

NeedleBench and the needle in the Haystack

Table of contents

- Motivation and Background
 - What is a context window?
 - Why we care about long context retrieval
- What is a “Needle in a Haystack” ?
 - Greg kamradt [1]
 - Initial findings

Table of contents

- Building on NIH Testing
 - Needle bench
 - Overview
 - Related Work
 - Methodology
 - Experiments
 - Findings
- Summary

Motivation & Background

Background – What is a context window?

- Context windows can be thought of as the working memory or RAM of models with the transformer architecture
- It contains things like the prompt, system prompt, conversation history and any additional documents relevant for the task at hand
- Since this space is limited, we often use RAG to pre-select sections to include in the context window

Motivation – Why long context retrieval?

- RAG: More content can be included
- Chatbots: Conversations can be longer before the model ‘forgets’ previous exchanges
- Eventually: Potentially replace RAG with ultra long context models

Question: How do we evaluate long-Context LLMs?

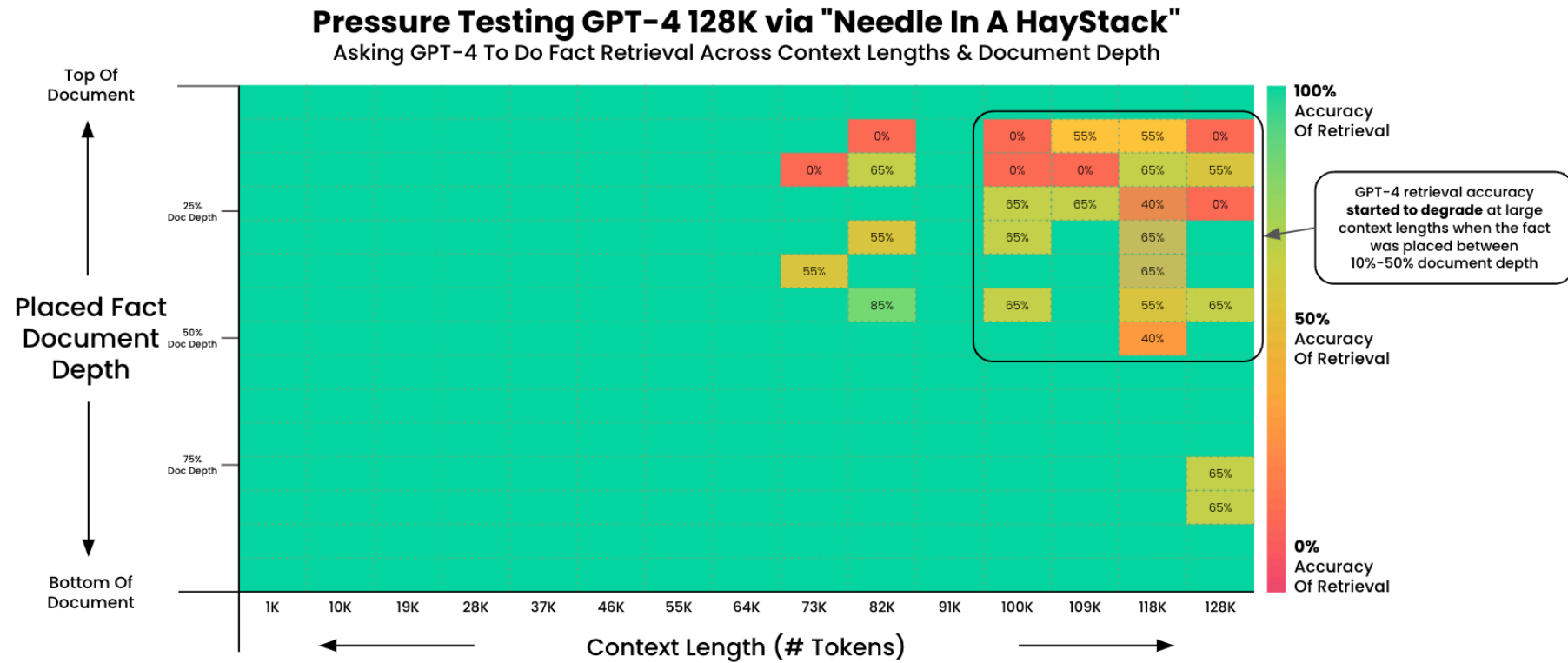
The needle in the haystack

- Developed by Greg Kamradt
- Initially tested on GPT4 with Paul Graham essays
- Aims to test retrieval ability of an LLM at various context lengths with varied needle placement

The needle in a haystack – The Test

1. Place a fact or statement (the needle) inside a chunk of a body of unrelated text (the haystack)
2. Provide the LLM with this text and a prompt asking about the needle
3. Increase the size of the haystack and change the needle position
4. Aggregate the results and evaluate accuracy

The needle in a haystack – Results (GPT-4)



Goal: Test GPT-4 Ability To Retrieve Information From Large Context Windows
A fact was placed within a document. GPT-4 (1106-preview) was then asked to retrieve it. The output was evaluated for accuracy. This test was run at 15 different document depths (top > bottom) and 15 different context lengths (1K > 128K tokens). 2x tests were run for larger contexts for a larger sample size.

The needle in a haystack – Limitations

- ~~• Only aims to retrieve a single needle with limited context size~~
- Needles are irrelevant filler content and require no reasoning ability
- Modern LLMs crush the basic NIHS test (Ruler)

Building on NIHS

NeedleBench: Can LLMs Do Retrieval and Reasoning in Information-Dense Context?

- Published by Mo Li et al. in <https://arxiv.org/abs/2407.11963>
- Aims to provide a more sophisticated Benchmarking framework

Needle bench – Related Work

- RULER (Hsieh et al.): shows that passing retrieval tests does not always mean robust understanding. Models can focus on certain key points to find the needle.
- LongBench(Bai et al.): more realistic tests run the risk of allowing the model to rely on its general knowledge to answer the test questions even without finding the needle

Needle bench – Methodology

- assess retrieval and reasoning performance in bilingual long-context tasks
- context lengths of (e.g., 32k, 128k, and beyond)
- Two types of tests
 - information-sparse, characterized by minimal relevant details embedded within extensive irrelevant text to simulate simpler real-world retrieval tasks;
 - information-dense, implemented as the Ancestral Trace Challenge, where relevant information is continuously distributed throughout the context to simulate more complex real-world reasoning tasks.

NeedleBench – Information-Sparse Tasks

- Single-Needle Retrieval Task (S-RT)
- Multi-Needle Retrieval Task (M-RT)
- Multi-Needle Reasoning Task (M-RS)

needles are synthetic, abstract, and fictional statements or relational facts

Single-Needle Retrieval Task (S-RT)

Single-Needle Retrieval (Needle First - Demonstration with English Version)

Prompt:

This is a test of long-text capability. You need to first read the long document below, and then answer the final question based on the information in the document.

The content of the long document is as follows

Hidden on Emerald Island is the legendary Stardust Shard.

—Paul Graham Essays— —Paul Graham Essays— —Paul Graham Essays— —Paul Graham Essays—
—Paul Graham Essays— —Paul Graham Essays— —Paul Graham Essays— —Paul Graham Essays— —Paul
Graham Essays— —Paul Graham Essays—
—Paul Graham Essays— —Paul Graham Essays— —Paul Graham Essays— —Paul Graham
Essays— —Paul Graham Essays— —Paul Graham Essays— —Paul Graham Essays— —Paul
Graham Essays— —Paul Graham Essays—

Based on the information in the document, now please answer: What legendary item is hidden on Emerald Island? Please answer in the format 'The legendary item hidden on the Emerald Island is _____.'

Multi-Needle Retrieval Task (M-RT)

Multi-Needle Retrieval (Demonstration with five Needles English Version Prompt)

Prompt:

This is a test of long-text capability. You need to first read the long document below, and then answer the final questions one by one based on the information in the document.

The content of the long document is as follows

—Paul Graham Essays— —Paul Graham Essays— —Paul Graham Essays—
Hidden on Forgotten Island is the legendary Stardust Shard.
—Paul Graham Essays— —Paul Graham Essays— —Paul Graham Essays—
Hidden on Mythical Island is the legendary Time-Space Key.
—Paul Graham Essays— —Paul Graham Essays— —Paul Graham Essays—
The ruler of the Alpha Bot star system is Cosmic Ruler Starshine.
—Paul Graham Essays— —Paul Graham Essays— —Paul Graham Essays—
Hidden on Storm Island is the legendary Goodness Heart.
—Paul Graham Essays— —Paul Graham Essays— —Paul Graham Essays—
The ruler of the Orion star system is Guardian of Time Lightspeed.
—Paul Graham Essays— —Paul Graham Essays— —Paul Graham Essays—

Based on the information in the document, now please answer: What legendary item is hidden on Forgotten Island? What legendary item is hidden on Mythical Island? Who is the ruler of the Alpha Bot star system? What legendary item is hidden on Storm Island? Who is the ruler of the Orion star system? Please answer in the format of “The legendary item hidden on the Forgotten Island is _____, The legendary item hidden on the Mythical Island is _____, The ruler of the Alpha Bot star system is _____, The legendary item hidden on the Storm Island is _____, The ruler of the Orion star system is _____.”

Multi-Needle Reasoning Task (M-RS)

Multi-Needle Reasoning (Demonstration with Three Needles English Version)

Prompt:

This is a test of long-text capability. You need to first read the long document below, and then answer the final question based on the information in the document.

The content of the long document is as follows

—*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*—

Jasmine Lane is not only James Hill's father but also James Hill's role model.

—*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*—

Janet Guzman is not only Carolyn Hicks's maternal grandmother but also Carolyn Hicks's role model.

—*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*—

James Hill, as Janet Guzman's paternal grandfather, has a significant impact on Janet Guzman's upbringing.

—*Paul Graham Essays*— —*Paul Graham Essays*— —*Paul Graham Essays*—

Based on the information in the document, now please answer: Given the context described above, who is the eldest relative that 'Carolyn Hicks' can trace back to in the context?

Multi-Needle Reasoning Task (M-RS)

For example:

Example 1: If James Hill's father is Jasmine Lane, and no further information about familial relationships is provided in the text, then the oldest relative James Hill can trace back to in the provided text is `\boxed{Jasmine Lane}`.

Example 2: If Andrew Williams's grandmother is Dan Newton, and Dan Newton's father is James Hill, and no further information about familial relationships is provided in the text, then the oldest relative Andrew Williams can trace back to in the provided text is `\boxed{James Hill}`.

Example 3: If Jeff White's father is Kevin Le, Dan Newton's grandmother is Jeff White, and Jeff White's father is Kevin Le, and Shelley Mills is Dan Newton's great-granddaughter, and no further information about familial relationships is provided in the text, then the oldest relative Shelley Mills can trace back to in the provided text is `\boxed{Kevin Le}`.

Notes:

1. You do not need to worry about the gender consistency of names in this test. For example, a name that is typically considered feminine can still be the father of another person. Our primary focus is on who is older.
2. Ignore surname inheritance issues. For instance, Andrew Williams could still be the biological father of Christopher Baker. We only care about who is older and do not need to consider whether a child should inherit the father's or mother's surname.
3. At the end of your response, remember to put your final answer within `\boxed{}`. For example: "So the oldest relative 'Carolyn Hicks' can trace back to in the provided text is `\boxed{(your answer here)}`."

Information-Sparse Tasks: Evaluation

NeedleBench – Information-Dense Tasks

Ancestral Trace Challenge (ATC):

-

Ancestral Trace Challenge (3-needle version example with DeepSeek R1 response)

Prompt:

Here is a test for multi-step reasoning ability called the Ancestral Trace Challenge. In this test, we will simulate different people's familial relationships, and your task is to continuously reason through them until you identify the eldest ancestor.

Now, the scrambled family relationships are provided below:

Wyatt James is the child of Maria Watson. Emily Barry, as Maria Watson's paternal grandfather, has a significant impact on Maria Watson's upbringing. Joseph Taylor is not only Emily Barry's maternal grandfather but also Emily Barry's role model.

Given the scrambled family relationships described above, who is the eldest relative that 'Wyatt James' can trace back to in the context?

For example:

Example 1: If James Hill's father is Jasmine Lane, and no further information about familial relationships is provided in the text, then the oldest relative James Hill can trace back to in the provided text is `\boxed{Jasmine Lane}`.

Example 2: If Andrew Williams's grandmother is Dan Newton, and Dan Newton's father is James Hill, and no further information about familial relationships is provided in the text, then the oldest relative Andrew Williams can trace back to in the provided text is `\boxed{James Hill}`.

Example 3: If Jeff White's father is Kevin Le, Dan Newton's grandmother is Jeff White, and Jeff White's father is Kevin Le, and Shelley Mills is Dan Newton's great-granddaughter, and no further information about familial relationships is provided in the text, then the oldest relative Shelley Mills can trace back to in the provided text is `\boxed{Kevin Le}`.

Information-Dense Tasks

Diversity in ATC:

-

Notes:

1. You do not need to worry about the gender consistency of names in this test. For example, a name that is typically considered feminine can still be the father of another person. Our primary focus is on who is older.
2. Ignore surname inheritance issues. For instance, Andrew Williams could still be the biological father of Christopher Baker. We only care about who is older and do not need to consider whether a child should inherit the father's or mother's surname.
3. At the end of your response, remember to put your final answer within `\boxed{}`. For example: "So the oldest relative xxx(the person you need to find the eldest relative for) can trace back to in the provided text is `\boxed{}`."

DeepSeek R1 Response (Correct Example):

Wyatt James is the child of Maria Watson, making Maria one generation above Wyatt. Maria's paternal grandfather is Emily Barry, placing Emily two generations above Maria and three generations above Wyatt. Emily Barry's maternal grandfather is Joseph Taylor, which adds two more generations above Emily (Emily's mother and then Joseph). Therefore, Joseph Taylor is five generations above Wyatt. Among the mentioned relatives, Joseph Taylor is the eldest ancestor.

So the oldest relative Wyatt James can trace back to in the provided text is `\boxed{Joseph Taylor}`.

Information-Dense Tasks: Evaluation

Needle bench – Experiments

- Evaluation of mainstream open source LLMs
- Information-sparse tasks at context lengths of 32K and 128K
- ATC task were also performed on close-source models such as GPT-4.1 and DeepSeek R1

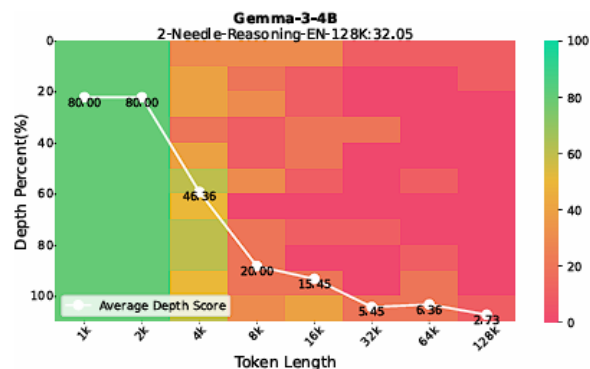
Performance of NeedleBench Information-Sparse Tasks

- Impact of Model Architecture and Technical Advances
- Effect of Model Scale on Multi-Needle Reasoning Performance
- Effect of Needle Count on Multi-Needle Reasoning Performance
- Impact of Language

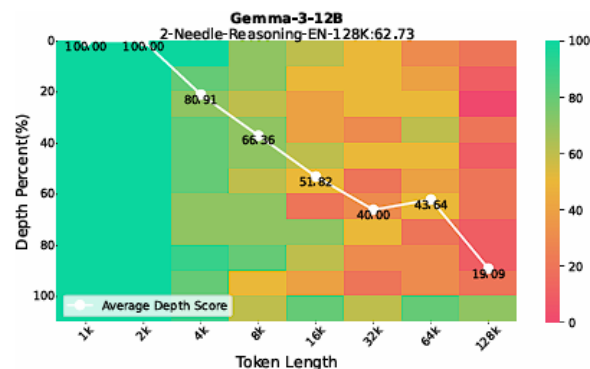
Table 1: **Main Results of NeedleBench 32K**. Overall denotes the mean score across all tasks. **Bold** denotes the best score among all models, and underline denotes the best score under the same model scale. The same notation applies in the following tables.

Model	Single-Retrieval			Multi-Retrieval			Multi-Reasoning			Overall
	Chinese	English	Overall	Chinese	English	Overall	Chinese	English	Overall	
Models with Fewer Than 7B Parameters										
Qwen-2.5-1.5B	96.67	94.44	95.56	95.39	97.29	96.34	0.00	15.63	7.82	66.57
Qwen-1.5-4B	95.66	99.60	97.63	95.76	97.01	96.38	2.68	7.05	4.86	66.29
ChatGLM3-6B-32K	93.64	98.89	96.26	90.83	94.38	92.61	0.18	9.07	4.62	64.50
Qwen-1.5-1.8B	78.99	71.11	75.05	54.26	52.93	53.60	0.00	0.00	0.00	42.88
Models with 7-20B Parameters										
Qwen-2.5-14B	99.19	98.79	98.99	99.07	99.23	99.15	29.65	17.90	23.78	73.97
Qwen-2.5-7B	100.00	99.80	99.90	97.70	99.31	98.51	12.65	18.64	15.64	71.35
Qwen-1.5-14B	99.60	99.49	99.55	92.57	99.15	95.86	0.58	10.20	5.39	66.93
Mistral-7B-Instruct-v0.2	92.73	96.36	94.55	87.23	96.97	92.10	11.57	14.27	12.92	66.52
Zephyr-7B-Beta	35.35	36.77	36.06	18.14	27.60	22.87	1.87	7.45	4.66	21.20
Models Larger Than 20B Parameters										
Qwen-2.5-72B	100.00	100.00	100.00	98.71	99.96	99.33	39.80	52.07	45.93	81.76
Qwen-2.5-32B	100.00	100.00	100.00	98.71	95.72	97.21	33.31	38.96	36.14	77.78
Qwen-1.5-32B	99.60	100.00	99.80	98.02	98.95	98.48	11.67	14.85	13.26	70.51
Mixtral-8x7B-Instruct-v0.1	95.76	99.60	97.68	94.63	99.43	97.03	5.93	15.88	10.91	68.54
Qwen-1.5-72B	97.37	89.60	93.48	93.49	92.24	92.87	9.75	7.35	8.55	64.97

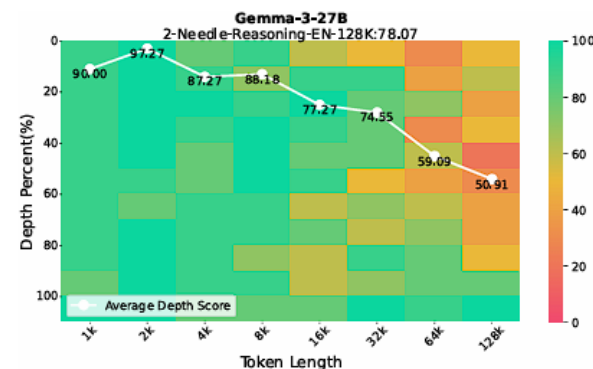
Performance – Generation and Needle Count



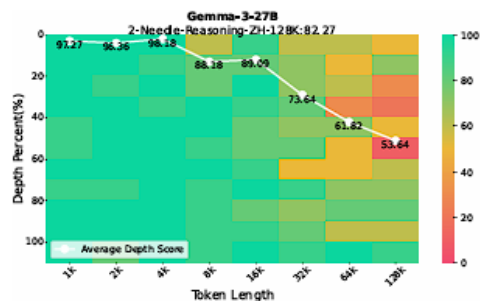
(a) Gemma-3-4B



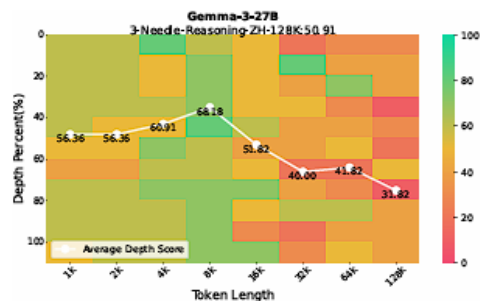
(b) Gemma-3-12B



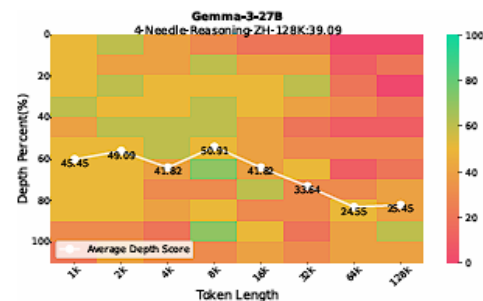
(c) Gemma-3-27B



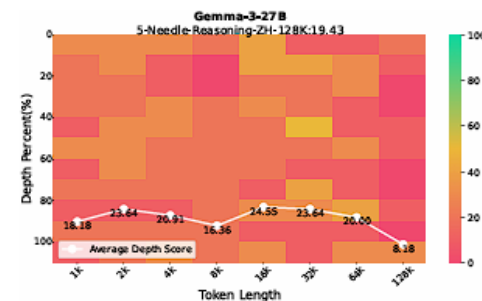
(a) 2-Needle



(b) 3-Needle

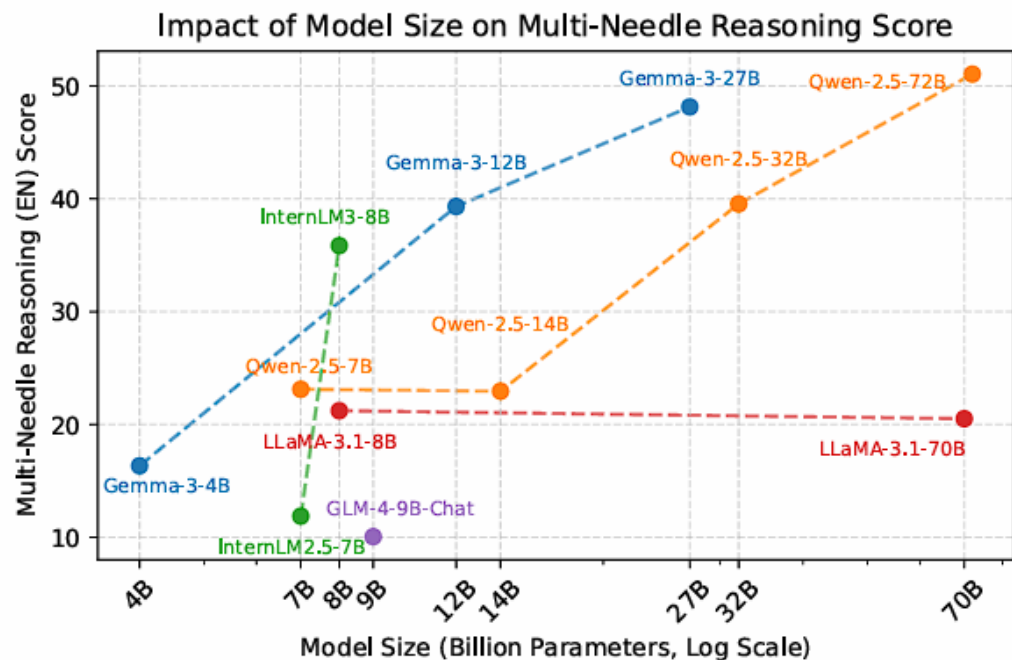


(c) 4-Needle

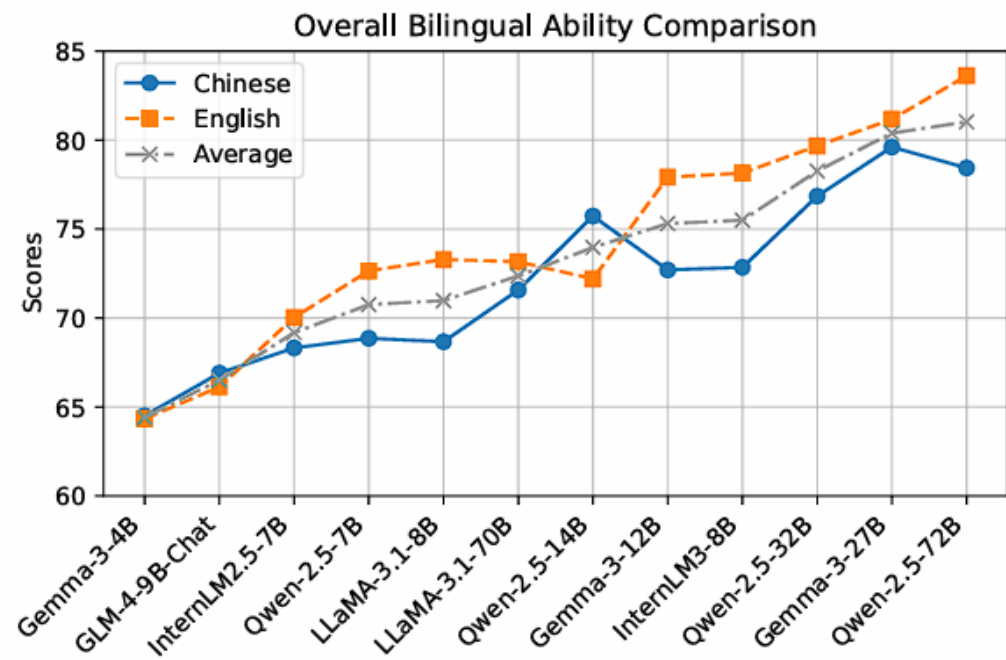


(d) 5-Needle

Performance – Size and Language



(a) Model size vs. long-context reasoning ability.



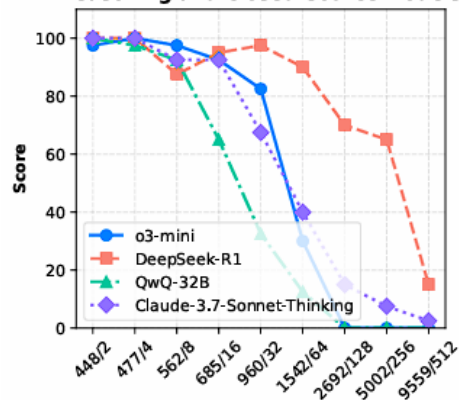
(b) English vs. Chinese performance.

Performance of NeedleBench Information-Dense Tasks

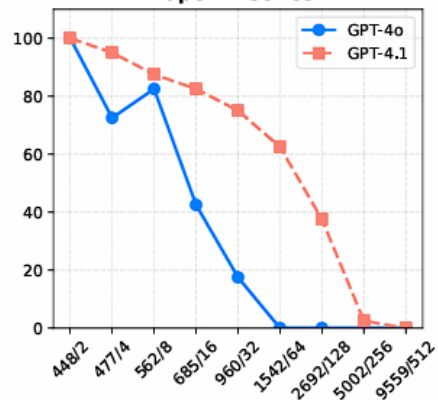
Table 3: ATC task results for all evaluated models. Models with stronger reasoning abilities generally achieve higher scores, with DeepSeek-R1 achieving the best overall performance (total score 44.01). In contrast, models with fewer parameters often achieve only single-digit scores.

Model	Needle Count									Evaluation Metric		
	2	4	8	16	32	64	128	256	512			
	Context Length (tokens)									Weighted Score	ENL	
	0.4K	0.5K	0.6K	0.7K	1.0K	1.5K	2.7K	5K	9.6K	≤ 2K	All	ENL-50
<i>Closed-Source and Reasoning Models</i>												
Claude-3.7-Sonnet-Thinking	100.0	100.0	92.5	92.5	67.5	40.0	15.0	7.5	2.5	59.84	12.39	32
DeepSeek-R1	100.0	100.0	87.5	<u>95.0</u>	97.5	90.0	<u>70.0</u>	<u>65.0</u>	15.0	92.86	44.01	256
GPT-4o	100.0	72.5	82.5	42.5	17.5	0.0	0.0	0.0	0.0	18.97	2.34	8
GPT-4.1	100.0	95.0	87.5	82.5	75.0	62.5	37.5	2.5	0.0	71.43	14.13	64
o3-mini	97.5	100.0	97.5	92.5	82.5	30.0	0.0	0.0	0.0	58.85	7.26	32
QwQ-32B	100.0	97.5	92.5	65.0	32.5	12.5	0.0	0.0	0.0	33.41	4.12	16
OREAL-32B	92.5	55.0	45.0	20.0	22.5	7.5	5.0	0.0	0.0	18.13	2.86	4
DeepSeek-R1-Qwen-32B	95.0	75.0	47.5	35.0	12.5	5.0	0.0	0.0	0.0	17.06	2.10	4
DeepSeek-R1-Qwen-14B	100.0	62.5	37.5	22.5	5.0	0.0	0.0	0.0	0.0	10.08	1.24	4
DeepSeek-R1-Qwen-7B	77.5	25.0	2.5	0.0	0.0	0.0	0.0	0.0	0.0	2.18	0.27	2
<i>Models with 20B or More Parameters</i>												
Qwen1.5-72B	70.0	25.0	10.0	0.0	0.0	0.0	0.0	0.0	0.0	2.54	0.31	2
Qwen2.5-72B	92.5	62.5	45.0	10.0	0.0	0.0	2.5	0.0	0.0	7.58	1.25	4
Qwen1.5-32B	75.0	20.0	7.5	7.5	5.0	0.0	0.0	0.0	0.0	4.52	0.56	2
Qwen2.5-32B	<u>97.5</u>	62.5	27.5	17.5	5.0	<u>2.5</u>	0.0	0.0	0.0	10.04	1.24	4
Gemma-3-27B	82.5	<u>70.0</u>	<u>67.5</u>	<u>47.5</u>	<u>30.0</u>	<u>2.5</u>	<u>5.0</u>	0.0	0.0	<u>22.74</u>	<u>3.43</u>	<u>8</u>
<i>Models with 7-20B Parameters</i>												
Qwen1.5-14B	52.5	17.5	10.0	2.5	2.5	0.0	0.0	0.0	0.0	2.98	0.37	2
Qwen2.5-14B	<u>72.5</u>	47.5	25.0	7.5	5.0	0.0	0.0	0.0	0.0	6.47	0.80	2
Gemma-3-12B	<u>72.5</u>	<u>55.0</u>	<u>45.0</u>	17.5	<u>10.0</u>	<u>2.5</u>	0.0	0.0	0.0	<u>11.79</u>	1.45	<u>4</u>
Mixtral-8x7B	50.0	12.5	15.0	7.5	0.0	0.0	0.0	0.0	0.0	3.10	0.38	2
GLM-4-9B	52.5	35.0	10.0	7.5	2.5	0.0	0.0	0.0	0.0	4.17	0.51	2
InternLM3-8B	55.0	32.5	27.5	<u>20.0</u>	2.5	0.0	0.0	<u>5.0</u>	0.0	6.83	<u>2.09</u>	2
<i>Models with Fewer Than 7B Parameters</i>												
Qwen1.5-1.8B	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.08	0.01	0
Qwen1.5-4B	35.0	10.0	7.5	0.0	0.0	0.0	0.0	0.0	0.0	1.35	0.17	0
Qwen2.5-1.5B	37.5	25.0	2.5	0.0	0.0	0.0	0.0	0.0	0.0	1.55	0.19	0
Qwen2.5-7B	<u>75.0</u>	<u>37.5</u>	7.5	5.0	<u>5.0</u>	0.0	0.0	0.0	0.0	4.76	0.59	<u>2</u>
Mistral-7B	40.0	17.5	2.5	2.5	0.0	0.0	0.0	0.0	0.0	1.67	0.21	0
Gemma-3-4B	57.5	27.5	<u>15.0</u>	<u>17.5</u>	2.5	0.0	0.0	0.0	0.0	<u>5.60</u>	0.69	<u>2</u>
ChatGLM3.6B-32K	27.5	7.5	10.0	5.0	2.5	0.0	2.5	0.0	2.5	2.58	1.88	0

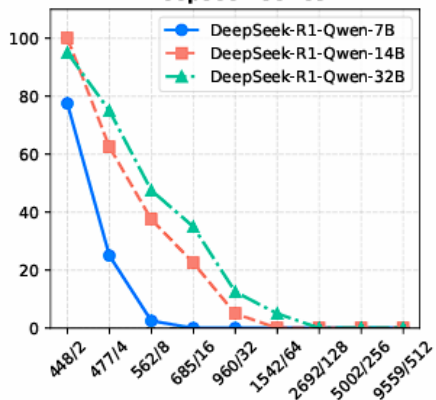
Reasoning and Closed-Source Models



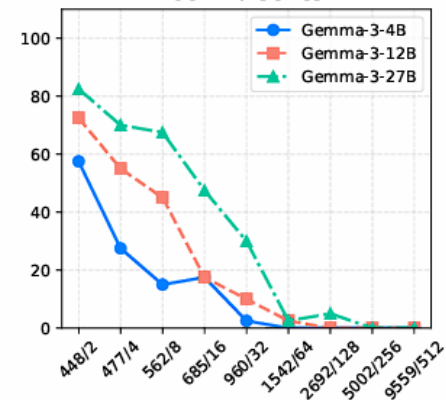
OpenAI Series



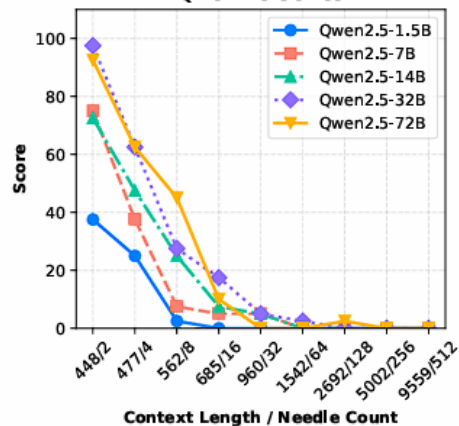
DeepSeek Series



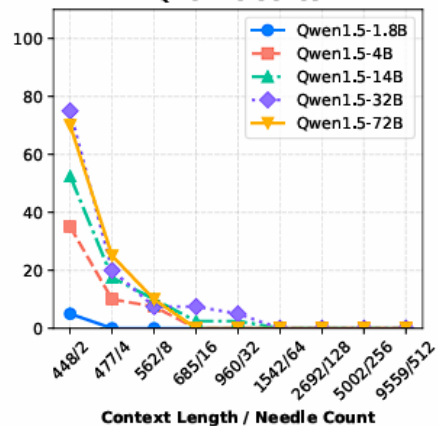
Gemma Series



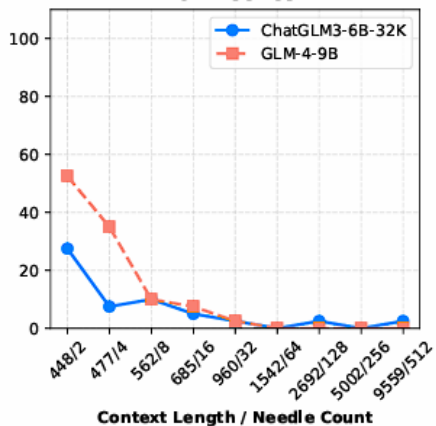
Qwen2.5 Series



Qwen1.5 Series



GLM Series



InternLM Series

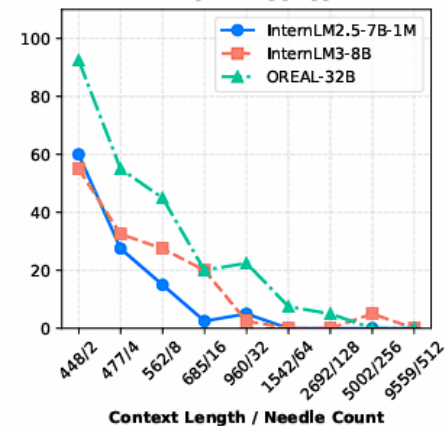


Table 4: Analysis of Common Error Types in ATC Task: Under-thinking is the most prevalent error, especially in strong models like DeepSeek R1 and o3-mini. Partial understanding errors indicate models only grasp part of key information. Smaller models frequently exhibit instruction following and repetitive output errors.

Error Type (%)	Explanation	Representative Models (non-exhaustive)
Under-thinking (60.2%)	The model prematurely halts reasoning, asserting that no further inference can be made despite the presence of remaining clues.	Reasoning models (e.g., DeepSeek R1, GPT-4.1, o3-mini, Claude-3-7-Sonnet, InternLM3-8B-Instruct, Qwen2.5-1.5B-Instruct)
Partial Understanding Error (12.9%)	Model only identifies part of the relationships mentioned.	Gemma3 27B, InternLM3-8B-Instruct, Qwen2.5-7B-Instruct
Instruction Following Error (7.5%)	Model fails to follow the required output format.	Qwen1.5-1.8B-Chat, Qwen2.5-1.5B-Instruct
Repetitive Output (7.5%)	Model either repeats reasoning steps or outputs meaningless text.	Deepseek-R1-Distill-Qwen-7B, Qwen2.5-1.5B-Instruct
Hallucination Error (7.5%)	Model introduces information that is not present in the original text.	DeepseekR1, Qwen1.5-1.8B-Chat, Deepseek-R1-Distill-Qwen-7B,
Other Errors (4.5%)	Miscellaneous errors not covered by the above categories.	Various models

Summary

[gkamradt/LLMTest_NeedleInAHaystack: Doing simple retrieval from LLM models at various context lengths to measure accuracy](#)