L'intégration de l'intelligence artificielle dans les usages d'un regroupement de recherche sur la persévérance et la réussite scolaire :

Développement et évaluation d'un robot conversationnel pour l'exploration du corpus de leurs écrits scientifiques

Serigne Mbacke Gueye

Post-doctorant, Faculté des sciences de l'éducation

Université Laval, Québec

Semestre 1 / 2024

Adjectif

RÉSUMÉ : *Depuis sa création en 2015, les chercheuses et chercheurs de PÉRISCOPE ont réalisé de nombreuses publications accessibles sur le site web du Réseau. Elles constituent un corpus et un métacadre d'analyse susceptible de conduire à des publications fort originales et aussi utiles pour d'autres domaines d'étude et de recherche ; d'où la pertinence scientifique et sociale d'appliquer le traitement du langage naturel (NLP) à la compréhension d'ensemble de nombre de travaux de recherche réalisés, en matière de persévérance et de réussite scolaires. Depuis ces dernières années, l'émergence des GPT (Generative Pre-trained Transformer) influence la manière dont l'IA est intégrée et utilisée dans divers domaines d'application avec la naissance des grands modèles de langue comme ChatGPT, offrant des capacités avancées en compréhension du langage naturel et en génération d'un texte similaire à celui écrit par un humain. Cependant, leur efficacité à usage général dans des domaines spécialisés reste limitée. Dans le cadre de cet article, nous avons mis à la disposition des chercheur·es et partenaires du réseau PÉRISCOPE un robot conversationnel (PÉRO) à l'instar de ChatGPT afin de leur permettre d'interroger l'ensemble des publications disponibles au niveau du site internet de PÉRISCOPE grâce au concept d'indexation (basé sur LlamaIndex) et de l'apprentissage en contexte (GPT-3.5). Pour évaluer les performances de PÉRO (notre solution), nous avons comparé ses réponses générées avec celles de ChatGPT et de Bing (entendre CoPilot et ChatGPT-4) pour les mêmes questions relatives au contenu des publications en rapport avec l'engagement et le langage écrit, deux concepts-clés en matière de persévérance et de réussite scolaires. Les résultats montrent que PÉRO arrive à faire quasiment aussi bien que les grosses IA génératives (ChatGPT et CoPilot), ce qui équivaudrait à une preuve de concept.*

MOTS-CLÉS : *PÉRISCOPE ; Chabot ; ChatGPT ; LlamaIndex.*

Introduction

Depuis 2015, le réseau PÉRISCOPE (Plateforme Échange, Recherche et Intervention sur la SCOlarité : Persévérance et réussitE) bénéficie d'un fonds d'infrastructure octroyé par le Fonds de Recherche du Québec – Société et Culture (FRQSC), avec le soutien de la Fondation Antoine-Turmel et du MÉES (ministère de l'Éducation et ministère de l'Enseignement supérieur). Ce réseau vise à favoriser les échanges entre chercheurs et intervenants des terrains en matière de scolarité, persévérance et réussite scolaires (PRS) en encourageant le croisement des différentes perspectives des acteurs de la PRS.

La persévérance tout comme la réussite scolaire sont deux objets de forte préoccupation au Québec. Le Gouvernement veut faire passer le taux de réussite scolaire à plus de 85 %[1]. Ce qui distingue

1 À noter que des variations importantes sont observées entre les genres de même que dans différentes régions du Québec.

l'activité du réseau PÉRISCOPE dans la poursuite de cet objectif, c'est son approche multidimensionnelle de la PRS et l'engagement des différents partenaires qui composent le Réseau. L'intensification de la participation des agents de l'éducation est sa stratégie-clé : participation de l'enfant dans son groupe et de l'élève dans la classe, participation du personnel enseignant de l'école, renforcement de la collaboration entre l'école, les parents et les organismes communautaires et participation dans les instances de décision. De nombreuses activités de recherche et d'intervention ont été déployées à cette fin, notamment concernant les technologies numériques comme outils de soutien à l'enseignement et à l'apprentissage ainsi qu'à diverses formes d'échanges entre agent·es de l'éducation.

Sur le site web du réseau PÉRISCOPE (https://www.periscope-r.quebec), plusieurs milliers de documents sont répertoriés, incluant résumés de publication, voire des textes entiers. Ces documents constituent un corpus qu'il est possible d'interroger partant de mots-clés, du type de document, des agents de la PRS et de leur contexte d'action ainsi que des six composants du triangle d'Engeström (1987, 2015) (agents, objet de l'activité, communauté, instruments/outils, rôles et règles). L'inclusion de ce triangle, qui illustre la troisième génération de la théorie de l'activité puisque la première génération est l'œuvre de Vygotsky (1978), dans la base de données permet une lecture originale et évolutive de l'activité du réseau PÉRISCOPE.

L'intégration d'un robot conversationnel à PÉRISCOPE, alimenté par quelques milliers de documents produits par ses membres, est désormais envisageable. Rendu possible grâce à l'évolution des modèles de langage en informatique, un tel ajout à l'infrastructure du réseau augmenterait vraisemblablement les échanges autant au sein du réseau qu'avec celui-ci, vu la pertinence scientifique et sociale d'appliquer le traitement du langage naturel (NLP) à la compréhension d'ensemble de nombre de travaux de recherche réalisés, en matière de persévérance et de réussite scolaire. En outre, le processus suivi est susceptible d'être utile à d'autres domaines d'étude et de recherche. Le présent article se penche sur la première année de design du robot PÉRO, celle du développement du prototype par voie de design participatif.

Contexte

Depuis ces dernières années, on assiste à l'émergence des techniques du traitement automatique du langage naturel (TALN) dont les transformateurs pré-entraînés générateurs (plus connus sous l'acronyme GPT (*Generative Pre-trained Transformer*)) (Floridi & Chiriatti, 2020 ; Lund & Wang, 2023). Ceci a considérablement influencé la façon dont l'IA est intégrée et utilisée dans divers domaines d'application. Les grands modèles de langue, plus connus sous l'appellation *Large Langage Models* (LLM), basés sur l'architecture des transformateurs (Vaswani et al., 2017), ont des capacités inédites, permettent, entre autres, de générer du texte semblable à celui écrit par un humain et aussi de comprendre le langage naturel afin de répondre aux questions qui lui sont posées dans un contexte donné. ChatGPT, un modèle de langage développé par OpenAI en novembre 2022, basé sur l'architecture GPT-3.5, en est un exemple.

Cependant, malgré la puissance de ces modèles pour des tâches de traitement du langage naturel, leur efficacité à usage général dans des domaines spécialisés reste limitée en raison d'un manque de connaissances spécifiques du domaine et de la compréhension du contexte ; et aussi du fait qu'ils se basent sur une approche statique. En effet, par exemple, l'entraînement du modèle GPT-3.5-Turbo a été fait avec des données allant jusqu'en septembre 2021. Ainsi, bien que GPT-4 soit disponible pour une personne abonnée ou via CoPilot, GPT-3.5, soit la version encore la plus utilisée, n'a pas accès aux connaissances plus récentes. Par conséquent, pour des questions portant sur des données récentes ou un domaine spécifique, ChatGPT3-5, et aussi ChatGPT4 et tout comme tout autre robot conversationnel, peut fournir des réponses incorrectes ou incomplètes.

La mise en place de grands modèles de langue (*Large Langage Model* (LLM)) se fait en deux étapes : le pré-entraînement et l'affinement (*fine-tuning*). Pendant le pré-entraînement, les modèles sont entrainés sur une quantité de données (du texte) importante provenant de diverses sources (livres, articles, sites web, etc.). Au cours de cette phase, le modèle apprend à prédire dans une phrase le mot suivant en fonction des mots précédents. Ainsi, il saisit la grammaire, la syntaxe et la sémantique et acquiert une quantité importante de connaissances. Cette phase nécessite d'importantes ressources de calcul, de stockage et d'énergie. Le raffinement (fine-tuning), quant à lui, consiste à entraîner à nouveau le modèle pré-entraîné pendant quelques périodes supplémentaires en utilisant des données étiquetées sur un domaine spécifique, ce qui permet au modèle d'être plus performant dans ce domaine. Ceci dans l'optique d'améliorer ses performances sur des données spécifiques au domaine tout en conservant la compréhension générale du langage acquise durant le pré-entraînement. Il faut cependant noter que *fine-tuner* un LLM dans un domaine spécifique requiert beaucoup de temps, de ressources informatiques et d'énergie (Taori et al., 2023 ; Touvron et al., 2023).

Dans le cadre de ce travail, nous comptons mettre à la disposition des partenaires du réseau PÉRISCOPE tout le potentiel de ces modèles d'IA très puissants, par la mise en place du robot conversationnel PÉRO, à l'instar de ChatGPT, afin de leur permettre d'interroger l'ensemble des publications disponibles au niveau du site internet de PÉRISCOPE. Ceci mettra en évidence les avancées du réseau pour la compréhension de la PRS et la résolution des problèmes qui lui sont afférents. Pour ce faire, l'apprentissage en contexte (*in-context learning* ou *Prompt engineering*) combiné avec le concept d'indexation, qui fait référence au processus d'organisation et de stockage des données d'une manière qui facilite et accélère la recherche et l'extraction d'informations pertinentes, est privilégié aux deux méthodes précédentes, étant donné la facilité de mise à jour d'un index par de nouvelles données, comparé au *fine-tuning* et au pré-entraînement qui, malgré la petitesse de la taille des nouvelles données, requièrent beaucoup de temps, de ressources informatiques et d'énergie (Taori et al. ; Touvron et al., 2023).

Le reste du document sera organisé comme suit. Dans la section 2, nous allons faire une revue de la littérature relative à l'application des grands modèles de langue, notamment ChatGPT, dans le domaine de l'éducation et plus particulièrement dans l'exploration des publications d'un centre de recherche afin de déceler leurs possibilités et limites par rapport à notre champ d'étude. Dans la section 3, nous allons détailler notre méthodologie de travail pour la mise en place de notre robot conversationnel. Dans cette partie, nous allons parler du fonctionnement de base de LlamaIndex. Nous allons ensuite, dans la section 4, évaluer ses performances par rapport à GPT-3.5 et à GPT-4 (Bing/CoPilot). Nous allons finir par une conclusion et dégager des perspectives de recherche.

Revue de littérature

Depuis son lancement en novembre 2022, ChatGPT a attiré une attention significative dans le milieu académique. Son impact dans divers domaines a été documenté à travers de nombreux articles scientifiques, qui explorent notamment ses applications potentielles dans la génération de texte, l'assistance virtuelle, la traduction automatique, et d'autres domaines connexes de l'intelligence artificielle. Son impact sur le monde universitaire et l'éducation est un domaine d'intérêt majeur. L'utilisation de ChatGPT dans le domaine de la recherche scientifique suscite un débat très animé entre universitaires du monde entier. Ainsi, les experts et les chercheurs s'interrogent sur l'utilité de ChatGPT et ses dérivés dans le domaine de la recherche scientifique et universitaire. Qasem (2023) a mis en exergue, dans ses travaux, les avantages et les craintes à propos de l'utilisation de ChatGPT aux niveaux scientifique et académique. Pour ce faire, il a interviewé sept experts en intelligence artificielle et en recherche scientifique. Ses résultats ont montré que ChatGPT a un potentiel important dans le domaine de la recherche scientifique et universitaire. En outre, Lund & Wang (2023) notent l'impact de ChatGPT dans le monde universitaire, notamment l'amélioration de la re-

cherche et de la découverte, la génération de contenu et de métadonnées, etc. Panda & Kaur (2023) ont exploré la viabilité de ChatGPT en tant qu'alternative aux systèmes de chatbot traditionnels dans les bibliothèques et les centres d'information. Leur étude donne un aperçu sur la portée de la mise en œuvre d'une telle solution et ses limites. Ils ont décrit également les possibilités futures pour l'utilisation de ChatGPT dans ces centres.

Selon eux, ChatGPT représente une avancée significative par rapport aux chatbots traditionnels, car il permet une conversation plus flexible en langage naturel. Les chatbots traditionnels s'appuient sur des règles et des réponses prédéfinies pour générer des réponses aux questions qu'on leur pose, ce qui peut limiter leur flexibilité (capacité d'un chatbot à s'adapter à une gamme variée de questions), leur évolutivité (capacité d'un chatbot, à s'adapter et à se développer pour prendre en charge une charge de travail croissante) et leurs capacités en matière de langage naturel. En revanche, ChatGPT est formé sur un large corpus de données et spécifiquement conçu pour générer des réponses en langage naturel, ce qui le rend plus flexible et adaptable aux différents besoins des utilisateurs. Les chatbots traditionnels ont également souvent du mal à répondre à des requêtes inattendues et non standard, alors que ChatGPT peut générer des réponses même si la question n'est pas formulée de manière habituelle. En outre, ChatGPT peut apprendre à partir de nouvelles données, alors que les chatbots traditionnels nécessitent une maintenance régulière pour s'assurer que les questions et les réponses soient à jour. Enfin, le coût de développement et de maintenance de ces derniers peut être important, alors que ChatGPT peut être formé à partir de données existantes et adapté à des tâches spécifiques, ce qui peut réduire les coûts. Sur la même lancée, Amber & Hashmi (2023) ont essayé d'identifier, dans leur étude, les efforts de recherche à propos de l'intégration de ChatGPT dans différents domaines de la vie. Cette revue de la littérature leur a permis d'identifier les domaines dans lesquels l'application de ChatGPT présente des lacunes en matière de recherche.

Parmi les domaines d'application, l'éducation se distingue particulièrement. Par conséquent, Lappalainen & Narayanan (2023) ont développé un chatbot personnalisé pour la bibliothèque de l'université Zayed des Émirats arabes unis. Leur étude est basée sur le langage de programmation Python et l'interface de programmation (*application programming interface* (API)) de ChatGPT. Leur chatbot, nommé Aisha, a été conçu pour fournir des services de référence et de soutien rapides et efficaces aux étudiants et aux professeurs en dehors des heures d'ouverture de la bibliothèque. Ils ont décrit le processus de développement, les capacités et les limites du robot. Ils ont aussi traité dans l'article les avantages des chatbots dans les bibliothèques universitaires et passé en revue les premières publications sur l'applicabilité de ChatGPT dans ce domaine. Adetayo (2023) a exploré, dans ses travaux, le potentiel des chatbots à base d'IA, en particulier ChatGPT, dans les bibliothèques universitaires. Selon lui, ChatGPT peut aider les usagers des bibliothèques à accéder aux documents pertinents en fonction de leurs besoins en recherche documentaire sans assistance humaine. Cependant, il a souligné, comme pour la plupart de ses pairs qui se sont intéressés à l'application de ChatGPT dans le monde universitaire et l'éducation, que des normes et des lignes directrices claires, utilisées de manière éthique et efficace, doivent être élaborées afin d'offrir la meilleure expérience possible aux utilisateurs et utilisatrices.

Malgré la pluralité des contributions dans l'application des IA génératives, plus particulièrement ChatGPT, dans le monde universitaire et en sciences de l'éducation, il n'existe pas de travaux, à notre connaissance, sur la mise en œuvre d'un chatbot basé sur ChatGPT et LlamaIndex dans l'optique de permettre aux partenaires d'un réseau de recherche de pouvoir explorer l'ensemble de leurs publications disponibles sur leurs plateformes. Dans ce qui suit, nous allons parler de notre méthodologie pour la mise en œuvre de notre chatbot basé sur ChatGPT (modèle 3.5) et LlamaIndex afin de permettre aux personnes de PÉRISCOPE de pouvoir interroger leurs publications, accessibles via le site du réseau, qui se consacrent à la persévérance et à la réussite scolaires.

Développement de PERO

Puisque le site du réseau PÉRISCOPE rassemble la grande majorité des publications des personnes engagées en recherche dudit réseau, sa plateforme recèle une grande quantité de données textuelles qui dépassent la limite de traitement de ChatGPT (8 192 tokens) (White et al., 2023). Pour surmonter cet obstacle et bien tirer profit de l'API de ChatGPT, nous avons utilisé Llamaindex (anciennement connu sous le nom de GPTIndex) (Liu, 2022) pour indexer l'ensemble des documents PDF disponibles au niveau de la plateforme PÉRISCOPE. LlamaIndex offre des outils permettant de récupérer des données à partir de différentes sources, telles que des fichiers PDF, des interfaces de programmation (API), des bases de données, et d'autres encore. Il crée des index, aussi bien sur les données structurées et non structurées, afin de faciliter l'apprentissage en contexte.

Extraction et indexation des données

Nous avons extrait toutes les publications en PDF disponibles sur le site suivant les thématiques de recherche du réseau PÉRISCOPE (participation dans la classe, participation dans l'école, relations école-famille-communauté et participation aux décisions de différentes instances).

Les publications extraites sont ensuite formatées en texte en enlevant les entêtes et pieds de page grâce à un script écrit en python. Nous nous sommes basés sur les travaux de Bast et Korzen (2017) qui ont fait un *benchmark* sur 12 098 articles scientifiques afin de comparer 14 outils de pointe qui permettent d'extraire le texte d'articles scientifiques en PDF. Leurs travaux ont montré que PDFact, l'outil développé par Bast & Korzen (2017), donne les meilleurs résultats en matière d'extraction de texte d'articles scientifiques en PDF.

Pour chaque fichier texte produit, les informations suivantes ont été ajoutées au tout début du fichier : 1) le titre de la publication ; 2) le type de publication (livre, chapitre de livre, article de colloque, article de revue, article de magazine, article de quotidien, communication, rapport, thèse ou document synthèse) ; 3) le nom des auteurs dans l'ordre (premier auteur, deuxième auteur, etc.) ; 4) l'année de publication ; et 5) les mots-clés associés à la publication. Pour trouver ces informations, nous avons recherché la publication concernée sur le site du réseau PÉRISCOPE dans la section « administrateur ». Cette section regroupe toutes les publications de la base de données du réseau. À noter que pour les mots-clés à ajouter au fichier texte, nous avons sélectionné les mots-clés cochés sur le site du réseau PÉRISCOPE en lien avec la publication en plus des mots-clés présents sur la publication elle-même, surtout dans le cas des articles de revues qui proposaient parfois des mots-clés supplémentaires à la liste de mots-clés du réseau PÉRISCOPE.

Avant de pouvoir interroger efficacement un LLM en langage naturel, il faut d'abord bien indexer son corpus (jeu de données). Les index jouent un rôle crucial dans l'organisation et l'extraction efficace des données, en particulier lorsque l'on travaille avec de grands ensembles de données non structurées ou de structures complexes. Dans ce contexte, nous avons utilisé LlamaIndex pour indexer nos données (les différents fichiers texte produits). Il offre différents types d'index (liste, vecteur, arbre, table de mots-clés, index de résumé, etc.), chacun ayant une fonction différente dans l'organisation des données et la recherche d'information. Nous allons, dans ce qui suit, faire une brève présentation de LlamaIndex afin de justifier notre choix du type d'index.

Brève présentation des types d'index de LlamaIndex

Connu également sous le nom de GPT Index, LlamaIndex (Liu, 2022) est une interface qui permet de connecter des données externes aux grands modèles de langue via des connecteurs qui peuvent s'intégrer à diverses sources et formats de données structurées ou non structurées (API, PDF, SQL, documents, etc.). Grâce à LlamaIndex, un large éventail de types de données, dont les documents PDF, sont indexés. Pour ce faire, il divise les documents en plusieurs nœuds qui sont des morceaux

de texte du document source. Ces morceaux contiennent aussi des métadonnées et des informations sur les relations avec d'autres nœuds. LlamaIndex offre plusieurs types d'index.

L'index de liste (*GPTListIndex*) est une structure de données simple où les nœuds sont stockés dans une séquence. Lors de la construction de l'index, les textes du document sont regroupés, convertis en nœuds et stockés dans une liste. Lors de la construction de l'index, le LLM n'est pas appelé pour générer l'incorporation (*embedding*), mais le génère au moment de la requête. Pour répondre à une question donnée, LlamaIndex charge tous les nœuds de la liste dans le module de synthèse de réponse. Par défaut, LlamaIndex utilise *text-davinci-003* pour synthétiser la réponse. Ce type d'index est généralement utilisé pour synthétiser une réponse qui combine des informations provenant de plusieurs sources de données.

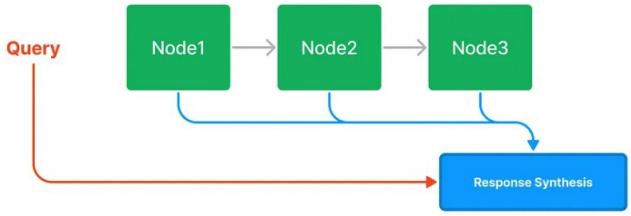

Figure 1: GPTListIndex

L'index vecteur (*GPTVectorStoreIndex*), quant à lui, permet de répondre à une requête sur un grand corpus de données. Contrairement à *GPTListIndex*, *GPTVectorStoreIndex* génère des intégrations (*embedding*) lors de la construction de l'index. Cela signifie que le point de terminaison (*endpoint*) du LLM sera appelé lors de la construction de l'index pour générer des données d'incorporation (*embedding data*). Pour répondre à une question, il récupère les nœuds les plus similaires et les transmet au module de synthèse de réponse.

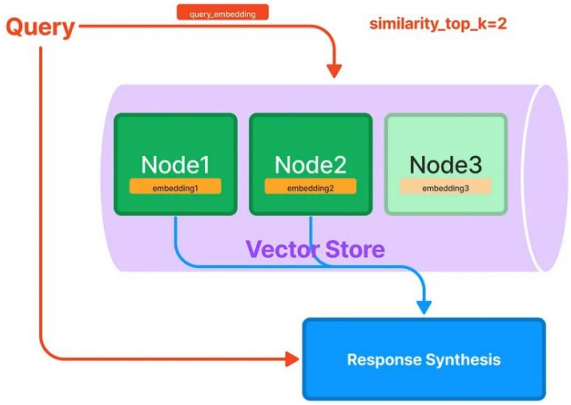

Figure 2: GPTVectoreStoreIndex

L'index d'arbre est structuré en arbre hiérarchique à partir d'un ensemble de nœuds. Chaque nœud est un résumé de ses nœuds enfants. Lors de la construction de l'index, l'arbre est construit de manière ascendante. Pour répondre à une requête, il traverse les nœuds racines jusqu'aux nœuds feuilles. Dépendamment de la valeur du *child branch factor*, on détermine, pour chaque nœud parent, le nombre de nœuds enfants qui seront pris en compte dans la réponse d'une requête. Contrairement à l'index vectoriel, l'index d'arbre génère l'incorporation (*embedding*) au moment de la requête. Les *embeddings* sont mises en cache si *retriever mode* = "*embedding*" est spécifié pendant la requête.

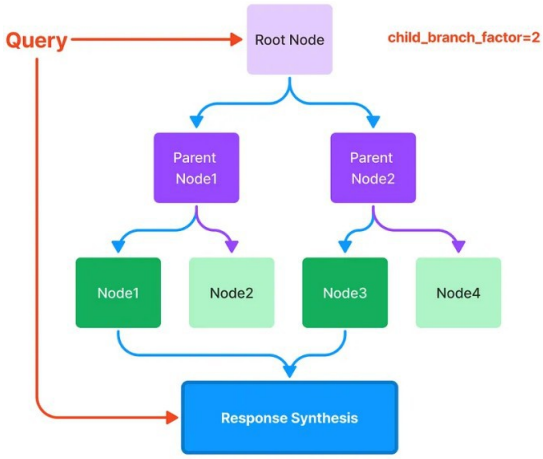

Figure 3:GPTTreeIndex

L'index de table de mots-clés (*GPTKeywordTableIndex*) extrait les mots-clés de chaque nœud et crée un mappage de chaque mot-clé vers les nœuds correspondants de ce mot-clé. Pour répondre à une question, il extrait les mots-clés pertinents de la requête et les fait correspondre aux mots-clés de chaque nœud afin de récupérer les plus pertinents qui seront, par la suite, transmis au module de synthèse de réponse. *GPTKeywordTableIndex* nécessite des appels LLM pendant la construction de l'index, mais, avec l'utilisation de *GPTSimpleKeywordTableIndex* qui se base sur une *regex* (expression régulière) pour extraire les mots-clés de chaque document, on peut s'en passer.

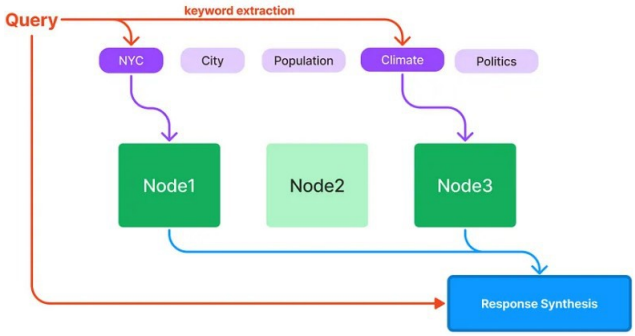

Figure 4 : GPTKeywordTableIndex

L'index résumés de documents *GPTDocumentSummaryIndex* est une nouvelle structure offerte par LlamaIndex. Elle est conçue pour des chatbots spécialisés sous un format question/réponse. Ce type d'index a été introduit pour pallier les limites qui peuvent impacter les performances des autres types d'index. Entre autres limites, on peut noter que :

- les nœuds (blocs de texte) n'ont pas un contexte global du corpus, ce qui peut limiter l'efficacité du processus de question/réponse;
- le choix du score de similarité nécessite un bon ajustement, car une valeur petite peut influencer la pertinence du contexte alors qu'une valeur grande peut augmenter le coût et la latence de la construction de l'index ; ça peut aussi influencer la pertinence du contexte ;
- pour une question donnée, les *embeddings* ne sélectionnent pas toujours le contexte le plus approprié.

Pour apporter des solutions à ces problèmes, certains développeurs ajoutent des filtres par mots-clés, ce qui reste toujours un défi, car il n'est pas facile de déterminer les mots-clés appropriés pour chaque document. Ça nécessite des efforts manuels ou l'utilisation de modèles NLP d'extraction de mots-clés. C'est ce qui a poussé les concepteurs de LlamaIndex à mettre en place l'index de résumé

de documents qui peut extraire et indexer le résumé de chaque document. Ainsi, l'index va contenir plus d'informations qu'un seul bloc de texte. Par conséquent, il a plus de signification sémantique. Pendant la construction de l'index, le LLM est utilisé pour extraire un résumé de chaque document. Pour répondre à une question donnée, il se base sur les résumés pour déterminer les documents pertinents.

Cette précédente étude nous a permis d'avoir une idée plus claire quant aux différents types d'index possibles avec LlamaIndex. Ceci nous a amenés à pouvoir déceler les points forts de chacun par rapport aux autres, cela afin de pouvoir motiver le type d'index utilisé dans le cadre de ce travail. Notre choix s'est porté sur l'index vecteur (*GPTVectorStoreIndex*), car il permet de retrouver efficacement des articles dont le contenu et le contexte sont similaires. Cela est particulièrement utile lors de la recherche d'articles liés à un sujet ou à un domaine de recherche spécifique.

Fonctionnement de LlamaIndex

Étant entrainés sur une grande quantité de données, les grands modèles de langue (LLM) fonctionnement très bien. Cependant, ils présentent des limites pour des questions portant sur des données récentes ou un domaine spécifique. Le *fine-tuning* est souvent utilisé pour permettre aux LLM de prendre en compte des données d'un domaine spécifique. Il consiste à entraîner les poids (paramètres) du modèle sur une nouvelle petite quantité de données spécifiques. Bien que la quantité de données soit petite, cela demeure un processus lourd, surtout s'il s'agit d'un LLM comme GPT-3 avec 175 milliards de paramètres. Une alternative à cette solution est l'apprentissage en contexte qui consiste à concevoir l'invite de commande (*prompt*), une technique utilisée pour spécifier le contexte ou la nature de la réponse attendue de la part du modèle de langage, d'une certaine manière afin de fournir les nouvelles données aux LLM pour qu'elles soient prises en compte par ces derniers dans leurs réponses, d'où son nom (*prompt engineering*). Le fonctionnement de LlamaIndex est basé sur cette solution. Il fournit des connecteurs à diverses sources de données qui peuvent être analysées et stockées en index sous forme de fichier ou de base de données vectorielle. Les index sont par la suite interrogés et les résultats servent de contexte au LLM dans sa réponse aux questions.

Implémentation technique de notre solution PÉRO

Dans le cadre de la présente démarche, les publications issues de la plateforme PÉRISCOPE constituent un grand corpus qui dépasse les limites de traitement actuel de ChatGPT (8 192 tokens pour GPT-4) (White et al., 2023). Ainsi, pour bien tirer profit du grand modèle de langage GPT-3.5 de ChatGPT, nous avons utilisé LlamaIndex comme solution alternative. Ceci nous a permis d'indexer cette grande quantité de données et de pouvoir les interroger via le modèle de ChatGPT sans aucune contrainte. En effet, pour que nos données puissent être compatibles avec l'API de ChatGPT, nous avons créé un index englobant toutes les publications issues de la plateforme PÉRISCOPE. Pour ce faire, nous nous sommes basé sur quatre classes de LlamaIndex: *set global service context*, *ServiceContext*, *StorageContext*, *load index from storage*. Nous avons généré des vecteurs pour chaque document avec la classe *VectorStoreIndex*. Pour ne pas définir manuellement les arguments des mots-clés (*kwargs*), nous avons utilisé la méthode ServiceContext.from defaults. Les vecteurs générés ont été stockés dans l'index en format JavaScript Object Notation (JSON). Ainsi, grâce au moteur de recherche de LlamaIndex qui exploite en arrière-plan la puissance de ChatGPT, les partenaires du réseau PÉRISCOPE, via un *prompt* mis à leur disposition, peuvent interroger l'ensemble des publications disponibles sur PÉRISCOPE. La figure 5 illustre le pipeline complet que nous avons mis en œuvre dans cette étude. Les détails d'implémentation sont disponibles sur ce dépôt GitHub (détails de l'implémentation).

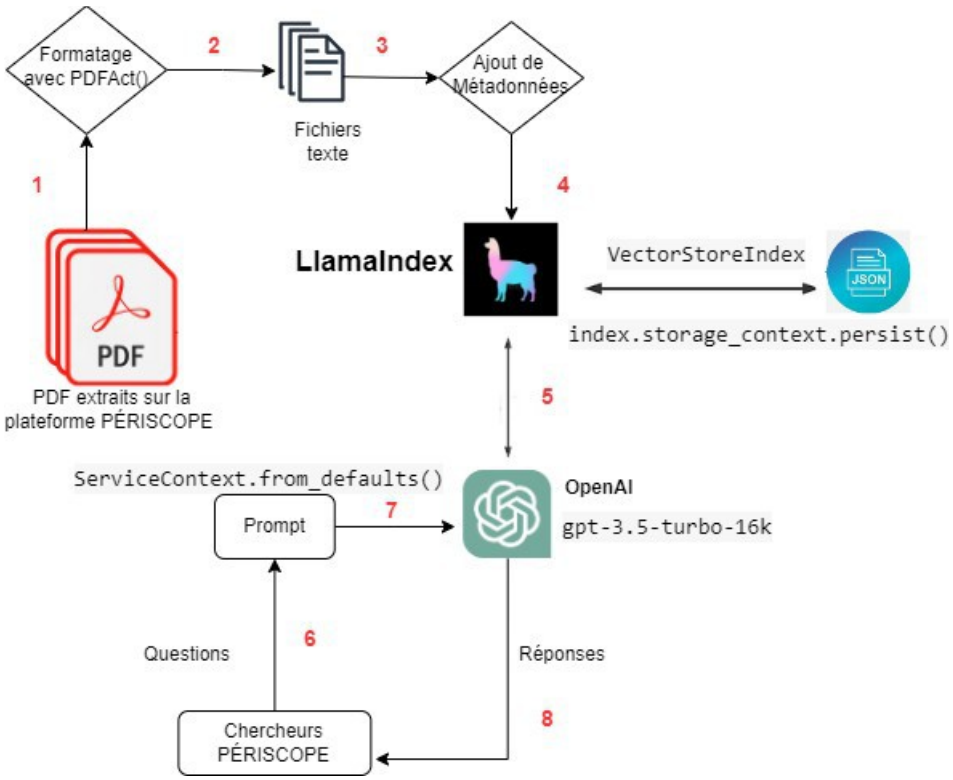

Figure 5 **:** ***Flux de travail (*****Workflow*****)***

Évaluation des performances

L'évaluation des performances d'un grand modèle de langue (LLM) implique de mesurer à quel point il réussit les tâches pour lesquelles il a été conçu. Il existe plusieurs méthodes et métriques pour évaluer les performances d'un LLM (Banerjee et al., 2023). Il est important de noter qu'utiliser une combinaison de méthodes et de métriques peut donner une image plus complète de ses performances. Dans le cadre de notre travail, afin de bien évaluer les performances de notre solution, nous avons combiné deux méthodes : 1) une analyse qualitative des réponses, qui consiste à examiner attentivement les sorties générées par le LLM pour détecter d'éventuelles erreurs, incohérences, informations incorrectes ou biais et qui peut nécessiter une évaluation humaine approfondie ; et 2) une comparaison avec d'autres modèles qui consiste à comparer les performances du LLM avec d'autres modèles disponibles ou des versions antérieures du modèle pour voir s'il y a eu une amélioration significative.

Nous avons comparé les réponses générées par notre chatbot, celles générées par ChatGPT-3.5 (https://chat.openai.com/) et celles générées par Bing (https://www.bing.com/) pour les mêmes questions, tout en considérant l'aspect crucial de l'évaluation de la performance des grands modèles de langage (LLM) dans la génération de réponses pertinentes, précises et contextuellement appropriées aux requêtes dqui sont formulées.

Après avoir mis en place notre robot conversationnel, nous avons demandé à la responsable du réseau PÉRISCOPE de préparer dix questions, à des fins de test, relatives au contenu des publications en rapport avec l'engagement et le langage écrit, soit un concept et un sous-domaine étudié s'agissant de persévérance et de réussite scolaires. Nous avons mis à sa disposition un démo de notre chatbot via une interface web réalisée avec Gradio, un *package* Python Open Source (Abid et al., 2019). Deux questions ont été retirées vu leur ambiguïté, les juges les ayant interprétées de manière opposée (par ex., en accordant une valeur très basse ou très élevée) à la réponse fournie par le chatbot.

Pour chaque réponse, les juges, qui avaient des profils diversifiés allant d'une connaissance approfondie de l'engagement de l'élève dans la classe à des professionnels de terrain, ont fourni un score de satisfaction en pourcentage dépendamment de leur degré de satisfaction sur la réponse générée par le chatbot interrogé.

Questions	Scores PÉRO (%)	Scores ChatGPT (%)	Scores Bing (%)
Q1	84	85	58
Q2	85	91	85
Q3	91	95	87
Q4	81	92	73
Q5	77	85	83
Q6	61	46	98
Q7	65	82	77
Q8	39	44	56

Tableau 1. Scores de satisfaction de notre solution (PÉRO, ChatGPT et BIng)

On observe que le degré de satisfaction au regard de PÉRO et de Bing, quatre questions ont obtenu un score moyen supérieur à 80 % et cinq dans le cas de ChatGPT. Trois questions posées à PÉRO ont obtenu un score de ou inférieur à 65 %, deux questions posées à ChatGPT et deux questions posées à Bing. Il ressort de cette évaluation de performances que PÉRO a des résultats du même ordre que les autres (ChatGPT et Copilot).

Toutefois, ces pourcentages présentent un problème de validité vu la taille restreinte du nombre d'experts consultés, composé uniquement de sept individus ayant exercé le rôle de juges). C'est là une limite de cette étude.

Les sept juges ont aussi fourni des remarques quant aux éléments à améliorer dans les réponses offertes par les chatbots : 1) références précises et pertinentes ; 2) définitions plus complètes ; 3) clarifications ou explicitations et 4) utilisation de la langue appropriée. En ce qui concerne PÉRO, il est principalement fait mention qu'il manquait d'argumentation dans ses réponses, surtout pour les questions fermées et qu'il avait tendance à répondre en anglais lorsqu'il ne trouvait pas de réponse sur le site PÉRISCOPE. ChatGPT ne répondait pas directement à la question. Bing fournissait des réponses exactes partant d'une ou de quelques références et quelquefois des réponses erronées.

Conclusion et perspectives

Dans cet article, nous avons rapporté le développement (phase 1) d'un robot conversationnel (PÉRO) pour le réseau PÉRISCOPE afin de permettre à toute personne intéressée d'interroger l'ensemble des publications disponibles sur le site internet de PÉRISCOPE. Les résultats sont suffisamment encourageants sous l'angle de la preuve de concept pour poursuivre ce développement. Ainsi, grâce à LlamaIndex, il a été possible de surmonter l'obstacle lié à la limite de traitement de données textuelles pouvant être soumises à ChatGPT via son API et le modèle GPT-3.5. Pour évaluer les performances de PÉRO (notre solution), les juges ont comparé ses réponses générées avec celles de ChatGPT et Bing pour les mêmes questions relatives au contenu des publications en rapport avec l'engagement et le langage écrit, soit un concept et un domaine de recherche qui

instruisent sur la manière dont les individus apprennent, communiquent, développent leurs compétences et, en conséquence, améliorent les méthodes d'enseignement et d'apprentissage. Les résultats montrent que PÉRO arrive déjà à faire quasiment aussi bien que les grosses IA génératives. Nous concluons qu'il s'agit-là d'une preuve de concept acceptable quant à la faisabilité de travailler avec des données plutôt restreintes et ancrées dans un domaine ou sous-domaine de recherche particulier.

En phase 2, il est prévu d'évaluer la performance de PÉRO sur l'ensemble des concepts et domaines de recherche associés à la persévérance et à la réussite scolaires. Étant donné que les travaux de la phase ont été basés sur l'API gpt-3.5- turbo-16ko d'OpenAI, nous prévoyons mettre en place notre propre modèle sur nos propres serveurs en se basant sur un grand modèle de langue Open Source (Mistral, Llama 2, etc.). Nous prévoyons aussi augmenter la taille de l'ensemble des données utilisées en incluant les vidéos disponibles sur le site de PÉRISCOPE, car elles regorgent d'informations à exploiter.

Références

Abid, A., Abdalla, A., Abid, A., Khan, D., Alfozan, A., & Zou, J. (2019). Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*.

Adetayo, A. J. (2023). Artificial intelligence chatbots in academic libraries: The rise of chat-gpt. *Library Hi Tech News*, *40*(3), 18–21.

Amber, Q., & Hashmi, F. A. (2023). Chatgpt revolution in different fields of life: Literature synthesis and research agenda. *Available at SSRN 4451821*.

Banerjee, D., Singh, P., Avadhanam, A., & Srivastava, S. (2023). Benchmarking llm powered chatbots: Methods and metrics. *arXiv preprint arXiv:2308.04624*.

Bast, H., & Korzen, C. (2017). A benchmark and evaluation for text extraction from pdf.

2017 ACM/IEEE joint conference on digital libraries (JCDL), 1–10.

Engeström, Y. (1987, 2015). *Learning by expanding*. Cambridge University Press.

Floridi, L., & Chiriatti, M. (2020). Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*, 681–694.

Lappalainen, Y., & Narayanan, N. (2023). Aisha: A custom ai library chatbot using the chatgpt api. *Journal of Web Librarianship*, 1–22.

Liu, J. (2022). LlamaIndex. https://doi.org/10.5281/zenodo.1234

Lund, B. D., & Wang, T. (2023). Chatting about chatgpt: How may ai and gpt impact academia and libraries? *Library Hi Tech News*, *40*(3), 26–29.

Panda, S., & Kaur, N. (2023). Exploring the viability of chatgpt as an alternative to traditional chatbot systems in library and information centers. *Library Hi Tech News*, *40*(3), 22–25.

Qasem, F. (2023). Chatgpt in scientific and academic research: Future fears and reassurances. *Library Hi Tech News*, *40*(3), 30–32.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., & Hashimoto, & T. B. (2023). Stanford alpaca: An instruction-following llama model.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozie`re, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

White, J., Hays, S., Fu, Q., Spencer-Smith, J., & Schmidt, D. (2023). Chatgpt prompt patterns for improving code quality, refactoring, requirements elicitation, and software design. *arXiv preprint arXiv:2303.07839*.