Document Haystack: A Long Context Multimodal Image/Document Understanding Vision LLM Benchmark

Goeric Huybrechts Srikanth Ronanki Sai Muralidhar Jayanthi Jack Fitzgerald Srinivasan Veeravanallur
Amazon AGI
{huybrech,ronanks,saimucja,jgmf,srivee}@amazon.com

Abstract

The proliferation of multimodal Large Language Models has significantly advanced the ability to analyze and understand complex data inputs from different modalities. However, the processing of long documents remains underexplored, largely due to a lack of suitable benchmarks. To address this, we introduce Document Haystack[1][2], a comprehensive benchmark designed to evaluate the performance of Vision Language Models (VLMs) on long, visually complex documents. Document Haystack features documents ranging from 5 to 200 pages and strategically inserts pure text or multimodal text+image "needles" at various depths within the documents to challenge VLMs' retrieval capabilities. Comprising 400 document variants and a total of 8,250 questions, it is supported by an objective, automated evaluation framework. We detail the construction and characteristics of the Document Haystack dataset, present results from prominent VLMs and discuss potential research avenues in this area.

1. Introduction

Large Language Models (LLMs) [1–4] have fundamentally transformed natural language processing, demonstrating unprecedented capabilities in understanding and generating human-like text. This transformation has recently expanded with the emergence of multimodal LLMs, including GPT-4 [5], Gemini 1.5 [6], LLama 3.2 [7], Claude 3 [8], and Nova [9]. These advanced models excel at diverse tasks, from image captioning [10] to video reasoning [11] and speech translation [12].

A particularly significant advancement has been in document understanding capabilities. This field is crucial across numerous domains, with particular importance in legal [13, 14], medical [15, 16], and financial sectors [17, 18], where accurate document interpretation directly impacts decision-making and regulatory compliance. The integration of LLMs has markedly improved key aspects of document processing, including information extraction accuracy, contextual understanding and question-answering capabilities based on document content.

However, document understanding presents distinct challenges that extend beyond conventional text processing. Documents frequently contain complex elements including tables, charts, and various visual components. These multilayered documents require sophisticated systems capable of processing diverse format types while efficiently managing large volumes of unstructured information.

While recent advances are promising, the effectiveness of Vision Language Models (VLMs) in processing multimodal documents remains an active area of research. Their capability to comprehend and analyze lengthy documents is not fully established as current evaluation frameworks focus on short documents, limiting our understanding of VLM performance on longer and more complex document analysis tasks.

In this work, we address this need by introducing Document Haystack, a novel benchmark designed to evaluate the performance of VLMs, in retrieving key multimodal information from long documents. The benchmark features 400 document variants ranging from 5 to 200 pages, with pure *text* or multimodal *text+image* needles inserted at different depths and in diverse formats. These needles, formatted as key-value pairs (e.g., "The secret sport is "basketball"."., where "basketball" could be written as text or be illustrated as an image), are strategically placed to test the VLMs' ability to locate and extract specific information within extensive textual and visual contexts. Document Haystack provides a fully objective and comprehensive evaluation framework of 8,250 questions.

Results on our dataset indicate that VLMs achieve over 90% accuracy in textual extraction from 200-page documents, but their performance drops by ∼30% when the same documents are provided as images. Accuracy even falls to ∼40% when retrieving combined text-image information, highlighting significant room for improvement in processing long visual documents.

[1] Dataset: https://huggingface.co/datasets/AmazonScience/document-haystack
[2] Code: https://github.com/amazon-science/document-haystack

2. Related Work

The evaluation of LLMs has traditionally focused on short texts and specific tasks such as question answering, summarization, and translation, with early benchmarks like GLUE [19] and SuperGLUE [20] playing instrumental roles in assessing natural language understanding capabilities. However, while these foundational benchmarks have been valuable, they primarily consist of short texts and tasks that do not require long-term context retention.

As LLMs advanced, the need for evaluating longer text comprehension became apparent, leading to the development of specialized long context benchmarks. Notable examples include the Needle in a Haystack [21], which tests models' ability to retrieve information from extended contexts, and LongBench [22], which introduced the first bilingual, multi-task framework for assessing long-form text understanding.

The emergence of multimodal VLMs necessitated new evaluation approaches. Benchmarks such as VQA [23], NLVR [24], and MileBench [25] were developed to assess models' capabilities in processing and reasoning about combined visual and textual information. Document understanding then emerged as a distinct challenge in multimodal evaluation. While early datasets focused on specific elements like charts [26] or single-page analysis [27], recent benchmarks such as DUDE [28], Loong [29], SlideVQA [30], and MMLongBench-Doc [31] attempted to tackle multi-page document comprehension.

Despite these advancements, there remains a significant gap in benchmarks that evaluate VLMs' performance on long, multimodal documents. Existing benchmarks either (1) lack the document length (<50 pages), (2) rely on independently pre-extracted text and images rather than original documents, or (3) don't allow a performance comparison of a similar task on different document lengths. While recently released benchmarks like MM-NIAH [32] and M-LongDoc [33] address some of these issues, they have their own constraints. MM-NIAH's limited prompt length (<72k tokens) makes it unsuitable for long document evaluation, while M-LongDoc, despite featuring documents spanning hundreds of pages, offers only 851 questions, employs a non-objective evaluation method, and doesn't support evaluating the same task with varying document lengths. Moreover, both rely on pre-extracting text and figures from the documents. This preprocessing approach prevents VLM providers from utilizing their native processing mechanisms on original PDF documents, which would provide a more accurate indication of their real-world performance. Although they are valuable contributions, there is a clear need for additional benchmarks. Document Haystack aims to fill this void.

3. Benchmark Characteristics

Document Haystack is a comprehensive benchmark designed to evaluate the performance of VLMs in retrieving key multimodal information from long documents. The following sections detail the essential characteristics of our benchmark.

# Pages	**5**	**10**	**25**	**50**	**75**	**100**	**150**	**200**	**Total**
***text* needles**									
# Documents	25	25	25	25	25	25	25	25	200
# Questions	125	250	625	625	625	625	625	625	4125
***text+image* needles**									
# Documents	25	25	25	25	25	25	25	25	200
# Questions	125	250	625	625	625	625	625	625	4125
Total									
Total # Documents	50	50	50	50	50	50	50	50	**400**
Total # Questions	250	500	1250	1250	1250	1250	1250	1250	**8250**

Table 1. Document Haystack characteristics

Needle in a Haystack - Document Haystack extends the conceptual foundation of the Needle in a Haystack benchmark [21], where language models are tasked with retrieving a specific statement (the "needle") embedded within a long context window (the "haystack").

Document Selection and Lengths - Document Haystack (Table 1) comprises 25 publicly available financial 10-K reports, each originally exceeding 200 pages. These documents were selected to provide a realistic and challenging dataset for evaluating VLMs. To create varied testing conditions, each report was trimmed to different sizes: 5, 10, 25, 50, 75, 100, 150, and 200 pages. This variation allows for assessing the VLMs' performance across different context lengths. In total, our benchmark encompasses 400 document variants.

Needle Design - Document Haystack consists of two parallel benchmark sets, distinguished by their different needle design (*text* needles vs *text+image* needles) as to evaluate different aspects of VLMs' capabilities. These parallel sets maintain identical document structures and distributions while varying only in how their target information is presented, enabling direct comparisons of VLMs' performance across different modalities.

(1) *text* needles - The first set of our benchmark incorporates *text* needles. To evaluate the VLMs' capability to extract specific text information from long documents, *text* needles were inserted into each document as overlaid text. These needles are key-value pairs formatted as "The secret KEY is "VALUE".". For example, "The secret sport is "basketball"." (Figure 1). The needles were inserted at various pages and spatial positions within the documents, with diverse colors, sizes, and fonts to add complexity.

(2) *text+image* needles - The second set of our benchmark introduces *text+image* needles, maintaining identical positioning and typographical variations as the *text* needles. The key distinction lies in the representation of "VALUE" as an image rather than text (Figure 2). This modification

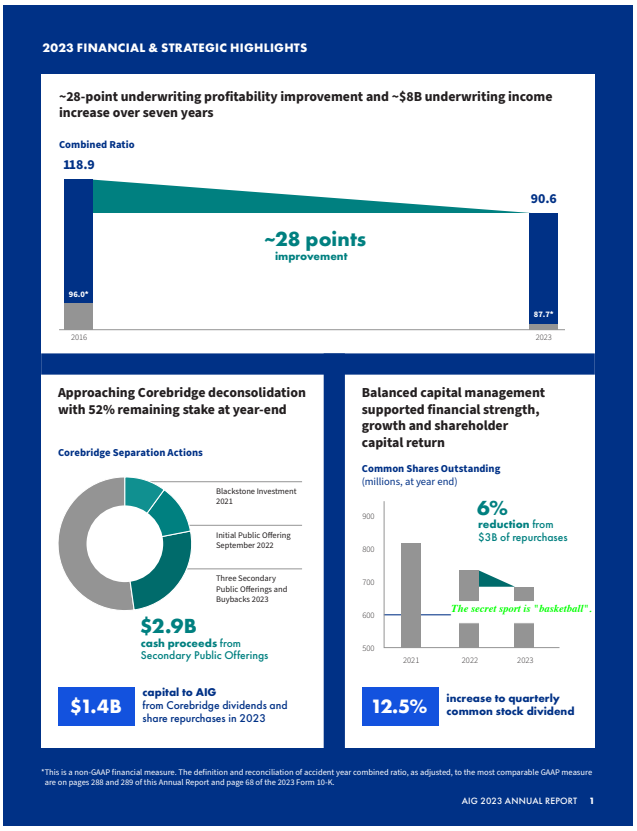


Figure 1. Example of a *text* needle: Document page containing the hidden text-based information "The secret sport is "basketball".".

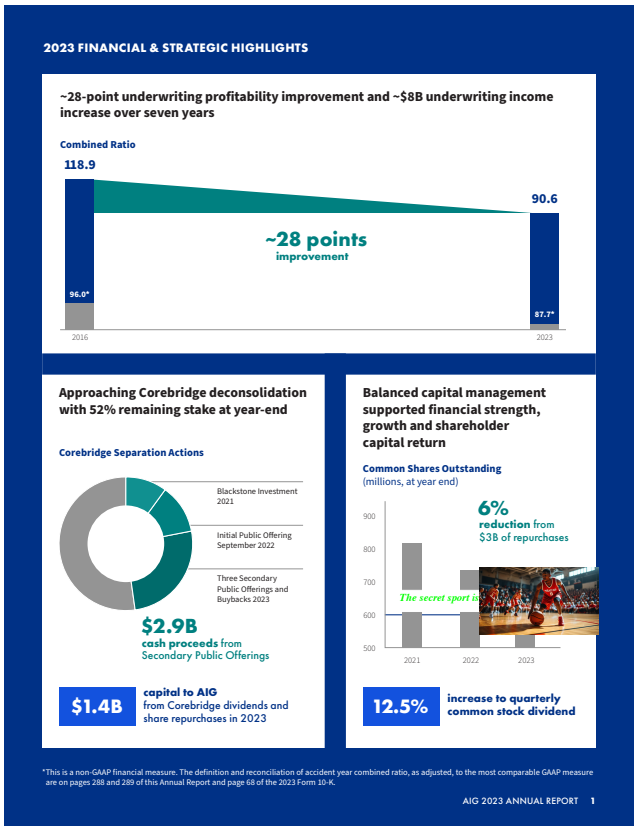


Figure 2. Example of a *text+image* needle: Document page containing the hidden multimodal information where the text "The secret sport is" is paired with an image of basketball.

enables assessment of VLMs' multimodal comprehension capabilities, requiring both textual and visual understanding for accurate information retrieval.

Multiple Formats - Document Haystack is available in three distinct formats (Table 2). First, both the *text* and *text+image* needle sets are provided in their original PDF format, preserving the complete document structure and formatting. Second, acknowledging that many VLM providers do not accept PDFs as input, we offer image versions of all documents, with each PDF page converted to a separate image file (200 DPI). Third, for scenarios requiring pure text processing, we provide a text-only version of the text needle set, extracted using the Python *PdfReader*[3] text extraction tool. This multi-format approach enables researchers to evaluate VLMs or text-only LLMs under various input conditions while maintaining benchmark consistency.

Needle Distribution - The distribution of needles was carefully designed to ensure they were placed at different depths within the documents. For documents with 5 and 10 pages, 5 and 10 needles were inserted, respectively. For documents with more than 25 pages, 25 needles were inserted. The pages for each needle were randomly selected from equal, non-overlapping page ranges. For instance, in a 100-page document, needle #1 would be inserted on a randomly chosen page between 1 and 4, needle #2 between pages 5 and 8, and so on.

Needle Sets - Five sets of 25 unique key-value pairs/needles were created (Table 3), spanning diverse categories such as objects, animals, instruments, and fruits. Each set was allocated to five different reports, ensuring a diverse and comprehensive evaluation. This setup allows for a robust assessment of the VLMs' ability to locate and extract the correct values. Each key-value pair has an associated question. The presence of multiple needles within each document serves as intentional distractors to add complexity.

Animal and Object Categories The inclusion of multiple animal and object sets (Animal #1-#5, Object #1-#5) serves a specific methodological purpose in evaluating VLMs' multimodal capabilities. While simpler categories like "drink" or "food" could potentially be answered by merely identifying relevant images in the document, animal and

[3]https://pypdf.readthedocs.io/en/stable/modules/PdfReader.html#the-pdfreader-class

Benchmark Set	Format	Description	Use Case
(1) *text* needles	PDF	Original document format	VLMs supporting PDF input
	Image	200 DPI page-wise images	VLMs requiring image input
	Text	Extracted plain text	Text-only LLMs
(2) *text+image* needles	PDF	Original document format	VLMs supporting PDF input
	Image	200 DPI page-wise images	VLMs requiring image input

Table 2. Document Haystack format variants

ID	Key	Values (Sets 1-5)
1	Shape	circle, triangle, rectangle, star, heart
2	Currency	euro, dollar, pound, rupee, ruble
3	Fruit	apple, banana, orange, grape, lemon
4	Vegetable	carrot, broccoli, onion, mushroom, cauliflower
5	Sport	basketball, tennis, boxing, skiing, surfing
6	Food	pizza, hamburger, sausage, fries, chocolate
7	Drink	coffee, tea, water, milk, smoothie
8	Instrument	guitar, piano, trumpet, drum, violin
9	Tool	hammer, wrench, saw, scissors, ruler
10	Transportation	car, boat, train, airplane, bike
11	Landmark	Eiffel Tower, Statue of Liberty, Taj Mahal, Colosseum, Big Ben
12	Kitchen appliance	blender, rice cooker, pan, toaster, microwave
13	Flower	rose, sunflower, tulip, lavender, daisy
14	Clothing	t-shirt, hat, glove, dress, sock
15	Office supply	pencil, paperclip, stapler, envelope, calculator
16	Animal #1	dog, cat, lion, elephant, giraffe
17	Animal #2	zebra, kangaroo, panda, koala, penguin
18	Animal #3	dolphin, shark, eagle, owl, spider
19	Animal #4	snake, frog, turtle, horse, cow
20	Animal #5	pig, bear, wolf, rabbit, squirrel
21	Object #1	book, table, chair, door, clock
22	Object #2	lamp, phone, key, watch, bottle
23	Object #3	spoon, fork, knife, plate, bowl
24	Object #4	umbrella, tree, bed, mirror, pillow
25	Object #5	comb, toothbrush, towel, candle, vase

Table 3. Five sets of 25 unique key-value pairs, where each value serves as a "needle" that models must locate when prompted with its corresponding key.

object categories require models to process both textual and visual information simultaneously. For instance, when multiple animal images appear in a document, the model must specifically identify which image corresponds to the "secret Animal #3" prompt, rather than simply detecting any animal image. This design ensures that VLMs truly combine their understanding of both modalities rather than relying solely on visual detection. These more challenging categories are exclusively used in documents of 25 pages or longer.

Evaluation Methodology - The purpose of Document Haystack is to retrieve the value associated with the secret key within the document. The evaluation question posed to the VLMs is: "What is the secret KEY in the document?". To determine whether the VLM successfully found the needle, the response is lowercased and searched for the VALUE associated with the KEY. For image needles, we developed a comprehensive alias list to accommodate valid alternative responses (e.g., "phone" vs. "mobile," "fries" vs. "chips"). This straightforward approach ensures clear and objective assessment criteria across the benchmark's total of 8,250 questions.

Objective - The overarching goal of Document Haystack is to provide a comprehensive and challenging benchmark for evaluating the long context capabilities of multimodal LLMs, namely VLMs. By simulating real-world document processing scenarios and inserting specific information needles, the benchmark offers a rigorous test of an VLM's ability to accurately retrieve key multimodal information from lengthy and visually complex documents.

Release - The dataset and benchmarking code are publicly available at https://huggingface.co/datasets/AmazonScience/document-haystack and https://github.com/amazon-science/document-haystack.

4. Results

In this section, we present the accuracy results from evaluating three prominent VLMs on our Document Haystack benchmark. We selected Nova, Gemini, and GPT as they are currently the only VLMs capable of processing more than 25 pages/images simultaneously. Other prominent VLMs have more limited image processing capabilities - for instance, Pixtral[4] [34] can only process 8 images at once, while Claude[5], despite providing support for 100 low-resolution images, is restricted to 20 images per API request for images larger than 2000x2000px.

Of these model families, we selected the following specific models for the benchmark evaluations: (1) Nova Lite, (2) Gemini Flash-2.0, and (3) GPT-4o-mini.

We report results across different formats (Table 2) of our benchmark to assess various capabilities of the VLMs. Specifically, we evaluate the models on following three sets:

1. **Text needles retrieval from document images** (4.2): This set assesses the models' visual capabilities in retrieving textual information from documents presented as images.
2. **Text needles retrieval from parsed document text** (4.3): This set offers an additional comparison by evaluating the models' capability to retrieve the same *text* needles from preprocessed, parsed document text.
3. **Text+Image needles retrieval from document images** (4.4): This set assesses the models' visual capabilities in retrieving multimodal information from documents presented as images.

[4]https://docs.mistral.ai/capabilities/vision

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

Model	Avg. tokens/image	#Pages							
		5	10	25	50	75	100	150	200
Nova Lite	1,626	8,130	16,260	40,650	81,300	121,950	162,600	243,900	325,200
Gemini 2.0-Flash	258	1,290	2,580	6,450	12,900	19,350	25,800	38,700	51,600
GPT-4o-mini	636	3,180	6,360	15,900	31,800	-	-	-	-

Table 4. Average image token consumption per model across different page lengths.

Model	#Pages							
	5	10	25	50	75	100	150	200
Nova Lite	**100.0**	**98.8**	85.0	76.6	**72.5**	**69.6**	**64.5**	**62.9**
Gemini Flash-2.0	83.2	74.8	82.7	64.0	63.2	58.4	46.9	51.8
GPT-4o-mini	96.0	98.0	**89.3**	**86.1**	-	-	-	-

Table 5. Accuracy on the **Text needles retrieval from document images** set - The task is to retrieve *text* needles from documents provided a query (e.g., "What is the secret fruit in the document?"). All documents are converted to images prior to VLM processing. The table shows the retrieval accuracy across different document lengths of three API providers: Nova Lite, Gemini Flash-2.0, and GPT-4o-mini.

Model	#Pages							
	5	10	25	50	75	100	150	200
Nova Lite	**100.0**	**100.0**	98.9	95.2	94.6	93.9	**94.1**	89.9
Gemini Flash-2.0	99.2	99.6	**99.5**	97.8	**96.8**	97.1	91.5	**91.8**
GPT-4o-mini	**100.0**	**100.0**	97.9	**98.4**	96.6	**97.5**	-	-

Table 6. Accuracy on the **Text needles retrieval from parsed document text** set - The task is to retrieve *text* needles from documents provided a query (e.g., "What is the secret landmark in the document?"). All documents are converted to text prior to VLM processing. The table shows the retrieval accuracy across different document lengths of three API providers: Nova Lite, Gemini Flash-2.0, and GPT-4o-mini.

Model	#Pages							
	5	10	25	50	75	100	150	200
Nova Lite	**84.0**	**84.0**	61.4	52.2	43.5	38.9	34.9	37.0
Gemini Flash-2.0	53.6	52.0	**67.4**	**56.8**	**48.6**	**43.5**	**37.9**	**38.7**
GPT-4o-mini	43.2	36.4	39.4	26.9	-	-	-	-

Table 7. Accuracy on the **Text+Image needles retrieval from document images** set - The task is to retrieve *text+image* needles from documents provided a query (e.g., "What is the secret flower in the document?"). All documents are converted to images prior to VLM processing. The table shows the retrieval accuracy across different document lengths of three API providers: Nova Lite, Gemini Flash-2.0, and GPT-4o-mini.

We exclude PDF format results from our analysis due to limited direct PDF support among current VLM providers. However, as VLMs evolve to handle PDFs natively, this format presents an interesting future evaluation dimension. The PDF processing approaches of the various providers, whether converting documents to images, extracting text and visual elements separately, or employing hybrid methods, could significantly impact performance. Such comparisons would reveal not only the models' core capabilities but also the effectiveness of their document preprocessing pipelines, providing valuable insights into which architectural choices best preserve and utilize document structure and formatting. This additional evaluation axis could help guide the development of more sophisticated document processing strategies in future VLM implementations.

4.1. Token Count

Table 4 presents the average token generation per image across the different VLMs, revealing significant variations in token consumption. Nova Lite generates approximately six times more tokens per image than Gemini (1,626 vs. 258 tokens). This difference becomes substantial when processing longer documents: for a 200-page document, Nova Lite processes 325,200 tokens compared to Gemini Flash-2.0's 51,600 tokens, while GPT-4o-mini theoretically falls in between at 127,200 tokens.

These token consumption patterns could significantly impact model performance, a topic that will be further explored in the following sections. Higher token counts, as seen in Nova Lite, may enable more detailed image

information extraction but could potentially complicate specific information retrieval due to the longer context window that must be processed. Conversely, lower token counts, as demonstrated by Gemini, might offer more efficient processing but could miss subtle image details.

4.2. Text Needles from Document Images

The performance metrics for the **Text needles retrieval from document images** set are presented in Table 5. Our analysis reveals a consistent pattern across all models: accuracy deteriorates as document length increases. This degradation reaffirms the fundamental challenge in VLM processing: the inverse relationship between input context size and model performance.

Nova Lite and GPT-4o-mini demonstrate superior performance, achieving similar accuracy scores. However, GPT-4o-mini has a notable limitation: unlike the other models, it cannot process documents exceeding 50 pages. Nova Lite surpasses Gemini Flash-2.0 across all document lengths. Gemini's significantly lower token count per image, while potentially more efficient in terms of processing, may have inadvertently omitted crucial information necessary for effective information retrieval. Other model-specific factors such as memory management strategies may also contribute to these performance differences.

Next, we report accuracy results across ten equidistant document depth ranges, each spanning 10% of the document length. Due to our systematic needle distribution strategy, each depth range contains roughly a similar number of needles/questions ($\sim$63), enabling consistent comparison across depths. Since Nova and Gemini are the only models capable of processing documents up to 200 pages, we present detailed heatmap visualizations for these two models only in Figures 3a and 3b, respectively. While there are some slight variations, both models maintain relatively consistent performance across document depths.

4.3. Text Needles from Parsed Document Text

Table 6 presents performance metrics for the **Text needles retrieval from parsed document text** set, which establishes an upper bound for the *Text needles retrieval from document images* set. The task of retrieving text information from pure text is inherently less challenging than extracting the same information from images, as it eliminates the complexity of visual processing [32, 35].

The results demonstrate consistent performance across models and high accuracy across different context lengths, underscoring the models' remarkable capability to handle long context text inputs. It's worth noting that we encountered technical limitations with GPT-4o-mini for the 150 and 200-page sets, as these documents exceeded the API provider's token limit. Comparing these results with those in Table 5 reveals a significant performance gap between text-based and image-based text information retrieval. This difference illustrates the challenges and potential areas for advancement in understanding document text from images as opposed to pre-extracted text, highlighting the need for further research and development in this domain.

4.4. Text+Image Needles from Document Images

Table 7 presents results from our most challenging benchmark set, **Text+Image needles retrieval from document images**, which reveals distinct performance patterns across document lengths. For shorter documents (5-10 pages), Nova Lite demonstrates remarkable superiority, outperforming other models by an absolute margin of 32+%, while GPT-4o-mini shows the lowest accuracy. However, this performance hierarchy shifts with increasing document length: Gemini Flash-2.0 emerges as the leading model for longer documents, surpassing Nova Lite, albeit by a small margin.

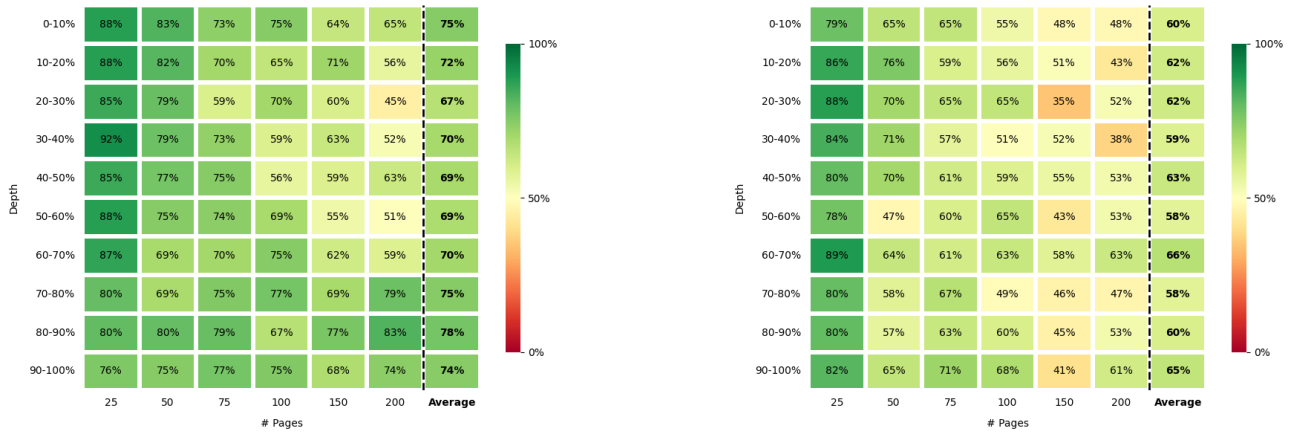
This performance pattern yields an interesting insight regarding token consumption. While Gemini's lower token count appeared to disadvantage it in the *Text needles retrieval from document images* set, it proves beneficial for *image* needle extraction. Although differences in architecture, training, inference strategies, and other model aspects likely contribute, we hypothesize that this difference stems from the nature of the information being processed: *text* needles in images require fine-grained detail preservation as they form a more subtle detail within the document page, whereas *image* needles present more readily distinguishable visual elements. Consequently, Gemini's more compact token representation appears well-suited for image-based needle tasks, potentially contributing to its slight superior performance with longer documents.

The heatmap analysis of our most challenging benchmark set reveals distinct patterns between models. Figures 4a and 4b present detailed performance visualizations for Nova Lite and Gemini Flash-2.0, respectively. Nova Lite exhibits a mild manifestation of the "lost in the middle" phenomenon described in [36], with slightly decreased accuracy in middle-depth ranges. In contrast, Gemini Flash-2.0 maintains consistent performance across all depth ranges, showing no discernible pattern of accuracy variation with document depth.

5. Further Work

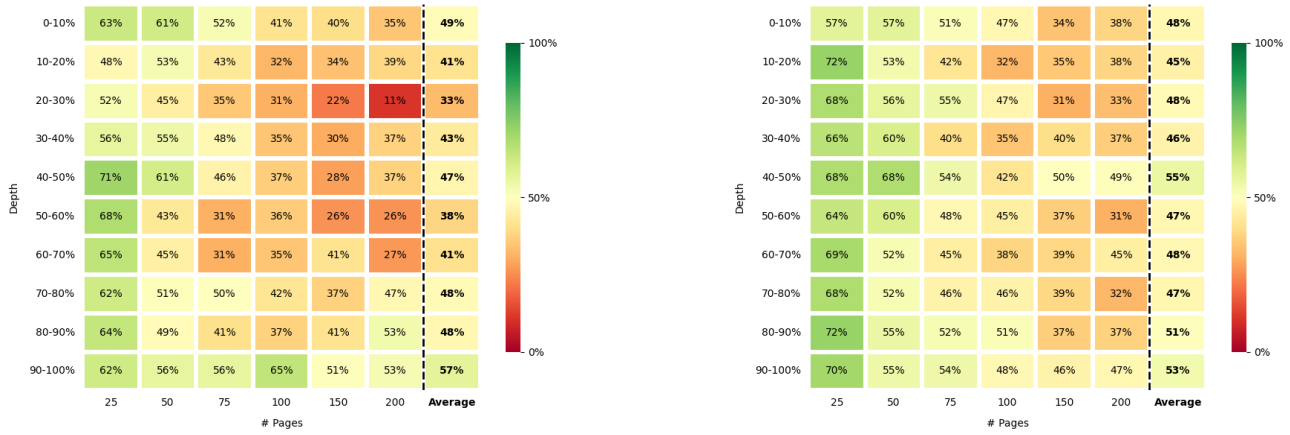
5.1. Document Haystack Latency

While our study emphasizes accuracy measurements, researchers could utilize our dataset to evaluate the latency characteristics of different VLMs in processing long documents. Such analysis would provide valuable insights into the computational efficiency and practical deployability of these models, particularly important for real-world applica-



(a) Nova Lite (b) Gemini Flash-2.0

Figure 3. Performance analysis on the **Text needles retrieval from document images** set across document depths: Accuracy scores for different models evaluated over ten equidistant depth ranges, each containing approximately 63 needles.



(a) Nova Lite (b) Gemini Flash-2.0

Figure 4. Performance analysis on the **Text+Image needles retrieval from document images** set across document depths: Accuracy scores for different models evaluated over ten equidistant depth ranges, each containing approximately 63 needles.

tions where response time is crucial. This analysis could also reveal important trade-offs between model accuracy and response time, helping practitioners make informed decisions based on their specific application requirements.

5.2. Advancing VLMs' Long Context Processing

Our benchmark results reveal a significant disparity in VLMs' performance between textual and visual information processing over extended contexts. While models demonstrate robust capabilities in handling long context textual information, their performance in identifying and extracting image-based information deteriorates as document length increases. Additionally, the accuracy of extracting multimodal information is notably lower compared to extracting pure textual information from document images. These performance gaps highlight a critical area for improvement in next-generation VLMs.

The challenge likely stems from the increased complexity of maintaining and processing rich visual information over longer sequences, as opposed to textual information. Future research should focus on developing more efficient architectures, training approaches, and inference algorithms specifically designed to maintain visual context over extended sequences. This might include novel attention mechanisms, better visual tokenization strategies, or more sophisticated methods for managing the interaction between textual and visual modalities in long documents. Addressing these limitations is crucial as real-world applications often require processing lengthy documents containing both text and images, from technical manuals to medical records.

5.3. Extended Analysis Framework

While our primary contribution is a benchmark dataset for long context multimodal retrieval challenges, we recognize the importance of enabling more fine-grained analysis tasks. To facilitate this, our dataset release includes comprehensive metadata for each needle, comprising: (1) page location within the document, (2) precise X-Y coordinates on the page, (3) image size, (4) color specifications, and (5) font type and size.

This rich metadata will help support various research directions beyond basic retrieval, such as location-aware information extraction, spatial relationship analysis, visual attribute understanding and document layout comprehension. These additional data points provide researchers with the flexibility to design and evaluate more sophisticated tasks that reflect different real-world document processing challenges. We envision this expanded functionality will drive innovation in both model architectures and evaluation methodologies for long context multimodal understanding.

6. Conclusion

Document Haystack marks a significant advancement in the evaluation of Visual Language Models (VLMs) by providing a comprehensive benchmark for assessing their performance on long, visually complex documents. Our findings highlight some limitations of current models, underscoring the need for continued research and development. We believe that Document Haystack will serve as a valuable resource for the research community, driving innovation and progress in the field of multimodal document understanding and paving the way for more effective long context VLMs.

References

[1] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 1

[2] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

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

[4] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1

[5] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[6] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1

[7] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 1

[8] Anthropic. Introducing the next generation of claude: Claude 3 model family. *https://www.anthropic.com/news/claude-3-family*, 2024. 1

[9] Amazon Artificial General Intelligence. The amazon nova family of models: Technical report and model card. *Amazon Technical Reports*, 2024. 1

[10] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[11] Sullam Jeoung, Goeric Huybrechts, Bhavana Ganesh, Aram Galstyan, and Sravan Bodapati. Adaptive video understanding agent: Enhancing efficiency with dynamic frame sampling and feedback-driven reasoning. *arXiv preprint arXiv:2410.20252*, 2024. 1

[12] Karel Mundnich, Xing Niu, Prashant Mathur, Srikanth Ronanki, Brady Houston, Veera Raghavendra Elluru, Nilaksh Das, Zejiang Hou, Goeric Huybrechts, Anshu Bhatia, et al. Zero-resource speech translation and recognition with llms. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 1

[13] Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36:44123–44279, 2023. 1

[14] Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *CoRR*, 2023. 1

[15] Cheng Peng, Xi Yang, Aokun Chen, Kaleb E Smith, Nima PourNejatian, Anthony B Costa, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, et al. A study of generative large language model for medical research and healthcare. *NPJ digital medicine*, 6(1):210, 2023. 1

[16] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023. 1

[17] Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. Large language models in finance: A survey. In *Proceedings*

of the fourth ACM international conference on AI in finance, pages 374–382, 2023. 1

[18] Zhiyu Zoey Chen, Jing Ma, Xinlu Zhang, Nan Hao, An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian McAuley, Linda Petzold, and William Yang Wang. A survey on large language models for critical societal domains: Finance, healthcare, and law. *arXiv preprint arXiv:2405.01769*, 2024. 1

[19] Alex Wang. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018. 2

[20] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019. 2

[21] Greg Kamradt. Needle in a haystack-pressure testing llms. *Github Repository*, page 28, 2023. 2

[22] Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*, 2023. 2

[23] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 2

[24] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018. 2

[25] Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. Milebench: Benchmarking mllms in long context. *arXiv preprint arXiv:2404.18532*, 2024. 2

[26] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 2

[27] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 2

[28] Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, et al. Document understanding dataset and evaluation (dude). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19528–19540, 2023. 2

[29] Minzheng Wang, Longze Chen, Fu Cheng, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, et al. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5627–5646, 2024. 2

[30] Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13636–13645, 2023. 2

[31] Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *arXiv preprint arXiv:2407.01523*, 2024. 2

[32] Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, et al. Needle in a multimodal haystack. *arXiv preprint arXiv:2406.07230*, 2024. 2, 6

[33] Yew Ken Chia, Liying Cheng, Hou Pong Chan, Chaoqun Liu, Maojia Song, Sharifah Mahani Aljunied, Soujanya Poria, and Lidong Bing. M-longdoc: A benchmark for multimodal super-long document understanding and a retrieval-aware tuning framework. *arXiv preprint arXiv:2411.06176*, 2024. 2

[34] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024. 4

[35] Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, and Heng Tao Shen. Cross-modal retrieval: A systematic review of methods and future directions. *arXiv preprint arXiv:2308.14263*, 2023. 6

[36] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. 6