

# Context or Retrieval? Evaluating RAG Methods for Art and Museum QA System

**Samuel Ramos-Varela, Jaime Bellver-Soler, Marcos Estecha-Garitagoitia,  
Luis Fernando D’Haro**

Speech Technology and Machine Learning Group (THAU) - ETSI de Telecomunicación  
Av. Complutense 30, 28040, Madrid, Spain - Universidad Politécnica de Madrid  
{s.rvarela, jaime.bellver, marcos.estecha.garitagoitia}@upm.es,  
{luisfernando.dharo}@upm.es

Corresponding author: [s.rvarela@upm.es](mailto:s.rvarela@upm.es)

## Abstract

Recent studies suggest that increasing the context window of language models could outperform retrieval-augmented generation (RAG) methods in certain tasks. However, in domains such as art and museums, where information is inherently multimodal, combining images and detailed textual descriptions, this assumption needs closer examination. To explore this, we compare RAG techniques with direct large-context input approaches for answering questions about artworks. Using a dataset of painting images paired with textual information, we develop a synthetic database of question-answer (QA) pairs for evaluating these methods. The focus is on assessing the efficiency and accuracy of RAG in retrieving and using relevant information compared to passing the entire textual context to a language model. Additionally, we experiment with various strategies for segmenting and retrieving text to optimise the RAG pipeline. The results aim to clarify the trade-offs between these approaches and provide valuable insights for interactive systems designed for art and museum contexts.

## 1 Introduction

The role of interactive question-answering (QA) systems in cultural activities, specifically art and museum contexts has grown significantly in recent years with the widespread of conversational chatbots (Schaffer et al., 2022). These systems must navigate the complexities of multimodal data, combining visual and textual elements to provide accurate and contextually relevant responses. Recent advances in Large Language Models (LLMs) with extended context windows (An et al., 2024) suggest new possibilities for such applications, potentially reducing reliance on classical retrieval mechanisms, but a series of problems arise due to the nature of these systems.

Along these advancements, retrieval-augmented generation (RAG) remains a promising approach,

especially for domains where information density exceeds the limits of even the largest context windows (Laban et al., 2024). Inspired by (Li et al., 2024), the present work investigates the trade-offs between RAG methods and direct large-context approaches for QA tasks in art and museum contexts, where the information is inherently multimodal and domain-specific.

By translating this visual context into detailed captions, and combining it with the original data, we will create a synthetic dataset for the evaluation different RAG pipelines. Comparing their outputs to a baseline generated using a large-context model, we aim to provide actionable insights for developing interactive systems in this domain.

Section 2 will perform an analyses of large-context models, RAG and their place within the landscape, as well as hot topics within the research community. We will lay down the foundations of the methods in Section 4, outlining the methodology and introducing the dataset we used for the experimentation. Section 5 covers the experiments performed with the different configurations and parameters, and finally a conclusion is given in Section 6, along with a few lines of research that might be interesting to explore.

## 2 Related

Retrieval-augmented generation (RAG) has emerged as a key paradigm for improving the performance of language models in tasks requiring external knowledge (Mansurova et al., 2024) and to reduce hallucinations (false or nonsensical outputs). Prior works have acknowledged the efficacy of RAG in both open and domain-specific applications, ranging from technical documentation to healthcare QA systems (Amugongo et al., 2024), (Wang et al., 2024b). Multimodal data comes as well with its own set of challenges (Chen et al., 2025). However, research on its applicability

in domain-specific tasks, such as art and museum contexts interaction, remains sparse.

Parallel to these developments, advancements in large-context models (even in the range of millions of tokens) have enabled direct ingestion of extensive input, offering an alternative to RAG for tasks involving dense information. Multimodal large language models (MLLMs) research has also been a hot topic lately (Jiang et al., 2024). With this in mind, the need for a representation of the visual context in the RAG systems arises, specially in visual domain applications (Yu et al., 2024).

Recent research builds around the use of RAG in conjunction with LLMs (Fan et al., 2024). Specifically in the RAG landscape, the chunking strategies of the indexed documents remains a research topic (Qu et al., 2024), as well as the embedding techniques for retrieval tasks (Warner et al., 2024). This paper builds on these foundations by comparing RAG with large-context approaches in a multimodal setting where the visual context is inferred to textual data, focusing on the art and museum domain.

### 3 Data preparation

When considering the interaction that visitors in a museum will have with a corporate chatbot, we can consider mainly two types of questions: those related with direct visual elements in the artwork, and those that transcend the artwork (i.e., name of artist, year, movement, technique, history, etc.). While designing our methodology, we consider both types of questions and the information we can have available: the images of the artworks and text describing the artwork (information usually available near the artwork or provided by the museum). Therefore, the dataset we used initially consists of painting images paired with detailed information sourced from museum databases and archives. The information that comes along the paintings at times references the author of the painting, and its historical context, but it rarely contains detailed descriptions of the painting's visual characteristics. For this reason, each image is processed using a vision language model to generate a descriptive caption, representing its visual context, divided in subject (depicted object(s)), form (visual organization: color, style, shape) and content fields (impact or meaning) (DiMaggio, 1987) (art, 2024). This caption is then concatenated with the original textual description to create a comprehensive context.

Using a language model, we developed a synthetic database of question-answer (QA) pairs based on this multimodal context. The dataset includes 1020 images and 12240 QA pairs, covering diverse artistic styles, historical periods, and cultural themes. This means that 12 questions are generated for each work of art. We wanted the questions and answers to be diverse, so we ask the large language model to formulate and answer 6 questions that a beginner user will ask, and other 6 as an expert in art. Furthermore, for each level, 3 questions are generated based on the original textual data, and the other 3 are based on the descriptive caption generated. This multimodal dataset serves as the foundation for evaluating both RAG and large-context approaches.

### 4 Methods

This section outlines the methodology developed and applied in this study to evaluate retrieval-augmented generation (RAG) approaches in contrast with large-context models in a multimodal art synthetic question-answer (QA) dataset.

First, a detailed caption of the image is generated with the "Qwen2-VL-7B-Instruct" model, a vision model with a size of 7B parameters (Wang et al., 2024a).

The RAG pipeline begins by segmenting the text provided by the museum related to each artwork, using several strategies. **Recursive splitting** splits the text based on the structure of its paragraphs, first splitting by double newline, then by newline and finally by period (Narimissa and Raithele, 2024). In the **semantic chunking** approach we divide documents based on semantic similarity, using the "all-MiniLM-L6-v2" (Transformers, 2021) embedding model we compute the similarity between sentences and group them together (Qu et al., 2024). **Sentence-based division** chunks texts into accurate and contextually meaningful chunks with the spacy tokenizer (Vasiliev, 2020), maintaining the integrity of sentences and their relationships within the text.

The resulting chunks will then be embedded using three different embedding types (dense, sparse and late interaction). Nomic's "modernbert-embed-base" (Nussbaum et al., 2024) model (embedding length of 768) is used as the dense embedding with cosine distance for comparisons. Contextualized Late Interaction over BERT (ColBERT) (Khattab and Zaharia, 2020) introduces a late interaction

architecture that encodes the query and the document using BERT and then employs an interaction step that models the similarity. Finally for the sparse representation, we will use "BM42" (qdr, 2024), which builds upon the foundations of TF-IDF (Term Frequency-Inverse Document Frequency).

The alternative approach involves feeding the full artwork context, i.e., comprising the original text information and the Qwen generated description captions, directly into a LLaMA 3.1 model of size 8B parameters (et al., 2024). We chose this model for its performance in "needle in a haystack" benchmarks, for its usage of context length and relatively small number of parameters (Hsieh et al., 2024). This model generates answers based on the entire input without relying on intermediate retrieval steps. The model is instructed to base its answers solely on the input text, to prevent hallucination, or answers already learned (see Annex A.3 for the prompt).

Synthetic QA pairs were created using "gpt-4-turbo" model with 128k of context window (OpenAI and et al., 2024). With the original text plus the generated captions as input, we generate pairs of synthetic questions along their corresponding answers by prompting the GPT-4-turbo model both the question and the answer (see Annex A.1 for the prompt). A random sample of around 1200 QA (aprox. 10% of the dataset) pairs were manually inspected to verify the quality of the generated dataset. These generated QA pairs will be then taken as ground truth, providing a benchmark baseline for evaluating the similarity of the answers generated by both RAG and large-context approaches.

We will then, for each question in the synthetic dataset, test two approaches: (a) first, retrieve the most relevant documents based on the cosine similarity score between the question embedding and different types of chunk embeddings., and (b) compare the performance against the responses of LLaMA3.1-8B with the full-context approach. In both cases, we will embed the original answer from the qa pairs, the full-context (LLaMA3.1-8B) generated answer and the RAG retrieved answer with the aforementioned "modernbert-embed-base" model, and compute cosine similarity between both generated answers and the ground truth answer. This metric provides a robust framework for comparing RAG outputs with the baseline answers vs the large-context model approach.

## 5 Experimentation

We conducted experiments using a multimodal synthetic QA dataset with the goal of evaluating the effectiveness of retrieval-augmented generation (RAG) versus large-context models in the task of properly answering questions about paintings. The experiments were run on a machine with an NVIDIA GeForce RTX 3090 (24GB of VRAM) using Python and the transformers library.

The RAG pipeline was built with the qdrant vector search library, for its ease of use and extensive community and support (Qdr, 2024). The final dataset consists of 1020 images with textual information and captions detailedly describing them. As it was previously described, each one of the images has 12 QA pairs associated generated with GPT4 (see Annex A.2). This synthetic dataset adds up to  $1020 \times 12 = 12,240$  total queries and their answers.

The Context-Based answers were generated by feeding the full context (original descriptions + Qwen captions) into the LLaMA 3.1-8B model. The model was instructed to base its responses exclusively on the provided context, mitigating potential hallucination (see Annex A.3). Table 3 shows the Context- mean similarity between the generated answers and the Ground Truth answers for the different groups of expertise and types of questions (about the description or original information). Annex A.2 showcases the generated answers.

Table 3 shows the results of the experimentation, showcasing the three chunking strategies, the three types of embeddings used and comparing the similarities with the baseline similarity.

As previously said, we designated several chunking strategies to segment the texts, depending on where we want to put the focus. Depending on the embedding size and structure of the data, the documents must be segmented in proper chunk sizes (number of characters in each chunk). As highlighted in Table 1 our dataset documents contain around 500-700 words each, amounting to 3500-4500 characters.

For the recursive splitting we set a maximum chunk size of 300 characters with an overlap of 50, which means that chunks will be as big as 300 characters, with the first 50 being from the previous chunk, and the last 50 from the next chunk. When performing the sentence-based chunking strategy, we state a chunk size of 200 characters that will leave room for adjoint short sentences to be grouped. If a sentence is bigger than 200 char-

	Mean Word Count	Mean Char Count
<b>Information</b>	398.85	2433.33
<b>Caption</b>	235.83	1424.96
<b>Total</b>	634.68	3858.28

Table 1: Mean word and character count in both the original information and the generated caption of each painting.

acters, it will still be treated as a chunk. Finally for the semantic approach, we will use a chunk size of 300 characters. Note that in semantic embedding, if a sentence is not similar enough to any other one, it will remain as a standalone chunk. Table 2 shows the number of chunks in which each chunking strategy divided the text corpus.

The "BM42" Sparse embedding model, grounded in TF-IDF principles, was evaluated for its efficiency in representing keyword-based similarities. Sparse embeddings excel in retrieval speed and computational simplicity, particularly for short, keyword-rich contexts, and here it's no different, it performs better with the caption texts, as it is more concise and contained. Particularly in "expert" questions with short answers containing keywords it performed best. On the other hand when facing more general or open ended questions, specially with questions regarding the original information which is inherently of diverse nature, the BM42 model achieved worse results.

We use the "modernbert-embed-base" model, which produces 768-dimensional embeddings optimized for capturing semantic similarity. These dense embeddings performed better in scenarios where the question and context were long. This time, the model performed well with both "expert" and "beginner" questions, but questions about the original information give worse results, in favor of questions about the captions.

In our experiments, ColBERT performance falls behind its two other opponents, but surprisingly its results are independent of the type (caption or information) and level (expert or beginner) of the question queried. The results show that the similarity is maintained along all of the questions no matter their structure, length or label.

Table 3 shows the different chunking and embedding configurations, along the computed similarity against the Ground Truth answers.

## 6 Conclusions and Future Work

After experimenting with different chunking approaches and parameters, the strategies chosen to

test on this dataset are: Semantic chunking with a chunk size of 300 characters, spacy sentence-based segmentation with a size of 200 characters, and recursive splitting with a maximum chunk size of 300.

The dense embedding approach with modernBERT generally works better than its two opponents, scoring higher regardless of the chunking mechanism used. At the same time its variance is higher than the other two, scoring higher in questions derived from the captions versus the original information. The late interaction ColBERT model on the other hand, although performing slightly worse, keeps higher consistency, with low variance along the different labeled fields. At last, BM42 falls right in between the other two embedding models. As showcased in Table 3 it performs better than ColBERT with Semantic chunking and Sentence chunking, and maintains a slight advantage in Recursive chunking.

On the comparison of RAG results with the baseline LLaMA answers, we can definitely say that we proved what was stated in (Li et al., 2024), in which they also found large-context model to outperform RAG based approaches in accuracy. Being this said, the approach that consistently closes the gap the most is the sentence-based document chunking strategy combined with modernBERT embeddings. Regarding the dataset, can observe trends that suggest that the most precise answers are generated for questions about the "Description" of the artwork. A slight advantage is present in the "expert" questions, over the "beginner" ones.

It must be said that we are measuring the raw answers directly from the RAG pipeline, without performing any modification/augmentation technique. The evaluation at this point is being carried out by cosine similarity, but many other metrics would need exploration. There are specific libraries for this purpose such as (rag, 2024), (Ru et al., 2024), which evaluate RAG and LLM pipelines in both quantitative (precision, similarity, recall, etc.) and qualitative categories (hallucination, noise, factual correctness, etc.). The most natural approach to



	Character	Recursive	Semantic	Spacy
Chunks	27,377	17,677	21,693	25,532

Table 2: Number of chunks generated for each chunking strategy.

	Beginner			Expert			Overall
	Inf	Desc	Both	Inf	Desc	Both	
<b>RAG-Recursive</b>							
ModernBERT	0.653	0.699	0.676	<b>0.680</b>	0.716	0.698	0.687
LateInteraction	0.616	0.649	0.633	0.647	0.658	0.652	0.642
Sparse	0.608	0.659	0.634	0.654	0.692	0.673	0.653
<b>RAG-Semantic</b>							
ModernBERT	0.646	0.695	0.670	0.671	0.713	0.692	0.681
LateInteraction	0.588	0.634	0.611	0.629	0.643	0.636	0.623
Sparse	0.592	0.663	0.628	0.635	0.691	0.663	0.645
<b>RAG-Sentence</b>							
ModernBERT	<b>0.671</b>	<b>0.710</b>	<b>0.691</b>	<b>0.680</b>	<b>0.729</b>	<b>0.705</b>	<b>0.698</b>
LateInteraction	0.623	0.629	0.626	0.641	0.643	0.642	0.634
Sparse	0.616	0.671	0.643	0.635	0.693	0.678	0.661
Context-Based	0.894	0.880	0.887	0.879	0.864	0.872	0.879

Table 3: Similarity Scores for RAG Methods with Retrieval-Embeddings, and Context-Based approach

follow up with this line of research would be to explore augmenting the retrieved document and feeding it to a large-context model along with the question. More RAG techniques can be tested, to perform re-ranking or combine embeddings through hybrid searches.

## Acknowledgements

This work is supported by the European Commission through Project ASTOUND (101071191 — HORIZON EIC-2021-PATHFINDERCHALLENGES-01), by project BEWORD (PID2021-126061OB-C43) funded by MCIN/AEI/10.13039/501100011033 and, as appropriate, by “ERDF A way of making Europe”, by the “European Union”, and by project INNOVATRAD-CM (PHS-2024/PH-HUM-52) from Comunidad de Madrid.

## References

2024. [BM42: New Baseline for Hybrid Search](#).
2024. [Components of an art work](#).
2024. [Qdrant: High-performance vector search at scale](#).
2024. [Ragas: Llm evaluation library](#).

Lameck Mbangula Amugongo, Pietro Mascheroni, Steven Geoffrey Brooks, Stefan Doering, and Jan Seidel. 2024. Retrieval augmented generation for large language models in healthcare: A systematic review.

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, and Jian-Guang Lou. 2024. [Make your llm fully utilize the context](#). *Preprint*, arXiv:2404.16811.

Boqi Chen, Anuj Khare, Gaurav Kumar, Arjun Akula, and Pradyumna Narayana. 2025. [Seeing beyond: Enhancing visual question answering with multi-modal retrieval](#). In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 410–421, Abu Dhabi, UAE. Association for Computational Linguistics.

Paul DiMaggio. 1987. [Classification in art](#). *American Sociological Review*, 52(4):440–455.

Aaron Grattafiori et al. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.

Wenqi Fan, Yajuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. [A survey on rag meeting llms: Towards retrieval-augmented large language models](#). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD ’24*, page 6491–6501, New York, NY, USA. Association for Computing Machinery.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang,

- and Boris Ginsburg. 2024. Ruler: What’s the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*.
- Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. 2024. [E5-v: Universal embeddings with multimodal large language models](#). *Preprint*, arXiv:2407.12580.
- Omar Khattab and Matei Zaharia. 2020. [Colbert: Efficient and effective passage search via contextualized late interaction over bert](#). *Preprint*, arXiv:2004.12832.
- Philippe Laban, Alexander R. Fabbri, Caiming Xiong, and Chien-Sheng Wu. 2024. [Summary of a haystack: A challenge to long-context llms and rag systems](#). *Preprint*, arXiv:2407.01370.
- Xinze Li, Yixin Cao, Yubo Ma, and Aixin Sun. 2024. [Long context vs. rag for llms: An evaluation and revisits](#). *Preprint*, arXiv:2501.01880.
- Aigerim Mansurova, Aiganyam Mansurova, and Aliya Nugumanova. 2024. [Qa-rag: Exploring llm reliance on external knowledge](#). *Big Data and Cognitive Computing*, 8(9).
- Esmaeil Narimissa and David Raithel. 2024. [Exploring information retrieval landscapes: An investigation of a novel evaluation techniques and comparative document splitting methods](#). *Preprint*, arXiv:2409.08479.
- Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. [Nomic embed: Training a reproducible long context text embedder](#). *Preprint*, arXiv:2402.01613.
- OpenAI and Josh Achiam et al. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Renyi Qu, Ruixuan Tu, and Forrest Bao. 2024. [Is semantic chunking worth the computational cost?](#) *Preprint*, arXiv:2410.13070.
- Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. [Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation](#). *Preprint*, arXiv:2408.08067.
- Stefan Schaffer, Aaron Ruß, Mino Lee Sasse, Louise Schubotz, and Oliver Gustke. 2022. Questions and answers: Important steps to let ai chatbots answer questions in the museum. In *ArtsIT, Interactivity and Game Creation*, pages 346–358, Cham. Springer International Publishing.
- Sentence Transformers. 2021. all-minilm-l6-v2. *Computer software*. Hugging Face. <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>.
- Yuli Vasiliev. 2020. *Natural language processing with Python and spaCy: A practical introduction*. No Starch Press.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. [Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution](#). *Preprint*, arXiv:2409.12191.
- Xiaopu Wang, Jianing Sun, and Chao Qi. 2024b. [Ceda-tqa: Context enhancement and domain adaptation method for textbook qa based on llm and rag](#). In *2024 International Conference on Networking and Network Applications (NaNA)*, pages 263–268.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. [Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference](#). *Preprint*, arXiv:2412.13663.
- Jun Yu, Yunxiang Zhang, Zerui Zhang, Zhao Yang, Gongpeng Zhao, Fengzhao Sun, Fanrui Zhang, Qingsong Liu, Jianqing Sun, Jiaen Liang, and Yaohui Zhang. 2024. [Rag-guided large language models for visual spatial description with adaptive hallucination corrector](#). In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM ’24, page 11407–11413, New York, NY, USA. Association for Computing Machinery.

## A Appendix

### A.1 GPT4 Prompt for QA generation

```
{ role: "system" content:" You are an expert in art history. First you will generate 6 question-answer pairs about a painting solely based on the provided "Description" text: the first 3 will be formulated by an expert in the domain and the last 3 will be formulated by a beginner. You will then generate 6 question-answer pairs about the painting solely based on the provided "Information" text: the first 3 will be formulated by an expert in the domain and the last 3 will be formulated by a beginner. Divide the data in these 6 columns: qa_id, qa_type, qa_level, question, answer. " }
```

```
{ role: "user" content:" Information: In the course of his life, El Greco painted numerous versions of the Annunciation, thus allowing his stylistic development to be traced through his changing treatment of this Biblical episode. This painting, dated around 1576, is thought to be one of the last versions executed in Italy, and is clearly influenced
```

by the Venetian style. From her prayer-stool at the left of the painting, the Virgin listens attentively to the message of the Archangel, a figure rendered very much in the style of Veronese. The light and the colouring owe much to Titian, a painter El Greco admired, while the arrangement of the figures and the treatment of the drapery strongly recall the work of Tintoretto. Here, El Greco places the figures within a simple architectural setting, loosely framing them to make the scene more realistic.

**Caption: ### Subject** The painting depicts a religious scene, likely from the Christian tradition, featuring two central figures: a seated woman and a standing angel. The woman, dressed in a blue robe, appears to be in a state of contemplation or prayer, with her hands raised in a gesture of supplication or meditation. The angel, dressed in a yellow robe with a white cloak, is depicted with wings and a halo, suggesting a divine or celestial presence. The angel is reaching out towards the woman, possibly conveying a message or blessing.

**### Form** The painting is executed in a Baroque style, characterized by its dramatic use of light and shadow, dynamic composition, and rich, vibrant colors. The use of chiaroscuro (the contrast of light and dark) creates a sense of depth and volume, emphasizing the three-dimensional forms of the figures and the surrounding elements. The background features a dramatic sky with clouds and a radiant light source, adding to the overall sense of divine intervention or revelation.

**### Content** The content of the painting is deeply religious, focusing on a moment of divine communication or revelation. The angel's gesture towards the woman suggests a divine message or a moment of divine intervention. The woman's posture and expression indicate a state of receptivity or contemplation, highlighting the spiritual nature of the scene. The presence of cherubs in the background adds to the celestial and divine atmosphere of the painting. The overall composition and the interplay of light and shadow create a sense of sacredness and awe, inviting the viewer to reflect on the divine nature of the moment depicted." }

## A.2 GPT4 QA generation results, and LLaMA answer

Here we showcase the QA generation results along with the answer inferred from the LLaMA-3.1-8B model.

greco\_(domenikos\_theotokopoulos)\_w1:

Description,expert,**What artistic technique is**

**predominantly featured to enhance the dramatic effect in the painting?**,"The use of chiaroscuro, the contrast of light and dark, is predominantly featured to enhance the dramatic effect in the painting.", "El Greco predominantly features the artistic technique of chiaroscuro to enhance the dramatic effect in the painting, creating a sense of depth and volume by contrasting light and dark."

Description,expert,**How does the baroque style manifest in the composition of the painting?**,"The Baroque style in the painting is manifested through the dynamic composition, dramatic use of light and shadow, and the rich, vibrant colors.", "In the composition of the painting, the Baroque style manifests through its dramatic use of light and shadow, dynamic arrangement of figures, and rich, vibrant colors, creating a sense of depth and volume. The use of chiaroscuro emphasizes the three-dimensional forms of the figures and the surrounding elements, while the background features a dramatic sky with clouds and a radiant light source, adding to the overall sense of divine intervention or revelation."

Description,expert,**How do the gestures and expressions of the figures contribute to the theme of the artwork?**,"The gestures and expressions of the figures, such as the woman's raised hands and contemplative posture and the angel's reaching out gesture, contribute to the theme by highlighting a moment of divine communication or revelation.", "The figures' gestures and expressions in this painting significantly contribute to its theme by conveying a sense of reverence and spiritual connection. The Virgin's raised hands and attentive posture convey her receptivity to the angel's message, emphasizing her devotion and faith. The angel's gentle gesture towards her, with his hands extended and a soft expression, suggests a benevolent and divine presence. This interplay of gestures and expressions creates a sense of intimacy and sacredness, drawing the viewer into the moment of divine communication. The overall effect is one of calm contemplation and spiritual awe, underscoring the painting's theme of divine revelation and connection."

Description,beginner,**What are the main colors used for the clothing of the angel and the woman?**,The angel is dressed in a yellow robe with a white cloak and the woman is dressed in a blue robe., "The main colors used for the clothing of the angel are yellow and white, while the woman is dressed in a blue robe."

Description,beginner,**What is the background of the painting like?**The background features a dramatic sky with clouds and a radiant light source., "The background of the painting features a dramatic sky with clouds and a radiant light source, adding to the overall sense of divine intervention or revelation, and creating a simple architectural setting that loosely frames the central figures of the Virgin and the Archangel."

Description,beginner,**What do the figures in the painting represent?**"The figures represent a seated woman in contemplation or prayer and a standing angel conveying a message or blessing, suggesting a divine or celestial presence.", "The figures in the painting represent the Virgin Mary and the Archangel, depicting a moment of divine communication, specifically the Annunciation, where the angel conveys a message from God to Mary."

Information,expert,**What stylistic influences are evident in El Greco's painting based on the information given?**"The painting shows influences by the Venetian style, specifically reminiscent of Veronese's style in the depiction of the archangel and owes much to Titian in its use of light and coloring, along with elements recalling Tintoretto's work in the arrangement of the figures and the treatment of drapery.", "El Greco's painting is influenced by the Venetian style, evident in its use of light and colouring, which owes much to Titian. The arrangement of the figures and treatment of drapery strongly recall Tintoretto's work. Additionally, the angel is rendered in the style of Veronese, and the overall Baroque style is characterized by dramatic use of light and shadow, dynamic composition, and rich colours."

Information,expert,**How does the painting reflect the evolution of El Greco's artistic style?**"This painting, being one of the last versions executed in Italy, reflects the evolution of El Greco's style through its mixture of Venetian influences seen in the handling of figures, light, and coloring, showing his transition in stylistic developments.", "This painting reflects El Greco's artistic evolution through its incorporation of various styles, showcasing his eclecticism and willingness to adapt. It is characterized by the Venetian style, evident in the use of light and color, reminiscent of Titian, and the arrangement of figures, recalling Tintoretto. The angel's depiction, however, is more in line with Veronese, highlighting El Greco's ability to blend different influences to create a unique work. The painting's use of chiaroscuro and dra-

matic composition, typical of the Baroque style, demonstrates El Greco's mastery of this period's artistic trends, while the simple architectural setting suggests a return to a more traditional approach, reflecting his stylistic development over time."

Information,expert,**Where does the scene of the painting take place based on the architectural elements described?**"El Greco places the scene within a simple architectural setting, loosely framing the figures to make the scene appear more realistic.", "The scene of the painting takes place in a simple architectural setting, likely an interior space, as there is no mention of exterior elements such as columns, arches, or a large window, instead it is loosely framed to make the scene more realistic."

Information,beginner,**What scene is depicted in this painting by El Greco?**"The painting depicts the Annunciation, with the Virgin Mary at her prayer-stool listening attentively to the Archangel's message.", "The scene depicted in this painting by El Greco is the Annunciation, a moment from the Christian tradition in which the Archangel conveys a divine message to the Virgin, who is seated in contemplation or prayer."

Information,beginner,**Who are the figures in the painting?**"The figures in the painting are the Virgin Mary and the Archangel, with the Virgin listening to the Archangel's message.", "The two central figures in the painting are a seated woman, likely the Virgin, and a standing angel, the Archangel, who is reaching out towards her with a message or blessing."

Information,beginner,**What year is this specific painting of the Annunciation by El Greco thought to be created?**This specific painting of the Annunciation by El Greco is thought to be created around 1576., This specific painting of the Annunciation by El Greco is thought to be created in 1576.

### A.3 LLaMA-3.1-8B Prompt for Question answering based on the texts

```
{"role": "system", "content": "You are an art bot that responds to questions about paintings. You will create concise responses solely based on the Information provided along the question. You will only respond to the asked question. Try not to enumerate. Stick to natural language and do not use new lines."}, {"role": "user", "content": "original information + description"}
```