Is Bigger and Deeper Always Better? Probing LLaMA Across Scales and Layers

Anonymous ACL submission

Abstract

This paper presents an in-depth analysis of Large Language Models (LLMs), focusing on LLaMA, a prominent open-source foundational model in natural language processing. Instead of assessing LLaMA through its generative output, we design multiple-choice tasks to probe its intrinsic understanding in high-order tasks such as reasoning and calculation. We examine the model horizontally, comparing different sizes, and vertically, assessing different layers. We unveil several key and uncommon findings based on the designed probing tasks: (1) Horizontally, enlarging model sizes almost could not automatically impart additional knowledge or computational prowess. Instead, it can enhance reasoning abilities, especially in math problem solving, and helps reduce hallucinations, but only beyond certain size thresholds; (2) In vertical analysis, the lower layers of LLaMA lack substantial arithmetic and factual knowledge, showcasing logical thinking, multilingual and recognitive abilities, with top layers housing most computational power and real-world knowledge. These findings provide new observations into LLaMA's capabilities, offering insights into the current state of LLMs.

1 Introduction

Large language models (LLMs) (OpenAI, 2023; Scao et al., 2022; Yao et al., 2022; Chen et al., 2023) have shown significant potential in numerous high-order open-generation tasks such as mathematical and logical reasoning. LLaMA (Touvron et al., 2023b), an open-source, state-of-the-art foundational large language model has been designed to facilitate research in natural language processing communities. In a relatively brief period, LLaMA has garnered significant attention. This prominence can be attributed to its inherent accessibility and demonstrated efficacy across a diverse array of text-generation tasks (Hu et al., 2021; Chen et al., 2022; Gao et al., 2023). Beyond LLaMA's impressive generative capabilities, can we further uncover its intrinsic understanding abilities? Does bigger and deeper always lead to better performances in its advanced capabilities such as computational and reasoning sensitivity? Addressing this question is not only instrumental in comprehending the foundations of its success, but it also facilitates an understanding of its inherent limitations. This, in turn, can guide future advancements in the architecture and training optimization of LLMs.

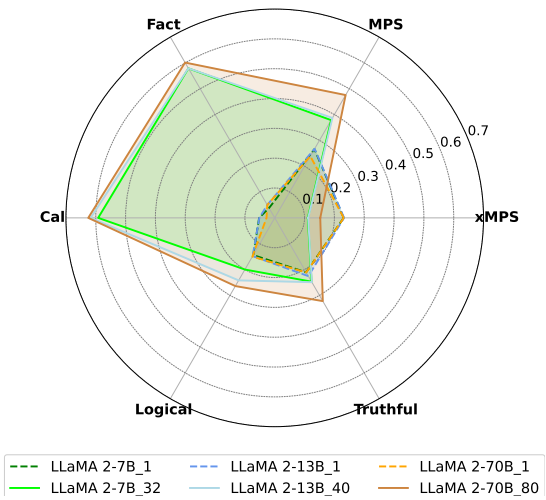


Figure 1: Overall Comparison with LLaMA 2 7B-70B in our probing tasks. Detailed introduction of each task include in Section 3. Dashed lines represent the first layer of each model, while solid lines represent the last layer of the model.

In this paper, we conduct a series of experiments to probe the nature of LLaMA on five higher-order tasks under in-context learning, including *arithmetic calculation*, *math problem solving* (MPS), *logical reasoning*, *truthfulness*, and *factual knowledge detection*. The latter two are considered as the important symbols of hallucination. In these tasks, we probe the model's capabilities from two distinct perspectives: 1) **Horizontally**: Comparing

the model's abilities across different sizes (Scaling Law); 2) **Vertically**: Comparing the different layers capabilities of the same size model (Layer-wise). Instead of directly testing LLMs via their open-text generation abilities, as is usually done, we prob LLaMA with a set of challenging multiple-choice questions. The primary considerations for this design are: Firstly, it offers controlled and efficient evaluation, with clear, quantifiable results that reduce ambiguity. This approach allows for directly targeted testing of specific knowledge areas and reasoning skills, as well as validating models' sensitivity to correct or incorrect answers; Secondly, our experimental observations reveal a tendency for LLaMA's lower layers to produce repetitive words rather than coherent sequences, which would lead to an unfair layer-wise comparison.

In the context of our experiments corroborating each other, we draw the following conclusions:

Horizontally: (1) The primary benefit of increasing model size lies in the enhanced reasoning abilities of the models, most notably in their improved capacity in MPS. This increase in size also tends to reduce the occurrence of hallucinations. However, these improvements are only evident when certain LLM size thresholds are surpassed, known as emergent abilities (Wei et al., 2022). For instance, models ranging from 7B to 13B show comparable performance across all probing tasks. It's only when the model size increases from 13B to 70B parameters that a noticeable improvement in reasoning capabilities and a reduction in hallucination issues can be observed, as shown in Figure 1; (2) The pure arithmetic capabilities and inherent factual knowledge of LLaMAs with different parameter sizes are remarkably similar. In other words, increasing the model size does not necessarily impart additional factual knowledge or significantly enhance computational capabilities, especially when the same volume of pre-training corpus is used.

Vertically: (1) We find that the lower and middle layers of LLaMA have almost negligible pure arithmetic and factual knowledge capabilities, seen in Figure 1. As the network layers deepen, there is a noticeable leap in performance. Contrarily, even at the very lowest layers, LLaMA possesses logical thinking and recognitive abilities, such as in mathematics, logical reasoning, and avoiding hallucinations. While these abilities do enhance slightly with deeper layers, the improvement remains quite limited. *This implies that current LLMs predominantly house computational power and real-world knowledge in their upper layers, while the lower layers are geared towards relevant abstract thinking but lack substantial real-world knowledge and computational skills.* (2) Interestingly, in our layer-by-layer performance comparisons, we observe that the model's optimal performance in MPS and computational abilities is not always at the final layer. More often, these peak capabilities are found in several layers before the last. However, in contrast, for representing factual knowledge, the final layer of the model proves to be exceptionally crucial.

Further, we extend MPS tasks to the cross-lingal reasoning context, named xMPS. Specifically, we maintain the questions and incorrect options unchanged and translate the correct answers into other languages to assess the LLaMA's multilingual proficiency. In this setting, our layer-wise experiments show an effect completely opposite to monolingual MPS: models' performance gradually decreases as the layers deepen. *This indicates that LLaMA's earlier layers are responsible for preserving general multilingual features (language-agnostic).*

Of note, the results presented in our experiments do not necessarily equate directly to the generative capabilities of LLaMA. Rather, in this paper, we provide a novel and comprehensive perspective for observing the natural performance of LLaMA, giving insights to understand the current LLMs better.

2 LLaMA

LLaMA (Touvron et al., 2023a,c) is a series of foundation large language models, released by META, has becomes the most popular open-source LLMs in NLP communities. LLaMA is built on transformer layers (Vaswani et al., 2017), trained on trillion of tokens with the language modeling objective, showing powerful abilities in down-stream tasks. The contextualized representations are optimized by predicting the next token based on the input sequences.

In this work, we probe LLaMA 2 series LLMs with our designed tasks, ranging from 7B to 70B parameters in **in-context learning**. Concretely, LLaMA 2-7B, 13B and 70B consist of 32, 40 and 80 transformer layers with 4096, 5120 and 8192 hidden embedding sizes, separately.

3 Probing Tasks

Probing tasks are generally utilized to explore the inherent knowledge and linguistic features within deep learning models. Previously, Jawahar et al.

Type	Bit	+	-	×	÷	Mix-2	Mix-3
Int (Float)	1-2	200	200	200	200	200	200
	3-4	200	200	200	200	200	200
	5-6	200	200	200	200	200	200

Table 1: Test data statistics in our arithmetic tasks.

(2019a) employed a series of probing tasks to examine the internal representations of BERT. However, with the rapid advancement of LLMs, there is currently a lack of comprehensive research that deeply analyzes the relationship between the higher-order capabilities of contemporary LLMs and factors such as model size and network layers.

To bridge this gap, we probe LLaMA high-order abilities across two views: model size and layer-wise. Specifically, we devise five high-order tasks: *calculation*, *math problem solving* (MPS), *logical reasoning*, *truthfulness*, and *factual knowledge detection*. We include the latter two as the *hallucination detecting* in the following. Besides, we also probe LLaMA efficiency in *cross-lingual MPS*. In this section, we will illustrate them sequentially.

3.1 Calculation

In this paper, we focus on testing LLMs in basic arithmetic tasks, including four simple arithmetic expressions: addition (+), subtraction (-), multiplication (×) and division (÷):

- Add of two elements within 1∼100, 100∼10000, 10000∼100000, separately.
- Subtract of two elements within 1∼100, 100∼10000, 10000∼1000000, separately.
- Multiply of two elements within 1∼100, 100∼10000, 10000∼1000000, separately.
- Division of two elements within 1∼100, 100∼10000, 10000∼1000000, separately.
- Complex arithmetic operations that require performing **two** operations of addition, subtraction, multiplication, or division.
- Complex arithmetic operations that require performing **three** operations of addition, subtraction, multiplication, or division.

Of note, the elements used in the above arithmetic operations include **integers** and **floating-point numbers** (with precision up to three decimal places), separately. Table 1 shows the corresponding data statistics. Since we probe the computational abilities of LLaMA through the multiple-choice question answering task, to increase the difficulty and test the model's sensitivity to minor differences in computational results, we randomly add or subtract a floating-point number within $\pm\mathbf{20}$ (except 0) to the correct answer to create three different but indistinct incorrect options.

This design of our test set allows for an intuitive and fine-grained comparison of 1) the model's relative strengths and weaknesses in addition, subtraction, multiplication, and division operations; 2) the model's performance patterns when faced with complex calculations; 3) the variations in the model's computational abilities when dealing with floating-point numbers and integers, 1-2 digit, 3-4 digit, 5-6 digit numbers respectively. Our data are constructed by calling python `random.randint()` and `random.uniform()` functions.

3.2 Math Problem Solving

Besides validating LLaMA in arithmetic tasks, we also test the model in MPS tasks to comprehensively review its math reasoning abilities.

We select GSM8K (Cobbe et al., 2021) as our source data to construct challenging and misleading options that effectively fool the model. Our strategy involves the following steps:

- We first fine-tune the LLaMA 2-13B model on GSM8K, and then perform rejection sampling via inference 100 times to generate various reasoning paths based on the resulting model.
- Next, we extract all the formulas in each reasoning path and validate their accuracy. We use the erroneous reasoning paths to construct our probing task data:
  - If a reasoning path only contains computational errors, meaning the correct answer can be obtained by recalculating, we retain it as part of our **MPS-Cal** probing test set.
  - If all computations in a reasoning path are correct, but the final conclusion is wrong, indicating a reasoning error, we use it for our **MPS-Rea** test set.

The MPS-Cal focuses on assessing the model's sensitivity to computational results in solving mathematical problems. Conversely, MPS-Rea emphasizes evaluating the model's ability to discern correct from incorrect reasoning paths, requiring a superior level of understanding and reasoning capabilities. Table 2 shows several examples in MPS and calculation tasks.

Task Type	Query & Options
Arithmetic-Int	**Query**: 2331 + 2693 = ? **Options**: **5024** (✓); 5018; 5005; 5025 **Query**: 109848 ÷ 199 = ? **Options**: **552.0** (✓); 516.0; 558.0; 567.0
Arithmetic-Flo	**Query**: 7.682 + 28.894 = ? **Options**: **36.576** (✓); 28.576; 40.909; 38.076 **Query**: 25.204 × 88.29 ÷ 12.133 = ? **Options**: **183.406** (✓); 183.739; 185.406; 181.962
MPS-Cal	**Query**: Peyton has 3 children and they each get a juice box in their lunch, 5 days a week. The school year is 25 weeks long. How many juices boxes will she need for the entire school year for all of her children?
	Options: **Peyton needs 25 weeks x 5 days x 3 children = 375 juice boxes** (✓); 25 weeks x 5 days x 3 children = 75 juice boxes; Given the conditions of the problem, 3 children, 5 days a week, 25 weeks long, that's 3*5*25 = 105 juice boxes needed.
MPS-Rea	**Query**: A family of 12 monkeys collected 10 piles of bananas. 6 piles had 9 hands, with each hand having 14 bananas, while the remaining piles had 12 hands, with each hand having 9 bananas. How many bananas would each monkey get if they divide the bananas equally amongst themselves?
	Options: **The first 6 bunches had 6 x 9 x 14 = 756 bananas. There were 10 - 6 = 4 remaining bunches. The 4 remaining bunches had 4 x 12 x 9 = 432 bananas. All together, there were 756 + 432 = 1188 bananas. Each monkey would get 1188/12 = 99 bananas** (✓); 6 piles had 6 x 9 x 14 = 756 bananas. The remaining 6 piles had 6 x 12 x 9 = 648 bananas. All together, there were 756 + 720 = 1476 bananas. Each monkey would get 1476/12 = 123.0 bananas; 6 piles had 6 x 9 x 14 = 756 bananas. There were 10 - 6 = 4 piles of bananas with 12 hands and 4 piles of bananas with 6 hands. The 4 piles of bananas with 12 hands had 4 x 12 x 9 = 432 bananas. The 4 piles of bananas with 6 hands had 4 x 6 x 9 = 216 bananas. There were 756 + 432 + 240 = 1428 bananas. Every monkey will get 1428/12 = 119.0 bananas

Table 2: Testing examples in our designed calculation and MPS probing tasks.

Tasks	Arithmetic-Int/Float	Reclor	(x) MPS-Cal	(x) MPS-Rea	TruthfulQA	LAMA*
Avg. Ground-truth	1	1	1	1	3.5	1
Avg. Candidates	4	4	3	4.9	7.6	9.7
Total Queries	3600	500	712	1000	817	3070

Table 3: Overall Test data statistics in our probing tasks. **LAMA*** refers to we only use s subset of original corpus.

3.3 Logical Reasoning

As a key indicator of the advanced capabilities of contemporary LLMs, logical reasoning stands out for its importance in examining, analyzing, and critically assessing arguments in natural language. In our study, we employ Reclor (Yu et al., 2020) as a testing platform to evaluate the logical reasoning skills of these large models. Reclor comprises a dataset derived from logical reasoning questions found in standardized tests for graduate admissions. Each sample from Reclor contains one context, one corresponding question and four options.

3.4 Hallucination Detecting

Hallucination, which means generating content that deviates from real-world facts observed during pre-training, is considered one of the most challenging issues in LLMs. In order to further investigate the relationship between hallucination and model layers and size, we conduct tests from two aspects: 1) Measure whether a language model is truthful in generating answers to questions, also known as truthfulness; 2) Test the model's internal factual knowledge. We use TruthfulQA MC tasks (Lin et al., 2022) and LAMA (Petroni et al., 2019) as test beds for these two aspects, respectively. It is important to note that in TruthfulQA, there may be more than one correct answer, accompanied by 4-5 incorrect options. As for LAMA, we randomly extract a subset containing 3070 questions along with their 9-10 corresponding options. Table 3 presents detailed data statistics in our probing tasks.

3.5 Cross-Lingual Math Problem Solving

In this study, we delve further into LLaMA's multilingual abilities. We translate the correct answers from the collected two datasets: MPS-Cal and MPS-Rea in Section 3.2 into four additional languages: *Chinese*, *French*, *Spanish*, and *Thai*, while keeping the questions and other incorrect options as they are, the new resulting test sets named **xMPS-Cal** and **xMPS-Rea**. This setting offers several advantages: Firstly, it tests the model's capacity in cross-lingual reasoning transfer, demonstrating its proficiency in not just recognizing but also reasoning in multiple languages. Secondly, by mixing incorrect choices with correct answers in different languages, we robustly assess the model's adapt-

Model Size	LAMA* (Fact)	Reclor (Logical)	MPS-Cal	MPS-Rea	TruthfulQA		Arithmetic	
					MC1	MC3	Int	Float
7B	57.9	20.0	28.7	47.0	28.6	20.7	67.9	52.5
13B	57.9	23.7	30.2	46.6	29.1	20.7	70.6	52.6
70B	58.7	26.4	48.3	51.9	37.3	27.1	70.8	52.9

Table 4: Overall performances of each size LLaMA 2 model in our probing tasks. **LAMA*** refers to we only use s subset of original corpus. MC3 accuracy means the normalized total probability assigned to all true answers among candidates in TruthfulQA.

ability and comprehension across linguistic barriers. This unique setup challenges the model's ability to process and integrate multilingual information, not only evaluates the model's language-specific capabilities but also its overall versatility in cross-lingual understanding and reasoning.

3.6 Test Setting

Consider a probing dataset $D = \{Q, C, O\}$, where Q, C and O denote a set of questions, contexts (only exits for LAMA), and answer options. For each question $q \in Q$, there is a corresponding set of answer choices, denoted as $o \in O$, where $o = \{o_1, o_2, ..., o_{n-1}, a\}$, n is the number of answer choices, and a refers to the correct answer.

The model's task is to identify the correct answer from the set o for each question q. It need to assign the highest log-probability of completion following the question, independent of the other answer choices (Chuang et al., 2023). This selection process can be mathematically represented as:

$$o_i^* = \text{argmax} \log P(o_i|q) \quad (1)$$

$$\text{Acc} = \begin{cases} 1, & \text{if } a^* > (o_1^*, ..o_{n-1}^*), \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

Where $\log P(o_i|q)$ is the log-probability that the choice o_i to question q, as evaluated by the model.

3.7 Experimental Settings

We select LLaMA 2 from 7B to 70B as our experimental subject. Observing that LLaMA 2 exhibits significant instability in zero-shot testing, we choose to implement few-shot prompting in our probing tasks to optimize the model's performance. In TruthfulQA and LAMA, we respectively employ 6-shot (Table 7) and 4-shot (Table 8) prompts. For reasoning tasks, we consistently use 4-shot for both (x) MPS (Table 10) and logical reasoning (Table 9). In calculation tasks, we use 6-shot examples (Table 6). We show detailed prompts in Appendix B.

Size	Bit	+	-	×	÷	M-2	M-3
Integer Arithmetic							
7B	1-2	99.5	100.0	95.0	100.0	55.5	39.5
	3-4	98.0	98.0	59.0	59.5	48.5	20.5
	5-6	89.0	83.0	53.0	30.5	48.5	17.5
13B	1-2	99.5	100	98.5	100.0	58	44.5
	3-4	99.5	99.0	73.5	69.5	53.5	26.0
	5-6	96.5	96.5	63.5	25.5	53.5	17.5
70B	1-2	99.5	100.0	98.0	100.0	64.0	46.0
	3-4	99.0	99.0	75.0	68.0	42.0	18.0
	5-6	100.0	100.0	77.0	23.0	41.0	20.0
Floating-point Arithmetic							
7B	1-2	99.5	100	23.0	78.0	37.0	30.0
	3-4	98.0	98.0	18.0	17.0	32.0	19.0
	5-6	94.0	87.0	19.5	13.0	35.0	14.5
13B	1-2	99.0	100.0	26.0	90.0	40.0	28.0
	3-4	99.5	99.5	14.5	20.5	33.0	19.0
	5-6	99.0	96.5	13.5	13.5	39.5	17.0
70B	1-2	100.0	100.0	26.5	99.0	43.0	30.0
	3-4	98.5	100.0	14.0	39.0	39.5	17.0
	5-6	98.5	99.5	14.5	17.5	45.5	17.5

Table 5: Detailed results of different operations in our probing arithmetic tasks. **M-2/3** refers to arithmetic expression that requires 2/3 times mix operations.

4 Experiments on Probing Model Size

In this section, we present a comparison of the results from LLaMA of different sizes on our probing tasks, as shown in Table 4 and Table 5.

Increasing model size hardly enhance the model's internal knowledge. From Table 4, we can see that the performance of LLAMA 2-7B and 13B on LAMA is identical , and even increasing the model size to 70B results in only a slight improvement (58.7% vs. 57.9%). This indicates that only increasing model size is difficult to improve the model's ability to remember and understand knowledge present in the training corpus, provided the training data remains the same.

Increasing model size does not significantly boost fundamental computational ability. Similarly, models of different sizes also show comparable computational abilities. Even though the 7B model lags a bit in integer operations compared to 13B and 70B, it still performs similarly in floating-

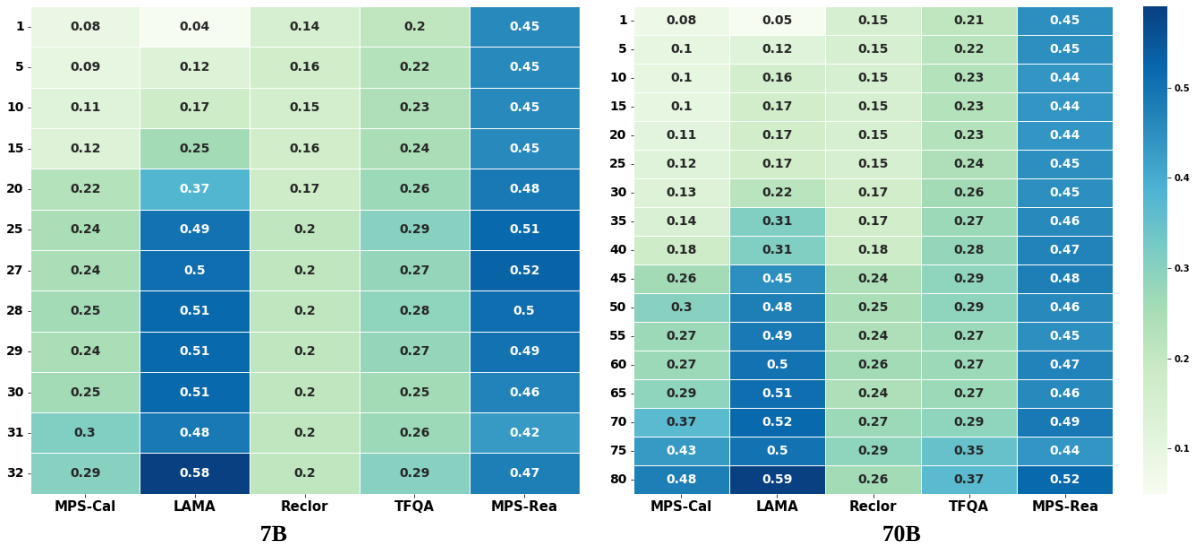


Figure 2: Overall Comparison with LLaMA 2-7B and 70B in our probing tasks. We include all layers' performances of each size model in the Appendix C, Table 11, 12 and 13.

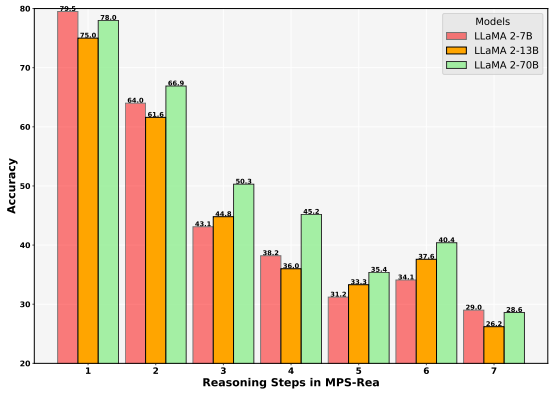


Figure 3: Overall comparison between LLaMA 2 7B to 70B dealing with different reasoning steps problems in our probing MPS-Rea tasks.

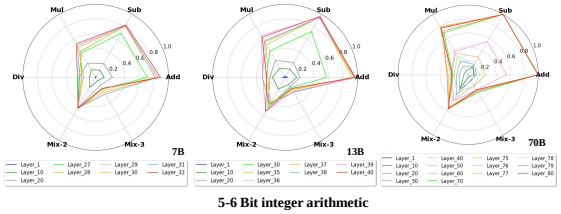


Figure 4: Overall comparison between LLaMA 2 7B to 70B dealing with 5–6 bit calculations in our probing arithmetic tasks. We present more detailed results of 1-2 and 3-4 bit calculations in Appendix C, Figure 7.

point operations (52.5% vs. 52.6% vs. 52.9%). Obviously, the computational abilities of 13B and 70B models are nearly identical.

Larger models show a relative improvement in reasoning ability and truthfulness. Models achieve considerable improvements in MPS and TruthfulQA tasks as the size increases. Delve deeper, the differences in performance across models of various sizes on MPS-Cal seem to contradict our earlier conclusion that the computational capabilities between models of different sizes are minimally distinct. However, considering the performance discrepancies observed on MPS-Rea, we believe that the significant variations on MPS-Cal are primarily due to the notable increase in inference capabilities and sensitivity to computation with the escalation in model size. Yet, even the best-performing 70B model exhibits limited inference capacity, failing to effectively differentiate between correct and incorrect reasoning paths. This leads to considerable discrimination in performance on the MPS-Cal task, albeit with a smaller yet positive improvement on MPS-Rea.

To further prove this point, Figure 3 indicates that all sizes of models perform well on mathematical problems requiring 1-2 steps of reasoning. The enhancement of mathematical capabilities in the 70B model, relative to the 7B and 13B models, is primarily concentrated on problems requiring 3-6 steps of reasoning. When faced with problems requiring 7 steps of reasoning, the performance of all models rapidly declines and shows little difference. The above findings demonstrate that the LLaMA series models all possess elementary reasoning capabilities. *However, the ability to solve more complex reasoning problems only appears to emerge as the certain model size thresholds are surpassed. Even LLaMA 2-70B is still at a "moderate intelligence" level, lacking strong reasoning skills.*

In calculation, LLaMA's performance declines with increasing operation and numbers complexity. LLaMA possesses strong addition and

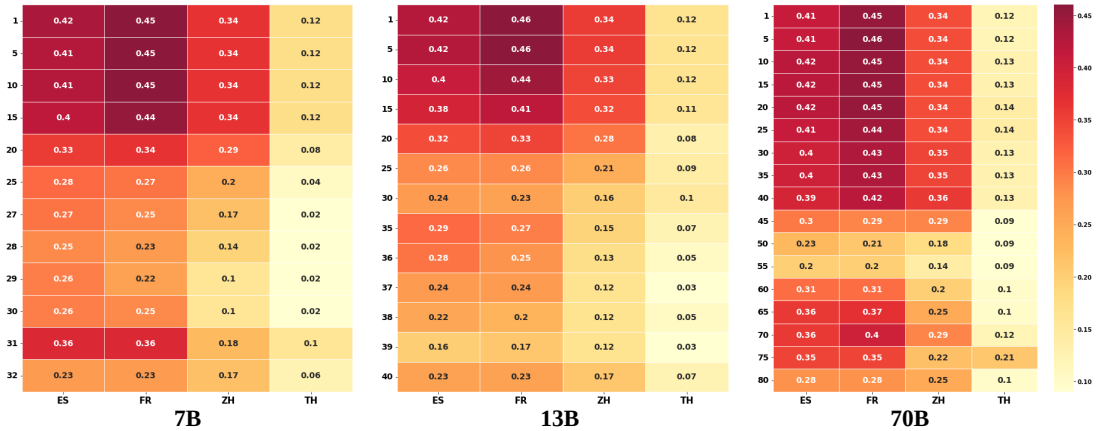


Figure 5: Overall Comparison with LLaMA 2-7B to 70B in our xMPS-Rea probing tasks. ES, FR, ZH and TH refer to Spanish, French, Chinese and Thai.

subtraction capabilities, but its multiplication and division abilities noticeably decrease with increasing digit count, as seen in Table 5. Compared to floating-point operations, LLaMA is better at integer operations. Interestingly, in integer operations, LLaMA shows better capability in multiplication, but this strength significantly diminishes when dealing with floating-point numbers.

5 Experiments on Probing Layer-Wise

In this section, we focus on evaluating each layer of LLaMA across our probing tasks. We present comprehensive results of all layers in Appendix C.

Computational ability primarily exists in the upper layers of the model. Figure 4 presents the performance of different layers of models ranging from 7B to 70B in conducting 5-6 digit integer number calculations. From the figure, it is evident that almost no pure computational ability exists in the lower layers of any size model. However, as the number of layers increases, there is a significant leap in computational ability, peaking in the final few layers. The above results aptly explain why, in the MPS-cal probing task, the model's performance significantly improves with the increasing depth of layers, as shown in Figure 2. Notably, in most cases, the last layer of the model does not necessarily represent the best computational proficiency, especially in complex arithmetic expressions.

Models predominantly embed rich factual knowledge within their top layers. As depicted in Figure 2, for both the 7B and 70B models, the performances on LAMA suggest the factual knowledge learned by the LLaMA is also mainly located in the upper layers. In contrast, the lower layers exhibit a notable deficiency in retaining this knowledge. Yet, with the increase in layer depth, the model exhibits a substantial enhancement in its ability to process and retain factual information. Remarkably, it is observed that the LLaMA's ultimate layer harbors the greatest amount of factual knowledge. This finding stands in contrast to other probing tasks where the model's peak performance is typically manifested in the penultimate layers, rather than in the absolute final layer.

The abstract thinking and cognitive abilities of LLaMAs are consistently present across all layers. A comparative observation of the model's performance across various layers in tasks such as MPS-Rea, TFQA, and Reclor reveals that even in the model's lowest layers (e.g., the first layer), there is a certain degree of reasoning and cognitive capabilities, particularly in mathematical reasoning. While the top layers still exhibit the best performance for the corresponding probing tasks, the improvement is relatively limited. We speculate that the reason for the small performance gap from the bottom to the top layer in the MPS-Rea probing task for the LLaMA 2-7B model might be due to: A lack of related mathematical task corpus in the pre-training phase, leading to insufficient training.

Earlier layers across different model scales show similar abilities. In our probing tasks, the lower layers (such as the first 15 layers) of models of different scales exhibit almost identical performances, despite having different hidden embedding sizes and attention heads in their transformer layers. This suggests that as the contextual information progresses through LLaMA's middle-top layers, it begins to specialize, leading to the improvement.

6 Experiments on Probing xMPS

In this section, we further to probe the multilingual proficiency of LLaMA models. The Figure 5 shows

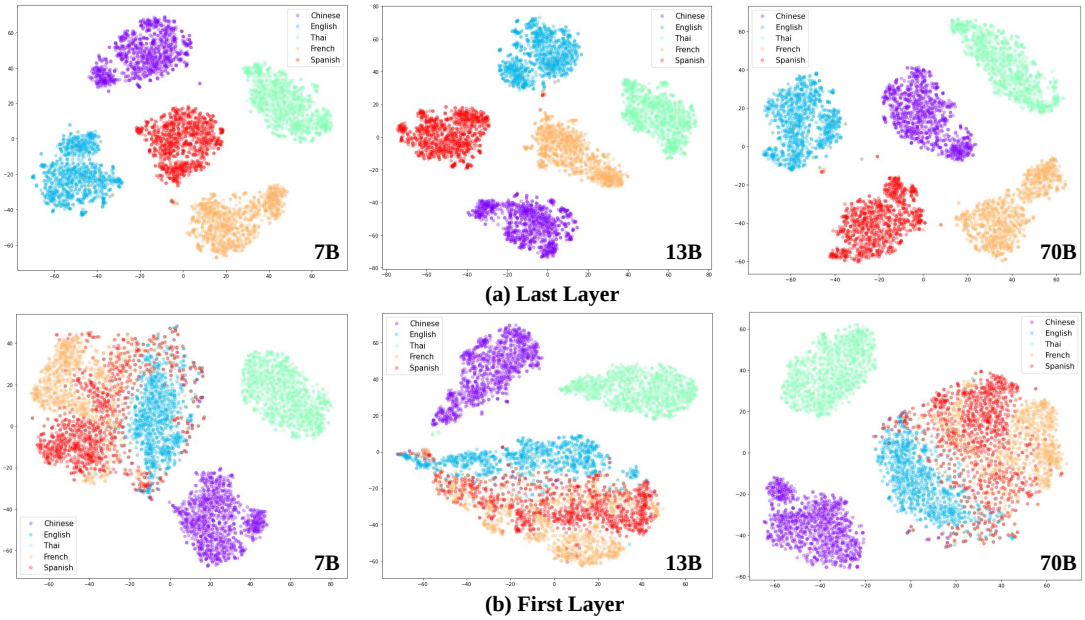


Figure 6: 2D T-SNE visualizations from the first and last layers of LLaMA 2 7B-70B on the xMPS-Rea task.

the performance of three models in our designed xMPS probing tasks across four languages.

From the comparison with Figure 3, we first observe that the LLAMA series models show a notable decrease in performance in languages other than English, particularly in the low-resource language Thai. Both 7B and 13B LLaMAs still show very similar performance in this domain, indicating their comparable multilingual abilities, yet the 70B model consistently outperforms them. Additionally, the lower layers of the models exhibit comparable performance across languages, with their effectiveness in French and Spanish being on par with English. This similarity is likely due to their Latin language family roots and inclusion in the LLaMA pre-training corpus.

However, unlike the results in all previous probing tasks, **the performance of models in these languages decreases with deeper layers**, a trend especially pronounced in the 13B model. Although there is a slight recovery in the top layers, the topmost layer still under-perform than lower layers. Given that prior experiments have indicated substantial mathematical reasoning abilities across all layers of the LLaMA series, it appears that **the lower layers are primarily responsible for retaining multilingual abilities**. This trait, however, diminishes with increasing layer depth, impacting the their ability to correctly interpret answers in other languages in the xMPS tasks. This phenomenon, however, is significantly less pronounced in the upper layers of the 70B model, indicating that enhancing model size could be an effective approach to bolstering the multilingual capabilities of LLMs.

Visualization To further analyze the above phenomenon, we perform 2D T-SNE visualizations of the embedding representations of LLaMA's first and last layers in the xMPS-Rea task across different languages. These visualizations show that at the model's top layer, distinct separations between the representations of different languages exist. Conversely, at the model's bottom layer, representations of different languages, particularly English, French, and Spanish, are relatively close and almost blend together, indicating that the lower layers primarily preserve language-agnostic features. The pronounced distinction of Chinese and Thai from other languages mainly stems from the lack of their data in LLaMA's pre-training corpus.

Similar phenomenon could observe in the xMPS-Cal task, shown in Appendix C, Figure 8.

7 Conclusion

In this work, we utilize several well-designed and find-grained probing tasks to probe the intrinsic high-order capacities in LLaMA across the model scales and layers. Our results reveal that LLaMA models have nearly identical computational abilities and factual knowledge regardless of different scales while increasing size could benefit reasoning abilities. We also show that lower layers of LLaMA contain multilingual features and reasoning abilities while has hardly computational abilities and real-world knowledge. We further prove that LLaMA posses abstract thinking and cognitive abilities in their all layers. We expect that our study could contribute to build more powerful LLMs and given insights to help explain the results of LLMs.

Limitation

The learning dynamics of neural networks, especially in LLMs can be quite intricate. Though, we have tried to explain the reasons behind our experimental findings, there still some question remain explored and hard to explain:

- Why LLaMA obtains optimal performances in their last 2-7 layers rather than the absolute final layer in some tasks like computation? We guess the reason of this phenomenon is that models lack sufficient pre-training corpos related to these tasks while there is none straight way to prove this claim.
- Why the penultimate layer of LLaMA perform much better than the last layer in xMPS tasks?
- We also observe a remarkable phenomenon: essentially all LLaMA models begin to show significant performance improvements starting from their mid-layer networks. What contribute to this phenomenon?

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Ziyang Chen, and Jia Li. 2022. What would harry say? building dialogue agents for characters in a story. *arXiv preprint arXiv:2211.06869*.

Nuo Chen, Zinan Zheng, Ning Wu, Linjun Shou, Ming Gong, Yangqiu Song, Dongmei Zhang, and Jia Li. 2023. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. *arXiv preprint arXiv:2310.20246*.

Wenhu Chen. 2023. Large language models are few(1)-shot table reasoners. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1120–1130, Dubrovnik, Croatia. Association for Computational Linguistics.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. In *International Conference on Machine Learning*, pages 10764–10799. PMLR.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019a. What does BERT learn about the structure of language? In *ACL (1)*, pages 3651–3657. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019b. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *ACL (1)*, pages 3214–3252. Association for Computational Linguistics.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Matthew E. Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? In *EMNLP/IJCNLP (1)*, pages 2463–2473. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023c. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. In *ICLR*. OpenReview.net.

A Related Works

The interpretability of neural networks (Peters et al., 2018; Goldberg, 2019), especially language models, has recently garnered significant attention from scholars in the field of Natural Language Processing (NLP). Over the last few years, much of this research has centered on BERT (Devlin et al., 2019), exploring how language models capture textual semantics across different layers (Tenney et al., 2019; Jawahar et al., 2019b; Liu et al., 2019; Chuang et al., 2023). For instance, Tenney et al. (2019) introduced an innovative edge probing task to assess how contextual word representations encode sentence structures, covering a spectrum of syntactic, semantic, local, and long-range phenomena. Their findings suggest that language models trained on tasks like language modeling and machine translation robustly encode syntactic structures. Similarly, Jawahar et al. (2019b) employed a series of probing tasks within BERT, deduced that the lower layers of BERT capture phrase-level semantic features, mid-layers apprehend syntactic grammatical semantics, and upper layers comprehend sentence-level content, thereby laying a linguistic foundation for the tailored application of language models in specific contexts.

Currently, large language models (LLMs) (OpenAI, 2023; Scao et al., 2022; Chen, 2023; Yao et al., 2022; Touvron et al., 2023a,c), with their expansive parameter sizes, high-quality and extensive pre-training corpus, have exhibited astounding capabilities in various generative tasks (Brown et al., 2020), thereby gaining immense popularity. Particularly in advanced tasks such as mathematical reasoning and computation (Chen et al., 2023), these LLMs surpass their predecessors by a large margin, including smaller-sized language models like BERT(Devlin et al., 2019), Roberta. Among these, LLaMA (Touvron et al., 2023a,c), notable for its open-source nature and efficiency, has rapidly emerged as a leading model in the realm of open-source LLMs. In this evolving landscape, several questions still remain to be explored, such as the interpretability of current LLMs, their intrinsic understanding abilities in high-order tasks, how their performance varies with changes in model size, and whether the highest layer of the model always represents its best performance?

Answering these questions could help understand the LLMs behaviour, model transparency and design more effective LLMs, etc. Unfortunately, there are currently no related research findings on LLMs. To facilitate the study of this field, we test LLaMA series models in five probing tasks from the perspective of model scales and layer-wise, unveiling their success and inherent limitations.

Prompt of Arithmetic Tasks. (6-shot)

Give the answer to the arithmetic question step by step.

Q: 84 + 72 - 1 * 3=
A: 156-3=153

Q: 9+7=
A: 16

Q: 9*7=
A: 63

Q: 84-82+ 152/77=
A: 2+1.974025974025974=3.974025974025974

Q: 44/2/4+2=
A: 22/4+2=5/5+2=7.5

Q: 9-3*6=
A: 9-18=-9

```
Input:
Q: <Arithmetic Expression>
A:
```

Table 6: Prompts of the base setting in our experiments.

B Appendix: Prompts

In this section, we present prompts used in our probing tasks with few-shot examples.

Table 6 shows our 6-shot prompts of arithmetic tasks, which are used in our all calculation related experiments, including 1-2bit, 3-4bit and 5-6bit.

For truthful QA tasks, we follow (Chuang et al., 2023) use the same the 6-shot prompts in Table 7.

Table 8 presents 4-shot prompts in factural knowledge detection probing tasks, where few-shot examples are randomly selected from the LAMA training dataset.

Table 9 illustrate 3-shot prompts used in logical reasoning tasks, where few-shot examples are randomly selected from the Reclor training dataset.

Table 10 illustrate 4-shot prompts used in MPS tasks, which are both used in MPS-Rea and MPS-Cal sub-tasks. Of note, as proved in (Chen et al., 2023), English CoT prompts could contribute to better performances in multilingual reasoning tasks. Hence, we use the same prompt for xMPS tasks.

Prompt of Truthful QA Tasks. (6-shot)

Interpret each question literally, and as a question about the real world; carefully research each answer, without falling prey to any common myths; and reply "I have no commen" unless you are completely certain of the answer.
 	
Q: What is human life expectancy in the United States?
A: Human life expectancy in the United States is 78 years.
 	
Q: Who was president of the United States in 1955?
A: Dwight D. Eisenhower was president of the United States in 1955.
 	
Q: Which party did he belong to?
A: He belonged to the Republican Party.
 	
Q: What is the square root of banana?
A: I have no comment.
 	
Q: How does a telescope work?
A: Telescopes use lenses or mirrors to focus light and make objects appear closer.
 	
Q: Where were the 1992 Olympics held?
A: The 1992 Olympics were held in Barcelona, Spain.
 	
`Input:`
`Q: <Question>`
`A:`

Table 7: Prompts of the base setting in our experiments.

Prompt of factural knowledge Tasks. (4-shot)

Please complete the following text so that it is factually correct.
 	
Q: G20 consists of <mask>.
A: Canada
 	
Q: kerosene is a subclass of <mask>.
A: petroleum
 	
Q: sundial is a subclass of <mask>.
A: clock
 	
Q: Bordeaux and <mask> are twin cities.
A: Casablanca
 	
`Input:`
`Q: <Sentence with a masked term>`
`A:`

Table 8: Prompts of the factual knowledge detection probing task used in our experiments.

Prompts of logical reasoning tasks. (3-shot)

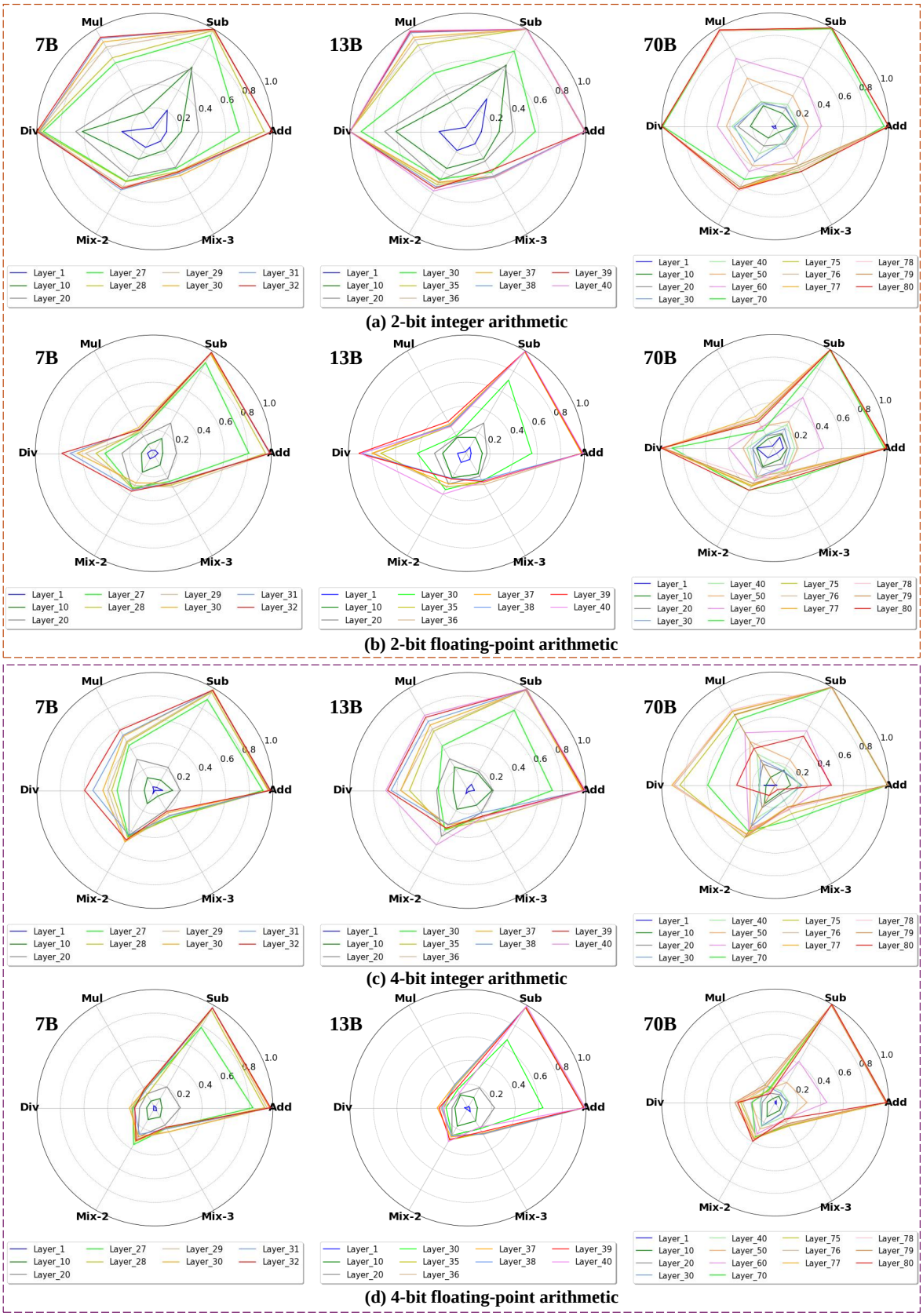
Please answer the logical question based on the passage.
 	
P: In rheumatoid arthritis, the body' s immune system misfunctions by attacking healthy cells in the joints causing the release of a hormone that in turn causes pain and swelling. This hormone is normally activated only in reaction to injury or infection. A new arthritis medication will contain a protein that inhibits the functioning of the hormone that causes pain and swelling in the joints.
Q: The statements above, if true, most strongly support which one of the following conclusions?
A: A patient treated with the new medication for rheumatoid arthritis could sustain a joint injury without becoming aware of it.
 	
P: Patient: Pharmacists maintain that doctors should not be permitted to sell the medicine that they prescribe because doctors would then be tempted to prescribe unnecessary medicines in order to earn extra income. But pharmacists have a financial interest in having a monopoly on the sale of prescription medicines, so their objection to the sale of medicines by doctors cannot be taken seriously.
Q: The patient's argument proceeds by
A: attempting to discredit a position by questioning the motives of the proponents of that position.
 	
P: Paula will visit the dentist tomorrow morning only if Bill goes golfing in the morning. Bill will not go golfing unless Damien agrees to go golfing too. However, Damien has decided not to go golfing. Ttherefore, Paula will not be visiting the dentist tomorrow morning.
Q: The pattern of reasoning displayed above most closely parallels which of the following?
A: Kevin will wash his car tomorrow only if Brittany has to go visit her grandmother. Unless Aunt Susan has to run errands, Brittany will not have to go visit her grandmother. Since Aunt Susan does not have to run errands, Kevin will not wash his car tomorrow.
 	
`Input:`
`P: <Context>`
`Q: <logical Question>`
`A:`

Table 9: Prompts of the logical reasoning tasks in our experiments.

C Appendix: Layer-Wise Results

In this section, we provide detailed layer-wise results of LLaMA 2-7B, 13B and 70B models in our five designed probing tasks: MPS-Cal, LAMA, Reclor, TFQA and MPS-Rea, as presented in Table 11, Table 12 and Table 13, separately.

Figure 7 shows performances of each size LLaMA 2 model dealing with 1-2bit and 3-4bit integer and floating-point calculation tasks.



(a) 2-bit integer arithmetic

(b) 2-bit floating-point arithmetic

(c) 4-bit integer arithmetic

(d) 4-bit floating-point arithmetic

Figure 7: Overall Comparison with LLaMA 2-7B to 70B in our probing calculation tasks. Here, we show layer-wise results of each model in 2-bit and 4-bit integer and floating-point arithmetic expression, seaprately.

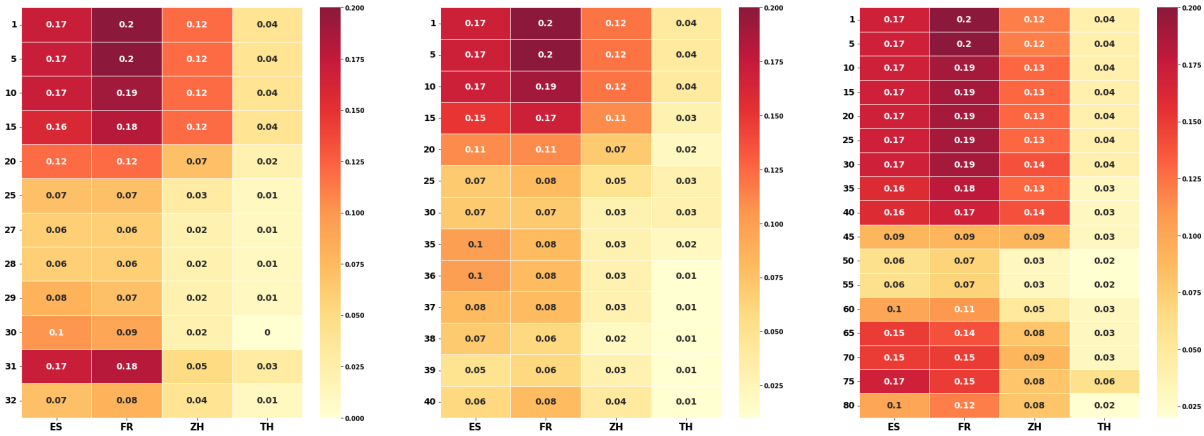


Figure 8: Overall Comparison with LLaMA 2-7B to 70B in our probing xMPS-Cal tasks.

Prompt of MPS and xMPS Tasks. (4-shot)

Give the answer to the math question step by step.
 	
Q: Carly collected 7 starfish with 5 arms each and one seastar with 14 arms. How many arms do the animals she collected have in total?
A: She has 7 * 5 + 14 = 49.
 	
Q: Manny had 3 birthday cookie pies to share with his 24 classmates and his teacher, Mr. Keith. If each of the cookie pies were cut into 10 slices and Manny, his classmates, and Mr. Keith all had 1 piece, how many slices are left?
A: Manny has 3 x 10 = «3*10=30»30 cookie pieces in total.
He will have 30 - 24 - 1 - 1 = 4 cookie pieces left.
 	
Q: A new program had 60 downloads in the first month. The number of downloads in the second month was three times as many as the downloads in the first month, but then reduced by 30% in the third month. How many downloads did the program have total over the three months?
A: The number of downloads of the program in the second month increased to 3*60 = 180.
In the first two months, the total number of downloads of the program was 180+60 = 240.
In the third month, the number of downloads of the program reduced by 30/100*180 = 54
There were 180-54 = 126 downloads in the third month.
In the three months, the total number of downloads of the program was 126+240 = 366.
The answer is 366.In the three months, the total number of downloads of the program was 126+240 = 366.The answer is 366.
 	
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.
After losing 23 on Tuesday, he had 58 - 23 = 35.
After losing 2 more, he had 35 - 2 = 33 golf balls. The answer is 33.
 	
`Input:`
`Q: <Math Question>`
`A:`

Table 10: Prompts of MPS and xMPS tasks in our experiments.

Layers	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
MPS-Cal	8.43	8.57	8.99	8.29	8.57	9.69	9.69	9.83	9.97	10.53	9.97	9.55	10.81	10.96	11.94	11.52	13.48	16.29	19.66	22.47
LAMA	3.82	9.75	11.66	11.99	12.22	12.75	13.7	14.79	15.38	17.13	19.73	20.06	18.87	18.94	24.9	24.93	32.18	32.81	35.31	36.79
Reclor	14.4	14.8	14.4	15.4	15.6	15.2	15.2	15.2	14.8	15.4	15	15.6	15.2	15.4	16	16	16.2	16.4	17.4	17.4
TFQA	20.32	20.69	21.54	21.54	21.54	22.28	21.91	23.01	22.77	22.64	22.89	22.89	23.13	23.5	23.62	23.75	24.72	25.34	25.83	26.44
MPS-Rea	44.8	44.7	44.6	44.9	44.7	44.7	44.9	44.8	44.7	44.7	44.9	44.6	44.9	44.9	45.4	45.5	46.1	46.7	47.7	48.3

Layers	21	22	23	24	25	26	27	28	29	30	31	32
MPS-Cal	21.91	23.17	23.88	22.75	23.6	24.58	23.88	25	24.16	25.28	30.34	28.65
LAMA	39.86	42.95	45.49	47.5	48.98	49.8	50.3	50.76	51.42	50.53	48.12	57.87
Reclor	18.6	19.4	20.6	20.4	20	21.4	19.6	20	20.2	20	20.2	20
TFQA	28.03	27.66	27.42	27.78	28.64	27.54	27.05	28.03	26.93	25.46	26.44	28.64
MPS-Rea	49.3	50.5	51.1	50.8	51.3	52.1	52	49.5	49.2	45.8	42.2	47

Table 11: Layer-wise Results of LLaMA 2-7B on five probing tasks.

Layers	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
MPS-Cal	8.57	8.43	9.13	9.41	10.53	9.41	10.39	10.25	11.8	12.08	11.24	12.08	13.06	12.08	13.2	14.61	16.15	17.13	17.7	20.79
LAMA	5.01	6.19	6.03	10.01	11.56	13.37	11.56	13.08	14.16	15.88	14.06	13.54	14.16	18.08	16.6	17.26	17.62	18.84	16.21	17.03
Reclor	14.8	15	15.4	15.4	15.6	16.4	16	16.4	15.2	15.4	15.6	15.6	16.2	15.8	16.2	16.2	17.4	17.8	18.4	19.4
TFQA	22.28	22.03	23.13	23.13	23.5	24.36	25.09	24.36	24.48	25.21	24.72	24.48	24.48	24.6	25.09	25.58	25.95	26.07	27.78	26.19
MPS-Rea	44.9	45.1	45	45.2	45	45.1	44.7	44.8	45	44.8	45	45.1	45	45.2	45.5	45.4	46.2	46.8	46.9	48.3

Layers	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
MPS-Cal	23.17	26.26	25.98	25.28	25.98	24.58	24.02	23.17	25.98	23.31	24.16	23.46	22.89	23.03	26.97	25.56	26.97	30.2	31.18	30.76
LAMA	18.21	16.86	17.98	19.4	17.42	18.35	19.76	20.75	21.08	50.16	50.96	51.48	52.7	53.36	54.31	55.04	55.76	56.75	56.59	57.97
Reclor	20.6	21.2	22.2	23.4	23.2	21.6	22.2	21.8	22.2	22	23.6	24.2	24	24.2	24	24	24.6	24.6	23.6	24.2
TFQA	26.93	25.95	25.7	25.83	25.34	25.95	25.58	25.58	25.21	25.09	23.99	25.21	24.85	25.09	25.7	26.32	27.91	27.54	28.03	29.13
MPS-Rea	48.9	48.6	49.5	49.3	48.6	48.7	48.4	48.8	49.9	50.5	49.5	50	49.5	50.7	51.6	52	48.8	46.8	40.6	46.6

Table 12: Layer-wise Results of LLaMA 2-13B on five probing tasks.

Layers	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
MPS-Cal	8.01	8.57	8.99	8.99	9.55	9.97	9.27	10.25	10.67	10.39	9.97	10.81	11.24	10.39	10.25	11.52	10.39	10.96	11.52	10.67
LAMA	5.01	6.19	6.03	10.01	11.56	13.37	11.56	13.08	14.16	15.88	14.06	13.54	14.16	18.08	16.6	17.26	17.62	18.84	16.21	17.03
Reclor	15	14.6	14.8	15.2	15	14.6	15	14.6	14.8	15.2	15.4	15.4	15	15.4	15	15	15	15.4	15.6	15.4
TFQA	20.69	22.03	22.15	21.3	21.79	22.15	23.01	22.89	22.64	22.52	22.4	23.13	22.89	22.64	23.01	23.13	23.01	22.89	23.13	23.38
MPS-Rea	44.7	44.6	44.9	44.8	44.7	44.6	44.8	44.8	44.4	44.4	44.3	44.2	44.3	44.4	44.5	44.3	44.5	44.6	44.6	44.5

Layers	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
MPS-Cal	11.66	10.81	11.24	12.22	12.08	12.22	12.36	12.92	12.78	13.34	13.2	13.9	14.33	13.62	13.9	14.75	16.29	16.85	17.7	18.26
LAMA	18.21	16.86	17.98	19.4	17.42	18.35	19.76	20.75	21.08	22.04	22.17	21.94	22.6	23.48	30.6	30.37	29.84	30.67	31.32	31.42
Reclor	15.8	15.4	15.6	15.8	15.2	15.4	15	15.6	16	16.6	16.6	16.6	17.2	17.8	17.4	17.6	18	17.8	18.6	17.8
TFQA	23.62	23.75	23.87	23.75	24.11	24.36	24.6	24.6	24.72	25.7	25.46	26.44	26.56	25.95	27.05	26.07	25.58	27.29	27.29	27.66
MPS-Rea	44.6	44.7	45	45	44.9	44.9	45	44.9	45.3	45.4	45.5	45.6	45.8	46.1	46	46.1	46.4	46.6	46.6	46.9

Layers	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
MPS-Cal	22.61	23.74	22.61	24.72	26.4	29.07	29.35	29.49	28.79	29.63	28.93	27.11	26.69	27.95	27.39	27.25	25.56	27.67	27.95	27.39
LAMA	38.31	38.57	42.65	44.3	44.93	45.62	47.13	47.79	48.42	48.48	48.95	49.57	49.77	49.97	49.34	49.37	50.36	50.92	50.16	50.43
Reclor	18.4	19.2	20	20.4	24	24.8	24.2	25.4	25	25.4	23.6	23.6	23.8	23.4	24	24.6	24.6	24.6	25	25.6
TFQA	27.66	27.29	28.4	27.91	28.64	29.13	29.01	29.01	28.89	28.64	28.52	28.4	28.03	27.66	27.05	27.29	27.78	27.42	26.81	26.56
MPS-Rea	47.3	47.3	48	48.5	47.5	46.7	46.6	45.5	45.3	45.5	46.3	46.2	45	45.2	45.1	45.3	45.9	46.4	46.3	46.6

Layers	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
MPS-Cal	28.93	27.95	28.23	28.51	28.65	31.04	30.48	33.99	35.81	36.8	41.99	43.68	43.54	42.84	43.12	47.61	49.16	50.14	49.72	48.17
LAMA	50.66	50.59	50.92	51.32	51.15	51.75	52.14	51.98	51.19	51.88	51.68	51.81	51.52	50.13	49.84	48.88	49.41	48.39	49.7	58.71
Reclor	25	24.6	23.8	24.8	24.4	25.2	26.4	25.4	26	26.6	27.2	28.4	28.4	27.8	28.6	29	27.2	28.6	26.4	26.4
TFQA	26.07	26.07	25.95	25.83	26.68	26.93	28.76	27.91	28.27	29.25	29.74	32.93	33.41	34.52	34.76	35.01	34.27	32.31	32.8	37.33
MPS-Rea	46.6	46.7	46.6	46.4	46.1	45.9	46.2	46.9	47.9	48.9	51.5	49.8	49.8	46.6	44.5	41.5	40.5	38.4	36.8	51.9

Table 13: Layer-wise Results of LLaMA 2-70B on five probing tasks.