

Anticiper l'évolution des capacités des IA à usage général

Des progrès rapides sur toutes les mesures de performance

Cette note vise à répondre à trois questions :

- Comment vont évoluer les performances des modèles d'IA à usage général d'ici 2030 ?
- En particulier, quand émergeront des capacités posant des problèmes de sécurité et de contrôle ?
- À quel point l'hétérogénéité des progrès observés est-elle une barrière au développement d'une AGI ?

I. Comment vont évoluer les performances des modèles d'IA à usage général d'ici 2030 ?

La décennie actuelle est marquée par le développement de systèmes d'intelligence artificielle à usage général (General-Purpose AI, GPAI), dont les performances progressent rapidement sur un large ensemble de tâches cognitives. Ces progrès restent toutefois très hétérogènes, avec d'un côté des GPAI qui dépassent les experts humains dans plusieurs disciplines, et de l'autre des limites encore visibles sur certaines tâches élémentaires. Cette combinaison rend l'évaluation des capacités actuelles et futures des GPAI particulièrement délicate et appelle une analyse structurée de leurs trajectoires d'évolution.

En particulier, les performances mesurées à un instant donné ne constituent pas, à elles seules, une base suffisante pour anticiper les capacités susceptibles d'émerger à court ou moyen terme, notamment celles pouvant poser des problèmes de sécurité et de contrôle¹. Il est donc indispensable de se concentrer, au-delà des capacités actuelles, sur la dynamique d'évolution des GPAI de frontière. **Cette note adopte ainsi une approche prospective consistant à analyser les trajectoires de progression observées sur soixante benchmarks standardisés, afin d'estimer l'évolution probable des capacités des GPAI dans les années à venir.**

L'objectif est d'identifier les tendances générales, les points de saturation des évaluations actuelles, ainsi que les limites intrinsèques de ces benchmarks pour éclairer l'interprétation des performances observées et leur portée pour l'évaluation du rapprochement vers des systèmes pouvant être qualifiés d'intelligence artificielle générale (AGI).

1. Contexte

a. Mesurer les capacités des IA

Depuis ses débuts, la communauté de l'intelligence artificielle a cherché à mesurer objectivement les progrès réalisés par les systèmes d'IA. Dès les années 1950, les pionniers du domaine ont proposé des tests pour évaluer les capacités des machines, le plus célèbre étant le test de Turing² qui visait à

¹ Cf. notre note précédente : "Artificial General Intelligence (AGI) : Anticipation des objectifs et implications pour la sécurité et le contrôle" (2025).

² Turing. "Computing Machinery and Intelligence." *Mind* (1950). [10.1093/mind/LIX.236.433](https://doi.org/10.1093/mind/LIX.236.433)

mesurer la capacité d'une machine à exhiber un comportement indistinguable de celui d'un humain à travers des interactions écrites. Cette approche s'est depuis considérablement sophistiquée avec le développement de benchmarks spécialisés.

Un benchmark est un ensemble standardisé de tâches conçues pour évaluer les performances d'un système d'IA de manière reproductible (voir l'encadré ci-dessous). Ces outils sont utilisés pour suivre les progrès de l'IA au fil du temps, fixer des objectifs concrets aux développeurs, comparer les performances de différentes approches, entraîner les modèles. Plus récemment, certains benchmarks ont été spécifiquement conçus pour évaluer des compétences critiques (piratage informatique, conception d'armes biologiques ou chimiques, accélération de la recherche en IA, etc.) ou la propension des modèles à présenter des comportements non-souhaitables (mensonge, tricherie, etc.), de manière à pouvoir anticiper d'éventuels problèmes de sécurité³.

Encadré 1 - Les benchmarks

Un *benchmark* est un ensemble de tâches standardisées utilisé pour évaluer les performances des systèmes d'IA. Il comprend généralement : (1) un corpus de questions ou de problèmes à résoudre, (2) une procédure d'évaluation, par exemple des solutions de référence considérées comme correctes, et (3) une métrique permettant de quantifier la performance (typiquement un pourcentage de bonnes réponses).

Les benchmarks peuvent mesurer des compétences très diverses : compréhension du langage naturel (*MMLU*, *HellaSwag*), raisonnement logique (*ARC-AGI*, *EnigmaEval*), capacités mathématiques (*MATH*, *FrontierMath*), programmation (*Aider Polyglot*, *SWE-Bench*), ou encore des compétences multimodales pour l'image, son ou vidéo (*GeoBench*, *VisualToolBench*). Certains benchmarks plus récents visent spécifiquement à tester des capacités considérées comme critiques, comme la cybersécurité (*Cybench*) ou les connaissances scientifiques avancées pouvant être utilisées de manière dangereuse⁴ (*LAB-Bench*, *WMDP*).

Les benchmarks présentent toutefois plusieurs limites. Par construction, ils forcent à restreindre l'évaluation à des tâches facilement spécifiables et évaluables (par exemple avec des réponses simples et uniques), très souvent peu réalistes par rapport aux situations complexes du monde réel. Par ailleurs, la incitations économiques à obtenir de meilleurs scores que les modèles concurrents sur les benchmarks les plus suivis pousse les développeurs de GAI de frontière à optimiser leurs modèles sur ces benchmarks spécifiques⁵. De même pour le problème connexe de la contamination des données d'entraînement : si les questions des benchmarks se retrouvent dans les corpus d'entraînement des modèles, les scores obtenus ne reflètent plus une véritable capacité de généralisation mais plutôt de la mémorisation⁶. Enfin, l'interprétation des scores est limitée par le fait que les évaluations ne parviennent jamais à mesurer les capacités maximales qu'aurait un modèle

³ Cf. notre note précédente : "Revue des Frontier AI Safety Frameworks pour se prémunir des enjeux de *Loss of Control*" (2025).

⁴ Dev et al. "Toward Comprehensive Benchmarking of the Biological Knowledge of Frontier Large Language Models" RAND Corporation (2025). https://www.rand.org/pubs/research_reports/RRA3797-1.html.

⁵ Ce fut le cas pour le modèle LLama 4, dont les performances partagées par Meta provenaient d'une variante du modèle optimisée sur ces benchmarks. "Meta got caught gaming AI benchmarks." *The Verge* (2025).

<https://www.theverge.com/meta/645012/meta-llama-4-maverick-benchmarks-gaming>. Plus largement, ce phénomène illustre la loi de Goodhart : lorsqu'une mesure devient un objectif, elle cesse d'être une bonne mesure, et le résultat obtenu a tendance à diverger drastiquement des buts souhaités. Nous avons détaillé les conséquences de la loi de Goodhart sur la sécurité et le contrôle des GAI dans une note précédente et dans Maier, Maier & David. "Take Goodhart Seriously: Principled Limit on General-Purpose AI Optimization." *ArXiv* (2025). [10.48550/arXiv.2510.02840](https://arxiv.org/abs/2510.02840).

⁶ Les benchmarks bien conçus limitent ce problème en conservant une partie de leurs jeux de données privés ou en renouvelant régulièrement les questions.

dans des conditions optimales (par exemple avec le bon cadre et les bons outils à disposition). On parle d'écart d'élicitation (*elicitation gap*) pour représenter cette différence entre les compétences latentes d'un modèle, et les compétences révélées par ses scores sur des benchmarks⁷. Malgré tout, les benchmarks restent parmi les meilleurs outils à notre disposition pour suivre et anticiper les progrès rapides en IA.

Au cours des dernières années, on observe une véritable course au développement de nouveaux benchmarks, pour évaluer des compétences de plus en plus sophistiquées. Cette accélération est due au fait que les benchmarks sont saturés (c'est-à-dire résolus) de plus en plus rapidement par les modèles de pointe. La Figure 1 illustre cette dynamique : de nombreux benchmarks récents ont atteint des niveaux de performance dépassant des panels humains seulement quelques années après leur publication. **Des tâches jugées très difficiles il y a encore trois ou quatre ans sont aujourd'hui résolues avec un taux de succès dépassant celui d'un adulte moyen, voire d'experts dans leur domaine⁸.**

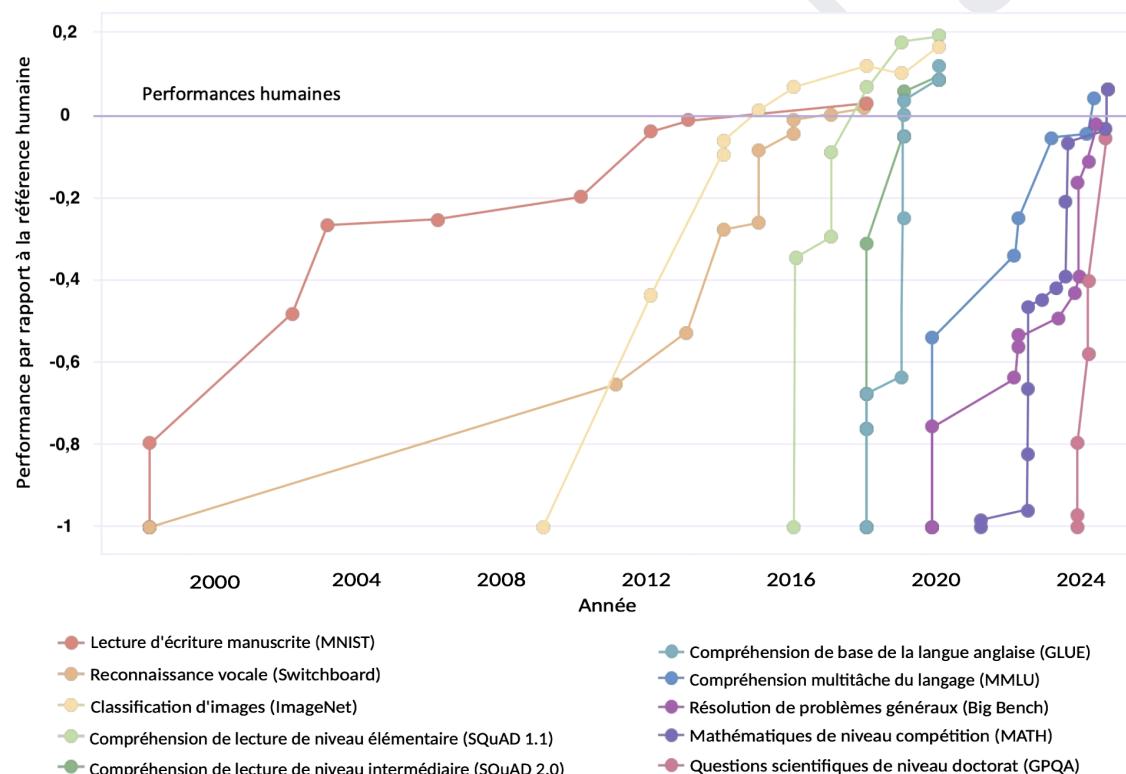


Figure 1 - De nombreux benchmarks récents ont dépassé les scores humains en quelques années.

Pour cette figure les benchmarks sont standardisés entre -1 (le meilleur score par un modèle dans la publication d'origine) et 0 (les performances humaines rapportées par les études d'origine). La référence humaine correspond au score obtenu par des non-experts pour les premiers benchmarks, puis par des experts du domaine pour les benchmarks plus récents et difficiles. Un score standardisé de 0,2 correspond ainsi à un modèle 20% meilleur que le panel humain. Adapté de Bengio et al. (2025).⁹

Parmi les benchmarks particulièrement difficiles récemment développés, on peut citer *Humanity's Last Exam*, qui compile des questions de très haute difficulté dans divers domaines académiques, ou

⁷ Weij et al. "AI Sandbagging: Language Models can Strategically Underperform on Evaluations." ArXiv (2024). [10.48550/arXiv.2406.07358](https://arxiv.org/abs/2406.07358).

⁸ Dev et al. "Toward Comprehensive Benchmarking of the Biological Knowledge of Frontier Large Language Models" RAND Corporation (2025). https://www.rand.org/pubs/research_reports/RRA3797-1.html.

⁹ Bengio et al. "International AI Safety Report." (2025). <https://www.gov.uk/government/publications/international-ai-safety-report-2025>

encore *FrontierMath*, qui présente des problèmes mathématiques originaux non résolus nécessitant des raisonnements avancés (ils prennent plusieurs heures à plusieurs jours à des mathématiciens pour les résoudre). Ces benchmarks n'ont pas encore été complètement résolus par les meilleurs modèles mais, comme nous allons le voir, la trajectoire observée suggère qu'ils le seront dans les deux à trois prochaines années.

b. Projeter l'évolution des capacités

L'évolution observée des performances appelle le développement de méthodes permettant d'en analyser et d'en projeter les trajectoires. La prévision quantitative (*forecasting*) des scores sur les benchmarks est un champ de recherche actif, et historiquement difficile (cf. Encadré 2). Un de ses résultats, pointé par plusieurs travaux récents, est que les progrès sur des benchmarks très divers peuvent se décomposer en¹⁰ :

- l'amélioration sur un axe principal de compétence générale¹¹ ; si un modèle est très bon sur une compétence, il est généralement aussi très bon sur toutes les autres compétences,
- l'amélioration sur diverses compétences spécialisées (code, mathématiques, suivi d'instructions). Par exemple, les modèles d'OpenAI sont meilleurs en mathématiques, tandis que les modèles d'Anthropic sont meilleurs en programmation¹².

Cette observation suggère qu'il existe une dynamique commune sous-jacente au développement des capacités d'IA, ce qui rend possible la projection de trajectoires futures¹³. Tout exercice de mesure et de projection reste toutefois imparfait : aucun ensemble de benchmarks ne capture la diversité des compétences ou des conditions d'usage réelles, et aucun modèle de prévision ne peut intégrer toutes les dynamiques impliquées. Néanmoins, **l'existence d'incertitudes inhérentes ne remet pas en cause la question examinée ici, à savoir si l'émergence de systèmes de GPAI présentant des capacités supérieures à celles d'experts humains constitue un scénario crédible dans les années à venir**. L'enjeu est moins d'assigner des dates précises que de quantifier notre niveau d'incertitude, et d'évaluer la plausibilité d'attendre ces seuils de capacités à court et moyen terme.

La méthodologie adoptée pour cette note consiste à analyser 60 benchmarks couvrant un large éventail de capacités cognitives et opérationnelles des modèles de GPAI, et régulièrement mis à jour. Nous introduisons trois éléments par rapport aux approches classiques : (i) l'usage d'un modèle de projection plus flexible que les courbes standards, afin de mieux représenter l'incertitude sur les trajectoires de progrès ; (ii) un recentrage sur les capacités maximales des modèles de frontière, plutôt que sur leurs performances moyennes observées, ce qui est plus pertinent pour l'analyse des enjeux de sécurité ; et (iii) une approche de projection conjointe qui utilise les benchmarks les plus complets pour informer les benchmarks dont les données sont limitées. Les détails de la méthodologie sont consultables en annexe de ce document.

¹⁰ Ruan et al. "Observational Scaling Laws and the Predictability of Language Model Performance." *NeurIPS* (2024). <https://arxiv.org/abs/2405.10938> ; Maia Polo et al. "Sloth: Scaling Laws for LLM Skills to Predict Multi-benchmark Performance Across Families." *NeurIPS* (2025). <https://arxiv.org/abs/2412.06540>.

¹¹ Ilić & Gignac. "Evidence of interrelated cognitive-like capabilities in large language models: Indications of artificial general intelligence or achievement?" *Intelligence* (2024). [10.1016/j.intell.2024.101858](https://doi.org/10.1016/j.intell.2024.101858) ; Kipnis et al. "metabench -- A Sparse Benchmark of Reasoning and Knowledge in Large Language Models." *ICLR* (2025). <https://arxiv.org/abs/2407.12844> ; Ho et al. "A Rosetta Stone for AI Benchmarks." *ArXiv* (2025). [10.48550/arXiv.2512.00193](https://arxiv.org/abs/2512.00193).

¹² Burnham et al. "Benchmark Scores = General Capability + Claudiness." *Epoch AI* (2025). <https://epoch.ai/gradient-updates/benchmark-scores-general-capability-claudiness>.

¹³ Pimpale et al. "Forecasting Frontier Language Model Agent Capabilities." *ArXiv* (2025). [10.48550/arXiv.2502.15850](https://arxiv.org/abs/2502.15850).

Encadré 2 - Historique de sur- et sous-estimation de la vitesse de progrès de l'IA

L'histoire de l'IA alterne entre sur-optimisme et sous-estimation de la vitesse des progrès. Aux débuts du domaine, dans les années 1950-1960, les pionniers ont drastiquement surestimé la vitesse à laquelle les machines parviendraient à des niveaux d'intelligence humaine. Cette sur-estimation, fondée sur les paradigmes de l'époque (systèmes symboliques et approches logiques), a conduit à plusieurs "hivers de l'IA" quand les promesses ne se sont pas concrétisées.

Plus récemment, la tendance s'est inversée. Lorsque des prédictions quantitatives précises ont été formulées dans les années passées, tant les experts du domaine que les *forecasters* (prévisionnistes) ont systématiquement eu tendance à sous-estimer la vitesse des progrès¹⁴. Les modèles ont notamment atteint des niveaux avancés en mathématiques bien plus rapidement que prévu (Figure 2), et ont aussi dépassé les attentes dans des domaines de recherche duale à risque comme la virologie¹⁵.

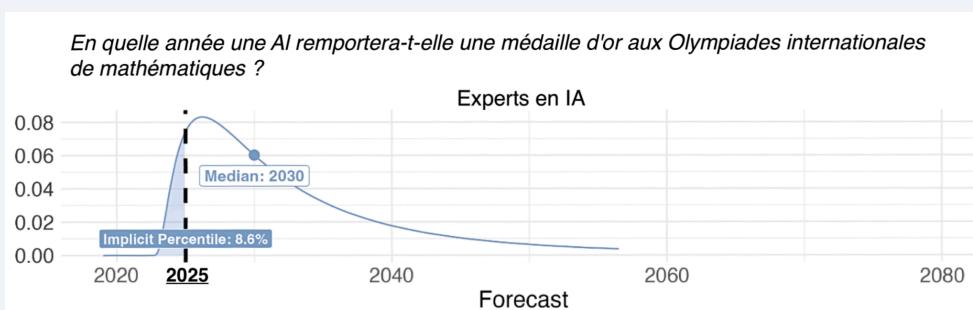


Figure 2 – Exemple de sous-estimation des progrès de l'IA en mathématiques. En 2022, le *Forecasting Research Institute* a organisé une série de prévisions incluant des avancées en IA. Les experts prédisaient une médaille d'or aux Olympiades internationales de mathématiques autour de 2030, attribuant 8,6% de chance avant 2025. Elle a été pourtant atteinte cette année-là par des modèles de Google DeepMind et OpenAI. Adapté de Kučinskas et al. (2025)¹⁶.

Cette sous-estimation récurrente s'explique en partie par la difficulté à anticiper les capacités émergentes qui apparaissent de manière parfois abrupte lorsque la taille des modèles et des données d'entraînement augmentent¹⁷. Il est donc important de prendre en compte ce biais vers la sous-estimation et d'adopter des marges de sécurité appropriées dans les prévisions.

¹⁴ Steinhardt. "AI Forecasting: One Year In." *Bounded Regret* (2022).

<https://bounded-regret.ghost.io/ai-forecasting-one-year-in/>

¹⁵ Williams et al. "Forecasting Biosecurity Risks from LLMs." *Forecasting Research Institute* (2025).

<https://forecastingresearch.org/ai-enabled-biorisk>. La conclusion du rapport fait l'observation suivante: "La prévision médiane des experts indique que si l'IA atteignait des seuils de performance spécifiques, comme égaler des équipes d'experts sur un test de résolution de problèmes en virologie, le risque annuel d'une épidémie causée par l'homme provoquant plus de 100 000 décès passerait de 0,3% à 1,5%. Cependant, les experts et superforecasters sous-estiment significativement les progrès de l'IA, prévoyant que ces capacités n'émergeront pas avant 2030 au plus tôt. En réalité, de nouvelles recherches menées en collaboration avec SecureBio indiquent que certaines de ces capacités ont déjà été atteintes."

¹⁶ Kučinskas et al. "Assessing Near-Term Accuracy in the Existential Risk Persuasion Tournament." *Forecasting Research Institute* (2025). <https://forecastingresearch.org/near-term-xpt-accuracy>.

¹⁷ Wei et al. "Emergent Abilities of Large Language Models." *ArXiv* (2022). [10.48550/arXiv.2206.07682](https://arxiv.org/abs/2206.07682)

2. Saturat^{ion} des benchmarks actuels avant 2030

Les projections issues de notre méthodologie convergent vers une conclusion : quasiment tous les benchmarks actuels, y compris les plus difficiles, devraient être saturés avant 2030¹⁸. La résolution de ces benchmarks ne signifie pas nécessairement l'atteinte d'une intelligence artificielle générale (AGI), mais indique que les modèles dépasseront les performances de la plupart des humains sur l'essentiel des tâches actuellement mesurables.

a. Résolution des benchmarks de sens commun

Parmi les premiers benchmarks largement saturés se trouvent ceux mesurant les connaissances de sens commun, c'est-à-dire la capacité à comprendre et raisonner dans des situations ordinaires. Encore récemment, les systèmes d'IA échouaient régulièrement sur des questions évidentes pour un enfant, un problème ciblé par des benchmarks comme *HellaSwag*¹⁹ (complétion de phrases dans des contextes de sens commun), *PIQA*²⁰ (raisonnement physique intuitif), ou *WinoGrande*²¹ (résolution d'ambiguités de pronoms). Leurs questions incluent : "Pour éviter les bulles dans un gâteau, est-ce qu'il faut taper le moule sur le plan de travail 1) avant la cuisson ou 2) après la cuisson ?", ou bien "Le mois dernier, Hélène est partie en vacances mais pas Christine car [Hélène / Christine] avait besoin de se reposer.". Ces benchmarks aujourd'hui sont saturés (Figure 3), avec des performances atteignant les niveaux humains de référence. **Les grands modèles de langage montrent ainsi une forme de compréhension implicite du monde physique et social.**

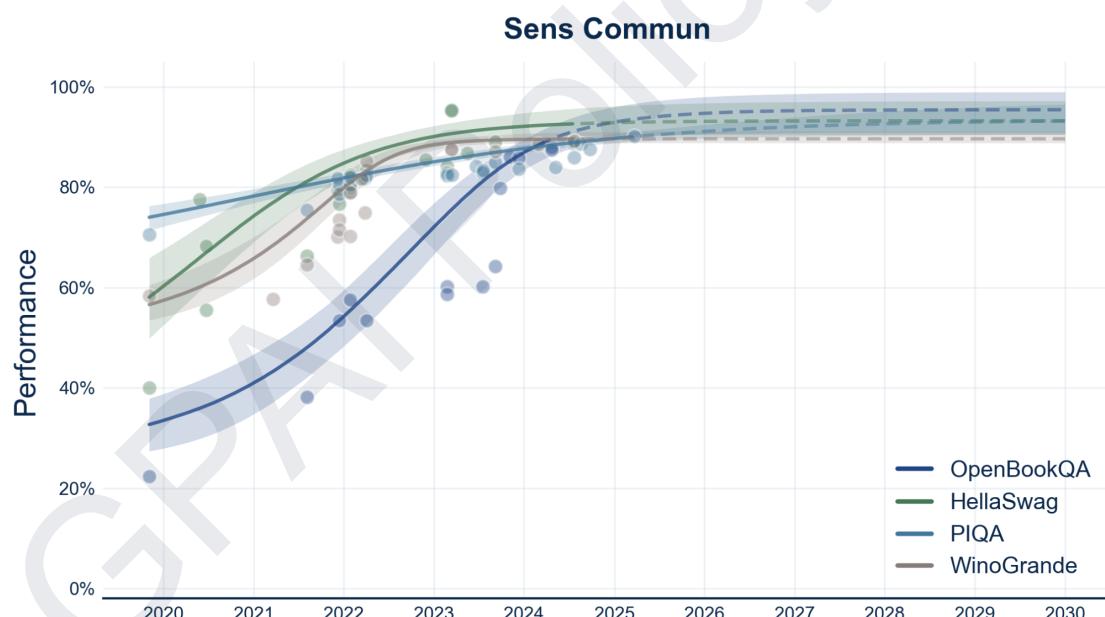


Figure 3 - Projections des performances d'IA sur des benchmarks de connaissances de sens commun. Les courbes représentent les trajectoires de progression estimées pour les benchmarks OpenBookQA, HellaSwag, PIQA et WinoGrande. Les points indiquent les scores empiriques à la frontière de performance (dans le top 3 à leur sortie), les traits pleins les médianes postérieures, et les zones ombrées les intervalles de crédibilité à 80%. Les projections en pointillés extrapolent jusqu'en 2030.

¹⁸ On considère ici qu'un modèle est saturé s'il a fait au moins 95% des progrès possibles, sur l'échelle entre le score obtenu par des réponses aléatoires et le score maximum estimé du benchmark. Dans notre modèle, près de 97% des benchmarks évalués atteignent ce seuil de saturation avant 2030 (cf. Annexe).

¹⁹ Zellers et al. "HellaSwag: Can a Machine Really Finish Your Sentence?" ACL (2019). [10.18653/v1/P19-1472](https://doi.org/10.18653/v1/P19-1472).

²⁰ Bisk et al. "PIQA: Reasoning about Physical Commonsense in Natural Language." AAAI (2020). [10.1609/aaai.v34i05.6239](https://doi.org/10.1609/aaai.v34i05.6239).

²¹ Sakaguchi et al. "WinoGrande: An Adversarial Winograd Schema Challenge at Scale." AAAI (2020). [10.1609/aaai.v34i05.6399](https://doi.org/10.1609/aaai.v34i05.6399).

b. Avancées rapides à travers les compétences et les modalités

Au-delà du sens commun, les progrès s'étendent à des domaines de plus en plus exigeants. Les benchmarks faisant appel à des capacités de raisonnement complexe ont notamment vu leurs scores augmenter particulièrement rapidement. Cette progression se traduit par une course au développement de benchmarks toujours plus difficiles.

On observe par exemple cette escalade continue de la difficulté des évaluations dans le domaine du raisonnement logique (Figure 4). Des benchmarks comme *Adversarial NLI*²² ou *SimpleBench*²³ ont été conçus pour tester la capacité des modèles à manipuler des concepts abstraits, ainsi que leur aptitude à résister à des formulations piégeuses, ambiguës ou adversariales. De nouveaux jeux de données comme *Balrog*²⁴ (raisonnement multi-étapes), *Chess Puzzles*²⁵ (problèmes d'échec) ou *EnigmaEval*²⁶ (longs puzzles multimodaux) augmentent encore la difficulté. **Les données empiriques et nos projections indiquent que les GPAI de frontière progressent sur les benchmarks de raisonnement de dizaines de points par an.**

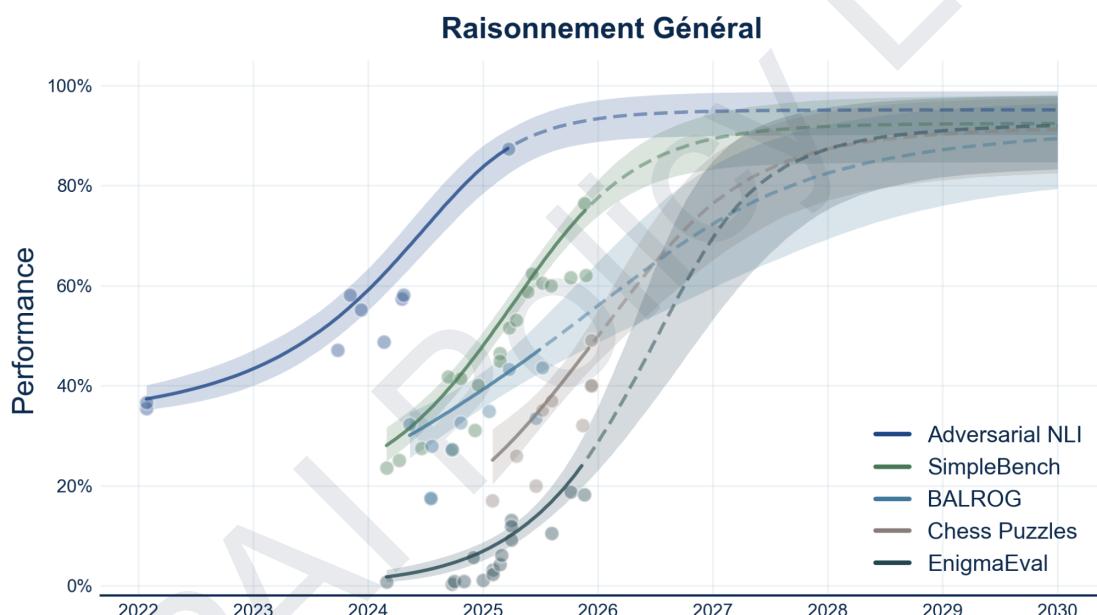


Figure 4 - Projections des performances d'IA sur des benchmarks de raisonnement. Conventions graphiques identiques à la Figure 3.

Les benchmarks spécialisés par discipline illustrent la même trajectoire (Figure 5). *ScienceQA*²⁷ et *AI2 Reasoning Challenge*²⁸ (ARC), parmi les premiers à tester les connaissances et le raisonnement scientifique élémentaire, sont aujourd'hui largement résolus. *GSM8K*²⁹, centré sur des problèmes

²² Nie et al. “Adversarial NLI: A New Benchmark for Natural Language Understanding.” ACL (2019).

[10.18653/v1/2020.acl-main.441](https://doi.org/10.18653/v1/2020.acl-main.441)

²³ “SimpleBench”. <https://simple-bench.com>.

²⁴ Paglieri et al. “BALROG: Benchmarking Agentic LLM and VLM Reasoning On Games.” ICLR (2024). [10.48550/arXiv.2411.13543](https://arxiv.org/abs/2411.13543).

²⁵ “Chess Puzzles.” Epoch AI. <https://epoch.ai/benchmarks/chess-puzzles>.

²⁶ “EnigmaEval.” Scale AI. https://scale.com/leaderboard/enigma_eval.

²⁷ Saikh et al. “ScienceQA: a novel resource for question answering on scholarly articles.” International Journal on Digital Libraries (2022). [10.1007/s00799-022-00329-v](https://doi.org/10.1007/s00799-022-00329-v).

²⁸ Clark et al. “Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge.” ArXiv (2018). [10.48550/arXiv.1803.05457](https://arxiv.org/abs/1803.05457).

²⁹ Cobbe et al. “Training Verifiers to Solve Math Word Problems.” ArXiv (2021). [10.48550/arXiv.2110.14168](https://arxiv.org/abs/2110.14168).

arithmétiques multi-étapes, est saturé à plus de 95%. Des benchmarks comme *GPQA*³⁰ (*Graduate-Level Google-Proof Q&A*), et son sous-ensemble encore plus difficile *GPQA Diamond*, mis au point par des experts pour tester une compréhension fine en biologie, physique et chimie, s'approchent également de la saturation. Les benchmarks récents ciblent des questions factuelles de plus en plus spécifiques (*SimpleQA Verified*³¹), ou bien font appel à des experts pour poser des problèmes pointus dans leur spécialité. C'est par exemple le cas pour les domaines financiers et légaux avec le *Professional Reasoning Benchmark* (*PRBench*³²), et dans une grande variété de disciplines (des sciences sociales aux mathématiques) avec *Humanity's Last Exam*³³.

Cette progression retrace une forme d'échelle académique : collège, lycée, licence, master en quelques années. Les modèles de raisonnement atteignent désormais des performances de niveau doctorat : sur *GPQA Diamond*, en 2024, des experts titulaires de doctorats dans les disciplines évaluées n'ont obtenu qu'un score de 69,7%, contre 78% pour le modèle o1³⁴ (Gemini 3 Pro atteint maintenant 93%). **On assiste à un dépassement des capacités humaines sur le terrain des connaissances, tous domaines confondus.**

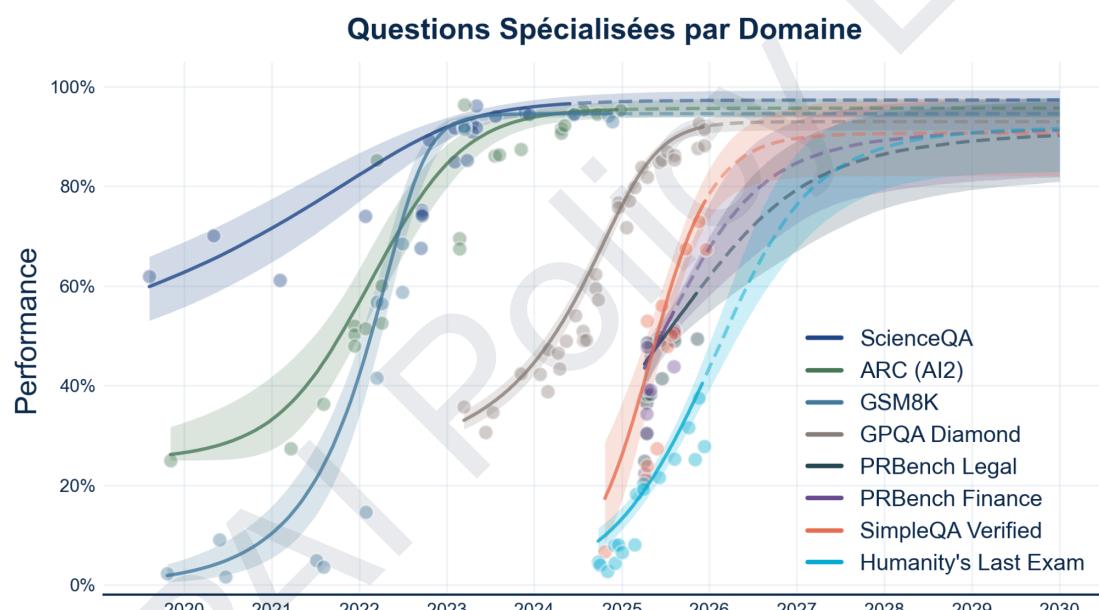


Figure 5 - Projections des performances d'IA sur des benchmarks spécialisés. Conventions graphiques identiques à la Figure 3.

L'entraînement des premiers grands modèles de langage (LLMs), consistant à prédire la suite de phrases issues d'internet, avait alimenté l'idée qu'ils étaient des "perroquets stochastiques"³⁵ reproduisant des motifs de texte sans vraiment comprendre. En réponse, de nombreux benchmarks textuels ont été développés pour distinguer cette simple forme de *pattern-matching* d'une compréhension plus profonde (*Fiction.liveBench*³⁶), dans plusieurs langues (*MultiNRC*³⁷), au cours de

³⁰ Rein et al. "GPQA: A Graduate-Level Google-Proof Q&A Benchmark." ArXiv (2023). [10.48550/arXiv.2311.12022](https://arxiv.org/abs/2311.12022).

³¹ "SimpleQA Verified." Epoch AI. <https://epoch.ai/benchmarks/simple-qa-verified>.

³² "Professional Reasoning Benchmark - Finance." Scale AI. <https://scale.com/leaderboard/prbench-finance> ; "Professional Reasoning Benchmark - Legal." Scale AI. <https://scale.com/leaderboard/prbench-legal>.

³³ "Humanity's Last Exam." Scale AI. https://scale.com/leaderboard/humanitys_last_exam.

³⁴ "L'apprentissage du raisonnement avec les LLM." OpenAI (2025) <https://openai.com/index/learning-to-reason-with-lms>.

³⁵ Bender et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜." Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (2021). [10.1145/3442188.3445922](https://doi.org/10.1145/3442188.3445922)

³⁶ "Fiction.liveBench." Epoch AI. <https://epoch.ai/benchmarks/fictionlivebench>.

³⁷ "MultiNRC." Scale AI. <https://scale.com/leaderboard/multinrc>.

conversations longues (*MultiChallenge*³⁸, *TutorBench*³⁹), ainsi que tester leur capacité à rédiger des textes cohérents et créatifs (*Lech Mazur Writing*⁴⁰). Pour prendre un exemple tiré de *MultiNRC* en français : “Si je suis un mot masculin seul, féminin au pluriel, je ne suis jamais bouclé. Qui suis-je?”⁴¹. Ces benchmarks sont aujourd’hui soit saturés, soit devraient l’être d’ici 2028 (Figure 6). **Les capacités de compréhension et de rédaction des modèles actuels démontrent une forme de maîtrise du langage qui se rapproche de celle d’humains éduqués.**

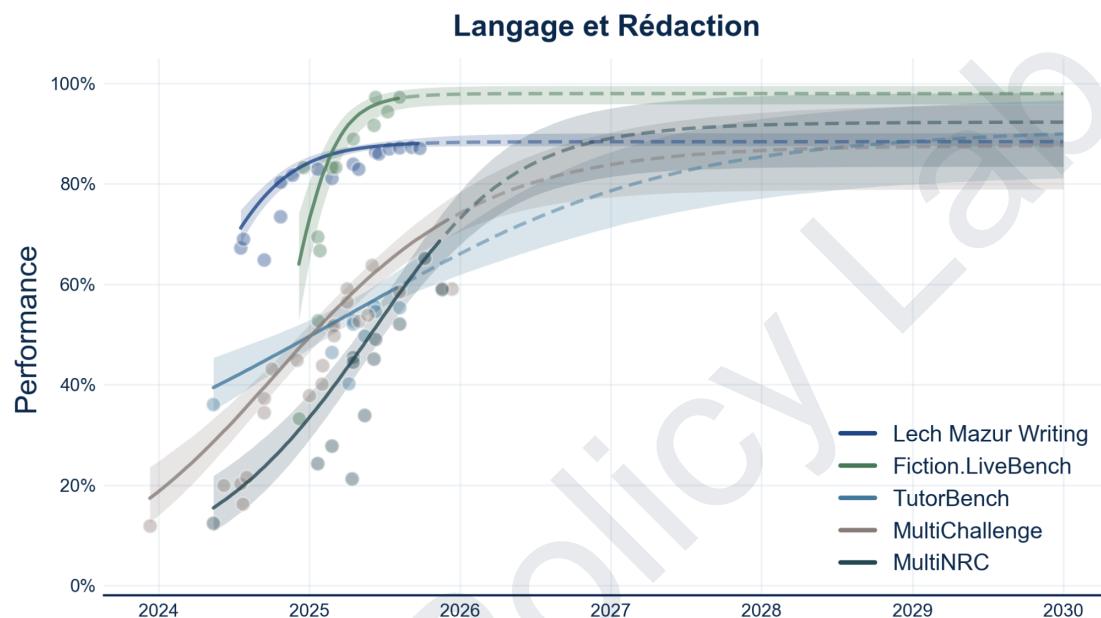


Figure 6 - Projections des performances d’IA sur des benchmarks de compréhension du langage et de rédaction. Conventions graphiques identiques à la Figure 3.

Les premiers LLMs étant basés uniquement sur du texte, une autre suspicion portait sur l’impossibilité de développer certaines capacités relevant de la perception visuelle, de la compréhension spatiale ou du traitement de l’information non textuelle. Cette question n’est plus d’actualité : les IA de frontière sont maintenant multimodales, capables de traiter et de générer du texte, des images, des vidéos et du son de manière intégrée. Les benchmarks qui mesurent cette maîtrise de la multimodalité progressent également rapidement. *CAD-Eval*⁴² pour la génération de formes 3D, *Visual Physics Comprehension Test (VPCT)*⁴³ pour la compréhension physique visuelle (prédir la trajectoire d’une balle sur une série de rampes), *GeoBench*⁴⁴ pour l’identification de lieux à partir de photos, *Visual Task Assessment (VISTA)*⁴⁵ pour les raisonnements combinant des informations visuelles et textuelles, *Audio MultiChallenge*⁴⁶ pour des conversations audio, ou encore *VisualToolBench*⁴⁷ pour l’édition de contenu visuel, montrent une progression continue. Nos projections suggèrent que ces benchmarks multimodaux seront largement résolus dans les deux ans

³⁸ “MultiChallenge.” Scale AI. <https://scale.com/leaderboard/multichallenge>.

³⁹ “TutorBench.” Scale AI. <https://scale.com/leaderboard/tutorbench>.

⁴⁰ “Lech Mazur Writing.” <https://github.com/lechmazur/writing>.

⁴¹ Solution : délice (masculin au singulier, féminin au pluriel, “jamais bouclé” = lisse).

⁴² “CAD-Eval.” Epoch AI. <https://epoch.ai/benchmarks/cad-eval>.

⁴³ “VPCT.” <https://cbrower.dev/vpct>.

⁴⁴ “GeoBench.” <https://geobench.org>.

⁴⁵ “VISTA.” Scale AI. https://scale.com/leaderboard/visual_language_understanding.

⁴⁶ “Audio MultiChallenge.” Scale AI. <https://scale.com/research/audiomc>.

⁴⁷ “VisualToolBench (VTB).” Scale AI. <https://scale.com/leaderboard/vtb>.

à venir (Figure 7). Les GPAI de frontière maîtrisent progressivement le traitement conjoint du texte, de l'image, du son et de la vidéo.

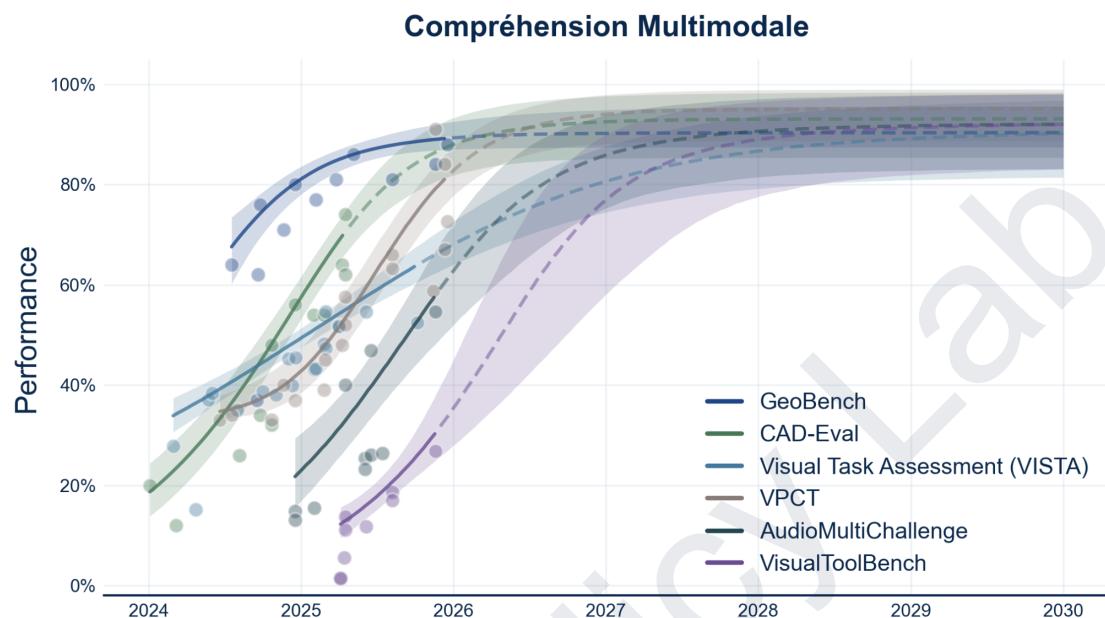


Figure 7 - Projections des performances d'IA sur des benchmarks de compréhension multimodale.
Conventions graphiques identiques à la Figure 3.

Les premiers résultats de notre analyse, présentés dans cette partie, confirment la prospective de saturation des benchmarks actuels avant 2030. Cette progression rapide devient particulièrement significative quand elle concerne l'émergence des capacités critiques, abordées dans la section suivante.

II. Quand émergeront des capacités posant des problèmes de sécurité et de contrôle ?

Certains benchmarks évaluent des capacités à potentiel de mésusage ou des capacités critiques pour la sécurité et le contrôle des systèmes d'IA. Leur maîtrise complète par des systèmes autonomes pourrait marquer un point de bascule, potentiellement irréversible, vers une prolifération de menaces inédites⁴⁸. Des benchmarks évaluant l'expertise des modèles d'IA pour la réalisation d'armes biologiques ou chimiques sont présentés en annexe, nous nous concentrerons ici sur les capacités cyber, sur l'automatisation de la recherche en IA et sur les mathématiques de pointe.

1. Cybersécurité et opérations informatiques agentiques

La cybersécurité est un exemple de domaine critique. L'écrasante majorité de systèmes informatiques comportent des failles encore non-identifiées ou non-corrigées, qui les rendent vulnérables à des cyberattaques. Les systèmes d'IA actuels sont déjà capables d'identifier de telles failles zero-day⁴⁹ et de contribuer significativement à des cyberattaques⁵⁰. Des benchmarks comme

⁴⁸ Cf. notre note précédente : "Artificial General Intelligence (AGI) : Anticipation des objectifs et implications pour la sécurité et le contrôle" (2025).

⁴⁹ En août 2025, lors du AI Cyber Challenge (AIxCC) de la DARPA, les systèmes d'IA présentés ont découvert 18 vulnérabilités zero-day réelles dans des logiciels open-source. <https://aicyberchallenge.com/Finals-winners-announcement>

⁵⁰ "Disrupting the first reported AI-orchestrated cyber espionage campaign." Anthropic (2025).
<https://www.anthropic.com/news/disrupting-AI-espionage>

*Cybench*⁵¹ ont été développés pour suivre l'évolution de ces capacités de hacking, et informer indirectement sur les capacités opérationnelles en conditions réelles. Pour établir un lien entre des benchmarks cyber et la capacité réelle à causer des dommages dans le monde réel, on peut notamment s'appuyer sur des travaux de l'organisation SaferAI⁵². Leur approche consiste à réunir des experts en IA et en cybersécurité pour estimer l'apport, pour des acteurs cyber malveillants, d'IA passant des seuils de performances sur des benchmarks comme *Cybench*.

D'autres benchmarks, comme *OS World*⁵³, *MCP Atlas*⁵⁴, *TerminalBench*⁵⁵ et *The Agent Company*⁵⁶ (Figure 8), évaluent plus largement la capacité d'agents IA à accomplir des tâches sur un ordinateur (emploi du terminal, workflows d'entreprise, utilisation d'outils). Nos projections indiquent une saturation dans les deux prochaines années.

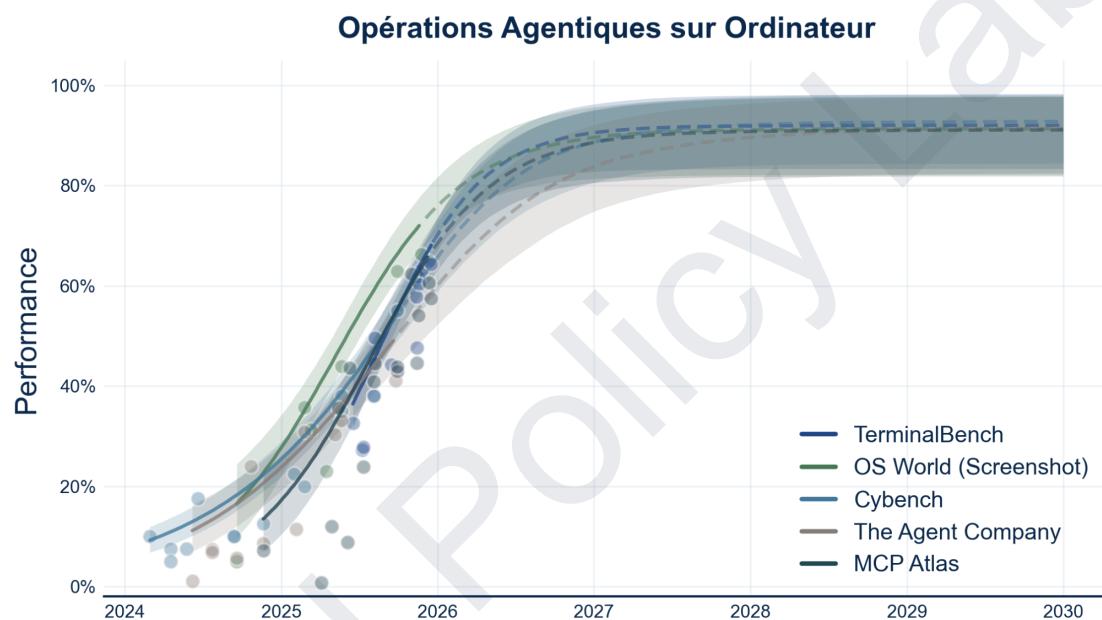


Figure 8 - Projections des performances sur des benchmarks mesurant les capacités d'agents IA à réaliser des tâches informatiques. Conventions graphiques identiques à la Figure 3.

La saturation des benchmarks cyber et informatiques va nous rapprocher, dans les années à venir, du stade où des IA pourront identifier et exploiter automatiquement les vulnérabilités des systèmes informatiques qui n'auront pas été sécurisés à temps.

2. Automatisation de la R&D en IA

L'automatisation de la recherche et développement (R&D) en IA crée une boucle de rétroaction : si les IA deviennent capables de contribuer de manière autonome à la recherche en IA, le rythme de progrès s'accélérera rapidement. On peut anticiper cette dynamique en suivant les capacités des

⁵¹ Zhang et al. "Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risks of Language Models." ICLR (2024). [10.48550/arXiv.2408.08926](https://arxiv.org/abs/2408.08926).

⁵² Barrett et al. "Toward Quantitative Modeling of Cybersecurity Risks Due to AI Misuse." ArXiv (2025).

[10.48550/arXiv.2512.08864](https://arxiv.org/abs/2512.08864) ; Murray et al. "A Methodology for Quantitative AI Risk Modeling." ArXiv (2025). [10.48550/arXiv.2512.08844](https://arxiv.org/abs/2512.08844)

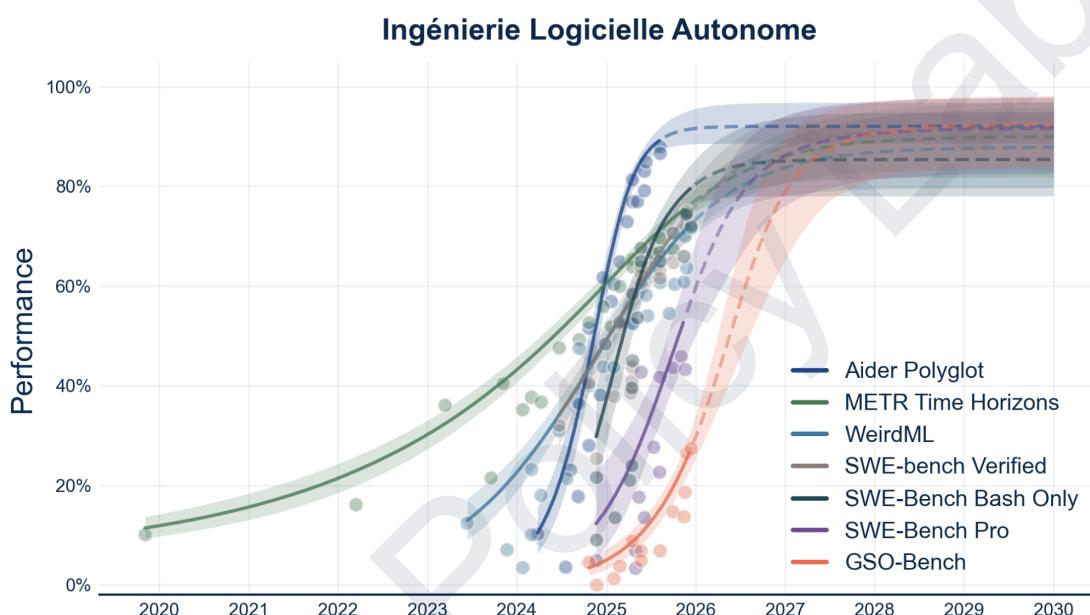
⁵³ Xie et al. "OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments." NeurIPS (2024). [10.48550/arXiv.2404.07972](https://arxiv.org/abs/2404.07972).

⁵⁴ "MCP Atlas." Scale AI. https://scale.com/leaderboard/mcp_atlas.

⁵⁵ "Terminal-Bench 2.0." <https://www.tbench.ai/leaderboard/terminal-bench/2.0>.

⁵⁶ Xu et al. "TheAgentCompany: Benchmarking LLM Agents on Consequential Real World Tasks." NeurIPS (2024). [10.48550/arXiv.2412.14161](https://arxiv.org/abs/2412.14161).

GPAI sur les compétences d'ingénierie logicielle autonome qui constituent les bases de la R&D en IA. Des benchmarks comme *SWE-bench Verified*⁵⁷, *Bash Only*⁵⁸ et *Pro*⁵⁹ (résolution autonome de problèmes logiciels réels), *GSO-Bench*⁶⁰ (optimisation de performances, face à des ingénieurs humains), *Aider polyglot*⁶¹ (problèmes de code dans plusieurs langages de programmation) ou *WeirdML*⁶² (tâches de *Machine Learning* non-standards) entrent dans cette catégorie. C'est également le cas pour les trois suites de tâches utilisées par l'organisation *METR*⁶³ pour mesurer l'horizon temporel des agents IA : *SWAA* (actions isolées de développement logiciel), *HCAST* (tâches autonomes en ingénierie logicielle, ML et cybersécurité) et *RE-Bench* (environnements de R&D en ML). Tous ces benchmarks présentent des progrès très rapides (Figure 9).



L'émergence d'IA capables de contribuer à leur propre développement, puis de l'automatiser, va progressivement accélérer les progrès en IA. Une telle accélération rend les trajectoires de progrès moins prévisibles, et réduit le temps disponible pour développer et déployer des mesures de sécurité face aux enjeux de sécurité et de contrôle que posent une AGI⁶⁴.

3. Raisonnements mathématiques de pointe

Les mathématiques avancées sont l'un des domaines dans lesquels l'IA se rapproche le plus vite des experts humains. *MATH*, qui compile des problèmes de compétitions de mathématiques pour lycéens, était hors de portée des modèles existants à sa sortie en 2021 ; les IA d'aujourd'hui

⁵⁷ "SWE-bench Verified." Epoch AI. <https://epoch.ai/benchmarks/swe-bench-verified>.

⁵⁸ "SWE-bench Bash Only" SWE-bench <https://www.swebench.com/bash-only.html>

⁵⁹ "SWE-Bench Pro." Scale AI. https://scale.com/leaderboard/swe_bench_pro_public.

⁶⁰ Shetty et al. "GSO: Challenging Software Optimization Tasks for Evaluating SWE-Agents." ArXiv (2025). [10.48550/arXiv.2505.23671](https://arxiv.org/abs/2505.23671).

⁶¹ "Aider Polyglot." <https://aider.chat/docs/leaderboards/#polyglot-leaderboard>.

⁶² "WeirdML." <https://htihle.github.io/weirdml.html>.

⁶³ Kwa et al. "Measuring AI Ability to Complete Long Tasks." ArXiv (2025). [10.48550/arXiv.2503.14499](https://arxiv.org/abs/2503.14499) et <https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks> pour les données mises à jour.

⁶⁴ Cf. notre note précédente : "Artificial General Intelligence (AGI) : Anticipation des objectifs et implications pour la sécurité et le contrôle" (2025).

atteignent désormais plus de 95% de réussite, y compris sur *MATH Level 5*⁶⁵, son dernier niveau de difficulté, et sur la compétition *AIME*⁶⁶. En réponse, de nouveaux benchmarks, tels que *FrontierMath*⁶⁷, ont été développés pour pousser les limites de ces systèmes en présentant des problèmes mathématiques originaux jamais publiés et nécessitant des raisonnements proches du niveau des chercheurs, en particulier pour son quatrième tiers de difficulté. Les trajectoires observées suggèrent que même ces benchmarks extrêmement difficiles pourraient être essentiellement résolus avant 2028 (Figure 10).

Cette progression s'explique par le caractère formel des mathématiques, qui facilite la vérification automatique et donc l'apprentissage par renforcement. Contrairement à des domaines comme la rédaction, où l'évaluation de la qualité d'une réponse reste subjective et coûteuse, les problèmes mathématiques ont des solutions vérifiables de manière déterministe. Les architectures récentes exploitent cette particularité en allouant davantage de temps de calcul à l'inférence pour explorer plusieurs pistes de raisonnement et vérifier la validité des résultats, ce qui permet de générer automatiquement de nouvelles données pour entraîner les modèles suivants.

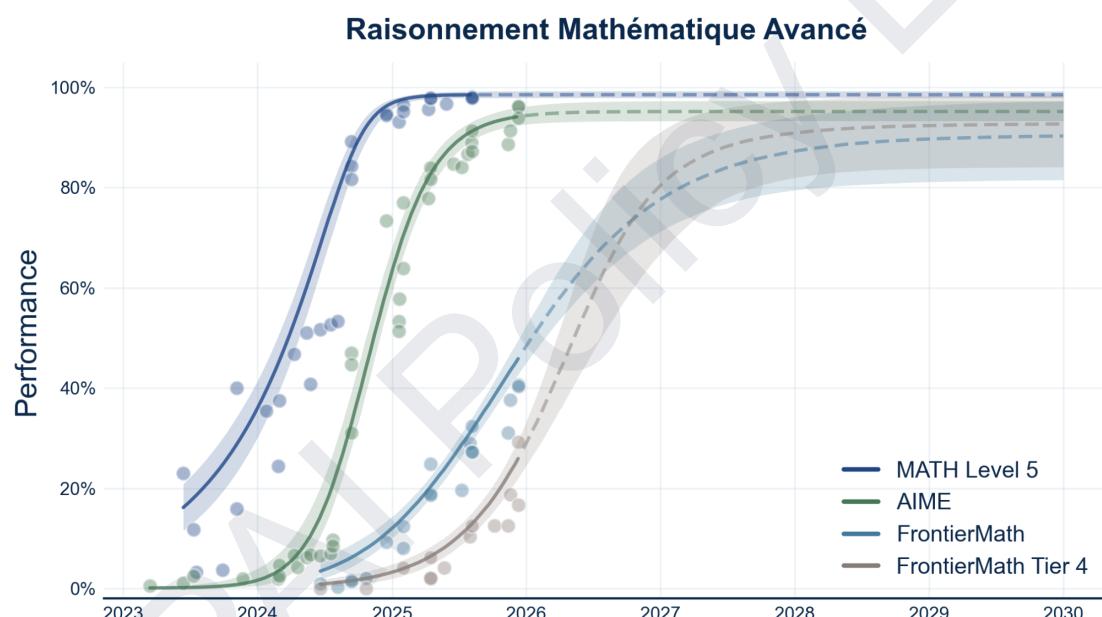


Figure 10 - Projections des performances d'IA sur des benchmarks de mathématiques avancées.
Conventions graphiques identiques à la Figure 3.

Les implications de ces améliorations s'étendent au-delà des mathématiques : dans la mesure où les capacités faciles à entraîner (mathématiques, programmation) généralisent à d'autres compétences nécessitant du raisonnement et de la planification, cela contribue à accélérer les progrès dans de nombreux autres domaines. **Enfin, le raisonnement mathématique avancé fait probablement partie des compétences nécessaires pour des innovations de R&D en IA**, contribuant à la boucle d'automatisation de la recherche en IA, et au problème de contrôle qu'elle implique.

⁶⁵ "MATH Level 5." Epoch AI. <https://epoch.ai/benchmarks/math-level-5>.

⁶⁶ "OTIS Mock AIME 2024-2025." Epoch AI. <https://epoch.ai/benchmarks/otis-mock-aime-2024-2025>.

⁶⁷ "FrontierMath." Epoch AI. <https://epoch.ai/benchmarks/frontiermath>.

III. À quel point l'hétérogénéité des progrès observés est-elle une barrière au développement d'une AGI ?

1. Avancées des benchmarks mesurant les progrès vers l'AGI

Une catégorie particulière de benchmarks vise spécifiquement à tester la généralité des capacités développées, soit en multipliant les compétences testées en parallèle (*LiveBench*⁶⁸, *Definition of AGI*⁶⁹), soit en ciblant des capacités faciles pour les humains mais difficiles pour les IA (*ARC-AGI-1*⁷⁰, *ARC-AGI-2*⁷¹). L'avantage de cibler les tâches accessibles à l'essentiel des humains est qu'elles n'évaluent pas l'accumulation de connaissances, mais plutôt le cœur de la généralité de la cognition humaine⁷². Ces tâches nécessitent de former rapidement de nouveaux concepts et d'acquérir de nouvelles compétences à partir d'exemples limités et de les appliquer à des problèmes inédits. Les modèles de raisonnement récents, notamment ceux utilisant des temps de réflexion étendus (comme o3 d'OpenAI), ont permis de faire des sauts de performance significatifs sur ces tâches⁷³, rapprochant *ARC-AGI-1* de la saturation (Figure 11). Le benchmark successeur *ARC-AGI-2*, au niveau de difficulté accru, est en passe d'être saturé courant 2026. Une troisième version du benchmark, basée sur la capacité à explorer des environnements inédits, a été annoncée⁷⁴.

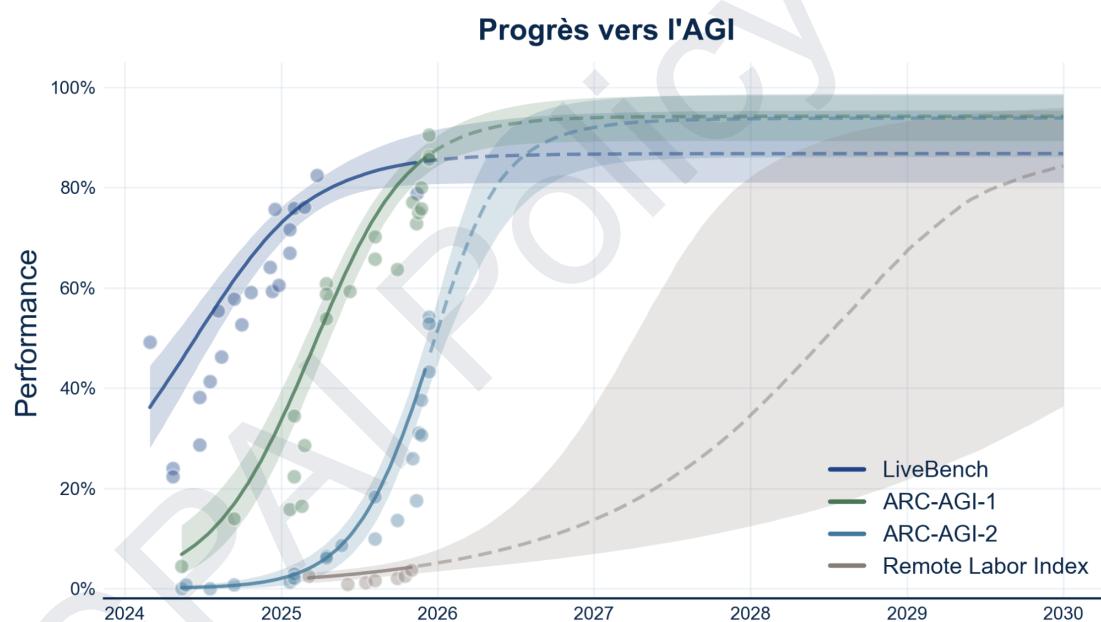


Figure 11 - Projections des performances d'IA sur des benchmarks visant à mesurer les progrès vers l'Intelligence Artificielle Génératrice (AGI). Conventions graphiques identiques à la Figure 3.

Une autre approche pour mesurer le progrès vers l'AGI est d'évaluer les modèles sur des tâches représentatives du travail humain, avec la complexité et les imprévus qui les caractérisent. Deux

⁶⁸ White et al. “LiveBench: A Challenging, Contamination-Free LLM Benchmark.” *ICLR* (2024). [10.48550/arXiv.2406.19314](https://arxiv.org/abs/2406.19314)

⁶⁹ Hendrycks et al. “A Definition of AGI.” *ArXiv* (2025). [10.48550/arXiv.2510.18212](https://arxiv.org/abs/2510.18212).

⁷⁰ “ARC-AGI-1.” *ARC Prize* (2019). <https://arcprize.org/arc-agi/1>.

⁷¹ “ARC-AGI-2.” *ARC Prize* (2025). <https://arcprize.org/arc-agi/2>.

⁷² Comme le souligne François Chollet, créateur du benchmark ARC-AGI, ces tâches testent la capacité d'adaptation en direct à des situations inconnues, une compétence fondamentale de l'intelligence générale. Voir Chollet. “On the Measure of Intelligence.” *ArXiv* (2019). <https://arxiv.org/abs/1911.01547>

⁷³ Chollet. “OpenAI o3 Breakthrough High Score on ARC-AGI-Pub.” *ARC Prize* (2024).

<https://arcprize.org/blog/oai-o3-pub-breakthrough>.

⁷⁴ “ARC-AGI-3.” *ARC Prize* (2025). <https://arcprize.org/arc-agi/3>.

benchmarks récents empruntent cette voie, dans le but de mesurer le potentiel d'impact des GPAI sur l'économie réelle :

- *GDPval*⁷⁵, développé par OpenAI, évalue des tâches utiles dans une gamme de métiers contribuant significativement au PIB étasunien, et montre que des IA arrivent à parité avec les professionnels humains (avec des variations importantes selon les métiers). Cependant, ce benchmark manque de réalisme : les instructions sont très détaillées, pour des tâches assez restreintes et peu représentatives du monde du travail⁷⁶.
- Le *Remote Labor Index* (RLI)⁷⁷ du *Center for AI Safety* apparaît plus réaliste, avec des tâches issues de plateformes de travail à distance et des instructions plus réduites. De fait, ce benchmark s'avère beaucoup plus difficile, puisque les meilleures IA n'égalent les professionnels humains que sur environ 2% des tâches. Nos projections à l'horizon 2030 sur ce benchmark restent très incertaines (cf. Figure 11).

2. Progrès hétérogènes de l'IA : une barrière fragile vers l'AGI

Une des caractéristiques de ces trajectoires de progrès, prises dans leur ensemble, est leur forte hétérogénéité (on parle aussi de *jaggedness*, ou de progrès non-uniforme, cf. Figure 12). Les modèles montrent des avancées spectaculaires sur certains types de tâches - par exemple des niveaux surhumains sur des questions scientifiques - tout en échouant encore sur des tâches triviales pour un enfant, comme certaines intuitions physiques. Cette hétérogénéité remet-elle en cause la trajectoire de progression vers une AGI ? **Bien qu'elle rende les projections plus incertaines, l'hétérogénéité ne garantit pas de disposer de décennies avant une AGI**, en particulier si l'on se trouve dans l'un des scénarios suivants :

1. *Croissance de la puissance de calcul dédiée à l'entraînement des GPAI, comblant progressivement les lacunes de capacités.* Le pari qui s'est avéré gagnant pour les développeurs d'IA, depuis 2018 (GPT-1), a été d'investir dans le passage à l'échelle (*scaling*)⁷⁸ des modèles pour débloquer de nouvelles capacités émergentes à chaque nouvelle génération. La poursuite de ce *scaling*, combinée à l'amélioration et la combinaison des méthodes d'entraînement actuelles (sans découverte révolutionnaire), pourrait continuer à réduire la liste des compétences dans lesquelles les IA échouent face aux humains, menant à terme à un modèle pouvant être qualifié d'AGI (Figure 12)⁷⁹.
2. *Dépassement des experts humains précisément sur les compétences nécessaires à la R&D en IA.* Si de nouveaux paradigmes, architectures ou méthodes d'entraînement sont nécessaires pour arriver à une AGI, le seuil à surveiller est celui où une GPAI égale un chercheur en IA. Une telle IA, sans dépasser les humains sur toutes les tâches cognitives, serait suffisante pour accélérer la découvertes de nouvelles approches plus performantes. Dans ce scénario, l'émergence d'AGI ne découlerait pas uniquement du *scaling* actuel, mais aussi des progrès algorithmiques permis par les GPAI elles-mêmes.

⁷⁵ Patwardhan et al. “GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks.” ArXiv (2025). [10.48550/arXiv.2510.04374](https://arxiv.org/abs/2510.04374).

⁷⁶ Mazeika et al. “Remote Labor Index: Measuring AI Automation of Remote Work.” ArXiv (2025). [10.48550/arXiv.2510.26787](https://arxiv.org/abs/2510.26787)

⁷⁷ “Remote Labor Index (RLI).” Scale AI. <https://scale.com/leaderboard/rli>.

⁷⁸ Kaplan et al. “Scaling Laws for Neural Language Models.” ArXiv (2020). [10.48550/arXiv.2001.08361](https://arxiv.org/abs/2001.08361).

⁷⁹ Cf. notre note précédente : “Artificial General Intelligence (AGI) : Faisabilité technique et horizons temporels” (2025).

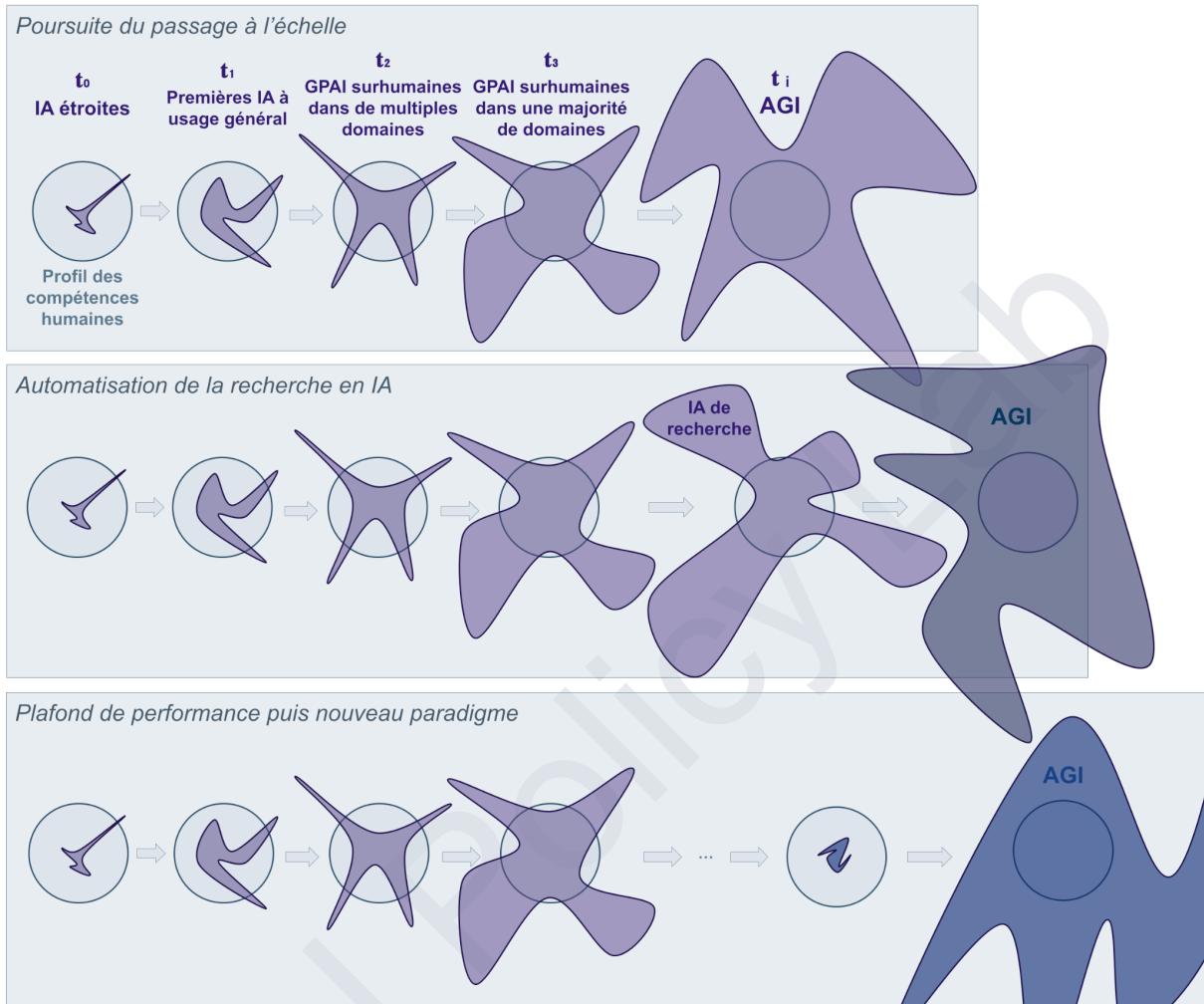


Figure 12 - Progression non-uniforme des IA de frontière, et scénarios pouvant mener à une AGI.

L'étendue des tâches cognitives que les IA peuvent réaliser (zone violette) s'étend rapidement dans de nombreux domaines, dépassant les experts humains sur certaines tâches tout en demeurant limitée par ailleurs. L'AGI représente la couverture complète du spectre cognitif humain. Dans le premier scénario, une AGI est atteinte en passant à l'échelle et en combinant les méthodes d'entraînement actuelles (adapté de Thomas Pueyo⁸⁰). Dans le deuxième scénario, les progrès actuels mènent à une IA capable d'automatiser la recherche en IA, qui trouve ensuite une nouvelle approche menant à l'AGI. Dans le troisième scénario, la progression des performances ralentit avant d'atteindre une AGI, mais celle-ci est développée quelques années plus tard à la suite d'une ou plusieurs percées algorithmiques.

En l'état, il n'est pas possible de prédire précisément combien de temps des lacunes locales vont persister dans les compétences des GPAI de frontière, ni quand elles seront capables de réaliser des tâches économiques réelles en complète autonomie. Cependant, si ces incertitudes rendent difficile la prévision du développement d'une AGI, elles n'impliquent pas de limitation intrinsèque des capacités des GPAI. **Du point de vue de la sécurité, à partir du moment où les ingrédients sont réunis pour développer une AGI, l'enjeu est moins de savoir si ce développement aboutira précisément en 2030 ou 2035 que d'anticiper les problèmes de contrôle qui en découlent**⁸¹.

⁸⁰ Thomas Pueyo. *Uncharted Territories Substack* <https://substack.com/@tomaspueyo/note/c-182052822>

⁸¹ Problèmes que nous avons détaillés dans notre note précédente : "Artificial General Intelligence (AGI) : Anticipation des objectifs et implications pour la sécurité et le contrôle" (2025).

Résumé

- **Quasiment tous les benchmarks actuels, y compris les plus difficiles, seront probablement saturés avant 2030.** Cette projection repose sur une analyse systématique des trajectoires de progrès sur 60 benchmarks, modélisés conjointement.
- **Les progrès sont rapides et généralisés :** du sens commun aux mathématiques, en passant par le raisonnement multimodal et la programmation agentiques, les modèles de frontière progressent de dizaines de points de pourcentage par an sur des tâches considérées comme hors de portée il y a peu. La plus grande incertitude réside dans l'automatisation de tâches complexes réelles, dont une partie pourrait tenir au-delà de 2030.
- **Les capacités critiques progressent au même rythme :** cybersécurité, automatisation de la R&D en IA, et mathématiques avancées suivent des trajectoires similaires, et poseront rapidement des problèmes de sécurité et de contrôle. Tous les benchmarks de ces catégories inclus dans notre analyse seront probablement saturés au plus tard courant 2028.
- **Il est difficile de prédire quand une AGI pourra être développée, mais l'hétérogénéité des progrès ne constitue pas une barrière structurelle.** Les lacunes locales actuelles pourraient ainsi être soit comblées par le passage à l'échelle des modèles, soit contournées si les capacités suffisantes pour automatiser la R&D en IA sont atteintes en premier.

Conclusion : Les systèmes d'IA à usage général vont, dans les prochaines années, atteindre ou dépasser les performances des experts humains sur l'essentiel des tâches cognitives actuellement mesurables. Ces projections suggèrent une progression vers l'AGI, et incitent à se concentrer sur l'évaluation des compétences nécessaires pour automatiser la recherche en IA.

Suite possible à cette note

1. Comment vont évoluer les capacités des GPAI de pointe au-delà des benchmarks actuels ? Quelles sont les propriétés des benchmarks que l'on peut extrapoler dans le temps et comparer aux capacités humaines ?
2. À quel point l'automatisation de la recherche et développement en IA peut-elle accélérer la progression des performances des GPAI ? Cette accélération est-elle fortement limitée par la quantité de puissance de calcul disponible ? A-t-on des données empiriques pour affiner ces estimations ?
3. Quels sont les risques concrets qui ont des chances d'émerger lorsque des GPAI auront saturé les benchmarks présentés dans cette note ? À partir de quels seuils de capacités doit-on anticiper un risque de perte de contrôle ?

Annexe

A. Méthodologie

Notre objectif est d'anticiper l'évolution des performances des meilleurs modèles d'IA sur un ensemble de benchmarks couvrant des compétences variées. Pour ce faire, nous modélisons l'évolution temporelle des scores de frontière par une trajectoire de croissance flexible, et extrapolons les données historiques vers le futur.

Les performances de frontière suivent empiriquement une trajectoire sigmoïde (en S) : scores initiaux proches de l'aléatoire, puis accélération, et saturation vers des réponses quasi-parfaites (cf. Figure S1). Certains benchmarks ne sont pas bornés de cette manière, et peuvent prendre des valeurs arbitrairement hautes (par exemple générer le plus d'argent possible), mais ne sont pas inclus dans cette analyse.

Cette évolution sigmoïde peut être extrapolée dans le futur en ajustant une fonction mathématique sur les résultats historiques des meilleurs modèles, pour ensuite la projeter dans les années à venir. Les approches classiques utilisent comme fonction d'évolution la courbe logistique, qui présente une dynamique sigmoïde symétrique, où l'accélération et la décélération se produisent à la même vitesse.

Notre méthodologie introduit trois innovations majeures par rapport aux méthodes standards :

1. À la place d'une courbe logistique, nous utilisons le modèle de Harvey⁸² (cf. Figure S1), qui autorise une asymétrie entre accélération et décélération. Cette flexibilité permet de mieux capturer l'incertitude sur la forme exacte de la courbe de progrès et d'éviter des prédictions trop confiantes sur le délai avant saturation.
2. **Plutôt que de prédire les performances moyennes des modèles de frontière, nous cherchons à nous approcher des capacités maximales au fil du temps**⁸³. Cette distinction est déterminante pour l'anticipation des enjeux de sécurité et de contrôle : les problèmes de sécurité émaneront principalement des modèles les plus avancés, lorsqu'ils sont utilisés avec les meilleures techniques d'élicitation, plutôt que du modèle moyen. Par conséquent, l'approche employée permet de réduire l'*elicitation gap* que nous avons évoqué, et qui constitue un angle mort dans les évaluations de sécurité.
3. Au lieu de réaliser des prévisions indépendantes pour chaque benchmark, nous employons une projection jointe de l'ensemble des benchmarks à travers un modèle unifié dit Bayésien hiérarchique⁸⁴. Les distributions des paramètres d'intérêt (vitesses de progression, points d'inflexion, asymétrie des courbes) sont informées simultanément par tous les benchmarks considérés. Cela permet d'affiner les projections pour les benchmarks avec peu de données, en s'appuyant sur les benchmarks plus complets.

⁸² Harvey. "Time Series Forecasting Based on the Logistic Curve." *Journal of the Operational Research Society* (1984). [10.1057/JORS.1984.128](https://doi.org/10.1057/JORS.1984.128) ; Young. "Technological growth curves. A competition of forecasting models." *Technological Forecasting and Social Change* (1993). [10.1016/0040-1625\(93\)90042-6](https://doi.org/10.1016/0040-1625(93)90042-6)

⁸³ Concrètement, nous modélisons les scores observés comme des réalisations imparfaites d'une capacité latente maximale. Pour chaque benchmark, nous supposons donc que les performances mesurées sont tirées dans une distribution asymétrique, qui implique que les scores sont le plus souvent inférieurs à la capacité maximale estimée.

⁸⁴ Un modèle bayésien hiérarchique (ou multiniveau) est une méthode statistique qui permet de représenter la structure présente dans les données, en modélisant des groupes distincts (ici les benchmarks) qui partagent des propriétés communes.

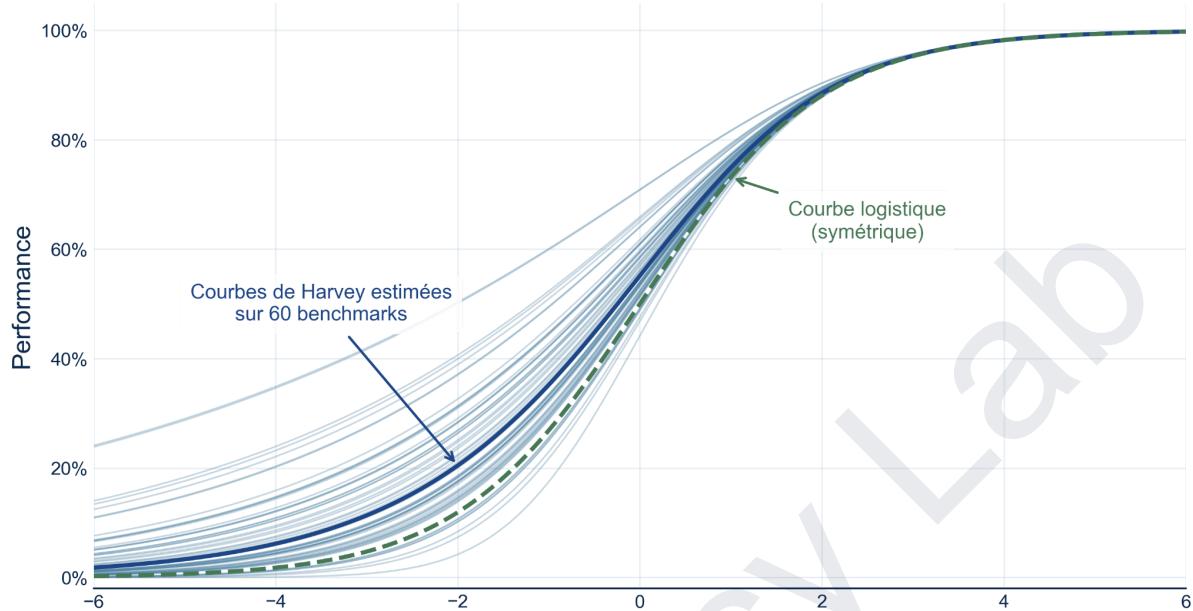


Figure S1 - Asymétrie des courbes de progrès sur des benchmarks d'IA : comparaison entre modèles de Harvey et fonction logistique. Les courbes bleues représentent les trajectoires de progrès approximées par des modèles de Harvey pour 46 benchmarks, et leur médiane (trait plus épais). Ces courbes montrent une asymétrie caractéristique avec une phase d'augmentation très graduelle des performances avant d'approcher de la saturation. La courbe en pointillés verts illustre pour référence la forme symétrique d'une fonction logistique classique. Toutes les courbes ont été standardisées, avec un point d'inflexion centré à 0 et un taux de croissance fixé à 1, de manière à comparer uniquement la forme des trajectoires.

Un dernier point méthodologique à noter est que tous les benchmarks ne s'étendent pas réellement de 0% à 100%. D'une part, le score aléatoire peut être non nul ; par exemple de 25% pour des questions à choix multiples avec 4 options de réponse. D'autre part, de nombreux benchmarks comportent des ambiguïtés dans certaines de leurs questions, ou alors des erreurs dans les "vraies réponses" annotées par des évaluateurs humains, ce qui empêche d'atteindre un score de 100%. Ces deux facteurs doivent être pris en compte pour des projections现实istes. Ils sont donc intégrés dans notre modèle en indiquant les scores aléatoires attendus, lorsqu'ils étaient disponibles, et en inférant le score maximal atteignable à partir des données observées.

B. Choix des benchmarks

Notre analyse s'appuie principalement sur les benchmarks répertoriés par *Epoch AI*⁸⁵, une organisation spécialisée dans la mesure et l'analyse des progrès de l'IA. Ces benchmarks présentent plusieurs avantages : ils sont disponibles librement avec des données uniformisées et standardisées, et ils couvrent un très large spectre de compétences, allant du sens commun élémentaire aux capacités de raisonnement avancé. Certains benchmarks ont été réévalués par *Epoch AI* avec des protocoles rigoureux, les autres proviennent d'évaluations par d'autres organisations. Nous avons également intégré les benchmarks de *Scale AI*, une entreprise qui réalise des évaluations de modèles d'IA, et a participé au développement de plusieurs benchmarks de pointe⁸⁶. Nous avons complété cette base avec des benchmarks additionnels particulièrement difficiles (ARC-AGI-2) ou présentant des enjeux de sécurité (biologie et chimie⁸⁷, cf. section C), et avons exclu huit

⁸⁵ "AI Benchmarking." *Epoch AI* (données téléchargées le 12 jan. 2025). <https://epoch.ai/benchmarks>.

⁸⁶ "LLM Leaderboards." *Scale AI* (données téléchargées le 12 jan. 2025). <https://scale.com/leaderboard>.

⁸⁷ Dev et al. "Toward Comprehensive Benchmarking of the Biological Knowledge of Frontier Large Language Models" *RAND Corporation* (2025). https://www.rand.org/pubs/research_reports/RRA3797-1.html.

benchmarks (*BIG-Bench Hard*, *BoolQ*, *LAMBADA*, *MMLU*, *SuperGLUE*, *TriviaQA*, *Video-MME*, *CommonsenseQA*) qui n'étaient plus mis à jour par *Epoch AI* avec les derniers modèles de frontière.

C. Validation des choix méthodologiques

L'analyse empirique valide nos choix méthodologiques. Les courbes de progression observées sont effectivement asymétriques : l'accélération est plus graduelle que ce qu'autoriserait une fonction logistique, et cette asymétrie varie d'un benchmark à l'autre (cf. Figure S1). Cela justifie l'utilisation d'un modèle flexible comme celui de Harvey. Par ailleurs, le taux d'erreur estimé pour le score maximal atteignable varie entre 3% et 15% selon les benchmarks (Figure S2 ci-dessous), confirmant que les questions et annotations humaines sont rarement parfaites.

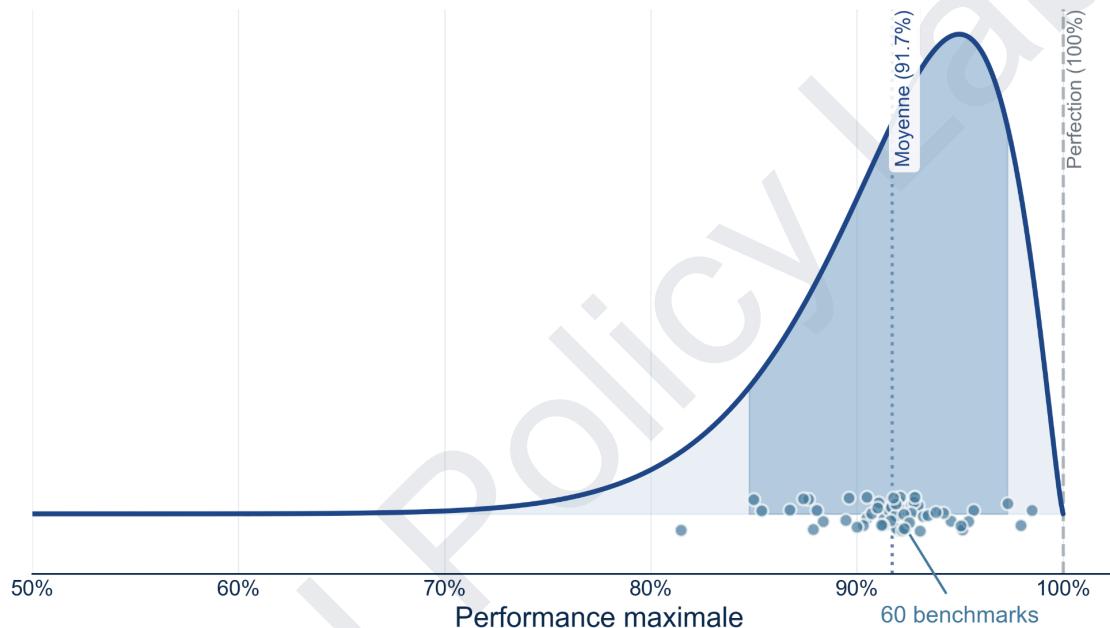


Figure S2 - Distribution estimée des performances maximales atteignables sur les benchmarks étudiés. La courbe représente la densité de probabilité estimée des plafonds de performance pour les benchmarks étudiés ; la zone ombrée délimite l'intervalle de crédibilité à 80%. Les points bleus indiquent les estimations médianes pour chaque benchmark individuel : certains benchmarks peuvent avoir des taux de réponse presque parfaits tandis que d'autres présentent des limites intrinsèques plus basses.

D. Convergence des projections vers une saturation avant 2030

Pour quantifier l'observation d'une saturation rapide et généralisée, nous avons calculé la distribution de la proportion de benchmarks saturés d'ici 2030 en agrégeant les trajectoires individuelles issues du modèle bayésien hiérarchique. Cette approche permet de synthétiser l'incertitude portant sur chaque benchmark en une mesure globale du degré de saturation attendu. La Figure S3 présente cette distribution postérieure.

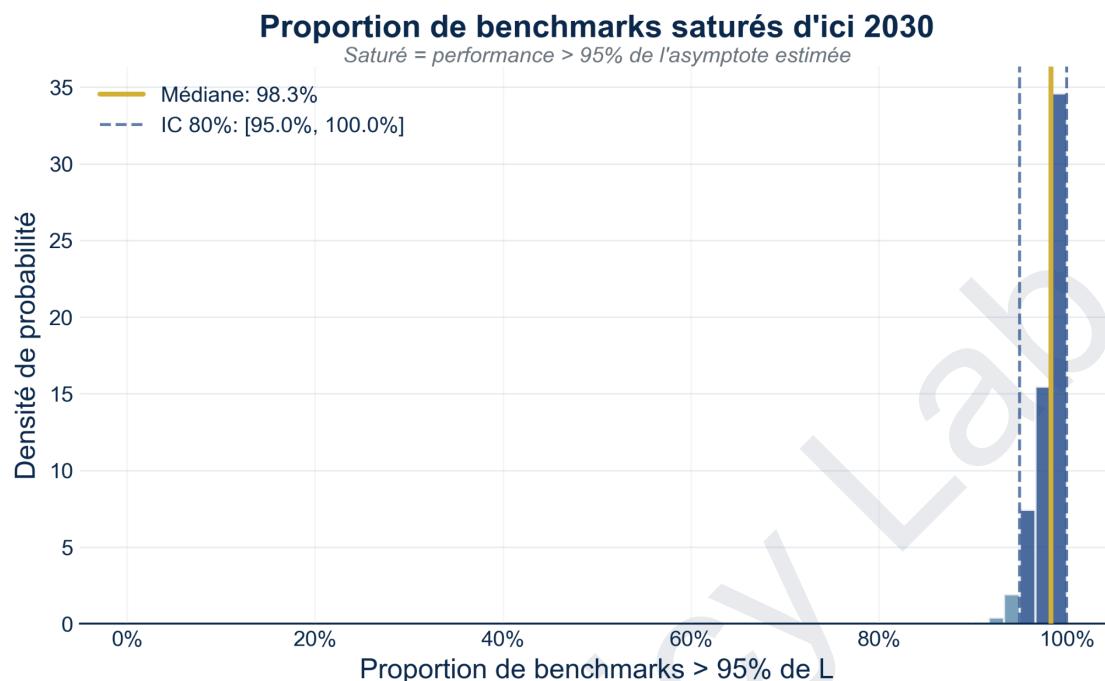


Figure S3 - Distribution estimée de la proportion de benchmarks saturés d'ici 2030. La saturation est définie comme l'atteinte d'au moins 95 % de l'asymptote de performance estimée. L'histogramme représente la densité de probabilité, la ligne verticale indique la médiane (98,3 %) et les lignes pointillées l'intervalle de crédibilité à 80 %. Cette concentration près de 100 % suggère une saturation généralisée des benchmarks actuels à l'horizon 2030, sous les hypothèses du modèle.

E. Capacités critiques en biologie et chimie

Les capacités en biologie et en chimie font partie des domaines les plus sensibles, du fait de leur potentiel de mésusage pour la conception d'armes biologiques ou chimiques. Les précédents historiques montrent que des acteurs malveillants ont tenté de développer de telles armes, mais ont généralement échoué par manque d'expertise technique. Des systèmes d'IA capables de fournir des connaissances biologiques et chimiques précises pourraient abaisser ces barrières techniques en aidant des acteurs inexpérimentés à mettre au point et réaliser des protocoles de laboratoire.

Plusieurs benchmarks ont été développés pour évaluer ces compétences⁸⁸, avec des niveaux de difficulté croissants. En biologie, les benchmarks couvrent un spectre allant des connaissances générales aux compétences opérationnelles de laboratoire. *MMLU Pro*⁸⁹ *Biology* et *GPQA*⁹⁰ *Diamond Biology* testent les connaissances théoriques avancées de niveau doctoral. La suite *LAB-Bench*⁹¹, développée par *FutureHouse*, évalue des compétences plus pratiques et directement applicables en laboratoire : compréhension de protocoles expérimentaux (*ProtocolQA*), analyse de séquences génétiques (*SeqQA*), conception de stratégies de clonage moléculaire (*CloningScenarios*), et exploitation de la littérature scientifique (*LitQAA2*). *BioLP-bench*, un benchmark récent de 800 questions à contexte long, évalue la capacité des modèles à identifier et corriger des erreurs introduites dans des protocoles biologiques réels. Le benchmark *WMDP Biology*, quant à lui, cible

⁸⁸ Dev et al. "Toward Comprehensive Benchmarking of the Biological Knowledge of Frontier Large Language Models" RAND Corporation (2025). https://www.rand.org/pubs/research_reports/RRA3797-1.html.

⁸⁹ Wang et al. "MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark." ArXiv (2024). [10.48550/arXiv.2406.01574](https://arxiv.org/abs/2406.01574).

⁹⁰ Rein et al. "GPQA: A Graduate-Level Google-Proof Q&A Benchmark." ArXiv (2023). [10.48550/arXiv.2311.12022](https://arxiv.org/abs/2311.12022).

⁹¹ Laurent et al. "LAB-Bench: Measuring Capabilities of Language Models for Biology Research." ArXiv (2024). [10.48550/arXiv.2407.10362](https://arxiv.org/abs/2407.10362).

spécifiquement les connaissances pouvant être détournées à des fins de développement d'armes de destruction massive : recherche sur les vecteurs viraux, virologie à double usage, mécanismes de déploiement.

Les projections indiquent que les benchmarks biologiques étudiés atteindront la saturation avant 2028 (Figure S4). Un rapport récent de la *RAND Corporation* établit que les modèles de raisonnement de début 2025 dépassaient déjà les performances d'experts humains sur *WMDP Biology*, *BioLP-bench* et approchent ou égalent les baselines expertes sur *LAB-Bench Protocol/QA* et *GPQA Main Biology*.

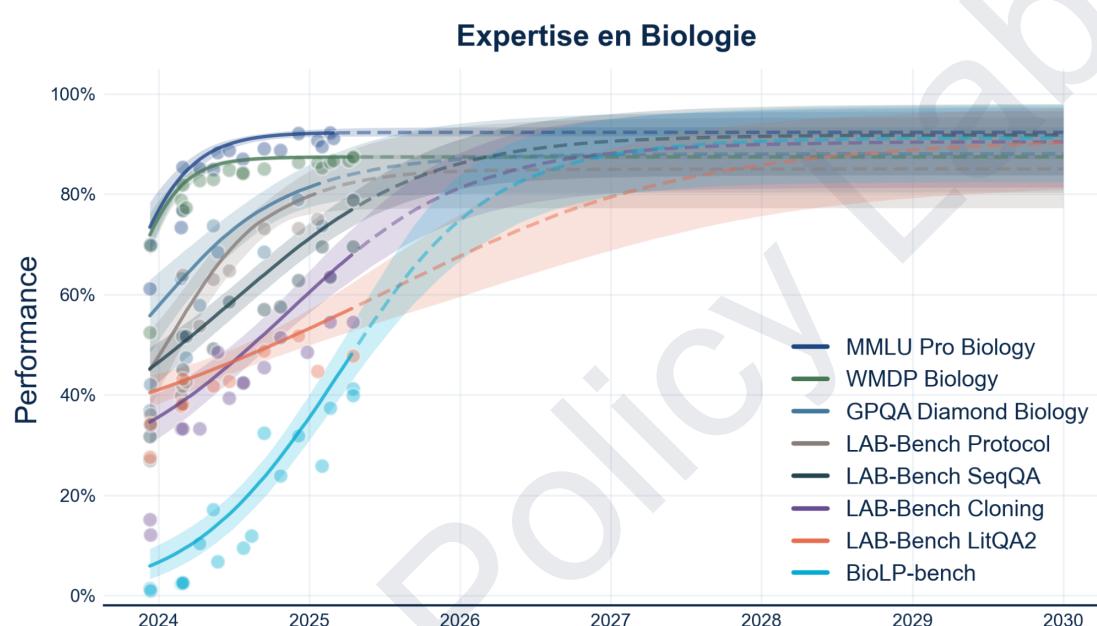


Figure S4 - Projections des performances d'IA sur des benchmarks visant à mesurer les capacités critiques en biologie. Conventions graphiques identiques à la Figure 3.

En chimie, les benchmarks suivent une dynamique similaire (Figure S5). *MMLU Pro Chemistry* et *GPQA Diamond Chemistry* évaluent les connaissances académiques de niveau doctoral. *WMDP Chemistry* cible les connaissances à potentiel de mésusage : synthèse et purification d'agents chimiques dangereux, acquisition de précurseurs, méthodes de vérification et d'analyse, mécanismes de dispersion. Les experts sollicités par RAND ont obtenu un score moyen de 43,3% sur ce benchmark, une performance à nouveau dépassée par les modèles de raisonnement de début 2025. *MMLU Pro Chemistry* et *WMDP Chemistry* sont déjà saturés, tandis que *GPQA Diamond Chemistry* devrait l'être en 2026.

La maîtrise complète de ces compétences par des systèmes autonomes pourrait significativement abaisser les barrières techniques à la conception d'agents pathogènes ou de substances chimiques dangereuses. Toutefois, d'autres barrières demeurent (accès aux équipements de laboratoire, aux matériaux biologiques, aux précurseurs chimiques réglementés), et le lien entre performance sur ces benchmarks et capacité réelle à causer des dommages dans le monde réel reste à établir plus précisément.

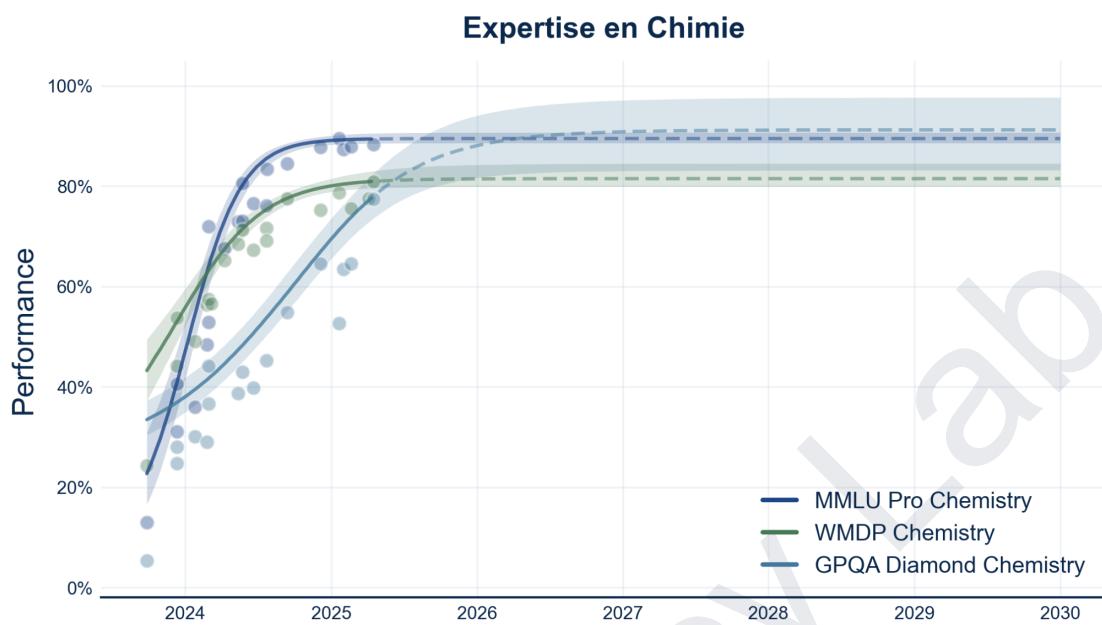


Figure S5 - Projections des performances d'IA sur des benchmarks visant à mesurer les capacités critiques en chimie. Conventions graphiques identiques à la Figure 3.