ACADÉMIE DES TECHNOLOGIES

POUR UN PROGRÈS RAISONNÉ, CHOISI ET PARTAGÉ

PROUESSES ET LIMITES DE L'IMITATION ARTIFICIELLE DE LANGAGES

LES AGENTS CONVERSATIONNELS INTELLIGENTS DONT CHATGPT

AVIS DE L'ACADÉMIE DES TECHNOLOGIES

– AVRIL 2023 –

Académie des technologies
Le Ponant – Bâtiment A
19 rue Leblanc
75015 PARIS
+33(0)1 53 85 44 44
secretariat@academie-technologies .fr
www .academie-technologies .fr

ISBN : 979-10-97579-45-6.

Illustration de couverture : Adobe Stock : Limitless Visions

ACADÉMIE DES TECHNOLOGIES

POUR UN PROGRÈS RAISONNÉ, CHOISI ET PARTAGÉ

PROUESSES ET LIMITES DE L'IMITATION ARTIFICIELLE DE LANGAGES

LES AGENTS CONVERSATIONNELS INTELLIGENTS DONT CHATGPT

AVIS DE L'ACADÉMIE DES TECHNOLOGIES

avril 2023

Remerciements

Gérard Roucairol, Président du pôle Numérique remercie chaleureusement les contributeurs à cet avis qui ont dû travailler dans un contexte scientifique, technique et médiatique rapidement évolutif.

Jean Claude André, Albert Benveniste, Yves Caseau, Thierry Chevalier , Nicolas Demassieux, Hervé Gallaire, Erol Gelenbe, Laurent Gouzenes , Stéphane Requena, Michèle Sebag, Joelle Toledano

L'agent conversationnel ChatGPT3[1] a été mis à la disposition du public en novembre 2022 par la Société OpenAI. De nombreux utilisateurs ont pu depuis explorer ses différents usages prévus ou imprévus. Cet engouement a suscité autant d'intérêt pour la multiplicité des possibilités offertes et la performance technologique que d'inquiétudes au sujet des limites de ce logiciel d'une part, et de son impact sociétal d'autre part[2].

Les principes de ChatGPT sont ceux de l'intelligence artificielle générative, c'est-à-dire capable d'apprendre des données fournies pour générer de nouvelles données « similaires ». Cette approche s'applique à de nombreux types de données : texte, image, vidéo, son, musique. Dans les domaines de l'image ou de la parole, on ne peut pas distinguer la production de l'IA générative (IA-G) de celle d'un humain : le système passe avec succès le test de Turing[3].

Qu'en est-il dans le domaine du langage naturel ? On observe que ChatGPT sait (souvent bien) répondre aux questions posées, quoique sans garantie de la véracité de la réponse (nous y reviendrons). Mais jusqu'à quel point comprend-il la question posée et sa propre réponse ?

L'Académie des technologies, dans le cadre de ses missions, a souhaité clarifier les possibilités et les limites de l'IA-G dans le cadre du langage. L'analyse présentée porte principalement sur ChatGPT3, le système le mieux connu en mars 2023, et pour lequel nous disposons d'un certain recul (en notant que la version suivante GPT4, plus puissante, pallie certains défauts de GPT3 sans toutefois remettre en cause les recommandations formulées par l'Académie).

Bien entendu, il s'agit ici d'un premier avis de l'Académie. La technologie de l'IA-G pour le langage est caractérisée actuellement par une rapidité inhabituelle d'évolution. Nous suivons donc ces évolutions et nous formulerons des mises au point, le cas échéant.

1 https://openai.com/blog/chatgpt

2 Voir deux annonces de dernière minute en date de publication de ce texte : une lettre ouverte (Pause Giant AI Experiments: An Open Letter) de scientifiques et industriels de renom, et des dépôts de plainte entreprises par diverses administrations ou États.

3 Voir https://en.wikipedia.org/wiki/Turing_test où il est écrit : The Turing test, originally called the imitation game by Alan Turing in 1950, [2] is a test of a machine's ability to exhibit intelligent behaviour equivalent to, or indistinguishable from, that of a human.

Sommaire

À QUOI SERT UN AGENT CONVERSATIONNEL INTELLIGENT ?

Les agents conversationnels intelligents ont le potentiel de modifier profondément les solutions techniques et les méthodes de travail actuellement utilisées dans de très nombreuses applications.

- La recherche d'information par requête en langage naturel, en complément ou remplacement d'outils comme Google search : pour le grand public ; dans un contexte professionnel spécifique (e.g. agents conversationnels pour le service après-vente, la maintenance, les agences de voyage) ; pour les chercheur.e.s et étudiant.e.s.
- La traduction automatique.
- L'aide à la génération de résumés ou mots-clefs pour des textes, ou de légendes pour des images ou vidéos.
- L'aide à la génération de contenus institutionnels, culturels, commerciaux… ou toxiques (spams, fake news).
- L'aide à la production de code logiciel à partir de descriptions de haut niveau, ou
- Le portage de logiciels existants vers de nouveaux langages (retargetting).

Pour toutes ces applications, lorsqu'elles sont utilisées de manière appropriée (compte tenu de leurs limites, voir 4. Étude de cas), les IA génératives constituent une rupture technologique majeure, à même de transformer la manière d'exercer de très nombreux métiers.

Réaliser un agent conversationnel intelligent et le mettre à disposition du public : comment ?

Les agents conversationnels s'appuient sur un modèle de langage, permettant de compléter une phrase incomplète. Ainsi « le chat a mangé… ? » peut devenir « le chat a mangé la souris/le poisson/l'oiseau/les croquettes » en fonction du contexte.

Étape 1, apprentissage d'un « modèle de langue » : un « langageur »

Le modèle est entraîné à partir d'un ensemble de textes (corpus). Pour chaque phrase du corpus, on crée des phrases incomplètes et on optimise le modèle pour reconstituer la phrase initiale. Ce qui est difficile est de prendre en compte le contexte de la phrase, du paragraphe, du document… Plus le contexte contient d'information, meilleure est la décision du système en général et plus grande est la base de textes nécessaire pour l'apprentissage. Les modèles de langue courants, appelés *Large Language Models* (grands modèles de langue, pour lesquels nous proposons le néologisme de ***langageur***) comprennent des centaines de milliards de paramètres (on peut les voir comme les coefficients d'un énorme polynôme). La prise en compte d'un mot peut être modifiée par la présence d'autres mots selon un mécanisme dit d'attention ; l'article fondateur de cette approche a été publié en 2017 par une équipe de Google[4].

4 Attention Is All You Need. A. Wasvani et al., 31st Conference on Neural Information *Processing Systems* (NIPS 2017).

Étape 2, consolidation

Cette étape consiste à munir le langageur de garde-fous. Le modèle ne doit pas reproduire les biais éventuellement présents dans les données (racisme, sexisme, etc.) ni permettre des usages dangereux (fabrication d'armes). L'apprentissage d'un bon contrôle a recours à l'expertise humaine : les données d'apprentissage du contrôle sont obtenues en demandant à des travailleurs du Web d'étiqueter des énoncés comme admissibles ou litigieux (la méthode utilisée est celle de l'apprentissage dit par renforcement avec feedback humain, *reinforcement learning with human feedback,* RLHF).

Étape 3, ouverture

La troisième étape consiste à réaliser une plate-forme permettant au langageur de dialoguer avec des millions d'êtres humains, et d'enregistrer les dialogues à des fins d'apprentissage en continu du modèle. La diffusion de ChatGPT3 a été beaucoup plus rapide que prévu par ses concepteurs (deux mois pour atteindre 100 millions d'utilisateurs – le système est, bien sûr, multilingue).

Les ressources nécessaires sont considérables

La mise à disposition auprès du grand public d'un agent conversationnel, et son appropriation par le plus grand nombre, nécessite de pousser les techniques relevant de l'informatique vers leurs limites.

En ce qui concerne la première étape, il faut disposer de données d'apprentissage en quantité et en diversité suffisantes (comprenant par exemple tout Wikipédia, soit plusieurs téraoctets). Le modèle cherché doit aussi être suffisamment complexe : expérimentalement, une bonne qualité de réponses demande un modèle d'au moins 60 milliards de paramètres. Actuellement, ChatGPT3 (OpenAI) ou Bloom (HuggingFace, entreprise franco-américaine) comprennent 175 milliards de paramètres pour l'un et 176 milliards pour le second ; Enrie (Baidu, Chine) en comprend 260 milliards, PaLM (Google) 540 milliards ; GPT4 (OpenAI) comprendrait 1 000 milliards de paramètres.

L'entraînement d'un langageur requiert l'utilisation de supercalculateurs capables de fonctionner plusieurs semaines, voire plusieurs mois, sans interruption. Ce type d'ordinateur comprend des centaines, voire des milliers, d'accélérateurs de calcul (GPU – *Graphic processor unit)* répartis en nœuds de traitement interconnectés indépendants afin d'exploiter le parallélisme inhérent des tâches. Le fait de conduire à bien l'entraînement du modèle (représentation et stockage des données, parallélisation des calculs et des architectures) nécessite donc d'une part des compétences en programmation extrêmement pointues et, d'autre part, de larges équipes de chercheur.e.s. et d'ingénieur.e.s (500 personnes pour GPT4) avec une organisation très efficace.

Le coût énergétique de l'entraînement est actuellement de l'ordre de 1 GWh.

L'étape de consolidation requiert une très importante intervention humaine, allant de l'annotation réalisée à bas coût dans des pays en voie de développement, à l'évaluation experte de la

plausibilité des réponses et de la présence de biais. Globalement, le coût de la consolidation se chiffre en centaines de millions de dollars.[5]

La troisième étape, la réalisation d'un service en ligne pour le grand public, demande également des ressources considérables. La quantité de ressources pour pouvoir dialoguer quasiment en temps-réel avec des milliers d'utilisateurs simultanés devient alors largement supérieure à celle nécessaire à la phase d'entraînement.

En termes de consommation énergétique, les dialogues avec ChatGPT3 ont nécessité selon les hypothèses entre 1 et 20 GWh pour le seul mois de janvier 2023, soit une consommation annuelle pouvant atteindre la centaine de GWh (sans compter la fabrication des équipements). L'usage massif d'un agent conversationnel demande une énergie encore plus importante que sa construction. Le fait d'ajouter un langageur aux moteurs de recherche existants augmenterait alors la consommation électrique mondiale de façon considérable au-delà du TWh[6]. Ce point doit être analysé de façon plus approfondie, et mérite une extrême vigilance.

Plus de détails sur les aspects financiers sont disponibles dans ([7]).

5 Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. Billy Perrigo *TIME Magazine*, January, 18 2023

6 À titre de comparaison, les deux EPR de Taishan fournissent au réseau électrique chinois jusqu'à 24 TWh d'électricité par an (https://www.edf.fr/groupe-edf/espaces-dedies/journalistes/tous-les-communiques-de-presse/le-premier-des-deux-epr-de-la-centrale-nucleaire-de-taishan-en-chine-entre-en-exploitation-commerciale). Pour sa part, le minage du bitcoin consomme de l'ordre de 110TWh/an, ce qui est bien pire. Toutefois, l'explosion de consommation porte actuellement plus sur les IA génératives que sur le bitcoin.

7 The Inference Cost Of Search Disruption. Large Language Model Cost Analysis. https://www.semianalysis.com/p/the-inference-cost-of-search-disruption

Étude de cas : usages et limites de ChatGPT

ChatGPT constitue une rupture dans le déploiement de l'IA. Cette rupture repose autant sur l'emploi de supercalculateurs et l'ouverture au grand public (permise par l'étape de la consolidation et le passage à l'échelle de la plate-forme) que sur les avancées du langageur. Un nombre considérable de dialogues ont été générés et enregistrés pour des usages parfois imprévus par quelques centaines de millions d'utilisateurs ayant dialogué avec ChatGPT depuis novembre 2022. Ils pourront être utilisés pour ré-entraîner et améliorer le système.

Cependant, les impacts économiques et sociétaux des usages de ChatGPT soulèvent des questions épineuses : qualité et impartialité des réponses ? droit d'auteur par rapport aux sources utilisées ? interdiction des textes générés par ChatGPT dans des contextes critiques (lesquels) ? capacité de détection automatique des textes générés ? etc.

L'analyse de ChatGPT3 est structurée selon le canevas classique : « forces, faiblesses, opportunités, menaces ».

Forces

– Les moteurs de recherche classiques proposent les pages où trouver la réponse à une question posée ; ChatGPT propose un texte de synthèse, répondant directement à la question posée. La longueur du texte ainsi que le niveau de langue peuvent être ajustés selon le souhait de l'utilisateur. L'échange tient également compte des phrases précédentes au cours du dialogue (session). Ceci peut permettre de franchir une nouvelle étape dans la démocratisation des savoirs liée au numérique. Par exemple, il est possible d'explorer un domaine en demandant une synthèse des documents disponibles sur ce sujet, en affinant les questions posées, en

changeant de point de vue ou de contexte.

- ChatGPT dispose d'un certain contrôle sur certaines de ses réponses, lui permettant de dire « Ceci m'est interdit » (en réponse à des demandes portant sur la fabrication d'armes, émettre des propos injurieux).
- ChatGPT est une IA-G, capable de générer des entités « à la manière » des données d'entraînement. Dans le contexte de la création, cette capacité créative est appréciée.

En résumé, la prouesse technique et son appropriation par un large public débouchent essentiellement sur la multiplication des usages possibles, dont le nombre continue d'augmenter.

Faiblesses

- Les réponses émises par ChatGPT ne se fondent pas sur la vérité, la logique ou le calcul, mais sur les statistiques. De fait, ChatGPT émet des réponses plausibles et rapides, mais non vérifiées ; ceci est caractéristique de ce que Daniel Kahneman, prix Nobel d'économie, appelle un « System 1 »[8].
- Le système souffre « d'hallucinations » (création de réponses de toute pièce), d'incohérences et d'indéterminismes (notamment entre les réponses fournies au cours d'un dialogue).
- Ces défauts sont d'autant plus graves qu'il s'agit de sujets pointus sur lesquels les données sont rares ou présentant des risques d'homonymie (demandez à ChatGPT de rédiger la biographie d'une personne peu présente sur internet ou dont le nom est assez commun).
- Par construction, les hallucinations générées sont similaires aux vraies données : à l'interlocuteur humain dedistinguer (parfois difficilement) le vrai de l'inventé.
- D'énormes efforts de consolidation (par RLHF) sont nécessaires en continu, pour améliorer la qualité des réponses et bloquer les interactions indésirées.
- La définition des interactions indésirées dépend entièrement de l'entreprise propriétaire du langageur, tout comme la « couleur » politique, économique, ou philosophique des réponses du langageur.
- Le contrôle de biais, d'excès de langage, d'interdictions peut être contourné en formulant ses demandes de manière subtile.
- Le corpus d'entraînement de ChatGPT ne respecte pas le règlement général sur la protection des données (RGPD), ce qui a conduit la Cnil italienne à limiter provisoirement son usage. Par ailleurs, la compatibilité des textes générés avec les législations en vigueur n'est pas établie, notamment en ce qui concerne le droit d'auteur et la propriété intellectuelle.

8 *Thinking fast and slow*, D. Kahneman, 2011.

Opportunités

On peut s'attendre à un foisonnement d'innovations fondées sur ChatGPT.

- ChatGPT fonctionne dans une boucle d'interaction : question/réponse/correction. Il permet d'accélérer considérablement la production de textes dans des domaines comme le marketing, la vente ou la relation avec des clients et prospects. Parce que sa réponse est souvent relativement exhaustive et structurée, ChatGPT est un bon outil pour défricher un sujet : il propose des pistes et l'utilisateur.e décide, ou non, de les approfondir avec de nouvelles questions. Les briques élémentaires de ChatGPT (langageur et RLHF) vont être intégrées dans les robots conversationnels (« chatbots ») et en améliorer l'efficacité en particulier dans la qualité du texte produit.
- Notons que le langageur permet aussi de générer des textes en langage non naturel, tels les langages de programmation, les documents structurés (tableurs, présentations, tableaux de visualisation de données) ou les nomenclatures industrielles. Ces outils vont arriver en 2023 dans les offres des acteurs tels que Microsoft, et ils vont permettre une augmentation significative de productivité. L'utilisation de GitHub copilot (un assistant de génération de code construit autour de GPT 3.5) démontre une vraie accélération de productivité lorsqu'il est utilisé par un programmeur compétent. Dans le même ordre d'idées, ChatGPT peut être utilisé pour faire communiquer des logiciels indépendants et réaliser des intégrations.
- ChatGPT peut être enrichi de capacités de raisonnement, améliorant sa cohérence et sa crédibilité et progressant vers une certaine notion de sens commun. GPT4 avancerait dans ce sens ; cependant l'hybridation des méthodes statistiques et des méthodes de preuve logique ou mathématique est un problème difficile, pour lequel une première proposition existe toutefois[9].

Menaces

- Une première menace vient de la facilité d'usage de ChatGPT, qui pourrait devenir un oracle fournissant une réponse par défaut à toute requête.
- Dans le cas où la vérité de la réponse importe, la menace consiste à créer des croyances arbitraires ou à pousser le demandeur à des actions inappropriée. La menace est de même nature que celle des *fake news*, mais les erreurs risqueraient d'être plus systématiques[10].

9 Voir l'annonce ChatGPT gets its "Wolfram superpower" par l'éditeur de logiciel scientifique Wolfram. Ce plugin permet de connecter chatGPT à Wolfram | Alpha, un outil de calcul formel mathématique doté d'un langage. Il reste à examiner de plus près les nouvelles possibilités offertes par cet assemblage. Voir aussi Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, un travail de Google.

10 La puissante association des Informaticiens américains *(Association for Computing Machinery)* et l'éditeur scientifique Cambridge University Press ont décidé de restreindre fortement l 'usage d'agents conversationnels.

- Dans les cas où l'objectif concerne plus une première ébauche et une mise en forme d'un texte (rédaction d'un état des lieux, écriture d'un fragment d'algorithme, expression d'idées personnelles) ChatGPT peut être très utile. Cependant, compte tenu des usages possibles, l'emploi des « travailleurs du savoir » (*knowledge workers*) risque à tout le moins d'être transformé. Une autre menace possible concerne les modes actuels d'évaluation des élèves ou étudiant.e.s. Ces types d'évaluation devraient être réexaminés de manière profonde pour pouvoir mesurer l'apport de l'élève par rapport à une réponse fournie par langageur.
- Enfin, une menace essentielle est que de tels systèmes sont excessivement énergivores tant par leurs usages que par la fabrication des infrastructures requises.

La position française et européenne

Les entreprises de premier plan en Intelligence artificielle (GAFAM, Nvidia, OpenAI) ont en général développé un langageur propriétaire, d'usage interne ou payant

Au moment de la finalisation de cet avis (avril 2023), l'entreprise la plus en pointe est OpenAI, créateur de ChatGPT. Fondée en 2015, OpenAI est soutenue par Microsoft (à hauteur d'un milliard en 2019 et de plusieurs milliards à partir de 2023). Microsoft voit dans le couplage avec ChatGPT une manière de rattraper le retard de son moteur de recherche Bing sur celui de Google.

En France, la société franco-americaine HuggingFace, PME de 150 personnes fondée par trois ingénieurs français, a créé en 2022 le projet BigScience qui regroupe une centaine d'institutions publiques et privées. Ce projet a conduit au langageur nommé Bloom (176 milliards de paramètres, 46 langues) en s'appuyant sur un supercalculateur du GENCI (Grand équipement national de calcul intensif) ; de plus, Bloom est ouvert, c'est-à-dire qu'il est utilisable par tous et ses biais peuvent être inspectés[11]. En France encore, l'entreprise LightOn déploie son langageur sur le marché des entreprises.

En Europe, la start-up AlephAlfa (Allemagne) dispose d'un modèle à cinq langues déjà commercialisé. La société Stability AI (UK) propose comme HuggingFace un modèle en logiciel libre.

En résumé, les compétences existent en Europe pour participer au meilleur niveau aux avancées scientifiques et technologiques liées aux agents conversationnels intelligents (étape 1). L'accès aux moyens de calcul est facilité notamment par l'initiative européenne EuroHPC qui permet à la recherche publique et privée en Europe d'avoir accès aux plus puissants des supercalculateurs.

11 https://huggingface.co/docs/transformers/model_doc/bloom

L'ingrédient manquant paraît être l'absence des capitaux nécessaires aux étapes 2 et 3 : consolidation et mise à disposition publique du langageur. De fait aux États-Unis, le capital-risque finançait jusqu'à présent assez peu l'IA générative. La situation est en train de changer avec un marché américain en plein essor.

Recommandations

L'Académie des technologies recommande d'anticiper les effets économiques et sociétaux des produits et services créés par les GAFAM : par expérience, le fait de remédier après coup à leurs conséquences indésirables s'avère difficile. En ce qui concerne les agents conversationnels intelligents, deux recommandations majeures ont été identifiées.

Soulignons que la régulation recommandée ne doit pas retarder **le développement de langageurs par les Européens. Elle doit au contraire les faciliter en visant notamment la création de biens communs, face à des pays traditionnellement plus réticents vis-à-vis des régulations.**

Création de langageurs libres et de confiance : exercer le levier européen

Un langageur reflète par construction les croyances de ses constructeurs, par le choix du corpus d'entraînement ou de ré-entraînement. Il est donc crucial dans un objectif de souveraineté au niveau des valeurs, de l'économie et de la recherche, de disposer de langageurs **libres**. L'exemple de la société HuggingFace démontre la faisabilité d'une telle réalisation.

L'action au niveau européen (alliant acteurs publics et privés, industriels et chercheurs, États et citoyens) est nécessaire pour répondre aux défis, regardant :

- La mise à disposition de langageurs pour le grand public et les acteurs industriels (étapes 2 et 3 de la construction des langageurs ; mutualisation d'équipements et/ou financement d'infrastructures).
- L'évaluation de la conformité d'un langageur (avec l'agilité requise).
- Les avancées scientifiques touchant à l'hybridation d'un langageur statistique et des méthodes de preuve logique ou mathématique.

– Les recherches permettant le développement d'outils de traçabilité permettant d'identifier un langageur comme étant le créateur d'un texte, outils pour lesquels nous proposons le terme de ***langageur inverse***[12]***.***

L'orientation des investisseurs vers le financement d'infrastructures de mise en ligne de langageurs doit être une priorité. Cela pourrait couvrir, tant des initiatives de portée générale que des développements plus particulièrement dédiés à des métiers, filières, ou administrations.

Enfin, l'organisation de l'accès à des langageurs libres par le plus grand nombre permettrait de contribuer au test et à l'amélioration de ce bien commun tout en préparant aux évolutions professionnelles et culturelles suscitées par cette technologie. En ce qui concerne le monde de l'Éducation, on pourrait aussi en attendre une formation très large des élèves aux langageurs et leur usage raisonné (ce ne sont pas des oracles). Du point de vue des éducateurs, ce serait aussi l'occasion de développer de nouvelles méthodes d'enseignement et des pratiques d'une *évaluation* des élèves rendue très difficile.

CRÉER UN CENTRE D'EXPERTISE SUR LA RÉGULATION DES LANGAGEURS

L'Académie note que plusieurs acteurs majeurs en IA appellent maintenant à une régulation du domaine[13]. Compte tenu de l'importance durable de cette technologie et de ses enjeux économiques et sociétaux, elle recommande :

– de définir légalement les responsabilités des offreurs[14] et des utilisateurs (en évitant les redondances avec les régulations déjà en place ou en cours de définition et qui incluent – implicitement – les langageurs dans leur domaine[15]) ;
– de mettre en place les moyens d'évaluer leur conformité aux contraintes et réglementations

12 Il s'agit de développer l'analogue de l'identification du locuteur en reconnaissance du langage parlé, par le biais de l'analyse statistique de caractéristiques du son émis. Dans le contexte des langageurs, il pourrait s'agir d'une analyse de caractéristiques appropriées. Ce thème du langageur inverse nous paraît important. Voir à ce sujet l'annonce https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text d'OpenAI qui déclare développer un tel outil en réponse à une pression des enseignants de certains états américains. D'autres approches de traçabilité devraient aussi être explorées.

13 Voir deux annonces de dernière minute en date de publication de ce texte : une lettre ouverte (Pause Giant AI Experiments: An Open Letter) de scientifiques et industriels de renom, et des dépôts de plainte entreprises par diverses administrations ou États.

14 S'agissant des offreurs, notre proposition s'aligne sur les principes adoptés par l'EU dans son action de régulation des réseaux sociaux, définissant un ensemble de contraintes auxquelles ceux-ci doivent se conformer.

15 Voir à ce sujet l'extrait suivant du texte suivant publié par le Peren : Dans son orientation générale, le Conseil de l'Union européenne propose d'ajouter une définition des *« systèmes d'IA à usage général »* au projet de règlement. Cette définition inclurait les LLM, et a fortiori les LLM conversationnels. Les fournisseurs de tels systèmes seraient soumis à des obligations particulières, distinctes des obligations applicables aux systèmes à haut risque. Le Parlement européen semble poursuivre ses travaux dans cette direction, en distinguant toutefois les systèmes d'IA à usage général reposant sur des « modèles de fondation » (ou Foundation models, c'est-à-dire tout modèle entraîné sur de larges volumes de données qui peut servir à un grand nombre de tâches) des systèmes reposant sur des modèles plus simples.

ainsi définies.

Pour cela, il y a besoin de créer un corpus de savoir-faire généraliste sur ce secteur (les aspects métiers seront réglés par les métiers après adaptation, et dans le cadre de leurs processus). Ainsi, on sera en mesure de formuler des demandes pertinentes et opératoires au regard des offreurs, avec une constante de temps faible. Par exemple, on pourra obliger les langageurs à produire leurs indices de confiance en accompagnement de chaque réponse ; on pourra exiger des offreurs qu'ils accompagnent un langageur de sa fonction de langageur inverse (cf. la recommandation précédente).

Il n'est pas dans notre propos de formuler une recommandation précise ni pour ce qui concerne l'instrument approprié pour cela, ni quant à son positionnement (national et/ou EU). Cet instrument devrait contribuer utilement à rendre plus techniquement étayés les travaux autour de l'IA Act.

Au stade actuel, l'Académie se limite à noter les points suivants :

- l'instrument qui doit servir de modèle et à terme de cible, est l'ANSSI (dédiée à la cybersécurité), et son domaine doit être à terme celui de l'IA anthropomorphique.
- les compétences requises sont larges (Droit, Histoire, Informatique, Économie, Linguistique, Littérature, Psychologie, Sociologie).
- le contexte institutionnel pour un tel instrument n'est pas vide (ainsi, s'agissant de la langue, on trouve la Cité internationale de la langue française et de la francophonie).

Académie des technologies
Le Ponant – Bâtiment A
19, rue Leblanc
75015 PARIS
+33(0)1 53 85 44 44
secretariat@academie-technologies.fr
www.academie-technologies.fr

ISBN : 979-10-97579-45-6.