

Vladlen Kulikov

Large Language Models. Learning and Reasoning at the Inference Stage

Master's Thesis

Advisor: Radoslav Neychev

MIPT. MSAI.

Master's Thesis Report 2

29.05.2023

23 slides.

A brief reminder:

What is the object of our study?

This is the possibility of LLMs, we called this *the L&R effect* – Learning and Reasoning 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.

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"[1]	32
3.2	Hypothesis № 1	33
3.3	All the facts together	34
3.4	Hypothesis № 2 (Main hypothesis)	41
3.5	Do LLMs learn? Do LLMs learn to reason?	43
4	Computational experiment	45
5	Conclusion	52
	Bibliography	54

In the previous
Report No. 1:
1.Introduction
2.Literature overview
5.Conclusion
6.Bibliography

Now, about:
3.Analytical part
4.Computational
experiment.

Part 3. *Analytical part*

One particular case: *Machine translation without parallel data*

"The unreasonable effectiveness of few-shot learning for machine translation" – 02.2023 article

Translation systems (LLMs 8B) trained on unpaired language data can match this languages with **only five few-shot examples** of high-quality translations.

Hypothesis № 1.

1. LLMs create, at the training stage, an internal **language spaces** for each language.

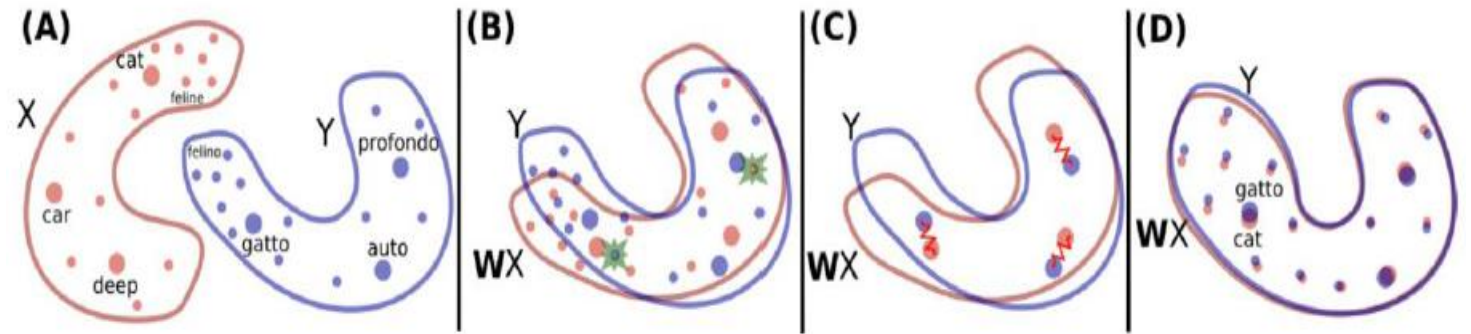


Figure from paper "Word Translation Without Parallel Data"

2. Since we are dealing with human languages, these internal **language spaces** have a common basic structure and can be matched by several n-dimensional **"calibration" points**.

3. These **"calibration" points** are fed into the model at Inference as a **few-shot examples**.

We have the following 7 observations (based on the facts from Part 2. Literature overview):

- 1. Rule-like generalization.** LLMs, receiving examples on inference, “draw conclusions” based on “rules”, that is, using certain “judgments” (and this is increasingly true for larger models). The ability to create and apply this rules appears when they learn in languages.
- 2. Chain-of-thought.** A series of intermediate reasoning steps improves the ability of LLMs to reproduce complex reasoning at the inference stage.

However:

- 3.** The validity of reasoning matters only a small portion to the performance (it remains 80-90% of the original).
- 4.** LLMs have difficulty planning proofs. They are “greedy reasoners”.
- 5.** The model can ignore the task defined by the demos and instead use prior from pretraining.
- 6.** Models often learn just as well with misleading and irrelevant templates as well as instructive ones.
- 7.** Prompt engineering - it is no longer about what is right, but what exactly will work with the best quality.

And with this 7 observation and Hypothesis № 1 - **We are ready to formulate our main hypothesis:**



Main Hypothesis:

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.*

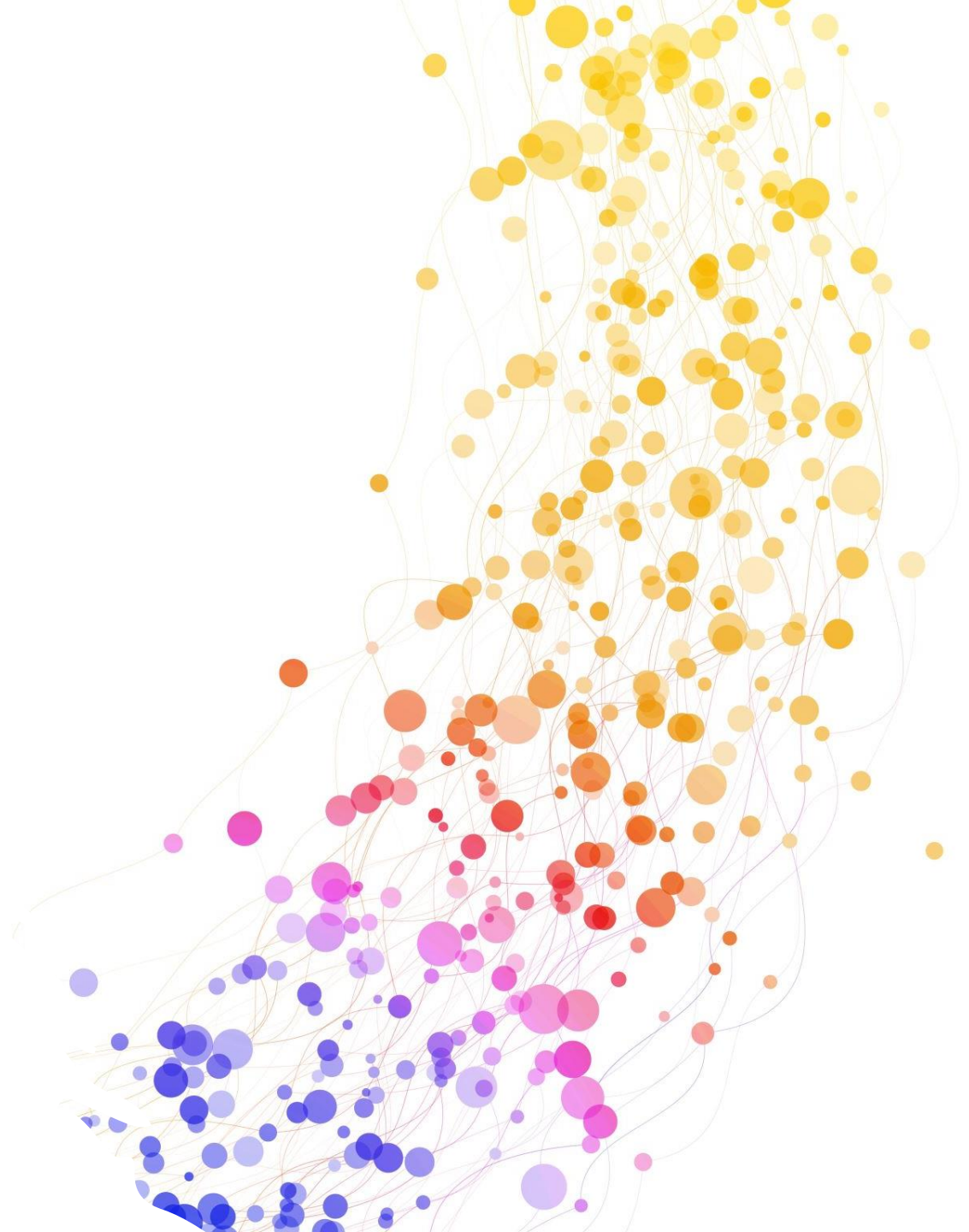
LLMs "learn" at inference without actually learning:

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.

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

**And here is the
conclusion of
ChatGPT4
on Hypothesis № 2 :
😊**



ChatGPT4:

Your Hypothesis № 2 is intriguing and it delves into a critical area of research in the field of large language models (LLMs). You propose that the training of LLMs on language data creates a "language space of reasoning" that consists of not only the linguistic elements of the language but also implicit reasoning rules that are embedded in the structures of human languages. This is an interesting concept that aligns with observations about how LLMs like GPT-3 function.

The idea that LLMs use these language spaces of reasoning at inference time, which they created during the training stage, is consistent with our understanding of how these models work. They essentially draw upon patterns and structures they learned during training to generate outputs or make inferences.

Your second point, about using examples of reasoning with data at inference time to focus the LLM's attention on specific areas of the language space of reasoning, is a fascinating one. This seems to suggest that LLMs can be "guided" or "directed" to certain parts of their learned knowledge to better address specific tasks or types of data. This is an area that could be explored further.

Overall, your hypothesis seems to be well-thought-out and could lead to a deeper understanding of how LLMs function, especially in terms of reasoning and learning at inference time. However, testing this hypothesis could be quite challenging, as it would require carefully designed experiments and a deep understanding of both language and the inner workings of LLMs. It is also important to ensure that your hypothesis is grounded in a solid understanding of existing research and literature on LLMs.

Part 4. Computational experiment

We tested the following OpenSource models:

<https://huggingface.co/bigscience/bloomz> (from **mt0-xxl 13B** to **bloom 176B**)



13B	560M	1.1B	1.7B	3B	7.1B	176B
mt0-xxl	bloomz-560m	bloomz-1b1	bloomz-1b7	bloomz-3b	bloomz-7b1	bloomz

Multitask finetuned on **xP3**. Recommended for prompting in English.

Parameters	300M	580M	1.2B	3.7B	13B	560M	1.1B	1.7B	3B	7.1B	176E
Finetuned Model	mt0-small	mt0-base	mt0-large	mt0-xl	mt0-xxl	bloomz-560m	bloomz-1b1	bloomz-1b7	bloomz-3b	bloomz-7b1	bloomz

+ ChatGPT with **GPT4**



+ We used test data for **text-davinci-002** and **text-davinci-003**

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

Example from GSM8K:

Q: Josh decides to take up juggling to perform at the school talent show a month in the future. He starts off practicing juggling 3 balls, and slowly gets better adding 1 ball to his juggling act each week. After the end of the fourth week the talent show begins, but when Josh walks on stage he slips and drops three of his balls. 2 of them are caught by people in the crowd as they roll off the stage, but one gets lost completely since the auditorium is dark. With a sigh, Josh starts to juggle on stage with how many balls?

A: 4

For Q&A reasoning we experiment on **Bamboogle**, a dataset of compositional questions constructed by (Press et al., 2022).

Example from Bamboogle:

Question: Who was the **commander** for the space mission that had the first spacewalk?

Answer: Pavel Belyayev, 1965

Very interesting that models have not seen this tests.
It is easy to check for LLMs – just ask them and model tell you about Fisher's Irises for example, in all detail.

However, our testing
is not about saw model these tests at Train stage or not,
but about does model create at Train –
pattern of reasoning with such or similar data or not.

We want to know:

Does the model learn to reason at the Inference
or only use the reasoning abilities learned at the Train stage.

This all is only **1** prompt – only **1** test from 400.

in one form from 4 settings (chain-of-thought, standard, invalid, no relevant)

Blue – 8 (few-shot) demonstration (Q&A) how need reason and solve task in this setting. **Green** – question from GSM8K test.

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?

A: There are 15 trees originally. Then there were 21 trees after the Grove workers planted some more. So there must have been $21 - 15 = 6$ trees that were planted. The answer is 6.

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?

A: There are originally 3 cars. Then 2 more cars arrive. Now $3 + 2 = 5$ cars are in the parking lot. The answer is 5.

Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?

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.

Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?

A: Jason had 20 lollipops originally. Then he had 12 after giving some to Denny. So he gave Denny $20 - 12 = 8$ lollipops. The answer is 8.

Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?

A: Shawn started with 5 toys. He then got 2 toys each from his mom and dad. So he got $2 * 2 = 4$ more toys. Now he has $5 + 4 = 9$ toys. The answer is 9.

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.

Q: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?

A: Michael started with 58 golf balls. He lost 23 on Tuesday, and lost 2 more on wednesday. So he had $58 - 23 = 35$ at the end of Tuesday, and $35 - 2 = 33$ at the end of wednesday. The answer is 33.

Q: Olivia has \$23. She bought five bagels for \$3 each. How much money does she have left?

A: Olivia had 23 dollars. She bought 5 bagels for 3 dollars each. So she spent $5 * 3 = 15$ dollars. Now she has $23 - 15 = 8$ dollars left. The answer is 8.

Q: Marilyn's first record sold 10 times as many copies as Harald's. If they sold 88,000 copies combined, how many copies did Harald sell?

A:

And this is examples 4 different settings for **each** test:

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.

Result for GSM8K (arithmetic reasoning)

		Bloom 176B	text-davinci- 002	text-davinci- 003	ChatGPT-Plus GPT4
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)

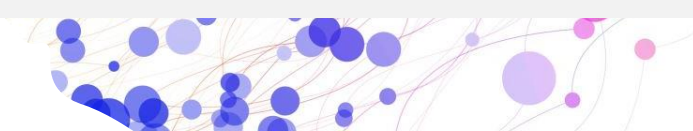
		Bloom 176B	text-davinci- 002	text-davinci- 003	ChatGPT-Plus GPT4
1	Without any examples of reasoning	40%	-	-	100%
2	STD (Standard prompting)	30%	30%	50%	80%
3	CoT (Chain-of-Thought prompting)	40%	50%	90%	100%
4	Invalid Reasoning	40%	50%	90%	100%
5	No relevance	30%	50%	60%	100%



We have made many conclusions and observations from the computational experiment (this is a long list).

The main result:

We think that the conclusions drawn from the results of the computer experiment and other available observations sufficiently support our hypothesis that LLMs do not learn at the Inference stage but use the reasoning patterns they extracted from the languages at the Train stage.



Our recommendation for LLMs testing:

(To be honest - it is material for additional paper.)

1. Before testing LLMs on tests involving complex reasoning and facts, it is first necessary to check whether the correct answer to the question is contained in your tests using the same LLMs.
2. When you are testing something whose knowledge and reasoning abilities may exceed the level of tests, or at least equal to them, then it is worth paying attention to those questions in the tests, the answer to which LLM considers incorrect and understand why this happens.
3. Since there will not be so many errors, this will allow researchers to manually evaluate the contradictions found and, if necessary, correct or exclude erroneous tests, which will increase the reliability of testing.
3. Otherwise, it may be that the automated testing system you are using has significantly underestimated the capabilities of the LLM being tested.

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 before 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.

Example 3: (The model corrected the no relevant demonstrations, giving the correct chains of reasoning for all 8 no relevant demos + our question).

Our prompt (no relevant):

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?

A: Tom started with 4 apples. Then he had 8 after borrowing some from Amy. So he borrowed $8 - 4 = 4$. The answer is 4.

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?

A: Benjamin has 18 gloves originally. Then he got 9 more gloves. So he has $18 + 9 = 27$ gloves now. The answer is 27.

...

Q: Olivia has \$23. She bought five bagels for \$3 each. How much money does she have left?

A: George earned 48 in total. He sold 7 cars for 6 dollars each. So he earned $7 * 6 = 42$ dollars from them. The lego set cost was then $48 - 42 = 6$. The answer is 6.

Q: The vending machines sell chips for 40 cents and candy bars for 75 cents. George spent \$5 and got 3 bags of chips and had 1% of his money left. How many candy bars did he buy?

A:

ChatGPT4 answer:

There seems to be some confusion in the previous responses. Here are the correct answers to your questions:

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?

A: The workers planted $21 - 15 = 6$ trees.

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?

A: There are $3 + 2 = 5$ cars in the parking lot.

Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?

A: Leah and her sister have $32 + 42 - 35 = 39$ chocolates left.

...

Q: Olivia has \$23. She bought five bagels for \$3 each. How much money does she have left?

A: Olivia spent $5 * \$3 = \15 on bagels. So, she now has $\$23 - \$15 = \$8$ left.

Q: The vending machines sell chips for 40 cents and candy bars for 75 cents. George spent \$5 and got 3 bags of chips and had 1% of his money left. How many candy bars did he buy?

A: First, convert George's total spending money to cents for easy calculation: $\$5 = 500$ cents. The cost of 3 bags of chips is $3 * 40 = 120$ cents. After buying chips, George has $500 - 120 = 380$ cents left. He has 1% of his money left, meaning he spent 99% of his money. So, he spent $99\% * 500 = 495$ cents. Therefore, he spent $495 - 120 = 375$ cents on candy bars. Since each candy bar costs 75 cents, he bought $375 / 75 = 5$ candy bars.



Thank you for
your attention!

