Long Context Benchmark

26.01.24

Agent Knowledge Benchmark

Parametric Knowledge

=> 1+1=2, 사과의 색깔은?, 한국의 수도

In-context Knowledge

=> 프롬프트에 주어진 문서를 보고 근거를 찾아 답변

External Memory / tools

=> RAG/Search/DB/Code 등 활용 능력

Why long context evaluation?

In-context에서 정보를 추출하는 방식

- Retrieval: 필요한 정보를 찾기

- Reasoning: 찾은 정보를 조합/추론

- Effective Length: 무관한 정보가 어떤 영향을 미치는지 확인

Long Context Benchmark

- Needle-in-a-Haystack (NIAH - https://github.com/gkamradt/LLMTest_NeedleInAHaystack)
- 현실 시나리오: LongBench v2: Towards Deeper Understanding and Reasoning on Realistic Long-context Multitasks (2025, ACL)
- 초장문 테스트: InfiniteBench: Extending Long Context Evaluation Beyond 100K Tokens (2024, ACL)
- 길이에 따른 성능 붕괴 RULER: What's the Real Context Size of Your Long-Context Language Models? (2024, COLM)
- 멀티모달 Long Context: Multimodal Needle in a Haystack: Benchmarking Long-Context Capability of Multimodal Large Language Models (2025, NAACL)

==========

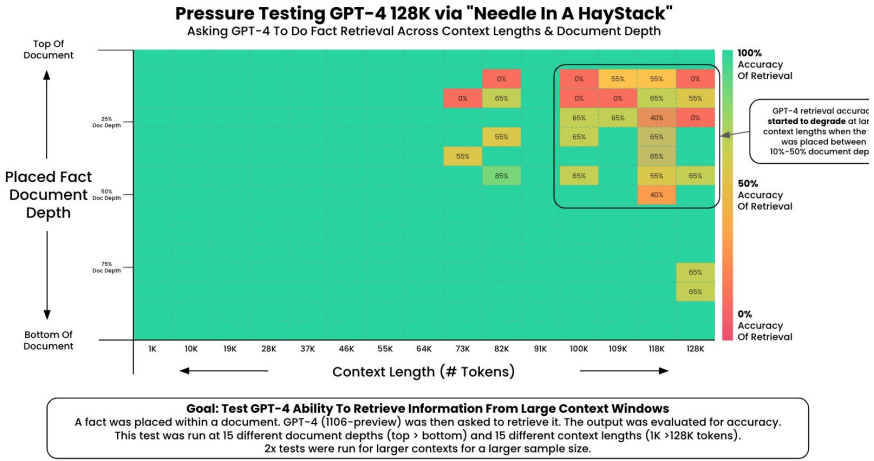
\+ LongBioBench: 현실적인 long bench

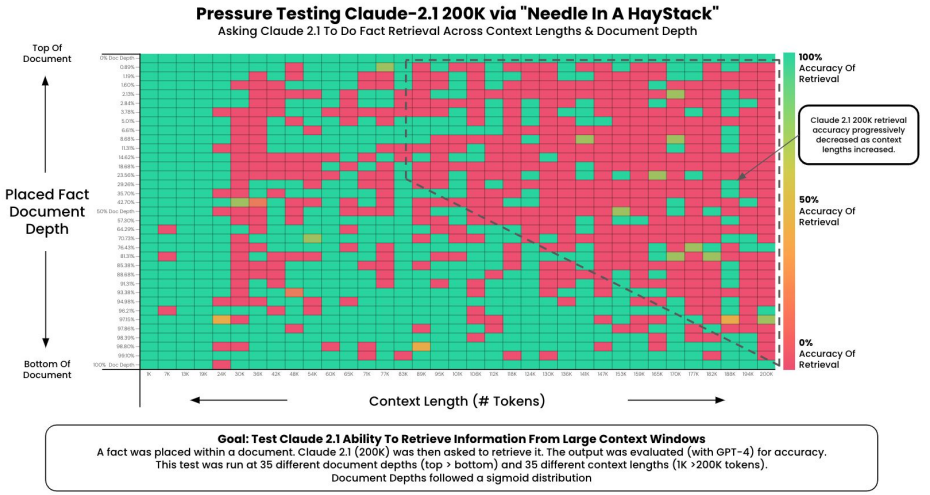
https://github.com/gkamradt/LLMTest_NeedleInAHaystack

Needle in a Haystack (NIAH) - Pressure Testing LLMs

1. 데이터 구성
- Retrieval Only
- Context Length, Document Depth

Needle: The access code is 7319
Haystack: News, Wiki etc
Q: In the document, what is the access code? Answer with digits only.





=> 길수록 성능 떨어짐, 두 모델 특정 위치에서 못 찾는 경향이 있음
GPT-4 128k, Claude 2.1 200K => Claimed context != Effective Context length

https://aclanthology.org/2025.acl-long.183.pdf

LongBench v2: Towards Deeper Understanding and Reasoning on Realistic Long-context Multitasks

1. 데이터 구성
- 503개의 객관식 문제로 구성 (4지선다)
- 컨텍스트 길이: 8k ~ 2M words
- retrieve + multi-hop/aggregation + reasoning

II.3. Multi-Document QA (Financial)

Task Description: Ask questions based on financial documents, requiring at least 2 documents. Questions must require information from each document to be answered, and there should be no irrelevant documents.
Example Questions: 1. How has the R&D investment of the enterprises changed in the past ten years?

III.2. Long In-context Learning (New language translation)

Task Description: Translation tasks involving the rare languages Zhuang (vocabulary book and translation corpus from Zhang et al. (2024a)) and Kalamang (vocabulary book and translation corpus from Tanzer et al. (2024)), requiring reading a vocabulary book to complete.
Example Questions: 1. Translate the following kalamang into English: Wa me kariak kaia kon untuk emumur kalo tumun amkeiret mu wara nanet.

IV.1. Long-dialogue History Understanding (Agent history QA)

Task Description: Based on the agent dialogue history as context, ask questions about the content of the history. Specifically, we provide annotators with LLMs' dialogue history on playing games, which is derived from the GAMA-Bench (tse Huang et al., 2025). This dataset includes eight classical multi-agent games categorized into three groups: Cooperative Games, Betraying Games, and Sequential Games. In our task, we use them as context and annotate questions for the agent interaction history.
Example Questions: 1. Which player is the most selfish one in the fourth round of the game?

Dataset	Source	#data	Length	Expert Acc	Expert Time*
I. Single-Document QA		175	51k	55%	8.9 min
Academic	Paper, textbook	44	14k	50%	7.3 min
Literary	Novel	30	72k	47%	8.5 min
Legal	Legal doc	19	15k	53%	13.1 min
Financial	Financial report	22	49k	59%	9.0 min
Governmental	Government report	18	20k	50%	9.5 min
Detective	Detective novel	22	70k	64%	9.3 min
Event ordering	Novel	20	96k	75%	9.4 min
II. Multi-Document QA		125	34k	36%	6.1 min
Academic	Papers, textbooks	50	27k	22%	6.1 min
Legal	Legal docs	14	28k	64%	8.8 min
Financial	Financial reports	15	129k	40%	7.0 min
Governmental	Government reports	23	89k	22%	6.0 min
Multi-news	News	23	15k	61%	5.3 min
III. Long In-context Learning		81	71k	63%	8.3 min
User guide QA	Electronic device, software, instrument	40	61k	63%	9.9 min
New language translation	Vocabulary book (*Kalamang*, *Zhuang*)	20	132k	75%	5.4 min
Many-shot learning	Multi-class classification task	21	71k	52%	8.0 min
IV. Long-dialogue History Understanding		39	25k	79%	8.2 min
Agent history QA	LLM agents conversation	20	13k	70%	8.3 min
Dialogue history QA	User-LLM conversation	19	77k	89%	6.5 min
V. Code Repository Understanding		50	167k	44%	6.4 min
Code repo QA	Code repository	50	167k	44%	6.4 min
VI. Long Structured Data Understanding		33	49k	73%	6.4 min
Table QA	Table	18	42k	61%	7.4 min
Knowledge graph reasoning	KG subgraph	15	52k	87%	6.2 min

Table 1: Tasks and data statistics in LongBench v2. 'Source' denotes the origin of the context. 'Length' is the *median* of the number of words. 'Expert Acc' and 'Expert Time' refer to the average accuracy and the *median* time spent on answering the question by human experts. *: We allow human experts to respond with "I don't know the answer" if it takes them more than 15 minutes. As a result, most expert times are under 15 minutes, but this doesn't necessarily mean that the questions are fully answered within such a time.

https://aclanthology.org/2025.acl-long.183.pdf

LongBench v2: Towards Deeper Understanding and Reasoning on Realistic Long-context Multitasks

2. 평가
- Accuracy (Multiple Choice Question - 4지선다)

Model	Avg	I	II	III	IV	V	VI
GLM-4-9B-Chat	30.2	30.9	27.2	33.3	38.5	28.0	24.2
w/o context	26.2	30.9	21.6	18.5	30.8	34.0	21.2
Llama-3.1-8B-Inst.	30.0	34.9	30.4	23.5	17.9	32.0	30.3
w/o context	25.8	31.4	26.4	24.7	23.1	22.0	6.1
Qwen2.5-72B-Inst.	39.4	40.6	35.2	42.0	25.6	50.0	42.4
w/o context	30.0	33.7	31.2	25.9	28.2	34.0	12.1
GLM-4-Plus	44.3	41.7	42.4	46.9	51.3	46.0	48.5
w/o context	27.6	33.7	27.2	25.9	10.3	38.0	6.1
GPT-4o	50.1	48.6	44.0	58.0	46.2	56.0	51.5
w/o context	33.1	40.0	25.6	32.1	38.5	34.0	18.2

Table 3: Scores (%) across 6 tasks: *I. Single-Doc QA*, *II. Multi-Doc QA*, *III. Long ICL*, *IV. Dialogue History*, *V. Code Repo*, and *VI. Structured Data*.

=> parametric knowledge 확인을 위해 context ablation 진행

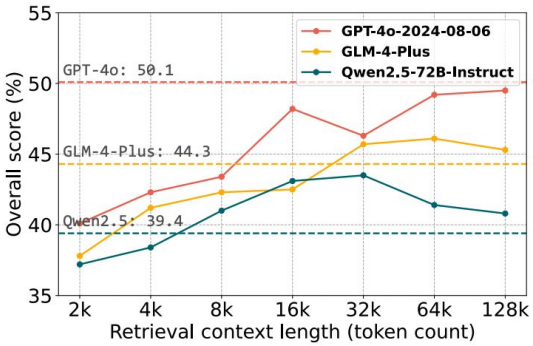


Figure 4: RAG performance across different context lengths, varied by including the top 4, 8, 16, 32, 64, 128, and 256 chunks of 512 tokens. The horizontal line show the overall score of each model without RAG at a full context length of 128k tokens.

=> RAG로 context를 줄일 때 효과

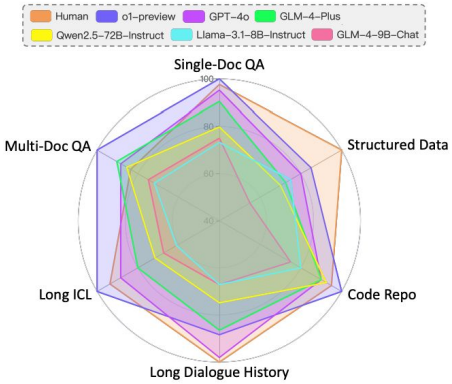


Figure 3: Average scores across tasks, normalized by the highest score on each task. All scores are evaluated in the zero-shot + CoT setting, except for o1-preview, since it latently performs CoT under zero-shot prompting.

=> Problem마다의 성능 비교

https://arxiv.org/html/2402.13718v3

InfiniteBench: Extending Long Context Evaluation Beyond 100K Tokens

1. 데이터 구성
- 200k 평균의 장문 벤치

Task	Annotation	# Ex.	Avg Len
Ret.PassKey	Auto	590	122.4K/2
Ret.Number	Auto	590	122.4K/4
Ret.KV	Auto	500	121.1K/22.7
En.Sum	Human	103	103.5K/1.1K
En.QA	Human	351	192.6k/4.8
En.MC	Human	229	184.4K/5.3
Zh.QA	Human	189	2068.6K/6.3
En.Dia	Auto	200	103.6K/3.4
Code.Debug	Human	394	114.7K/4.8
Code.Run	Auto	400	75.2K/1.3
Math.Calc	Auto	50	43.9K/43.9K
Math.Find	Auto	350	87.9K/1.3

Table 2: Data statistics. The columns indicate whether the annotation was auto-generated or done by humans, the number of examples, and the average length (input/output) in tokens.

<= Retrieval 노이즈 안에서 정보 정확하게 찾기

<= Reasoning: Summarization, QA (추론), Multiple Choice, Dialogue (발화자 맞추기)

<= Reasoning: Code 버그 함수 찾기 (Choice), Return Value prediction
<= Reasoning: 수식 계산, 배열에서 정보 찾기 (minmax)

https://arxiv.org/html/2402.13718v3

InfiniteBench: Extending Long Context Evaluation Beyond 100K Tokens

*rouge: recall-oriented understudy for gisting evaluation
=> rouge-1는 bog of word 평가, rougeLsum은 최장 길이 토큰 순서 평가

2. 평가

- Accuracy: Math, Code, Retrieve, En.Dia, En.MC
- ROUGE F1: 서술형 QA {Zh, En}.QA
- rougeLsum: En.Sum

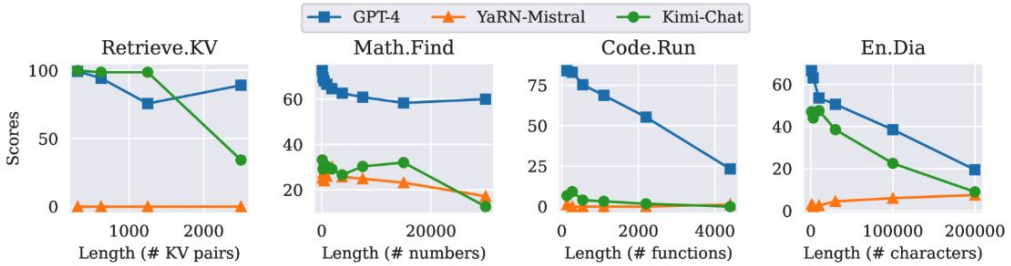


Figure 4: Baseline performance as a function of input length.

=> 길수록 성능 떨어짐

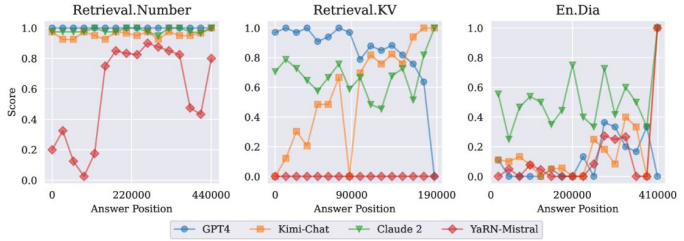


Figure 5: Performance as a function of the answer position (in the number of characters). The steep drop in performance for Kimi-Chat in the middle on Retrieval.KV is caused by the answer being removed by truncation.

=> context내의 정답의 위치가 성능에 영향을 미침

https://arxiv.org/pdf/2404.06654

RULER: What's the Real Context Size of Your Long-Context Language Models?

1. 데이터 구성

Retrieval: (NIAH)

Task	Configuration	Example
Single NIAH (S-NIAH)	type_key = word type_value = number type_haystack = essay size_haystack ∝ 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 ∝ 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 ∝ 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 ∝ 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

Multi-hop Tracing

Variable Tracking (VT)	num_chains = 2 num_hops = 2 size_noises ∝ 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

Aggregation

Common Words Extraction (CWE)	freq_cw = 2, freq_ucw = 1 num_cw = 10 num_ucw ∝ 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 ∝ 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

QA

Question Answering (QA)	dataset = SQuAD num_document ∝ context length	Document 1: ..... aaa ..... Document 2: ..... bbb ..... Document 3: ..... ccc ..... Question: question Answer: bbb

https://arxiv.org/pdf/2404.06654

RULER: What's the Real Context Size of Your Long-Context Language Models?

2. 평가

Models	Claimed Length	Effective Length	4K	8K	16K	32K	64K	128K	Avg.	wAvg. (inc)	wAvg. (dec)
Llama2-7B (chat)	4K	-	96.9								
Gemini-1.5	1M	>128K	99.8	99.9	99.6	99.7	99.7	99.6	99.7	$99.7_{(1st)}$	$99.7_{(1st)}$
Llama3.1 (8B)	128K	64K	99.9	99.9	99.8	99.6	98.7	92.6	98.4	$97.5_{(3rd)}$	$99.4_{(2nd)}$
GLM4 (9B)	1M	64K	99.4	99.2	99.5	99.4	97.3	94.4	98.2	$97.5_{(2nd)}$	$98.9_{(3rd)}$
Llama3.1 (70B)	128K	64K	100.0	100.0	99.9	99.6	98.5	78.9	96.1	$93.5_{(5th)}$	$98.8_{(4th)}$
GPT-4	128K	32K	99.9	99.9	98.7	98.3	90.9	84.8	95.4	$92.9_{(6th)}$	$97.9_{(5th)}$
Command-R-plus (104B)	128K	32K	99.9	99.9	99.4	97.9	89.6	65.7	92.1	$87.3_{(8th)}$	$96.9_{(6th)}$
GradientAI/Llama3 (70B)	1M	16K	99.0	98.8	98.3	94.5	91.2	84.9	94.4	$92.1_{(7th)}$	$96.8_{(7th)}$
Yi (34B)	200K	16K	98.2	96.8	97.3	95.1	93.0	90.2	95.1	$93.8_{(4th)}$	$96.4_{(8th)}$
Qwen2 (72B)	128K	32K	100.0	99.9	99.9	99.4	84.5	48.0	88.6	$81.3_{(11th)}$	$95.9_{(9th)}$
Phi3-medium (14B)	128K	8K	98.7	98.5	96.6	95.4	91.9	51.3	88.7	$82.6_{(10th)}$	$94.9_{(10th)}$
Mixtral-8x22B (39B/141B)	64K	16K	99.3	99.1	97.7	96.7	89.9	23.8	84.4	$74.8_{(12th)}$	$94.1_{(11th)}$
LWM (7B)	1M	<4K	92.5	92.1	87.6	83.7	84.1	83.4	87.2	$85.5_{(9th)}$	$89.0_{(12th)}$
Mistral-v0.2 (7B)	32K	4K	98.1	96.2	94.3	85.5	51.1	10.7	72.6	$58.8_{(13th)}$	$86.5_{(13th)}$
DBRX (36B/132B)	32K	8K	99.4	99.0	93.5	73.4	0.5	0.0	61.0	$41.6_{(14th)}$	$80.3_{(14th)}$
Together (7B)	32K	<4K	96.2	89.9	82.3	80.2	0.0	0.0	58.1	$40.2_{(15th)}$	$76.0_{(15th)}$
LongChat (7B)	32K	<4K	93.3	92.2	81.1	67.3	0.0	0.0	55.7	$37.6_{(16th)}$	$73.7_{(16th)}$
LongAlpaca (13B)	32K	<4K	74.9	72.2	70.8	53.2	0.0	0.0	45.2	$30.7_{(17th)}$	$59.7_{(17th)}$
Llama2-7B (base)	4K	-	90.9								
Mixtral-base (8x7B)	32K	32K	99.9	99.7	98.4	94.8	72.1	29.1	82.3	$71.8_{(2nd)}$	$92.8_{(1st)}$
Mistral-base (7B)	32K	16K	99.3	97.5	95.7	89.8	56.8	10.2	74.9	$61.2_{(4th)}$	$88.6_{(2nd)}$
Jamba-base (52B)	256K	<4K	86.4	80.5	73.7	72.3	68.1	56.9	73.0	$68.5_{(3th)}$	$77.4_{(5th)}$
LWM-base (7B)	1M	<4K	88.5	87.7	84.5	79.6	76.1	74.2	81.8	$79.1_{(1st)}$	$84.4_{(4th)}$
LongLoRA-base (7B)	100K	16K	95.3	95.6	92.7	81.5	76.2	0.0	73.5	$60.6_{(5th)}$	$86.5_{(3rd)}$
Yarn-base (7B)	128K	<4K	89.9	86.1	78.4	59.0	49.5	17.5	63.4	$51.7_{(6th)}$	$75.1_{(7th)}$
Together-base (7B)	32K	8K	95.4	91.5	86.1	75.1	0.0	0.0	58.0	$39.9_{(7th)}$	$76.2_{(6th)}$

Table 13: Performance of selected aligned and base models across length 4K to 128K by averaging 8 task scores in Retrieval (NIAH) of RULER.

=> Claimed != Effective (even at Retrieval)

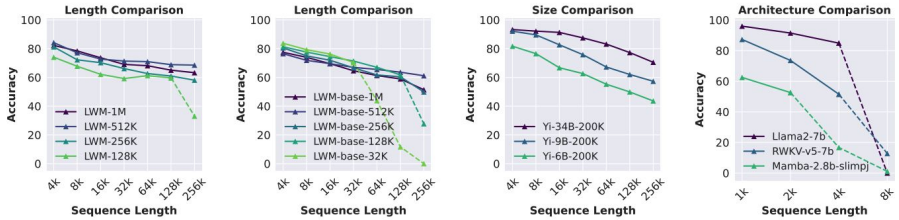


Figure 4: (**Left & middle left**): Comparison of LargeWorldModel (LWM) series trained up to various context sizes with fixed parameter size of 7B. (**Middle right**): Comparison of Yi suite models with different parameter sizes with controlled training context length of 200K. (**Right**): Performance of non-Transformer architectures lags behind the Transformer baseline Llama2-7B by large margin. Length extrapolation is presented with dashed lines.

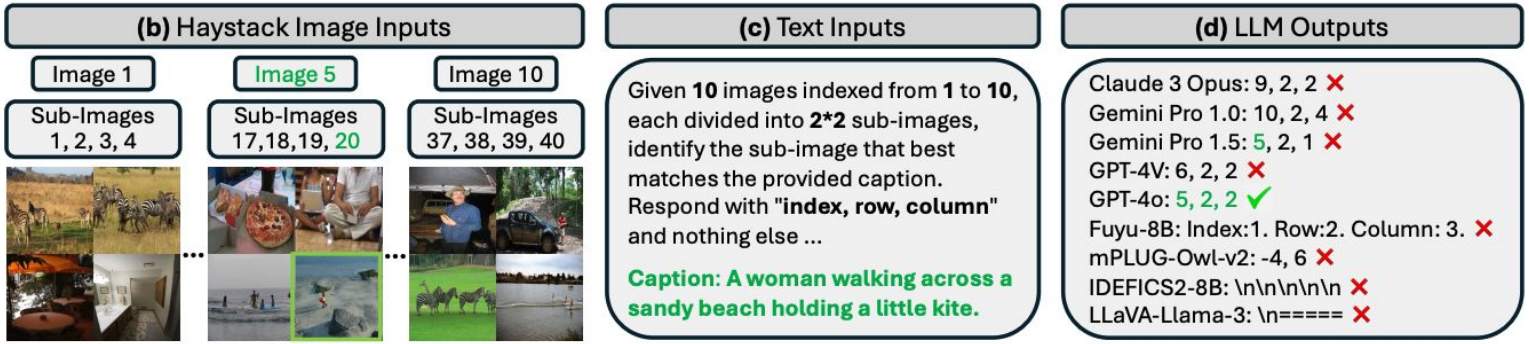
=> 모델 크기, 아키텍쳐마다 effective length 다름

https://aclanthology.org/2025.naacl-long.166/

Multimodal Needle in a Haystack: Benchmarking Long-Context Capability of Multimodal Large Language Models

1. 데이터 구성

*MM-NIAH는 카운팅 등의 복잡한 태스크도 실험



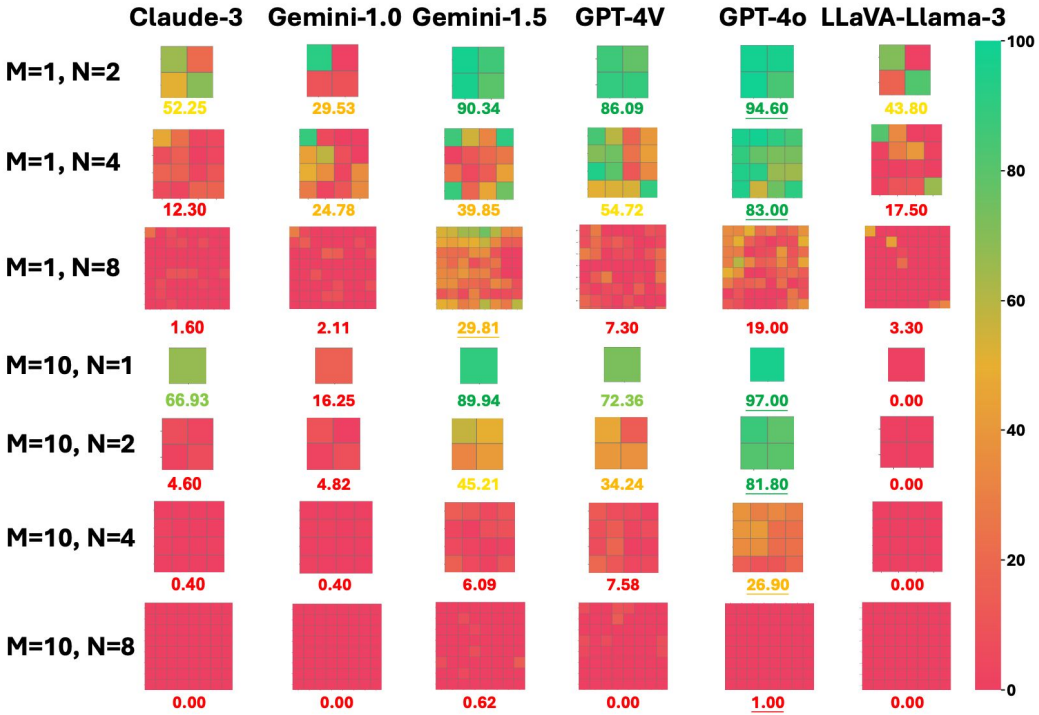
=> 이미지를 grid list로 제공하고, 이미지 위치를 Retrieval + “없음"도 정답에 포함

- Single Needle, Multi Needle
- image grid와 이미지 개수가 ablation

https://aclanthology.org/2025.naacl-long.166/

Multimodal Needle in a Haystack: Benchmarking Long-Context Capability of Multimodal Large Language Models

2. 평가 - Accuracy



M: 이미지 장수, N: Grid

1. GPT 계열이 Vision에 강함

2. 이미지 장수, Grid가 커지면 성능 악화

3. 특정 포지션을 잘 맞추는 경향이 있음

https://aclanthology.org/2025.naacl-long.166/

Multimodal Needle in a Haystack: Benchmarking Long-Context Capability of Multimodal Large Language Models

2. 평가 - Existence, Index, Exact (sub index까지 맞추는지)

Table 3: Accuracy (%) for the $M = 10$ setting. We mark the best results with **bold face**. Note that the existence accuracy is measured by whether the model outputs "-1".

	Stitching	1 × 1			2 × 2			4 × 4			8 × 8		
	Metrics	*Existence*	*Index*	*Exact*	*Existence*	*Index*	*Exact*	*Existence*	*Index*	*Exact*	*Existence*	*Index*	*Exact*
API-Based Models	Claude 3 Opus	83.77	67.23	66.93	66.60	9.90	4.60	64.78	6.46	0.40	54.13	5.93	0.00
	Gemini Pro 1.0	83.66	33.90	16.25	81.63	10.74	4.82	58.92	4.81	0.40	18.11	1.61	0.00
	Gemini Pro 1.5	97.08	90.04	89.94	98.84	53.42	45.21	96.17	17.26	6.09	89.02	9.86	0.62
	GPT-4V	95.11	75.59	72.36	98.32	52.10	34.24	99.80	24.87	7.58	99.50	10.57	0.00
	GPT-4o	99.00	**97.00**	**97.00**	99.60	**87.20**	**81.80**	**100.00**	**45.00**	**26.90**	99.80	**17.80**	**1.00**
Open-Source Models	Fuyu-8B	**100.00**	0.00	0.00	**100.00**	0.00	0.00	**100.00**	0.00	0.00	**100.00**	0.00	0.00
	mPLUG-Owl-v2	15.90	5.60	0.40	70.10	5.20	0.10	88.50	8.10	0.00	86.10	6.30	0.00
	IDEFICS2-8B	71.10	0.30	0.00	93.80	0.70	0.00	99.60	6.40	0.00	96.60	2.40	0.00
	LLaVA-Llama-3	**100.00**	0.20	0.00	**100.00**	0.10	0.00	**100.00**	0.00	0.00	**100.00**	0.00	0.00

=> 존재 여부는 잘 맞춰도, Index, Exact는 오픈소스 모델들이 현저히 떨어짐
– 오픈소스 모델 한계: structure output 실패가 다수

https://arxiv.org/pdf/2506.02921

A Controllable Examination for Long-Context Language Models

기존 Long Context 한계
- Real-world 벤치는 오류 해석이 어렵고, 데이터 오염 존재
- Synthetic은 무관한 haystack이 needle을 찾기 쉽게 만듦

좋은 Long Context
- Seamless Context: 문맥상 연결되는 needle => shortcut을 방지할 수 있음
- Controllable Settings: 길이/난이도를 통제 가능
- Sound evaluation: 재현 가능한 평가 (exact match 등)

Table 1: Comparison of long-context Benchmarks. Here cheap means that the benchmark is cheap to construct.
*: We only consider the non-synthetic task within the benchmark.

		Seamlessness		Controllability		Soundness	
Benchmark	Cheap	Fluent	Coherent	Configurable	Extendable	Leakage Prev.	Reliable Metric
L-Eval [1]	✗	✓	✓	✗	✗	✗	✗
LongBench-v2 [2]	✗	✓	✓	✗	✗	✗	✓
NoCha [3]	✗	✓	✓	✗	✗	✗	✓
∞-Bench*	✗	✓	✓	✗	✗	✗	✗
Helmet* [4]	✗	✓	✓	✗	✗	✗	✗
BABILong [5]	✓	✓	✗	✗	✓	✓	✓
RULER [6]	✓	✗	✗	✓	✓	✓	✓
Michelangelo [7]	✓	✗	✗	✓	✓	✓	✓
NoLiMa [8]	✓	✓	✗	✓	✗	✓	✓
MRCR(OpenAI) [9]	✓	✓	✓	✓	✗	✓	✓
LongBioBench	✓	✓	✓	✓	✓	✓	✓

https://arxiv.org/pdf/2506.02921

A Controllable Examination for Long-Context Language Models

데이터 구성 (LongBio Bench)
- 한 사람의 bio (이름, 생일, 주소, 등)를 Context로 하고 이 정보를 기반으로 문제 파생
=> parametric knowledg[illegible] shortcut을

Example Context (Standard)

Context:
[bios_1]: ...
...
[bios_i]: Below is the bio of Andrew Xavier Jimenez. The birthday of Andrew Xavier Jimenez is 1972-11-09. Andrew Xavier Jimenez was born in Dembī Dolo. The hobby of Andrew Xavier Jimenez is radio-controlled model collecting...
This is the end of this bio.
...
[bios_n]: ...

Question: What is the birthday of Andrew Xavier Jimenez?

Answer: The birthday of Andrew Xavier Jimenez is 1972-11-09.

Table 2: Task Overview for the LongBioBench. Here {Pn} refers to the name of the n-th person. Acc is the accuracy of the exact match.

Task	Description	Metric	Example
Understanding			
Standard	Retrieve a specific attribute of one person.	Acc	**Attribute:** The hobby of {P1} is dandyism. **Question:** What's the hobby of {P1}?
Multi_standard	Retrieve multiple attributes of different people.	All-or-Nothing Acc	**Attribute:** The hobby of {P1} is dandyism. {P2} is mycology. **Question:** What's the hobby of {P1} and {P2}?
Paraphrase	Attribute expressions are paraphrased.	Acc	**Attribute:** {P1} worked in Dhaka. **Question:** Which city did {P1} work in?
Pronoun	Bio written from first-person view.	Acc	**Attribute:** I was born on 1993-06-26. **Question:** What is the birthday of {P1}?
Reasoning			
Calculation	Compute age difference between two people.	Acc	**Attribute:** {P1} is 61, {P2} is 43. **Question:** What's their age difference?
Rank	Rank people by age.	Acc	**Attribute:** {P1} is 61, {P2} is 43. **Question:** Rank from youngest to oldest.
Multihop	Retrieve an attribute via cross-person reference.	Acc	**Attribute:** {P1} born in Santa Paula. {P2} born same place as {P1}. **Question:** Birthplace of {P2}?
Twodiff	Identify two people with specific age difference.	Acc	**Attribute:** {P1} is 61, {P2} is 43. **Question:** Who has 18 years age difference?
Trustworthy			
Citation	Answer plus source citation.	Citation Acc	**Attribute:** Bio [1]: {P1} born in Santa Paula. **Question:** Which university did Isabel graduate from?
IDK	No-answer case detection.	Refuse while Answer Acc	**Attribute:** Attribute removed. **Question:** What's the hobby of {P1}?

https://arxiv.org/pdf/2506.02921

A Controllable Examination for Long-Context Language Models

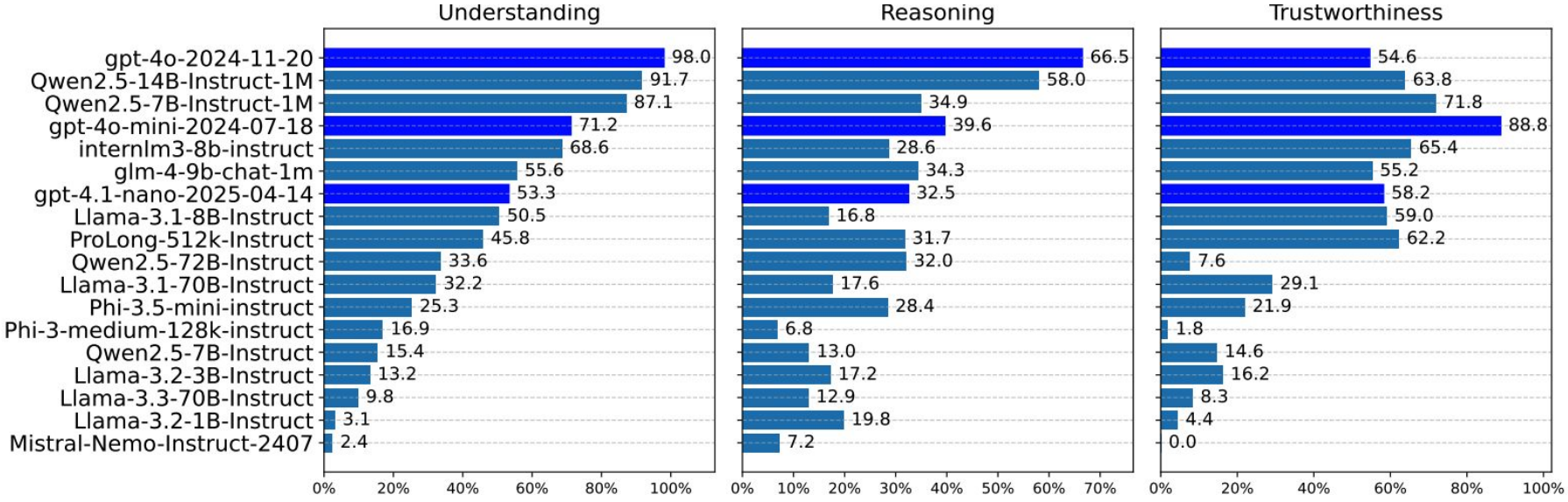


Figure 3: The average performance of all models on Understanding, Reasoning, and Trustworthiness categories.

=> Understanding, Reasoning이 좋아도 Trustworthiness에서는 역전

Real-World Long Benchmarks

- (medical) MedOdyssey: (4K, 8K, …200K)

- (law) CaseHOLD

- (manufacturing) DesignQA: (70k + images)