Université Hassan 2 De Casablanca **ENSAM Casablanca**

SUJET

ANALYSE DES TENDANCES DE RECHERCHE EN INTELLIGENCE ARTIFICIELLE GENERATIVE VIA LA FOUILLE DE TEXTES

Master Big Data et Internet Des Objets

Réalisé par

Fanid Yassmine

Majdany Ikram

Encadrant

Mr Bahassine Said

Année universitaire : 2024 / 2025

RESUME

L'essor rapide de l'intelligence artificielle générative (IAG), porté par des avancées majeures telles que ChatGPT, GPT-3 et DALL-E, a suscité une multitude de recherches dans des domaines variés, allant du traitement du langage naturel à la vision par ordinateur, en passant par l'éducation et le droit. Cette explosion rend difficile le suivi des évolutions et des tendances scientifiques. Notre étude vise à exploiter des techniques de fouille de texte sur un corpus scientifique construit à partir des bases ouvertes arXiv et OpenAlex, afin d'analyser les tendances de recherche en IAG, d'identifier les thématiques émergentes, de suivre leur évolution entre 2019 et 2024, et de repérer les pays, institutions et auteurs les plus influents.

Nous avons collecté, fusionné et nettoyé un corpus de 3437 articles uniques. Après un prétraitement et une vectorisation TF-IDF avec sélection des caractéristiques, nous avons appliqué deux méthodes complémentaires de modélisation thématique (LDA et BERTopic) ainsi qu'un clustering K-Means pour structurer les données. Des analyses exploratoires ont été conduites pour examiner la dynamique temporelle, géographique et institutionnelle, ainsi que le réseau de collaboration entre auteurs et les citations.

L'analyse révèle une croissance rapide des publications et l'émergence de nouveaux concepts tels que generative, chatgpt, llm et diffusion model, accompagnée d'une interdisciplinarité croissante vers la médecine et les sciences sociales. Six thématiques principales ont été identifiées : applications médicales, systèmes d'IA générative, génération d'images, grands modèles linguistiques, réseaux neuronaux et enjeux sociaux liés à ChatGPT. BERTopic a également mis en lumière des sujets innovants liés aux matériaux et à l'énergie renouvelable. Le clustering a distingué un groupe de publications techniques spécialisées et un autre plus généraliste.

Sur le plan géographique, les États-Unis et la Chine dominent largement, avec des institutions phares telles que Stanford, MIT, Tsinghua University, Google Research et Microsoft Research. L'analyse des auteurs met en avant des chercheurs centraux comme Dacheng Tao, et révèle l'existence de communautés de recherche fortement connectées. Enfin, les articles les plus cités concernent principalement les avancées en modèles de diffusion multimodale et en grands modèles linguistiques, confirmant leur rôle moteur dans l'évolution de l'IAG.

ABSTRACT

The rapid rise of generative artificial intelligence (GAI), driven by major advances such as ChatGPT, GPT-3 and DALL-E, has triggered a massive research effort in fields ranging from natural language processing to computer vision, education, and law. This explosive growth makes it difficult to track scientific trends and developments. Our research aims to apply text mining techniques to scientific corpora in the open-access archives arXiv and OpenAlex to analyze research trends in GAI, identify emerging themes, track its development between 2019 and 2024, and identify the most influential countries, institutions, and authors. We collected merged, 3437 articles. Then we applied preprocessing to the text, resulting with 3437 unique articles. Furthermore, in order to prepare the text in a format that NLP models can manipulate, we used TF-IDF vectorization along with feature selection and dimension reduction. Moreover, we applied two complementary topic modeling approaches (LDA and BERTopic) and K-Means clustering to structure the data. Our study combined different types of analysis such as : exploratory analyses, temporal analyses, geographic analyses, and institutional analyses as well as collaboration networks among authors and citations.

The analysis reveals a rapid growth in publications and the emergence of new concepts such as "generative","chatgpt","llm,"and "diffusion model," accompanied by increasing interdisciplinarity towards medicine and social sciences. Six main topic have been identified by LDA: medical applications, generative AI systems, image generation, large language models, neural networks, and social impacts related to ChatGPT. BERTopic has also highlighted innovative topics related to material Sciences and renewable energy. The clustering distinguished a group of specialized technical publications and a more general one.

Geographically, the United States and China dominate, with leading institutions such as Stanford, MIT, Tsinghua University, Google Research, and Microsoft Research. The analysis of authors highlights central researchers like Dacheng Tao and reveals the existence of strongly connected research communities. Finally, the most cited articles mainly concern advances in multimodal diffusion models and large language models, confirming their driving role in the evolution of GAI.

LISTE DES FIGURES

LISTE DES TABLES

TABLE DES MATIERES

I. INTRODUCTION

1. CONTEXTE DU PROBLEME

Le jaillissement qu'a connu le domaine de l'intelligence artificielle générative (IAG) notamment avec l'apparition de ChatGPT, GPT-3 et DALL-E, constitue le sujet principal d'une multitude de recherches scientifiques qui s'y intéressent vu que l'IAG touche plusieurs domaines tel que NLP, Computer Vision, éducation, droit...etc, et s'applique à divers types de données (texte, image, son...etc). Ce jaillissement rend difficile le suivi et l'organisation de ce tas d'informations afin de comprendre les tendances et l'évolution technique de ce sujet.

2. OBJECTIFS

Conjointement à l'explosion des recherches scientifiques relatives au domaine de l'intelligence artificielle générative, les objectifs de cette étude concernent l'application des techniques de fouille de texte sur un corpus scientifique créé à partir de bases de données ouvertes à savoir arXiv et OpenAlex afin d'analyser les tendances de recherche dans ce domaine, de détecter les thématiques émergentes en exploitant les métadonnées disponibles, d'analyser l'évolution temporelle des sujets de recherche de l'année 2019 à 2024 et d'identifier les pays, les institutions ainsi que les auteurs prépondérants.

3. PROBLEMATIQUE ET HYPOTHESES

Considérant les hypothèses suivantes :

-La croissance exponentielle du volume de publications sur l'IAG notamment après l'essor des grands modèles de langage GPT-2, GPT-3...etc.

-Les publications dans ce sujet sont produites principalement par une minorité de pays et d'institutions régnantes.

Nous aborderons les problématiques suivantes :

RQ1 : Quels sont les domaines ou sous-domaines qui connaissent une croissance rapide en IA générative ?

RQ2 : Comment les sujets de recherche évoluent-ils dans le temps depuis l'année 2019 à l'année 2024 ?

RQ3 : Quels sont les mots-clés, concepts ou modèles les plus fréquents ou émergents?

RQ4 : Existe-t-il des clusters thématiques ou des communautés de recherche identifiables ?

RQ5 : Qui pilote cette recherche et quels sont les pays, institutions ou auteurs dominants ?

RQ6 : Peut-on détecter des topics émergents ou déclinants ?

4. METHODOLOGIE GENERALE

Dans l'intention d'atteindre les objectifs de cette étude et de répondre à ses problématiques, nous avons procédé initialement par la collecte des données à partir de deux sources de bases de données populaires à noter arXiv et OpenAlex. Le processus de leur prétraitement - qui englobe le nettoyage, la vectorisation TF-IDF, la sélection des features les plus pertinents ainsi que la réduction de la dimensionnalité - nous a permis d'obtenir après leur fusion un corpus de 3437 articles, sur lesquels nous avons appliqué une approche hybride de méthodes qualitatives à savoir les topic modeling Latent Dirichlet Allocation (LDA) et BERTopic en plus de l'analyse des citations et du nuages de mots et de méthodes quantitatives telles que l'analyse temporelle, la méthode TF-IDF et le clustering à l'aide du modèle K-Means.

II. ÉTAT DE L'ART

1. TRAVAUX PRECEDENTS SUR LE DOMAINE

Article 1: "*Text Mining Approaches for Exploring Research Trends in the Security Applications of Generative Artificial Intelligence"*

Le premier article que nous avons étudié est l'intitulé « Text Mining Approaches for Exploring Research Trends in the Security Applications of Generative Artificial Intelligence » de Jinsick Kim, Byeongsoo Koo, Moonju Nam and Youngseo Song, Kukjin Jang, Jooyeoun Lee Myoungsug Chung, publié en 2025, et tiré de la revue scientifique « Applied Sciences ». Il aborde le sujet d'analyse des dernières tendances de recherches sur les implications de sécurité de l'intelligence artificielle générative dans divers domaines, notamment l'éducation, le commerce, la finance, la santé…etc.
Certes, avec l'émergence de l'intelligence artificielle générative ces dernières années, et leurs applications multidisciplinaires, plusieurs préoccupations ont été soulevées concernant le déploiement sécurisé des systèmes d'IA. Cela souligne la nécessité de mettre en place des cadres de sécurité solides combinant des solutions techniques, une responsabilité éthique et une conformité réglementaire. De ce fait, de nombreuses recherches ont été élaborées afin de répondre efficacement aux questions de sécurité liées à ces technologies.
La présente étude vient pour compléter les anciennes recherches qui n'ont utilisé que des méthodes qualitatives et conceptuelles pour leurs analyses, chose qui a limité leur capacité à évaluer de manière exhaustive de grands volumes de recherche.
Pour cela, cette étude exploite les techniques de la fouille de texte en l'occurrence TF-IDF et Keyword Centrality Analysis, pour analyser 1047 articles sur la sécurité des LLMs (Large Language Models), issus de la base de données SCOPUS. Ces méthodes scientométriques sont des outils quantitatifs qui offrent une approche structurée et axée sur les données pour comprendre les relations entre les articles de recherches par des représentations graphiques. Ainsi, LDA Topic Modeling est aussi utilisée afin d'extraire des insights sémantiques à partir de grands corpus de texte, en utilisant des méthodes statistiques et d'optimisation. La combinaison de ces techniques ensembles a pour

objectif d'identifier les grands clusters de recherche, les structures thématiques, et les zones peu explorées dans les recherches sur la sécurité des LLMs.
Les résultats ont montré qu'au cours des deux dernières années, l'intérêt des recherches en matière de sécurité de l'intelligence artificielle générative a augmenté de 2 publications en 2022, à 540 publications en 2023, et 504 publications en première moitié de 2024. En ce qui concerne la répartition géographique, Les Etats-Unis, la Chine et l'Inde, sont les grands contributeurs dans cette matière de recherche. La participation globalisée des grandes universités de différents pays et continents, favorisent le développement des systèmes IA générative respectant des normes de sécurités mondiales. Des institutions comme « The Chinese Academy Of Sciences » avec plus de 20 publications, et « Ministère de l'Éducation de la République populaire de Chine », se distinguent par leur dominance dans les recherches sur l'IA générative.
De plus, l'analyse des mots clés a montré que les termes « ChatGPT », « AI », « Model », « Large Language Models » sont au centre des discussions académiques sur la sécurité. En outre, l'étude a identifié les six thèmes suivants :
L'IA dans l'éducation : l'impact sur l'apprentissage et l'intégrité académique, la triche assistée par l'IA générative.
La sécurité des modèles linguistiques : les attaques par injection de prompt et la fuite des informations sensibles.
Le développement logiciel sécurisé : la génération de code en s'assurant de la robustesse et l'absence des failles.
Gestion du risque en système IA générative : Les stratégies proactives de détection et d'atténuation des menaces sont essentielles pour gérer les risques.
Confidentialité et protection des données personnelles : problématiques liées au traitement des données sensibles.
Sécurité dans le domaine de la santé : l'utilisation des modèles pour le diagnostic médical tout en protégeant la vie privée des patients.
Enfin, l'article a mentionné quelques recommandations comme l'adoption d'une approche interdisciplinaire, l'expansion des sources de données notamment des documents non anglophones, et le suivi d'une approche combinant à la fois des méthodes quantitatives et qualitatives afin d'offrir une vision plus complète des défis de sécurité liés à l'IA générative.

Limites :
L'étude présente plusieurs limitations, notamment l'utilisation des articles en langue anglaise, extraits d'une seule revue SCOPUS, ce qui pourrait exclure des recherches pertinentes publiées dans d'autres bases, ou d'autres langues. De plus, la période d'analyse ne tient compte que des recherches récentes (de juin 2022 à juin 2024) sur la sécurité de l'IA. Par ailleurs, l'intelligence artificielle générative évolue très rapidement ce qui pourraient rendre certaines conclusions obsolètes. En outre, les méthodes quantitatives seules peuvent manquer de nuance dans l'interprétation des contenus riches et complexes. Enfin, l'analyse aurait gagné en richesse en intégrant des sources issues de domaines juridiques, sociétaux et industriels.

Article 2: "*Landscape of Generative AI in Global News: Topics, Sentiments, and Spatiotemporal Analysis"*

Le deuxième article rédigé par Lu Xian, Lingyao Li, Yiwei Xu, Ben Zefeng Zhang et Libby Hemphill, intitulé "Landscape of Generative AI in Global News: Topics, Sentiments, and Spatiotemporal Analysis", publié en 2024 et tiré de la plateforme en ligne de prépublications scientifiques arXiv, traite de la manière dont les médias couvrent l'intelligence artificielle générative et l'influence de cette technologie sur le public.

L'étude objet de cet article comble les lacunes observées dans les recherches antérieures qui se concentraient sur la couverture médiatique de l'IA uniquement et ce en collectant des articles issus de la base de données Newsstream de ProQuest parus de Janvier 2018 à Novembre 2023 en cohérence avec la sortie de BERT. La collecte de ces données a permis la récupération de 24 827 articles (initialement 38 199) après filtrage linguistique - en ne conservant que les articles en anglais -, suppression de doublons et filtrage de pertinence en associant PaLM avec Google pour éviter de confondre certains acronymes et mots communs (palm et PaLM). Ces articles ont été profilés en trois grandes catégories : journaux nationaux américains, journaux locaux et spécialisés américains et journaux internationaux.

Dans cette étude trois grandes questions ont été soulevées relatives à quand et où les articles de différents sujets ont été publiés, quel est le sentiment des articles et de quoi ils parlent. Diverses techniques ont été adoptées : BERTopic pour sa capacité à capturer le contexte des mots, UMAP pour la réduction de la dimensionnalité des vecteurs produits par BERT, la méthode du coude combinée à K-Means pour garantir une granularité fine des thématiques avec k=100.Chaque cluster étant ensuite représenté par ses mots clés les plus significatifs à l'aide de c-TF-IDF. Ensuite, ces 100 clusters ont été regroupés en 10 grands thèmes à l'aide du codage qualitatif combinant l'approche bottom-up et top-down et afin d'évaluer l'attitude générale des articles envers ces thèmes, une analyse de sentiment a été appliquée à l'aide du modèle RoBERTa-base.

Les techniques suscitées ont produit des résultats diversifiés avec une concentration sur le développement technologique des entreprises (19%), sur la régulation et sécurité (16%), sur l'éducation (13%) et sur les affaires (10%). En réponse à la première question les chercheurs ont constaté que le volume des articles représente des pics lors de lancement de produits – pic de 12 au 19 Février 2023 en liaison avec l'introduction de Bard par Google (6 Février) et Bing par Microsoft (7 Février) et ont noté que la distribution géographique se concentre sur cinq principaux pays : Etats-Unis, Inde, Royaume-Uni, Australie et Canada. Quant à la deuxième question, la majorité des textes sont neutres (66%) ou positifs (28%) avec une minorité d'articles exprimant un sentiment négatif notamment ceux en liaison avec la réglementation et la sécurité de plus les médias locaux américains sont plus positifs que les internationaux confirmés par un test de Kruskal-Wallis. En analysant les réseaux sémantiques, les chercheurs ont pu répondre à la dernière question, cette analyse a révélé des clusters distincts pour les affaires, le développement technologique (grandes entreprises et innovation), la

régulation et sécurité (débats éthniques et cybersécurité) et l'éducation.

Limites :
Bien que complète, cette étude connait des limites en premier lieu dans le domaine des données côté linguistique en se concentrant seulement sur les articles en anglais, côté accès aux données qui proviennent de ProQuest et qui sont limitées par les abonnements universitaires, et côté représentativité des nouvelles nationales qui est partielle et limitée aux grands journaux nationaux. En second lieu, dans le domaine de la modélisation des sujets par la simplification excessive des contenus complexes via BERTopic qui suppose que chaque article traite d'un seul sujet, et en dernier lieu dans le domaine de l'analyse de sentiments où le modèle RoBERTa-base peut réduire la précision des résultats car il n'est pas entraîné spécialement pour l'analyse des sentiments des articles de presse.

Article 3: "*Characterizing generative artificial intelligence applications: Textmining-enabled technology roadmapping"*

Cet article intitulé "Characterizing generative artificial intelligence applications: Textmining-enabled technology roadmapping" oeuvre de Shiwangi Singh, Surabhi Singh, Sascha Kraus, Anuj Sharma et Sanjay Dhir édité au journal "Journal of Innovation & Knowledge" en 2024, traite de l'analyse des tendances technologiques en intelligence artificielle générative à l'aide d'une fouille de textes basée sur les brevets. Les rédacteurs de cet article ont opté pour l'utilisation du Structural Topic Model (STM) comme technique pour l'extraction de sujets latents à partir d'un corpus de 2398 brevets déposés de 2017 à 2023, sachant que l'année 2017 choisie comme date de départ coïncide avec des innovations croissantes dans le domaine des modèles génératifs. Le processus de collecte des brevets extraits à partir de diverses bases de données comme USPTO, EPO, JPO a débuté par l'identification de 2985 brevets qui ont été réduits après filtrage et sélection à partir de mots-clés tel que "GPT", "DALL-E", "LLMs" etc à 2398 brevets. Cette opération de filtrage a été basée sur la suppression des caractères non anglais, des mots vides et des noms de pays, et l'adoption de la méthode du Tokenizer n-gram qui a permis de conserver les relations sémantiques en réduisant les bigrammes et trigrammes en unigrammes.
La technique du STM a été privilégiée à celle du LDA pour sa capacité à intégrer des covariables documentaires. Après évaluation des scores d'exclusivité, de cohérence sémantique et de la vraisemblance des documents non inclus, le nombre de sujets a été fixé à six vu que la cohérence sémantique chutait lorsque le nombre de sujets dépassait six ce qui a assuré précision et cohérence. Enfin, une feuille de route technologique a été élaborée en unissant les processus technologiques identifiés à travers le STM aux tendances du marché, conformément aux délais de dépôt pour modéliser les évolutions technologiques à court, moyen et long terme.
Dans les 2398 brevets analysés, les résultats ont montré l'extraction de 5102 mots-clés dont "network" (9135), "data" (6995), "adversarial" (5565), "learning" (2577) et "neural" (2283), reflétant les architectures de réseaux neuronaux utilisés dans les

modèles d'IA générative, ainsi que l'identification de six sujets principaux : applications financières et de sécurité de l'information (22,8%), génération et traitement d'images (21,7%), applications médicales (14,6%), systèmes cyber-physiques (14,3%), agents conversationnels intelligents (13,9%) et détection et identification d'objets (12,7%).

Limites :

Les limites dont souffre cette étude on peut les résumer dans les deux points suivants : L'unique utilisation d'une seule approche de fouille pour la classification des données de brevets ce qui introduit des biais et l'exclusion des autres sources de données telles que les publications académiques et les études de marché.

2. MÉTHODES ET OUTILS EXISTANTS

a. SOURCES DE DONNEES :

SCOPUS : Une grande base de données qui contient des millions de document scientifiques notamment des livres, des brevets et des articles couvrant diverses disciplines scientifiques : médecine, arts, sciences, technologie...etc.

ProQuest : Une grande plateforme agrégeant un ensemble de contenus scientifiques et académiques à savoir des thèses, des rapports, des journaux, des articles et plus encore.

b. Outils d'analyse textuelle et de modélisation :

NetMiner : Un logiciel qui a pour objectif d'aider les analystes et les chercheurs à mieux explorer, analyser et visualiser des réseaux sociaux (ensemble d'entités reliées par des relations).

Document-Term Matrix (DTM) : Une structure utilisée en Text Mining pour illuster un corpus de documents de manière numérique où chaque ligne représente un document du corpus et chaque colonne correspond à un terme. La cellule expose la fréquence du terme dans le document.

UMAP : C'est une technique de réduction de dimensionnalité utilisée dans l'apprentissage automatique et la science des données afin de réduire le nombre de variables tout en gardant l'essentiel de l'information utile.

K-Means Clustering : C'est une méthode de l'apprentissage automatique non supervisé qui permet de regrouper des données semblables en un ensemble de groupes appelé clusters.

c-TF-IDF : Il s'agit d'une version modifiée du TF-IDF où chaque cluster est considéré comme un seul document. Elle est très utilisée avec BERTopic afin d'identifier les mots-clés de chaque thème.

RoBERTa-base (Robustly Optimized BERT Approach - base): Il s'agit d'un modèle de NLP basé sur BERT qui permet l'analyse de sentiment c'est à dire de classer le texte en un ton : positif, neutre ou négatif

c. METHODES DE TRAITEMENT ET DE PRETRAITEMENT DES DONNEES :

Prétraitement textuel : Les techniques de prétraitement classiques à savoir le nettoyage, tokenisation, lemmatisation, suppression des stop-words.

Systematic Literature Review : est une méthode scientifique qui vise à répondre à une question de recherche en identifions, sélectionnons et évaluons les recherches antérieures sur un sujet précis.

d. METHODES ANALYTIQUES :

TF-IDF : Méthode utilisée en Text Mining et en NLP afin de repérer les mots importants dans un document relativement à l'ensemble du corpus.

Keyword Centrality Analysis : Méthode du Text Mining utilisée pour mesurer l'importance des mots-clés dans un réseau de cooccurrence c'est-à-dire un réseau où les mots apparaissent ensemble dans les textes. Elle est dotée de trois principales mesures de centralité : centralité de degré, centralité de proximité et centralité d'intermédiarité.

LDA (Latent Dirichlet Allocation) : Une méthode de modélisation de sujets qui sert à détecter des thèmes cachés dans un large corpus de textes en regroupant les mots qui apparaissent souvent ensemble.

BERTopic : C'est une technique moderne de modélisation de sujets. Elle permet d'extraire automatiquement les thèmes prépondérants dans un large corpus de textes en utilisant le modèle BERT.

STM (Structural Topic Modeling) : Une méthode de modélisation de sujets qui en plus de détecter les thèmes dans les textes, elle identifie comment ces thèmes changent en fonction de métadonnées (année, localisation, auteur...etc).

Analyse de réseaux de mots-clés et prédiction de liens : Cette technique permet de comprendre les liens existants entre les mots-clés et de prédire des cooccurrences futures.

Réseau Bayésien : C'est un modèle graphique probabiliste qui représente les dépendances conditionnelles entre les variables à l'aide d'un graphe orienté acyclique. Il permet alors de prédire en présence d'incertitudes.

Analyse morphologique : C'est un processus par lequel la machine comprend la structure d'un mot : sa racine, ses affixes et sa forme fléchie (genre, nombre...etc).

e. Méthodes qualitatives :

Technique Delphi : C'est une méthode largement utilisée dans le domaine de la recherche scientifique, qui vise à consulter itérativement et de façon anonyme un panel d'experts pour prendre des décisions.

Codage qualitatif : C'est une technique d'analyse qui attribue des étiquettes ou des codes à des informations clés d'un texte afin d'extraire les thèmes et les concepts majeurs.

Analyse de sentiment : C'est une méthode qui a pour objectif d'extraire et de classifier les émotions, les attitudes et les opinions exprimés dans un texte.

f. METHODES SPECIFIQUES A LA CARTOGRAPHIE TECHNOLOGIQUE :

TRM (Technology Roadmapping) : C'est un outil de planification qui permet d'identifier des objectifs futurs ainsi que les technologies nécessaires pour les atteindre.

III. METHODOLOGIE

Pour implémenter notre projet, nous avons opté pour le langage Python, qui est largement utilisé grâce à sa polyvalence, sa rapidité, et son application courante au traitement des données et aux représentations graphiques.

Le tableau ci-dessous montre les bibliothèques clés que nous avons utilisé :

Accès aux API Arxiv et OpenAlex	arxiv, requests
Nettoyage et gestion des données	pandas, numpy
Prétraitement linguistique	Nltk, re
Vectorisation	scikit-learn
Entrainement du modèle LDA	gensim
Clustering	sklearn, umap-learn
Visualisation	matplotlib, seaborn, word-cloud, plotly

TABLEAU 1 - BIBLIOTHEQUES UTILISEES

Ces outils ont permis de construire un pipeline exhaustif pour l'extraction et l'analyse des tendances dans la littérature scientifique.

La méthodologie adoptée dans notre étude se décompose en plusieurs étapes et suit un processus rigoureux d'analyse de données textuelles, basée sur les étapes d'un projet de text mining. Ce processus est illustré dans le flowchart ci-dessous :

FIGURE 1 - FLOWCHART DU PIPELINE DU PROJET

Tout d'abord, nous avons collecté les données depuis deux sources majeures : la plateforme Arxiv et la base ouverte OpenAlex.

Ensuite, nous avons fusionné ces deux jeux de données afin d'aligner leurs structures hétérogènes, en passant par un nettoyage primaire afin d'unifier les colonnes et d'éliminer les doublons.

Après cela, nous avons effectué un prétraitement du texte allant de la conversion en minuscule, la suppression des nombres, de la ponctuation, des caractères spéciaux, des URL, des espaces superflus, des stopwords anglais, la tokenisation, jusqu'à la lemmatisation.

Une fois les textes nettoyés, nous avons procédé à la vectorisation TF-IDF, afin d'avoir une matrice dont les lignes représentent les articles et les colonnes sont les mots de chaque article. La vectorisation nous permet de comprendre la diversité de vocabulaire, en se basant sur le score TF-IDF attribué à chaque cellule de la matrice.

De plus, afin de minimiser le bruit et de réduire la complexité, nous avons sélectionné les features en appliquant dans un premier temps la méthode de variance threshold, ce qui élimine les mots très rares et ceux constants, et dans un deuxième temps la réduction de la dimensionnalité avec TruncatedSVD qui est adaptée aux matrices creuses comme celle obtenue par TF-IDF effectué. Cette méthode préserve les relations sémantiques entre les documents, garde l'essentiel de l'information et produit un corpus réduit et prêt à être classifié et visualisé.

Après, nous avons mené une analyse exploratoire détaillé sur le corpus obtenu, afin de le mieux comprendre, en commençant par un diagramme de barres sur l'évolution du nombre de publications par année, puis la fréquence des mots-clés par années, les Bigrams et Trigrams fréquents, les concepts pertinents, et enfin un nuage de mot global et d'autres par année.

Ensuite, nous avons analysé les tendances en utilisant deux approches complémentaires : la modélisation LDA et BERTopic. Le premier modèle est probabiliste, nous a permis d'identifier les topics dominants et leur évolution par année, tandis que le deuxième est basé sur les embeddings sémantiques de BERT, ce qui a enrichi l'analyse avec une modélisation thématique plus fine et contextuelle. En complément de cette étape, nous avons appliqué un clustering avec l'algorithme KMeans, précédé d'une projection des données via UMAP, pour regrouper les articles similaires dans un espace réduit.

En outre, nous avons fait une analyse temporelle, afin d'observer l'évolution des sujets au fil des années, accompagnée d'une cartographie des collaborations entre auteurs, d'une analyse sur les pionniers institutionnels, d'une analyse géographique qui souligne les pays dominants dans les recherches sur l'IAG, et d'une analyse des citations.

Enfin, une évaluation quantitative et qualitative des modèles LDA et KMeans a été réalisée à l'aide des métriques adaptées : Coherence Score et log-Vraisemblance pour LDA, Silhouette Score et Davies-Bouldin Index pour HDBSCAN.

Ces étapes ont permis de construire un corpus propre, interprétable et exploitable pour identifier les tendances thématiques et temporelles en IA générative.

IV. ÉTUDE ET EXPERIMENTATION

1. JEU DE DONNEES UTILISE

Pour cette étude nous avons choisie deux sources de données académiques :

- **Arxiv** : est une plateforme de prépublications académiques en plusieurs disciplines comme l'informatique, les mathématiques et les sciences physiques. Elle permet aux chercheurs de publier leurs travaux avant leur acceptation dans des revues scientifiques, ce qui offre un accès rapide aux avancées dans le domaine de l'intelligence artificielle générative.
- **OpenAlex** : est une base de données bibliographique ouverte qui indexe une large collection de publications scientifiques, et permet une exploration approfondie grâce à son architecture et ses outils avancés.

Les caractéristiques de ces deux sources sont détaillées dans les sections suivantes.

a. ARXIV

arXiv est une archive en ligne ouverte et gratuite lancé en 1991, qui permet aux chercheurs la déposition de leurs prépublications scientifiques dans des domaines de sciences dures, ce qui incite et encourage le partage des connaissances au sein de la communauté scientifique mondiale. Il s'agit donc d'une ressource précieuse pour la veille scientifique.

- **Caractéristiques principales :**

Type	Archive en libre accès pour les articles scientifiques
Création	Lancé en 1991 par Paul Ginsparg
Nature des données	Articles scientifiques de prébublication accompagnés de métadonnées détaillées
Total des données (en 2024)	Plus de 2 millions de prépublications
Format des fichiers	JSON
Période couverte	De 1991 jusqu'à présent
Méthode d'extraction	API arXiv

Langues du contenu	Majoritairement en anglais avec un nombre minoritaires d'articles dans d'autres langues.

TABLEAU 2 - CARACTÉRISTIQUES D'ARXIV

b. OPENALEX

OpenAlex est un graphe de connaissances scientifiques ouvert, demeure une alternative gratuite des grandes bases de données commerciales telles que *Web of Science* de *Clarivate* et *Scopus* d'*Elsevier*. Il couvre une vaste collection de publications, aussi bien les articles publiés que les prépublications, provenant de plusieurs sources notamment *Crossref*, *PubMed*, *ORCID* et bien d'autres. Il offre des métadonnées détaillées sur les travaux scientifiques, les auteurs, les sources, les institutions, les thématiques des travaux, les informations géographiques...etc

- **Caractéristiques principales :**

Type	Une base de données bibliographique et bibliométrique à accès libre.
Création	3 Janvier 2022
Nature des données	Métadonnées des articles publiés et prépublications.
Total des données (en 2024)	250 millions de travaux scientifiques, 90 millions d'auteurs, 100 000 d'institutions, 250 000 sources.
Format des fichiers	JSON et CSV.
Période couverte	Données issues de Microsoft Academic Graph (MAG) et enrichies depuis 2022.
Méthode d'extraction	API OpenAlex.
Langues du contenu	Multilingue.
Avantages	Accès libre aux publications scientifiques, à leurs métadonnées, ainsi que l'exploration des réseaux de collaboration entre auteurs et institutions.

TABLEAU 3 - CARACTERISTIQUES D'OPENALEX

c. FUSION DES DEUX DATASETS

- **Méthode de collecte de données d'Arxiv**

La méthode de collecte de données d'Arxiv repose sur son API publique ce qui a permis l'accés aux métadonnées et aux documents en texte complet.

Le processus d'extraction des publications respecte les exigences suivantes :

- **Période** : de Janvier 2019 à Décembre 2024.
- **Langue** : Anglais.
- **Domaine ciblé** : Intelligence artificielle générative et ses sous domaines.

La requête utilisée est la suivante : **'ti:"generative AI"** OR **ti:"generative model"** OR **ti:"generative adversarial"** OR **''ti:"diffusion model"** OR **ti:"large language model"** OR **ti:"text-to-image"** OR **ti:"text to image"** OR **''ti:"image generation"** OR **ti:"text generation"** OR **ti:"stable diffusion"** OR **''ti:"GPT"** OR **ti:"LLM"** OR **ti:"generative pre-trained"** OR **''abs:"generative AI"** AND (**cat:cs.AI** OR **cat:cs.CL** OR **cat:cs.CV** OR **cat:cs.LG)**'. Elle combine des filtres sur les champs titre, résumé et catégorie.

Les colonnes choisies pour chaque article récupéré d'arXiv sont :

- **id :** identifiant unique de l'article
- **title :** titre de l'article
- **authors :** liste des auteurs
- **abstract :** résumé de l'article
- **categories :** liste des catégories associées à l'article
- **primary_category :** catégorie principale de l'article
- **published_date :** date de publication au format YYYY-MM-DD
- **year :** année extraite à partir de la date
- **month :** mois extrait à partir de la date
- **comment :** commentaires des auteurs
- **doi :** identifiant DOI de l'article

L'issu de cette requête a donné lieu à un nombre total de 1447 articles. Les statistiques que nous avons réalisés concernent la distribution des articles par année ainsi que le classement des 10 catégories les plus fréquentes parmi les articles collectés.

Les résultats du volume de publications au fil du temps sont : 5 articles pour l'année 2019, 11 articles en 2020, 24 articles en 2021, 22 articles en 2022, 400 articles en 2023 et 985 en 2024.Ce qui montre une augmentation exponentielle de nombre d'articles publiés dans les dernières années.

Le top 10 des catégories inclut principalement cs.AI avec 228 articles, cs.CV avec 214 articles et cv.LG ainsi que cv.CL avec 204 articles.

- **Avantages de la source :**

Les avantages d'arXiv demeurent nombreux et on peut citer : l'accès libre et gratuit sans exigence d'abonnement, la disponibilité de l'API chose qui facilite la collecte de données, le volume élevé d'articles dans une large couverture disciplinaire ainsi que la bonne structuration des métadonnées.

- **Inconvénients de la source :**

Au-delà des forces d'arXiv, nous avons relevé quelques contraintes lors de la collecte des données :

- Métadonnées incomplètes notamment pour la colonne doi (1288) et pour la colonne comment (656).
- Manque de colonnes comme Institutions et Countries.
- Presque tous les articles sont en anglais.
- Redondance de quelques articles en d'autres versions.
- Concentration que sur les sciences dures.

- **Méthode de collecte de données d'OpenAlex**

La collecte des données depuis OpenAlex est réalisée grâce à une requête de filtrage avancée à travers son API REST. Elle a permis d'extraire les publications respectant les critères suivants :

- **Période** : de Janvier 2019 à Décembre 2024.
- **Langue** : Anglais.
- **Domaine ciblé** : tous travaux ayant une relation directe avec l'intelligence artificielle générative.

La requête est formulée comme suit : "**generative AI**" OR "**generative model**" OR "**diffusion model**" OR "**large language model**" OR "**LLM**" OR "**GANs**" OR "**transformer model**" OR "**stable diffusion**" OR "**text-to-image**" OR "**GPT**" OR "**DALL-E**" OR "**image generation**" **OR "text generation**".

Elle a récupéré **2000** articles, couvrant divers aspects des modèles génératifs. Voici les colonnes choisies avec leur description :

- **id :** identifiant unique de l'article
- **title :** titre de l'article
- **doi :** Digital Object Identifier qui permet de retrouver rapidement un document en ligne, indépendamment des changements d'URL ou de l'emplacement physique du document.
- **authors :** liste des auteurs
- **publication_date :** date de publication au format YYYY-MM-DD
- **year :** année extraite à partir de la date
- **abstract :** résumé de l'article
- **open_access :** indique si l'article est accessible librement
- **concepts :** les 5 concepts principaux associés à l'article
- **concepts_scores :** les scores associés aux concepts principaux
- **categories :** liste des concepts associés à l'article
- **cited_by_count :** nombre de citations reçues par l'article
- **referenced_works_count :** nombre de travaux cites dans l'article
- **type :** type de publication
- **author_countries :** pays d'affiliation des auteurs
- **author_institutions :** institutions des auteurs

Comme une première exploration des données, nous avons fait quelques statistiques sur la distribution par année des articles obtenus, pour cela nous avons écrit un script

Python qui a donné les résultats suivants : 792 articles en 2019, 497 articles en 2020, 316 articles en 2021, 162 articles en 2022, 215 articles 2023, et 18 articles en 2024. Ceci montre que la majorité des articles provient des années 2019 à 2021.

- **Avantages de la source :**

La richesse historique des données d'OpenAlex, offre l'accès libre et gratuit à des publications de différentes types (articles, conférence, revue…), multilingues, incluant ceux antérieures à 2019, avec des métadonnées complètes comme les citations, les institutions, les pays et les concepts, ce qui enrichie l'étude par des analyses bibliométriques, géographiques et institutionnelles.

- **Inconvénients de la source :**

Malgré les atouts d'OpenAlex, nous avons observé quelques limites lors de la collecte des données :

- Certaines colonnes présentes des valeurs manquantes notamment 358 pour abstract, 63 pour author_countries, 61 pour author institutions, 15 pour doi et 1 pour authors.
- Le champ abstract ne peut pas être exploité directement depuis l'API sans avoir passé par un traitement spécifique sur un autre champ abstract_inverted_index.

➢ **Fusion**

Dans le but d'avoir une vue plus complète sur les tendances en IA générative, une étape essentielle consiste à fusionner les **1447** articles extraits d'arXiv avec les **2000** obtenus d'OpenAlex, ce qui va combiner la richesse thématique et technique d'arXiv avec les métadonnées bibliométriques et institutionnelles d'OpenAlex.

Tous d'abord, après avoir chargé les deux fichiers csv contenant les données collectées dans des Dataframes distingués, nous avons défini une colonne **source** afin de garder la traçabilité de l'origine des articles. Ensuite pour unifier la structure des deux sources, nous avons en premier temps sélectionné les colonnes pertinentes pour chaque dataset, les renommé, et les défini comme un ensemble commun à conserver : **['id', 'source', 'source_id', 'title', 'authors', 'abstract', 'publication_date', 'year', 'doi', 'categories', 'concepts', 'cited_by_count', 'countries', 'institutions']**. Les colonnes non présentes dans un dataset sont remplies par NaN, en l'occurrence arXiv ne dispose pas des colonnes sur les institutions, les pays, ou les citations. Par la suite, nous avons obtenu comme premier résultat **3447** articles qui contiennent des doublons et des valeurs manquantes. Pour cela nous avons effectué un nettoyage préliminaire par lequel nous avons supprimé les caractères spéciaux, les sauts de ligne, les tabulations, les doublons basés sur le titre et l'année, la normalisation des dates au format standard YYYY-MM-DD, l'extraction des années dans la colonne **year,** et la gestion des valeurs manquantes : 0 pour les entiers ou réels, NaN pour les objets**.** Ce processus a donné **3437** articles uniques, dont **1439** articles provenant d'arXiv, et **1998** articles d'OpenAlex.

2. PRETRAITEMENT DES DONNEES TEXTUELLES

Le prétraitement NLP est une étape cruciale pour la préparation du texte brut en format exploitable par un modèle d'apprentissage automatique. Elle consiste à passer par plusieurs opérations :

- Le nettoyage du texte implique la suppression des caractères spéciaux, de la ponctuation et de tout élément indésirable, cette phase varie selon la source des données.
- La standardisation convertit le texte en un format uniforme comme la mise en minuscule.
- La normalisation réduit les mots à leur forme de base selon des méthodes comme le stemming ou la lemmatisation.
- La tokenisation divise un texte en parties plus petites, appelées tokens ou jetons, ceci permet aux modèles d'apprentissage automatique de manipuler des données plus faciles à gérer et plus spécifique.
- L'élimination de bruit comme la suppression des jetons trop courts.

En Python, plusieurs bibliothèques sont utilisées pour faciliter le processus de cette étape, en l'occurrence SpaCy, NLTK et TextBlob. Dans notre cas, nous avons choisi de travailler avec la bibliothèque **NLTK**, pour sa richesse fonctionnelle qui nous a facilité le prétraitement. La méthodologie que nous avons suivi commence tout d'abord par la conversion du texte en minuscules afin d'unifier les occurrences de chaque mot, puis la suppression des URLs, des emails, des chiffres, des caractères spéciaux, de la ponctuation, et des espaces superflus, grâce à des expressions régulières traitées par la bibliothèque **re**. Ensuite, la segmentation du texte en unités lexicales de longueur supérieure à 2 lettres, avec l'élimination des mots vides anglais, qui ne portent aucune signification réelle, et sont très fréquents. Enfin, la réduction des mots à leur forme de base en prenant en considération leur contexte linguistique.
Ce processus est appliqué principalement sur les deux colonnes **title** et **abstract**, qui vont être combinées en une seule colonne appelée **cleaned_text**, garantissant une meilleure couverture thématique. Grâce à cette étape nous avons obtenu un corpus de **3437** articles uniques, prêt à l'exploitation dans les étapes suivantes.

3. VECTORISATION ET SELECTION DE FEATURES

a. VECTORISATION TF-IDF

Après le prétraitement du texte, les données textuelles doivent être transformées en données numériques, afin de permettre l'application des algorithmes d'apprentissage automatique.
La vectorisation joue un rôle important dans cette transformation, et assure que ces textes sont bien en un format compréhensible par les modèles. Parmi ses techniques les plus utilisées, il existe :

- Bag-Of-Words (BoW) : convertit un texte en un vecteur de fréquence d'apparition des mots dans un document. Chaque dimension correspond à un mot du corpus, sans prendre en compte l'ordre ou le contexte des termes.
- Term Frequency – Inverse Document Frequency (TF-IDF) : évalue l'importance d'un mot dans un document par rapport au corpus. Elle la pondère en fonction de sa fréquence d'apparition dans un document, et sa rareté dans le corpus.

- Word Embeddings : utilisent des vecteurs denses pour capturer les relations sémantiques et contextuelles entre les mots. Parmi les modèles les plus couramment utilisées, on retrouve Word2Vec, Glove, et FastText, chacun ayant ses propres caractéristiques et avantages selon l'application envisagée.

Nous avons choisi de vectoriser notre corpus à l'aide de TF-IDF en raison de sa capacité à identifier les termes les plus significatifs dans un corpus de texte. Contrairement à BoW qui ne considère que la fréquence des mots sans distinction de leur importance, TF-IDF pondère chaque mot selon sa rareté dans le corpus ce qui réduit l'impact des mots trop fréquents. De plus, les matrices creuses obtenues par TF-IDF sont bien adaptées pour les modèles linéaires, le clustering et certaines méthodes de modèle thématique (topic modeling) comme LDA.

Nous avons appliqué cette vectorisation TF-IDF sur la colonne **cleaned_text**, issue du prétraitement NLP effectué précédemment. L'ajustement des paramètres se présentent comme suit :

- **max_df = 0.85** : pour ignorer les mots apparaissant dans plus de **85%** des documents.
- **min_df = 2 :** pour garder uniquement les mots présents dans au moins **2** documents.
- **stop_words = 'english' :** pour supprimer les mots vides anglais.
- **Lowercase = True :** pour la convertion en minuscules automatique.
- **max_feature = 10000 :** pour limiter le vocabulaire à **10000** termes au maximum.

Le résultat obtenu est une matrice creuse de dimensions (**3437 × 9343**), dont chaque ligne correspond à un document, chaque colonne correspond à un mot du vocabulaire, et chaque cellule correspond à un score TF-IDF indiquant l'importance du mot dans le document par rapport au reste du corpus. Ce résultat montre que le corpus est très riche lexicalement, et capture les mots pertinents, pour cela une première requête a été effectuée pour extraire les mots plus fréquents, révélant les termes suivants : **'model', 'image', 'generative', 'learning', 'data', 'network', 'language', 'method', 'based', 'deep'**. Ces mots reflètent les thèmes dominants de l'IA générative, une analyse thématique approfondie détaille ce résultat dans les sections suivantes.

Néanmoins, cette richesse de vocabulaire, elle soulève un défi de dimensionnalité qui rend l'exploitation de ces données difficile, pour cette raison nous avons envisagé deux autres étapes essentielles : la sélection de features et la réduction de dimensionnalité.

b. SELECTION DES FEATURES

Dans l'objectif de réduire la complexité computationnelle et d'éliminer les termes peu informatifs, nous avons choisi comme technique pour la sélection des features la **Variance Threshold** qui exclue les mots ayant une faible variance selon un seuil précis, cette suppression de caractéristiques n'affecte pas les performances des modèles puisque leurs valeurs sont presque constantes à travers plusieurs échantillons.

Le fonctionnement de cette technique passe par plusieurs phases, elle débute par la mesure de dispersion des valeurs de chaque feature dans le dataset, puis compare la variance calculée de chaque caractéristique avec un seuil prédéfini par le paramètre

threshold, dans notre cas sa valeur égale à **0.0001**, si cette variance est inférieure à celui-ci, alors la caractéristique est considérée comme non pertinente, et sera supprimée. Après ce passage nous avons obtenu une nouvelle matrice de dimension (**3437 × 2394**), cela signifie que **6949** termes ont été éliminé, soit environ **74%** des mots initiaux, sans perte significative d'informations thématiques, ce qui permet de réduire la dimensionnalité et d'améliorer la performance des modèles.

c. REDUCTION DE DIMMENSION

Malgré la sélection des caractéristiques, la dimension du la matrice reste encore vaste pour qu'il soit exploité efficacement par un modèle d'apprentissage automatique, et nécessite une deuxième réduction, pour cela nous avons opté pour une approche appelée **TruncatedSVD** qui est adaptée aux matrices creuses comme celle produit par TF-IDF, et projette les données dans un espace de dimension moins réduit, tout en gardant l'essentiel de l'information et préservant les relations sémantiques entre les documents. L'avantage de cette technique par rapport aux d'autres comme l'ACP, c'est qu'il ne centre pas les données avant la décomposition en valeurs singulières, elle applique cette décomposition directement à notre matrice, pour la décomposer en trois sous matrices, et sélectionne les k premières valeurs singulières, dans notre cas **300** composantes. Cette configuration du paramètre **n_components** à **300**, vient après une visualisation de la courbe de la variance expliquée cumulée par rapport au nombre de composantes (fig.2), ce qui a montré que cette valeur explique plus de 50% de la variance. Notre matrice originale est donc projetée dans un nouvel espace de dimension réduite (**3437 × 300**) facilitant ainsi l'analyse et l'exploitation par les modèles.

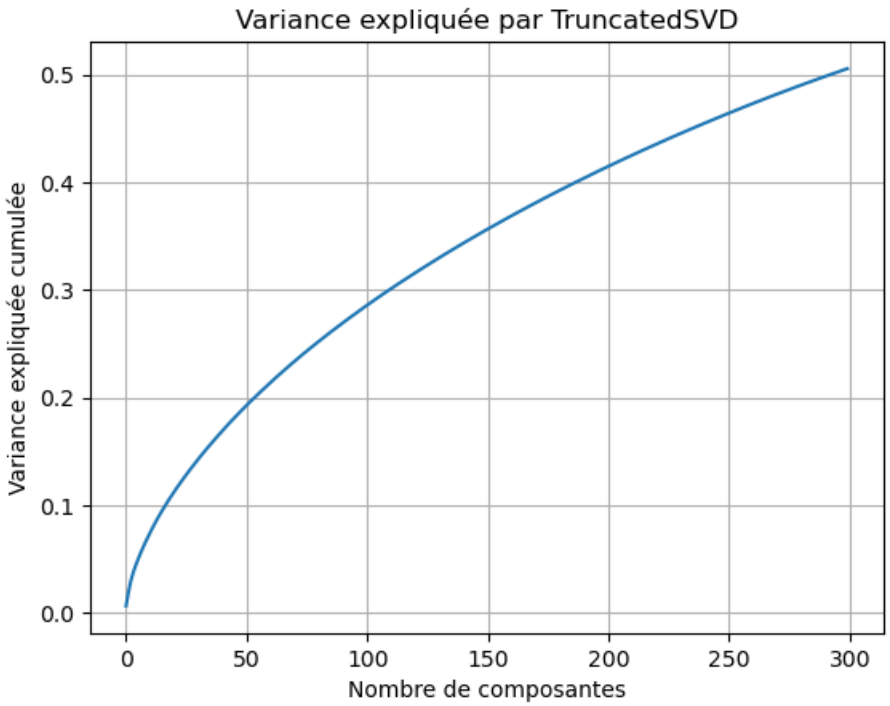

FIGURE 2 - VARIANCE EXPLIQUEE PAR TRUNCATEDSVD

4. EXPLORATION ET VISUALISATION DES TEXTES

Afin d'explorer notre corpus consolidé de **3437** articles, publiés entre **2019-2024**, nous avons mené une analyse exploratoire qui a permis d'extraire des insights significatifs. Cette étape donne une vue globale sur les tendances dans le domaine de l'IA générative, et constitue une base essentielle pour guider les analyses ultérieures.

a. EVOLUTION DE NOMBRE DE PUBLICATIONS PAR ANNEE

A l'aide de la bibliothèque **matplotlib**, nous avons observé la croissance globale des publications au fil des années en utilisant un diagramme en barres. Les résultats obtenus se présentent comme suit :

2019	796 articles
2020	508 articles
2021	340 articles
2022	184 articles
2023	607 articles
2024	1002 articles

TABLEAU 4 - EVOLUTION DE NOMBRE DE PUBLICATIONS PAR ANNEE

La distribution des publications par année (fig.3) montre que globalement entre 2019 et 2024, les recherches en intelligence artificielle générative ont connu une croissance exponentielle, et ce, met en évidence l'évolution constante en ce domaine.
Entre **2020-2022**, le nombre de publications a envisagé une baisse relative qui peut être causée par un facteur externe comme la pandémie.
En **2023**, une croissance notable est marquée allant de **607** articles à **1002** articles en **2024**, reflétant l'essor rapide de l'intelligence artificielle générative.

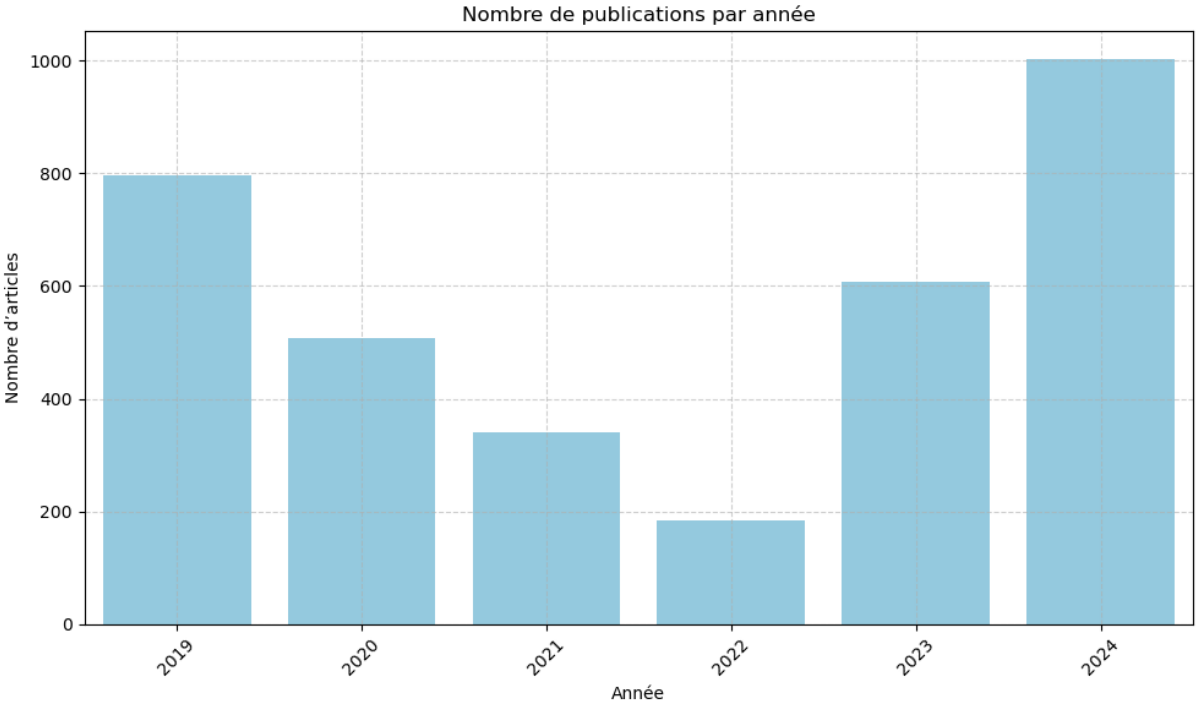

FIGURE 3 - NOMBRE DE PUBLICATIONS PAR ANNEE

b. FREQUENCE DES MOTS-CLES PAR ANNEE

Afin d'avoir une vision sur les grands thématiques de recherches en IA entre **2019-2024**, nous avons extrait les 20 mots les plus fréquents par année, en utilisant la méthode **CountVectorizer()** appliquée sur les textes de la colonne nettoyée **cleaned_text**, en éliminant les mots vides anglais et en se limitant sur 500 mots au maximum pour une meilleur lisibilité. Les résultats obtenus montrent une évolution des sujets étudiés, en passant du général au spécifique.

- En **2019**, les mots clés-dominants sont :

  [model: 766, learning: 752, image: 730, network: 714, method: 584, data: 556, deep: 447, based: 439, approach: 293, paper: 292, using: 292, performance: 287, neural: 269, state: 265, high: 257, result: 257, feature: 255, training: 245, application: 240, task: 240]

  Ces mots-clés forment la base de l'apprentissage profond (deep learning) et des modèles génératifs.

- En **2020**, les mots clés-dominants sont :

  [model: 712, learning: 518, image: 450, network: 446, data: 444, method: 406, based: 390, deep: 326, covid:311, approach: 229, using: 212, state: 209, task: 205, language: 204, result: 192, training: 187, neural: 184, research: 184, paper: 180, domain: 176]
  En prenant compte la pandémie covid-19, les recherches scientifiques ont été ralenties, tout en favorisant des nouvelles utilisations dans des contextes pratiques.

- En **2021**, les mots clés-dominants sont :

  [learning: 536, model: 445, data: 301, image: 260, method: 246, deep: 244, network: 238, based: 226, research: 167, task: 166, approach: 141, application: 135, detection: 121, using: 119, machine: 118, paper: 117, language: 115, performance: 112, survey: 110, used: 110]

  On observe que la mise en œuvre des modèles concerne des domaines plus spécifiques notamment la vision par ordinateur (computer vision) et la détection automatique.

- En **2022**, les mots clés-dominants sont :

  [model: 344, image: 262, learning: 229, data: 152, task: 149, method: 142, based: 131, language: 125, network: 112, deep: 102, text: 93, high: 88, using: 84, transformer: 82, challenge: 80, feature: 79, research: 79, application: 76, large: 73, paper: 73]

On remarque que cette année marque une transition vers les transformers multimodaux.

- En **2023**, les mots clés-dominants sont :

  [model: 1594, generative: 1141, chatgpt: 735, data: 585, language: 581, image: 550, text: 406, large: 404, human: 379, research: 359, learning: 357, task: 354, based: 350, llm: 328, using: 328, study: 327, potential: 311, generation: 296, intelligence: 288, use: 287]

  Cette année est témoin d'une explosion rapide des grands modèles de langage (LLM), de l'IA générative, et surtout du modèle ChatGpt.

- En **2024**, les mots clés-dominants sont :

  [model: 2437, generative: 2116, data: 1023, llm: 909, image: 898, language: 760, based: 718, generation: 617, human: 615, method: 607, large: 600, learning: 536, paper: 535, study: 520, approach: 512, research: 510, task: 502, using: 490, generated: 488, text: 486]

  On constate que le vocabulaire devient plus centré sur les applications concrètes de l'intelligence artificielle générative notamment l'utilisation humaine.

En conclusion, la fréquence des mots-clés au fil des années montre une transition progressive de la recherche en IA générative, d'un cadre purement technique vers l'innovation.

c. BIGRAMS ET TRIGRAMS FREQUENT

L'analyse des bigrams et trigrams fréquents est une technique de traitement de langage naturel, qui vise à examiner les combinaisons de mots les plus fréquentes dans un texte, en tenant compte leurs relations sémantiques et le contexte.

Pour capturer les paires et les triplets de mots qui se suivent souvent dans notre corpus, nous avons effectué une analyse des bigrams/trigrams sur la colonne **cleaned_text** en utilisant la méthode **CountVectorizer()**, en se limitant sur les **20** expressions les plus fréquentes.

Les résultats (Tableau 3) montrent que « language model » domine au niveau des bigrams avec 996 occurrences, et « large language model » domine au niveau des trigrams avec 675 occurrences.

Bigrams \ Trigrams	**Fréquence**
language model	996
deep learning	771
state art	737
large language	681
large language model	675

artificial intelligence	667
generative model	652
neural network	593
machine learning	573
diffusion model	422
natural language	394
model llm	327
language model llm	322
text image	275
real world	260
generative adversarial	248
adversarial network	237
large scale	236
generative adversarial network	232
language processing	232

TABLEAU 5 - LES 20 BIGRAMS/TRIGRAMS FREQUENTS

Cette analyse a clarifié les tendances thématiques dominantes. En effet, la paire **language model** et le triplet **large language model** montrent la centralité des LLMs dans les recherches scientifiques. De plus, les termes **deep learning**, **neural network**, et **machine learning**, reflètent toujours les fondations méthodologiques de ces recherches. Ainsi, l'expression **diffusion model** marque l'apparition des architectures basées sur les modèles de diffusion. En outre, certaines bigrams rappellent les bases théoriques de l'IA générative, par exemple **generative model**, **generative adversarial** et le trigram **generative adversarial network**. Par ailleurs, **State art** est le troisième bigram fréquemment utilisé, ce qui montre que beaucoup d'articles font des comparaisons entre différents modèles.
Cette étape nous amène à conclure que l'IA générative évolue continuellement, passant des fondements technologiques aux applications concrètes.

d. LES CONCEPTS DOMINANTS

Grâce à la colonne concept nous avons pu extraire les concepts dominants des recherches en ce domaine, offrant ainsi une catégorisation disciplinaire ce qui facilite l'analyse des tendances et thématiques.

Cette catégorisation est obtenue à travers le découpage des listes de concepts associés à chaque publication, ensuite de compter les occurrences de chaque concept dans tout le corpus, et enfin de sélectionner les 20 concepts les plus fréquents.

Concept	**Nombre d'occurrence**
Computer science	1374
Artificial intelligence	757
Deep learning	145
Machine learning	126
Generative grammar	101
Natural language processing	95
Materials science	89

Transformer	88
Convolutional neural network	80
Data science	75
Context (archaeology)	68
Field (mathematics)	63
Benchmark (surveying)	57
Computer vision	56
Medicine	55
Pattern recognition (psychology)	55
Image (mathematics)	53
Segmentation	52
Artificial neural network	49
Inference	48

TABLEAU 6 - LES 20 CONCEPTS FRÉQUENTS

L'analyse révèle que les disciplines Computer Science (1374 occurrences) et Artificial Intelligence (757 occurrences) sont largement dominantes, illustrant le coeur des Travaux portant sur l'IA générative. De plus, les concepts deep learning (145 occurrences), machine learning (126 occurrences), natural language processing (95 occurrences) et computer vision (56 occurrences), traduisent leur importance comme des piliers méthodologiques en figurant parmi les fondements techniques les plus fréquents. En outre, les architectures spécifiques comme les transformers (88 occurrences) et les convolutional neural network (80 occurrences) soulignent aussi leur rôle dans la conception des modèles d'IA générative. En outre, les concepts techniques comme generative grammar (101occurrences), data science (75 occurrences), segmentation (52 occurrences) et inference (48 occurrences) accentuent l'intégration de l'IA générative dans des domaines analytiques et appliqués, et indiquent une attention attribuée aux règles structurelles, à la génération cohérente et à la segmentation précise des contenus multimodaux. Enfin, certains domaines transversaux notamment material science (80 occurrences), medicine (55 occurrences) et pattern recognition (psychology) (55 occurrences) montrent que les applications de l'intelligence artificielle générative s'étendent pour impacter de multiples domaines scientifiques.

e. NUAGE DE MOTS

En profitant des représentations visuelles intuitives offertes par les nuages de mots, nous avons visualisé les termes les plus fréquents dans le corpus.

- **Global**

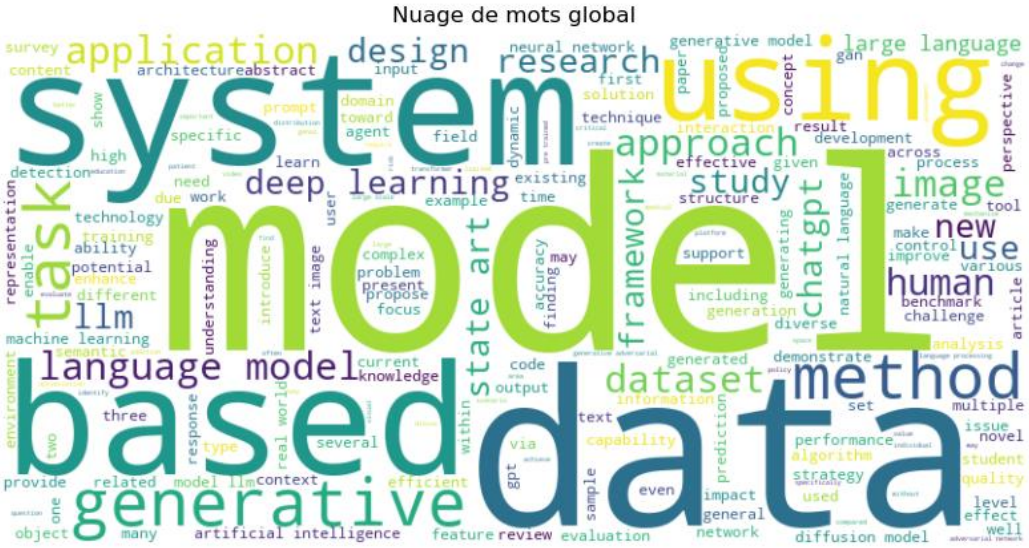

FIGURE 4- NUAGE DE MOTS GLOBAL

On remarque que les expressions : **model**, **data**, **system**, **based**, **generative**, et **using** sont les plus dominants dans le corpus en sa globalité, ce qui reflète leur importance centrale dans le domaine.

- **Par année**

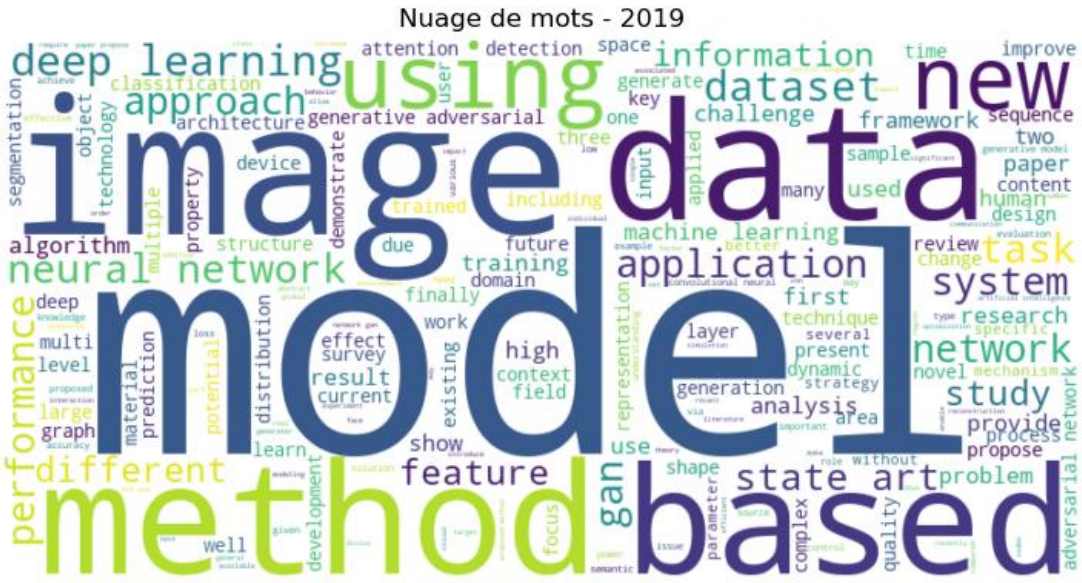

FIGURE 5 - NUAGE DE MOTS 2019

En **2019,** le nuage des mots les plus fréquents met en évidence des termes techniques tels que **model**, **learning**, **neural**, **network**, **deep**, **image**, **data,** mais aussi des expressions comme **approach**, **research**, **method**, **study**. Ce résultat marque une phase de consolidation des approches classiques et de leur utilisation dans divers contextes.

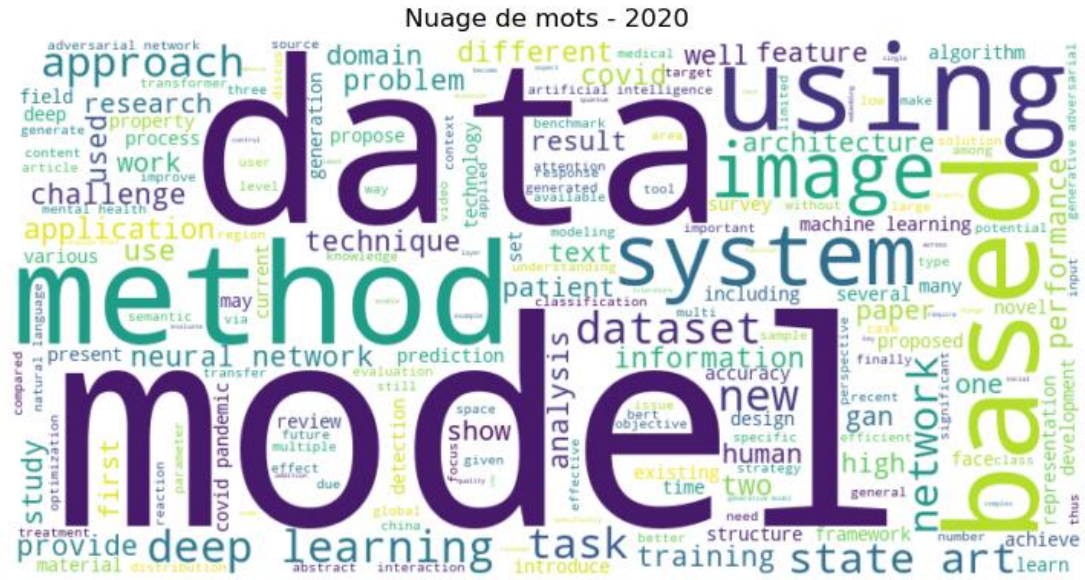

FIGURE 6- NUAGE DE MOTS 2020

En **2020,** on observe la présence des termes suivants : **covid, detection, deep learning, human, analysis, domain, medical** qui révèlent qu'une partie importante des recherches scientifiques pendant la pandémie a été focalisé sur les différentes applications de l'IA générative en santé ou en analyse des données.

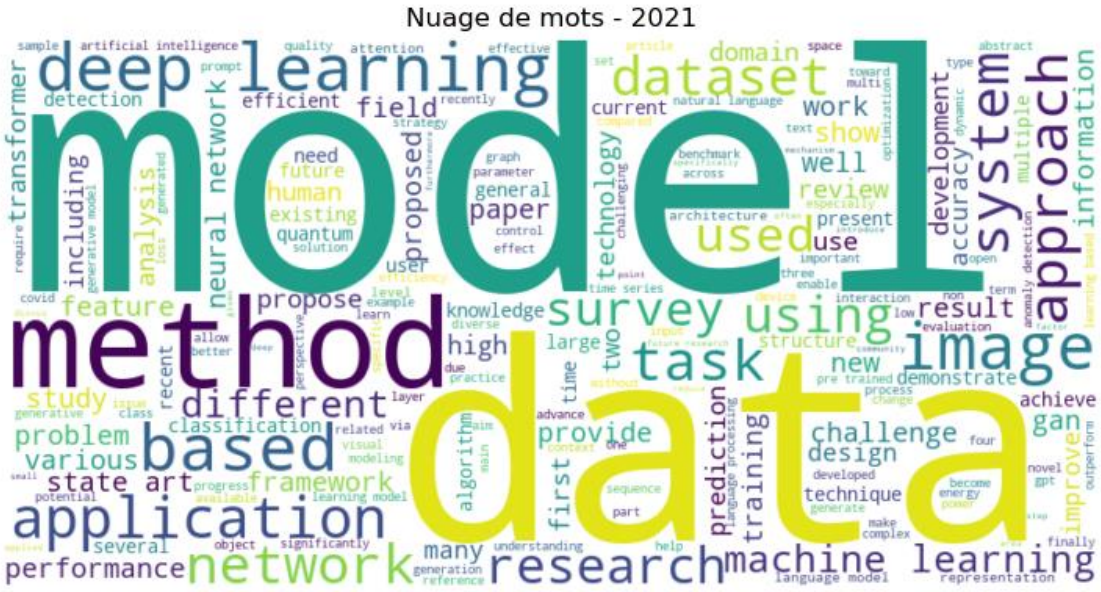

FIGURE 7- NUAGE DE MOTS 2021

En **2021,** ce nuage de mot illustre une transition entre les architectures classiques et les applications émergentes. Les termes **transformer, generative, application, based** sont souvent liés à des avancées technologiques. D'une autre part, les termes **natural language processing** et **computer vision** indique une diversification dans les domaines d'application.

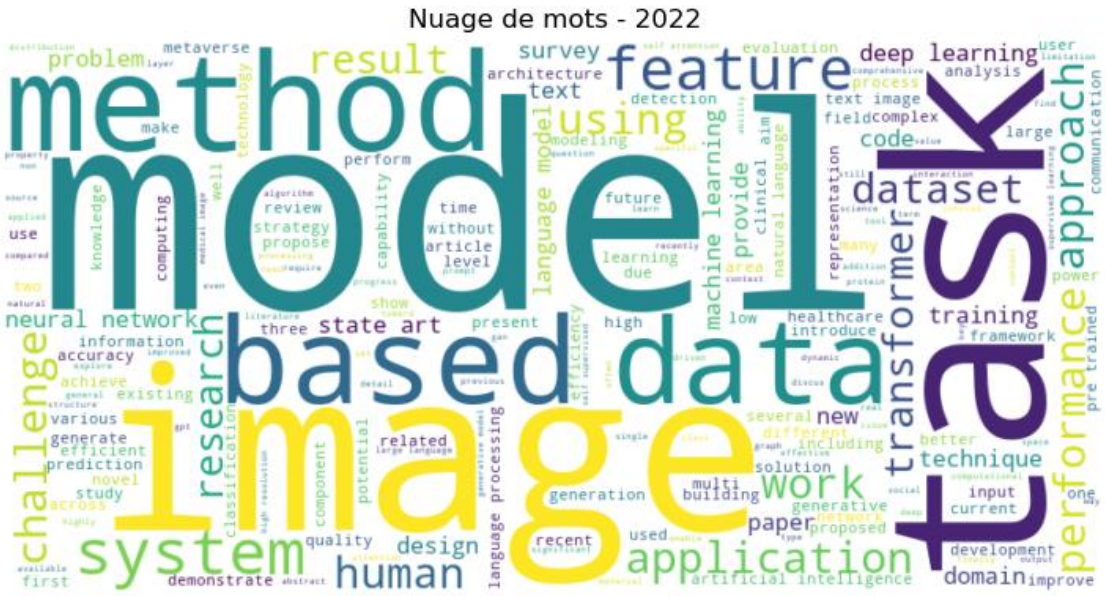

FIGURE 8- NUAGE DE MOTS 2022

En **2022,** l'apparition de nouveaux concepts tel que **large language model** marque l'essor des grands modèles de langage. En effet, les termes **model, image, transformer,** apparaissent en grand, ce qui illustre l'intensification des recherches autour les architectures de génération d'image et de langage.

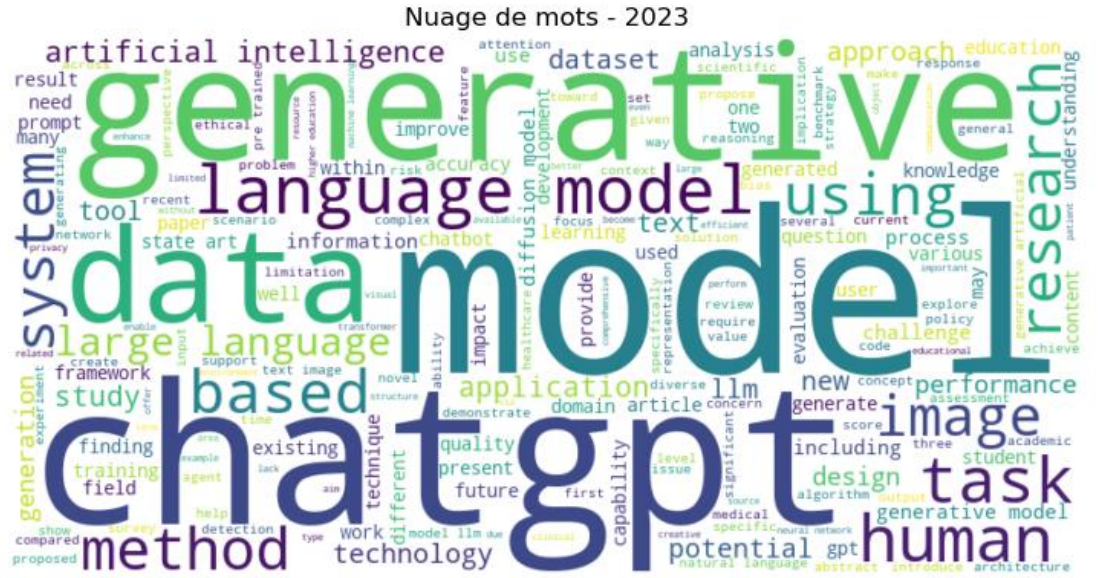

FIGURE 9 - NUAGE DE MOTS 2023

En **2023,** nous observons que parmi les termes qui apparaissent en grands, existe **generative, chatgpt, model, data, large language model, image, task, human, research, prompt, system, artificial intelligence** ceci souligne la dominance de ces technologies dans les recherches surtout avec l'émergence de Chatgpt. Les expressions

study, analysis, use, potential, montre la croissance des implications sociales et éthiques en l'IA générative.

Par ailleurs, cette année a connu une accélération rapide des publications concernant les applications sociales des LLMs et Chatgpt.

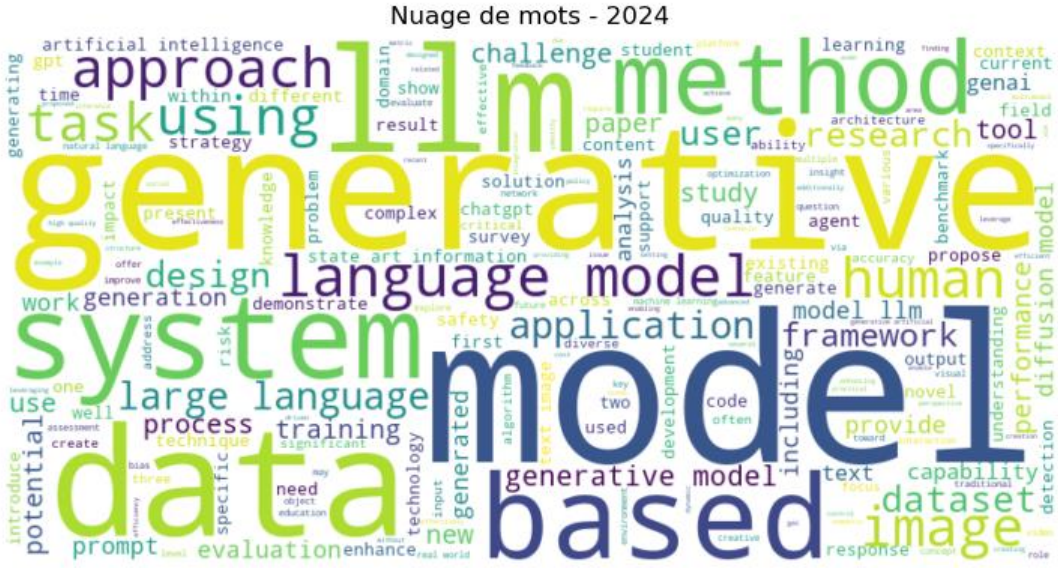

FIGURE 10- NUAGE DE MOTS 2024

En **2024,** le nuage de mots met en avant des termes comme **generative, llm, system, model, method, human, research, study, genai, prompt, performance, gpt, diffusion model, analysis, potential, critical, risk, using, biais,** qui reflètent une consolidation des tendances de 2023, et marquent une évolution des recherches vers une analyse critique de l'intelligence artificielle générative, particulièrement portée à l'interaction homme-machine, à la sécurité, à l'éthique, et aux usages concrets.

En somme, les nuages de mots des années entre 2019-2024, nous a permis d'identifier l'évolution des sujets étudiés dans le domaine de l'IA générative. Passant des fondations techniques en 2019, des applications liées à la crise sanitaire en 2020, de l'émergence de l'architecture transformer en 2021, des transitions vers les modèles multimodaux en 2022, de l'explosion de Chatpgt et des LLMs en 2023, à la réflexion sur les biais, les risques et les implications sociales d'usage de l'IA générative en 2024.

5. APPLICATION D'ALGORITHMES DE TEXT MINING

Après avoir exploité notre corpus constitué de 3437 articles uniques pour l'exploration de l'évolution de nombre de publications au fil de temps, et d'extraire les mots-clés fréquents, ainsi que d'avoir une idée générale sur les thématiques dominantes et les sujets pionniers à travers plusieurs méthodes complémentaires notamment les bigrams/trigrams fréquents, les concepts émergents, et les nuages de mots. Nous avons mené une analyse des tendances en intelligence artificielle générative couvrant les années entre 2019 et 2024, et utilisant des techniques diverses afin de capturer à la fois les thèmes dominants et les acteurs clés.

a. MODELISATION THEMATIQUE

Afin de répondre aux questions qui soulèvent à propos des thèmes et sujets dominants, la modélisation thématique est une étape essentielle qui permet d'extraire, d'étudier et d'analyser les grands piliers des tendances de la recherche scientifique. Pour cela, plusieurs méthodes se présentent, qui sont choisies selon le besoin de l'étude. Dans notre analyse, nous avons usé deux approches principales, chacune est caractérisée par ses avantages :

➢ LDA

La technique Latent Dirichlet Allocation abréviée LDA, est une méthode statistique et probabiliste qui suppose que chaque document est composé de plusieurs sujet, et chaque sujet est constitué de plusieurs mots. Elle repose sur la distribution des mots dans les documents pour identifier les thèmes sous-jacents.

- **ETAPES PRINCIPALES ET PARAMETRAGE**

  La bibliothèque **gensim** est adoptée largement pour le traitement du langage naturel, elle dispose du modèle probabiliste LDA, pour cela nous l'avons utilisé dans notre analyse. Le processus a débuté par la tokenisation du texte nettoyé de la colonne **cleaned_text** à l'aide de la fonction **simple_process()** présente dans la librairie **gensim**, permettant d'obtenir une liste de tokens normalisés et sans ponctuations. Voici un exemple du résultat obtenu après la tokenisation : **['survey', 'image', 'data', 'augmentation', 'deep', 'learning', 'deep', 'convolutional', 'neural', 'network']**. Ensuite, pour garder que les mots pertinents, un dictionnaire gensim a été construit pour contenir ces tokens en filtrant les mots trop rares par le paramètre **no_below=0** et ceux trop fréquents par **no_above=0.8**. Par ailleurs, chaque document a été représenté sous forme de sac de mots (BoW) grâce à la méthode **doc2vec()**.

  Les paramètres de l'entrainement du modèle ont été choisi pour garantir un bon équilibre entre la qualité des topics extraits et le temps d'exécution. Voici l'ajustement fait :

  – Nombre de topics : 6
  – Graine aléatoire : random_state = 42 pour assurer la reproductibilité.
  – Taille des lots : chunksize = 100
  – Mise à jour après chaque lot : update_every = 1
  – Nombre de passes sur le corpus : passes = 10
  – Estimation automatique des hyperparamètres : alpha = 'auto'
  – Calcul des distribuions par mot : per_word_topics = True

- **ENTRAINEMENT DU MODELE**

  Le modèle LDA a été entraîné sur l'ensemble du corpus vectorisé. Une fois l'entrainement terminé, chaque document a été extrait avec ses **10** mots-clés les plus probables, qui sont les termes les plus représentatifs du sujet auquel appartient le document.

➢ BERTOPIC

C'est une méthode plus récente qui exploite le modèle BERT (Bidirectional Encoder Representations from Transformers) pour capturer les relations sémantiques entre les mots. BERTopic utilise des embeddings de mots qui sont des vecteurs numériques représentés dans un espace multidimensionnel, ce qui permet au modèle d'extraire les relations sémantiques et contextuelles entre les mots. Il utilise des techniques de clustering pour regrouper les documents en fonction de leur similarité sémantique.

- **ETAPES PRINCIPALES ET PARAMETRAGE**
  Nous avons appliqué l'algorithme BERTopic sur la colonne nettoyé cleaned_text, sa configuration se présente comme suit :
  - Le modèle d'encodage utilisé est all-MiniLM-L6-v2 qui est un Sentence Transformer léger, utilisé pour capturer les relations sémantiques entre les mots.
  - Le vectorisateur choisi est **CountVectorizer()** avec les paramètres **stop_words = "english"** pour supprimer les mots vides anglais, **min_df = 6** qui indique le seuil minimal de fréquence des mots, et **ngram_range = (1,2)** pour la prise en compte des unigrammes et bigrammes.
  - Le nombre de topics est configuré pour le calcul automatique, basé sur la diversité des corpus.

- **ENTRAINEMENT DU MODELE**
  L'entrainement du modèle a été sur l'ensemble des textes nettoyés de la colonne **cleaned_text**, une fois terminé, chaque document a été attribué à un topic via la méthode **fit_transform()**.

b. CLUSTERING KMEANS

L'algorithme KMeans a été utilisé pour regrouper les documents selon leur similarité sémantique. Pour identifier le nombre optimal des clusters, nous avons analysé le score Silhouette attribué à chaque valeur de k (intervalle entre 2-14).

- **ETAPES PRINCIPALES ET PARAMETRAGE**
  Pour qu'on puisse appliquer KMeans sur notre corpus, nous devons disposer d'une matrice réduite, comme celle générée précédemment avec TruncatedSVD, cela permet à l'algorithme de projeter les données dans un espace à 300 dimensions.

- **ENTRAINEMENT DU MODELE**
  Ensuite, pour chaque valeur de nombre de clusters n_clusters, le modèle est entrainé avec une graine de reproductibilité égale à 42 (random_state=42) et n_init=10 pour mesurer le score Silhouette avec (range_clusterrs = range(2,15)). La meilleure configuration est déterminée en choisissant celle qui correspond à la valeur maximale de Silhouette score. Une fois le nombre optimal de clusters identifié, chaque document est assigné à un cluster par la suite.

c. ANALYSE GEOGRAPHIQUE

Pour analyser les contributions géographiques à la recherche en IA générative, nous avons extrait les informations relatives aux pays d'origine des articles publiés entre 2019 et 2024 depuis le champ **countries** de notre Dataframe. Tout d'abord, nous avons extrait les code pays à partir des chaînes de caractères séparés par des points-virgules dans la colonne countries. Chaque code pays a été traduit en son nom complet via un dictionnaire pour des raisons de lisibilité. Ensuite, notre analyse s'est portée sur l'identification des 15 pays les plus contributifs en matière de recherches sur l'IA générative, puis elle s'est focalisée sur les 5 premiers afin d'étudier les tendances de leurs publications entre 2019-2024.

d. ANALYSE INSTITUTIONNELLE

L'analyse institutionnelle permet d'identifier le rôle des différents acteurs dans la publication des articles ou en recherche sur l'IA générative. Pour réaliser cette analyse, nous avons extrait les noms des institutions à partir du champ **institutions,** puis de compter le nombre d'occurrences de chaque institution pour identifier celles les plus actives en termes de publications sur l'IA générative. De plus, pour associer chaque institution à ses clusters respectifs, nous avons utilisé les informations de clustering fournies par KMeans, ce qui a complémenté l'analyse des tendances thématiques.

e. ANALYSE DES AUTEURS

L'objectif de l'analyse des auteurs est d'étudier la structure du réseau d'auteurs, d'identifier des communautés, et les collaborations entre eux. Cette analyse est réalisée en suivant les étapes suivantes :

Tout d'abord, afin de construire un réseau d'auteurs, nous avons extrait leurs noms à partir du champs **authors** de notre corpus, puis nous avons filtré ces auteurs pour ne garder que ceux ayant participés à au moins deux articles, pour garder la pertinence des collaborations. Par ailleurs, un graphe a été construit à l'aide de la fonction **Graph()** présente dans la librairie **networkx**, où chaque nœud représente un auteur et chaque arête représente une collaboration entre deux auteurs. Le poids de l'arête correspond au nombre de collaborations entre eux. Ensuite, nous avons utilisé l'algorithme de Louvain présent dans la bibliothèque **community,** pour détecter les communautés de recherche au sein du réseau, ce qui permet de regrouper les auteurs ayant des interactions fréquentes entre eux. De plus, nous avons calculé la centralité pour chaque auteur, en mesurant le nombre de collaborations qu'il a établies, et la stockée dans la variable **centrality** du graphe. Les auteurs ayant une haute centralité sont considérés comme des acteurs clés dans le réseau.

f. ANALYSE DES CITATIONS

L'analyse des citations vise à comprendre l'impact scientifique des articles publiés dans le domaine de l'IA générative, en se basant sur le nombre de fois où chaque article a été cité, ces valeurs sont tirées de la colonne **cited_by_count,** ce qui va nous permettre d'évaluer l'influence des travaux de recherche.

Nous avons débuté par l'identification des 20 articles les plus influents, puis nous avons étudié la distribution globale des citations, ensuite, nous avons analysé leur évolution dans le temps, ainsi qu'une exploration des liens entre les citations et les clusters

thématiques de KMeans, et enfin l'extraction des mots-clés fréquents dans les titres des articles très cité.

6. ÉVALUATION

Dans l'intention de garantir que les modèles que nous avons utilisés sont à la fois efficaces, performants et capables de généraliser correctement à de nouvelles situations – et donc éviter le surapprentissage – nous avons abordé leurs évaluations en s'appuyant sur des métriques bien choisies.

a. EVALUATION DU MODELE LDA

Pour évaluer le modèle LDA, nous avons utilisé deux métriques :

- Coherence Score :
  Ce métrique permet de mesurer la cohérence sémantique entre les mots les plus représentatifs d'un sujet, plus ce score est élevé plus cette cohérence est forte.

Score	Cohérence	Interprétation
<0.3	Faible	Sujets mélangés
0.3 à 0.5	Moyenne	Sujets relativement cohérents
>0.7	Bonne	Sujets fortement liés

TABLEAU 7 - COHERENCE SCORE

Dans notre étude, nous avons trouvé le score suivant : 0.4478 qui représente une cohérence modérée et des topics relativement cohérents.

- Log-Vraisemblance :
  C'est une mesure statistique qui indique dans quelle mesure le modèle s'ajuste aux données d'entrées. Nous avons exprimé cette mesure à l'aide du log-perplexity qui quantifie la surprise moyenne du modèle face aux mots des documents. Il s'agit d'un grand nombre négatif, plus ce nombre est proche de zéro - donc moins négatif - plus le modèle s'adapte mieux aux données.
  Au cours de nos expérimentations, le meilleur résultat obtenu est une log-perplexité de –7.4546 pour un modèle avec 6 topics.

b. EVALUATION DU MODELE K-MEANS

Dans le but de juger notre modèle de clustering K-Means, nous avons utilisé les trois métriques suivantes :

- Silhouette Score:
  C'est une mesure qui identifie à quel point un point est bien regroupé avec les autres tout en garantissant qu'il est bien séparé des points des autres clusters. Il s'agit d'une valeur entre –1 et 1.
  Si cette mesure est proche de 1, cela signifie que le point est bien dans le cluster, si elle est proche de 0, ce point est entre deux clusters et si elle est négative cela indique que le point est mal regroupé.

Nous avons mesuré plusieurs Silhouette Score pour différents k, la valeur maximale trouvé est de 0.043 pour un k=2.

- Calinski-Harabasz Index:
  Cette métrique mesure le rapport entre la dispersion inter-cluster et intra-cluster. Plus le score est élevé - et donc petites variances intra-clusters et grandes distances inter-cluster – meilleur est le clustering. Le score obtenu est de 62.418.

- Davies-Bouldin Index:
  Le Davies-Bouldin Index mesure la similarité entre chaque cluster et celui qui lui est le proche.Plus il est faible, meilleur est le clustering car un DBI faible témoigne de clusters qui sont compacts et bien séparés. Le score que nous avons obtenu est de 5.316.

V. RESULTATS ET INTERPRETATION

a. RESULTAT DE L'ANALYSE THEMATIQUE

➢ Résultat du LDA

▪ TOPICS IDENTIFIES

Pour des raisons de lisibilité les topics identifiés ont été nommé et décrit manuellement en fonction des mots-clés dominants.

Topic	Nom	Description	Mots-clés
0	Medical and Structural Applications	Des recherches focalisées sur les applications multimodales en imagerie médicale, soins des patients, et des matériaux avec des relations structure-propriété.	["multimodal", "editing", "design", "patient", "medical", "structure", "property", "imaging", "screening","ranging"]
1	Generative AI Systems and Research	Recherches générales sur l'IA générative, ses applications, et ses défis.	["generative", "system", "research", "intelligence", "challenge", "paper", "application", "artificial", "technology", "tool"]
2	Image Generation and Diffusion Models	Recherche sur les modèles génératifs d'images, les modèles de diffusion, et les technologies de génération texte à image.	["image", "generative", "model", "diffusion", "text", "generation", "method", "generated", "quality", "synthetic"]
3	Large Language Models and Natural	Des travaux portant sur les LLMs, les interactions homme-machine, les tâches linguistiques, et prompt engineering.	["model", "language", "llm", "large", "task", "human", "generation", "agent", "prompt", "gpt"]

	Language Generation		
4	Neural Networks and Learning Methods	Des recherches fondamentales sur les architectures des réseaux de neurones, les méthodes d'apprentissage, et les approches d'entrainement.	["model", "data", "learning", "network", "based", "training", "method", "approach", "task", "algorithm"]
5	ChatGPT and Human-AI Interaction Studies	Recherches spécifiques sur l'utilisation de Chatgpt, expérience utilisateur, et implications sociales de l'IA générative.	["study", "chatgpt", "user", "generated", "human", "analysis", "social", "use", "using", "used"]

TABLEAU 8 -TOPICS IDENTIFIES LDA

Pour déterminer le sujet dominant de chaque année entre 2019-2024, nous avons calculé pour chaque document les distributions des topics et puis sélectionner celui qui a la probabilité la plus élevée.

- **VISUALISATION DES TOPICS**

  Après avoir regroupé les documents par année et topic dominant, un graphe d'évolution temporelle (fig.11) est élaboré, pour illustrer l'évolution des sujets au fil d'année 2019-2024.

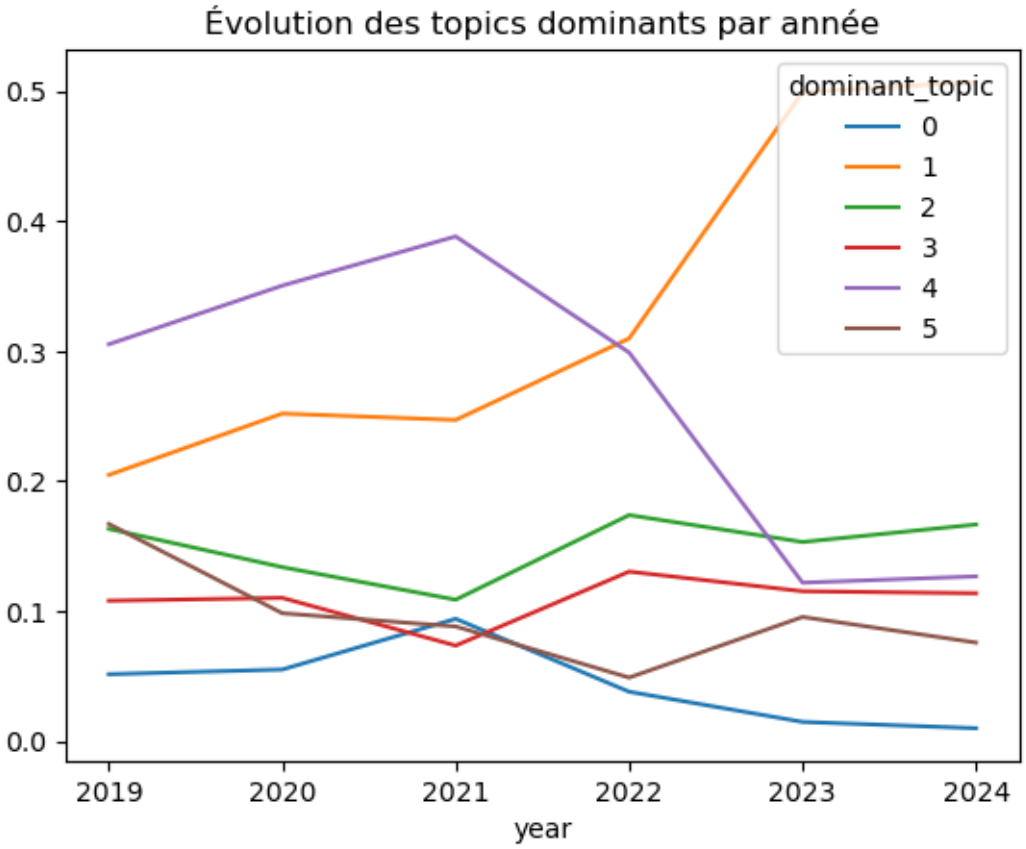

FIGURE 11 - EVOLUTION DES TOPICS DOMINANTS PAR ANNEE

On observe que jusqu'au 2021 les recherches sur les réseaux neuronaux et méthodes d'apprentissage demeurent dominantes. A partir de 2022, les recherches se concentrent principalement sur les applications de l'IA générative et ses défis. En outre, les modèles de diffusion et les modèles génératifs d'image prennent de plus en plus de place. En 2023-2024, les recherches concernant l'utilisation de Chatgpt ont connu un pic.

- **Résultat du BERTopic**

- **TOPICS IDENTIFIES**

Les topics principaux identifiés par le modèle BERTopic sur un corpus constitué de 3437 articles, reflète la richesse et la diversité du vocabulaire.

Ce modèle nous a permi d'identifier 14 topics, voici les 10 premiers avec leurs descriptions et représentation sémantique :

Topic	Nom	Total
0	-1_model_generative_language_data	897
1	0_model_image_learning_generative	2055
2	1_material_device_high_light	88
3	2_removal_metal_carbon_composite	67
4	3_protein_metal_carbon_composite	63
5	4_cell_gene_single_data	39
6	5_market_supply_chain_risk	34
7	6_health_patient_disease_case	34
8	7_quantum_circuit_classical_network	27
9	8_user_social_graph_network	21

TABLEAU 9 -TOPICS IDENTIFIES BERTOPIC

On remarque que le topic 0 est dominé par les concepts liés aux modèles génératifs et au langage naturel, ce qui reflète l'importance croissante des LLMS. De plus, le topic 1 se focalise sur les modèles génératifs de génération d'image, ainsi que des méthodes d'apprentissage liées aux modèles de diffusion. Par ailleurs, les topics 2 à 9 couvrent des domaines variés tels que la science des matériaux, la biotechnologie, et les applications industrielles.

- **VISUALISATION DES TOPICS**

En générant une visualisation interactive avec la fonction **visualize_topics**, nous avons pu d'observer clairement la distribution des topics dans un espace à deux dimensions, où chaque topic est représenté par un cercle de taille proportionnelle à son importance relative dans le corpus.

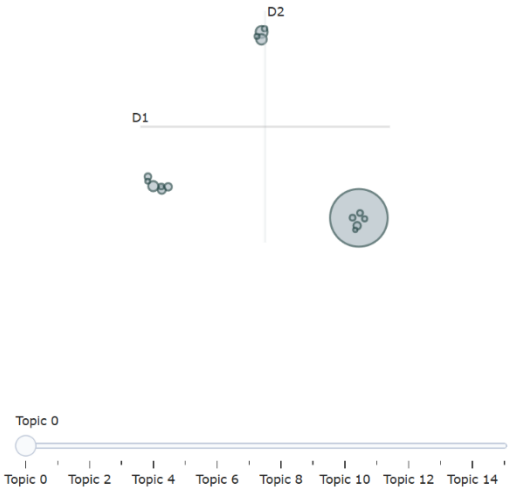

FIGURE 12 - VISUALISATION DES TOPICS: BERTOPIC

- **EVOLUTION TEMPORELLE DES TOPICS**

La fonction topics_over_time nous a permis de suivre l'évolution temporelle des sujets extraits par le modéle BERTopic, nous avons obtenu un tableau de 40 lignes. Les résultats sont illustrés par le graphe ci-dessous, cette courbe représente pour chaque topic la fréquence normalisée de ses mots-clés entre 2019-2024.

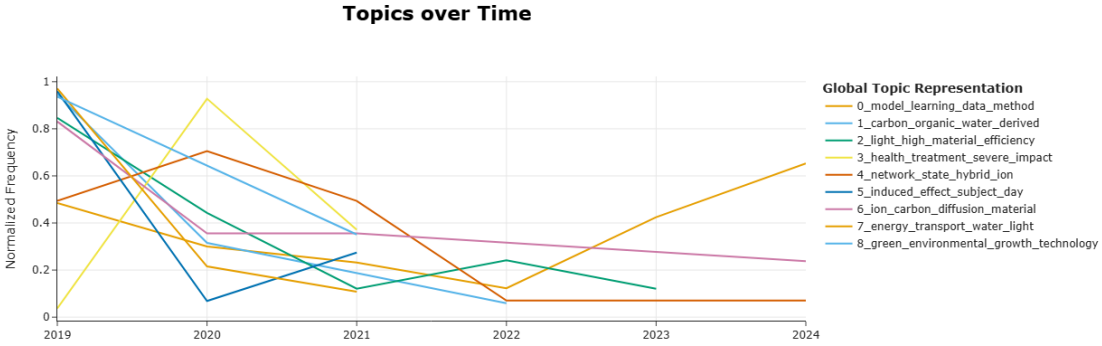

FIGURE 13 - EVOLUTION TEMPORELLE DES TOPICS

Le topic 0 reste stable jusqu'en 2024, avec une baisse légère relative, ceci reflète une continuité des recherches fondamentales autour des modèles classiques comme GANs et les réseaux neuronaux.

Le topic 1 apparaît en 2019 avec des expressions liées à la chimie organique et l'eau, au fil des années les travaux en ce domaine en intégration avec l'IA générative diminuent progressivement.

Le topic 2 présent dès 2019, lié à des applications physiques et environnementales, montre une stabilité relative jusqu'en 2021.

Le topic 3 apparaît en 2019, concerne les applications médicales malgré une forte expansion en 2020 due à la pandémie, ce sujet a diminué rapidement.

Le topic 4 présent dès 2019 et relié à des architectures avancées de réseaux neuronaux, montre une évolution fluctuante avec une diminution notable après 2021.
Le topic 5 est lié à des analyses causales et des expériences contrôlées. Cette variabilité pourrait refléter des recherches ponctuelles portant sur l'impact d'un phénomène spécifique sur un groupe donné, dans un cadre expérimental ou observationnel.
Le topic 6 présent dès 2019, s'appuie sur des concepts clés tels que les ions, le carbone, la diffusion et les matériaux. Son évolution montre une présence significative jusqu'en 2020.
Le topic 7 apparaît en 2019, repose sur des concepts clés notamment l'eau, l'énergie, le transport et la lumière. Il a connu une stabilité relative jusqu'à 2021 avant de commencer à décliner. Cette évolution peut être liée à une spécialisation vers des applications environnementales et énergétiques plus ciblées, tels que le transport durable, l'optimisation énergétique et de l'intégration des énergies renouvelables.
Le topic 8 repose sur des concepts comme le développement durable, l'environnement, la croissance verte, et les technologies innovantes.
Nous avons observé que parmi les résultats existe un topic –1 qui peut correspondre aux articles non attribués à un sujet spécifique.

En somme, cette analyse a révélé que l'IA générative ne se limite plus aux seuls domaines techniques, mais s'étend à des applications transversales et impact sociétaux.

b. RESULTAT DU CLUSTERING KMEANS

➢ Choix du nombre de clusters

TABLEAU 10 - CHOIX DU NOMBRE DE CLUSTERS

Afin d'assurer la qualité de la segmentation du corpus en clusters, l'algorithme doit être configuré par le bon nombre de clusters. Pour cela, la mesure de la métrique Silhouette Score a permis de choisir la valeur optimale. Les résultats correspondant à chaque valeur de k sont présentés dans le tableau suivant :

Silhouette Score	Nombre de clusters (k)
0.043	2
0.020	3
0.015	4
0.037	5
0.006	6
0.027	7
0.020	8
0.022	9
0.031	10

0.023	11
0.029	12
0.010	13
0.015	14

TABLEAU 11 - CHOIX DU NOMBRE DE CLUSTERS

D'après le tableau, et en observant la courbe ci-dessous, nous constatons que la valeur optimale de nombre de clusters est k=2, car elle correspond au maximum global du Silhouette Score.

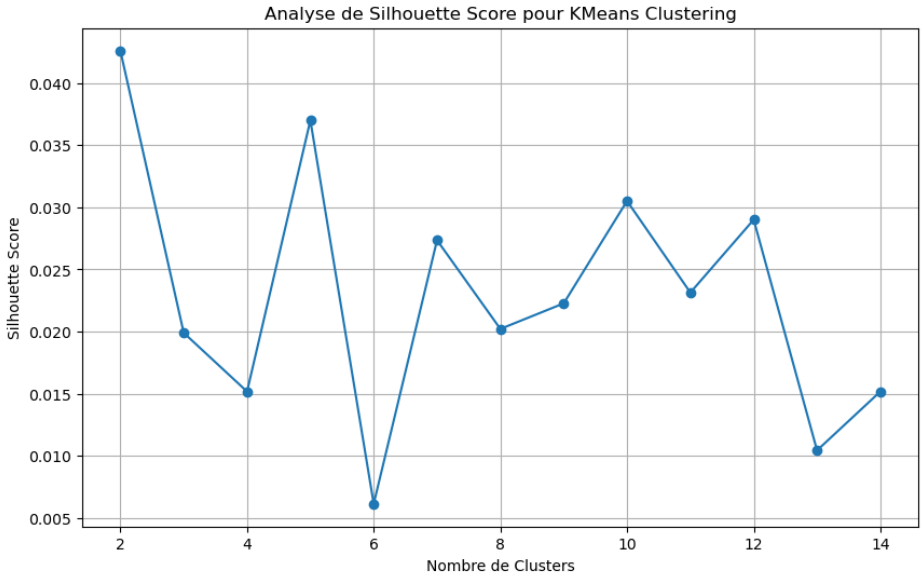

FIGURE 14 - CHOIX DU NOMBRE DE CLUSTERS

➢ Taille des clusters

Notre corpus se divise en deux clusters, chacun contient les documents qui présentent des similarités entre eux. Voici la taille des clusters obtenus :

Cluster	**Nombre d'éléments**
0	534
1	2903

FIGURE 15 - TAILLE DES CLUSTERS

Cela montre une imbalance significative entre les deux clusters.

➢ Visualisation 2D avec UMAP

Après avoir appliqué l'algorithme aux documents, et pour comprendre leur distribution spatiale dans l'espace réduit par la méthode TruncatedSVD, nous avons utilisé **UMAP** (Uniform Manifold Approximation and Projection) qui est une technique d'analyse non dimensionnelle non linéaire, et sert à projeter les données dans un espace à deux dimensions. Elle permet de visualiser les similarités entre les documents tout en préservant leur structure globale.

FIGURE 16 - VISUALISATION 2D AVEC UMAP

Les points verts représentent le cluster 1 et les points bleus correspondent au cluster 0, on observe qu'il y a une séparation nette entre les deux groupes, bien que certains points soient proches de frontières, ce qui suggère un chevauchement entre certains docments. En fait, malgré que le cluster 0 est plus petit, il est densément concentré, et peut représenter des articles spécialisés, des domaines moins fréquents ou des sous-groupes spécifiques de recherches. Par ailleurs, le cluster 1 est dominant, et contient une grande variété de sujets, qui reflètent probablement les tendances principales de la recherche en IA générative. De plus, certaines zones de chaque cluster contiennent des points proches, suggérant l'existence de sous-thèmes bien définis.

- **Résultat de l'analyse temporelle des clusters**

Pour analyser l'évolution temporelle des clusters identifiés par KMeans, nous avons mené un diagramme illustrant la proportion de chaque cluster pour chaque année entre 2019 et 2024.

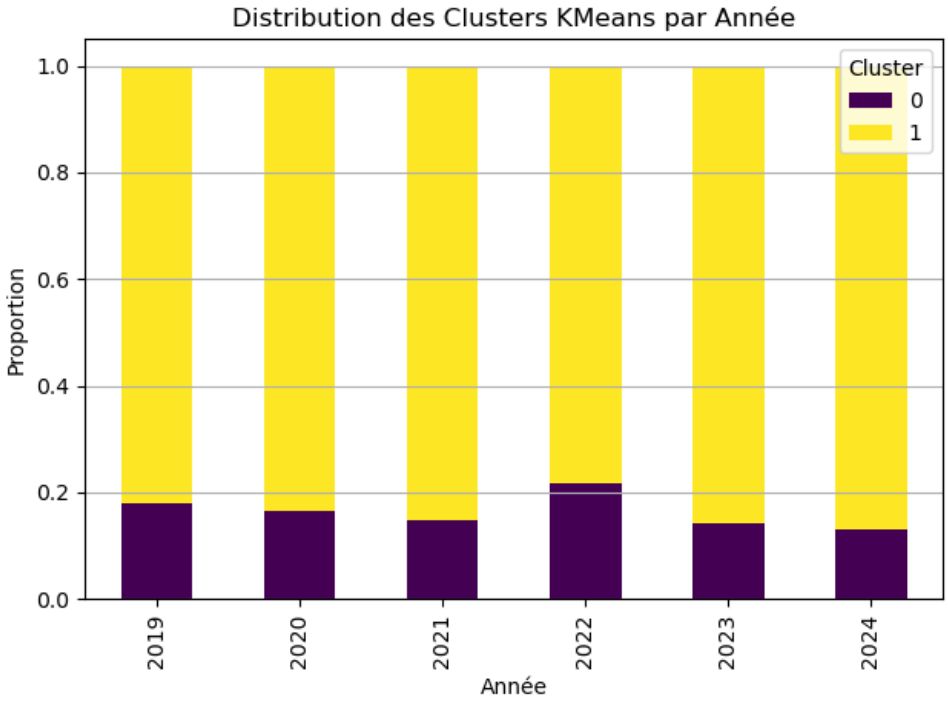

FIGURE 17 - DISTRIBUTION DES CLUSTERS KMEANS PAR ANNEEE

On observe que le cluster 0 a une présence minoritaire variant légèrement entre 15% et 22% selon les années. Il reste presque stable, indiquant qu'un sous-groupe de recherches demeure intacte tout au long de la période étudiée. En revanche, le cluster 1 montre une proportion majoritaire, plus de 80% des documents publiés appartient à ce cluster, ce qui correspond probablement aux tendances principales de l'IA générative.

La prédominance constante du cluster 1 témoigne une convergence des préoccupations scientifiques autour de l'IA générative, probablement centrée sur des thèmes majeurs comme les LLMs ou la génération des images. Par ailleurs, le maintien d'un cluster minoritaire suggère toutefois l'existence d'un courant de recherches complémentaire émergent, en l'occurrence le pic relatif qu'a connu ce cluster en 2022 indique une brève croissance des sous-domaines ou des problématiques ayant suscité un intérêt ponctuel.

c. RESULTAT DE L'ANALYSE GEOGRAPHIQUE

- **Les 15 pays les plus contribuants en recherches sur l'IA générative**

  Afin de visualiser les 15 pays les plus actifs en ce domaine, nous avons mené un diagramme en barres horizontalement, qui montre pour chaque pays son nombre de publications couvrant la période de notre étude.

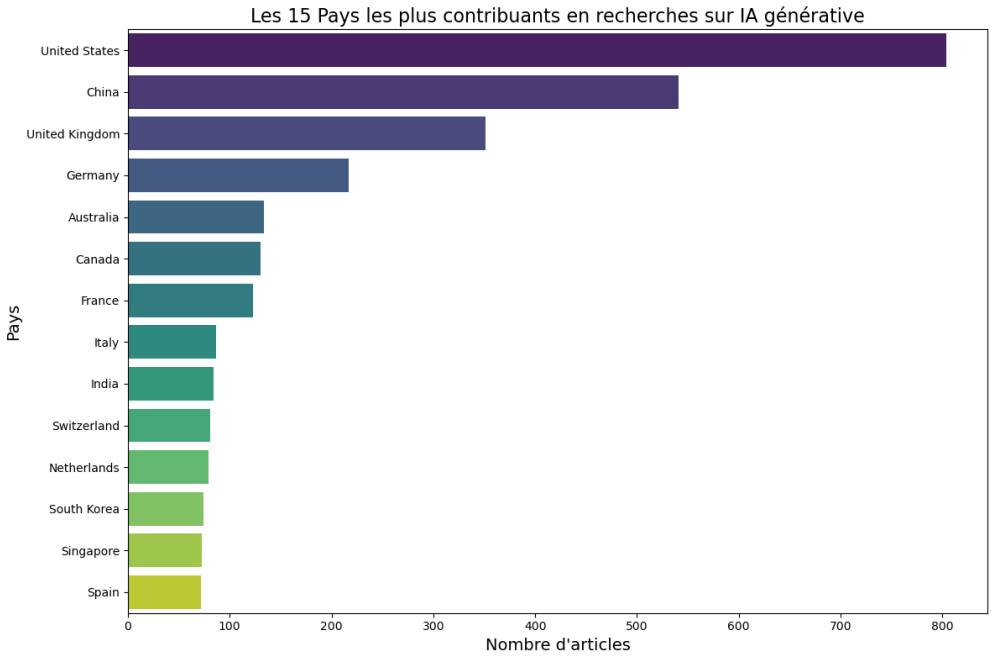

FIGURE 18 - LES 15 PAYS CONTRIBUANTS

On remarque clairement que les Etats-Unis dominent largement avec 800 articles, suivi de la Chine et du Royaume-Uni. Ces trois pays constituent 70% des contributions totales.

- **Les tendances de recherche des 5 pays les plus contributifs**

  L'évolution annuelle des contributions des 5 premiers pays, est illustrée par le diagramme en courbes ci-dessous. L'axe des abscisses représente la période 2019-2024, l'axe des ordonnées indique le nombre d'articles, et chaque courbe symbolise un pays selon une couleur attribuée. La légende choisie simplifie la distinction entre les courbes des pays :

  – Bleu pour l'Australie (AU)
  – Orange pour la Chine (CN)
  – Vert pour l'Allemagne (DE)
  – Rouge pour le Royaume-Uni (GB)
  – Violet pour les Etats-Unis (US)

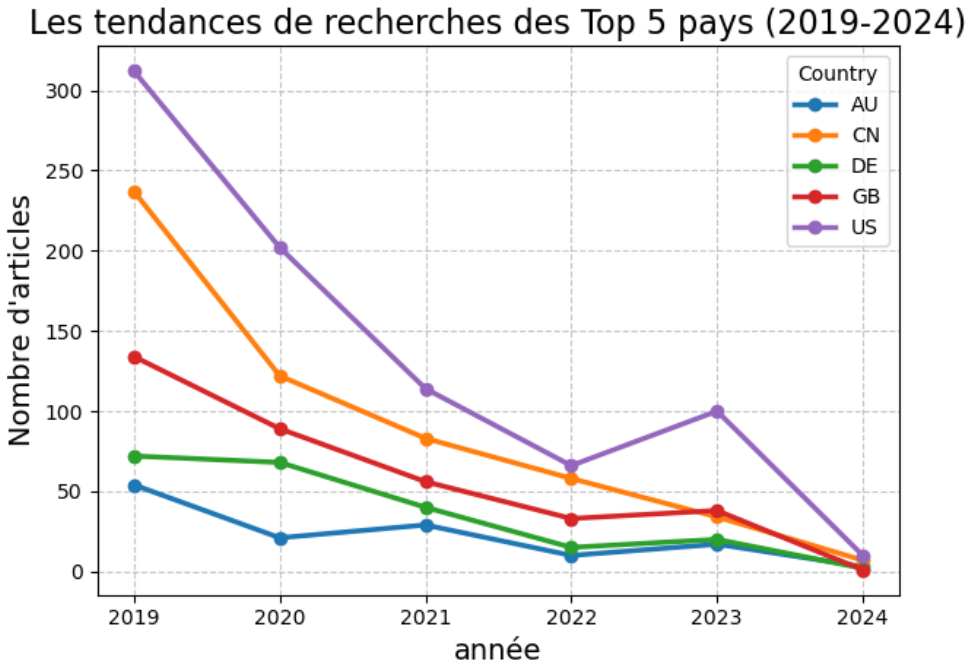

FIGURE 19 - LES 5 PAYS PRINCIPAUX

On observe que la contribution des Etats-Unis demeure dominante tout au long de la période, malgré la baisse significative qu'a connu son nombre de publications. La Chine présente une forte croissance jusqu'à 2020, puis le nombre de ses publications chute linéairement pour atteindre une valeur très bas en 2024. Le Royaume Uni présente une tendance générale à la baisse de 2019 à 2024, bien que sa contribution reste moins que la Chine, elle connaît une certaine augmentation entre 2022 et 2023. L'Allemagne montre une activité relativement stable entre 2019 et 2020, suivie d'une diminution progressive jusqu'en 2024. L'Australie maintient le nombre le plus bas parmi les cinq pays tout au long de la période. Elle présente une certaine fluctuation, qui reste marginale par rapport aux autres.

Néanmoins, ce diagramme reflète la diminution de l'activité de recherche en IA générative dès 2019, bien que certains pays aient connu une stabilisation ou une légère baisse après 2023, les Etats-Unis restent nettement en tête, reflétant leur dominance historique dans la recherche scientifique, qui peut être expliqué par l'infrastructure académique solide, des financements importants, ou une forte collaboration internationale. La Chine se positionne quant à elle comme le deuxième contributeur majeur, confirmant son rôle dans le paysage mondial de l'IA générative.

d. RESULTAT DE L'ANALYSE INSTITUTIONNELLE

- **Les 20 principales institutions**

  Nous nous sommes servies d'un diagramme en barres horizontales pour présenter les 20 principales institutions en termes de recherches sur l'IA générative.

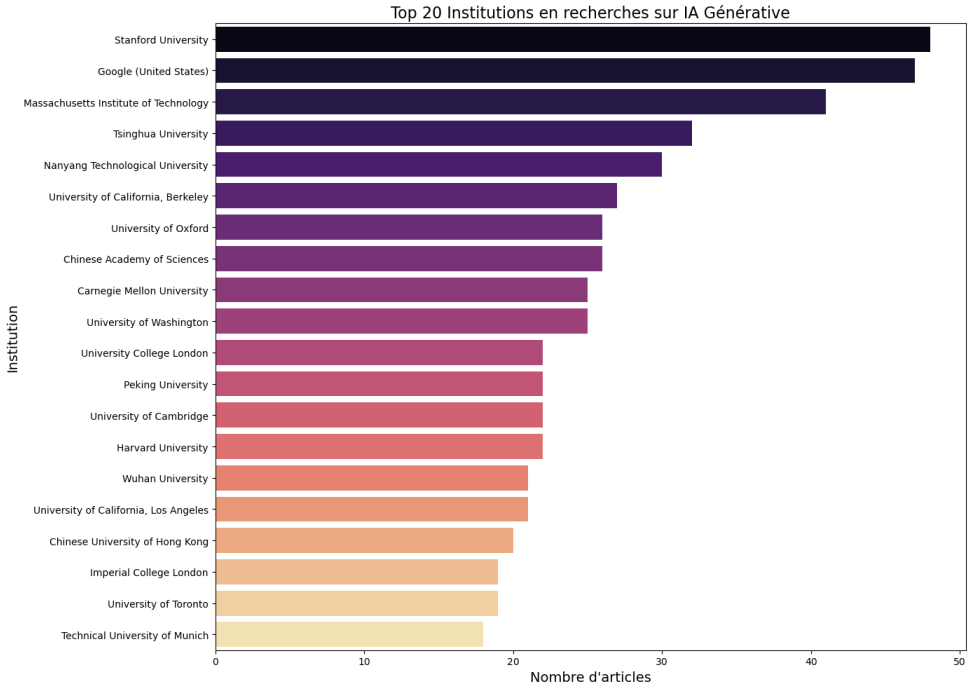

FIGURE 20 - LES 20 PRINCIPALES INSTITUTIONS

Nous observons que l'institution Stanford University est la plus présente avec près de 50 articles publiès, suivi de Google (Etats-Unis), Massachusetts Institut of Technology (MIT), et l'université Tsinghua University de la Chine avec un nombre d'articles élevé également. Les autres institutions ont entre 10 et 30 articles chacune. On peut constater que les principales institutions impliquées dans la recherche en IA générative sont des universités renommées notamment en Etats-Unis et en Chine.

Ce graphique montre que les recherches en ce domaine sont menées de manière intense avec une concurrence serrée. Les institutions leaders dans ces recherches sont celle disposant de moyens et de talents importants, leur permettant de produire une quantité élevée de publications.

- **Les tendances de recherches des 10 principales institutions**

Le Heatmap suivant montre la répartition des recherches des 10 institutions sur les deux clusters identifiés par l'algorithme KMeans. Les institutions sont listées sur l'axe verticale, tandis que les clusters sont représentés sur l'axe horizontal. La couleur et l'intensité indiquent le nombre d'articles publiés par chaque institution dans chaque cluster. On peut ainsi observer les tendances de recherche de chaque institution selon ces deux clusters.

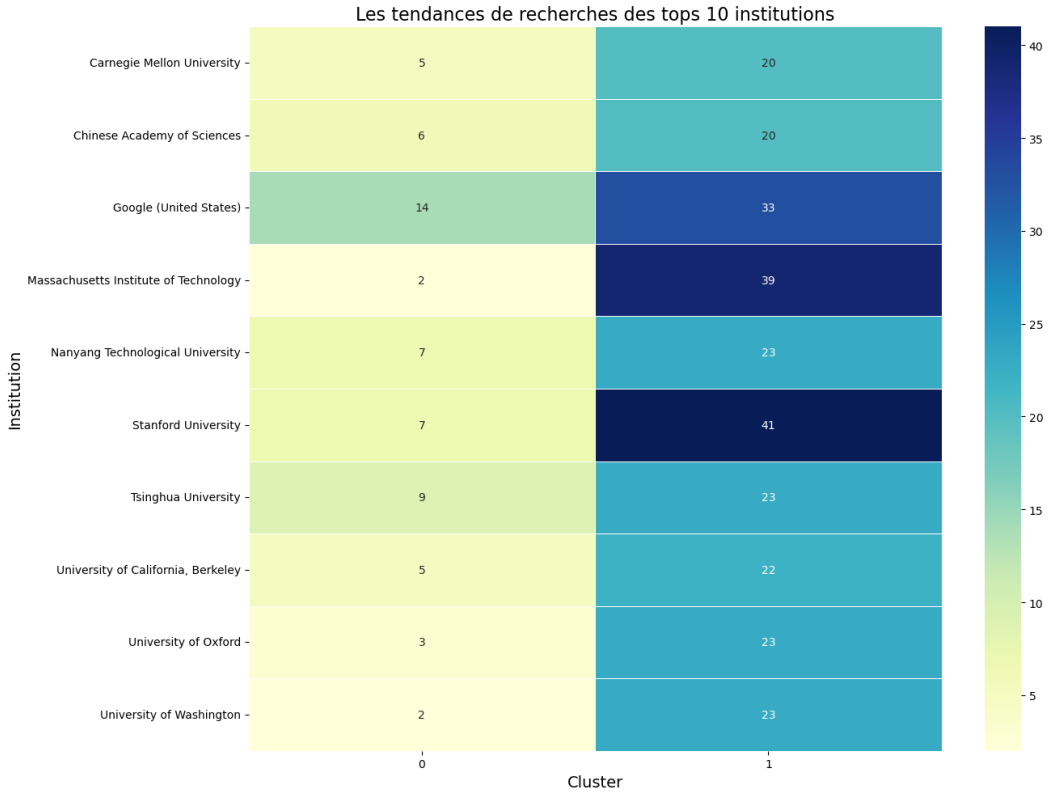

FIGURE 21 - HEATMAP DES TENDANCES DE RECHERCHES DES 10 PRINCIPALES INSTITUTIONS

On remarque que les institutions Standford University et MIT se concentrent principalement sur le cluster 1 avec plus de 39 publications, montrant ainsi un fort intérêt sur des domaines spécifiques en IA générative .
Bien que la majorité des institutions sont actives dans le cluster 1, les institutions Google et Tsinghua University semblent être les principaux contributeurs dans le cluster 0. Cela suggère une certaine complémentarité des centres de recherches permettant d'identifier des collaborations possibles.
Cette cartographie institutionnelle met en lumière la diversité des approches adoptées par les leaders mondiaux de la recherche en IA générative, et met en évidence des synérgies potentielles entre certaines institutions et certains clusters, ce qui illustre la richesse des travaux menés dans ce champs en pleine expansion.

- **Nuage de mots des institutions**

En profitant de ce nuage de mots, on peut distinguer les différentes institutions actives en ce domaine.
On remarque que parmi les mots les plus grands existent University, Technology, et Institute, reflétant ainsi que les principaux acteurs sont des établissements universitaires et technologiques. En outre, de nombreuses universités ont été renommées mondialement, notamment MIT, Stanford University, Harvard, Nanjing University, Tsinghua Universty, Oxford, Nanyang Technological ...etc. De plus, certains autres types d'institutions sont également présents, comme Google United, Medical Center, Microsoft Research...etc, ainsi que des laboratoires en l'occurrence National Laboratory.

FIGURE 22 - WORDCLOUD DES INSTITUTIONS DE RECHERCHES

Ce nuage de mots met en évidence les principaux acteurs de la recherche scientifique, avec une prédominance des grandes universités en domaine de l'innovation et la technologie. La présence forte des établissements américains et chinois reflète leur leadership dans ce secteur. Ainsi, la diversité des institutions impliquées, souligne l'existence d'un écosystème riche et varié de la recherche où différentes organisations contribuent aux avancées scientifiques et technologiques.

e. RESULTAT DE L'ANALYSE DES AUTEURS

Le réseau créé (fig.) contient **1046** auteurs qui sont symbolisés par des nœuds, et **3587** collaborations qui sont représentées par des arêtes. Les nœuds sont colorés selon leur communauté, et leur taille est proportionnelle à leur centralité de degré, ainsi que les arêtes sont ajustées en fonctions des fréquences de collaboration (poids). Ce réseau ne contient que les auteurs qui ont publié au moins deux publications.

- **Les 20 auteurs les plus centraux dans le réseau de collaboration**

Nous avons identifié les 20 auteurs prolifiant dans ce domaine de recherche et les triés par ordre décroissant selon la centralité. Le tableau ci-dessous contient les résultats obtenus :

Auteur	Centralité	Communauté
Dacheng Tao	0.0364	12
Stefan C. Dentro	0.0325	78
Ignaty Leshchiner	0.0325	78
Kerstin Haase	0.0325	78
Maxime Tarabichi	0.0325	78
Jeff Wintersinger	0.0325	78
Amit G. Deshwar	0.0325	78
Kaixian Yu	0.0325	78
Yulia Rubanova	0.0325	78

Geoff Macintyre	0.0325	78
Jonas Demeulemeester	0.0325	78
Kortine Kleinheinz	0.0325	78
Dimitri Livitz	0.0325	78
Salem Malikić	0.0325	78
Nilgun Donmez	0.0325	78
Pavana Anur	0.0325	78
Clemency Jolly	0.0325	78
Marek Cmero	0.0325	78
Daniel Rosebrock	0.0325	78
Yu Fan	0.0325	78

TABLEAU 12- LES 20 AUTEURS LES PLUS CENTRAUX DANS LE RESEAU DE COLLABORATION

On remarque que Dacheng Tao apparaît comme l'auteur le plus central, avec une valeur de centralité de 0.0364, ce qui montre qu'il est un acteur clé dans ce domaine de recherche. Les 19 suivants ont la même centralité 0.0325 ce qui indiquent qu'ils forment un groupe d'auteurs très connectés et collaborant étroitement, ce qui est justifié par leur appartenance à la même communauté 78.

- **Le réseau de collaboration des auteurs**

Voici le réseau de collaboration entre les auteurs travaillant sur l'IA générative.

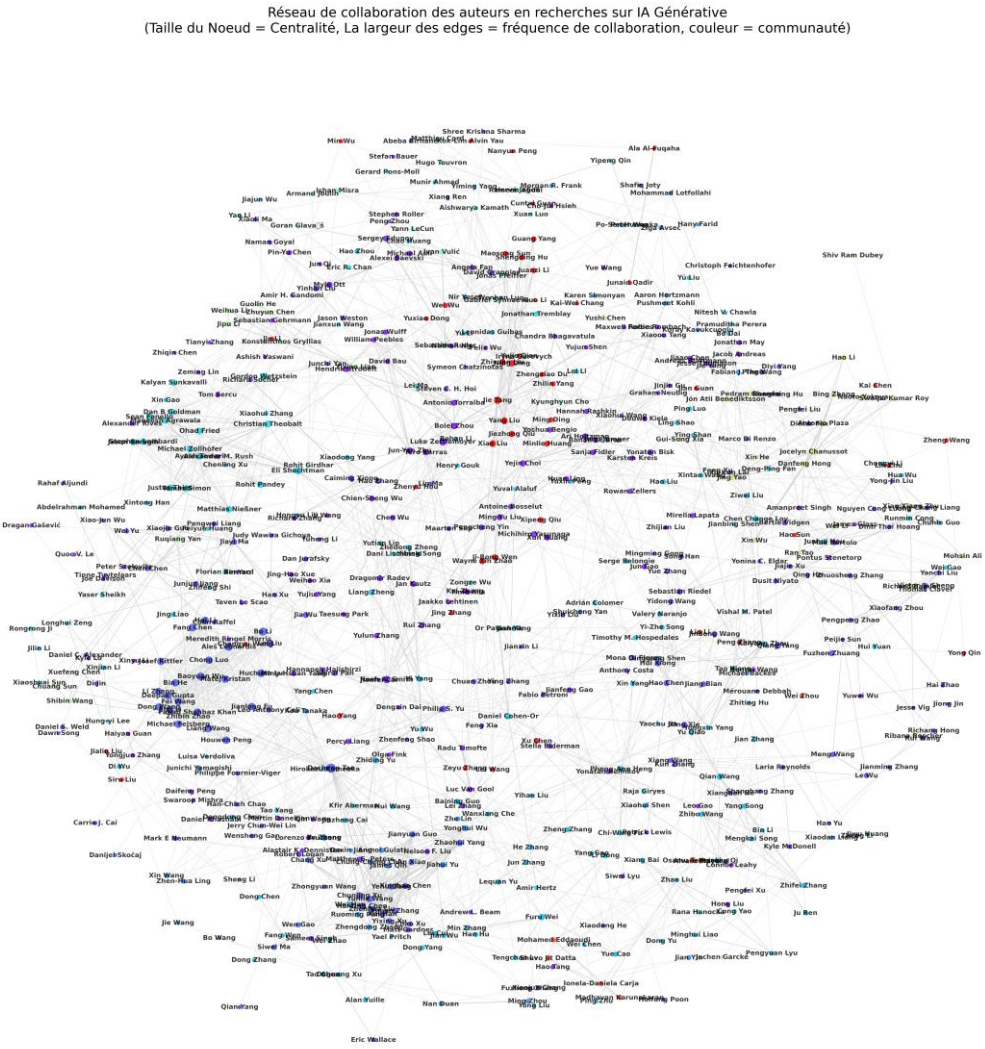

FIGURE 23 - LE RESEAU DE COLLABORATION DES AUTEURS

La taille importante du réseau indique une activité collaborative intense dans ce domaine.

- **Les communautés d'auteurs les plus importantes**

Les communautés les plus larges sont listées comme suit :

- La communauté 4 constituée de108 auteurs.
- La communauté 26 constituée de 68 auteurs.
- La communauté 8 constituée de 66 auteurs.
- La communauté 1 constituée de 56 auteurs.
- La communauté 28 constituée de 52 auteurs.

Ces résultats indiquent qu'il existe des groupes d'auteurs collaborant de manière particulièrement étroite au sein de ce réseau.

- **Sous graphe des communautés de recherches**

Afin de voir les structures internes des quatre principales communautés de recherche, nous avons créé des sous-graphes pour chacune.

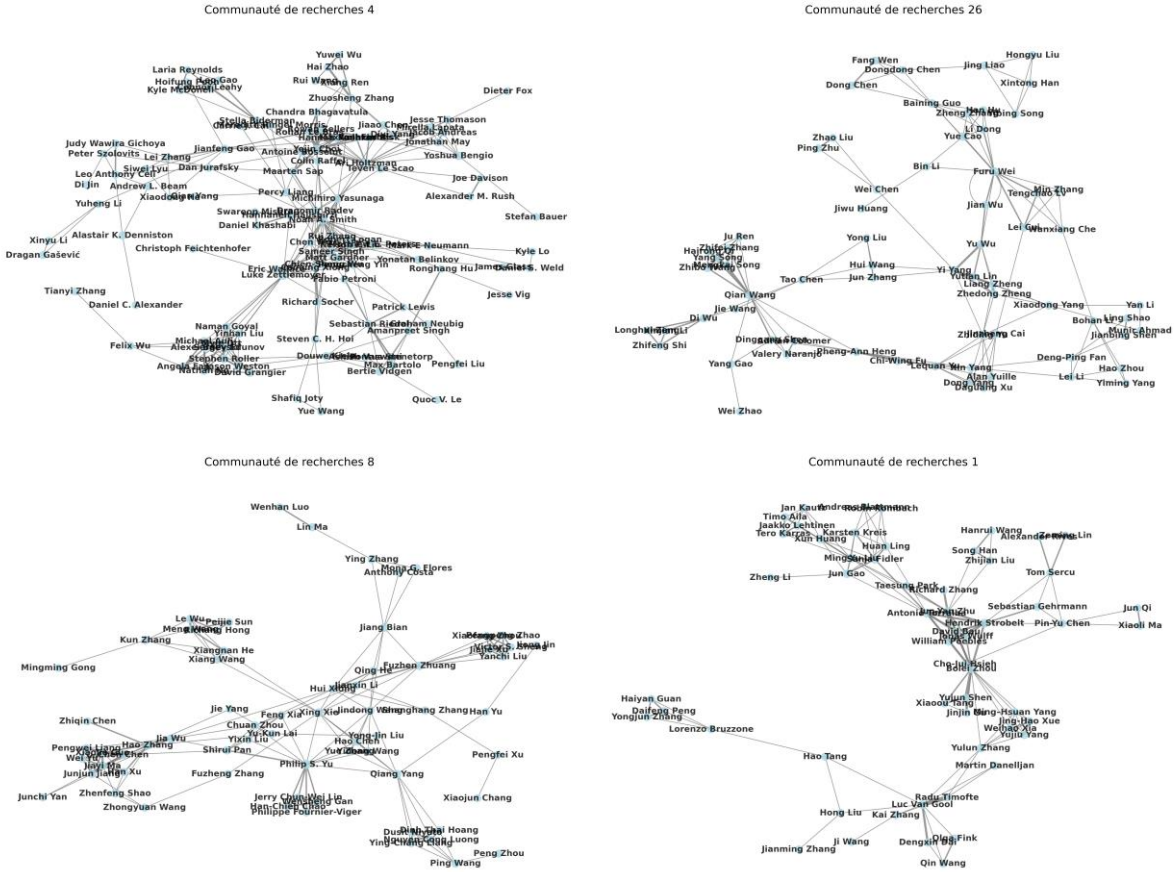

FIGURE 24 - SOUS GRAPHE DES COMMUNAUTES DE RECHERCHES

Nous observons que chaque communauté présente des clusters d'auteurs fortement connectés, avec des collaborations intenses au sein de ces groupes. Pour chaque cluster existe un auteur central qui est visible par son nœud plus grand que les autres nœuds.
La présence de ces communautés bien structurées suggère qu'il existe des domaines thématiques scientifiques plus spécialisé au sein de l'IA générative. Et les auteurs centraux jouent un rôle de catalyseurs de la recherche dans leur domaine spécifique.

f. RESULTAT DE L'ANALYSE DES CITATIONS

- **Statistiques descriptives des citations**

Nombre total d'articles	3437
Moyenne (mean)	~243 citations
Écart-type (std)	~451 citations
Minimum	0 citation
Médiane (50%)	196 citations
Maximum	9215 citations

TABLEAU 13 - STATISTIQUES DESCRIPTIVES DES CITATIONS

Le jeu de données contient **3437** observations, avec une moyenne de **242,53** de citations de chaque article. La dispersion des valeurs est assez élevée, indiquant une grande

variabilité dans le nombre de citations reçues par les articles, la valeur minimale de citation est 0, tandis que l'article le plus cité a reçu 9215 citations. En effet, 25% des articles n'ont reçu aucune citation, 50% des articles ont reçu 196 citations, et 75% ont 303 ou moins.

Nous constatons donc que la distribution des citations est asymétrique, avec de nombreux articles peu cités et d'autres fortement cités. Ainsi, les articles qui se démarquent nettement avec un très grand nombre de citations, indiquent leur influence dans ce domaine.

- **Les 20 papiers les plus influents**

Pour compléter l'analyse descriptive déjà menée, nous avons exploré les 20 articles les plus cités dans le domaine de l'IA générative (Tableau), ceci va permettre d'identifier les travaux fondateurs qui ont façonné ce domaine.

Rang	Titre	Année	Citation
1	A survey on Image Data Augmentation for Deep Learning	2019	9215
2	Generative adversarial networks	2020	8822
3	High-Resolution Image Synthesis with Latent Diffusion Models	2022	6378
4	Transformers: State-of-the-Art Natural Language Processing	2020	6126
5	Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI	2019	5967
6	Dynamic Graph CNN for Learning on Point Clouds	2019	5237
7	BioBERT: a pre-trained biomedical language representation model for biomedical text mining	2019	5176
8	Analyzing and Improving the Image Quality of StyleGAN	2020	4729
9	OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields	2019	4114
10	Unsupervised Cross-lingual Representation Learning at Scale	2020	4040
11	BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis	2019	3478
12	On the Dangers of Stochastic Parrots	2021	2965
13	Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting	2021	2929
14	Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy	2019	2485
15	Semantic Image Synthesis With Spatially-Adaptive Normalization	2019	2484
16	A survey of the recent architectures of deep convolutional neural networks	2020	2476
17	fairseq: A Fast, Extensible Toolkit for Sequence Modeling	2019	2464
18	SciBERT: A Pretrained Language Model for Scientific Text	2019	2414
19	Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models	2023	2388
20	ChatGPT for good? On opportunities and challenges of large language models for education	2023	2371

TABLEAU 14 -LES 20 PAPIERS LES PLUS INFLUENTS

D’après le tableau, on constate que les GANs, le StyleGAN, les Transformers et les Diffusion Models apparaissent clairement comme des piliers du domaine. Les modèles de génération d’image sont aussi parmi les très hauts scores de citation. Ainsi, des sujets transversaux et sociétaux notamment les sujets 12, 19 et 20 soulignent la prise en conscience éthique autour des grands modèles linguistiques et les applications de Chatgpt dans l’éduction. De plus, on remarque que la majorité des articles très cités datent des années 2019-2021, ce qui correspond à la phase de l’innovation technologique en IA générative. Cependant, les articles de 2023 montrent une évolution vers les grands questions éthiques et sociétales liées aux modèles de langage.

- **Les mots clés fréquents dans les titres des 20 articles les plus cités**

  En appliquant la fonction CountVectorizer sur les titres de ces 20 articles, nous obtenons les résultats suivants :

Terme	Fréquence
language	5
image	4
ai	3
challenges	3
learning	3
models	3
opportunities	3
artificial	2
biomedical	2
chatgpt	2
deep	2
education	2
intelligence	2
large	2
model	2

TABLEAU 15 - LES MOTS CLES FREQUENTS DANS LES TITRES DES 20 ARTICLES LES PLUS CITES

Ces résultats montrent l’importance autour de certains thèmes:

- Le terme language est dominant, soulignant l’imprtance des modèles linguistiques comme Chatgpt.
- Le terme Image indique une forte présence de travaux liés à la génération visuelle comme les Diffusion Models.
- Les termes AI, Deep Learning, Intelligence reflètent le cœur technique du domaine.
- Les termes Chatgpt, education, opportunities, challenges montrent l’intérêt croissant pour les implications sociétales et pédagogiques de l’IA générative.

- **Distribution des citations par article**

Le diagramme suivant montre la distribution du nombre de citations pour un ensemble de publications sur l'échelle logarithmique pour visualiser correctement cette distribution.

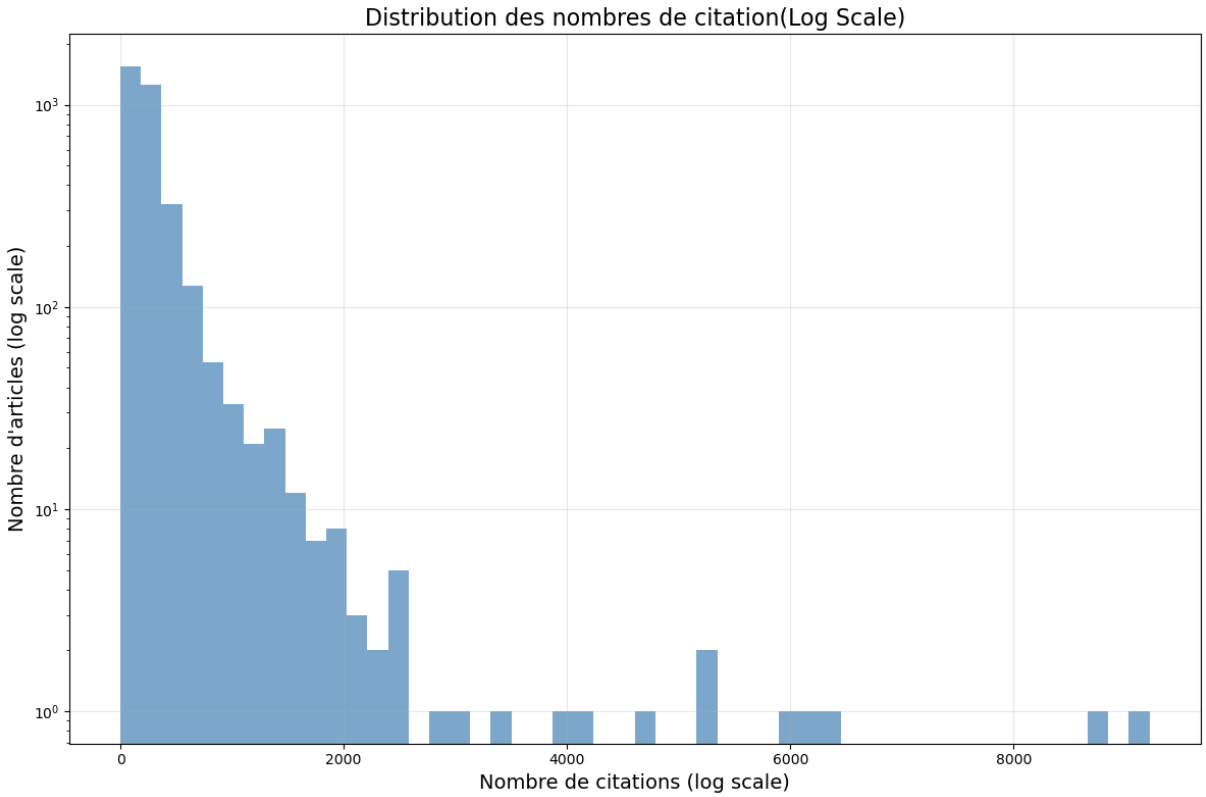

FIGURE 25 - DISTRIBUTION DES CITATIONS PAR ARTICLE

On remarque que la distribution des citations est asymétrique, la plupart des articles reçoivent peu de citations, en revanche certains deviennent des références majeures.

- **Distribution des citations par année**

Pour explorer l'évaluation de la moyenne des citations reçues par les articles au fil des années, nous avons utilisé un diagramme en barres ayant comme axe d'abscisse les années 2019-2024 et en ordonnée le nombre des citations.

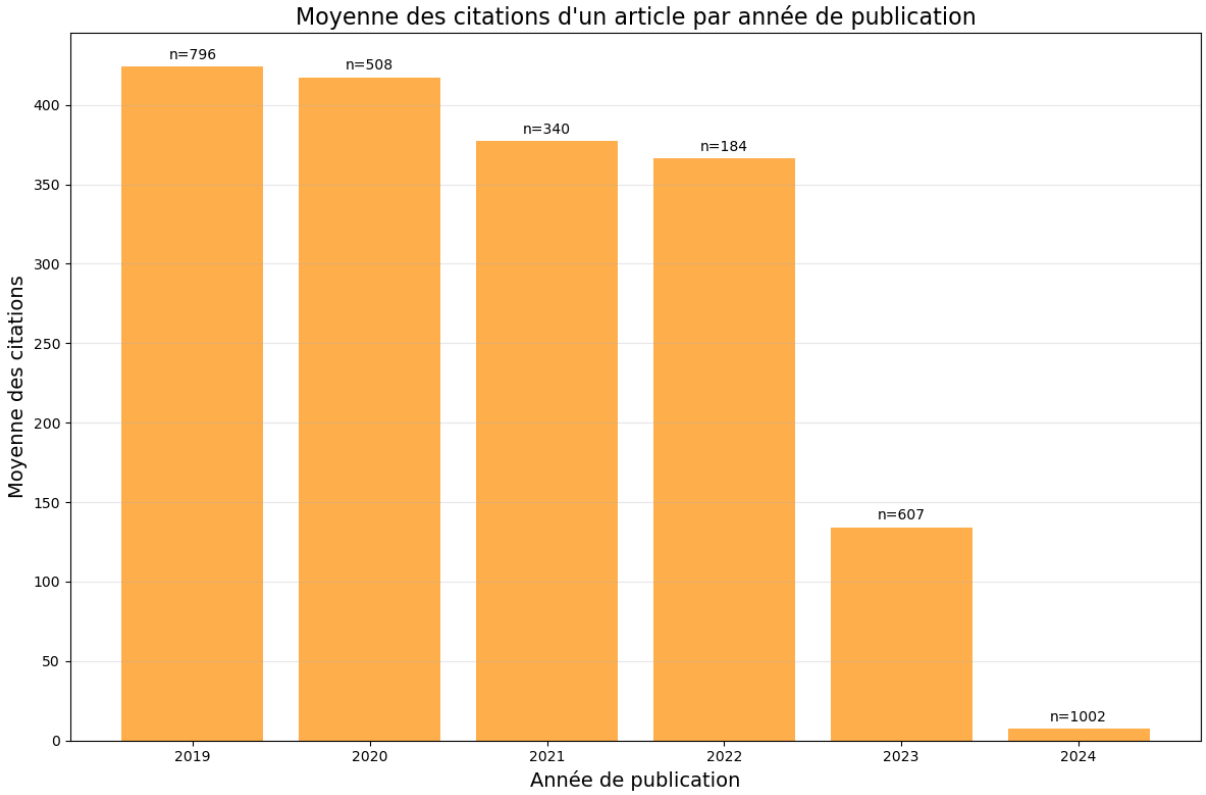

FIGURE 26 - DISTRIBUTION DES CITATIONS PAR ANNEE

En analysant ce diagramme, on observe que la moyenne des citations est la plus élevée pour les publications de 2019, avec environ 400 citations pour 796 articles. Ensuite, la moyenne des citations diminue progressivement pour les années suivantes, atteignant environ 140 citations pour 607 articles en 2023.

Cette représentation permet de visualiser clairement la baisse de la moyenne des citations a fur et à mesure que le temps passe depuis la publication des articles. Cela suggère que les publications récentes (2022-2024) ont en moyenne un impact moindre que celles publiées en 2019.

- **Distribution des citations par cluster**

Le diagramme montre deux clusters de recherche distincts, représentés sur l'axe horizontal, ainsi que la moyenne des citations d'un article sur l'axe vertical. Le premier cluster, représenté à gauche, a une moyenne de citations d'environ 310 par article. Le second, a une moyenne de citations d'environ 240 par article.

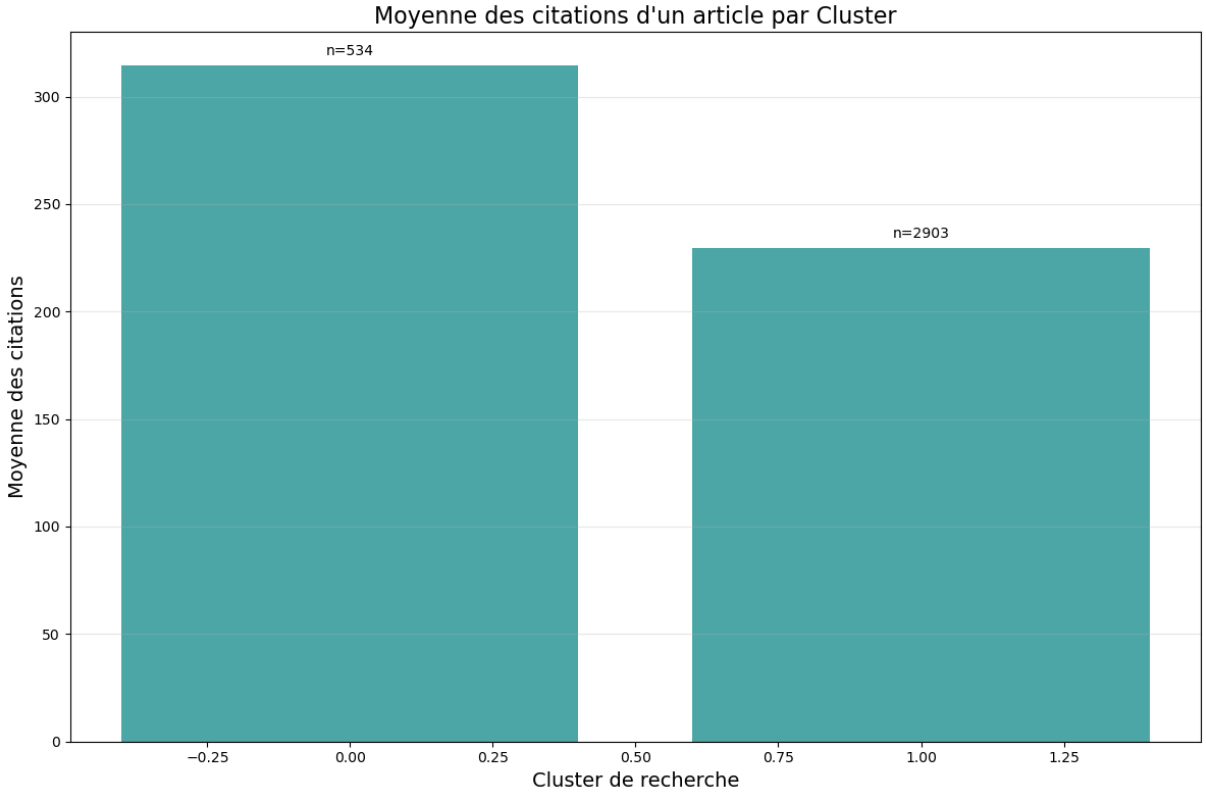

FIGURE 27 - DISTRIBUTION DES CITATIONS PAR CLUSTER

Cette représentation permet de visualiser clairement la différence de moyennes de citations entre les deux groupes identifiés par KMeans, cela suggère que les publications du premier cluster ont en moyenne une reconnaissance plus importantes que ceux du second cluster.

On peut expliquer cette différence par plusieurs facteurs, notamment les thématiques de recherches différentes, l'impact des articles scientifiques, les différentes approches utilisées, et l'ancienneté des publications.

g. LIMITES RENCONTREES

- Certaines métadonnées importantes sont manquantes ou incomplètes notamment les abstracts pour certaines articles OpenAlex, les citations, les institutions, et les pays pour arXiv.
- Le nombre de clusters choisi demeure acceptable mais peut être augmentée pour une séparation plus fine.
- Les modèles LDA sont sensibles au prétraitement, et peuvent produire des résultats redondants ou mal définis.
- Les articles étudiés sont tous écrit en anglais.

VI. CONCLUSION

Dans le cadre de cette étude, nous avons mené une analyse approfondie des tendances en intelligence artificielle générative, couvrant la période 2019-2024, en passant par un pipeline précis : collecte des données, le prétraitement des textes, la vectorisation, la sélection des features, la modélisation thématique, le clustering, l'analyse exploratoire, l'analyse géographique et institutionnelle, l'analyse des auteurs et citations, et enfin l'évaluation des modèles. Cette démarche nous a permis de répondre aux problématiques posées. Nous avons collecté deux jeux de données, 1447 articles d'arXiv et 2000 articles d'OpenAlex via leurs APIs officielles. En appliquant un prétraitement rigoureux sur le dataset fusionné, et en alignant ses colonnes, nous avons obtenu un corpus de 3437 articles uniques. La vectorisation TF-IDF et la sélection des features (Variance Threshold et TruncatedSVD) ont été essentielles pour rendre les textes manipulables par les modèles, aboutissant à une matrice finale de dimensions (3437, 300). Une analyse exploratoire a mis en évidence une croissance exponentielle des publications de 796 articles en 2019 à 1002 en 2024. De plus, nous avons constaté l'émergence de nouveaux concepts notamment generative, chatgpt, llm et diffusion model, ainsi qu'une transition vers des publications interdisciplinaires, particulièrement la médecine, l'environnement, les sciences sociales. En outre, pour extraire les sujets émergents et explorer profondément les tendances de recherche en IA générative, nous avons appliqué deux méthodes complémentaires, LDA qui est une technique statistique et probabiliste, identifiant ainsi 6 topics principaux : applications médicales, systèmes d'IA générative, génération d'image, grands modèles linguistiques, réseaux neuronaux et méthodes d'apprentissage, études sociales autour de Chatgpt. Par ailleurs, BERTopic, qui se distingue par sa prise en compte du contexte et des relations sémantiques entre les mots, a permis d'identifier des concepts non capturés par LDA, notamment des applications dans les sciences des matériaux, l'énergie renouvelable, et l'impact social de l'IA générative. Le clustering KMeans quant à lui, a validé la structure thématique identifiée par LDA et BERTopic en regroupant le corpus en deux clusters, le premier symbolisé par cluster 0, contient des publications spécialisées, souvent orientées aux fondations techniques, et l'autre symbolisé par cluster 1, plus vaste et contient des publications généralistes. En ce qui concerne la répartition géographique, les Etats-Unis dominent avec plus de 800 articles publiés, la Chine suit de près avec plus de 500 articles. En outre, la majorité des institutions contribuants dans les recherches en IA générative sont des grandes universités notamment Stanford University, MIT, Tsinghua University, et des laboratoires de recherches en l'occurrence Google Research, Microsoft Research et National Laboratory. En outre, l'analyse des auteurs a permis d'identifier ceux les plus prolifiant, par exemple Dacheng Tao avec la centralité la plus élevé égale à 0.0364 ce qui indique qu'il est un acteur clé en ce domaine. De même, cette analyse a montré la dominance de certaines communautés comme la communauté 4 qui rassemblent 108 auteurs, montrant ainsi une forte collaboration et présence en matière de recherche scientifique en IA générative. De plus, les articles les plus cités (plus de 9000 citations) sont des catalyseurs dans l'évolution des architectures génératives, en particulier les modèles de diffusion multimodaux et LLMs.

PERSPECTIVES FUTURES

On peut envisager plusieurs pistes d'améliorations :

- Une analyse multilingue.
- Application d'une analyse de sentiments.
- Application de topic modeling guidé comme cBERTopic ou Guided LDA

BIBLIOGRAPHIE

[1] S. S. S. K. ,. S. D. A. S. Shiwangi Singh, «Characterizing generative artificial intelligence applications: Text-mining-enabled technology roadmapping,» [En ligne]. Available: https://www.sciencedirect.com/science/article/pii/S2444569X24000702?utm_source=chatgpt.com#sec0005.

[2] L. L. 1. Y. X. 2. Lu Xian 1*, «Landscape of Generative AI in Global News: Topics, Sentiments, and Spatiotemporal Analysis,» [En ligne]. Available: https://arxiv.org/pdf/2401.08899.

[3] «OpenAlex,» [En ligne]. Available: https://openalex.org/. [Accès le 10 Avril 2025].

[4] «arXiv.org,» [En ligne]. Available: https://arxiv.org/. [Accès le 10 Avril 2025].

[5] «CeRIS-Institut Pasteur,» [En ligne]. Available: https://openscience.pasteur.fr/2024/03/11/openalex/. [Accès le 17 Mai 2025].

[6] «Wikipedia.org,» [En ligne]. Available: https://fr.wikipedia.org/wiki/OpenAlex. [Accès le 17 Mai 2025].

[7] 2. ,. B. K. 1. M. N. 1. Jinsick Kim 1, «Text Mining Approaches for Exploring Research Trends in the, applied sciences,» [En ligne]. Available: https://www.mdpi.com/2076-3417/15/6/3355.

[2]

ANNEXES

ANNEXE 1 - CODE SOURCE

LIEN DU CODE SOURCE SUR GITHUB : HTTPS://GITHUB.COM/YASSF9/TEXT-MINING