

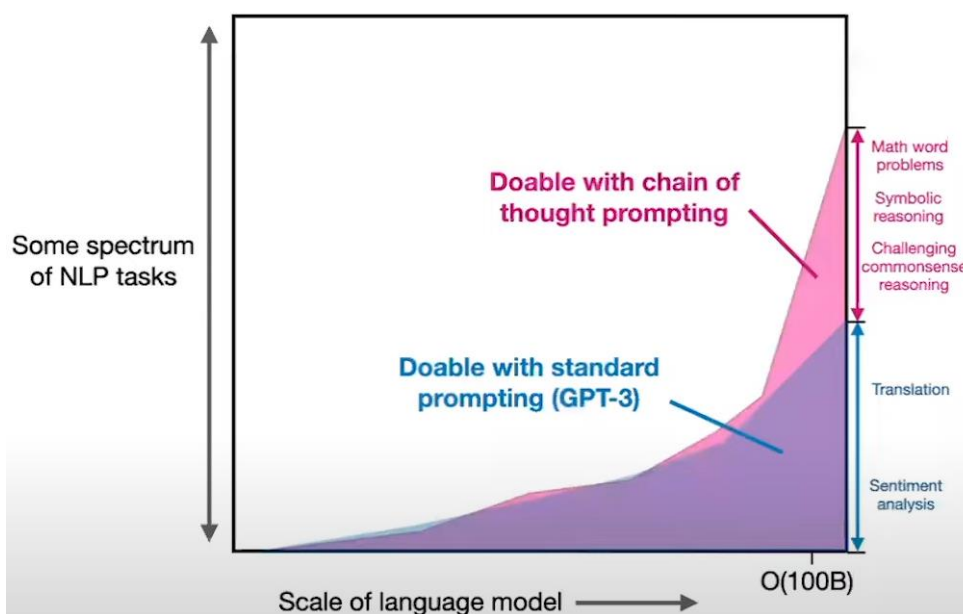
## Part 4. Computational Experiment

Принятое в этом тексте сокращение: CoT – chain-of-thought.

Располагая Open Source линейкой Large Language Models от BigScience, доступной на Huggingface,

<https://huggingface.co/bigscience/bloomz>

мы смогли проверить отдельные факты и наблюдения касательно chain-of-thought - как форме L&R effect-а наиболее сложной и требовательной к размеру модели, начинающей значимо наблюдаться (с ассигасу порядка 40-50%), в моделях размера большего 50B – Fig 1.



Fig, 1 From CS25 I Stanford Seminar - Transformers United 2023: Emergent Abilities and Scaling in LLMs January 24, 2023 Jason Wei (Emergent Abilities and Scaling in LLMs Jason Wei)

Работа, методологии и тестам которой мы следуем [21], была выполнена с использованием InstructGPT-175B (Ouyang et al., 2022) text-davinci-002 (and addition text-davinci-003). Поэтому особенно интересно было сравнить результаты этой работы – как с OpenSource линейкой LLMs Bloom, так и с последней на данный момент версией ChatGPT - GPT 4.

Bloom 176B использовался нами через бесплатно доступный online интерфейс (запуск локально требует порядка 800Гб GPU):

<https://huggingface.co/bigscience/bloom>

Однако boomZ 176B – то есть multitask finetuned версия bloom была отключена для публичного API, однако это как раз пошло на пользу - оказалось более показательным сравнить результаты non multitask finetuned версии opensource bloom 176B модели и самой продвинутой на данный момент closed source модели GPT4.

GPT4 на момент проведения экспериментов не имел публичного API и мы использовали платную подписку ChatGPT Plus в версии GPT4.

Ввиду всех этих ограничений все тесты пришлось проделать в ручном режиме на небольшом количестве примеров – всего было выполнено 400 тестов, которые были также проверены вручную, визуально, что привело к неожиданным дополнительным выводам, изложенным в конце этого раздела. В целом результаты тестирования хорошо коррелируют с результатами из [21] и прочими фактами из литературного обзора.

For arithmetic reasoning, we experiment on GSM8K (Cobbe et al., 2021), one of the most challenging mathematical reasoning benchmarks.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian,  
Jacob Hilton, Reiichiro Nakano, Christopher  
Hesse, and John Schulman. 2021.

Training verifiers to solve math word problems. arXiv preprint  
arXiv:2110.14168.

For Q&A reasoning we experiment on Bamboogle, a dataset of compositional questions constructed by

Press et al. (2022).

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt,  
Noah A Smith, and Mike Lewis. 2022.

Measuring and narrowing the compositionality gap in language  
models. arXiv preprint arXiv:2210.03350.

Мы тестировали chain-of-thought, на всей линейке доступных LLMs, в варианте с 9 few-shot примерами вопросов-ответов и одним последующим вопросом, ответ, на который и ожидается от модели.

Кроме того, демонстрируемые 9 примеров были в 4 вариантах(в наших тестах, в исходных в 9). Приведем по одному примеру из каждого варианта, для прояснения картины демонстрируемых модели тестовых данных:

**STD (Standard prompting):**

Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?

A: 29

**CoT (Chain-of-Thought prompting):**

Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?

A: There were originally 9 computers. For each day from monday to thursday, 5 more computers were installed. So  $4 * 5 = 20$  computers were added. Now  $9 + 20 = 29$  computers are now in the server room. The answer is 29.

**Invalid Reasoning:**

Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?

A: There were originally 9 computers. For each day from monday to thursday, 5 more computers were installed. Now  $9 * 5 = 45$  computers. Since  $4 * 4 = 16$ , now  $45 - 16 = 29$  computers are now in the server room. The answer is 29.

**No\_relevance:**

Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?

A: Haley is currently 23 inches tall. She grows at the rate of 10 inches every year for 4 years. So she will have grown by  $10 * 4 = 40$  inches. Her height after 4 years will be  $23 + 40 = 63$  inches. The answer is 63.

В итоге на моделях:

1) bloomz-560M, bloomz-1B1, bloomz-1.7B, bloomz-3B - эффект CoT не наблюдался. Ответом модели были просто числа, без chain-of-thought, причем и ответ был не верен.

2) bloomz-7B1 – ответом модели были короткие обрывки рассуждений и(или) не верные ответы. Вероятно, это позволяет констатировать первые признаки проявления эффекта CoT, но на модели этого размера - с нулевым по качеству результатом.

3) mt0-xxl 13B – эффект не наблюдался, но сама модель была из другой линейки. Ответом модели были только числа или факты, без chain-of-thought, ответы были не верны.

4) bloom 176B – эффект CoT наблюдался в полном объеме. Рассуждения, доходящие до ответа - но часто ответ был не верен, а рассуждения порой неадекватны.

5) ChatGPT Plus with GPT 4 (размер модели не известен, вероятнее всего не меньше 175B как было в предыдущей версии ) – эффект CoT наблюдался в полном объеме, правильные рассуждения и верный результат в почти во всех случаях.

Сами наши тесты и их результаты можно найти: <https://github.com/VladGKulikov>

Приводим здесь сводную таблицу с нашими тестами – Fig. \_\_\_\_

#### **GSM8K (arithmetic reasoning)**

		Bloom 176B	text-davinci-002	text-davinci-003	ChatGPT4
1	Without any examples of reasoning	0%	-	-	100%
2	STD (Standard prompting)	0%	0%	10%	80%
3	CoT (Chain-of-Thought prompting)	10%	40%	60%	90%
4	Invalid Reasoning	10%	40%	80%	100%
5	No relevance	0%	10%	10%	100%

#### **Bamboogle (Q&A reasoning about facts)**

		Bloom 176B	text-davinci-002	text-davinci-003	ChatGPT4
1	Without any examples of reasoning	40%	-	-	100%
2	STD (Standard prompting)	30%	30%	50%	80%
3	CoT (Chain-of-Thought prompting)	40%	50%	90%	100%
4	Invalid Reasoning	40%	50%	90%	100%
5	No relevance	30%	50%	60%	100%

А также, для полноты картины и для сравнения, таблицы из [21] с тестами для InstructGPT-175B text-davinci-002 and text-davinci-003.

Отметим, что в [21] кроме 4 типов тестов, были еще 5 дополнительных, но мы ограничились 4-мя наиболее показательными. Примеры этих 5 типов демонстраций можно увидеть как в репозитории к работе [21], так и в нашем репозитории.

	GSM8K			Bamboogle	
	Inter. Recall	Inter. F1	Answer Acc.	Inter. Recall	Answer F1
STD (Standard prompting)	N/A	N/A	15.4	N/A	20.6
CoT (Chain-of-Thought prompting)	43.9	48.3	48.5	45.2	45.2
① Invalid Reasoning	39.8	43.9	39.5	44.4	39.4
② No <i>coherence</i> for <i>bridging objects</i>	35.3	39.2	35.8	40.8	37.4
③ No <i>relevance</i> for <i>bridging objects</i>	21.4	26.2	27.5	39.6	34.0
④ No <i>coherence</i> for <i>language templates</i>	24.1	28.3	25.8	35.2	32.1
⑤ No <i>relevance</i> for <i>language templates</i>	29.5	34.0	32.8	40.4	29.4
⑥ No <i>coherence</i>	25.2	29.4	23.1	39.6	33.8
⑦ No <i>relevance</i>	9.6	11.9	11.0	36.8	23.9

Table 2: Intrinsic and extrinsic evaluation results under InstructGPT (text-davinci-002) for all settings in our experiments. Results for text-davinci-003 could be found in Table 6.

	GSM8K			Bamboogle	
	Inter. Recall	Inter. F1	Answer Acc.	Inter. Recall	Answer F1
STD (Standard prompting)	N/A	N/A	15.2	N/A	25.1
CoT (Chain-of-Thought prompting)	48.4	53.1	54.5	61.6	59.5
① Invalid Reasoning	50.2	53.5	51.5	60.8	56.4
② No <i>coherence</i> for <i>bridging objects</i>	46.5	51.5	50.4	59.2	55.2
③ No <i>relevance</i> for <i>bridging objects</i>	32.5	38.3	47.2	60.4	56.9
④ No <i>coherence</i> for <i>language templates</i>	37.8	43.3	41.9	57.2	51.4
⑤ No <i>relevance</i> for <i>language templates</i>	44.6	49.9	51.8	62.4	59.3
⑥ No <i>coherence</i>	34.5	39.4	31.0	57.6	55.2
⑦ No <i>relevance</i>	15.5	17.8	16.2	50.0	49.0

Table 6: Intrinsic and extrinsic evaluation results under InstructGPT (text-davinci-003) for all settings. Discussions are included in Appendix A.3.

Figure:

Arithmetic Reasoning	Multi-hop QA	<p><b>Relevance.</b> A component of the rationale has relevance if it is based on the corresponding component from the query. For bridging objects, this could be formally defined as using the exact same objects mentioned in the query (numbers for arithmetic reasoning and entities for factual QA); for language templates, they have relevance if they are still about the same set of entities/relations as the query, and allude to the question being asked. For example, a template about “Patricia” and “hair” would not have relevance to a query about “Leah” and “Chocolates”, and similarly, a template that attempts to find the “brother-in-law” of the topic entity does not have relevance to a query which seeks the “grandchild” (Table 4).</p> <p><b>Coherence.</b> A component of the rationale has coherence if it is in the correct order, i.e., later steps could not be pre-conditions for earlier steps and reversely, earlier steps could not be based on later steps. For example, a rationale where “32 + 42 = 74” appears before the introduction of “32” or “42” would not have coherence on bridging objects, and similarly for language templates.</p> <p>In what follows, we design a set of ablation settings to examine the impact of these two aspects for different components of a CoT-like rationale.</p>
<p>Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?</p> <p>A: Originally, Leah had 32 chocolates and her sister had 42. So in total they had 32 + 42 = 74. After eating 35, they had 74 - 35 = 39 pieces left in total. The answer is 39.</p>	<p>Q: Who is the grandchild of Dambar Shah?</p> <p>A: Dambar Shah (? - 1645) was the father of Krishna Shah. Rudra Shah was the child of Krishna Shah (? - 1661). So the final answer (the name of the grandchild) is: Rudra Shah.</p>	

Table 1: Bridging objects and language templates of a Chain-of-Thought rationale. Here we illustrate with one in-context exemplar for each task we experiment with.

Как можно видеть из данных тестов, эффект CoT:

- улучшает результат на задачах, которые модель ранее не видела(Bloom) причем в вариантах как правильного CoT, так и *invalid reasoning*(адекватность рассуждений, пусть и с ошибками, важна).
- не мешает, когда модель ранее сталкивалась с такого типа задачами, в вариантах как правильного CoT, так и *invalid reasoning*
- существенно ухудшает результат при **no relevance**(бредовые рассуждения) и **standard prompting**(запрет рассуждений)

### Выводы:

1.Эффект CoT наблюдается нами как на OpenSource bloomz-176B, так и на ChapGPT with GPT 4.0

2.В полном согласии с обзором литературы — эффект CoT эмерджентен и не наблюдается в моделях, до определенного размера модели.

3. Нижняя граница наблюдения признаков эффекта CoT зависит не только от размера модели, но и от самой модели. Признаки эффекта, хотя и слабые, наблюдаемые в bloomz-7B1, не проявляются в mt0-xxl 13B. Как и было описано в изученных источниках.

4. Эффект CoT – помогает в случаях, когда модели ранее не видели подобные задачи.

Но как нами и предполагалось ранее, в разделе о гипотезах, если в модели нет уже сформированного на train шаблона рассуждений, эффект будет проявляться слабо(в случае попадания на схожую область рассуждений внутри языкового пространства рассуждений), либо не будет наблюдаться вообще – что будет проявляться в том, что ответы и рассуждения будут неадекватны вопросу (все это мы и наблюдаем в случае Bloom на GSM8K).

5. Наиболее продвинутая модель, ChatGPT4 – чаще всего просто игнорирует демонстрируемые примеры, причем чем хуже пример (то есть из категории *no relevance*), тем лучше ассигасу. Соответственно - модель явно использует априорное знание и способы рассуждений, которые изначально уже достигают предельного уровня(без необходимости CoT на тестовых примерах).

6. Значимо мешает всем моделям запрет на явные рассуждения, которым и является **standard prompting**

7. Совершенно не адекватные примеры **no relevance** – вредят всем моделям, кроме ChatGPT4.

8. ChatGPT4 хотя и игнорирует полный бред(*no relevance*), но может ухудшить результат при правильных CoT, которые вероятно меньше соответствуют задаче, чем ее собственные цепочки рассуждений. То есть правильные CoT из примеров – позиционируют ChatGPT4 (в 10%случаев) на менее подходящих паттернах рассуждений, чем ее собственные, априорные.

В итоге верные примеры CoT иногда мешают ChatGPT4 на GSM8K.

Исходя из чего, мы считаем, что полученные выводы и прочие имеющиеся наблюдения в достаточной мере свидетельствуют в пользу нашей гипотезы о том, что LLMs не учатся на Inference, а используют шаблоны рассуждений, извлеченные ими из языков на Train stage.

**Кроме того, в процессе вычислительного эксперимента были замечены следующие проблемы с тестированием LLMs:**

1. С помощью ChatGPT4 были выявлены как отдельные ошибки, неоднозначности и не точности формулировок в GSM8K, так и не верные сведения в Q&A тестах. То есть ChatGPT4 демонстрирует, если можно так сказать, экспертный уровень, не только в 100% случаев проходя тесты(если не мешать ей через *standard prompting* CoT), но и показывая ошибки в тестовом материале.

**Example 1**(Outdated or incorrect data):

Question: In what year was the tallest lighthouse completed?

Answer in the test: 1902, Vierge (271 ft)

ChatGPT4's answer:

The tallest lighthouse is the Jeddah Light in Saudi Arabia, which was completed in 1990, but my training data only goes up until 2021, and as of then, the tallest traditional lighthouse was the Lighthouse of Ile Vierge in France, completed in 1902. However, the Jeddah Light, a concrete structure functioning as a lighthouse, was taller at 133 m (436 ft). Please verify this information as it may have changed post 2021.

**Example 2** (Controversial wording of the question):

Question: In a school activity, fifty-four students are to be separated into six groups of equal size. If the activity requires 12 groups, how many more groups are needed to form?

Answer in the test: 3

ChatGPT4's answer:

First, let's find out how many students are in each of the initial groups. If fifty-four students are separated into six equal groups, each group has  $54 / 6 = 9$  students.

The activity requires 12 groups. If six groups are already formed,  $12$  (required groups) -  $6$  (formed groups) =  $6$  more groups are needed.

Please note that this assumes that the additional groups also have 9 students each, so more students may need to be added to the activity.

2. При анализе результатов тестов из [21] и ручном тестировании всех моделей выявилось, что автоматическая проверка на различных CoT для LLMs может быть существенно не точна, поскольку как верные априорные рассуждения моделей могут существенно отличаться от продемонстрированных, так и ответы могут моделей быть более точными, чем сведения в тестах – что приводит к существенному занижению ассурасу в подобных тестах.

3. В одном случае ChatGPT4 – в режиме по relevance, модель расценила ошибочные примеры рассуждений как подлежащие исправлению, правильно исправила их все и правильно ответила на заданный после этих примеров вопрос, что при автоматическом тестировании могло бы быть расценено как ошибочный ответ, поскольку ответ был слишком длинным и корректное решение упомянуто в конце длинного ответа с исправлениями 9-ти по relevance примеров.

**Example 3.**

Ответ ChatGPT4 на вопрос 6 в файле no\_relevance\_davinci-002-003-GPT40-BloomZ.txt

**Исходя, из чего мы формулируем следующее предложение по тестированию LLMs, которое дополнительно будет вынесено в заключение:**

*Перед тестированием LLMs на тестах, включающих сложные рассуждения и факты, предварительно необходимо проверить правильный ли ответ на вопрос содержится в ваших тестах, с помощью этой же LLMs.*

*Когда вы тестируете что-то чьи знания и способности рассуждать могут превосходить уровень тестов, или как минимум равны им, то стоит уделить внимание, тем вопросам в тестах, ответ на которые LLM считает неверным и понять почему это происходит.*

*Поскольку ошибок будет не столь много, это позволит исследователям вручную оценить найденные противоречия и если будет необходимо, то исправить или исключить ошибочные тесты, что повысит достоверность тестирования.*

*Иначе может отказаться, что используемая вами автоматическая система тестирования существенно недооценит возможностей тестируемой LLM.*