

Vladlen Kulikov

Large Language Models. Learning and Reasoning at the Inference Stage

Master's Thesis

Advisor: Radoslav Neychev

MIPT. MSAI.

Master's Thesis Report 1

30.04.2023

I'm sorry – 7 min instead of 5. *25 slides.*

First, briefly, for listeners unfamiliar with the Large Language Models (LLMs) field and terminology:

What is the object of our study?

This is the possibility of LLMs (we called this *the L&R effect*), based on instructions and examples **at the Inference stage**, to solve certain types of problems, without any additional training of LLMs for these tasks.

That is, a model that *hasn't been explicitly trained* to solve certain types of tasks, suddenly becomes capable of solving those tasks under certain conditions.

This is the L&R effect in the form of **in-context learning**:

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 cheese => ..... ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 sea otter => loutre de mer ← example
3 cheese => ..... ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 sea otter => loutre de mer ← examples
3 peppermint => menthe poivrée ←
4 plush girafe => girafe peluche ←
5 cheese => ..... ← prompt
```

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.



Figure 1: (Figure 2.1 from [1]): Zero-shot, one-shot and few-shot, contrasted with traditional fine-tuning.

This is the
L&R effect in
the form of
*chain-of-
thought*:

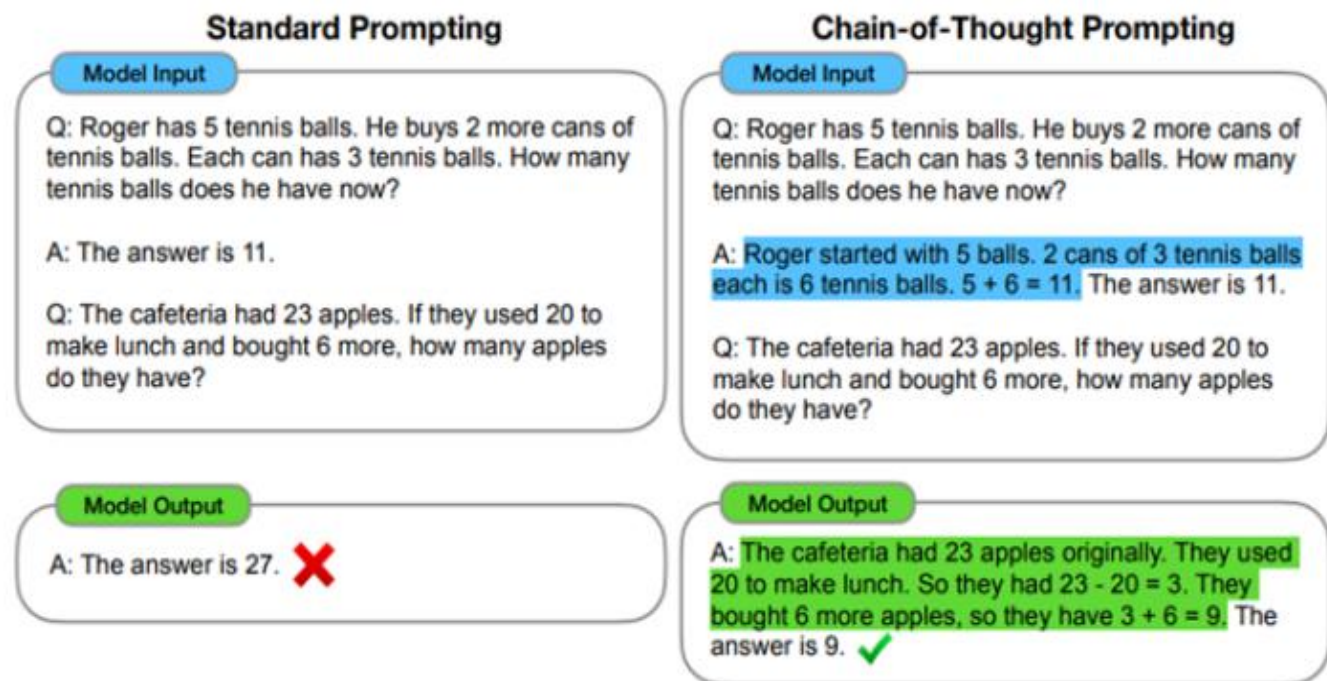


Figure 3: (Figure 1 from (Wei et al., 2022)). Chain-of-thought reasoning processes are highlighted.

Problem statement:

We were interested in the following questions:

I. What is necessary for the emergence (occurrence) of the L&R effect?

i.e.,

What conditions are required for L&R effect in LLMs?

II. L&R effect - What is it?
(No one knows this today.).

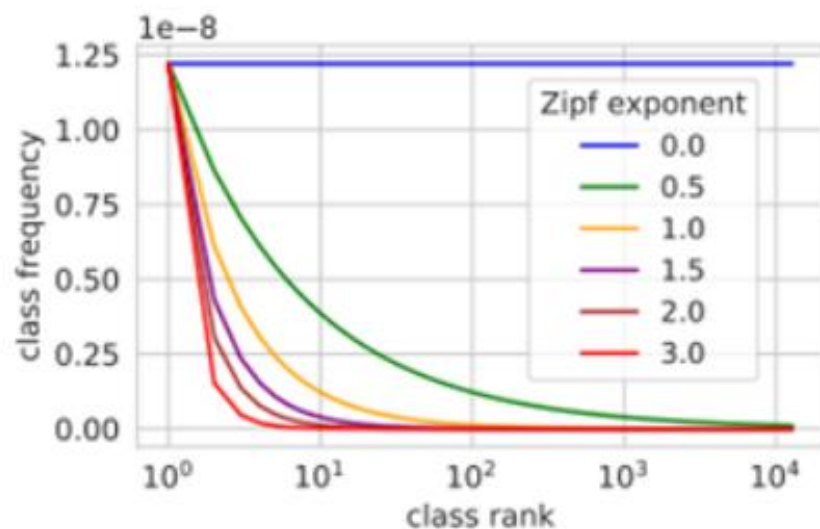
III. Do LLMs learn and reason at the Inference?
Or is it something else?



To answer these questions, we studied:

1. How does Language as a training dataset affect?

(a) Examples of Zipfian distributions.



(b) Distribution of tokens in a natural language corpus.

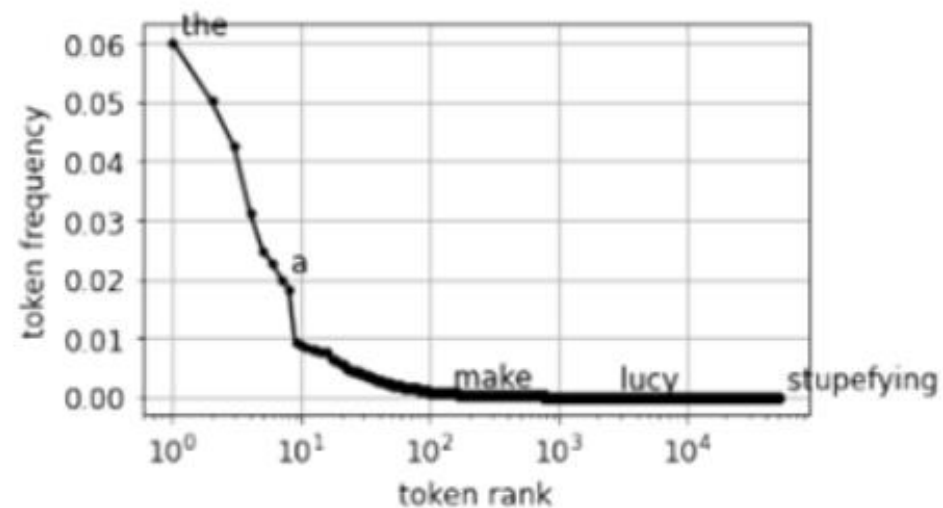


Figure 4: (Figure 6a, 6b from (Chan et al., 2022))

2. How does Model's type affect?

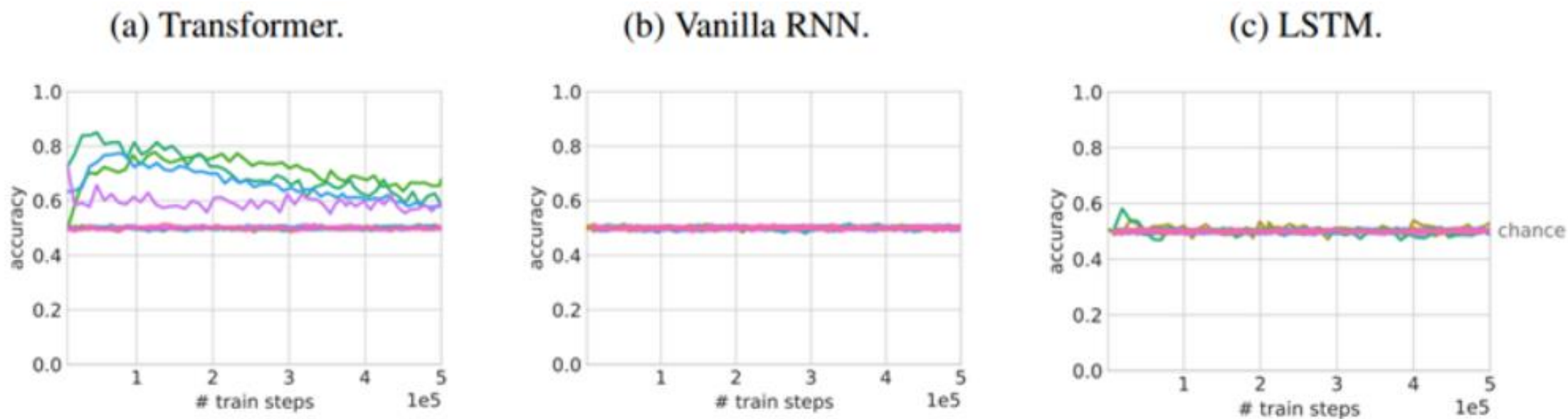


Figure 5: (Fig. 7 from (Chan et al., 2022)) In-context learning in transformers vs. recurrent architectures.

3. How does Number of examples affect?

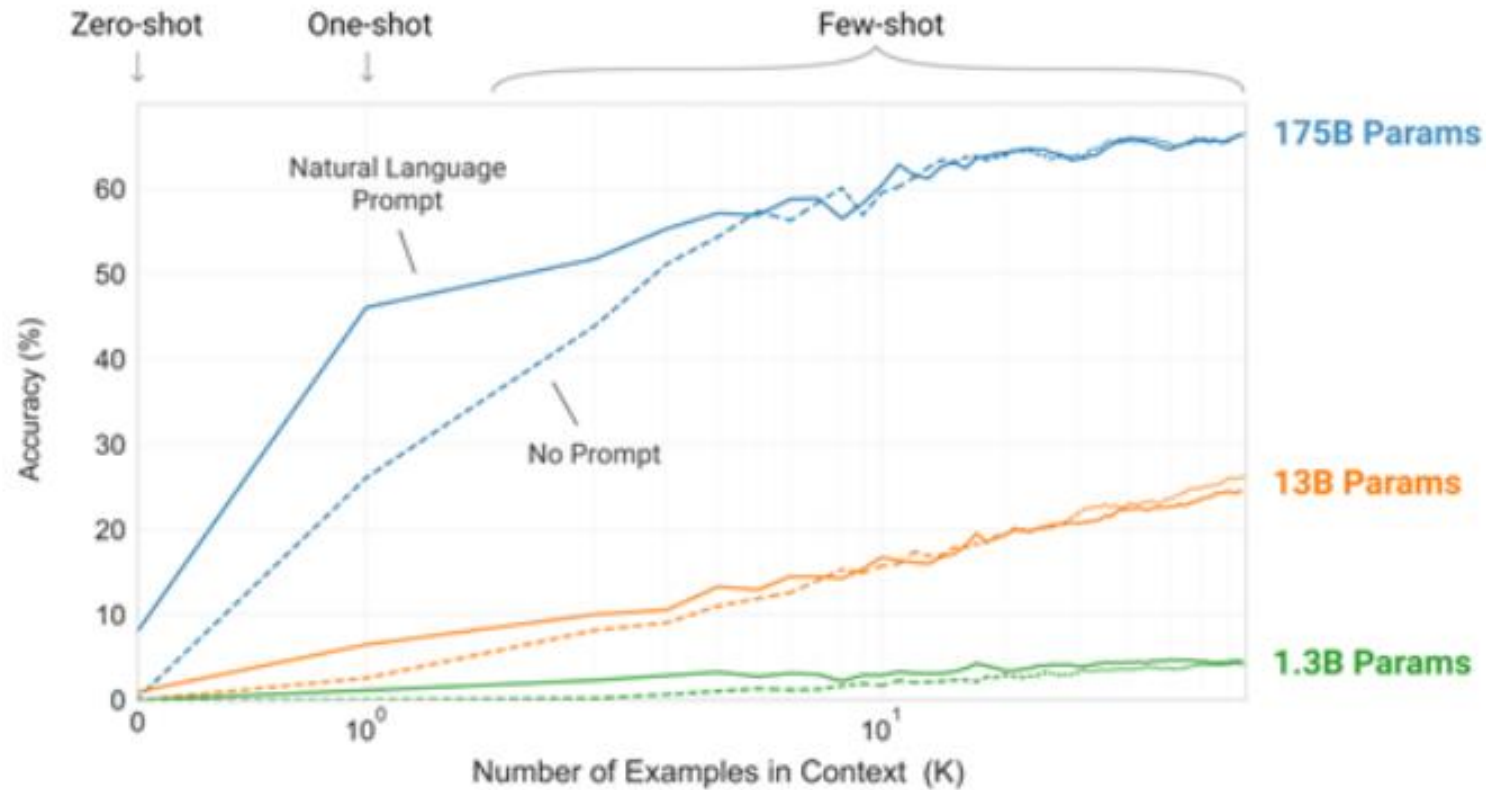
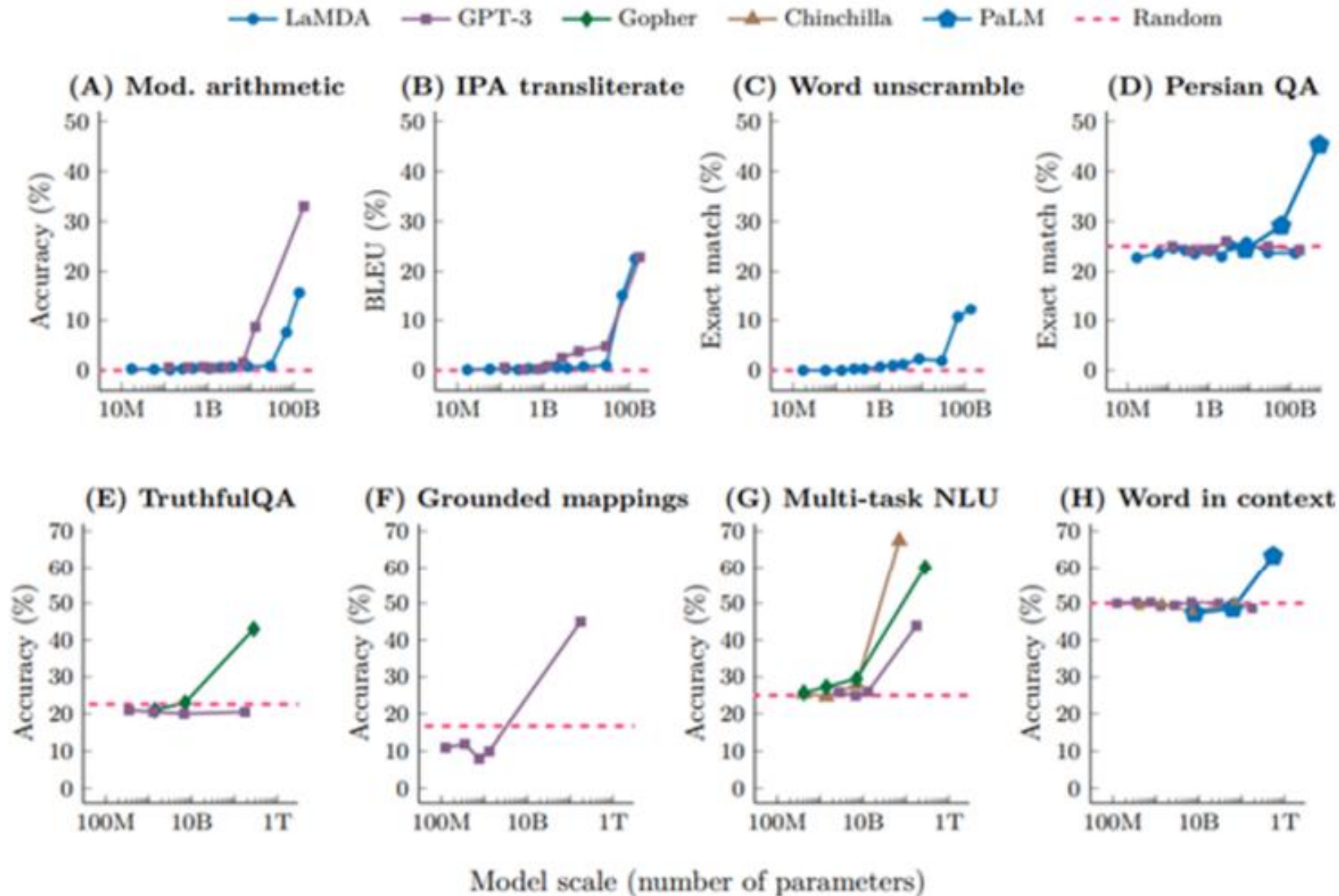


Figure 7: (Figure 1.2. from (Brown et al., 2020)) Larger models make increasingly efficient use of in-context information.

4. How does Model's size affect?



5. How does Training dataset's size affect?

Model	Parameters	Train tokens	Train FLOPs
GPT-3	125M	300B	2.25E+20
	350M	300B	6.41E+20
	760M	300B	1.37E+21
	1.3B	300B	2.38E+21
	2.7B	300B	4.77E+21
	6.7B	300B	1.20E+22
	13B	300B	2.31E+22
	175B	300B	3.14E+23
LaMDA	2.1M	262B	3.30E+18
	17M	313B	3.16E+19
	57M	262B	8.90E+19
	134M	170B	1.37E+20
	262M	264B	4.16E+20
	453M	150B	4.08E+20
	1.1B	142B	9.11E+20
	2.1B	137B	1.72E+21
	3.6B	136B	2.96E+21
	8.6B	132B	6.78E+21
	29B	132B	2.30E+22
	69B	292B	1.20E+23
Gopher	137B	674B	5.54E+23
	417M	300B	7.51E+20
	1.4B	300B	2.52E+21
	7.1B	300B	1.28E+22
Chinchilla	280B	325B	5.46E+23
	417M	314B	7.86E+20
	1.4B	314B	2.63E+21
	7.1B	[sic] 199B	8.47E+21
PaLM	70B	1.34T	5.63E+23
	8B	780B	3.74E+22
	62B	780B	2.90E+23
Anthropic LM	540B	780B	2.53E+24
	800M	850B	4.08E+21
	3B	850B	1.53E+22
	12B	850B	6.12E+22
	52B	850B	2.65E+22

Figure 12: (Tab. 2 from (Wei,Tai et al., 2022)) Parameters, training examples, and training FLOPs of LLMs.

6. How does Dataset's composition affect?

2.6 Dataset composition

Let's analyze the composition of the dataset for some of the models that appear in Figure 10 and Figure 11.

GPT-3:

The model that most often shows the best results in Figure 11, was trained on a large unlabeled text corpus – Figure 13. To what extent this factor can determine the manifestation of emergence is the material for a separate large study.

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Figure 13: (Table 2.2 from [1]) Datasets used to train GPT-3

PaLM:

PaLM train dataset include large multilingual corpus, text from more than 100 languages – Figure 14.

Total dataset size = 780 billion tokens	
Data source	Proportion of data
Social media conversations (multilingual)	50%
Filtered webpages (multilingual)	27%
Books (English)	13%
GitHub (code)	5%
Wikipedia (multilingual)	4%
News (English)	1%

Figure 14: (Table 2 from [5]): Proportion of data from each source in the training dataset.

LaMDA:

It is initially focused on dialogue and therefore has a specific composition of dataset – Appendix E from [7]:

«Pre-training data composition of LaMDA: The pre-training data, called In-finiset, is a combination of dialog data from public dialog data and other public web documents. It consists of 2.97B documents and 1.12B dialogs with 13.39B utterances. The composition of the data is as follows: 50% dialogs data from public forums; 12.5% C4 data [11]; 12.5% code documents from sites related to programming like Q&A sites, tutorials, etc.; 12.5% Wikipedia (English); 6.25% English web documents; and 6.25% non-English web documents. The total number of words in the dataset is 1.56T.»

According to Figure 11 (Table 1 from [2]) – LaMDA as well as GPT-3, it often demonstrates good results for augmented prompting abilities.

Gopher:

Gopher MassiveText data makeup - Figure 15.

	Disk Size	Documents	Tokens	Sampling proportion
MassiveWeb	1.9 TB	604M	506B	48%
Books	2.1 TB	4M	560B	27%
C4	0.75 TB	361M	182B	10%
News	2.7 TB	1.1B	676B	10%
GitHub	3.1 TB	142M	422B	3%
Wikipedia	0.001 TB	6M	4B	2%

Figure 15: (Table 2 from [3]) Gopher MassiveText data makeup

For each subset of MassiveText, authors list its total disk size, its number of documents, and its number of SentencePiece tokens. During training authors sample from MassiveText non-uniformly, using the sampling proportion shown in the right-most column.

7. When LLMs are given examples in context, they draw conclusions based on "rules" (using some "reasoning"). Why? What is it?

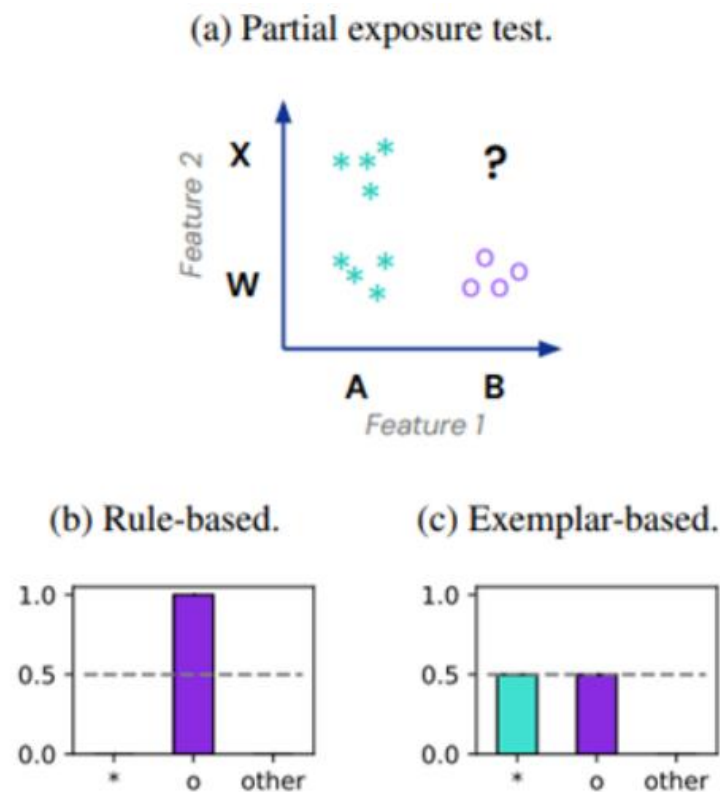


Figure 17: (Figure 1 from (Chan et al., 2022b)) Partial exposure test for differentiating rule-based vs exemplar-based generalization. Stimuli have two features.

8. Why does Chain-of-thought work?

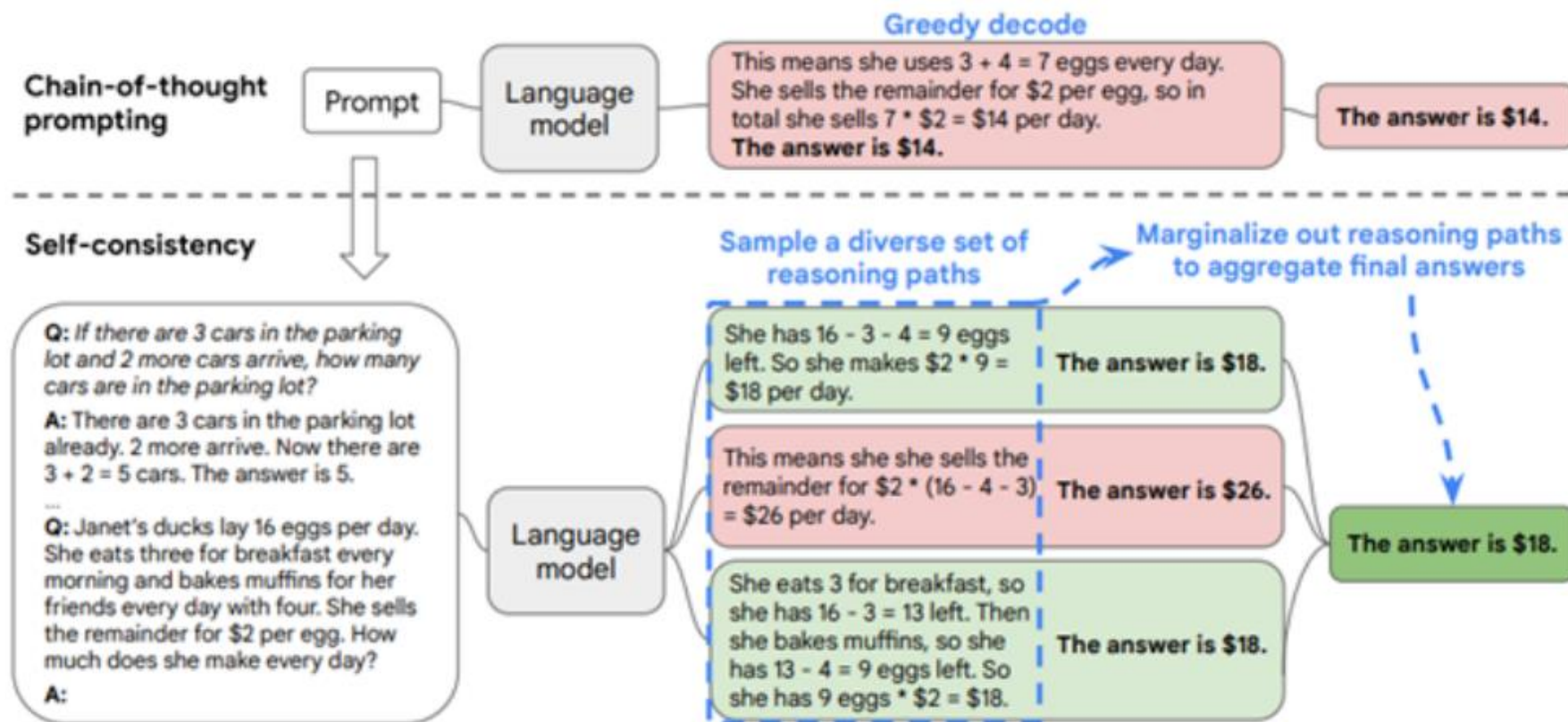


Figure 20: (Fig. 1 from (Wang et al., 2022b)) CoT and the self-consistency method.

9. Why do LLMs "learn" with invalid examples of reasoning?

	In-context Demonstration	Inference by LLM
Query	Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?	Julie is reading a 120-page book. Yesterday, she was able to read 12 pages and today, she read twice as many pages as yesterday. If she wants to read half of the remaining pages tomorrow, how many pages should she read?
Standard	The answer is 39.	The answer is 18. ✗
CoT	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.	Julie is reading a 120-page book. Yesterday, she read 12 pages and today, she read 24 pages. So she read a total of $12 + 24 = 36$ pages. Now she has $120 - 36 = 84$ pages left. Since she wants to read half of the remaining pages, she should read $84 / 2 = 42$ pages. The answer is 42. ✓
Invalid Reasoning	Originally, Leah had 32 chocolates and her sister had 42. So her sister had $42 - 32 = 10$ chocolates more than Leah has. After eating 35, since $10 + 35 = 45$, they had $45 - 6 = 39$ pieces left in total. The answer is 39.	Yesterday, Julie read 12 pages. Today, she read $12 * 2 = 24$ pages. So she read a total of $12 + 24 = 36$ pages. Now she needs to read $120 - 36 = 84$ more pages. She wants to read half of the remaining pages tomorrow, so she needs to read $84 / 2 = 42$ pages tomorrow. The answer is 42. ✓

Method	Answer Acc. (Blue)	Inter. F1 (Dark Blue)
Standard	~15	~15
CoT	~48	~48
Invalid Reasoning	~40	~44

Figure 21: ((Fig. 1 from (Wang et al., 2022))) Results of standard prompting, Chain-of-Thought (CoT) prompting, and our ablation setting with invalid reasoning (§4). We show one demonstration example and one inference example for

If anyone is
confused 😊 – here
is the contents of
our work:

Everything is
almost done.

We just need to finish the
Computational experiment
and added it in the thesis file.

Contents

1	Introduction	3
2	Literature overview	7
2.1	Language as a training set	8
2.2	Model's type	10
2.3	Zero, One, Few-shot	12
2.4	Model's size	14
2.5	Training dataset's size	18
2.6	Dataset composition	20
2.7	Rule-like generalization	24
2.8	Chain-of-thought	26
2.9	What do models understand in their prompts?	28
3	Analytical part	31
3.1	About one particular case: The unreasonable effectiveness of few-shot learning for machine translation	32
3.2	Hypothesis № 1	33
3.3	All the facts together	34
3.4	Hypothesis № 2 (Main hypothesis)	39
3.5	Do LLMs learn? Do LLMs learn to reason?	41
4	Computational experiment	43
5	Conclusion	44
	Bibliography	46

And I'm going to skip
the details from *Analytical part*
(*let there be something to tell on the Report 2*)
and give you the ***Conclusion*** right now,
just show you at first 2 examples
from the *Computer experiment* part.

1.2.ChatGPT with chain-of-thought:

IR

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?



A: The cafeteria started with 23 apples. They used 20 to make lunch, so they have $23 - 20 = 3$ apples left. Then, they bought 6 more, so they have $3 + 6 = 9$ apples now. The answer is 9.



2.2. ChatGPT uses the reasoning pattern (about odd and even numbers) from Train stage, which is not in the demonstration itself, at the Inference stage .



We have numbers from 1 to 10.
This numbers belong to 2 classes:
(1 , class A)
(2, class A)
(3 , class A)
(5, class A)
(6 , class B)
(7 , class B)
(8 , class B)
(10 , class B)

(4, class ?)
give me most likely answer.
Explain



Based on the given information, the most likely answer is that the number 4 belongs to "class A".



The numbers in "class A" are all consecutive odd integers between 1 and 5, while the numbers in "class B" are all consecutive even integers between 6 and 10. Therefore, based on this pattern, we can conclude that 4 is more likely to belong to "class A" because it is a consecutive odd integer between 1 and 5.

However, it is important to note that the classification of the numbers may depend on the specific context or criteria being used. Without additional information or context, we cannot be certain about the class of 4.

Part 5. Conclusion

5.1. For getting L&R effect in the LLMs we need:

1. Training dataset: Language-like dataset.
2. Type of a model: Transformer.
3. Size of a model: From 7 to 8 billion (B) parameters for certain rare tasks (but for accuracy $>50\%$ $> 40-60B$, on the average).
4. Number of examples to train: From zero to few-shot with increasing quality.
5. Size of Dataset: $\sim 300B$ of training tokens and the size of the model is more significant.

5.2. Also, we summarize:

1. Observations and experiments, showing the predominant importance of train stage learning for LLMs.
2. Facts that validity of reasoning is only a small part of the value for performance – pathologically incorrect examples are usually no worse than true ones, while relevance to the input query and following the order of reasoning steps are the key to efficiency.

Following other researchers ([15], [16], [19]), we state, that LLMs at the inference stage don't learn (in the classical sense of the notion of learning), but their **ability** to learn and reasoning is limited by the **abilities** acquired during the train stage.

5.3. None of this, however, diminishes the extremely high importance of:

1. Studying and applying prompt engineering methods for extracting from LLMs at the inference stage language reasoning patterns (formed in LLMs at the train stage).
 2. Research toward a real ability to learn and reason in models at the inference stage, which can contribute to the construction of new models.
- Also, it's worth noting with the size and power of modern LLMs, even the simplest ability to learn or reason will become a full-fledged ability.

*However, this area (point 2) already carries the highest social risks – even the theoretical emergence of the ability of future models to self-learn at the Inference stage **will have global and unpredictable consequences.***

References: 37 on 30.04.2023

References

- (Brown et al., 2020) Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Gñish Sastry, Amanda Askell, et al.: Language models are few-shot learners. *NeurIPS*, 2020. <https://arxiv.org/abs/2005.14165>
- (Wei, Tai et al., 2022) Jason Wei, Yi Tay, Rishi Bommanani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Boer, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, William Fedus.: Emergent Abilities of Large Language Models. Published in *Transactions on Machine Learning Research* (08/2022). <https://arxiv.org/abs/2206.07682>
- (Rae et al., 2021) Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Suzannah Young, et al.: Scaling language models: Methods, analysis & insights from training Gopher. <https://arxiv.org/abs/2112.11446>
- (Hoffmann et al., 2022) Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al.: Training compute-optimal large language models. *NeurIPS*, 2022. <https://arxiv.org/abs/2203.15556>
- (Chowdhery et al., 2022) Ankitha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Boer, Gaurav Mishra, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, et al.: PaLM: Scaling language modeling with Pathways. <https://arxiv.org/abs/2204.02311>
- (Dong et al., 2023) Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, Zhifang Sui.: A Survey on In-context Learning. <https://arxiv.org/abs/2301.00234>
- (Thoppilan et al., 2022) Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Nishkesh, Heng-Tze Cheng, Alicia Ju, Taylor Bos, Leslie Baker, Yu Du, et al.: LaMDA: Language models for dialog applications. <https://arxiv.org/abs/2201.08239>
- (Chan et al., 2022) Stephanie C.Y. Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aditya Singh, Pierre Richemond, James McClelland, Felix Hill.: Data Distributional Properties Drive Emergent In-Context Learning in Transformers. <https://arxiv.org/abs/2205.05055>
- (Wei et al., 2022) Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Boer, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou.: Chain of thought prompting elicits reasoning in large language models. <https://arxiv.org/abs/2201.11903>
- (Wei et al., 2022b) Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou.: Self-consistency improves chain of thought reasoning in language models. <https://arxiv.org/abs/2203.11171>
- (Chan et al., 2022b) Stephanie C.Y. Chan, Ishita Dasgupta, Junhyung Kim, Dhruv Kumar, Andrew K. Lampinen, Felix Hill.: Transformers generalize differently from information stored in context vs weights. <https://arxiv.org/abs/2210.05675>
- (Huang et al., 2023) Shaozhan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuang Ma, Tengchao Lv, Lei Cui, Owam Khan Mohammed, Barun Patra, Qiang Lin, Kati Aggarwal, Zewen Chi, Johan Björck, Vishrav Chaudhary, Subhojit Som, Xia Song, Furu Wei.: Language Is Not All You Need: Aligning Perception with Language Models. <https://arxiv.org/abs/2302.14045>
- (Dries et al., 2023) Danny Dries, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Weizhong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, Fets Florence.: PaLM-E: An Embodied Multimodal Language Model. <https://arxiv.org/abs/2303.03378>
- (OpenAI (2023)) GPT-4 Technical Report. <https://openai.com/papers/gpt-4.pdf>
- (Wang et al., 2022) Bozhi Wang, Sewon Min, Xiang Deng, Jianming Shen, You Wu, Luke Zettlemoyer, Huan Sun.: Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters. <https://arxiv.org/abs/2212.10001>
- (Min et al., 2022) Sewon Min, Minni Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hamaneh Hajishir, and Luke Zettlemoyer.: Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? <https://arxiv.org/abs/2202.12837>
- (Webson and Pavlick, 2022) Albert Webson and Ellie Pavlick.: Do prompt-based models really understand the meaning of their prompts? <https://arxiv.org/abs/2109.01247>
- (Saparov and He., 2023) Abulhair Saparov & He He.: Language Models Are Greedy Reasoners: A Systematic Formal Analysis of Chain-of-Thought. <https://arxiv.org/abs/2210.01240>
- (Razeghi et al., 2022) Yassaman Razeghi, Robert L. Logan IV, Matt Gardner, and Sameer Singh.: Impact of Pretraining Token Frequencies on Few-Shot Reasoning. <https://arxiv.org/abs/2202.07206>
- (Garcia et al., 2023) Xavier Garcia, Yannis Bannal, Colin Cherry, George Foster, Maxim Krikun, Fangxiao Fu, Melvin Johnson, Orhan Firat.: The unreasonable effectiveness of few-shot learning for machine translation. <https://arxiv.org/abs/2302.01398>
- (Connex et al., 2018) Alexis Connex, Guillaume Lample, Marc-Aurèle Ranato, Ludovic Denoyer, Hervé Jégou.: Word Translation Without Parallel Data. <https://arxiv.org/abs/1710.04087>
- (Akytas et al., 2022) Ekin Akytas, Dale Schuurmans, Jacob Andreas, Tengyu Ma, Denny Zhou.: What learning algorithm is in-context learning? Investigations with linear models. <https://arxiv.org/abs/2211.15661>
- (Oswald et al., 2022) Johannes von Oswald, Eyvind Niklasson, Enrico Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, Max Vladymyrov.: Transformers learn in-context by gradient descent*. <https://arxiv.org/abs/2212.07677>
- (Garg et al., 2022) Shivam Garg, Dimitris Tsipras, Percy Liang, Gregory Valiant.: What can transformers learn in-context? a case study of simple function classes. <https://arxiv.org/abs/2208.01066>
- (Dai et al., 2022) Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Zhifang Sui, Furu Wei.: Why Can GPT Learn In-Context? <https://arxiv.org/abs/2212.10559>
- (Irie et al., 2022) Kazuki Irie, Robert Csordás, Jürgen Schmidhuber.: The Dual Form of Neural Networks Revisited: Connecting Text Time Predictions to Training Patterns via Spotlights of Attention. <https://arxiv.org/abs/2202.05798>
- (Patel & Pavlick (2022)) Roma Patel and Ellie Pavlick.: Mapping language models to grounded conceptual spaces. *ICLR*, 2022. <https://openreview.net/forum?id=gjcEM6rdHK>
- BiQ-Bench.: Beyond the imitation game: Measuring and extrapolating the capabilities of language models. <https://arxiv.org/abs/2206.04515>
- (Wei et al., 2022) Jason Wei, Maarten Boer, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le.: Finetuned language models are zero-shot learners. *ICLR*, 2022. <https://arxiv.org/abs/2109.01652>
- (Nye et al., 2021) Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Boer, David Luan, et al.: Show your work: Scratchpads for intermediate computation with language models. <https://arxiv.org/abs/2112.00114>
- (Kadavath et al., 2022) Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield Dodds, Nova DasSarma, Eli Tran-Johnson, et al.: Language models (mostly) know what they know. <https://arxiv.org/abs/2207.05221>
- (Lin et al., 2021) Stephanie Lin, Jacob Hilton, and Owain Evans.: TruthfulQA: Measuring how models mimic human falsehoods. <https://arxiv.org/abs/2109.07958>
- (Tay et al., 2022b) Yi Tay, Vinh Q. Tran, Mostafa Dehghani, Jimmy Ni, Dora Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, et al.: Transformer memory as a differentiable search index. <https://arxiv.org/abs/2202.06991>
- (Lampinen et al., 2022) Andrew K. Lampinen, Ishita Dasgupta, Stephanie C.Y. Chan, Kory Matthews, Michael Heury Tessier, Antonia Creswell, James L. McClelland, Jane X. Wang, and Felix Hill.: Can language models learn from explanations in context? Findings of EMNLP, 2022. <https://arxiv.org/abs/2204.02329>
- (Zhou et al., 2022) Denny Zhou, Nathaniel Schiødt, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Breugnot, Quoc Le, and Ed Chi.: Least-to-most prompting enables complex reasoning in large language models. <https://arxiv.org/abs/2205.10615>
- (Shi et al., 2022) Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivastava, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei.: Language models are multilingual chain-of-thought reasoners. <https://arxiv.org/abs/2210.03057>
- (Arora et al., 2022) Shantanu Arora, Avnika Narayan, Mayya F Chan, Laurel J Orr, Neel Guha, Kush Bharia, Isha Chaudh, Frederic Sala, and Christopher Ré.: Ask me anything: A simple strategy for prompting language models. <https://arxiv.org/abs/2210.02441>



Thank you for
your attention!





Spoiler about Analytical part:

This hypothesis is in
good agreement with
all the facts and
experiments.

Hypothesis № 2.

*Teaching LLMs based on languages data leads to the creation of a **language space of reasoning**, which includes not only the language itself linguistically, but also some patterns of rules of reasoning implicitly embedded in the structures of human languages.*

LLMs "learn" at inference without actually learning:

- 1. LLMs use (at the inference time) language spaces of reasoning that LLMs created earlier(at the training stage).*
- 2. Demonstrating examples of reasoning with data (at inference time) focuses the LLM's "attention" on those areas of the language space of reasoning where already there are similar "rules" that allow models to work with this and similar data.*