Intelligence artificielle générative au service de la médiation culturelle et patrimoniale : une approche interactive et personnalisée

Paul Pina-Gherardi, Thierry Antoine-Santoni, Marie-Michèle Venturini[1]

[1] Université de Corse Pascal Paoli, UMR CNRS 6240 LISA

{pina-gherardi_p, antoine-santoni_p, venturini_mm}@univ-corse.fr

Résumé

La découverte du patrimoine culturel ne passe plus uniquement par des visites sur place, mais est étendue grâce aux outils numériques. De nouvelles utilisations émergent, avec un nouveau public, cherchant plus d'interactions avec ces plateformes culturelles. Nous proposons une ébauche de médiation culturelle numérique basée sur des agents IA, qui génèrent, posent des questions de lecture, et évaluent la réponse de l'utilisateur. Cette médiation s'effectue sur des morceaux de documents écrits, souvent longs, qu'on retrouve sur ces plateformes, dont celle de la Médiathèque Culturelle de la Corse et des Corses (M3C) de l'UMR CNRS 6240 Lieux Identités eSpaces et Activités.

Mots-clés

Patrimoine, IA générative, culture, LLM, évaluation automatisée

Abstract

The discovery of cultural heritage is no longer limited to on-site visits but is extended through digital tools. New uses are emerging, attracting audiences seeking more interaction with cultural platforms. We propose a digital cultural mediation framework based on AI agents that generate reading questions and evaluate user responses. This mediation focuses on excerpts from long written documents available on platforms such as the Médiathèque Culturelle de la Corse et des Corses (M3C)

Keywords

Cultural heritage, generative AI, culture, LLM, automated evaluation

1 Introduction

Du latin *mediatio* (entremise) ou *medius* (milieu), la médiation désigne initialement une pratique faisant intervenir une tierce personne neutre pour résoudre des conflits, rétablir des relations, ou bien faciliter la circulation d'informations. La médiation fait également référence au fait que les communications entre deux acteurs les transforment, au-delà de la communication réalisée [12]. Les interventions à caractère culturel pour les visiteurs des musées ou autres structures culturelles sont des formes de médiation, car elles vont faire le lien entre le patrimoine et le public, en favorisant le plaisir de la découverte et en facilitant l'appropriation des connaissances. La médiation peut s'opérer à travers un dispositif, qui correspond à une substitution d'actions ou relations humaines, créant alors une asynchronie dans la communication. Ce dispositif se déploie dans des organisations, les écoles, les hôpitaux, etc. [5] Le dispositif est conçu par un ensemble de méthodes et techniques qui lui sont propres, par une ou plusieurs personnes et prévu ou être exploité par d'autres personnes, il a donc une dimension langagière et technique.

La massification des appareils numériques chez le grand public comme les smartphones et les ordinateurs qui a débuté à partir des années 2000 laisse transparaître de nouvelles méthodes de sauvegarde, de transmission et de communication entre les institutions. Ces nouveaux outils ont pour objectif d'«*enrichir l'expérience visiteurs*» en y ajoutant une dimension ludique, de visualiser ce qui a disparu, en «*créant un continuum numérique entre la visite elle-même*» [15]. Parmi ces outils, on retrouve l'intelligence artificielle générative (IAG), avec des capacités de traitement du langage naturel bluffantes[14]. Il devient possible de simuler des conversations, de générer du code, d'extraire des informations dans d'un texte, de générer du texte.

Cela représente cependant un travail considérable à mettre en œuvre, pour des résultats qui peuvent être perçus comme médiocres. L'apparition de comptes publics des institutions dans les réseaux sociaux, ainsi que les fonctionnalités de communication directement sur leurs site web ne génère que peu d'utilisation, et ce, pour ceux qui ont décidé de franchir le pas du numérique [8]. Les structures des musées, hiérarchisées et compartimentées, séparant les experts des techniciens informatiques et limitant la communication, mais également le manque de moyens sont des freins au développement d'outils numériques publics.

Dans le cadre de notre projet, nous proposons une implémentation des IAG pour favoriser l'interaction avec un visiteur d'une plateforme culturelle numérique [1]. Elle visera à générer des questions/réponses sur des documents, poser

1. Code source disponible à l'adresse https://github.com/ObaulG/M3C-LLM/tree/main

ces questions au visiteur, et évaluer ses réponses. Nous présenterons le contexte et l'état de l'art, puis nous détaillerons l'implémentation et son expérimentation, avant de commenter les résultats obtenus et conclure avec les perspectives futures.

2 Contexte et problématique

L'UMR LISA met à la disposition au public la Médiathèque Culturelle de la Corse et des Corses (M3C) dont l'objectif est «*de diffuser des savoirs scientifiques et culturels à destination du plus grand nombre*». Cette médiathèque a également une plateforme numérique[2], diffusant les oeuvres numérisées en accès libre. Elle compte plus de 200 visites récurrentes (au moins une fois par mois), avec des publics variés : des enseignants-chercheurs, doctorants et étudiants expérimentés, déjà familiers du contenu ; des étudiants et internautes qui arrivent via un moteur de recherche pour un thème, lieu ou auteur, et naviguent d'un document à l'autre ; et enfin, un public occasionnel découvrant la plateforme sans objectif défini, par simple curiosité culturelle. Ce dernier public dispose de la rétention la plus faible. Seulement, ce type de plateforme est souvent conçu pour être une vitrine laissant un accès ouvert à une quantité conséquente des ressources disponibles, sans accompagnement spécifique[16].

Nous pensons qu'implémenter des solutions basées sur les IAG améliorera l'expérience utilisateur pour la consultation de documents numérisés. Nous proposons ici un système expérimental où un utilisateur intéressé par un document serait accompagné par une IAG qui lui pose des questions sur des parties de documents. Ce système vise au départ le public occasionnel, et éventuellement étudiant.

3 État de l'art

L'intelligence artificielle générative s'est dévoilée au grand public en fin 2022, avec des *modèles de langage pré-entraînés* principalement sur l'architecture *Transformer* [20] : des réseaux de neurones profonds avec un système d'auto-attention[3]. On la distinguera des modèles d'IA "prédictifs" comme BERT[9], qui sont basés sur la partie *encoder* du Transformer[4], et dont l'utilisation peut être différente : analyse de sentiments, classification de textes, etc.

Ces systèmes produisent des données à partir de caractéristiques tirées d'un corpus initial de très grande taille, en déterminant le mot ou la suite de mots le plus probable qui viendra à la suite d'un mot déterminé[5]. La brique de base est le *token*, une séquence de caractères unique, déterminée par rapport à un corpus initial avec un algorithme comme *Byte-Pair Encoding*[4].

De manière générale, en notant x une séquence de *tokens*, où x_t est le t-ième *token* et $x_{<t}$ la séquence de *tokens* allant de x_1 à x_{t-1}, on peut modéliser l'approche autorégressive de la modélisation de langage, en calculant consécutivement les probabilités d'obtention du prochain token[17] :

$$p(x) = \prod_{t=1}^{T} p(x_t | x_{<t}) \tag{1}$$

La particularité de ces modèles est qu'ils sont *few-shot learners*[4] : avec un petit nombre d'exemples de textes décrivant la tâche et des exemples, les résultats augmentent significativement.

On peut retrouver une multitude d'actions réalisables par les IA génératives dans les milieux culturels. On retrouve principalement des *chatbots* conversationnels, dont le but est de changer le mode de communication des informations sur les objets dans les musées[18], ou sur des plateformes Web[19]. Des musées offrent la possibilité de simuler des conversations avec des personnalités historiques : le musée Orsay propose d'échanger avec Van Gogh, qui est en réalité une IAG *fine-tuned* avec 900 lettres qu'il a écrites à son frère[6]. Pour illustrer le quotidien de travailleurs du XVIIIe siècle, *The Workers Museum* crée des personæ interactifs, qui peuvent fournir la source de leur propos que les visiteurs peuvent ensuite consulter[10]. Les visiteurs viennent en général avec des questions sur le contenu, ou pour trouver du contenu selon une requête et l'IA apporte ces informations demandées[7]. Il existe aussi des applications qui génèrent des QCM à la volée selon les œuvres visitées par l'utilisateur [11].

Ces approches sont possibles avec le *Retrieval-Augmented Generation* (RAG), qui combine une mémoire paramétrique, représentée par les millions/milliards de coefficients du modèle de langage, avec une mémoire non-paramétrique, issue de divers documents denses, tels qu'une base de données de Wikipédia[13]. Formellement, à partir de la séquence d'entrée x, on cherche à construire une séquence de texte z pour l'utiliser en tant que contexte supplémentaire pour générer la séquence y. Un récupérateur d'éléments p_η va parcourir l'ensemble des documents D, et déterminer les k documents les plus pertinents :

$$z = \text{Top-}k\left(p_\eta(x, D)\right) \tag{2}$$

p_η attribue un score de pertinence, le plus classique étant la similarité cosinus entre les *embeddings* des séquences textuelles.

D'après Marion Carré, le visiteur devient acteur de sa visite car ces types de dispositifs ne peuvent fonctionner que sur commande[6]. Les contenus peuvent donc facilement s'adapter aux visiteurs : ces derniers peuvent moduler le niveau de détail, et le niveau de langue utilisé pour la description d'une œuvre ; et le multilinguisme est à l'honneur car les IA génératives peuvent plus ou moins facilement passer d'une langue à une autre dans l'exécution des prompts.

2. Voir https://m3c.universita.corsica/s/fr/page/home
3. https://www.tensorflow.org/text/tutorials/transformer?hl=fr
4. Tandis que les modèles d'IA génératives s'appuient sur la partie *decoder*
5. Firth J. R. «*You shall know a word by the company it keeps*» (1957)
6. https://www.musee-orsay.fr/fr/articles/numerique-bonjour-vincent-275618

4 Expérimentations

Nous posons l'hypothèse que *l'intégration d'un agent conversationnel génératif, posant des questions ciblées sur des extraits de documents numérisés, offre une nouvelle modalité d'interaction avec le patrimoine écrit, susceptible de modifier la manière dont les utilisateurs s'approprient, comprennent et perçoivent ces contenus*. Les quiz d'autoévaluation ou ludiques sont des possibilités existantes sur place, ou sur le site Web de l'institution [7] [8] [9].

Pour mettre en place un prototype, nous avons d'abord sélectionné des textes candidats pour l'expérimentation, en privilégiant des sujets accessibles au grand public. Deux livres ont été retenus : *Lochi mondu* [10] et *Atlas de la Corse contemporaine* [11]. Le premier présente des chroniques, courtes, sur des éléments culturels et historiques divers ; le deuxième propose un aperçu détaillé sous plusieurs angles de la région. Ensuite, nous avons découpé le contenu de ces livres par morceaux de 900 tokens environ ce qui représente à peu près une page de texte. Pour chaque morceau, nous générons 3 questions de lectures et leur réponses, d'une longueur moyenne de 40 mots, à l'aide du modèle `mistral-medium` [12]. Ces informations sont sauvegardées dans une base de données PostGreSQL. Les questions restent indépendantes entre elles. Contrairement à une précédente approche [11], les réponses sont rédigées et non des choix multiples. Nous pensons que laisser le temps à l'utilisateur d'écrire sa réponse est plus stimulant que de choisir parmi des réponses prédéfinies, et de plus, cela exploite les capacité de traitement du langage offert par les modèles.

Le prototype Web est ensuite proposé aux utilisateurs en leur présentant d'abord l'ensemble des documents disponibles . Le document sélectionné est ensuite téléchargé au format PDF et chargé sur la moitié gauche de la fenêtre, et le *chatbot* occupe l'autre moitié(voir figure 2). Pour l'expérimentation, 5 questions sont présélectionnées pour chaque livre. La première est affichée en tant que message, la page correspondant au morceau utilisé pour générer la question est indiquée à l'utilisateur, et le système attend sa réponse, qui est rédigée en langage naturel.

Nous employons deux instances différentes de modèles de langage pour traiter la réponse : une première, basée sur `mistral-small`, évalue si le message envoyé par l'utilisateur est bien une réponse à la question actuelle. Le prompt contient la question, la réponse attendue et le message, et produit une classe parmi `reponse`, `hors-sujet` ou `autre`. Dans le cas où le message est estimé comme étant une réponse, alors un modèle `mistral-small` est prompté pour noter de 1 à 10 la réponse, en vérifiant notamment que les différents éléments présents dans la réponse

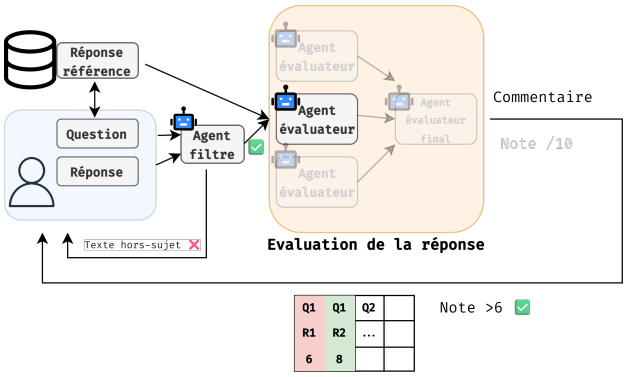

FIGURE 1 – Schéma du processus d'évaluation. Il peut y avoir un ou plusieurs agents évaluateurs, organisés hiérarchiquement.

de référence soient également présents dans la réponse de l'utilisateur. Un commentaire sera généré et transmis à l'utilisateur, indiquant alors la pertinence du propos fourni par rapport au texte. Dans une optique de médiation culturelle, la note est uniquement utilisée comme résultat numérique du *LLM-as-a-judge* à des fins statistiques et algorithmiques et ne sera pas présentée à l'utilisateur. Si la note obtenue est <7, alors l'utilisateur devra de nouveau répondre à la question.

L'application est implémentée en Python, avec un serveur Web FastAPI, et la librairie AtomicAgents [13] comme intermédiaire pour réaliser les appels aux modèles de langage. Elle a pour philosophie de déclarer des agents réalisant chacun une seule opération *atomique*, avec des types d'entrée et sortie définis à l'avance. Nous avons privilégié les modèles de Mistral pour leur accessibilité via une API gratuite, et nous élargirons les tests à d'autres modèles dans le futur.

Après avoir essayé le prototype, les utilisateurs ont répondu à un questionnaire, totalement anonyme. Il contient les questions du *SUS Questionnaire*, prévu pour évaluer le système de manière subjective[3]. Ces questions ont des réponses numériques variant de 1 (désaccord total) à 5 (accord total). Des questions supplémentaires spécifiques au prototype ont également été posées.

5 Résultats

5.1 Tests internes

Nous nous sommes d'abord intéressés aux résultats proposés avec une liste de réponses déjà préparées, pour évaluer la pertinence de l'évaluation des réponses. Pour les 5 questions du livre *Lochi mondu*, nous avons prompté les modèles `ministral-3b`, `ministral-8b`, `mistral-small`, `mistral-medium` et `mistral-large` ainsi :

```
# Background
Cet agent est spécialisé dans l'évaluation
des réponses des utilisateurs à des
questions de compréhension. Il vérifie si
```

7. https://www.cite-sciences.fr/fr/ressources/juniors/bidouilles-et-manips
8. https://www.muma-lehavre.fr/fr/collections/decouverte-ludique/quiz
9. https://musee.eurometropolemetz.eu/fr/quiz.html
10. https://m3c.universita.corsica/s/fr/item/116738
11. https://m3c.universita.corsica/s/fr/item/116782
12. https://mistral.ai/news/mistral-medium-3
13. https://github.com/BrainBlend-AI/atomic-agents/tree/main

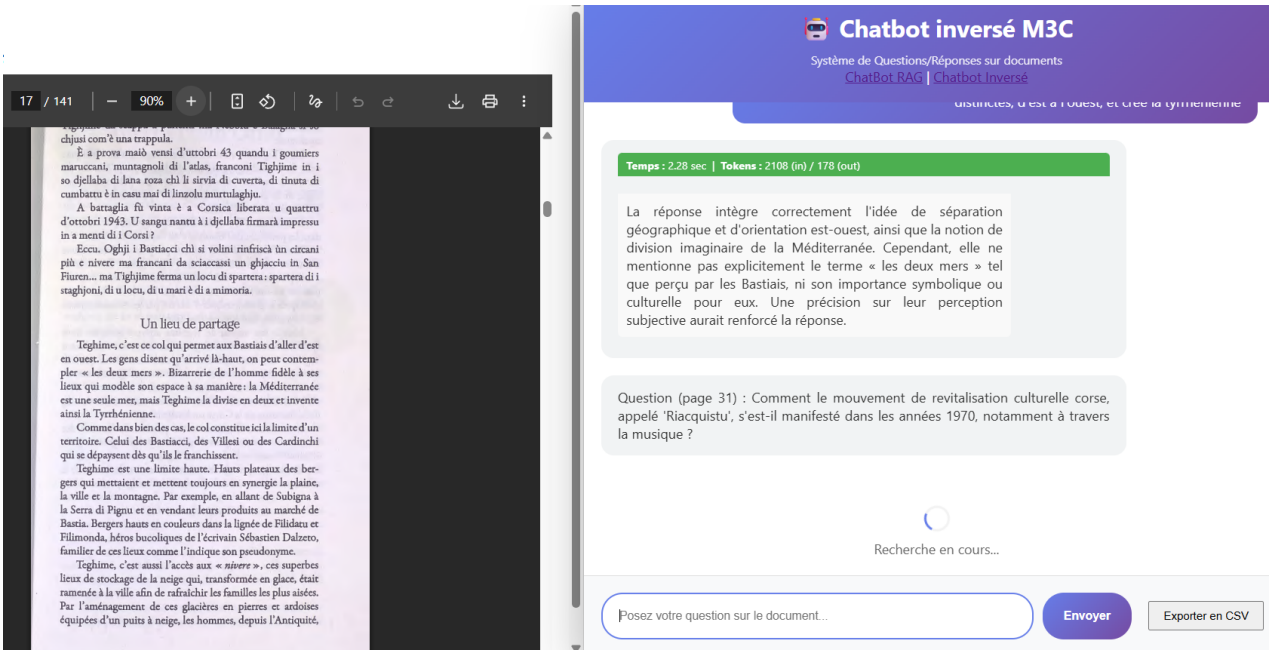

FIGURE 2 – Capture d'écran de l'interface du prototype de questions/réponses.

```
la réponse de l'utilisateur correspond à
la réponse attendue.

# Etapes
Analyser la question et la réponse
attendue. Comparer la réponse de
l'utilisateur avec la réponse attendue.
Évaluer la pertinence de la réponse.
Attribuer une note de 1 à 10
(1 = complètement incorrect,
10 = complet).
Répondre à l'utilisateur de manière
naturelle.
# Output instructions
La note doit être un entier
entre 1 et 10. Le commentaire doit être
clair, constructif et en français.
L'évaluation doit être légère. Ne pas
pénaliser si l'utilisateur donne une
réponse cohérente. Ne pas pénaliser si
l'utilisateur rajoute du contexte si
cela est pertinent. La réponse doit
être rédigée. Pas de réponse en
mots-clefs. Ne pas mentionner la note
dans la réponse rédigée.
# Question : {qst}
# Réponse référence : {rep_ref}
# Réponse utilisateur : {rep_ut}
```

Ces prompts sont fournis directement via l'API de Mistral sur le *Free Tier*. Nous avons également pu faire tourner localement le même modèle ministral-3b avec Ollama sur une machine disposant d'une Nvidia RTX 4060 portable avec 8Go de VRAM.

D'après la table 1, on voit sans surprise que les modèles les plus légers fournissent bien plus rapidement leur réponse que les modèles plus lourds. Ce temps est largement acceptable pour une telle application. Nous relevons également le nombre de *tokens* consommés par l'utilisation, avec en moyenne 1360 tokens par évaluation, ce qui permet d'avoir une estimation vulgaire du prix à payer dans le cas d'une utilisation d'API payante. Les fournisseurs donnent leur tarifs en général par million de *tokens*, variant de 0,1/M pour ministral-3b à 1,5/M pour mistral-large [14]. Le million de *tokens* serait donc atteint en moyenne avec 736 évaluations de question.

La qualité du retour des évaluations est globalement similaire pour l'ensemble des modèles. `mistral-large` est beaucoup plus généreux, autant sur le nombre de points, que sur les éloges [15] ajoutés à la fin du texte. Nous choisissons de retenir le modèle `ministral-8b` pour les tests utilisateurs.

Nous avons également essayé le paradigme multi-agent IA pour obtenir des types d'évaluation différentes : strict, bienveillant, pédagogique, créatif, minimaliste. Une série de prompts a été préparé à cet effet. Nous avons ensuite regroupé les évaluations par groupes de 2 et 3 agents, et créé un agent d'évaluation finale, qui compile les retours de chaque agents, ainsi que leur note, pour produire un retour et une note finale.

En parallélisant les appels, le temps de réponse moyen est environ 2 fois supérieur à celui d'une évaluation classique. Cependant, les réponses finales obtenues n'étaient pas particulièrement plus pertinentes qu'avec un seul agent, pour une consommation en *tokens* qui est environ 3 à 4 fois plus

14. Prix relevés le 24 février sur https://mistral.ai/pricing#api.
15. *"c'est une réponse qui mérite tous les éloges !"; "dans l'ensemble, c'est une excellente réponse "*

TABLE 1 – Résultats de l'exécution de la session questions/réponses pour chaque modèle. P est le nombre de paramètres en milliards, t est le temps total de la session, S_{moy} indique le score moyen donné à une réponse.

Modèle	**P(Mds)**	**t(s)**	S_{moy}	**Tokens**
ministral-3b	3	6,40	7,4	6897
ministral-3b (local)	3	17,48	9	6303
ministral-8b	8	7,06	7,8	7419
mistral-small	24	10,71	7,2	7152
mistral-medium	(?)	16,19	8,8	6358
mistral-large	41	19,93	9,2	6627

élevée. Nous garderons donc pour ces essais utilisateurs un seul modèle.

5.2 Par des utilisateurs

8 personnes ont pris part à l'expérience, toutes issues du monde universitaire. Parmi eux, 2 sont des doctorants en informatique, 3 doctorants en histoire contemporaine, 1 ingénieur de recherche au CNRS, 1 administratrice et 1 étudiante en Langue et Culture Corses. Tous sont familiers avec le fonctionnement d'une IA générative, mais pas forcément avec le contenu des livres. Le fonctionnement global du prototype leur a été expliqué en amont, puis chacun a choisi l'un des deux livres, et ont répondu aux questions qui ont été proposées. 1 personne seulement avait déjà lu le livre avant l'expérimentation.

Les personnes familières avec les sujets des livres étaient en mesure de répondre très rapidement aux questions, même s'ils n'ont pas lu ces livres. Les autres ont pris environ une minute pour lire le texte avant de commencer à répondre. Les réponses fournies par les utilisateurs sont de longueurs variables :

- pour les réponses courtes, le système a très souvent estimé que la réponse était trop incomplète. Cela survient lorsque la réponse références contient plusieurs informations et que l'utilisateur n'en fournit qu'une ; dans ce cas, le retour formule les indications pour compléter la réponse ;
- pour les réponses longues (parfois plus de 100 mots), le système a systématiquement considéré qu'elles étaient complètes. Mais pour un cas, pour la question *Pourquoi le col de Teghime est-il considéré comme un lieu symbolique pour les Bastiais ?*, la réponse attendue[16] ne contient qu'une affirmation parmi d'autres. La personne avait mentionné tous les autres éléments, sauf celui-ci, et le modèle a considéré la réponse comme incomplète.

Lors de tests complémentaires sur cette question précisément, nous avons remarqué que la réponse *Le col de Teghime sépare la méditérranée en deux mers distinctes, d'est à l'ouest, et crée la tyrrhénienne* est considérée comme incomplète, mais que la même réponse en ajoutant le mot "symbole" ou "symbolique" est considérée comme valide.

Certains utilisateurs ont cherché volontairement à contourner le fonctionnement prévu pour répondre aux questions. En copiant les paragraphes correspondant à la question, les informations demandées y sont nécessairement présentes, et donc le système accepte la réponse. Ou bien, certains ont volontairement donné un seul mot comme réponse, en espérant que le retour contienne toutes les informations, mais ces réponses ont été classées comme *hors-sujet*. Pour compenser, ils ont donc donné une petite phrase contextuellement correcte mais insuffisante, et ont copié le contenu du retour pour obtenir les bonnes réponses. Ce profil d'utilisateur peut se qualifier de *technocurieux* [2], adeptes de trouvailles fortuites et ayant *le goût de l'outil*.

Les utilisateurs étaient globalement satisfaits de leur expérience, ce qui est reflété par le score SUS moyen de 84. Ce score se calcule ainsi :

$$\text{Score SUS} = 2{,}5 \times \sum_{i=1}^{10} \begin{cases} x_i - 1 & \text{si } i \text{ est impair,} \\ 5 - x_i & \text{si } i \text{ est pair.} \end{cases}$$

6 Discussions

Le faible effectif ne permet pas de généraliser les résultats obtenus. Ce choix s'inscrit dans une démarche exploratoire, visant à identifier des pistes d'amélioration avant un déploiement à plus grande échelle. Les retours recueillis offrent des aperçus précieux sur l'ergonomie du système, la pertinence des questions générées, et l'adéquation entre les attentes des utilisateurs et les capacités de l'IA. Ils soulignent également l'importance d'un accompagnement humain. Les utilisateurs ont posé des questions sur la conception de l'application, y compris ceux qui n'ont pas de notions avancées.

La question financière mérite d'être mise en avant. Cette application ne serait pas la seule à exploiter l'utilisation de l'IAG dans le futur, et dans le cas où l'utilisation de l'API serait privilégiée, cette dépense même faible se rajouterait, dans un cadre où les dotations ont tendance à diminuer. Nous privilégierions des petits modèles de langage[17], dont les coûts des API, la consommation énergétique sont plus faibles, mais également car la philosophie de leur utilisation est de ne pas déléguer l'entièreté des opérations à l'IAG, mais seulement des tâches spécifiques et bien ciblées, en coordination avec un expert humain [1].

7 Conclusion

Les résultats du questionnaire sont encourageants et montrent bien l'intérêt que peuvent porter les visiteurs pour un tel outil, même si le nombre de participants reste limité. Les questions/réponses sont pour le moment générées, mais nous aimerions proposer une méthode de validation et correction humaine qui suivrait la logique des sciences participatives, pour que le contenu présenté devienne authentique. Des études en utilisant le paradigme inverse, basé sur le RAG, sont en cours, et pourront compléter les résultats en terme d'expérience utilisateur.

16. *Extrait : «il leur permet de traverser d'est en ouest et offre une vue sur ce qu'ils appellent « les deux mers » »*

17. Belcak. et al estiment comme «petits» les modèles de -10Mds de paramètres[1].

Enfin, nous souhaiterions proposer une personnalisation et une recommandation de documents selon le profil et les préférences de l'utilisateur, en s'appuyant par exemple sur les données de navigation et d'utilisation de la M3C.

Références

[1] Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, and Pavlo Molchanov. Small Language Models are the Future of Agentic AI, September 2025. arXiv :2506.02153 [cs].

[2] Marie-Laure Bernon. Les publics de musées sur internet : une typologie de visiteurs par leurs profils et usages des expositions en ligne. *Revue française des sciences de l'information et de la communication*, (27), December 2023.

[3] John Brooke. SUS : A 'Quick and Dirty' Usability Scale. In *Usability Evaluation In Industry*. CRC Press, 1996. Num Pages : 6.

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners, May 2020.

[5] Aurélie Canizares and Cécile Gardiès. Médiation, dispositif, médiatisation : un triptyque conceptuel à interroger. *Sciences de la société*, (107), August 2021. Number : 107.

[6] Marion Carré. Que peut l'IA générative pour la médiation culturelle ? *NECTART*, 20(3) :102–110, 2024.

[7] Angelo Geninatti Cossatin, Noemi Mauro, Fabio Ferrero, and Liliana Ardissono. Tell me more : integrating LLMs in a cultural heritage website for advanced information exploration support. *Information Technology & Tourism*, January 2025.

[8] Noémie Couillard and Maylis Nouvellon. À la rencontre des adolescents en ligne ? L'enjeu du numérique pour la médiation culturelle. *Cahiers de l'action*, 38(1) :43–49, 2013.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. arXiv :1810.04805 [cs].

[10] Maria Padilla Engstrøm and Anders Sundnes Løvlie. Using a Large Language Model as Design Material for an Interactive Museum Installation. In *Companion Publication of the 2025 ACM Designing Interactive Systems Conference*, pages 386–390. Association for Computing Machinery, New York, NY, USA, July 2025.

[11] Alessio Ferrato, Carla Limongelli, Fabio Gasparetti, Giuseppe Sansonetti, and Alessandro Micarelli. Exploring the Potential of Multimodal Large Language Models for Question Answering on Artworks. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, UMAP Adjunct '25, pages 432–436, New York, NY, USA, June 2025. Association for Computing Machinery.

[12] Patrick Fraysse. La médiation numérique du patrimoine : quels savoirs au musée ? *Distances et médiations des savoirs. Distance and Mediation of Knowledge*, 3(12), December 2015. Number : 12.

[13] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, April 2021. arXiv :2005.11401 [cs].

[14] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A Comprehensive Overview of Large Language Models, October 2024. arXiv :2307.06435 [cs].

[15] Dominique Pagès. La démocratisation culturelle et les promesses des médiations culturelles numériques : mirage ou tournant ? *Quaderni. Communication, technologies, pouvoir*, (99-100) :97–112, January 2020. Number : 99-100.

[16] Éva Sandri. Les ajustements des professionnels de la médiation au musée face aux enjeux de la culture numérique. *Études de communication*, 46(1) :71–86, July 2016.

[17] C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3) :379–423, 1948. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1002/j.1538-7305.1948.tb01338.x.

[18] Georgios Trichopoulos. Large Language Models for Cultural Heritage. In *Proceedings of the 2nd International Conference of the ACM Greek SIGCHI Chapter*, CHIGREECE '23, pages 1–5, New York, NY, USA, September 2023. Association for Computing Machinery.

[19] Iva Vasic, Hans-Georg Fill, Ramona Quattrini, and Roberto Pierdicca. LLM-Aided Museum Guide : Personalized Tours Based on User Preferences. In Lucio Tommaso De Paolis, Pasquale Arpaia, and Marco Sacco, editors, *Extended Reality*, pages 249–262, Cham, 2024. Springer Nature Switzerland.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, August 2023. arXiv :1706.03762 [cs].