Vir-Curora: Prototyping a Bilingual Retrieval-Augmented Virtual Curator from Oral Interviews at a Cultural Heritage Site*

Phasit Ruangmak[1], Taito Yoshimura[2], Hirotaka Tsutsui[3], Eriko Ozaki[4], Pattara Leelaprute[5], Yuki Matsuda[6]

[1] *Faculty of Engineering*, *Kasetsart University*, Bangkok, Thailand, phasit.r@ku.th

[2] *Graduate School of Env., Life, Nat. Sci. and Tech.*, *Okayama University*, Okayama, Japan, taito.yoshimura@cocolab.jp

[3] *Okayama Prefectural Kurashiki Seiryo Senior High School*, Okayama, Japan

[4] *Shinsenryoku Co., Ltd.*, Okayama, Japan

[5] *Faculty of Engineering*, *Kasetsart University*, Bangkok, Thailand, pattara.l@ku.ac.th

[6] *Faculty of Env., Life, Nat. Sci. and Tech.*, *Okayama University*, Okayama, Japan, yukimat@okayama-u.ac.jp

***Abstract*—Cultural heritage sites preserve rich curatorial knowledge that is often transmitted through expert curators but remains difficult to reuse digitally, especially across languages. This paper presents Vir-Curora, a prototype bilingual virtual curator for Ohara House Katalyzer in Kurashiki, Japan. The system transforms Japanese oral interviews with on-site curators into structured knowledge units for retrieval-augmented generation. Each unit is linked to hierarchical spatial and object metadata, enabling retrieval to be scoped according to the visitor's current location or selected artifact. Vir-Curora combines a local vector store, bilingual embeddings, and a locally served language model to provide grounded responses in Japanese and English through a browser-based conversational interface. To reduce hallucination risk, the system applies source-grounded prompting and abstention behavior when retrieved context is insufficient. Preliminary automated and manual evaluations indicate that the prototype can provide faithful responses while revealing remaining challenges in knowledge-base coverage, response latency, citation presentation, and language consistency. This work demonstrates a practical pipeline for converting curator interviews into a context-aware conversational learning resource for cultural heritage sites.**

***Index Terms*—retrieval-augmented generation, cultural heritage, virtual curator, bilingual chatbot, knowledge extraction, oral interview, local language model**

I. Introduction

The Ohara House Katalyzer (語らい座大原本邸, Katarai-za Ohara Hontei) in Kurashiki, Okayama Prefecture, is a nationally designated Important Cultural Property and the former residence of the Ohara family, whose history spans more than 300 years. The estate preserves historical materials and site-specific knowledge related to the family, its architecture, and its cultural activities. Since opening as Ohara House Katalyzer in 2018, it has presented these materials and related research to the public through exhibitions and curatorial interpretation.

As part of this project, structured oral interviews were recorded between local high school students and curators at Ohara House Katalyzer, capturing diverse questions about the building's architecture, specific exhibits, and spatial history. These recordings represent a valuable but currently inaccessible asset: the data currently exists as raw, unstructured Japanese-language multimedia (audio, video, and photos), requiring significant effort to convert into a reusable, searchable form.

Fig. 1. On-site interview session at Ohara House Katalyzer. High school students interviewed curators about the site's architecture, exhibits, and spatial history to build the Vir-Curora knowledge base.

With the rapid advancement of Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) [1], an opportunity has emerged to transform these multimedia interview records into an interactive, bilingual learning resource for visitors.

Relying primarily on on-site curators to interpret Ohara House Katalyzer poses challenges for long-term and multilingual knowledge access. First, curatorial knowledge is held by a limited number of experts, and subtle site-specific knowledge may become difficult to reuse if it is not recorded and structured. Second, because the source interviews and most explanations are in Japanese, international visitors face a language barrier, while manual translation remains labor-intensive and may lose local nuance. Third, heritage knowledge is closely tied to physical locations, rooms, and artifacts,

This work was supported in part by the SCAT Research Grant and JSPS KAKENHI Grant Number JP24K20763. We also thank Kurashiki Seiryo High School Broadcasting Club, Ohara House Katalyzer for their invaluable cooperation.

requiring retrieval to be scoped by the visitor's current spatial or object context. Finally, open-domain LLMs may generate unsupported historical explanations; therefore, a virtual curator must be grounded in a verified knowledge base and abstain when retrieved evidence is insufficient [2].

To navigate the conservation constraints of the Ohara House Katalyzer, which as a National Important Cultural Property discourages invasive hardware installations or structural modifications, this project introduces Vir-Curora, a non-invasive, browser-based virtual curator application. The system transforms on-site recorded curator interviews into an interactive, bilingual learning assistant.

The technical architecture is built as a private, local RAG system integrated with a bilingual language model running entirely on a local PC via Ollama. Unstructured audio, video, and photographic assets are processed into structured knowledge units (KUs) for system integration. The backend manages this knowledge through a five-layer schema, enabling natural language queries in both English and Japanese to be dynamically mapped to physical locations, artifacts, and related visual assets.

II. Related Work

A. Retrieval-Augmented Generation and Evaluation

Retrieval-Augmented Generation (RAG), first proposed by Lewis *et al.* [1], addresses core limitations of parametric LLMs by grounding generated responses in dynamically retrieved external evidence. Gao *et al.* [2] survey the evolution of RAG architectures, including retrieval refinement, reranking, and modular retrieval pipelines. Asai *et al.* [3] further introduce Self-RAG, in which the model learns to retrieve, critique, and revise its responses based on retrieved evidence.

Evaluation is also critical for RAG-based systems. Saad-Falcon *et al.* [4] propose ARES, an automated framework for evaluating faithfulness, answer relevance, and context relevance using LLM-as-a-judge prompts. These evaluation perspectives are particularly relevant for cultural heritage applications, where generated explanations must remain faithful to verified source materials.

B. Robustness and Multilingual Retrieval

A key challenge in public-facing RAG systems is avoiding unsupported answers when the local corpus does not contain sufficient evidence. Thakur *et al.* [5] address this issue through NoMIRACL, which evaluates whether multilingual RAG systems can recognize unanswerable queries and abstain appropriately, building on the MIRACL multilingual retrieval benchmark [6].

The choice of embedding architecture determines cross-lingual retrieval precision. Feng *et al.* [7] introduce LaBSE (Language-agnostic BERT Sentence Embeddings), which maps 109 languages into a shared vector space. Wang *et al.* [8] propose the E5 embedding family, employing asymmetric query-passage prefix training to improve discriminative search precision. For specialized monolingual Japanese retrieval over historical terminology, Clavié [9] introduces JaColBERT, a Japanese-specific adaptation of the multi-vector ColBERT architecture, achieving an average Recall@10 of 0.813 over Japanese retrieval benchmarks.

C. Bilingual Generation and Spatial Retrieval

When source documents are Japanese but user queries are in English, standard architectures may suffer from citation drift, where output tokens do not align with retrieved source text; query translation is vulnerable to translation errors, while translating all retrieved documents before generation strips localized nuance.

Metadata pre-filtering constrains vector search to a subset of eligible documents before similarity ranking, guaranteeing that top-k results strictly match the visitor's spatial context rather than surfacing semantically similar but physically irrelevant passages. Wang *et al.* [10] introduce VirtuWander, a framework enhancing multi-modal virtual tour guidance using LLMs to generate situated, path-aware narratives, demonstrating the feasibility of spatially contextualized conversational museum guidance.

III. Proposed System

A. System Architecture

Vir-Curora is implemented as a four-layer system, illustrated in Fig. 2: a Vue.js 3 presentation layer, a FastAPI application layer, a Chroma retrieval layer with hierarchical metadata pre-filtering, and an Ollama-served model layer.

Presentation Layer: The frontend is a five-page Vue.js 3 single-page application with a persistent right-side chat panel on every page: *Landing* (hero carousel and zone preview), *Main/Explore* (six-zone grid: Soil Floor, Detached Room, Middle Warehouse, Inner Warehouse, Book Cafe, Main Garden), *Point of Interest* (zone detail with clickable artifact cards), *Artifact* (detail view with a metadata sidebar showing Knowledge Unit ID, source transcript, verification status, and a source citation block), and *About* (project background and credits). The chat panel is context-aware: it updates its RAG scope according to the page in view and surfaces NoMIRACL-style abstention messages for off-topic queries. The visual design uses a near-black ground with warm white text and a crimson accent, set in Shippori Mincho B1, Noto Sans JP, and DM Mono.

Application Layer: A FastAPI backend exposes navigation, search, and session-management endpoints that back the frontend's Place $\rightarrow$ POI $\rightarrow$ Artifact hierarchy, together with a core streaming endpoint (`POST /api/query/stream`) that accepts a question with optional place/POI/artifact/session identifiers and streams tokens back via `StreamingResponse` as they are generated by the RAG pipeline. The vector database is loaded once at server startup to keep query latency low, and blocking calls are dispatched to a worker thread pool to avoid stalling the async event loop.

Retrieval Layer: Retrieval uses a local Chroma vector store (collection `ohara_corpus`). Each knowledge unit's `location_ids` and `object_ids` are expanded into single-valued metadata fields at indexing time, enabling native

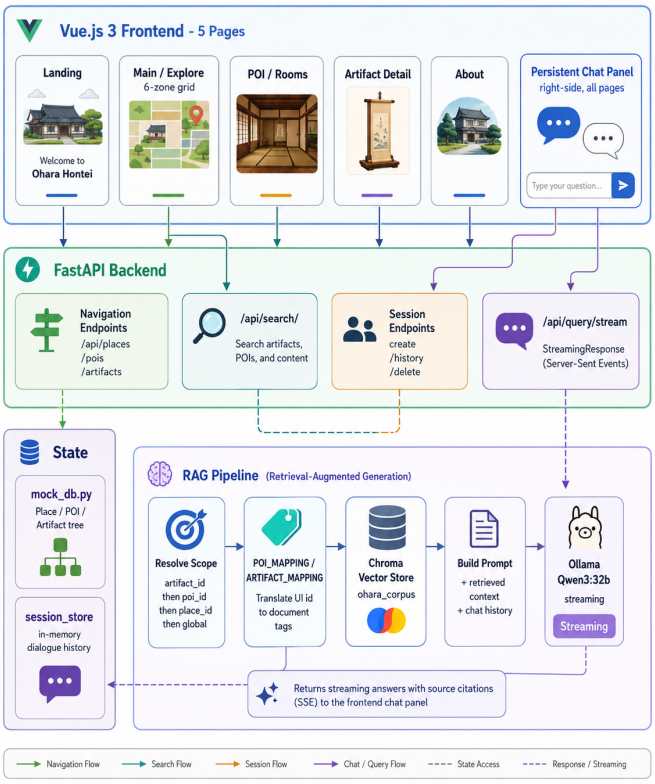

Fig. 2. System architecture and RAG pipeline flow of Vir-Curora. The five-page Vue.js 3 frontend, the FastAPI application layer, the Chroma retrieval layer with hierarchical metadata pre-filtering, and the Ollama (Qwen3:32b) inference layer are connected in a single query-time flow: the active UI context resolves a hierarchical metadata filter (artifact → POI → place → global), Chroma retrieves the top-k matching passages under that filter, the prompt is assembled with optional session history, and Ollama streams a grounded, source-cited response with an instructed abstention fallback.

Chroma metadata filtering. At query time, the active UI context resolves a metadata filter following a strict precedence: an `artifact_id`, if present, restricts retrieval to that artifact's `object_id`; otherwise a `poi_id` restricts to that POI's `location_id`; otherwise a `place_id` restricts to that place; if none are supplied, retrieval runs globally. Two translation tables expand a single UI-level identifier into the set of underlying document tags it should match, reconciling cases where UI-level zones and document-level tags do not correspond one-to-one.

Model and Inference Layer: The local Ollama engine hosts the Qwen3:32b dense model (temperature $= 0.5$, top_p $= 0.9$). Bilingual document embeddings are generated using BAAI/bge-m3. Generation is streamed token-by-token from Ollama through the FastAPI backend to the frontend chat panel.

B. Knowledge Unit Schema

The structural core of the database is the Knowledge Unit (KU), a self-contained JSON object parsed from verified curator transcripts. Each KU maps Japanese oral interview segments to aligned bilingual question/answer/summary text and hierarchical identifiers. An example KU is shown below:

```
{
  "unit_id": "ku_0001",
  "unit_type": "qa_pair",
  "location_ids": ["entrance", "doma_1"],
  "object_ids": ["roof_tile",
    "ohara_family_crest"],
  "question_ja": "この灰色の瓦のようなものは何ですか。",
  "question_en": "What is this gray
    tile-like object?",
  "answer_en": "It is a roof tile
    bearing the Ohara family crest,
    Mukai-matsu...",
  "confidence": "high",
  "verified_by_japanese_speaker": true
}
```

At ingestion, each KU's `location_ids` and `object_ids` lists are expanded into one Chroma document per (location, object) pair. The broader five-layer data schema comprises: (1) `raw_media_index`; (2) `initial_transcript`, generated via Gemini and ChatGPT Business; (3) `reviewed_transcript`, human-corrected; (4) `knowledge_units`, structured RAG-ready units as shown above; and (5) `visual_context`, photo and object metadata.

C. Abstention Policy

To reduce hallucination risk, Vir-Curora implements an abstention mechanism at the prompt level: the LLM's system instructions explicitly direct it to answer only from retrieved context, respond "*I don't know*" when the context does not support an answer, avoid inventing details, and cite source metadata for traceability. A quantitative extension of this mechanism is planned, bypassing generation entirely when the maximum retrieved cosine similarity falls below a threshold $s < 0.3$, consistent with the NoMIRACL evaluation paradigm [5]:

$$\text{respond\_with\_fallback} \iff \max_i \cos(q, d_i) < 0.3 \qquad (1)$$

where q is the query embedding vector and d_i are the document embedding vectors returned by the vector store. Implementing this quantitative threshold alongside the existing prompt-level instruction is identified as near-term future work (Section V).

IV. Implementation

A. Data Collection and Transcription

The dataset consists of 11 audio/video interview files recorded at the Ohara House Katalyzer on June 5, 2026, totaling approximately 180 minutes of processed audio after format conversion from MP4/MOV to M4A (99.6% size reduction). The interviews feature high school students querying curators about the building's architecture, exhibits, and spatial history in Japanese.

Initial transcription was performed using Gemini and ChatGPT Business, with long files segmented into 10–15 minute chunks. Raw transcripts were then corrected by a Japanese

graduate student to resolve proper nouns, Meiji-era historical events, regional architectural terms (e.g., *Kurashiki-mado*, *Namako-kabe*), and Ohara-family-specific concepts (e.g., *Nisan no Oshie*). The verified text was parsed into QA-structured knowledge units with bilingual summaries, keyword lists, and coordinate tags mapping each unit to physical locations within the estate.

B. RAG Pipeline

The RAG pipeline core is shared between the FastAPI backend's live streaming endpoint (Section III-A) and a standalone CLI tool used for offline prompt and retrieval tuning. A loader reads knowledge units from JSONL sources and expands each unit's `location_ids`/`object_ids` into single-valued `place_id`/`location_id`/`object_id` metadata, alongside source fields for citation tracing. Documents are embedded with BAAI/bge-m3 and indexed into a local Chroma collection; the database handle is initialized once at FastAPI startup, loading a persisted index directly if one exists.

At query time, the backend resolves a Chroma metadata filter from the active UI context using the artifact → POI → place → global precedence described in Section III-A, with two translation tables reconciling UI-level zone identifiers against document-level tags. The retriever returns the top-k matching passages ($k = 2$ in the present configuration), which are formatted into numbered context chunks and combined with a fixed bilingual system instruction directing the model to answer only from context, respond "*I don't know*" when insufficient, and cite source metadata. When an active session is supplied, prior dialogue turns are interpolated into the prompt as a history block for multi-turn awareness. The composed prompt is sent to Ollama (Qwen3:32b, temperature $= 0.5$, top_p $= 0.9$) via its streaming API, with retrieval and generation dispatched to a worker thread pool to avoid stalling the event loop. On completion, the response is appended to session history. A complementary CLI pipeline logs every standalone tuning run with a unique test ID, prompt, response, and cited sources, supporting offline prompt-tuning review independent of the live API.

C. Frontend Interface

A five-page Vue.js 3 single-page application has been completed as a working interactive mockup, consuming the FastAPI backend described in Section III-A. Across all five pages, a persistent right-side chat panel issues queries to `POST /api/query/stream` carrying the page's active `place_id`/`poi_id`/`artifact_id` context, displays streamed tokens as they arrive, renders source citations in the `ku_000x / audio_xxx_reviewed` format, and surfaces a NoMIRACL-style decline message for off-topic queries.

D. Automated Hyperparameter Evaluation

System verification relies on two metrics from the RAG Triad [4]: *Answer Relevance* (Rel), measuring whether the generated text directly addresses the query, and *Faithfulness* (Faith), measuring factual consistency against the retrieved knowledge units. A combined score (Comb) is computed as their equal-weight average. An automated evaluation sweep was conducted across four temperature values (0.2, 0.4, 0.6, 0.8) and three nucleus sampling values (top_p $\in \{0.7, 0.8, 0.9\}$), yielding 12 configurations, all running Qwen3:32b with $k = 2$ retrieved passages, using English-language evaluation prompts. All 108 individual runs (12 configurations $\times$ 3 categories $\times$ 3 iterations) completed successfully at the pipeline level, though automated-judge scoring failed to parse a small number of iterations, reducing the valid sample size for the affected rows in Table I (marked *). Average latencies for the configurations shown in Table I ranged from 34.3s to 60.1s. Table I reports the top results per category.

TABLE I
TOP HYPERPARAMETER CONFIGURATIONS PER TOPIC CATEGORY (3 RUNS EACH UNLESS NOTED)

Category	temp	top_p	Rel	Faith	Comb	Latency (s)
book_cafe	0.4	0.7	0.70	1.00	0.85	48.2
book_cafe	0.2	0.7	0.70	1.00	0.85	60.1
book_cafe	0.2	0.8	0.60	1.00	0.80	47.5
garden	0.2	0.8	1.00	1.00	**1.00**	47.3
garden	0.4	0.9	1.00	1.00	**1.00**	43.6
garden	0.6	0.7	1.00	1.00	**1.00**	45.8
garden	0.4	0.8	0.90	0.90	0.90	46.3
main_hall	0.6	0.7	0.85	1.00	**0.93**	36.0*
main_hall	0.8	0.7	0.85	1.00	**0.93**	37.6*
main_hall	0.6	0.9	0.80	1.00	0.90	34.3
main_hall	0.2	0.9	0.13	1.00	0.57	36.2
main_hall	0.4	0.9	0.20	1.00	0.60	36.9*

*Fewer than 3 valid scored runs; automated judge failed to parse output for the remaining iteration(s).

Faithfulness remained high across the tested configurations, and no major unsupported historical explanations were observed in this evaluation set. Answer Relevance therefore served as the main differentiating metric. The *garden* category showed stable performance in this evaluation: several configurations achieved Combined = 1.00 and retrieved the same relevant source rows, citing the planting palette and the Kurama stone transported from Kyoto. The *main_hall* category revealed a critical failure zone: top_p = 0.9 combined with low temperature (0.2 or 0.4) caused relevance to collapse to 0.13–0.20. One possible explanation is that high top_p under low-temperature settings allowed responses to drift off-topic while still citing valid sources. In these cases, the responses remained grounded in retrieved context but did not fully address the question asked. The *book_cafe* category did not exceed Relevance = 0.70 across the tested configurations: the evaluation prompt asked how the space is *arranged for visitors*, but the retrieved knowledge units only describe collection size and Ohara Soichiro's reading interests. In the tested responses, the system often abstained on the uncovered portion, suggesting that the prompt-level abstention instruction can be effective under partial-context conditions. The limited relevance score appears to be mainly due to a knowledge-base coverage gap rather than only to generation parameters.

Table II compares a shortlisted set of configurations across

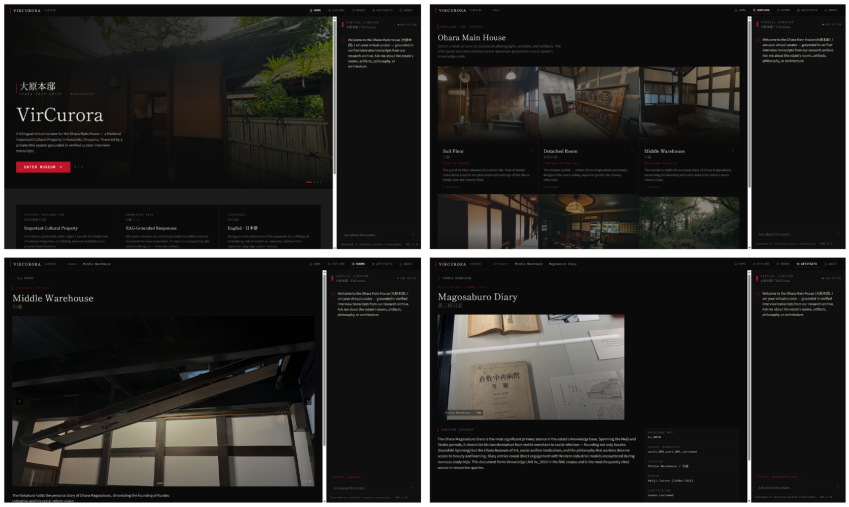

Fig. 3. Vir-Curora web application interface. Left: the Main (Explore) page presenting the six-zone estate grid. Right: the persistent context-aware chat panel showing a bilingual curator response with source citations. The interface uses a near-black ground (`#0a0a0a`), warm white body text (`#e8e2da`), and crimson accent (`#c01428`), set in Shippori Mincho B1, Noto Sans JP, and DM Mono.

TABLE II
CROSS-CATEGORY COMBINED SCORES FOR CANDIDATE GLOBAL CONFIGURATIONS

temp	top_p	book_cafe	garden	main_hall	Avg
0.4	0.7	0.85	0.95	0.85	**0.88**
0.6	0.7	0.70	1.00	0.93	0.88
0.2	0.8	0.80	1.00	0.85	0.88
0.2	0.9	0.75	0.95	0.57	0.76
0.4	0.9	0.70	1.00	0.60	0.77

TABLE III
MANUAL EVALUATION: MEAN SCORES BY CATEGORY (1–5 SCALE, JAPANESE-LANGUAGE PROMPTS, 3 EVALUATORS)

Category	Factual	Tone	Citation
Garden	4.7	4.3	4.7
Main Hall	5.0*	3.7	2.4
Book Cafe	4.9	4.5	4.3

*Inflated by abstentions: several "I don't know" responses were marked Factual=5 with a qualifier.

all three categories to identify a single global setting. Three configurations tied for the highest average combined score of 0.88: temp = 0.4/top_p = 0.7, temp = 0.6/top_p = 0.7, and temp = 0.2/top_p = 0.8. Among this tied set, temperature = 0.4 and top_p = 0.7 were selected as a practical default configuration for subsequent evaluation and deployment, rather than as a uniquely highest-scoring option. This choice reflects the best overall trade-off among the tied candidates: it achieves the highest book_cafe combined score (0.85) of the three, maintains strong garden performance (0.95), and avoids the collapse in main_hall relevance observed under low-temperature, high-top_p settings.

E. Manual Evaluation

While the automated ARES sweep (Section IV-D) was conducted entirely in English, the manual evaluation described in this section was conducted with *Japanese-language* prompts, providing an independent check on the pipeline's bilingual behavior that the English-only automated sweep cannot surface. Three bilingual evaluators independently reviewed the same 15-question set spanning the three topic categories (5 questions per category) submitted in Japanese to the deployed pipeline. Each response was scored on a 1–5 scale across three dimensions: *Factual Correctness*, *Tone, Clarity & Formatting*, and *Citation Accuracy*, and evaluators recorded notes on repetitive loops, cross-zone context bleeding, and language mixing.

Table III reports the mean score per category and dimension, averaged across the three evaluators. Averaging the three manual dimensions per category yields a partially different picture from the automated sweep: Garden and Book Cafe scored similarly (mean 4.57 across Factual, Tone, and Citation for both categories), while Main Hall scored lowest (mean 3.70), driven primarily by weak Citation Accuracy (2.4/5) rather than Factual Correctness. This suggests the automated and manual evaluations agree that Main Hall is the weakest category, but disagree on whether Garden or Book Cafe performs best – a discrepancy plausibly explained by the different rubric and language used in each evaluation, and by the Main Hall Factual score being inflated by abstentions, as noted above.

Language mixing. Evaluators repeatedly flagged Japanese prompts returned substantially or entirely in English, most heavily in main_hall. This bilingual-generation issue was not visible in the English-only automated evaluation and should be prioritized in future improvements.

Abstention overlaps with automated faithfulness. Several main_hall responses were "*I don't know*" abstentions; evaluators marked these with a qualifying question mark rather than a flat 5, since a correct abstention is not a verified correct answer. This is consistent with the high automated Faithfulness scores:

in these cases, the model tended to avoid unsupported answers, while the main hall knowledge units did not sufficiently cover the questions asked in either language.

Citation format ambiguity. Citation Accuracy was weakest for main_hall (2.4/5) and book_cafe (4.3/5), well below garden (4.7/5). This was driven less by incorrect citations than by unclear presentation (e.g., "Chunk 1" labels with no visible mapping to the `ku_000x` format used elsewhere), so the citation *format*, not its correctness, needs improvement.

This manual pass should be read as a small-scale pilot rather than a statistically powered study; a larger bilingual evaluation, and resolution of the language-mixing failure mode, are identified as future work (Section V).

V. Conclusion

This paper presented Vir-Curora, a bilingual virtual curator system for the Ohara House Katalyzer. The implemented system comprises a five-page Vue.js 3 frontend with a persistent context-aware chat panel, a FastAPI backend exposing hierarchical navigation, unified search, and a stateful streaming RAG chat endpoint, and a retrieval core combining a local Chroma vector store, BAAI/bge-m3 bilingual embeddings, and a quantized Qwen3:32b model served via Ollama. Hierarchical metadata pre-filtering, resolving artifact, POI, place, and global retrieval scopes through UI-to-document translation tables, directly addresses the spatial context sensitivity challenge identified in Section I, while session-aware multi-turn memory and an instructed abstention policy address conversational continuity and hallucination risk respectively.

An automated hyperparameter sweep across 12 temperature and nucleus-sampling configurations showed high faithfulness across the tested settings, suggesting that the source-grounded prompting and abstention instruction helped reduce unsupported answers in this evaluation set. Answer relevance varied by category, and temperature = 0.4 with top_p = 0.7 provided the best overall trade-off among the tested configurations, with an average combined score of 0.88. The results suggest that a data coverage gap in the book_cafe knowledge units was a major factor limiting relevance in that zone. A subsequent manual pilot evaluation, conducted in Japanese rather than the English used for the automated sweep, corroborated Main Hall as the weakest category and additionally surfaced a language-mixing failure mode and a citation-presentation ambiguity that the English-only automated evaluation could not detect.

The system's browser-based architecture requires no physical installation or hard-wiring within the estate, complying fully with the conservation requirements of a designated National Important Cultural Property.

Near-term future work includes implementing the quantitative cosine-similarity abstention threshold described in Section III-C, migrating the mock dataset and in-memory session store to persistent relational storage, resolving the language-mixing and citation-presentation issues surfaced by the manual evaluation, and extending that evaluation to a larger, statistically powered bilingual rater pool alongside a RAGAS-based automated Japanese evaluation. Latency is also a concern: the reported response times were measured against the current small knowledge base and will need re-profiling as it scales. Longer-term work will investigate graph-based retrieval for relational historical queries, VLM integration for phtograph-grounded explaination, voice-based interaction via automatic speech recognition to support hands-free, non-invasive queries within the gallery setting, and VUI persona alignment with an on-site curator's conversational register. Further latency reduction is also planned, including model quantization, prompt caching, and hybrid retrieval strategies such as approximate nearest-neighbor indexing, which will become increasingly necessary as the knowledge base is expanded beyond the current small corpus.

Acknowledgment

This work was supported in part by the SCAT Research Grant and JSPS KAKENHI Grant Number JP24K20763. The authors gratefully acknowledge the staff and researchers at the Ohara House Katalyzer for their participation in the recorded interviews, and Broadcasting Club at Okayama Prefectural Kurashiki Seiryo Senior High School for their invaluable cooperation in the data collection process.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 9459–9474.

[2] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, "Retrieval-Augmented Generation for Large Language Models: A Survey," 2023, available: arXiv:2312.10997.

[3] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection," in *International Conference on Learning Representations*, vol. 2024, 2024, pp. 9112–9141.

[4] J. Saad-Falcon, O. Khattab, C. Potts, and M. Zaharia, "ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2024, pp. 338–354.

[5] N. Thakur, L. Bonifacio, C. Zhang, O. Ogundepo, E. Kamalloo, D. Alfonso-Hermelo, X. Li, Q. Liu, B. Chen, M. Rezagholizadeh, and J. Lin, ""Knowing When You Don't Know": A Multilingual Relevance Assessment Dataset for Robust Retrieval-Augmented Generation," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 12 508–12 526.

[6] X. Zhang, N. Thakur, O. Ogundepo, E. Kamalloo, D. Alfonso-Hermelo, X. Li, Q. Liu, M. Rezagholizadeh, and J. Lin, "MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1114–1131, 2023.

[7] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-Agnostic BERT Sentence Embedding," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 878–891.

[8] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, and F. Wei, "Text Embeddings by Weakly-Supervised Contrastive Pre-training," 2022, available: arXiv:2212.03533.

[9] B. Clavié, "Towards Better Monolingual Japanese Retrievers with Multi-Vector Models," 2023, available: arXiv:2312.16144.

[10] Z. Wang, L.-P. Yuan, L. Wang, B. Jiang, and W. Zeng, "VirtuWander: Enhancing Multi-modal Interaction for Virtual Tour Guidance through Large Language Models," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, ser. CHI '24, 2024. [Online]. Available: https://doi.org/10.1145/3613904.3642235