HUB FRANCE IA

CHOISIR UN MODELE D'IA GENERATIVE POUR SON ORGANISATION

Juin 2024

TABLE DES MATIERES

INTRODUCTION

Depuis le succès rencontré par ChatGPT fin 2022, les entreprises se posent la question de l'appropriation de l'IA générative. Au cours des mois écoulés, de nombreux cas d'usage ont émergé. Pour en savoir plus, n'hésitez pas à consulter le rapport du Hub France IA sur ce sujet[1] publié en janvier 2024. Il y a également eu une augmentation du nombre de fournisseurs de modèles de langage, avec des solutions aussi bien en open source que propriétaires, et une adaptation des plateformes cloud qui proposent ces modèles. Se pose alors la question pour les entreprises de comment choisir parmi tous ces fournisseurs. Et c'est justement pour répondre à cette question que le Hub France IA a décidé de lancer fin 2023 un groupe de travail sur cet aspect.

Ce groupe de travail, composé d'une quinzaine de membres du Hub France IA s'est réuni de façon hebdomadaire pour traiter la question. L'objectif du groupe était de fournir un livrable écrit permettant d'éclairer le sujet du choix des modèles de type « *Large Language Models* » (LLM) dans les organisations (entreprises, collectivités, …). Pour rappel, on parle de modèles de langage de grande taille ou « *Large Language Models* » pour les modèles possédant un grand nombre de paramètres (généralement de l'ordre de plusieurs milliards de paramètres ou poids). C'est ce type de modèle, qui a été popularisé par OpenAI via le déploiement de ChatGPT.

Afin de répondre au plus proche des attentes des organisations, nous avons commencé par réaliser une enquête auprès d'elles. Cette enquête a été relayée sur LinkedIn et auprès des membres du Hub France IA. Elle nous a permis de comprendre les critères de choix les plus importants pour les organisations et d'orienter les travaux du groupe. La méthodologie utilisée pour réaliser cette enquête ainsi que ses résultats font l'objet de la première partie de ce livrable.

Parallèlement à la réalisation de cette enquête, la difficulté de maintenir à jour une analyse des performances des modèles a vite été identifiée. En effet, il y a une évolution très rapide dans ce domaine. Notre attention s'est donc plutôt portée sur l'analyse et le décryptage des comparatifs de performances (dits « *benchmarks* »)

[1] Hub France IA. Les usages de l'IA générative. Janvier 2024. https://www.hub-franceia.fr/wp-content/uploads/2024/02/Livre-blanc_Les-usages-de-lia-generative-01.2024.pdf

existants. Ce sujet est traité dans la deuxième partie de ce document et inclut des liens vers les différents *benchmarks* mentionnés.

Ensuite, nous avons constaté que ces *benchmarks* ne permettaient pas de comparer tous les critères importants identifiés dans l'enquête auprès des organisations. D'autres aspects importants pour le choix étaient mentionnés : aspects juridiques, financiers, infrastructure, … Or, ces éléments ne sont pas toujours facilement accessibles pour les différents fournisseurs de LLM. Nous avons donc, dans le cadre du groupe de travail, pris contact avec les principaux fournisseurs de LLM en France pour les collecter. Nous présentons dans la troisième partie la méthodologie employée, nous décrivons les critères investigués et nous présentons les différents acteurs que nous avons contactés.

Enfin, la quatrième et dernière partie comporte l'analyse détaillée des réponses qui nous ont été fournies par les différents acteurs contactés. Nous avons ordonné les réponses suivant les thématiques majeures identifiées lors de l'enquête préliminaire auprès des organisations. Nous fournissons aussi dans cette partie des liens utiles qui vous permettront de vous rendre sur les pages web ou documents pertinents des fournisseurs de modèle si vous souhaitez creuser certains aspects plus en détails.

HUB FRANCE IA

PARTIE 1 Les critères de choix de LLM dans les organisations

CHOISIR UN MODELE D'IA GENERATIVE POUR SON ORGANISATION

Juin 2024

PARTIE 1 : Les critères de choix de LLM dans les organisations

Il était difficile pour nous d'imaginer sélectionner a priori les critères que nous jugions importants pour les organisations. Nous avons donc décidé très tôt au sein du groupe de travail de lancer une enquête auprès de ces dernières. Afin de viser un panel assez large, nous avons relayé cette enquête sur LinkedIn et au sein des canaux liant les membres du Hub France IA. Ceci nous a permis de récupérer plus de 60 réponses provenant d'interlocuteurs variés. Dans cette partie, nous présentons la méthodologie puis les résultats de l'enquête réalisée. Enfin, nous abordons les questions importantes à se poser pour choisir un fournisseur LLM pour son entreprise.

PARTIE 1.1 - Méthodologie de l'enquête auprès des entreprises

Afin de maximiser le nombre de répondants à l'enquête, nous nous sommes restreints à quelques questions essentielles. Cela a permis de rendre l'enquête assez courte avec des réponses possibles en moins de 5 minutes. Les questions posées étaient les suivantes :

- Pour vous, quels sont les 5 critères principaux dans le choix de modèles d'Intelligence Artificielle Générative ?
- Pour vous, quels sont les 5 modèles qui doivent absolument être traités dans notre livrable ?
- Quel est le nombre d'employés de votre entreprise ?
- Quel est le secteur de votre entreprise ?
- Indiquez enfin ici des attentes particulières ou des remarques par rapport au livrable de notre groupe de travail si vous en avez.
- Si vous souhaitez recevoir le livrable par mail, quel est votre mail ?

Une question clé pour nous est celle sur les critères de choix des modèles. En effet, les entreprises ont des contextes et des problématiques spécifiques qui se traduisent par des critères de choix aussi bien pour le modèle choisi que pour l'architecture informatique qui permet de le faire fonctionner.

Afin de guider les réponses nous avons suggéré une liste de critères possibles dans différents domaines (juridique, financier, écologique, …). Voici l'ensemble des critères proposés :

- Conformité légale et réglementaire : données personnelles (ex : RGPD, …), … ;
- Niveau de transparence sur les données d'entraînement ;
- Sécurité des données (flux de données prompts/réponses, données d'entraînement du modèle) ;
- Intégration avec l'infrastructure existante : API, On Premise, cloud (SecNumCloud), … ;
- Open source ou propriétaire ;
- Scalabilité : niveau de scalabilité et coûts associés ;
- Coûts initiaux, de maintenance, de traitement et d'évolution ;
- Risques légaux et financiers (ex : utilisation de données copyrightées, …)
- Facilité d'utilisation : présence d'une interface utilisateur, nécessité d'un développeur ;
- Pérennité de la solution (investisseurs, levées de fond, années d'existence, …) ;
- Options de fine tuning ;
- Support : disponibilité, coût et qualité du support technique ;
- Formation : modalités de formations au modèle ;
- Présence d'une communauté active ;
- Services d'accompagnement pour l'installation et le déploiement ;
- Interopérabilité avec d'autres systèmes ;
- Mises à jour et fréquence d'évolution du modèle ;
- Nombre maximal de tokens pour les prompts/réponses ;
- Nombre maximum de requêtes par jour ;
- Durabilité et considérations écologiques (énergie, eau, …) ;
- Spécialisation du modèle sur un domaine particulier (finance, légal, …) ;
- Multimodalité ;
- Plan de continuité d'activité et de secours (en cas de problème) ;
- Stockage et export des conversations.

Il n'y a pas dans ces critères les notions de performance, de qualité, de taux d'hallucinations. En effet, ces critères sont importants pour les entreprises mais sont déjà très largement couverts par plusieurs *benchmarks* (ou « comparatifs »). De plus, une analyse des performances des modèles nécessite une mise à jour très régulière. Nous avons plutôt centré nos efforts sur les critères de choix les plus stables. En revanche, comme nous le verrons dans la partie 2, nous donnons des clés de lecture des *benchmarks* de référence et tentons d'expliquer comment tester les performances des modèles sur ces aspects.

PARTIE 1.2 - Résultats de l'enquête

Suite à la collecte de plus de 60 réponses, nous avons analysé les résultats de l'enquête. Nous avons commencé par regarder plus en détails les profils des répondants. Ceci nous a permis de constater qu'il y a une forte représentation des entreprises de moins de 10 salariés et des entreprises de 10 à 200 salariés (Figure 1). Il y a également un peu plus d'une dizaine de personnes issues de groupes de plus de 10 000 salariés. Les résultats issus de cette enquête représentent donc une large variété d'entreprises.

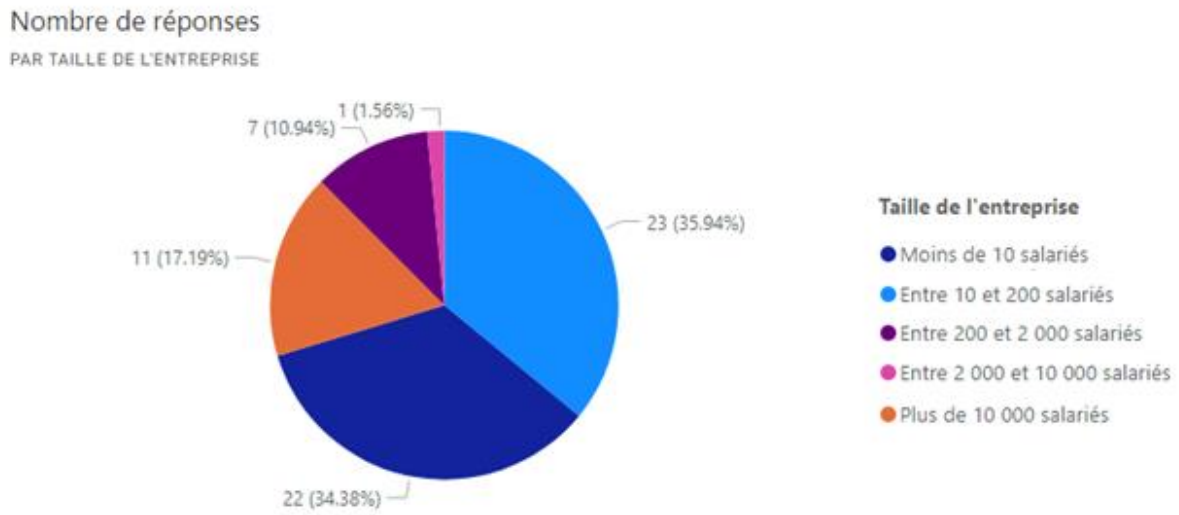

Figure 1 : Nombre de réponses en fonction de la taille de l'entreprise du répondant

Les secteurs d'activité des entreprises qui se sont exprimées sont également très variés. Les secteurs les plus représentés sont liés à l'information, la communication, le commerce, la gestion, le management, l'enseignement, la formation, l'industrie, le droit, la santé et l'agroalimentaire.

Détaillons maintenant les résultats obtenus lors de l'analyse des résultats. Certains critères ont été largement plus plébiscités que les autres (Figure 2). En particulier, tous les aspects concernant la sécurité des données et la conformité légale et réglementaire ont recueilli beaucoup de suffrages. Et ceci reste vrai indépendamment de la taille de l'entreprise du répondant.

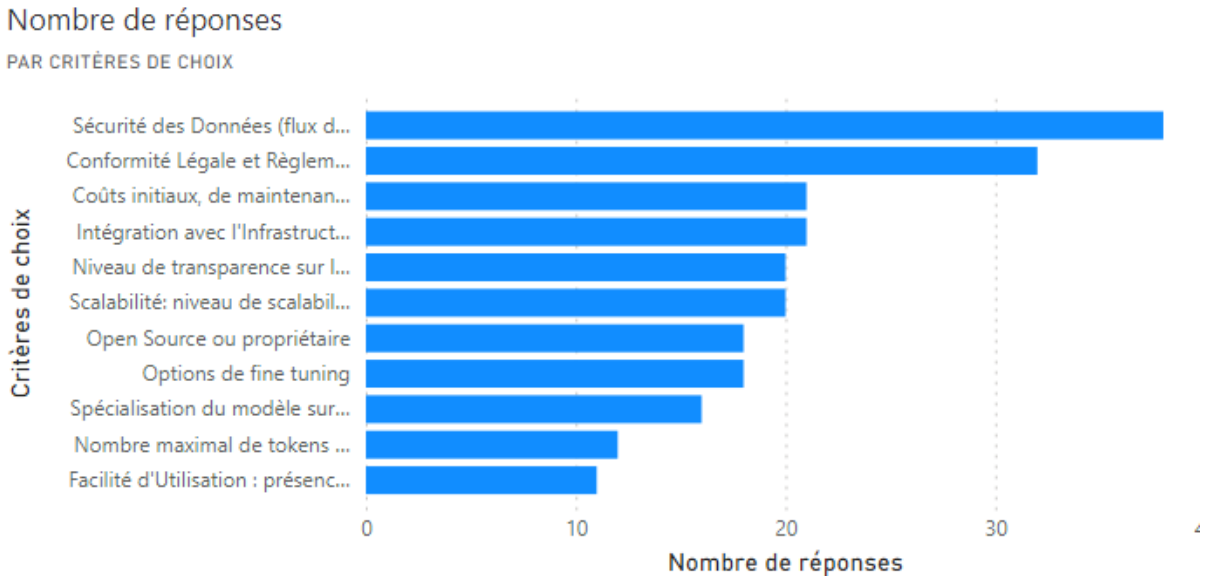

Figure 2 : Liste des critères présentant le plus grand nombre de répondants

Le tableau ci-dessous regroupe ces critères selon leur ordre d'importance pour les répondants.

Critères	Pourcentage de citation
• Conformité Légale et Réglementaire : Données personnelles (ex : RGPD, …), … • Sécurité des Données (flux de données prompts/réponses, données d'entraînement du modèle)	Plus de 50%
• Intégration avec l'Infrastructure existante : API, On Premise, cloud (SecNumCloud), … • Coûts initiaux, de maintenance, de traitement et d'évolution	Entre 30% et 40%
• Niveau de transparence sur les données d'entraînement • Open Source ou propriétaire • Scalabilité : niveau de scalabilité et coûts associés • Options de fine tuning • Spécialisation du modèle sur un domaine particulier (finance, légal, …)	Entre 20% et 30%
• Risques légaux et financiers (ex : Utilisation de données copyrightées, …) • Facilité d'Utilisation : présence d'une interface utilisateur, nécessité développeur • Interopérabilité avec d'Autres Systèmes • Nombre maximal de tokens pour les prompts/réponses	Entre 10% et 20%

• Durabilité et Considérations Écologiques (énergie, eau, …) • Multimodalité • Mises à Jour et fréquence d'évolution du modèle	
• Pérennité de la solution (investisseurs, levées de fond, années d'existence, …) • Support : disponibilité, coût et qualité du support technique • Formation : modalités de formations au modèle • Présence d'une communauté active • Services d'accompagnement pour l'installation et le déploiement • Nombre maximum de requêtes par jour • Plan de Continuité d'Activité et de Secours (en cas de problème) • Stockage et export des conversations	Moins de 10%

Les résultats de l'enquête provenant de plus de 60 organisations nous ont confirmé l'intérêt de ces dernières pour des critères non directement liés à la performance. De plus, nous avons constaté que certains critères essentiels (sécurité, juridique par exemple) ne sont pas toujours présents ou facilement accessibles sur les sites web des différents fournisseurs.

PARTIE 1.3 - Questions à se poser pour choisir un modèle

Nous souhaitons à travers ce document parler à toutes les organisations qui souhaitent intégrer l'IA générative à leur activité. Pour ce faire, nous avons, au-delà de l'enquête réalisée dont nous venons de présenter les résultats, analysé la situation actuelle. Nous la résumons dans cette partie, tout en essayant de vous aider à identifier les questions clés à vous poser en fonction de vos cas d'usage.

Dans une étude réalisée par BPI France[2] au cours du premier trimestre 2024 auprès de 3 077 dirigeants de TPE et PME françaises, il a été constaté que seuls 3 % ont fait un usage régulier de l'IAG et 12% un usage occasionnel. Ceci exprime la difficulté que peuvent avoir des organisations à trouver les usages et le mode opératoire pertinent pour intégrer ces technologies à leurs activités.

[2] BPI France. Enquête BPI France. L'IA Générative dans les TPE et PME. 14 mars 2024. https://lelab.bpifrance.fr/Etudes/ia-generatives-opportunites-et-usages-dans-les-tpe-et-pme

Rappelons, que Hub France IA, dans son livre blanc sur les usages de l'IA générative[3], a décrit les cas d'usages qui se développent très rapidement autour des six grands domaines que sont la cybersécurité, les industries culturelles et créatives, les ressources humaines, le développement informatique, l'éducation et le marketing. Dans le livrable que nous vous proposons aujourd'hui, nous répondons plutôt aux questionnements concernant le choix du modèle et du fournisseur associé quel que soit le cas d'usage choisi.

En effet, pour choisir un modèle d'IAG (Intelligence Artificielle Générative), il est nécessaire de se poser un certain nombre de questions qui pourront faire émerger des critères de choix. Tout d'abord, il s'agit d'identifier les cas d'usage auxquels l'IAG peut répondre en interne. Il est important d'analyser non seulement le besoin mais aussi sa finalité pour pouvoir choisir le modèle adéquat. Cela permet notamment de s'assurer que le recours à l'IAG est utile et qu'il n'existe pas de briques plus simples et plus pertinentes à mettre en œuvre dans ce cas.

Selon la finalité du cas d'usage et les données utilisées, plusieurs questions seront à se poser quant au choix des modèles pertinents. Sachant que le modèle choisi pour un cas d'usage donné peut associer des catégories de données en entrée comme du texte, du son, de la vidéo, de l'image, les données d'entraînement du modèle vont s'appuyer sur des données supplémentaires, complémentaires qui peuvent être conservées. Il faudra veiller à analyser si des données personnelles, confidentielles ou sensibles seront utilisés dans le cas d'usage. Dans ce cas, il faudra veiller à choisir les modèles permettant de garantir un maximum de sécurité et une gestion adéquate de ces données.

Un autre élément important à prendre compte est la fiabilité du modèle. En effet, les modèles IAG peuvent être sujet à des « hallucinations », c'est-à-dire qu'ils peuvent fournir des réponses factuellement fausses mais qui semblent plausibles. Ce mécanisme est de plus en plus testé sur les différents modèles. Mais il y a un degré de criticité différent des hallucinations selon que le modèle IAG est utilisé pour générer des prototypes de mail qui seront revus et corrigés ou s'il est utilisé pour faire des propositions commerciales à des clients.

[3] Hub France IA. Les usages de l'IA générative. Janvier 2024. https://www.hub-franceia.fr/wp-content/uploads/2024/02/Livre-blanc_Les-usages-de-lia-generative-01.2024.pdf

Nous retrouvons tous ces questionnements, ainsi que d'autres concernant les besoins d'accompagnement, les coûts, et bien d'autres dans les résultats de l'enquête menée auprès des 60 répondants. Nous avons regroupé dans le tableau ci-dessous quelques-unes des questions qui nous semblent importantes à se poser lors de l'identification des cas d'usages mobilisant des modèles IAG.

Questions à se poser	Pourquoi est-ce important
Quel est le cas d'usage ? Quelle est la complexité de la tâche ?	Déterminer le but principal du cas d'usage et valider la pertinence d'un modèle IAG en en définissant les caractéristiques.
Le cas d'usage utilise-t-il des données à caractère personnel, des données sensibles ou des données confidentielles ?	S'assurer de la bonne gestion des données. Pour cela, il faudra veiller particulièrement aux critères liés à la sureté et la sécurité des données mais aussi aux aspects juridiques et légaux.
Quel est l'impact du cas d'usage ? Quel degré de fiabilité doit-on obtenir ?	Déterminer si le cas d'usage va être utilisé en interne ou exposé en externe et l'impact qu'il peut avoir. Il faudra se concentrer sur les aspects de fiabilité et évaluer les modèles au travers des *benchmarks* présentés en partie 2 mais aussi éventuellement par des tests plus poussés.
Quel est le degré de réactivité demandé au modèle ? Utilisation en direct, ou bien en asynchrone ?	Déterminer le type de modèles à utiliser ou l'infrastructure nécessaire. Pour des cas avec un fort besoin de réactivité, une infrastructure performante ou un modèle plus léger pourront permettre ce type de fonctionnement. Pour un fonctionnement en asynchrone, il sera possible d'utiliser des modèles plus lourds ou une infrastructure moins performante, permettant aussi de réduire le coût énergétique.
Quel budget pour mettre en place le cas d'usage ?	Déterminer l'infrastructure et le type de modèles à utiliser. En effet, les modèles open source ou propriétaires tels qu'ils seront présentés dans la suite du livrable ont chacun leurs avantages, leurs inconvénients et leurs propres

	coûts. Il faudra donc identifier le budget potentiel pour calibrer les modèles adéquats et l'architecture informatique associée.
Quel est le profil des personnes qui implémenteront et/ou utiliseront le cas d'usage ?	Déterminer les besoins en formation, support et interface nécessaires pour le cas d'usage. En effet, si les personnes en charge ne sont pas familières des modèles IAG, il pourra être intéressant de se baser sur les critères d'accompagnement ou sur la présence d'une interface pour faciliter l'implémentation et le déploiement du cas d'usage.

La suite de ce livrable permettra d'éclairer les différents critères mis en lumière dans cette première partie.

HUB FRANCE IA

PARTIE 2 Les benchmarks existants

PARTIE 2 - Les *benchmarks* existants

Lorsqu'il faut choisir un LLM, le grand nombre de *benchmarks* existants peut rendre la tâche d'évaluation des performances complexe. Comprendre clairement ces *benchmarks* est crucial, car ils orientent vers le LLM le plus adapté à vos besoins spécifiques. Cette section sert de guide pour naviguer dans le paysage varié des *benchmarks*, offrant un éclairage sur leur rôle essentiel dans l'évaluation actualisée des capacités et performances des modèles. Nous aborderons l'importance de choisir le bon *benchmark* en fonction de l'application visée et comment cette décision peut influencer l'efficacité et la réussite de vos projets d'intelligence artificielle générative.

PARTIE 2.1 - Comment choisir le *benchmark* pertinent ?

La première étape du processus de sélection des *benchmarks* consiste à définir clairement la tâche ou l'ensemble des tâches que vous souhaitez que votre LLM exécute. En comprenant les tâches, vous pourrez naviguer plus efficacement à travers l'écosystème des *benchmarks*.

Il est important de distinguer les *benchmarks open-domain* et les *benchmarks closed-domain*. Les *benchmarks open-domain* évaluent les LLMs sur des questions qui ne font pas partie de l'ensemble de données d'entraînement et sont représentatives de la réalité utilisateur, offrant une mesure plus réaliste des capacités du modèle en situation réelle (c'est l'équivalent de la capacité de généralisation de l'IA prédictive). En revanche, les *benchmarks closed-domain* se concentrent sur un domaine spécifique et peuvent inclure des données issues de l'ensemble d'entraînement. Notre attention se porte principalement sur les *benchmarks open-domain* car ce qui nous intéresse vraiment, c'est leur capacité à représenter la réalité elle-même. En d'autres termes, nous privilégions les *benchmarks* qui posent des questions extérieures à l'ensemble de données d'entraînement et qui sont véritablement représentatives des scénarios réels auxquels les utilisateurs peuvent faire face.

Ci-dessous, nous vous proposons une *mindmap* (ou carte mentale) qui trace un vaste éventail de *benchmarks*, chacun lié à une certaine fonction ou domaine de capacité du LLM. Cette représentation visuelle est comme une boussole pour naviguer dans le terrain complexe des *benchmarks* LLM, garantissant que vous sélectionnez un test qui examine les capacités les plus pertinentes pour l'usage

prévu de votre LLM. Elle augmente vos chances de trouver le LLM le plus adapté à votre usage.

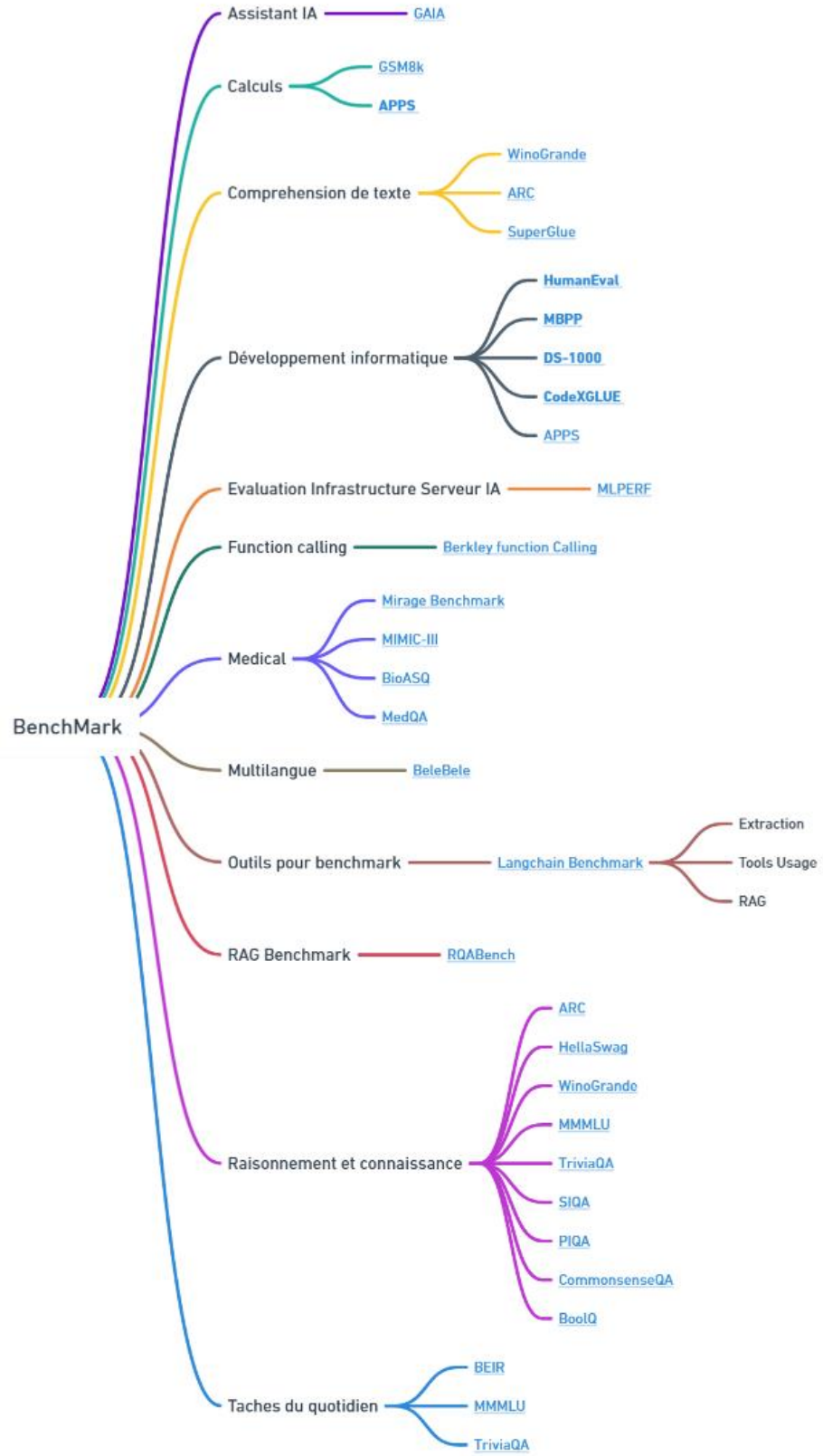

Figure 3 : *MindMap*

Le choix du *benchmark* devrait être étroitement aligné avec les tâches spécifiques à accomplir — que ce soit le raisonnement et la compréhension linguistiques, les fonctionnalités quotidiennes, ou les tâches computationnelles, parmi d'autres. Chaque tâche correspond à un ou plusieurs *benchmarks* qui peuvent mesurer précisément la capacité du LLM à exécuter cette tâche, assurant ainsi une évaluation adéquate.

Pour un aperçu actuel et complet de la performance de divers LLM à travers ces *benchmarks*, des classements à jour sont disponibles, bien qu'ils ne couvrent que partiellement l'ensemble des *benchmarks* mentionnés ici. Des plateformes telles que le classement de Hugging Face[4] et le classement CRFM HELM de Stanford[5] compilent et mettent continuellement à jour les métriques de performance d'un large éventail de LLM. Ces ressources sont inestimables pour rester informé des dernières avancées et pour prendre une décision fondée sur les données concernant le bon LLM pour vos besoins.

Hugging Face *Open Leaderboard*

C'est une plateforme où chercheurs et développeurs peuvent comparer les performances de différents modèles d'apprentissage automatique, en se concentrant particulièrement sur le traitement du langage naturel (NLP). Ce classement permet aux utilisateurs de soumettre leurs modèles et de les faire évaluer par rapport à des *benchmarks* établis. Il fait partie de l'initiative plus large de Hugging Face visant à promouvoir la science ouverte et la transparence dans le développement de l'IA.

Les modèles dans le classement sont généralement évalués sur diverses tâches telles que la classification de texte, la réponse aux questions, la génération du code et plus encore. Les performances de chaque modèle sont rapportées en utilisant des métriques standard pertinentes pour chaque tâche, ce qui fournit une comparaison claire et directe entre les modèles.

[4] Hugging Face. Open Leaderboard. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard
[5] Stanford Center for Research on Foundation Models. HELM. https://crfm.stanford.edu/helm/

CRFM HELM de Stanford

Le *Stanford Center for Research on Foundation Models* (CRFM) propose HELM (*Holistic Evaluation of Language Models*), un ensemble de *benchmarks* conçu pour évaluer les modèles de langage fondamentaux sur une variété de tâches complexes. En comparaison avec le Hugging Face *Open Leaderboard*, qui se concentre uniquement sur des modèles *open-source* évalués par la communauté, HELM se distingue par sa capacité à inclure également des modèles propriétaires dans ses évaluations.

Le *benchmark* Stanford CRFM HELM comprend six évaluations distinctes (*leaderboards*), chacune ciblant différents aspects des modèles de langage et de vision-langage :

- *Lite* : évaluations légères et larges des capacités des modèles de langage utilisant l'apprentissage en contexte.
- *Classic* : évaluations approfondies des modèles de langage basées sur les scénarios décrits dans l'article original de HELM.
- HEIM : Évaluations holistiques des modèles de génération de texte vers image.
- *Instruct* : évaluations des modèles suivant des instructions, avec des notations absolues.
- MMLU : évaluations de la compréhension du langage sur de multiples tâches (*Massive Multitask Language Understanding*) utilisant des invites de commande standardisées.
- VHELM : évaluations holistiques des modèles vision-langage.

Pour ceux qui sont intéressés à utiliser des modèles du classement, il est important de considérer que, bien qu'un modèle puisse bien performer sur des tâches de *benchmark*, les performances dans des applications réelles peuvent varier. Ainsi, une validation et une adaptation supplémentaires peuvent être nécessaires pour répondre à des besoins opérationnels spécifiques et nous vous suggérons de faire une validation supplémentaire sur vos cas d'usage spécifiques.

PARTIE 2.2 - Description des *benchmarks* existants

La sélection d'un grand modèle de langage (LLM) commence par une étape fondamentale : **identifier les tâches** les plus critiques pour le succès de votre projet. Connaître les tâches que vous avez besoin que votre LLM accomplisse est essentiel, car cela détermine quels *benchmarks* seront les plus pertinents pour votre processus de décision. Avec un objectif de tâches clair en tête, l'immense gamme

de *benchmarks* disponibles devient un guide sur mesure, vous orientant vers le LLM le mieux adapté à vos exigences. Cette section fournit une explication du paysage des *benchmarks*, vous dotant des connaissances nécessaires pour évaluer et choisir un LLM spécifique.

Approfondissement des *benchmarks* spécifiques aux tâches

Les *benchmarks* spécifiques aux tâches agissent comme des instruments finement réglés pour évaluer la compétence des LLM dans des domaines de performance distincts. Ces *benchmarks* offrent des aperçus sur la capacité d'un LLM à gérer des fonctions spécialisées — que ce soit maîtriser les nuances de la langue, trouver des solutions à des problèmes complexes, ou valider la fiabilité d'une information — ce qui est crucial lorsque votre application exige une exécution de tâches précises et efficaces.

Le tableau ci-dessous présente une variété de tâches qu'il est possible de réaliser à l'aide d'un LLM. Les *benchmarks* présentés ici proposent des séries de tests pour s'assurer de la bonne réalisation de ces tâches telles que le raisonnement, la compréhension de texte, etc. Ces tests s'étendent de questions triviales jusqu'à des tâches de compréhension de texte avancée et des calculs arithmétiques, en passant par des évaluations de la performance d'infrastructures de serveurs IA. Chaque ligne du tableau ci-dessous présente le nom d'un *benchmark* (avec un lien vers celui-ci), ses tâches principales, et la description de l'évaluation menée.

2.2.1 - Assistant IA

Nom du *benchmark* (avec lien)	Descriptif de l'évaluation	Dataset
GAIA	Se concentre sur l'évaluation des assistants IA généraux à travers des tâches du monde réel qui mettent à l'épreuve leur capacité de raisonnement, d'interaction multimodale, et d'utilisation d'outils web. Il vise à combler le fossé entre les performances humaines et celles de l'IA.	Ensemble de tâches diverses et représentatives des défis réels, soulignant l'importance d'une approche polyvalente et adaptative par les modèles d'IA pour répondre aux exigences du monde réel.

2.2.2 - Calculs

Nom du *benchmark* (avec lien)	Descriptif de l'évaluation	Dataset
GSM8K	Se concentre sur le raisonnement arithmétique avec des questions dérivées de problèmes de mathématiques de niveau primaire.	8 500 problèmes de mathématiques variés et de haute qualité créés par des auteurs humains, divisés en 7 500 problèmes d'entraînement et 1 000 problèmes de test. Ces problèmes nécessitent de 2 à 8 étapes de raisonnement et impliquent des opérations arithmétiques élémentaires.

2.2.3 - Compréhension de texte

Nom du *benchmark* (avec lien)	Descriptif de l'évaluation	Dataset
SuperGlue	Propose un ensemble de *benchmarks* conçu pour évaluer des compétences complexes en compréhension de texte et en raisonnement pour les modèles de langage avancés, surpassant les défis posés par des *benchmarks* antérieurs comme GLUE.	Plusieurs tâches de raisonnement et de compréhension qui nécessitent des interactions complexes avec le texte, mettant les modèles au défi de démontrer des capacités avancées en traitement du langage naturel.
WinoGrande	Teste la capacité des LLM à comprendre et à raisonner dans des contextes où il est nécessaire d'utiliser le bon sens pour résoudre des problèmes.	44 000 problèmes inspirés par le défi original de *Winograd Schema Challenge*, mais ajustés pour améliorer l'échelle et la difficulté du dataset.
Arc	Contient des questions de science complexes à choix multiples conçues pour les niveaux du CE2 à la 3ème. Il y a un ensemble "Facile" et un ensemble "Défi" qui nécessite un raisonnement plus avancé.	7 787 questions de science à choix multiples issues de diverses sources pour des tests standardisés de niveau collège, ainsi qu'un corpus de 14 millions de phrases pertinentes pour l'entraînement ou l'ajustement fin des modèles dans le domaine scientifique.

2.2.4 – Développement informatique

Nom du *benchmark* (avec lien)	Descriptif de l'évaluation	Dataset
HumanEval	Évalue la compréhension du langage et la compétence en programmation des modèles en leur demandant de résoudre des problèmes de programmation.	164 problèmes de programmation soigneusement conçus qui évaluent la compréhension du langage, les algorithmes et les mathématiques simples. Les problèmes sont variés, certains étant comparables à des problèmes de maintenance courante.
MBPP (*Mostly Basic Python Problems*)	Résoudre des problèmes de codage Python sur des niveaux basiques et sur les fondamentaux.	1 000 problèmes de programmation Python, se concentrant sur des fonctionnalités de base et des tâches de programmation courantes.
DS-1000 (*DeepSource Python Bugs Dataset*)	Évalue la capacité des modèles à détecter des bugs dans le code Python.	1 000 fonctions Python annotées qui testent les modèles sur leur capacité à identifier et comprendre les erreurs courantes dans le codage.
CodeXGLUE	Inclut des tâches telles que la complétion de code, la traduction de code et la réparation de code dans plusieurs langages de programmation.	14 Dataset qui englobent une variété de défis de programmation et de langages, offrant une évaluation complète de la compréhension et de la génération de code des modèles.
APPS (*Automated Programming Progress Standard*)	Teste la capacité des modèles à comprendre et résoudre des problèmes de programmation complexes qui nécessitent une compréhension approfondie des concepts algorithmiques et des structures de données.	Une collection de 10 000 problèmes de programmation, allant de questions introductives à des défis de niveau compétition universitaire, avec des solutions de référence et des cas de test pour évaluer la correction et l'efficacité des solutions générées par les modèles.
SWE-Bench (*Software Engineering Benchmark*	Évalue la capacité des modèles de langage à résoudre des problèmes logiciels réels en générant des correctifs pour les issues sur GitHub.	Un ensemble de problèmes de programmation issus de GitHub, où chaque problème nécessite que le modèle propose une solution pour rectifier l'issue soulevée.

2.2.5 - Evaluation infrastructure serveur IA

Nom du *benchmark* (avec lien)	Descriptif de l'évaluation	Dataset
MLPerf	Mesure la performance du matériel, du logiciel et des services d'apprentissage machine à travers une gamme de tâches et de scénarios.	Divers datasets en fonction des scénarios spécifiques, comme ImageNet pour la classification d'images, SQuAD pour le traitement du langage naturel, et LibriSpeech pour la reconnaissance vocale.

2.2.6 - *Function Calling*

Nom du *benchmark* (avec lien)	Descriptif de l'évaluation	Dataset
Berkeley function calling	Teste la capacité des LLM à interpréter des instructions en langage naturel pour effectuer des appels aux outils externes, ce qui est crucial pour l'interaction avec des API externes.	Collection structurée de scripts d'évaluation et d'API pour tester les capacités d'appel de fonctions des modèles, en se concentrant sur la conversion efficace des instructions en exécutions de fonction.

2.2.7 – Médical

Nom du *benchmark* (avec lien)	Descriptif de l'évaluation	Dataset
Mirage *Benchmark*	Évalue les systèmes de génération augmentée par la récupération (RAG) dans le domaine médical, testant leur capacité à intégrer des connaissances externes pour la génération de réponses précises.	Cinq ensembles de données couramment utilisés dans les Q/R médicales, avec des questions allant de l'examen médical aux recherches biomédicales, visant à tester la capacité des systèmes RAG à répondre de manière informée et précise.
PubMedQA	Teste l'extraction et la vérification d'informations factuelles à partir de textes scientifiques (PubMed) en répondant par Oui, Non ou peut-être.	1 000 instances d'assurance qualité étiquetées par des experts, 2 000 non étiquetées et, 3 000 instances d'assurance qualité générées artificiellement.
MIMIC-III	Utilisé pour prédire les résultats des patients, extraire des informations cliniques et générer des notes cliniques.	Notes, résultats de tests de laboratoire, signes vitaux, etc. anonymisés de patients en soins intensifs.
BLUE (*Biomedical Language Understanding Evaluation*)	Se compose de cinq tâches différentes de *text-mining* biomédical avec dix corpus. Ces tâches couvrent un large éventail de genres de textes (littérature biomédicale et notes cliniques), de tailles d'ensembles de données et de degrés de difficulté et, plus important encore, mettent en évidence les défis courants de l'exploration de textes en biomédecine.	Divers corpus biomédicaux, y compris des résumés de PubMed et des rapports d'essais cliniques.
BioASQ	Évalue l'indexation sémantique biomédicale via la récupération et la génération d'informations biomédicales précises.	Questions et réponses par des humains et résumés d'articles de PubMed par exemple.

MedQA (USMLE)	Évalue la compréhension et l'application des connaissances médicales dans un contexte d'examen.	Réponses à des questions à choix multiples basées sur les examens de la licence médicale des États-Unis (USMLE). L'ensemble des données est collecté à partir des examens professionnels de médecine. Il couvre trois langues : anglais, chinois simplifié et chinois traditionnel, et contient respectivement 12 723, 34 251 et 14 123 questions pour les trois langues.

2.2.8 - Multilingue

Nom du *benchmark* (avec lien)	Descriptif de l'évaluation	Dataset
BeleBele	Teste la compréhension et la génération de langage naturel multilingue à travers des tâches diverses, mettant en évidence l'importance de l'apprentissage contextuel et de la pertinence dans la génération de texte.	67 500 échantillons de formation et 3 700 échantillons de développement, principalement issus du dataset RACE. Il inclut des questions en 122 variantes de langue, avec 900 questions par variante, couvrant 488 passages distincts. Chaque question est accompagnée de quatre réponses à choix multiples, dont une seule est correcte.

2.2.9 - Outils pour *benchmark*

Nom du *benchmark* (avec lien)	Descriptif de l'évaluation	Dataset
Langchain Benchmark	Langchain est conçu pour évaluer les tâches liées aux modèles de langage en utilisant divers cas d'usage de bout en bout. Ce *benchmark* évalue la capacité des modèles à interagir avec des outils et à traiter des tâches spécifiques en utilisant LangSmith pour le stockage et l'évaluation des datasets.	Chaque *benchmark* est accompagné d'un dataset associé, qui est utilisé pour tester et évaluer les performances des modèles de langage sur des tâches spécifiées, avec un accent sur l'utilisation pratique et l'évaluation des architectures (GitHub) (LangChain AI) (LangChain AI) (LangChain AI) (PyPI).

2.2.10 - Rag *Benchmark*

Nom du *benchmark* (avec lien)	Descriptif de l'évaluation	Dataset
RetrievalQA	Se concentre sur l'évaluation de la capacité des modèles à exécuter des tâches de question-réponse où la récupération d'informations est cruciale. Il teste la performance des systèmes de question-réponse en utilisant des techniques de récupération avancées pour améliorer la précision des réponses.	Le *benchmark* n'indique pas spécifiquement un dataset unique mais teste les capacités des modèles sur des ensembles de données réels avec des métriques telles que le rang réciproque moyen (MRR) et le rappel à N pour une évaluation précise

2.2.11 - Raisonnement et connaissance

Nom du *benchmark* (avec lien)	Descriptif de l'évaluation	Dataset
Arc	Contient des questions de science complexes à choix multiples conçues pour les niveaux du CE2 à la 3ème. Il y a un ensemble "Facile" et un ensemble "Défi" qui nécessite un raisonnement plus avancé.	7 787 questions de science à choix multiples issues de diverses sources pour des tests standardisés de niveau collège, ainsi qu'un corpus de 14 millions de phrases pertinentes pour l'entraînement ou l'ajustement fin des modèles dans le domaine scientifique
HellaSwag	Vise le raisonnement de bon sens avec des défis de complétion de contexte qui sont faciles pour les humains mais difficiles pour les LLM.	70 000 questions à choix multiples basées sur des situations concrètes issues de deux domaines principaux : ActivityNet et WikiHow. Chaque question est accompagnée de quatre propositions de réponse, où la bonne réponse est la suite logique de l'événement décrit, et les trois autres sont stratégiquement conçues pour tromper les modèles tout en restant plausibles pour les humains.
WinoGrande	Teste la capacité des LLM à comprendre et à raisonner dans des contextes où il est nécessaire d'utiliser le bon sens pour résoudre des problèmes et résoudre des ambiguïtés.	44 000 problèmes inspirés par le défi original de *Winograd Schema Challenge*, mais ajustés pour améliorer l'échelle et la difficulté du dataset.
MMLU	Fournit une évaluation large sur plusieurs tâches pour mesurer la connaissance générale et le raisonnement.	57 sujets dans divers domaines allant des sciences, technologies, ingénierie et mathématiques aux sciences humaines et sociales, en passant par le droit et l'éthique. Il mesure les connaissances acquises pendant la phase de pré-entraînement des modèles en évaluant exclusivement dans des contextes de *zero-shot* (prompts sans exemples de résultats) et de *few-shot* (en donnant quelques exemples de résultats désirés dans les prompts), ce qui le rend similaire à la manière dont les humains sont évalués.
TriviaQA	Contient des questions et réponses pour le questionnement de culture générale type *Trivial Pursuit*.	Plus de 650 000 triples question-réponse-évidence. Il inclut 95 000 paires de questions-réponses rédigées par des amateurs de trivia et des documents de preuve indépendamment collectés, fournissant une supervision distante de qualité pour répondre aux questions. Les paires de questions-réponses comprennent à la fois des sous-ensembles vérifiés par l'homme et générés par la machine
BoolQ	*Benchmark* pour répondre à des questions de type vrai ou faux basées sur des extraits de Wikipédia.	15 942 questions basées sur les recherches Google et articles de Wikipedia correspondants.
CommonsenseQA (CQA)	Évalue la capacité des modèles à raisonner sur des connaissances de bon sens.	12 247 Questions à choix multiples nécessitant du bon sens pour répondre.
Physical Interaction Question Answering (PIQA)	Teste la compréhension des propriétés physiques à travers des scénarios de résolution de problèmes.	Questions nécessitant un raisonnement sur des interactions physiques du quotidien (ex : séparer le blanc du jaune d'œuf avec une bouteille).

Social Interaction Question Answering (SIQA)	Teste la navigation des modèles dans des situations sociales à travers des questions à choix multiples.	37 000 questions/réponses Scénarios impliquant des interactions humaines.

2.2.12 - Tâche du quotidien

Nom du *benchmark* (avec lien)	Descriptif de l'évaluation	Dataset
BEIR	Évalue une gamme variée de tâches IR (Recherche d'information) au niveau des phrases ou des passages pour une évaluation *zero-shot* (sans exemples insérés dans les prompts).	18 datasets de 10 tâches hétérogènes de récupération d'informations, offrant une diversité de tâches, de domaines et de stratégies d'annotation.
MMLU	Fournit une évaluation large sur plusieurs tâches pour mesurer la connaissance générale et le raisonnement.	57 sujets dans divers domaines allant des sciences, technologies, ingénieries et mathématiques aux sciences humaines et sociales, en passant par le droit et l'éthique. Il mesure les connaissances acquises pendant la phase de pré-entraînement des modèles en évaluant exclusivement dans des contextes de *zero-shot* et de *few-shot*, ce qui le rend similaire à la manière dont les humains sont évalués.
TriviaQA	Contient des questions et réponses pour le questionnement de culture générale type *Trivial Pursuit*.	Plus de 650 000 triples question-réponse-évidence. Il inclut 95K paires de questions-réponses rédigées par des amateurs de trivia et des documents de preuve indépendamment collectés, fournissant une supervision distante de qualité pour répondre aux questions. Les paires de questions-réponses comprennent à la fois des sous-ensembles vérifiés par l'homme et générés par la machine.

PARTIE 2.3 - Quelques exemples simples pour choisir les LLM

Lors de la sélection initiale de grands modèles de langage (LLM) pour une application spécifique, l'utilisation de *benchmarks* établis constitue une première étape solide pour cibler les candidats susceptibles de bien correspondre au contexte souhaité. Voici des exemples détaillés de la manière dont différents *benchmarks* peuvent être appliqués pour adapter le choix d'un LLM à des tâches spécialisées.

Création de contenu éducatif (Utilisation du *benchmark* Arc)

- **Application** : Développement de logiciels éducatifs conçus pour générer des questions de quizz scientifiques pour les élèves de l'école primaire et du collège.
- **Tâche** : Le logiciel doit créer des questions qui mettent au défi le raisonnement et la compréhension des élèves à différents niveaux de difficulté.
- ***Benchmark* choisi** : Le *benchmark* Arc, qui comprend des questions à choix multiples complexes conçues pour les niveaux du CE2 à la 3ème.

- **Raison du choix** : Ce *benchmark* est choisi car son accent sur le raisonnement et la connaissance scientifique s'aligne parfaitement avec les objectifs éducatifs du logiciel, assurant que les questions sont à la fois précises et suffisamment stimulantes.

Raisonnement contextuel dans la génération d'histoires (Utilisation du *benchmark* HellaSwag)

- **Application** : Un outil de génération de récits cohérents et logiquement progressifs, utilisé à des fins de divertissement et éducatives.
- **Tâche** : L'outil doit garantir que les récits maintiennent une cohérence logique et une progression réaliste tout au long de l'histoire.
- ***Benchmark* choisi** : HellaSwag, qui évalue les capacités de raisonnement du modèle avec des défis de complétion de contexte.
- **Raison du choix** : Il est particulièrement efficace pour évaluer la capacité du modèle à continuer les histoires de manière logique, ce qui est crucial pour produire des narrations captivantes et crédibles.

Ces exemples démontrent que l'utilisation de *benchmarks* adaptés permet de sélectionner des LLM qui sont non seulement compétents en théorie, mais aussi efficaces en pratique pour des applications spécialisées. Cependant, pour répondre précisément aux besoins spécifiques d'une application, il est souvent nécessaire d'adopter une méthode d'évaluation plus personnalisée. Bien que ce guide ne couvre pas en détails les processus d'optimisation internes, tels que l'ajustement des paramètres ou le choix final des modèles, il est essentiel de comprendre que la validation finale de l'application nécessitera une évaluation interne complète utilisant des jeux de données spécifiques à l'organisation.

PARTIE 2.4 - Attention à la contamination des *Benchmark*s

Les *leaderboards* [*benchmarks*] de grands modèles de langage (LLM) fournissent une comparaison pratique de différents modèles sur des tâches spécifiques, telles que la génération de texte, la compréhension linguistique et la traduction. Toutefois, il est crucial de comprendre que ces classements peuvent parfois être affectés par la manière dont les *benchmarks* sont structurés et exécutés.

Dans l'évaluation des LLM, il est non-négligeable de considérer les risques associés à l'utilisation des *benchmarks* pour les classements, tels que la contamination des données. Ce phénomène se produit lorsque les données de test, destinées à

évaluer les modèles, sont aussi incluses dans l'ensemble des données d'entraînement, que ce soit intentionnellement ou non[6, 7]. Cette inclusion peut entraîner une reconnaissance des exemples de test par le modèle au lieu de leur résolution par une compréhension authentique, faussant ainsi les résultats des évaluations.

Il est donc recommandé d'utiliser les *leaderboards* avec discernement et de considérer le développement de *benchmarks* spécifiques à des tâches pour mieux évaluer les capacités des modèles dans des scénarios d'application réels. Cette approche permettra d'obtenir des évaluations plus robustes et significatives, essentielles pour le développement technologique dans le domaine des LLM.

PARTIE 2.5 - Comment aller plus loin que les *benchmarks* ?

Pour garantir l'efficacité d'une application basée sur les LLM, nous recommandons vivement d'adopter une démarche d'évaluation réfléchie. Cela commence par la création d'un ensemble de données d'évaluation, adapté aux entrées spécifiques que l'application est susceptible de recevoir. Cet ensemble peut inclure des questions et réponses attendues, enrichies de contextes pertinents.

Pour une mesure complète, il est recommandé d'utiliser une combinaison diversifiée de métriques, à la fois statistiques et basées sur des modèles. Ces métriques devraient être adaptées aux tâches spécifiques de l'application, reflétant des aspects tels que la cohérence factuelle, la pertinence des réponses, la cohérence logique, la toxicité des contenus générés, et les biais potentiels. Pour cela, on pourra se référer à la littérature sur le sujet[8].

---

[6] Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen and Jiawei Han. Don't Make Your LLM an Evaluation Benchmark Cheater. arXiv preprint. November 2023. https://arxiv.org/pdf/2311.01964.

[7] Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E. Gonzalez, Ion Stoica. Rethinking Benchmark and Contamination for Language Models with Rephrased Samples. arXiv preprint. November 2023. https://arxiv.org/pdf/2311.04850

[8] Trois références sur le sujet :

- Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, Jianfeng Gao. Large Language Models : A survey. arXiv preprint. arXiv preprint. February 2024. https://arxiv.org/abs/2402.06196v2
- Taojun Hu, Xiao-Hua Zhou. Unveiling LLM Evaluation Focused on Metrics: Challenges and Solutions. arXiv preprint. April 2024. https://arxiv.org/abs/2404.09135
- Tinh Son Luong, Thanh-Thien Le, Linh Ngo Van, Thien Huu Nguyen. Realistic Evaluation of Toxicity in Large Language Models. arXiv preprint. May 2024. https://arxiv.org/abs/2405.10659v2

L'étape suivante implique l'implémentation d'un système de notation pour calculer les scores selon ces métriques. Par exemple, l'utilisation de modèles d'inférence du langage naturel pour évaluer la cohérence factuelle, ou des encodeurs croisés pour la pertinence des réponses.

Enfin, il est crucial d'intégrer ces évaluations comme des tests unitaires dans ce qu'on appelle les pipelines CI/CD (*Continuous Integration/ Continuous Delivery*), permettant des évaluations automatiques régulières. Cela aide à identifier et à améliorer les réponses insatisfaisantes de manière proactive, assurant ainsi que l'application reste performante et fiable dans le temps.

En résumé, évaluer une application LLM est un processus continu et itératif, indispensable pour développer des solutions robustes et fiables. Il est crucial de développer des *benchmarks* spécifiques à des tâches pour obtenir des évaluations plus robustes et significatives. Les *benchmarks* devraient rester un outil d'aide à la décision, facilitant une évaluation plus complète et contextualisée des technologies de langage avancées.

HUB FRANCE IA

PARTIE 3 Echanges avec les fournisseurs

CHOISIR UN MODELE D'IA GENERATIVE POUR SON ORGANISATION

Juin 2024

PARTIE 3 : Échanges avec les fournisseurs

Pour rappel, nous avons commencé dans la première partie par analyser une enquête réalisée auprès des organisations utilisatrices potentielles de LLM. Celle-ci a permis de guider notre travail sur un certain nombre de critères importants. Ensuite, nous avons fait un état des lieux de plusieurs comparatifs de performances existants en les décryptant et précisant pour chacun d'eux ce que représentent les différents scores affichés. Dans cette troisième partie, nous allons décrire comment nous avons obtenu les informations sur les autres critères identifiés comme importants. Pour cela, nous avons pris contact avec différents fournisseurs œuvrant dans le domaine des modèles de langage LLM : **Microsoft (incluant OpenAI), Google, Meta, Mistral, LightOn, la DiNum, Anthropic, Amazon, HuggingFace.**

PARTIE 3.1 - Critères issus de l'enquête auprès des organisations utilisatrices

Nous avons créé une grille de questions à la suite des résultats de l'enquête. Celle-ci comporte 7 thématiques principales qui permettent d'articuler le document que nous avons envoyé aux différents fournisseurs de modèles. Voici les thématiques en question ainsi que les questions associées :

- **Sécurité & sûreté**

Question	Votre réponse
Quelles mesures prenez-vous pour sécuriser les flux de données entrants et sortants (prompts et réponses) ?	Liste de mesures
Quelles mesures prenez-vous pour sécuriser les données d'entrainement utilisées ?	Liste de mesures
Comment votre système est-il protégé contre les attaques potentielles, comme le piratage ou l'ingénierie sociale, le prompt injection?	Liste de mesures
Comment sont utilisées les données apportées par les prompts des utilisateurs?	Non utilisées, utilisées pour l'entrainement, stockées quelque part, … quelle politique et options ?

- **Légal & juridique**

Question	Votre réponse
Avez-vous adopté des normes au sein de votre organisation (ex : ISO 42001, ISO 9001, ISO27001, ISO14001,...)	Liste de normes adoptées + champ autre
Prévoyez-vous d'avoir recours aux normes pour obtenir une présomption de conformité avec l'AI Act ?	Oui, Partiellement, Non + explication
A quels cadre réglementaires vous conformez vous ? (Européen, USA, Chine...)	Listes de cadre réglementaires
Prévoyez-vous de prendre part à l'élaboration et l'implémentation des codes de conduites prévus dans le cadre de l'AI Act concernant les GPAI ?	Oui, Partiellement, Non + explication
Où peut-on consulter votre politique RGPD ?	Lien vers la politique RGPD
Partagez-vous des informations sur vos données d'entraînement?	Oui, Partiellement, Non + explication
Prenez-vous la responsabilité juridique en cas d'utilisation de données copyrightées non autorisées lors de votre phase d'apprentissage ?	Oui, Partiellement, Non + explication

- **Modèles**

Question	Votre réponse
Votre modèle peut-il être requêté via une API? Ou une interface utilisateur ?	Oui, Non
Votre modèle s'intègre-t'il déjà dans des environnements de développement prêt à l'emploi ?	Oui, Non + explication
Quelle est votre fréquence de mise à jour et évolution des modèles ?	Fréquence approximative
Quel est le nombre maximal de tokens dans les prompts/réponses et vos projections futures ?	Nombre actuel
Quel est le nombre maximal de requêtes par jour ?	Nombre actuel
Y a t'il des limites particulières au modèle en terme d'augmentation du nombre d'utilisateurs ou du nombre de requêtes (et des limites par jour ?) ?	Oui, Non + explication
Avez-vous déjà publié des modèles multimodaux ?	Oui, Non + explication
Certains de vos modèles sont-ils spécialisés dans certains domaines ou tâches ? Si oui lesquels ?	Oui, Non + liste
Quelles sont les possibilités de fine tuning de vos modèles ?	Explication
Le modèle est-il intégré à des solutions logicielles (Copilot, ...)	Oui, Non + explication
Quelle valeur ajoutée par rapport aux autres modèles ? Qu'est-ce qui démarque ?	Explication

- **Infrastructure**

Question	Votre réponse
Votre modèle est-il utilisable sur des plateformes Cloud (cloud public ou privé voire souverain) ?	Oui, Non
Quelle configuration minimale pour faire tourner le modèle ?	Configuration minimale nécessaire
Sur quelles plateformes votre modèle est-il déjà disponible ? utilisable ?	liste de plateformes
Un plan entreprise (prix de groupe, sécurisation des données commerciales) est-il possible ?	Oui, Non + explication
Quelles sont les SLAs du service ?	Détail à fournir

- **Business model**

Question	Votre réponse
Quels mode de tarification proposez-vous (essai ? au prompt ? au nombre de tokens ? ...) ?	Lien vers le coût par token ? Autre ?
Y a t'il d'autres coûts à prévoir (maintenance, évolution, …) ?	Explication + Lien vers document de pricing
Pouvez-vous nous donner une idée du nombre d'utilisateurs de votre modèle ?	Ordre de grandeur

- **Accompagnement des clients**

Question	Votre réponse
Réalisez-vous des formations sur l'utilisation de vos modèles ? Si oui sous quelles conditions ?	Oui, Non + explication
Avez-vous une communauté d'entraide active sur votre modèle ? Forum d'entraide ?	Oui, Non + explication
Avez-vous des services d'accompagnement à l'installation ou l'utilisation, y compris support ? Si oui sous quelles conditions ?	Oui, Non + explication

- **Considérations écologiques**

Question	Votre réponse
Mesurez-vous l'impact écologique de vos modèles ? Sur quels aspects (énergie, eau, impact carbone, ...) ?	Oui, Partiellement, Non + explication
Quels documents ou chiffres pouvez-vous communiquer sur les consommations en carbone, en électricité ou en eau sur la phase d'entrainement?	Lien vers des documents sur les aspects écologiques
Quels documents ou chiffres pouvez-vous communiquer sur les consommations en carbone, en électricité ou en eau sur la phase d'inférence?	Détails à fournir

PARTIE 3.2 - Méthodologie pour récupérer les informations

La grille de questions a été envoyée dans son intégralité aux fournisseurs de modèles qui ont un modèle propriétaire. Cette grille a été raccourcie pour les fournisseurs de modèles *open source*, ces derniers laissant la main aux utilisateurs pour un certain nombre d'aspects dont l'hébergement. Enfin, en cours de route, nous avons également vu apparaître une catégorie d'acteurs qui se positionnent plus comme des plateformes capables d'héberger plusieurs modèles en mode SaaS (*Software as a Service*), ce qui signifie qu'elles donnent accès via leur plateforme à différents modèles qu'il est possible d'utiliser directement sur cette même plateforme.

Après l'envoi de la grille de questions, chaque fournisseur disposait de plusieurs semaines pour répondre. Ensuite, une session d'échanges en visioconférence a été

programmée entre les fournisseurs répondants et le groupe de travail pour élaborer plus en détails certaines réponses fournies.

Voici le bilan de la participation à l'enquête :

Acteur concerné	Réponse
Amazon	N'a pas souhaité répondre à l'enquête, indiquant préférer être vu comme une plateforme d'accès aux modèles
Anthropic	N'a pas souhaité répondre à l'enquête, indiquant préférer mettre à jour son site avec les informations demandées
DiNum	A répondu à l'enquête et participé à l'échange
Google	A répondu à l'enquête et participé à l'échange
HuggingFace	A répondu à l'enquête et participé à l'échange
LightOn	A répondu à l'enquête et participé à l'échange
Meta	N'a pas souhaité répondre à l'enquête, indiquant un manque de temps pour fournir les réponses aux différentes questions posées
Microsoft (incluant OpenAI)	A répondu à l'enquête et participé à l'échange
Mistral	A répondu à l'enquête et participé à l'échange

En ce qui concerne les quelques fournisseurs n'ayant pas voulu participer à l'enquête, les critères ont été analysés par le groupe de travail sur la base des informations publiques disponibles sur le site du fournisseur.

PARTIE 3.3 - Présentation détaillée des différents acteurs contactés

Détaillons maintenant les différents acteurs contactés. Ces descriptions ont été élaborées à partir d'informations disponibles sur les sites des différents acteurs et d'échanges avec ceux qui nous ont répondu.

Amazon

Amazon Web Services (AWS) met à disposition une infrastructure *cloud*, conçue pour l'entraînement et l'inférence de modèles d'IA générative à grande échelle. AWS

propose Amazon Bedrock, permettant d'accéder à des modèles de fondation[9] (*foundation models*), provenant d'entreprises comme Anthropic, Cohere, Meta, etc. A ce titre, AWS noue des partenariats avec des sociétés comme Mistral AI, dont le dernier modèle *Mistral Large* est désormais disponible sur Bedrock.

Les clients peuvent personnaliser ces modèles avec leurs propres données et bénéficier de la sécurité, de la confidentialité et du contrôle d'accès offerts par AWS. Pour répondre aux enjeux de souveraineté des données, AWS a récemment rendu Bedrock disponible dans la région Europe (Paris), permettant aux entreprises françaises et européennes d'exploiter l'IA générative en conformité avec les réglementations locales.

AWS n'a pas de modèle majeur d'IA générative propre à Amazon mais propose des applications d'IA générative prêtes à l'emploi comme Amazon Q.

Anthropic

Anthropic est une entreprise américaine fondée en 2021 par d'anciens membres d'OpenAI, dédiée à la recherche et au développement de systèmes d'intelligence artificielle. Elle se définit comme une *AI & safety company*.

Les offres principales d'Anthropic incluent des modèles de langage avancés comme Claude utilisables à travers leur chatbot en ligne, via des plateformes cloud comme AWS ou via une interface API.

DiNum

L'offre d'ALBERT, développée par la Direction Interministérielle du Numérique (DINUM) en France, dans le cadre du Datalab, est une initiative visant à déployer les technologies de LLM au sein des services publics français.

[9] Le terme a été défini comme « models (e.g., BERT, DALL-E, GPT-3) trained on broad data (generally using self-supervision at scale) that can be adapted to a wide range of downstream tasks" par le centre de recherche CRFM de Stanford dans :

- R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx et al. On the opportunities and risks of foundation models. arXiv preprint. July 2022. https://arxiv.org/pdf/2108.07258

L'objectif est de favoriser la souveraineté technologique, l'indépendance vis-à-vis des fournisseurs étrangers et la lutte contre le phénomène de *Shadow GPT*[10] dans les administrations.

Cette offre souveraine repose sur des modèles *open source* pré-entraînés, comme les modèles Llama, Falcon, Mistral, … qui ont été optimisés via le fine-tuning pour répondre aux besoins spécifiques des services publics afin d'assister les agents dans des tâches administratives répétitives, comme la réponse aux questions des usagers, la rédaction de rapports et la gestion des demandes complexes.

Initiée en juin 2023, l'offre est actuellement en cours de déploiement (Mai 2024).

Google

L'offre de Google en matière de LLM est articulée autour de Vertex AI et de Gemini :

- Vertex AI est la plateforme de développement d'IA entièrement gérée de Google, offrant aux entreprises et aux développeurs un accès à plus de 130 modèles de base (*Open Source* et propriétaires), y compris les modèles de la série Gemini, exclusif à Google. Vertex AI permet de personnaliser et de gérer ces modèles de manière intégrée.
- Gemini est la série de LLM exclusive à Google. Gemini 1.5 Pro, par exemple, est leur modèle le plus avancé à date (mai 2024), actuellement en prévisualisation publique pour les clients *Cloud* et les développeurs. Ce modèle excelle dans la compréhension contextuelle longue, capable de traiter jusqu'à 1 million de tokens, ce qui ouvre de nouvelles possibilités pour les entreprises (comme l'analyse vidéo et les rapports d'incidents).
- AI Studio de Google Cloud est une plateforme qui permet aux développeurs de construire et de déployer des applications utilisant les modèles de la série Gemini. AI Studio facilite l'expérimentation avec des fonctionnalités avancées comme les capacités multimodales (traitement de l'audio, de la vidéo, du texte, du code, etc.).

[10] En mai 2023, près de 70 % des salariés français se servaient de ChatGPT sans le dire à leur responsable, d'après :
- Jacques Cheminat. Le shadow GPT s'installe dans les entreprises françaises. Le Monde Informatique. 12 mai 2023. https://www.lemondeinformatique.fr/actualites/lire-le-shadow-gpt-s-installe-dans-les-entreprises-francaises-90415.html

- Partenariats et solutions intégrées : Google collabore avec divers partenaires pour étendre les capacités de ses modèles d'IA. Par exemple, des solutions comme Gemini Code Assist sont développées pour aider les développeurs avec la génération de code et d'autres types d'assistance.

En résumé, Google offre une solution d'hébergement agnostique des LLM avec en plus des outils et modèles exclusifs comme la série Gemini.

Hugging Face

L'entreprise américano-française Hugging Face occupe une place particulière dans le paysage des LLMs. Exclusivement positionnée sur l'Open Source, Hugging Face propose une plateforme mettant gratuitement à disposition (avril 2024) :

- Plus de 600 000 modèles d'IA et de 133 000 datasets ;
- Des ressources pédagogiques ;
- Des outils pour tester des modèles directement depuis son navigateur ;
- Une large communauté ;
- Des *benchmarks* sur les performances (voir notre partie dédiée sur le sujet).

Fort de son expérience sur l'Open Source, Hugging Face offre **pour les entreprises des solutions expertes payantes comme l'Entreprise Hub** qui propose des solutions d'hébergement des modèles Open Source présents sur sa plateforme communautaire. Hugging Face se veut agnostique et propose notamment des solutions compatibles avec les *hyperscalers* tels que OVH, Microsoft Azure, AWS ou encore GCP.

L'entreprise met également à disposition un support Experts unique pour accompagner les entreprises dans le choix, l'intégration et le déploiement des modèles *open source*. Hugging Face a su en effet construire au fil des années une expertise unique sur le triptyque : hébergement / données / modèles IA *open source*.

A l'image d'autres grands acteurs dans la course aux LLM comme Mistral et Meta, Hugging Face parie sur la puissance de la communauté *open source* pour rivaliser avec les solutions propriétaires des éditeurs. Ces derniers ont été pionniers dans la démocratisation des LLM mais pourraient prochainement être rattrapés, selon Hugging Face, par les solutions *open source*, comme cela a été déjà le cas par le passé pour d'autres technologies numériques, notamment pour des modèles spécialisés.

Etant donné le positionnement spécifique d'Hugging Face, nous avons préféré lui consacrer ce paragraphe et ne pas l'inclure dans l'analyse par critères. En effet, cette dernière n'était pas pertinente pour eux.

LightOn

LightOn est une entreprise française fondée en 2016 qui se positionne comme un acteur stratégique de l'IA générative et des LLM pour les entreprises en tenant compte de l'enjeu de souveraineté nationale. Depuis 2020, l'équipe dédiée de LightOn a développé 12 modèles LLM, et la solution de plateforme Paradigm. Leur modèle phare s'appelle Alfred .

Le modèle Alfred et la plateforme Paradigm ont été développées par LightOn pour mettre à disposition des grandes entreprises et institutions publiques des solutions d'IA générative efficaces, sur mesure et facilement intégrables à leur infrastructure, garantissant la confidentialité des données.

Meta

L'offre de Meta s'articule autour de la série de modèles de langage Llama. Meta s'est distingué d'OpenAI et Google en publiant les modèles Llama sous une licence *open source* spécifique permettant les utilisations notamment dans le cadre de la recherche.

Les spécificités de Llama incluent :

- Des versions de modèles de différentes tailles, optimisées pour divers besoins et contraintes de ressources ;
- Une architecture basée sur les dernières avancées en apprentissage profond, permettant une compréhension contextuelle fine et une génération de texte fluide.

Microsoft (incluant OpenAI)

L'offre de Microsoft se compose ainsi :

- Azure AI est une place de marché mettant à disposition de nombreux modèles (*closed source* ou *open source*) ;
- Microsoft a une exclusivité pour opérer les modèles d'OpenAI pour les entreprises sur Azure ;
- Microsoft développe ses propres *Small Language Model* SLM comme Phi3 ;

- Microsoft a des partenariats non-exclusifs comme avec Mistral.ai ;
- Microsoft propose Azure AI studio, une plateforme de développement et déploiement d'applications IA générative pour les développeurs ;
- Copilot Pro est une solution payante qui intègre directement GPT4 ou GPT4 turbo pour les utilisateurs finaux, notamment sur l'offre Office 360 (Word, Excel, Powerpoint, Outlook, …). Copilot a un studio pour permettre au développeur de personnaliser la solution.

Les modèles sont disponibles pour les entreprises sur la plupart des géographies Azure et notamment en France. Les SLM peuvent être déployés partout. Microsoft opère les modèles avec des accord de niveau de service (*service-level agreement* SLA), des niveaux de sécurité importants et dans le respect de la RGPD (*Trust center* et *Azure Compliance*) Le modèle économique peut être soit au token soit sous la forme d'engagement d'unités de débit approvisionnées (*Provisioned Throughput Units* PTU). Microsoft propose des outils d'informatique décisionnelle (Power BI) permettant aux entreprises de suivre leurs émissions de gaz à effet de serre associées à leur usage d'Azure. Les clients de Microsoft peuvent se former via *Microsoft Learning Center*. Microsoft accompagne également ses clients via ses architectes d'entreprises, via ses distributeurs et son écosystème de partenaires de services.

Mistral

Mistral AI est une entreprise française cofondée en avril 2023 par Arthur Mensch, Guillaume Lample et Timothée Lacroix. Les principaux produits de Mistral sont Le Chat et La Plateforme. Ils sont conçus pour fournir aux utilisateurs des capacités de génération et de personnalisation de textes de premier ordre.

Disponibilité des produits et dates de sorties :

- Modèles commerciaux : *large*, *medium*, *small*, *embedding*. Peut être déployé sur le site de Mistral via La plateforme et sur Le chat (tous deux hébergés par Mistral) mais aussi via le *cloud* (Azure, AWS, Snowflake …) ;
- Modèles *open source* : Mistral 7B (sorti en septembre 2023), Mixtral 8x7B (novembre 2023), Mixtral 8x22B (avril 2024) ;
- La Plateforme permet d'effectuer des requêtes API sur tous les modèles de Mistral depuis décembre 2023 : *large*, *next*, *medium*, *small*, *tiny* (Mistral 7b), Mixtral 8x7B, Mixtrall 8x22B, *embedding*. C'est un produit B2B qui permet aux développeurs de créer des modèles personnalisés et d'exploiter les API des

modèles. Son service de mise au point permet aux utilisateurs de mettre au point, de tester et de déployer leurs modèles en un seul endroit.
- Le Chat est une assistant conversationnelle sorti en février 2024 et qui supporte les modèles *large* et *small*. Il permet d'interagir facilement avec les différents modèles Mistral.

Mistral est aussi disponible via des partenariats avec des distributeurs *cloud* tiers :

- Azure AI déploie le modèle *large* avec les fonctionnalités *function calling* et mode Json ;
- AWS déploie le modèle *large* (sans *Function Calling* et mode Json) ;
- Snowflake : modèles OS et *large* dans leurs fonctions Cortex et dans leur outil co-pilote.

Les produits de Mistral sont les modèles LLM. Tout ce qui est construit autour des modèles est un échafaudage pour les rendre plus accessibles et performants.

- Modèles génératifs texte à texte, capables de compléter du texte et de dialoguer. Ces modèles peuvent être modifiés et transformés arbitrairement par les clients ;
- Un service de déploiement pour faire tourner des modèles génératifs sur n'importe quelle infrastructure, de manière hermétique et sécurisée. Ce service de déploiement sera en mesure de recueillir des commentaires humains sur les générations de modèles, en vue d'une amélioration ultérieure du modèle ;
- Un service de spécialisation, capable de transformer un modèle en un nouveau modèle résolvant une tâche spécifique à une entreprise. Pour ce faire, certaines données spécifiques à la tâche sont nécessaires (avec les premiers clients, Mistral fournit des recommandations sur la forme que devraient prendre les données). Le service peut être hébergé sur l'infrastructure de l'entreprise et utilisé de manière hermétique.
- Un service de contextualisation capable d'indexer le contenu des connaissances et de l'exposer pour contextualiser les compléments ou les réponses fournis par les modèles génératifs. Cela implique à la fois un modèle d'intégration et un service de base de données vectorielles. Le service peut être hébergé sur l'infrastructure de l'entreprise et utilisé de manière hermétique.

Dans le cadre de l'étude en cours qui avait déjà été lancée lors des dernières annonces de Mistral, les réponses qui ont été demandées se focalisent sur les modèles *open source* de Mistral.

HUB FRANCE IA

PARTIE 4 Analyse détaillée des réponses

CHOISIR UN MODELE D'IA GENERATIVE POUR SON ORGANISATION

Juin 2024

PARTIE 4 : Analyse détaillée des réponses

PARTIE 4.1 - Sécurité et Sureté

Au sein de cette thématique, seuls les fournisseurs de modèles proposant des modèles propriétaires ont été interrogés sur les aspects sécurité et sûreté. En effet, les modèles *open source* devront être hébergés et utilisés sur les serveurs de l'entreprise utilisatrice ou sur des plateformes tierces que l'entreprise aura choisies. Il incombe donc directement à l'organisation ou à la plateforme tierce choisie de gérer les aspects sécuritaires (sécurisation des flux de données, des attaques potentielles, du piratage, …).

Ainsi, les réponses apportées ne le sont que pour les modèles propriétaires interrogés soit **Anthropic, Google, Microsoft (incluant OpenAI).**

Il faut aussi noter que bien que Microsoft et OpenAI aient été traités conjointement dans cette étude, des différences existent. Par exemple, si on utilise directement les API et interfaces d'OpenAI, les données se retrouveront directement sur les serveurs américains de l'entreprise. En revanche, si on utilise OpenAI à travers la suite Microsoft Azure, les données resteront hébergées sur les serveurs prévus dans l'abonnement.

Question : Quelles mesures prenez-vous pour sécuriser les flux de données entrants et sortants (prompts et réponses) ?		
Anthropic	**Google**	**Microsoft (incluant OpenAI)**
Pas d'élément trouvé	Chiffrement de bout en bout	Liste de mesures
Question : Quelles mesures prenez-vous pour sécuriser les données d'entrainement ?		
Anthropic	**Google**	**Microsoft (incluant OpenAI)**
Pas d'élément trouvé	Mesures confidentielles non dévoilées	Liste de mesures
Question : Comment votre système est-il protégé contre les attaques potentielles, comme le piratage ou l'ingénierie sociale, le prompt injection ?		
Anthropic	**Google**	**Microsoft (incluant OpenAI)**
Quelques propositions pour gérer la sécurité	Mesures confidentielles non dévoilées	Liste de mesures
Question : Comment sont utilisées les données apportées par les prompts des utilisateurs ?		
Anthropic	**Google**	**Microsoft (incluant OpenAI)**
Pas d'élément trouvé	Les données ne sont ni stockées ni utilisées	Les prompts sont gardés 30 jours pour analyse (*opt out* possible)

PARTIE 4.2 - Légal et Juridique

Pour cette thématique, l'ensemble des fournisseurs de modèles ont été interrogés. Il s'agit dans cette partie de bien comprendre les points importants autour de la protection des données personnelles, mais aussi la préparation à la mise en œuvre de l'EU AI Act.

Question : Avez-vous adopté des normes au sein de votre organisation (ex : ISO 42001, ISO 9001, ISO27001, ISO14001, ...) ?						
Anthropic	**DiNum**	**Google**	**LightOn**	**Meta**	**Microsoft (incluant OpenAI)**	**Mistral**
Pas d'élément trouvé	**Oui** Homologation des API via les protocoles très rigoureux de l'administration. Développements en lien étroit avec l'ANSII et la CNIL.	**Oui** Vertex Compliance	**En cours** Préparation CIS *Benchmark* et ISO27001	Pas d'élément trouvé	**Oui** Azure Compliance	**En cours** ISO27001 en cours d'adoption
Question : Prévoyez-vous d'avoir recours aux normes pour obtenir une présomption de conformité avec l'AI Act ?						
Anthropic	**DiNum**	**Google**	**LightOn**	**Meta**	**Microsoft (incluant OpenAI)**	**Mistral**
Pas d'élément trouvé	**Oui** Il est explicitement précisé qu'un LLM ne peut pas être utilisé dans la prise de décision, la responsabilité revient à 100% à l'utilisateur	**Oui** Google s'assure de la compliance avec le régulateur européen avant la sortie des modèles (pouvant impliquer des retards de déploiement)	**Oui**	Pas d'élément trouvé	**Oui** Lien fourni	**Oui**
Question : A quels cadres réglementaires vous conformez vous ? (Européen, USA, Chine...) ?						
Anthropic	**DiNum**	**Google**	**LightOn**	**Meta**	**Microsoft (incluant OpenAI)**	**Mistral**
Pas d'élément trouvé	Europe	Plusieurs régions Security & Compliance	Europe	Pas d'élément trouvé	Plusieurs régions Trust Center	Europe

Question : Prévoyez-vous de prendre part à l'élaboration et l'implémentation des codes de conduites prévus dans le cadre de l'AI Act concernant les GPAI ?						
Anthropic	**DiNum**	**Google**	**LightOn**	**Meta**	**Microsoft (incluant OpenAI)**	**Mistral**
Pas d'élément trouvé	Oui	Oui	Oui	Pas d'élément trouvé	Oui Lien fourni	Oui
Question : Où peut-on consulter votre politique RGPD ?						
Anthropic	**DiNum**	**Google**	**LightOn**	**Meta**	**Microsoft (incluant OpenAI)**	**Mistral**
Pas d'élément trouvé	N/A	Privacy & GPDR	RGPD	Pas d'élément trouvé	Trust Center	RGPD
Question : Partagez-vous des informations sur vos données d'entraînement ?						
Anthropic	**DiNum**	**Google**	**LightOn**	**Meta**	**Microsoft (incluant OpenAI)**	**Mistral**
Pas d'élément trouvé	Certaines utilisations de modèles sources Llama 13/7b et Mistral 13/7. Pour le fine-tuning, fiches DILA de service-public.fr (libre accès à tous) et des réponses d'agents de la fonction publique (données privées).	Non	Certaines Mix entre des bases de données publiques et des prompts validés par Lighton	Certaines Lien fourni	Non Pas de complément vs OpenAI	Certaines
Question : Quelles licences proposez-vous (Apache, MIT, GPL) ?						
Anthropic	**DiNum**	**Google**	**LightOn**	**Meta**	**Microsoft (incluant OpenAI)**	**Mistral**
Pas d'élément trouvé	N/A	Modèles propriétaires	Apache 2.0	Meta Llama Community	Modèles Phi en licence MIT	Modèles open sources en Apache 2.0

Question : Prenez-vous la responsabilité juridique en cas d'utilisation de données copyrightées non autorisées lors de votre phase d'apprentissage ?						
Anthropic	**DiNum**	**Google**	**LightOn**	**Meta**	**Microsoft (incluant OpenAI)**	**Mistral**
Pas d'élément trouvé	**Non** Pas pour les données issues du modèle source (Llama ou Mistral)	**Oui** Protecting customers with GenAI Indemnification	**Non** Utilisation majoritaire-ment à des fins privées et non publiques	Pas d'élément trouvé	**Oui** Customer Copyright Commitment	**Oui** Mistral prend la responsabilité du fait que les modèles respectent les lois applicables

PARTIE 4.3 - Modèles

L'ensemble des fournisseurs a également été interrogé pour cette thématique sur les modèles. L'idée est de faire un état des lieux des différences entre les différents modèles proposés. Ce sont les informations les plus relayées et accessibles sur Internet. Ce sont aussi celles qui évoluent le plus vite. Dans la mesure du possible nous incluons donc des liens vers les documentations sur les modèles des différents fournisseurs sur la version électronique de ce document.

Question : Votre modèle peut-il être requêté via une API ? Ou une interface utilisateur ?						
Anthropic	**DiNum**	**Google**	**LightOn**	**Meta**	**Microsoft (incluant OpenAI)**	**Mistral**
Oui Claude et API	**Oui** Une solution clé en main (front) et API	**Oui** Gemini et API	**Oui** Paradigm (pas d'API)	N/A	**Oui** Copilot et API Azure	**Oui** Le Chat et API
Question : Votre modèle s'intègre-t 'il déjà dans des environnements de développement prêts à l'emploi ?						
Anthropic	**DiNum**	**Google**	**LightOn**	**Meta**	**Microsoft (incluant OpenAI)**	**Mistral**
Oui Amazon Bedrock	**Oui** *Cloud* privé commercial et souverain (et *on premise*)	**Oui** Vertex AI	**Oui** Paradigm	N/A	**Oui** AI Studio / Pro-code, Copilot Studio / Low-code	**Oui** Azure, AWS, Snowflake
Question : Quelle est votre fréquence de mise à jour et évolution des modèles ?						
Anthropic	**DiNum**	**Google**	**LightOn**	**Meta**	**Microsoft (incluant OpenAI)**	**Mistral**
Pas d'élément trouvé	N/A	Environ 2 fois par an Liste des modèles.	N/A	**N/A**	Selon les releases d'OpenAI	N/A

Question : Quel est le nombre maximal de tokens dans les prompts/réponses et vos projections futures ?						
Anthropic	**DiNum**	**Google**	**LightOn**	**Meta**	**Microsoft (incluant OpenAI)**	**Mistral**
+200k tokens	N/A	PaLM => 8k ou 32K Gemini 1.0 => 32K Gemini 1.5 => 1 million	N/A	N/A	Selon les modèles d'OpenAI	N/A
Question : Quel est le nombre maximal de requêtes par jour ?						
Anthropic	**DiNum**	**Google**	**LightOn**	**Meta**	**Microsoft (incluant OpenAI)**	**Mistral**
Quotas	N/A	Quotas Dépend des modèles et de la localisation des *endpoints*	N/A	N/A	Quotas, extension possible	N/A
Question : Y a-t-il des limites particulières au modèle en termes d'augmentation du nombre d'utilisateurs ou du nombre de requêtes (et des limites par jour ?) ?						
Anthropic	**DiNum**	**Google**	**LightOn**	**Meta**	**Microsoft (incluant OpenAI)**	**Mistral**
Quotas	N/A	Quotas Possibilité de demander plusieurs milliers d'appels par minute.	N/A	N/A	Quotas, extension possible	N/A
Question : Avez-vous déjà publié des modèles multimodaux ?						
Anthropic	**DiNum**	**Google**	**LightOn**	**Meta**	**Microsoft (incluant OpenAI)**	**Mistral**
Oui Claude 3 Vision	**Non** Mais à venir (son, image, vidéo)	**Oui** Gemini est nativement multi modal : texte, image, vidéo et sons (mp3)	**Non**	**Non**	**Oui** Dall-E, GPT4-Vision	**Non**

Question : Certains de vos modèles sont-ils spécialisés dans certains domaines ou tâches ? Si oui lesquels ?						
Anthropic	**DiNum**	**Google**	**LightOn**	**Meta**	**Microsoft (incluant OpenAI)**	**Mistral**
Pas d'élément trouvé	**Oui** Accès services publics.	**Oui** MedLM => Santé SecPalm => Cyber Sécurité Imagen => Génération d'images Codey => Génération / explication de code	**Non**	Pas d'élément trouvé	**Oui** Exemple : orca-math)	**Non**
Question : Quelles sont les possibilités de fine tuning de vos modèles ?						
Anthropic	**DiNum**	**Google**	**LightOn**	**Meta**	**Microsoft (incluant OpenAI)**	**Mistral**
Aucune	**Aucune** A venir pour le RAG mais à date cela reste entre les mains du Datalab	**Plusieurs** Supervised tuning (PEFT), RLHF, Model distillation	**Plusieurs** Fine tuning et RAG	**Plusieurs** RHLF	**Plusieurs sur certains modèles** Voir les informations	**Plusieurs** Modèles fine tunables *on premise*. Bientôt sur leur plateforme et sur les *clouds*
Question : Le modèle est-il intégré à des solutions logicielles (Copilot, …) ?						
Anthropic	**DiNum**	**Google**	**LightOn**	**Meta**	**Microsoft (incluant OpenAI)**	**Mistral**
Oui Dans Google Sheets	N/A	**Oui** Gemini 1.0 : Latence et pricing model, multimodalité : Image ET Videos, Accuracy	N/A	Pas d'élément trouvé	**Oui** GPT4, GPT4-Turbo dans copilot	N/A
Question : Quelle valeur ajoutée par rapport aux autres modèles ? Qu'est-ce qui démarque ?						
Anthropic	**DiNum**	**Google**	**LightOn**	**Meta**	**Microsoft (incluant OpenAI)**	**Mistral**
Pas d'élément trouvé	Produit souverain et fine-tuné sur les services publics	Gemini 1.0 : Latence et pricing model, multimodalité : Image ET Videos, Accuracy Gemini 1.5 : contexte de 1M de token	Au-delà du modèle, il y a la plateforme Paradigm qui permet d'utiliser différents modèles *open source*	Pas d'élément trouvé	Pas de réponse spécifique fournie par le fournisseur de modèles	Coût, efficience des modèles, haute performance, déployable sur tous les *clouds*, déployable *on premise*

PARTIE 4.4 - Infrastructure

Pour cette thématique, seuls les modèles propriétaires sont concernés. En effet, les modèles *open source* doivent être hébergés et utilisés sur la propre infrastructure d'entreprise ou sur des infrastructures tierces.

Ainsi, les réponses apportées ne le sont que pour les modèles propriétaires interrogés soit **Anthropic, Google, Microsoft (incluant OpenAI).**

Question : Le modèle est-il utilisable sur des plateformes Cloud (cloud public ou privé voire souverain) ?		
Anthropic	**Google**	**Microsoft (incluant OpenAI)**
Oui Amazon Bedrock	**Oui**	**Oui** Public pour OpenAI, les SLM (Phi3 peuvent être déployés n'importe où)
Question : Quelle est la configuration minimale pour faire tourner le modèle ?		
Anthropic	**Google**	**Microsoft (incluant OpenAI)**
Pas d'élément trouvé	Modèles managés par Google, intégration par API, pas de configuration nécessaire	*Model as a Service*, pas d'infrastructure propre à prévoir
Question : Sur quelles plateformes votre modèle est-il déjà disponible ? utilisable ?		
Anthropic	**Google**	**Microsoft (incluant OpenAI)**
Amazon Bedrock	API requêtable depuis n'importe où	Azure
Question : Un plan entreprise (prix de groupe, sécurisation des données commerciales) est-il possible ?		
Anthropic	**Google**	**Microsoft (incluant OpenAI)**
Oui Pricing	**Oui** Par défaut et obligatoire	**Oui** PTU (Here) – *Price* & Perf
Question : Quelles sont les SLAs du service ?		
Anthropic	**Google**	**Microsoft (incluant OpenAI)**
Pas d'élément trouvé	SLA of Vertex AI (Google Cloud AI/ML Platform)	Tous les SLA sont disponibles

PARTIE 4.5 - Business Model

Encore une fois, seuls les modèles propriétaires ont été comparés pour cette thématique. En effet, les modèles *open source* ont des coûts internes liés à l'infrastructure utilisée, la puissance de calcul nécessaire, mais pas directement liés à leur utilisation.

Ainsi, les réponses apportées ne le sont que pour les modèles propriétaires interrogés soit **Anthropic, Google, Microsoft (incluant OpenAI).**

Question : Quels modes de tarification proposez-vous (essai ? au prompt ? au nombre de tokens ? …) ?		
Anthropic	**Google**	**Microsoft (incluant OpenAI)**
Par Token	Par Nombre de caractères	Par Token ou prix fixe avec PTU
Question : Y a-t-il d'autres coûts à prévoir (maintenance, évolution, …) ?		
Anthropic	**Google**	**Microsoft (incluant OpenAI)**
Pas d'élément trouvé	Non	Non

PARTIE 4.6 - Accompagnement des clients

Pour cette thématique, tous les fournisseurs de modèles ont été consultés. En effet, aussi bien les modèles propriétaires *open source* que propriétaires peuvent être proposés avec des tutoriels, des formations et peuvent s'organiser autour de communautés d'utilisateurs. L'idée ici est bien de déterminer de quel type d'accompagnement peut bénéficier l'entreprise utilisatrice en fonction des fournisseurs.

Question : Réalisez-vous des formations sur l'utilisation de vos modèles ? Si oui sous quelles conditions ?						
Anthropic	**DiNum**	**Google**	**LightOn**	**Meta**	**Microsoft (incluant OpenAI)**	**Mistral**
Pas d'élément trouvé	**Oui** Pour les services internes via le Datalab. A noter que le Datalab est une entité sur l'innovation et non la production donc cela pourrait évoluer.	**Oui**	**Oui** Formations et workshops sur l'utilisation de Paradigm inclus pour les clients LightOn	Pas d'élément trouvé	**Oui** Microsoft Learn	**Oui** Documentation sur le site et vidéos sur YouTube
Question : Avez-vous une communauté d'entraide active sur votre modèle ? Forum d'entraide ?						
Anthropic	**DiNum**	**Google**	**LightOn**	**Meta**	**Microsoft (incluant OpenAI)**	**Mistral**
Pas d'élément trouvé	**Oui** Même réponse que précédem-ment	**Oui** Plusieurs	**Oui** Echanges en 1-1 avec les clients et parfois des séminaires online	**Oui** AI-Research Community	**Oui** Microsoft Learn	**Oui** Chaine Discord Mistral AI
Question : Avez-vous des services d'accompagnement à l'installation ou l'utilisation, y compris support ? Si oui sous quelles conditions ?						
Anthropic	**DiNum**	**Google**	**LightOn**	**Meta**	**Microsoft (incluant OpenAI)**	**Mistral**
Pas d'élément trouvé	**Oui** Même réponse que précédemment	**Oui**	**Oui** Support et accompagnement pour les clients de LightOn. Rien pour les modèles *open source*	Pas d'élément trouvé	**Oui** Microsoft Learn	**Oui** Chaine Discord Mistral AI

PARTIE 4.7 - Considérations écologiques

Pour cette thématique, le sujet étant un enjeu majeur, tous les fournisseurs de modèles ont été interrogés. Le but était de savoir quelle était leur approche sur la question et de pouvoir collecter les éléments pertinents sur le sujet pour pouvoir les exposer au travers de liens.

Question : Mesurez-vous l'impact écologique de vos modèles ? Sur quels aspects (énergie, eau, impact carbone, ...) ?						
Anthropic	**DiNum**	**Google**	**LightOn**	**Meta**	**Microsoft (incluant OpenAI)**	**Mistral**
Pas d'élément trouvé	**Oui** Un papier va prochaine-ment sortir sur le sujet. Bonnes pratiques d'IA frugales mises en place comme la mise en place de petits modèles spécialisés et orchestrés en fonction de leur expertise, afin d'éviter l'utilisation d'un seul gros modèle énergivore.	**Oui**	**Oui** Aspect carbone	Pas d'élément trouvé	**Oui**	**Oui** Consommation électrique des clusters
Question : Quels documents ou chiffres pouvez-vous communiquer sur les consommations en carbone, en électricité ou en eau sur la phase d'entraînement ?						
Anthropic	**DiNum**	**Google**	**LightOn**	**Meta**	**Microsoft (incluant OpenAI)**	**Mistral**
Pas d'élément trouvé	Même réponse que précédem-ment	Apprentis-sage réalisé avec des TPU, 90% via énergies renouvela-bles	Une étude avait été réalisée ici	Pas d'élément trouvé	Emission Impact Dashboard	Pas de document fourni mais une étude en cours de réalisation

Question : Quels documents ou chiffres pouvez-vous communiquer sur les consommations en carbone, en électricité ou en eau sur la phase d'inférence ?						
Anthropic	**DiNum**	**Google**	**LightOn**	**Meta**	**Microsoft (incluant OpenAI)**	**Mistral**
Pas d'élément trouvé	Utilisation de modèle open source SLM avec de la quantisation. Possibilité pour rappel de faire tourner certains modèles sur un Macbook Pro M3.	Les impacts en émission de CO2 de l'utilisation de la totalité de la plateforme sont offertes	Une étude avait été réalisée ici	Pas d'élément trouvé	Emission Impact Dashboard	Pas de document fourni mais une étude en cours de réalisation

HUB FRANCE IA

Conclusion

Conclusion

Alors que les fournisseurs d'IA générative sont déjà nombreux, il est crucial d'aider les organisations à faire leurs choix dans cette offre. Comme l'a montré l'enquête réalisée auprès des organisations utilisatrices, certains critères non liés aux performances sont importants. C'est notamment le cas de la sécurité des données, du respect des réglementations, de l'adaptation facile à l'infrastructure ou encore du budget à prévoir.

A l'heure actuelle, il est possible de comparer les modèles sur les aspects liés aux performances. Ce document liste les comparatifs de référence (dits « *benchmarks* ») et décrypte les différents critères qu'ils utilisent. Ceci fournit alors les clés de lecture pour identifier par rapport aux cas d'usage pressentis quels *benchmarks* utiliser pour comparer les modèles qui intéressent plus particulièrement. Comme on l'aura certainement compris, nous suggérons également de mener sur les quelques modèles qui pourraient convenir des tests complémentaires pour valider leur pertinence.

Malgré l'existence de ces nombreux *benchmarks*, nous avons remarqué que des éléments clés exprimés par les organisations utilisatrices dans notre enquête sont difficiles voire impossibles à trouver sans contact direct avec les fournisseurs. C'est pourquoi nous avons programmé des échanges avec ces derniers pour pouvoir condenser et retranscrire les informations relatives à ces critères importants. Comme on l'aura constaté dans la partie 3, nous avons contacté tous les principaux acteurs du domaine et avons restitué dans la partie 4 les informations collectées, avec le concours des acteurs qui ont accepté de se prêter à l'exercice.

Bien sûr, il est impossible de dire quel est le meilleur fournisseur de modèles de façon globale : il n'existe pas d'acteur qui se démarque sur tous les points. En revanche, la bonne façon d'utiliser ce livrable est, en partant d'un cas d'usage, de lister les critères importants associés, comme indiqué dans la partie 1.3 puis de regarder les *benchmarks* existants pertinents de la partie 2 avant de finaliser l'analyse en inspectant les critères listés en partie 4. Cela devrait permettre à l'organisation d'éclairer son choix.

Nous avons pris beaucoup de plaisir à réaliser ce livrable et mener les échanges avec les différents acteurs de l'IAG. Ce sujet évoluant très vite, nous avons pris le parti de fournir un maximum de liens cliquables dans la version numérique de ce livrable. Les liens de la partie 2 pointent vers les pages pertinentes et à jour des différents *benchmarks*. Les liens de la partie 4 pointent vers les pages pertinentes sur les sites web des différents fournisseurs lorsqu'elles nous ont été communiquées. Nous encourageons le lecteur à les utiliser pour faire perdurer la pertinence de ce qui a été écrit dans ce document.

REMERCIEMENTS

Le Hub France IA remercie l'ensemble des participants au groupe de travail IAG, et tout particulièrement les contributeurs de ce livrable. Des remerciements sont aussi adressés à tous les acteurs contactés qui ont accepté de nous accorder du temps et de répondre à nos questions.

Le coordinateur

- **Jean-François DELDON** – Yakadata

Les contributeurs

- **Belkacem LAÏMOUCHE** – Direction Générale de l'Aviation Civile (DGAC)
- **Ophélie GUENOUX**
- **Laurence RELMY**
- **Sacha MARTINI** – FFB Occitanie
- **Luc TRUNTZLER** – Spoon AI
- **Bertrand LAFFORGUE** – Konverso
- **Kevin PACI** – Mediaco Vrac
- **Kajetan WOJTACKI** – DecisionBrain
- **Jean-François DELDON** – Yakadata
- **Miguel SOLINAS** – Neovision
- **Marie-Aude AUFAURE** – Datarvest
- **Françoise SOULIE-FOGELMAN** – Conseiller Scientifique – Hub France IA

Les relecteurs

- **Françoise SOULIE-FOGELMAN** – Conseiller Scientifique – Hub France IA
- **Cyril NICOLOTTO** – Chef de projet – Hub France IA

La touche finale

- **Mélanie ARNOULD** – Responsable des opérations – Hub France IA

CHOISIR UN MODELE D'IA GENERATIVE POUR SON ORGANISATION

Juin 2024

HUB FRANCE IA