Emergent Abilities of Large Language Models

Выполнил:
Разин Арслан Дмитриевич, БПМИ202

1. **Эмерджентность**
2. Метрики
3. Примеры
4. Последствия
5. Критика другой статьи
6. Источники

Эмерджентность

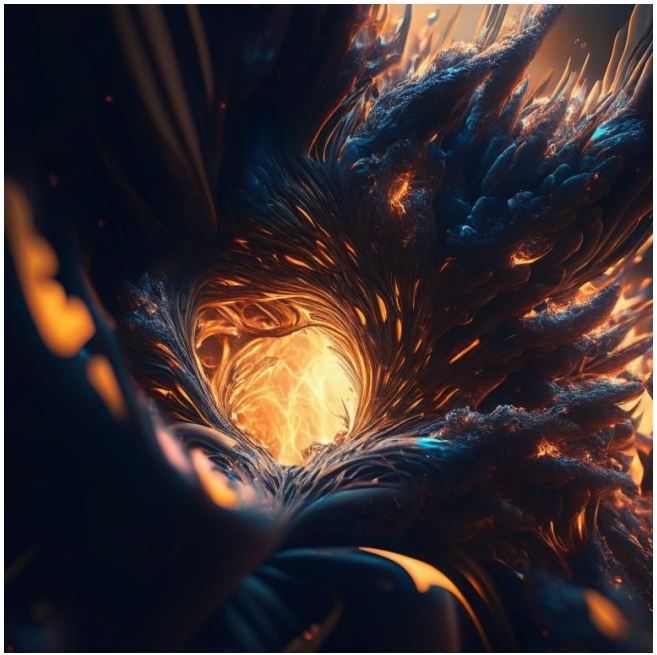
Emergence is when quantitative changes in a system result in qualitative changes in behavior.

An ability is emergent if it is not present in smaller models but is present in larger models.

Эмерджентность в 4k от Kandinsky 3.0

Эмерджентность

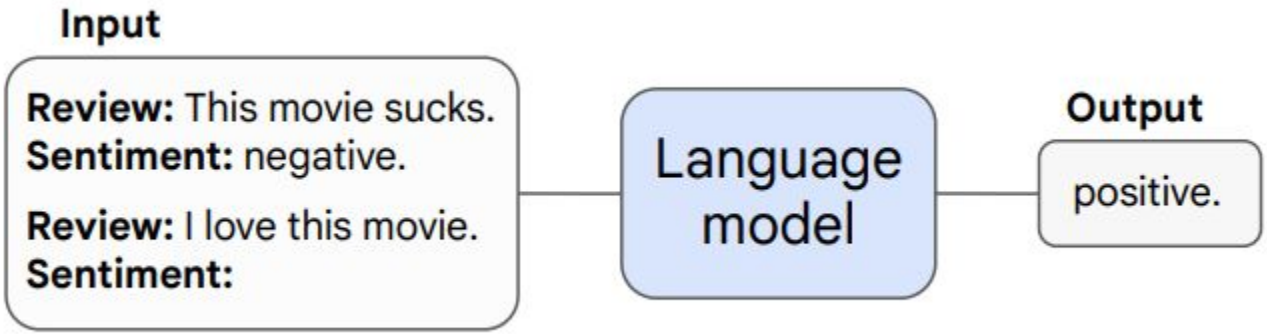
Kandinsky 2.2



Kandinsky 2.1

1. Эмерджентность
2. **Метрики**
3. Примеры
4. Последствия
5. Критика другой статьи
6. Источники

Метрики



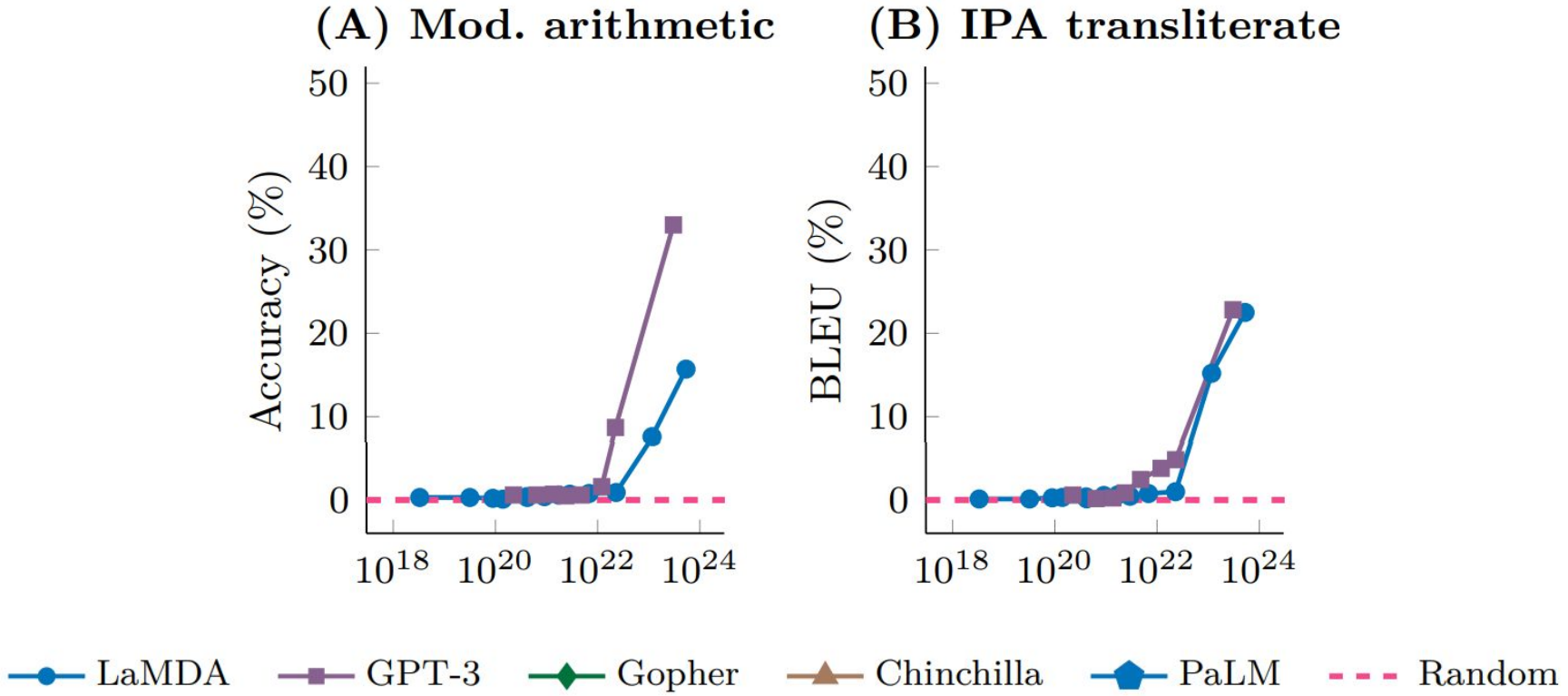
Использование few-shot prompting

Метрики

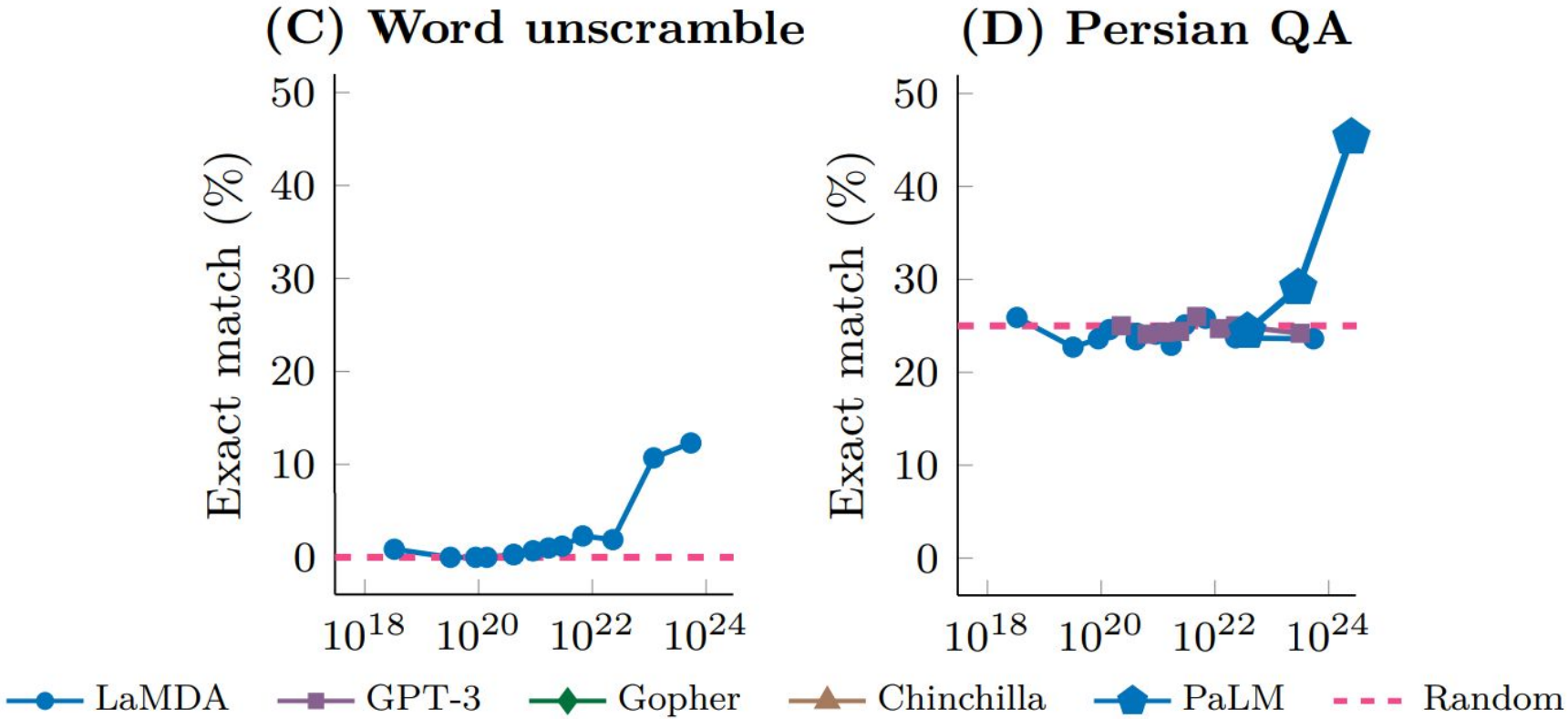
- Mod. arithmetic – простые арифметические операции (2-3 знака, сложение, умножение)
- IPA transliterate – транслитерация
- Word unscramble – расшифровка слова
- Persian QA – вопрос/ответ на персидском
- TruthfulQA – честный вопрос/ответ
- Grounded mappings – выделение концепции
- Multi-task NLU – мультизадачный тест на разные области знаний
- Word in context – понимание значения слова в контексте

1. Эмерджентность
2. Метрики
3. **Примеры**
4. Последствия
5. Критика другой статьи
6. Источники

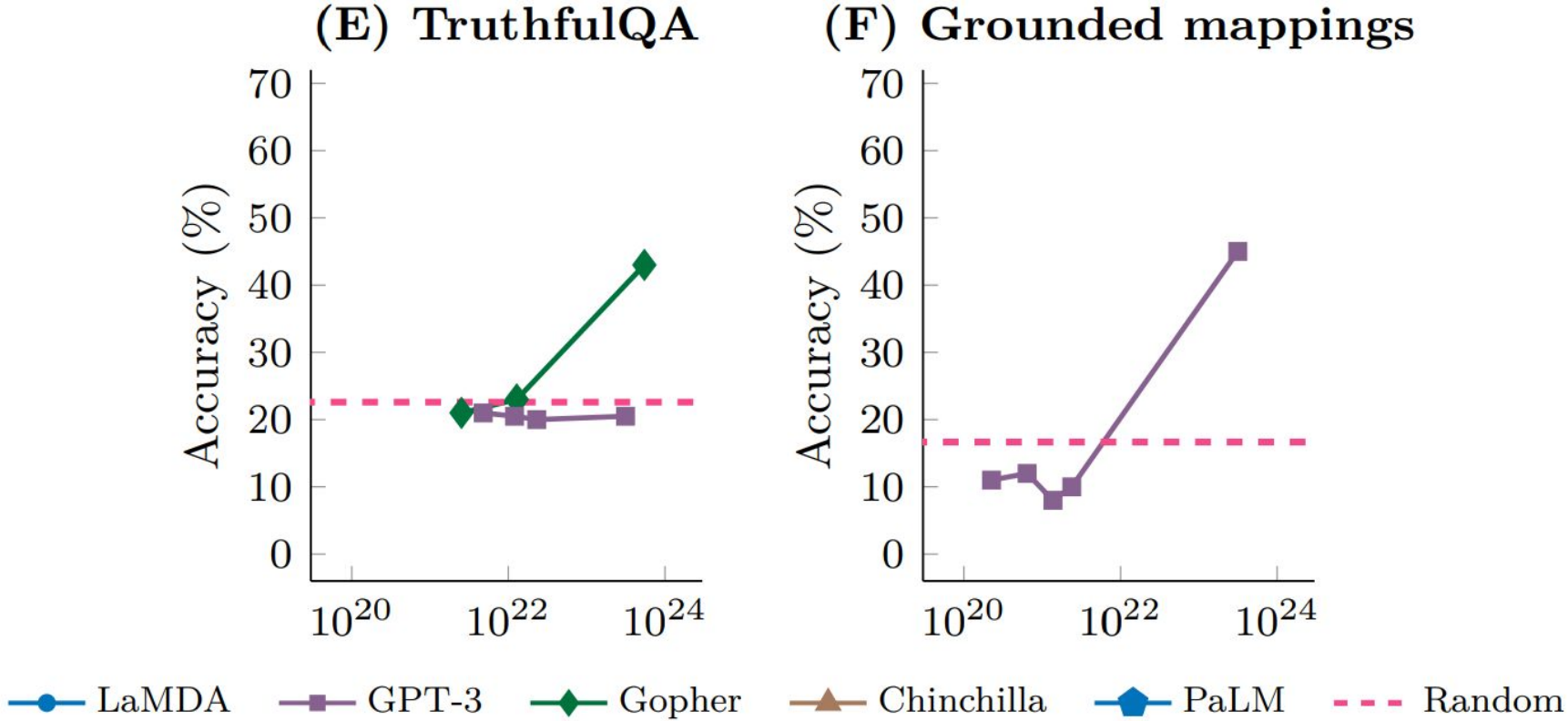
Примеры


(A) Mod. arithmetic
(B) IPA transliterate
Accuracy (%)
BLEU (%)
50
40
30
20
10
0
10^{18}
10^{20}
10^{22}
10^{24}
LaMDA
GPT-3
Gopher
Chinchilla
PaLM
Random

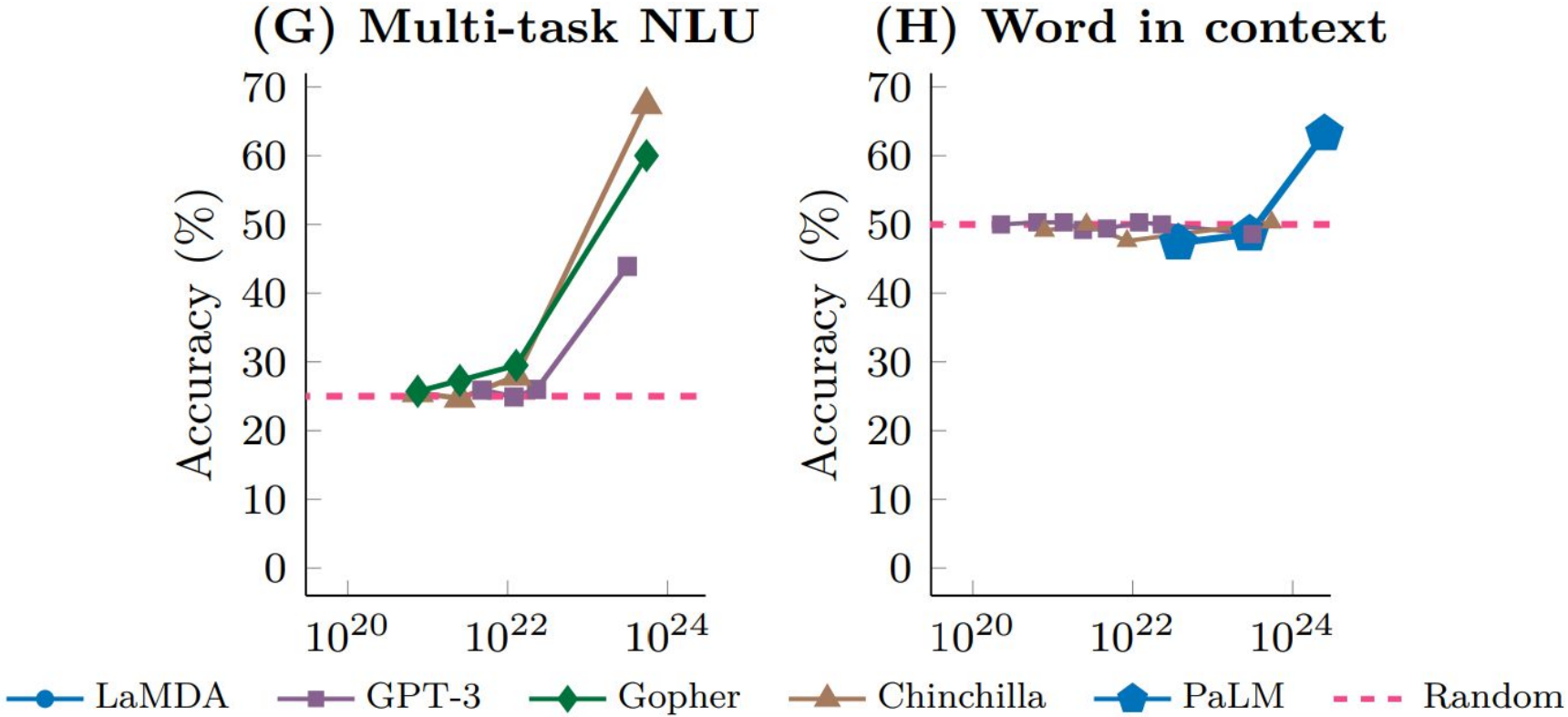
Примеры


(C) Word unscramble
(D) Persian QA
Exact match (%)
50
40
30
20
10
0
10^{18}
10^{20}
10^{22}
10^{24}
LaMDA
GPT-3
Gopher
Chinchilla
PaLM
Random

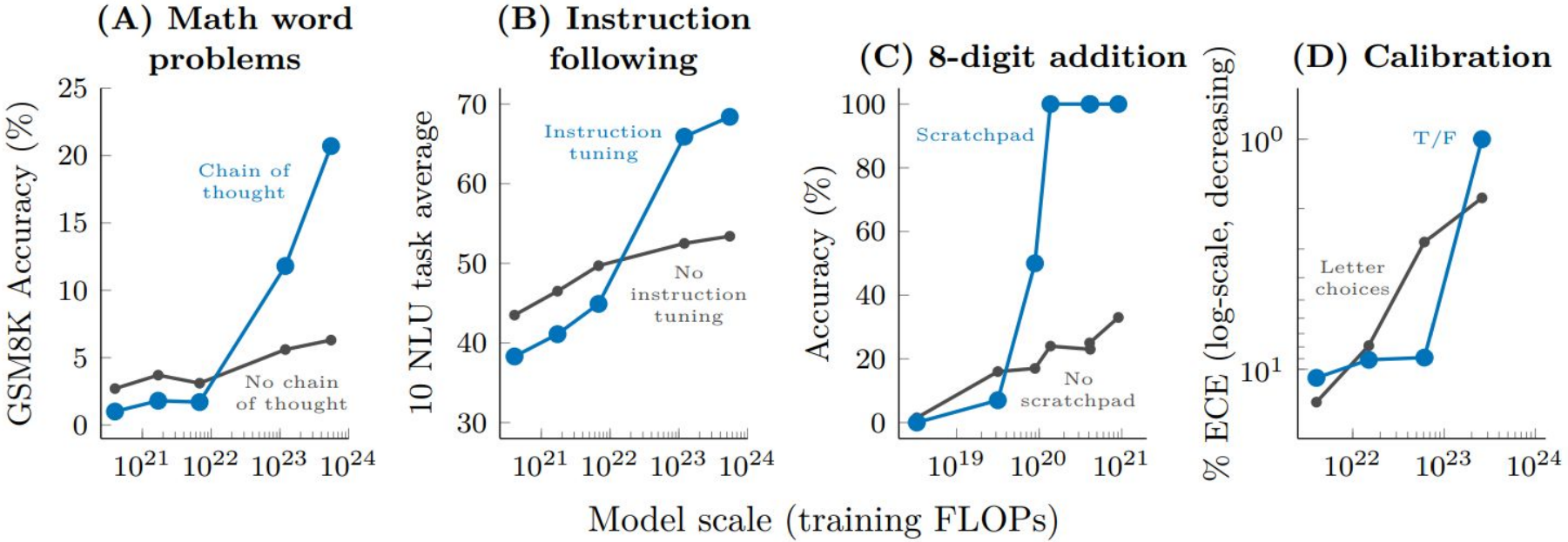
Примеры


(E) TruthfulQA
(F) Grounded mappings
Accuracy (%)
70
60
50
40
30
20
10
0
10^{20}
10^{22}
10^{24}
LaMDA
GPT-3
Gopher
Chinchilla
PaLM
Random

Примеры


(G) Multi-task NLU
(H) Word in context
Accuracy (%)
70
60
50
40
30
20
10
0
10^{20}
10^{22}
10^{24}
LaMDA
GPT-3
Gopher
Chinchilla
PaLM
Random

Продвинутые примеры (Augmented Prompting)



Рассуждение по шагам

Следование инструкциям

Имитация интерпретатора

Самооценка

Обнаруженные авторами примеры

	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)
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)

1. Эмерджентность
2. Метрики
3. Примеры
4. **Последствия**
5. Критика другой статьи
6. Источники

Последствия

Основные тезисы авторов статьи по оценке перспектив:

1. Улучшение существующих приложений и создание новых – эмерджентность позволяет улучшить взаимодействие человека и ИИ, наделяя алгоритмы человеческими качествами, а также заменяя человека в выполнении бытовых и типовых задач
2. Этические и социальные вызовы – проблемы конфиденциальности данных, предвзятости моделей, влияние на рынок труда и потенциальное использование для создания вводящего в заблуждение или манипулятивного контента
3. Необходимость регулирования и стандартов – создание мер по обеспечению прозрачности, ответственности и безопасности
4. Необходимость продолжать исследования – улучшение технических аспектов моделей (например, эффективность, масштабируемость)

1. Эмерджентность
2. Метрики
3. Примеры
4. Последствия
5. **Критика другой статьи**
6. Источники

Критика

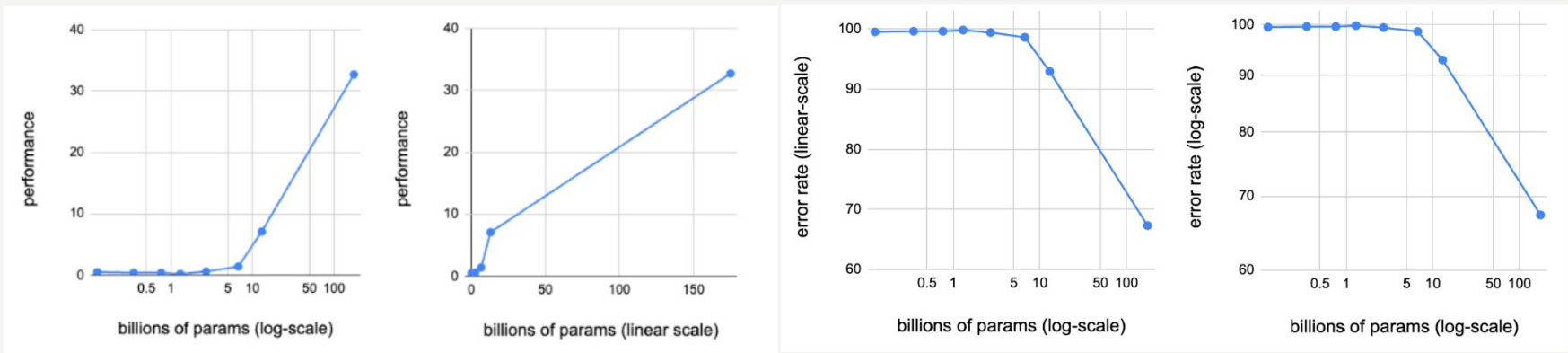
Аргумент: возникающие способности, это следствие неправильных метрик, слишком топорно оценивающих результат. Для оценки качества генерации нужно использовать более сглаженные метрики.

Ответ: единственный ожидаемый ответ на вопрос “Сколько будет 15 + 23?” – “38”. Никакой другой ответ нас не устраивает, поэтому хоть ответ 37 был бы ближе, чем -2.591, но нам важен только конечный результат. С другой стороны использование сглаженных суррогатных метрик тоже важно, так как они позволяют оценить прогресс в достижении нужного эффекта.

Критика

Аргумент: плохо рисовать графики, где по ось X логарифмическая, а Y нет.

Ответ: Ок, легче стало? (это не влияет никак на появление эффекта)



Критика

Аргумент: в статье не исследован вопрос того, является ли повышение точности непрерывным и плавным. Например, кажется маловероятным, что модель с 1 000 000 параметрами будет иметь 50% (случайную) точность, а модель с 1 000 001 параметром будет иметь точность 90%.

Ответ: у нас нет таких моделей, чтобы можно было так точно поймать этот момент. И даже сравнить качество ответов для близких моделей с помощью более простых моделей не всегда возможно, так как не все модели можно одинаково оценивать.

1. Эмерджентность
2. Метрики
3. Примеры
4. Последствия
5. Критика другой статьи
6. **Источники**

Источники

- Статья “Emergent Abilities of Large Language Models”: https://arxiv.org/pdf/2206.07682.pdf
- Статья “Are Emergent Abilities of Large Language Models a Mirage?”: https://arxiv.org/pdf/2304.15004.pdf
- Комментарии авторов первой статьи на вторую: https://www.jasonwei.net/blog/common-arguments-regarding-emergent-abilities