Towards a Curatorial Agent for Heritage Institutions: Web Source Credibility Verification for Grounding Domain-Specific LLMs

Aleksandra Pshenova[1], Jaehong Ahn[1]

[1] Korea Advanced Institute of Science and Technology - Daejeon, South Korea – spshenova@kaist.ac.kr, ahnjh@kaist.ac.kr

Keywords: Hallucinations, BERT, BART, URL Classification, Source Credibility, Question Generation

Abstract

Hallucination problem is the main cause of the weak reliability of Large Language Models (LLMs) for their use in cultural institutions, such as museums and galleries. One proposed solution to the hallucination problem is to ground the LLM in the real data found on the Web. However, since the cultural heritage domain requires factual accuracy, cultural institutions cannot fully rely on the data obtained from the Web. To make the data suitable for the heritage domain use case, additional source filtering and verification must be applied. In this paper, we propose a potential source verification pipeline for verifying web sources, as well as a question-generating agent designed to guide heritage experts in collecting the right sources for their needs. Upon evaluation, the proposed system successfully filters the web-scraped sources given a search keyword, achieving moderate results in both classification tasks. In addition, our contributions include the curation of a custom dataset for training both models and estimation of an optimal training & dataset configuration for the proposed ’curatorial question generation’ task.

1. Introduction

Hallucination problem is the main cause of the weak reliability of Large Language Models (LLMs) for their use in cultural institutions, such as museums and galleries. It can lead to distortion and misinterpretation of cultural values, meaning, and historical context, as a result perpetuating inaccurate narratives within society (Bu et al., 2025).

In order to avoid the usage of false data, institutions are actively using local AI agents in their operations and rely on an internal database. While reliable, this approach limits the information to which these institutions have access and discourages inter-institutional cooperation and data exchange. The Web might seem like the best means of data exchange and cooperation between institutions around the world. Meanwhile, a large amount of unverified sources and misinterpreted information prevents institutions from using the Web.

To make data retrieval from the Web verifiable, there is a demand for an algorithm for scraping the Web for relevant sources and classifying them as credible or not. This approach is well researched, and various Machine Learning (ML) techniques have been used to detect malicious or phishing Uniform Resource Locators (URLs)(Aljabri et al., 2022). With a change in classification criteria, it is possible to apply similar techniques to classify web sources as credible or not credible, allowing them to be used in the highly factual culture heritage domain.

In this paper, we address the hallucination issue and propose a framework for web source verification, adapting a malicious URL classification model to the cultural institutions’ use case. This approach allows the user to not only leverage the local database to gather information but also make use of the information from the web sources validated by the framework. The framework is composed of two modules: VeriBERT, a BERT-based heritage domain-specific source verifier and ExPlanBART, a BART-based curatorial brainstorming agent, helping the user find the right sources for exhibition planning application.

We aim for this framework to become the first component of the potential curatorial agent model. We define the curatorial agent as a smart agent grounded in information retrieved from trusted sources that is able to guide both heritage institution visitors and experts in the heritage domain, and serve a curatorial or an assistive role depending on the target user.

Our main contributions are as follows. First, our approach attempts to adapt a malicious URL classification model to create a filtering framework for the verification of web sources for use in cultural institutions. To prepare the source classification model for downstream application, we fine-tune a brainstorming agent and build a pipeline that makes use of the collected URLs. In this way, we are moving towards a full-cycle curatorial agent for heritage institutions, one that is not just grounded in real-life knowledge, but considers only the sources verified in accordance with the domain standards.

Second, in order to train our custom domain-specific web source validation system, we curated a labeled dataset of web-scraped culture heritage pages using a fine-grained labeling pipeline using curated criteria. For the curatorial question generation agent, we preprocessed a domain-specific Q&A dataset in order to tailor it for further applications in training chatbots and domain-specific dialog systems, and experimented with several training setups to propose the best configuration so far for this particular task.

2. Related Work

2.1 Source Credibility Verification

By source credibility, most researchers mean ’source believability’, that is, a perceived quality made up of multiple dimensions, such as expertise and trustworthiness (Fogg and Tseng, 1999). Thus, it is a measure of how accurate the source presented is to the known-credible material. Source credibility has been a prominent area of research, but the concept and format of ’source’ changed with time in accordance with the emergence

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-M-9-2025-1213-2025 | © Author(s) 2025. CC BY 4.0 License.

and development of new technology. This way, in one of the first works on the topic, (Hovland and Weiss, 1951), the authors discussed the general influence of credibility from a psychological point of view, using a trustworthy person as a source.

The authors of (Fogg and Tseng, 1999) discuss a then-emerging field of 'computer credibility' and present two perspectives on what computer users evaluate when assessing credibility, namely, a systems perspective, emerged with the development of technology, and a psychological perspective, reflecting the subjective nature of credibility. In the broader culture domain, many works cover the verification of sources in media. In particular, (Kiousis, 2001), (Sundar and Nass, 2001), (Choi et al., 2006), (Cassidy, 2007) address the issue of source verification and credibility of media outlets in the digital era, highlighting different domain-specific aspects of credibility and its impact on the media consumer.

When it comes to source credibility and verification in the heritage domain, research is much more scarce. Several works, such as (Duff et al., 2004) and (Amin et al., 2008) report the importance of being able to analyze the credibility of the source to the domain experts in the digital age, before using the sources in their operations. In (Amin et al., 2008), the authors conducted a user study to identify the most important information seeking needs of the heritage domain experts. It has been concluded that the majority of use cases can be classified as information gathering tasks related to activities such as exhibition or publication preparation and collection documentation management. In particular, a major difficulty the experts face is inevitable manual collection, examination and alignment of relevant pieces of information, which is not currently supported by their tools. These tasks are exactly what we are trying to address and automate in our paper.

Finally, in (Amin et al., 2009) the authors conducted an empirical user study to assess the effect of source credibility ratings on the scope of the cultural heritage information aggregator. It has been concluded that, while it does give users greater confidence in the information they select, there still remains the need to define the exact credibility criteria unique for each domain, like culture heritage.

Many research works on credibility are rooted in social science, which is reflected in the formulation of their problem statement and their approach to quantitative and qualitative evaluation of proposed methods. In this paper, we try to find a more technical approach to assessing credibility.

2.2 URL Classification

URL classification task usually deals with binary or multiclass classification for filtering of trustworthy links or grouping them2 with defined group labels. When it comes to URL classification, the applications of the task are numerous, for example, (Rajalakshmi et al., 2024) introduces an attention-based classifier for adult websites. However, data security remains one of the most active domains, and numerous studies have been conducted to apply URL classification to this domain. The basic methodology for malicious URL detection involves extracting features from each link and analyzing them for malicious patterns. In addition to direct URL feature extraction, malicious URL detection methods include HTML-based and JavaScript-based methods, which involve analysis of page content alongside the URL (Tian et al., 2025).

Multiple frameworks for malicious URL detection have been proposed. Transformer-based methods, such as BERT, are widely used to extract semantic features from URLs and analyze them for classification. (Su and Su, 2023) achieved particularly exceptional results in applying BERT to malicious URL detection, reporting around 98.5% average accuracy across three public URL datasets in binary classification and 99.78% in multi-label classification. In recent years, there has been a significant rise in works exploring the potential of BERT in URL classification. (Li et al., 2024) introduces URLBERT, a pre-trained model for malicious URL detection suitable for fine-tuning for downstream task adaptation, followed by M-BERT (Yu et al., 2024), and DomURLs_BERT (Mahdaouy et al., 2024), which is the current state-of-the-art model. Its flexibility for downstream adaptation made it a suitable backbone candidate for our domain-specific task adaptation through transfer learning. In this paper, we attempt to apply similar principles of the data security domain while keeping the system tailored for culture heritage in-domain use.

2.3 Question generation

Question Generation (QG) is the task of automatically generating questions based on a given input, often formulated as an answer to a previous question (conversational question generation) or a piece of information to ask questions about (clarifying question generation). QG is an application of LLMs that has always been addressed, but has started gaining a lot of traction in recent years with the advancement of NLP research. Although general QG is a topic of active research, its domain-specific applications remain much less addressed, especially in the heritage domain. This is due to the extreme narrow application of this task in areas like exhibition narrative brainstorming.

Overall, numerous efforts have been made in the field of QG in recent years. (Zhou et al., 2018) introduces the Neural Question Generation (NQG) framework to generate answer-specific questions as a preliminary study. Following this study, numerous studies experimented with model architectures and training strategies for the QG task. Among those, sequence-to-sequence (seq2seq) models such as BART are actively explored for their multitask flexibility and adaptation potential.

(Park et al., 2022) applied post-training to KoBART, a version of BART pre-trained on Korean language corpora, for Korean question generation. The usage of BART as a seq2seq backbone for QG tasks became a prominent trend in recent years. (Majumder et al., 2021) introduced a novel BART-based approach to generate clarification questions based on missing information in the given text. The problem formulation in the paper is similar to our current task of formulating clarification questions for information gathering from the user and goal-oriented dialog initialization. Notably, the model was trained on 'missing information – question (about this information)' pairs.

In this paper, our objective is to perform a preliminary exploration study on a QG agent to obtain specific information from the user. Due to the structural differences of these tasks (missing content-based clarification question generation and suggestive question generation based on previously given answers), creating a custom dataset and training pipeline for this task is necessary.

3. Method

In this paper, we propose two agents aimed at aiding the heritage domain experts in filtering and utilizing the web sources, VeriBERT and ExPlanBART. Both models can be used independently as they share the same local database for writing and retrieval of search results.

While the second agent is a straightforward question generator finetuned for a specific domain, VeriBERT (Fig. 1) addresses a less trivial task of web source validation. It is composed of two BERT-based classifiers, first, the Domain Verifier, takes 10 web search results as input and classifies if the domain is trustworthy. Filtered domains are passed down to the deep scraper for internal page scraping, where each of the 10 domains is recursively scraped for pages that contain the initial search keyword. Raw internal-scraped pages are passed into the Keyword Relevance Classifier that assesses if the content of the given page is truly relevant to the search keyword, collecting the classified relevant pages into a final output list. We curate a single complex labeled dataset to train both VeriBERT models. For ExPlanBART, we modify an already existing heritage domain Q&A dataset.

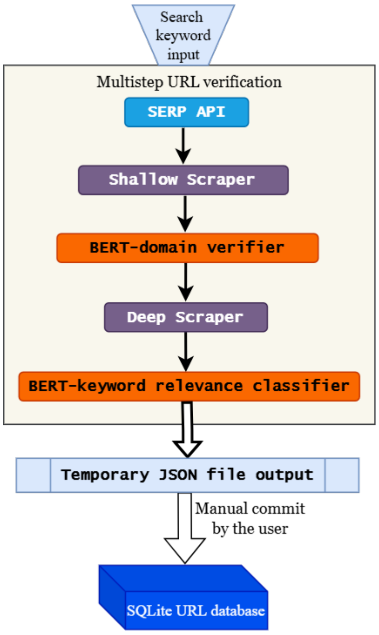
Search keyword input
Multistep URL verification
SERP API
Shallow Scraper
BERT-domain verifier
Deep Scraper
BERT-keyword relevance classifier
Temporary JSON file output
Manual commit by the user
SQLite URL database

Figure 1. General proposed pipeline of VeriBERT

3.1 Data Collection

3.1.1 VeriBERT To collect data for training VeriBERT, we use a curated list of search keywords as initial input and perform an automated Web search via the SERP API to obtain a list of candidate URLs for labeling. We end up with a list of roughly 34,125 web pages paired with their initial search keyword. We then run the resulting raw URL dataset through a multi-filter labeling pipeline to extract a content snippet from each page and label each URL as valid or invalid based on a set of curated criteria, as well as calculate keyword relevance score for each page.

Inspired by (Pattanaphanchai et al., 2013), we use the outbound link count and cosine similarity to anchor institution pages as the main validation criteria for each page. We collect a list of known valid institution web pages and embed them to get a list of anchor embeddings. We then compare each new URL snippet embedding to this list to classify whether the page is structured semantically similar to the web pages of trusted heritage institutions. We also count the outbound links of each page to estimate how many valid sources does the page cite and refer to. If the page passes both checks, it gets a ’valid’ label, and ’invalid’ if it contains invalid content or fails the checks.

After initial labeling, we perform data cleaning and remove entries with non-English content, invalid or absent snippets, duplicates, and truncate the snippets to bring all entries to uniform length. After cleaning, the total number of entries in the dataset is 24,097.

To collect a unique test set for evaluating both VeriBERT models, we perform the same steps described above with a unique list of search keywords, resulting in a dataset of 3,919 samples for final model evaluation.

3.1.2 ExPlanBART We use the VISCOUNTH dataset (Becattini et al., 2023) as our prime source of data for curating a custom question generation dataset for ExPlanBART. VISCOUNTH is a large-scale dataset consisting of 6.5M question-answer pairs of 43 question types referring to different aspects of Italian cultural assets. For pre-processing, we select only 7 relevant question types, resulting in roughly 770k Q&A pairs before pre-processing.

During the first pre-processing step, we remove trailing numbers and junk symbols, remove obscure entries, and reverse the pairs to turn answer into an input and the question into a generation target, resulting in roughly 262k valid entries. We also add a prompt guidance to BART in this step, marking each input entry as ’Exhibition planner input:[input]’. In order to support a variety of possible input styles, we diversify the dataset by sampling long answers, short answers, and single-word inputs from the original dataset with the 50:30:20 ratio.

The next pre-processing step involves running the entries through a FLAN-T5-large model in order to turn neutral content-based questions into rich curatorial suggestions and clarifying questions. We utilize prompt engineering and provide examples to carefully guide FLAN-T5 to generate appropriate questions for each question type. We then clean the resulting augmented dataset to remove the misgenerated inputs that did not end in a question. The final output of data preparation is a dataset of 261,932 augmented ’input answer’-’target question’ pairs.

3.2 Model Training

For VeriBERT, we use DomURLs_BERT (Mahdaouy et al., 2024) as the backbone for the domain verification classifier and BERT-base (Devlin et al., 2018) for the keyword relevance classifier. The choice of base models for fine-tuning is dictated by the inputs and data processing requirements. The domain classifier takes the URL of the web page and the scraped content snippet as input. It is thus beneficial to use DomURLs_BERT as it already learned the basic URL features during pre-training and can serve as an additional filter for potential malicious URLs picked up by automatic scraper.

The keyword relevance classifier addresses the more intricate task of classifying whether the keyword appears in a context semantically relevant to the keyword. We use BERT-base as the backbone, since the task requires semantic understanding of the page content.

During ExPlanBART training, we experiment with different dataset setups to determine the dataset structure that yields the best quantitative and qualitative results of the trained model. We use BART-base (Lewis et al., 2019) as a backbone and train three models: first one on a 'baseline' dataset with both columns taken from the original dataset unchanged, second on the dataset with FLAN-T5-augmented outputs and unmodified inputs, and third one on a dataset with both inputs and outputs modified: inputs with an added 'Exhibition planner input: [input]' guidance prefix, and outputs modified by FLAN-T5. After this round of training and comparison, we can make a strong assumption about the best training strategy and dataset composition for this specific task.

3.3 Pipeline Assembly

As presented in Fig. 1, the final output of VeriBERT is a temporary JSON file containing a list of filtered URLs, each entry paired with a short snippet preview and an initial search keyword. The user can make modifications to the list by manually removing certain entries or correcting misspelled keywords. After the user is satisfied with the resulting list, they can commit the list to the main database, and the content of the file is automatically appended to the main SQLite URL database in the format of [URL] : [search keyword]. The content snippet is dropped in the final output.

After aggregating enough URLs through VeriBERT, the user can utilize the collected URLs through the proposed URL selection pipeline. Fig. 2 illustrates the entire downstream pipeline of the filtering system. Since VeriBERT can be used independently from the other parts of the pipeline, the total URL count in the URL database can reach thousands of entries. It is thus necessary to implement a logic to automatically retrieve URLs relevant to the current user needs, e.g. planning an exhibition project. For this purpose, ExPlanBART serves as a starting point, involving the user in a brainstorming session to narrow down the exact themes of the planned exhibition.

After a short chat, the chat content is passed to KeyBERT for keyword extraction. Using the extracted keywords as a search query, the system queries the local URL database and runs the results through a finetuned approximate string matching algorithm to extract assigned keywords, from the URL database and compare them to the keywords extracted from chat content. This way we narrow down the keyword pool and select only matching keywords from both lists. The final output is a temporary list of URLs relevant for the current user need (e.g. generation of a narrative for an exhibition on a specific topic). The list can be saved and modified for further inquiry. However, in the system itself, the output list is not stored, and is rewritten every session.

4. Results

4.1 VeriBERT

Due to the domain-specific application of VeriBERT, comparative evaluation with other out-of-domain URL classification models would produce misleading results. Thus, only a quantitative evaluation with basic metrics was conducted on a standalone test set.

Table 1 illustrates the evaluation results of VeriBERT on a curated test set. According to (Abad et al., 2023), *Precision* is a measure that demonstrates the proportion of true positive instances within all the instances that the model classified as positive. *Recall*, on the other hand, computes the proportion of true positive instances that the model correctly identified. *F1 Score* combines *Precision* and *Recall* in a single balanced metric. *Accuracy* measure refers to the proportion of true positive results among the total number of results.

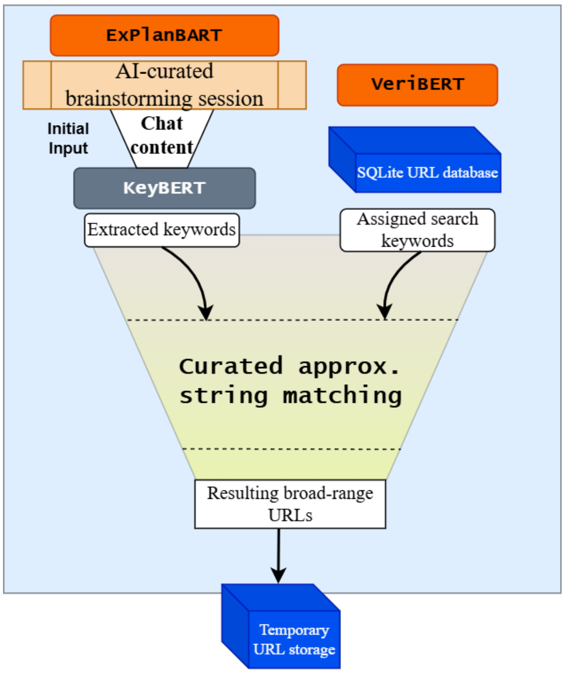

Figure 2. Overview of the downstream pipeline

	Domain Validity Classifier	Keyword Relevance Classifier
Accuracy	86.68	70.13
Precision	80.61	66.54
Recall	86.25	72.74
F1 Score	83.33	69.50

Table 1. Evaluation of both VeriBERT models on a curated test set

For a prototype model, the domain validity classifier achieves a high accuracy score, effectively filtering invalid domains with a high confidence score. However, the keyword relevance classifier model demonstrates valid, but comparatively suboptimal, performance. This is likely due to the dataset limitation, constrained by the intricate nature of the keyword relevance estimation task. However, during the downstream pipeline testing, the model successfully filters relevant pages, albeit with a lower confidence score.

4.2 ExPlanBART

According to numerous sources (Pan et al., 2019) (Zhang et al., 2022), (Kurdi et al., 2020), human evaluation remains the best strategy for evaluating the performance for QG systems. However, since human evaluation is resource intensive, we evaluate our proposed model on a number of common evaluation metrics, such as ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), BERTScore (Zhang et al., 2019), and back these metrics with rule-based statistical measures, such as *Diversity* expressed by number of distinct n-grams and average length of generations as well as % of outputs generated with a question mark, in par with the original prompt.

Table 2 demonstrates the results of the conducted comparative evaluation of dataset setups. For experiment clarity, we keep the training hyperparameters the same for all three models, only changing the dataset input. The results of the comparative evaluation demonstrate that the FLAN-augmented dataset with unchanged input entries achieves the best performance results on a held out test set. It is thus beneficial to use this dataset configuration as a starting point for the curation of a full fledged curatorial chatbot. Further enhancement of the current dataset with human-authored answer-question pairs present a promising future work direction.

	Baseline	FLAN-augment only	Prefix + FLAN-augment
ROUGE1 ↑	16.36	**72.55**	70.21
ROUGE2 ↑	4.34	**66.21**	63.15
ROUGE-L ↑	14.87	**70.61**	67.97
ROUGE-Lsum ↑	15.22	**71.72**	69.25
BERTScore_F1 ↑	86.76	**94.53**	94.06
METEOR ↑	10.71	**69.03**	67.36
Diversity (Distinct-1)	0.06	**4.98**	4.88
Diversity (Distinct-2)	0.011	**13.11**	12.39
Average length (words)	5.561	15.054	**15.75**
Ends with question(%)	**100**	99	99

Table 2. Comparative evaluation of ExPlanBART trained on different dataset configurations, tested on a held out test set

5. Discussion

During training of the domain validation classifier of VeriBERT, the dataset imbalance presented the biggest limitation: the number of samples was significantly biased toward the 'invalid' label due to the raw and heterogeneous nature of scraped web sources. However, this fact might provide a strong signal to the model about what exactly is an invalid domain, which is beneficial to our final goal.

We address this limitation in data pre-processing for the keyword relevance classifier model, successfully creating a balanced binary-labeled dataset. However, this balancing comes at a cost of dropping entries, which leads to dataset sparsity, since most entries in the curated compound dataset did not include keyword-related data (we omit keyword labels for entries labeled 'invalid', since the keyword relevance classifier only gets valid domains as input).

Overall, while the curated dataset could benefit from more diversified entries, which, apart from scraping more pages, could theoretically be achieved through data augmentation, the scarce and raw nature of the current dataset, sourced from organic, web-scraped pages, might serve as an advantage for the downstream application of the trained models, as the retrieval target of the system is the Web, heterogeneous and unstable by nature.

However, constant access to the Web is also one of the main limitations of the system. Since the system heavily relies on SERP API and several web-scraping algorithms in its operation, some degree of data loss is inevitable due to internal errors occurring during web search or snippet scraping. In this sense, web search and scraping present as a black box which cannot be influenced directly by the user or the system. Thus, when the scraper gets blocked by the website security system or the page returns a 404 error, the system just skips the fault page, possibly missing a valuable and valid information source.

Another limitation of the current system is the suboptimal performance of both VeriBERT models. Although the domain validation classifier does demonstrate strong performance with high accuracy for most cases, it can sometimes validate such domains as Wikipedia, which will be passed down the pipeline. We addressed this issue of occasional false positives by enabling the user to modify the output JSON file before committing it to the main database. However, to avoid such risks in the future, it is beneficial to collect a bigger dataset with more diverse domain pool and adjusted labeling pipeline to handle corner cases which might lead to the appearance of false positives and false negatives. In any case, we consider avoiding the usage of synthetic data for this task, since it may hinder both model performance for the in-the-wild web sources. This limitation makes the task of collecting a larger dataset particularly difficult, since we have to feed the labeling pipeline a unique set of keywords each time.

While talking about automated web search and scraping, the discussion of the ethical implications of such actions and the potential risks of scraping sensitive information and information from websites that forbid scraping is unavoidable. Mitigating such risks is another potential future work direction, although it seems rather challenging at the moment considering the black-box and separate nature of the agents utilized in the system in its current state.

Considering prior research on the topic of source credibility in the heritage domain discussed in Section 2.1, we can assume that the overall results presented by the proposed framework are promising, and the system is a step towards a fully automated source validation for downstream use, since even at the current state the performance speed allows for quasi-real-time usage. Training on a diverse pool of web-scraped sources makes the system flexible to various domains within culture heritage, such as art curation and historical research. Thus, further fine-tuning on a bigger set of more specific data, tailored to a specific use case, seems beneficial to boost the system's ability to generalize to specific research areas.

When it comes to curatorial question generation, it is difficult to speculate further potential use of this model as this task presents to be too specific and narrow for the current use case. However, specifying the exact dataset structure required for fine-tuning such an agent might become a valuable contribution to the development of museum conversational agents tailored for domain experts and curators, and the future potential research into the development of heritage domain curatorial agents. Switching the current BART backbone to a more chatbot-tailored model presents a potential future work direction and might turn a simple brainstorming session into a complete multifunctional human-computer interaction experience. In the scope of the current research objective, we only experimented with the dataset structure, applying the same backbone model to all dataset formats. As described above, we augmented the entries by reformulating them using explicit question templates. However, as noted in previous studies (Pan et al., 2019), learning from hand-crafted templates lacks real understanding and reasoning over input. At this point the model learns the patterns in the data, e.g. when it detects a name or location in the input, it invokes a pre-learned structure to wrap this token in it. Future research should focus on developing of a model with a deeper understanding of the input context to formulate more intricate questions and truly initialize a feasible interaction.

6. Conclusion

In this paper, we introduced a complex pipeline for filtering the web sources for their further use in the culture heritage domain. Although this paper focused on using web-retrieved sources for exhibition narrative preparation, the actual application scope of the proposed pipeline is much wider. The initial evaluation of both proposed models shows potential in further development of the system, and with a richer dataset and backbone experimenting, the pipeline has potential for in-domain adoption.

The introduced web source filtering framework is the first step towards the development of a multifunctional curatorial agent for the heritage domain, and research aimed at further improving the system and enhancing the current pipeline with generation and conversation features is currently ongoing.

References

Abad, S., Gholamy, H., Aslani, M., 2023. Classification of Malicious URLs Using Machine Learning. *Sensors 2023, Vol. 23, Page 7760*, 23, 7760. https://www.mdpi.com/1424-8220/23/18/7760.

Aljabri, M., Altamimi, H. S., Albelali, S. A., Al-Harbi, M., Alhuraib, H. T., Alotaibi, N. K., Alahmadi, A. A., AlHaidari, F., Mohammad, R. M. A., Salah, K., 2022. Detecting Malicious URLs Using Machine Learning Techniques: Review and Research Directions. *IEEE Access*, 10, 121395-121417.

Amin, A., Ossenbruggen, J. V., Hardman, L., Nispen, A. V., 2008. Understanding cultural heritage experts' information seeking needs. *Proceedings of the ACM International Conference on Digital Libraries*, 39-47. https://dl.acm.org/doi/pdf/10.1145/1378889.1378897.

Amin, A., Zhang, J., Cramer, H., Hardman, L., Evers, V., 2009. The effects of source credibility ratings in a cultural heritage information aggregator. *WICOW'09 - Proceedings of the 3rd Workshop on Information Credibility on the Web, Co-located with WWW 2009*, 35-42. https://dl.acm.org/doi/pdf/10.1145/1526993.1527003.

Banerjee, S., Lavie, A., 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. 1. https://dl.acm.org/doi/10.5555/1626355.1626389.

Becattini, F., Bongini, P., Bulla, L., Bimbo, A. D., Marinucci, L., Mongiovì, M., Presutti, V., 2023. VISCOUNTH: A Large-scale Multilingual Visual Question Answering Dataset for Cultural Heritage. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19. https://dl.acm.org/doi/pdf/10.1145/3590773.

Bu, F., Wang, Z., Wang, S., Liu, Z., 2025. An Investigation into Value Misalignment in LLM-Generated Texts for Cultural Heritage. https://arxiv.org/abs/2501.02039v1.

Cassidy, W. P., 2007. Online News Credibility: An Examination of the Perceptions of Newspaper Journalists. *Journal of Computer-Mediated Communication*, 12, 478-498. https://dx.doi.org/10.1111/j.1083-6101.2007.00334.x.

Choi, J. H., Watt, J. H., Lynch, M., 2006. Perceptions of News Credibility about the War in Iraq: Why War Opponents Perceived the Internet as the Most Credible Medium. *Journal of Computer-Mediated Communication*, 12, 209-229. https://dx.doi.org/10.1111/j.1083-6101.2006.00322.x.

Devlin, J., Chang, M. W., Lee, K., Toutanova, K., 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1, 4171-4186. https://arxiv.org/pdf/1810.04805.

Duff, W., Craig, B., Cherry, J., 2004. Historians' Use of Archival Sources: Promises and Pitfalls of the Digital Age. *The Public Historian*, 26, 7-22. /tph/article/26/2/7/89840/Historians-Use-of-Archival-Sources-Promises-and https://dx.doi.org/10.1525/tph.2004.26.2.7.

Fogg, B. J., Tseng, H., 1999. The elements of computer credibility. *Conference on Human Factors in Computing Systems - Proceedings*, 80-87. https://dl.acm.org/doi/pdf/10.1145/302979.303001.

Hovland, C. I., Weiss, W., 1951. The Influence of Source Credibility on Communication Effectiveness. *Public Opinion Quarterly*, 15, 635-650. https://dx.doi.org/10.1086/266350.

Kiousis, S., 2001. Public Trust or Mistrust? Perceptions of Media Credibility in the Information Age. *Mass Communication Society*, 4, 381-403. https://www.tandfonline.com/doi/pdf/10.1207/S15327825MCS0404$_4$.

Kurdi, G., Leo, J., Parsia, B., Sattler, U., Al-Emari, S., 2020. A Systematic Review of Automatic Question Generation for Educational Purposes. *International Journal of Artificial Intelligence in Education*, 30, 121-204. https://link.springer.com/article/10.1007/s40593-019-00186-y.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L., 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 7871-7880. https://arxiv.org/pdf/1910.13461.

Li, Y., Wang, Y., Xu, H., Guo, Z., Cao, Z., Zhang, L., 2024. URLBERT:A Contrastive and Adversarial Pre-trained Model for URL Classification. https://arxiv.org/abs/2402.11495v1.

Lin, C.-Y., 2004. Rouge: A package for automatic evaluation of summaries.

Mahdaouy, A. E., Lamsiyah, S., Idrissi, M. J., Alami, H., Yartaoui, Z., Berrada, I., 2024. DomURLs_BERT: Pretrained BERTbased Model for Malicious Domains and URLs Detection and Classification. https://arxiv.org/abs/2409.09143v1.

Majumder, B. P., Rao, S., Galley, M., McAuley, J., 2021. Ask what's missing and what's useful: Improving Clarification Question Generation using Global Knowledge. *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, 4300-4312. https://arxiv.org/pdf/2104.06828.

Pan, L., Lei, W., Chua, T.-S., Kan, M.-Y., 2019. Recent Advances in Neural Question Generation. https://arxiv.org/pdf/1905.08949.

Park, G.-M., Hong, S.-E., Park, S.-B., 2022. Post-Training with Interrogative Sentences for Enhancing BART-based Korean Question Generator. *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2, 202-209. https://aclanthology.org/2022.aacl-short.26/.

Rajalakshmi, R., Raymann, J., Prabhu, A., Karthik, R., Aravindan, C., 2024. AI-UNet: Attention Information-based deep URL Network for adult webpage classification. *Neural Computing and Applications*, 37, 2597-2615. https://link.springer.com/article/10.1007/s00521-024-10408-7.

Su, M. Y., Su, K. L., 2023. BERT-Based Approaches to Identifying Malicious URLs. *Sensors 2023, Vol. 23, Page 8499*, 23, 8499. https://www.mdpi.com/1424-8220/23/20/8499/htm https://www.mdpi.com/1424-8220/23/20/8499.

Sundar, S. S., Nass, C., 2001. Conceptualizing Sources in Online News. *Journal of Communication*, 51, 52-72. https://dx.doi.org/10.1111/j.1460-2466.2001.tb02872.x.

Tian, Y., Yu, Y., Sun, J., Wang, Y., 2025. From Past to Present: A Survey of Malicious URL Detection Techniques, Datasets and Code Repositories. https://arxiv.org/pdf/2504.16449.

Yu, B., Tang, F., Ergu, D., Zeng, R., Ma, B., Liu, F., 2024. Efficient Classification of Malicious URLs: M-BERT - A Modified BERT Variant for Enhanced Semantic Understanding. *IEEE Access*, 12, 13453-13468.

Zhang, R., Guo, J., Chen, L., Fan, Y., Cheng, X., 2022. A Review on Question Generation from Natural Language Text. *ACM Transactions on Information Systems*, 40, 14. https://dl.acm.org/doi/pdf/10.1145/3468889.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., Artzi, Y., 2019. BERTScore: Evaluating Text Generation with BERT. *8th International Conference on Learning Representations, ICLR 2020*. https://arxiv.org/pdf/1904.09675.

Zhou, Q., Yang, N., Wei, F., Tan, C., Bao, H., Zhou, M., 2018. Neural question generation from text: A preliminary study. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10619 LNAI, 662-671. https://link.springer.com/chapter/10.1007/978-3-319-73618-1$_5$6.