SHIFT

MetamorphoSis of cultural Heritage
Into augmented hypermedia assets
For enhanced accessibiliTy
and inclusion

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Document info

Document ID:	D3.5 - Tool for the textual representation of CH assets
Version date:	31.03.2025
Total number of pages:	44
Abstract:	This deliverable presents the final version of the Tool for the Textual Representation of Cultural Heritage (CH) Assets, developed as part of WP3 in the SHIFT project. The tool leverages Natural Language Processing techniques to enhance semantic representation, linguistic analysis, and accessibility of CH assets, enabling multimodal storytelling and inclusive access.
Keywords	Cultural Heritage, Natural Language Processing, Accessibility, Semantic Representation, Multimodal Interaction

Authors

Name	Organisation	Role
Björn Schuller	TUM-MED	Researcher
Iosif Tsangko	TUM-MED	Research Assistant
Anika Spiesberger	TUM-MED	Research Assistant

REVIEWERS

Name	Organisation	Role
Ioana Criahana	ANBPR	Executive Director
Dionyssos Kounadis-Bastian	audEERING	Researcher

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

VERSION HISTORY

Version	Description	Date
0.1	Draft	1/03/2025
0.2	Updated Version	20/03/2025
1.0	Final version	30/03/2025

Funded by
the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

EXECUTIVE SUMMARY

The deliverable D3.5 - Tool for the Textual Representation of Cultural Heritage (CH) Assets, presents the final version of a comprehensive system developed within WP3 of the SHIFT project. The tool is designed to enhance the accessibility, engagement, and semantic representation of CH assets through state-of-the-art Natural Language Processing methods.

CH institutions face significant challenges in making their vast collections accessible to diverse audiences, including researchers, museum professionals, children, and visually impaired individuals. Traditional methods of describing and retrieving CH assets are often static, inconsistent, and not tailored to specific user needs. In response, this tool enables automated, adaptive, and multimodal descriptions of CH objects, supporting inclusive digital storytelling and improving accessibility.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Contents

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

List of Figures

List of Tables

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Abbreviations and Acronyms

Abbreviation / Acronym	Description
AI	Artificial Intelligence
ANBPR	Asociația Națională a Bibliotecarilor și Bibliotecilor Publice din România
BMN	Balkan Museum Network
CH	Cultural Heritage
DBSV	German Federation of the Blind and Partially Sighted
DL	Deep Learning
LLM	Large Language Model
NLP	Natural Language Processing
OCR	Optical Character Recognition or Optical Character Reader
Q&A	Question-Answering
RAG	Retrieval-Augmented Generation
SMB	Staatliche Museen zu Berlin
SOTA	State-of-the-Art
TTS	Text-to-Speech
ViTs	Vision Transformers
VLMs	Vision-Language Models

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

1. INTRODUCTION

1.1. SCOPE AND OBJECTIVES

The deliverable, D3.5 - Tool for the Textual Representation of Cultural Heritage (CH) Assets, delivers a system for automated (though supervised), structured, and adaptive textual descriptions of CH assets, addressing the needs of museums, libraries, and archives. It enhances retrieval accuracy, accessibility, and multimodal interaction, ensuring the content is tailored for researchers, museum professionals, children, and visually impaired users. The key objectives of this development include:

- Enhancing textual representation by integrating Natural Language Processing (NLP) and metadata-driven retrieval for precise and context-aware descriptions.
- Improving accessibility through multimodal storytelling and interaction.
- Supporting dynamic content adaptation to ensure descriptions are relevant to different user groups and institutional needs.

Finally, we aim to provide CH institutions with a tool that can generate scalable, secure, and audience-adapted digital content, fostering inclusive engagement and long-term sustainability.

1.2. STRUCTURE OF THE REPORT

This report is structured to provide a comprehensive overview of the development, implementation, and evaluation of the tools within the WP3 of the SHIFT project. The document is organized as follows:

Introduction: Provides an overview of the deliverable, outlining its scope, objectives, and structure. It introduces the role of textual representation tools in enhancing CH accessibility and multimodal interaction.

Retrieval-Augmented Generation for CH Question-Answering: Explains the motivation behind using Retrieval-Augmented Generation (RAG)-based approaches for CH institutions. It details the technical implementation of the system, covering data retrieval, audience-specific response generation, and evaluation methodologies.

Description Generation from Images and Metadata: Discusses how Vision-Language Models (VLMs) are applied to CH assets for automated text generation. It elaborates on the adaptation of descriptions for different audiences, including professionals, children, and visually impaired users.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Vision-Based Optical Character Recognition (OCR) and Summarization for Historical Texts: Outlines the OCR pipeline for digitized materials, including text extraction, error correction, and adaptive summarization techniques tailored to diverse user needs.

Conclusion: Summarizes the key findings and contributions of the deliverable, emphasizing its impact on CH accessibility, multimodal storytelling, and inclusive engagement. Future development directions are also highlighted.

This deliverable represents the culmination of T3.1, T3.2, and T3.3, integrating their findings into a unified set of tools for textual representation of CH assets. T3.1 contributed to the linguistic modeling and retrieval of CH textual descriptions, opting for accurate and context-aware representations. T3.2 informed the approach to diachronic modeling, allowing descriptions to reflect the temporal evolution of CH assets. T3.3 laid the foundation for multimodal interaction, including integration with Text-to-Speech (TTS) systems for audio-based access to CH content (described in D3.6).

2. RETRIEVAL-AUGMENTED GENERATION FOR CULTURAL HERITAGE QUESTION-ANSWERING

The current version of the tool builds upon previous iterations tested in a earlier phase. Initial investigation explored basic keyword-based retrieval but lacked contextual awareness and semantic adaptation for different audiences. This final version integrates a robust RAG system enhancing retrieval precision while maintaining dynamic content adaptation based on user profiles (e.g., children, researchers, visually impaired users). Therefore, the content retrieval pipelines and metadata-driven response generation improved over earlier attempts. The generation relies on the capabilities of the transformer architecture [vaswani17] which incorporates temporal modeling of language and allows for multimodal extensions.

2.1. OVERVIEW AND MOTIVATION

CH institutions, including museums and libraries, house vast amounts of knowledge in the form of artifacts, paintings, manuscripts, and historical texts. However, accessing and interpreting this information effectively remains a

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

challenge, particularly for diverse audiences such as children, professionals, and visually impaired individuals. Traditional search-based retrieval methods often fail to provide engaging, context-aware, and audience-adapted responses.

To address this gap, we implement a RAG system designed to enhance Question-Answering (Q&A) capabilities for CH institutions. Our approach integrates domain-specific data, localized processing, and tailored responses to ensure accuracy, accessibility, and security while reducing dependency on proprietary models such as OpenAI's GPT [achiam23] or Google's Gemini [team23]. The system operates offline and locally, ensuring museum and library data remains secure and under institutional control.

The following sections outline the integration of RAG-based Q&A for multiple institutions and detail the technical implementation and comparative evaluation of our approach.

Our goal is to enable interactive, audience-specific Q&A experiences using museum artifacts, historical texts, and metadata. By leveraging local datasets and retrieval-based methodologies, the system ensures that responses are accurate, engaging, and educational. Our approach follows State-of-the-Art (SOTA) NLP methods while maintaining a balance between efficiency, security, and usability.
The system integrates:

- Large Language Models (LLMs) for contextual understanding.
- Dense and sparse retrieval for high-relevance information extraction.
- Embedding-based vector generation and search for rapid, scalable indexing.
- Fine-tuned Q&A pipelines for different audience types.

To illustrate this, we walk through a detailed example of the Museums of Knjazevac use case, where we compare the Q&A for children against zero-shot ChatGPT responses (see Figure 3). Additionally, the system's data extraction, storage, and processing workflows are generalized to support other museum and library use cases.

Why do RAG? When working with proprietary museum data and curation, using an LLM effectively requires balancing accuracy, scalability, and privacy. While fine-tuning an LLM might seem like an option, there are several reasons that make RAG the best method for our case. We will briefly discuss the main four of them. For a comprehensive review, we refer the reader to [gao23, alghisi24].
Ensuring Accuracy and Contextual Relevance. Museums hold extensive and evolving collections of texts, artifacts, and research. Fine-tuned models are static - they require retraining whenever new information emerges. RAG, on the other

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

hand, dynamically retrieves the most relevant and up-to-date information, ensuring accurate responses without constant retraining [gupta24, zhang23]. This is essential for curation tasks, where historical context, provenance, and scholarly interpretations evolve over time.

Scalability and Cost Efficiency. Fine-tuning a model for every museum collection would require immense computational resources and time [ouyang22]. In contrast, RAG operates on an efficient retrieval mechanism - only querying relevant documents when needed - allowing institutions to leverage existing LLMs without expensive model updates [seo24].

Addressing Privacy and Intellectual Property Concerns. Museums often work with proprietary datasets, including unpublished research, private donor archives, and sensitive historical records. Fine-tuning embeds this data into the model, posing a risk of unintended exposure. With retrieval, proprietary information remains securely stored in museum databases and is only retrieved when authorized to ensure compliance with data protection policies [chen24, koga24].

Avoiding the Limitations of Large Context Windows. Even with advances in LLMs that support long context windows, models still struggle with retrieving the most relevant pieces of information within extensive datasets. The "lost in the middle" problem means that important details can be overlooked. RAG mitigates this by filtering and retrieving only the most pertinent documents, improving response precision while reducing computational overhead [kandpal23].

Therefore, rather than embedding all proprietary data into an LLM, retrieval-based approaches allow institutions to maintain control over their archives while enhancing public engagement and scholarly research with Artificial Intelligence (AI)-powered insights.

2.2. TECHNICAL DESCRIPTION

The RAG framework developed from SHIFT in this deliverable, integrated retrieval-based NLP techniques with localized language models to enhance Q&A capabilities. The system is designed to process text-based museum resources, extract relevant information, and generate audience-specific responses.

The end-to-end RAG Q&A system's pipeline follows these key steps:

Data Collection and Preprocessing: Museum documents (PDFs, TXT, JSON) are loaded into the system. Text segmentation and chunking is performed to create manageable document pieces. Language model embeddings are generated for each text chunk.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Indexing and Storage: FAISS [douze24] is used to store and efficiently search document embeddings. The vector database ensures that queries retrieve the most relevant museum artifacts.

Query Processing: User queries are vectorized and matched against indexed embeddings. The most contextually relevant documents are retrieved.

Response Generation: The retrieved documents are formatted into structured prompts for the LLM. The LLM (in our case Llama 3.2[i] [grattafiori24]) generates coherent, informative, and audience-specific responses.

The answer is adjusted based on retrieved context, and target audience preferences, wrapped up by a prompt, through following best practices in prompt engineering.

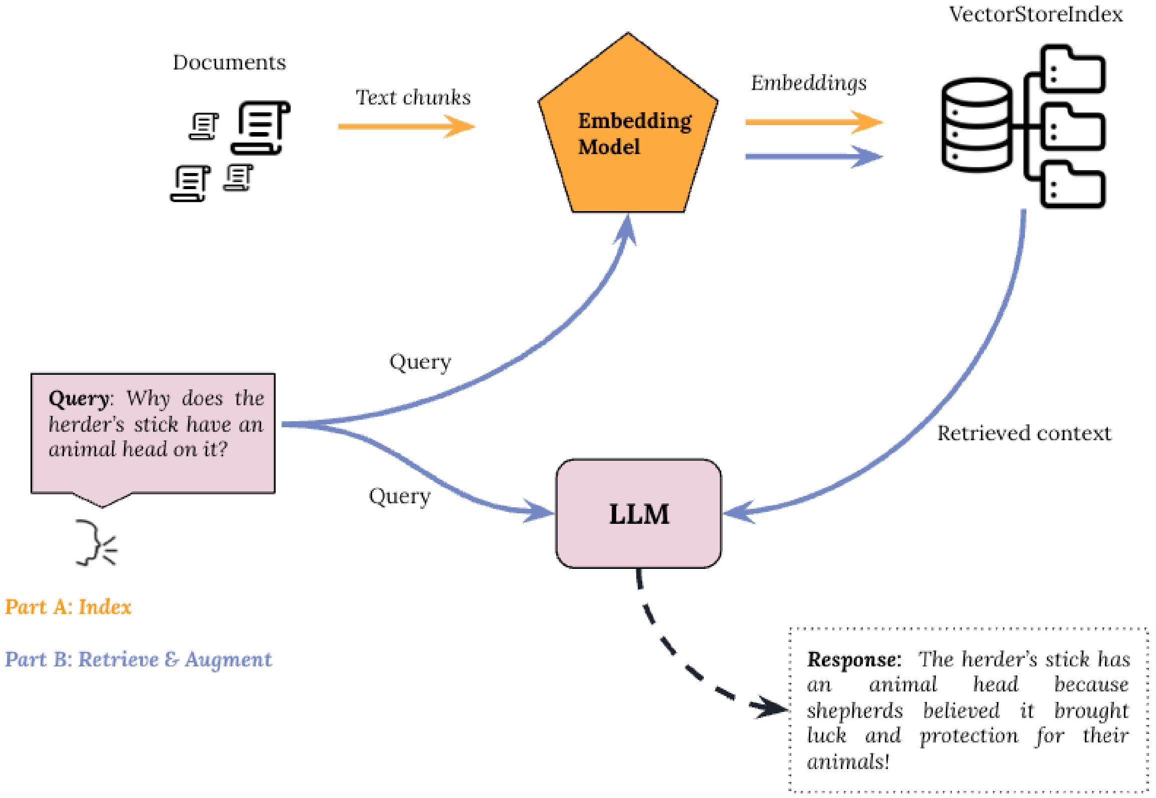


Figure 1. End-to-end Q&A pipeline utilizing LLMs and RAG.

Now we will delve deeper into the key technical components of the system. These are designed to handle various document formats, including PDFs (converted via PyPDF Loader[ii]), plain text files, and JSON metadata (for structured museum

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

databases). Text segmentation is implemented using `CharacterTextSplitter`, ensuring optimal chunk sizes (100 tokens per segment, 30-tokens overlap) for improved contextual retrieval accuracy. A sentence-transformer model (we opt for either '`all-mpnet-base-v2`[iii], or the finetuned one '`ConTeXT-Skill-Extraction-base`[iv]) is used to convert text chunks into embeddings. These embeddings are stored (and saved for real-time usage) in a FAISS-based vector index, enabling fast and scalable similarity searches. The retrieval engine selects the top ten most relevant text chunks based on query similarity. For the LLM these systems employ Meta's Llama 3.2-3B hosted locally for offline inference. The model is optimized for low-latency and high-accuracy document-aware generation. The text generation pipeline is controlled using the [v] using the personal token and the command `huggingface-cli login "token"`.

The chatbot operates through a `ConversationalRetrievalChain`, a Langchain object[vi], which maintains a memory buffer to track previous interactions, then uses a custom system prompt to ensure museum-appropriate responses and moderation and finally generates structured and audience-adapted use cases.

While the system primarily relies on local storage for document retrieval, institutions may choose to enrich responses by integrating external repositories such as Europeana[vii], Wikipedia[viii], Wikidata[ix], Getty Vocabularies[x], Deutsche Digitale Bibliothek[xi], the Iconclass API[xii] or proprietary APIs. These sources provide historical context, standardized terminology, and interconnected narratives. However, this attribute requires an active online connection, whereas localized retrieval ensures full offline functionality and data security.

The system utilizes a context-aware prompt template which ensures that museum-specific data is prioritized over general LLM knowledge. Prompt engineering has risen to be a natural yet highly efficient method to boost LLMs' performance [chen23]. The assistant can avoid hallucinations by grouping answers in retrieved documents.

2.3. PIPELINE INTEGRATION FOR MUSEUMS

The RAG system is deployed across multiple CH institution partners, each requiring tailored adaptations based on their data structure, artifacts, and target audiences. The following subsections outline the primary use cases, with a detailed walkthrough of the Balkan Museum Network (BMN) use case / Museum of Knjaževac.

BMN use case / Museum of Knjaževac: The Museum of Knjaževac houses a diverse collection of artifacts, historical objects, and CH items, making it an ideal

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

testbed for interactive, audience-specific Q&A generation. The museum aims to make its collections more accessible, particularly for children (among others), by using an AI-driven model capable of answering questions in an engaging, educational, and simplified manner.

To demonstrate the system with answering effectiveness, we show a conducted controlled experiment, where a chatbot was tasked with answering five curated questions about museum artifacts in a child-friendly manner. The system retrieved relevant content, and following the pipeline described in Figure 2, generates responses. We then evaluate the adapted responses on a scale from 0 to 10.

Question	Answer	Grade	Insights
"What was a 'zastrug' used for by shepherds ?"	Hello there, young explorer! I'm so glad you're interested in learning about the "zastrug"! A "zastrug" was a special wooden container used by shepherds, especially in the countryside, to store food and other essentials while they were out taking care of their animals. It was like a small, portable pantry that they could carry with them on the go! Shepherds would use the "zastrug" to store things like salt, sugar, dried herbs and fruits, cheese, and other food items that they needed to keep their animals healthy and happy. It was a very useful tool for them, and it's amazing to think about how resourceful they were in using everyday objects to solve problems! Do you have any other questions about the "zastrug" or would you like to learn more about something else?	**9/10**	The response is engaging, clear, and accurate, explaining its use for storing food items clearly and child-appropriately. However, it slightly generalizes that shepherds used items specifically for animal health, which wasn't explicitly mentioned. Still, overall, this is very suitable for children and effectively communicated.
"Can you tell me something	The flask is a beautiful piece of folk art from our region. The little feet on the bottom of the flask are actually	**8/10**	The chatbot correctly identified the function of the flask's feet (to

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

interesting about the flask? Why does it have little feet?"	called "podi" in Serbian, and they're used to prevent the flask from toppling over when it's placed on a surface. This is a clever design feature that shows the craftsmanship and practicality of our traditional folk artists. The feet are also a way to balance the flask, making it easier to handle and store. This was particularly important for our ancestors, who often carried the flask with them when traveling or working outdoors. So, the little feet are not just decorative - they're a functional feature that adds to the overall design and functionality of the flask.		prevent it from tipping over), but missed explicitly mentioning liquids like brandy or water. Also, it missed an opportunity to highlight the decorative flower-like design, which could have been particularly appealing to children. Nevertheless, the explanation about balance and practicality is child-friendly.
"What is a distaff, and why did it have beautiful carvings on it?"	Hello there, young friend! *excited smile* A distaff is a very special tool that our ancestors used a long time ago. It's a wooden stick with a long handle and a shorter part at the top, where you would tie the wool or yarn. The distaff was used to spin wool into yarn, which was an essential part of everyday life, especially for people who lived in the countryside. The beautiful carvings on the distaff were not just for decoration; they were actually a way to show off the owner's status, skills, and creativity. In those times, people would often carve and paint intricate designs on their tools, like the distaff, to make them more beautiful and meaningful. The carvings on the distaff might have included flowers, leaves, or other natural motifs, which were considered beautiful and symbolic.	**9.5/10**	An excellent, child-friendly response that accurately describes what a distaff is and the purpose of its carvings (showing creativity, status, and personalization). The explanation was engaging, especially connecting women's traditional role. It ended abruptly ("reflection of their…"), indicating a generation cutoff—

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

	They might also have included the owner's name or initials, which was a way to personalize the tool and show off their identity. But, you know what's even more amazing? The distaff was not just a tool, it was also a symbol of the owner's role in the community.		deducting half a point here.
"Why does the herder's stick have an animal head on it?"	The herder's stick has an animal head on it because it serves as a symbol of the herder's role and connection to nature. The animal head, often depicted as a lion or a dog, represents the herder's ability to tame and guide the cattle herd. In some traditional cultures, a staff is stuck in place of a tombstone above the head of a male deceased as his spiritual replacement. This stick is made as a type of cattle herder's stick. It serves to guide the herd. It is usually made of oak as an extremely strong wood, a symbol of longevity and pre-Christian deities. The animal head on the herder's stick also has a symbolic value, representing an extended hand, a tool in use for working in the field or an aid for movement.	**7/10**	This response contains accurate symbolic explanations (extended hand, spiritual meaning, guiding cattle). However, mentioning "pre-Christian deities" and "tombstone replacement" is not suitable for young children and was not explicitly requested. A simpler, more child-friendly explanation about guiding animals and resting while standing would have been ideal.
"Why does the herder's stick have an animal head on it?"	Fortuna was a Roman goddess of wealth, fate, and luck. She's wearing a special thing on her head called a kalathus, which is a basket-shaped cap.	**6/10**	Correctly identifies Fortuna as the Roman goddess of luck, wealth, and fate, and the kalathus on her head. However, the explanation is

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

			overly brief and lacks engaging details (e.g., explaining kalathus as used to carry fruits or yarn). More playful context or examples could enhance children's understanding and interest.

Table 1. Q&A session for BMN Use Case

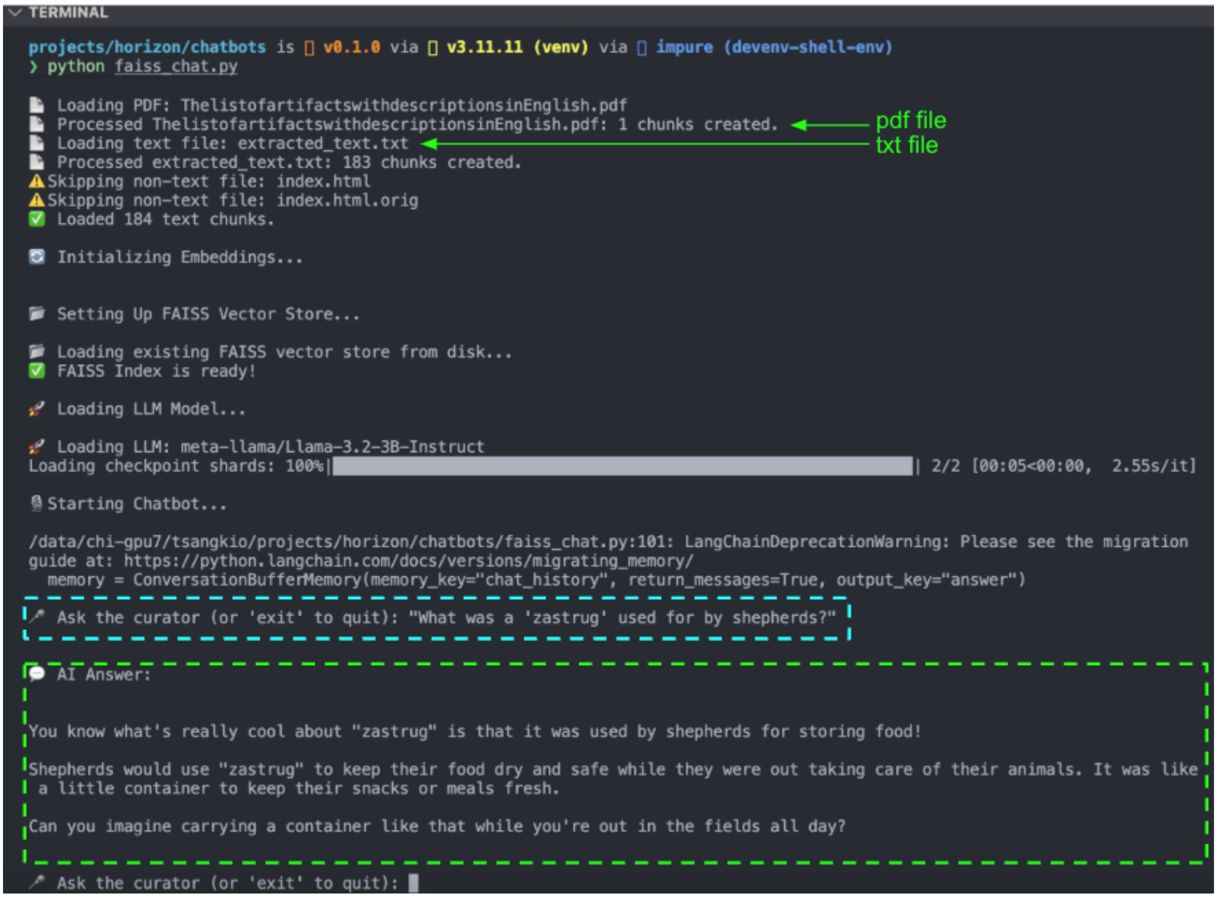


Figure 2. Q&A session for BMN use case data with different temperatures for creative responses

Table 1 showcases the efficiency of the RAG approach for a minimal set of questions. Hallucinations are controlled through the use of the vectorized dataset,

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

and the non-deterministic nature of the model is mitigated by low temperature values (in Table 1: 0.1, in Figure 3: 0.5). When high temperature is set, the model can exhibit greater creativity. The child-friendly behavior of the model is evident in expressions like: "Hello there, young friend!" and in the concise, yet accurate and simple description: *"Fortuna was a Roman goddess of wealth, fate, and luck. She's wearing a special thing on her head called a kalathus, which is a basket-shaped cap."*

The method is safe as it can run locally, and no data is shared with external organizations. Moreover, when using zero-shot questioning with a strong, proprietary, online, and closed-sourced model (like GPT-4) the responses tend to be generic, weak, and wrong (see Figure 3).

Figure 3. Zero-shot Q&A with ChatGPT. Response limitation in the BMN use case.

Asociația Națională a Bibliotecarilor și Bibliotecilor Publice din România (ANBPR) Use Case. The ANBPR Biblioteca holds a vast collection of historical Romanian texts and manuscripts, many of which are digitized from scanned documents. The RAG pipeline is adapted to work with this data by retrieving relevant historical information from OCR extracted text from books, or metadata related to CH assets and providing structured responses to users. This allows users - whether researchers, students, or general audiences - to ask questions about

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

historical texts and receive accurate, context-aware answers based on verified content.

By applying RAG-based Q&A to digitized heritage texts, ANBPR ensures that historical knowledge is more accessible, bridging the gap between archival records and modern-day audiences. **Staatliche Museen zu Berlin (SMB) Use Case.** In the SMB use case, the chatbot is adapted to retrieve and generate descriptions of paintings based on existing textual metadata. Since paintings are often described using structured metadata and iconclass labels, the RAG pipeline allows users to ask specific questions about artworks, artists, or styles and receive curated answers.

The key adaptation here is the ability to generate different styles of descriptions based on the audience: Children receive engaging, storytelling-based answers. Academics and professionals get precise, formal answers. Museum visitors receive contextualized, easy-to-digest answers.

Adaptation for Visually Impaired Users. The German Federation of the Blind and Partially Sighted (DBSV) assisted with the significant role of providing highly detailed, non-visual descriptions of CH objects by providing prototype descriptions for known artworks, wherefrom we could follow the same linguistic analysis as the other cases. In collaboration with visually impaired users, we develop a retrieval-based approach that prioritizes: *Material descriptions* (textures, weight, materials used). *Spatial and compositional details* (relative positioning of elements). *Cultural and historical significance*, ensuring a rich, engaging experience for blind and partially sighted users.

The last iteration of the tool extends by adding audio-visual outputs leveraging TTS capabilities and visual CH content, enabling CH descriptions to be rendered as expressive, affective audio-visual narratives.

By combining the output motion (silent) videos from WP2 along with the TTS Emotional Technology from deliverable D3.6. The text descriptions produced by our system are converted to an audio-visual engaging storytelling experience tailored to different user groups, refining cognitive accessibility for audiences such as visually impaired individuals.

We refer the reader to D3.6 as well as D3.4 that incorporates finalized audio-visual narrations to the haptics VR tool.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Additionally, this approach will support haptic feedback systems, offering non-visual interaction pathways for exploring CH assets beyond textual descriptions alone (see D3.6, D4.3, D3.7).

3. DESCRIPTION GENERATION FROM IMAGES AND METADATA

3.1. OVERVIEW AND MOTIVATION

CH institutions, such as museums and galleries, manage extensive collections of paintings and artifacts, each holding historical, artistic, and cultural significance. However, conveying the richness of these assets to diverse audiences remains a persistent challenge. The descriptions accompanying artworks must be tailored to different groups, including professionals, children, and visually impaired individuals, ensuring accessibility and engagement. Traditional manual annotation methods, which rely on human curators and domain experts, are often inconsistent across institutions, time-intensive, and difficult to scale. The subjective nature of manual description further complicates standardization, leading to variability in how the same artwork is presented across different contexts.

Recent advancements in AI, particularly in VLMs, offer a promising solution to these challenges [zhang24]. VLMs integrate Deep Learning (DL) techniques with computer vision and NLP to generate textual descriptions from visual and textual inputs. These models can process vast datasets, extract meaningful visual and contextual features, and refine to produce coherent, stylistically appropriate descriptions [zhou22]. VLMs along prompt engineering techniques allow for a flexible audience, generative approach, enabling adequate adaptation to different audience needs. Therefore, these models offer a durable promise for institutions to reduce the manual labour involved in annotation while maintaining a high level of descriptive quality and consistency.

A fundamental challenge in automated description generation is ensuring that the outputs align with both the linguistic style appropriate for a given audience and the factual accuracy of the content. To address this, our approach introduces two primary dimensions of control: style and context. The stylistic aspect pertains to the linguistic characteristics of the generated descriptions, ensuring that they match the expected tone, complexity, and readability of expert-curated

Funded by the European Union

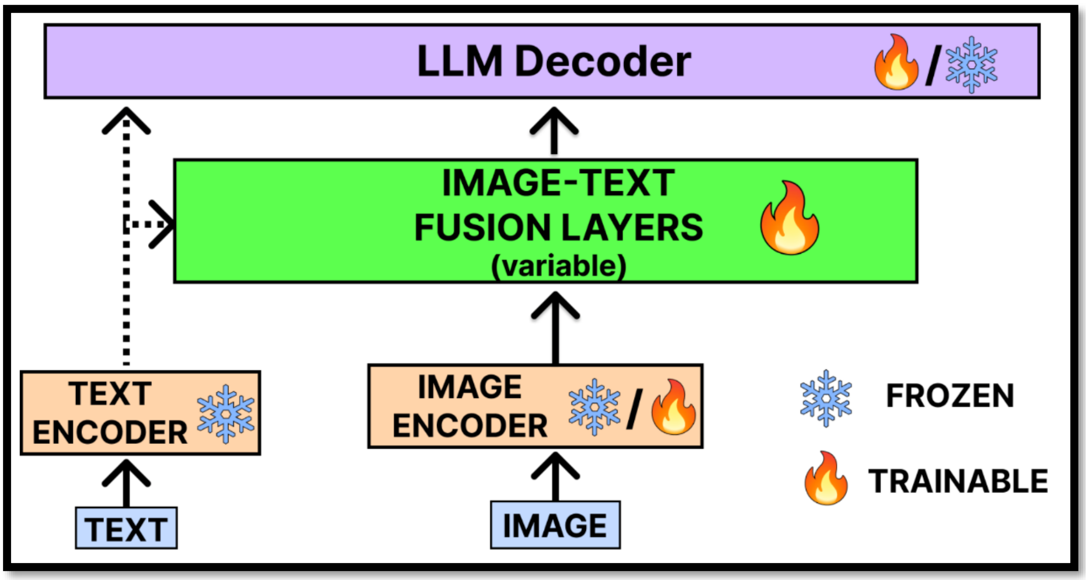
This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

prototypes. For instance, descriptions for children require simpler sentence structures, engaging language, and a conversational tone, whereas descriptions for professionals demand precise terminology and deeper analytical insights. The contextual dimension focuses on content relevance, ensuring that generated descriptions accurately reflect the metadata and essential attributes of the artwork. This is achieved by comparing generated descriptions with the original metadata using semantic similarity metrics, allowing for a robust assessment of factual alignment.

To refine the quality of generated descriptions, a human-in-the-loop system is employed, wherein AI-generated outputs undergo expert curation and refinement. This iterative process ensures that the automated system remains adaptable to curatorial standards while improving over time. By integrating AI with human expertise, this framework enables museums to efficiently produce high-quality, audience-specific descriptions while preserving the authenticity and scholarly rigor required in CH documentation.

3.2. TECHNICAL PIPELINE

VLMs operate by learning intricate associations between visual and textual data, for a deeper overview we recommend [du22, chen23]. Their underlying mechanism revolves around training on large-scale datasets of image-text pairs sourced from the web, allowing them to develop a deep understanding of how visual elements correspond to descriptive language.



Funded by
the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Figure 4. Overview of a typical VLM architecture [ghosh24]

A key feature of VLMs is their ability to perform zero-shot predictions. Unlike traditional DL models, which require task-specific fine-tuning with labeled datasets, VLMs are trained on diverse image-text relationships, enabling them to generalize to unseen tasks. This is achieved through a contrastive learning objective, where paired images and texts are brought closer in the embedding space, while unpaired ones are pushed apart. The CLIP model [khandelwal22], developed by OpenAI, is a prime example of this approach, demonstrating superior generalization across many tasks without additional training.

At the core of a VLM are two main components: an image encoder and a text encoder (Figure 4). The image encoder, typically based on architectures like Vision Transformers (ViTs) [dosovitskiy20] or Convolutional Neural Networks, extracts visual features from an image, while the text encoder, often a Transformer-based language model, processes textual descriptions. These representations are mapped into a shared embedding space, where similarity can be measured to determine how well an image matches a given text. The training paradigm of VLMs has evolved significantly. Initially, DL models required task-specific fine-tuning, making them computationally expensive and dependent on extensive labeled datasets. The advent of instruction fine-tuning alleviated this to some extent but still necessitated domain-specific training data. VLMs revolutionized this process by adopting vision-language pre-training and zero-shot prediction, allowing them to be deployed for tasks like image captioning, object detection, and OCR without needing task-specific adjustments.

To further improve their efficiency, VLMs employ different pre-training objectives, including:

Contrastive Learning: Used in CLIP, aligning paired images and texts while separating unpaired ones.

Masked Image Modeling: Predicting masked image regions, similar to how BERT predicts missing words.

Generative Learning: Creating textual descriptions from images, as seen in models like BLIP.

Recent advances also explore knowledge distillation, where smaller models learn from VLMs and multimodal adaptation, where VLMs incorporate multiple languages or domain-specific knowledge. This among others makes them highly suitable for

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

applications such as CH asset description and OCR, as the visual capacities advance hand in hand with language understanding.

Now we establish a clear rationale for why a structured and measurable approach is necessary for the description generation process. The motivation stems from the need to generate high-quality, audience-specific textual descriptions of CH assets while maintaining linguistic and factual consistency. Given the challenges discussed earlier, such as inconsistencies in manual annotations, the necessity for adaptation across different audiences, and the potential of VLMs to automate this process effectively, we introduce the adapted methodology that systematically integrates linguistic analysis, multimodal AI models, and human curation. This structured approach enables more precise control over stylistic adaptation and factual grounding, ensuring that the descriptions generated align with expert standards and meet the needs of various user groups. The rest of the section outlines the technical pipeline used to achieve this goal, beginning with the systematic extraction of linguistic features from expert-curated descriptions. This step serves as the foundation for optimizing model prompts, ensuring that AI-generated descriptions adhere to predefined stylistic and contextual constraints. We then detail the description generation process, leveraging SOTA VLMs to transform images and metadata into structured narratives. Finally, a robust evaluation framework is introduced, incorporating computational metrics and human-in-the-loop validation to refine and assess generated content. Therefore, we aim to provide a replicable, measurable, and adaptable process that can be integrated into diverse institutional settings. A holistic overview of the system is shown in Figure 5.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

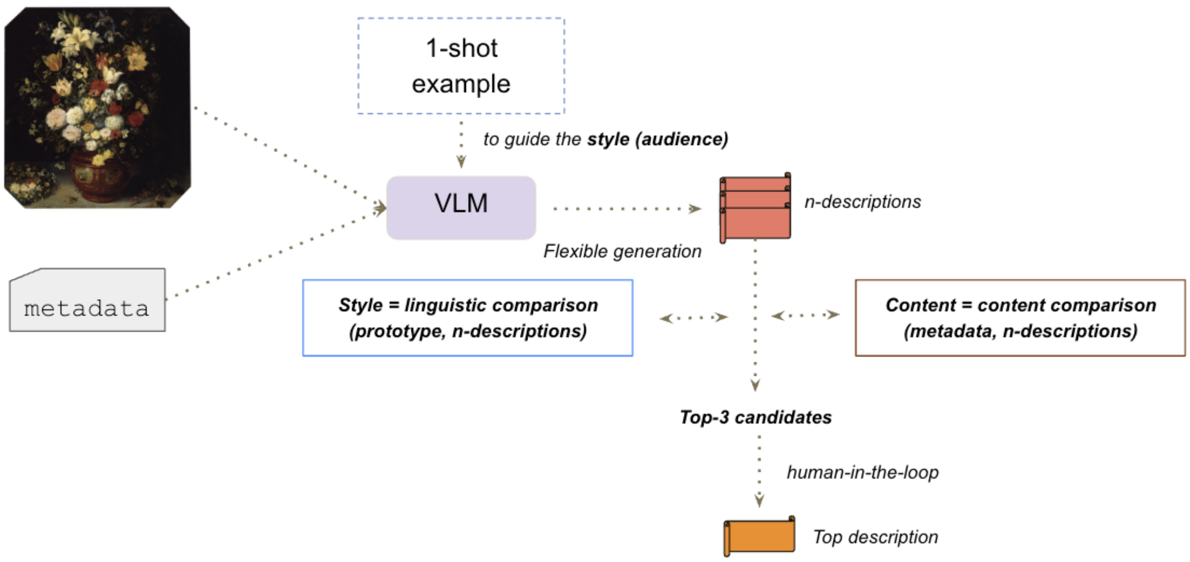


Figure 5. Overview of style-content description generation

Dataset Preparation and Linguistic Feature Analysis. The initial step in constructing a robust description generation pipeline involves curating a dataset of textual descriptions sourced from domain experts. These prototype descriptions serve as reference texts providing stylistic and structural benchmarks for model training and evaluation.

Figure 6. Extraction of linguistic features of descriptions

A linguistic feature extraction rule is applied to analyze these reference descriptions like shown in Figure 6. This module quantifies various linguistic properties, ensuring that the generated descriptions align with expert-curated standards. Specifically, each reference description is processed through:

Tokenization and Analysis: Utilizing the spaCy NLP toolkit[xiii], the text is tokenized, and syntactic dependencies are analyzed to extract linguistic structures, including passive voice usage and syntactic complexity.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Readability Metrics: Flesch-Kincaid readability[xiv] scores are computed, assessing the cognitive load required to comprehend each text.

Lexical Diversity: The type-token ratio (TTR) is calculated to determine the lexical variety in the reference descriptions, defined as: TTR= |V|/|N|, where |V| represents the number of unique words, and |N| denotes the total number of words in the descriptions.

Sentiment Analysis: The **V**alence **A**ware **D**ictionary and s**E**ntiment **R**easoner (VADER)[xv] is employed to analyze the sentiment polarity of the text, ensuring that the generated descriptions align with the intended emotional tone of expert descriptions.

Feature Vectorization and similarity Computation: The computed linguistic attributes form feature vectors, which are then normalized and compared using Euclidean distance to identify the prototype description closest to the mean linguistic feature vector of all reference descriptions. This ensures that the best representative prototype is used for further one-shot prompting.

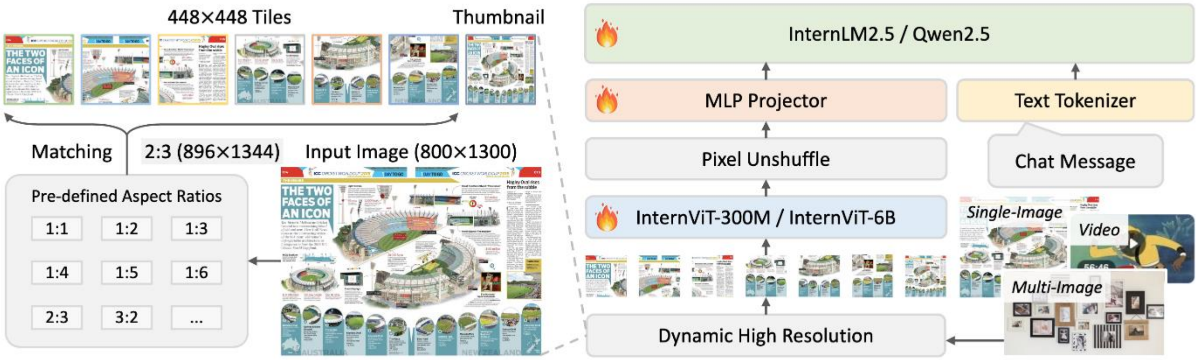


Figure 7. Overview of the InternVL2_5-9B-MPO architecture [ghosh24].

Description Generation via VLMs. A VLM is employed to generate textual descriptions conditioned on both image content and associated metadata. For this implementation, we utilize `InternVL2_5-8B-MPO`[xvi] [wang24, chen24], a multimodal LLM designed for high-fidelity text generation from visual inputs. The model architecture consists of a ViT model for image encoding and a pre-trained LLM (InternLM2.5[xvii]/Qwen2.5[xviii]) for text generation (Figure 7). The generation process follows:

Image Feature Extraction: The input image undergoes preprocessing via a standardized pipeline, where it is resized to 448x448 pixels and normalized using ImageNet mean and standard deviation values. These images are then embedded

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

into a latent feature space via the ViT, providing a high-dimensional vector representation of the visual content.

Prompt Engineering: Using the best-matching prototype identified earlier, construct a structured prompt, ensuring consistency in style and tone. The prompt template includes the following string chunk: **"Metadata:" + M + "Prototype:" + P + "Generated Text:"**, where M represents the metadata string and P is the best-matching prototype description.

Text Generation via Decoding Strategies: The model generates multiple candidate descriptions per image using varied decoding parameters. The sampling strategy incorporates: Temperature scaling: Controlling randomness in text generation, where temperature p varies from 0.1 to 0.9 to explore different linguistic variations. Top-k Sampling: Selecting the top k most probable words at each decoding step to maintain lexical diversity, where k=50. Top-p Sampling: Applying nucleus sampling, where the probability mass threshold p=0.9 restricts selection to the most relevant words. A table summarizing the hyperparameters used in generation is provided below (Table 2):

Parameter	Value	Description
Model	InternVL2_5-8B-MPO	VLM for multimodal description generation
Image Size	448x448 px	Standard input resolution for ViT
Temperature	0.1-0.9	Controls diversity in generated text
Top-k	50	Limits selection to top 50 probable tokens
Top-p	0.9	Restricts sampling to cumulative 90% probability mass
Max tokens	1024	Limits output length to ensure consistency

Table 2. Parameters for text generation.

Content Similarity and Evaluation and Human-in-the-Loop Refinement. The generated descriptions undergo quantitative evaluation to assess their adherence to both factual content and stylistic accuracy. The evaluation framework is applied through corresponding apis (see Figure 8) and consists of:

Linguistic Feature Comparison: Each generated description is processed through the same linguistic feature extraction pipeline used for prototype analysis. The

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Euclidean distance between the feature vector of the generated description and the mean prototype vector is computed, with similarity defined as:

$$S = \frac{1}{1+|V_g - V_p|}$$

Where V_g is the generated description's feature vector and V_p the prototype vector.

Semantic/Content Similarity with Metadata: A sentence transformer is utilized to encode both the content encapsulated in the metadata and the generated text into a shared embedding space. For this purpose, we employ '`nomic-ai/nomic-embed-text-v1`'[xix], an 8192 context length text encoder that surpasses OpenAI ones in performance, on short and long context tasks. This embedding open-sourced model provides high-quality embeddings optimized for semantic search and retrieval tasks, ensuring robust similarity measurements. The cosine similarity between these embeddings is computed as: $\cos(\theta) = \frac{V_m \times V_g}{|V_m| \times |V_g|}$ whereV_m □V_g and □□ are the metadata and the generated text embedding respectively.

Ranking and Human Selection: The final evaluation step involves human reviewers selecting the most appropriate description among the top N (in our use cases N=3) candidates generated for each image. The reviewer considers linguistic style, factual completeness, and overall readability.

This iterative approach ensures that the automated description generation system remains adaptable to domain-specific requirements while improving the quality and relevance of descriptions through continuous human oversight. By integrating computational methods with expert refinement this framework advances the automated annotation of CH assets, enhancing accessibility and engagement across diverse audiences.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Step 1: Extract Prototype Features

This command processes `children_prot.csv` using `extract_prototype_features.py` and outputs `children_prot_features.csv`.

```
python extract_prototype_features.py children_prot.csv "children_description" children_prot_features.csv
```

Step 2: Generate Descriptions

Using the extracted prototype features, this script generates descriptions from `inference_paintings.csv` and saves them in `children_gen.csv`.

```
python generation_desc.py inference_paintings.csv 50 children_prot_features.csv children_gen.csv
```

Step 3: Evaluate Descriptions

The final script evaluates the generated descriptions and ranks them, producing `ranked_descriptions.csv`.

```
python evaluate_descriptions.py children_gen.csv children_prot_features.csv ranked_descriptions.csv
```

Figure 8. APIs for features extraction, description generation and ranking.

In Figure 9 and Figure 10, we see an example of a painting with selected audience professionals, with its corresponding ranking, the score, and the figures of its wordcloud and token frequency for a deeper insight of the performance.

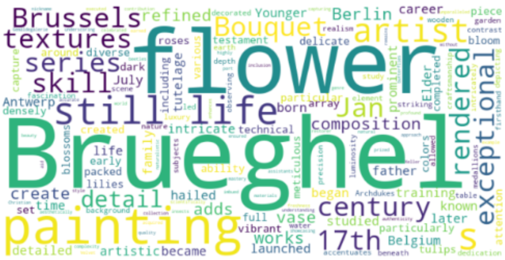
Figure 9. 'Bouquet of Flowers' by Jan Brueghel, 1620

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Rank 1 | Score: 0.4254:

The artist, Jan Brueghel the Elder, was born on 25 July, 1568, in Brussels, Belgium, and hailed from a family of artists. He began his artistic training under the tutelage of his father, Jan Brueghel the Younger, and later studied under prominent artists of his time. Brueghel launched his career in Antwerp, where he became known for his intricate and detailed still-life paintings, particularly of flowers. In the early 17th century, Brueghel created a series of flower still-lifes, with this particular piece, "Bouquet of Flowers," completed around 1620. The painting is a testament to Brueghel's exceptional technical skill and his ability to capture the delicate textures and vibrant colors of various flowers in full bloom. The bouquet is densely packed with a diverse array of blossoms, including lilies, tulips, and roses, each rendered with meticulous attention to detail. The composition is set against a dark background, which accentuates the luminosity of the flowers and creates a striking contrast. The vase, intricately decorated with medallions depicting water and earth, adds an element of luxury and craftsmanship to the scene. The wooden table beneath the vase is rendered with such precision that it adds depth and realism to the painting.
Brueghel's fascination with nature led him to study flowers in the garden of the Archdukes in Brussels. This dedication to observing his subjects firsthand allowed him to create works that were both aesthetically refined and scientifically accurate. The inclusion of insects such as butterflies and beetles further enhances the naturalistic quality of the painting. The Berlin painting is part of a series of flower still-lifes by Brueghel, all executed from life without the aid of assistants. This approach ensured that his works were imbued with a freshness and authenticity that was highly prized during the 17th century. The painting's exceptional detail and refined style earned Brueghel the nickname "Velvet Brueghel," underscoring his mastery of capturing the textures and materials in his compositions. In 1862, the painting was acquired from the collection of the restorer Christian Xeller for the Gemäldegalerie in Berlin. It remains a celebrated example of Brueghel's contribution to the genre of flower still-life painting, showcasing his unparalleled skill and his profound understanding of the beauty and complexity of the natural world.



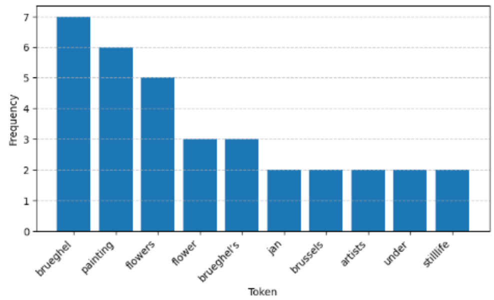


Figure 10. Selected (rank=1) description, wordcloud and token(word) distribution

4. VISION-BASED OPTICAL CHARACTER RECOGNITION AND SUMMARIZATION FOR HISTORIC TEXTS

4.1. OVERVIEW AND MOTIVATION

Figure 11 shows the pipeline of our framework. The ANBPR Biblioteca holds a vast collection of historical Romanian texts and manuscripts, many of which have been digitized from scanned documents (Figure 12). However, many manuscripts such as these suffer from quality degradation, missing segments, and inconsistencies due to varying print styles and formats. To enhance accessibility and comprehension, we introduce a pipeline that integrates OCR technology, error correction with LLMs, and audience-adapted summarization for historical texts.

Our approach ensures that digitized texts are not only transcribed accurately but also rewritten in contemporary and simplified language, making them accessible to broader audiences, including children. With this structured pipeline we aim for automated content restoration, style adaptation, and multi-step summarization, ensuring that CH materials remain accurate and relevant.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

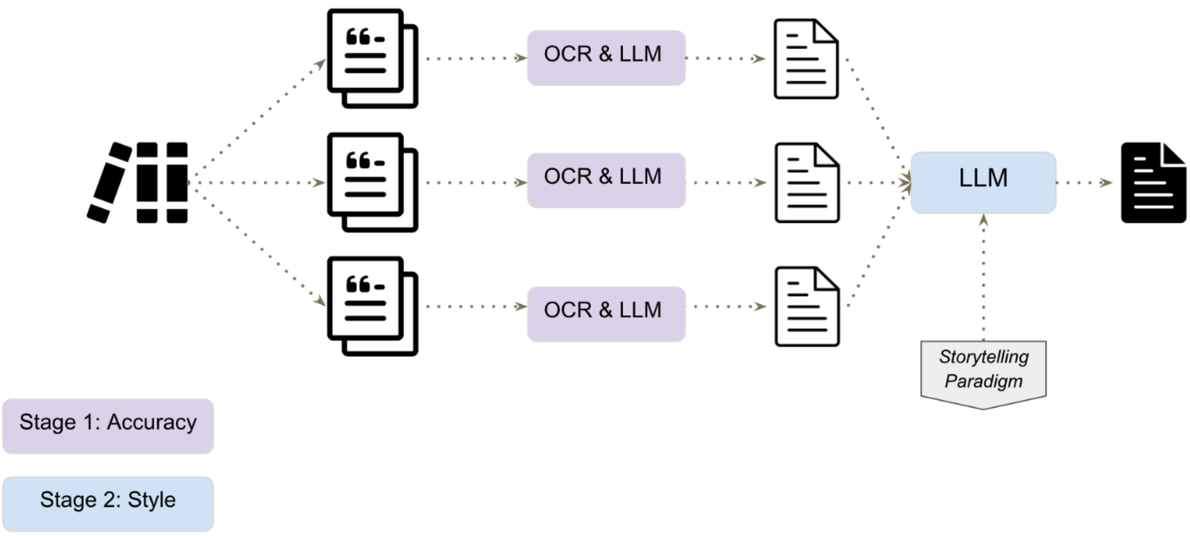


Figure 11. Overview of OCR with LLM correction tool for books summarization

4.2. OPTICAL CHARACTER RECOGNITION PIPELINE FOR CULTURAL HERITAGE TEXTS

Figure 12. Overview of the state of processed books, and synthesized samples for training simulating realistic condition ANBPR

OCR Model selection and fine-tuning. To achieve high transcription accuracy, we use EasyOCRxx as our baseline model, which performs average in qualitative

Funded by the European Union
This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

samples (Figure 14). However, to improve its performance on historical Romanian texts, we fine-tune it with a custom dataset that mimics historical print styles. The training dataset includes:

Scanned Romanian books provided by ANBPR, covering different print eras (i.e., background, fonts).

Artificially generated text samples with variations in fonts, paper textures, and noise levels to simulate degraded documents, see Figure 12 (using https://github.com/Belval/TextRecognitionDataGenerator).

Cyrillic and Latin-script Romanian texts, ensuring robust recognition across historical periods.

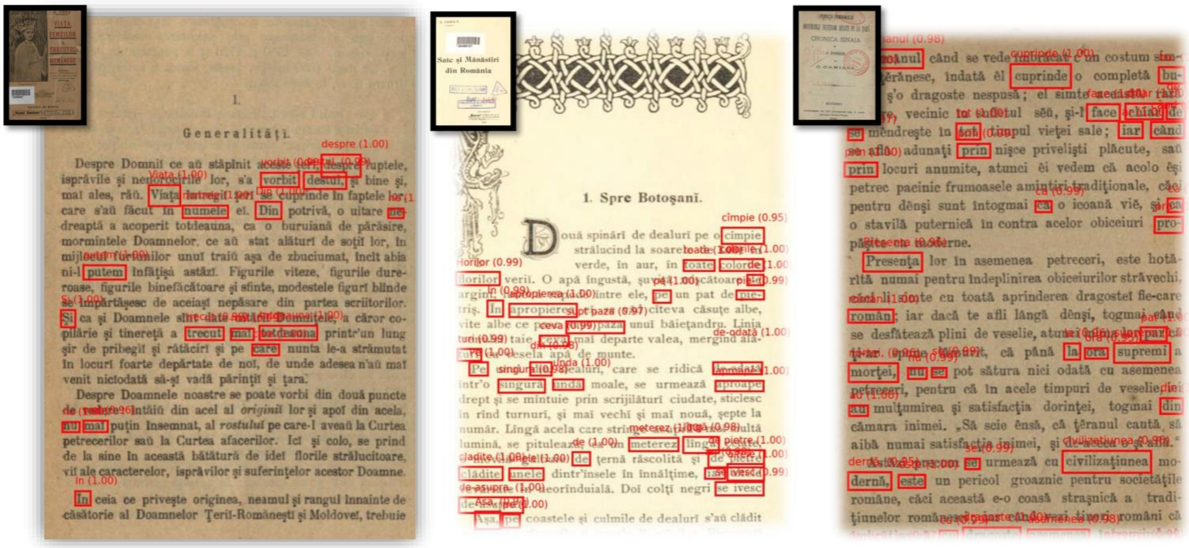


Figure 13. Word instances of high performance of OCR

Fine-tuning involves supervised training with character-level annotations and domain-specific augmentations, including background noise simulation, ink bleed effects, and faded text reconstruction. After training the model (Figure 14) on the generated data, we append it to our framework.

Post-Processing and Summarization. Once OCR transcriptions are obtained, the LLM-powered correction and summarization system refines the output. As seen in

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Figure 11, first, a book is divided into N chunks of K pages for iterative factual summarization.

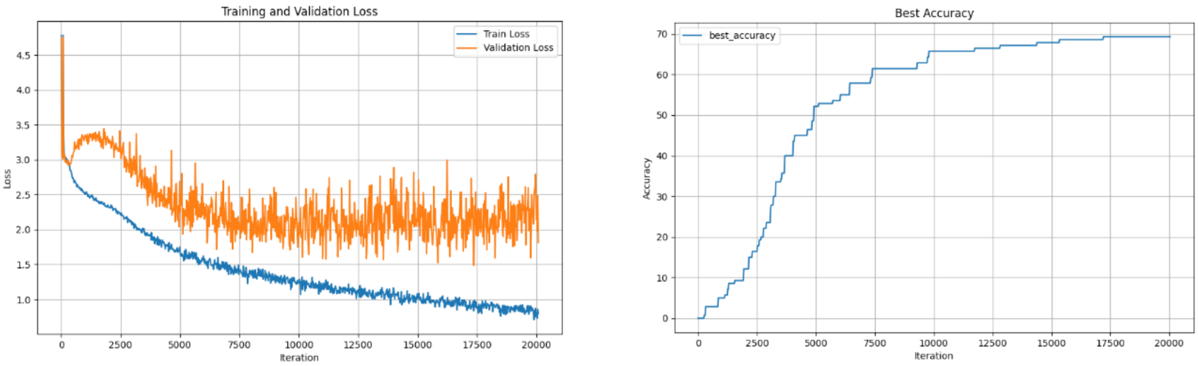


Figure 14. Loss decreases while training while accuracy increases.

For every set of pages, we apply Error Correction and Normalization with a fine-tuned Romanian LLM [masala24] (RoLlama2-7b-Instruct model[xxi]) on Romanian text to correct OCR errors. This step ensures accurate text restoration. Then the content is extracted recursively until a 10-page summary is reached. Finally, an {audience}-friendly summarization step applies the one-shot prototype style description. In the case of children as the audience we showcase the questionnaire implemented for the results.

4.3. CHILDREN-FRIENDLY SUMMARIZATION ADAPTATION

To ensure that the final summaries are engaging and appropriate for young readers, we incorporate findings from our children’s prototype questionnaire (see Annex). This study gathered user feedback on different description styles, evaluating readability, engagement, and linguistic complexity. Key takeaways included:

- Simplifying vocabulary to match different age groups.
- Using interactive and engaging storytelling elements.
- Encouraging imagination instead of over-explaining concepts.

The goal is the same; historical texts must be adapted to resonate with young audiences while maintaining factual accuracy.

We then utilize this prototype by *refining* it and form the one-shot example for our description generation with the LLM. In Table 3, we present the refinement

Funded by the European Union

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement no 101060660.

process, outlining how user feedback from the questionnaire was incorporated into the final version of the children's description. The descriptions refer to the painting Nighthawks (1942) by Edward Hopper which depicts four people in a brightly lit diner at night, as viewed through its large glass windows, contrasting with the empty city streets outside.

Since this process occurs only once and given the subjectivity of defining an *ideal* prototype curators play a crucial role in validating and refining the final selection. Their expertise ensures that the generated descriptions align with institutional guidelines, audience expectations, and age-appropriate engagement strategies.

To highlight key refinements made on the prototype, we have color-coded the differences: 🟢 Added parts (new details)

🔴 Removed parts (unnecessary or misleading details)

🔵 Modified parts (reworded for clarity & accuracy)

Original Description (selected one)	Refined Description
"Have you ever been awake late at night when everything is quiet? That's what Nighthawks is all about! You can see a shiny diner with bright lights and big windows. Inside, there are only a few people—two are talking quietly, but one man is sitting all alone. The streets outside are completely empty, like everyone has gone to bed. Do you think these people are lonely or just enjoying a peaceful night? Maybe they're waiting for something exciting to happen!"	"This painting is called Nighthawks. It was painted by Edward Hopper in 1942 and you can see it at the Art Institute of Chicago. It shows a bright diner with big glass windows on a quiet, empty street. Inside, four people sit at the counter. A man and woman are together, another man sits alone, and the server is behind the counter. The world outside is silent. What do you think they are thinking about? Do they know each other, or are they strangers? What would you say if you walked inside?"
"Have you ever been awake late at night when everything is quiet?"	🟢 "This painting is called *Nighthawks*. It was painted by Edward Hopper in 1942 and you can see it at the Art Institute of Chicago."
"That's what Nighthawks is all about!"	🔴 *Removed: redundant and vague*

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

"You can see a shiny diner with bright lights and big windows"	🔵 "It shows a bright diner with big glass windows on a quiet, empty street." *(simplified wording for clarity)*
"Inside, there are only a few people—two are talking quietly, but one man is sitting all alone."	🔵 "Inside, four people sit at the counter. A man and woman are together, another man sits alone, and the server is behind the counter." *(now includes all four characters, avoids assumptions about talking)*
"The streets outside are completely empty, like everyone has gone to bed."	🔵 "The world outside is silent" *(simplified & avoids unnecessary speculation)*
"Do you think these people are lonely or just enjoying a peaceful night? "	🔴 *(Removed: forces a specific interpretation of emotions.)*
"Maybe they're waiting for something exciting to happen!"	🔴 *(Removed: suggests a dramatic event that may not be present in the painting.)*
	🟢 "What do you think they are thinking about? Do they know each other, or are they strangers? What would you say if you walked inside?" *(Encourages imagination & open-ended interpretation instead of suggesting feelings*

Table 3. How user feedback from the questionnaire was incorporated into the final version of the children's description.

Now this will be the prototype we use in the process described in Figure 11 as a 'Storytelling Paradigm'.

To operationalize the pipeline, we developed a Streamlit-based interface (Figure 14) that allows users to process and summarize scanned historical texts efficiently. This interface provides an intuitive way to select OCR and LLM models, choose summarization styles, and upload scanned materials for automatic processing. The key features are:

- Model Selection: Users can choose the LLM (e.g., RoLlama2-7b-Base) for text correction and summarization.
- Summarization Task Type**:** The system supports different audience-specific styles:

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

   - *Children*: Simplified, engaging descriptions based on the one-shot prototype.
   - *Curator*: Scholarly summaries with precise historical and factual accuracy.
   - *Contemporary*: Modernized language for broader public accessibility.
- File Upload and Processing: Users can upload scanned PDFs or images, and the system performs OCR, correction, and summarization automatically.
- Output and Export Options: Summarized texts can be reviewed, exported, or converted into speech for enhanced accessibility

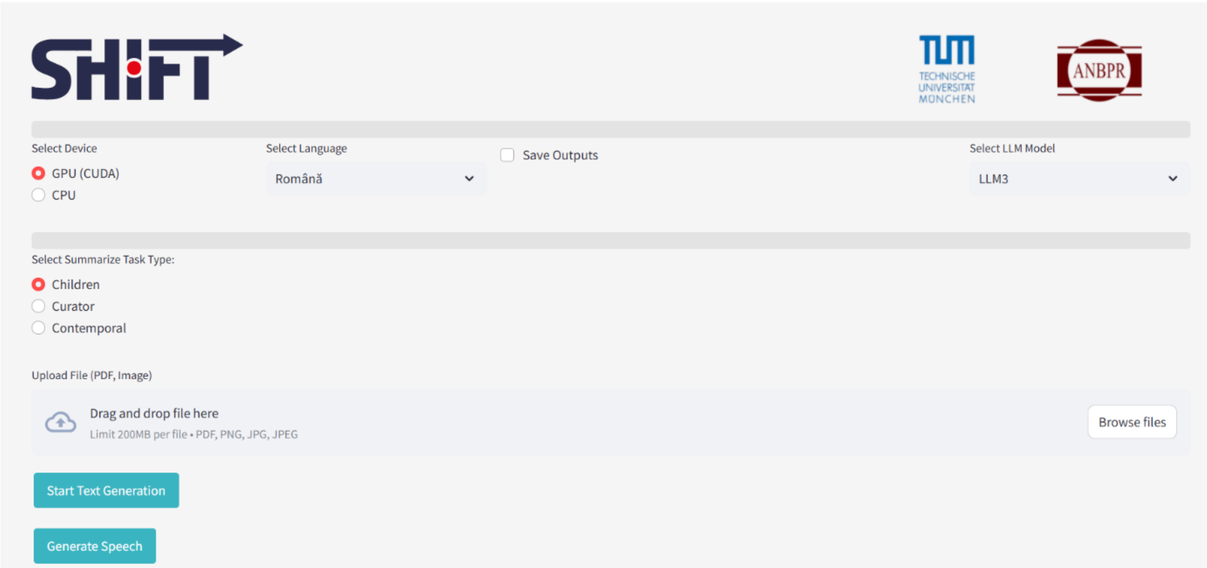


Figure 14. Streamlit application

Integrating the storytelling paradigm, the system enhances engagement while maintaining historical fidelity, ensuring that Romanian CH texts are accessible and engaging for different audiences.

5. CONCLUSIONS

This deliverable, D3.5, presents the final version of the Tool for the Textual Representation of CH Assets, a key outcome of the SHIFT project within WP3. The tool integrates SOTA NLP techniques, RAG, and multimodal storytelling methodologies to enhance accessibility, engagement, and semantic representation of CH assets across various user groups.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Through iterative development and testing, the tool has demonstrated its capacity to generate high-quality, audience-adapted textual descriptions of CH assets, supporting museums, libraries, and CH institutions in delivering enriched and inclusive experiences. By leveraging local, domain-specific datasets and minimizing reliance on proprietary models, our approach ensures data security, contextual relevance, and long-term adaptability.

The implementation of a RAG-based Q&A system has provided dynamic, context-aware responses tailored to different audiences. Multimodal adaptations have customized CH descriptions for diverse user groups, including children, researchers, and visually impaired individuals. The integration of VLMs has enabled automated generation of descriptions from images and metadata, improving scalability and efficiency. Furthermore, the OCR-enhanced summarization pipeline has facilitated access to digitized historical texts through AI-assisted transcription, correction, and adaptive summarization.

The tool has been successfully deployed and evaluated across multiple CH institutions, demonstrating its effectiveness in preserving cultural narratives while enhancing digital accessibility. Comparative studies, such as the BMN Museum of Knjaževac use case, validate the system's ability to generate engaging, audience-specific content while mitigating hallucinations and ensuring factual integrity.

Looking ahead, future development will focus on refining linguistic adaptation, integrating additional language models for underrepresented cultural contexts, and enhancing multimodal interactions through speech synthesis and haptic feedback. Continuous iteration based on user feedback and expanded collaboration with CH institutions will ensure the tool remains a valuable resource in the ongoing digital transformation of CH.

Ultimately, this deliverable marks a significant milestone in achieving SHIFT's broader objective of fostering inclusive, AI-driven cultural engagement while upholding ethical standards, data security, and institutional autonomy.

Funded by
the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

6. REFERENCES

[vaswani17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need.* Advances in Neural Information Processing Systems, 30.

[achiam23] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., et al. (2023). *GPT-4 Technical Report.* arXiv preprint arXiv:2303.08774.

[team23] Team, G., Anil, R., Borgeaud, S., Alayrac, J. B., Yu, J., Soricut, R., et al. (2023). *Gemini: A Family of Highly Capable Multimodal Models.* arXiv preprint arXiv:2312.11805.

[gao23] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., et al. (2023). *Retrieval-Augmented Generation for Large Language Models: A Survey.* arXiv preprint arXiv:2312.10997.

[alghisi24] Alghisi, S., Rizzoli, M., Roccabruna, G., Mousavi, S.M., & Riccardi, G. (2024). *Should We Fine-Tune or RAG? Evaluating Different Techniques to Adapt LLMs for Dialogue.* arXiv preprint arXiv:2406.06399.

[gupta24] Gupta, S., Ranjan, R., & Singh, S.N. (2024). *A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions.* arXiv preprint arXiv:2410.12837.

[zhang23] Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., & Wang, L. (2023). *Siren's song in the AI ocean: a survey on hallucination in large language models.* arXiv preprint arXiv:2309.01219.

[ouyang22] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., & Schulman, J. (2022). *Training language models to follow instructions with human feedback.* Advances in Neural Information Processing Systems, 35, 27730-27744.

[seo24] Seo, M., Baek, J., Thorne, J., & Hwang, S. J. (2024). *Retrieval-augmented data augmentation for low-resource domain tasks.* arXiv preprint arXiv:2402.13482.

[chen24] Chen, X., Tang, S., Zhu, R., Yan, S., Jin, L., Wang, Z., Su, L., Zhang, Z., Wang, X., & Tang, H. (2024). *The Janus interface: How fine-tuning in large*

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

language models amplifies the privacy risks. Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security, 1285-1299.

[koga24] Koga, T., Wu, R., & Chaudhuri, K. (2024). *Privacy-Preserving Retrieval Augmented Generation with Differential Privacy.* arXiv preprint arXiv:2412.04697.[kandpal23] Kandpal, N., Deng, H., Roberts, A., Wallace, E., & Raffel, C. (2023). *Large language models struggle to learn long-tail knowledge.* International Conference on Machine Learning, PMLR, 15696-15707.

[douze24] Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P. E., Lomeli, M., Hosseini, L., & Jégou, H. (2024). *The FAISS library.* arXiv preprint arXiv:2401.08281.

[grattafiori24] Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., & Yang, A. (2024). *The LLaMA 3 herd of models.* arXiv preprint arXiv:2407.21783.

[chen23] Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2023). *Unleashing the potential of prompt engineering in large language models: a comprehensive review.* arXiv preprint arXiv:2310.14735.

[zhang24] Zhang, J., Huang, J., Jin, S., & Lu, S. (2024). *Vision-Language Models for Vision Tasks: A Survey.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(8), 5625-5644. https://doi.org/10.1109/TPAMI.2024.3369699

[zhou22] Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022). *Learning to prompt for vision-language models.* International Journal of Computer Vision, 130(9), 2337-2348.

[du22] Du, Y., Liu, Z., Li, J., & Zhao, W. X. (2022). *A survey of vision-language pre-trained models.* arXiv preprint arXiv:2202.10936.

[chen23] Chen, F.-L., et al. (2023). *VLP: A survey on vision-language pre-training.* Machine Intelligence Research, 20(1), 38-56.

[ghosh24] Ghosh, A., Acharya, A., Saha, S., Jain, V., & Chadha, A. (2024). *Exploring the frontier of vision-language models: A survey of current methodologies and future directions.* arXiv preprint arXiv:2404.07214.

[khandelwal22] Khandelwal, A., Weihs, L., Mottaghi, R., & Kembhavi, A. (2022). *Simple but Effective: CLIP Embeddings for Embodied AI.* Proceedings of the

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 14829-14838.

[dosovitskiy20] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale.* arXiv preprint arXiv:2010.11929.

[wang24] Wang, W., Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., et al. (2024). *Enhancing the reasoning ability of multimodal large language models via mixed preference optimization.* arXiv preprint arXiv:2411.10442.

[chen24] Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., et al. (2024). *Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling.* arXiv preprint arXiv:2412.05271.

[triantafyllopoulos24] Triantafyllopoulos, A., Tsangko, I., Gebhard, A., Mesaros, A., Virtanen, T., & Schuller, B. (2024). *Computer audition: From task-specific machine learning to foundation models.* arXiv preprint arXiv:2407.15672.

[masala24] Masala, M., Ilie-Ablachim, D. C., Dima, A., Corlatescu, D., Zavelca, M., Olaru, O., Terian-Dan, S., Terian-Dan, A., Leordeanu, M., Velicu, H., & Popescu, M. (2024). *"Vorbeşti Româneşte?" A Recipe to Train Powerful Romanian LLMs with English Instructions.* arXiv e-prints, arXiv-2406.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

ANNEX

ANNEX 1. SUMMARIZATION ADAPTATION-SURVEY RESULTS

Overview. To ensure that generated summaries align with audience expectations, a structured user feedback study was conducted. The study evaluated different textual descriptions of Nighthawks (1942) by Edward Hopper, focusing on readability, engagement, and appropriateness for different target groups. Participants selected the most effective descriptions and provided qualitative feedback.

Key Findings. The responses revealed important insights into audience-specific preferences:

- ❖ Children's descriptions benefited from engaging, imaginative storytelling with simple vocabulary.
- ❖ Professional descriptions required analytical depth, focusing on artistic interpretation and historical context.
- ❖ Visually impaired-friendly descriptions prioritized sensory-rich narratives, emphasizing spatial awareness and atmosphere

Results. Among 37 participants, 16.2% identified as visually impaired. The evaluation involved three alternative descriptions for each audience group. The selection results were as follows

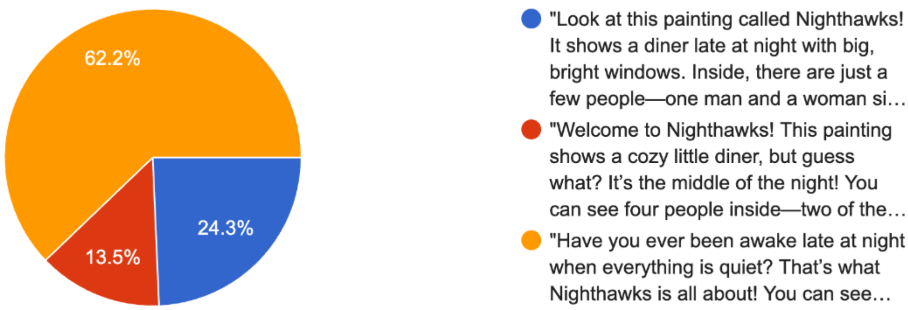
Audience Type	Preferred Description	Percentage
Children's Audience	Engaging and vivid storytelling	62.2
	Structured but slightly formal narrative	24.3
	Encouraging self-reflection	13.5
Professional Audience	Detailed art-historical analysis	40.5
	Focus on urban disconnection	32.4
	Emphasis on stylistic choices	27
Visually Impaired Audience	Vivid depiction of diner's lighting and atmosphere	40.5
	Strong sensory details and emotional depth	35.1
	Structured and spatially detailed interpretation	24.3

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Please select the best description for Nighthawks (Children Audience)

37 responses



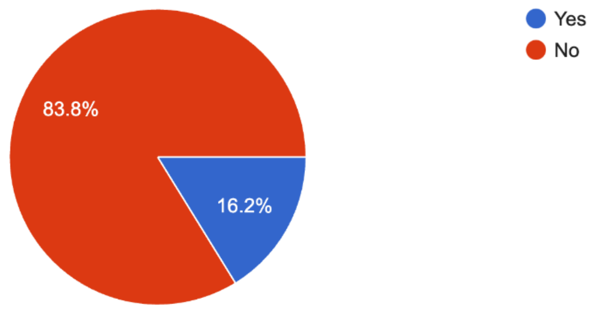
Please select the best description for Nighthawks (Professional Audience)

37 responses

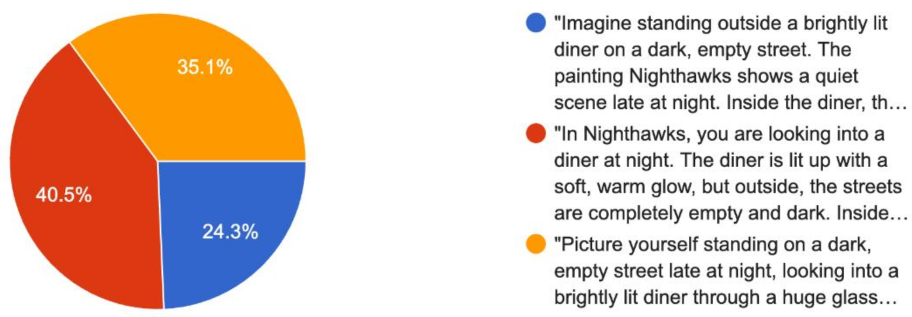
Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Do you identify as a person with a vision impairment (partially sighted or blind)?
37 responses



Please select the best description for Nighthawks (Visually Impaired Audience)
37 responses



Please provide any additional comments or suggestions. Feel free to suggest an alternate description that may be more appropriate for the audience (***Visually Impaired*** *Audience*).

[i] https://huggingface.co/collections/meta-llama/llama-32-66f448ffc8c32f949b04c8cf
[ii] https://pypi.org/project/pypdf/
[iii] https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[iv] https://huggingface.co/TechWolf/ConTeXT-Skill-Extraction-base

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

[v] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct
[vi] https://www.langchain.com/
[vii] https://www.europeana.eu/
[viii] https://www.wikipedia.org/
[ix] https://www.wikidata.org/wiki/Wikidata:Main_Page
[x] https://www.getty.edu/research/tools/vocabularies/
[xi] https://www.deutsche-digitale-bibliothek.de/
[xii] https://iconclass.org/en/_
[xiii] https://spacy.io/
[xiv] https://pypi.org/project/py-readability-metrics/
[xv] https://github.com/cjhutto/vaderSentiment
[xvi] https://huggingface.co/OpenGVLab/InternVL2_5-8B-MPO
[xvii] https://github.com/InternLM/InternLM
[xviii] https://github.com/QwenLM/Qwen2.5
[xix] nomic-ai/nomic-embed-text-v1
[xx] https://github.com/JaidedAI/EasyOCR
[xxi] https://huggingface.co/OpenLLM-Ro/RoLlama2-7b-Instruct

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.