Titre: Title:	GenQA : Génération et validation d'un ensemble de couples de Questions/Réponses générés à partir de données journalistiques
Auteur: Author:	Théo Jean Maurice Lecardonnel
Date:	2024
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Lecardonnel, T. J. M. (2024). GenQA : Génération et validation d'un ensemble de couples de Questions/Réponses générés à partir de données journalistiques [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/59470/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/59470/
Directeurs de recherche: Advisors:	Thomas Hurtut, & Christophe Hurter
Programme: Program:	Génie informatique

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

GenQA : Génération et validation d'un ensemble de couples de Questions/Réponses générés à partir de données journalistiques

THÉO JEAN MAURICE LECARDONNEL
Département de génie informatique et génie logiciel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
Génie informatique

Septembre 2024

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Ce mémoire intitulé :

GenQA : Génération et validation d'un ensemble de couples de Questions/Réponses générés à partir de données journalistiques

présenté par **Théo Jean Maurice LECARDONNEL**
en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment accepté par le jury d'examen constitué de :

Jinghui CHENG, président
Thomas HURTUT, membre et directeur de recherche
Christophe HURTER, membre et codirecteur de recherche
Amal ZOUAQ, membre

DÉDICACE

"Se junulo ne lernis, maljunulo ne scias." - L.L. Zamenhof
À tous les enseignants du secondaire comme
du supérieur qui m'ont mené jusqu'ici . . .

REMERCIEMENTS

Je tiens tout d'abord à exprimer ma gratitude envers *M. Thomas HURTUT*, mon directeur de maîtrise, pour m'avoir offert l'opportunité d'acquérir une expérience au sein du laboratoire jData et de Polytechnique Montréal. Je remercie aussi *M. Christophe HURTER*, mon co-directeur de maîtrise, pour nous avoir rejoint sur ce projet ainsi que l'ensemble de son laboratoire pour l'accueil qui m'a été fait à l'ENAC. Leurs conseils éclairés et leurs soutiens constants ont grandement contribué à enrichir mon parcours. Leurs enseignements resteront ainsi des atouts précieux pour ma carrière professionnelle à venir.

Mes remerciements vont également aux membres du laboratoire que j'ai pu croiser. Je pense tout particulièrement à *Mme Qianq XU* et *M. Louri COMPLAIN* pour leur expertise apportée sur différents aspects du projet, rendant cette expérience plus enrichissante.

Cette recherche a été financée par une subvention Alliance du CRSNG ALLRP-561132-20, en collaboration avec *Le Devoir*, et une subvention IVADO en collaboration avec *Radio-Canada*. Je suis également reconnaissant pour l'aide et la contribution apportées par *Le Devoir* et *Radio-Canada* lors des différentes étapes de ce projet.

RÉSUMÉ

De nos jours, les visualisations de données sont de plus en plus utilisées dans les articles de presse en ligne, notamment au sein de *data-driven storie*. Néanmoins, du fait de leur nature visuelle, ce type de contenu n'est que très peu accessible aux utilisateurs atteints de déficience visuelle. Ces utilisateurs doivent ainsi utiliser différents intermédiaires pour rendre audibles ces visualisations de données et accéder à l'information. C'est ainsi qu'un agent conversationnel ou un lecteur d'écran peut être utilisé.

La rédaction de texte alternatif est actuellement le standard d'accessibilité le plus communément utilisé pour fournir une description textuelle d'une image. Ces descriptions générales restent néanmoins peu utilisées par les salles des nouvelles de manière générale, et plus spécifiquement, pour les visualisations de données. De plus, lorsqu'elles sont existantes, elles sont généralement jugées trop simplistes par les utilisateurs atteints de déficience visuelle car lacunaires.

Plusieurs facteurs, humains comme économiques, peuvent être à l'origine de cette situation. Le nombre limité de journalistes disponibles pour rédiger ces descriptions détaillées, le manque de règles de rédaction précises et standardisées ainsi que la potentielle courbe d'apprentissage de la salle des nouvelles sont autant de limites liées à ce contexte journalistique.

Pour accroître cette accessibilité, nous proposons une nouvelle approche afin d'assister les journalistes dans leur production de description de visualisation de données, basée sur un ensemble de paires de Question/Réponse (Q/A) générés par IA.

Du fait des limites journalistiques précédemment listées, notre méthodologie génère ces Q/As en utilisant un modèle de Traitement Automatique en Langage Naturel (TALN) basé sur une IA générative. Cette approche atténue la charge de travail de la rédaction des Q/As en l'homogénéisant, permettant ainsi une exploration plus systématique et exhaustive des paires possibles pour une visualisation de données spécifique. Néanmoins, l'utilisation d'outils à base d'IA générative dans un contexte journalistique représente un risque quant à la publication d'informations peu fiables voire biaisées.

Cet écueil est contrebalancé par le grand degré de contrôle accordé au journaliste sur l'ensemble Q/As généré. Pour permettre et optimiser cette tâche de validation obligatoire, nous avons conçu une interface où les paires de Q/As sont regroupées autant sémantiquement que lexicalement mais aussi en terme d'intérêt lié à l'accessibilité. Des aides à la décision visuelles sont également utilisées afin d'améliorer la prise de décision du journaliste.

Pour évaluer cette méthodologie, baptisée *GenQA*, une étude comparative réunissant des journalistes de différents médias québécois a été menée. Cette étude a mis en avant la capacité de l'interface à assister les journalistes dans la production de description détaillée de visualisation de données. Ce constat repose notamment sur la sérendipité de *GenQA*, permettant de couvrir des thématiques non anticipées par les journalistes. De ces observations ont également émergé deux profils de journalistes distincts, fonction du nombre des couples de Q/A sélectionnés. Le premier, validant un nombre restreint de couples, consiste à choisir quelques couples sans en considérer l'intégralité. À l'opposé, le second profil propose une vérification systématique des Q/As, conduisant ainsi un temps de validation plus conséquent.

ABSTRACT

Data visualizations are now commonly used in online press articles and so-called data-driven stories. However, due to its visual nature, this type of content inherently lacks accessibility (e.g. when one wants to consume those visualizations using conversational agents, hearing them in audible formats, or using screen readers). Writing alternative texts is the recommended standard in order to provide text descriptions associated to an image. However, newsrooms rarely produce them for data visualizations, or when they do, these are overly simplistic. Several intertwined limitations explain that situation: the limited amount of time journalists have to produce these expected detailed descriptions, a lack of precise and standardized writing guidelines for describing visualizations, and a potential learning curve in the newsroom.

To improve this situation, we propose a new approach to help journalists produce a visualization description, based on a set of generated question and answer pairs (hereafter called Q/A). Due to the previously enumerated limitations, our method first generates those Q/As using a generative NLP AI model. This approach alleviates and homogenizes the writing task workload and allows for a systematic and more exhaustive exploration of the possible Q/As for a given visualization. However, among the critical challenges of using AI-based generative tools in a journalism context is the risk of publishing unreliable or biased information. Therefore, the methodology proposed in this paper gives the journalist user a high level of control over the AI-generated Q/As. To enable and optimize this mandatory validation task, we design an interface where Q/As are grouped in terms of semantic and textual content, and accessibility interest. Visual cues are also displayed to improve the journalist's decision-making.

To evaluate this proposed methodology, that we call *GenQA*, we conducted a comparative design study that gathered journalists from two different Canadian newsrooms. We observed that GenQA was efficiently used by those users and helped them to produce detailed visualization descriptions that met their expectations in terms of quality and workload. This study also showed that GenQA triggered significant serendipity potential, allowing users to explore and produce Q/As that cover aspects they might not have considered. Additionally, from these observations, two distinct profiles of journalists have also emerged, depending on the number of Q/A pairs selected. The first profile, validating a limited number of pairs, involves choosing a few pairs without considering the entirety. In contrast, the second profile suggests a systematic verification of Q/As, resulting in a more substantial validation time.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

HIC	Supervision globale du système - Human-In-Command -
HITL	Intervention humaine pour chaque décision - Human-In-The-Loop -
XAI	AI explicable - Explainable AI -
LLM	Grand Modèle de Langage - Large Langage Model -
LVLM	LLM visuel - Large Vision-Language Model -
TALN	Traitement Automatique en Langage Naturel
CQA	Interrogation de graphique - Chart Question Answering -
QA	Interrogation de texte - Question Answering -
Q/A	Question/Réponse - Question/Answer -
NASA-TLX	NASA Task Load Index - Mesure de charge de travail NASA -
FAQ	Foire Aux Questions - Frequent Asked Questions -
RAG	Génération Augmentée de Récupération - Retrieval-Augmented Generation -

LISTE DES ANNEXES

CHAPITRE 1 INTRODUCTION

1.1 Définitions et concepts de base

Ce sujet répond à un mandat de *Radio-Canada* et de *LeDevoir* visant à améliorer l'accessibilité des visualisations de données incluses dans les articles de presse en ligne. Ainsi, avant de présenter plus en détails la méthodologie *GenQA* qui a été développée, quelques éléments de contexte doivent être posés.

Visualisation de données Une visualisation de données (data-visualisation ou encore dataviz) est une représentation graphique d'un ensemble de valeurs, généralement difficile à interpréter sous forme tabulaire ou textuelle. Dans le cadre d'articles de presse en ligne, ces graphiques permettent d'illustrer les propos du journaliste. Ainsi comme les articles de presse, les visualisations de données font l'objet de choix éditoriaux afin de mettre l'emphase sur des faits particuliers comme une tendance, une valeur isolée et autre.

Éthique journalistique Les journalistes, comme toutes les professions réunies en fédération au Québec, se doivent de respecter un ensemble de règles dans la réalisation de leur travail. Pour le journalisme factuel, ces directives visent notamment à s'assurer de la qualité de l'information transmise aux lecteurs. Celui-ci s'axe principalement sur les cinq points ci-dessous, extrait du guide de déontologie journalistique du *Conseil de Presse du Québec*.

Ces points doivent être respectés dans tous les projets ayant une application journalistique. L'exactitude ainsi que la rigueur du raisonnement doivent donc particulièrement être scrutées.

Matériel journalistique

Les journalistes et les médias d'information ne transmettent pas leur matériel journalistique à des tiers, sauf si la loi leur en impose l'obligation ou s'il existe un intérêt public prépondérant justifiant de le faire.

Qualités de l'information

Les journalistes et les médias d'information produisent, selon les genres journalistiques, de l'information possédant les qualités suivantes [1] :

a) `exactitude` : fidélité à la réalité ;

b) `rigueur de raisonnement` ;

c) `impartialité` : absence de parti pris en faveur d'un point de vue particulier ;

d) `équilibre` : présentation d'une juste pondération du point de vue des différentes parties ;

e) `complétude` : présentation des éléments essentiels à sa bonne compréhension, tout en respectant la liberté éditoriale du média.

FIGURE 1.1 Extrait du guide de déontologie journalistique du CPQ

1.2 Problématique

Le Web est une porte d'entrée vers une mine d'informations et de ressources de tout type. Au cours des dernières années, l'utilisation et la diffusion des visualisations de données dans les médias en ligne (médias d'information, médias sociaux, blogs, etc.) a connu une croissance fulgurante, entraînant ainsi la nécessité d'une meilleure accessibilité de ce type de contenu. En effet, les graphiques, diagrammes et autres visualisations de données véhiculent le plus souvent des informations complexes qui restent largement inaccessibles aux personnes dépendant de lecteurs d'écran et d'autres technologies d'assistance (comme les agents conversationnels ou les formats audibles). Malgré la sensibilisation croissante à ce manque d'accessibilité par différents organismes (dont ISO [2], le ministère américain de la Justice [3] ou encore W3C [4]), l'utilisation de texte alternatif (ou alt-text) reste faible pour les images. Son emploi est encore plus marginal pour les visualisations de données. Une étude menée en 2019 a révélé que seulement 0,1% des images partagées sur Twitter étaient accompagnées d'alt-texts [5], en violation des différentes normes conçues pour améliorer l'accessibilité du contenu Web. Pourtant, ce faible taux d'utilisation des textes alternatifs n'est pas l'apanage des réseaux sociaux, il concerne également les contenus journalistiques en ligne. Différentes actions en justice pour manque d'accessibilité ont ainsi été entamées à l'encontre d'entreprises médiatiques. L'action

collective Burbon c. Fox News Network LLC [1] a ainsi mise en avant la négligence et la non-application de ces normes. Cette affaire judiciaire a souligné l'importance des alt-texts pour les médias en ligne, mettant en exergue les conséquences d'une telle absence sur l'exclusion et la discrimination numérique.

Plusieurs limites propres aux salles de rédaction peuvent expliquer cette situation déficiente. Comme chaque secteur économique, les médias sont soumis à des pressions financières importantes et, par conséquent, à de fortes contraintes de temps. Celles-ci sont de plus en plus présentes du fait d'une digitalisation accrue de l'information, entraînant une réduction du personnel pour de nombreuses salles de nouvelles. Selon les discussions menées avec différents médias québécois, seules quelques minutes peuvent ainsi être consacrées à améliorer l'accessibilité d'une telle visualisation. Ces restrictions financières ont également pour conséquence une réduction du personnel et donc du nombre de journalistes d'une salle des nouvelles. Or, confier cette tâche à un rédacteur en chef ou un pupitreur — qui est en plus chargé de préparer et corriger les articles avant publication — n'est pas non plus une solution idéale, puisque les objectifs de communication d'une visualisation de données sont propres à chaque article et à chaque journaliste. L'amélioration de l'accessibilité doit ainsi être effectuée par le journaliste qui a rédigé l'article de presse.

Le manque de directives rédactionnelles précises et normalisées sur la manière de décrire efficacement et précisément une visualisation de données avec du texte est également un problème pour les médias.

Une solution potentielle pourrait être l'automatisation de ce processus. Malgré les avancées technologiques actuelles en matière de TALN, l'automatisation totale de la rédaction de texte alternatif reste hors de portée. Les sous-sections de haut niveau des visualisations de données, telles que la mise en contexte ou celles liées aux connaissances générales, ne peuvent pas être générées automatiquement en raison de leur dépendance à la perception et au raisonnement humain [6]. Par conséquent, l'automatisation ne peut pas être utilisée pour générer une telle description.

1. Voir `https://www.classaction.org/media/burbon-v-fox-news-network-llc.pdf`

1.3 Hypothèses

Ce projet propose d'utiliser des couples de Questions/Réponses (Q/As) pour véhiculer l'information des visualisations de données auprès du grand public. Leur emploi a pour objectif d'optimiser le processus de validation de l'information pour les journalistes pour ainsi permettre d'obtenir des descriptions de meilleure qualité. Seul l'aspect journalistique est exploré dans ce mémoire. La présentation de ces couples (Section 5.2.5) ainsi que leur intérêt (Section 5.2.3) ne sont donc pas abordés.

Reposant sur l'utilisation de Q/As, ce projet nécessite le **développement d'un générateur de Q/As** à partir d'une visualisation de données et d'un article de presse. Des **algorithmes de regroupement sémantique comme lexical** des Q/As seront mis en place. Les journalistes pourront les utiliser en complément d'**aides à la décision** afin de valider les données plus efficacement. Enfin, une **interface** présentera de manière précise et concise l'ensemble des informations disponibles.

Le développement de l'interface permettant la réalisation de l'interface repose ainsi sur les hypothèses suivantes qui seront à valider :

- $\mathcal{H}_1$: les regroupements des couples mis en place améliorent la navigation et facilite la sélection des Q/As.
- $\mathcal{H}_2$: la génération automatique des Q/As est de bonne qualité et permet une forme de sérendipité.
- $\mathcal{H}_3$: les aides à la décision accélèrent la prise de décision et renforce la confiance envers le système.

1.4 Plan du mémoire

Le deuxième chapitre propose une revue de la littérature mettant l'emphase sur la fiabilité de sortie des LLMs ainsi que sur les différentes méthodes existantes pour décrire une visualisation de données. Le chapitre 3 présente, de manière générale, l'interface avant de détailler la succession d'algorithmes (génération des Q/As et présentation des couples). Le chapitre 4 présente le protocole adopté ainsi que les observations liées à la méthodologie *GenQA*. Finalement, la chapitre 5 résume l'ensemble du projet, en ouvrant notamment avec les futurs travaux à mener ainsi que les différentes possibilités de présentation des Q/As.

CHAPITRE 2 REVUE DE LITTÉRATURE

Afin d'améliorer l'accessibilité des visualisations de données, plusieurs pistes ont été explorées et expérimentées. Ces approches descriptives, bien que variées, peuvent être regroupées en deux catégories.

La première concerne la génération d'une description d'une visualisation de données dans son intégralité. Cette approche globale couvre idéalement l'ensemble des questions que pourrait avoir l'utilisateur final lors de sa lecture de l'article et de la visualisation du graphique (Section 2.1). Ce type de description est relativement long, nécessitant ainsi une organisation interne, peu présente dans les faits [6–8]. De part sa définition et sa proximité avec les textes alternatifs usuels (aussi appelés alt-texts), cette catégorie est compatible avec les différentes normes d'accessibilité. Ces descriptions globales conviennent ainsi aux utilisateurs sans question spécifique en tête, mais qui souhaitent explorer les informations disponibles d'une visualisation donnée.

À l'opposé, la seconde catégorie réunit l'ensemble des méthodes issues du Chart Question Answering (CQA). Ces algorithmes permettent à un utilisateur d'interroger, de façon précise, une visualisation de données et ainsi d'obtenir une réponse en retour (Section 2.2). Dans ce cas, l'utilisateur final n'a donc pas accès immédiatement à l'ensemble mais seulement à une sous-partie de l'information de la visualisation de données.

Quelle que soit la catégorie considérée, la génération automatique de ces descriptions textuelles repose largement sur les modèles neuronaux de type *LLM*. Tel que discuté précédemment dans l'introduction, il est ainsi nécessaire de réduire ces risques associés à leur utilisation, d'une manière générale, mais aussi spécifiquement dans ce contexte médiatique. La section 2.3 traitera des méthodes qui abordent ce problème de fiabilité.

2.1 Description générale d'une visualisation de données

Parmi l'ensemble des descriptions possibles de visualisations de données, la description générale représente l'approche la plus commune. Cette approche générale, basée sur une description globale de la visualisation, permet aux utilisateurs d'accéder à l'ensemble de l'information sans pour autant avoir besoin de formuler de requête. Du fait de sa proximité avec les textes alternatifs, il s'agit de la méthode usuellement utilisée pour rendre les images, et en particulier les visualisations de données, accessibles. Néanmoins, bien qu'étant la méthode la plus couramment employée, cette description générale d'une visualisation de données est généralement mal ou incorrectement appliquée. Par exemple, moins de la moitié des textes alternatifs inclus dans les publications HCI (Conférence sur les Interactions Homme-Machine) sont considérés comme complets, une proportion faible et relativement stable depuis une décennie [8]. Paradoxalement, ces sections aux descriptions lacunaires sont perçues par les lecteurs atteints de déficience visuelle ou non, comme les plus importantes [6]. La partie contextualisation, bien que populaire auprès des utilisateurs voyants [7] n'est présente que dans moins de 5% de ces textes alternatifs [8]. Dans ce contexte, plusieurs directives de rédaction et d'outils d'aide à la rédaction ont été expérimentés afin d'améliorer la qualité des descriptions.

2.1.1 Outils d'aide à la rédaction

Ces outils d'aide à la rédaction visent à permettre une rédaction plus efficace et plus qualitative des descriptions générales de visualisations de données tout en conservant une tâche de rédaction. Certaines de ces approches peuvent notamment fournir un retour aux auteurs sur la qualité de la description rédigée. *S. S. Chintalapati*, *J. Bragg* et *L. L. Wang* proposent ainsi un outil détectant les sections manquantes d'une description [8].

Au-delà de cette détection, une forme d'apprentissage par l'exemple a également été expérimentée. Spécifiquement dédié à améliorer l'accessibilité de visualisations de données inclues dans des articles scientifiques, *Alt4Blind* utilise la mise en relation de visualisations-références avec celle dont la description est à rédiger [9]. Cet algorithme utilise comme référence un ensemble composé de quelques centaines de visualisations et de leurs descriptions de « grande qualité » [1] ainsi qu'une comparaison d'une représentation vectorielle associée par cosinus-similarité. Lors de la rédaction de la description, l'auteur a ainsi sous les yeux différentes visualisations avec leur description, ces exemples étant relativement proches de la visualisation à décrire.

1. Traduction du terme *high-quality*

À l'opposé des algorithmes jusqu'alors présentés, *GenAssist* se propose d'améliorer la description d'une image [10]. Une première description sommaire est ainsi fournie en entrée. Suite à cela, l'image est interrogée par une série de questions. Celles-ci servent plusieurs objectifs comme la vérification de l'information de la description originale ou l'incorporation d'information liée au style et au visuel de l'image. Des questions sont également directement générées depuis l'image afin de permettre l'ajout d'information. Finalement, les réponses à chacune de ces questions sont agrégées afin de créer la description augmentée.

2.1.2 Règles de rédaction

Parallèlement à l'élaboration d'outils d'aide à la rédaction, un ensemble de règles de rédaction ont été conçues. Ces règles couvrent diverses caractéristiques telles que la langue, l'ordre des informations ou la longueur du texte [11]. Des patrons ont également été établis. Par exemple, pour un *barchart*, différents formats sont préconisés [12, 13] :

- `This is a horizontal bar chart titled` [TITRE]`, measuring` [UNITÉ X-AXIS] `for` [NOMBRE] ([Y-LABEL] | [`bars / clusters of bars`]).
  `A caption reads:` "[LÉGENDE]." `The data range from` [VALEUR X-MIN] `to` [X-MAX VALEUR] `in increments of` [PAS X-AXIS]. [TENDANCE]

- `The barchart represents` [TITRE] `where` [X-LABEL] `is plotted against` [Y-LABEL]. `This chart features the categories:` [X-AXIS TICK VALEUR].
  `The highest category is` [CATÉGORIE Y-MAX] `with` [VALEUR Y-MAX].
  `The lowest category is` [CATÉGORIE Y-MIN] `with` [VALEUR Y-MIN].

FIGURE 2.1 Exemples de patron d'une description globale d'un *barchart*

Ces deux exemples véhiculent des messages différents, le premier se focalisant sur les valeurs tandis que le second met l'emphase sur les labels et catégories. De plus, ces exemples ne prennent pas en compte des évolutions locales remarquables (comme une augmentation brutale ou une grande stabilité) ni d'éventuelles valeurs particulières telle que la valeur médiane. Cette profusion de règles pour décrire un même type de visualisation met également en avant l'absence de consensus sur ce sujet. En effet, aucune règle ni patron n'a jusqu'à présent été unanimement accepté par l'ensemble des organismes de normalisation. Par conséquent, leur utilisation dépendrait du journaliste sans aucune possibilité de normalisation, les rendant inutilisables.

2.1.3 Génération automatique

La génération d'une description générale d'une visualisation peut également être effectuée de manière automatique, sans aucune intervention humaine. Cette génération automatique de texte visant à décrire une visualisation de données, et plus généralement une image, a fait l'objet de différentes méthodologies dont la complexité du texte généré est variable.

Par exemple, les systèmes *AAT* [14] et *AIDE* [15] reposent sur la détection d'objets au sein des images à décrire. Dans le cas du processus *AAT*, la description générée, relativement simple, est ainsi présentée sous le format "`The image may contains :` [LIST_OBJETS]" - où `LIST_OBJETS` est la liste des objets détectés au sein de l'image - tandis que le système *AIDE* utilise une seconde couche, basée sur l'IA, afin de rédiger un texte avec une meilleure lisibilité.

Néanmoins, malgré ces prototypes, une rédaction automatisée de descriptions ne peut pas actuellement être employée. Une étude-utilisateur menée par *Lundagard & Satyanarayan* [6] conclut que les algorithmes IA ne peuvent pas décrire la contextualisation ou le contenu basé sur la cognition. Ainsi, les systèmes utilisant la rédaction automatique d'une description globale permettent l'édition du texte généré [16] pour faire face à ce défaut.

Finalement, la génération automatique de descriptions textuelles pour les visualisations n'est pas suffisamment mature pour atteindre la rigueur nécessaire à une pratique journalistique et les quelques directives de rédaction existantes ne sont pas entièrement applicables, du fait d'un manque de consensus.

2.2 Chart Question Answering

Contrairement à une description de l'intégralité d'une visualisation de données, une seconde approche préconise de mettre l'utilisateur final au centre. Cette approche, nommée CQA, fournit une description spécifique à partir d'une requête-utilisateur particulière. L'utilisateur peut et doit ainsi formuler une question sur des aspects particuliers de la visualisation. Cette stratégie est notamment plébiscitée par les individus atteints de déficience visuelle, permettant une analyse des données, la découverte de faits particuliers ainsi qu'un renforcement des connaissances [17].

Ce domaine, jusqu'à présent peu employé, a connu une récente expansion du fait de la généralisation des LLMs. Ces algorithmes reposent en grande partie sur ce type de réseaux neuronaux pour générer les réponses aux questions posées par l'utilisateur [18], faisant ainsi la force et la faiblesse du CQA. Par exemple, l'approche proposée par *S.K.C*, *P. Joshi* & *L.A.* consiste à interroger un tableau de données, extrait de la visualisation, à l'aide d'un modèle de Question Answering (QA) sur les Tableaux (TQA). Le processus ainsi développé souffre d'un faible taux de réussite, restant systématiquement inférieur à 70%, quel que soit l'algorithme d'extraction de données utilisé [19].

Si l'apparition de nouveaux modèles neuronaux basée sur la vision a réduit le taux d'erreur, il reste néanmoins trop im portant pour pouvoir être utiliser dans un contexte journalistique. Par exemple, la méthode *VProChart*, faisant partie *Vision Model Language* (ou VLM), a un taux d'erreur de l'ordre de 25% [20].

Les travaux existants portent principalement sur des questions factuelles [18] - question résolue par des opérations logiques ou arithmétiques - négligeant ainsi toutes les questions ouvertes qui pourraient pourtant fournir une contextualisation de la visualisation de données. Nonobstant le manque d'études sur ce dernier type de question, il demeure la catégorie la plus utilisée par les utilisateurs atteints de déficience visuelle [21]. Plus spécifiquement, les questions portant sur des détails spécifiques ou visant à acquérir de l'information supplémentaire constituent approximativement un tiers de l'ensemble des questions posées par ces utilisateurs.

Par conséquent, malgré son potentiel significatif et l'intérêt des utilisateurs, les méthodes issues du CQA sont peu utilisées pour améliorer l'accessibilité des visualisations de données.

Ainsi, que ce soit une description générale du graphique d'une visualisation de données ou de son interrogation, aucun des algorithmes présentés n'est applicable du fait du contexte journalistique strict.

2.3 Fiabilité de la sortie des LLMs

Bien que le recours aux réseaux neuronaux soit de plus en plus commun du fait de la récente diffusion de systèmes Large Langage Model (LLM) auprès du grand public, son usage comporte de nombreux défis, comme le risque d'hallucination. Cet artefact inhérent aux LLMs [22] se définit comme la production de texte grammaticalement correct mais factuellement inexact ou faux. Cela peut se traduire par des sorties générées discriminatoires et biaisées [23]. Ce phénomène hallucinatoire est particulièrement problématique dans un contexte journalistique où l'exactitude de l'information est d'une importance capitale, si ce n'est la plus importante. Ces hallucinations ouvrent également la voie à de possibles manipulations externes, de légères variations dans le prompt pouvant conduire à la génération de sorties trompeuses et erronées [24]. Cette faiblesse peut être à l'origine d'attaques basées sur l'injection de texte dans le prompt [25], rendant les systèmes LLMs particulièrement vulnérables.

2.3.1 Quantification de l'hallucination

Pour limiter l'impact de ces hallucinations, une première étape est de quantifier ce risque afin de mieux le comprendre. De nombreux indicateurs ont ainsi été conçus pour estimer l'importance de ce phénomène. Ces métriques peuvent être divisées selon leur domaine d'application : sur un couple de Question/Réponse spécifique ou sur l'intégralité du système LLM génératif.

Hallucination d'un couple de Question/Réponse

La mesure du risque hallucinatoire d'un couple de Question/Réponse repose sur l'utilisation d'autres systèmes LLMs. Cela peut être effectué à partir des états cachés (ou *hidden states*) d'un tel réseau neuronal [26]. Cette dernière approche fournit un couple de Question/Réponse à un LLM au format `"<Question> question <Answer> réponse"`. L'état caché final du segment lié à la question ainsi que l'état caché final du prompt sont regroupés afin d'établir un score d'hallucination. Si ce score permet d'estimer le risque d'hallucination, il ne permet pas pour autant de qualifier le type d'hallucination observé. Une seconde méthodologie consiste à comparer les différentes réponses obtenues par un ensemble de LLMs de référence pour une même question [23]. Ces réponses sont finalement agrégées par le biais d'une pondération basée sur l'expertise de chaque réseau neuronal. Ces deux approches permettent ainsi d'obtenir un score d'hallucination pour un couple de Question/Réponse donné, quantifiant donc le risque d'incohérence entre la question et la réponse d'une même couple.

Hallucination du système

Par opposition d'une mesure spécifique à une entrée, le risque d'hallucination de l'ensemble d'un système LLM peut être évalué. Ceci a notamment pour objectif de sélectionner le modèle avec le plus faible risque. *Y. Liu et al.* l'évaluent dans le cas d'un Large Vision-Language Model (LVLM) par le biais d'une interrogation du système sur la présence (ou non) d'objets au sein d'une image [27]. Les objets testés sont directement issus de l'image ou échantillonnés négativement. Cette approche peut néanmoins être affinée par le biais de métriques plus précises. Ainsi, *G. Hong et al.* propose de qualifier les modèles LLMs grâce à une métrique de factualité [2] et d'un score de fidélité [3] [28]. Toutes deux se basent sur la réalisation d'un ensemble de tâches, respectivement QA, vérification de faits, détection d'hallucination et résumé, compréhension écrite, réalisation d'instruction, détection d'hallucination.

Toutes ces évaluations soulignent l'importance de ces hallucinations, mais surtout de sa prise en compte dans des systèmes plus complexes. Néanmoins, malgré les efforts effectués pour les comprendre et les quantifier, *Z. Xu*, *S. Jain* et *M. Kankanhalli* ont mathématiquement démontré qu'éliminer totalement ces artefacts hallucinatoires est irréalisable [22]. Par conséquent, seuls ces symptômes peuvent être minimisés grâce à un système d'atténuation sous supervision humaine afin d'assurer la solidité, la transparence, la fiabilité ainsi que la sécurité du système [29].

2.3.2 Le facteur humain : un facteur d'atténuation

Au sein d'un système, le degré d'implication humaine peut fortement varier en fonction du niveau d'automatisation, allant ainsi de Human-In-The-Loop (HITL), nécessitant une intervention humaine à chaque prise de décision, à Human-In-Command (HIC), caractérisé par une supervision humaine globale du système. Conformément à l'éthique journalistique stricte (Figure 1.1), seul le plus haut degré de contrôle, HITL, sera considéré comme approprié. L'intégration de cette supervision humaine peut être effectuée à divers stades du processus de génération.

2. Traduction du terme *factuality score*
3. Traduction du terme *faithfulness score*

Interaction humaine au sein du système IA

Tsiakas et *Murray-Rust* proposent une architecture où les humains interagissent directement avec le système génératif lui-même [30]. L'intégration des principes HITL selon le modèle Explainable AI (XAI) permet au système de fournir et de recevoir des informations d'utilisateurs, autorisant ainsi la prise en compte des expériences d'utilisations. Néanmoins, le schéma proposé nécessite d'allouer des rôles spécifiques comme un *AI Designer* responsable de l'adaptation du modèle IA (re-entraînement, changement de configurations IA, etc.) et un *supervisor*, responsable d'ajuster le modèle au contexte d'application. Interagir directement avec l'IA exige donc de nouvelles capacités qui vont au-delà des champs de compétence des salles de rédaction. De plus, le modèle XAI diminue paradoxalement la confiance des utilisateurs dans le système en réduisant la compréhensibilité du système [31]. La supervision humaine ne peut donc intervenir directement sur le système LLM.

Post-supervision humaine

Une telle supervision doit ainsi être effectuée à la suite de la génération de la sortie par le LLM. Centrée sur une vérification postérieure, cette approche nécessite une expertise quant à la véracité de la sortie générée. Suivant ce principe, *Biloborodova* et *Skarga-Bandurova* ont conçu un *framework* [32] basé sur une collaboration parallèle entre une IA entraînée à cette tâche spécifique et les contributions humaines, aboutissant à une décision collective sur l'acceptation ou non de la sortie du LLM. Néanmoins, le facteur humain n'a pas ici la prédominance sur la décision finale, ce qui est contradictoire avec l'éthique journalistique. Cette supervision doit donc uniquement être effectuée par un expert du domaine.

Absence de supervision humaine

À la marge des systèmes utilisant une réelle supervision humaine, différents algorithmes acceptent le risque d'hallucination de la sortie générée en communiquant l'incertitude auprès de l'utilisateur final, pour peu que celui-ci possède une expertise sur le sujet considéré. Ainsi l'algorithme *AAA* [14] utilise le terme anglophone `may` et *AIDE* [15] les termes `may` et `might`, traduisant l'incertitude, dans chacune des phrases générées afin de mettre en garde l'utilisateur final quant à la véracité de l'information transmise par ce texte. Néanmoins, cette communication ne peut à elle seule suffire pour améliorer l'accessibilité à cause du non-respect de l'éthique journalistique du fait de l'absence de vérification des faits du texte soumis à l'utilisateur final.

Afin d'être efficace tout en permettant un réel contrôle humain sur le système, seule une supervision humaine post-génération est envisageable. Toutefois, son intégration dépend également du type de description utilisé.

2.4 Objectifs et Hypothèses

Aucun des algorithmes présentés dans cette revue de littérature ne permet à lui seul d'améliorer l'accessibilité dans notre contexte journalistique. Ainsi, si chacune des modalités de présentation possède des faiblesses (description rarement existante versus absence de contrôle de l'information) une combinaison des deux peut permettre d'en exploiter les avantages. Ce consensus passe par l'usage de couples de Q/As - propre au CQA - pré-générés. Cette pré-génération, indépendante de l'utilisateur final, permet de couvrir l'ensemble de la visualisation, comme pourrait le faire une description générale. Malgré cela, la tâche de rédaction des Q/As, longue et fastidieuse, doit être déléguée à un réseau neuronal. Sous contrôle HITL, le plus conservateur, chacun de ces couples doit être validé par le journaliste pour être présenté aux utilisateurs finaux.

2.4.1 Objectifs de recherche

Le but de ce projet de recherche est ainsi de concevoir une nouvelle méthodologie journalistique améliorant l'accessibilité des visualisations de données incluses dans les articles de presse en ligne. Ce processus, nommé *GenQA*, se situe ainsi entre le matériel journalistique et la présentation auprès du grand public de la description textuelle ainsi générée. *GenQA* repose notamment sur l'utilisation de couples de Questions/Réponses pour la validation de l'information par les journalistes et sa transmission auprès des utilisateurs finaux. L'interface de cette méthodologie, implémentant l'outil de validation côté salle des nouvelles, se propose d'intégrer des aides à la décision ainsi que différentes modalités de regroupement.

Ce projet constitue ainsi une preuve de concept de la capacité des journalistes à sélectionner et valider un nombre important de couples de Questions/Réponses pour ainsi mener à de futurs travaux sur l'utilisation d'un tel format de présentation de l'information.

Des Questions/Réponses pour véhiculer l'information

Dans l'approche proposée dans ce mémoire, une méthodologie assistée par ordinateur est présentée afin d'aider les journalistes à produire des descriptions textuelles des visualisations de données incluses dans des articles de presse en ligne. Ces descriptions s'appuient sur un ensemble de paires de Q/A. Centré sur l'utilisateur final, ces Q/As offrent une large flexibilité quant à leur utilisation et leur modalité de présentation. Ces couples peuvent en effet répondre à différents besoins, au-delà de l'accessibilité des personnes atteintes de déficience visuelle.

Ainsi, si un agent conversationnel basé sur ces paires accroît la compréhension et l'exploration des données de la visualisation [33], un lecteur d'écran permet un parcours de ces mêmes données avec un plus faible effort cognitif [34].

Cette approche génère de manière automatique les couples à partir du matériel journalistique grâce à un modèle d'IA génératif TALN. Cette approche neuronale atténue ainsi la charge de travail des journalistes tout en homogénéisant la rédaction de ces couples, permettant une exploration plus exhaustive et systématique de l'ensemble des Q/As possible pour une visualisation donnée [35]. Néanmoins, l'usage de l'Intelligence Artificielle dans un contexte médiatique soulève de nombreux enjeux. Le risque de publier des informations peu fiables ou biaisées est une préoccupation majeure des médias, au centre de l'éthique journalistique et des standards actuels.

Une interface comme outil de validation

Pour optimiser, et ainsi rendre faisable cette tâche de validation, nous proposons une interface dont les couples de Q/As sont regroupés sémantiquement comme lexicalement. Plusieurs indices visuels y sont représentés afin d'améliorer la prise de décision du journaliste et d'accroître sa confiance en l'interface [36].

CHAPITRE 3 INTERFACE

Cette interface journalistique constitue une preuve de concept de la méthodologie *GenQA*, proposée dans ce mémoire. Dans cette partie, elle est présentée du point de vue de l'utilisateur (Section 3.1) avant une description détaillée des différents algorithmes utilisés (Sections 3.2 et 3.3).

3.1 Présentation générale

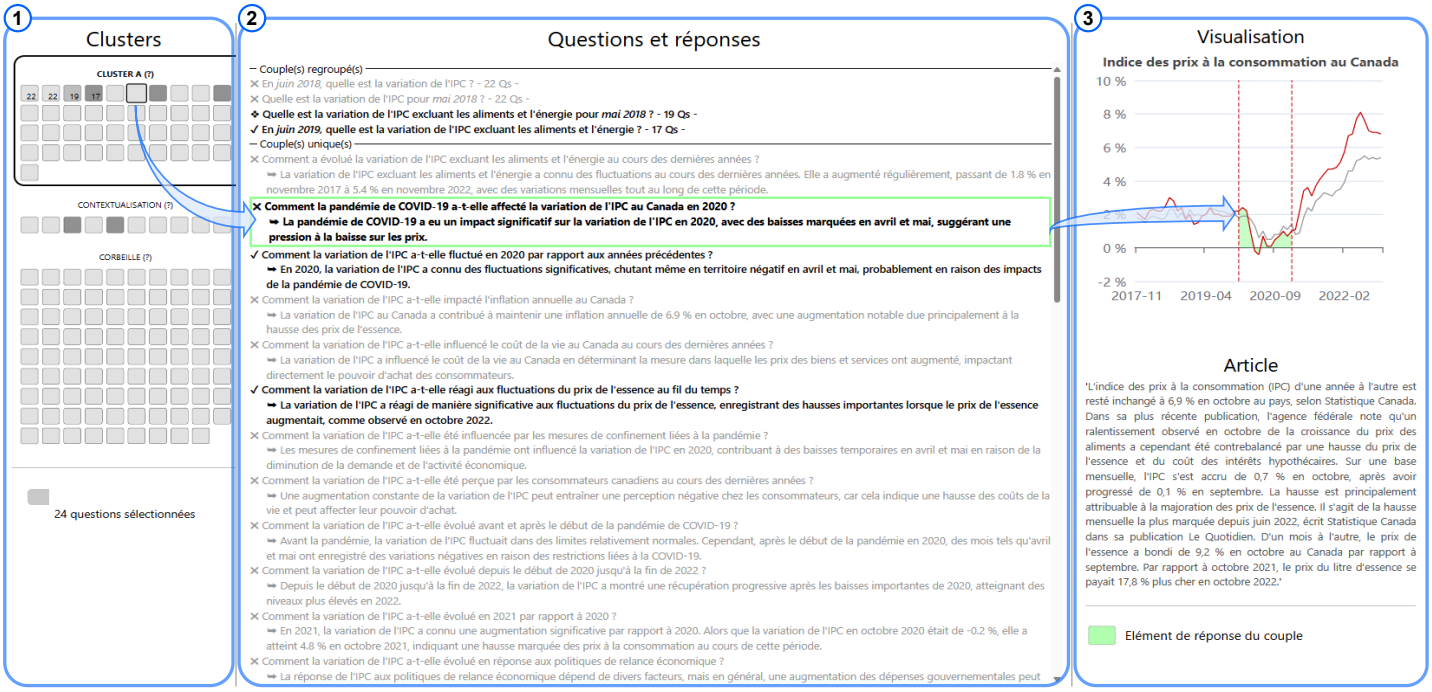


FIGURE 3.1 Présentation générale de l'interface

Divisée en trois colonnes, cette interface regroupe l'ensemble de couples de Questions/Réponses qui ont été générés ainsi que les aides à la décision associées au matériel journalistique. Les deux premières colonnes représentent tous les Q/As sous différents formats (graphique pour la colonne 1 et textuelle pour la colonne 2) tandis que la dernière met en avant les éléments d'aides à la décision (3). La figure 3.1 met également en avant le parcours du journaliste au sein de l'interface. Celui-ci est détaillé dans la suite de ce mémoire.

Afin d'optimiser le tâche de validation du journaliste, les Q/As ont été regroupés. Le premier se base sur la sémantique afin de limiter les ruptures de sujet lors du passage d'un couple à l'autre. Le second utilise la proximité lexicale. Cela permet d'accorder un degré de confiance plus important à ces ensembles du fait de la répétition de mêmes structures lexicales.

(1) Agrégation des Q/As Cette première colonne contient la vue la plus générale, grâce à la représentation de l'ensemble des couples qui ont été automatiquement générés. Cette vue d'ensemble subdivise les couples de Question/Réponses en différents regroupements. Tout en haut figurent ainsi les clusters jugés les plus pertinents, étiquetés par une lettre. À l'intérieur de chacun de ces clusters, les Q/As partagent des similarités sémantiques comme un thème commun ou une même source au sein des données journalistiques. Le regroupement suivant, nommé « Contextualisation », regroupe tous les couples qui ne sont pas associés au matériel journalistique. Finalement, le cluster « Corbeille » contient l'ensemble des Q/As de type « données manquantes » (Section 3.3.2).

(2) Passage en revue et sélection des Q/As Comme une extension de la première colonne, cette colonne centrale représente le contenu textuel du regroupement sémantique courant (ou associé comme « Contextualisation » et « Corbeille »). Par exemple, la figure 3.1 représente le contenu du CLUSTER A. Le journaliste peut alors passer en revue l'ensemble des couples de Question/Réponse individuellement et ainsi les sélectionner ou non. Dans cette section, les Q/As sont organisés en cluster selon la proximité lexicale de leurs questions. Lorsqu'un tel regroupement lexicale existe, ces ensembles de Q/As sont affichés sous le label « Couple(s) regroupé(s) », par opposition aux autres couples étiquetés « Couple(s) unique(s) ». Les couples lexicalement regroupés, ceux dont les questions sont similaires et ne varient que de quelques mots, sont affichés dans la partie supérieure (Figure 3.9) afin de les répartir selon leur potentiel journalistique, c'est-à-dire, selon l'intérêt que pourrait porter le journaliste sur ces Q/As.

(3) Données journalistiques Liée aux deux autres parties, cette dernière section met l'accent sur l'élément de réponse le plus probable associé à un couple de Question/Réponse afin de faciliter et d'accélérer la prise de décision. Cet élément de réponse est mis en avant au sein de la visualisation ou de l'article. Le journaliste peut ainsi utiliser cette information pour valider la véracité d'un Q/A. Cet élément permet aussi de qualifier des sous-ensembles de Q/As spécifiques ou de se focaliser sur une plage de données particulière, par la représentation de distribution (Figure 3.10).

L'ensemble des couples de Questions/Réponses, affichés dans l'interface précédemment présentée, sont générés lors d'une phase de pré-traitement (Section 3.2). Suite à cela, les paires de Q/As sont organisées et regroupées (Section 3.3) au sein des différents gaufriers.

3.2 Pré-traitement

Cette première phase consiste à générer l'ensemble des couples de Questions/Réponses à partir du matériel journalistique (Figure 3.2). Pour permettre leur exploitation, les données issues de l'article de presse et de la visualisation sont normalisées (Section 3.2.1) puis liées entre elles (Section 3.2.2) avant d'être utilisées dans la génération des questions et des réponses (Section 3.2.3).

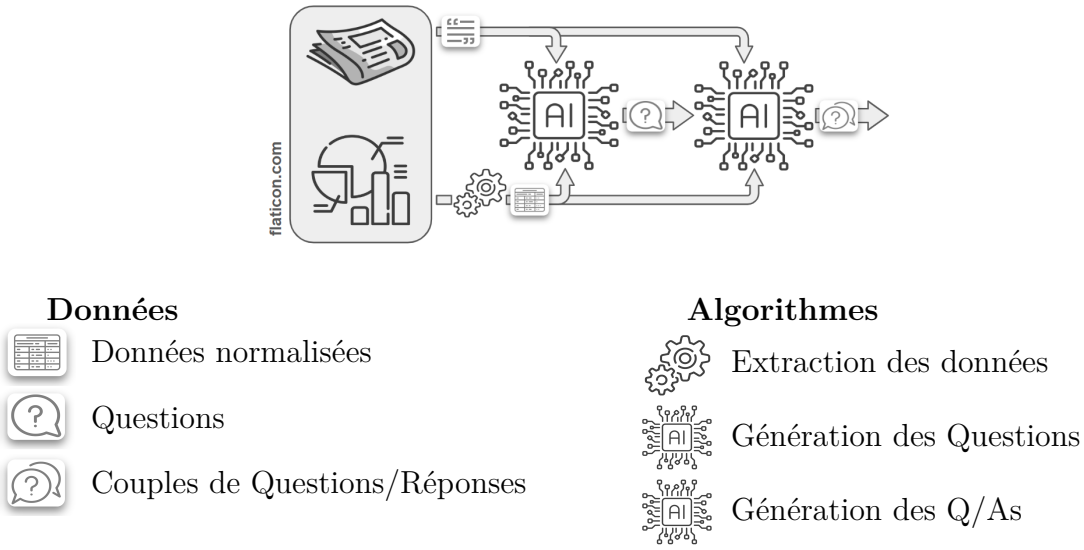


FIGURE 3.2 Algorithmes de pré-traitement

3.2.1 Extraction des données de la visualisation

Cette première phase de ce pré-traitement consiste à extraire les données véhiculées par la visualisation afin de les rendre utilisables pour la suite du processus.

L'utilisation du DOM (format XML) permet d'extraire les valeurs numériques réelles, et non une approximation d'une image « raster »[1] [37]. Dans le cas de visualisation de données, cette approximation due au format peut conduire à des taux d'erreur de 7.7% pour un *linechart*[2] [38]. Néanmoins, le DOM ne peut être utilisé tel quel. Les processus qui utilisent

1. image raster : image définie pixel par pixel (ex : PNG), par opposition à la vectorialistion (ex : SVG)
2. *linechart* : diagramme en ligne

un tel format en entrée sont relativement imprécis. Par exemple, Xu et Wall propose une AI entraînée à partir de visualisations au format SVG (issu du XML) pour y effectuer une analyse visuelle [39]. Si cette approche permet d'identifier des valeurs remarquables, elle peine en revanche à en extraire des valeurs spécifiques. Une transformation est nécessaire afin de le rendre exploitable et compréhensible par un LLM.

Cette extraction utilise deux modalités différentes pour mettre l'emphase à la fois sur l'aspect visuel et sur les données numériques. Un algorithme de *Web-scraping*[3] est exécuté sur le DOM de la visualisation pour en extraire les propriétés visuelles (couleurs, labels) et certaines métadonnées comme le type de graphique, la source ou l'auteur. Ces dernières sont plus particulièrement extraites des sections *<header>* et *<footer>*. Les données numériques associées sont, elles, extraites à partir le fichier CSV original de la visualisation.

3.2.2 Lien entres les données

Une fois ces données extraites et exploitables, une mise en relation est réalisée entre les données numériques et l'article de presse. Cela passe par l'édition de champs par le journaliste au travers d'une fenêtre appelée « *Complément d'information* » (Figure 3.3). L'objectif étant de créer un lien entre les termes employés dans l'article et les attributs de la visualisation, il est ainsi possible de renommer le nom des axes, le nom des attributs ainsi que le nom des couleurs de la visualisation. Tous ces champs sont initialement pré-remplis avec les données fournies issues de la visualisation. Cette possibilité d'édition permet au journaliste d'avoir l'ultime contrôle sur cette étape cruciale pour la suite du processus.

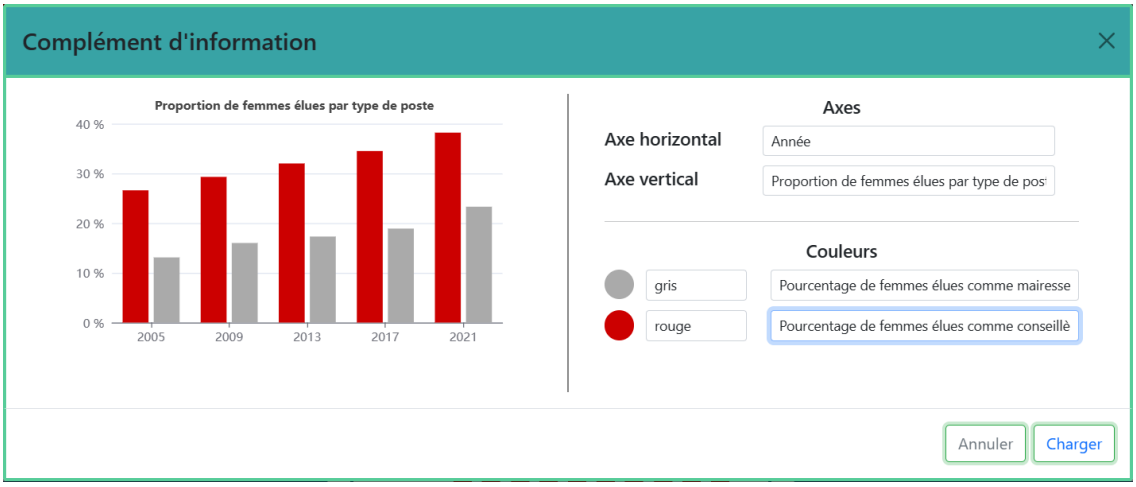


FIGURE 3.3 Fenêtre contextuelle

3. *Web-scraping* : moissonnage du Web

3.2.3 Génération des couples de Questions/Réponses

Une fois les données sur la visualisation et l'article réunies, il est possible de générer l'ensemble des couples de Question/Réponse. Ce processus est réalisé par un modèle génératif (dans ce cas ChatGPT-3.4) de type LLM. Le choix s'est porté sur ce modèle puisqu'il s'agit, lors de la réalisation de cette maîtrise, du LLM le plus populaire et le plus facilement accessible car disponible en ligne. Néanmoins, dans un objectif d'optimisation du processus, d'autres modèles pourraient être envisagés, notamment afin de permettre un affinage.

Afin de permettre l'incorporation de questions-heuristiques lors de ce processus, la création des Q/As est scindée en deux étapes distinctes : la génération des questions suivie par celle des réponses. Ces questions-heuristiques, uniquement dépendantes du type de la visualisation, peuvent par exemple porter sur la tendance des données ou sur les valeurs extremums dans le cas d'un *linechart*[4] ou sur l'attribut majoritaire pour un *piechart*[5].

La génération de l'ensemble des questions est réalisée depuis les prompts ci-dessous. Ces deux requêtes mettent en exergue des parties différentes des données. Les questions sont ensuite extraites des sorties textuelles obtenues par le biais d'expressions régulières. Ce choix permet une grande souplesse quant au format de sortie du LLM, qui est, par nature, très variable.

- $\mathcal{P}_1$ `Génère-moi un array de` [QUEST_NOMBRE] `questions sur les données` [DATA] `de l'article` [ARTICLE].
- $\mathcal{P}_2$ `Génère-moi un array de` [QUEST_NOMBRE] `questions à partir du` [CONTEXTE]. [VISUEL] `Elle lie les données` [DATA] `avec l'article` [ARTICLE].

où `QUEST_NOMBRE` est le nombre de questions à générer, `ARTICLE` l'article de presse, `DATA` les données issues de la visualisation et `VISUEL` les données liées au visuel du graphique.

Le nombre de Q/As a été ajusté afin de fournir un compromis entre découvrabilité, favorisant un nombre important de Q/As, et le risque de redondance (nombre faible de Q/As). De plus, ces couples doivent également être manipulables par le journaliste, même à travers notre interface. Ce nombre a ainsi été fixé à 200 pour la majorité des visualisation de données et 100 pour les plus petites (avec un nombre restreint de valeurs).

4. *linechart* : diagramme en ligne
5. *piechart* : diagramme circulaire ou camembert

À partir de cette liste de questions (`LIST_QUEST`), les réponses sont générées avec la requête suivante :

> $\mathcal{P}_3$ `Réponds à chacune de ces questions` [LIST_QUEST] `grâce à ce texte` [ARTICLE] `et ces données` [DATA].

où `LIST_QUEST` est la liste de questions générées, `ARTICLE` l'article de presse et `DATA` les données issues de la visualisation de données.

Les différents prompts ont été définis de manière à obtenir des couples Q/A extractibles depuis le texte généré, sans pour autant faire l'objet d'un réel *prompt engineering*. Les performances de ce modèle dépendant fortement des différentes mises à jour [40], ces techniques comme les différentes expressions régulières utilisées ne sont pertinentes que pour une version précise, les rendant inopérantes sur les suivantes. De plus, utiliser des Q/As générés sans optimisation particulière permet d'éprouver la méthodologie *GenQA*, en tant qu'outil de validation.

3.3 Phase de validation

Deuxième partie de l'algorithme, cette phase de validation consiste à regrouper les Q/As générés lors de la phase de pré-traitement (Section 3.2) et à les relier au matériel journalistique fourni. L'ensemble des couples générés (nommé Q/A^*) constitue ainsi l'entrée de cette partie tandis qu'une organisation de ce même ensemble est retournée. Les paragraphes suivants présentent les différents sous-algorithmes utilisés, dans un ordre non-chronologique (Figure 3.4).

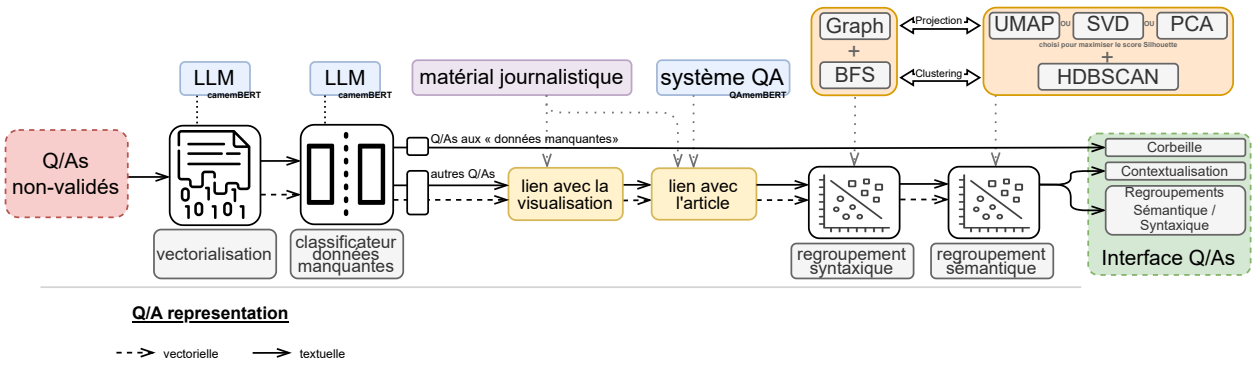


FIGURE 3.4 Algorithmes de regroupement

Dans un premier temps, les couples Q/A sont vectorialisés. Les couples de type « données manquantes » sont ensuite écartés du processus, menant ainsi au regroupement « Corbeille » (Section 3.3.2). Les couples de Q/As restants sont liés avec le matériel journalistique, formant notamment le regroupement « Contextualisation » (Section 3.3.1). Les paires de Questions/Réponses sont ensuite regroupées lexicalement (Section 3.3.2) et finalement sémantiquement (Section 3.3.2).

3.3.1 Algorithmes de liaison avec le matériel journaliste

Les deux algorithmes suivants ont pour objectif de créer un lien entre les paires de Questions/Réponses et le matériel journalistique. Ces liens ne sont pas créés lors de la génération des Q/As car ce dernier système peut halluciner. Ainsi, effectuer cette mise en relation par un second modèle neuronal renforce le degré de confiance quant à ce lien.

Le lien construit vise à renforcer le degré de confiance des journalistes en ces couples en mettant en avant leur possible provenance [36]. L'objectif est ici d'améliorer et d'accélérer la prise de décision du journaliste. Ils seront par la suite représentés comme aides à la décision que ce soit au sein de l'article ou de la visualisation.

Lien avec la visualisation de données

Pour chaque Q/A, les attributs associés sont déterminés par une recherche de sous-chaîne de caractères au sein de la question comme de la réponse. En cas de reformulation, un attribut est néanmoins associé grâce à une comparaison avec la représentation vectorielle des différents attributs. Ces attributs ont été définis par le journaliste avec la fenêtre « Complément d'information » (Section 3.2.2). La représentation vectorielle de ces attributs est fournie par la dernière couche du modèle francophone *camemBERT* [41]). Malgré cela, le degré de confiance accordé est inférieur à celui accordé par une recherche de caractères.

Un parcours des valeurs correspondant à ces attributs est par la suite effectué pour rechercher ces valeurs au format textuel au sein du Q/A considéré. Si cette dernière recherche est infructueuse, le couple de Question/Réponse est considéré comme n'ayant pas de lien avec la visualisation. Dans le cas contraire, on obtient ainsi un lien entre une paire attribut / valeur et la Q/A, et donc un lien à la visualisation.

Algorithm 1 Algorithme de liaison avec la visualisation de données

Require: $data$ ▷ Données numériques de la visualisation de données
Require: (qst, nsw) ▷ Couple courant (question, réponse)
Require: $attrs$ ▷ Liste d'attributs
  $links \leftarrow \emptyset$
  for all $row \in data$ **do**
    for all $(value, attr) \in (row, attrs)$ **do**
      if $value \in nsw$ AND ($attr \in nsw$ OR $attr \in qst$) **then**
        $links \leftarrow links \cup (attr, value)$
      end if
    end for
  end for

Cette recherche par sous-chaîne de caractères est rendue possible grâce aux propriétés du LLM employé pour générer les couples de Question/Réponse, ChatGPT3.4.

En effet, la similarité des couples générés permet d'appliquer les mêmes traitements à l'ensemble des Q/As. L'un d'entre eux consiste à adapter les dates aux formats "AAAA", "mois AAAA" ou "JJ mois AAAA", seuls formats utilisés en sortie du LLM. De plus, une stratégie spécifique a été mise en place pour gérer les regroupements par mois et par années. Par exemple, lorsqu'une année XXXX est présente dans un couple de Question/Réponse, la plage temporaire associée à mettre en avant dans la visualisation doit inclure tous les mois de janvier XXXX à décembre XXXX. Le même raisonnement est appliqué pour les jours.

Lien avec l'article de presse

À la suite de cette mise en relation avec la visualisation, un second algorithme est exécuté pour, cette fois, lier les Q/As à l'article de presse. Celui-ci consiste à déterminer au sein d'un texte, la potentielle origine d'un couple de Question/Réponse. Pour cela, à une question donnée, cette phase associe un élément de réponse issu de l'article de presses, pratique relevant ainsi du QA.

Ce type de modèle étant propre à chaque langage, le modèle *QAmemBERT* [42] a été utilisé pour la version française. Ce modèle a été entraîné à partir de données du type *SQuAD2.0*. Contrairement à sa version antérieure *SQuAD1.0*, le type d'ensemble *SQuAD2.0* prend en compte une possible absence de réponse au sein du texte, renforçant ainsi la pertinence d'un tel modèle dans notre contexte. En effet, le regroupement « Contextualisation » correspond aux couples qui ne sont ni associés à l'article, ni à la visualisation. La possibilité qu'un Q/A ne soit pas lié avec le visualisation de données est ainsi prise en compte. Un couple est ainsi considéré comme étant lié à l'article de presse si l'indice de confiance fourni par système QA est supérieur à un seuil fixé par expérience.

3.3.2 Algorithmes de regroupement

Conjointement avec la mise en relation des couples avec le matériel journalistique, les couples Q/A obtenus sont agrégés au sein de différents clusters afin d'améliorer l'expérience utilisateur de l'interface. Différents algorithmes de clustering ont ainsi été mis en place pour effectuer un regroupement lexical (regroupement des Q/As dont les questions n'ont que quelques mots de différence) et sémantique (regroupement des Q/As de sens proches).

Regroupement « Corbeille »

Classification réalisée juste après la génération des Q/As, cet algorithme vise à effectuer un premier tri de l'ensemble des couples de Question/Réponse. Il cible tout particulièrement les Q/As aux « données manquantes ». Il s'agit ici de mettre de côté, l'ensemble des couples dont les réponses indiquent un manque de données. En effet, certains des couples générés reflètent une absence de données dans la réponse. Quelques exemples sont listés ci-dessous :

- ➢ Existe-t-il une corrélation entre les jours spécifiques de la semaine et l'augmentation du nombre de morts ?
  - ➥ Il n'y a pas suffisamment de données fournies pour déterminer s'il existe une corrélation entre les jours spécifiques de la semaine et l'augmentation du nombre de morts. Une analyse plus approfondie pourrait être nécessaire pour identifier de telles tendances.
- ➢ Quel mois a enregistré la plus grande augmentation du taux directeur en 2023 ?
  - ➥ La plus grande augmentation du taux directeur en 2023 n'est pas spécifiée dans les données fournies.
- ➢ Y a-t-il une corrélation entre les déplacements de personnes et les changements climatiques dans des régions tropicales et subtropicales ?
  - ➥ Le texte ne fournit pas d'information directe sur la corrélation entre les déplacements de personnes et les changements climatiques dans les régions tropicales et subtropicales.
- ➢ Existe-t-il une corrélation entre les interceptions de Cubains et les événements politiques ou économiques à Cuba ?
  - ➥ Cette information n'est pas fournie dans les données ou le texte fourni.

Diverses stratégies ont été expérimentées pour réaliser cette division. Dans chaque cas, les ensembles d'entraînement et de test sont issus d'une partition stratifiée de même taille. Elles sont présentées dans la suite de cette partie. Après leur identification, les couples Q/A qualifiés de type « données manquantes » sont toujours disponibles depuis l'interface finale dans une catégorie propre (voir la figure 3.8), l'objectif étant de représenter l'ensemble des couples générés. Cette catégorie est cependant moins mise en évidence que les autres, de façon à favoriser les paires de Questions/Réponses potentiellement les plus pertinentes pour véhiculer le message du journaliste.

Néanmoins, quelque soit la stratégie employée, des ensembles de test et d'entraînement de même taille (soit 572 Q/As chacun) ont été créés à partir d'une annotation manuelle. Sur les 1144 couples générés pour l'expérimentation, 189 Q/As ont été labellisé comme ayant des « données manquantes », ce qui représente environ 16,5% de l'ensemble. Dans les rapports d'entrainement des modèles, les couples aux « données manquantes » sont représentés par la classe 1 tandis les autres couples le sont par la classe 0 .

Comparaison cosinus Les premières observations de ces couples aux « données manquantes » ont mis en avant la forte présence des formulations suivantes (nommées $\mathcal{S}_{naives}$ par la suite) au sein de leurs réponses :

- `Les données fournies ne permettent pas de répondre à la question`
- `Les données ne fournissent pas d'informations supplémentaires`
- `Les données fournies ne mentionnent pas`

Ainsi, la première approche a été d'utiliser une proximité sémantique entre les réponses et chacune des assertions listées ci-dessus. Cette approche naïve classifie ainsi un couple Q/A de « données manquantes » si sa réponse est « proche » des $\mathcal{S}_{naives}$. La notion de proximité est ici représentée par la similarité cosinus minimale entre la réponse d'un couple et chacune des $\mathcal{S}_{naives}$. Les représentations vectorielles sont issues de la dernière couche du modèle *camemBERT* [41].

Un entraînement de ce modèle a par la suite été effectué afin d'ajuster la valeur de coupure. En dessous de celle-ci, les couples seront considérés comme étant du type « données manquantes ». Pour cela, le F1-score, une métrique de la qualité d'une classification binaire par rapport à un ensemble labellisé, a été calculé pour différentes valeurs seuil (voir Figure 3.5).

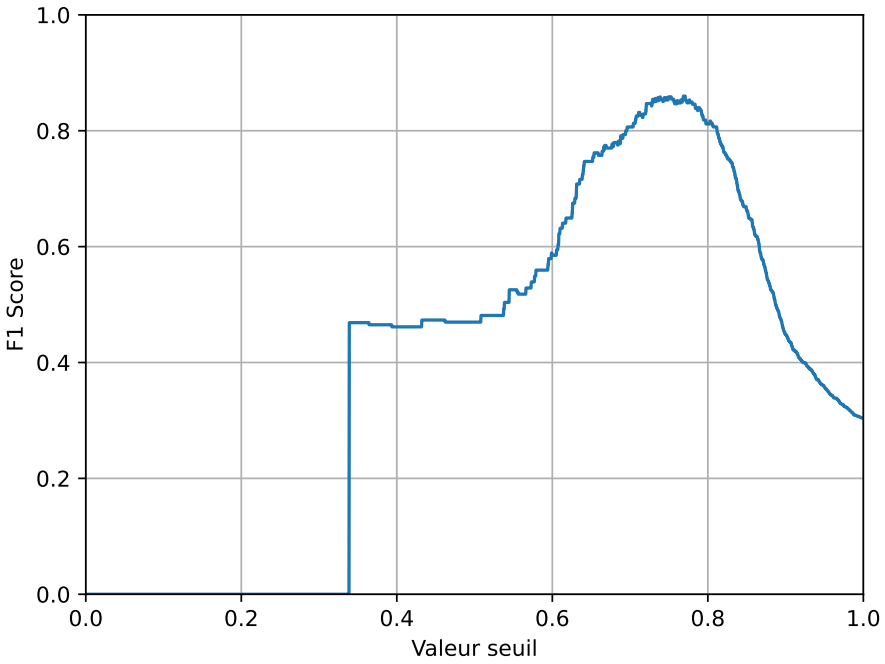


FIGURE 3.5 F1-score selon la valeur seuil sur l'ensemble d'entraînement

La valeur seuil a ainsi été ajusté à 0.768 afin de maximiser le F1-score (à 0.860) de l'ensemble d'entraînement. Cet ajustement de l'hyperparamètre permet d'obtenir un F1-score de 0.91 sur l'ensemble de test (voir Figure 3.6).

	précision	rappel	f1-score	support
0	0.97	0.96	0.97	477
1	0.83	0.87	0.85	95
exactitude			0.95	572
moyenne macro	0.90	0.92	0.91	572
moyenne pond.	0.95	0.95	0.95	572

FIGURE 3.6 Rapport de l'entraînement du modèle naïf

Classification par BERT Afin de mettre en regard les performances de ce premier modèle avec un modèle neuronal, une comparaison est effectuée avec un modèle de classificateur neuronal. Ce modèle-ci repose sur l'affinage d'un modèle BERT (*'camembert-base'*) dont l'entraînement a été réalisé avec les 572 Q/As de l'ensemble d'entraînement - *batches* de taille 8 avec 8 *epochs* -. Ce modèle a ensuite été évalué par le biais d'une *cross-validation*[6] (Figure 3.7). À cause du faible taux de Q/As d'entraînement de type « données manquantes », l'approche stratifiée a été utilisée (avec 5 *folds*), s'assurant ainsi de la même répartition des deux classes dans tous les sous-blocs de la validation croisée. Malgré ce faible nombre de Q/As donnés lors de l'affinage du modèle, ce dernier obtient un F1-score de 0.96. Cela peut s'expliquer par la nature du modèle et notamment la similarité de cette tâche de classification avec la tâche d'apprentissage *Next Sentence Prediction*[7]. Il s'agit en effet de déduire si, pour une réponse donnée, l'affirmation « Car les données ne permettent pas de répondre à la question. » est une suite logique de la réponse.

	précision	rappel	f1-score	support
0	0.99	0.99	0.99	481
1	0.93	0.93	0.93	91
exactitude			0.98	572
moyenne macro	0.96	0.96	0.96	572
moyenne pond.	0.98	0.98	0.98	572

FIGURE 3.7 Rapport de l'entraînement du modèle BERT

Cette dernière méthode ayant permis d'obtenir un F1-score supérieur d'environ 5% par rapport à la version naïve, elle est utilisée lors de la création du regroupement « Corbeille ». Cette classification par BERT permet en outre une meilleure représentation des couples aux « données manquantes ».

Regroupement lexical

Ce regroupement vise à réunir les ensembles des Q/As dont les formulations lexicales sont proches. Les clusters formés sont représentés par une unique question au sein de l'interface, offrant ainsi la possibilité de sélectionner l'ensemble des couples de Questions/Réponses sous-jacents d'un seul geste, optimisant le travail de validation du journaliste.

6. *cross-validation* : validation croisée
7. *Next Token Prediction* : tâche consistant à déduire la conséquence logique entre deux phrases

La formation de ces regroupements nécessite donc de disposer d'une mesure de proximité entre chaque couple de Question/Réponse, idéalement une distance sur l'ensemble des Q/As dont chaque question est unique. Pour cela, une variation de la distance de *Levenshtein* a été considérée. Contrairement à d'autres métriques comme ROUGE ou BLEU, une distance d'édition permet de mettre en avant les sections variables d'un texte à l'autre. Dans une optique de représentation des couples, ce distance permet donc de différencier la partie commune des Q/As d'un même regroupement lexical des parties variantes.

Contrairement aux distances d'édition usuellement exécutées sur les caractères, celle-ci utilise les mots comme atomes afin d'obtenir une distance sur l'ensemble des phrases, et par extension, sur l'ensemble Q/A^*. Elle est mathématiquement définie de la manière suivante :

$$\mathrm{dist}_{\mathrm{lev}}(a,b) = \begin{cases} \max(|a|,|b|) & \text{si } \min(|a|,|b|) = 0, \\ \mathrm{dist}_{\mathrm{lev}}(w^a_{1:|a|}, w^b_{1:|b|}) & \text{si } w^a_0 = w^b_0, \\ 1 + \min \begin{cases} \mathrm{dist}_{\mathrm{lev}}(w^a_{1:|a|}, b) \\ \mathrm{dist}_{\mathrm{lev}}(a, w^b_{1:|b|}) \\ \mathrm{dist}_{\mathrm{lev}}(w^a_{1:|a|}, w^b_{1:|b|}) \end{cases} & \text{sinon.} \end{cases}$$

où w^x_i est le i-ème mot de la phrase x

À partir de cette distance est construit un graphe non-orienté sur l'ensemble des couples avec une arête entre chaque élément de Q/A^* [8]. Formellement, ce graphe $G_{Lev} = (V, E)$ est défini de la façon suivante :

$$\begin{cases} V = Q/A^*, \\ E = \{(a,b) \in (Q/A^*)^2 \mid dist_{lev}(a_{quest}, b_{quest}) \leq seuil\} \end{cases}$$

Par la suite, un parcours en largeur est effectué afin d'extraire toutes les composantes connexes de G_{Lev}. Finalement, chacun de ces ensembles est étiqueté avec la question du couple le plus central, c'est-à-dire celui qui minimise la somme des distances aux autres couples du même ensemble. C'est ce dernier qui sera ainsi affiché sur le panneau central de l'interface.

Regroupement sémantique

Suite au regroupement lexical, un regroupement sémantique est mis en place. Contrairement au précédent, ce dernier vise à réunir les Q/As selon leur sémantique. Il est ainsi complémentaire aux regroupements lexicaux. L'objectif est ici de diminuer les ruptures de sujets trop importantes d'un couple à l'autre auprès des utilisateurs de l'interface.

8. Q/A^* : Ensemble des Q/As issus du processus de génération

Pour éviter toute hallucination du système, une approche basée sur une réduction de la dimensionnalité suivie d'un clustering est utilisée. Les algorithmes de réduction UMAP, SVD et PCA sont d'abord exécutés sur les représentations vectorielles des Q/A (dernière couche de *camemBERT* [41]). Ces algorithmes reposant sur différentes modalités, ils sont utilisés à des fins de comparaison pour ne finalement considérer que le meilleur modèle, dépendamment de l'ensemble Q/A^*. Suite à cela, l'algorithme de regroupement HDBSCAN est appliqué.

Parmi les trois regroupements obtenus (avec UMAP, SVD ou PCA), seul celui dont le score Silhouette est le plus élevé est finalement considéré. Cette métrique Silhouette constitue une mesure de la qualité du partitionnement prenant en compte à la fois de la cohésion intra-cluster et de la séparation inter-cluster. Pour chaque ensemble nouvellement formé, des textes descriptifs ont été générés par ChatGPT, via le prompt $\mathcal{P}_4$ afin de les qualifier. Graphiquement, ces descriptions sont accessibles par un survol du titre de chaque regroupement.

$\mathcal{P}_4$ `Résume l'information transmise par les couples de Questions/Réponses suivants en quelques mots :` [Q/A].

La création de ces regroupements n'est cependant pas appliquée pour les Q/As issus de l'ensemble « Corbeille » (Section 3.3.2). Ces couples de Question/Réponse sont ainsi affichés individuellement (Figure 3.8).

3.3.3 Design de l'interface

Une fois l'ensemble de ces regroupements effectués, l'information véhiculée par tous les couples doit être présentée auprès de l'utilisateur de l'interface. Pour cela, plusieurs techniques ont été développées afin de représenter l'ensemble des Q/As tout en permettant une sélection individuelle de chacun d'entre eux.

Vue générale de l'ensemble des couples de Questions/Réponses

Section principalement dédiée à la navigation, les différents gaufriers de la colonne de gauche permettent de représenter l'ensemble des couples Q/A générés. Il constitue ainsi une vue d'ensemble du système. L'état de validation des couples de Question/Réponse y est aussi représenté grâce à différentes nuances de gris (Figure 3.8).

Chaque gaufrier représente ainsi un regroupement sémantique tandis que les regroupements lexicaux sont illustrés par un indice au sein de quelques carrés.

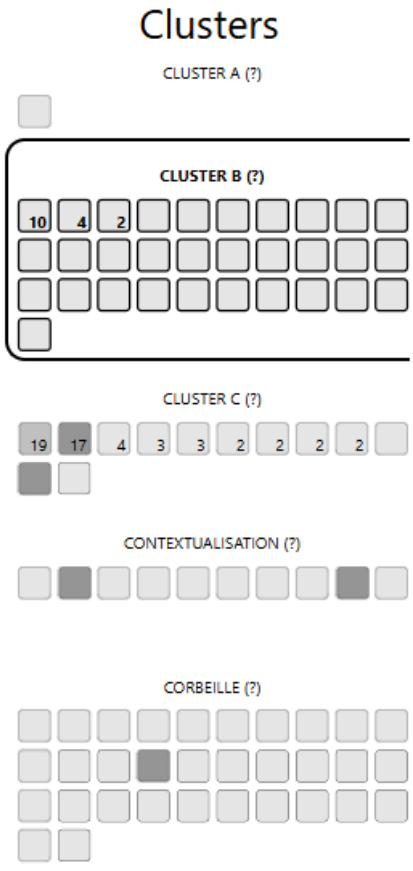


FIGURE 3.8 Vue d'ensemble des couples générés

Par exemple, le premier regroupement lexical du CLUSTER B de la Figure 3.8 réunit 10 couples de Question/Réponse. Le regroupement sémantique courant est, lui, encadré d'un liseré noir. À chaque titre est associée une infobulle, accessible par survol de la souris et résumant le contenu de chaque cluster.

Un ensemble de gaufriers a été choisi pour représenter ces données puisqu'il permet la navigation au sein de ces centaines de Q/As tout en les représentant dans leur l'intégralité. Il offre également une hiérarchisation avec différents niveaux : regroupements sémantiques, regroupements lexicaux et Q/As individuels. L'utilisation de gaufriers a été le fruit d'un long processus itératif. Initialement, une projection des Q/As sur un plan 2D avait été envisagé. Cette proposition n'a pas tété conservé du fait de l'absence de signification de la distance entre chacun de ces Q/As.

Représentation textuelle des couples de Questions/Réponses

Les couples de Questions/Réponses associés à un regroupement sémantique sont représentés sur la colonne centrale de l'interface. Parcourables grâce aux flèches directionnelles ainsi que via la souris (clic, double-clic et défilement), l'affichage de ces Q/As est scindé en deux parties. La partie supérieure, intitulée « `Couple(s) regroupé(s)` », concentre l'ensemble des regroupements lexicaux tandis que sous le label « `Couple(s) unique(s)` » sont présents les couples ne faisant partis d'aucuns de ces regroupements lexicaux.

Pour ces derniers, chaque couple de Q/A est directement représenté avec sa question et sa réponse. Le statut de validation y est représenté par un coche s'il est sélectionné (✔) ou une croix dans le cas contraire (✖). Une augmentation de l'épaisseur ainsi que de la noirceur de la police d'écriture utilisée met un peu plus en avant ces couples sélectionnés.

Affichés en haut du fait d'un degré de confiance plus important accordé à ces couples, les Q/As des regroupements lexicaux sont ainsi plus facilement accessibles et visibles par les utilisateurs que les autres couples. Un troisième état de validation y est également associé, représenté par un diamant (❖). Ceci correspond à la sélection de quelques-uns des couples contenus dans le regroupement lexical concerné (voir le regroupement *"Quelle année a vu le plus grand nombre d'immatriculations de voitures électriques au Québec ?"* de la Figure 3.9).

Initialement, seuls les labels des différents groupements lexicaux sont représentés sur l'interface (Figure 3.9). Pour chacun de ces labels, les mots variants, déterminés par la distance d'édition de *Levenshtein*, sont affichés en italique. Le nombre de Q/As au sein de chaque regroupement y est également présent.

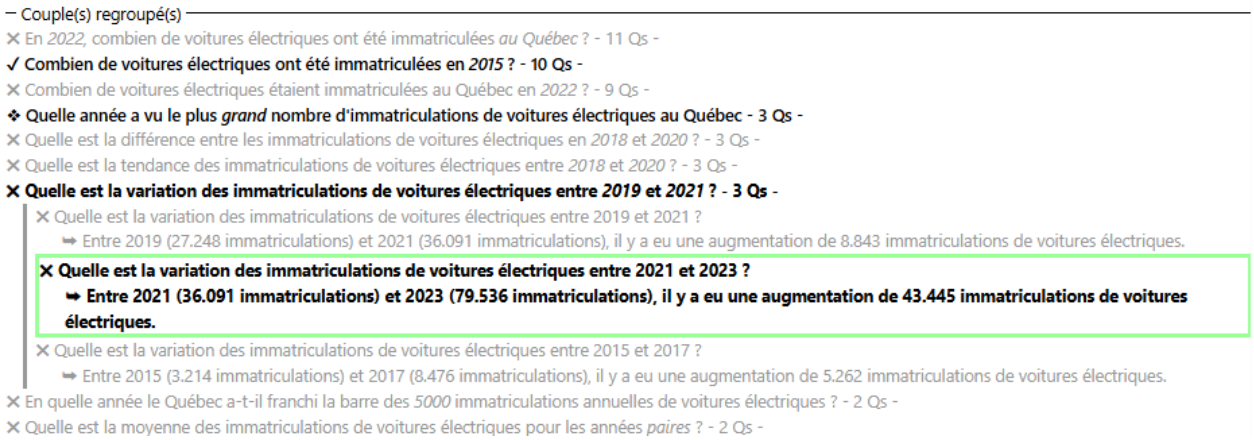


FIGURE 3.9 Représentation des regroupements lexicaux

Comme une prolongation de la navigation globale, le journaliste peut accéder à l'ensemble des couples sous-jacents (voir le regroupement *"Quelle est la variation des immatriculations de voitures électriques entre 2019 et 2021 ?"* de la Figure 3.9). Le couple courant reste ainsi cerclé de vert. L'utilisateur peut ainsi sélectionner un sous-couple individuellement ou considérer l'ensemble des Q/As d'un regroupement lexical.

Aides à la décision visuelles

Ces aides représentent une partie importante de ce projet. Représentées sur la colonne de droite, elles ont pour objectif d'accélérer la prise de décision de l'utilisateur tout en renforçant sa confiance envers le système. Elles permettent d'établir une correspondance entre chaque couple Q/A et l'information qu'il véhicule.

Leur intérêt est double. L'utilisateur peut utiliser cette aide pour vérifier la véracité d'un couple particulier en mettant en avant la section (plage de valeur ou extrait de l'article) la plus probablement liée à un Q/A individuel. Ces indices visuels peuvent aussi être à l'origine de la caractérisation d'ensembles spécifiques par la représentation d'une distribution d'aides à la décision (voir la figure 3.10).

FIGURE 3.10 Distribution d'aides à la décision

Grâce à cela, les utilisateurs peuvent ainsi cibler un intervalle ou un extrait d'article particulier pour, par exemple, obtenir une répartition uniforme des Q/As validés ou une représentation de phénomènes particuliers. Néanmoins, l'affichage des différents aides visuelles sur un même support nécessite un pré-traitement.

Cette représentation fait suite à une agrégation des différents intervalles à afficher. Du fait du chevauchement des intervalles, il n'est pas possible de les utiliser tel quel au sein de la visualisation ou de l'article. L'ensemble des intervalles à représenter est ainsi subdivisé de façon à obtenir une fréquence d'apparition constante pour chaque intervalle résultant. Cet algorithme est détaillé dans l'algorithme 2. La fréquence de couverture de chaque section de l'article et de la visualisation y est encodée par l'opacité.

Algorithm 2 Algorithme de construction des intervalles

Require: $sections$ ▷ Liste d'intervalles List((int, int))
  $sects \leftarrow []$
  for all $(begin, end) \in sections$ **do**
    $sects \leftarrow sects \cup (end, CLOSE)$
    $sects \leftarrow sects \cup (begin, OPEN)$
  end for
  $sects.sort()$ ▷ Tri inversé sur le premier élement
  $val_{prec} \leftarrow sects[0]$
  $freq \leftarrow 1$ ▷ Nombre d'apparition de l'intervalle $[val_{cur}, val_{prec}]$
  $segs \leftarrow []$
  for all $(val_{cur}, status_{cur}) \in sects$ **do**
    if $freq > 0$ **then**
      $segs \leftarrow segs \cup [\{"freq" : freq, "seq" : [val_{cur}, val_{prec}]\}]$
    end if
    if $status_{cur} == OPEN$ **then**
      $freq \leftarrow freq + 1$
    else
      $freq \leftarrow freq - 1$
    end if
    $val_{prec} \leftarrow val_{cur}$
  end for

Validation finale

Phase finale de la validation, un récapitulatif de l'ensemble des couples de Questions/Réponses validés est disponible. Cette partie permet une dernière validation des Q/As par l'utilisateur de l'interface. L'intégralité des couples jusqu'alors sélectionnés y est représentée indépendamment du regroupement lexical ou sémantique d'origine (Figure 3.11). L'utilisateur peut alors retirer certains de ces couples (touche ENTER ou SPACE). Les aides à la décision sont également présentes. Il s'agit de la même information que celle accessible par le biais de la touche R du clavier.

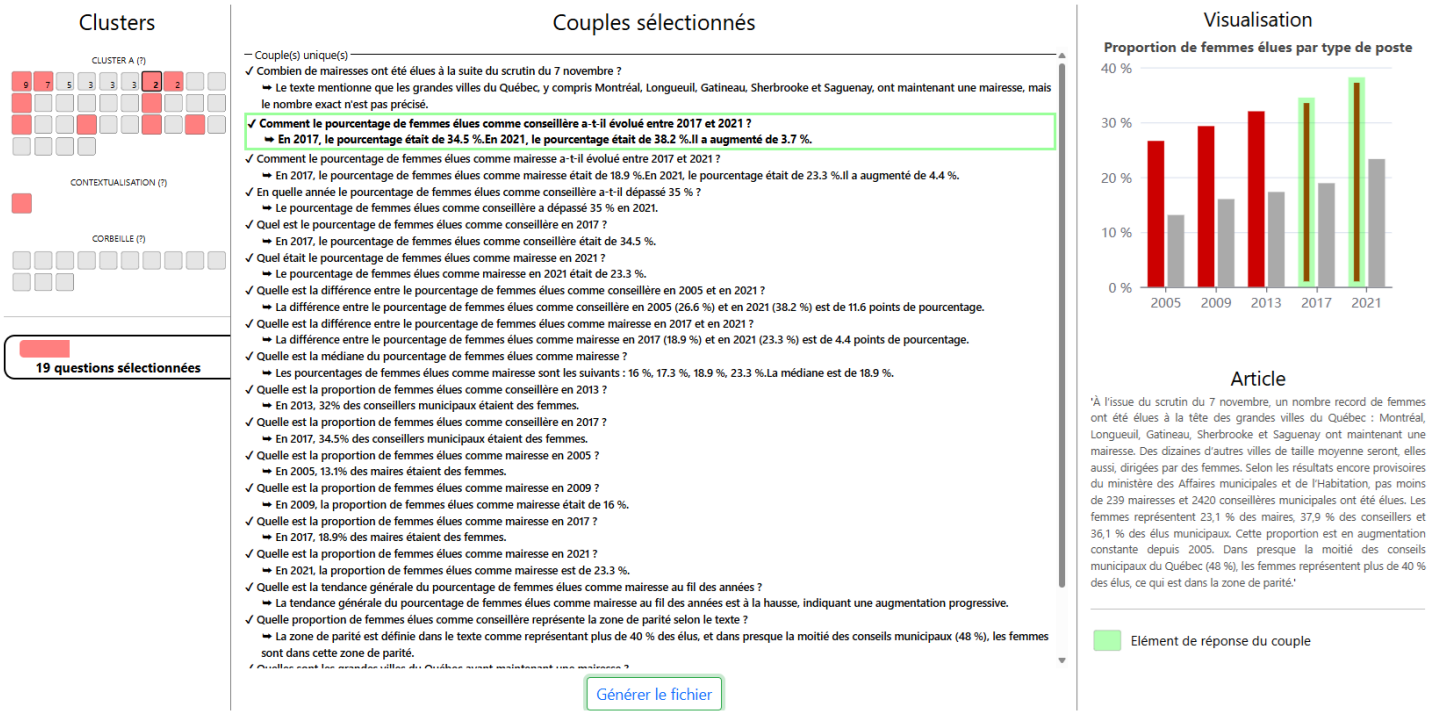


FIGURE 3.11 Phase finale du processus, avec les deux premières colonnes de l'interface

Interactions et navigation

Communication Une série de « Toasters » a été mise en place pour informer l'utilisateur de l'initialisation des différents serveurs et services utilisés ainsi que de la bonne création du fichier `.faq` contenant l'ensemble des Q/As validés.

Ces éléments visent tout particulièrement à vérifier les droits d'accès des utilisateurs à la visualisation et à l'article. Afin de respecter les différentes législations notamment sur le moissonnage des visualisations de données, l'extraction des données de la visualisation avec cette méthode est limitée aux utilisateurs disposant d'un token d'identification associé.

Navigation Afin de faciliter la navigation, trois niveaux ont été définis. Ils reposent notamment sur les regroupements lexicaux comme sémantiques. Passer d'un niveau à l'autre nécessite l'usage des flèches directionnelles GAUCHE et DROITE tandis que les flèches HAUT et BAS permettent la navigation au sein de chaque niveau. Le pointeur offre également la possibilité de passer d'un niveau à l'autre en ciblant des couples particuliers. La sélection / dé-sélection d'un couple ou d'un ensemble s'effectue par les touches SPACE ou ENTER.

Les touches R et G sont respectivement associées à l'ensemble des Q/As générés et à l'ensemble des couples jusqu'à présent sélectionné. Leur pression permet de représenter la répartition des aides à la décision associée. Ce choix a été fait de manière à passer rapidement d'un contexte local à un contexte global, et inversement.

- **Niveau sémantique** : Niveau le plus large de l'interface, l'utilisateur accède ici à chaque regroupement sémantique (contextualisation incluse) individuellement. La répartition de ces aides à la décision est ainsi disponible sur la colonne de droite tandis que les couples sont textuellement représentés au centre. Le journaliste peut ici sélectionner l'ensemble des couples inclus au sein d'un regroupement sémantique. Ce niveau permet également l'accès à l'ensemble des couples qui ont jusqu'à présent été sélectionnés.

- **Niveau lexical** : Niveau existant uniquement dans le cas d'un regroupement lexical, ces ensembles sont considérés individuellement. Il est ainsi possible de représenter les aides à la décision associées. Comme pour le niveau sémantique, l'utilisateur peut sélectionner l'ensemble des Q/As contenus dans un regroupement lexical. Il est combiné avec le niveau individuel des couples uniques.

- **Niveau individuel** : Dernier niveau, il considère chaque couple de Question/Réponse individuellement, permettant ainsi une sélection plus fine et précise. De même, l'aide à la décision associée à chaque Q/A est alors représentée.

3.4 Version anglophone

Avec un objectif de soumission d'un article portant sur la méthodologie *GenQA*, il a été nécessaire de généraliser l'interface à la langue anglaise. Ce projet a ainsi été rendu bilingue selon deux modalités. Pour ce faire, il y ainsi fallu traduire l'interface en tant que telle, ainsi que permettre la prise en compte d'article anglophone. Pour cette dernière, les modèles d'IA utilisés ont dû être adaptés.

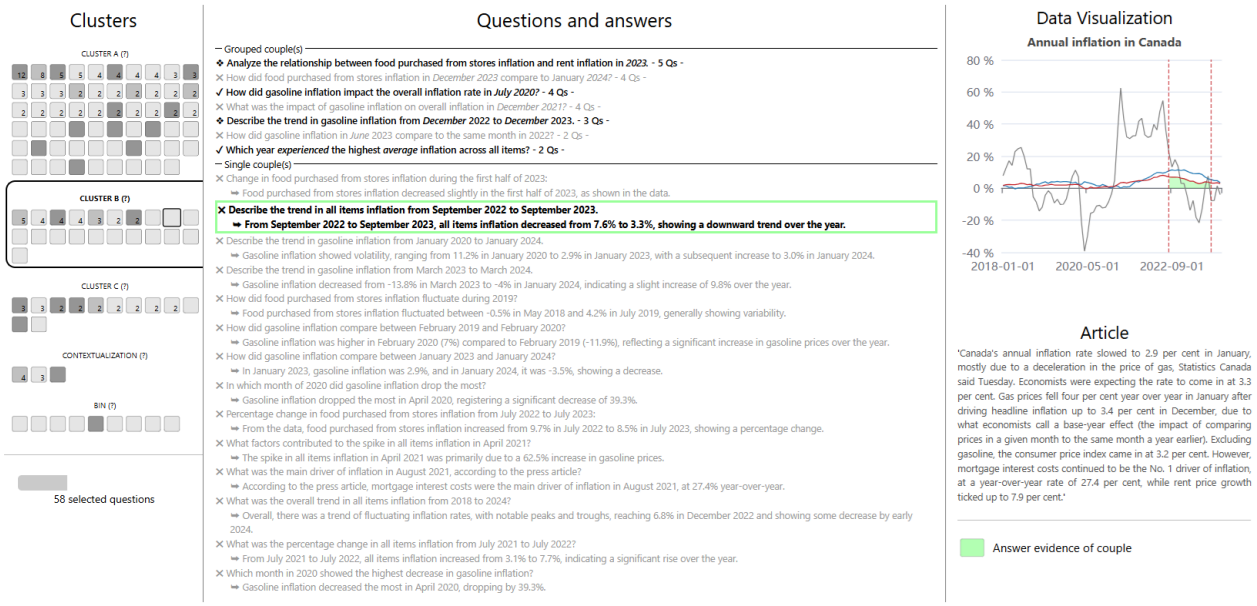


FIGURE 3.12 Interface générale anglophone

Ainsi, le lien entre les différents Q/As et l'article de presse (voir section 3.3.1) est réalisé avec un modèle ROBERTA entraîné sur un ensemble de type SQUAD2.0 (*roberta-base-squad2*).

Visualisation	Article de presse
05U1o	Canada's inflation rate slowed to 2.9% in January as gas prices fell

TABLEAU 3.1 Articles anglophones

En raison du nombre limité de couples de Q/As générés, il n'est pas possible d'entraîner un modèle de classification. Ainsi, le regroupement « Corbeille » des couples de la version anglophone est réalisé en utilisant l'approche de la similarité cosinus, comme décrit dans la section 3.3.2.

CHAPITRE 4 EXPÉRIMENTATION

Cette évaluation vise à mesurer l'intérêt d'une telle méthodologie du point de vue de la salle de rédaction (Hypothèse $\mathcal{H}_2$). Elle a également pour objectif d'évaluer la pertinence de l'interface pour effectuer une telle tâche de validation (Hypothèses $\mathcal{H}_1$ et $\mathcal{H}_3$). Cette étude a fait l'objet d'une accréditation de la part du Comité d'Éthique de la Recherche de Polytechnique Montréal (CER-2324-62-D).

Ces expérimentations ont été précédées par un ensemble de pré-tests plus informels réalisés auprès de doctorants et de post-doctorants de l'ENAC et d'enseignants en IA et visualisation de données. Celles-ci ont eu pour objectif de dimensionner les expérimentations suivantes ainsi que d'acquérir un premier retour sur le projet.

4.1 Recrutement des participants

Les trois participants ont été recrutés auprès de nos salles de rédaction partenaires - *Radio-Canada* et *LeDevoir* -. Une liste des participants potentiels a été établie suite à de précédentes expérimentations au sein du laboratoire.

		Profession	Age	Sexe	Niveau d'éducation	Expérience
Part.	A	Journaliste	30-40	H	Baccalauréat	5 ans
	B	Journaliste	40-50	F	Baccalauréat	12 ans
	C	Enseignant	30-40	H	Baccalauréat	4 ans

Tableau 4.1 Caractéristiques socio-démographiques des participants

L'ensemble de ces journalistes a été contacté via le courriel fourni en annexe B puis, le cas échéant, leur consentement a été formalisé par le formulaire de consentement disponible en annexe C. Afin d'augmenter le nombre de participants, un courriel a également été envoyé aux étudiants en journalisme de l'UQÀM, sans succès.

Idéalement, la répartition des articles auprès des différents participants devait être telle qu'indiquée dans la table 4.3. Celle-ci a pour objectif de balayer les articles de presse par l'ensemble des participants, permettant ainsi une mesure de l'impact de l'interface grâce à des mesures issues de différents participants.

Néanmoins, le nombre de participants étant faible et les temps de validation variables d'un journaliste à l'autre, il a fallu adapter ce cadre. C'est ainsi que le participant C n'a validé que deux articles tandis que le journaliste A en a effectué le triple (voir la Table 4.4).

4.2 Articles de presse utilisés

Un ensemble de six articles a été considéré, tous issus de la salle des nouvelles de *Radio-Canada*. Leur liste est présentée dans la Table 4.2. Ils balayent différents sujets allant de la politique à l'économie en passant par l'évolution du parc automobile, tout en couvrant plusieurs types de visualisation de données parmi les plus utilisés - *barchart*[1] et *linechart*[2] avec de simples ou multiples attributs -. Ils sont également de difficultés variables du fait de données portant sur des valeurs numériques ou catégorielles mais aussi des couples des Q/As générés portant sur des données plus ou moins agrégées - moyenne, variance, différence -.

Id	Article de presse
0	Après six mois de guerre, Gaza en cartes et en graphiques
I	Plus de femmes au pouvoir au Qc, mais pas encore de parité
II	Le parc automobile augmente au Qc, toutes catégories confondues
III	L'inflation fait du surplace à 6,9% en octobre au pays
IV	L'inflation a légèrement rebondi à 3,4% au Canada en décembre
V	Nombre record de personnes déplacées par les désastres climatiques
VI	Cuba : un exode prévisible

Tableau 4.2 Articles utilisés lors des tests-utilisateurs

1. *barchart* : diagramme en barre
2. *linechart* : diagramme en ligne

4.3 Protocole

Le protocole associé à cette étude est scindé en trois parties distinctes. La manipulation de l'interface, section centrale de ces tests, est précédée par une phase de prise en main tandis qu'un entretien semi-directif clôt ce protocole. L'interface *GenQA* a été mise en ligne afin de faciliter son accès aux participants de cette étude.

Prise en main Chaque expérimentation est menée individuellement et indépendamment des autres. Lors de la première phase de ce protocole, le participant prend en main l'interface sur l'article O, un exemple dédié (voir la Table 4.2). Une série de tâches correspondant à l'ensemble du processus y est réalisée - sélection de certains Q/As et sous-ensembles, affichage des aides à la décision et validation finale - afin de vérifier la bonne compréhension et utilisation de l'interface. Cette partie est également mise à profit pour informer le participant sur l'ensemble du processus *GenQA* et lui exposer les différents objectifs, en particulier, l'amélioration de l'accessibilité des personnes atteintes de déficience visuelle.

Manipulation de l'interface Suite à cette initiation, les participants effectuent les tâches de validation à partir de différents articles. Durant cette phase, ils peuvent avoir accès à l'ensemble des outils de l'interface (configuration `Avec interface` ou O) ou seulement la liste de l'ensemble des Q/As (configuration `Sans interface` ou S), sans les regroupements ni les aides à la décision. Réalisé en autonomie, seule une assistance technique est éventuellement fournie auprès des participants. Pour chaque article, le même ensemble de Q/As est utilisé par tous les participants afin de comparer les différents résultats obtenus. Les articles ont été attribués aux participants selon les tables suivantes : prévue (Table 4.3) et effective (Table 4.4).

		Participants					
		A	B	C	D	E	F
Articles	I	S	S	O	O	O	O
	II	O	S	S	O	O	O
	III	O	O	S	S	O	O
	IV	O	O	O	S	S	O
	V	O	O	O	O	S	S
	VI	S	O	O	O	O	S

Tableau 4.3 Répartition idéale Article x Participant

		Part.		
		A	B	C
Articles	I	O	O	-
	II	O	S	-
	III	O	-	S
	IV	S	O	-
	V	S	-	O
	VI	O	-	-

Tableau 4.4 Répartition effective Article x Participant

Lors d'une validation avec l'interface, le participant a accès à l'ensemble des outils de l'interface tandis que la validation sans interface n'utilise pas les différents regroupements ni les aides à la décision, de sorte que seule une liste de l'ensemble des couples est représentée. L'ensemble des couples Q/As validés sont collectés en vue d'analyses ultérieures. Chacune de ces validations est suivie d'une évaluation de la charge de travail du journaliste par le biais de la version française du *NASA Task Load Index (NASA-TLX)* [43].

Entretien final Un entretien semi-directif général (guide d'entretien en annexe E) est mené afin de collecter les impressions des participants. Il concerne notamment la pertinence des ensembles initiaux de Q/As, leur génération automatique ainsi que de l'intérêt de l'interface pour réaliser cette tâche.

Phase de l'expérimentation	Hypothèses		
	$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{H}_3$
Tâche de validation par le participant	✓		
Test NASA-TLX	✓		✓
Entretien semi-directif	✓	✓	✓

Tableau 4.5 Association des phases de l'expérimentation avec les hypothèses formulées

4.3.1 NASA-TLX

Ce test, sous forme d'un questionnaire normalisé, évalue la charge de travail grâce à une échelle discrète allant de 0 à 11 et un ordonnancement des six dimensions suivantes afin d'établir un profil caractérisant l'interface :

- exigence mentale,
- exigence temporelle,
- effort,
- exigence physique,
- performance,
- frustration

À chacune des ces dimensions, une série de questions est associée :

Exigence mentale

- Quelle activité mentale et perceptive était requise (par exemple, penser, décider, calculer, se souvenir, regarder, chercher, etc.) ?
- La tâche était-elle facile ou exigeante, simple ou complexe, indulgente ou exigeante ?

Exigence physique

- Quelle quantité d'activité physique était requise (ex : tourner, activer, etc.) ?
- La tâche était-elle facile ou exigeante, lente ou rapide, détendue ou exténuante, reposante ou laborieuse ?

Exigence temporelle

- Quelle pression temporelle avez-vous ressentie en raison de la vitesse ou du rythme auquel les tâches ou les éléments de la tâche se sont déroulés ?
- La tâche était-elle facile ou exigeante, simple ou complexe, indulgente ou exigeante ?

Performance

- ans quelle mesure pensez-vous avoir réussi à atteindre les objectifs de la tâche fixés par l'expérimentateur (ou vous-même) ?
- Dans quelle mesure avez-vous été satisfait de votre performance dans l'accomplissement de ces objectifs ?

Effort

- Dans quelle mesure avez-vous dû travailler (mentalement et physiquement) pour atteindre votre niveau de performance ?

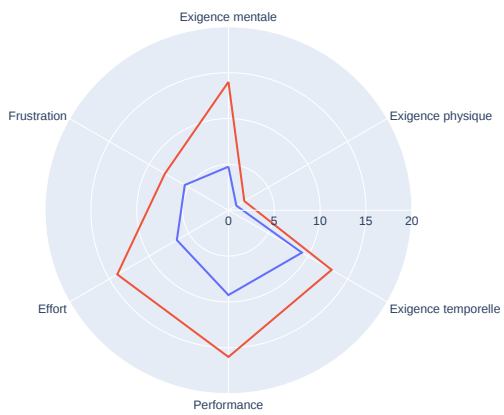
Frustration

- Dans quelle mesure avez-vous ressenti de l'insécurité, du découragement, de l'irritation, du stress et de l'agacement par rapport à de la sécurité, de la satisfaction, du contentement, de la détente et de la complaisance pendant la tâche ?

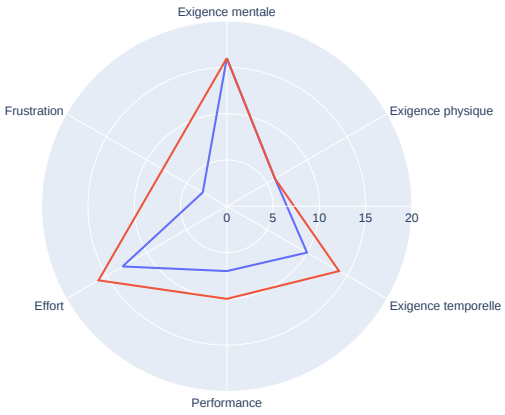
Principalement utilisé pour le design UX, le score TLX permet ainsi d'évaluer la charge de travail, mais surtout de la comparer à une autre tâche effectuée par le même participant. Les six dimensions peuvent aussi être utilisées afin de dresser un profil partiel de la charge de travail, toujours en vue d'une comparaison. Le formulaire utilisé lors des tests-utilisateurs est disponible en annexe D. Pour cette étude, ce score permet de comparer l'impact de l'interface pour un même participant.

4.4 Observations

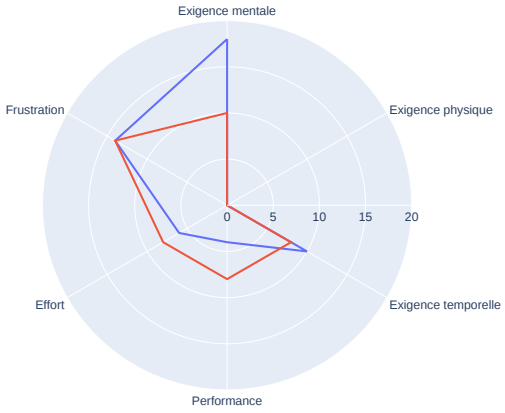
Cette étude a permis de collecter un ensemble de valeurs numériques lié au test NASA-TLX, au temps de validation ainsi qu'au nombre de couples validés. Ces résultats, agrégés au sein des figures suivantes, sont également disponibles en annexe de façon détaillée (Annexe A).



(a) Participant A



(b) Participant B

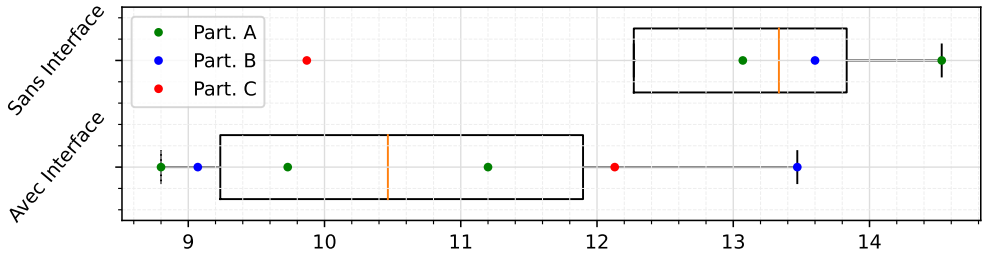


(c) Participant C

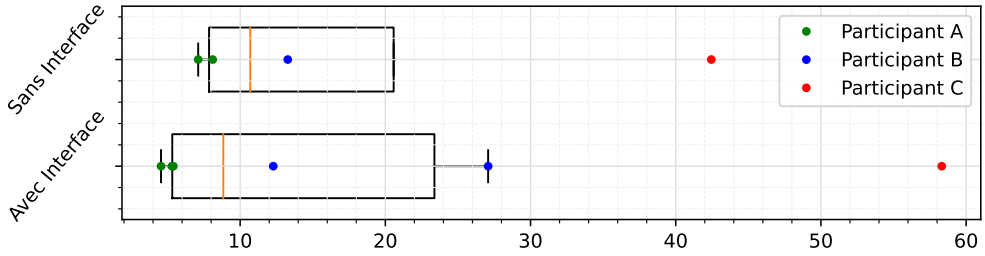
FIGURE 4.1 Résultats agrégés des dimensions du test TLX

Ligne rouge : sans interface / Ligne bleue : avec interface

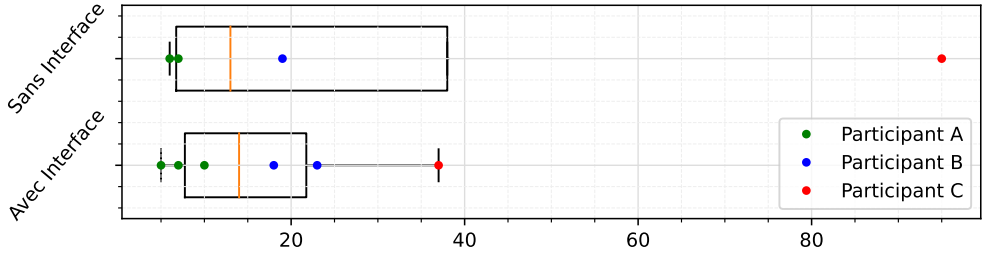
Diminution de la charge de travail Pour chaque dimension considérée par le test NASA-TLX, l'interface permet de diminuer son importance, conduisant également à un score TLX plus faible (Figure 4.1). Néanmoins, ce constat n'est pas partagé par le participant C. Du fait d'un nombre important de Q/As validés, s'assurer l'unicité sémantique des Q/As conduit à une charge mentale plus importante pour le participant C que pour les autres. D'autres faits ont été observés comme l'apparition de profils utilisateur de cette interface.



(a) Charge de travail



(b) Temps de validation



(c) Nombre de couples validés

FIGURE 4.2 Résultats agrégés des différentes métriques

Autres métriques En plus d'une diminution globale de la charge de travail 4.2a, une diminution du temps de validation a également été observé (Figure 4.2b). Malgré cette charge de travail moins importante et cette diminution du temps de validation, une légère augmentation du nombre de Q/As validés.

4.4.1 Deux profils distincts de journalistes

Les observations réalisées sur le temps de validation (Figure 4.2b) et sur le nombre de couples validés (Figure 4.2c) ainsi que l'observation directe de l'utilisation de l'interface ont permis de faire émerger deux profils d'utilisateurs différents. Cette distinction s'effectue principalement selon le nombre de couples validés.

Sélection d'un nombre restreint de couples

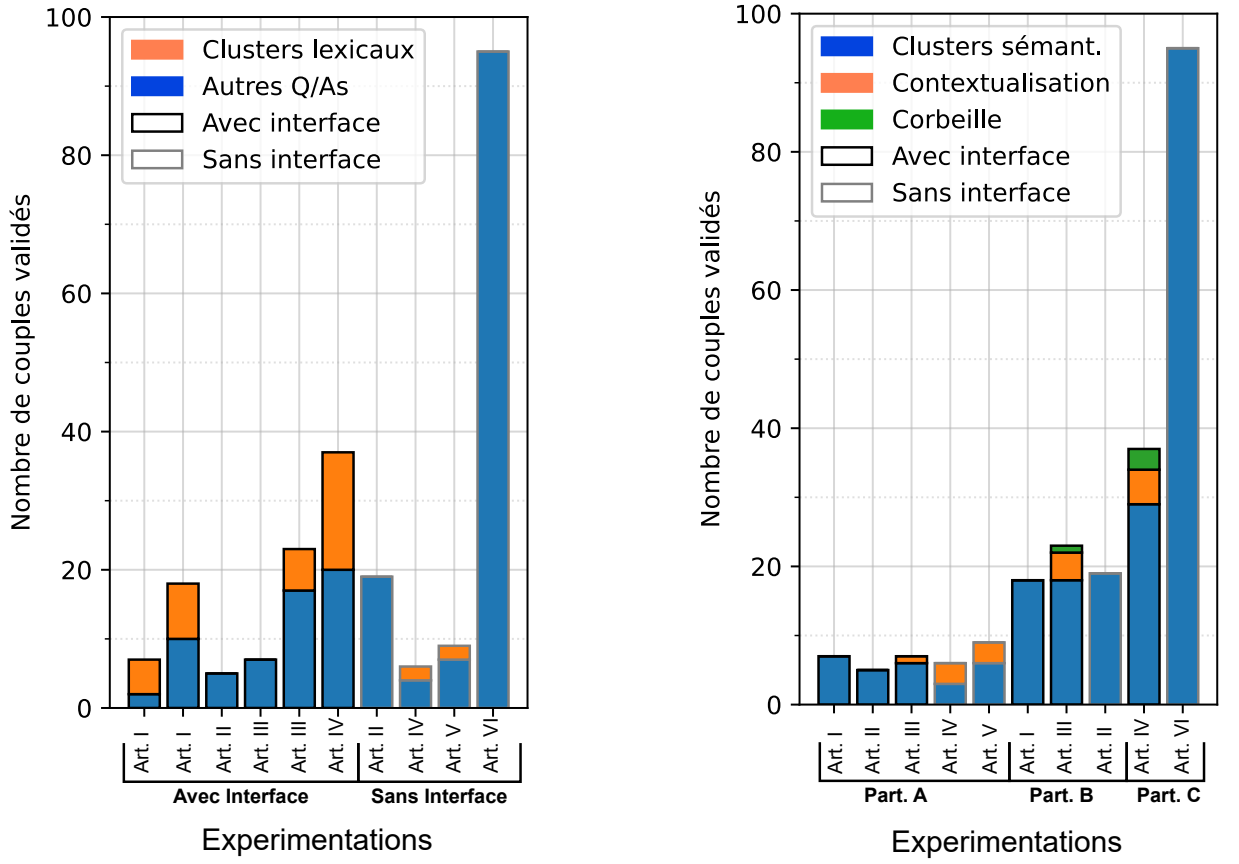
Le journaliste pioche uniquement une petite dizaine de couples sans prendre connaissance de l'ensemble des Q/As. L'utilisateur se focalise ici sur certaines questions et sujets précis, notamment grâce à l'emploi des aides à la décision et des différents regroupements. Cela a pour conséquence de renforcer l'impact de l'interface sur la tâche de validation, avec une diminution de la charge de travail d'environ 20% (Figure 4.2a). Ainsi, lors de l'expérimentation, le journaliste A a tout particulièrement ciblé des couples portant sur une section particulière de l'article ou de la visualisation grâce aux aides à la décision. La validation y est donc relativement rapide, autour des cinq minutes. Malgré cela, le temps de validation pour chacun des couple est relativement faible, de l'ordre de 0.7s par Q/A.

Sélection d'un nombre plus important de couples

Ce second profil se base sur une vérification systématique de l'intérêt et de la véracité des couples générés. Il consiste à valider un nombre plus important de couples, supérieure à la vingtaine. Pour ce faire, l'ensemble des couples est parcouru et leur véracité vérifiée individuellement. Profil des utilisateurs B et C, ce profil conduit à des temps de validation plus conséquents, supérieurs à la dizaine de minutes (Figure 4.2b) avec une moyenne de 1.25s par Q/A validé. Du fait de cette approche systématique, l'impact de l'interface est moindre pour cette tâche de validation. Son impact sur la charge de travail y est ainsi négligeable (Figure 4.2a). Ceci est particulièrement visible pour le participant C. Pour ce dernier, la vérification de l'unicité sémantique de chaque Q/A a entraîné une forte augmentation de la charge de travail ainsi que du temps de validation. De plus, le passage sur l'ensemble des couples générés entraîne une caducité des regroupements en tant que facteur d'accélération de la validation.

4.4.2 Outils de l'interface

En plus de ces deux profils distincts, plusieurs autres observations ont également été faites sur l'utilisation de l'interface, notamment l'usage des regroupements proposés ainsi que des aides à la décision disponibles. Tous ces outils ont permis de ne valider que des couples lexicalement et sémantiquement corrects. Ainsi aucun des couples validés ne contient d'hallucination, ce qui est l'un des objectifs de l'interface.



(a) Regroupements lexicaux et autres

(b) Regroupements sémantiques et associés

FIGURE 4.3 Répartition des couples validés selon les regroupements d'origine

Aides à la décision

L'ensemble des participants a utilisé ces aides visuelles à la décision. Elles ont été utilisées afin de caractériser les différents regroupements (participant B) mais aussi pour cibler certains couples de Question/Réponses (participant A).

Regroupements lexicaux

L'ensemble des journalistes ont mis en avant leur intérêt pour les regroupements lexicaux lors de l'entretien final, leur permettant ainsi de ne prendre en compte que les « *augmentations ou diminutions les plus significatives* » ou la « *statistique la plus récente* ». En effet, dans certains cas, regrouper les Q/As selon leur proximité textuelle a permis un même sujet avec différentes dates. La figure 4.3a illustre cet intérêt. Ainsi, contrairement à l'objectif initial qui était de sélectionner un regroupement lexical au complet, les journalistes A et B ont détourné son usage afin de ne prendre en compte que quelques couples particuliers. De plus, la prise en compte des couples de Q/As issus des regroupements lexicaux semble liée au nombre total de Q/As validés. Néanmoins, pour le participant C ces regroupements sont des « *source de frustration* », complexifiant la vérification de l'unicité sémantique des questions des Q/As validés.

Regroupements sémantiques et associés

La figure 4.3b met en avant un intérêt pour les regroupements « Corbeille » et « Contextualisation ». Leur sous-représentation dans les Q/As validés suggère une bonne discrimination de ces couples. En effet, lors des tests réalisés sans interface, le nombre de couples associés au regroupement « Corbeille » et « Contextualisation » est marginal. Ces observations confirment donc l'intérêt de la mise en avant des Q/As inclus dans des regroupements sémantiques par rapport aux autres couples.

De plus, on peut noter une augmentation de la proportion des couples ayant comme origine les regroupements « Corbeille » ou « Contextualisation » grâce à l'utilisation de l'interface. Ainsi, son emploi permet d'obtenir une meilleure contextualisation, défaut majeur des textes alternatifs actuels.

Tous les participants ont souligné leur intérêt pour le regroupement des Q/As en différentes catégories (hypothèse $\mathcal{H}_1$). Toutefois, certains participants plébiscitent les regroupements sémantiques au détriment des regroupements lexicaux tandis que d'autres mettent en avant le contraire. Finalement, l'absence d'utilisation par le participant C des regroupements et son temps de validation important mettent en avant l'impact de ces regroupements sur le temps de réalisation de la tâche.

4.4.3 Méthodologie globale

Au-delà de l'interface en elle-même, une partie de l'étude porte sur l'intégralité du processus *GenQA*. Des entretiens semi-directifs finaux ont ainsi ressortis les thématiques suivantes :

Usage de l'Intelligence Artificielle Cette étude a révélé un fort intérêt pour l'intégration de l'Intelligence Artificielle au sein d'une salle des nouvelles. Cependant, cet enthousiasme est tempéré par des préoccupations concernant la charge de validation qui pourrait être trop lourde et conséquente. Il s'agit donc davantage d'une curiosité que d'un véritable engouement. En effet, seul le participant A semble disposé à adopter cette méthodologie de manière quotidienne et professionnelle. Néanmoins, son usage s'accompagne de la crainte d'une lourdeur trop importante lors de la tâche de contre-vérification de l'information, impérative dans ce contexte journalistique.

Redondance du travail L'intégration de l'Intelligence Artificielle suscite également des préoccupations quant à la redondance dans le travail des journalistes. Le participant B exprime notamment des inquiétudes concernant les visualisations de données sans contexte significatif. Il craint que la tâche de contre-vérification des Q/As soit alors similaire au processus qui a conduit à rédiger l'article de presse à partir du graphique. Afin de limiter cet effet tout en limitant le risque de biais, la tâche de validation peut être effectuée par un journaliste autre que l'auteur de l'article considéré.

4.4.4 Qualité des Questions/Réponses générées

Cette étude a mis en avant la sérendipité et la découvrabilité propres à la génération des Q/As. La majorité des participants a ainsi admis qu'il n'aurait pas pensé à certaines questions sans cet outil. Le participant C a ainsi qualifié les questions de « *bonnes questions, [ou] proches d'être bonnes* » (hypothèse $\mathcal{H}_2$). Nuançant ces propos, le participant B a cependant souligné des questions générées trop générales ainsi qu'une faible contextualisation. Le participant A, quant à lui, met en avant une perte de pertinence des Q/As lorsque la « *temporalité est loin de l'aujourd'hui* ».

De plus, certains de ces couples ont fait l'objet d'hallucinations. C'est ainsi qu'un certain nombre des couples utilisés lors de cette étude sont incorrects. Ces derniers peuvent être catégorisés selon le type d'erreur. Le tableau suivant (Table 4.6) regroupe l'ensemble des erreurs observées pour les articles utilisés lors des expérimentations. Ces statistiques ont été obtenues suite à une annotation manuelle de chacun de ces Q/As.

Type d'erreur	Répartition	
	Brute	% des Q/As
Absence de réponse malgré l'existence des données	64	5.59
Question extrapolée	73	6.38
Réponse à la question incorrecte	9	0.79
Formulation incorrecte ou ne faisant que peu de sens	12	1.05
Autres types	12	1.05

Tableau 4.6 Répartition des erreurs de génération des couples

L'ensemble des erreurs ici recensées restent donc marginal par rapport à l'ensemble des couples Q/As générés ce qui tend à valider l'hypothèse $\mathcal{H}_2$.

Les erreurs observées sont caractérisables de manière plus précises, notamment par le biais d'exemples. La suite de cette section présentent ainsi quelques uns de ces couples afin d'illustrer les différents types d'erreurs rencontrés par les participants. La majorité est issue de l'article V dont les données tabulaires sont fournies par la table 4.7, et plus marginalement, de l'article IV.

Attributs	Valeurs (en millions)									
Année	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Personnes déplacées	15.0	30.1	22.1	19.1	19.2	24.2	18.8	17.2	24.9	30.7

Tableau 4.7 Données extraites de la visualisation de données V

Absence de réponse malgré l'existence des données

Malgré la présence des données permettant de répondre à la question, la réponse est du type « données manquantes » (voir partie 3.3.2). Les deux exemples de Q/As suivants illustrent ce phénomène puisque les données portaient sur la période 2011-2020, incluant notamment 2019 (article V) :

- ➢ En 2019, quel était le pourcentage d'augmentation du nombre de personnes déplacées ?
  - ➥ En 2019, le pourcentage d'augmentation du nombre de personnes déplacées par rapport à l'année précédente n'est pas fourni dans les données fournies.
- ➢ Quelle année a enregistré la plus forte augmentation par rapport à l'année précédente avant 2017 ?
  - ➥ Les données ne sont pas fournies pour répondre à cette question spécifique.

Question extrapolée

La question générée porte sur des données qui n'existent pas. Généralement observé comme le prolongement d'une série temporelle, il n'y a aucune réponse possible. Cela est notamment le cas avec l'article IV. Par exemple, le couple de Question/Réponse suivant porte sur l'année 2025 alors que les données de cette article s'échelonne sur la période janvier 2021 - janvier 2024 :

- ➢ Quel mois a enregistré le taux d'inflation le plus élevé en 2025 ?
  - ➥ D'après le tableau, le mois avec le taux d'inflation le plus élevé en 2025 était le mois de juin, avec un taux d'inflation de 8.1%.

Réponse à la question incorrecte

Si le texte généré est lexicalement et sémantiquement correct, la réponse fournie est en contradiction avec les données. Ainsi, selon l'article V, le minimum de personnes déplacées est atteinte en 2011 tandis que l'accroissement le plus important a été constaté pour l'année 2012 avec une augmentation de 200% ce qui est contradictoire avec les deux couples générés suivants :

- ➢ Y a-t-il une année où le nombre de personnes déplacées a atteint son point le plus bas ?
  - ➥ En 2017, le nombre de personnes déplacées a atteint son point le plus bas au cours de cette période.
- ➢ Quelle année a enregistré le pourcentage le plus élevé d'augmentation du nombre de personnes déplacées ?
  - ➥ L'année qui a enregistré le pourcentage le plus élevé d'augmentation du nombre de personnes déplacées est 2020, avec une augmentation de 23.29%.

Formulation incorrecte ou ne faisant que peu de sens

Contrairement aux précédentes catégories, le Q/A généré, que ce soit la question et/ou la réponse, est sémantiquement fausse. Par exemple, le premier exemple suivant met en avant une possible corrélation entre un phénomène et une organisation, ce qui ne fait pas de sens. Quant au second exemple, la formulation "variation moyenne annuelle" ne permet pas de définir de quelle métrique il s'agit :

- ➢ Y a-t-il une corrélation entre les déplacements de personnes et l'Organisation internationale des migrations ?
  - ➥ Oui, le texte indique que l'Organisation internationale des migrations (OIM) s'inquiète de l'augmentation de la température dans des zones très chaudes, ce qui pourrait entraîner des migrations importantes.
- ➢ Quelle est la variation annuelle moyenne des personnes déplacées au cours de cette période ?
  - ➥ La variation annuelle moyenne des personnes déplacées au cours de cette période est de 1.88 millions.

4.4.5 Limitations de l'étude

Cette étude ne porte que sur un nombre restreint de participants et de couples de Question/Réponse. Toutes les observations effectuées ne sont ainsi que difficilement généralisables. Des tests-utilisateurs seront ainsi nécessaires pour confirmer ces résultats avec une plus grande certitude.

De plus, le faible nombre de couples validés par les journalistes A et B - 17 Q/As en moyenne par article - semble dénoter une mauvaise compréhension de la problématique originelle et/ou des objectifs visés. L'accessibilité des personnes atteintes de déficience visuelle étant la principale cible, le nombre de couples sélectionnés devrait, a priori, être plus important. En effet, ce public cherche à acquérir le plus d'information possible [6] sur ce qui ne peuvent pas percevoir et tout particulièrement sur les données de la visualisation [17]. Par conséquent, idéalement, une grande partie des couples véridiques relatifs aux données devraient être inclus dans l'ensemble validé final.

Cela s'oppose au point de vue des participants, qui semble être celui d'un utilisateur voyant. Ainsi pour le participant B, « *un graphique peut se résumer [...] en maximum 5-6 questions. Ça me semble efficace et pertinent à la fois.* ».

Ce constat peut s'expliquer par une mauvaise compréhension du sujet, ou plus largement, une méconnaissance des attentes des publics atteints de déficience.

4.4.6 Synthèse des observations

Ces tests-utilisateurs ont ainsi mis en évidence une diminution globale de la charge de travail sur la tâche de validation grâce à l'usage de l'interface (hypothèses $\mathcal{H}_1$ et $\mathcal{H}_3$). Si le processus et l'interface sont appréciés par les participants dans leur globalité, différentes interrogations ont été émises sur la qualité des couples générés.

Les participants ont reporté une forme de séréndipité des Q/As générées, bien que ces couples puissent sembler trop généraux pour les utilisateurs finaux atteints de déficience visuelle. Par conséquent, l'hypothèse $\mathcal{H}_2$ n'est que partiellement confirmée. En revanche, les regroupements proposés des Q/As ont été largement approuvés par tous les participants lorsqu'ils ont été utilisés, validant ainsi l'hypothèse $\mathcal{H}_1$. Diverses améliorations ont néanmoins été suggérées, telles que l'implémentation d'une recherche par terme ou la possibilité de modifier l'ordre d'affichage des paires sur l'interface. Quant à l'hypothèse $\mathcal{H}_3$, elle a été confirmée par l'usage significatif des aides à la décision durant tous les tests.

De plus, une limite de la connaissance des attentes des personnes atteintes de déficience visuelle en termes d'accessibilité et/ou de l'objectif visé semble exister auprès des journalistes interrogés.

CHAPITRE 5 CONCLUSION

Les visualisations de données sont au cœur de la communication, comme dans les articles de presse en ligne. Il importe donc que ces graphiques puissent être accessibles à un plus large nombre et notamment aux personnes atteintes de déficience visuelle. Face à la déficience d'accessibilité actuelle, l'un des moyens est de faciliter la génération de ces descriptions par les journalistes en établissant une nouvelle méthodologie.

5.1 Synthèse des travaux

La méthodologie proposée dans ce mémoire repose sur l'utilisation de couples de Questions/Réponses générés par l'Intelligence Artificielle pour véhiculer l'information du matériel journalistique. Ce processus devant se conforter à l'éthique journalistique, une phase de validation est nécessaire avant toute publication. Celle-ci est réalisée par le biais d'une interface. Des aides visuelles sont disponibles afin de vérifier la véracité de l'information tandis que les différents regroupements des Questions/Réponses visent à limiter la rupture de sujets entre chaque couple. Ces différentes fonctionnalités ont pour but de rendre la tâche de vérification de ces couples fiable tout en étant rapide.

Afin de valider l'utilisabilité d'une telle interface ainsi que la méthodologie dans son intégralité, une évaluation a été réalisée par quelques journalistes du Devoir et de Radio-Canada. Cette expérimentation visait à effectuer le travail de validation de ces couples de Questions/Réponses pour un ensemble d'articles. Ce protocole a eu pour objectif d'évaluer les trois hypothèses de recherche suivantes :

- $\mathcal{H}_1$: les regroupements des Q/As mis en place améliorent la navigation et facilite la tâche de sélection.
- $\mathcal{H}_2$: la génération automatique des Q/As est de bonne qualité et permet une forme de sérendipité.
- $\mathcal{H}_3$: les aides à la décision accélèrent la prise de décision et renforce la confiance au système.

Les observations montrent que, lorsqu'ils sont utilisés, les regroupements des couples de Questions/Réponses permettent d'optimiser le processus de validation en diminuant le temps de réalisation. Cela valide ainsi l'hypothèse 1 ($\mathcal{H}_1$). Ces expériences ont également mis en évidence un recours massif aux aides à la décision lors de la réalisation de cette tâche, validant l'hypothèse 3 ($\mathcal{H}_3$). Si la sérendipité est soulignée par l'ensemble des participants, les avis sont relativement partagés sur la qualité des couples de Questions/Réponses générés. L'hypothèse 2 ($\mathcal{H}_2$) n'est ainsi que partiellement validée.

5.2 Améliorations futures

Ce mémoire présente une nouvelle méthodologie visant à améliorer l'accessibilité des visualisations de données. Toutefois, en raison de sa nature conceptuelle, plusieurs aspects pourraient être améliorés, notamment lors de la génération des Q/As et de leurs regroupements.

5.2.1 LLM utilisé

Lors de la rédaction de ce mémoire, ce processus de génération est réalisé avec le modèle *ChatGPT-3.4*. Néanmoins, son utilisation soulève une multitude de questions aussi bien éthiques qu'environnementales. De nombreuses controverses existent quant à la confidentialité des données et à l'entraînement d'une telle architecture. Du fait de la spécificité de notre tâche de génération, un modèle de taille plus modeste, entraîné spécifiquement sur ces tâches, devra être envisagé avant une quelconque diffusion. On pourrait ainsi s'intéresser à des modèles concurrents tels que *Llama3.1* [44]. Ce dernier possède le double avantage d'être exécutable en local (et ainsi de garder le contrôle sur les données) et open-source.

5.2.2 Généralisation de l'implémentation

De nombreuses restrictions existent quant à l'usage de cette interface. Du fait de cette preuve de concept, seuls les articles avec des visualisations de données générées par l'API *DataWrapper* sont pris en compte. C'est pourquoi, pour pérenniser son utilisation, un plus grand nombre de types de visualisations de données et de format doit être permis. Cela pourrait passer par l'utilisation d'autres librairies et outils.

5.2.3 Études supplémentaires

Les expérimentations actuelles concernant uniquement des rédactions québécoises francophones de Montréal, une étude avec un panel plus large et varié permettrait d'obtenir des résultats plus pertinents sur l'utilisation de l'interface puisque les performances des différents LLMs et algorithmes de regroupement varient grandement selon la langue considérée. Un nombre plus important de participants aurait également comme objectif de mener une étude quantitative sur l'usage de l'interface. Une seconde étude, sur la pertinence de l'utilisation de couples Q/As auprès des utilisateurs finaux pourra également être menée.

5.2.4 Prompts et entraînements des LLMs

Le prompt actuel n'ayant pas fait l'objet d'une attention particulière, il peut être affiné. Les algorithmes autour du Retrieval-Augmented Generation (RAG) semblent particulièrement prometteurs [45]. Leur utilisation pourrait aboutir à une baisse du taux d'hallucination grâce à une meilleure utilisation des données de la visualisation, et ainsi à une diminution du nombre de couples rejetés. Cette approche permettrait de choisir les données externes les plus pertinentes pour les intégrer dans le prompt, en se basant étroitement sur les informations disponibles de l'article et de la visualisation de données. Cela fournirait ainsi des éléments de contexte tirés des articles déjà rédigés sur le même sujet, améliorant ainsi la qualité des Q/As contextuelles.

Les modèles neuronaux actuellement utilisés dans ce processus sont des LLMs. Ils permettent d'effectuer de nombreuses tâches dans différents domaines. Néanmoins, les besoins étant ici spécifiques et définis, il serait intéressant d'utiliser un modèle entraîné sur un ensemble de données dédié. De tels ensembles d'entraînement existent actuellement [46, 47] mais uniquement en langue anglaise.

5.2.5 Présentation des couples de Questions/Réponses

Une fois validées, les Q/As doivent être présentées aux utilisateurs finaux. Pour ce projet, deux possibilités ont été explorées : une Foire Aux Questions (FAQ) ou un agent conversationnel (chatbot). Les FAQs ne permettent de présenter qu'un nombre restreint de couples pour rester lisible, tandis qu'un agent conversationnel autorise et nécessite un ensemble plus important de Q/As. Ces deux options sont présentées dans les sections suivantes. Leur choix dépend du paradigme adopté par l'utilisateur de l'interface : qualité ou quantité. Le journaliste a ainsi la possibilité de valider uniquement quelques-uns des couples qu'il juge comme étant les plus importants ou, au contraire, de sélectionner toutes les Q/As correctes.

Foire Aux Questions textuelle

Il s'agit d'afficher les couples sous forme d'une liste, éventuellement subdivisée selon les thématiques abordées. Cette liste peut, par la suite, être lue par l'utilisateur ou par le biais d'un lecteur d'écran. Une étude comparant ces deux options met en avant une préférence des utilisateurs voyants pour les FAQs par rapport aux agents conversationnels basés sur ces mêmes FAQs [34]. Néanmoins, cette préférence est très ténue puisqu'aucune différence significative n'a été observée pour la qualité de service ou pour la satisfaction. Seul un plus faible nombre de contraintes explique cette préférence.

Certaines considérations et adaptations doivent cependant être prises en compte. Afin de minimiser la charge de travail des utilisateurs finaux, un nombre de 16 à 20 mots par phrase est conseillé [48].

Ainsi, cette modalité devrait être utilisée avec un objectif de communication envers des utilisateurs voyants. Néanmoins, du fait du nombre restreint de Q/As présentables et de sa simplicité, cette modalité de présentation ne fait l'objet que de peu de publications scientifiques.

Agent conversationnel

Contrairement à une FAQ, cette modalité de présentation oblige les utilisateurs à formuler des questions au système. En cela, elle est assimilable, dans un contexte éducatif, à la génération de questions par étudiants. Cette approche pédagogique consiste à faire rédiger des questions sur un texte par les étudiants avant de leur faire résoudre les questions des autres étudiants. La première phase a été étudiée isolément par *J. M. Bugg* et *M. A. McDaniel* [49] avec des étudiants formés à la génération de ces questions. Ils concluent à une meilleure compréhension du sujet sur le court et long terme par rapport à une simple lecture du texte. Néanmoins, la majorité des études portent sur l'ensemble du processus (avec les réponses) [50]. Dans ce cas, l'attention et les efforts cognitifs nécessaires à la réalisation de l'ensemble du processus sont plus importants. Cette sollicitation améliore les performances des élèves, notamment pour les questions de haut niveau - relations entre différents objets et explications complexes - [51]. Cependant, cela dépend fortement de la qualité des questions. L'utilisation d'un agent conversationnel semble améliorer la compréhension des sujets et la mémorisation. Un agent conversationnel pourra donc être employé pour améliorer l'accessibilité, qui est le but principal de cette maîtrise, mais aussi avec un objectif d'éducation.

Souvent orientées vers le domaine médical, les FAQs sont employés afin de résumer [52] et réunir une quantité importante de données textuelles [53]. Différentes approches ont été expérimentées afin de créer de tels systèmes. Ce type de système réclamant un nombre important de Q/As afin d'alimenter une IA, la principale tâche consiste à collecter un ensemble de couples de Questions/Réponses. La majorité des projets utilise le moissonnage du Web, comme des forums, pour construire une telle base de données [54, 55]. Parallèlement, de nombreux travaux à propos des agents conversationnels basés sur une FAQ ont été publiés [56, 57]. Quelle que soit la stratégie adoptée, l'ensemble de ces algorithmes ne prennent pas en compte les formats visuels comme les visualisations de données, acceptant uniquement en entrée des formats textuels.

Cette modalité a été testée par le biais du LLM *ChatGPT* dont les données ont été incluses lors d'un "apprentissage en contexte". L'usage d'un LLM comme celui-ci autorise la personnalisation des réponses ainsi que la création de conversations permettant ainsi une amélioration de l'expérience utilisateur [56]. L'ensemble des Q/As validés par les journalistes lui ont été fournis au format suivant :

$\mathcal{P}_4$ `Crée un chatbot à partir des couples de Questions/Réponses suivants :` [LIST_COUPLES].

où `LIST_COUPLES` est le contenu du fichier de sortie `.faq`.

Il a été constaté qu'un LLM pouvait faire des recoupements entre les différentes questions et ainsi couvrir un nombre de questions d'utilisateurs finaux plus important.

RÉFÉRENCES

[1] C. de Presse du Québec, *Guide de déontologie journalistique*, Conseil de Presse du Québec Standard, 2017. [En ligne]. Disponible : https://conseildepresse.qc.ca/wp-content/uploads/2017/12/Guide-de-deontologie-journalistique_CPQ.pdf

[2] ISO Central Secretary, *Information technology – User interface component accessibility – Part 11 : Guidance on text alternatives for images*, International Organization for Standardization Standard ISO/IEC 20 071-11, 2019. [En ligne]. Disponible : https://www.iso.org/standard/74345.html

[3] Civil Rights Division - U.S. Department of Justice, *Fact Sheet : New Rule on the Accessibility of Web Content and Mobile Apps Provided by State and Local Governments*, Civil Rights Division - U.S. Department of Justice Standard Title II of ADA, 2024. [En ligne]. Disponible : https://www.ada.gov/resources/2024-03-08-web-rule/

[4] World Wide Web Consortium, *Web Content Accessibility Guidelines (WCAG) 2.1*, World Wide Web Consortium Standard WCAG 2.1, 2023. [En ligne]. Disponible : https://www.w3.org/TR/WCAG21/

[5] C. Gleason *et al.*, ""it's almost like they're trying to hide it" : How user-provided image descriptions have failed to make twitter accessible," dans *The World Wide Web Conference*, ser. WWW '19. New York, NY, USA : Association for Computing Machinery, 2019, p. 549–559. [En ligne]. Disponible : https://doi.org/10.1145/3308558.3313605

[6] A. Lundgard et A. Satyanarayan, "Accessible visualization via natural language descriptions : A four-level model of semantic content," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, nº. 1, p. 1073–1083, jan 2022. [En ligne]. Disponible : https://doi.org/10.1109/TVCG.2021.3114770

[7] K. Mack *et al.*, "Designing tools for high-quality alt text authoring," dans *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, ser. ASSETS '21. New York, NY, USA : Association for Computing Machinery, 2021. [En ligne]. Disponible : https://doi.org/10.1145/3441852.3471207

[8] S. S. Chintalapati, J. Bragg et L. L. Wang, "A Dataset of Alt Texts from HCI Publications : Analyses and Uses Towards Producing More Descriptive Alt Texts of Data Visualizations," dans *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*, ser. ASSETS '22. New York, NY, USA : Association for Computing Machinery, 2022. [En ligne]. Disponible :

https://doi.org/10.1145/3517428.3544796

[9] O. Moured *et al.*, “Alt4Blind : A User Interface to Simplify Charts Alt-Text Creation,” 2024.

[10] M. Huh, Y.-H. Peng et A. Pavel, “GenAssist : Making Image Generation Accessible,” dans *ACM Special Interest Group on Computer Graphics and Interactive Techniques (SIGGRAPH) ; ACM Special Interest Group on Computer-Human Interaction (SIGCHI)* -, San Francisco, CA, United states, 2023.

[11] C. Jung *et al.*, “Communicating Visualizations without Visuals : Investigation of Visualization Alternative Text for People with Visual Impairments,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, n°. 1, p. 1095–1105, Jan 2022.

[12] V. S. Morash *et al.*, “Guiding novice web workers in making image descriptions using templates,” *ACM Trans. Access. Comput.*, vol. 7, n°. 4, nov 2015. [En ligne]. Disponible : https://doi.org/10.1145/2764916

[13] A. Belle *et al.*, “Alt-Texify : A Pipeline to Generate Alt-text from SVG Visualizations,” dans *International Conference on Evaluation of Novel Approaches to Software Engineering*, 01 2022, p. 275–281.

[14] S. Wu *et al.*, “Automatic alt-text : Computer-generated image descriptions for blind users on a social network service,” dans *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, ser. CSCW ’17. New York, NY, USA : Association for Computing Machinery, 2017, p. 1180–1192. [En ligne]. Disponible : https://doi.org/10.1145/2998181.2998364

[15] R. Sreedhar *et al.*, “AIDE : automatic and accessible image descriptions for review imagery in online retail,” dans *Proceedings of the 19th International Web for All Conference*, ser. W4A ’22. New York, NY, USA : Association for Computing Machinery, 2022. [En ligne]. Disponible : https://doi.org/10.1145/3493612.3520453

[16] N. Singh, L. L. Wang et J. Bragg, “Figura11y : Ai assistance for writing scientific alt text,” dans *Proceedings of the 29th International Conference on Intelligent User Interfaces*, ser. IUI ’24. New York, NY, USA : Association for Computing Machinery, 2024, p. 886–906. [En ligne]. Disponible : https://doi.org/10.1145/3640543.3645212

[17] J. Kim *et al.*, “Exploring Chart Question Answering for Blind and Low Vision Users,” dans *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’23. New York, NY, USA : Association for Computing Machinery, 2023. [En ligne]. Disponible : https://doi.org/10.1145/3544548.3581532

[18] E. Hoque, P. Kavehzadeh et A. Masry, “Chart Question Answering : State of the Art and Future Directions,” *Computer Graphics Forum*, vol. 41, n°. 3, p. 555–572, 2022. [En ligne]. Disponible : https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14573

[19] S. K. C, P. Joshi et L. A, “Data Extraction and Question Answering on Chart Images Towards Accessibility and Data Interpretation,” *IEEE Open Journal of the Computer Society*, vol. 4, p. 314–325, 2023.

[20] M. Huang *et al.*, “VProChart : Answering Chart Question through visual perception alignment agent and programmatic solution reasoning,” 2024. [En ligne]. Disponible : https://arxiv.org/abs/2409.01667

[21] M. N. Hoque *et al.*, “Towards designing a question-answering chatbot for online news : Understanding questions and perspectives,” 2024.

[22] Z. Xu, S. Jain et M. Kankanhalli, “Hallucination is inevitable : An innate limitation of large language models,” 2024.

[23] J. Wei *et al.*, “Measuring and reducing llm hallucination without gold-standard answers via expertise-weighting,” 2024.

[24] J.-Y. Yao *et al.*, “Llm lies : Hallucinations are not bugs, but features as adversarial examples,” 2023.

[25] Y. Liu *et al.*, “Prompt injection attack against llm-integrated applications,” 2024.

[26] H. Duan, Y. Yang et K. Y. Tam, “Do llms know about hallucination ? an empirical investigation of llm’s hidden states,” 2024.

[27] Y. Li *et al.*, “Evaluating object hallucination in large vision-language models,” 2023.

[28] G. Hong *et al.*, “The hallucinations leaderboard – an open effort to measure hallucinations in large language models,” 2024.

[29] J. Mahilraj *et al.*, “Evaluation of the robustness, transparency, reliability and safety of ai systems,” dans *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1, March 2023, p. 2526–2535.

[30] K. Tsiakas et D. Murray-Rust, “Using human-in-the-loop and explainable ai to envisage new future work practices,” dans *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*, ser. PETRA ’22. New York, NY, USA : Association for Computing Machinery, 2022, p. 588–594. [En ligne]. Disponible : https://doi.org/10.1145/3529190.3534779

[31] D. R. Honeycutt, M. Nourani et E. D. Ragan, “Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy,” 2020.

[32] T. Biloborodova et I. Skarga-Bandurova, “Human-ai collaboration in decision making : An initial reliability study and methodology,” dans *2023 IEEE 12th International Conference on Intelligent Data Acquisition and Advanced Computing Systems : Technology and Applications (IDAACS)*, vol. 1, Sep. 2023, p. 1151–1155.

[33] D. Song, “Student-generated questioning and quality questions : A literature review,” *Research Journal of Educational Studies and Review*, vol. 2, p. 58–70, 01 2016.

[34] S. Han et M. K. Lee, “Faq chatbot and inclusive learning in massive open online courses,” *Computers & Education*, vol. 179, p. 104395, 2022. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S0360131521002724

[35] J. A. Jiang *et al.*, “Supporting serendipity : Opportunities and challenges for human-ai collaboration in qualitative analysis,” *Proc. ACM Hum.-Comput. Interact.*, vol. 5, nº. CSCW1, apr 2021. [En ligne]. Disponible : https://doi.org/10.1145/3449168

[36] R. Borgo *et al.*, “Trust junk and evil knobs : Calibrating trust in ai visualization,” dans *IEEE PacificVis conference proceedings*, ser. PacificVis conference proceedings. IEEE, 2024.

[37] K. Davila *et al.*, “Chart mining : A survey of methods for automated chart analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, p. 1–1, 05 2020.

[38] M. Savva *et al.*, “Revision : automated classification, analysis and redesign of chart images,” dans *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, ser. UIST ’11. New York, NY, USA : Association for Computing Machinery, 2011, p. 393–402. [En ligne]. Disponible : https://doi.org/10.1145/2047196.2047247

[39] Z. Xu et E. Wall, “Exploring the capability of LLMs in performing low-level visual analytic tasks on svg data visualizations,” 2024.

[40] L. Chen, M. Zaharia et J. Zou, “How is ChatGPT’s behavior changing over time ?” 2023.

[41] L. Martin *et al.*, “CamemBERT : a tasty French language model,” dans *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online : Association for Computational Linguistics, juill. 2020, p. 7203–7219. [En ligne]. Disponible : https://www.aclweb.org/anthology/2020.acl-main.645

[42] B. Albar, P. Bedu et L. Bourdois, “QAmembert (Revision 9685bc3),” Centre Aquitain des Technologies de l’Information et Electroniques, 2023. [En ligne]. Disponible : https://huggingface.co/CATIE-AQ/QAmembert

[43] J. Cegarra et N. Morgado, “Étude des propriétés de la version francophone du NASA-TLX,” *EPIQUE 2009 : 5ème Colloque de Psychologie Ergonomique*, p. 233–239, 01 2009.

[44] A. Dubey *et al.*, “The Llama 3 Herd of Models,” 2024. [En ligne]. Disponible : https://arxiv.org/abs/2407.21783

[45] K. Muludi *et al.*, “Retrieval-augmented generation approach : Document question answering using large language model,” *International Journal of Advanced Computer*

Science and Applications, vol. 15, nº. 3, 2024. [En ligne]. Disponible : http://dx.doi.org/10.14569/IJACSA.2024.0150379

[46] S. E. Kahou *et al.*, "FigureQA : An Annotated Figure Dataset for Visual Reasoning," 2018.

[47] R. Chaudhry *et al.*, "LEAF-QA : Locate, Encode ; Attend for Figure Question Answering," dans *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Los Alamitos, CA, USA : IEEE Computer Society, mar 2020, p. 3501–3510. [En ligne]. Disponible : https://doi.ieeecomputersociety.org/10.1109/WACV45572.2020.9093269

[48] B. B. Kadayat et E. Eika, "Impact of Sentence Length on the Readability of Web for Screen reader Users," dans *Universal Access in Human-Computer Interaction. Design Approaches and Supporting Technologies*, M. Antona et C. Stephanidis, édit. Cham : Springer International Publishing, 2020, p. 261–271.

[49] J. M. Bugg et M. A. McDaniel, "Selective Benefits of Question Self-Generation and Answering for Remembering Expository Text," *Journal of Educational Psychology*, vol. 104, p. 922–931, 2012. [En ligne]. Disponible : https://api.semanticscholar.org/CorpusID:9469208

[50] M. Ebersbach, M. Feierabend et K. B. B. Nazari, "Comparing the effects of generating questions, testing, and restudying on students' long-term recall in university learning," *Applied Cognitive Psychology*, vol. 34, nº. 3, p. 724–736, 2020. [En ligne]. Disponible : https://onlinelibrary.wiley.com/doi/abs/10.1002/acp.3639

[51] D. Song, "Student-generated questioning and quality questions : A literature review," *Research Journal of Educational Studies and Review*, vol. 2, p. 58–70, 01 2016.

[52] S. Jeyaraj et T. Raghuveera, "A deep learning based end-to-end system (F-Gen) for automated email FAQ generation," *Expert Systems with Applications*, vol. 187, 2022. [En ligne]. Disponible : http://dx.doi.org/10.1016/j.eswa.2021.115896

[53] X. F. Zhang *et al.*, "COUGH : A Challenge Dataset and Models for COVID-19 FAQ Retrieval," dans *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, Virtual, Punta Cana, Dominican republic, 2021, p. 3759 – 3769.

[54] X. Guo *et al.*, "The construction of a Diabetes-oriented FAQ Corpus for Automated Question-Answering services," dans *ACM International Conference Proceeding Series*, Taiyuan, China, 2020, p. 60 – 66. [En ligne]. Disponible : http://dx.doi.org/10.1145/3433996.3434008

[55] W. Astuti *et al.*, "Predicting FAQ on the COVID-19 Chatbot using the DIET Classifier," dans *3rd 2021 East Indonesia Conference on Computer and Information*

Technology, EIConCIT 2021, Virtual, Surabaya, Indonesia, 2021, p. 25 – 29. [En ligne]. Disponible : http://dx.doi.org/10.1109/EIConCIT50028.2021.9431913

[56] F. Khennouche *et al.*, "Revolutionizing generative pre-traineds : Insights and challenges in deploying ChatGPT and generative chatbots for FAQs," *Expert Systems with Applications*, vol. 246, p. 123224, 2024. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S0957417424000897

[57] A. Chatterjee, M. Gupta et P. Agrawal, "FAQAugmenter : Suggesting Questions for Enterprise FAQ Pages," dans *Proceedings of the 13th International Conference on Web Search and Data Mining*, ser. WSDM '20. New York, NY, USA : Association for Computing Machinery, 2020, p. 829–832. [En ligne]. Disponible : https://doi.org/10.1145/3336191.3371862

ANNEXE A RÉSULTATS DÉTAILLÉES DES TESTS-UTILISATEURS

Part.	Id article	Interface	Temps (min)	Nb Q/As	Score TLX
A	I	Avec	4.55	7	11.2
A	II	Avec	5.40	5	9.73
A	III	Avec	5.30	7	8.8
A	IV	Sans	8.11	6	13.07
A	V	Sans	7.11	9	14.53
B	I	Avec	12.28	18	9.07
B	II	Sans	13.28	19	13.6
B	III	Avec	27.02	23	13.47
C	IV	Avec	58.32	37	13.47
C	VI	Sans	42.45	95	9.87

Tableau A.1 Résultats détaillés

ANNEXE B COURRIEL DE RECUTEMENT

Bonjour [Mme | M.] [NOM],

En tant que journaliste à la rédaction du journal *Le Devoir* / du site des nouvelles *Radio-Canada*, vous êtes convié à participer à l'évaluation d'un outil de validation de couples de Questions/Réponses.

Ce projet de recherche en partenariat entre *Polytechnique Montréal*, *Le Devoir* et *Radio-Canada* a pour objectif de proposer l'utilisation de Questions/Réponses (sur le modèle des FAQs) afin de véhiculer l'information transmise par les articles de presse en ligne avec graphique.

Si vous souhaitez participer à l'évaluation, cliquez sur le lien suivant : `https://forms.gle/VWYqQ4WzuYGFWCYU7`. Il vous redirigera vers le formulaire de consentement et confirmera votre inscription à l'étude. Il est à noter que vous pourrez en tout temps mettre fin à votre participation si vous ne souhaitez plus participer à l'évaluation.

Nous vous remercions à l'avance pour votre participation à l'évaluation de notre outil.

Cordialement,

Théo LECARDONNEL
Étudiant à la maîtrise Recherche en génie logiciel
Polytechnique Montréal

ANNEXE C FORMULAIRE DE CONSENTEMENT

Le formulaire de consentement prend la forme d'un document Google Form, consultable via le lien suivant : `https://forms.gle/VWYqQ4WzuYGFWCYU7`. Il est également fourni dans la suite de ce document.

Formulaire d'information et de consentement

* Indique une question obligatoire

1. Adresse e-mail *

_____

***GenQA* : un outil de validation d'ensemble de couples de Questions/Réponses**

Responsable de l'activité de recherche
Théo LECARDONNEL
Étudiant à la maîtrise Recherche
C.P. 6079, succ. Centre-ville
Montréal (Québec)
H3C 3A7
Tél. (514) 340-4711 poste 7109
Fax. (514) 340-5139
Adresse courriel: theo.lecardonnel@polymtl.ca

Sous la direction de
Thomas HURTUT
Professeur
École Polytechnique de Montréal
C.P. 6079, succ. Centre-ville
Montréal (Québec)
H3C 3A7
Tél. (514) 340-4711 poste 7109
Fax. (514) 340-5139
Adresse courriel : thomas.hurtut@polymtl.ca

Financement de l'activité de recherche
Ce projet fait l'objet d'un financement par une subvention de recherche. Cette subvention a été octroyée par CRSNG Alliance / Prompt.

Conflit d'intérêts
L'équipe de recherche n'est pas en situation de conflit d'intérêts dans le contexte de la présente activité de recherche.

Présentation

GenQA : un outil de validation d'ensemble de couples de Questions/Réponses

Préambule

Nous vous invitons à participer à une activité de recherche qui vise à aider des journalistes lors de la validation de couples de Questions/Réponses destinés à véhiculer l'information transmise par les articles avec graphiques.

Cependant, avant d'accepter de participer à cette activité et de signer le présent formulaire d'information et de consentement, veuillez prendre le temps de lire l'information présentée.

Nous vous invitons à poser toutes les questions que vous jugerez utiles à la responsable ou au responsable de l'activité de recherche ou à tout autre membre de l'équipe de recherche et à leur demander de vous expliquer tout mot ou renseignement qui ne serait pas clair. Nous vous invitons également à prendre conseil auprès de toute autre personne de qui vous aimeriez obtenir un avis à propos de votre éventuelle participation.

Présentation générale du projet de recherche

Dans le cadre de notre projet de recherche, nous nous intéressons à l'usage de couples de Questions/Réponses (noté par la suite Q\As) pour véhiculer l'information transmise par les graphiques. Ces couples devant être fournis en grande quantité pour que cette démarche soit pertinente, la génération de ces couples est dévolue à un modèle LPLM de type décodeur. Il devient ainsi nécessaire de contrôler la qualité de celles-ci.

L'objectif de ce projet est ainsi de proposer un outil permettant de **valider un grand nombre de ces Q\As en un temps minimum**. Cette interface devra permettre aux journalistes de filtrer un ensemble de couples avant publication.
Les principaux objectifs de l'étude que nous réalisons sont les suivants :

- Proposer une alternative à l'utilisation de text-alternatif pour véhiculer l'information de ces articles avec graphiques
- Proposer un outil interactif permettant d'effectuer une validation de ces couples

Afin d'atteindre nos objectifs, nous effectuerons les activités suivantes:

- Utilisation en contexte sur différents articles de l'outil proposé
- Entretien destinée à collecter des retours sur l'expérience du participant

Nature et durée de votre participation à l'activité de recherche

Votre participation dans le cadre du présent projet sera constitué d'une unique séance en ligne d'environ deux heures, effectuée via le logiciel de vidéoconférence de Zoom sous licence institutionnelle.
Elle sera divisé en quatre parties :

- Présentation générale du projet et de l'interface développée
- Validation de différents ensembles de couples de Questions/Réponses avec l'interface
- Validation de différents ensembles de couples de Questions/Réponses sans l'interface
- Entretien final

L'intégralité de cet entretien final sera enregistré (aussi bien vidéo que audio) afin de permetre une prise de notes ultérieure.

2. **Habilités générales** *

Plusieurs réponses possibles.

- [ ] Je suis à l'aise avec les ordinateurs.
- [ ] Je n'ai pas de défauts visuels m'empêchant de travailler avec des visualisations de données

Activité de Recherche

GenQA : un outil de validation d'ensemble de couples de Questions/Réponses

Avantages pouvant découler de votre participation à l'activité de recherche

Vous ne retirerez aucun bénéfice personnel de votre participation à la présente activité de recherche. Toutefois, votre participation permettra de faire avancer l'état des connaissances.

Risques pouvant découler de votre participation à l'activité de recherche

La présente activité de ne devrait pas entraîner des risques plus grands que ceux que vous rencontrés dans votre vie de tous les jours.

Inconvénients pouvant découler de votre participation à l'activité de recherche

Votre participation au projet de recherche nécessitera du temps, ponctuellement lors de la phase d'utilisation en contexte, et pour une durée maximale de deux heures.

Participation volontaire et possibilité de retrait

Votre participation à la présente activité de recherche est volontaire. Vous êtes donc libre de refuser d'y participer et pouvez à tout moment décider de vous en retirer sans avoir à motiver votre décision et sans risquer d'en subir de préjudice. Vous n'avez qu'à en informer la personne-ressource de l'équipe de recherche et ce, par simple avis verbal.

En cas de retrait, vous pouvez demander la destruction des données vous concernant. Cependant, il sera impossible de retirer vos données ou votre matériel des analyses menées une fois ces dernières publiées ou diffusées.

Tout au long des activités de recherche, vous recevrez en temps opportun l'information pertinente en lien avec votre participation.

L'équipe de recherche et le comité d'éthique de la recherche se réservent le droit de vous retirer de l'étude si vous ne respectez pas les consignes, s'il existe des raisons administratives d'abandonner l'activité, ou pour toutes autres raisons concernant la faisabilité de l'étude. Si une telle situation survient, l'équipe de recherche vous en informera dès que possible.

Indemnisation en cas de préjudice et droits des participant(e)s

Si vous deviez subir quelque préjudice que ce soit par suite de votre participation à cette activité de recherche, vous ne renoncez à aucun de vos droits ni ne libérez les chercheurs, l'organisme de financement ou Polytechnique Montréal de leurs responsabilités légales et professionnelles

Gestion et diffusion des données

Confidentialité et protection de vos données

Ce test-utilisateur, dans son intégralité, a fait l'objet d'une approbation de la commission d'éthique de Polytechnique Montréal sous le label "**Projet CER-2324-62-D**".

L'équipe de recherche recueillera et consignera toutes vos données de manière sécuritaire de façon à en protéger le caractère confidentiel.

Voici comment nous protégerons vos données **lors de la collecte :**

- Les entretiens seront organisés en ligne avec un logiciel de visioconférence, et seront enregistrés.
- Ces enregistrements seront conservées dans un dossier google drive restreint aux membres de l'équipe de recherche cités plus haut et dans un disque dur de secours séparé .

Voici comment nous protégerons vos données **lors des analyses et du transfert des données** entre les membres de l'équipe :

- Les transferts se feront au cas par cas, uniquement entre les personnes listées dans la section "Équipe de recherche". L'analyse sera faite uniquement par ces personnes, et vos données ne seront pas transférées a autrui.

Voici comment nous protégerons vos données **lors des publications** :

- La publication référeront uniquement à des extraits d'entretien anonymisés.
- Si une références directe aux participants dans les publications se feront sous la forme "Participant 1"

Enfin, voici comment nous protégerons vos données **après le projet de recherche** :

- Les enregistrement vidéos des entretiens seront détruits après avoir été transcrits.
- Les transcriptions des entretiens seront conservés pour un durée de 7 ans après le projet de recherche.
- Ces données ne seront pas réutilisées au delà du projet et des publications liées. Elles seront conservées a des fins d'archive.

Vous avez le droit de consulter votre dossier de recherche pour vérifier l'exactitude des renseignements recueillis aussi longtemps que l'équipe de recherche ou Polytechnique Montréal détiendront ces informations. Cependant, afin de préserver l'intégrité scientifique du projet de recherche, certaines informations seront accessibles seulement à la fin du projet de recherche.

Diffusion des résultats de la recherche

Le chercheur responsable utilisera les données du projet de recherche pour les simples fins du projet de recherche. Les résultats du projet pourront être publiés dans des documents tels qu'une revue scientifique.

Personnes-ressources

Si vous avez des questions sur les **aspects scientifiques** du projet de recherche ou pour vous **retirer de l'étude**, vous pouvez contacter Thomas Hurtut au (514) 340-4711, poste 7109 ou encore par courriel à thomas.hurtut@polymtl.ca.

Pour toute préoccupation sur vos droits ou sur les responsabilités de l'équipe de recherche concernant votre participation à ce projet, vous pouvez contacter le Comité d'éthique de la recherche de Polytechnique Montréal au (514) 340-4711, poste 4420 ou encore par courriel à ethique@polymtl.ca

Consentement

1. J'ai pris connaissance de la documentation ci-jointe, décrivant la nature et le déroulement du projet de même que les risques et les inconvénients qui pourraient survenir.

2. Je comprends que j'ai droit à des réponses satisfaisantes aux questions que je poserais quant à mon implication dans ce projet tout au long de ma participation.

3. Je consens à participer librement à ce projet, après avoir obtenu et pris le temps d'y réfléchir à ma satisfaction et sans avoir subi de pression à cet effet.

4. Je comprends qu'en participant à ce projet de recherche, je ne renonce à aucun de mes droits ni ne dégage les chercheurs de leurs responsabilités.

5. Je comprends que je peux consulter le dossier que l'équipe de recherche constitue sur moi.

6. Je pourrai à tout moment, sur simple avis de ma part, revenir sur ma décision de participer et serai alors immédiatement libéré de mon engagement.

7. J'ai reçu une copie du présent document.

3. **Consentement à la participation au projet de recherche** *

Une seule réponse possible.

- ◯ J'accepte de participer à ce projet de recherche aux conditions qui y sont énoncées.
- ◯ Je refuse de participer à ce projet de recherche aux conditions qui y sont énoncées.

4. **Consentement à la prise d'images vidéo** *

Une seule réponse possible.

- ◯ J'accepte que l'entrevue lors de laquelle seront collectés vos retours sur le projet sera enregistrée
- ◯ Je refuse que l'entrevue lors de laquelle seront collectés vos retours sur le projet sera enregistrée

5. **Nom du participant** *

Veuillez indiquer votre nom et courriel afin d'être recontacté pour d'organiser la session d'expérimentation sur ce projet.

ANNEXE D FORMULAIRE NASA-TLX

Le formulaire de réponse prends la forme d'un document Google Form, accessible par le lien suivant : `https://forms.gle/t6nrDVvPj4TJKAiKA`. Ce document est également disponible dans les pages suivantes. Il est constitué de la version française du test original.

Questionnaire NASA-TLX

L'évaluation que vous êtes sur le point d'effectuer est une technique qui a été développée par la NASA pour évaluer l'importance relative de six facteurs dans la détermination de la charge de travail que vous avez ressentie lors de l'exécution d'une tâche que vous avez récemment effectuée. Les six facteurs sont définis à la page suivante. Lisez-les pour vous assurer que vous comprenez la signification de chaque facteur. Si vous avez des questions, n'hésitez pas à les poser.

* Indique une question obligatoire

1. Identifiant du participant *

2. Identifiant de l'article-exemple *

Définitions

Exigence mentale (faible/élevée)

Quelle activité mentale et perceptive était requise (par exemple, penser, décider, calculer, se souvenir, regarder, chercher, etc.) La tâche était-elle facile ou exigeante, simple ou complexe, indulgente ou exigeante ?

Exigence physique (faible/élevée)

Quelle quantité d'activité physique était requise (par exemple, pousser, tirer, tourner, contrôler, activer, etc.) La tâche était-elle facile ou exigeante, lente ou rapide, détendue ou exténuante, reposante ou laborieuse ?

Exigence temporelle (faible/élevée)

Quelle pression temporelle avez-vous ressentie en raison de la vitesse ou du rythme auquel les tâches ou les éléments de la tâche se sont déroulés ? Le rythme était-il lent et tranquille ou rapide et effréné ?

Performance (bonne/mauvaise)

Dans quelle mesure pensez-vous avoir réussi à atteindre les objectifs de la tâche fixés par l'expérimentateur (ou vous-même) ? Dans quelle mesure avez-vous été satisfait de votre performance dans l'accomplissement de ces objectifs ?

Effort (faible/élevé)

Dans quelle mesure avez-vous dû travailler (mentalement et physiquement) pour atteindre votre niveau de performance ?

Frustration (faible/élevé)

Dans quelle mesure avez-vous ressenti de l'insécurité, du découragement, de l'irritation, du stress et de l'agacement par rapport à de la sécurité, de la satisfaction, du contentement, de la détente et de la complaisance pendant la tâche ?

Évaluation des six dimensions

3. **Exigence mentale** *

Quelle activité mentale et perceptive était requise (par exemple, penser, décider, calculer, se souvenir, regarder, chercher, etc.) La tâche était-elle facile ou exigeante, simple ou complexe, indulgente ou exigeante ?

Une seule réponse possible.

	0	1	2	3	4	5	6	7	8	9	10	
Faib	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Forte

4. **Exigence physique** *

Quelle quantité d'activité physique était requise (par exemple, pousser, tirer, tourner, contrôler, activer, etc.) La tâche était-elle facile ou exigeante, lente ou rapide, détendue ou exténuante, reposante ou laborieuse ?

Une seule réponse possible.

	0	1	2	3	4	5	6	7	8	9	10	
Faib	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Forte

5. **Exigence temporelle** *

Quelle pression temporelle avez-vous ressentie en raison de la vitesse ou du rythme auquel les tâches ou les éléments de la tâche se sont déroulés ? Le rythme était-il lent et tranquille ou rapide et effréné ?

Une seule réponse possible.

	0	1	2	3	4	5	6	7	8	9	10	
Faib	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Forte

6. **Performance** *

Dans quelle mesure pensez-vous avoir réussi à atteindre les objectifs de la tâche fixés par l'expérimentateur (ou vous-même) ? Dans quelle mesure avez-vous été satisfait de votre performance dans l'accomplissement de ces objectifs ?

Une seule réponse possible.

	0	1	2	3	4	5	6	7	8	9	10	
Bon	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Mauvaise

7. **Effort** *

Dans quelle mesure avez-vous dû travailler (mentalement et physiquement) pour atteindre votre niveau de performance ?

Une seule réponse possible.

	0	1	2	3	4	5	6	7	8	9	10	
Faib	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Forte

8. **Frustration** *

Dans quelle mesure avez-vous ressenti de l'insécurité, du découragement, de l'irritation, du stress et de l'agacement par rapport à de la sécurité, de la satisfaction, du contentement, de la détente et de la complaisance pendant la tâche ?

Une seule réponse possible.

	0	1	2	3	4	5	6	7	8	9	10	
Faib	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Forte

Ordonnancement des dimensions

Choisissez ci-dessous le facteur qui contribue le plus à la charge de travail pour la tâche spécifique que vous avez récemment effectuée.

9. Choix n°1 *

Une seule réponse possible.

- ◯ Exigence physique
- ◯ Exigence mentale

10. Choix n°2 *

Une seule réponse possible.

- ◯ Performance
- ◯ Exigence mentale

11. Choix n°3 *

Une seule réponse possible.

- ◯ Effort
- ◯ Exigence physique

12. Choix n°4 *

Une seule réponse possible.

- ◯ Effort
- ◯ Performance

13. Choix n°5 *

Une seule réponse possible.

- ◯ Performance
- ◯ Frustration

14. Choix n°6 *

Une seule réponse possible.

- ◯ Exigence temporelle
- ◯ Exigence physique

15. Choix n°7 *

Une seule réponse possible.

- ◯ Effort
- ◯ Exigence mentale

16. Choix n°8 *

Une seule réponse possible.

- ◯ Exigence temporelle
- ◯ Peformance

17. Choix n°9 *

Une seule réponse possible.

- ◯ Exigence mentale
- ◯ Frustration

18. Choix n°10 *

Une seule réponse possible.

- ◯ Exigence mentale
- ◯ Exigence temporelle

19. Choix n°11 *

Une seule réponse possible.

- ◯ Peformance
- ◯ Exigence physique

20. Choix n°12 *

Une seule réponse possible.

- ◯ Frustration
- ◯ Exigence physique

21. Choix n°13 *

Une seule réponse possible.

- ◯ Effort
- ◯ Frustration

22. Choix n°14 *

Une seule réponse possible.

- ◯ Effort
- ◯ Exigence temporelle

23. Choix n°15 *

Une seule réponse possible.

- ◯ Frustration
- ◯ Exigence temporelle

ANNEXE E GUIDE D'ENTRETIEN

Présentation de l'objectif de l'entretien final

Tout d'abord merci d'avoir accepté de participer à cette expérience. L'objectif de cet échange final est de collecter des retours sur l'outil qui vous avez pu manipuler avec différents article-exemples.

Consentement à l'enregistrement

Avant de commencer, je souhaite vous demander votre autorisation pour enregistrer cette entrevue (audio comme vidéo). L'enregistrement a pour but de garantir une prise de notes précise et une analyse de notre conversation. Êtes-vous d'accord pour que j'enregistre notre entretien ?

Droits de la personne interviewée et sécurité

Vous pouvez à tout moment demander à mettre fin à l'entretien, sans condition ou justification. Toutes les informations seront anonymisées dans l'analyse finale des données. Le stockage des données d'entretien sur une plateforme cloud fermée, exclusivement accessible aux membres du laboratoire jData participant à l'expérience, c'est-à-dire moi même, *Christophe HURTER* et *Thomas HURTUT*.
Vos réponses sont sur une base volontaire, et vous pouvez à tout moment refuser de répondre ou mettre fin à l'entretien.
Avant de débuter, avez-vous des questions ?

Questions sur l'utilisation de Q/As

Ouverture

Que pensez-vous de l'usage de Q/As pour véhiculer l'information issue d'un article et de son graphique ?

Taille de l'ensemble des Q/As

Quel serait, pour vous, le nombre de Q/As idéal à présenter aux utilisateurs finaux ?

Combien de couples sélectionneriez-vous pour un article donné avec l'outil ? Sans l'outil ?

Combien de couples généreriez-vous pour un article avant d'effectuer la validation avec l'outil ? Sans l'outil ?

Questions générées par IA

Les Q/As vous ont-ils semblé de bonne qualité avant une quelconque selection ? Après ?

L'utilisation de l'IA dans ce processus vous parait-elle acceptable ? Sous quelles conditions ?

Quelles questions / sujets vous semblent manquantes / manquants ?

Quelles questions / sujets n'auriez vous pas inclus sans cet outil ?

Questions sur l'outil proposé

Ouverture

Passons à l'outil en général. Quelle a été votre impression sur cette interface ?

Stratégie

Quelle stratégie avez-vous adopté pour valider ces ensembles ? Etait-elle la même pour tous les articles ?

Pouvez vous me dire comment l'outil a pu vous limiter ou vous contraindre dans votre validation des Q/As ?

Regroupement

Pensez-vous que les regroupements proposés ont influencé la façon dont vous avez valider les ensembles ?

Comment modifierez-vous ces regroupements pour les rendre plus compréhensible ?

Interaction

Qu'avez-vous pensé des interactions proposées ?

Pensez-vous que les interactions ont influencé la façon dont vous avez valider ces ensembles ?

Comment changeriez-vous les options d'interaction (clavier / souris) pour les rendre plus intéressantes ?

Ergonomie

Qu'avez-vous pensé de l'ergonomie de l'outil ?

Y-a-t-il des choses qui vous ont posé des difficultés pour utiliser l'outil ? Lesquelles ?

Comment changeriez-vous l'outil pour le rendre plus facile à utiliser ?

Exemples utilisés

Qu'avez-vous pensé de l'intérêt et de la qualité des articles ?

Quels impacts ces articles ont-il pu avoir sur votre validité ?

Selon-vous, quels changements seraient nécessaires pour améliorer la pertinence de ces articles ?

— Autres

Quels changements apporteriez-vous à l'outil pour le rendre plus utile ?

Quel impact pensez-vous qu'utiliser cet outil a eu sur votre temps et vos efforts de travail ?

Questions de clôture

Idées additionnelles

Y-a-t-il des sujets qui n'ont pas été couverts par cet entretien et qui devraient être discutés ?

Souhaitez-vous ajouter quelque chose avant de terminer ?

Conclusion

Merci pour votre temps et votre contribution.

Relances

Générique

Pourriez-vous entrer dans les détails de votre raisonnement ?

Pourriez-vous approfondir votre réponse ?

Serait-il possible d'avoir plus de détails ?

Pourriez-vous m'en dire plus à ce sujet ?

Pourriez-vous me donner un exemple ?

Pourriez-vous étoffer ?

Spécifique

Est-ce que vous pourriez développer davantage à propos de [sujet] ?

Pourriez-vous m'en dire plus à propos de [sujet] ?

Vous parlez de [sujet], pouvez-vous détailler ?