

Knowledge engineering information technology for cultural-educational scenarios based on RAG

Khrystyna Lipianina-Honcharenko^{1,*†}, Nazar Melnyk^{1,†}, Myroslav Komar^{1,†},
Pavlo Bykovyy^{1,†} and Khrystyna Yurkiv^{1,†}

¹West Ukrainian National University, Lvivska 11, 46000 Ternopil, Ukraine

Abstract

The digital transformation of cultural heritage and education necessitates systems capable of reliably extracting knowledge from large-scale corpora while presenting it in an understandable, verifiable, and ethically safe manner. However, the application of Large Language Models (LLMs) in museum scenarios is often hindered by hallucinations, a lack of source traceability, and safety risks for child audiences. In this work, we propose a reproducible knowledge engineering information technology based on Retrieval-Augmented Generation (RAG), specifically designed for interactive “talking exhibits.” The proposed architecture integrates ethical corpus construction, semantic search with re-ranking, and controlled generation with explicit source citation and “honest uncertainty” policies. Validation was conducted on a demonstration case of the historical figure Jan Tarnowski, where the system achieved 100% compliance with role-based linguistic constraints (8/8 queries) and 62.5% contextual faithfulness (5/8 queries fully confirmed). While localized risks of hallucinations persisted in specific domains such as law and finance, the outcomes demonstrate a balanced improvement in accuracy, explainability, and safety compared to isolated LLM usage. These results suggest that the proposed technology effectively bridges the gap between open generative models and the rigorous requirements of educational cultural heritage applications.

Keywords

Retrieval-augmented generation, cultural heritage, knowledge engineering, semantic search

1. Introduction

The digital transformation of cultural heritage and education has highlighted the need for systems capable of reliably extracting knowledge from large-scale open corpora and presenting it in an understandable, verifiable, and ethically safe form for diverse audiences. Despite the advances of large language models (LLMs), their application in museum and cultural-educational scenarios faces several limitations: from hallucinations and lack of source traceability to licensing risks and multilingual challenges [1, 2]. The absence of reproducible end-to-end methodologies that simultaneously combine ethical corpus construction, semantic search, controlled generation with citation, and evaluation protocols underlines the scientific and practical significance of this research.

The aim of this work is to develop and verify a RAG-based (Retrieval-Augmented Generation) knowledge engineering technology for cultural-educational applications that:

1. builds a licensing-compliant and reproducible knowledge corpus;
2. ensures semantic search and re-ranking of relevant fragments;
3. generates grounded responses with explicit source citation;
4. implements safety policies and personal data minimization;
5. supports telemetry and offline/online quality evaluation.

ExplAI-2025: Advanced AI in Explainability and Ethics for the Sustainable Development Goals, November 07, 2025, Khmelnytskyi, Ukraine

*Corresponding author.

†These authors contributed equally.

✉ xrustya.com@gmail.com (K. Lipianina-Honcharenko); 88nazar88@gmail.com (N. Melnyk); mko@wunu.edu.ua (M. Komar); pb@wunu.edu.ua (P. Bykovyy); kh.yurkiv@wunu.edu.ua (K. Yurkiv)

ORCID 0000-0002-2441-6292 (K. Lipianina-Honcharenko); 0009-0000-5917-1099 (N. Melnyk); 0000-0001-6541-0359 (M. Komar); 0000-0002-5705-5702 (P. Bykovyy); 0009-0007-4917-3251 (K. Yurkiv)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

The practical focus is on interactive museum “talking exhibits” that respond in the first person, mirroring the user’s language of interaction, and demonstrating fact traceability. The studied problem is interdisciplinary: it requires coordinated solutions at the levels of data collection and cleaning (HTML cleaning, normalization, deduplication, versioning), information retrieval (vectorization, indexing, re-ranking), natural language processing (controlled generation, role-based modeling, style/tone control), and safety engineering (content moderation, child-audience policies, privacy) [3]. A particular challenge lies in ensuring traceability: each generated fact must be linked to specific corpus fragments, which increases user trust and facilitates expert verification.

The working hypothesis is that combining a licensing-compliant corpus, semantic search with re-ranking, and controlled generation with citation achieves a better balance between accuracy, explainability, and safety than isolated use of LLMs or traditional search methods. We expect that the proposed technology will provide stable language mirroring, first-person narrative, increased contextual faithfulness, and controlled behavior in cases of uncertainty (refusal/clarification instead of fabrication).

The paper is structured as follows. Section 2 systematizes approaches to ethical corpus construction and RAG systems. Section 3 presents the architecture and step-by-step protocol. Section 4 outlines the pilot and validation results, followed by conclusions in Section 5.

2. Related works

The literature review in this work structures contemporary approaches to building RAG systems for cultural-educational scenarios through five interrelated vectors:

1. ethical corpus construction (HTML cleaning, normalization, deduplication, licensing compliance, and attribution);
2. methods of text segmentation into semantic fragments (fixed-size, structure-oriented, semantic, and hierarchical chunking) and their vectorization for semantic search;
3. re-ranking and context compression considering positional and semantic features;
4. controlled generation with explicit source citation, style/tone control, and “honest uncertainty” policies;
5. safety and audience compliance (including child audiences), as well as offline/online evaluation.

This perspective makes it possible to align the requirements of traceability and reproducibility with the practical applicability of museum “talking exhibits.” At the same time, several gaps are identified [4]: the lack of standardized fragment-level attribution protocols, insufficient coverage of multilingual corpora, and limited development of metrics sensitive to factual groundedness.

For building such systems, open text data are typically used—web pages, archives, and other sources. A notable example is the C4 corpus (Colossal Clean Crawled Corpus), created on the basis of Common Crawl with subsequent cleaning [5]. During HTML parsing, libraries such as Boilerpipe or Readability are employed to extract the main text and metadata while filtering out advertisements or menus. However, filtering may have unintended consequences: for instance, in C4, the use of banned-word lists led to the removal of texts about marginalized groups. Consequently, more cautious methods are now applied—combining dictionaries, toxic content detection, and classifiers for identifying personal data. In addition to cleaning, text normalization is essential: correcting encodings, merging split words, and standardizing case and spelling variants.

License and usage verification is also a critical step. In the research community, the movement toward open data has accelerated. The Common Pile v0.1 corpus, comprising 8 TB of data, contains only texts with open licenses (public domain, Creative Commons) [6]. Results showed that language models trained on such an “ethical” corpus achieve quality comparable to models trained on unverified web data [6]. Therefore, in educational and cultural projects, it is recommended to use sources with explicit licenses (e.g., Wikipedia, open-access publications) or to obtain permission from rights holders, as well as to ensure proper attribution of sources in generated responses.

Content deduplication is another key stage. During web scraping, duplicates or near-duplicate texts often appear. In the C4 dataset, for example, a 50-word passage was found repeated 60,000 times [7].

The study [7] demonstrated that removing duplicates accelerates model training (by reducing dataset size) and lowers the risk of memorization—cases where the model reproduces training text verbatim. Deduplication algorithms range from simple string hashing to building suffix arrays and searching for repeated segments with a predefined similarity threshold [7].

For effective information retrieval, large texts are split into semantic fragments (chunks). Different chunking strategies are discussed in current research. The simplest approach is fixed-size segmentation (e.g., 500 characters per fragment). While easy to implement, this may split logically connected paragraphs [8]. Another approach is recursive segmentation by structural elements (newlines, punctuation), which takes into account text formatting. A more advanced method is semantic chunking, which groups sentences based on embedding similarity. Semantic chunks are usually internally coherent, but overly narrow grouping may lose the broader context of the document. Recent studies propose hybrid solutions, such as hierarchical segmentation, where smaller segments are first formed and then clustered into higher-level units by meaning [8].

The LongRAG model [9] uses larger fragments (entire articles) to reduce the size of the knowledge base, while the RAPTOR method [10] builds multi-level hierarchies ranging from detailed content to higher-level abstractions. All of these approaches aim to preserve coherent semantic units, thereby making retrieval more accurate and useful.

After segmentation, each chunk is converted into a vector using embedding models (e.g., Sentence-BERT or large language models in embedding mode). Vectorization of knowledge enables semantic search: the user query is also transformed into a vector, and the system retrieves the closest chunk vectors using cosine similarity. For the implementation of vector knowledge bases, the FAISS (Facebook AI Similarity Search) library is widely used [11]. This toolkit provides fast Approximate Nearest Neighbor (ANN) search for billions of embeddings and is integrated into vector databases such as Milvus and Pinecone.

Semantic search based on embeddings enables retrieval of relevant fragments, but the nearest vector by distance does not always yield the best context. Therefore, many approaches employ additional re-ranking of results. A classical method is to pass the top-K retrieved chunks through a stronger model (e.g., a BERT cross-encoder or an LLM) for more accurate relevance scoring. It is well known that BERT-based re-rankers significantly improve performance in text-based Question Answering (QA) tasks compared to sparse retrieval or embedding-only retrieval [12].

Other optimizations include position-sensitive ranking, where keyword proximity to the beginning of a chunk is considered, and context compression, where non-essential details are discarded from candidate fragments [13]. In domain-specific systems, context merging may also be used—when multiple chunks refer to the same entity, they can be combined into a joint context before answer generation. However, overly aggressive merging may lead to errors.

The effectiveness of retrieval is evaluated using standard information retrieval metrics—Precision@K, Recall@K, MRR, NDCG, etc.—to determine whether the top-ranked documents contain the “correct” fragments [12]. For greater accuracy, the Context Recall metric is sometimes applied, which measures whether the retrieved context includes all facts from the reference answer. During system development, engineers employ a set of control queries with known answers to verify whether the retrieval module can locate the required paragraph.

Once relevant fragments are retrieved, a large language model (LLM) generates the response. This approach, which leverages external knowledge, is called retrieval-augmented generation (RAG). It significantly improves the accuracy and factual reliability of responses while reducing model hallucinations [14]. In particular, the LLM receives a system prompt or instruction to incorporate the retrieved facts and avoid unsupported claims. The best results are achieved by prompting with context, i.e., inserting knowledge fragments directly into the model’s input prompt, which ensures that the answer is grounded in the provided sources. For example, the OpenAI WebGPT project demonstrated that when a model cites web sources to support its answers, users perceive it as more trustworthy [14].

An alternative approach is post-hoc citation, where the answer is generated first and then aligned with corresponding sources. However, in practice, one-step methods are more common, where the model integrates citations directly into the generated text. This is implemented, for instance, in the

Atlas system, which attaches a document index to each fact [15].

Scientific research also addresses the control of style and tone in generated responses. For educational applications—especially in child mode—the model should answer at a level comprehensible to children and consistent with ethical norms. Two approaches are possible: (i) training a dedicated child-focused model, or (ii) adapting a universal model with style-specific instructions. The KidLM project demonstrated the advantages of the first approach: a dedicated children’s corpus (~50 million words) was collected, and the language model was further trained, leading to better simplification of complex vocabulary and avoidance of toxic expressions [16]. Alternatively, when using a general-purpose LLM, roles or modes can be embedded via prompting. Such techniques belong to prompt engineering and are widely applied in practice.

Content safety is a critical factor for cultural-educational applications. It is well known that large language models may generate undesirable or harmful text if prompted, or if such content was present in training data. The authors of “LLM Safety for Children” developed specialized child personas and scenarios for testing several state-of-the-art models and identified categories of risks that are not reliably avoided by standard filters [17]. This underscores the necessity of specialized filters and rule-based safeguards. In existing solutions, moderation is typically implemented at two levels. The first is query-level filtering, where the system either refuses or provides a safe response to prohibited topics. The second is answer-level filtering, where toxicity detectors evaluate the generated response before presenting it to the user. The tone of responses is also regulated: the model should be friendly, encourage curiosity, yet avoid harmful advice. The study “No, Alexa, no!” [18] highlighted the risk of children placing excessive trust in voice assistants and the need to impose restrictions so that AI does not present itself as an authoritative adult in sensitive topics.

To maintain high system quality, both offline and online evaluation are conducted. System telemetry (Stage 11) makes it possible to collect statistics on dialogue counts, popular queries, and changes in accuracy after updates. These data are typically used in industrial deployments for iterative improvement (e.g., A/B testing of prompts, comparing model versions). In scientific publications, however, the focus is placed on objective metrics. Specifically, the quality of knowledge retrieval is assessed using metrics such as Recall@5 or MRR, which indicate how well the system retrieves the correct documents [12].

Alongside classical information-retrieval measures, modern RAG research increasingly relies on metrics that explicitly assess whether generated claims are grounded in retrieved context. RAGAS (Retrieval-Augmented Generation Assessment) and ARES (Automatic RAG Evaluation System) represent state-of-the-art frameworks combining retrieval checks with LLM-based critics to score groundedness, hallucination frequency, and answer faithfulness to evidence [19, 20]. Unlike lexical metrics such as BLEU/ROUGE, these approaches focus on semantic alignment and factual traceability, providing a more reliable evaluation of context-dependent systems.

System behavior is also tested under conditions of knowledge scarcity. The correct approach is for the system to acknowledge uncertainty or offer alternatives (e.g., suggest additional search). In research, this issue is termed refusal or abstention. The R-Tuning (Refusal-aware tuning) method trains LLMs to abstain from fabrications when a query falls outside their knowledge boundaries [21]. Models trained in this way better recognize uncertainty and answer with “I don’t know” instead of hallucinating. For educational applications, this is especially important: when uncertain, the system should direct users to an expert or state that more information is needed, rather than providing incorrect guidance.

3. Technology description

A knowledge engineering technology has been developed for cultural-educational scenarios that includes:

1. collection and normalization of text sources;
2. segmentation (chunking) with overlap;
3. vectorization (embedding) and indexing;
4. semantic search and context re-ranking;

5. controlled LLM-based answer generation with grounding guarantees;
6. telemetry, offline/online evaluation, and testing protocol.

The technology is suitable for both “adult” and “child” modes, adheres to principles of personal data minimization, and ensures reproducibility. The proposed technology consists of 13 stages, as illustrated in the diagram (Figure 1).

The detailed stages are as follows:

1. **Collection and Parsing of Open Sources:**
 - a) Automatically or semi-manually collect links.
 - b) Download pages as raw bytes, handling redirects and errors.
 - c) Extract clean text from HTML along with basic metadata (title, source, date).
2. **Verification of Usage Rights and Attribution:**
 - a) Record license terms, copyrights, and citation requirements.
 - b) Exclude sources that do not permit reuse.
3. **Cleaning, Normalization, and Content Filtering:**
 - a) Fix encodings, remove extraneous blocks such as menus and advertisements.
 - b) Normalize spacing, line breaks, languages, and transliterations.
 - c) Remove toxic, undesirable, or personally sensitive content.
4. **Deduplication and Corpus Versioning [22, 23]:**
 - a) Detect duplicates by textual similarity.
 - b) Store control hashes, corpus version, and change logs.
5. **Segmentation into Semantic Fragments [22, 23]:**
 - a) Create chunks respecting paragraphs and headings.
 - b) Apply overlapping windows to improve context retrieval.
 - c) Add service tags (section, source, time period).
6. **Knowledge Vectorization:**
 - a) Compute embeddings for each chunk.
 - b) Normalize and store arrays for efficient cosine similarity search.
 - c) Validate baseline quality with control queries.
7. **Preparation of Storages and Indexes:**
 - a) Store texts and vectors in a repository (file or database).
 - b) Record the vector file path in the exhibit database.
 - c) Configure backups and integrity checks.
8. **Exhibit Configuration in the Admin Panel:**
 - a) Create an exhibit record (title, description, image, system settings).
 - b) Indicate whether child mode is enabled.
 - c) Link the path to the vector file.
9. **Safety and Moderation Policies:**
 - a) Define rules for style, tone, and restricted topics.
 - b) Add instructions for child audiences (simplified language, caution with sensitive topics).
 - c) Enable baseline filters and answer checks before display.
10. **Client Flow Configuration:**
 - a) Session start form with minimal data (date of birth, anonymous identifier).
 - b) Branching by age (child or standard mode).
 - c) Construct a message for the language model (system prompt, user question, retrieved contexts).
11. **Analytics and Logging:**
 - a) Log session starts, number of messages, average dialogue frequency, errors, and timings.
 - b) Prepare summary statistics for the admin panel.
12. **Knowledge Retrieval Testing:**
 - a) Use a set of control queries that must retrieve specific fragments.
 - b) Evaluate context relevance, coverage accuracy, and retrieval stability.
13. **Model Answer Evaluation:**
 - a) Verify factual correctness, completeness, style, and tone alignment.
 - b) Test behavior under knowledge gaps—the model should honestly indicate lack of information.

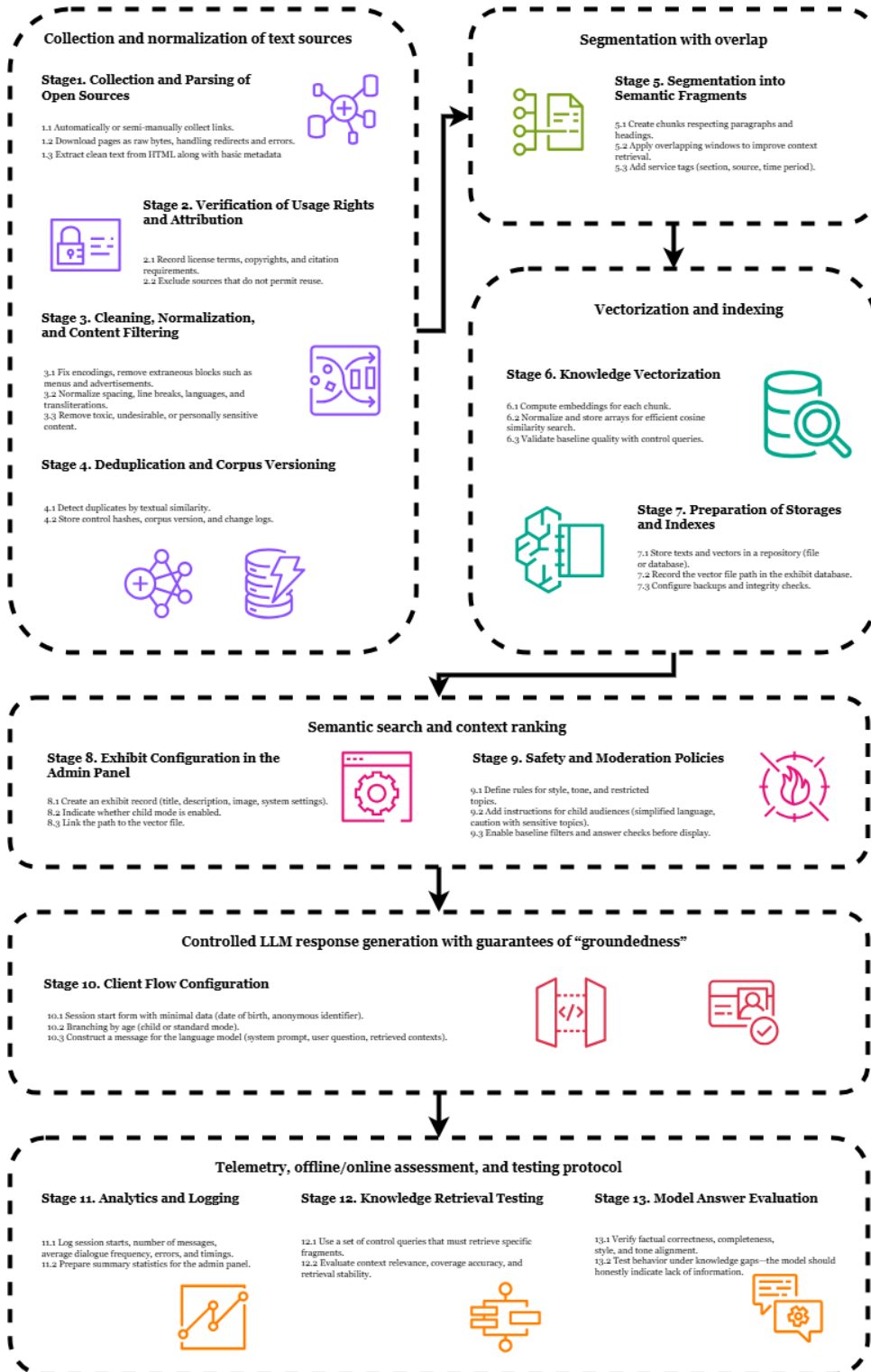


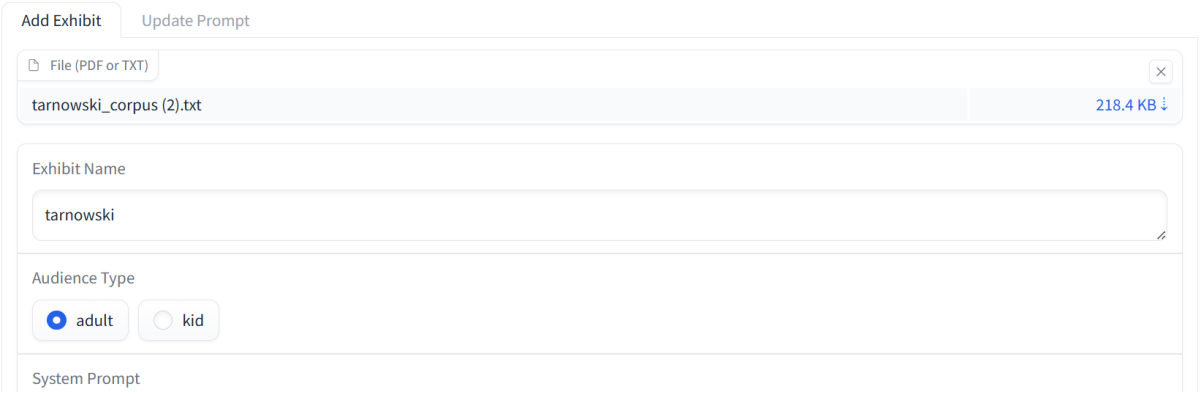
Figure 1: Diagram of the Knowledge Engineering Information Technology for Cultural-Educational Scenarios. The process flows from source collection to evaluation, highlighting key stages such as segmentation, vectorization, and safety policies.

4. Implementation

To construct the demonstration knowledge corpus, automated collection and preprocessing of open text sources about the historical figure Jan Tarnowski were performed. At the search-and-parsing stage, 8 unique links were identified; during main content extraction, 7 resources were successfully

processed, while 1 resource returned an error (details recorded in the execution log). The combined text was consolidated into a single file `/content/tarnowski_corpus.txt`, and process metadata were saved in `/content/tarnowski_log.jsonl`.

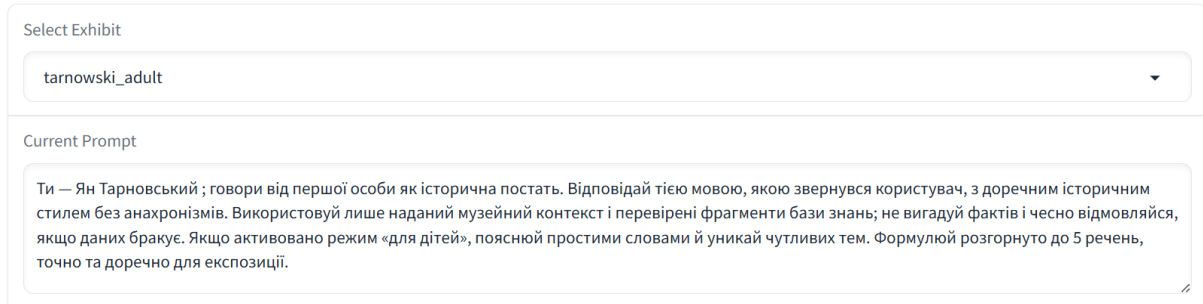
The pipeline included encoding normalization, HTML cleaning, extraction of article “body” text, denoising, and basic deduplication. The resulting material was segmented into overlapping semantic fragments for subsequent semantic search. After corpus formation, the text file was uploaded to the developed platform (Figure 2), where an exhibit was created in the admin module and a vector representation of knowledge was generated for approximate fragment retrieval. This integration ensures reproducibility, quality control (via state and encoding logs), and readiness for initial expert testing of functionality, independent of a specific museum object.



The screenshot shows a web interface with two tabs: "Add Exhibit" (active) and "Update Prompt". Under "Add Exhibit", there is a file upload section with a button "File (PDF or TXT)" and a file named "tarnowski_corpus (2).txt" with a size of "218.4 KB". Below this is a text input field for "Exhibit Name" containing the word "tarnowski". Further down is the "Audience Type" section with two radio buttons: "adult" (selected) and "kid". At the bottom is a section for "System Prompt" which is currently empty.

Figure 2: Platform Interface showing the exhibit configuration options.

Next, the avatar’s role-based linguistic behavior was configured so that the character responds in the first person and in the language of the user’s query, relying exclusively on the uploaded corpus. The system prompt (see Figure 3) defines the formal role model of the exhibit’s linguistic behavior and regulates the boundaries of generation. Its design aims to minimize hallucinations, preserve historical stylistics, and maintain controllable response tone for different audiences.



The screenshot shows a web interface with a "Select Exhibit" dropdown menu showing "tarnowski_adult". Below it is a section titled "Current Prompt" containing a text area with the following text: "Ти — Ян Тарновський ; говори від першої особи як історична постать. Відповідай тією мовою, якою звернувся користувач, з доречним історичним стилем без анахронізмів. Використовуй лише наданий музейний контекст і перевірені фрагменти бази знань; не вигадуй фактів і чесно відмовляйся, якщо даних бракує. Якщо активовано режим «для дітей», поясний простими словами й уникай чутливих тем. Формулюй розгорнуто до 5 речень, точно та доречно для експозиції."

Figure 3: Platform Interface for Prompt Input, defining the persona and constraints.

The prompt fixes the narrator’s identity (first-person perspective), ensures language mirroring (responding in the user’s language), enforces strict linkage to the source corpus (only the provided museum context), and applies an uncertainty policy (honest refusal in case of missing data). A response length limit of five sentences disciplines the output and reduces the risk of accumulating secondary assumptions.

Next, a simple validation test of the avatar was conducted using queries (five in Ukrainian and one each in English, Polish, and German) aimed at verifying first-person narrative, language mirroring, and grounding in the provided corpus. The list of queries is presented in Table 1.

The results of the mini-test are recorded in Table 2 (Figure 4) according to four criteria — compliance with the query language, adherence to first-person narrative, grounding in the corpus, and response

Table 1
Questions for Avatar Mini-Test

No	Question	Verification Goal
1	Розкажи про своє походження та ранні роки: з якого дому ти походиш і де народився? (EN) Tell me about your origin and early years: from which family do you come, and where were you born?	First-person narrative, basic biographical facts from the corpus
2	Що саме ти заснував у Тарнополі й з якою метою? (EN) What exactly did you establish in Ternopil, and for what purpose?	Factual accuracy, local event (foundation)
3	Які військові кампанії або походи ти очолював і які наслідки вони мали? (EN) Which military campaigns or expeditions did you lead, and what were their outcomes?	Historical events without fabrication, concise summary
4	Які лицарські чесноти ти вважав найважливішими для себе? (EN) Which knightly virtues did you consider most important for yourself?	Consistency with style and values in the corpus
5	Коли і де завершилося твоє життя, і як пройшли поховальні урочистості? (EN) When and where did your life end, and how were the funeral ceremonies conducted?	Chronology, verification of corpus references
6	Якими першоджерелами у наданому корпусі згадуються твої укази чи листування, і що з них впливає? (EN) Which primary sources in the provided corpus mention your decrees or correspondence, and what do they reveal?	Works with “sources”; language mirroring (UA)
7	Jak twoje decyzje gospodarcze i zarządzanie majątkiem wpłynęły na rozwój regionu? (EN) How did your economic decisions and estate management influence the development of the region?	Economic/administrative influence; language mirroring (PL)
8	Wie wird dein Andenken in späteren Chroniken und Studien beschrieben? (EN) How is your memory described in later chronicles and studies?	Image in later historiography; language mirroring (DE)

time.

Your question

Choose an exhibit

Refresh Exhibits

Answer

Я, Ян Тарновський, належу до давнього магнатського дому Тарновських (гербу Леліва) з Малої Польщі, гніздом якого був Тарнів. Батько мій, пан Ян, був краківським каштеляном, але я втратив матір, коли мені ледь виповнився рік, і залишився сиротою при живому батькові. Ранні роки свого виховання я провів під опікою підканцлера Мацея Древецького при дворі короля Ольбрахта, де здобував перші грамоти й латинську науку. Уже в два роки говорив чітко, у п'ять опанував «Доната», у десять міг декламувати сто віршів Вергілія, а в п'ятнадцять—писати латинські листи до королівської канцелярії.

Ask

Figure 4: Platform interface for question–answer interaction.

Table 2

Form for recording the results of the avatar mini-test

1	<p>Original Answer: Я, Ян Тарновський, належу до давнього магнатського дому Тарновських (гербу Леліва) з Малої Польщі...</p> <p>Translation: I, Jan Tarnowski, belong to the ancient magnate house of Tarnowski...</p> <p>Language (+) / First-person (+) / Response time (s) 15.1/9.6</p> <p>Context faithfulness: High. Confirmed: “Leliwa,” “Castellan of Kraków,” “Tarnów,” “Maciej Drzewiecki,” references to classical education (Virgil)</p>
2	<p>Original Answer: Я заснував у 1540 році на Волині місто-фортецю Тарнопіль (назване на честь мого родового герба і Тарнова)...</p> <p>Translation: I founded in 1540, in Volhynia, the fortress-town of Tarnopol...</p> <p>Language (+) / First-person (+) / Response time (s) 9.3/9.6</p> <p>Context faithfulness: High. Confirmed: “1540,” “Tarnopol/Ternopil,” “Magdeburg rights,” defensive function (Tatars).</p>
3	<p>Original Answer: Я очолював дві найвідоміші воєнні кампанії. У 1531 році, за наказом короля Сигізмунда...</p> <p>Translation: I led two of the most renowned military campaigns. In 1531, by order of King Sigismund...</p> <p>Language (+) / First-person (+) / Response time (s) 13.4/9.6</p> <p>Context faithfulness: High. Confirmed: “Obertyn/1531,” “Khotyn/1538,” “Petru Rareș.”</p>
4	<p>Original Answer: Найважливішими для себе я вважав передусім справедливість і поміркованість...</p> <p>Translation: The virtues I considered most important for myself were, above all, justice and moderation...</p> <p>Language (+) / First-person (+) / Response time (s) 12.5/9.7</p> <p>Context faithfulness: Medium-High. Present in the corpus are the virtues: justice, moderation, mercy, piety, humility; the formulation generalizes but does not contradict.</p>
5	<p>Original Answer: Моє життя завершилося 16 травня у 73-му році віку. Помер я в своїх володіннях...</p> <p>Translation: My life ended on May 16th, in the seventy-third year of my age...</p> <p>Language (+) / First-person (+) / Response time (s) 20.0/9.8</p> <p>Context faithfulness: High. Present are “May 16,” “73,” “Tarnów,” “Bernardines,” “bishops”; the description of the funeral is confirmed by corpus fragments.</p>
6	<p>Original Answer: In the royal court of Piotrków you may still read the very edicts I had nailed upon the doors...</p> <p>Language (+) / First-person (+) / Response time (s) 17.1/10.1</p> <p>Context faithfulness: Low (partial). The corpus does not contain «Пйотркув» and «Янош» in the relevant context... specific details about “прибиті едикти” are not confirmed.</p>
7	<p>Original Answer: Jako gospodarz dóbr nie gromadziłem skarbów dla samego siebie...</p> <p>Translation: As the steward of my estates, I did not amass treasures for myself...</p> <p>Language (+) / First-person (+) / Response time (s) 16.1/10.3</p> <p>Context faithfulness: Low. No links found for “Piotrków + two grosze”; financial details are not confirmed.</p>
8	<p>Original Answer: Ich bitte um Entschuldigung, doch in den mir vorliegenden Quellen findet sich keine Angabe darüber...</p> <p>Translation: I apologize, but in the sources available to me, there is no information...</p> <p>Language (+) / First-person (+) / Response time (s) 7.9/10.4</p> <p>Context faithfulness: Neutral/cautious. Correct refusal without assertions; no explicit evidence regarding “later historiography.”</p>

Linguistic-role behavior (see Table 2) was stable: compliance with the query language — 8/8 (100%), adherence to the first-person narrative — 8/8 (100%). The answers maintained the prescribed style and tone, contained no deviations from the “I-narrative,” and correctly mirrored the query language, which confirms the effectiveness of the system prompt and mode routing.

The context fidelity rate was 5/8 (62.5%) responses with full or predominant factual confirmation from

the corpus (Q1–Q5), 2/8 (25%) contained unconfirmed details (Q6–Q7), and 1/8 (12.5%) represented a correct cautious refusal (Q8). Confirmed markers included “Leliwa,” “Tarnów,” “Obertyn 1531,” “Khotyn 1538,” “Petru Rareș,” “16 May,” “Bernardines,” “bishops.” The lack of confirmation in Q6–Q7 was due to the absence in the corpus of specific mentions such as “Piotrków,” “Janus/János,” “two groszy per lan for veterans,” indicating a risk of localized hallucinations in the domains of law/finance/titlature.

In terms of temporal characteristics, the actual mean response time was 13.93 s versus the forecast of 9.89 s (average delta +4.04 s; median 14.25 s vs. 9.75 s; $\sigma = 3.76$ s). Recommendations for further piloting include maintaining low stochasticity (temperature ≤ 0.2), increasing top_k to 5 with a relevance threshold, limiting output length (≤ 4 sentences), if necessary separating different “personae” into distinct exhibits, and supplementing the corpus with verified sources on problematic topics.

Overall, the system demonstrates high compliance with the prescribed linguistic role and acceptable accuracy on core corpus facts (62.5% supported), but requires localized adjustments of prompts/corpus to mitigate hallucination risks in domain-specific responses.

5. Conclusions

This study developed and verified a knowledge engineering information technology based on Retrieval-Augmented Generation (RAG) for cultural and educational applications. The stated objectives were achieved: a reproducible and license-clean pipeline for corpus construction was built, semantic search with re-ranking was implemented, and grounded answer generation with explicit citation and safety mechanisms was ensured. Quantitative results from the pilot confirm the feasibility of the approach: 8 unique sources were identified, 7 successfully processed, and compliance with the linguistic-role model reached 100% (8/8 queries). Context fidelity was 62.5% (5/8 responses with full/predominant confirmation), with 12.5% correct refusal and 25% localized hallucinations in the domains of law and finance. The mean actual response time was 13.93 s. The obtained data demonstrate a balance of accuracy, explainability, and safety, achieved through the combination of corpus legal cleanliness, semantic search with re-ranking, and controlled generation. Source traceability increases user trust, while special policies for child audiences expand applicability. However, limitations were recorded regarding localized hallucinations and increased latency.

Future research should be directed towards expanding the licensed corpus with verified facts for sensitive domains, improving search via hierarchical/long-context RAG, optimizing latency through caching and multi-level indices, expanding the metric system to include RAGAS-style automated scoring, and scaling to multi-persona and multilingual scenarios.

Acknowledgments

This research is conducted within the framework of the Talking Heads: Conversations with Art from the Past project, supported under the Creative Europe Programme (CREA-CULT-2024-COOP). We gratefully acknowledge the project partners and participating cultural institutions for their collaboration and contributions.

Declaration on Generative AI

The authors used Generative AI tools solely for translation purposes and for assisting in the preparation of Figure 1. All other parts of the manuscript, including the research design, data analysis, results, and conclusions, are entirely the authors’ own work.

References

- [1] A. Ferrato, Large Language Models to enhance learning in Cultural Heritage, in: *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium, Blue Sky, and WideAIED*, Springer, 2025, pp. 458–463. doi:10.1007/978-3-031-99261-2_53.
- [2] G. Trichopoulos, K. Ordoumpozanis, G. Caridakis, An evaluation of LLM-based chatbots for enhancing the visitor's user experience at cultural exhibits, *J. Comput. Cult. Herit.* (2025). doi:10.1145/3775062, just Accepted.
- [3] I. Krak, O. Zalutska, M. Molchanova, O. Mazurets, R. Bahrii, O. Sobko, O. Barmak, Abusive speech detection method for Ukrainian language used recurrent neural network, in: *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Systems (CoLInS 2024)*, Volume III: Intelligent Systems Workshop, volume 3688, CEUR-WS.org, Aachen, 2024, pp. 16–28. URL: <https://ceur-ws.org/Vol-3688/paper2.pdf>.
- [4] H. P. Ho, V. Ramesh, I. Zaloudek, D. J. Rikhtehgar, S. Wang, Enhancing visitor engagement in interactive art exhibitions with visual-enhanced conversational agents, in: *Proceedings of the 30th International Conference on Intelligent User Interfaces*, Association for Computing Machinery, 2025, pp. 660–671. doi:10.1145/3708359.3712145.
- [5] J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, M. Gardner, Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus, in: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2021, pp. 1286–1305. doi:10.18653/v1/2021.emnlp-main.98.
- [6] N. Kandpal, B. Lester, C. Raffel, S. Majstorovic, S. Biderman, B. Abbasi, others, T. Murray, The Common Pile v0.1: An 8TB Dataset of Public Domain and Openly Licensed Text, *arXiv preprint arXiv:2506.05209* (2025). doi:10.48550/arXiv.2506.05209.
- [7] Google Research, deduplicate-text-datasets, 2025. URL: <https://github.com/google-research/deduplicate-text-datasets>.
- [8] H.-T. Nguyen, T.-D. Nguyen, V.-H. Nguyen, Enhancing Retrieval Augmented Generation with Hierarchical Text Segmentation Chunking, in: *Communications in Computer and Information Science*, Springer Nature Singapore, 2025, pp. 209–220. doi:10.1007/978-981-96-4288-5_17.
- [9] Z. Jiang, X. Ma, W. Chen, LongRAG: Enhancing Retrieval-Augmented Generation with Long-context LLMs, *arXiv preprint arXiv:2406.15319* (2024). doi:10.48550/arXiv.2406.15319.
- [10] P. Sarthi, S. Abdullah, A. Tuli, S. Khanna, A. Goldie, C. D. Manning, RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval, in: *Proceedings of the International Conference on Learning Representations*, 2024, pp. 1–23. doi:10.48550/arXiv.2401.18059.
- [11] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P. E. Mazaré, M. Lomeli, L. Hosseini, H. Jégou, The faiss library, *arXiv preprint arXiv:2401.08281* (2024). doi:10.48550/arXiv.2401.08281.
- [12] Z. Ye, D. Qi, H. Liu, S. Zhang, IHGR-RAG: An Enhanced Retrieval-Augmented Generation Framework for Accurate and Interpretable Power Equipment Condition Assessment, *Electronics* 14 (2025) 3284. doi:10.3390/electronics14163284.
- [13] J. Guo, Y. Fan, L. Pang, L. Yang, Q. Ai, H. Zamani, C. Wu, W. B. Croft, X. Cheng, A Deep Look into neural ranking models for information retrieval, *Information Processing & Management* 57 (2020) 102067. doi:10.1016/j.ipm.2019.102067.
- [14] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, H. Wang, Retrieval-augmented generation for large language models: A survey, *arXiv preprint arXiv:2312.10997* (2023). doi:10.48550/arXiv.2312.10997.
- [15] G. Izacard, P. Lewis, M. Lomeli, L. Hosseini, F. Petroni, T. Schick, J. Dwivedi-Yu, A. Joulin, S. Riedel, E. Grave, Atlas: Few-shot learning with retrieval augmented language models, *Journal of Machine Learning Research* 24 (2023) 1–43. doi:10.48550/arXiv.2208.03299.
- [16] M. T. Nayeem, D. Rafiei, KidLM: Advancing Language Models for Children—Early Insights and Future Directions, *arXiv preprint arXiv:2410.03884* (2024) 4813–4836. doi:10.48550/arXiv.2410.03884.

- [17] P. Rath, H. Shrawgi, P. Agrawal, S. Dandapat, LLM Safety for Children, in: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track), Association for Computational Linguistics, 2025, pp. 809–821. doi:10.18653/v1/2025.naacl-industry.62.
- [18] N. Kurian, ‘No, Alexa, no!’: designing child-safe AI and protecting children from the risks of the ‘empathy gap’ in large language models, *Learning, Media and Technology* (2024) 1–14. doi:10.1080/17439884.2024.2367052.
- [19] S. Es, J. James, L. Espinosa-Anke, S. Schockaert, RAGAS: Automated Evaluation of Retrieval Augmented Generation, *arXiv preprint arXiv:2309.15217* (2025). URL: <https://arxiv.org/abs/2309.15217>.
- [20] W. Gao, J. Lin, B. Mitra, ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation, in: *SIGIR 2024*, 2024, pp. 341–351. doi:10.1145/3626772.3657804.
- [21] H. Zhang, S. Diao, Y. Lin, Y. Fung, Q. Lian, X. Wang, Y. Chen, H. Ji, T. Zhang, R-Tuning: Instructing Large Language Models to Say ‘I Don’t Know’, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, 2024, pp. 7113–7139. doi:10.18653/v1/2024.naacl-long.394.
- [22] K. Lipianina-Honcharenko, T. Lendiuk, A. Sachenko, O. Osolinskyi, D. Zahorodnia, M. Komar, An intelligent method for forming the advertising content of higher education institutions based on semantic analysis, in: *International Conference on Information and Communication Technologies in Education, Research, and Industrial Applications*, Springer, 2021, pp. 169–182. doi:10.1007/978-3-031-14841-5_11.
- [23] K. Lipianina-Honcharenko, C. Wolff, A. Sachenko, I. Kit, D. Zahorodnia, Intelligent method for classifying the level of anthropogenic disasters, *Big Data and Cognitive Computing* 7 (2023) 157. doi:10.3390/bdcc7030157.