MMLongBench: Benchmarking Long-Context Vision-Language Models Effectively and Thoroughly

Anonymous Authors[1]

Abstract

The rapid extension of context windows in large vision-language models has given rise to *long-context vision-language models* (LCVLMs), which are capable of handling hundreds of images with interleaved text tokens in a single forward pass. In this work, we introduce MMLongBench, the first benchmark covering a diverse set of long-context vision-language tasks, to evaluate LCVLMs effectively and thoroughly. MMLongBench is composed of 13,331 examples spanning five different categories of downstream tasks with broad coverage of image types. To assess models on different input lengths, all examples are delivered at five standardized input lengths via a cross-modal tokenization scheme that combines vision patches and text tokens. By thoroughly benchmarking 46 LCVLMs, we find that all models face challenges in our multimodal long-context tasks. Then, we also provide a comprehensive analysis of the current models' long-context ability. With wide task coverage, various image types, and rigorous length control, MMLongBench provides the missing foundation for developing the next generation of LCVLMs.

1. Introduction

Recent advances in long-context modeling unlocked a wide array of new capabilities for both large language models (LLMs; Dubey et al., 2024; Yang et al., 2024) and large vision–language models (LVLMs; Kamath et al., 2025; Meta, 2025). In particular, long-context vision–language models (LCVLMs) represent an important step forward by enabling LVLMs to process hundreds of images and thousands of interleaved text tokens in a single forward pass. This allows applications such as document-level visual question answering (Ma et al.), multi-hop reasoning across web pages (He et al., 2024), and instruction following in complex visual contexts (Shridhar et al., 2020; Li et al., 2022).

To support such capabilities, researchers have proposed various techniques to extend the context windows of LVLMs (Chen et al., 2024a; Hurst et al., 2024). However, the development of effective evaluation benchmarks is lagging behind. It remains unclear how well current LCVLMs perform in long-context settings, what types of tasks they struggle with, and how robust they are to input length variation. Here, we take a closer look and find that existing benchmarks suffer from the following shortcomings and provide key feature comparisons in Table 1:

- **Limited task coverage**: Existing benchmarks often target a single long-context vision-language task, like needle-in-a-haystack (Wang et al., 2024a) or long-document VQA (Ma et al.), which fails to capture the broader capabilities needed for diverse applications (Hsieh et al.).
- **Insufficient image-type coverage**: Most existing benchmarks are limited to either natural images (Wang et al., 2024a; Wu et al., 2024b) – such as photographs of everyday scenes, objects, or people – or synthetic images (Ma et al.; Deng et al., 2024), such as scanned documents, web pages, or app screenshots. This limitation leads to an incomplete understanding of model performance across diverse image types.
- **Lack of context length control**: Existing benchmarks miss a consensus on cross-modality length control, especially image tokens. For example, MM-NIAH (Wang et al., 2024c) follows InternVL1.5 (Chen et al., 2024c) to compute text and image tokens together, but other works (Wu et al., 2024b; Deng et al., 2024) only report the number of images as the context length. This inconsistency makes it difficult to compare model performance across different benchmarks.
- **Single context length**: Many text-only long-context benchmarks use a few standardized context lengths (e.g., 32K, 128K), providing examples at each standard length (Hsieh et al.; Yen et al., 2024). Hence, model developers can easily know the performance change with different lengths. However, such practice is not followed in LCVLM evaluations. For example, MM-NIAH (Wang et al., 2024c) builds contexts with web pages of randomly varying lengths, complicating systematic analysis of context length effects.

In this paper, we introduce MMLongBench, a comprehensive benchmark covering a diverse set of long-

Table 1. Benchmark comparison for LCVLMs: MM-NIAH, Visual Haystack, MMNeedle, MMLongBench-Doc (MMLB-Doc), M-Longdoc, LongDocURL, and our MMLongBench. "Summ" and "DocVQA" refer to summarization and long-document VQA. "L" means the number of input tokens. "Mixed" indicates that the dataset includes both natural and synthetic images.

	Type of tasks					Benchmark features		
	VRAG	NIAH	ICL	Summ	DocVQA	Image Type	L Control	Multiple L
MM-NIAH (Wang et al., 2024c)	✗	✓	✗	✗	✗	Mixed	✓	✗
Visual Haystack (Wu et al., 2024b)	✗	✓	✗	✗	✗	Natural	✗	✓
MMNeedle (Wang et al., 2024a)	✗	✓	✗	✗	✗	Natural	✗	✓
MMLB-Doc (Ma et al.)	✗	✗	✗	✗	✓	Synthetic	✗	✗
M-Longdoc (Chia et al.)	✗	✗	✗	✗	✓	Synthetic	✗	✗
LongDocURL (Deng et al., 2024)	✗	✗	✗	✗	✓	Synthetic	✗	✗
MMLongBench (Ours)	✓	✓	✓	✓	✓	Mixed	✓	✓

context vision-language tasks across five different categories. Specifically, in addition to multimodal *needle-in-a-haystack* (NIAH) and *long-document VQA* (DocVQA) tasks, we also include *visual retrieval-augmented generation* (VRAG), *many-shot in-context learning* (ICL), and *summarization* in our benchmark. VRAG examples are from knowledge-based VQA (Chen et al.) with Wikipedia passages as context, and ICL examples involve on-the-fly image classification. For summarization, models are required to summarize image-based PDF documents. Overall, our benchmark includes diverse downstream tasks and image types to enable a comprehensive evaluation. In addition, we use a unified token-count method that counts image tokens based on the number of patches produced by current vision encoders, followed by a 2×2 pixel unshuffle. This approach is consistent with most recent models, such as Qwen2.5-VL (Bai et al., 2025), making it well-suited for long-context evaluation. Furthermore, we equip all examples with five standardized context lengths, from 8K to 128K tokens. This enables a thorough analysis of performance changes as the context length increases.

Finally, we thoroughly evaluated 46 frontier LCVLMs and find that: i) performance on a single task poorly reflects overall long-context ability; ii) long-context vision-language tasks present significant challenges for all frontier models; and iii) models with stronger reasoning ability tend to exhibit better long-context capabilities. Furthermore, our error analysis reveals that Optical Character Recognition (OCR) and cross-modality retrieval abilities remain bottlenecks in current LCVLMs.

2. Our Benchmark — MMLongBench

In our work, we seek to address the limitations of current benchmarks by meeting the following criteria: i) broad coverage of both vision-language downstream tasks and image types, ii) a unified token-count method across different modalities and datasets, and iii) multiple standardized context lengths for each example, from 8K to 128K tokens. In this section, we describe the task categories in MMLongBench, highlighting the improvements over existing benchmarks. A benchmark overview is provided in Table 2, and several ***concrete examples*** are given in Appendix F.

2.1. Diverse Long-Context Applications for LCVLMs

Visual retrieval-augmented generation (VRAG) evaluates an LCVLM's ability to locate relevant information from a large corpus. Here, we use the factual knowledge VQA, which requires answering questions about the named entity shown in an image, such as "Who designed the building in this picture?" We include InfoSeek (Chen et al.) and ViQuAE (Lerner et al., 2022) in this category.

We insert the gold passage(s) among a large set of distracting passages retrieved from Wikipedia. For ViQuAE, we use gold passages from KILT (Petroni et al., 2021), as it is constructed upon TriviaQA (Joshi et al., 2017). For InfoSeek, we choose the lead section of the named entity's Wikipedia page as the gold reference and remove all examples for which the answer cannot be found. We split Wikipedia pages into 100-word passages and add retrieved passages that do not contain the answer or the named entity until the input length L. We use the substring exact match (SubEM) as the metric, following previous work (Asai et al., 2023). See more details in Appendix B.1. We also test different needle positions to eliminate bias (Appendix D).

Needle-in-a-haystack (NIAH) measures how well an LCVLM can recall a piece of information embedded within a long, unrelated multimodal input. We adopt tasks from Visual Haystack (VH; Wu et al., 2024b) and MM-NIAH (Wang et al., 2024c). VH requires retrieving images of target objects in an image haystack, with VH-Single and VH-Multi for finding one or multiple images, respectively. MM-NIAH includes retrieval (Ret), counting (Count), and reasoning (Reason) tasks with interleaved text and images; each featuring both text and image needles.

In VH, needle images and target objects are from the original datasets. We accompany needle images with negative images until a given input length L. We report accuracy following the original work (Wu et al., 2024b). In MM-NIAH,

Table 2. Overview of MMLongBench. We include 16 datasets and 13,331 examples in total. Image types are shown per dataset; "Mixed" indicates both natural and synthetic images. SubEM and Acc indicate substring exact match and accuracy.

Category	Dataset	Metrics	Image	Size	Description
Visual RAG	InfoSeek	SubEM	Natural	1,128	Long-tail entity question answering
	ViQuAE	SubEM	Natural	1,144	Question answering based on TriviaQA
Needle-in-a-Haystack	VH-Single	Acc	Natural	1,000	Retrieve an image from an album
	VH-Multi	Acc	Natural	1,000	Retrieve multiple images from an album
	MM-NIAH-Ret	SubEM/Acc	Mixed	1,200	Retrieve text/image needles in web pages
	MM-NIAH-Count	Acc	Mixed	1,178	Count text/image needles in web pages
	MM-NIAH-Reason	SubEM/Acc	Mixed	1,158	Reason about text/image needles in web pages
Many-Shot In-Context Learning	Stanford Cars	Acc	Natural	458	50-category car classification
	Food101	Acc	Natural	500	50-category food classification
	SUN397	Acc	Natural	500	50-category scene classification
	iNat2021	Acc	Natural	500	50-category species classification
Summarization	GovReport	Model-based	Synthetic	241	Summarizing government reports in PDF
	Multi-LexSum	Model-based	Synthetic	146	Summarizing multiple legal documents in PDF
Long-Document VQA	MMLongBench-Doc	SubEM/Acc	Synthetic	961	Long PDF document VQA
	LongDocURL	SubEM/Acc	Synthetic	1,153	Long PDF document VQA
	SlideVQA	SubEM/Acc	Synthetic	1,064	Slide deck understanding and reasoning

the haystacks are composed of web pages (Laurençon et al., 2023), and we include all three tasks of retrieval, counting, and reasoning. Similar to VRAG, we split text in web pages into 100-word passages and add passages and images to reach the input length. Following the original work (Wang et al., 2024c), we use SubEM and accuracy for retrieval and reasoning. For counting, we use the needle count accuracy due to better robustness. See Appendix B.2 for details.

Many-shot in-context learning (ICL) tests models' ability to adapt to new tasks on the fly using in-context examples. Following prior work (Yen et al., 2024; Li et al., 2024c; Bertsch et al.), we focus on image classification with large label spaces. We collect four datasets with diverse domains: Stanford Cars (Krause et al., 2013), Food101 (Bossard et al., 2014), SUN397 (Xiao et al., 2010), and iNat2021 (Van Horn et al., 2021). We adjust the number of shots to reach the length L, and exemplars in each class are balanced. We randomly sample 50 classes per dataset to ensure sufficient shots, since the 128K context window accommodates 500 images. We report accuracy as the metric.

Summarization (Summ) evaluates the ability to generate concise outputs of salient information from long multimodal documents. We choose GovReport (Huang et al., 2021) and Multi-LexSum (Shen et al., 2022), as their PDF-formatted documents are long and easily accessible. Our evaluation provides models with PDF-formatted documents rather than OCR-extracted text used in previous works (Tay et al.; Gao et al., 2024). We truncate the document from the end based on the input length L. Following previous work (Yen et al., 2024), we use LLM-based evaluation for both datasets instead of the commonly used ROUGE-L. More details are provided in Appendix B.4.

Long-document VQA assesses models' aptitude to answer questions that require reasoning over multiple images and text segments. We include commonly adopted datasets: SlideVQA (Tanaka et al., 2023), MMLongBench-Doc (Ma et al.), and LongDocURL (Deng et al., 2024). For documents longer than L, we truncate them evenly from both sides while keeping the answer pages. For shorter documents, we alternately pad both sides with random negative documents up to L. Padding documents may occasionally contain information related to the question and change the answer. To ensure the validity, we preface each question with the prompt "Based on the Document <*Original Doc ID*>, answer the following question." We follow the metrics used in LongDocURL but remove questions with long answers, thereby avoiding LLM-based answer extraction. We list specific details in Appendix B.5.

2.2. Cross-Modality Token Counting

Various long-context tasks of LCVLMs involve varying text-to-image ratios. For example, VRAG contains only one image related to a named entity, whereas the context in Long-Document VQA primarily consists of images. A key challenge in benchmarking LCVLMs lies in standardizing the context length of diverse datasets with different text-image combinations. In this work, we count both text and visual tokens together as the input length L, unlike prior works (Wang et al., 2024a; Wu et al., 2024b; Lu et al., 2024c) that simply use the image number as context length. We use the Llama2 tokenizer (Touvron et al., 2023) to calculate the number of text tokens following previous practice (Yen et al., 2024). Then, we divide each image into 14×14 patches and apply a 2×2 pixel unshuffle to compress the visual token number. The patch size and pixel unshuffle are both commonly adopted in current LVLMs (Chen et al., 2024c; Bai et al., 2025; Zhu et al., 2025; Abouelenin et al., 2025). This method ensures compatibility with modern LVLMs, making it well-suited for evaluation.

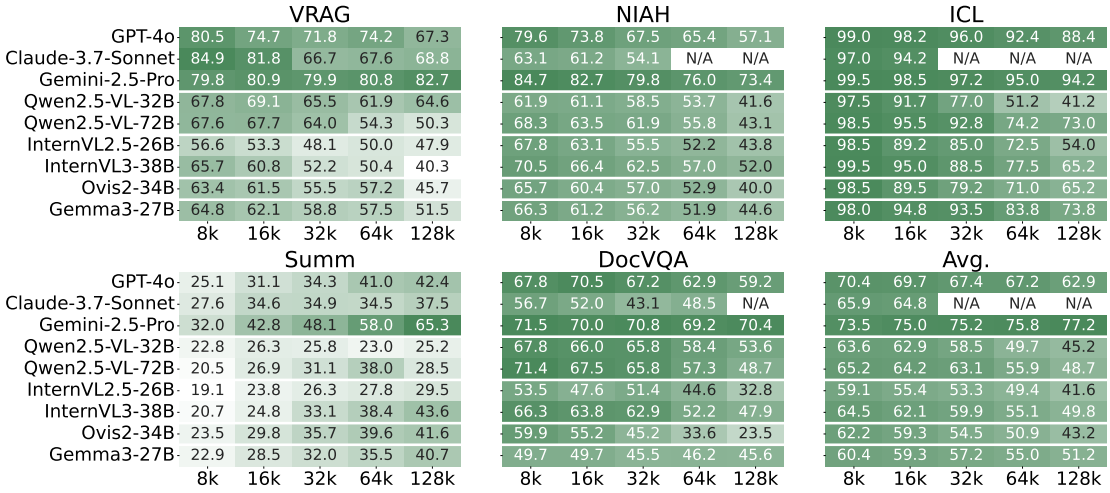


Figure 1. Performance on MMLONGBENCH for selected frontier models, and the full results of all models are provided in Figure 22. Note that Claude-3.7-Sonnet supports at most 100 images, and we mark the results as N/A for cases with more images (More in Appendix E.8)

2.3. Standardized Input Length

Input length L is crucial in evaluating long-context ability, as longer inputs offer more information but also introduce distracting information. As aforementioned in Section 2.1, we can control the input length L for each dataset either by adjusting the number of passages, images, or exemplars, or by truncating the PDF-formatted documents. This allows us to present each example at multiple standardized input lengths and better understand how performance changes as the context length increases. Specifically, our benchmark provides five input lengths L: 8K, 16K, 32K, 64K, and 128K tokens, using binary prefixes $K = 2^{10}$.

3. Evaluation and Analysis

In total, we evaluate 46 LCVLMs on MMLONGBENCH. As we know, our evaluation provides the most thorough and controlled comparison of the vision-language long-context ability. We list all models evaluated in Table 10. More experimental details are in Appendix D. We provide the full results and analysis in Appendix E.

We present the performance of selected frontier LCVLMs in Figure 1, and the full results of all 46 models are reported in Figure 22. We analyze model performance from multiple perspectives and summarize our main findings as follows:

All models struggle, but closed-source models perform better. Considering the performance at 128K tokens, all models struggle on our long-context tasks. Even GPT-4o only achieves 62.9 on average, while open-source models perform even worse. We find that Gemini-2.5-Pro is the strongest LCVLM, outperforming open-source models by roughly 15–20 points across all tasks. While other closed-source models generally outperform open-source ones, the margin is usually less than 10 points. Notably, Qwen2.5-VL-32B almost matches GPT-4o on VRAG, and InternVL3-38B even surpasses GPT-4o on summarization, showing the competitiveness of open-source models.

Different models exhibit different strengths. We find that model performance varies considerably across different tasks. For example, Qwen2.5-VL-32B outperforms InternVL3-38B on VRAG, but underperforms on NIAH. Ovis2-34B excels at summarization but struggles on DocVQA. These findings support the necessity of a comprehensive benchmark covering diverse downstream tasks.

Reasoning can improve long-context ability. We include Gemini-2.0-Flash-T in Figure 22, the thinking variant of Gemini-2.0-Flash. From the results, we observe that the reasoning ability can consistently improve the performance on all tasks. In particular, summarization and DocVQA exhibit improvements of 25.3% and 10.1%, respectively. Then, Gemini-2.5 models exhibit even stronger performance, which are natively designed as thinking models.

4. Conclusion

In this work, we have introduced MMLONGBENCH, the first comprehensive benchmark for evaluating long-context vision–language models (LCVLMs) across a wide spectrum of downstream tasks. By covering five distinct task categories—while unifying cross-modal token counting and standardizing context lengths, MMLONGBENCH provides a rigorous, extensible foundation for diagnosing the strengths and weaknesses of frontier LCVLMs. We conducted a comprehensive evaluation of 46 LCVLMs, yielding several insightful findings—for example, that evaluation on a single task does not reliably predict overall long-context capability. Looking forward, we hope MMLONGBENCH will serve as a standard yardstick for the community, driving research on more efficient vision-language token encodings, more robust position-extrapolation schemes, and improved multi-modal retrieval and reasoning capabilities.

Impact Statement

The long-context ability of LVLMs unlocked a large range of applications, including understanding documents with hundreds of pages and reasoning over dozens of web pages automatically. This ability can also help users to summarize a long document or revise a large-scale code repository. Meanwhile, there are a large number of instruction-following scenarios grounded in complex vision-language contexts, such as long-term dialogue with humans or dialogue-based navigation for robots. Looking ahead, our MMLONGBENCH will serve as a standard evaluation for the whole community to benchmark new LCVLMs and to stimulate the development of models with more efficient vision-language token encodings, more robust position-extrapolation schemes, and improved OCR, multi-modal retrieval, and reasoning capabilities.

References

Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Abouelenin, A., Ashfaq, A., Atkinson, A., Awadalla, H., Bach, N., Bao, J., Benhaim, A., Cai, M., Chaudhary, V., Chen, C., et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

Agrawal, P., Antoniak, S., Hanna, E. B., Bout, B., Chaplot, D., Chudnovsky, J., Costa, D., De Monicault, B., Garg, S., Gervet, T., et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024.

Anthropic. Claude 3.7 sonnet. *URL https://www.anthropic.com/news/claude-3-7-sonnet*, 2024.

Arora, S., Eyuboglu, S., Timalsina, A., Johnson, I., Poli, M., Zou, J., Rudra, A., and Re, C. Zoology: Measuring and improving recall in efficient language models. In *The Twelfth International Conference on Learning Representations*.

Asai, A., Wu, Z., Wang, Y., Sil, A., and Hajishirzi, H. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2023.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Bai, Y., Lv, X., Zhang, J., Lyu, H., Tang, J., Huang, Z., Du, Z., Liu, X., Zeng, A., Hou, L., et al. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137, 2024a.

Bai, Y., Tu, S., Zhang, J., Peng, H., Wang, X., Lv, X., Cao, S., Xu, J., Hou, L., Dong, Y., et al. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *arXiv preprint arXiv:2412.15204*, 2024b.

Beltagy, I., Peters, M. E., and Cohan, A. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Bertsch, A., Ivgi, M., Alon, U., Berant, J., Gormley, M. R., and Neubig, G. In-context learning with long-context models: An in-depth exploration. In *First Workshop on Long-Context Foundation Models@ ICML 2024*.

Bertsch, A., Alon, U., Neubig, G., and Gormley, M. Unlimiformer: Long-range transformers with unlimited length input. *Advances in Neural Information Processing Systems*, 36:35522–35543, 2023.

Bossard, L., Guillaumin, M., and Van Gool, L. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pp. 446–461. Springer, 2014.

Cattan, A., Roit, P., Zhang, S., Wan, D., Aharoni, R., Szpektor, I., Bansal, M., and Dagan, I. Localizing factual inconsistencies in attributable text generation. *arXiv preprint arXiv:2410.07473*, 2024.

Chang, Y., Lo, K., Goyal, T., and Iyyer, M. Booookscore: A systematic exploration of book-length summarization in the era of llms. In *The Twelfth International Conference on Learning Representations*.

Chang, Y., Narang, M., Suzuki, H., Cao, G., Gao, J., and Bisk, Y. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16495–16504, 2022.

Chen, S., Wong, S., Chen, L., and Tian, Y. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.

Chen, Y., Hu, H., Luan, Y., Sun, H., Changpinyo, S., Ritter, A., and Chang, M.-W. Can pre-trained vision and language models answer visual information-seeking questions? In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Chen, Y., Xue, F., Li, D., Hu, Q., Zhu, L., Li, X., Fang, Y., Tang, H., Yang, S., Liu, Z., et al. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024a.

Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.

Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., Ma, J., Wang, J., wen Dong, X., Yan, H., Guo, H., He, C., Jin, Z., Xu, C., Wang, B., Wei, X., Li, W., Zhang, W., Zhang, B., Lu, L., Zhu, X., Lu, T., Lin, D., and Qiao, Y. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024c.

Chia, Y. K., Cheng, L., Chan, H. P., LIU, C., Song, M., Aljunied, M., Poria, S., and Bing, L. M-longdoc: A benchmark for multimodal super-long document understanding and a retrieval-aware tuning framework.

Dao, T. Flashattention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations*.

Deng, C., Yuan, J., Bu, P., Wang, P., Li, Z.-Z., Xu, J., Li, X.-H., Gao, Y., Song, J., Zheng, B., et al. Longdocurl: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. *arXiv preprint arXiv:2412.18424*, 2024.

Deutsch, D., Dror, R., and Roth, D. Re-examining system-level correlations of automatic summarization evaluation metrics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 6038–6052, 2022.

Ding, Y., Zhang, L. L., Zhang, C., Xu, Y., Shang, N., Xu, J., Yang, F., and Yang, M. Longrope: extending llm context window beyond 2 million tokens. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 11091–11104, 2024.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A. S., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., tiste Rozière, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E. A., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I. M., Misra, I., Evtimov, I., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J.-Q., Alwala, K. V., Upasani, K., Plawiak, K., Li, K., neth Heafield, K.-., Stone, K. R., El-Arini, K., Iyer, K., Malik, K., ley Chiu, K., Bhalla, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M. H. M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., lay Bashlykov, N., Bogoychev, N., Chatterji, N. S., Duchenne, O., cCelebi, O., Alrassy, P., Zhang, P., Li, P., Vasić, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Girdhar, R., Patel, R., main Sauvestre, R., nie Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., hana Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S. C., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., ginie Do, V., Vogeti, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., ney Meers, W., Martinet, X., Wang, X., Tan, X. E., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A. K., Grattafiori, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Vaughan, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Franco, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, P.-Y. B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Civin, D., Beaty, D., Kreymer, D., Li, S.-W.,

Wyatt, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Ozgenel, F., Caggioni, F., Guzm'an, F., Kanayet, F. J., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Thattai, G., Herman, G., Sizov, G. G., Zhang, G., Lakshminarayanan, G., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Molybog, I., Tufanov, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., KamHou, U., Saxena, K., Prasad, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Huang, K., Chawla, K., Lakhotia, K., Huang, K., Chen, L., Garg, L., Lavender, A., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Tsimpoukelli, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Laptev, N. P., Dong, N., Zhang, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., dro Rittner, P., Bontrager, P., Roux, P., Dollár, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Li, R., Hogan, R., Battey, R., Wang, R., Maheswari, R., Howes, R., Rinott, R., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Shankar, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Cho, S.-B., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Kohler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V. A., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wang, X., Wu, X., Wang, X., Xia, X., Wu, X., Gao, X., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Wang, Y., Hao, Y., Qian, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., and Zhao, Z. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Fu, Y., Panda, R., Niu, X., Yue, X., Hajishirzi, H., Kim, Y., and Peng, H. Data engineering for scaling language models to 128k context. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 14125–14134, 2024.

Gao, T., Wettig, A., Yen, H., and Chen, D. How to train long-context language models (effectively). *arXiv preprint arXiv:2410.02660*, 2024.

Ge, J., Chen, Z., Lin, J., Zhu, J., Liu, X., Dai, J., and Zhu, X. V2pe: Improving multimodal long-context capability of vision-language models with variable visual position encoding. *arXiv preprint arXiv:2412.09616*, 2024.

Google. Introducing gemini 2.0: our new ai model for the agentic era, 2024. URL https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/#ceo-message.

Google. Gemini 2.5: Our most intelligent ai model. *URL https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/*, 2025.

Goyal, T., Li, J. J., and Durrett, G. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*, 2022.

Gu, A. and Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*.

He, H., Yao, W., Ma, K., Yu, W., Dai, Y., Zhang, H., Lan, Z., and Yu, D. Webvoyager: Building an end-to-end web agent with large multimodal models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6864–6890, 2024.

Hsieh, C.-P., Sun, S., Kriman, S., Acharya, S., Rekesh, D., Jia, F., and Ginsburg, B. Ruler: What's the real context size of your long-context language models? In *First Conference on Language Modeling*.

Hu, H., Luan, Y., Chen, Y., Khandelwal, U., Joshi, M., Lee, K., Toutanova, K., and Chang, M.-W. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12065–12075, 2023.

Huang, L., Cao, S., Parulian, N., Ji, H., and Wang, L. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1419–1436, 2021.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b, 2023.

Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

kaiokendev. Things i'm learning while training superhot, 2023. *URL https://kaiokendev.github.io/til*, 2023.

Kamath, G. T. A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ram'e, A., Rivière, M., Rouillard, L., Mesnard, T., Cideron, G., Grill, J.-B., Ramos, S., Yvinec, E., Casbon, M., Pot, E., Penchev, I., Liu, G., Visin, F., Kenealy, K., Beyer, L., Zhai, X., Tsitsulin, A., Busa-Fekete, R. I., Feng, A., Sachdeva, N., Coleman, B., Gao, Y., Mustafa, B., Barr, I., Parisotto, E., Tian, D., Eyal, M., Cherry, C., Peter, J.-T., Sinopalnikov, D., Bhupatiraju, S., Agarwal, R., Kazemi, M., Malkin, D., Kumar, R., Vilar, D., Brusilovsky, I., Luo, J., Steiner, A., Friesen, A., Sharma, A., Sharma, A., Gilady, A. M., Goedeckemeyer, A., Saade, A., Kolesnikov, A., Bendebury, A., Abdagic, A., Vadi, A., Gyorgy, A., Pinto, A. S., Das, A., Bapna, A., Miech, A., Yang, A., Paterson, A., Shenoy, A., Chakrabarti, A., Piot, B., Wu, B., Shahriari, B., Petrini, B., Chen, C., Lan, C. L., Choquette-Choo, C. A., Carey, C., Brick, C., Deutsch, D., Eisenbud, D., Cattle, D., Cheng, D., Paparas, D., Sreepathihalli, D. S., Reid, D., Tran, D., Zelle, D., Noland, E., Huizenga, E., Kharitonov, E., Liu, F., Amirkhanyan, G., Cameron, G., Hashemi, H., Klimczak-Pluci'nska, H., Singh, H., Mehta, H., Lehri, H. T., Hazimeh, H., Ballantyne, I., Szpektor, I., Nardini, I., Pouget-Abadie, J., Chan, J., Stanton, J., Wieting, J. M., Lai, J., Orbay, J., Fernandez, J., Newlan, J., Ji, J., Singh, J., Black, K., Yu, K., Hui, K., Vodrahalli, K., Greff, K., Qiu, L., Valentine, M., Coelho, M., Ritter, M., Hoffman, M., Watson, M., Chaturvedi, M., Moynihan, M., Ma, M., Babar, N., Noy, N., Byrd, N., Roy, N., Momchev, N., Chauhan, N., Bunyan, O., Botarda, P., Caron, P., Rubenstein, P. K., Culliton, P., Schmid, P., Sessa, P. G., Xu, P., Stańczyk, P., Tafti, P. D., Shivanna, R., Wu, R., Pan, R., Rokni, R. A., Willoughby, R., Vallu, R., Mullins, R., Jerome, S., Smoot, S., Girgin, S., Iqbal, S., Reddy, S., Sheth, S., Põder, S., Bhatnagar, S., Panyam, S. R., Eiger, S., Zhang, S., Liu, T., Yacovone, T., Liechty, T., Kalra, U., Evci, U., Misra, V., Roseberry, V., Feinberg, V., Kolesnikov, V., Han, W., Kwon, W., Chen, X., Chow, Y., Zhu, Y., Wei, Z., Egyed, Z., Cotruta, V., Giang, M., Kirk, P., Rao, A., Lo, J., Moreira, E., Martins, L. G., Sanseviero, O., Gonzalez, L., Gleicher, Z., Warkentin, T., Mirrokni, V. S., Senter, E., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Matias, Y., Sculley, D., Petrov, S., Fiedel, N., Shazeer, N. M., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Alayrac, J.-B., Anil, R., Lepikhin, D., Borgeaud, S., Bachem, O., Joulin, A., Andreev, A., Hardin, C., Dadashi, R., and Hussenot, L. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Kamoi, R., Goyal, T., Rodriguez, J. D., and Durrett, G. Wice: Real-world entailment for claims in wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7561–7583, 2023.

Kamradt, G. Needle in a haystack-pressure testing llms, 2023. *URL https://github.com/gkamradt/LLMTest_NeedleInAHaystack*, 2024.

Kim, Y., Chang, Y., Karpinska, M., Garimella, A., Manjunatha, V., Lo, K., Goyal, T., and Iyyer, M. Fables: Evaluating faithfulness and content selection in book-length summarization. In *First Conference on Language Modeling*.

Kočiskỳ, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K. M., Melis, G., and Grefenstette, E. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018.

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

Kwan, W. C., Zeng, X., Wang, Y., Sun, Y., Li, L., Jiang, Y., Shang, L., Liu, Q., and Wong, K.-F. M4le: A multi-ability multi-range multi-task multi-domain long-context evaluation benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15568–15592, 2024.

Laurençon, H., Saulnier, L., Tronchon, L., Bekman, S., Singh, A., Lozhkov, A., Wang, T., Karamcheti, S., Rush, A., Kiela, D., et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36:71683–71702, 2023.

Laurençon, H., Marafioti, A., Sanh, V., and Tronchon, L. Building and better understanding vision-language models: insights and future directions. In *Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models*, 2024a.

Laurençon, H., Tronchon, L., Cord, M., and Sanh, V. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37: 87874–87907, 2024b.

Lerner, P., Ferret, O., Guinaudeau, C., Le Borgne, H., Besançon, R., Moreno, J. G., and Lovón Melgarejo, J. Viquae, a dataset for knowledge-based visual question answering about named entities. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pp. 3108–3120, 2022.

Levy, M., Jacoby, A., and Goldberg, Y. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15339–15353, 2024.

Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Li, M., Zhang, S., Liu, Y., and Chen, K. Needlebench: Can llms do retrieval and reasoning in 1 million context window? *arXiv preprint arXiv:2407.11963*, 2024b.

Li, T., Zhang, G., Do, Q. D., Yue, X., and Chen, W. Long-context llms struggle with long in-context learning. *arXiv preprint arXiv:2404.02060*, 2024c.

Li, Y., Li, W., and Nie, L. Mmcoqa: Conversational question answering over text, tables, and images. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4220–4231, 2022.

Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., and Zhang, M. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023.

Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 11:157–173, 2024a.

Liu, Z., Zhu, L., Shi, B., Zhang, Z., Lou, Y., Yang, S., Xi, H., Cao, S., Gu, Y., Li, D., et al. Nvila: Efficient frontier visual language models. *arXiv preprint arXiv:2412.04468*, 2024b.

Lu, H., Liu, W., Zhang, B., Wang, B., Dong, K., Liu, B., Sun, J., Ren, T., Li, Z., Yang, H., et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024a.

Lu, S., Li, Y., Chen, Q.-G., Xu, Z., Luo, W., Zhang, K., and Ye, H.-J. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*, 2024b.

Lu, Y., Li, X., Fu, T.-J., Eckstein, M., and Wang, W. Y. From text to pixel: Advancing long-context understanding in mllms. *arXiv preprint arXiv:2405.14213*, 2024c.

Ma, Y., Zang, Y., Chen, L., Chen, M., Jiao, Y., Li, X., Lu, X., Liu, Z., Ma, Y., Dong, X., et al. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Mathew, M., Karatzas, D., and Jawahar, C. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

Meta. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation, 2025. *URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/*, 2025.

Mishra, A., Shekhar, S., Singh, A. K., and Chakraborty, A. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pp. 947–952. IEEE, 2019.

Pan, J., Gao, T., Chen, H., and Chen, D. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8298–8319, 2023.

Peng, B., Quesnelle, J., Fan, H., and Shippole, E. Yarn: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*.

Petroni, F., Piktus, A., Fan, A., Lewis, P., Yazdani, M., De Cao, N., Thorne, J., Jernite, Y., Karpukhin, V., Maillard, J., et al. Kilt: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2523–2544, 2021.

Shaham, U., Ivgi, M., Efrat, A., Berant, J., and Levy, O. Zeroscrolls: A zero-shot benchmark for long text understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7977–7989, 2023.

Shen, Z., Lo, K., Yu, L., Dahlberg, N., Schlanger, M., and Downey, D. Multi-lexsum: Real-world summaries of civil rights lawsuits at multiple granularities. *Advances in Neural Information Processing Systems*, 35:13158–13173, 2022.

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., and Wang, Z. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1874–1883, 2016.

Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., Zettlemoyer, L., and Fox, D. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10740–10749, 2020.

Song, D., Chen, S., Chen, G. H., Yu, F., Wan, X., and Wang, B. Milebench: Benchmarking mllms in long context. *arXiv preprint arXiv:2404.18532*, 2024a.

Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., Chi, H., Guo, X., Ye, T., Zhang, Y., et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18221–18232, 2024b.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Tanaka, R., Nishida, K., Nishida, K., Hasegawa, T., Saito, I., and Saito, K. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 13636–13645, 2023.

Tay, Y., Dehghani, M., Tran, V. Q., Garcia, X., Wei, J., Wang, X., Chung, H. W., Bahri, D., Schuster, T., Zheng, S., et al. Ul2: Unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations*.

Team, G., Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Touvron, H., Martin, L., Stone, K. R., Albert, P., Almahairi, A., Babaei, Y., lay Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D. M., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A. S., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I. M., Korenev, A. V., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M. H. M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Van Horn, G., Cole, E., Beery, S., Wilber, K., Belongie, S., and Mac Aodha, O. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12884–12893, 2021.

Wang, H., Shi, H., Tan, S., Qin, W., Wang, W., Zhang, T., Nambi, A., Ganu, T., and Wang, H. Multimodal needle in a haystack: Benchmarking long-context capability of multimodal large language models. *arXiv preprint arXiv:2406.11230*, 2024a.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Wang, W., Zhang, S., Ren, Y., Duan, Y., Li, T., Liu, S., Hu, M., Chen, Z., Zhang, K., Lu, L., et al. Needle in a multimodal haystack. *Advances in Neural Information Processing Systems*, 37:20540–20565, 2024c.

Wang, X., Song, D., Chen, S., Zhang, C., and Wang, B. Longllava: Scaling multi-modal llms to 1000 images efficiently via a hybrid architecture. *arXiv preprint arXiv:2409.02889*, 2024d.

Wang, Z., Zhang, H., Fang, T., Tian, Y., Yang, Y., Ma, K., Pan, X., Song, Y., and Yu, D. Divscene: Benchmarking lvlms for object navigation with diverse scenes and objects. *arXiv preprint arXiv:2410.02730*, 2024e.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Wu, H., Li, D., Chen, B., and Li, J. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024a.

Wu, T.-H., Biamby, G., Quenum, J., Gupta, R., Gonzalez, J. E., Darrell, T., and Chan, D. M. Visual haystacks: A vision-centric needle-in-a-haystack benchmark. *arXiv preprint arXiv:2407.13766*, 2024b.

Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024c.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Ye, J., Xu, H., Liu, H., Hu, A., Yan, M., Qian, Q., Zhang, J., Huang, F., and Zhou, J. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *CoRR*, 2024.

Yen, H., Gao, T., Hou, M., Ding, K., Fleischer, D., Izsak, P., Wasserblat, M., and Chen, D. Helmet: How to evaluate long-context language models effectively and thoroughly. *arXiv preprint arXiv:2410.02694*, 2024.

Young, A. A., Chen, B., Li, C., Huang, C., Zhang, G., Zhang, G., Li, H., Zhu, J., Chen, J., Chang, J., Yu, K., Liu, P., Liu, Q., Yue, S., Yang, S., Yang, S., Yu, T., Xie, W., Huang, W., Hu, X., Ren, X., Niu, X., Nie, P., Xu, Y., Liu, Y., Wang, Y., Cai, Y., Gu, Z., Liu, Z., and Dai, Z. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024.

Zhang, S. and Bansal, M. Finding a balanced degree of automation for summary evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6617–6632, 2021.

Zhang, X., Chen, Y., Hu, S., Xu, Z., Chen, J., Hao, M., Han, X., Thai, Z., Wang, S., Liu, Z., et al. Infbench: Extending long context evaluation beyond 100k tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15262–15277, 2024.

Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Duan, Y., Tian, H., Su, W., Shao, J., Gao, Z., Cui, E., Cao, Y., Liu, Y., Xu, W., Li, H., Wang, J., Lv, H., Chen, D., Li, S., He, Y., Jiang, T., Luo, J., Wang, Y., He, C., Shi, B., Zhang, X., Shao, W., He, J., Xiong, Y., Qu, W., Sun, P., Jiao, P., Wu, L., Zhang, K., Deng, H., Ge, J., Chen, K., Wang, L., Dou, M., Lu, L., Zhu, X., Lu, T., Lin, D., Qiao, Y., Dai, J., and Wang, W. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A. Related Work

Long-context vision-language models (LCVLMs). The context window of LLMs has experienced a fast growth from less than 8K (Touvron et al., 2023; Jiang et al., 2023) to 128K tokens (Hurst et al., 2024; Dubey et al., 2024) or more (Team et al., 2024; Anthropic, 2024). To support this, techniques such as longer pre-training length (Dubey et al., 2024; Fu et al., 2024; Young et al., 2024), position extrapolation (Peng et al.; Ding et al., 2024; Chen et al., 2023), and efficient architectures (Beltagy et al., 2020; Bertsch et al., 2023; Gu & Dao) have been developed. With this progress, recent literature also investigated how to extend the context length of LVLMs to build LCVLMs, such as Gemini-2.5 (Google, 2025), Qwen2.5-VL (Bai et al., 2025), and others (Abouelenin et al., 2025; Zhu et al., 2025; Kamath et al., 2025; Ye et al., 2024). In addition, several recent works on LVLMs made efforts to compress vision tokens to accommodate longer input sequences (Shi et al., 2016; Wang et al., 2024b; Wu et al., 2024c; Laurençon et al., 2024b;a; Abdin et al., 2024). Meanwhile, a growing body of works has adopted various techniques from LLMs to extend LVLMs' context length, such as position extrapolation (Ge et al., 2024) and more efficient model architectures (Wang et al., 2024d). With extended context lengths, LCVLMs can support various applications, such as multi-hop reasoning across web pages (He et al., 2024) and instruction following grounded in complex visual contexts (Shridhar et al., 2020; Li et al., 2022; Wang et al., 2024e).

Long-context benchmarks. Needle-in-a-haystack (NIAH) (Kamradt, 2024) is one of the first commonly adopted tasks to evaluate the text-pure long context ability of LLMs, as it can be procedurally generated with arbitrarily long lengths and needle position (Liu et al., 2024a). This task inserts a "needle" at specific depths of a long essay and tests models' ability to recall it. Recent works have also extended the NIAH task to more complex versions (Li et al., 2024b; Levy et al., 2024; Arora et al.). However, several benchmarks (Yen et al., 2024; Hsieh et al.) discover that using a single NIAH task only partially reflects LLMs' overall long-context ability. As a result, numerous benchmarks with broad coverage of diverse downstream applications have been constructed (Bai et al., 2024a; Shaham et al., 2023; Hsieh et al.; Yen et al., 2024; Bai et al., 2024b; Zhang et al., 2024; Kwan et al., 2024) to provide a comprehensive evaluation.

In contrast, the evaluation of LVLMs' long-context capability remains limited. Existing benchmarks only involve either NIAH (Wang et al., 2024c;a; Wu et al., 2024b; Lu et al., 2024c) or long-document VQA (Ma et al.; Deng et al., 2024), lacking comprehensive coverage across diverse vision-language applications. As a result, frontier LCVLMs (Hurst et al., 2024; Team et al., 2024) only report long-context performance on other modalities, such as video (Chen et al., 2024a; Wu et al., 2024a; Song et al., 2024b) or audio (Team et al., 2024; Hurst et al., 2024), and neglect the prevalent use cases of long-context vision-language inputs. While recent MileBench (Song et al., 2024a) claims to be a comprehensive long-context benchmark with various text-image tasks, our closer inspection reveals that it actually contains a lot of short-context tasks, and the average length is only about 9K tokens. Datasets like DocVQA(Mathew et al., 2021), WebQA (Chang et al., 2022), and OCRVQA (Mishra et al., 2019) contain only one image per sample and minimal context, making MileBench unqualified as a true long-context benchmark. In this work, we introduce the first comprehensive benchmark that evaluates a wide range of vision-language downstream applications across five standardized input lengths.

B. Dataset Details

In this appendix, we provide more details on how to build long-context examples based on existing datasets.

B.1. Visual Retrieval-Augmented Generation

Gold Passage. InfoSeek (Chen et al.) is a large-scale dataset for factual knowledge-based VQA featuring long-tail entities from Wikipedia (Hu et al., 2023). For InfoSeek, we use the lead section of the Wikipedia page of the named entity in the question image as gold passages, which is the first section on each page and serves as a summary of the whole page. The lead section may be long, so we chunk it into multiple 100-word passages. We remove all the queries whose corresponding lead section does not contain the correct answer.

ViQuAE (Lerner et al., 2022) repalces the named entities in questions from TriviaQA (Joshi et al., 2017) with corresponding entity images from Wikimedia Commons[1]. We obtain gold passages for each question from the KILT benchmark (Petroni et al., 2021), which provides human annotations of gold passages for queries in TriviaQA.

[1] https://commons.wikimedia.org/

Length Control. We populate the context with hard negative passages from Wikipedia, and the version we used is the Wikipedia 2019-08-01 dump (Petroni et al., 2021). We follow the KILT benchmark to preprocess Wikipedia articles into 100-word passages. For retrieval, we adopt a retrieval-and-rerank pipeline, where BM25 is first used for coarse retrieval, followed by reranking with dense embedding from `Alibaba-NLP/gte-large-en-v1.5` (Li et al., 2023). Here, we replace the image in each question with its original entity name for better retrieval accuracy, because text-based retrieval achieves higher recall and provides harder distractors.

Previous work (Yen et al., 2024) shows that this pipeline presents a significantly greater challenge than randomly sampled passages. Also, using a real embedding model for retrieval is consistent with various downstream visual retrieval-agumented generation tasks and can better reflect downstream application performance.

B.2. Needle-in-a-Haystack

Length Control. For the Visual Haystack dataset (Wu et al., 2024b), we directly use the needles and target objects from the original dataset. Then, we change the number of negative distractors to build long-context examples with a given input length L. Note that the original dataset simply reports the image number as context length, ignoring different image sizes. Here, we count image tokens based on patches split by vision encoders as discussed in Section 2.2. There are two tasks in the dataset: VH-Single and VH-Multi, where the target objects are contained in a single needle image or multiple needle images, respectively.

For MM-NIAH (Wang et al., 2024c), the contexts in the original dataset are composed of entire web pages from OBELICS (Laurençon et al., 2023). However, using full web pages makes it difficult to control input length *at a fine granularity* since web pages typically contain tens of thousands of tokens. To solve the issue, we chunk the text content of web pages into 100-word passages, as we did for the Wikipedia corpus in VRAG. Meanwhile, the images in MM-NIAH contain only a few hundred tokens. Thus, we can achieve fine-grained control over the context length and incrementally add text passages and images to reach a given length L.

Metrics. We report the accuracy on Visual Haystack, exactly the same as in the original work. The MM-NIAH dataset contains *three* different tasks: needle retrieval, counting, and reasoning. We use MM-NIAH-Ret, MM-NIAH-Count, and MM-NIAH-Reason as their abbreviations, respectively. In each task, there are both text-needle and image-needle examples. In MM-NIAH-Ret and MM-NIAH-Reason, we use substring exact matching (SubEM) for text-needle examples and accuracy for image-needle examples, exactly following the original paper (Wang et al., 2024c). In MM-NIAH-Count, we find that the soft accuracy metric proposed in the original paper (Wang et al., 2024c) can be exploited: simply predicting a list of zeros ([0, 0, ...]) results in a score over 30 on image-needle counting. Thus, we report the accuracy of the total count of the needle in the haystack instead of comparing the list of needle counts, which we find is more reliable. Last but not least, we sample text-needle and image-needle examples evenly in all three tasks.

Please refer to the original MM-NIAH paper (Wang et al., 2024c) for comprehensive details of all three tasks and two needle modalities.

B.3. Many-Shot In-Context Learning

Class Sampling. We include Stanford Cars (Krause et al., 2013), Food101 (Bossard et al., 2014), SUN397 (Xiao et al., 2010), and iNat 2021 (Van Horn et al., 2021). Since the 128K context length can accommodate only about 500 images, 50 different classes are randomly sampled from each dataset. With 50 classes, we can ensure that there are about 10 exemplars from each class, which is sufficient. For iNat 2021, since the dataset contains substantially more classes (over 10,000 species), we randomly sample 50 classes from the "Birds" supercategory and 50 classes from the "Plants" supercategory. For every single example, all the exemplars and the test image are either from the "Birds" classes or the "Plants" classes, ensuring the task remains a 50-way classification problem. Meanwhile, for shorter input lengths, we need to reduce the class number to ensure sufficient shots per class. Specifically, we randomly sample 5, 10, 20, and 40 classes for the input length of 8K, 16K, 32K, and 64K tokens. With those class numbers, we find that the number of exemplars per class is similar to that when there are 128K tokens.

Label mapping and length control. We employ a label mapping strategy to ensure that models perform classification based on in-context exemplars instead of relying on their pre-trained knowledge. Each label is randomly mapped to an integer $i \in \{0, 1, \ldots, N-1\}$, where N is the number of classes, following established practices (Pan et al., 2023; Yen et al.,

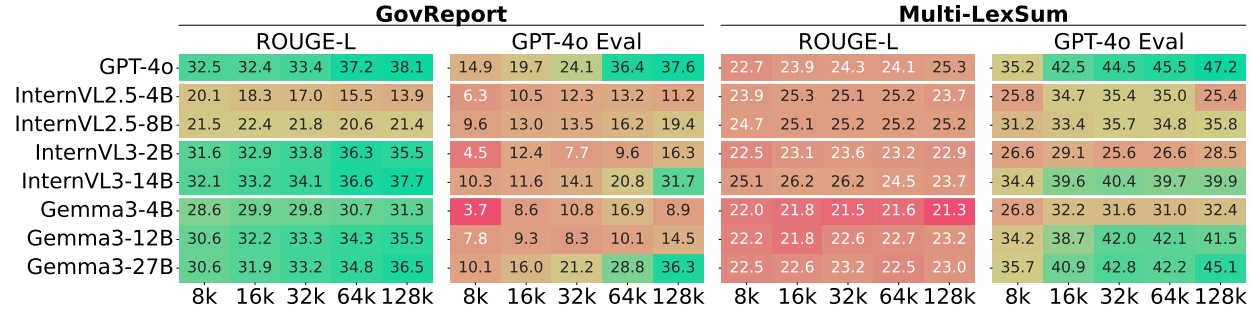


Figure 2. Comparison between ROUGE-L and the GPT-4o evaluation on summarization datasets. GPT-4o evaluation reflects the performance gain on Gemma3-27B with increased input length, and it also clearly sets apart open-source models with different sizes. In comparison, ROUGE-L remains almost the same for all models and input lengths.

2024). Throughout the evaluation, we provide models with images and their corresponding integer labels. Following Li et al. (2024c), we arrange exemplars into demonstration rounds, each of which includes exactly one exemplar per label in a random order. We concatenate these demonstration rounds, with the last round truncated if needed, to build examples of input length L. Thus, the label distribution is balanced in all datasets and input lengths.

B.4. Summarization

Preprocessing. GovReport (Huang et al., 2021) consists of reports written by the U.S. Government Accountability Office (GAO)[2] and the Congressional Research Service (CRS)[3]. GAO reports constitute the majority of the dataset (more than 12K) and provide enough coverage for evaluation. Since CRS reports have a different format from GAO reports and there are only a few CRS reports available, we only use GAO reports in our benchmark. Summaries of GAO reports are written by experts and are structured into three aspects: "Why GAO did this study," "What GAO found," and "What GAO recommends." Those summaries are written at the beginning pages of the PDF-formatted GAO documents. We use PyMuPDF [4] to detect those answers and remove the corresponding pages to ensure no answer leakage in the inputs.

Multi-LexSum (Shen et al., 2022) consists of multi-document summarization problems about civil rights lawsuits, and the summaries are written by domain experts (i.e., lawyers and law students).

Both datasets are constructed using the OCR-extracted plain text as the input. In our evaluation, we replace the OCR-extracted plain text with the original PDF-formatted documents. We screenshot each page of a PDF-formatted document with 144 DPI, following common practices (Ma et al.). Different from previous works (Ma et al.; Deng et al., 2024), we *do not concatenate* images to reduce the token numbers and instead directly feed them into LCVLMs since we are stress-testing the model's long-context capability.

Length Control. To control the input length L, we truncate document pages from the end. When there are multiple documents in Multi-LexSum, we truncate each document evenly from the end. Additionally, we discard examples that exceed the 128K context length by more than 24K tokens, as adding them would require truncating too many pages to fit within the context window of 128K. In this way, we can avoid confounding effects on model performance caused by the loss of key information during page truncation.

Data Scale. In long-form generation tasks, each summary typically contains many atomic claims to be verified, in contrast to short outputs of other categories, such as VRAG. There are 15,951 claims in 387 examples in these two datasets, indicating a large scale for evaluation.

Model-Based Metric The N-gram overlap metrics, such as ROUGE-L (Lin, 2004), have long been condemned for their poor correlation with human judgment for long-form generation (Goyal et al., 2022; Deutsch et al., 2022). To ensure reliable evaluation for summarization, we adopt the reference-based LLM evaluation method proposed in HELMET (Yen et al.,

[2] www.gao.gov
[3] crsreports.congress.gov
[4] https://pymupdf.readthedocs.io

2024). Specifically, we first break down the gold reference summary into a set of atomic claims with GPT-4o, following prior work (Kamoi et al., 2023; Zhang & Bansal, 2021; Cattan et al., 2024). Next, we ask the model to check for three properties of model predictions: precision, recall, and fluency. We utilize GPT-4o to assess if each sentence in the generated summary is supported by the gold reference (precision) and if each claim in the gold reference is present in the generated summary (recall). The F1 score is computed from the recall and precision. We also prompt GPT-4o to assess the fluency of the generated summary. The fluency is assigned a value of 0 if the output is incoherent, incomplete, or repetitive, and a value of 1 if it is fluent and coherent. The final score, Fluency-F1, is the product of fluency and F1 score.

Our empirical study in Figure 2 demonstrates that InternVL3-2B achieves ROUGE-L scores comparable to GPT-4o. Moreover, ROUGE-L exhibits minimal difference across different input lengths from 8K to 128K tokens. These observations reveal that ROUGE-L has low discriminative capacity and often fails to effectively distinguish between the quality of generated texts. In contrast, GPT4-o evaluation shows a significant gap across different input lengths and shows lower scores for models with shorter context windows, such as InternVL2.5.

Atomic Claims Verification. We manually checked 100 atomic claims from 25 Multi-LexSum summaries and another 100 atomic claims from 25 GovReport summaries. We found that only one claim was not factually accurate. Then, we checked the coverage of the claims and found no key facts were missing. This manual verification shows that GPT-4o is virtually always reliable for the decomposition task. For Multi-LexSum, we follow HELMET (Yen et al., 2024) and use the short summary to obtain atomic claims, where the dataset also provides a long and tiny summary for each case.

GPT-4o Judgment Verification. We show the detailed prompts for evaluating the fluency, precision, and recall in Tables 4 to 9, following previous works (Kamoi et al., 2023; Yen et al., 2024; Chang et al.; Kim et al.). We further conduct human analysis to verify the evaluation metric.

Quantitatively, we found that GPT-4o can consistently distinguish fluent and non-fluent outputs. The agreement between the model judgements and human judgements is 100% for randomly sampled outputs from GovReport and Multi-LexSum. Then, we sample 10 generated summaries for both GovReport and Multi-LexSum (20 in total) and check 5 atomic claims evaluations for each summary. The models we used are Gemini-2.5-Pro and Qwen2.5-VL-32B. We follow a similar procedure to the GPT-4o evaluation and manually check the precision and recall of those sampled summaries. For precision, we observed Cohen's $\kappa = 0.90$ for GovReport and $\kappa = 0.89$ for Multi-LexSum, suggesting almost perfect agreement. Meanwhile, for recall, we observed Cohen's $\kappa = 0.90$ for GovReport and $\kappa = 0.93$ for Multi-LexSum, which are also near perfect.

Qualitatively, inspecting the disagreements, we find that most disagreements come from partially supported cases. We identified two common underlying reasons for partially supported cases when measuring precision and recall, respectively. First, a sentence in a generated summary may contain two points: "While agencies generally documented their review, inconsistencies and documentation gaps existed." We found the reference summary only supports "inconsistencies and documentation gaps existed," and the "agencies generally documented" part is an entailment inferred by the model. *Such inferred (entailed) information causes a lot of partially supported cases when measuring precision.* Second, the claims in the gold reference may include specific details, such as some geographic locations or organization names. These details may not be explicitly mentioned in the generated summary, *causing the partially supported cases for recall.*

B.5. Long-Document VQA

Preprocessing. MMLongBench-Doc (Ma et al.) and LongDocURL (Deng et al., 2024) contain questions on various kinds of documents, such as financial reports, guidebooks, and academic papers, and the answer formats include string, integer, float, and list. More importantly, the rule-based evaluation method commonly adopted on those datasets depends on answer formats.

First, we find that there is a proportion of noisy answer format annotations. For example, a list answer like `['Top 10 File Categories Sorted By Disk Space', 'Last 12 Months Modified Disk Space History']` is annotated as being in string format. Conversely, answers in string format are also annotated in list format, such as `['PRIVACY SCREEN OPTIONS']`. Therefore, our first step with these datasets is to correct the mislabeled answer formats and discard the instances for which the correct answer format cannot be recovered.

Second, both datasets rely heavily on LLMs, such as GPT-4o, to extract the answer from model predictions. This leads to *high evaluation costs* and poses *challenges for large-scale evaluation*, like 46 models in our work. Then, we take a closer

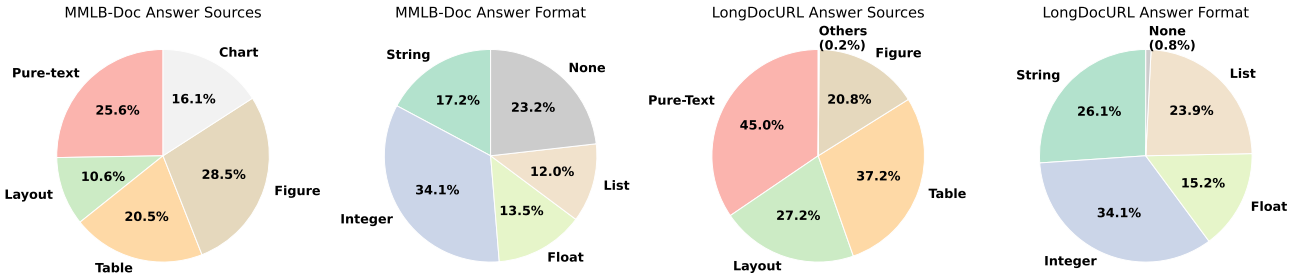


Figure 3. Data distribution of MMLongBench-Doc and LongDocURL after our pre-processing. Both datasets remain well-distributed, and their distributions are similar to the ones in the original paper.

examination of different formats of answers: (1) For integer and float answers, we find that numbers can be extracted with regular expressions; (2) For string answers, if the answer is short (less than 5 words), we find that model predictions are also short. Thus, we can directly use automatic metrics like ROUGE F1 without the need for answer extractions; (3) As a result, only *long-form string answers* require LLM-based extraction; Since long-form string answers ($>$ 5 words) constitute only a small proportion of these datasets, we simply discard those instances to enable scalable evaluation without relying on GPT-4o for answer extraction. We find that the retained short string answers are mostly entity names; (4) Note that for list answers, we evaluate each element in the list (i.e., integer, float, or string) separately (then task average), and the answer evaluation is determined for each element by its type.

After all the filtering, we find both datasets remain *well-distributed* as shown in Figure 3.

Evaluation Metrics. We follow previous works (Ma et al.; Deng et al., 2024) and employ the same rule-based scoring method that applies different strategies depending on the format of the reference answer: (1) For **String** format answers, we initially use regular expressions to determine whether the answers require exact matching (e.g., telephone numbers, email addresses, website addresses, filenames, times, dates, etc.). If the answer needs an exact match, we perform a substring exact matching (SubEM) with a score of 0 or 1. Otherwise, we follow previous works (Tanaka et al., 2023; Zhang et al., 2024; Kočiskỳ et al., 2018; Yen et al., 2024) and calculate ROUGE F1 scores; (2) For integer answers, we perform an exact match comparison, and the score is either 0 or 1; (3) For float answers, we treat the model prediction and gold reference as the same if the relative error is less than 1%; (4) For list answers, we evaluate each element separately based on its answer type and take the average. Here, we follow LongDocURL to use the Greedy List Match: for each element in the reference list, we compute its score against all elements in the prediction list and *greedily* select the highest score as its matching score. This metric does not require the predicted list to follow the same element order as the reference list, thereby providing greater tolerance in evaluation.

Different from them, SlideVQA (Tanaka et al., 2023) features questions based on 20-page slide decks, which contain rich layout information and less dense text. The answer formats in the dataset are string, integer, and float, and do not cover list answers. We use the same rule-based scoring method as described for MMLongBench-Doc and LongDocURL.

Length Control. The input lengths of DocVQA tasks are also easy to control. If an example exceeds a given length L, we truncate the document evenly from both sides while preserving the answer pages. If the document cannot fill the length L, we alternately pad the left and right sides with randomly sampled negative documents until the required length is reached. Notably, we may also truncate a few pages of the last padding document as needed to control the length at the granularity of pages, instead of documents.

The randomly sampled padding documents are not guaranteed to be truly irrelevant or negative. They may occasionally contain information related to the question, which could potentially change the answer. To ensure models attend to the original document, we preface each question with the prompt "Based on the Document <*Original Doc ID*>, answer the following question."

Table 3. Average number of images per example in all datasets and input lengths. The values in subscript denote the standard deviations. (T) and (I) represents the text and image needle in each task of MM-NIAH.

	Data Length	8K	16K	32K	64K	128K
VRAG	InfoSeek	$1.0_{0.0}$	$1.0_{0.0}$	$1.0_{0.0}$	$1.0_{0.0}$	$1.0_{0.0}$
	ViQuAE	$1.0_{0.0}$	$1.0_{0.0}$	$1.0_{0.0}$	$1.0_{0.0}$	$1.0_{0.0}$
NIAH	VH-Single	$21.7_{0.9}$	$44.7_{1.3}$	$90.9_{1.9}$	$183.4_{2.9}$	$368.2_{4.3}$
	VH-Multi	$21.7_{0.9}$	$44.8_{1.3}$	$91.0_{1.9}$	$183.4_{2.8}$	$368.2_{4.3}$
	MM-NIAH-Ret (T)	$3.8_{1.4}$	$7.6_{2.0}$	$15.4_{2.7}$	$30.4_{3.8}$	$59.3_{5.5}$
	MM-NIAH-Count (T)	$3.7_{1.4}$	$7.6_{2.0}$	$15.4_{2.7}$	$30.3_{3.9}$	$59.3_{5.7}$
	MM-NIAH-Reason (T)	$3.8_{1.4}$	$7.6_{2.0}$	$15.5_{2.7}$	$30.5_{3.9}$	$59.3_{5.7}$
	MM-NIAH-Ret (I)	$8.1_{1.0}$	$11.8_{1.5}$	$19.3_{2.2}$	$34.2_{3.4}$	$63.8_{5.5}$
	MM-NIAH-Count (I)	$5.7_{1.2}$	$9.3_{1.6}$	$16.9_{2.2}$	$31.7_{3.2}$	$61.5_{5.6}$
	MM-NIAH-Reason (I)	$6.5_{1.0}$	$10.2_{1.4}$	$17.6_{2.1}$	$32.5_{3.4}$	$62.3_{5.7}$
ICL	Stanford Cars	$36.1_{1.3}$	$72.6_{2.3}$	$156.3_{0.9}$	$324.0_{5.3}$	$628.8_{9.0}$
	Food101	$25.0_{0.8}$	$52.0_{0.8}$	$106.1_{1.4}$	$215.5_{2.3}$	$432.5_{0.9}$
	SUN397	$37.0_{2.0}$	$80.1_{3.7}$	$161.3_{4.2}$	$326.8_{2.2}$	$656.6_{8.2}$
	Inat2021	$31.2_{0.6}$	$66.1_{0.8}$	$134.6_{1.4}$	$271.0_{1.6}$	$543.8_{1.8}$
Summarization	GovReport	$2.0_{0.1}$	$6.0_{0.0}$	$12.0_{0.0}$	$25.0_{0.0}$	$50.7_{0.5}$
	Multi-LexSum	$3.0_{0.1}$	$6.0_{0.2}$	$12.1_{0.5}$	$25.2_{0.9}$	$51.3_{1.8}$
DocVQA	MMLongBench-Doc	$3.3_{1.3}$	$6.9_{2.4}$	$13.8_{4.8}$	$28.0_{8.2}$	$56.4_{12.1}$
	LongDocURL	$3.6_{2.1}$	$7.2_{4.5}$	$14.2_{8.1}$	$28.6_{14.4}$	$55.3_{18.3}$
	SlideVQA	$7.5_{0.9}$	$16.2_{2.1}$	$33.2_{2.7}$	$67.2_{3.5}$	$135.2_{5.2}$

B.6. Image Resizing and Statistics

The number of tokens per image is determined by the image size in our benchmark, as we discussed in Section 2.3. In MM-NIAH, we find that many images from the OBELICS dataset are unnecessarily large (up to 8000×6000 pixels) and are not text-rich. Then, we resize those images' longer edge to 1024 pixels while preserving their aspect ratio.

We calculate the average number of images per example in all the datasets and input lengths in Table 3. From the table, we can find that our benchmark covers various text-to-image ratios. For example, VRAG tasks are text-centric and only contain one image per example, while ICL represents image-centric tasks with hundreds of images. MM-NIAH tasks are intermediate and feature both substantial text and multiple images.

B.7. License

All the data collected are based on previously open-sourced datasets, and all licenses are publicly available.

C. Full Model List

We list all 46 models (Hurst et al., 2024; Google, 2025; 2024; Anthropic, 2024; Wang et al., 2024b; Bai et al., 2025; Chen et al., 2024b; Zhu et al., 2025; Lu et al., 2024b; Kamath et al., 2025; Laurençon et al., 2024b;a; Abdin et al., 2024; Abouelenin et al., 2025; Liu et al., 2024b; Agrawal et al., 2024) we evaluated in Table 10. All 46 models have a pixel unshuffle operation to reduce the token counts of images. This is consistent with our token counting methods (Section 2.2). The only exception is Pixtral-12B, but we can resize its image (to 0.5× on each side) to reduce the image tokens. Thus, we can fit Pixtral-12B on our GPU server and avoid extremely long input sequences.

Table 4. GovReport Fluency Evaluation Prompt

Task: GovReport **Metric: Fluency**

Please act as an impartial judge and evaluate the fluency of the provided text. The text should be coherent, non-repetitive, fluent, and grammatically correct.

Below is your grading rubric:
- Score 0 (incoherent, repetitive, or incomplete): Incoherent sentences, repetitive sentences (even if not by exact words), incomplete answers, or gibberish. Note that even if the answer is coherent, if it is repetitive or incomplete, it should be given a score of 0.
- Examples:

- Incomplete: `"Summary:"`
- Incoherent: `"Summary: U.S. agencies engaged export and controls controls controls controls diversion prevent items U.S. activities compliance allies transshipment risk misuse exported misuse misuse illicit illicit against interests or."`
- Repetitive: `"Summary:The audit focused on determining the cost and schedule performance of selected programs. The audit focused on determining the cost and schedule performance of selected programs. The audit focused on determining the cost and schedule performance of selected programs. The audit focused on determining the cost and schedule performance of selected programs."`

- Score 1 (coherent, non-repetitive answer): Coherent, non-repetitive, fluent, grammatically correct answers. If the text is coherent, non-repetitive, and fluent, but the last sentence is truncated, it should still be given a score of 1.
Examples:

- `"Why GAO Did This Study: Tobacco use is the leading cause of preventable death and disease in the United States. In 2009, the Family Smoking Prevention and Tobacco Control Act (Tobacco Control Act) granted FDA, an agency within the Department of Health and Human Services (HHS), authority to regulate tobacco products, including marketing and distribution to youth. The act established CTP, which implements the act by educating the public on the dangers of tobacco use; developing the science needed for tobacco regulation; and developing and enforcing regulations on the manufacture, marketing, and distribution of tobacco products. The act authorized FDA to assess and collect user fees from tobacco manufacturers and importers. The Tobacco Control Act mandated that GAO review the authority and resources provided to FDA for regulating the manufacture, marketing, and distribution of tobacco products. This report examines (1) how FDA spent tobacco user fees for key activities using its authorities granted in the act, and (2) any challenges FDA encountered in using its authorities. GAO analyzed data on tobacco user fees collected and spent on key activities by FDA as of March 31, 2014; reviewed documents related to FDA's key activities, as well as relevant laws, regulations, and guidance; and interviewed CTP, public health, and tobacco industry officials... [about 150 more words]"`

Now, read the provided text, and evaluate the fluency using the rubric. Then output your score in the following json format: `{"fluency": 1}`.
Text: `"{text}"`

Table 5. Multi-LexSum Fluency Evaluation Prompt

Task: Multi-LexSum **Metric: Fluency**

Please act as an impartial judge and evaluate the fluency of the provided text. The text should be coherent, non-repetitive, fluent, and grammatically correct.

Below is your grading rubric:
Score 0 (incoherent, repetitive, or incomplete): Incoherent sentences, repetitive sentences (even if not by exact words), incomplete answers, or gibberish. Note that even if the answer is coherent, if it is repetitive or incomplete, it should be given a score of 0.
- Examples:

- Incomplete: `"Summary:"`
- Incoherent: `"Summary: The plaintiff the the the the able the the the the the the the the the the able the the the the the Ã?\n"`
- Repetitive: `"Summary: The U.S. government brought a criminal case against four defendants. Summary: The U.S. government brought a criminal case against four defendants. Summary: The U.S. government brought a criminal case against four defendants. Summary: The U.S. government brought a criminal case against four defendants."`

Score 1 (coherent, non-repetitive answer): Coherent, non-repetitive, fluent, grammatically correct answers. If the text is coherent, non-repetitive, and fluent, but the last sentence is truncated, it should still be given a score of 1.
- Examples:

- `"This case is about an apprenticeship test that had a disparate impact on Black apprenticeship applicants. The Equal Employment Opportunity Commission (EEOC) filed this lawsuit on December 27, 2004, in U.S. District Court for the Southern District of Ohio."`
- `"The plaintiffs sought declaratory and injunctive relief, as well as attorneys' fees and costs, under the Americans with Disabilities Act, the Rehabilitation Act of 1973, the Social Security Act, and the Nursing Home Reform Act. The case was certified as a class action on behalf of all Medicaid-eligible adults with disabilities in Cook County, Illinois, who are being, or may in the future be, unnecessarily confined to nursing facilities and with appropriate supports and services may be able to live in a community setting. The defendants denied the allegations and argued that the plaintiffs' claims were not typical of the class and that the class definition was too broad. The case is ongoing, with discovery and expert testimony scheduled for the fall of"`

Now, read the provided text, and evaluate the fluency using the rubric. Then output your score in the following json format: `{"fluency": 1}`.

Text: `"{text}"`

Table 6. GovReport Precision Evaluation Prompt

Task: GovReport **Metric: Precision**

Please act as an impartial judge and evaluate the quality of the provided summary of a government report from U.S. Government Accountability Office (GAO). The summary should discuss one or more of the following: why GAO did this study, what GAO found, and what GAO recommends.

Below is your grading rubric:
Precision:

- Evaluate the provided summary by deciding if each sentence in the provided summary is supported by the information provided in the expert summary. A sentence is still supported even if some minor details (e.g., dates, entity names, or locations) are not explicitly mentioned in the expert summary. A sentence is not supported if its major facts are not mentioned, contradicted, or introduce new information not present in the expert summary (e.g., extra analysis or commentary).

- **Score:** the number of sentences in the provided summary that are supported by the expert summary.

- **Examples:** use the following examples to guide your evaluation.

Example 1:

Expert summary: <start of summary>Why GAO Did This Study: The Congressional Budget Office projects that federal deficits will reach $1 trillion in 2020 and average $1.2 trillion per year through 2029, further adding to the more than $16 trillion in current debt held by the public. As a result, Treasury will need to issue a substantial amount of debt to finance government operations and refinance maturing debt. To support its goal to borrow at the lowest cost over time, Treasury must maintain strong demand from a diverse group of investors for Treasury securities. GAO prepared this report as part of continuing efforts to assist Congress in identifying and addressing debt management challenges. This report (1) identifies factors that affect demand for Treasury securities and (2) examines how Treasury monitors and analyzes information about the Treasury market to inform its debt issuance strategy. GAO analyzed data on investor holdings of Treasury securities; surveyed a non-generalizable sample of 109 large domestic institutional investors across 10 sectors (67 responded); reviewed Treasury analysis and market research; and interviewed market participants across sectors, experts on foreign investors, and Treasury officials... [about 300 more words] <end of summary>

Provided summary: <start of summary>The U.S. Government Accountability Office (GAO) conducted a performance audit from June 2018 to December 2019 to assess the management of federal debt by the Department of the Treasury. The audit aimed to evaluate how Treasury manages its debt to finance the federal deficit and refinances maturing debt while minimizing costs. Treasury issues various types of securities, including Treasury bills, notes, bonds, and inflation-protected securities, with maturities ranging from a few weeks to 30 years, to attract a diverse investor base and maintain a healthy secondary market. The audit found that Treasury's regular and predictable framework for issuing securities supports reliable demand, but changes in market conditions and policies pose risks to the liquidity, depth, and safety of Treasury securities. Treasury uses market outreach, auction and market metrics, and analytical models to inform its debt issuance decisions but lacks policies for bilateral market outreach and quality assurance for analytical models. The report recommends Treasury finalize its market outreach policy and establish a quality assurance policy for analytical models to ensure transparency and appropriate documentation. Treasury agreed with the recommendations and plans to implement them.<end of summary>

Reasoning: Sentence 1 is not supported (audit dates and "performance audit" not mentioned). Sentence 2 is supported (aligns with Treasury's goal of borrowing at lowest cost). Sentence 3 is not supported (specific security types and maturity ranges not listed). Sentence 4 is supported (risks to liquidity, depth, safety are mentioned). Sentence 5 is supported (mentions the three inputs and missing policies). Sentence 6 is supported (matches the recommendations). Sentence 7 is supported (Treasury agreed). Therefore, the precision score is 5.

Output: `{"precision": 5, "sentence_count": 7}`

Example 2: ...

Now, read the provided summary and expert summary, and evaluate the summary using the rubric. First, think step-by-step and provide your reasoning and assessment on the answer. Please keep your response concise and limited to a single paragraph. Then output your score in the following json format: `{"precision": 7, "sentence_count": 20}`.
Expert summary: <start of summary>`{expert_summary}`<end of summary>
Provided summary: <start of summary>`{summary}`<end of summary>

Table 7. Multi-LexSum Precision Evaluation Prompt

Task: Multi-LexSum **Metric: Precision**

Please act as an impartial judge and evaluate the quality of the provided summary of a civil lawsuit. The summary is based on a set of legal documents, and it should contain a short description of the background, the parties involved, and the outcomes of the case.

Below is your grading rubric:
Precision:

- Evaluate the provided summary by deciding if each sentence in the provided summary is supported by the information provided in the expert summary. A sentence is considered supported if its major facts align with the information in the expert summary. A sentence is still considered supported even if some of its minor details, such as dates, entity names, or the names of laws and previous court cases, are not explicitly mentioned in the expert summary. A sentence is not supported if its major facts are not mentioned or contradicted in the expert summary.

- Score: the number of sentences in the provided summary that are supported by the expert summary.

- Examples: use the following examples to guide your evaluation.

Example 1:

Expert summary: "This lawsuit, brought in the the U.S. District Court for the Central District of California, was filed on June 3, 2020. The plaintiffs were represented by attorneys from the ACLU of Southern California. This lawsuit followed nation-wide protests that occurred in response to the killing of George Floyd by a police officer in Minneapolis. While most protests were peaceful, some ended in violence, property destruction, rioting, and looting. Many cities, including Los Angeles and San Bernardino, issued curfews in an attempt to quell these riots. This action challenged these curfews as violations of free speech and assembly, free movement, due process, and challenged the San Bernardino curfew as a violation of the establishment clause (the San Bernardino curfew included a provision that exempted attendants of religious meetings from the curfew.)... [about 100 more words]."

Provided summary: "In June 2020, Black Lives Matter - Los Angeles and several individuals filed a lawsuit in the U.S. District Court for the Central District of California against Los Angeles Mayor Eric Garcetti, other city officials, and the City of San Bernardino, challenging the constitutionality of curfew orders imposed during protests against police violence. The plaintiffs, represented by the ACLU of Southern California, argued that the curfews violated their First Amendment rights to free speech and assembly, as well as their freedom of movement, by suppressing political protests and other activities. The lawsuit also claimed that the curfews were not narrowly tailored to address any emergency and lacked sufficient notice. However, the plaintiffs voluntarily dismissed the case in July 2020 after the defendants lifted the curfew orders and did not reinstate them in the following weeks."

Reasoning: The first sentence in the provided summary is well supported by the expert summary even though some entity names are not explicitly mentioned. The second sentence is also well supported by the expert summary, as it mentions the ACLU of Southern California and the First Amendment rights. The third sentence is not supported by the expert summary, as it does not mention the lack of narrow tailoring or sufficient notice. The fourth sentence is well supported by the expert summary, as it mentions the voluntary dismissal of the case in July 2020. Therefore, the precision score is 3.

Output: `{"precision": 3, "sentence_count": 4}`

Example 2: ...

Now, read the provided summary and expert summary, and evaluate the summary using the rubric. First, think step-by-step and provide your reasoning and assessment on the answer. Please keep your response concise and limited to a single paragraph. Then output your score in the following json format: `{"precision": 2, "sentence_count": 6}`.

Expert summary: `"{expert_summary}"`

Provided summary: `"{summary}"`

Table 8. GovReport Recall Evaluation Prompt

Task: GovReport **Metric: Recall**

Please act as an impartial judge and evaluate the quality of the provided summary of a government report from U.S. Government Accountability Office (GAO). The summary should discuss one or more of the following: why GAO did this study, what GAO found, and what GAO recommends. The text should contain all the major points in the expert-written summary, which are given to you.

Below is your grading rubric:
Recall:

- Evaluate the provided summary by deciding if each of the key points is present in the provided summary. A key point is considered present if its factual information is mostly-supported by the provided summary. If a key point contains multiple facts, it is considered supported if most of the facts are present.
- Score: the number of key points mostly-supported by the provided summary.
- Examples: Use the following example to guide your evaluation.

Example 1:

Key points:
1. The Future Combat System (FCS) program is the centerpiece of the Army's effort to transition to a lighter combat force.
2. The FCS program is the centerpiece of the Army's effort to transition to a more agile combat force.
3. The FCS program is the centerpiece of the Army's effort to transition to a more capable combat force.
4. By law, GAO is to report annually on the FCS program.
5. Law requires the Department of Defense (DOD) to hold a milestone review of the FCS program.
6. This milestone review is now planned for 2009.
7. This report addresses (1) what knowledge will likely be available in key areas for the review.
8. This report addresses (2) the challenges that lie ahead following the review.
9. To meet these objectives, GAO reviewed key documents and performed analysis.
10. GAO attended demonstrations and design reviews to meet these objectives.
11. GAO interviewed DOD officials to meet these objectives.
12. The Army will be challenged to demonstrate the knowledge needed to warrant an unqualified commitment to the FCS program.
13. This challenge will occur at the 2009 milestone review.
14. The Army has made progress.
15. Knowledge deficiencies remain in key areas. [31 more points]

Summary: <start of summary>Why GAO Did This Study: The Future Combat System (FCS) program is the centerpiece of the Army's effort to transition to a lighter combat force. By law, GAO is to report annually on the FCS program. This report addresses (1) what knowledge will likely be available in key areas for the review, and (2) the challenges that lie ahead following the review. To meet these objectives, GAO reviewed key documents and interviewed DOD officials.

What GAO Found: The Army will be challenged to demonstrate the knowledge needed to warrant an unqualified commitment to the FCS program. While the Army has made progress, knowledge deficiencies remain in key areas. Specifically, all critical technologies are not currently at a minimum acceptable level of maturity. Actual demonstrations of FCS hardware and software have been limited. Network performance is also largely unproven. DOD could have at least three programmatic directions to consider for shaping investments in future capabilities. [106 more words]<end of summary>

Reasoning: The summary covers: FCS as Army's transition centerpiece (point 1), GAO's reporting requirement (point 4), report objectives (points 7, 8), GAO's methods (points 9, 11), Army's challenges (point 12), progress and deficiencies (points 14, 15), technology issues (points 16, 19, 21), three programmatic directions (points 27, 29, 31, 33, 34, 36, 38, 41-43). It omits: "more agile/capable" (points 2, 3), 2009 milestone review (points 5, 6, 13), demonstrations attendance (point 10), design requirements issues (points 17, 18), small-scale concepts (point 20), program immaturity explanation (points 22, 23), funding competition (points 24-26), challenges after review (point 28), production before design demonstration (points 30, 32), technology testing issues (point 35), $50 billion funding (point 37), surrogate systems (points 39, 40), and increment justification (points 44-46). The summary supports 22 key points.

Output: `{"supported_key_points": [1, 4, 7, 8, 9, 11, 12, 14, 15, 16, 19, 21, 27, 29, 31, 33, 34, 36, 38, 41, 42, 43], "recall": 22}`

Now, read the provided summary and key points, and evaluate the summary using the rubric. First, think step-by-step and provide your reasoning and assessment on the answer. Please keep your response concise and limited to a single paragraph. Then output your score in the following json format: `{"supported_key_points": [1, 4, 7, 8, 9, 11, 12, 14, 15, 16, 19, 21, 27, 29, 31, 33, 34, 36, 38, 41, 42, 43], "recall": 22}`, where "supported_key_points" contains the key points that are present in the summary and "recall" is the total number of key points present in the summary.

Key points:
`{keypoints}`

Summary: <start of summary>`{summary}`<end of summary>

Table 9. Multi-LexSum Recall Evaluation Prompt

Task: Multi-LexSum **Metric: Recall**

Please act as an impartial judge and evaluate the quality of the provided summary of a civil lawsuit. The summary is based on a set of legal documents, and it should contain a short description of the background, the parties involved, and the outcomes of the case. The text should contain all the major points in the expert-written summary, which are given to you.

Below is your grading rubric:
Recall:

- Evaluate the provided summary by deciding if each of the key points is present in the provided summary. A key point is considered present if its factual information is well-supported by the provided summary.
- Score: the number of key points present in the provided summary.
- Examples: use the following examples to guide your evaluation.

Example 1:

Key points:
1. The case challenged curfews in Los Angeles and San Bernardino, California.
2. The curfews were issued in response to the nationwide protests following the police killing of George Floyd in Minneapolis.
3. The complaint argued that the curfews violated free speech, free assembly, free movement, and Due Process.
4. The complaint also argued that the San Bernardino curfew violated the Establishment Clause.
5. The complaint sought injunctive and declaratory relief.
6. The plaintiffs voluntarily dismissed the case on July 7, 2020.
7. The dismissal occurred because the city had rescinded the curfews and not attempted to reinstate them.

Summary: In June 2020, Black Lives Matter - Los Angeles and several individuals filed a lawsuit in the U.S. District Court for the Central District of California against Los Angeles Mayor Eric Garcetti, other city officials, and the City of San Bernardino, challenging the constitutionality of curfew orders imposed during protests against police violence. The plaintiffs, represented by the ACLU of Southern California, argued that the curfews violated their First Amendment rights to free speech and assembly, as well as their freedom of movement, by suppressing political protests and other activities. The lawsuit also claimed that the curfews were not narrowly tailored to address any emergency and lacked sufficient notice. However, the plaintiffs voluntarily dismissed the case in July 2020 after the defendants lifted the curfew orders and did not reinstate them in the following weeks.

Reasoning: The summary states that the plaintiffs challenged the constitutionality of curfew orders against Los Angeles and San Bernadino, so key point 1 is present. The summary does not mention that the curfew orders were issued in response to the nationwide protest that resulted from the police killing of George Floyd in Minneapolis, so key point 2 is missing. The summary does mention that the complaint argued that the curfews violated the First Amendment rights to free speech and assembly, so key point 3 is present. The summary does not mention that the complaint argued that the San Bernardino curfew violated the Establishment Clause, so key point 4 is missing. The summary does not mention that the complaint sought injunctive and declaratory relief, so key point 5 is missing. The summary mentions that the plaintiffs voluntarily dismissed the case in July 2020 after the defendants lifted the curfew orders and did not reinstate them in the following weeks, so key point 6 and 7 are present. Finally, key points 1, 3, 6, and 7 are present in the summary, so the recall score is 4.

Output: `{"recall": 4}`

Example 2: ...

Now, read the provided summary and key points, and evaluate the summary using the rubric. First, think step-by-step and provide your reasoning and assessment on the answer. Please keep your response concise and limited to a single paragraph. Then output your score in the following json format: `{"recall": 2}`.

Key points: `{keypoints}`

Summary: `"{summary}"`

Table 10. Length means the training length (default) or claimed context window (denoted by †). All LCVLMs are instruction-tuned. "Image Porc." stands for Image Processing, which is mainly Dynamic Resolution ViT (Wang et al., 2024b) or Dynamic Tiling (Wu et al., 2024c). The positional embedding includes RoPE (Su et al., 2024), M-RoPE (Wang et al., 2024b), Linear Scaling (Chen et al., 2023; kaiokendev, 2023) LongRoPE (Ding et al., 2024), Dynamic-NTK, NTK-by-parts or YaRN (Peng et al.).

Name	Length	Image Proc.	Positional Emb.	# Params
Proprietary (No model details except the claimed context lengths.				
gpt-4o-2024-11-20	128,000†	?	?	?
claude-3-7-sonnet-20250219	200,000†	?	?	?
gemini-2.0-flash-001	1,048,576†	?	?	?
gemini-2.0-flash-thinking-exp-01-21	1,048,576†	?	?	?
gemini-2.5-flash-preview-04-17	1,048,576†	?	?	?
gemini-2.5-pro-preview-03-25	1,048,576†	?	?	?
Qwen2-VL & Qwen2.5-VL				
Qwen2-VL-2B-Instruct	32,768	Dynamic-Resolution ViT	M-RoPE	2B
Qwen2-VL-7B-Instruct	32,768	Dynamic-Resolution ViT	M-RoPE	7B
Qwen2-VL-72B-Instruct-AWQ	32,768	Dynamic-Resolution ViT	M-RoPE	72B
Qwen2.5-VL-3B-Instruct	32,768	Dynamic-Resolution ViT	M-RoPE	3B
Qwen2.5-VL-7B-Instruct	32,768	Dynamic-Resolution ViT	M-RoPE	7B
Qwen2.5-VL-32B-Instruct	32,768	Dynamic-Resolution ViT	M-RoPE	32B
Qwen2.5-VL-72B-Instruct-AWQ	32,768	Dynamic-Resolution ViT	M-RoPE	72B
InternVL2, InternVL2.5, & InternVL3				
InternVL2-1B	8,192	Dynamic Tiling	RoPE	0.9B
InternVL2-2B	8,192	Dynamic Tiling	Dynamic-NTK	2.21B
InternVL2-4B	8,192	Dynamic Tiling	LongRoPE	4.15B
InternVL2-8B	8,192	Dynamic Tiling	Dynamic-NTK	8.08B
InternVL2_5-1B	16,348	Dynamic Tiling	RoPE	0.9B
InternVL2_5-2B	16,348	Dynamic Tiling	Dynamic-NTK	2.2B
InternVL2_5-4B	16,348	Dynamic Tiling	RoPE	4.2B
InternVL2_5-8B	16,348	Dynamic Tiling	Dynamic-NTK	8.1B
InternVL2_5-26B	16,348	Dynamic Tiling	Dynamic-NTK	25.5B
InternVL3-1B	32,768	Dynamic Tiling	Dynamic-NTK	0.9B
InternVL3-2B	32,768	Dynamic Tiling	Dynamic-NTK	1.9B
InternVL3-8B	32,768	Dynamic Tiling	Dynamic-NTK	8.1B
InternVL3-14B	32,768	Dynamic Tiling	Dynamic-NTK	15.1B
InternVL3-38B	32,768	Dynamic Tiling	Dynamic-NTK	38.4B
Ovis2				
Ovis2-1B	32,768	Dynamic Tiling	RoPE	1B
Ovis2-2B	32,768	Dynamic Tiling	RoPE	2B
Ovis2-4B	32,768	Dynamic Tiling	RoPE	4B
Ovis2-8B	32,768	Dynamic Tiling	RoPE	8B
Ovis2-16B	32,768	Dynamic Tiling	RoPE	16B
Ovis2-34B	32,768	Dynamic Tiling	RoPE	34B
Gemma-3				
gemma-3-4b-it	131,072†	Dynamic Tiling	Linear Scaling	4B
gemma-3-12b-it	131,072†	Dynamic Tiling	Linear Scaling	12B
gemma-3-27b-it	131,072†	Dynamic Tiling	Linear Scaling	27B
Idefics2				
idefics2-8b	8,192	Dynamic-Resolution ViT	RoPE	8B
idefics2-8b-C (chatty)	8,192	Dynamic-Resolution ViT	RoPE	8B
Mantis-8B-Idefics2	8,192	Dynamic-Resolution ViT	RoPE	8B
Idefics3				
Idefics3-8B-Llama3	10,240	Dynamic Tiling	NTK-by-parts	8B
Phi-based				
Phi-3-vision-128k-instruct	131,072	Dynamic Tiling	LongRoPE	4.2B
Phi-3.5-vision-instruct	131,072	Dynamic Tiling	LongRoPE	4.2B
Phi-4-multimodal-instruct	131,072	Dynamic Tiling	LongRoPE	5.6B
NVILA				
NVILA-Lite-2B-hf-preview	32,768	Dynamic Tiling	RoPE	2B
NVILA-Lite-8B-hf-preview	32,768	Dynamic Tiling	RoPE	8B
Pixtral				
pixtral-12b	131,072	Dynamic-Resolution ViT	RoPE	12B

Table 11. The number of tokens produced by models *without pixel unshuffle* for the inputs (64K and 128K tokens) of the ICL and Visual Haystack (VH) datasets. The numbers are in thousands (K). These models cannot process long sequences of images efficiently.

Model	ICL		VH	
	64K	128K	64K	128K
Llama-3.2-11B	1,821K	3,626K	1,174K	2,358K
Llava-onevision-7b	558K	1,142K	496K	996K
mPLUG-Owl3-7B	1,398K	2,786K	930K	1,870K

C.1. LVLMs Beyond our Evaluation

Token Efficiency. Beyond the models in our evaluation, there are also a lot of excellent models, such as Llava-onevision (Li et al., 2024a), Llama3.2 (Dubey et al., 2024), mPLUG-Owl3 (Ye et al., 2024). However, we find that those models don't have pixel unshuffle operations. While Pixtral-12B's ViT can take images with dynamic resolution, the ViTs of these three models use dynamic tiling, such as 560×560 tiles for Llama3.2-11B. Thus, unlike Pixtral-12B, we cannot reduce the number of image tokens by simply resizing the input image, since these models do not accept images smaller than the predefined tile size. As shown in Table 11, these models cannot efficiently process high-resolution images as a whole, generating a number of tokens that is at least ten times the pre-defined input length (64K or 128K). *This causes extremely long input sequences, and we cannot fit them on our GPU server.*

Challenge for Model Integration. DeepSeek-VL (Lu et al., 2024a), DeepSeek-VL2 (Wu et al., 2024c), and LongLlava (Wang et al., 2024d) are also excellent, but they present additional challenges. These models do not provide a standard API in the HuggingFace Transformers framework (Wolf et al., 2019), which we used for inference. Therefore, these models cannot be loaded directly via Transformes, and we need to develop them based on their GitHub repositories. As a result, integrating these models requires a *prohibitive* amount of engineering effort to adapt their codebases, making it impractical within our current scope. We leave their integration for future work.

D. Experimental Setup

As previously described, we evaluate all 46 models across different input lengths $L \in \{8192, 16384, 32768, 65536, 131072\}$. We evaluate the proprietary models using their API. The specific versions we used are as follows:

- GPT-4o: `gpt-4o-2024-11-20`
- Claude-3.7-Sonnet: `claude-3-7-sonnet-20250219`
- Gemini-2.0-Flash: `gemini-2.0-flash-001`
- Gemini-2.0-Flash-T: `gemini-2.0-flash-thinking-exp-01-21`
- Gemini-2.5-Flash: `gemini-2.5-flash-preview-04-17`
- Gemini-2.5-Pro: `gemini-2.5-pro-preview-03-25`

For all open-source models, we evaluate them on an 8×A100 (80GB) GPU server. We use the HuggingFace Transformers framework (Wolf et al., 2019) to deploy models and generate outputs. Since all models are instruction-tuned, we apply the chat templates to all datasets. We load models in BF16 with FlashAttention2 (Dao) for faster inference. The largest open-source models tested in our work have 72B parameters. Our computational resources are limited to 8×A100 GPUs; thereby, we cannot evaluate models with over 100B parameters, such as Llama4 (Meta, 2025), at 128K tokens.

We sampled 100 examples from each dataset to evaluate models. This amount actually results in 600 examples for single-needle tasks: ViQuAE, VH-Single, and MM-NIAH-Ret, and 300 examples for multi-needle tasks: InfoSeek, VH-Multi, and MM-NIAH-Count. This is because we test 6 different depths (i.e., [0, 0.2, 0.4, 0.6, 0.8, 1.0]) for single-needle examples and 3 different permutations for multi-needle examples to mitigate the positional bias. Note that in MM-NIAH, we sample the text-needle and image-needle examples evenly, with 50 of each type. The MM-NIAH-Reason is more complex since

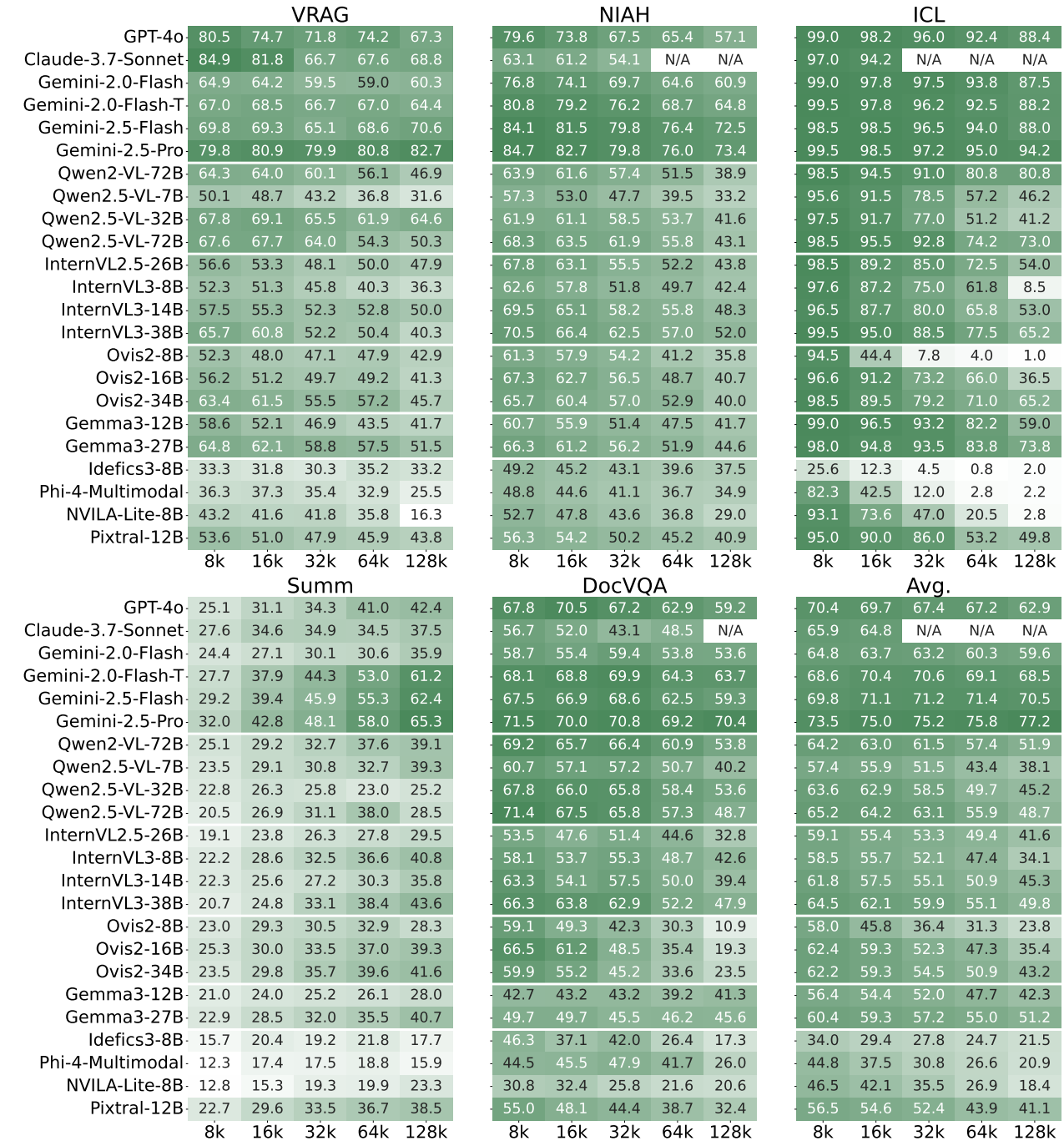


Figure 4. Performance on MMLongBench. We report results for selected frontier models, and the full results of all models are provided in Figure 22. Note that Claude-3.7-Sonnet supports at most 100 images, and we mark the results as N/A for cases with more images (More in Appendix E.8)

image-needle (I) examples have a single needle, while text-needle (T) ones have multiple needles. There are 300 examples for MM-NIAH-Reason (I) (50 $\times$ 6 depths) and 150 examples for MM-NIAH-Reason (T) (50 $\times$ 3 permutations). Due to this depth imbalance in MM-NIAH-Reason, we compute the average score for each subset separately and report their mean as the final result. Together, we evaluate each model on 4,050 examples across five different input lengths, resulting in a total of 20,250 examples.

E. Evaluation and Analysis

With broad task coverage, unified token counting, and standardized input length, we are now able to thoroughly examine LCVLMs' long-context ability across multiple dimensions. In total, we evaluate 46 LCVLMs on MMLongBench. To the best of our knowledge, our evaluation provides the most thorough and controlled comparison of the vision-language long-context ability on broad real-world applications. These models include closed-source models GPT-4o (Hurst et al., 2024), Claude-3.7 (Anthropic, 2024), and Gemini 2 and 2.5 (Google, 2024; 2025), as well as open-source model families, such as Qwen2.5-VL (Bai et al., 2025), InternVL3 (Zhu et al., 2025), and Gemma3 (Kamath et al., 2025). We also consider position extrapolation methods, such as YaRN (Peng et al.) and V2PE (Ge et al., 2024) (See Appendix E.9). The full list of evaluated models is provided in Table 10. Following existing works (Yen et al., 2024), we use greedy decoding for all models for consistency and randomly sample 100 examples from each dataset. More details are in Appendix D.

E.1. Evaluation on MMLongBench across Tasks and Context Lengths

We present the performance of selected frontier LCVLMs in Figure 4, and the full results of all 46 models are reported in Figure 22. We analyze model performance from multiple perspectives and summarize our main findings as follows:

All models struggle, but closed-source models perform better. Here, we consider the performance at the longest input length of 128K tokens. In general, we observe that all models struggle on our vision-language long-context tasks. For example, even GPT-4o only achieves 62.9 on average, while open-source models perform even worse. We find that Gemini-2.5-Pro stands out as the strongest LCVLM. Other than ICL, Gemini-2.5-Pro outperforms open-source models by about 20 absolute points. On ICL, although the gap is relatively smaller, due to the strong performance of Qwen2-VL-72B, there is still a difference of about 14 points. While the other closed-source models continue to surpass open-source models, the margin is often under 10 points. Further, Ovis2-34B achieves a score of 41.6 on summarization, similar to GPT-4o (42.4). Qwen2.5-VL-32B achieves a SubEM score of 64.6 on VRAG, even better than Gemini-2.0-Flash. These findings show that while current closed-source models generally perform better, open-source ones are also competitive.

Models can generalize to longer context lengths. Another interesting observation is that some models can generalize to longer context lengths than they are officially designed for. For example, although the context window for Qwen2-VL-72B during training is only 32K tokens, the model can generalize to a 128K input length and still achieve an average score of 51.9. We also observed similar effects on other models, such as Ovis2-34B and InternVL2.5-26B. This phenomenon is likely because the underlying LLMs of those LCVLMs have been trained with longer context windows (Bai et al., 2025). We leave further investigation to future work.

Reasoning can improve multimodal long-context ability. We include Gemini-2.0-Flash-T in our evaluation, which is the thinking variant of Gemini-2.0-Flash. From the results, we observe that the reasoning ability can consistently improve the Gemini-2.0-Flash on all tasks. While the changes for VRAG, Recall, and ICL are modest, summarization and DocVQA exhibit marked improvements of 25.3% and 10.1%, respectively. Then, Gemini-2.5 models exhibit even stronger performance, which are natively designed as thinking models.

Different models exhibit different strengths. Generally, we find that model performance varies considerably across different tasks. For instance, Qwen2.5-VL-32B outperforms InternVL3-38B on VRAG, but underperforms on NIAH. Similarly, Ovis2-34B excels at summarization but struggles on DocVQA. These findings further support the necessity of a comprehensive benchmark covering diverse downstream tasks.

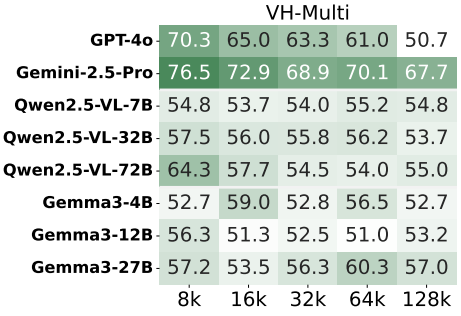


Figure 5. Model performance on VH-Multi dataset. Random guess yields 50% accuracy, highlighting its difficulty.

E.2. Can Needle-in-a-Haystack Task Reflect LCVLM's Overall Long-context Ability?

The needle-in-a-haystack (NIAH) task has been primarily used to evaluate LCVLMs' long-context abilities. However, it remains unclear whether strong performance on NIAH reliably reflects overall long-context capability on diverse tasks. In this section, we first analyze the difficulty of existing NIAH benchmarks and find that current NIAH tasks are challenging, resulting in limited differentiation between models. Further, we compute Spearman's rank correlation (ρ) between NIAH performance and that on other tasks. Our results show that none of these NIAH tasks consistently correlates with performance

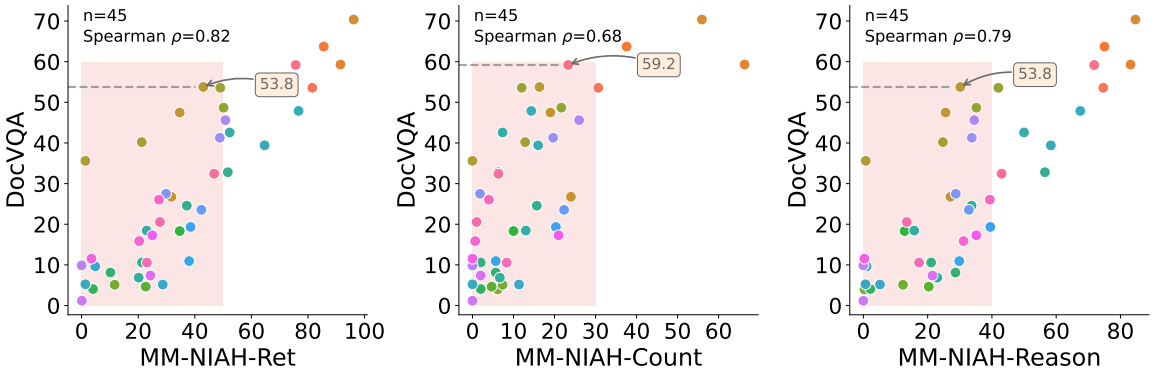


Figure 6. Distribution of long-document VQA (DocVQA) with respect to performance on MM-NIAH variants. We find that the models are concentrated in the coral-shaded areas.

across diverse, practical scenarios.

Text-image interleaved NIAH tasks are challenging. In Figure 5, we find that even state-of-the-art models like GPT-4o and Gemini-2.5 struggle to surpass 80% accuracy on VH-Multi when the context length is just 8K tokens (approximately 22 images). Most models are just slightly better than a random guess (50%). This demonstrates that locating objects in a large set of images is still hugely challenging for current LCVLMs. See more discussion in Appendix E.5. Then, we plot the performance of different models on the retrieval, counting, and reasoning tasks of MM-NIAH against their performance on long-document VQA in Figure 6. We find that most models achieve low performance on the counting and reasoning tasks, with scores below 30 and 40, respectively. The difficulty of the tasks and low performance result in poor separability between models. While the retrieval is an easier task, it still does not align well with DocVQA tasks. In short, we find that *both VH and MM-NIAH present significant challenges to current LCVLMs*, thus showing limited differentiation between models and weak alignment with other tasks.

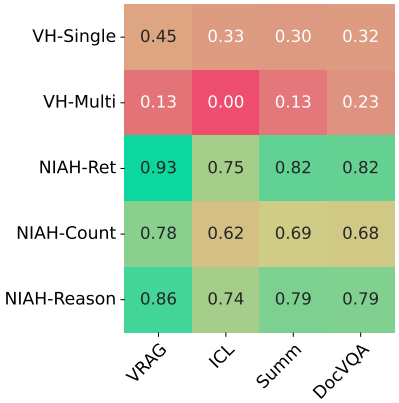


Figure 7. Spearman's ρ across all 46 models at 128K tokens.

NIAH tasks fail to reflect overall long-context abilities. As shown in Figure 7, none of the NIAH tasks exhibit strong correlation with the broader set of long-context tasks. This suggests that performance on NIAH tasks may not be a reliable indicator of general long-context capabilities. In particular, Visual Haystack (VH) tasks show especially low correlations due to their high difficulty, as discussed above, which results in limited ability to distinguish between models. In MM-NIAH, counting and reasoning tasks show weak correlations with several downstream tasks, with coefficients below 0.8. The retrieval task also shows weak alignment with ICL performance. Interestingly, simpler tasks – like retrieval with a single needle in unrelated essays – tend to correlate better with diverse task categories, which is consistent with our findings in Figure 6. We further examine the differences between text-based and image-based needles in Appendix E.6.

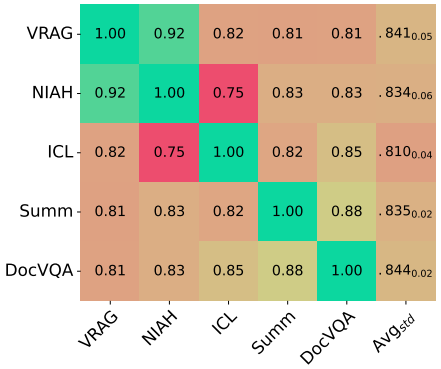


Figure 8. Spearman's ρ between all categories with L =128K. For each category, the *Avg* excludes the correlation with itself.

E.3. Cross-Category Correlations Identify Long-Document VQA as a Reliable Proxy

We perform a cross-category correlation analysis of model performance. We find that different categories do not consistently show strong correlation ($<$ 0.85) with each other, as shown in Figure 8. Specifically, VRAG and NIAH closely correlate because retrieval is the central capability of both tasks. A further investigation shows that VRAG achieves its highest correlation (of 0.93) with the retrieval task (MM-NIAH-Ret) in Figure 14, reinforcing the shared emphasis on retrieval. Meanwhile, summarization and long-document VQA show a high correlation of 0.88, likely due to their shared input format – image-formatted PDF documents. This suggests that image types affect category correlations. In contrast, ICL tasks show relatively

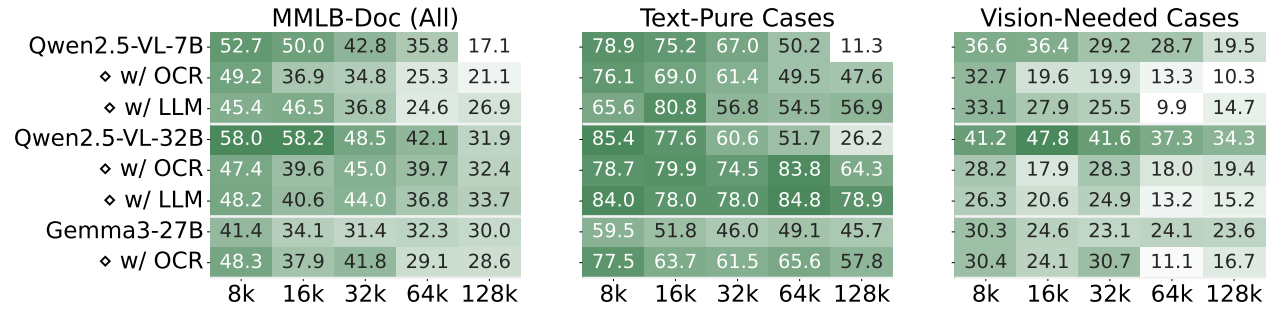


Figure 9. Error Analysis on MMLongBench-Doc. Instead of PDF-formatted documents, we feed OCR-extracted plain text to LCVLMs (◇ w/ OCR) and also test corresponding LLMs: Qwen2.5-7B and Qwen2.5-32B (◇ w/ LLM). We also show scores on examples with different answer sources.

weak correlations with other categories. The ICL tasks evaluate models' ability to induce new classification rules from numerous exemplars, a skill orthogonal to recalling facts in long contexts. This further demonstrates that model developers should consider various long-context skills to draw a more holistic picture of LCVLMs. See Appendix E.7 for detailed dataset-level correlations and additional category-wise insights.

Long-document VQA as a reliable proxy for long-context capabilities. As shown in Figure 8, long-document VQA achieves the highest average correlation with other categories, indicating that it is more aligned with the broader range of long-context tasks. For example, questions from LongDocURL (Deng et al., 2024) cover not only simple retrieval but also complex understanding and reasoning. Meanwhile, long-document VQA exhibits the smallest standard deviation, showing that it is also stable and balanced. Taken together, these findings suggest long-document VQA is a more representative and reliable proxy than the commonly adopted NIAH for reflecting overall system performance, allowing model developers to iterate more rapidly without the overhead of full-scale evaluation.

E.4. Error Analysis

Our evaluation shows that current LCVLMs have significant room for improvement. To better understand their limitations, we analyze model predictions in detail.

In Figure 9, we show the performance of using another pipeline for DocVQA on MMLongBench-Doc. Here, we convert PDF-formatted documents to plain text with OCR (◇ w/ OCR) and feed them to LCVLMs. There is no clear winner between the PDF-formatted and OCR-extracted pipelines across all models. While Qwen2.5-VL models perform better with imaged-formatted PDF documents in most cases, Gemma3-27B prefers plain text for shorter input lengths ($\leq$ 32K). Furthermore, we performed a fine-grained analysis by categorizing examples according to their answer sources into two groups: *text-pure* and *vision-needed*. As expected, using PDF documents leads to higher scores in vision-needed cases, whereas plain text yields better performance in text-pure cases, especially with longer inputs (64K and 128K). This suggests that *OCR capability remains a bottleneck for current LCVLMs* when handling long-context inputs. Future work could explore combining both pipelines to further enhance performance. Meanwhile, when using OCR-extracted text, replacing LCVLMs with the corresponding LLMs, Qwen2.5-7B and Qwen2.5-32B (◇ w/ LLM), yields better results in text-pure cases of DocVQA.

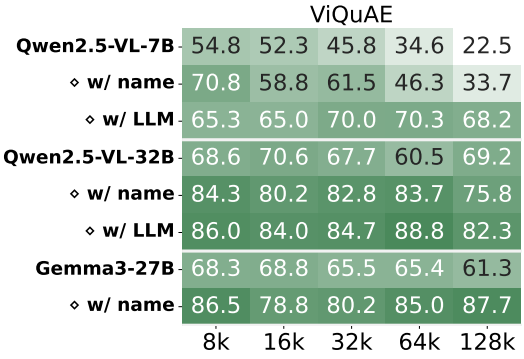


Figure 10. Error analysis on ViQuAE. We replace the image with its original entity name (◇ w/ name) and also test text-only counterparts: Qwen2.5-7B and Qwen2.5-32B (◇ w/ LLM).

We also examine the sources of errors in the VRAG category in Figure 10. Since ViQuAE is built on TriviaQA (Joshi et al., 2017), we replace all images in ViQuAE questions with their corresponding entity names and feed those text-only questions into LCVLMs. All models show varying degrees of improvement, with Gemma3-27B achieving the largest gain of 26.4 points (at 128K), suggesting that *a bottleneck of LCVLMs lies in cross-modality information retrieval.* Besides, providing entity names as input to corresponding LLMs improves model performance. These results illustrate a common trade-off between multimodal and text-only long-context abilities during the training of LCVLMs.

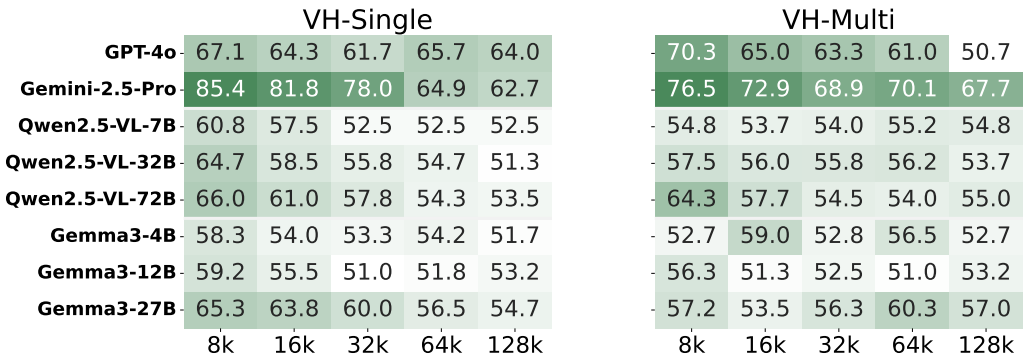


Figure 11. The performance of selected models on VH-Single and VH-Multi. The accuracy of a random guess is 50%. We find that these two tasks are very challenging.

E.5. The Difficulty of Visual Haystack

We discussed the difficulty of Visual Haystack in **??**. As shown in Figure 5, current LCVLMs achieve the performance only slightly higher than random guessing on VH-Multi. Here, in Figure 11, we present the performance of the selected models on both VH-Single and VH-Multi, providing a complete view. We find that models also perform poorly on VH-Single.

Task Correctness. We manually checked a number of examples from the Visual Haystack dataset and didn't find any errors in the task labels. As shown in Figure 11, Gemini-2.5-Pro achieves an accuracy of 85.4 on VH-Single, and GPT-4o achieves an accuracy of 70.3 on VH-Multi. These results are higher than a random guess (50%), which demonstrates the correctness of the dataset labels and our implementation.

E.6. Correlation between NIAH and Various Downstream Tasks

We discussed the correlation between NIAH tasks and various downstream applications in Appendix E.2. Here, we provide a detailed version of the task correlation in Figure 12. For the three tasks in MM-NIAH, we also report the correlations on the subsets containing only text-needle or only image-needle examples. We can find that subsets of image-needle examples correlate less with various downstream tasks compared to text-needle examples, especially MM-NIAH-Count (I) and MM-NIAH-Reason (I)

We further show the performance of the selected models on the text-needle and image-needle subsets in Figure 13. From it, we observe that models exhibit weak performance on MM-NIAH-Count (I) and MM-NIAH-Reason (I). This challenging nature leads to a low degree of separability between different models.

E.7. Correlation between Datasets

We plot the correlation between all MMLongBench datasets and category averages in Figure 14. Generally, the datasets in each category strongly correlate with each other. The VH-Single and VH-Multi are exceptions, due to their high difficulty. Also, MM-NIAH-Count exhibits relatively weak correlations with MM-NIAH-Ret and MM-NIAH-Reason, suggesting that counting is a different skill from retrieving needles (key information) and subsequently reasoning over them.

E.8. Performance of Claude

At the time of our evaluation, the Claude 3 family of models can take up to 100 images[5] in a single request. However, a few datasets, such as Food101, VH-Single, or SlildeVQA, contain hundreds of images at input lengths of 64K and 128K tokens. As a result, it is impossible to process all images in a single pass through the model. For each input length, if one or more datasets within a category contain samples with more than 100 images, we exclude that category from evaluation at that input length. We provide the statistics about the average image number per example in each dataset at all five input lengths in Table 3 and Appendix B.6. When testing `Claude-3.7-Sonnet`, we mark those untestable cases as "N/A" in

[5] https://docs.anthropic.com/en/docs/build-with-claude/vision#basics-and-limits

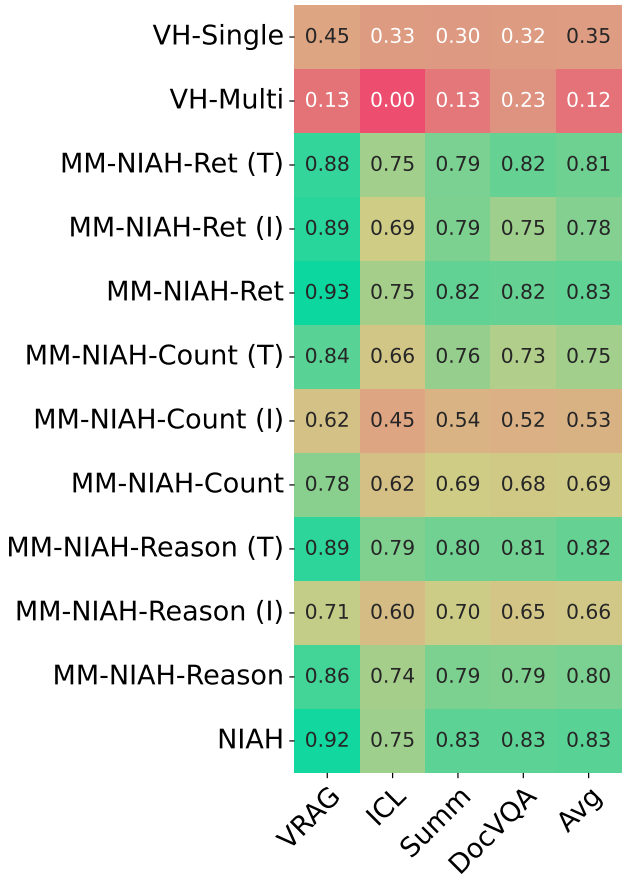


Figure 12. Spearman's correlation at 128K input length, calculated across 46 LCVLMs, between all NIAH and other downstream tasks.

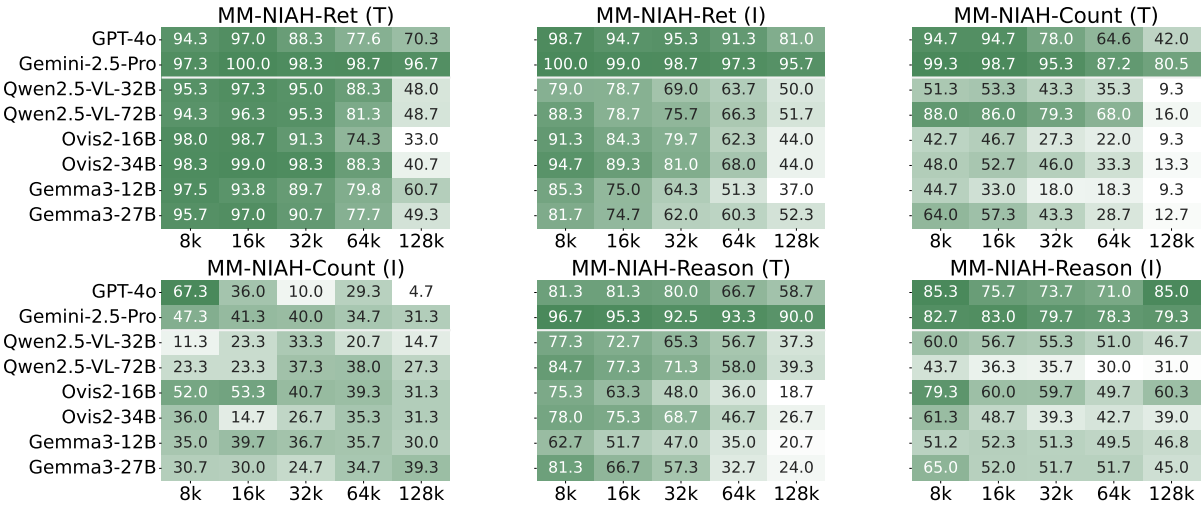


Figure 13. Results of selected models on the subsets of MM-NIAH containing only text-needle or image-needle examples. (T) and (I) represent text-needle and image-needle examples, respectively.

our results to distinguish them from genuine model failures (which receive a score of 0).

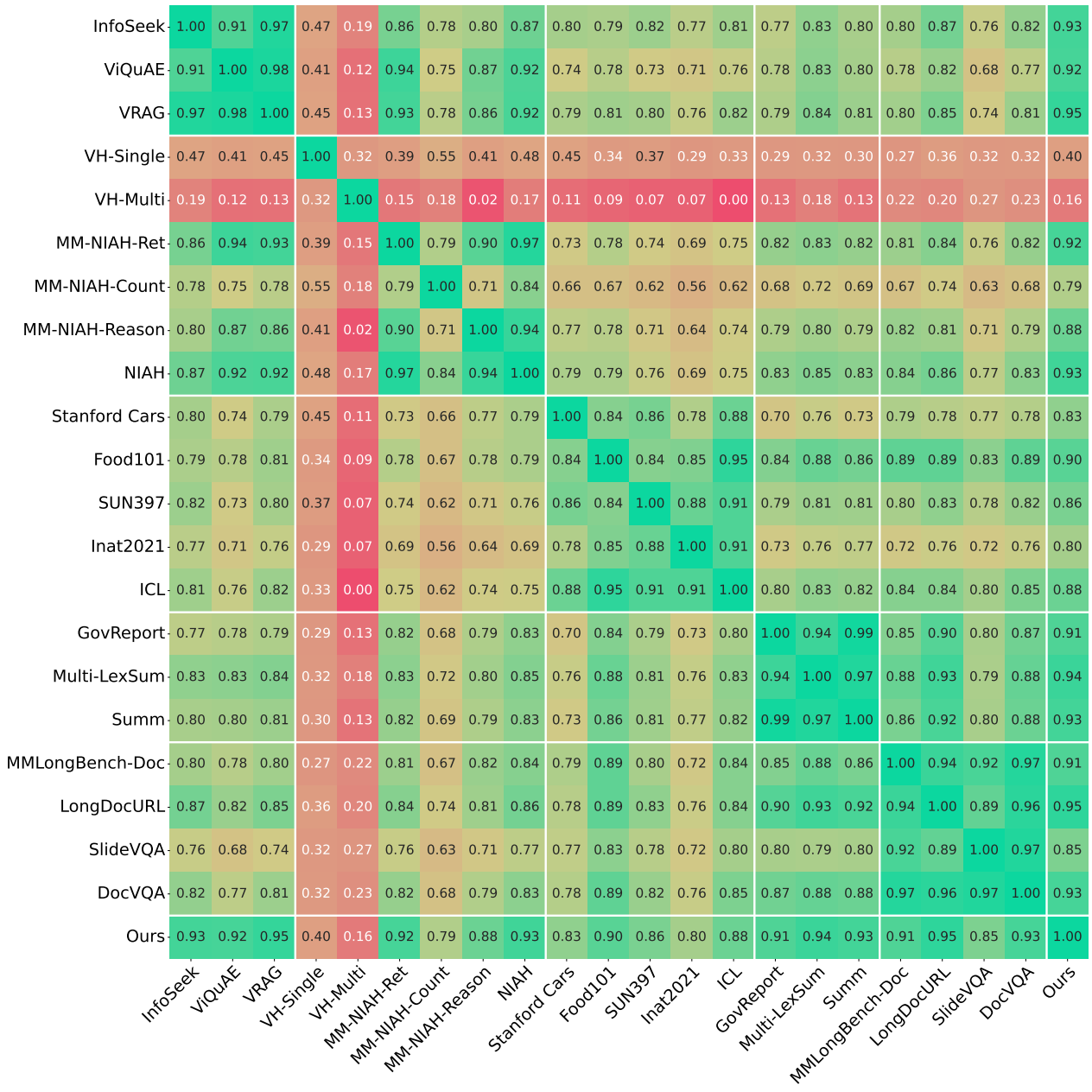


Figure 14. Spearman's correlation at 128K input length, calculated across 46 LCVLMs, between all MMLongBench datasets and category averages.

E.9. Positional Embedding Extrapolation Experiments

In this section, we evaluate two positional embedding extrapolation methods, namely YaRN (Peng et al.) and V2PE (Ge et al., 2024). Experimental results indicate that the current positional extrapolation methods still pose significant challenges for effectively extending the context window of LCVLMs.

Adding YaRN to Qwen2.5-VL. According to its technical reports (Bai et al., 2025), Qwen2.5-VL models are pre-trained with a context length of 32K tokens during the "Long-Context Pre-Training" stage. Meanwhile, its HuggingFace model card

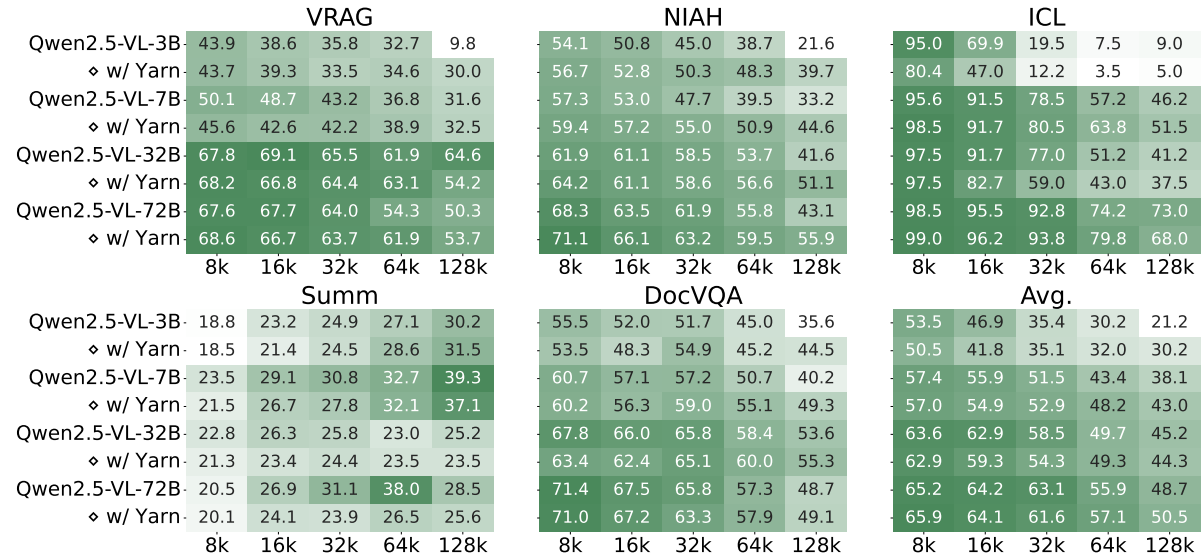


Figure 15. Results of applying YaRN (Peng et al.) to Qwen2.5-VL models. We find that YaRN only improves the performance of Qwen2.5-VL-3B substantially. However, the SoTA performance from larger models (i.e., 32B and 72B) only fluctuates slightly.

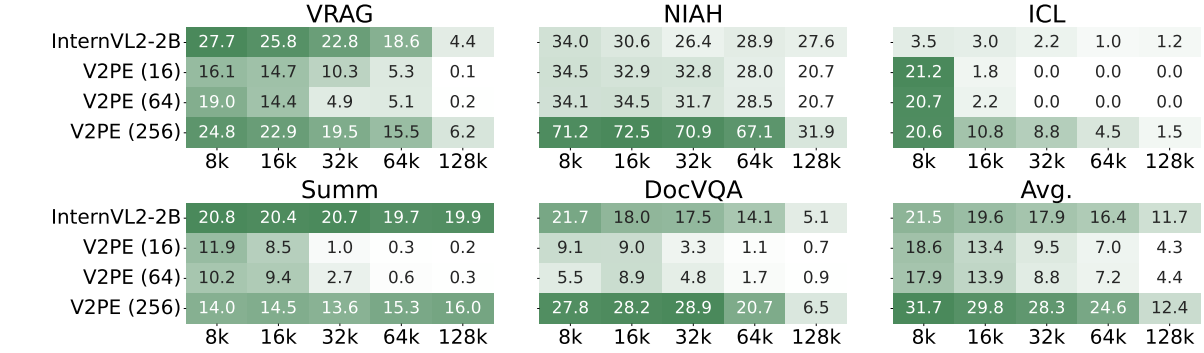


Figure 16. Results of applying V2PE (Ge et al., 2024) to InternVL2-2B. We find V2PE is very sensitive to the visual increment δ and overfitted to NIAH tasks. Note that we use ROUGE-L for summarization here, since it is already sufficient to distinguish between models. Thus, there is no need to use the costly GPT-4o evaluation. The numbers in parentheses (i.e., 16, 64, and 256) correspond to the visual increment $\delta \in \{\frac{1}{16}, \frac{1}{64}, \frac{1}{256}\}$, respectively.

shows that we can use YaRN (Peng et al.), with a scaling factor of 4, to extend its context length to 128K tokens[6]. Note that YaRN is evaluated in a zero-shot manner, as there is no continual training for YaRN on Qwen2.5-VL. In Figure 15, we test the performance of using YaRN. We have two observations: (1) Using YaRN may hurt the performance on shorter input lengths. For example, at 8K tokens, the DocVQA score of Qwen2.5-VL-32B decreases from 67.8% to 63.4%; (2) On average across the entire MMLongBench, YaRN only substantially improves the performance of Qwen2.5-VL-3B (from 21.2 to 30.2 at 128K). However, the SoTA performance from large models (i.e., 32B and 72B) only fluctuates slightly.

To ensure a fair comparison, we do not apply YaRN to Qwen2.5-VL in our main evaluations. Since YaRN is used in a zero-shot way here, applying it would lead to an unfair comparison over other models.

Adding V2PE to InternVL2. Ge et al. (2024) proposed a positional embedding extrapolation method called V2PE, where it assigns smaller positional increments to visual tokens than textual tokens. They further applied V2PE to InternVL2 and trained the model to enhance its performance on MM-NIAH-Ret (I). In this experiment, we evaluate the V2PE-256K

[6] https://huggingface.co/Qwen/Qwen2.5-VL-3B-Instruct#processing-long-texts

checkpoints[7] with the visual increments $\delta \in \{\frac{1}{16}, \frac{1}{64}, \frac{1}{256}\}$. As shown in Figure 16, we find that (1) V2PE is very sensitive to different visual increments. For example, when we use $\frac{1}{16}$ and $\frac{1}{64}$, the performance is even worse than InternVL2-2B, which leads to extra hyperparameter tuning; (2) The V2PE (256) shows extremely high performance on NIAH tasks, which can be attributed to the fact that it was trained on MM-NIAH. The sharp performance difference on NIAH tasks versus other categories suggests that V2PE is strongly overfitted to the MM-NIAH dataset.

E.10. Lost in the Middle

Existing works found that text-pure LLMs often struggle to recall needles in the middle of the input sequence, named lost in the middle (Liu et al., 2024a). On our benchmark, we extend the previous analysis to long-context vision-language tasks with input length up to 128K tokens. We place the needle at six different evenly spaced depths in the context (i.e., $[0, 0.2, 0.4, 0.6, 0.8, 1.0]$) and evaluate the LCVLMs' ability to retrieve it. In our study, the needle may be a gold passage, an image, or a key sentence. We show the results in Figure 17 for ViQuAE, Figure 18 for VH-Single, Figure 19 for MM-NIAH-Ret (T), Figure 20 for MM-NIAH-Ret (I), Figure 21 for MM-NIAH-Reason (I).

We observe a similar phenomenon in many LCVLMs on long-context vision-language tasks. For example, the InternVL3-14B in Figure 19 and Ovis2-34B in Figure 20 both exhibit much better performance when the needle is at depths 0 and 1.0. Furthermore, as we extend the context to longer lengths (e.g., 128K tokens), we observe cases where the model tends to favor either the very beginning or the very end of the context, but not both simultaneously. For example, as shown in Figure 17, InternVL3-8B prefers the very beginning of the context (depth 0) at 128K tokens, whereas Qwen2.5-VL-72B favors the very end (depth 1.0).

E.11. Error Analaysis Details

We conducted two error analyses in Appendix E.4. We provide more details of those two analyses in this section. First, for MMLongBench-Doc, we used PyMuPDF [8] to extract the plain text from PDF-formatted documents. For ViQuAE, the entity names are already provided in the dataset, since it is constructed based on TriviaQA. The text-pure LLMs we used, corresponding to Qwen2.5-VL models, are Qwen2.5-7B-Instruct and Qwen2.5-32B-Instruct, which are instruction-tuned versions.

E.12. Full Model Evaluation Results

In Figure 22, we provide the results of all 46 models. We also plot the performance of all 46 models on each dataset in Figures 23 to 26.

E.13. Idefics2 Performance

The Idefics2-8B and Idefics-8B-C only have a training context window of 2K tokens (Laurençon et al., 2024b). We find this leads to very poor long-context generalization. Also, the LLM used in Idefics2 is Mistral-7B-v0.1 (Jiang et al., 2023), whose training length is only 8K tokens. From Figure 19, we observe that Idefics2 models perform well only when the needle depth is 1.0 and the context is short (8K or 16K tokens). Additionally, we conduct a sanity check by removing all negative images and retaining only the needle images in Visual Haystack (i.e., one image for single-needle examples and two or three images for multi-needle examples). As shown in Table 12, we observe that both models achieve performance much higher than a random guess (50%), indicating the correctness of the implementation.

Table 12. Sanity check of Idefics2-8B and Idefics2-8B-C. Here we use the Visual Haystack dataset. We remove all negative images and only retain needle images (i.e., one image for single-needle examples and two or three images for multi-needle examples).

Model	VH-Single	VH-Multi
Idefics2-8B	79.33	67.67
Idefics2-8B-C	69.00	58.67

[7] https://huggingface.co/OpenGVLab/V2PE/tree/main/V2PE-256K
[8] https://pymupdf.readthedocs.io

F. Prompts and Data Examples

We list a few data examples with prompts in Figures 27 to 32. For NIAH tasks, we provide examples of both VH and MM-NIAH, as their input formats are very different.

G. Limitation

For limitation of evaluated models, while we already provide an extensive coverage of 46 frontier LCVLMs, there are still some models that we cannot cover due to token efficiency or integration challenges of codebases, as we discussed in Appendix C.1. We leave those works for future study. Meanwhile, the largest open-source models we evaluated are up to 72B in size (Qwen2-VL-72B and Qwen2.5-VL-72B). As we discussed in Appendix D, our computational resources are limited to 8×A100 (80G) GPUs; thereby it is hard to deploy and evaluate larger models with over 100B parameters at the input length of 128K tokens, such as Llama4 (Meta, 2025).

For evaluating summarization, we use a model-based metric (See Appendix B.4) that can provide much better alignment with human judgment than N-gram overlap metrics, such as ROUGE-L. However, we find using GPT-4o to provide the evaluation is expensive, which prevents the long-context community from conducting evaluations with hundreds or even thousands of models. Therefore, it is necessary to find an alternative evaluation method with a lower cost.

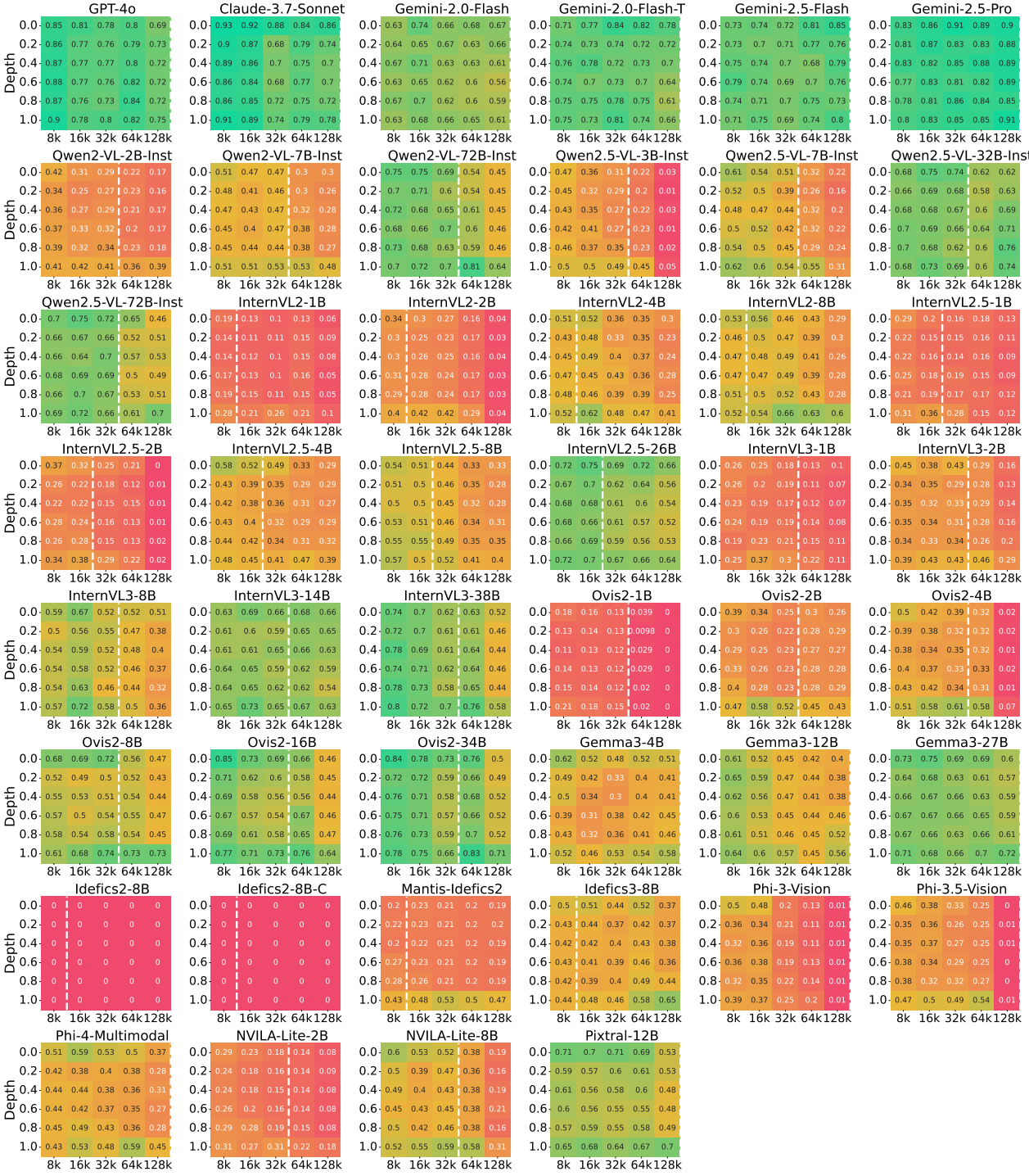


Figure 17. Performance of models on ViQuAE at different depths. Depth is the position of the gold passage, and its values are $[0.0, 0.2, 0.4, 0.6, 0.8, 1.0]$, where 0.0 is the beginning of the context (the top of each heatmap) and 1.0 is the end (the bottom of each heatmap).

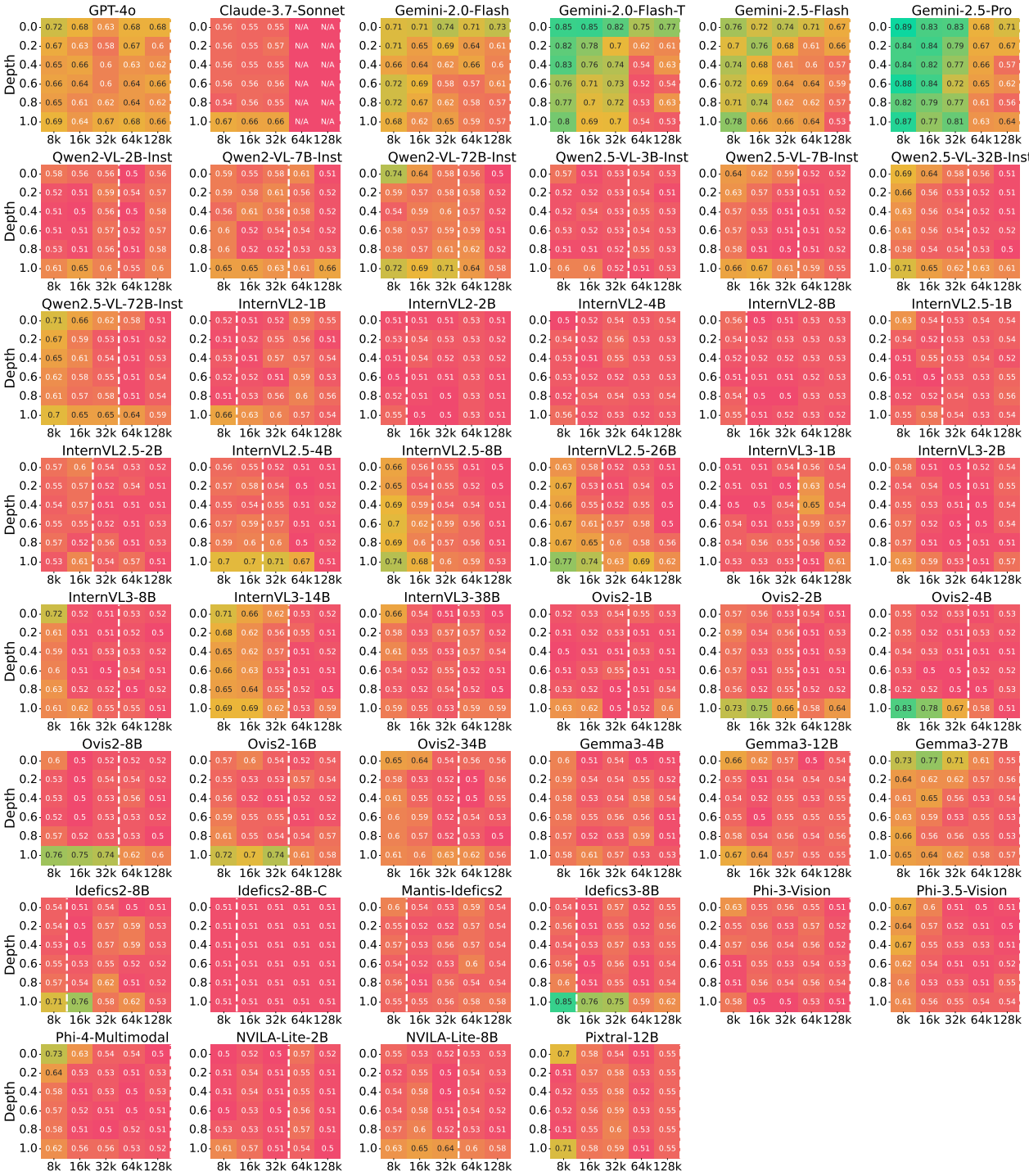


Figure 18. Performance of models on VH-Single at different depths. Depth is the position of the image containing the target object, and its values are $[0.0, 0.2, 0.4, 0.6, 0.8, 1.0]$, where 0.0 is the beginning of the context (the top of each heatmap) and 1.0 is the end (the bottom of each heatmap).

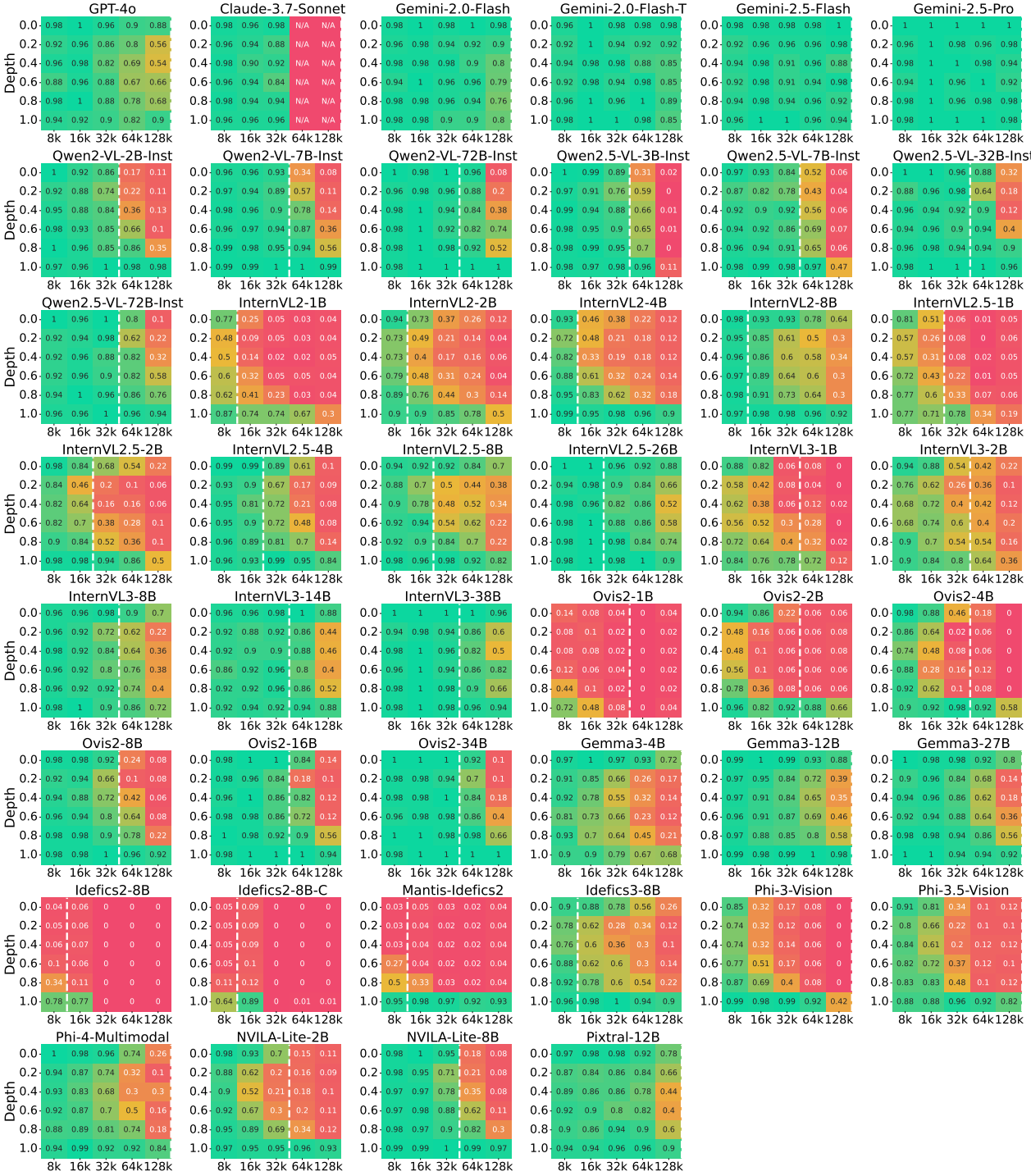


Figure 19. Performance of models on MM-NIAH-Ret (T) at different depths. Depth is the position of the text needle, and its values are $[0.0, 0.2, 0.4, 0.6, 0.8, 1.0]$, where 0.0 is the beginning of the context (the top of each heatmap) and 1.0 is the end (the bottom of each heatmap).

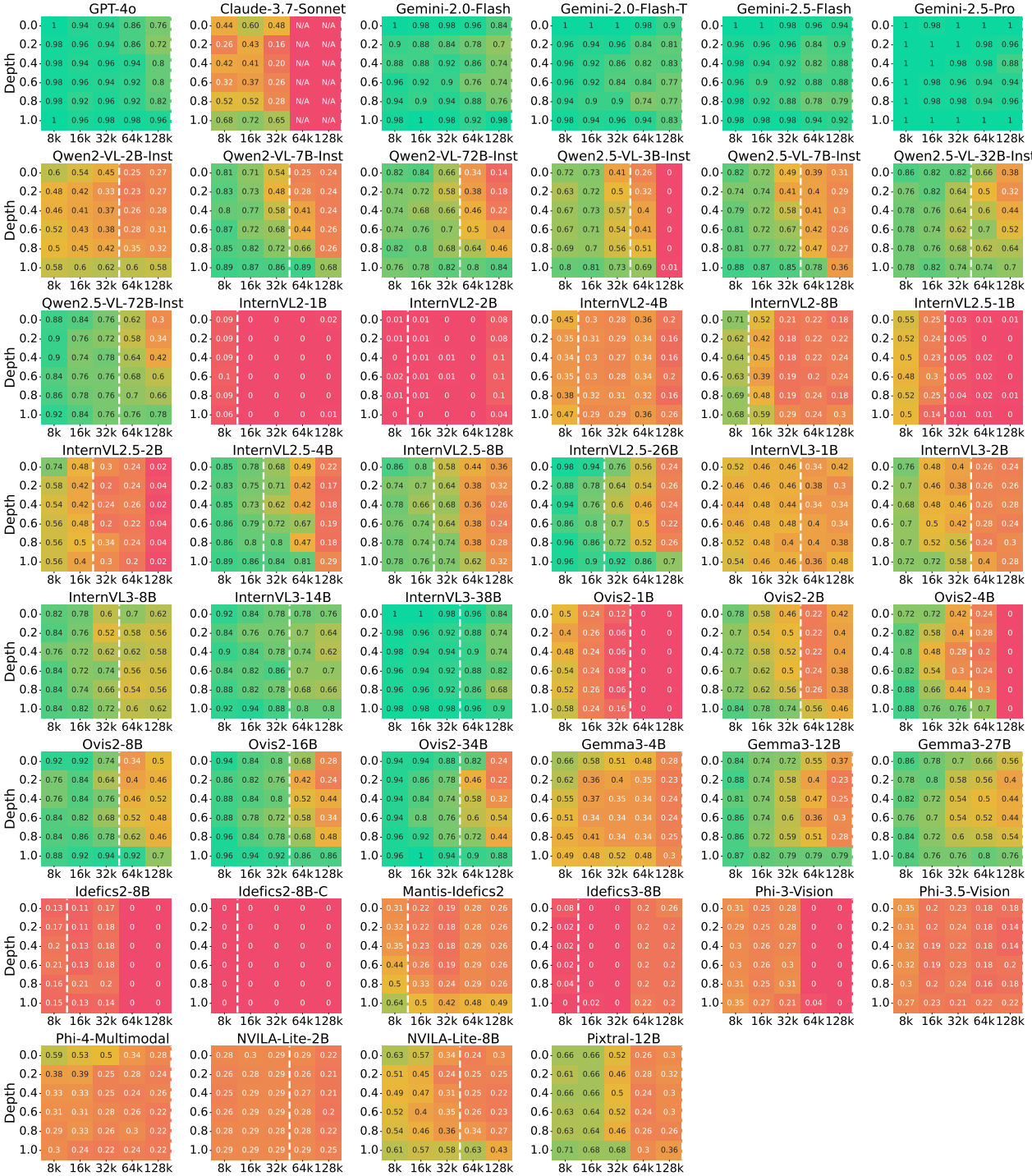


Figure 20. Performance of models on MM-NIAH-Ret (I) at different depths. Depth is the position of the image needle, and its values are $[0.0, 0.2, 0.4, 0.6, 0.8, 1.0]$, where 0.0 is the beginning of the context (the top of each heatmap) and 1.0 is the end (the bottom of each heatmap).

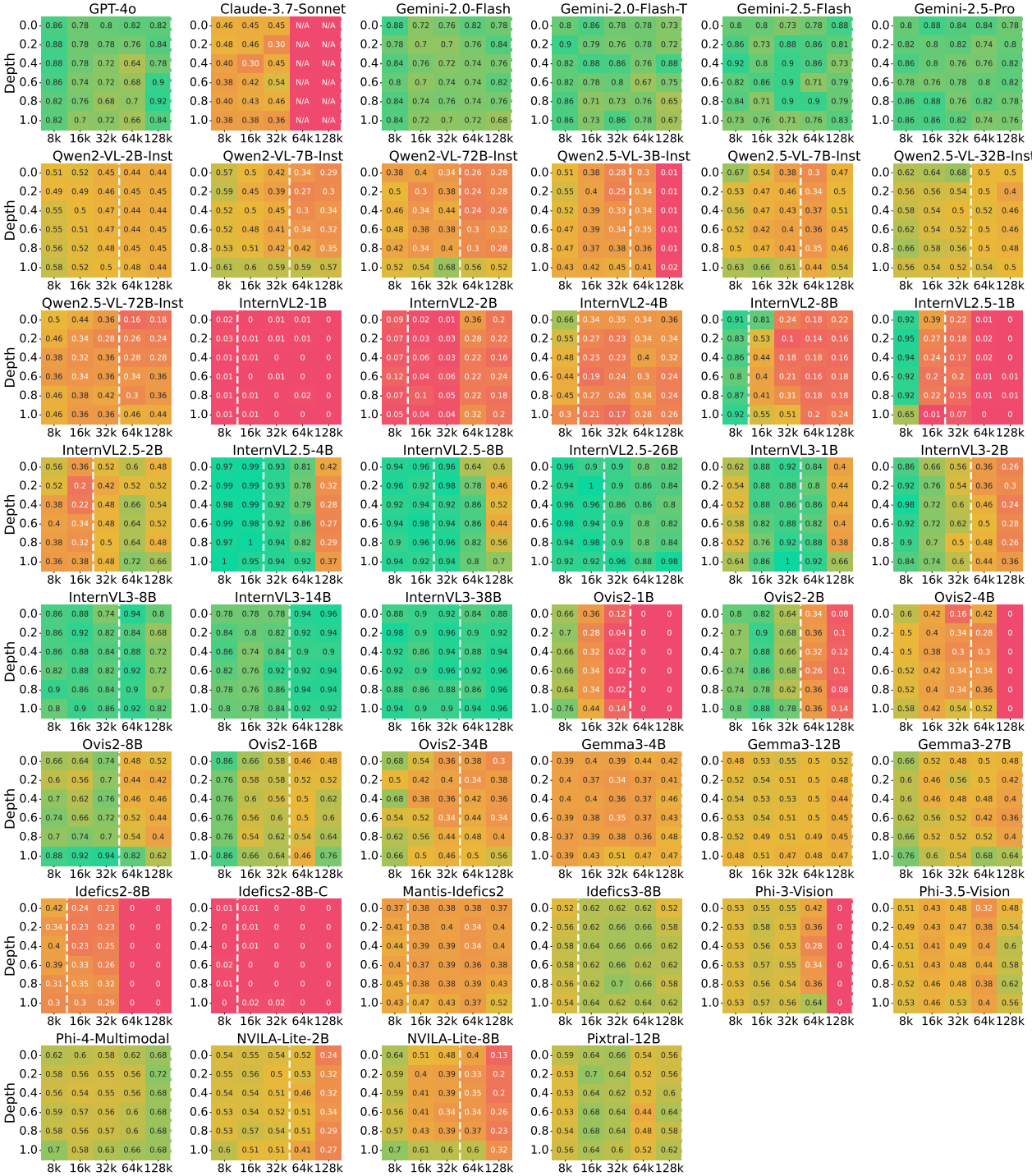


Figure 21. Performance of models on MM-NIAH-Reason (I) at different depths. Depth is the position of the needle image used for reasoning, and its values are $[0.0, 0.2, 0.4, 0.6, 0.8, 1.0]$, where 0.0 is the beginning of the context (the top of each heatmap) and 1.0 is the end (the bottom of each heatmap).

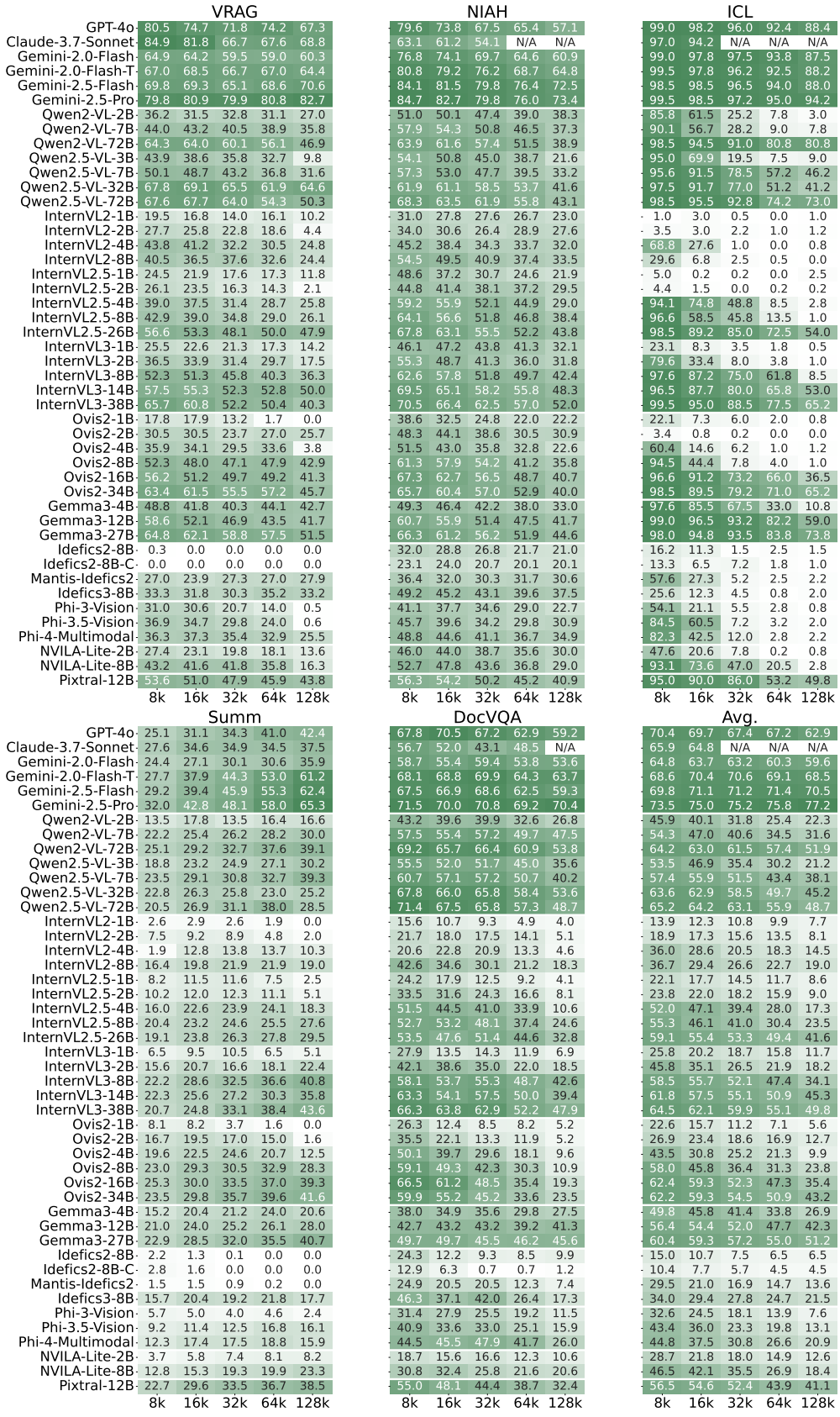


Figure 22. Results of all 46 models on MMLongBench at various lengths.

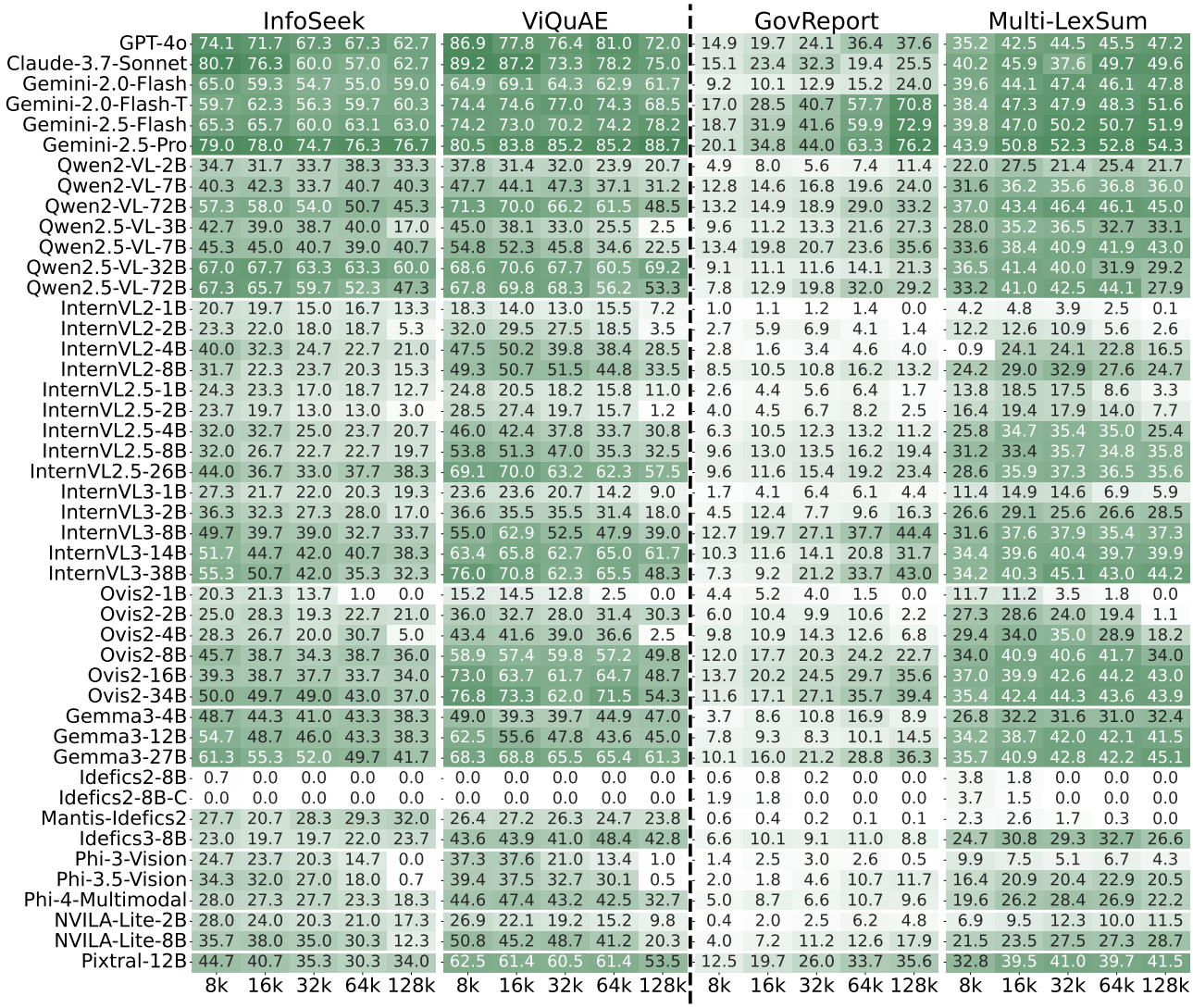


Figure 23. Results of 46 models on categories VRAG and Summ at various lengths.

	VH-Single					VH-Multi					MM-NIAH-Ret (T)					MM-NIAH-Ret (I)				
	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k
GPT-4o	67.1	64.3	61.7	65.7	64.0	70.3	65.0	63.3	61.0	50.7	94.3	97.0	88.3	77.6	70.3	98.7	94.7	95.3	91.3	81.0
Claude-3.7-Sonnet	50.5	51.8	52.0	N/A	N/A	57.3	54.7	50.3	N/A	N/A	97.0	94.0	91.9	N/A	N/A	44.0	50.8	33.7	N/A	N/A
Gemini-2.0-Flash	69.7	66.3	65.1	62.5	61.5	62.0	61.8	58.1	54.8	56.4	97.0	98.7	95.3	93.6	83.7	94.3	92.7	92.7	86.0	79.3
Gemini-2.0-Flash-T	80.1	74.8	73.4	58.2	62.0	66.5	67.0	64.6	63.7	63.8	94.7	99.3	96.5	95.5	89.2	96.5	93.5	91.8	86.0	81.9
Gemini-2.5-Flash	73.6	70.9	65.7	63.8	59.9	78.6	72.4	67.5	65.4	61.4	94.2	98.0	95.0	95.8	94.8	97.6	94.9	93.9	87.0	88.2
Gemini-2.5-Pro	85.4	81.8	78.0	64.9	62.7	76.5	72.9	68.9	70.1	67.7	97.3	100.0	98.3	98.7	96.7	100.0	99.0	98.7	97.3	95.7
Qwen2-VL-2B	53.3	54.0	57.3	50.3	57.7	55.5	54.5	52.2	51.8	51.0	97.0	92.2	85.7	54.2	29.7	52.3	47.5	42.8	32.8	33.8
Qwen2-VL-7B	59.8	57.2	53.3	53.5	51.3	54.8	53.5	51.8	51.7	55.7	97.5	96.7	93.5	75.0	37.3	84.2	77.0	64.3	48.8	32.0
Qwen2-VL-72B	62.5	60.5	61.2	59.3	50.8	60.7	57.7	55.7	53.3	54.0	98.0	99.3	96.7	90.3	48.7	77.0	76.3	68.3	52.0	37.3
Qwen2.5-VL-3B	54.2	50.3	52.0	53.5	51.7	51.8	52.2	54.2	51.7	54.2	98.3	95.8	89.3	64.5	2.5	69.7	73.3	55.2	43.2	0.2
Qwen2.5-VL-7B	60.8	57.5	52.5	52.5	52.5	54.8	53.7	54.0	55.2	54.8	94.0	91.7	88.2	63.7	12.7	80.8	75.7	62.0	47.8	29.8
Qwen2.5-VL-32B	64.7	58.5	55.8	54.7	51.3	57.5	56.0	55.8	56.2	53.7	95.3	97.3	95.0	88.3	48.0	79.0	78.7	69.0	63.7	50.0
Qwen2.5-VL-72B	66.0	61.0	57.8	54.3	53.5	64.3	57.7	54.5	54.0	55.0	94.3	96.3	95.3	81.3	48.7	88.3	78.7	75.7	66.3	51.7
InternVL2-1B	52.5	53.7	51.8	55.7	51.2	50.5	51.0	54.8	51.8	53.0	64.0	32.5	19.0	13.8	8.5	8.7	0.0	0.0	0.0	0.5
InternVL2-2B	50.3	50.8	51.2	53.0	51.3	50.8	50.5	51.7	53.5	55.2	83.0	62.7	39.2	31.3	15.0	0.8	0.8	0.3	0.0	8.3
InternVL2-4B	51.0	51.5	54.0	52.8	53.3	55.0	57.8	55.3	56.0	58.8	88.2	61.0	45.0	35.0	26.3	39.0	30.3	28.7	34.3	19.0
InternVL2-8B	51.3	50.3	51.2	52.8	52.8	53.3	60.7	59.3	57.0	57.0	96.8	90.3	74.8	67.7	46.7	66.2	47.5	20.7	22.3	22.7
InternVL2.5-1B	54.0	53.5	53.3	53.3	50.8	56.0	56.2	55.5	56.2	50.3	70.2	47.0	25.8	7.5	7.7	51.2	23.5	3.8	1.3	0.3
InternVL2.5-2B	55.2	57.0	53.2	52.8	51.5	52.5	53.7	54.7	52.7	51.3	89.0	74.3	48.0	38.3	17.3	59.0	45.0	26.3	23.3	3.0
InternVL2.5-4B	59.0	59.3	58.2	53.3	50.3	56.0	53.5	55.0	55.5	50.2	95.7	90.3	80.0	52.0	22.2	85.7	78.5	72.8	54.7	20.5
InternVL2.5-8B	68.8	59.8	56.7	54.7	50.3	57.3	53.5	50.7	52.5	55.3	92.3	86.7	71.0	67.3	44.7	78.3	73.3	67.0	42.7	29.7
InternVL2.5-26B	67.8	61.0	55.8	57.8	52.8	62.8	59.0	56.0	56.7	51.8	97.7	99.0	92.0	88.7	71.3	93.0	84.0	72.3	57.3	32.0
InternVL3-1B	50.8	51.0	52.2	58.3	55.3	52.8	54.7	55.3	53.8	55.2	70.0	59.0	28.0	26.0	2.7	49.0	47.0	45.0	37.0	37.7
InternVL3-2B	57.3	53.3	50.3	50.5	54.2	54.7	52.7	52.3	51.7	53.0	81.0	75.0	52.3	46.3	19.3	71.0	52.7	46.7	28.3	26.7
InternVL3-8B	62.7	52.8	50.7	51.2	51.0	53.5	50.2	50.0	56.0	51.5	96.7	94.0	86.0	75.3	46.3	82.3	75.7	64.3	60.0	58.3
InternVL3-14B	67.3	64.3	57.5	51.5	51.8	57.7	57.2	54.8	51.0	50.5	91.7	92.0	94.0	89.0	59.7	88.3	83.7	80.7	73.3	69.7
InternVL3-38B	58.5	53.5	52.5	52.0	50.7	57.0	55.3	54.7	51.3	50.7	97.3	99.7	96.7	90.0	74.7	97.7	96.3	94.3	90.7	78.7
Ovis2-1B	53.3	53.8	52.2	52.2	53.3	55.2	55.0	54.8	55.3	55.5	26.3	15.0	3.7	0.0	2.7	50.3	24.7	9.0	0.0	0.0
Ovis2-2B	59.8	56.8	56.7	52.8	54.0	55.0	55.0	54.0	53.7	55.2	70.0	40.0	23.3	19.7	16.7	74.3	62.0	54.7	28.7	40.7
Ovis2-4B	58.7	56.2	53.7	50.8	52.0	55.5	55.2	54.8	54.3	55.0	88.0	63.7	30.0	23.7	9.7	82.0	62.3	43.3	32.7	0.0
Ovis2-8B	58.5	54.5	56.5	51.0	50.5	54.7	54.5	55.0	54.5	55.0	96.0	95.0	83.3	52.3	24.0	83.3	86.7	75.7	54.3	52.0
Ovis2-16B	60.0	57.5	54.5	50.8	54.5	57.0	52.7	54.7	51.0	50.7	98.0	98.7	91.3	74.3	33.0	91.3	84.3	79.7	62.3	44.0
Ovis2-34B	60.8	58.0	51.5	54.2	52.3	59.3	54.0	53.7	53.3	50.0	98.3	99.0	98.3	88.3	40.7	94.7	89.3	81.0	68.0	44.0
Gemma3-4B	58.3	54.0	53.3	54.2	51.7	52.7	59.0	52.8	56.5	52.7	90.7	82.7	71.2	47.7	34.0	54.7	42.3	41.0	38.8	25.7
Gemma3-12B	59.2	55.5	51.0	51.8	53.2	56.3	51.3	52.5	51.0	53.2	97.5	93.8	89.7	79.8	60.7	85.3	75.0	64.3	51.3	37.0
Gemma3-27B	65.3	63.8	60.0	56.5	54.7	57.2	53.5	56.3	60.3	57.0	95.7	97.0	90.7	77.7	49.3	81.7	74.7	62.0	60.3	52.3
Idefics2-8B	57.3	53.0	57.2	54.8	50.7	53.5	51.2	52.0	53.5	54.2	22.8	18.8	0.0	0.0	0.0	17.0	13.7	17.5	0.0	0.0
Idefics2-8B-C	50.0	50.0	50.0	50.0	50.0	50.3	50.3	50.3	50.3	50.3	15.8	23.0	0.0	0.2	0.2	0.0	0.0	0.0	0.0	0.0
Mantis-Idefics2	56.2	52.8	54.3	57.5	54.7	55.2	55.0	50.2	52.7	50.3	30.2	24.7	18.2	17.0	18.8	42.7	29.3	23.3	31.8	29.8
Idefics3-8B	60.8	55.3	59.2	50.0	56.0	54.0	54.0	52.8	50.3	50.3	86.7	74.7	60.3	49.7	29.0	3.0	0.3	0.0	20.3	21.0
Phi-3-Vision	55.8	54.7	53.3	55.2	52.2	50.3	54.0	56.2	55.3	57.7	82.7	52.3	33.2	21.0	7.0	31.0	25.7	27.5	0.7	0.0
Phi-3.5-Vision	63.5	55.8	52.5	50.0	51.5	55.5	52.2	52.0	50.2	50.8	84.7	75.2	42.8	24.3	23.0	31.2	20.2	22.8	18.3	17.7
Phi-4-Multimodal	62.0	53.3	52.8	51.7	51.0	53.3	52.5	52.2	50.8	52.7	93.5	90.5	80.2	58.7	30.7	36.7	35.5	29.3	27.7	24.0
NVILA-Lite-2B	52.7	53.8	50.0	55.7	51.0	50.8	54.0	50.5	50.5	50.3	93.3	76.3	50.8	33.2	24.3	27.0	29.0	28.8	27.8	21.8
NVILA-Lite-8B	55.8	58.2	52.5	51.3	50.5	53.7	51.0	53.3	54.3	52.2	97.8	97.7	87.0	52.8	27.0	55.0	48.7	36.3	32.8	28.3
Pixtral-12B	56.5	51.2	55.0	51.8	55.0	55.3	52.8	50.2	53.3	53.2	91.5	89.7	90.0	87.0	63.0	65.0	65.7	52.3	26.3	30.7

	MM-NIAH-Count (T)					MM-NIAH-Count (I)					MM-NIAH-Reason (T)					MM-NIAH-Reason (I)				
	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k
GPT-4o	94.7	94.7	78.0	64.6	42.0	67.3	36.0	10.0	29.3	4.7	81.3	81.3	80.0	66.7	58.7	85.3	75.7	73.7	71.0	85.0
Claude-3.7-Sonnet	96.0	90.0	72.0	N/A	N/A	40.0	33.3	23.2	N/A	N/A	96.6	90.6	73.0	N/A	N/A	41.7	40.6	42.4	N/A	N/A
Gemini-2.0-Flash	78.7	76.7	64.0	50.0	34.0	62.7	60.7	48.7	34.0	27.3	88.7	84.0	76.9	72.3	72.0	83.0	72.3	73.0	75.3	77.3
Gemini-2.0-Flash-T	84.7	77.8	72.8	59.7	46.1	63.3	66.0	58.5	46.5	29.1	91.3	92.7	86.4	80.3	76.9	84.0	79.2	79.6	75.2	73.2
Gemini-2.5-Flash	95.3	96.5	94.6	90.3	87.7	71.3	71.3	75.3	63.9	45.0	95.3	90.7	89.1	87.8	85.7	82.7	77.6	83.7	81.0	80.6
Gemini-2.5-Pro	99.3	98.7	95.3	87.2	80.5	47.3	41.3	40.0	34.7	31.3	96.7	95.3	92.5	93.3	90.0	82.7	83.0	79.7	78.3	79.3
Qwen2-VL-2B	17.7	14.7	16.7	11.0	13.7	35.0	40.7	36.0	26.0	34.3	36.3	38.0	26.7	17.0	9.7	54.0	51.0	47.2	45.0	44.5
Qwen2-VL-7B	33.3	28.0	21.0	16.7	6.7	29.0	19.0	28.3	41.3	31.3	49.7	50.7	46.3	35.0	15.0	55.7	50.7	44.7	37.7	36.2
Qwen2-VL-72B	69.3	64.0	49.3	40.0	13.3	22.7	20.7	13.3	20.0	19.3	79.3	80.7	69.3	56.0	28.0	46.0	38.3	43.7	31.7	32.3
Qwen2.5-VL-3B	36.7	34.3	22.3	8.3	0.0	28.0	21.0	2.3	9.3	0.0	47.7	39.3	34.7	16.3	0.3	49.2	39.2	33.8	35.0	1.2
Qwen2.5-VL-7B	33.7	26.0	19.3	5.3	9.3	21.7	14.0	7.0	3.3	16.3	54.7	50.0	43.0	24.0	0.7	56.8	50.5	44.8	36.0	48.8
Qwen2.5-VL-32B	51.3	53.3	43.3	35.3	9.3	11.3	23.3	33.3	20.7	14.7	77.3	72.7	65.3	56.7	37.3	60.0	56.7	55.3	51.0	46.7
Qwen2.5-VL-72B	88.0	86.0	79.3	68.0	16.0	23.3	23.3	37.3	38.0	27.3	84.7	77.3	71.3	58.0	39.3	43.7	36.3	35.7	30.0	31.0
InternVL2-1B	12.0	7.7	6.0	8.3	4.0	6.7	18.0	30.7	26.7	8.3	11.0	9.7	6.7	2.0	0.7	1.5	0.7	0.5	0.7	0.0
InternVL2-2B	14.0	10.0	2.3	2.0	2.7	6.3	5.7	3.0	5.3	12.0	26.0	19.7	10.3	10.0	4.7	7.8	4.8	3.7	27.0	20.0
InternVL2-4B	21.0	13.7	6.7	3.3	9.3	3.3	1.3	1.0	0.0	0.0	40.3	33.3	18.7	13.3	11.3	48.0	25.2	24.7	33.3	29.3
InternVL2-8B	22.0	16.7	10.0	4.0	6.7	15.3	24.3	25.7	28.0	13.3	48.7	41.3	31.3	14.7	6.7	86.5	52.3	25.8	17.3	19.0
InternVL2.5-1B	15.7	13.3	6.7	4.7	1.3	29.0	32.7	28.3	7.7	2.7	11.7	14.3	8.3	4.3	4.0	88.3	22.2	16.5	1.2	0.3
InternVL2.5-2B	5.3	3.3	3.3	9.3	4.7	2.7	11.3	28.7	16.0	6.7	33.3	28.0	10.7	11.3	4.0	43.3	30.3	48.0	63.0	53.3
InternVL2.5-4B	28.7	22.3	14.7	6.7	1.7	4.3	4.7	5.7	19.3	2.3	49.0	38.7	28.3	16.0	9.7	98.3	98.3	93.0	83.0	32.5
InternVL2.5-8B	32.0	20.0	4.0	9.3	10.0	38.7	27.3	31.3	32.0	21.3	54.0	37.3	35.3	22.7	12.7	93.3	94.7	95.0	79.3	54.7
InternVL2.5-26B	54.7	46.7	22.7	16.0	10.0	14.7	14.7	6.0	10.7	2.7	61.3	52.0	48.0	36.7	28.0	95.0	95.0	90.3	83.3	85.0
InternVL3-1B	18.0	12.0	12.7	12.7	6.7	35.3	37.3	36.0	21.3	6.7	24.7	20.7	10.7	6.7	0.7	57.0	84.7	91.0	85.3	45.3
InternVL3-2B	16.7	17.3	18.7	14.7	12.0	29.3	28.0	13.3	16.0	14.0	40.7	30.0	20.0	7.3	3.3	90.3	71.7	57.0	43.3	28.3
InternVL3-8B	42.7	49.3	34.0	21.3	8.7	21.3	14.0	9.3	7.3	6.0	65.3	50.0	40.7	28.7	26.0	85.0	88.7	82.0	90.0	74.0
InternVL3-14B	68.7	61.3	37.3	33.3	14.7	35.3	22.0	8.7	18.0	17.3	79.3	72.0	52.7	46.7	23.3	81.3	77.0	84.3	92.3	93.3
InternVL3-38B	74.0	71.3	52.7	38.7	23.3	29.3	17.3	8.7	8.7	5.3	84.0	72.7	66.0	46.0	41.3	92.0	88.7	92.7	89.0	93.7
Ovis2-1B	3.3	5.3	7.3	0.0	0.0	16.7	22.0	3.3	0.7	0.0	4.0	5.3	4.7	4.0	1.3	68.0	34.7	6.0	0.0	0.0
Ovis2-2B	5.3	9.3	6.0	5.3	6.0	4.0	2.7	3.3	1.3	16.7	24.7	18.0	9.3	3.3	0.0	74.7	85.3	67.7	33.3	10.3
Ovis2-4B	20.7	16.7	11.3	6.0	0.0	0.0	0.0	0.0	2.0	0.0	42.0	24.0	22.7	16.0	2.0	53.7	40.7	33.7	37.0	0.0
Ovis2-8B	36.0	35.3	19.3	7.3	6.7	42.7	37.3	25.3	10.7	4.7	57.3	37.3	39.3	22.0	12.0	71.3	69.0	76.0	54.3	47.7
Ovis2-16B	42.7	46.7	27.3	22.0	9.3	52.0	53.3	40.7	39.3	31.3	75.3	63.3	48.0	36.0	18.7	79.3	60.0	59.7	49.7	60.3
Ovis2-34B	48.0	52.7	46.0	33.3	13.3	36.0	14.7	26.7	35.3	31.3	78.0	75.3	68.7	46.7	26.7	61.3	48.7	39.3	42.7	39.0
Gemma3-4B	21.3	15.3	7.3	3.7	3.0	20.7	14.3	15.7	5.0	0.7	44.3	43.7	35.7	23.3	13.0	39.0	39.5	38.8	39.7	44.5
Gemma3-12B	44.7	33.0	18.0	18.3	9.3	35.0	39.7	36.7	25.7	30.0	62.7	51.7	47.0	35.0	20.7	51.2	52.3	51.3	49.5	46.8
Gemma3-27B	64.0	57.3	43.3	28.7	12.7	30.7	30.0	24.7	34.7	39.3	81.3	66.7	57.3	32.7	24.0	65.0	52.0	51.7	51.7	45.0
Idefics2-8B	5.3	2.3	0.0	0.0	0.0	8.7	8.7	5.7	0.0	0.0	8.7	7.7	0.0	0.0	0.0	36.0	28.0	26.3	0.0	0.0
Idefics2-8B-C	4.3	1.3	0.0	0.0	0.0	4.3	7.7	6.3	0.3	0.0	5.3	6.3	0.0	0.0	0.0	0.7	0.8	0.3	0.0	0.0
Mantis-Idefics2	3.7	2.0	1.7	4.3	3.0	4.7	1.3	5.0	4.0	1.0	18.3	7.3	6.7	3.0	1.3	41.7	39.5	39.5	36.3	41.7
Idefics3-8B	33.3	26.0	14.0	9.3	6.0	36.7	40.7	38.7	38.0	36.0	47.3	29.3	28.7	13.3	11.3	55.7	62.7	65.3	64.3	59.0
Phi-3-Vision	3.0	2.0	1.7	0.0	0.0	0.0	1.7	0.3	0.0	0.0	29.0	21.7	10.3	7.3	0.7	53.0	56.5	54.3	40.0	0.0
Phi-3.5-Vision	7.0	4.0	1.3	8.0	0.7	0.0	1.7	1.3	0.7	0.7	45.0	35.7	15.3	7.3	6.0	51.2	43.7	48.8	38.7	56.3

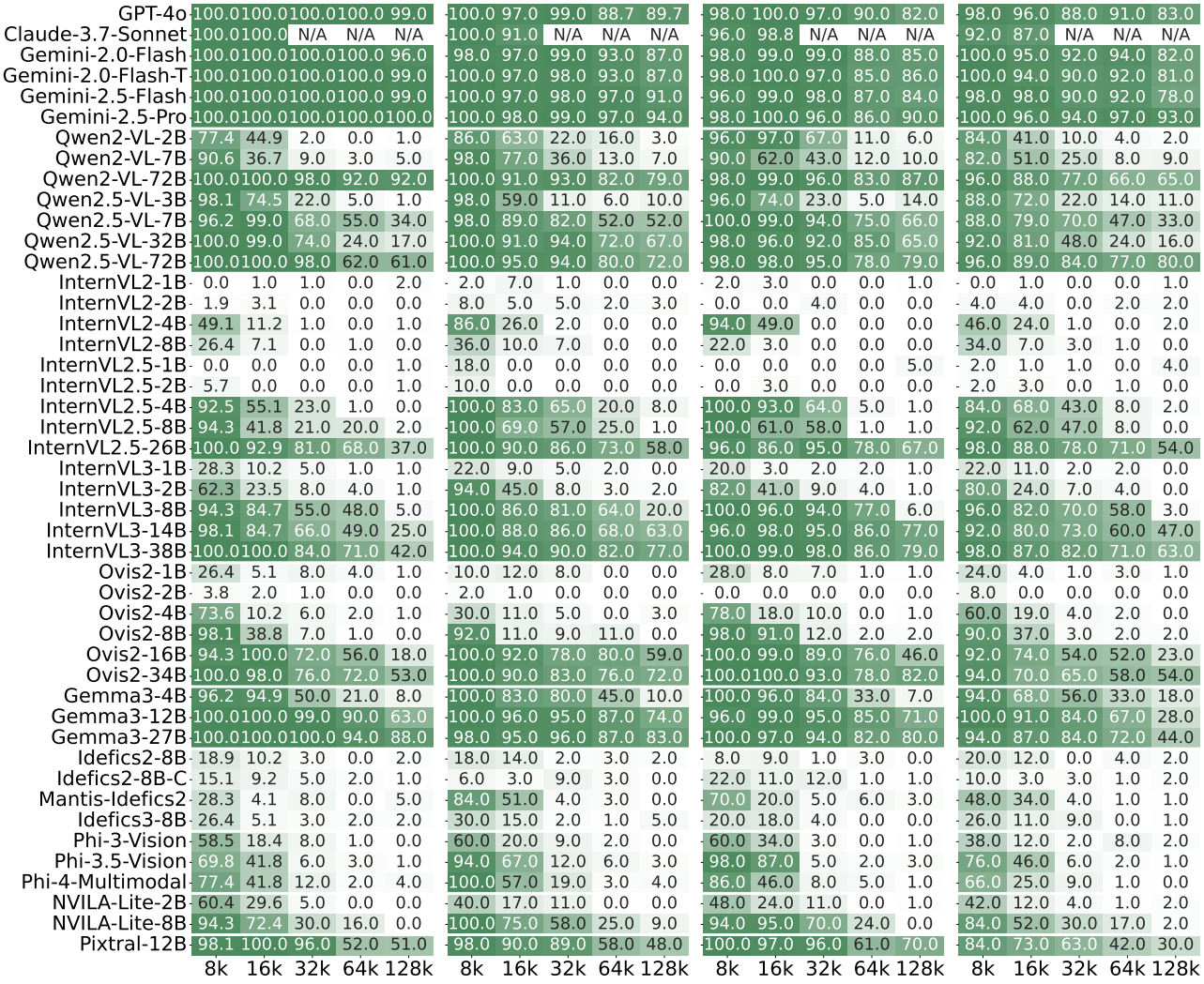


Figure 25. Results of 46 models on the category ICL at various lengths.

	MMLongBench-Doc					LongDocURL					SlideVQA				
GPT-4o	54.1	66.1	50.8	47.1	43.0	72.3	67.9	75.9	68.2	63.2	77.0	77.5	75.0	73.5	71.4
Claude-3.7-Sonnet	51.8	39.2	35.1	33.0	N/A	61.0	56.4	63.4	55.2	N/A	57.4	60.5	30.7	57.4	N/A
Gemini-2.0-Flash	47.3	47.9	47.1	39.7	38.8	59.6	52.0	56.8	50.4	56.2	69.0	66.5	74.1	71.1	65.7
Gemini-2.0-Flash-T	63.5	61.1	57.0	52.1	53.2	63.3	62.0	69.4	65.1	60.5	77.5	83.1	83.2	75.5	77.5
Gemini-2.5-Flash	53.8	55.6	57.4	51.3	50.9	66.2	61.2	68.5	59.6	52.0	82.6	83.9	79.8	76.7	74.9
Gemini-2.5-Pro	60.2	62.2	55.9	56.0	60.0	68.9	63.5	69.2	66.2	65.4	85.3	84.3	87.3	85.3	85.6
Qwen2-VL-2B	28.3	21.5	21.7	21.6	13.6	52.5	38.6	49.3	31.2	25.9	48.7	58.7	48.8	45.1	40.7
Qwen2-VL-7B	43.9	37.9	32.6	27.4	23.9	60.4	53.3	65.0	48.5	49.1	68.0	75.0	74.1	73.3	69.5
Qwen2-VL-72B	59.8	50.8	45.8	43.7	32.1	69.5	63.7	71.2	64.4	55.6	78.2	82.7	82.1	74.7	73.7
Qwen2.5-VL-3B	40.6	41.3	30.9	22.2	14.7	57.8	45.1	51.4	41.9	41.7	68.0	69.7	72.8	71.1	50.4
Qwen2.5-VL-7B	52.7	50.0	42.8	35.8	17.1	61.9	52.8	62.1	48.0	46.6	67.6	68.5	66.8	68.3	56.9
Qwen2.5-VL-32B	58.0	58.2	48.5	42.1	31.9	67.5	62.2	70.7	57.0	54.5	77.8	77.7	78.2	76.1	74.3
Qwen2.5-VL-72B	65.6	58.7	55.5	42.1	28.8	65.8	62.6	65.0	52.4	43.3	82.8	81.3	77.0	77.5	73.9
InternVL2-1B	11.3	6.9	5.3	1.8	3.4	15.6	14.5	13.5	5.0	3.5	19.9	10.7	9.0	7.9	5.0
InternVL2-2B	15.3	12.7	9.1	7.0	2.7	24.5	19.1	19.1	12.8	4.4	25.3	22.2	24.4	22.5	8.3
InternVL2-4B	13.3	14.3	13.4	4.2	4.2	17.9	25.0	23.7	14.9	8.8	30.5	29.1	25.7	20.8	1.0
InternVL2-8B	37.1	33.7	18.7	11.3	10.5	37.1	30.2	34.6	22.2	19.9	53.5	39.9	36.8	30.0	24.6
InternVL2.5-1B	17.7	14.8	8.7	6.8	3.1	24.1	19.4	18.9	11.4	5.0	30.9	19.5	9.9	9.3	4.1
InternVL2.5-2B	23.0	23.3	12.3	13.2	3.8	34.4	32.7	27.7	16.0	8.9	43.0	38.9	32.8	20.7	11.6
InternVL2.5-4B	44.2	38.0	28.8	22.9	11.3	48.1	41.4	40.3	37.3	20.1	62.1	54.2	53.9	41.6	0.3
InternVL2.5-8B	40.4	42.6	35.5	29.4	17.1	54.9	51.8	53.3	35.1	30.7	62.8	65.1	55.4	47.6	26.0
InternVL2.5-26B	38.1	36.4	32.7	30.0	16.3	54.5	44.0	54.3	43.8	34.0	67.9	62.3	67.1	60.1	48.0
InternVL3-1B	19.3	9.2	6.6	4.5	2.7	31.8	16.3	17.1	12.4	9.4	32.6	14.9	19.2	19.0	8.5
InternVL3-2B	32.6	31.3	24.3	15.6	13.2	40.9	34.1	42.2	19.7	20.1	52.7	50.5	38.6	30.7	22.0
InternVL3-8B	48.2	39.2	33.6	28.4	25.0	57.3	52.4	61.5	48.7	43.2	68.9	69.6	70.8	69.1	59.6
InternVL3-14B	55.8	51.2	48.0	37.8	26.7	57.4	54.7	54.8	48.8	47.3	76.7	56.2	69.8	63.3	44.3
InternVL3-38B	52.1	55.2	46.5	35.2	29.4	61.5	57.6	68.0	51.5	53.6	85.2	78.6	74.2	69.8	60.7
Ovis2-1B	19.5	13.2	6.3	5.0	2.2	39.9	13.9	7.6	9.0	6.9	19.4	10.0	11.5	10.5	6.5
Ovis2-2B	29.0	18.9	7.0	6.7	2.4	40.5	21.8	14.2	12.8	8.1	37.0	25.5	18.7	16.2	5.0
Ovis2-4B	38.4	36.7	17.3	12.9	9.0	53.2	32.7	31.9	22.4	11.3	58.7	49.9	39.6	19.1	8.5
Ovis2-8B	44.2	37.4	24.0	16.6	8.1	55.2	51.9	51.5	36.4	16.7	77.8	58.6	51.5	37.9	8.0
Ovis2-16B	56.5	53.4	39.3	28.0	12.6	66.7	58.6	58.0	41.8	29.8	76.2	71.5	48.3	36.4	15.7
Ovis2-34B	47.3	44.5	32.0	25.3	13.1	61.4	57.3	51.0	42.6	31.5	71.2	63.8	52.5	33.0	26.0
Gemma3-4B	35.0	32.9	26.8	19.5	17.1	38.8	32.6	35.6	29.1	25.3	40.1	39.3	44.4	40.8	40.1
Gemma3-12B	34.2	33.3	31.7	24.1	22.6	45.0	41.9	45.9	41.9	44.4	48.9	54.5	51.9	51.6	56.8
Gemma3-27B	41.4	34.1	31.4	32.3	30.0	47.2	47.3	44.9	44.4	49.9	60.6	67.9	60.3	61.8	56.9
Idefics2-8B	16.9	8.8	4.8	4.8	7.8	22.8	15.0	10.2	6.1	4.1	33.2	12.7	12.9	14.6	17.8
Idefics2-8B-C	10.4	5.3	1.0	1.0	1.1	16.5	13.0	1.0	0.8	1.5	11.8	0.7	0.0	0.2	1.0
Mantis-Idefics2	15.2	14.6	12.4	9.1	2.1	23.6	13.0	19.0	7.7	4.7	36.1	33.9	30.3	20.1	15.4
Idefics3-8B	30.5	27.1	28.0	15.5	12.3	48.9	37.3	49.3	37.1	24.0	59.5	46.9	48.8	26.5	15.5
Phi-3-Vision	15.3	19.1	12.5	9.8	7.6	32.8	23.7	24.3	15.5	6.0	46.1	40.9	39.6	32.4	21.0
Phi-3.5-Vision	23.4	20.6	18.9	20.9	15.2	42.2	29.8	36.2	20.3	14.4	57.0	50.4	43.8	34.0	18.1
Phi-4-Multimodal	23.7	25.1	26.0	23.4	20.0	44.3	43.7	54.3	39.8	14.6	65.5	67.6	63.6	61.8	43.6
NVILA-Lite-2B	11.0	5.8	9.7	5.7	5.7	13.9	11.4	13.2	8.0	7.9	31.3	29.7	26.9	23.3	18.1
NVILA-Lite-8B	16.9	24.5	13.4	12.0	9.9	26.2	24.9	20.8	15.6	16.2	49.5	47.7	43.1	37.2	35.6
Pixtral-12B	34.7	34.6	24.0	26.5	15.9	52.8	41.0	49.6	36.6	36.6	77.5	68.8	59.6	52.9	44.8
	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k	8k	16k	32k	64k	128k

Figure 26. Results of 46 models on the category DocVQA at various lengths.

Use the given documents to write a concise and short answer to the question about the entity shown in the image. Write your answer in the following format:
Answer: [answer]

Document (Title: Tropidacris collaris): Tropidacris collaris is a species of grasshopper in the family Romaleidae. A large South American grasshopper, it is also known as the blue-winged grasshopper although they vary greatly in coloration. It is common in both forests and dry areas of South America from Colombia to Argentina. In parts of northern Argentina, they are considered a pest. They are also popular among insect and terrarium enthusiasts.

Document (Title: Anarta myrtilli): [Warren] from Sintra, Portugal, the whole forewing is suffused with blackish, leaving only the white blotch on vein 2 conspicuous, and the orange of the hindwing, both above and below, is pale lemon yellow; as the insect is decidedly larger than average typical "myrtilli", it may prove a distinct species; at present I have seen only one - taken in the spring of 1909 by Mr N. C. Rothschild, and now in the Tring Museum.

Document (Title: Nipponaclerda biwakoensis): This species has become established (as of 2017) in the United States in the state of Louisiana, where it has rapidly become a serious pest of roseau cane, damaging over 80% of the reeds in some areas such as the Pass a Loutre Wildlife Management Area, where it is referred to by the older common name Phragmites scale insect or the more recently-coined name, roseau cane mealybug.

Document (Title: Dioctria atricapilla): The violet black-legged robber fly, Dioctria atricapilla, is a species of robber fly in the subfamily Dasypogoninae. This 9- to 12-millimeter long insect has a wingspan of roughly 7 to 9 mm and short, three-segmented antennae. It's a predatory insect, feeding mainly on smaller flies and predatory hymenopterans. It primarily thrives in grassland, and is seen from May to July.

Document (Title: Fauna of New Guinea): Notable endemic insect species include "Ornithoptera paradisea", "Ornithoptera chimaera", "Papilio weymeri", "Graphium weiskei", "Ideopsis hewitsonii", "Taenaris catops", "Parantica rotundata", "Parantica clinias", "Rosenbergia rufolineata", "Mecopus doryphorus", "Mecopus serrirostris", "Sphingnotus mirabilis", "Sphingnotus insignis", "Belionota aenea", "Poropterus solidus", "Poropterus gemmifer", "Aesernia splendens", "Aporhina bispinosa", "Eupholus petitii", "Eupholus bennetti", "Schizoeupsalis promissa", "Barystethus tropicus", "Eupholus geoffroyi", "Rhinoscapha loriai", "Rhinoscapha funebris", "Rhinoscapha insignis" "Alcides exornatus", "Alcides elegans", "Xenocerus lacrymans", "Arachnobas sectator", "Arrhenodes digramma", "Eupholus magnificus", "Mecopus bispinosus", "Callictita" spp.. Also known from New Guinea are "Batocera wallacei", "Ithystenus curvidens", "Meganthribus pupa", "Sipalinus gigas", "Pelargoderus rubropunctatus", "Rhynchophorus bilineatus", "Gasterocercus anatinus", "Acalolepta australis", "Actinus imperialis", "Megacrania batesii".

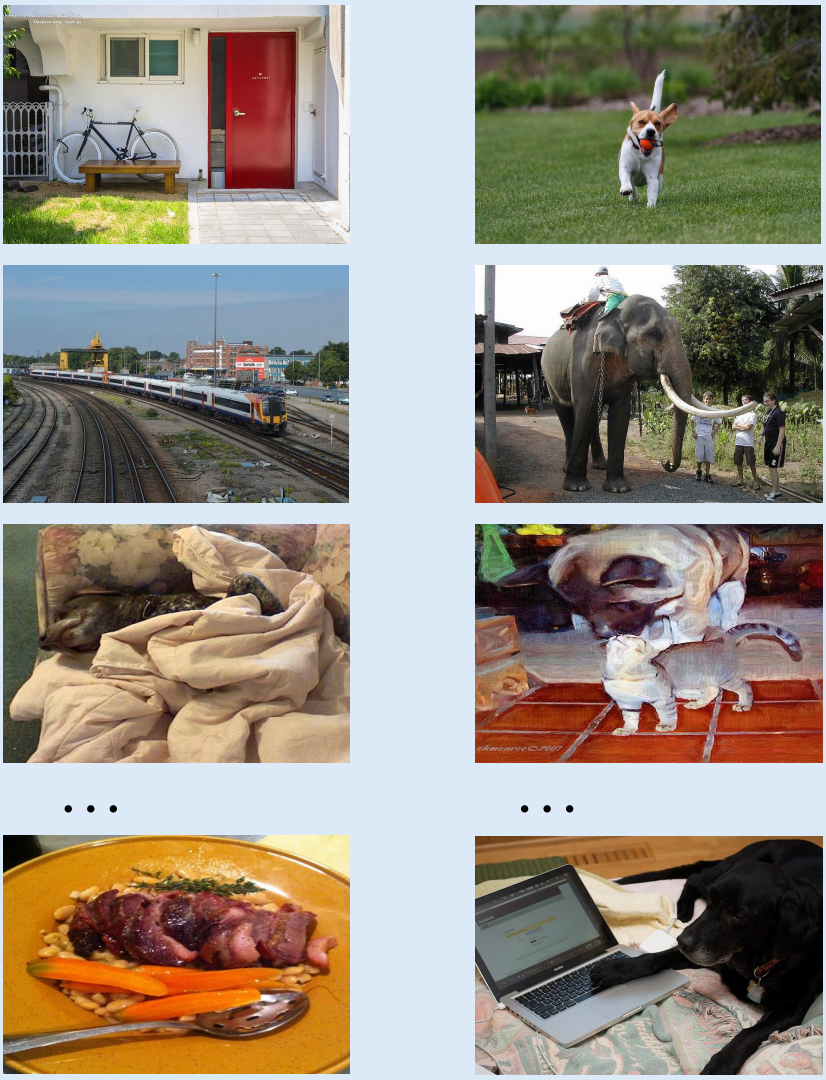
...

Document (Title: Melanopsis brevicula): Melanopsis brevicula is a small species of gastropod endemic to small streams near Agourai, Morocco. It is distinctive due to its minute size, flattened sculpture, low spire, and small aperture. It is known from a single location 10 km in area(Oued Ain Maarouf) which has been well surveyed, and found to be threatened by increasing human population, droughts of increasing extremity, water diversion, and pastoralization. Shell collecting presents a minor threat to populations. The species has been classified as Critically endangered by the IUCN.

Question:  Which place is this insect endemic to?

Figure 27. Example of InfoSeek dataset in the VRAG category.

You are given a set of images. Please answer the question in Yes or No based on the given images. Write your answer in the following format: Answer: [answer]



Question: For the image with an elephant, is there a dog?

Figure 28. Example of Visual Haystak-Single dataset in NIAH category. *Note: The input image list is shown in two columns for display clarity; in the actual input, the images are arranged in a single sequence.*

You are given interleaved text and images. Please answer the question with the option's letter (A, B, etc.) based on the given text and images. Write your answer in the following format:
Answer: [answer]

He also featured as Cooper in the worldwide tv show Game of Thrones in the episode “The Watchers on the Wall.”He is the first son of the union between Tim Roth and Nikki Butler. He was named Timothy Hunter Roth; Timothy after his father and the Hunter after the popular journalist Hunter S.

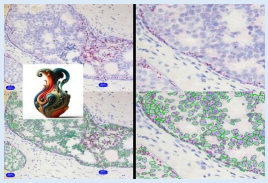
•••

I like my scones seasoned. Whether it's sweet or savoury, always add some salt to it. In terms of cooking, when you make your dough don't play with it. Just fold all the crumbs together, and it doesn't matter if it's bubbly. A lot of people play with the dough because they think it makes it smoother, but when a scone falls to pieces, it's because you've played with the dough too much."

He was a musician with original compositions, skilled at playing the guitar. In November, 2021 Cormac was diagnosed with 3 germ cell cancer. George of the Jungle Star Brendan Fraser and his family story
Pamela Adlon: The Stardom FamilySmart Watches – Honest review based on my experience

•••

Caretaker Sporting boss Tiago Fernandes said: ‘The players did exactly what I asked them to do. In our game plan we know we had to be rigorous and they were almost perfect on that. ‘We were aware of the opponent’s quality but we, knowing our capacity and being creative and aggressive with and without ball, could try to surprise here.’Bruce Willis has reprised his iconic role as John McClane for a new Die Hard video.



Question: Which of the following images appears in a certain image of the above document?

A. 

B. 

C. 

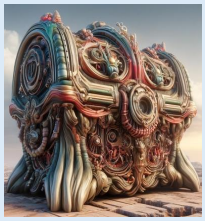
D. 

Figure 29. Example of MM-NIAH-Ret dataset in NIAH category.

You need to recognize entities in images. Use the provided mapping from the image to label to assign a label to the test image. Only output "label: {label}" and nothing else.

Training examples:

label: 2

label: 4

label: 0

...

label: 0

label: 3

label: 1

Now classify this image:

Figure 30. Example of the Stanford Cars dataset in the ICL category. *Note: The input image list is shown in three columns for display clarity; in the actual input, the images are arranged in a single sequence.*

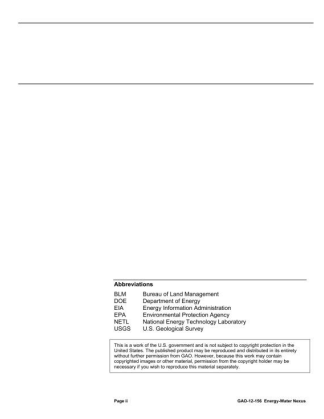
You are given a government report from U.S. Government Accountability Office (GAO), and you are tasked to summarize the report. Write a concise summary (around 550 words) organized in multiple paragraphs. Where applicable, the summary should contain a short description of why GAO did this study, what GAO found, and what GAO recommends.

Government Report:

Document gao-12-156 (page 0):

Contents

Document gao-12-156 (page 1):



Abbreviations

BLM	Bureau of Land Management
DOE	Department of Energy
EIA	Energy Information Administration
EPA	Environmental Protection Agency
NETL	National Energy Technology Laboratory
USGS	U.S. Geological Survey

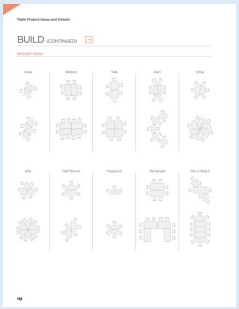
...

Now please summarize the report.

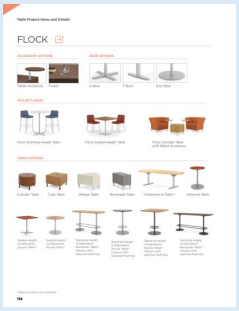
Figure 31. Example of GovReport in the summarization category. We only show two pages due to limited space.

You are given a document with text and images, and a question. Answer the question as concisely as you can, using a single phrase or sentence if possible. If the question cannot be answered based on the information in the article, write 'Not answerable.' Write your answer in the following format:
Answer: [answer]

Document 4057524 (page 113):



Document 4057524 (page 114):



Document 4057524 (page 115):

...

Question: Based on Document 4057524, answer the following question. Enumerate the available height-adjustable base options listed under "Coordinate" section.

Figure 32. Example of LongDocURL dataset in the DocVQA category. We only show three pages due to limited space.