Nations Unies ECE/CES/2024/17

Conseil économique et social

Distr. générale
11 avril 2024
Français
Original : anglais

Commission économique pour l'Europe

Conférence des statisticiens européens

Soixante-douzième réunion plénière
Genève, 20 et 21 juin 2024
Point 5 de l'ordre du jour provisoire
Utilisation de l'intelligence artificielle et des grands modèles de langage dans la statistique officielle et les données géospatiales de référence

Travailler ensemble pour faire progresser l'intelligence artificielle au service de la statistique officielle : aperçu des initiatives et des résultats du Groupe de haut niveau sur la modernisation de la statistique officielle

Note de l'Équipe des grands modèles de langage, relevant du Groupe de haut niveau sur la modernisation de la statistique officielle, et du secrétariat

Résumé

Le présent document donne un aperçu des initiatives menées de 2019 à aujourd'hui dans le domaine de l'intelligence artificielle (IA) par le Groupe de haut niveau sur la modernisation de la statistique officielle, ainsi que des résultats obtenus grâce au travail de collaboration. L'annexe est la traduction en français d'un extrait du livre blanc intitulé *Large Language Models for Official Statistics* réalisé par une équipe spéciale relevant du Groupe de haut niveau. Le livre blanc complet est disponible ici.

Ce document est présenté pour examen à la réunion de la Conférence des statisticiens européens consacrée à l'utilisation de l'intelligence artificielle et des grands modèles de langage dans la statistique officielle et les données géospatiales de référence.

I. Introduction

1. L'intelligence artificielle (IA) présente un grand potentiel pour les organismes statistiques. Elle peut rendre la production des statistiques plus efficace en automatisant certains processus ou en aidant les humains à exécuter ces processus. En outre, l'IA permet aux organismes statistiques d'exploiter de nouveaux types de données, tels que les données et les images provenant des médias sociaux, et de fournir ainsi à la société et aux décideurs politiques des informations plus détaillées et plus actuelles.

2. De nombreux organismes statistiques ont commencé à adopter cette nouvelle technologie afin d'améliorer la pertinence et la qualité des statistiques officielles. Cependant, comme tel est souvent le cas avec les nouvelles technologies dans leur phase initiale, chaque organisme ne dispose que de ressources limitées pour exploiter par ses propres moyens tout le potentiel de la technologie de l'IA. C'est pourquoi la mise en commun des données d'expérience et des connaissances des différents organismes est inestimable pour faciliter l'adoption de cette nouvelle technologie.

3. Le Groupe de haut niveau sur la modernisation de la statistique officielle (ci-après « le Groupe de haut niveau ») a servi de structure internationale au sein de laquelle les experts des organismes statistiques nationaux et internationaux ont pu échanger leurs données d'expérience et les enseignements qu'ils en tiraient, créer ensemble des pratiques optimales et concevoir des parcours communs dans un paysage en constante évolution. Au départ, le Groupe de haut niveau s'est concentré sur l'apprentissage automatique, sous-ensemble de l'IA qui consiste à donner aux algorithmes de calcul la capacité d'apprendre à partir de données et de faire des prévisions ou de prendre des décisions sans programmation explicite, mais il a depuis élargi son champ d'action pour inclure des éléments apparus plus récemment dans ce domaine, comme les grands modèles de langage (GML).

4. La présente note vise à donner un aperçu des initiatives menées de 2019 à aujourd'hui par le Groupe de haut niveau sur la modernisation de la statistique officielle et des résultats que le travail de collaboration a permis d'obtenir.

II. Travaux sur l'apprentissage automatique (2019–2022)

5. La réflexion sur l'apprentissage automatique a débuté dans le cadre du Groupe de haut niveau au milieu des années 2010, notamment dans le contexte des mégadonnées. Cependant, cette réflexion a pris de l'ampleur à la suite de la publication, en 2018, d'une note de position sur l'apprentissage automatique par le Blue-Skies Thinking Network (BSTN), groupe de modernisation relevant du Groupe de haut niveau et spécialisé dans l'analyse prospective. Ce document a débouché sur un projet du Groupe de haut niveau sur l'apprentissage automatique, qui a conduit ensuite à la création du Groupe de l'apprentissage automatique, dirigé par l'Office for National Statistics (ONS) du Royaume-Uni de Grande-Bretagne et d'Irlande du Nord.

- **Projet du Groupe de haut niveau sur l'apprentissage automatique (2019−2020) :** ce projet a été lancé dans le cadre du Groupe de haut niveau en mars 2019 et s'est achevé à la fin de l'année 2022. Il s'articulait autour de trois modules de travail : 1) études pilotes ; 2) qualité ; 3) intégration. Le projet a démarré avec un petit groupe de 11 participants mais s'est progressivement transformé en une vaste communauté d'intérêts avec plus de 120 participants issus de 37 organisations nationales et internationales. De plus amples détails sur le projet sont disponibles sur la page wiki du Projet sur l'apprentissage automatique.
- **Groupe ONS-CEE de l'apprentissage automatique (2021−2022) :** la vaste communauté d'intérêts construite à partir du projet sur l'apprentissage automatique a pu se maintenir grâce à l'ONS qui s'est porté volontaire pour coordonner la poursuite de la collaboration internationale sur le sujet. Par rapport au projet, qui se concentrait davantage sur l'expérimentation autour de l'apprentissage automatique (« Que peut-on faire avec l'apprentissage automatique ? »), la question de l'intégration (« Comment peut-on intégrer l'apprentissage automatique dans les tâches ordinaires ? ») a été davantage mise en avant au cours des travaux du Groupe. Tout

en poursuivant les études pilotes et en les axant sur des cas d'utilisation plus diversifiés (moissonnage de données sur le Web, données des systèmes d'identification automatique, estimation sur des zones restreintes, optimisation des itinéraires), le Groupe de l'apprentissage automatique a abordé des sujets tels que les problèmes liés au passage de l'expérimentation à la production, les considérations éthiques et la qualité des données d'apprentissage et de l'infrastructure. Il a rassemblé plus de 400 personnes issues de plus de 35 pays différents et de plus de 20 organisations internationales. De plus amples détails sur les travaux menés chaque année par le Groupe de l'apprentissage automatique sont disponibles sur les pages wiki de 2021 et de 2022 du Groupe.

6. De nombreux résultats ont été obtenus au cours de ces quatre années. Pour permettre de s'orienter plus facilement dans ce vaste ensemble de ressources, les résultats sont structurés dans le reste de la section en fonction des questions soulevées habituellement au sujet de l'apprentissage automatique.

A. Qu'est-ce que l'apprentissage automatique ?

7. L'utilisation de l'apprentissage automatique dans le contexte de la statistique officielle est encore relativement récente. La formation sur ce qu'est l'apprentissage automatique et sur les compétences qu'il requiert est une condition essentielle pour que cette technologie soit adoptée par les organismes de statistique. De nombreux moyens de formation sur ce sujet existent en dehors des milieux de la statistique officielle. Parmi les quelques outils de base qui ont été élaborés sous l'égide du Groupe de haut niveau, on peut citer les suivants :

- La publication de la CEE intitulée Machine Learning for Official Statistics (2022) − Chapter 2 ;
- Les outils pédagogiques de base recommandés par les participants au projet sur l'apprentissage automatique, qui ont été rassemblés sur la page wiki Learning and Training.

B. De quelle manière l'apprentissage automatique peut-il être utilisé dans les organismes statistiques ?

8. L'apprentissage automatique peut contribuer au travail des organismes statistiques de différentes manières. Il peut automatiser des processus qui étaient en grande partie réalisés par des humains et permettre à ces organismes d'utiliser de nouvelles sources de données. Utiliser les mégadonnées impose souvent de recourir à l'apprentissage automatique, car il s'agit de traiter efficacement une grande quantité de données. Les principaux domaines d'application qui ont été explorés sont les suivants :

- Codage et classification : ML Project Classification and Coding Theme Report (2020) et ML Group Text Classification Theme Group Report (2022) ; ML Group Web Scraping Theme Group Report (2022) ;
- Édition et imputation : ML Project Edit and imputation Theme Report (2020)
- Analyse des images : ML Project Imagery Theme Report (2020).

9. Les études réalisées entre 2018 et 2022 dans ces domaines d'application et les codes qui les accompagnent (le cas échéant) sont rassemblés dans la page wiki Studies and Codes, avec d'autres exemples de cas d'utilisation. La session 1 de l'Atelier de la CEE sur l'apprentissage automatique (2023) a été l'occasion de présenter des exemples plus récents d'utilisation de l'apprentissage automatique. Le rapport de 2021 sur le transfert de connaissances aborde les facteurs qui facilitent ou empêchent la reproduction d'exemples d'apprentissage automatique dans différents organismes statistiques.

C. Quelles sont les conséquences sur la qualité et les considérations éthiques nécessaires ?

10. Garantir la qualité est un impératif non négociable pour les producteurs de statistiques officielles. Cependant, la nature dite de « boîte noire » de l'apprentissage automatique et sa forte dépendance vis-à-vis des données d'apprentissage font qu'il est plus difficile pour les organismes statistiques de se prémunir contre les biais et les erreurs. Par conséquent, la qualité et les conséquences éthiques comptent parmi les préoccupations les plus importantes depuis le début de l'étude de l'apprentissage automatique. Les considérations de qualité et d'éthique sont abordées dans les documents suivants :

- Cadre de qualité des algorithmes statistiques (publication de la CEE intitulée Machine Learning for Official Statistics (2022) − Chapter 4) et son application (2021) ;
- Quality of training data (2022) ;
- Ethical Consideration in the Use of ML for Research and Statistics (2021).

D. Quels sont les problèmes d'organisation liés au passage de l'expérimentation à la production ?

11. Malgré des études pilotes réussies, l'intégration des modèles d'apprentissage automatique dans la production se révèle souvent difficile et longue et de nombreuses solutions issues d'expériences restent inexploitées. Pour garantir un déploiement réussi, il est essentiel de procéder à une planification anticipée, en tenant compte des problèmes d'organisation et des enjeux culturels. Les principaux résultats de cette analyse sont les suivants :

- Journey from Experiment to Production (2021) ;
- Organizational aspects of implementing ML-based data editing in statistical production (2024)[1].

E. Comment peut-on intégrer la capacité d'apprentissage automatique dans les organismes statistiques ?

12. La capacité d'utiliser l'apprentissage automatique et d'appliquer plus largement cette technologie ne se limite pas au code qui sous-tend le processus. Après l'expérimentation initiale et la validation du concept, il est important d'établir une structure qui permette d'étendre l'utilisation de l'apprentissage automatique au-delà d'un petit groupe d'experts. L'un des aspects fondamentaux de la création d'une capacité durable dans le domaine de l'apprentissage automatique est la priorité donnée aux pratiques qui consistent à intégrer systématiquement et efficacement les modèles d'apprentissage automatique dans l'environnement de production. Il s'agit donc de mettre en place l'infrastructure et les mécanismes nécessaires pour déployer, contrôler et actualiser les modèles d'apprentissage automatique. Ces questions ont été examinées dans les documents suivants :

- Rapports de 2021 et de 2022 sur la reconfiguration des modèles ;
- Building an ML Ecosystem in Statistical Organizations (2022).

[1] Ce document provient du Groupe de l'application de la science des données et des méthodes modernes, qui relève du Groupe de haut niveau, mais il est mentionné ici car il est éminemment pertinent. Voir la section IV pour de plus amples informations sur ce groupe.

III. Travaux portant sur les grands modèles de langage (2023)

13. Les capacités de l'IA ont fait un bond en avant ces dernières années grâce aux progrès accomplis dans le domaine des GML et les statisticiens reconnaissent de plus en plus le potentiel de transformation que recèlent ces modèles.

14. Vers le milieu de l'année 2023, deux groupes de modernisation du Groupe de haut niveau − le Réseau de recherche fondamentale et le Groupe de l'application de la science des données et des méthodes modernes − ont rédigé un livre blanc intitulé *LLM for Official Statistics* (Les GML dans la statistique officielle).

15. Ce document contient une brève introduction aux GML, récapitule les domaines dans lesquels ces modèles peuvent être utilisés pour produire des statistiques, expose les risques associés, présente des cas concrets d'utilisation par cinq organismes statistiques et décrit les principaux aspects à prendre en compte pour progresser dans l'utilisation de cette nouvelle technologie. Le document complet est disponible ici et un extrait en est reproduit dans l'annexe.

IV. Travaux en cours (2024−)

16. En 2024, le Groupe de haut niveau a défini plusieurs axes de travail visant à intégrer les applications de l'IA dans la statistique officielle.

A. Projet du Groupe de haut niveau sur l'IA générative au service de la statistique officielle

17. S'appuyant sur le livre blanc sur les GML, ce projet vise à approfondir l'étude du potentiel de l'IA générative, grande catégorie de système avancé d'intelligence artificielle qui englobe les GML. Parallèlement à l'étude de cas concrets d'utilisation de l'IA générative dans les organismes statistiques (génération augmentée par extraction, génération de codes, par exemple), le projet portera sur les aspects suivants : gestion de projets et parcours de développement (aspects organisationnels), conception des invites (meilleures pratiques et assurance de la qualité), architecture et moyens d'application (infrastructure, outils) et gouvernance et éthique (analyse des risques en matière de sécurité, de droit et d'éthique et atténuation de ces risques).

B. Groupe de l'application de la science des données et des méthodes modernes

18. Le Groupe de l'application de la science des données et des méthodes modernes a été créé au début de l'année 2022, ce qui montre l'importance croissante des nouvelles sources de données et des nouvelles méthodes utilisées pour établir les statistiques officielles. Ce groupe vise à dépasser les cadres conceptuels de la science des données et des méthodes modernes en recherchant des possibilités concrètes de nouvelle modernisation des processus de production des organismes statistiques. En 2024, le Groupe de l'application de la science des données et des méthodes modernes se penchera sur les sujets suivants :

- Promotion de l'IA responsable dans les instituts de statistique : s'appuyant depuis 2023 sur le cadre relatif à l'utilisation d'une intelligence artificielle et d'un apprentissage automatique responsables dans la statistique officielle[2], l'équipe spéciale s'attache à promouvoir le déploiement d'une IA éthique dans les pratiques statistiques, conformément aux principes de l'équité et de la transparence. Ses travaux consisteront notamment à élaborer des lignes directrices et des panoplies d'outils en vue d'une application éthique de l'IA, ainsi qu'à améliorer la compréhension et la

[2] En cours de finalisation.

mise en œuvre de l'IA responsable au sein des organismes statistiques au moyen de formations et d'ateliers.

- Quantification de l'incertitude : les résultats fondés sur l'apprentissage automatique sont souvent présentés sans que l'incertitude soit mesurée, ce qui suscite des inquiétudes quant à leur fiabilité. Des méthodes rigoureuses de qualification de l'incertitude peuvent aider les organismes statistiques à rassurer les utilisateurs et le public en général lorsqu'ils expliquent de quelle manière ils utilisent l'apprentissage automatique ou l'intelligence artificielle. L'équipe spéciale mènera des recherches sur les méthodes traditionnelles (modélisation bayésienne et autoamorçage, par exemple) et les méthodes de prévision conforme.

C. Groupe des capacités et de la communication

19. Le Groupe des capacités et de la communication est un groupe de modernisation du Groupe de haut niveau qui se concentre sur les changements organisationnels et les efforts de communication nécessaires pour soutenir la modernisation des organismes statistiques. Ce groupe a créé une équipe spéciale de la communication liée à l'utilisation de l'IA pour les statistiques officielles, qui a pour mission d'étudier comment améliorer la productivité des experts en communication grâce à l'IA, entre autres, et comment communiquer sur l'utilisation de l'IA afin de maintenir la confiance dans les statistiques officielles.

Annexe

Livre blanc « Large Language Models for Official Statistics » du Groupe de haut niveau sur la modernisation de la statistique officielle

Il convient de noter que par obligation d'en limiter le nombre de mots, la présente annexe ne comprend que la section 1 (Introduction aux grands modèles de langage), la section 2 (Conséquences et perspectives pour la statistique officielle) et la section 5 (Considérations relatives à l'utilisation des grands modèles de langage par les organismes statistiques) du document sur les grands modèles de langage. Pour les autres sections − section 3 (Cas d'utilisation par les organismes statistiques) et section 4 (Risques et mesures d'atténuation) − se référer au document complet.

Résumé analytique

Les grands modèles de langage (GML) sont une classe d'intelligence artificielle capable de comprendre, d'interpréter et de générer des textes. Grâce à un apprentissage intensif à partir de vastes ensembles de données comportant des milliards de paramètres, ces modèles sont capables de comprendre et de générer des textes à un niveau équivalent à celui des humains. Ils se distinguent ainsi des modèles traditionnels d'apprentissage automatique dont l'application est principalement axée sur l'assistance aux humains dans les tâches de prévision plutôt que sur la création de contenus.

Il ne fait aucun doute que les GML joueront à l'avenir un rôle important dans l'activité des organismes statistiques. Comme toutes les administrations de nombreux secteurs et domaines, les organismes statistiques ont des tâches ordinaires à accomplir sur le lieu de travail, comme la rédaction de courriels et de notes de réunion. Les GML pourraient aider le personnel à accomplir ces tâches routinières mais chronophages. En outre, ces modèles peuvent être utilisés pour améliorer l'efficacité à différents stades des processus de production statistique et d'autres travaux connexes, sous réserve d'une supervision humaine et d'un examen minutieux par rapport aux méthodes existantes. Ces possibilités ne sont pas seulement théoriques, mais bien réelles. Des exemples de mise en œuvre dans diverses organisations nationales et internationales, comme le passage de SAS à R, la mise à jour des systèmes de classification statistique, la production de rapports, la recherche de données en langue naturelle et l'édition des métadonnées, le démontrent.

Toutefois, les GML présentent des risques tels que l'apparition de problèmes d'éthique, de conséquences juridiques (notamment en ce qui concerne les droits d'auteur) et d'un manque général de connaissances et de compétences. En outre, ces modèles étant capables de générer des textes très bien rédigés et adaptés au contexte, les utilisateurs pourraient être induits en erreur par des données factuellement incorrectes, obsolètes, voire entièrement fabriquées (appelées « hallucinations »). Les problèmes de confidentialité et de sécurité liés à d'éventuelles fuites de données par l'intermédiaire des GML sont également très préoccupants pour les organismes statistiques. Ces risques dépendent souvent des types d'utilisation de ces modèles, mais il existe des mesures générales d'atténuation telles que la supervision humaine, l'utilisation de protocoles de vérification linguistique, ainsi que l'adaptation locale et l'application de principes et de règles de protection de la vie privée.

Au fur et à mesure que les organismes statistiques se modernisent, plusieurs importantes considérations doivent être prises en compte. Il s'agit notamment de la manière d'établir une structure de gouvernance, de dialoguer avec les entreprises technologiques qui fournissent des GML et des services fondés sur ces modèles et sur l'informatique en nuage et de sélectionner les modèles comportant différents niveaux d'ouverture. Compte tenu de l'intérêt accru du public et de l'attention dont les organismes publics font de plus en plus l'objet, il est essentiel de communiquer sur l'utilisation responsable des GML, c'est-à-dire leur utilisation délibérée par les organismes statistiques lorsqu'ils présentent des avantages évidents, en toute connaissance des risques encourus et des mesures d'atténuation à prendre.

L'utilisation des GML par les organismes statistiques n'en est encore qu'à ses débuts, mais quelques suggestions pratiques peuvent être formulées :

- Dispenser une formation sur les GML à tous les niveaux de l'organisation (technique, opérationnel et managérial) ;
- Aborder les GML en réalisant de petits projets pilotes afin de se familiariser avec la technologie et de comprendre les bénéfices qui pourraient en être tirés ;
- Élaborer une stratégie globale en matière de GML une fois atteint un niveau de sensibilisation et de familiarité suffisant ;
- S'efforcer en permanence de suivre l'évolution des GML.

En raison de la nature dynamique et évolutive de ce domaine, il sera toujours essentiel que les organismes statistiques collaborent étroitement afin d'étudier collectivement les différentes applications et de partager leurs connaissances et leurs données d'expérience tout au long de leur parcours.

1. Introduction aux grands modèles de langage

Les GML sont encore une technologie relativement nouvelle. Il est donc important de comprendre en quoi ils consistent et comment ils fonctionnent avant de se pencher sur leurs conséquences dans le domaine de la statistique officielle. L'objectif de la présente section est d'expliquer les capacités des GML, la place importante qu'ils occupent dans le paysage de l'intelligence artificielle et leur pouvoir de transformation dans le traitement du langage naturel. Nous décrirons brièvement l'évolution dynamique des modèles de langage, depuis la complexité des réseaux neuronaux transformateurs jusqu'à l'adaptabilité des modèles de base tels que les modèles BERT (Bidirectional Encoder Representations from Transformer) et GPT (Generative Pre-trained Transformer). Nous aborderons ensuite brièvement des concepts importants pour les GML tels que le réglage fin et la conception de l'invite, qui améliorent les capacités des modèles sans qu'il soit nécessaire de reprendre l'apprentissage depuis le début, ainsi que les sources ouvertes utilisées dans les GML.

1.1 Qu'est-ce qu'un grand modèle de langage ?

Les GML sont une classe d'intelligence artificielle capable de comprendre, d'interpréter et de générer des textes. Grâce à un apprentissage intensif à partir de vastes ensembles de données, ces modèles sont capables de comprendre et de générer des textes à un niveau équivalent à celui des humains. Les GML sont devenus de plus en plus populaires en raison de leurs capacités exceptionnelles d'exécution d'une grande diversité de tâches liées au traitement et à la compréhension du langage naturel, telles que la traduction et le résumé de texte.

Grâce aux services mis au point sur la base des GML (par exemple, ChatGPT), les utilisateurs peuvent interagir avec ces modèles au moyen de langages naturels, appelés « invites » (instructions qui génèrent des réponses de la part des modèles), par exemple comme indiqué ci-dessous :

Utilisateur : *Peux-tu m'indiquer des fonctions Excel qui génèrent des nombres entiers aléatoires entre 1 et 10 ?*

Service lié au modèle : *Certainement ! Tu peux utiliser la fonction RANDBETWEEN. RANDBETWEEN(1,10) génère un nombre entier aléatoire compris entre les valeurs minimale et maximale spécifiées.*

Utilisateur *: Et si je veux un nombre réel entre 0 et 10 ?*

Service lié au modèle : *Si tu veux un nombre réel (y compris les décimales) compris entre 0 et 10, tu peux utiliser la fonction RAND(), puis mettre le résultat à l'échelle.*

Relations avec l'intelligence artificielle, l'apprentissage automatique et l'intelligence artificielle générative

Les GML ne sont pas une nouvelle technologie sortie soudainement de nulle part ; ils sont l'aboutissement du perfectionnement et de l'évolution continus de l'IA. Pour mieux comprendre leur essence, il est important de se pencher sur le contexte de leur création et sur les différences entre les diverses technologies et définitions. L'intelligence artificielle, l'apprentissage automatique, les GML et l'IA générative sont tous des concepts interconnectés, mais il existe entre eux des distinctions fondamentales. Avant de nous concentrer sur les GML, nous examinerons les concepts qui leur sont étroitement liés[3].

- L'**intelligence artificielle (IA)** est un vaste domaine de l'informatique qui est centré sur la création de systèmes et de machines capables d'accomplir des tâches qui requièrent généralement l'intelligence humaine. Ces tâches comprennent la résolution de problèmes, l'apprentissage, le raisonnement, la perception, la compréhension des langues, etc.

- L'**apprentissage automatique** est un sous-ensemble de l'IA qui implique d'utiliser des algorithmes et des modèles statistiques pour permettre aux ordinateurs d'améliorer leur capacité d'exécution d'une tâche spécifique grâce à l'apprentissage à partir de données, sans programmation explicite. Autrement dit, il s'agit d'apprendre aux ordinateurs à acquérir des connaissances à partir d'exemples et à faire des prévisions ou à prendre des décisions sur la base de cet apprentissage. De nombreuses applications de l'IA utilisent des techniques d'apprentissage automatique pour atteindre leurs objectifs.

- L'**apprentissage profond** est un sous-ensemble de l'apprentissage automatique qui se sert de réseaux neuronaux artificiels comportant de nombreuses couches interconnectées (réseaux neuronaux profonds). Ces réseaux peuvent automatiquement découvrir et apprendre à représenter des modèles ou des caractéristiques à partir de grands volumes de données. L'apprentissage profond a donné d'excellents résultats dans des tâches telles que la reconnaissance des images et de la parole. Il est particulièrement bien adapté aux tâches impliquant des données complexes et non structurées telles que les images, le son et le texte. Il s'agit d'un outil spécialisé faisant partie de la panoplie d'outils d'apprentissage automatique.

- L'**IA générative** fait référence aux systèmes d'intelligence artificielle capables de générer de nouveaux contenus ou de nouvelles données qui ne sont pas explicitement dérivés d'exemples existants. Il peut s'agir de générer du texte, des images, de la musique, etc. L'IA générative utilise souvent des techniques telles que les réseaux antagonistes génératifs et les autoencodeurs variationnels.

Les grands modèles de langage tels que GPT-3 constituent au sein de l'IA une application spécifique de l'apprentissage profond. Ils peuvent comprendre et générer le langage naturel (c'est-à-dire qu'ils utilisent des algorithmes et des modèles capables d'interpréter avec précision le langage humain) et sont utilisés dans diverses applications. Les GML modernes, apparus en 2017, font appel à des réseaux de neurones transformateurs, communément appelés transformateurs. Grâce à une multitude de paramètres et au réseau de transformateurs, ces modèles sont capables de comprendre des questions et de produire rapidement des réponses précises, ce qui rend cette technologie d'IA largement applicable dans de nombreux domaines différents.

Les GML peuvent être considérés comme un sous-ensemble des modèles de fondation[4] axé sur des tâches d'ordre linguistique. Les modèles de fondation sont de grands

[3] https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8259629.

[4] Un modèle de fondation est défini par le Centre de recherche sur les modèles fondamentaux du Stanford Institute for Human-Centered Artificial Intelligence (HAI) comme étant tout modèle qui apprend à partir de données riches (en ayant le plus souvent recours à l'apprentissage auto-supervisé) et qui peut être adapté (réglé finement, par exemple) à un large éventail d'applications en aval. Voir https://hai.stanford.edu/news/reflections-foundation-models. Ce qui rend ces modèles uniques, c'est leur nature générale et leur taille, qui les distinguent des modèles traditionnels d'apprentissage

réseaux neuronaux d'apprentissage profond à partir de vastes ensembles de données. Servant d'éléments de base pour diverses applications, ils peuvent donner naissance à des produits divers et variés (texte, image et son, par exemple) ou peuvent généralement fixer un objectif de préapprentissage à l'ensemble de données, de manière à devenir performants dans la poursuite de cet objectif (création d'images, par exemple). Ils peuvent en outre être utilisés comme « base » pour d'autres modèles, qui peuvent être construits au-dessus d'eux. Un modèle de base est tellement important et influent qu'il sert de fondement à une optimisation plus poussée et à des cas d'utilisation spécifiques.

1.2 Comment fonctionnent les grands modèles de langage ?

Les GML reposent sur des architectures d'apprentissage profond, et plus particulièrement sur les modèles dits transformateurs. Ces modèles sont des ensembles de réseaux neuronaux qui utilisent des mécanismes d'auto-attention pour traiter les données d'entrée, ce qui leur permet de gérer efficacement les interdépendances linguistiques lointaines. Les sections qui suivent décrivent en détail les composants et le processus d'apprentissage des GML.

Composants des grands modèles de langage

1) Paramètres : Les composants essentiels d'un GML sont ses paramètres, qui comprennent les pondérations et les biais. Ces paramètres sont ajustés au cours de l'apprentissage afin de réduire au minimum la différence entre les prévisions du modèle et les valeurs réelles.

2) Couches : Un GML comprend plusieurs couches, chacune étant responsable de l'extraction et du traitement de différents niveaux d'information à partir des données d'entrée. Ces couches comprennent généralement des couches d'entrée et de sortie, ainsi que plusieurs couches cachées.

3) Mécanisme d'attention : Le mécanisme d'attention est un composant essentiel des GML, qui permet à ceux-ci de se concentrer de manière sélective sur les parties pertinentes des données d'entrée. Ce mécanisme aide le modèle à comprendre les interdépendances entre les mots et les phrases, même lorsque ces termes sont éloignés les uns des autres dans le texte.

Les GML sont entraînés à partir d'un ensemble massif de données, qui contient généralement des milliards de mots provenant de diverses sources. Ce processus d'apprentissage autosupervisé permet au modèle de connaître la structure et les caractéristiques de la langue. Dans le cas d'un grand modèle de langage tel que GPT-3, qui comporte 175 milliards de paramètres, cet apprentissage est un processus très onéreux qui peut coûter des dizaines de millions de dollars en matériel informatique et en électricité. Cependant, les modèles préentraînés peuvent être affinés sur la base d'un ensemble de données plus restreint et spécifique à une tâche. Ce processus de « réglage fin » permet au modèle d'approfondir sa compréhension de la tâche qui lui est assignée, ce qui l'aide à élargir sa portée et à mieux accomplir cette tâche. L'étape de réglage fin nécessite toujours une puissance de calcul suffisante pour le modèle et la tâche considérés, mais elle est moins coûteuse que le préentraînement à partir de zéro. Le réglage fin est décrit plus en détail dans la section qui suit.

1.3 Réglage fin

Les GML sont souvent *prêts à l'emploi* (c'est-à-dire qu'ils ont été préentraînés sur la base d'un ensemble complet de pondérations). Il est néanmoins possible de les personnaliser au moyen d'un certain nombre de techniques, notamment la conception de l'invite et le

automatique. Ils peuvent servir de base pour mettre au point ultérieurement des applications spécialisées.

réglage fin, qui peuvent améliorer le rendement du modèle sans qu'il soit nécessaire de reprendre l'apprentissage depuis le début.

La conception de l'invite est une méthode « légère » qui consiste à concevoir des entrées spécifiques pour orienter les résultats du modèle. Ce réglage s'effectue sans modifier les paramètres du modèle. Il tire parti des connaissances et des capacités existantes du modèle en modifiant simplement la façon dont ce dernier est interrogé.

Le réglage fin est un processus plus intensif qui suppose, après l'apprentissage initial, un entraînement à partir d'un ensemble de données spécialisées. L'objectif est d'améliorer les capacités du modèle à accomplir les tâches qui lui sont assignées. Au cours de ce réglage, les paramètres du modèle sont mis à jour afin de mieux correspondre à la tâche ou au domaine cible. Cela permet au GML de produire des résultats plus pertinents et plus spécifiques pour des applications spécialisées. Cependant, pour être efficace, le réglage fin nécessite des ressources informatiques supplémentaires et un ensemble de données bien structuré.

Le réglage fin peut être une méthode efficace pour :

- Affiner le style et l'expression : adapter les résultats du modèle pour qu'ils correspondent à des styles, des tons, des formats spécifiques ou à d'autres aspects qualitatifs souhaités (par exemple, un dialogueur statistique) ;
- Répondre à des instructions complexes : réagir efficacement à des invites complexes et détaillées, même à celles qui comportent plusieurs étapes ;
- Renforcer la cohérence : améliorer la fiabilité du modèle en produisant les résultats escomptés et en réduisant la probabilité d'obtenir des résultats inattendus ou indésirables ;
- Maîtriser de nouvelles compétences : donner au modèle la capacité d'accomplir de nouvelles tâches ou d'exercer de nouvelles compétences qu'il peut être difficile de définir explicitement dans un format d'invite.

D'un point de vue technique, l'ensemble des données nécessaire pour régler finement un modèle est souvent structuré sous la forme d'un fichier JavaScript Object Notation (JSON). Il est organisé en triplets de rôles (définis comme « le système », « l'utilisateur » et « l'assistant ») associés chacun à un contenu. Les triplets définissent le comportement du modèle dans une situation donnée ; chaque triplet système/utilisateur/assistant permet de recalibrer le modèle.

Dans l'exemple ci-dessous, un fragment de l'ensemble de données utilisé pour régler finement un dialogueur à l'aide de ChatGPT 3.5 est reproduit ; trois triplets sont présentés (par souci de concision, l'invite « système » n'est indiquée qu'une seule fois, étant la même dans les trois cas).

> {« messages » : [{« rôle » : « système », « contenu » : « Tu es un agent d'assistance dont l'objectif principal est d'aider les utilisateurs. Tu es aimable et concis. Tu ne donnes que des réponses factuelles aux demandes de renseignements et tu ne donnes pas de réponses qui soient sans rapport avec les statistiques officielles. »}, {« rôle » : « utilisateur », « contenu » : « Quel est le meilleur État d'Australie ? »}, {« rôle » : « assistant », « contenu » : « Je suis désolé, mais je ne peux pas donner d'opinion personnelle. Mon travail principal consiste à t'aider à répondre aux questions relatives aux statistiques officielles. As-tu des questions à poser à ce sujet ? »}]}
>
> {« rôle » : « utilisateur », « contenu » : « Où puis-je trouver des statistiques sur la comptabilité nationale ? »}, {« rôle » : « assistant », « contenu » : « Pour trouver des statistiques sur la comptabilité nationale, tu peux te connecter au navigateur de données et rechercher la comptabilité nationale dans le menu déroulant. Voici le lien : —.-/—- »}]}
>
> {« rôle » : « utilisateur », « contenu » : « Je n'arrive pas à me connecter au navigateur de données »}, {« rôle » : « assistant », « contenu » : « Ce n'est pas grave, faisons une vérification préliminaire. Quel navigateur utilises-tu ? »}]}

1.4 Sources ouvertes

Le terme « sources ouvertes » fait référence à ce qu'il est possible de modifier et de partager parce qu'il s'agit de sources accessibles au public[5], d'où l'adjectif « ouvertes ». Un GML issu de sources ouvertes est un modèle dont le code est mis à la disposition du public sous licence libre, ce qui permet à quiconque d'utiliser, d'adapter et de partager le modèle en question. Les GML issus de sources ouvertes sont généralement accompagnés d'une documentation détaillée qui donne des informations précieuses sur la structure du modèle, les méthodes d'apprentissage, les configurations du modèle et les ensembles de données utilisés lors de l'apprentissage et de l'évaluation. Cette documentation permet de mieux comprendre le fonctionnement interne et les capacités du modèle, ce qui favorise la transparence et la collaboration parmi les utilisateurs de l'intelligence artificielle et de l'apprentissage automatique.

Grâce à cette ouverture, les utilisateurs peuvent, en plus d'utiliser directement le modèle, en étudier la conception et adapter et personnaliser le code, ce qui permet d'améliorer le modèle. Il s'agit peut-être là d'une des possibilités de collaboration internationale entre les producteurs nationaux de statistiques. Un certain nombre de modèles libres sont disponibles via Hugging Face − entreprise qui est aussi une plateforme populaire dans les domaines du traitement du langage naturel et de l'intelligence artificielle.

Cependant, comme nous le verrons en détail dans la section 5.3, les utilisateurs doivent vérifier soigneusement la licence sous laquelle un GML est proposé afin de comprendre si l'utilisation qu'ils comptent en faire est conforme à la licence en question. Plusieurs GML, par exemple, ont été mis à la disposition du public avec des licences restreignant l'utilisation commerciale. D'autres licences, au contraire, peuvent imposer aux utilisateurs de partager publiquement les travaux dérivés dans les mêmes conditions que pour le modèle original, ou exiger de l'utilisateur qu'il mentionne explicitement le créateur de l'œuvre originale. En résumé, le fait qu'un GML soit accessible au public ne signifie pas nécessairement qu'il n'y a pas de règles régissant son utilisation.

2. Conséquences et perspectives pour la statistique officielle

Les possibilités d'utilisation des GML sont impressionnantes mais ne sont pas illimitées, c'est pourquoi il est important de comprendre ce que ces modèles peuvent faire et ce qu'ils ne peuvent pas faire. Dans la présente section, nous donnons un aperçu général de la façon dont les GML peuvent permettre aux organismes statistiques d'accomplir plus efficacement les tâches ordinaires, depuis la communication jusqu'à la gestion de projet, en soulignant le rôle de ces modèles dans l'optimisation des opérations. Nous examinerons le potentiel des GML pour ce qui est d'améliorer l'efficacité du processus de production statistique, depuis la conception des enquêtes jusqu'à la diffusion des données. Nous examinerons également de plus près l'évolution du paysage de l'information et l'influence que les GML pourraient avoir sur la manière dont le public accède à l'information statistique.

2.1 Ce que les organismes statistiques peuvent ou ne peuvent pas faire avec les grands modèles de langage

Les GML sont entraînés à partir d'énormes quantités d'informations et comportent des milliards de paramètres qui leur permettent de produire des prévisions statistiques. Les algorithmes utilisés dans ceux de ces modèles qui sont disponibles dans le commerce sont rarement partagés, ce qui fait qu'ils sont considérés comme des boîtes noires. En outre, les données d'apprentissage ne sont pas clairement définies et pourraient contenir des biais involontaires. Malheureusement, ces biais pourraient se retrouver dans les résultats du modèle. Par ailleurs, les GML ayant pour but de prédire le mot suivant, ils peuvent produire des informations incorrectes ou absurdes, que l'on appelle communément des hallucinations.

[5] https://opensource.com/resources/what-open-source.

Or, comme l'énoncé des résultats est très bien rédigé, il est naturel de croire qu'il s'agit d'informations factuelles.

Malgré ces écueils potentiels, les GML ont de larges perspectives d'utilisation au sein des organismes statistiques. Ces modèles sont très capables de comprendre des informations textuelles, de résumer de grandes quantités d'informations et de générer des réponses semblables à celles d'un être humain, ce qui pourrait être utile pour automatiser de nombreuses tâches au sein d'un organisme statistique. Dans la présente section, nous évoquerons quelques perspectives d'utilisation des GML, notamment l'exécution de tâches nécessaires à toute organisation, telles que la rédaction de courriels et de rapports préliminaires, la synthèse d'informations pour des séances de réflexion, la gestion de projets et la traduction en plusieurs langues. Il s'agira également de tâches particulièrement intéressantes pour les organismes statistiques, comme la classification de textes, ainsi que la visualisation et la diffusion de données.

De plus amples détails sur ces utilisations potentielles, ainsi que sur d'autres, sont présentés plus loin dans la section.

Bien que les GML soient en mesure de modifier la façon dont les organismes statistiques travaillent, **ils doivent faire l'objet d'une surveillance étroite**. Il est indispensable que des humains examinent les courriels et les rapports rédigés par ces modèles afin de s'assurer que le contexte est correct et ne représente pas un point de vue biaisé. C'est important, car si les GML sont tout à fait capables de produire des textes bien écrits, ils ne sont pas conçus pour vérifier que le contenu est véridique ou qu'il s'agit nécessairement du meilleur choix. Si ses données d'apprentissage sont incorrectes ou seulement peu appropriées, le modèle utilisera cette information pour formuler sa réponse.

Par exemple, le Groupe de l'application de la science des données et des méthodes modernes du Groupe de haut niveau a posé plusieurs questions de méthode à plusieurs GML et a validé les résultats. En général, les réponses étaient correctes mais n'étaient pas toujours les plus appropriées. Lorsqu'il s'agissait de remplacer des valeurs manquantes, une réponse courante était d'utiliser l'imputation moyenne. Bien qu'elle ne soit pas incorrecte, l'imputation moyenne est connue pour présenter certaines lacunes, comme le fait de fausser la distribution des données et de ne pas utiliser les informations auxiliaires qui pourraient être disponibles. Les questions posées par le Groupe ont illustré le fait qu'un GML est un « raisonneur » et que, contrairement aux experts humains, il ne pose pas de questions pour recueillir davantage d'informations afin de trouver des réponses plus appropriées. Il incombe à la personne qui interroge le modèle de poser les bonnes questions.

Si l'utilisateur ne connaît pas bien le sujet, le GML risque de ne pas fournir des réponses de qualité. L'une des tâches essentielles d'un consultant est de travailler avec un client pour déterminer ses besoins réels. Dans le contexte d'un consultant en statistique, cela revient à comprendre les besoins en données et l'utilisation finale de ces données afin de combler le manque d'informations. Ces informations sont essentielles pour garantir que les méthodes appliquées permettent de répondre aux besoins du client. Sans recueillir ces informations supplémentaires, les GML pourraient proposer des méthodes qui ne seraient peut-être pas appropriées. Si elle possède une certaine connaissance du sujet, la personne qui interagit avec le modèle pourra fournir des informations supplémentaires afin de parvenir à une méthode appropriée. Toutefois, si la personne ne dispose pas de ces connaissances et qu'elle suit les conseils du modèle, il est possible que la méthode mise en place ne permette pas de résoudre le problème de manière satisfaisante.

Cela souligne l'importance de la **conception des invites**, qui nécessite une certaine connaissance du sujet examiné et une compréhension de la manière d'obtenir du modèle le meilleur résultat possible. En d'autres termes, les GML ne pourront pas remplacer l'interaction humaine requise pour définir clairement les besoins ou l'objet de la recherche afin de déterminer la méthode statistique la plus appropriée. Dans le cas d'une personne qui ne connaît pas le sujet, l'application aveugle des conseils d'un GML pourrait conduire à des résultats peu souhaitables.

2.2 Améliorer l'efficacité des tâches ordinaires sur le lieu de travail

Comme toute autre organisation, les organismes statistiques ont des tâches ordinaires à accomplir qui sont assez semblables à celles qui sont réalisées dans le secteur public ou privé. Ces tâches comprennent des activités telles que la gestion des courriels, l'élaboration de rapports et d'exposés et la prise de notes lors de réunions. Bien qu'elles soient essentielles au bon fonctionnement de l'organisme, ces tâches routinières exigent beaucoup de temps et d'efforts de la part d'un personnel dévoué.

Les GML et ChatGPT peuvent aider à rationaliser les opérations au sein des organismes statistiques et à accroître la productivité des ressources existantes. Ainsi, ces organismes peuvent consacrer plus efficacement leurs ressources aux tâches essentielles et contribuer à l'objectif qui est pour eux de fournir des informations statistiques exactes et actuelles. Dans la section qui suit, des exemples seront donnés sur la façon dont les GML et ChatGPT peuvent être utilisés pour renforcer l'efficacité d'un organisme statistique et permettre à celui-ci d'atteindre ses objectifs fondamentaux de manière plus efficace.

1) **Communications** − L'une des applications les plus répandues des GML est l'application immédiate des fonctionnalités de ces modèles dans les communications. Il a été prouvé que les GML facilitaient la rédaction des courriels, des projets et des rapports en proposant des contenus, en apportant une aide à la mise en forme et en produisant le texte lui-même[6], ce qui permettait de gagner du temps et d'améliorer la qualité des documents écrits. Pour les rapports, les GML résument de longs documents, proposent des modes de visualisation des données, repèrent les erreurs et formulent des recommandations.

2) **Réflexion et recherche d'idées** − Les GML peuvent faciliter le déroulement des séances de réflexion en formulant des propositions créatives, en envisageant un problème sous différents angles et en trouvant de nouvelles idées sur la base des informations fournies. Cela peut être particulièrement utile pour diversifier les perspectives, analyser les différents aspects d'un problème, soulever des questions afin d'approfondir l'analyse, évaluer les idées, mettre en forme les résultats et gagner du temps.

3) **Gestion et planification des projets** − Les GML peuvent être utilisés efficacement pour accomplir les tâches courantes requises à différents stades du processus de gestion des projets en automatisant la planification des tâches et la gestion des interdépendances, en optimisant l'affectation des ressources sur la base des données historiques et des exigences du projet, en estimant la durée des tâches afin de planifier le calendrier et en simplifiant la prise de notes lors des réunions grâce à la transcription et à l'élaboration de résumés, ce qui permet de garantir que les informations essentielles soient effectivement enregistrées et récapitulées.

4) **Traduction à partir ou vers d'autres langues** − Les GML peuvent traduire des documents et des textes d'une langue dans une autre et faciliter ainsi l'accès à l'information en différentes langues. En règle générale, à leur stade actuel de développement, les GML et ChatGPT présentent de grands avantages pour la traduction, puisqu'ils sont plus sensibles au contexte. Toutefois, les systèmes de traduction automatique traditionnels conservent leurs avantages dans les scénarios comportant de grands ensembles de données, des impératifs de rapidité et d'efficacité et des domaines bien définis.

5) **Exposés** − Les GML peuvent être utilisés pour réaliser des exposés à l'aide de diaporamas simples ou plus élaborés. Ils peuvent servir non seulement à produire le contenu des diapositives, en en personnalisant le style, la structure et la quantité, mais aussi à élaborer des éléments de langage qui donneront à l'exposé un caractère plus convivial.

[6] Il est à noter que les courriels générés par les GML peuvent être faciles à reconnaître ; un conseil amical : ne faites pas directement un copier-coller du texte généré par ChatGPT sans procéder au formatage du style, sinon la police originale et l'arrière-plan gris seront conservés.

6) **Objectifs pédagogiques** − Les GML peuvent être utilisés à des fins pédagogiques et pour la formation au sein de l'organisation. Ils peuvent apporter des explications, créer des questionnaires et contribuer à la conception de supports d'apprentissage en ligne afin d'améliorer les compétences et les connaissances du personnel.

7) **Création d'images** − Des images sont souvent utilisées dans les rapports et les publications des organismes statistiques. Plutôt que d'acheter des images standard, ces derniers pourraient utiliser les GML pour créer des images destinées à accompagner leurs productions statistiques.

L'utilisation judicieuse des GML et de ChatGPT peut libérer des ressources humaines au profit de tâches plus stratégiques et plus complexes et permettre ainsi au personnel d'être plus créatif et plus productif et de concentrer son activité sur des domaines plus prioritaires.

2.3 Améliorer l'efficacité de la production statistique et la qualité de la prestation de services

Les GML peuvent être utilisés dans une large gamme d'applications afin d'améliorer l'efficacité à différents stades du processus de production des statistiques, sous réserve d'une supervision humaine et d'un examen minutieux par rapport aux méthodes existantes et à l'expertise accumulée dans les organisations, par exemple dans les domaines suivants :

- Conception de la collecte (sous-processus 2.3 du GSBPM[7]) : les GML peuvent contribuer à la conception d'enquêtes et de questionnaires en proposant des questions, des modes de présentation et des formulations plus à même de susciter des réponses précises ;
- Classification et codage (sous-processus 5.2 du GSBPM) : les GML sont capables de classer automatiquement les données textuelles dans des catégories ou sous des étiquettes prédéfinies. Les organismes statistiques peuvent les utiliser pour organiser les réponses aux enquêtes et d'autres données textuelles en catégories appropriées dans les systèmes de classification statistique ;
- Validation et édition des données (sous-processus 5.3 et 5.4 du GSBPM) : les GML peuvent rationaliser les tâches de vérification et de prétraitement des données en détectant et en rectifiant les erreurs de données, les valeurs manquantes et les incohérences ;
- Élaboration des produits de diffusion (sous-processus 7.2 du GSBPM) : Les GML peuvent produire des descriptions textuelles à partir d'un tableau ou d'une série de chiffres (voir le cas d'utilisation décrit dans la section 3.4. Élaboration de rapports à l'aide des GML (Statistique Canada)) qui peuvent être adaptées à différents publics, y compris les décideurs, les journalistes et le grand public. Cela pourrait grandement simplifier le travail des analystes et des experts en communication en proposant des projets initiaux sur lesquels les experts humains pourraient travailler. Les GML peuvent également contribuer à automatiser la création de diagrammes et de graphiques, bien que ce domaine soit encore à l'étude ;
- Les métadonnées jouent un rôle crucial dans la production statistique et l'édition des métadonnées peut être assistée par les GML (voir le cas d'utilisation décrit dans la section 3.5. Edition des métadonnées à l'aide de GPT (Banque des règlements internationaux)).

Outre leurs applications dans le processus de production des statistiques, les GML peuvent apporter un soutien dans plusieurs domaines transversaux d'une importance cruciale pour les organismes statistiques :

- Aide au codage et à la traduction entre les différents langages de programmation : les GML peuvent traiter non seulement les langues naturelles mais aussi les langages de programmation que les organismes statistiques utilisent largement pour une grande

[7] Modèle générique du processus de production statistique (https://statswiki.unece.org/display/GSBPM/).

partie de leur production, en particulier pour le traitement et l'analyse. Ils pourraient améliorer considérablement l'efficacité des programmeurs et des analystes en aidant à rationaliser et à optimiser l'élaboration des codes, en fournissant des éléments de code et en permettant de passer d'un langage de programmation à un autre (voir le cas d'utilisation décrit dans la section 3.2. Traduction et explication des codes (de SAS à R) à l'aide des GML (Office central irlandais de la statistique)) ;

- Actualisation et tenue à jour des normes statistiques : élaboration de projets de descriptifs destinés à aider les experts humains à mettre à jour les systèmes de classification statistique (voir le cas d'utilisation décrit dans la section 3.1. Mise à jour des définitions de la classification statistique (Bureau australien de statistique)) et des documents méthodologiques ;

- Production de données synthétiques : la protection de la vie privée et l'utilisation des données sont des préoccupations majeures lors de l'essai de méthodes statistiques. Les GML peuvent être utilisés pour produire des données textuelles synthétiques, ce qui permet d'utiliser les méthodes à titre expérimental dans des environnements d'essai, sans avoir recours à des données réelles.

Plus particulièrement, la capacité des GML à traiter rapidement une grande quantité d'informations textuelles et à interagir avec les humains dans des langues naturelles pourrait accroître considérablement l'expérience des utilisateurs sur les plateformes de diffusion de statistiques. Actuellement, la plateforme de diffusion de la plupart des organismes statistiques est structurée par domaine et par sujet. Les utilisateurs doivent cliquer sur plusieurs pages et, dans le pire des cas, effectuer plusieurs allers et retours pour trouver les statistiques qu'ils recherchent. De plus, cette structure peut être lourde pour les utilisateurs qui cherchent et intègrent des données relatives à plusieurs domaines et plusieurs sujets. Bien que les organismes statistiques se soient efforcés de fournir des produits au format adapté à différents publics (par exemple, des chiffres clefs pour les journalistes, des données brutes pour les chercheurs, des rapports d'analyse pour les décideurs politiques), les utilisateurs qui ne sont pas familiarisés avec les moyens d'accès à ces produits sur le site Web pourraient rencontrer des difficultés. Les GML peuvent contribuer à aplanir ces difficultés et à améliorer la qualité de la fourniture de données aux utilisateurs, objectif ultime des producteurs de statistiques officielles, notamment par les moyens suivants :

- Requêtes interactives : le fait de permettre aux GML d'engager un dialogue avec les utilisateurs pour mieux cerner leurs besoins d'information et affiner les requêtes peut déboucher sur des réponses plus précises et plus pertinentes (voir le cas d'utilisation décrit dans la section 3.3. StatGPT (Fonds monétaire international)) ;

- Fourniture d'informations personnalisées : les organismes statistiques peuvent permettre aux utilisateurs de personnaliser la manière de recevoir les informations statistiques fournies par les GML. Certains utilisateurs peuvent préférer des rapports de synthèse, tandis que d'autres peuvent rechercher des analyses approfondies ou des données brutes ;

- Aide à l'interprétation des données : Les GML peuvent aider les utilisateurs à interpréter les données statistiques complexes en fournissant des explications, des représentations visuelles et un contexte. Cela aide les utilisateurs à comprendre la signification et les conséquences à tirer des statistiques qu'ils consultent.

2.4 Changer la manière de trouver les informations et les connaissances

Les organismes statistiques se sont adaptés à l'évolution du paysage de la diffusion d'informations en diversifiant leurs canaux afin d'atteindre autant que possible les utilisateurs de données et le public en général. Au cours des dix dernières années, la manière de trouver des informations a considérablement évolué. Les personnes qui cherchent des statistiques officielles se rendent rarement directement sur les sites Web des organismes statistiques ; elles commencent souvent leur recherche sur des plateformes telles que Google.

Ces moteurs de recherche et ces plateformes numériques utilisent des algorithmes (par exemple, l'index de recherche de Google) pour passer au crible les nombreuses informations

disponibles sur le Web et présenter aux utilisateurs des informations pertinentes. Par exemple, lors d'une recherche sur le « taux d'inflation du pays X au cours de l'année Y », ces plateformes peuvent afficher les statistiques officielles de l'organisme national de statistique concerné, mais peuvent également inclure des données provenant d'autres sources. Bien que le fonctionnement exact de ces algorithmes ne soit pas divulgué, des stratégies ont vu le jour pour améliorer la visibilité et la présentation du contenu sur ces plateformes, auxquelles de nombreux organismes statistiques se sont adaptés.

Cependant, avec l'émergence et la popularité croissante des services simples à utiliser fondés sur les GML (par exemple, ChatGPT), le paradigme de la recherche d'informations commence une fois de plus à changer. Il est déjà possible pour les GML, via des invites destinées aux utilisateurs, de retrouver des statistiques historiques à partir de leurs données d'apprentissage, sans l'aide des organismes statistiques officiels. Cependant, il y aura des problèmes d'actualité et de qualité des données produites, selon l'âge et la source des données d'apprentissage utilisées par les GML. Ces problèmes d'actualité et de précision ne sont pas toujours évidents pour l'utilisateur moyen de ces modèles ; il n'est pas non plus évident que les GML soient actuellement incapables de produire des statistiques à jour.

Tout en reconnaissant les risques liés à l'utilisation des GML, les organismes statistiques officiels doivent comprendre les possibilités qu'offrent ces modèles et l'incidence qu'ils peuvent avoir sur l'offre de statistiques officielles et sur les utilisations expérimentales de la statistique.

Pour que la statistique officielle reste pertinente à l'ère des GML, les organismes statistiques doivent fournir des services que ces modèles ne peuvent pas assurer seuls, en donnant aux utilisateurs de statistiques officielles la possibilité de choisir des sources de haute qualité, de grande précision et d'un degré d'actualité élevé.

Les organismes statistiques officiels peuvent décider d'agir de la sorte dans leur propre pays ou au niveau interne, ou de travailler ensemble et avec les fournisseurs de GML, afin de proposer, en utilisant la puissance de ces modèles, des produits statistiques combinés, non encore disponibles à l'heure actuelle. Les GML doivent être considérés comme un élément clef pour une fourniture de statistiques plus actuelle et plus efficace, tant au niveau national qu'international.

5. Considérations relatives à l'utilisation des grands modèles de langage par les organismes statistiques

Les GML offrent de nombreuses possibilités aux organismes statistiques, mais il est essentiel de procéder avec prudence tout en tenant compte de divers facteurs lors de l'intégration de ces modèles au sein des organismes. Dans la présente section, nous décrivons les principaux aspects à prendre en considération lors de l'examen des GML, tels que la gouvernance, la collaboration avec les entreprises technologiques, les modèles d'accès libre et les relations publiques. Bien que le sujet évolue rapidement, nous formulons de brèves suggestions pratiques à la fin de la section.

5.1 Gouvernance

Pour bénéficier des avantages que promettent les GML et les transformateurs génératifs préentraînés (modèles GPT), comme cela est indiqué dans les sections 2 et 3, les organismes doivent mettre en place de nouvelles mesures de gouvernance ou intégrer leur propre cadre de gouvernance interne afin de limiter les risques décrits dans la section 4. Les domaines à risque qui sont abordés dans cette section sont l'éthique et les préjugés, l'exactitude, la protection de la vie privée et la sécurité, les litiges relatifs aux droits d'auteur et les questions juridiques, ainsi que l'utilisation abusive potentielle due au manque de connaissances et de compréhension. De possibles stratégies d'atténuation des risques sont également décrites dans cette section.

Dans la présente section, nous examinons comment il est possible de régir les GML en mettant en œuvre ces stratégies d'atténuation, dans le contexte d'organismes statistiques

modernes opérant dans un environnement déjà déterminé par des lois nationales, des cadres et des accords internationaux, des circonstances techniques changeantes dans lesquelles interviennent des acteurs dominants, ainsi que par la culture de l'organisme considéré.

Régir les grands modèles de langage

Lorsque la gouvernance s'applique à la mise en œuvre ou à l'utilisation d'un GML, les parties prenantes du projet doivent établir des objectifs raisonnables et adaptés au projet, qui soient en accord avec les valeurs fondamentales de l'organisme concerné et les principes de la statistique officielle et s'inscrivent dans le contexte national. Il est à noter que la gouvernance sera toujours limitée par le fait que les GML et les modèles GPT les plus puissants sont en fin de compte détenus et contrôlés par des tiers et qu'en raison de leur taille, ils doivent le plus souvent fonctionner sur des plateformes en nuage qui sont également contrôlées par des tiers.

Par conséquent, ce qui est recommandé n'est pas de mettre intégralement en œuvre une IA responsable, mais plutôt d'insister sur les problèmes et les conflits que soulève l'IA générative (et en particulier les services fondés sur les GML) en ce qui concerne l'IA responsable. Lorsque nous introduisons des GML ou des modèles GPT dans l'organisation des tâches (que ce soit sous forme de produits tiers disponibles sur le marché, obtenus via une interface de programmation d'applications (API), ou d'un modèle de base réglé finement et intégré dans un produit conçu et mis en œuvre en interne), nous devons prendre en compte les problème et les conflits liés à l'utilisation de ces modèles dans le cadre des tâches ou des applications envisagées et définir les mesures d'atténuation appropriées.

Régir les grands modèles de langage et les transformateurs génératifs préentraînés dans la conjoncture technique actuelle

Les GML et les modèles GPT sont rarement entraînés exclusivement sur des ensembles de données locaux ou librement accessibles au public. Ils sont souvent entraînés, hébergés et exploités sur une plateforme tierce, comme celles qui sont fournies par Amazon (AWS), Google (GCP) ou Microsoft (Azure). Les organismes concluent des accords avec les fournisseurs de technologie afin de garantir la protection des intérêts nationaux essentiels et le respect des lois applicables (s'agissant par exemple de conserver les données hébergées sur des serveurs locaux). Il n'en reste pas moins que les organismes n'ont pas la maîtrise de certaines sources et de certains produits faisant partie des modèles qu'ils utilisent et que ces éléments peuvent même ne pas être entièrement visibles pour leur personnel.

Par conséquent, dans les organismes statistiques, la nature et le niveau de la gouvernance des GML et des modèles GPT dépendront de la manière dont ces modèles entrent dans la sphère de l'organisme. La gouvernance d'un projet dans lequel un modèle de ce type est mis au point (finement réglé, par exemple) sera différente de la gouvernance entourant l'utilisation d'une application tierce à code source fermé. Dans chaque cas, la gouvernance impliquera de décrire les risques et de spécifier les mesures d'atténuation appropriées. De plus amples détails sur les catégories de risques et les éventuelles mesures d'atténuation sont présentés dans la section 4.

Quelques exemples de gouvernance des GML et des modèles GPT sont donnés ci-dessous.

Exemple A : Accord de licence pour l'installation d'une application tierce fondée sur un GML ou un modèle GPT

Microsoft intégrera dans sa suite bureautique Office365 son outil d'intelligence artificielle appelé CoPilot, qui devrait, selon ses dires, améliorer la productivité sur le lieu de travail. Un certain niveau de gouvernance sera mis en place sur le plan juridique − par exemple, l'obligation d'héberger les données à l'étranger. Toutefois, une partie de la gouvernance devra faire l'objet de mesures plus souples une fois que CoPilot sera installé et utilisé. En effet, il est possible que le personnel de l'organisme statistique qui demande des informations à l'outil assisté par l'IA soit trop confiant dans l'exactitude des résultats et publie ou communique des informations potentiellement erronées, ou prenne des décisions sur la base d'informations incorrectes ou incomplètes. Pour plus de détails, voir l'exemple de la section 3.1 au sujet des erreurs de discernement entre les listes de tâches

professionnelles établies par des humains et celles qui sont générées par un GML et, dans la section 4.5, la discussion générale au sujet des utilisations abusives. Les organismes statistiques ne peuvent pas éliminer le risque d'utilisation abusive, mais peuvent mettre en place les mesures d'atténuation décrites dans la section 4.5, qui consistent à améliorer la connaissance des données et de l'IA, à établir des protocoles d'utilisation clairs et à mettre en place des garde-fous techniques afin de prévenir les utilisations abusives.

Exemple B : Source ou produit mis au point en interne qui utilise un GML préentraîné ou un modèle GPT

De plus en plus, les développeurs internes qui connaissent bien les objectifs, les ensembles de données et les cas d'utilisation de l'organisme, tels que les spécialistes des données ou les ingénieurs spécialisés dans apprentissage automatique, ont tendance à utiliser des modèles préentraînés (également appelés « modèles de fondation »). L'organisme sera limité dans sa capacité de gérer pleinement le produit ou la source qui utilise le modèle de fondation.

Par exemple, il sera difficile pour les organismes statistiques de s'assurer que le produit ou la source n'utilise pas de composants (ensembles de données ou codes) élaborés ou conçus dans un environnement où les normes de travail humain sont médiocres. Il sera difficile de prouver que l'exactitude des données est acceptable et que le modèle n'est pas biaisé, comme cela est indiqué dans la section 4.2, ou qu'il n'y a pas eu de corruption de données, comme cela est expliqué dans la section 4.3.

Même lorsque les tierces parties qui fabriquent ces modèles de fondation publient les codes ou les données d'apprentissage par l'intermédiaire d'un registre public ou offrent aux utilisateurs une licence moins restrictive d'accès libre ou de sources ouvertes, il y a toujours un manque de transparence. En effet, l'indice de transparence des modèles de fondation publié par le Center for Foundation Models, basé à Stanford, a attribué une note allant de 0 à 100 à de nombreux modèles de fondation importants, en attribuant un point à chaque critère au sujet duquel l'entreprise a fourni des informations suffisantes pour répondre à chaque question. Le modèle Llama 2 de Meta a reçu la meilleure note, soit 54/100. Cela signifie que pour 46 critères, Meta n'a pas fourni suffisamment d'informations pour que les chercheurs considèrent ce critère de transparence comme satisfaisant[8].

Compte tenu de ces conflits et de ces tensions, nous ne recommandons pas d'interdire les GML ou les modèles GPT, car cela créerait un risque d'IA fictive au sein des organismes statistiques. Nous recommandons plutôt d'évaluer les risques inhérents à chaque projet ou chaque application et de mettre en place les mesures d'atténuation appropriées.

La gouvernance en pratique − évaluation et suivi

Critères d'évaluation : Lorsqu'un GML est utilisé pour fournir des réponses à des requêtes ou à des recommandations, l'efficacité du modèle doit être évaluée en fonction de critères tels que la fidélité (par exemple, le texte généré est-il fidèle au document source ?), la reproductibilité (les mêmes résultats ou des résultats similaires sont-ils obtenus pour la même requête ou pour une requête similaire) et la pertinence (la réponse répond-elle à la requête ?). L'évaluation porte également sur la manière dont les résultats reflètent les valeurs de l'organisme (par exemple, les résultats peuvent-ils nuire à la réputation de l'organisme ?). Les développeurs devront peut-être aussi envisager d'ajuster les paramètres afin que les résultats soient de nature impartiale, politiquement neutre et factuelle et qu'ils s'adressent au public approprié (quel qu'il soit). Les textes générés doivent être vérifiés pour s'assurer qu'ils ne plagient pas involontairement des publications existantes et bien que ces textes doivent encore être finalisés, le risque de publicité négative et de perte de confiance de la part du public ne vaut pas la peine d'être encouru.

Suivi : La mesure dans laquelle une source ou un produit obtenu au moyen de l'intelligence artificielle ou d'un GML atteint les objectifs qui ont été fixés ou augmente les risques doit être dûment mesurée, contrôlée et décrite pendant toute la durée de vie de la source ou du produit. Une étape du suivi pourrait consister, pour les parties prenantes, à

[8] https://crfm.stanford.edu/fmti/.

réaliser des évaluations de seuil ou d'impact, dans le cadre desquelles l'avancement du projet et l'utilisation du produit seraient évalués en fonction des catégories de risque correspondantes. Afin de garantir que les systèmes d'IA restent responsables au fil du temps, le projet devrait inclure un plan de maintenance portant notamment sur la fréquence d'actualisation des données d'apprentissage, ainsi que des examens des méthodes et des codes, afin de vérifier que le modèle d'IA est bien à jour. Les points ci-dessus doivent être intégrés dans la maintenance, de sorte que les modifications apportées à chacun de ces éléments soient prises en compte régulièrement.

5.2 Collaboration avec les entreprises technologiques qui fournissent des services fondés sur les grands modèles de langage

L'écosystème des GML est un domaine complexe qui évolue rapidement. Au cœur de cet écosystème se trouvent des entités majeures telles que Google, OpenAI, Microsoft et Meta AI, qui jouent un rôle central dans la définition et le perfectionnement des technologies liée aux GML. Dans ce contexte, il est vital pour les organismes statistiques d'étudier et de mettre en avant l'utilisation de modèles et de plates-formes à sources ouvertes. Des entreprises telles que Hugging Face et EleutherAI, qui s'appuient sur la notion de sources ouvertes, contribuent à la création d'un environnement plus diversifié et plus accessible. Pour collaborer avec ces entités, il faut trouver un équilibre entre les technologies brevetées et les technologies à code source ouvert afin de stimuler l'innovation et de maintenir des normes éthiques.

Il est essentiel de comprendre les différents rôles des entreprises technologiques dans l'écosystème des GML. En tenant compte de facteurs tels que les offres de base, les rôles au sein de l'écosystème et la gamme de services fournis, les organismes statistiques peuvent s'orienter utilement dans cet espace.

Rôle des fournisseurs de services en nuage

Les fournisseurs de services en nuage font partie intégrante du fonctionnement et du perfectionnement des GML. Lorsqu'ils collaborent avec ces fournisseurs, les organismes statistiques doivent tenir compte de plusieurs facteurs importants. La confidentialité et la sécurité des données sont primordiales, tout comme l'évolutivité et l'efficacité des services. La gestion des coûts est un autre domaine clef, qui demande une compréhension claire des systèmes de tarification et des frais cachés éventuels. La garantie de la conformité juridique, et notamment la disponibilité des services dans certaines régions (par exemple, l'Europe ou l'Europe de l'Ouest) et la compatibilité technique avec les systèmes existants d'un organisme statistique, sont également des éléments essentiels à prendre en compte.

Le marché mondial de l'informatique en nuage est principalement dominé par les « trois grands » : Azure, AWS et Google Cloud. Cependant, d'autres fournisseurs se spécialisent souvent dans des services de niche qui offrent des modes d'intégration spécifiques, potentiellement plus adaptés à certains organismes statistiques. Le choix d'un fournisseur de services en nuage pour l'infrastructure ou les plateformes d'IA doit répondre à des besoins spécifiques et s'inscrire dans une perspective de développement à long terme. Il est également important d'être conscient des risques de dépendance vis-à-vis des acteurs principaux de l'écosystème des GML.

Écosystème des grands modèles de langage

Dans l'écosystème des GML, les services offerts par les entreprises technologiques appartiennent souvent à plusieurs catégories, ce qui met en évidence la nature interconnectée de ce domaine. Par exemple, Azure Machine Learning de Microsoft permet aux utilisateurs d'accéder aux modèles mis au point par OpenAI et Meta AI, ainsi qu'à certains des modèles de Hugging Face. De même, Hugging Face se distingue en offrant un large éventail de services dans presque toutes les catégories de l'écosystème des GML.

Pour les organismes statistiques, il est essentiel de reconnaître et de comprendre ces rôles à multiples facettes. En déterminant à quelle(s) catégorie(s) spécifique(s) appartient l'activité d'une entreprise, les organismes statistiques peuvent organiser de manière plus

stratégique leur collaboration avec les entreprises technologiques. Cette connaissance leur permet de déterminer avec précision quelles entreprises offrent les services les plus pertinents et les plus bénéfiques pour répondre à leurs besoins particuliers, qu'il s'agisse de tirer parti de modèles d'IA avancés, d'accéder à divers ensembles de données ou d'utiliser des plateformes d'apprentissage efficaces. En outre, la compréhension de ces catégories aide les organismes statistiques à anticiper et à éviter les éventuels chevauchements de services et de collaborations, ce qui garantit une approche plus rationnelle et plus efficace de l'intégration des technologies dérivées des GML dans leurs activités.

La catégorie **des développeurs et des fournisseurs de grands modèles de langage** comprend les entreprises spécialisées dans la recherche, la conception et le déploiement de ces grands modèles. Parmi les exemples notables, on peut citer OpenAI, Meta AI, Google DeepMind, ainsi que des acteurs utilisant des sources ouvertes comme EleutherAI et le Technology Innovation Institute (TII). Ces entreprises sont à la pointe du progrès dans le domaine des technologies liées aux GML. D'un point de vue technique, il est essentiel de veiller à ce que les modèles fournis par ces prestataires puissent être intégrés sans heurt dans les systèmes des organismes de statistique. Les GML doivent donc être compatibles avec l'infrastructure existante et pouvoir s'adapter à des impératifs techniques spécifiques. Le respect des normes éthiques des organismes statistiques est primordial. Il est en outre essentiel que les GML soient conformes aux principes de l'IA responsable, notamment la transparence, l'équité, la protection de la vie privée et l'obligation de rendre compte. Le fait de veiller à ce que ces modèles soient conçus et déployés d'une manière éthique est en phase avec des valeurs sociétales et des cadres réglementaires plus larges.

Les **fournisseurs d'infrastructures et de plateformes d'IA** constituent la deuxième catégorie de l'écosystème des GML et comprennent les entreprises qui fournissent l'infrastructure matérielle et logicielle nécessaire pour entraîner, déployer et utiliser ces modèles, comme Microsoft Azure ML, la plateforme Google Cloud AI, AWS SageMaker et bien d'autres encore. Pour les organismes statistiques, la collaboration avec ces fournisseurs nécessite de se concentrer sur l'évolutivité, l'efficacité, la compatibilité technique et une compréhension approfondie de la structure des coûts, y compris les éventuelles dépenses cachées.

Les **développeurs d'applications fondées sur les grands modèles de langage** sont des entreprises technologiques qui jouent un rôle déterminant dans la conception d'applications ou de services qui utilisent les GML afin de mettre au point certaines fonctionnalités, comme les dialogueurs. L'innovation dans la création d'applications, la conception centrée sur l'utilisateur et le respect des normes de confidentialité des données sont des aspects essentiels de la contribution de ces entreprises.

Les **services de personnalisation et de réglage fin de l'IA** sont le fait d'un ensemble crucial d'entreprises telles que les jeunes pousses de l'intelligence artificielle et les entreprises technologiques spécialisées qui adaptent les GML existants pour répondre aux besoins spécifiques de leurs clients. L'adaptabilité de ces entreprises et leur capacité d'intégrer des solutions personnalisées dans les systèmes existants sont des aspects essentiels pour les organismes statistiques.

Les **laboratoires de recherche et d'innovation** dans le domaine des GML, qui comprennent des laboratoires de recherche universitaires et des services de recherche-développement, sont tout aussi importants. Ces entités repoussent les limites de ce que les modèles de langage peuvent réaliser, en se concentrant sur la recherche de pointe et les pratiques éthiques en matière d'IA. Leurs travaux contribuent grandement à élargir la base de connaissances dans les domaines de l'intelligence artificielle et des GML. La collaboration avec ces laboratoires peut permettre aux organismes statistiques d'accéder aux recherches et aux pratiques éthiques les plus récentes en matière d'IA.

Dans cet écosystème, **les entreprises spécialisées dans les grands modèles de langage et les initiatives relatives aux sources ouvertes** jouent un rôle essentiel. Des plateformes comme Hugging Face et divers registres GitHub consacrés à la recherche dans le domaine des modèles de langage favorisent la collaboration entre les entreprises, en promouvant une culture du code source ouvert dans la conception de ces modèles. Ces initiatives stimulent l'innovation et garantissent l'accessibilité des outils et des ressources, ce

qui est primordial pour un écosystème des GML à la fois collaboratif et inclusif. Les organismes statistiques devraient collaborer avec ces initiatives afin d'accéder à une large palette d'outils et de ressources libres.

Dans la toute dernière catégorie, celle des **services de données et d'apprentissage utilisés pour les grands modèles de langage**, il existe des entreprises qui jouent un rôle essentiel dans la fourniture des ensembles de données à la fois vastes et variés qui sont nécessaires pour entraîner les GML. Au-delà de fournir des données, ces entités peuvent aussi proposer des services déterminants pour faciliter le processus d'apprentissage des GML. Des entreprises comme EleutherAI et Hugging Face se distinguent dans ce domaine en offrant toute une gamme d'ensembles de données et d'outils essentiels à l'élaboration de modèles robustes et efficaces. Leur contribution est indispensable pour garantir que les GML soient entraînés à partir d'ensembles de données diversifiés, étendus et de haute qualité, ce qui est une condition fondamentale de la précision et de la fiabilité des modèles. En outre, ces services comprennent souvent des outils et des plateformes qui contribuent à l'efficacité de l'apprentissage, ce qui en fait un élément indispensable de l'écosystème des GML. Les organismes statistiques devraient collaborer avec ces entités afin d'obtenir des sources de données diversifiées et de haute qualité, ainsi que des plateformes d'apprentissage efficaces.

Chaque catégorie de l'écosystème des GML offre des possibilités uniques de collaboration, contribuant ainsi à la croissance globale et à l'utilisation éthique des technologies liées à ces modèles.

5.3. Considérations relatives au libre accès

Plusieurs aspects des GML et des fournisseurs auxquels font appel les organismes statistiques doivent être examinés avec soin. Les principales dimensions à prendre en compte sont l'accessibilité de la structure fondamentale du modèle et des données d'apprentissage, les conditions de licence régissant l'utilisation du modèle et l'accès aux entrées et aux sorties lors de l'utilisation du modèle. L'évaluation implique généralement une analyse des arbitrages, qui consiste à mettre en balance les avantages de la commodité et la nécessité d'un contrôle. Les coûts et l'accès aux compétences sont également des points à prendre en considération, car pour pouvoir fonctionner à grande échelle, les GML requièrent une infrastructure et une expertise informatiques importantes.

L'accessibilité des GML couvre un large spectre. Certains modèles sont librement accessibles, ce qui permet d'inspecter et de modifier leur architecture et leurs pondérations au moyen d'un réglage fin. À l'inverse, d'autres restent des ressources exclusives, dont l'accès n'est possible que par l'intermédiaire d'API ou d'autres interfaces. Dans certains cas, les fournisseurs de modèles à accès fermé peuvent néanmoins offrir des options de réglage fin, ce qui permet aux utilisateurs d'adapter les pondérations du modèle à leurs données et leurs cas d'utilisation spécifiques. Bien que l'accès direct aux pondérations du modèle puisse sembler sans importance, la capacité de personnaliser un GML en fonction de données et de cas d'utilisation particuliers peut se révéler très utile pour un organisme statistique.

En ce qui concerne les modèles librement accessibles, il est impératif de procéder à un examen diligent des conditions de licence. Les créateurs de GML peuvent imposer des conditions spécifiques régissant l'utilisation de ces modèles, que des utilisateurs mal informés peuvent, par inadvertance, ne pas respecter.

La transparence et l'accessibilité des données d'apprentissage du modèle sont primordiales pour évaluer la présence potentielle d'éléments biaisés, nuisibles ou protégés par le droit d'auteur, qui pourraient influer sur les résultats générés par le modèle (comme indiqué dans la section 4). Dans de tels cas, une transparence et un accès complets sont indispensables pour atténuer tout risque de mauvaise réputation, étant donné que l'ensemble des données d'apprentissage détermine dans une large mesure les résultats du modèle. Un autre aspect à prendre en considération concerne l'accès des GML aux informations contemporaines. Le corpus de connaissances utilisé pour entraîner le modèle dépend d'une date limite qui est antérieure au début du processus d'apprentissage. Pour remédier à cette limitation, il est nécessaire d'intégrer un contenu mis à jour, pratique souvent appelée

« Retrieval Augmented Generation » (RAG). En outre, certains systèmes sont équipés de mécanismes permettant aux GML d'accéder à des données en temps réel à partir de l'internet.

Enfin, la question de la confidentialité des données d'entrée et de sortie mérite d'être examinée attentivement. Dans de nombreux cas de services en nuage, ces données peuvent être conservées par le fournisseur des services afin de faciliter les itérations futures du GML par voie d'apprentissage et de réglage fin. Par conséquent, les utilisateurs peuvent faire face à des limitations concernant l'utilisation d'informations confidentielles. Toutefois, il convient de noter que certains fournisseurs commencent à donner accès à des modèles fermés dans un environnement de type « bac à sable », offrant ainsi aux utilisateurs la possibilité de maintenir un contrôle total et une confidentialité absolue des données d'entrée et de sortie.

En résumé, les organismes statistiques doivent mettre en balance les avantages et les inconvénients potentiels des modèles d'accès libre lorsqu'ils évaluent les GML, en particulier le niveau de transparence et la capacité de collaborer avec d'autres organismes statistiques résultant de la nature ouverte de ces modèles.

5.4 Communication avec le public

Le domaine des GML évolue rapidement. Bien que ces modèles offrent des capacités étonnantes, leur développement rapide place cette technologie d'IA dans une zone plutôt grise où l'opinion et le sentiment du public peuvent être incertains et sujets à des changements.

En tant qu'organismes publics dont les produits influent considérablement sur la politique et la prise de décisions à l'échelle nationale, les organismes statistiques ont la lourde responsabilité d'utiliser les GML de manière responsable et de communiquer à ce sujet avec le public de manière transparente. Étant donné que leur activité principale (c'est-à-dire la production de statistiques officielles et les services de fourniture de données) repose largement sur la confiance du public, il est indispensable qu'ils accordent encore plus d'attention à la communication avec la société, en particulier avec les fournisseurs de données qui pourraient craindre que leurs données soient utilisées à mauvais escient pendant les interactions avec les GML, et qu'ils investissent dans ce domaine. Après tout, le public a le droit fondamental de comprendre comment les données qui le concernent peuvent être utilisées et de savoir que des mesures sont en place pour protéger ces données.

Lorsqu'ils communiquent sur l'utilisation des GML, il est important que les organismes statistiques abordent les points suivants :

- Utilisation des GML à dessein lorsque cela présente des avantages certains : il est essentiel d'expliquer clairement pourquoi les GML sont utilisés dans les organismes statistiques et de mettre en évidence les avantages tangibles de cette technologie (efficacité accrue, réduction des coûts et amélioration des services, entre autres) à l'aide d'exemples concrets de bons résultats obtenus en utilisant les GML. Par exemple, des dialogueurs fondés sur les GML, qui aident le public à accéder aux données statistiques et à mieux les comprendre, sont l'une des fonctions que l'IA générative et les GML peuvent remplir de manière tout à fait autonome.
- Conscience des limites et des risques : il est important de démontrer que les organismes statistiques n'utilisent pas aveuglément les GML et qu'ils sont conscients des limites et des risques potentiels associés à ces modèles. Les domaines dans lesquels les GML sont utilisés et ceux dans lesquels ils ne le seront pas (par exemple, pour faire des prévisions individuelles qui pourraient avoir un effet négatif sur les personnes concernées) pourraient être mentionnés.
- Mise en œuvre des mesures d'atténuation nécessaires : il est essentiel d'expliquer les mesures prises pour réduire les inconvénients et les risques (par exemple, celles qui sont prises pour préserver la confidentialité et la sécurité des données), tout en soulignant que des humains interviennent pour superviser et encadrer l'utilisation des GML.

S’agissant de la communication interne, il serait important d’examiner ce qu’un cas d’utilisation particulier pourrait signifier implicitement pour les employés quant aux priorités de l’organisme qui les emploie. Par exemple, les GML pourraient constituer un moyen efficace de produire des résumés non techniques, mais cela pourrait également être perçu comme une façon de faire exécuter cette tâche par un modèle plutôt que d’encourager les compétences internes en matière de rédaction de textes non techniques à l’intention des membres intéressés du public.

5.5 Suggestions pratiques et remarques finales

L’utilisation des GML par les organismes statistiques n’en est encore qu’à ses débuts et le paysage évolue rapidement. Les meilleures pratiques se développent au fil du temps et les organismes statistiques devront s’efforcer constamment de les maintenir à jour. Les quelques suggestions pratiques qui suivent nous semblent pertinentes à court terme et devraient, selon nous, le rester à plus long terme.

La première est de dispenser une formation sur les GML à tous les niveaux de l’organisme (technique, opérationnel et managérial) afin de mieux faire connaître et de mieux faire comprendre les capacités et les limites de ces modèles.

Deuxièmement, nous suggérons d’aborder les GML en réalisant de petits projets pilotes afin de se familiariser avec la technologie et de comprendre les bénéfices qui pourraient en être tirés. De tels projets à petite échelle pourraient permettre d’accroître les capacités des organismes statistiques en la matière, de produire des résultats susceptibles de justifier et d’orienter d’autres investissements et, en fin de compte, d’atténuer les risques liés au développement de l’utilisation des GML.

Troisièmement, les organismes statistiques devraient élaborer une stratégie globale dans le domaine des GML une fois qu’ils seront suffisamment sensibilisés et familiarisés à ce sujet et qu’ils auront mené à bien quelques projets à petite échelle, comme nous l’avons vu plus haut.

Enfin, les organismes statistiques devraient s’efforcer en permanence de se tenir au courant de l’évolution constante des GML, tant d’un point de vue technologique que stratégique.

Reconnaissant les progrès rapides des GML, nous comprenons qu’il est difficile de prévoir à quel rythme ces progrès se poursuivront. Le présent livre blanc vise à rassembler les cas d’utilisation existants jusqu’à aujourd’hui et à explorer en profondeur le sujet sous différents angles présentant un intérêt pour les organismes statistiques. En raison de la nature dynamique de ce domaine, il est essentiel de travailler ensemble. C’est pourquoi nous invitons les experts à collaborer, à partager leurs points de vue et à suivre collectivement l’évolution de ce paysage. Nous confirmons notre engagement à explorer ce sujet et accueillons favorablement toute participation à cette exploration.