

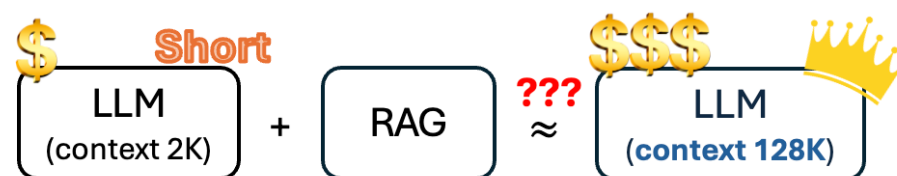


GSM-Infinite: How Do Your LLMs Behave over Infinitely Increasing Context Length and Reasoning Complexity?

Yang Zhou*, Hongyi Liu*, Zhuoming Chen, Yuandong Tian, Beidi Chen

Agenda for the talk

- Long-context LLMs are essential for AI finding solutions for intellectual challenges of humanity
- Existing long-context benchmarks are insufficient
 - Low Complexity
 - Detectable Noise
 - Scarce Quantity
- Using Computational Graphs to generate higher-quality long-context reasoning problems
- Generating the synthetic benchmark of GSM-Infinite
- Evaluation and Findings



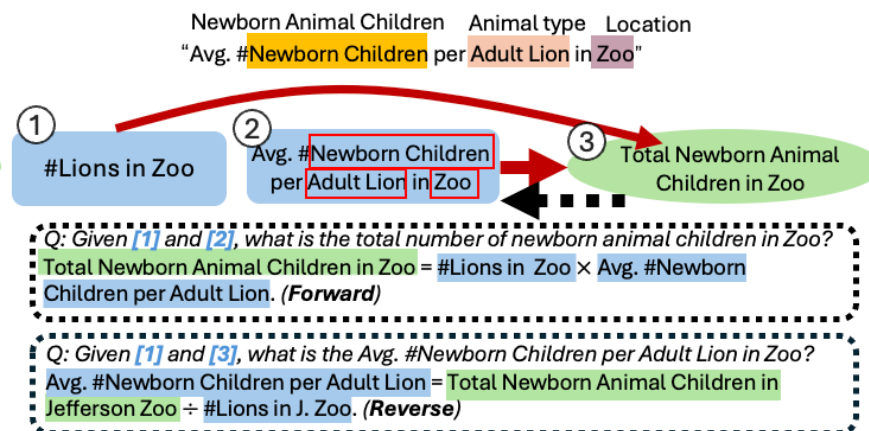
Datasets	Context RAG	LC-LLM
RULER (64k)	84.6	85.9
LongBenchv2	33	30
LOFT (128k)	54.1	41.9
GSM-Infinite* (Ours)	13.1	66.3

* GSM-Infinite: we select Medium and report avg. acc. ops from 2 to 10.

Agenda for the talk

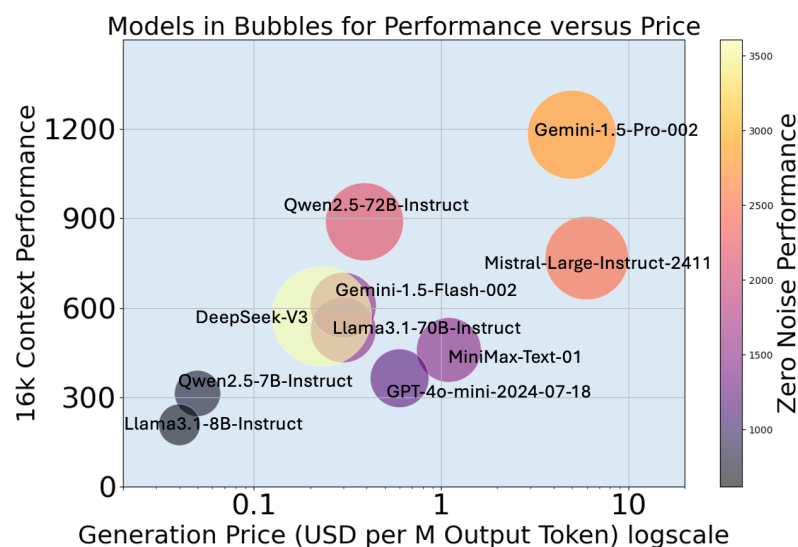
- Long-context LLMs are essential for AI finding solutions for intellectual challenges of humanity
- Existing long-context benchmarks are insufficient
 - Low Complexity
 - Detectable Noise
 - Scarce Quantity
- Using Computational Graphs to generate higher-quality long-context reasoning problems
- Generating the synthetic benchmark of GSM-Infinite
- Evaluation and Findings

Three-entity variables additionally induces hidden operations \times and \div



Agenda for the talk

- Long-context LLMs are essential for AI finding solutions for intellectual challenges of humanity
- Existing long-context benchmarks are insufficient
 - Low Complexity
 - Detectable Noise
 - Scarce Quantity
- Using Computational Graphs to generate higher-quality long-context reasoning problems
- Generating the synthetic benchmark of GSM-Infinite
- Evaluation and Findings



Long-context LLMs are Getting Amazingly Strong

Gemini 1.5 Pro is getting almost Perfect Score on 10M Context Retrieval

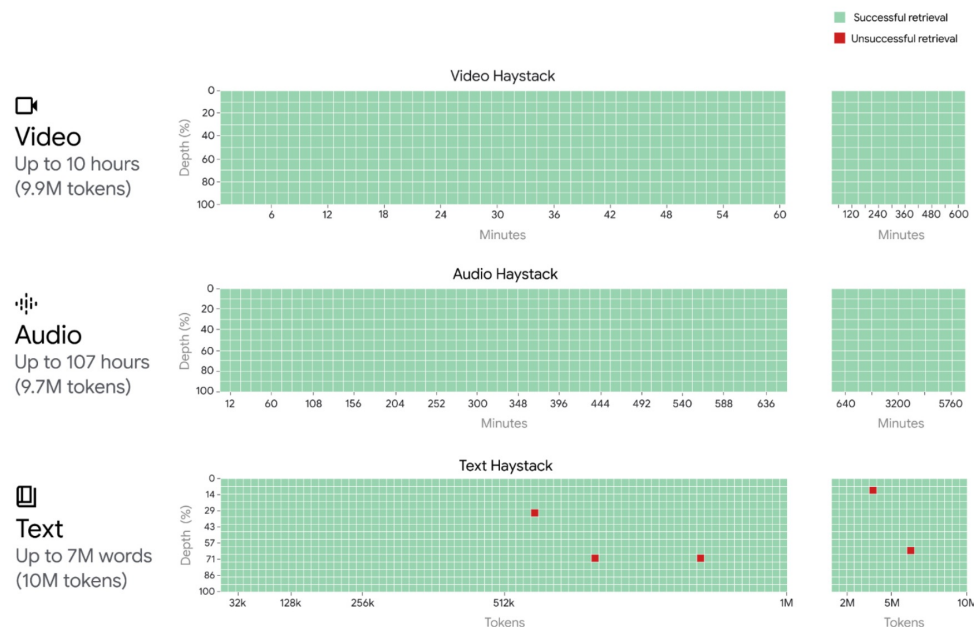



Figure 1 | Gemini 1.5 Pro achieves near-perfect “needle” recall (>99.7%) up to 1M tokens of “haystack” in all modalities, i.e., text, video and audio. It even maintains this recall performance when extending to 10M tokens in the text modality (approximately 7M words); 9.7M tokens in the audio modality (up to 107 hours); 9.9M tokens in the video modality (up to 10.5 hours). The x-axis represents the context window, and the y-axis the depth percentage of the needle placed for a given context length. The results are color-coded to indicate: green for successful retrievals and red for unsuccessful ones. Note that the performance for all modalities is obtained with the previously reported Gemini 1.5 Pro version from February.

Long-context LLMs are promising

- Can AI agents one-day solve humanity's most challenging intellectual problems, like advanced mathematics or scientific discovery?
 - **Note:** 88K tokens in Fermat's Last Theorem Proof
- Context-level methods (RAG) cannot capture deep, interconnected logics and semantics
- Long-context LLMs are the **only** viable path right now towards a true intellectual agent
- We need high-quality long-context benchmark with fine-grained control of both complexity and context length and sufficient test examples


Annals of Mathematics, 141 (1995), 443-551



**Modular elliptic curves
and
Fermat's Last Theorem**

By ANDREW JOHN WILES*

For Nada, Claire, Kate and Olivia



Cubum autem in duos cubos, aut quadratoquadratum in duos quadratoquadratos, et generaliter nullam in infinitum ultra quadratum potestatum in duos ejusdem nominis fas est dividere: cujus rei demonstrationem mirabilem sane detexi. Hanc marginis exiguitas non caperet.

- Pierre de Fermat ~ 1637

Abstract. When Andrew John Wiles was 10 years old, he read Eric Temple Bell's *The Last Problem* and was so impressed by it that he decided that he would be the first person to prove Fermat's Last Theorem. This theorem states that there are no nonzero integers a, b, c, n with $n > 2$ such that $a^n + b^n = c^n$. The object of this paper is to prove that all semistable elliptic curves over the set of rational numbers are modular. Fermat's Last Theorem follows as a corollary by virtue of previous work by Frey, Serre and Ribet.

Introduction

An elliptic curve over \mathbf{Q} is said to be modular if it has a finite covering by a modular curve of the form $X_0(N)$. Any such elliptic curve has the property that its Hasse-Weil zeta function has an analytic continuation and satisfies a functional equation of the standard type. If an elliptic curve over \mathbf{Q} with a given j -invariant is modular then it is easy to see that all elliptic curves with the same j -invariant are modular (in which case we say that the j -invariant is modular). A well-known conjecture which grew out of the work of Shimura and Taniyama in the 1950's and 1960's asserts that every elliptic curve over \mathbf{Q} is modular. However, it only became widely known through its publication in a paper of Weil in 1967 [We] (as an exercise for the interested reader!), in which, moreover, Weil gave conceptual evidence for the conjecture. Although it had been numerically verified in many cases, prior to the results described in this paper it had only been known that finitely many j -invariants were modular.

In 1985 Frey made the remarkable observation that this conjecture should imply Fermat's Last Theorem. The precise mechanism relating the two was formulated by Serre as the ϵ -conjecture and this was then proved by Ribet in

Limitations of Current Long-context Benchmarks

• Three major limitations:

- Tasks are too simple (text retrieval, text summarization, QA)
- Short-context problems bloated to longer context (detectable filler text)
- Too scarce

• A close look at RULER

Task	Configuration	Example
Single NIAH (S-NIAH)	type_key = word type_value = number type_haystack = essay size_haystack \propto context length	(essays) One of the special magic numbers for long-context is: 12345. What is the special magic number for long-context mentioned in the provided text? Answer: 12345
Multi-keys NIAH (MK-NIAH)	num_keys = 2 type_key = word type_value = number type_haystack = essay size_haystack \propto context length	(essays) One of the special magic numbers for long-context is: 12345. One of the special magic numbers for large-model is: 54321. What is the special magic number for long-context mentioned in the provided text? Answer: 12345
Multi-values NIAH (MV-NIAH)	num_values = 2 type_key = word type_value = number type_haystack = essay size_haystack \propto context length	(essays) One of the special magic numbers for long-context is: 12345. One of the special magic numbers for long-context is: 54321. What are all the special magic numbers for long-context mentioned in the provided text? Answer: 12345 54321
Multi-queries NIAH (MQ-NIAH)	num_queries = 2 type_key = word type_value = number type_haystack = essay size_haystack \propto context length	(essays) One of the special magic numbers for long-context is: 12345. One of the special magic numbers for large-model is: 54321. What are all the special magic numbers for long-context and large-model mentioned in the provided text? Answer: 12345 54321
Variable Tracking (VT)	num_chains = 2 num_hops = 2 size_noises \propto context length	(noises) VAR X1 = 12345 VAR Y1 = 54321 VAR X2 = X1 VAR Y2 = Y1 VAR X3 = X2 VAR Y3 = Y2 Find all variables that are assigned the value 12345. Answer: X1 X2 X3
Common Words Extraction (CWE)	freq_cw = 2, freq_ucw = 1 num_cw = 10 num_ucw \propto context length	aaa bbb ccc aaa ddd eee ccc fff ggg hhh iii iii What are the 10 most common words in the above list? Answer: aaa ccc iii
Frequent Words Extraction (FWE)	$\alpha = 2$ num_word \propto context length	aaa bbb ccc aaa ddd eee ccc fff ggg aaa hhh aaa ccc iii iii What are the 3 most frequently appeared words in the above coded text? Answer: aaa ccc iii
Question Answering (QA)	dataset = SQuAD num_document \propto context length	Document 1: aaa Document 2: bbb Document 3: ccc Question: question Answer: bbb

Table 2: Task examples with flexible configurations in RULER. We use different colors to highlight queries, keys, values, and distractors in our examples.

Limitations of Current Long-context Benchmarks

- There are three facets of reasons:
 - Tasks are too simple (text retrieval, text summarization, QA)
 - Short-context problems bloated to longer context (detectable filler text)
 - Too scarce

Task	Configuration	Example
Single NIAH (S-NIAH)	type_key = word type_value = number type_haystack = essay size_haystack \propto context length	(essays) One of the special magic numbers for long-context is: 12345. What is the special magic number for long-context mentioned in the provided text? Answer: 12345
Multi-keys NIAH (MK-NIAH)	num_keys = 2 type_key = word type_value = number type_haystack = essay size_haystack \propto context length	(essays) One of the special magic numbers for long-context is: 12345. One of the special magic numbers for large-model is: 54321. What is the special magic number for long-context mentioned in the provided text? Answer: 12345
Multi-values NIAH (MV-NIAH)	num_values = 2 type_key = word type_value = number type_haystack = essay size_haystack \propto context length	(essays) One of the special magic numbers for long-context is: 12345. One of the special magic numbers for long-context is: 54321. What are all the special magic numbers for long-context mentioned in the provided text? Answer: 12345 54321
Multi-queries NIAH (MQ-NIAH)	num_queries = 2 type_key = word type_value = number type_haystack = essay size_haystack \propto context length	(essays) One of the special magic numbers for long-context is: 12345. One of the special magic numbers for large-model is: 54321. What are all the special magic numbers for long-context and large-model mentioned in the provided text? Answer: 12345 54321
Variable Tracking (VT)	num_chains = 2 num_hops = 2 size_noises \propto context length	(noises) VAR X1 = 12345 VAR Y1 = 54321 VAR X2 = X1 VAR Y2 = Y1 VAR X3 = X2 VAR Y3 = Y2 Find all variables that are assigned the value 12345. Answer: X1 X2 X3
Common Words Extraction (CWE)	freq_cw = 2, freq_ucw = 1 num_cw = 10 num_ucw \propto context length	aaa bbb ccc aaa ddd eee ccc fff ggg hhh iii iii What are the 10 most common words in the above list? Answer: aaa ccc iii
Frequent Words Extraction (FWE)	$\alpha = 2$ num_word \propto context length	aaa bbb ccc aaa ddd eee ccc fff ggg aaa hhh aaa ccc iii iii What are the 3 most frequently appeared words in the above coded text? Answer: aaa ccc iii
Question Answering (QA)	dataset = SQuAD num_document \propto context length	Document 1: aaa Document 2: bbb Document 3: ccc Question: question Answer: bbb

Table 2: Task examples with flexible configurations in RULER. We use different colors to highlight queries, keys, values, and distractors in our examples.

Limitations of Current Long-context Benchmarks

- There are three facets of reasons:
 - Tasks are too simple (text retrieval, text summarization, QA)
 - Short-context problems bloated to longer context (detectable filler text)
 - Too scarce

Task	Configuration	Example
Single NIAH (S-NIAH)	type_key = word type_value = number type_haystack = essay size_haystack \propto context length	(essays) One of the special magic numbers for long-context is: 12345. What is the special magic number for long-context mentioned in the provided text? Answer: 12345
Multi-keys NIAH (MK-NIAH)	num_keys = 2 type_key = word type_value = number type_haystack = essay size_haystack \propto context length	(essays) One of the special magic numbers for long-context is: 12345. One of the special magic numbers for large-model is: 54321. What is the special magic number for long-context mentioned in the provided text? Answer: 12345
Multi-values NIAH (MV-NIAH)	num_values = 2 type_key = word type_value = number type_haystack = essay size_haystack \propto context length	(essays) One of the special magic numbers for long-context is: 12345. One of the special magic numbers for long-context is: 54321. What are all the special magic numbers for long-context mentioned in the provided text? Answer: 12345 54321
Multi-queries NIAH (MQ-NIAH)	num_queries = 2 type_key = word type_value = number type_haystack = essay size_haystack \propto context length	(essays) One of the special magic numbers for long-context is: 12345. One of the special magic numbers for large-model is: 54321. What are all the special magic numbers for long-context and large-model mentioned in the provided text? Answer: 12345 54321
Variable Tracking (VT)	num_chains = 2 num_hops = 2 size_noises \propto context length	(noises) VAR X1 = 12345 VAR Y1 = 54321 VAR X2 = X1 VAR Y2 = Y1 VAR X3 = X2 VAR Y3 = Y2 Find all variables that are assigned the value 12345. Answer: X1 X2 X3
Common Words Extraction (CWE)	freq_cw = 2, freq_ucw = 1 num_cw = 10 num_ucw \propto context length	aaa bbb ccc aaa ddd eee ccc fff ggg hhh iii iii What are the 10 most common words in the above list? Answer: aaa ccc iii
Frequent Words Extraction (FWE)	$\alpha = 2$ num_word \propto context length	aaa bbb ccc aaa ddd eee ccc fff ggg aaa hhh aaa ccc iii iii What are the 3 most frequently appeared words in the above coded text? Answer: aaa ccc iii
Question Answering (QA)	dataset = SQuAD num_document \propto context length	Document 1: aaa Document 2: bbb Document 3: ccc Question: question Answer: bbb

Table 2: Task examples with flexible configurations in RULER. We use different colors to highlight queries, keys, values, and distractors in our examples.

Limitations of Current Long-context Benchmarks

- There are three facets of reasons:
 - Tasks are too simple (text retrieval, text summarization, QA)
 - Short-context problems bloated to longer context (detectable filler text)
 - Too scarce
 - SWE-bench on average consists of 438k lines of code
 - Baseline using RAG + LLM (BM25 as retriever) performs poorly

Github Code Completion

- SWE-bench (2024) – 2294 Problems from 12 repos
- DafnyBench (2024) – 782 ground-truths

Table 5: We compare models against each other using the BM25 retriever as described in Section 4.

Model	SWE-bench		SWE-bench Lite	
	% Resolved	% Apply	% Resolved	% Apply
Claude 3 Opus	3.79	46.56	4.33	51.67
Claude 2	1.97	43.07	3.00	33.00
ChatGPT-3.5	0.17	26.33	0.33	10.00
GPT-4-turbo	1.31	26.90	2.67	29.67

- Data requires **huge** human labor to clean + deduped + verified (**Cannot be scaled up Easily**)
- Problems cannot be quantitatively categorized in incremental difficulty, nor controlled context length (**Cannot be Controlled with Fine-grainularity**)

We argue that current long-context benchmarks cannot sufficiently evaluate the value of long-context LLMs.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, & Karthik R Narasimhan (2024). SWE-bench: Can Language Models Resolve Real-world Github Issues?. In The Twelfth International Conference on Learning Representations.

Chloe Loughridge, Qinyi Sun, Seth Ahrenbach, Federico Cassano, Chuyue Sun, Ying Sheng, Anish Mudide, Md Rakib Hossain Misu, Nada Amin, & Max Tegmark (2024). DafnyBench: A Benchmark for Formal Software Verification. arXiv preprint arXiv:2406.08467.

Simple-to-build RAG Systems are Surprisingly Robust

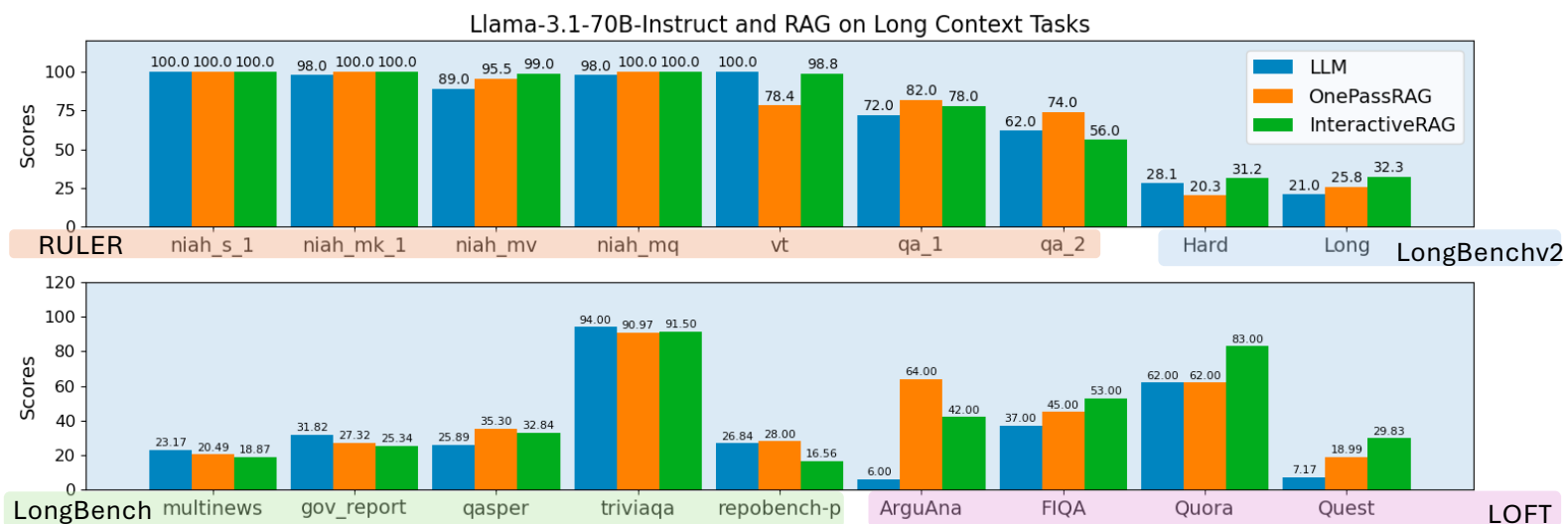


Our experiments show that RAG (retriever + decoder) is surprisingly strong and robust on **existing popular long-context benchmarks** reaching scores comparable to long-context LLMs!

However, if that is true, long-context LLMs (LC-LLMs) are almost useless on these tasks

- (prefilling) RAG only needs to tokenize **chunked** context, compute distance, topk select **chunks**
- (prefilling) LC-LLMs encodes **every token in the context**
- (decoding) RAG only asks the LLM decoder to attend to the short selected context
- (decoding) LC-LLMs needs to attend to every token in the long context

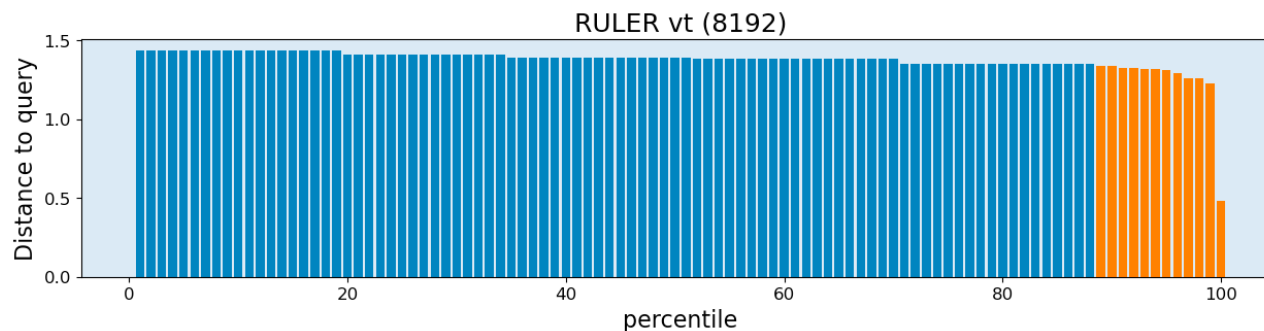
Simple-to-build RAG Systems are Robust



* Full table in the paper Appendix C.

- We use **four** common long-context benchmarks for our studies: RULER, Long-Bench, Long-Benchv2, LOFT
- **Llama-3.1-70B-Instruct (2K) + al-mpnet-base-v2** vs. **Llama-3.1-70B-Instruct (full context)**
- Here we consider a passive RAG and an active RAG [1] as the two RAG reference methods
- Full results shown in paper, we found for most tasks, **RAG matches or even surpasses the LC-LLMs**
- (Caveat) **There exists tasks where RAGs are sub-optimal**: CWE from RULER and PassageCount from LongBench
 - though these tasks don't really need NN to solve

Simple-to-build RAG Systems are Robust



- **Here we take a closer look**
- We use the retriever to compute the distance between each context chunk and query
- We rank them from large (left) to small (right) – From least relevant to most relevant deemed by the retriever
- We also manually labeled the chunks are needed to be pulled out labeled in **coral**
- Other chunks are noise so labeled in **blue**
- **The filler text are cleanly separated from the necessary context**

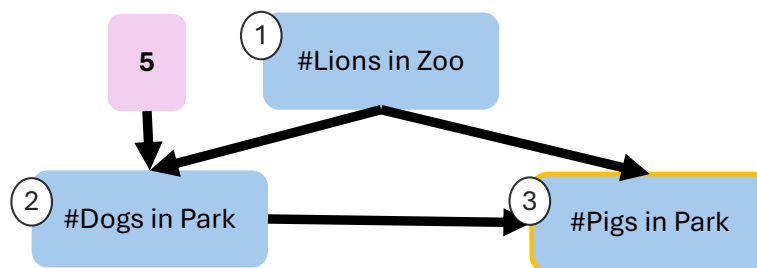
Problem Statement

Ideally, we want a long-context benchmark to have the following characteristics:

- Offers **Controllable and Scalable Complexity**
 - **Hard-to-distinguish Noise** (Only LC-LLM solvable)
 - **Infinite quantity** (or at least at large amount readily available)
-
- **How** can we develop a benchmark that contains sufficient problems at every fine-grained level of reasoning difficulty, from easy retrieval tasks to infinitely hard challenges, while providing infinitely customizable context length with high information density?

Reasoning Questions Through Computational Graph: Explicit Operations

Q: There are 7 lions in South Zoo. # dogs at North Park equals 5 + # Lions in South Zoo. # pigs in South Zoo equals the sum of # dogs at North Park and # Lions in South Zoo. What is # pigs in South Zoo.



Q: Given [1]. [2] equals 5 + [1]. [3] equals sum of [1] and [2]. [3] is?

You can see that if every Ops are stated explicitly. The conversion from and to a computational graph and a problem in natural language seems to be very easy.

Say if every Ops are explicitly stated in the problem. Then, we can do it backward. We first **generate** a graph, we randomly perturb the graph (or randomly generate new graph), and **convert** back to natural language (infinite scale up the difficulty and problems quantity for every op)

Reasoning Questions Through Computational Graph: Implicit Operations

[user] For any location mentioned in the problem, if a type of animal is never mentioned for that location, assume its inexistence.

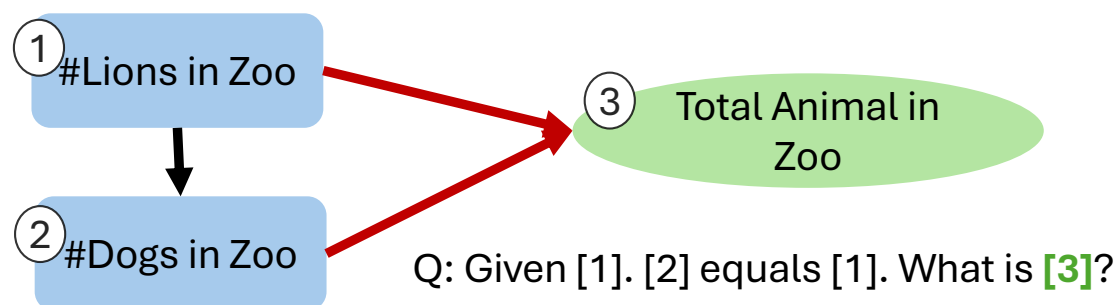
[user] # **Dogs** in Jefferson Zoo equals 5. # **Lions** in Jefferson Zoo equals # **Dogs** in Jefferson Zoo. What is the total number of animal in Jefferson Zoo?

[Answer] # Lions in Jefferson Zoo = # Dogs in Jefferson Zoo = 5. **Total number of animal in Jefferson Zoo** = # Dogs in Jefferson Zoo + # Lions in Jefferson Zoo = 5 + 5 = 10.

Note: the problem never mentions using addition to compute # Animals, but the solution requires such operation to compute the answer. The addition operations here are implicit operations, which is implied through **common sense knowledge** and natural language.

In GSM-8K, all four operations have the corresponding relationships in questions that ask for the answer to perform an implicit operations. **These operations severely impact the model performance.**

Reasoning Questions Through Computational Graph: Implicit Operations

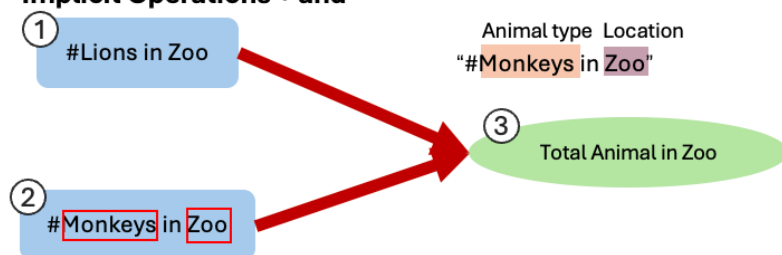


Following Physics of LM 2.1, we use the abstract – instance parameter construct. The abstract parameters are usually variables that are summative or more abstract in naming, while the instance parameters are named concretely.

The edges from instance to abstract parameters are red. The essence of implicit operations is that we omit the description of red edges, forcing the LLM to rely on its own commonsense reasoning to complete the operations.

What about implicit multiply operator?

Questions Contains Only Two-entity Variables Induces Implicit Operations + and -



It turns out that if every nodes in the computational graph are with the following naming pattern “something” in “something” (2-entity), we cannot generate multiplication.

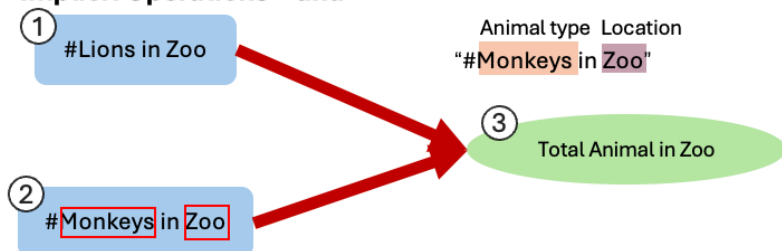
Q: Given [1], [2], what is the total number of animals in Zoo?

Total Animal in Zoo = #Lions in Zoo + #Monkeys in Zoo. (**Forward**)

Multiplication usually happens at instance dependency, or when instance has its own instance. Therefore, we needs one additional layer of dependency for generating the additional dependency (3-entity variables).

What about implicit multiply operator?

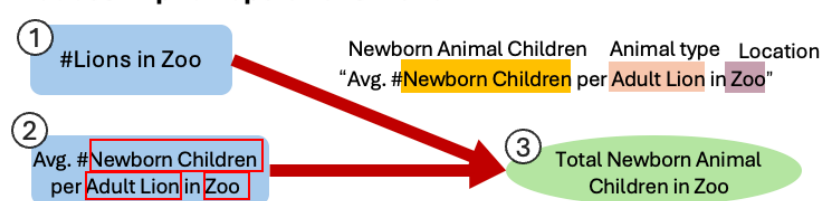
Questions Contains Only Two-entity Variables Induces Implicit Operations + and -



Q: Given [1], [2], what is the total number of animals in Zoo?

Total Animal in Zoo = #Lions in Zoo + #Monkeys in Zoo. (Forward)

Questions Containing Three-entity Variables Additionally Induces Implicit operations \times and \div



Q: Given [1] and [2], what is the total number of newborn animal children in Zoo?

Total Newborn Animal Children in Zoo = #Lions in Zoo \times Avg. #Newborn Children per Adult Lion. (Forward)

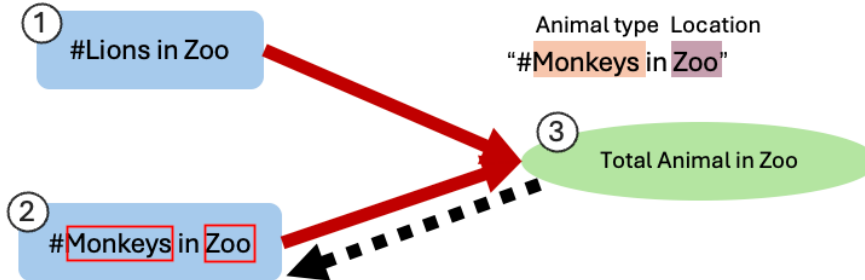
Q: # Lions in Zoo equals 7. The average # Newborn Children per Adult Lion in Zoo equals 2. What is the total # Newborn Animal Children in Zoo?

How to generate implicit – and ÷

- Though ubiquitous in GSM-8K, missing in prior works
- Naively, we still can generate these implicit operations using formulating dependency as we have shown before
- However, randomly generated “–” easily introduce negative results, floating numbers from “÷”, which cannot be easily solved without adding restrictions to the original problems
- Here we introduce **Reverse Mode**, a simple fix for the problem
 - Generate minus using plus, and divide using multiply as before. But perturb the sequence which they appear in

How to generate implicit – and ÷

Questions Contains Only Two-entity Variables Induces Implicit Operations + and -



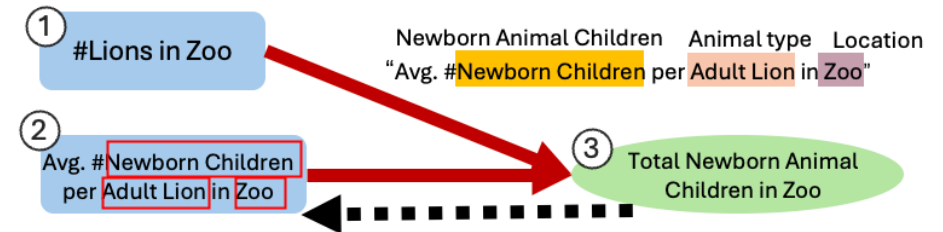
Q: Given [1], [2], what is the total number of animals in Zoo?

$\text{Total Animal in Zoo} = \text{\#Lions in Zoo} + \text{\#Monkeys in Zoo}$. (Forward)

Q: Given [1], [3], what is the Number of Monkeys in Zoo?

$\text{\#Monkeys in Zoo} = \text{Total Animal in Zoo} - \text{\#Lions in Zoo}$. (Reverse)

Questions Containing Three-entity Variables Additionally Induces Implicit operations \times and \div



Q: Given [1] and [2], what is the total number of newborn animal children in Zoo?

$\text{Total Newborn Animal Children in Zoo} = \text{\#Lions in Zoo} \times \text{Avg. \#Newborn Children per Adult Lion}$. (Forward)

Q: Given [1] and [3], what is the Avg. #Newborn Children per Adult Lion in Zoo?

$\text{Avg. \#Newborn Children per Adult Lion} = \text{Total Newborn Animal Children in Jefferson Zoo} \div \text{\#Lions in J. Zoo}$. (Reverse)

Mapping from Graph to natural language problems

- Principle: LLMs should be failing because of **reasoning deficiency** the complex logic, less about the arithmetic
 - Enforce Integer Arithmetic, all variables less than $1e3$ for ops < 30
- Linguistic familiarity (purely prompt engineering): LLMs don't like A's B. B in A. For three-entity, LLMs don't like A's B's C. Use C per B in A.
- **Avoid realistic concepts** (locations, person's full name, festival name): LLM may confused by its own memory.
- Unit alignment: "C per B in A" and "B in A" needs to have the same unit, so that the limitation to the random generation of computation graph is minimal (encouraging more diverse pattern)

Increasing Diversity Through Natural Language Templates

Template: "Crazy Zootopia"

Problem: The number of adult racoon in South Zoo equals 1 plus the total number of adult animals in Mayer Aquarium. The number of adult fox in Mayer Aquarium equals 2.

Question: What is the total number of adult animals in South Zoo?

Solution: Define adult fox in Mayer Aquarium as t ; so $t = 2$. Define total number of adult animals in Mayer Aquarium as l ; so $l = t = 2$. Define adult racoon in South Zoo as h ; $n = l = 2$; so $h = 1 + n = 1 + 2 = 3$. Define total number of adult animals in South Zoo as Y ; so $Y = h = 3$. Answer: 3.

Template: "Teachers in School"

Problem: The number of regional medical school in Brightford equals 1. The number of elementary school in Hawkesbury equals 1 plus the total number of schools in Brightford.

Question: What is the total number of schools in Hawkesbury?

Solution: Define regional medical school in Brightford as B ; so $B = 1$. Define total number of schools in Brightford as y ; so $y = B = 1$. Define elementary school in Hawkesbury as T ; $m = y = 1$; so $T = 1 + m = 1 + 1 = 2$. Define total number of schools in Hawkesbury as q ; so $q = T = 2$. Answer: 2.

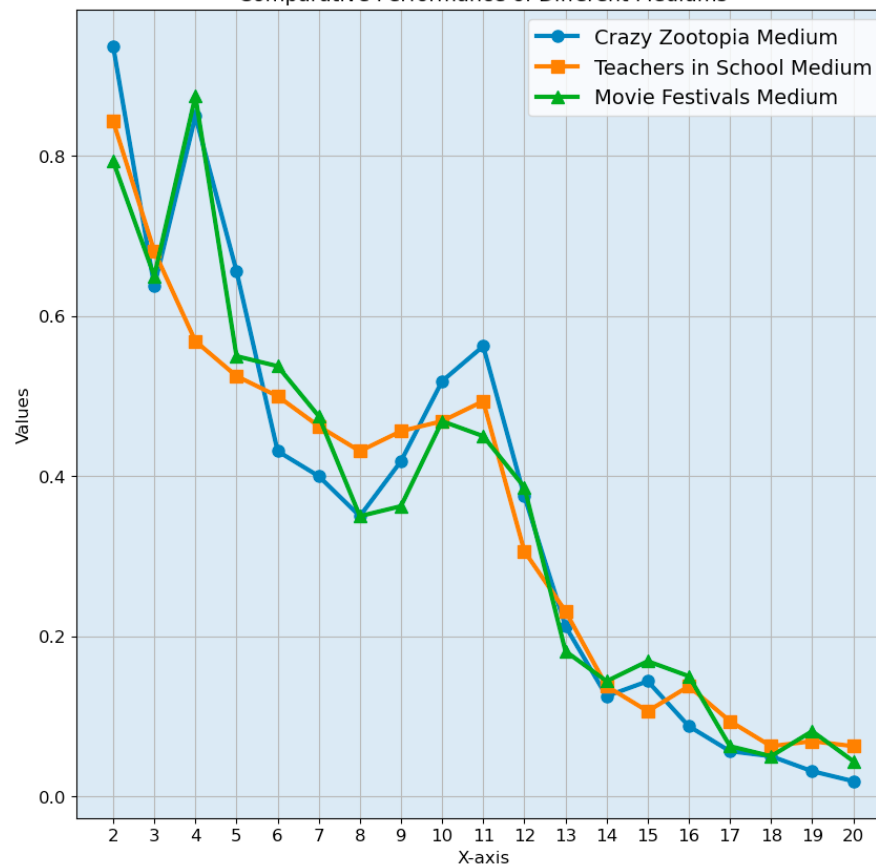
Template: "Movie Festival Awards"

Problem: The number of solemn period drama in Festival de Clairmont equals 1 plus the total number of movies in Festival de Saint-Rivage. The number of calm road movie in Festival de Saint-Rivage equals 3.

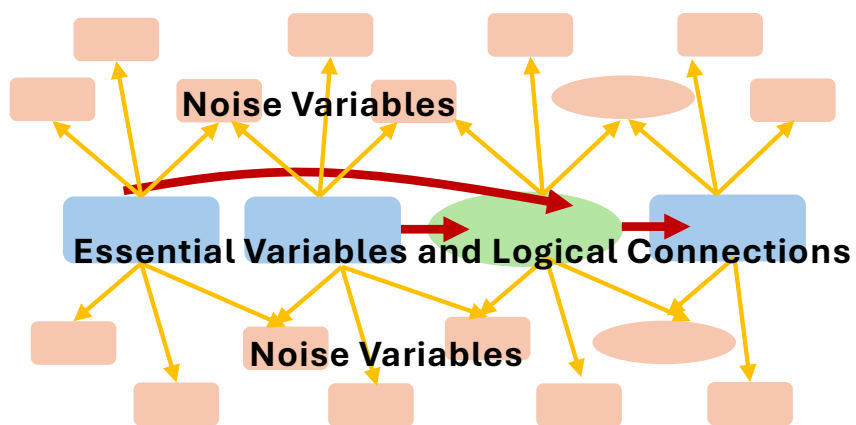
Question: What is the total number of movies in Festival de Clairmont?

Solution: Define calm road movie in Festival de Saint-Rivage as Z ; so $Z = 3$. Define total number of movies in Festival de Saint-Rivage as x ; so $x = Z = 3$. Define solemn period drama in Festival de Clairmont as e ; $o = x = 3$; so $e = 1 + o = 1 + 3 = 4$. Define total number of movies in Festival de Clairmont as G ; so $G = e = 4$. Answer: 4.

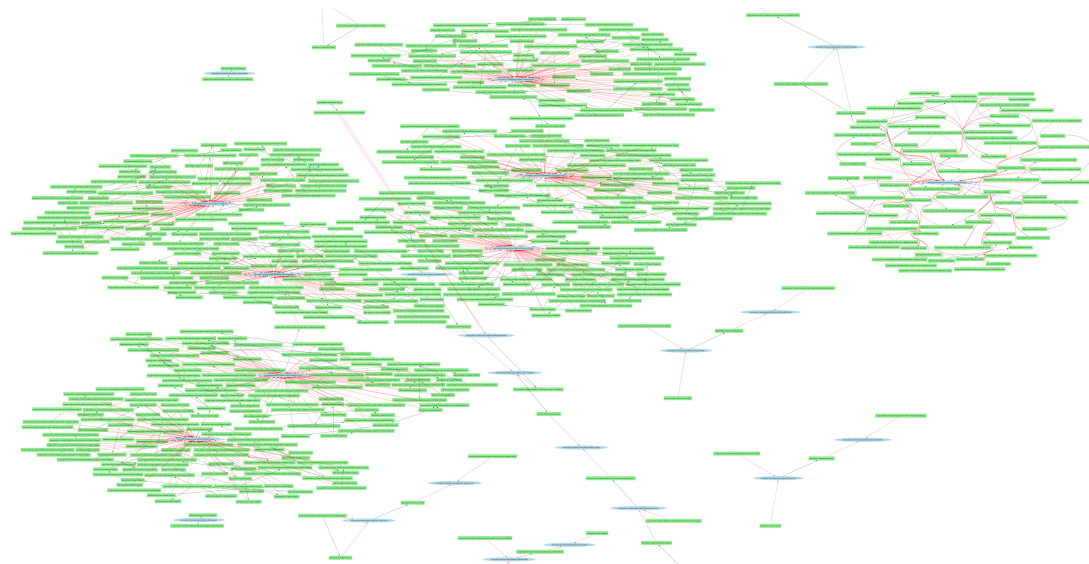
Comparative Performance of Different Mediums



Long Context through Extending Core Computational Graph



Design of Noise Addition



Actual Visualization of 8K Problem

Adding noise to boost it to longer context

The problem can be boosted to 8K, 16K, 32K, 64K, 128K



Noise addition



The Necessary Statements

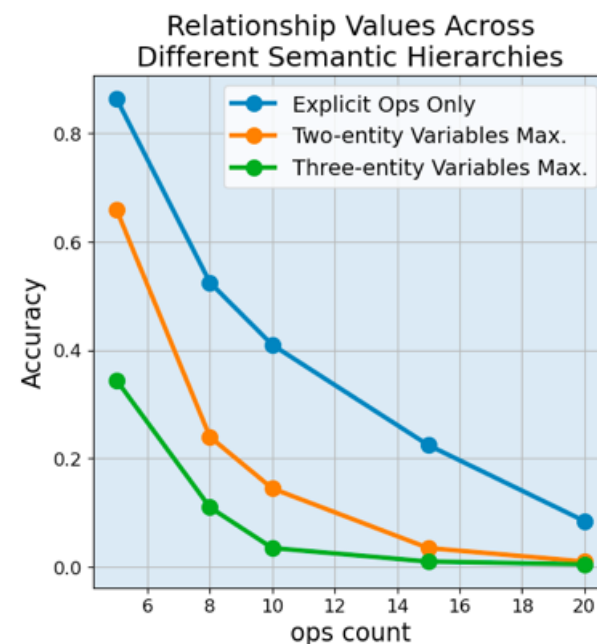
Problem: The number of adult glass frog in Moonshadow Current equals the difference between the number of adult crow in Oakridge Riverside and the total number of adult animals in Pine Ridge. The number of adult axolotl in Moonshadow Current equals the sum of the total number of adult animals in Oakridge Riverside and the total number of adult animals in Crystalbrook Stream. The number of adult emperor penguin in Crystalbrook Stream equals the sum of the total number of adult animals in Pine Ridge and the number of adult crow in Oakridge Riverside. The number of adult red-eyed tree frog in Moonshadow Current equals the number of adult crow in Oakridge Riverside. The number of adult scarlet macaw in Moonshadow Current equals the difference between the number of adult blue jay in Oakridge Riverside and the number of adult crow in Cedar Valley. The number of adult wood frog in Moonshadow Current equals the number of adult crow in Oakridge Riverside. The number of adult crow in Oakridge Riverside equals 3. The number of adult rainbow lorikeet in Moonshadow Current equals the sum of the total number of adult animals in Cedar Valley and the total number of adult animals in Pine Ridge. The number of adult marbled salamander in Moonshadow Current equals the number of adult crow in Oakridge Riverside. The number of adult tiger salamander in Moonshadow Current equals the sum of the number of adult crow in Oakridge Riverside and the number of adult blue jay in Oakridge Riverside. The number of adult fire salamander in Moonshadow Current equals the difference between the number of adult blue jay in Cedar Valley and the total number of adult animals in Pine Ridge. The number of adult coqui frog in Moonshadow Current equals the number of adult blue jay in Oakridge Riverside. The number of adult crow in Pine Ridge equals the difference between the total number of adult animals in Cedar Valley and the number of adult blue jay in Cedar Valley. The number of adult poison dart frog in Moonshadow Current equals the total number of adult animals in Cedar Valley. The number of adult hellbender in Moonshadow Current equals the total number of adult animals in Oakridge Riverside. The number of adult golden mantella in Moonshadow Current equals the total number of adult animals in Cedar Valley. The number of adult blue jay in Cedar Valley equals 1. The number of adult surinam toad in Moonshadow Current equals the sum of the number of adult crow in Oakridge Riverside and the number of adult blue jay in Cedar Valley. The number of adult cane toad in Moonshadow Current equals the difference between the total number of adult animals in Pine Ridge and the number of adult blue jay in Oakridge Riverside. The number of adult pacific tree frog in Moonshadow Current equals the number of adult crow in Oakridge Riverside. The number of adult african clawed frog in Moonshadow Current equals the total number of adult animals in Oakridge Riverside. The number of adult crow in Cedar Valley equals 3. The number of adult blue jay in Oakridge Riverside equals the total number of adult animals in Pine Ridge. The number of adult dwarf african frog in Moonshadow Current equals the total number of adult animals in Cedar Valley. The number of adult eastern newt in Moonshadow Current equals the total number of adult animals in Oakridge Riverside. The number of adult toucan in Moonshadow Current equals the total number of adult animals in Cedar Valley. The number of adult giant salamander in Moonshadow Current equals the difference between the number of adult blue jay in Cedar Valley and the total number of adult animals in Cedar Valley. The number of adult leopard frog in Moonshadow Current equals the number of adult blue jay in Cedar Valley. The number of adult bullfrog in Moonshadow Current equals the number of adult blue jay in Oakridge Riverside. The number of adult golden pheasant in Moonshadow Current equals the sum of the number of adult crow in Oakridge Riverside and the total number of adult animals in Oakridge Riverside. The number of adult peacock in Moonshadow Current equals the total number of adult animals in Cedar Valley.

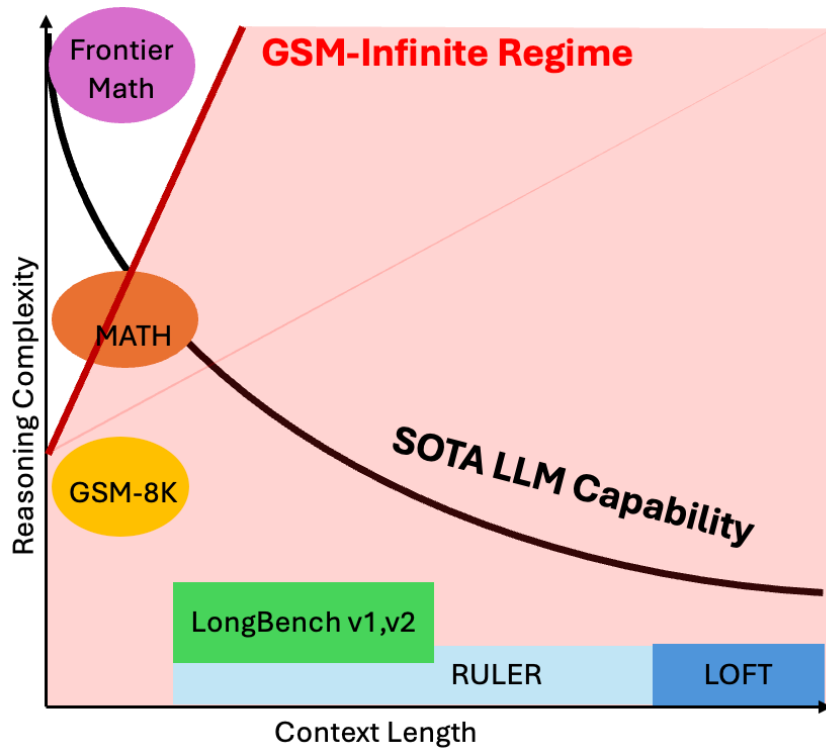
Question: What is the total number of adult animals in Oakridge Riverside?

Solution: Define adult blue jay in Cedar Valley as z ; so $z = 1$. Define adult crow in Cedar Valley as D ; so $D = 3$. Define total number of adult animals in Cedar Valley as f ; so $f = z + D = 1 + 3 = 4$. Define adult crow in Pine Ridge as o ; so $o = f - z = 4 - 1 = 3$. Define total number of adult animals in Pine Ridge as X ; so $X = o = 3$. Define adult blue jay in Oakridge Riverside as g ; so $g = X = 3$. Define adult crow in Oakridge Riverside as F ; so $F = 3$. Define total number of adult animals in Oakridge Riverside as t ; so $t = g + F = 3 + 3 = 6$. Answer: 6.

GSM-Infinite

- Generated through our generator
- Our release contains three different sub-partition of data
 - Symbolic (only with explicit Ops)
 - Medium (explicit Ops + Implicit +/-)
 - Hard (explicit Ops + Implicit +/-)
- Zero-noise + Long-context variants
- LLMs show natural performance separation on the three difficulty partitions
 - Llama-3.1-8B-Instruct performance (right)



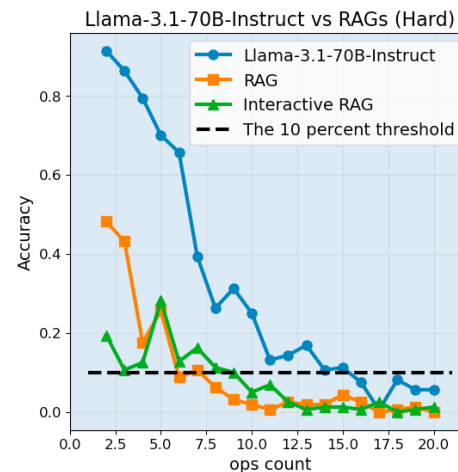
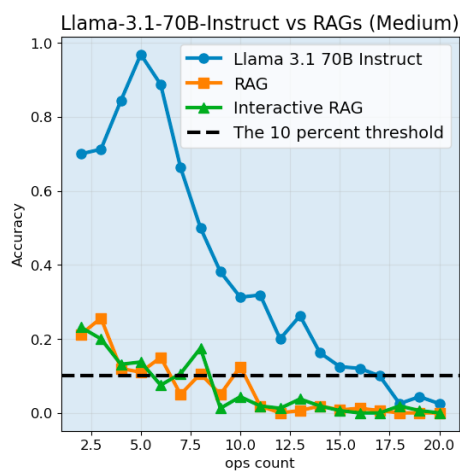
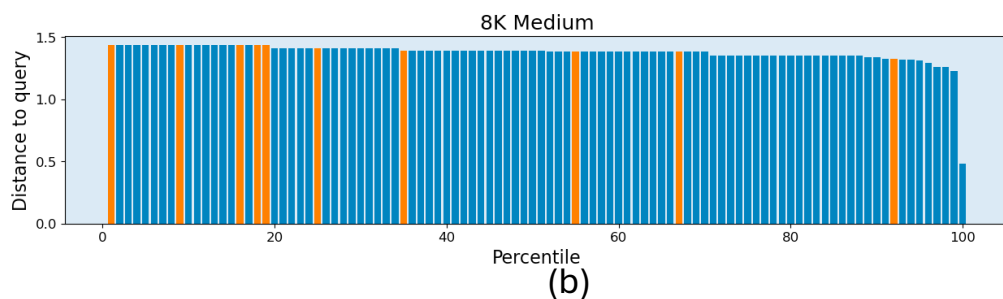
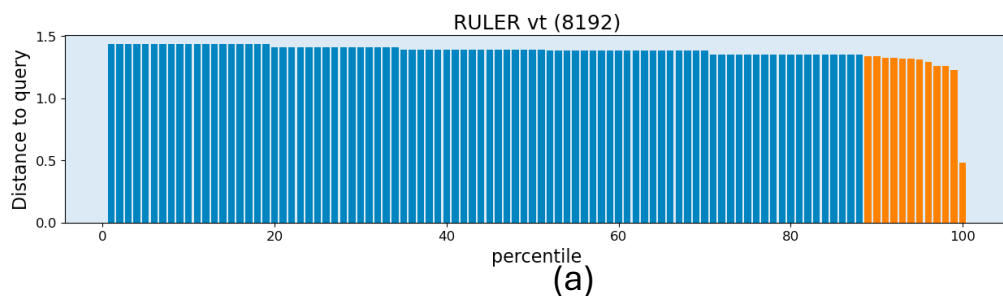


Positioning of GSM-Infinite

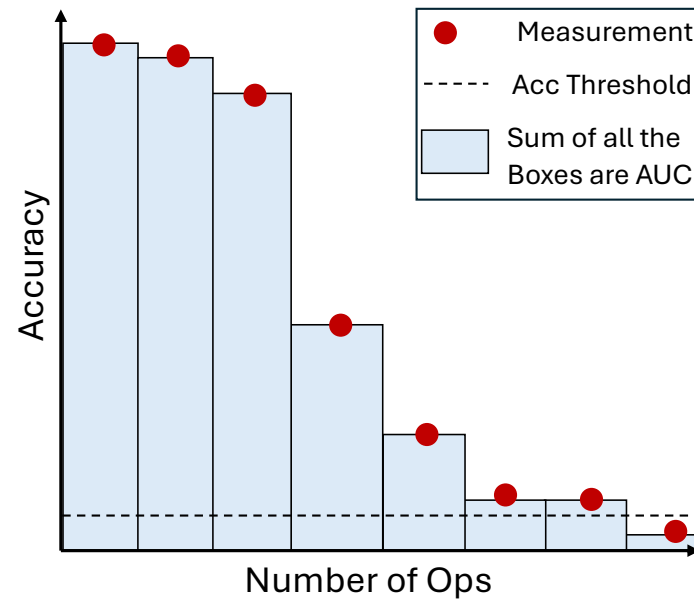
Table 12 Illustrative Problems from Each Subset

Three example problems one for each subtask	
Problem	<ul style="list-style-type: none"> Symbolic (op=5): <code><context>\nassign V705804 = V437110 + 1. assign V986916 = V705804. assign V873548 = 6. assign V684196 = V873548. assign V437110 = V873548.\n </context> \n\nThe context contains relationships between variables. These relationships are independent mathematical equations that are all satisfied simultaneously.\n Using only these relationships, determine which variables (if any) from which values can be derived are equal to 7.\nShow your step-by-step reasoning and calculations, and then conclude your final answer in a sentence. Answer: V705804,V986916.</code> Medium (op=5): Problem: The number of adult owl in Bundle Ranch equals 2 times the number of adult eagle in Bundle Ranch. The number of adult eagle in Hamilton Farm equals the difference between the total number of adult animals in Bundle Ranch and the number of adult eagle in Bundle Ranch. The number of adult owl in Hamilton Farm equals 4 times the number of adult owl in Bundle Ranch. The number of adult eagle in Bundle Ranch equals 3. Question: What is the total number of adult animals in Bundle Ranch? Answer: 9. Hard (op=5): The average number of newborn children per adult blue jay in Bundle Ranch equals 2. The number of adult parrot in Bundle Ranch equals 2. The number of adult blue jay in Bundle Ranch equals 2 times the average number of newborn children per adult blue jay in Bundle Ranch. The number of adult eagle in Bundle Ranch equals 2 times the average number of newborn children per adult blue jay in Bundle Ranch. The number of adult parrot in South Zoo equals 4 times the sum of the average number of newborn children per adult eagle in Hamilton Farm, the number of adult eagle in Hamilton Farm, and the average number of newborn children per adult eagle in Hamilton Farm. The average number of newborn children per adult eagle in Hamilton Farm equals the number of adult eagle in Bundle Ranch. The number of adult eagle in Hamilton Farm equals 3. The average number of newborn children per adult parrot in Bundle Ranch equals the total number of adult animals in Hamilton Farm. The number of adult eagle in South Zoo equals 1. The average number of newborn children per adult parrot in South Zoo equals the average number of newborn children per adult parrot in Bundle Ranch. The average number of newborn children per adult eagle in Bundle Ranch equals 3 plus the average number of newborn children per adult parrot in Bundle Ranch. The average number of newborn children per adult eagle in South Zoo equals the sum of the number of adult blue jay in Bundle Ranch, the average number of newborn children per adult blue jay in Bundle Ranch, the average number of newborn children per adult parrot in Bundle Ranch, and the number of adult parrot in Bundle Ranch. Question: What is the average number of newborn children per adult eagle in Bundle Ranch? Answer: 6.

Semantically Close Noise is Challenging for RAG



Area-Under-Curve (AUC) is the core metric used to compare between LLMs



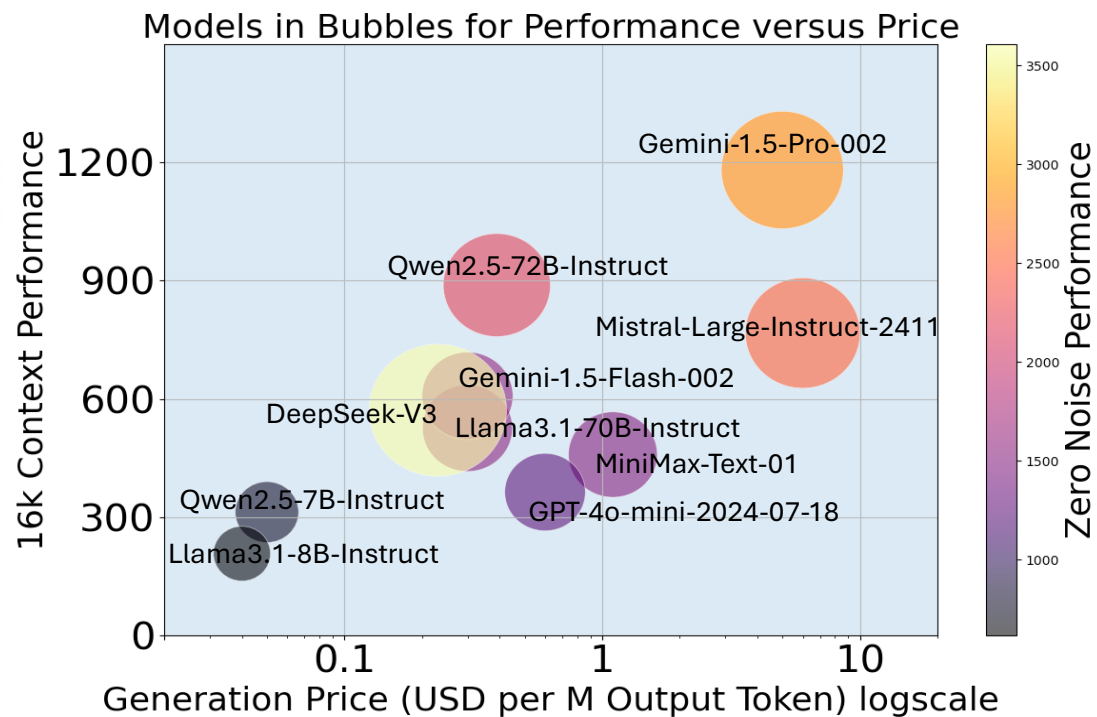
Zero-noise Leaderboard

Models	Three Subtasks			Detailed Statistics on Hard Subtask			Score
	Symbolic	Medium	Hard	1st<50% op	1st<10% op	Avg. Acc op≤30	Avg.↑
DeepSeek-R1	7280.0	9750.85	8573.8	100	>130	0.9427	8534.88
GPT-o3-mini	6690.0	8335.66	5769.96	70	110	0.9423	6931.88
GPT-o1-mini	5060.0	6054.91	3738.43	50	90	0.8397	4951.11
DeepSeek-V3	4310.0	4100.81	2407.86	24	55	0.6669	3606.22
QwQ-32B-preview	3530.0	3205.75	1846.19	21	50	0.5403	2860.65
Gemini-1.5-Pro-002	2547.0	3659.59	2318.28	26	45	0.6924	2841.62
Claude-3.5-Sonnet	2161.0	3281.8	2115.79	26	40	0.6758	2519.53
Mistral-Large	2332.5	2879.92	2310.49	24	50	0.6645	2507.64
Qwen2.5-72B-Instruct	2048.0	2496.81	2016.38	21	40	0.5433	2187.06
GPT-4o	2379.0	2457.37	1451.54	18	30	0.5064	2095.97
Gemini-1.5-Flash-002	1970.0	1478.75	1274.25	13	30	0.4460	1574.33
Llama3.1-70B-Instruct	1769.0	1650.25	1205.25	15	30	0.4314	1541.50
MiniMax-Text-01	1618.5	1712.64	1178.51	14	30	0.4213	1503.22
GPT-4o-mini	1389.0	1406.5	913.89	12	22	0.3094	1236.46
Claude-3.5-Haiku	897.0	1053.16	784.34	10	22	0.2910	911.50
Qwen2.5-7B-Instruct	786.95	886.75	618.5	7	19	0.2257	764.07
Llama3.1-8B-Instruct	462.0	786.5	606.5	6	17	0.2186	618.30
Jamba-1.5-Large	856.0	485.13	466.4	6	26	0.1828	602.51

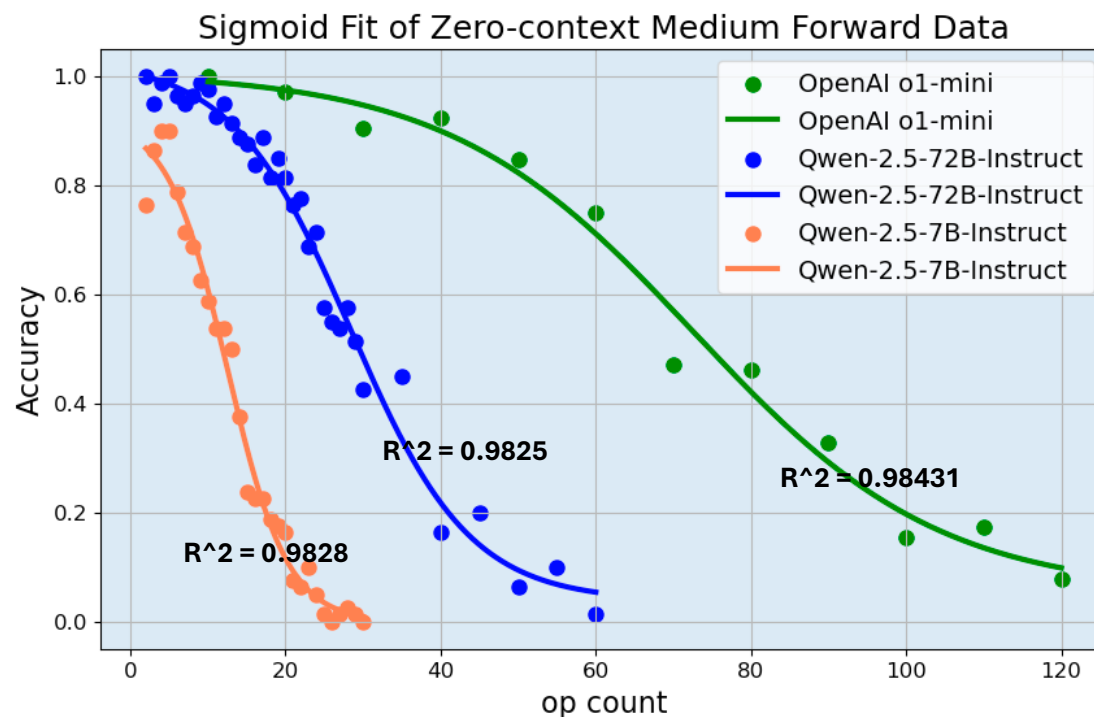
<https://infiniailab-gsm-infinite-leaderboard.hf.space/>

Long-context Leaderboard

Model	8K	16K	32K	Average↑
gemini-1.5-pro-002	1182.43	896.31	812.96	963.9
qwen-2.5-72b-instruct	927.33	681.53	563.65	724.17
mistral-large-2411	914.49	563.73	319.21	599.14
deepseek-v3	935.10	477.02	313.66	575.2
gemini-1.5-flash-002	673.88	476.72	377.38	509.3
llama-3.1-70b-instruct	479.00	394.50	355.5	409.67
minimax-text-01	481.32	359.56	325.95	388.94
gpt-4o-mini	401.00	337.81	275.63	338.15
qwen-2.5-7b-instruct	248.00	211.50	196.17	218.56
llama-3.1-8b-instruct	183.67	149.50	109.45	147.54

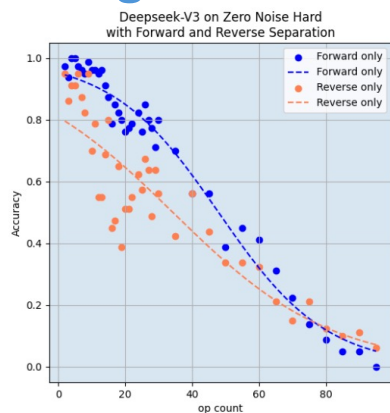


Performance Degradation of LLMs

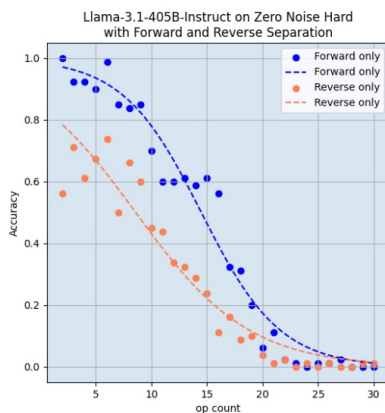


- LLM performance degradation is exponential; can be fitted well using sigmoid function on Realistic subsets;
- We also can see the significant difference between cot model vs. non-cot models and model of different sizes

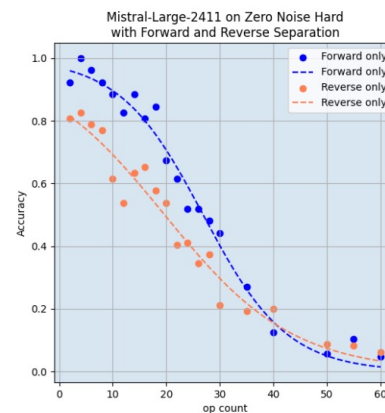
Reverse-thinking Problems have LLM Performance consistently lower than on Forward-thinking Ones



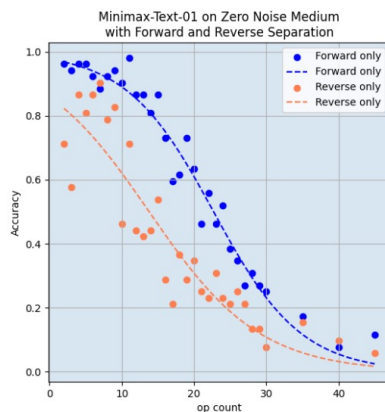
(a) Deepseek-V3



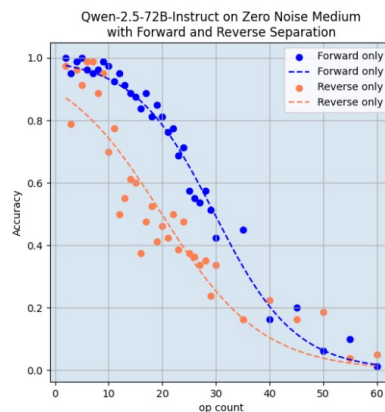
(b) Llama-3.1-405B-Instruct



(c) Mistral-Large-2411

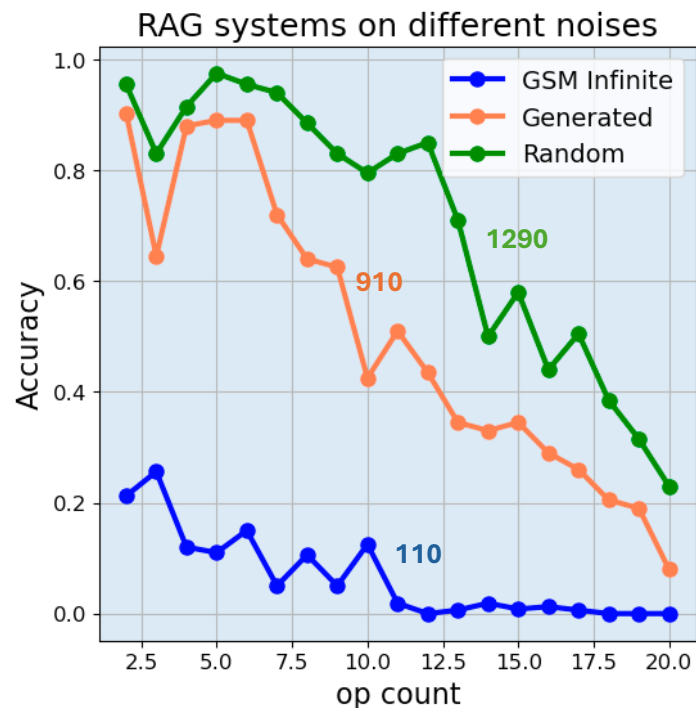
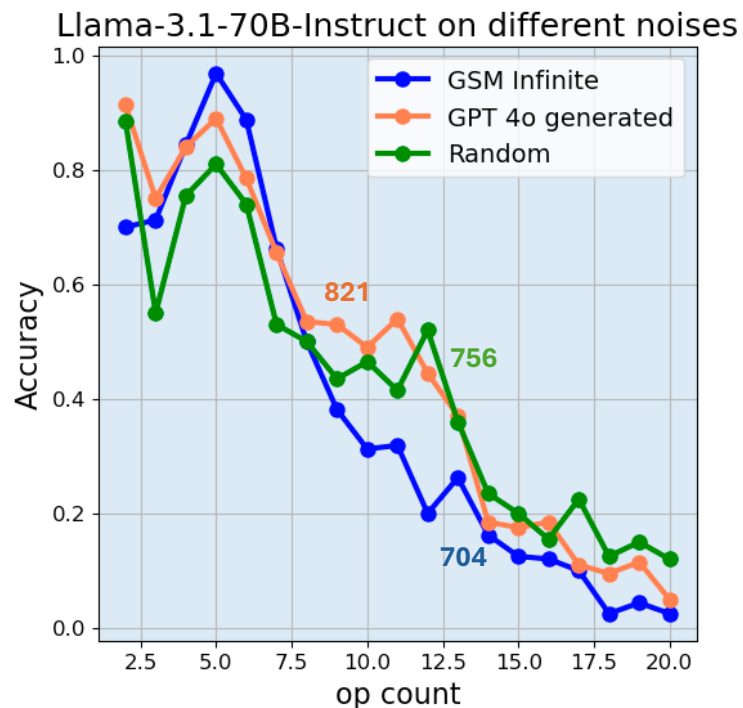


(d) Minimax-Text-01



(e) Qwen-2.5-72B-Instruct

Noise Ablations – Close Noise is Crucial for RAG-Insolvable



- Different noise actually leads to increase the performance of RAG
- Striking difference between other noise and the close noise

LLM Performance drops sharply as context length increases (Diverse Pattern)

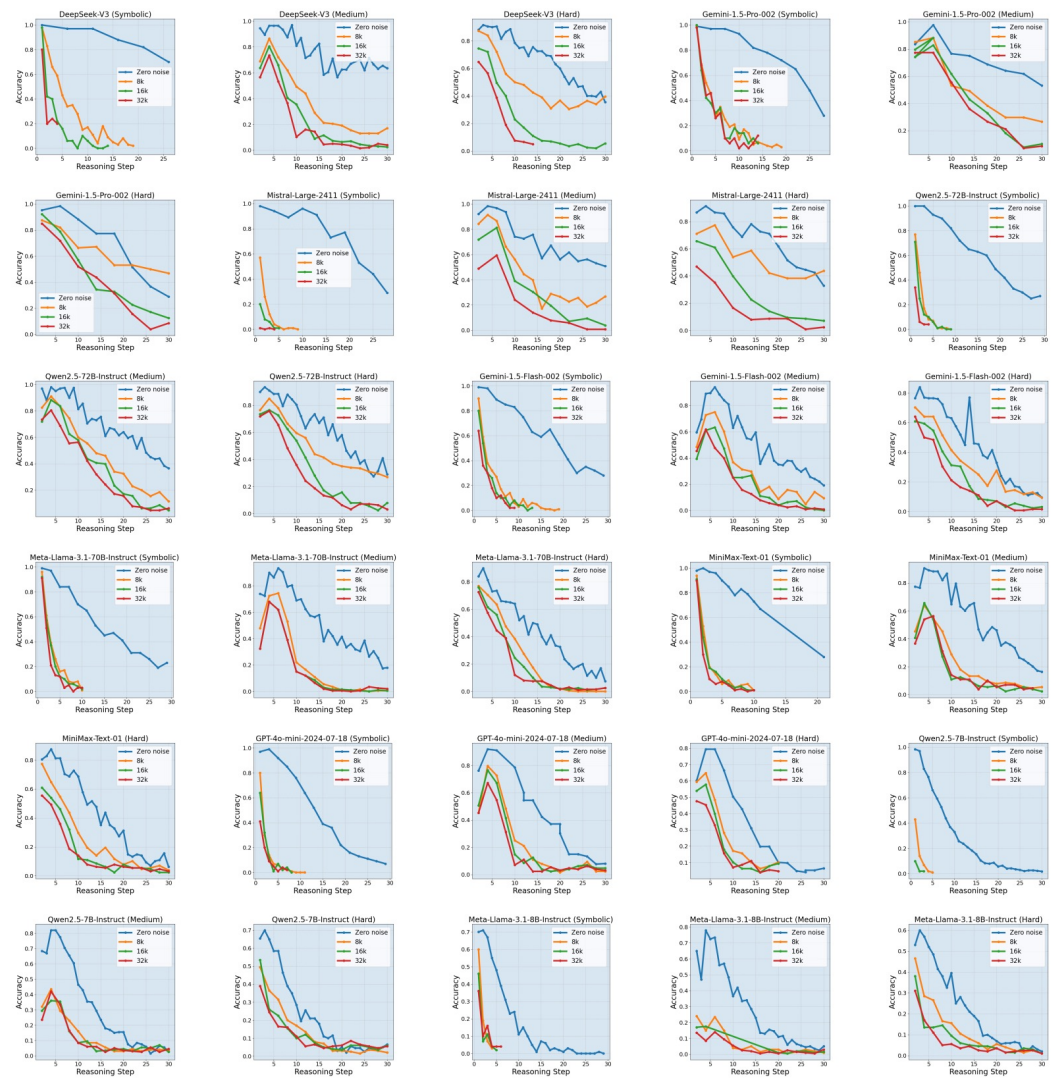
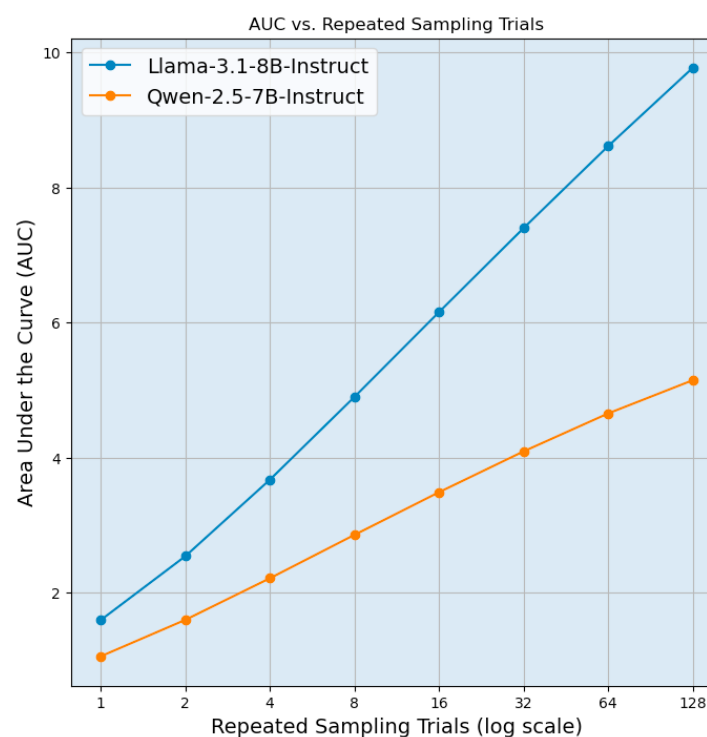
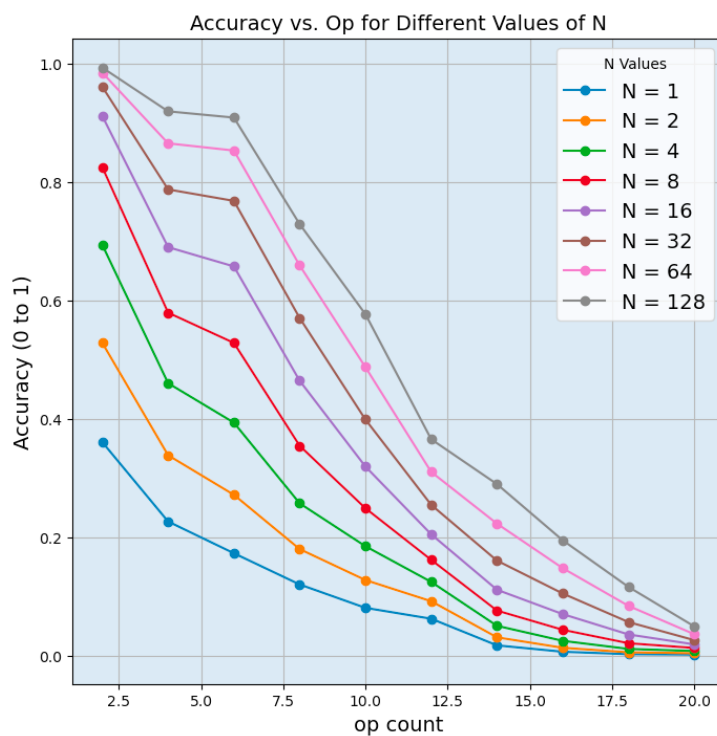


Figure 15 Accuracy decay with context length for different models

Repeated Sampling

Linear Scaling in Performance with Exponential Increase in Compute Inference

”Intelligence is the log of compute.” Satya Nadella (Microsoft CEO)



Future Works

- Where are we going from GSM-Infinite? [Training/finetuning](#)
- Where is the unique advantage of using synthesized dataset?
 - Precise control of training data difficulty ([Curriculum Training](#))
 - Intermediate reward without the need for training a PRM ([Augmenting GRPO](#))
- Synthetic construction can also be applied to study other LLM abilities
 - [Memorization/Hallucination](#) (through pretraining LMs on synthesized bios)
 - Games play (chess, mazes, etc.)
 - Coding

KIMI K1.5: SCALING REINFORCEMENT LEARNING WITH LLMs

TECHNICAL REPORT OF KIMI K1.5

Kimi Team

ABSTRACT

Language model pretraining with next token prediction has proved effective for scaling compute but is limited to the amount of available training data. Scaling reinforcement learning (RL) unlocks a new axis for the continued improvement of artificial intelligence, with the promise that large language models (LLMs) can scale their training data by learning to explore with rewards. However, prior published work has not produced competitive results. In light of this, we report on the training practice of Kimi k1.5, our latest multi-modal LLM trained with RL, including its RL training techniques, multi-modal data recipes, and infrastructure optimization. Long context scaling and improved policy optimization methods are key ingredients of our approach, which establishes a simplistic, effective RL framework without relying on more complex techniques such as Monte Carlo tree search, value functions, and process reward models. Notably, our system achieves state-of-the-art reasoning performance across multiple benchmarks and modalities—e.g., 77.5 on AIME, 96.2 on MATH 500, 94-th percentile on Codeforces, 74.9 on MathVista—matching OpenAI’s o1. Moreover, we present effective long2short methods that use long-CoT techniques to improve short-CoT models, yielding state-of-the-art short-CoT reasoning results—e.g., 60.8 on AIME, 94.6 on MATH500, 47.3 on LiveCodeBench—outperforming existing short-CoT models such as GPT-4o and Claude Sonnet 3.5 by a large margin (up to +550%).

Curriculum Learning training recipe seems to help RL policy convergence.

Thank you all for listening

I am here to take questions