in Context (WiC) benchmark	2.5E+24	540B	PaLM	Chowdhery et al. (2022)
• Many BIG-Bench tasks (see Appendix E)	Many	Many	Many	BIG-Bench (2022)

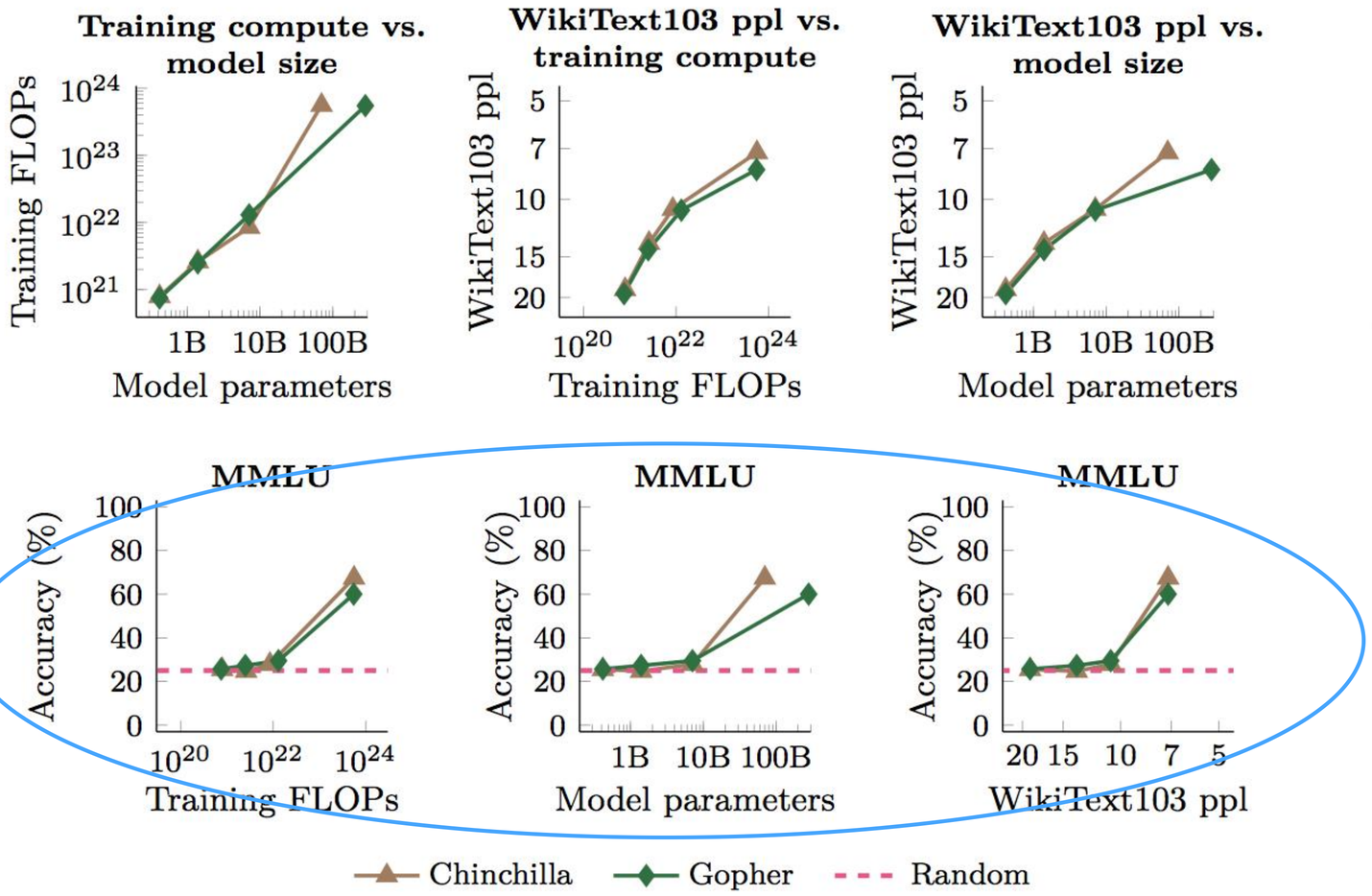
模型规模和涌现现象的关系：CoT等其它涌现能力

与具体任务 / 具体模型相关：有些任务40M即可，有些任务需要达到280B，大部分需要达到50B

	Emergent scale			
	Train. FLOPs	Params.	Model	Reference
Augmented prompting abilities				
• Instruction following (finetuning)	1.3E+23	68B	FLAN	Wei et al. (2022a)
• Scratchpad: 8-digit addition (finetuning)	8.9E+19	40M	LaMDA	Nye et al. (2021)
• Using open-book knowledge for fact checking	1.3E+22	7.1B	Gopher	Rae et al. (2021)
• Chain-of-thought: Math word problems	1.3E+23	68B	LaMDA	Wei et al. (2022b)
• Chain-of-thought: StrategyQA	2.9E+23	62B	PaLM	Chowdhery et al. (2022)
• Differentiable search index	3.3E+22	11B	T5	Tay et al. (2022b)
• Self-consistency decoding	1.3E+23	68B	LaMDA	Wang et al. (2022b)
• Leveraging explanations in prompting	5.0E+23	280B	Gopher	Lampinen et al. (2022)
• Least-to-most prompting	3.1E+23	175B	GPT-3	Zhou et al. (2022)
• Zero-shot chain-of-thought reasoning	3.1E+23	175B	GPT-3	Kojima et al. (2022)
• Calibration via P(True)	2.6E+23	52B	Anthropic	Kadavath et al. (2022)
• Multilingual chain-of-thought reasoning	2.9E+23	62B	PaLM	Shi et al. (2022)
• Ask me anything prompting	1.4E+22	6B	EleutherAI	Arora et al. (2022)

把模型做小影响LLM的涌现能力吗：小模型代表Chinchilla

Chinchilla：DeepMind发布于2021年的LLM，SOTA性能，模型规模70B，训练数据量1400B



貌似Chinchilla在MMLU任务是具备涌现能力的

可能的推论：

1. 减小模型大小增加训练数据数量，可能不影响小模型的涌现能力
2. 如果上述假设成立，那么我们目前应该把模型先做小，再推大（保险起见，不应小于70B）

把模型做小影响LLM的涌现能力吗：小模型代表LLaMA

LLaMA：Meta发布的开源LLM，规模从7B到65B，本质上是开源的Chinchilla

		Humanities	STEM	Social Sciences	Other	Average
GPT-NeoX	20B	29.8	34.9	33.7	37.7	33.6
GPT-3	175B	40.8	36.7	50.4	48.8	43.9
Gopher	280B	56.2	47.4	71.9	66.1	60.0
Chinchilla	70B	63.6	54.9	79.3	**73.9**	67.5
PaLM	8B	25.6	23.8	24.1	27.8	25.4
	62B	59.5	41.9	62.7	55.8	53.7
	540B	**77.0**	**55.6**	**81.0**	69.6	**69.3**
LLaMA	7B	34.0	30.5	38.3	38.1	35.1
	13B	45.0	35.8	53.8	53.3	46.9
	33B	55.8	46.0	66.7	63.4	57.8
	65B	61.8	51.7	72.9	67.4	63.4

Table 9: **Massive Multitask Language Understanding (MMLU).** Five-shot accuracy.

貌似LLaMA在MMLU任务也是具备涌现能力的

效果比Chinchilla略好些 +1.8

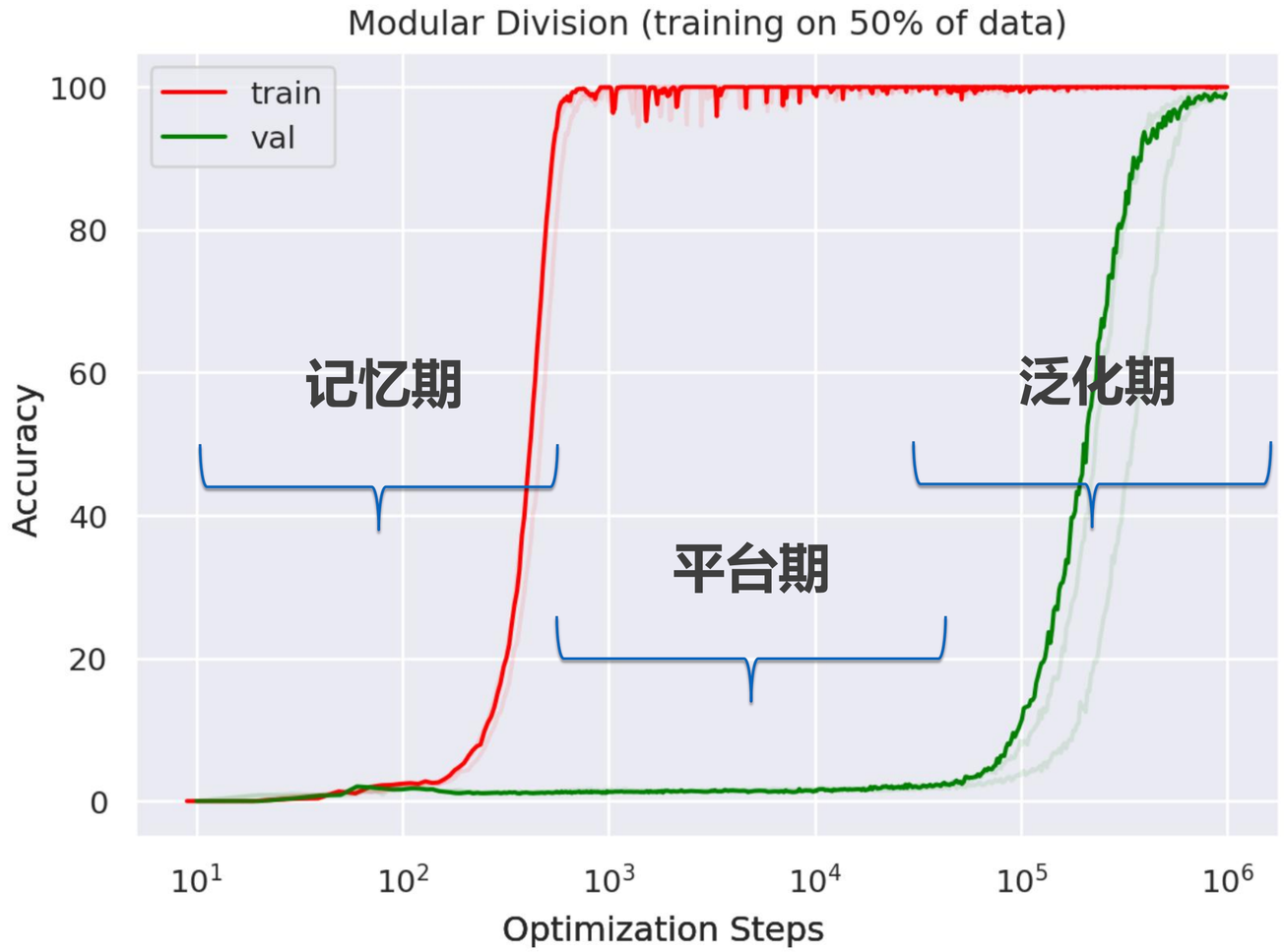
因为涌现能力跟任务相关
所以可能需要对更多任务评估

Outline

模型训练过程中的顿悟现象：Grokking

Grokking：训练数据较少的某些数学任务（取余问题），训练过程三阶段（记忆期、平台期、泛化期）

(a+b) mod c =?



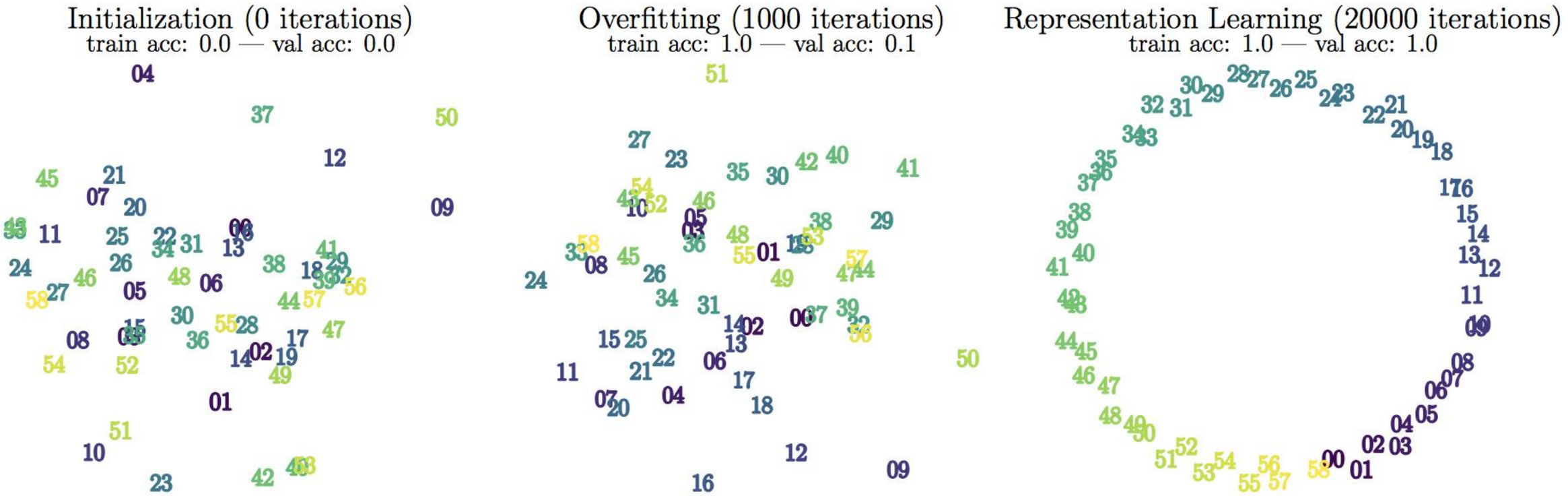
记忆期：这些例子我记住了

平台期：我再想一想

泛化期：我明白规律了

模型训练过程中的顿悟现象：Grokking

Grokking的泛化：目前看是在学习一种输入的好的表征，这种表征可以体现当前学习任务的任务结构



Grokking现象对应的训练动态

模型训练过程中的顿悟现象：Grokking

Grokking可以解释LLM的涌现现象吗？

目前个别研究暗示两者应该存在某种关联，但是并未有研究建立两者之间的直接关系

Grokking看着更像是在训练过程中出现的涌现现象，和LLM的规模怎样建立起联系？

使用Grokking解释LLM涌现的核心在于说清楚：为何规模小的LLM模型不能出现Grokking?

后面会给出一种个人猜测，试图从Grokking角度来解释LLM的涌现现象

Outline

涌现能力原因猜想一：任务评价指标不够平滑

Emoji_movie任务：输入Emoji，要求LLM给出完全正确的电影名称

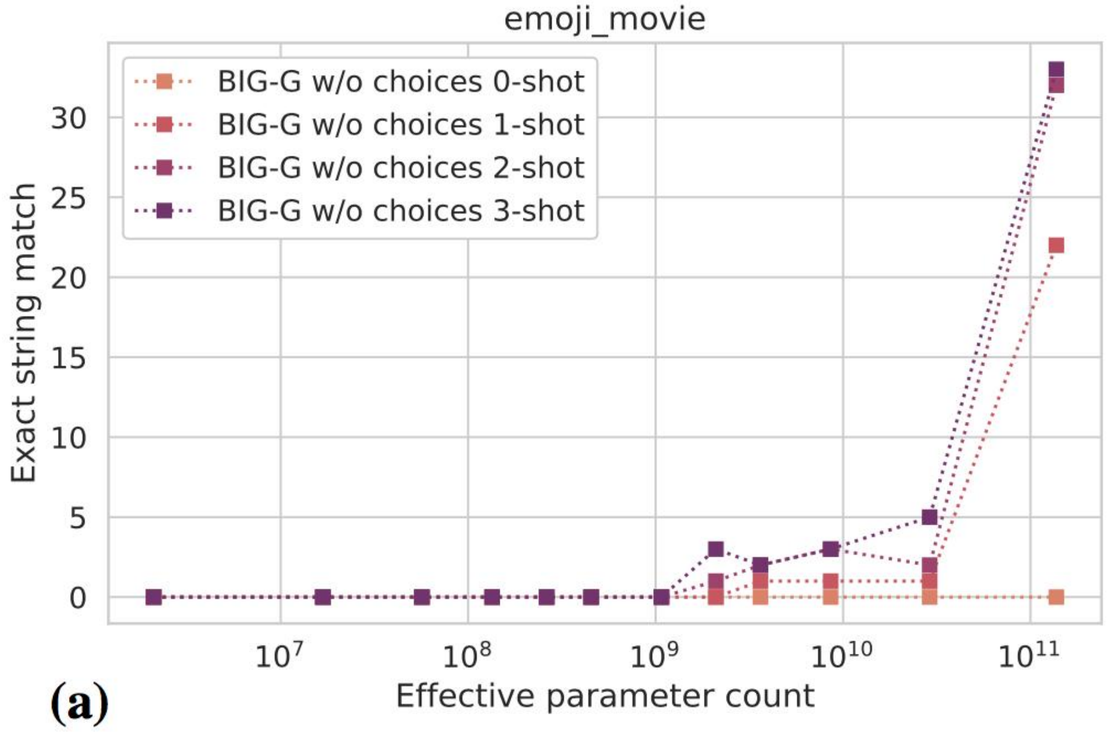
Q: What movie does this emoji describe? 👧🐟🐠🐡

2m: i'm a fan of the same name, but i'm not sure if it's a good idea
16m: the movie is a movie about a man who is a man who is a man …
53m: the emoji movie 🐟🐠🐡
125m: it's a movie about a girl who is a little girl
244m: the emoji movie
422m: the emoji movie
1b: the emoji movie
2b: the emoji movie
4b: the emoji for a baby with a fish in its mouth
8b: the emoji movie
27b: the emoji is a fish
128b: finding nemo

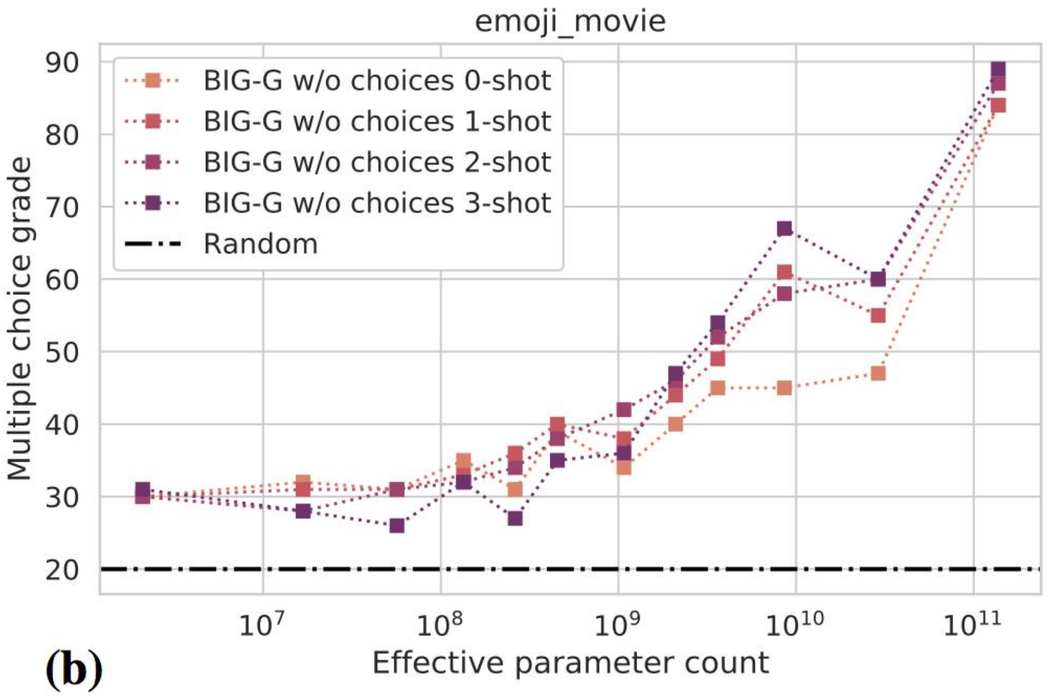
例子:finding nemo(海底总动员)

涌现能力原因猜想一：任务评价指标不够平滑

Emoji_movie任务：输入Emoji，要求LLM给出完全正确的电影名称



精确匹配:涌现现象

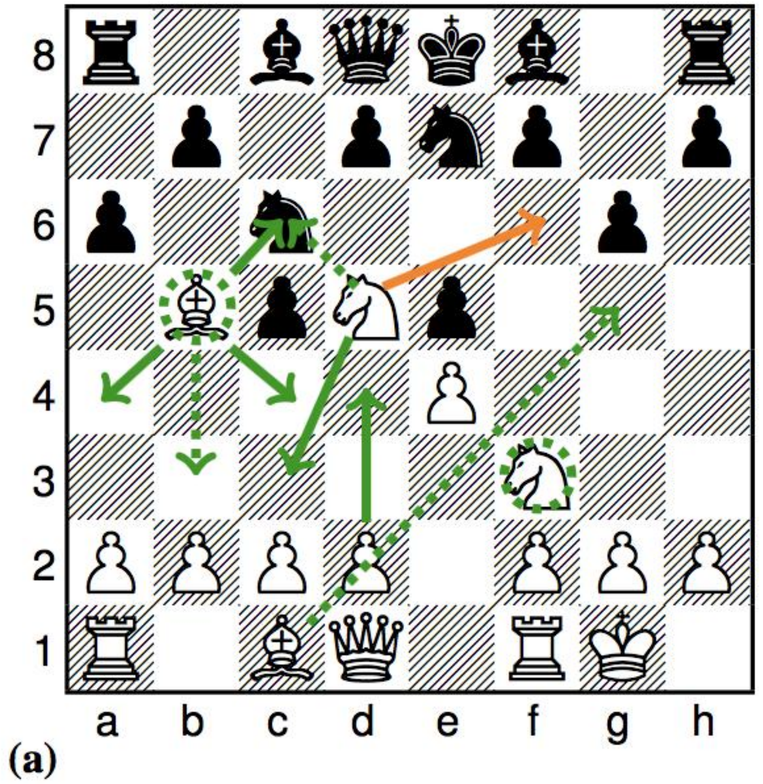


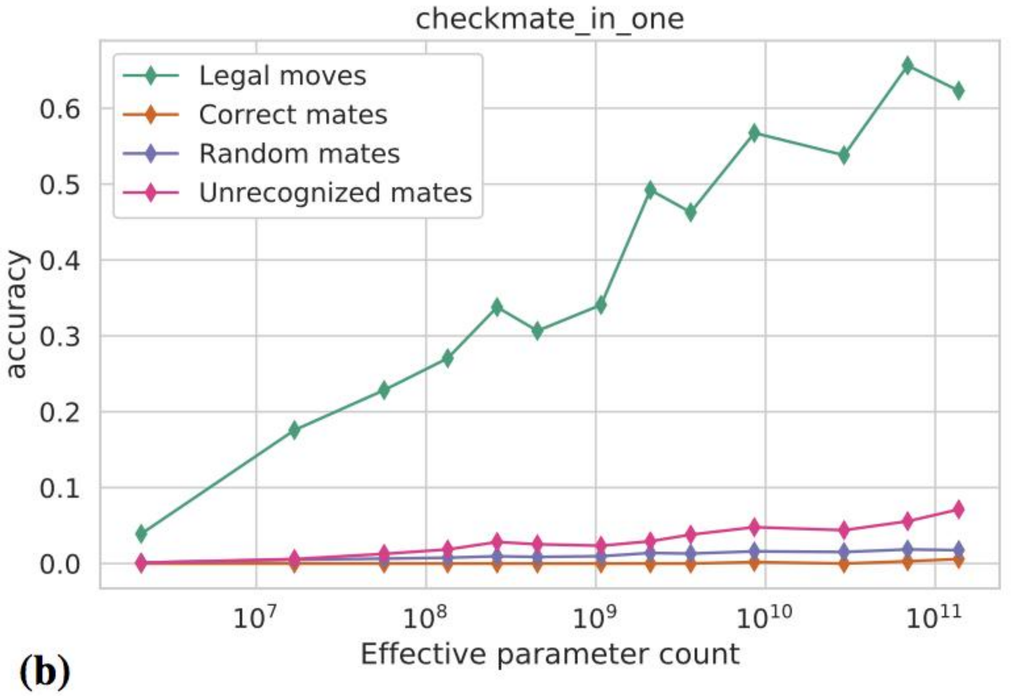
多项选择:scaling law现象

涌现能力原因猜想二：复杂任务 vs 子任务

猜想：最终任务过于复杂，由多个子任务构成，子任务符合Scaling Law，最终任务体现为涌现

假设：最终任务T由5个sub-T构成，每个sub-T从40%提升到60%，最终任务从1.1% 提升到 7.8%



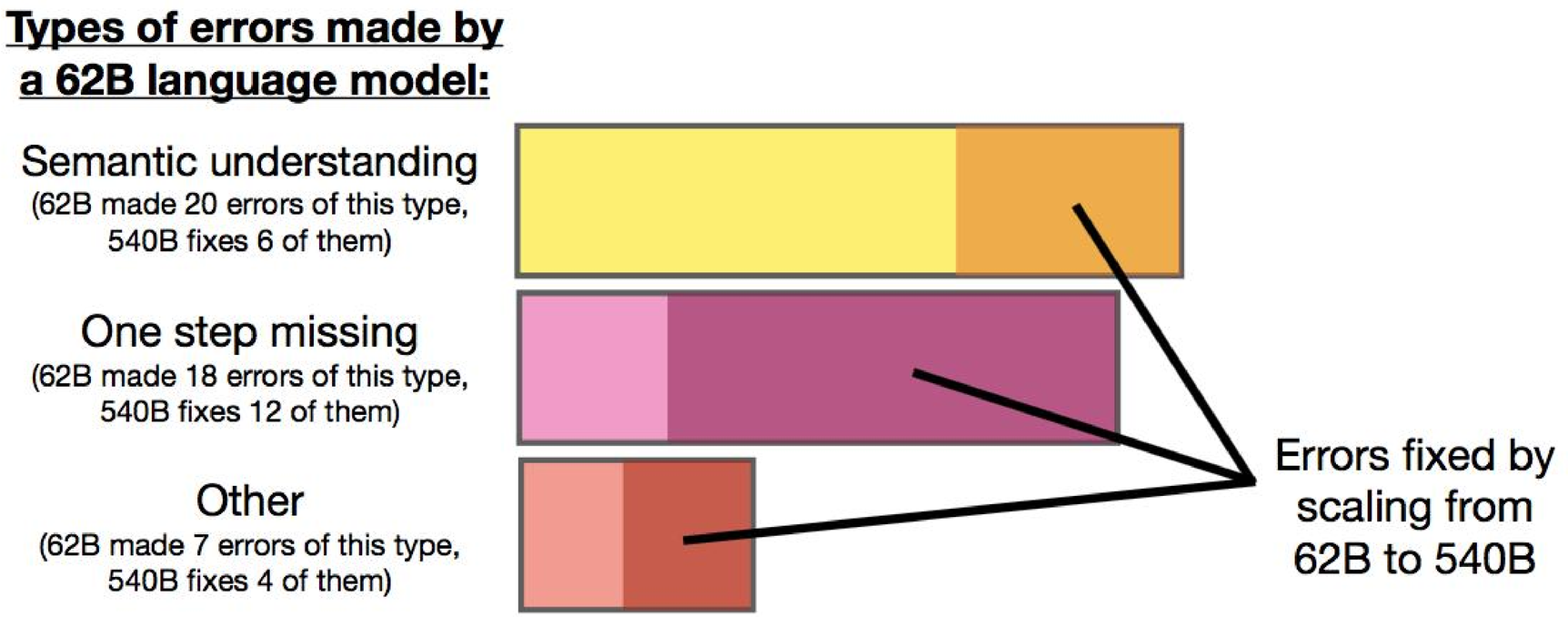


国际象棋:合法移动 vs 将死

合法移动：Scaling Law vs 将死：涌现

涌现能力原因猜想二：复杂任务 vs 子任务

证据：CoT任务，62B和540B模型错误例子分析

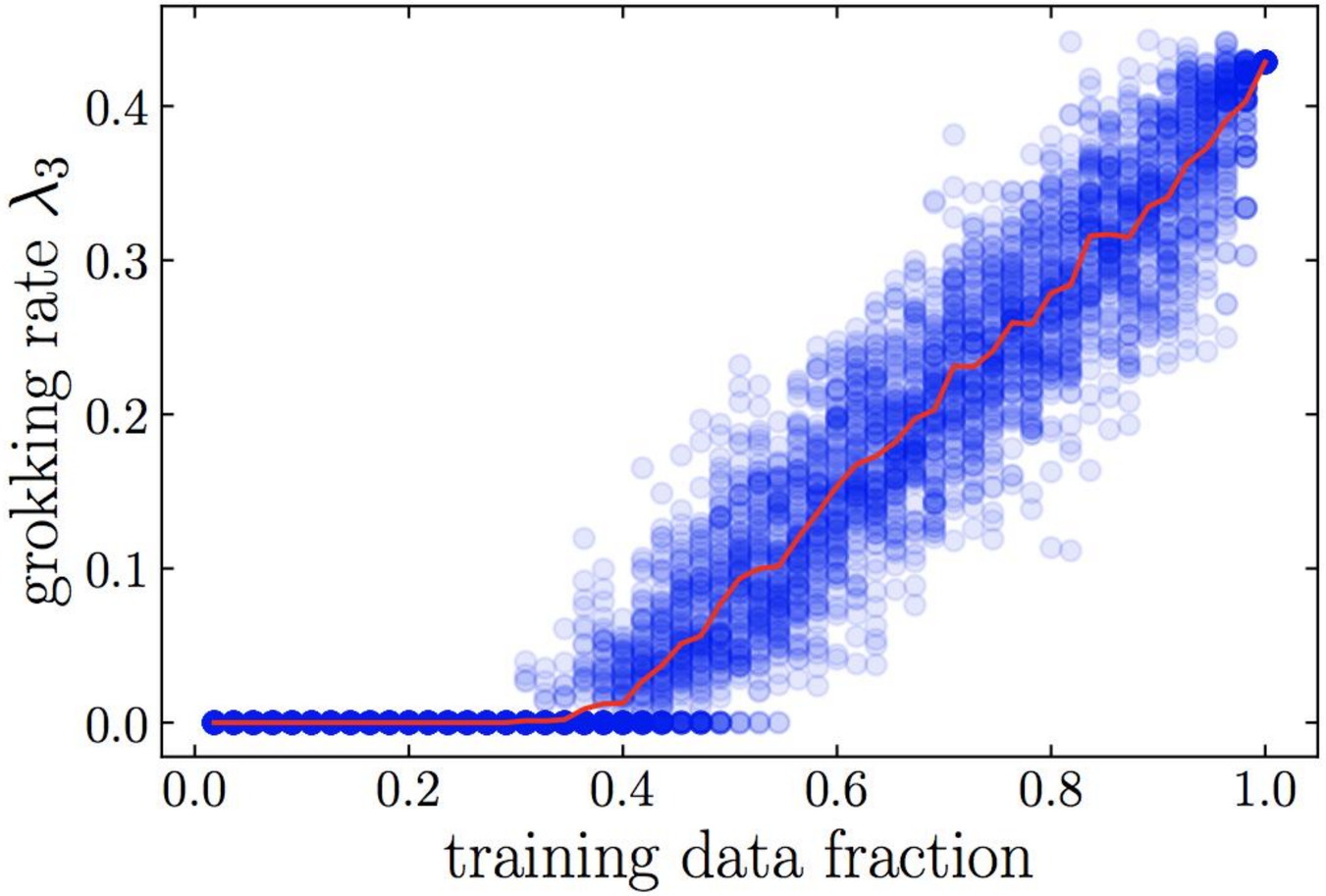


单步推理错误：540B PaLM模型修正了62B模型18个错误中的12个

语义理解错误：540B PaLM模型修正了62B模型20个错误中的6个

涌现能力原因猜想三（个人意见）：用Grokking来解释涌现

事实一：Grokking现象是描述数据量较少ML任务的，但是最少数据量需要达到阈值

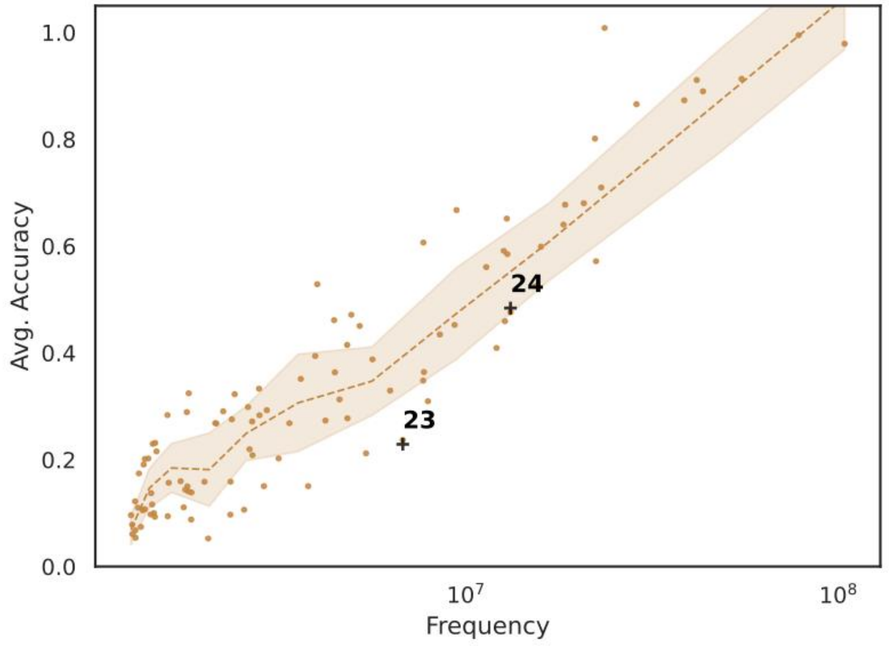


训练集合使用比例对Grokking现象的影响

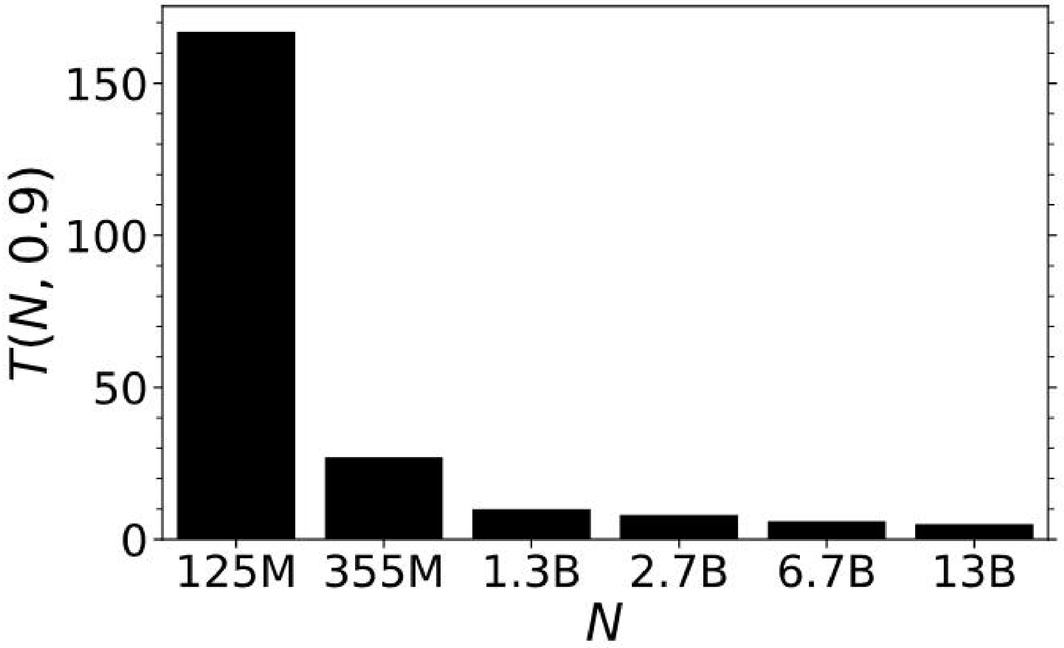
涌现能力原因猜想三（个人意见）：用Grokking来解释涌现

事实二：LLM模型规模越大，记忆数据的能力越强

Q: What is 24 times 18? A: ___ *Model:* 432 ✓
Q: What is 23 times 18? A: ___ *Model:* 462 ✗



乘法运算:出现频次的影响



模型越大记忆能力越强：
T(N,0.9)记住90%训练数据需要看到训练数据次数

涌现能力原因猜想三（个人意见）：用Grokking来解释涌现

问题：为什么当模型规模变大后，有些任务会出现涌现现象？

简单的解释：只利用事实1即可，就是说Grokking的出现要求：最少数据量需要达到阈值

↓

某个任务T，尽管LLM总的训练数据量是足够大的，但是具体到任务T本身，相关数据量其实很少

↓

大规模LLM相对规模小些的LLM来说，增加了训练数据量，所以任务T的训练数据量增加，达到最小阈值

↓

对于任务T来说，当LLM规模达到某个值→T的训练数据超过特定值→发生Grokking→学会规律→效果变好

涌现能力原因猜想三（个人意见）：用Grokking来解释涌现

问题：为什么当模型规模变大后，有些任务会出现涌现现象？

复杂的解释：假设训练数据量保持不变，如何解释？同时利用事实1和事实2

↓

某个任务T，尽管LLM总的训练数据量是足够大的，但是具体到任务T本身，相关数据量S其实很少

↓

大规模LLM相对规模小些的LLM来说，记忆能力强，能有效记忆S中更多训练数据，当模型规模足够大，记忆S中的有效训练数据量，达到最小阈值

↓

对于任务T来说，当LLM规模达到某个值→记忆T的训练数据超过特定值→发生Grokking→学会规律→效果变好

THANKS