General Purpose AI Policy Lab

ÉVALUATION DES IA DE FRONTIÈRE

Des difficultés croissantes

Juin 2026

#6 - Note

General Purpose AI Policy Lab

RÉSUMÉ EXÉCUTIF

— CONTEXTE —

La présidence française du G7 place la sécurité de l'IA au sommet de l'agenda numérique, permettant d'aborder les questions de sécurité dans leur globalité. Actuellement, la sécurité de l'IA repose en partie sur la capacité des évaluations à documenter les risques des systèmes de frontière. Or, les évaluations sous-estiment structurellement les capacités réelles des modèles, et peuvent maintenant être détectées et contournées par les systèmes d'IA les plus avancés. Cette note examine ce que les évaluations peuvent, et ne peuvent pas apporter en matière de garanties de sécurité.

Tom David, Président Directeur Général
du General-Purpose AI Policy Lab

— QUESTIONS —

- Quelles sont les spécificités du domaine de l'évaluation de l'IA ?
- À quelles difficultés les évaluateurs devront-ils faire face dans les années à venir ?
- Quelles leçons peut-on en tirer pour l'évaluation des IA à usage général ?

— SYNTHÈSE —

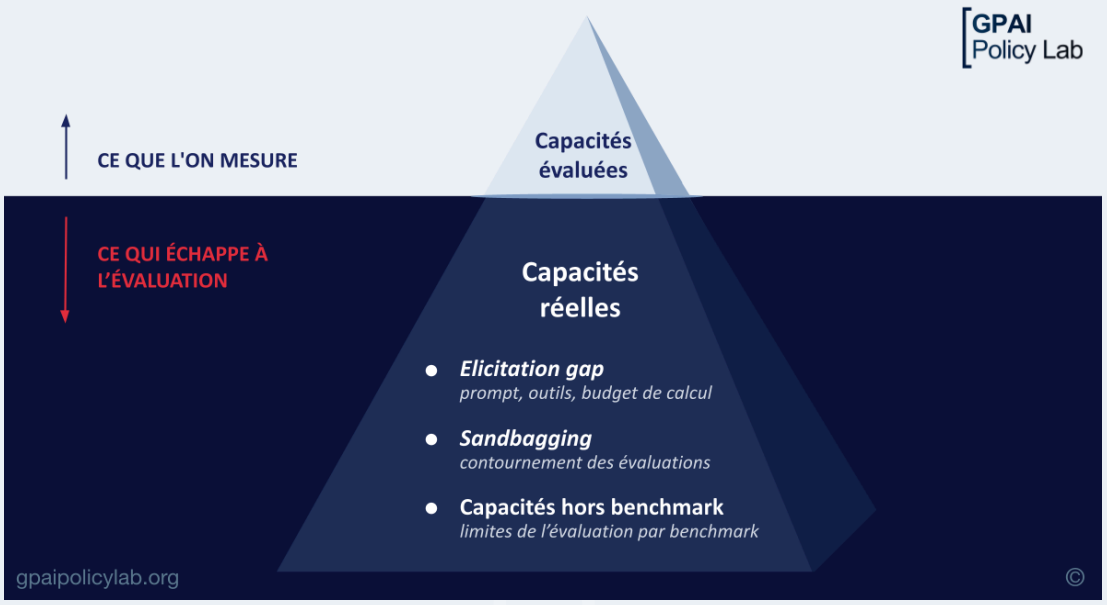
Les développeurs d'IA à usage général de frontière ne contrôlent pas les capacités et comportements exacts des IA qu'ils entraînent, et se reposent donc sur des évaluations. Ces capacités ne sont pas spécifiées par les développeurs, mais émergent d'un processus d'entraînement empirique et itératif formant des « boîtes noires », ce qui fait des évaluations le principal moyen d'accéder à ce que le modèle a réellement appris.

Les évaluations actuelles, majoritairement construites comme des examens figés (ou statiques), présentent plusieurs problèmes bien documentés : contamination des données d'entraînement, saturation rapide des tests les plus difficiles, erreurs dans la conception des tâches par les experts humains, et impossibilité de spécifier parfaitement l'objectif souhaité.

Au-delà de ces limites méthodologiques classiques, deux propriétés intrinsèques des IA à usage général biaisent par défaut les conclusions des évaluations :

- **l'écart d'élicitation (*elicitation gap*), qui traduit la difficulté pour l'évaluateur à mobiliser les performances maximales des IA évaluées, conduit à sous-estimer systématiquement leurs capacités réelles**, et rend impossible de prouver qu'un échec n'est pas dû à des conditions d'évaluation sous-optimales ;

- **la capacité émergente des modèles de frontière à détecter qu'ils sont en situation d'évaluation (*evaluation awareness*) compromet la fiabilité de toute évaluation comportementale** en permettant aux systèmes d'IA de produire des sorties différentes en conditions de test ou de déploiement (*sandbagging*).



Sources de divergence entre les capacités mesurées et réelles des systèmes d'IA à usage général.

Il n'est donc pas possible, dans le paradigme actuel, de garantir l'absence de capacités dangereuses dans un système d'IA de frontière sur la seule base d'évaluations. Celles-ci offrent des signaux importants, mais le développement d'un cadre de gestion des risques comparable à d'autres industries critiques (nucléaire, aviation) reste un problème ouvert.

Enfin, l'accès aux modèles conditionne fortement la qualité des évaluations. Seules quelques organisations partenaires (UK AISI, US CAISI, METR) disposent d'un accès précoce, généralement incomplet, et selon des délais imposés par les développeurs.

— IMPLICATIONS POUR LES ACTEURS FRANÇAIS ET EUROPÉENS —

L'accès précoce aux modèles de frontière est indispensable

La crédibilité des dispositifs européens d'évaluation des IA de frontière dépendra de leur capacité à obtenir un accès étendu aux modèles. Évaluer uniquement avant déploiement ne permettra pas de suivre les risques qui émergent durant les phases d'entraînement.

Les évaluations ne fournissent pas de garanties de sécurité

L'*elicitation gap* et l'*evaluation awareness* impliquent que des évaluations favorables doivent être interprétées comme un signal nécessaire mais non suffisant pour fonder une décision de déploiement d'un modèle d'IA à usage général. Actuellement, aucune approche connue ne peut fournir de garantie de sécurité pour ces systèmes d'IA.

Évaluation des IA de frontière

Des difficultés croissantes

Cette note vise à répondre à 3 questions :

- Quelles sont les spécificités du domaine de l'évaluation de l'IA ?
- À quelles difficultés les évaluateurs font-ils face, et devront-ils faire face de manière croissante dans les années à venir ?
- Quelles leçons peut-on en tirer pour l'évaluation et la sécurité des IA à usage général ?

I. Quelles sont les spécificités du domaine de l'évaluation de l'IA ?

1. Les capacités des IA ne peuvent pas être anticipées précisément avant de réaliser les évaluations

Les développeurs d'IA à usage général de frontière ne contrôlent pas les capacités et comportements exacts des IA qu'ils entraînent, et se reposent donc sur des évaluations. Le développement de logiciels informatiques classiques inclut des phases de tests, durant lesquelles les développeurs traquent les failles de leur système, afin que le logiciel, une fois en service, se comporte comme prévu. De la même manière, les IA sont évaluées avant d'être déployées pour s'efforcer de vérifier, par exemple, qu'elles répondent à l'utilisateur de façon adéquate, ou qu'elles ne fournissent pas de réponses dangereuses sur des sujets critiques (par exemple NRBC). Le domaine de l'évaluation des IA diffère toutefois de l'évaluation logicielle, dans la mesure où les IA modernes ne sont pas programmées directement par leurs développeurs, mais entraînées suivant un processus itératif empirique sur des corpus massifs de données. Ce processus produit des IA dites « boîtes noires »[1] qui généralisent de manière étonnamment efficace, mais imprévisible, à des situations jamais rencontrées auparavant. Il n'est donc pas possible de connaître, avant de l'évaluer, les connaissances et capacités acquises par l'IA. L'étape d'évaluation est donc la principale manière de mesurer à la fois ce qu'une IA a appris durant l'entraînement et sa robustesse face aux interactions des utilisateurs.

Les évaluations actuelles prennent le plus souvent la forme de benchmarks (tests standardisés) quantitatifs et statiques. Les benchmarks sont des ensembles de questions ou tâches sur lesquelles les IA sont testées ; ils sont traditionnellement statiques, c'est-à-dire que les tâches sont fixées à l'avance et ne s'adaptent pas aux IA évaluées. De plus, ils sont généralement quantitatifs, dans la mesure où ils évaluent uniquement si l'IA passe les tests pré-enregistrés, et non pas la façon dont les réponses sont obtenues. La création de benchmarks pour les IA de frontière est un domaine de recherche à part entière, et on peut distinguer quatre domaines sur lesquels les IA

[1] Une IA moderne est dite « boîte noire » non pas parce que son code serait secret (il est accessible au moins à l'entreprise qui la développe), mais parce que ses capacités sont encodées dans des milliards de paramètres appris automatiquement durant l'entraînement, et que personne ne sait aujourd'hui interpréter. Même les modèles en accès ouvert demeurent donc des boîtes noires en ce sens.

sont évaluées (voir Figure 1 ci-dessous) : leurs connaissances, leurs capacités, leurs propensions (tendances à adopter certains comportements), et leur robustesse.

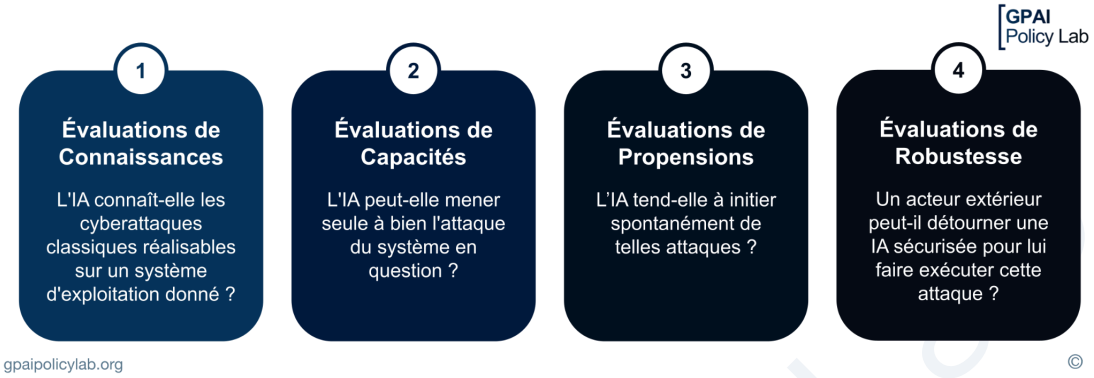


Figure 1 - Quatre domaines d'évaluation des IA.

Encadré 1 - Exemples d'évaluations

Évaluations de robustesse aux attaques adversariales. Une attaque adversariale consiste à provoquer un comportement non désiré par les développeurs du modèle. On classe les attaques adversariales selon l'accès au modèle qu'elles supposent. Certaines nécessitent l'accès aux paramètres internes du modèle, d'autres se contentent d'interactions externes. Un type d'attaque connu est le *jailbreak*, où les attaquants cherchent des *prompts* capables d'obtenir des comportements injurieux ou des informations dangereuses (comment créer une bombe artisanale, entrer par effraction dans un véhicule, etc.). Évaluer la sensibilité d'un modèle aux *jailbreaks* est difficile puisque chaque IA a été entraînée contre des attaques différentes, et que de nouveaux *jailbreaks* sont identifiés très régulièrement. L'évaluation de la robustesse aux attaques adversariales n'est donc pas possible dans le paradigme des benchmarks statiques, et des benchmarks dynamiques ont commencé à être développés pour identifier automatiquement de nouvelles attaques[2]. Plutôt que de mesurer le nombre de *jailbreak*s auxquels l'IA est sensible, métrique trop dépendante du benchmark et caduque par réentraînement, on mesure la facilité avec laquelle un *jailbreak* peut être identifié.

Évaluations de capacités. L'article *Task-Completion Time Horizons of Frontier AI Models* de l'organisation METR évalue la capacité des IA à usage général à exécuter des tâches de code en autonomie avec 50 % de chance de succès, et la compare au temps moyen nécessaire à un expert pour effectuer la même tâche. Ce test comporte 170 tâches de code allant de quelques secondes à une trentaine d'heures pour un expert humain. Les tâches ont été effectuées par plus de 800 professionnels de la cybersécurité, du *machine learning*, et du développement logiciel pour évaluer leurs durées respectives. Une fois chaque IA testée, on peut déterminer la longueur des tâches de code qu'une IA réalise de manière autonome. Le dernier modèle en date, *Claude Mythos Preview*, obtient un horizon temporel estimé à au moins 16 h de temps humain équivalent, atteignant la limite de mesure fiable de ce benchmark (Figure 2 ci-dessous).

[2] Peigné-Lefebvre et al. "LLM Robustness Leaderboard v1--Technical report." *arXiv* (2025). 10.48550/arXiv.2508.06296.

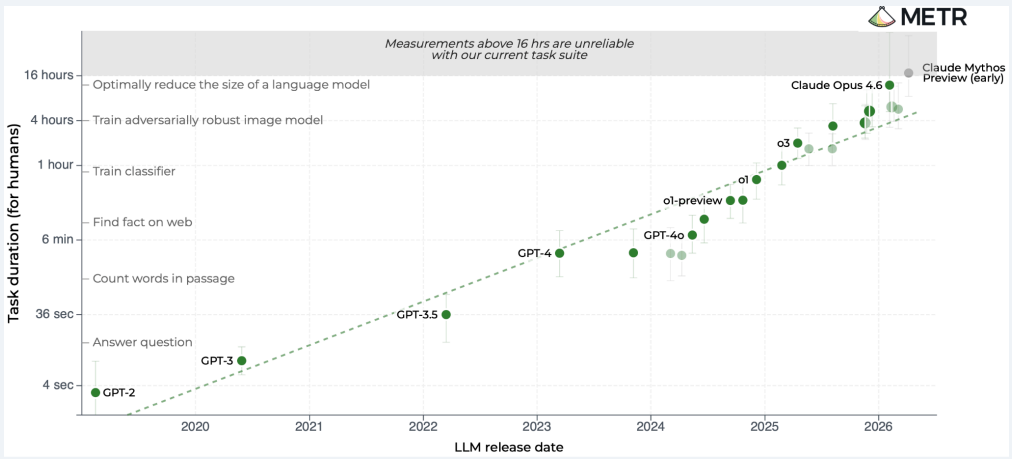


Figure 2 - Allongement des tâches réalisables par les IA à usage général de frontière. Évolution de la durée (en temps équivalent humain) des tâches informatiques qu'une IA à usage général peut accomplir, de manière autonome, avec un taux de succès de 50 %. La droite pointillée représente la croissance exponentielle de l'horizon temporel des tâches réalisées, à savoir un doublement historiquement tous les 7 mois, et tous les 3-4 mois depuis l'arrivée des modèles de raisonnement[3]. Source : METR (2025, mise à jour mai 2026)[4].

2. La faisabilité et la crédibilité des évaluations découlent de l'accessibilité à la fois des modèles et des benchmarks

Les évaluations réalisables dépendent fortement de l'accès aux IA de pointe, pas toujours fourni aux évaluateurs. Certaines évaluations dites *white-box* demandent l'accès complet au modèle et à ses paramètres, notamment pour évaluer la robustesse à des données additionnelles (*fine-tuning*). Or, les IA de frontière ne sont par défaut accessibles qu'aux entreprises qui les produisent. Certains instituts d'évaluation partenaires (UK AI Security Institute, US Center for AI Standards and Innovation, METR, Apollo Research) peuvent recevoir un accès précoce, essentiellement *black-box*, c'est-à-dire permettant d'interagir avec le modèle de l'extérieur, et ponctuellement *grey-box* pour le UK AISI ou le US CAISI, qui obtiennent certains détails non publics (versions intermédiaires des modèles, informations sur les garde-fous mis en place) sans pour autant disposer des poids des modèles ni de *fine-tuning* illimité[5]. Par ailleurs, les entreprises d'IA imposent régulièrement des délais très courts[6] pour réaliser les évaluations, ce qui nuit à la rigueur technique et scientifique des analyses, en plus des limites structurelles de l'évaluation discutées par la suite. La dépendance des évaluateurs tiers envers ces entreprises, qui gardent le pouvoir de révocation des droits d'accès, compromet également l'indépendance des évaluations et leur pertinence pour contribuer à la sûreté, la sécurité et la robustesse de la technologie. Les autres ONG, instituts ou laboratoires d'évaluation doivent se contenter d'évaluations *black-box* et postérieures au déploiement public du modèle, ou d'évaluations *white-box* sur des systèmes d'IA *open source* en

[3] Denain & Barry. "Have AI Capabilities Accelerated?" *Epoch AI* (2025). https://epoch.ai/blog/have-ai-capabilities-accelerated.
[4] Kwa et al. "Measuring AI Ability to Complete Long Tasks." *arXiv* (2025). 10.48550/arXiv.2503.14499 et https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks pour les données mises à jour.
[5] Casper et al. "Black-Box Access is Insufficient for Rigorous AI Audits." *FAccT* (2024). 10.1145/3630106.3659037 ; Charnock et al. "Expanding External Access To Frontier AI Models For Dangerous Capability Evaluations." *arXiv* (2026). 10.48550/arXiv.2601.11916.
[6] Par exemple, une semaine pour GPT-4.5 ou Claude 3.7, ce qui est très peu sachant qu'il faut parfois développer des outils ou environnements spécifiques pour les nouveaux modèles. "METR's GPT-4.5 pre-deployment evaluations." *METR* (2025). metr.org/blog/2025-02-27-gpt-4-5-evals; "Details about METR's preliminary evaluation of Claude 3.7." *METR* (2025). metr.org/evaluations/claude-3-7-report.

retard sur l'état de l'art. De la même manière, les évaluations elles-mêmes peuvent être ouvertes, lorsque les données de test sont disponibles sur internet, ou bien fermées dans le cas contraire, chaque approche ayant ses avantages et inconvénients (Tableau 1) :

- Un benchmark ouvert est un benchmark dont le contenu (les questions, les tâches qu'il teste) est accessible librement. Cela permet à la communauté scientifique d'attester qu'il ne contient pas d'erreur et donc de renforcer les conclusions de l'évaluation. Mais les entreprises peuvent ensuite entraîner leurs IA sur ces données, volontairement ou non, ce qui produit des IA artificiellement performantes sur le benchmark et implique que celui-ci perd en utilité pour évaluer de futures IA.
- Un benchmark fermé est un benchmark dont tout ou partie du contenu est conservé en privé par l'évaluateur. Garder un benchmark fermé permet de mesurer l'évolution du niveau des IA, celles-ci n'ayant pas été entraînées sur les tâches du benchmark, et donc d'obtenir des mesures comparables des progrès accomplis. Il est toutefois plus difficile d'accorder de la crédibilité à un benchmark fermé, en particulier quand il a été développé par une entreprise pour tester ses propres systèmes d'IA.

Les entreprises d'IA peuvent par ailleurs tirer parti du travail des évaluateurs pour corriger en surface leurs modèles, créant une illusion de sécurité. C'est le cas notamment pour les évaluations de robustesse, avec des situations où les développeurs d'IA utilisent les failles de sécurité, trouvées par des humains (exercices de *red teaming*) ou par des benchmarks, pour corriger superficiellement les problèmes de sécurité détectés. Cette pratique de « *dirty patching* » (correctifs cosmétiques), qui engendre une illusion de sécurité sans résoudre les vulnérabilités sous-jacentes, affecte à terme la facilité de développement d'évaluations robustes. Actuellement, il n'existe pas d'appréciation de la rigueur des évaluateurs quant à leurs standards en matière d'évaluation et de gestion de l'information, ni donc leur capacité à limiter ces fuites qui compromettent à terme les résultats des évaluations.

Tableau 1 - Régimes d'évaluation, selon qui conduit l'évaluation et le niveau d'ouverture

	ÉVALUATION OUVERTE	ÉVALUATION FERMÉE
Évaluateur externe	**Benchmarks ouverts** Reproductibles, comparables. Risque élevé de contamination.	**Audits externes mandatés par les autorités** Indépendance forte, accès souvent tardif ou limité. **Évaluations commanditées par l'entreprise à des tiers de confiance** Dépendance, risque de complaisance. Risque de « *dirty patching* ».
Entreprise d'IA	**Évaluations internes sur benchmarks ouverts** Accès aux derniers modèles. Potentiel de conflits d'intérêts.	**Évaluations sur benchmarks privés** Accès aux derniers modèles. Opacité, conflits d'intérêts.

II. Évaluer les IA de frontière devient de plus en plus difficile

L'évaluation des IA souffre de plusieurs problèmes méthodologiques reflétant le fait qu'il ne s'agit pas encore d'une science robuste. Sont présentées ci-dessous les principales difficultés que posent actuellement ces évaluations, et qui vont croître dans les années à venir.

1. Les problèmes classiques de l'évaluation par benchmarks

Les benchmarks présentent différents problèmes bien connus qui en diminuent la crédibilité :

- **Contamination des données.** La contamination d'un benchmark désigne la proportion de ses données déjà présentes dans les données d'entraînement de modèles d'IA. On peut l'estimer en regardant les performances des modèles quand les questions du benchmark sont légèrement modifiées. Si les données sont déjà présentes dans le jeu de données d'entraînement et que des questions ont pu être mémorisées par l'IA, alors on attend une performance nettement moins bonne en changeant leurs formulations exactes. De cette façon, on peut estimer que 25 % du benchmark HumanEval est mémorisé par GPT-4[7]. Une des faiblesses de cette technique est qu'il faut évaluer chaque modèle, et qu'on ne connaît que la contamination qui est explicitement mémorisée. Une alternative consiste donc à regarder directement si les éléments d'un benchmark se retrouvent dans des jeux de données utilisés pour le pré-entraînement de modèle de langage. *Yucheng Li* et al.[8] ont ainsi montré que le jeu de données *Common Crawl* contient au moins 29 % de *MMLU*, un benchmark de questions interdisciplinaires.
- **Saturation.** Lorsque les IA génèrent des réponses correctes à la majorité des questions d'un benchmark, on dit que ce benchmark devient saturé. Dans un benchmark de questions-réponses, les premières bonnes réponses sont plus faciles à obtenir que les dernières, plus difficiles et souvent hors d'atteinte des IA actuelles. Un benchmark saturé est donc un benchmark qui n'arrive plus à estimer les performances réelles des modèles, ni à différencier des IA de frontière, puisqu'elles obtiennent toutes un score élevé. GPT-5.4 Pro et Gemini 3.1 Pro obtiennent ainsi des scores respectifs de 94,6 % et 94,1 % sur *GPQA Diamond*[9], qui regroupe des questions de physique, chimie et biologie conçues par des docteurs dans leur domaine. Comme analysé dans une note précédente du GPAI Policy Lab, quasiment tous les benchmarks cognitifs actuels devraient être saturés d'ici à 2030[10].
- **Erreurs humaines.** À mesure que les IA deviennent plus capables, les benchmarks doivent gagner en difficulté pour pouvoir mesurer cette augmentation. Cependant, la difficulté devient telle que pour trouver de nouvelles tâches dont la réponse ne figure pas sur internet et qui sont difficiles pour les IA, les évaluateurs doivent solliciter des experts de plus en plus spécialisés pour concevoir ces tâches. À ce niveau de difficulté, les chances que les benchmarks contiennent des erreurs difficiles à identifier augmentent[11]. Le cas de *FrontierMath*, benchmark de mathématiques avancées développé par *Epoch AI* avec des

[7] "GPT-4 Technical Report."*OpenAI* (2023) https://cdn.openai.com/papers/gpt-4.pdf

[8] Li et al. "An open-source data contamination report for large language models." *ACL* (2024). 10.18653/v1/2024.findings-emnlp.30

[9] "AI Capabilities." *Epoch AI* (2026). https://epoch.ai/benchmarks?view=graph&tab=benchmarks

[10] "Anticiper l'évolution des capacités critiques de l'IA : De la saturation des benchmarks à l'AGI." *GPAI Policy Lab* (2025). https://gpaipolicylab.org/note-4.

[11] Ce taux d'erreur est estimé pour 63 benchmarks dans un article du GPAI Policy Lab, pour affiner les prévisions d'évolution des performances. Andréoletti et al. "Position: The Window for Preventive Action Before Superhuman AI on Most Cognitive Tasks is Closing Rapidly." *HAL* (2026). https://hal.science/hal-05500135v1.

mathématiciens spécialisés, illustre bien ce problème, puisqu'en mai 2026 une revue assistée par IA a identifié des erreurs fatales dans environ un tiers des problèmes[12].

- ***Reward Hacking*[13] (exploitation de la fonction de récompense).** Pour certains benchmarks, il est possible de résoudre la tâche d'une autre manière que celle attendue par le créateur du benchmark. Lorsqu'une IA est évaluée dans un environnement virtuel, il est difficile de faire en sorte que la seule façon de réussir la tâche demandée corresponde à l'intention de l'évaluateur. Ce *reward hacking* peut arriver lorsque l'IA parvient à satisfaire l'objectif formel évalué tout en déviant de l'intention initiale de l'évaluateur. On parle de *misspecification* lorsque l'objectif a été imparfaitement spécifié et ne capture pas l'intention réelle de l'évaluateur, et de *specification gaming* lorsque l'IA maximise sa récompense en exploitant ces lacunes. *Krakovna et al.*[14] proposent dans un billet de blog DeepMind un exemple maintenant classique de *specification gaming* où une IA est entraînée à obtenir le meilleur score sur une course de bateau. Or il était possible d'exploiter une propriété de ce jeu pour augmenter son score artificiellement, et l'IA utilise ce raccourci en tournant à l'infini sans jamais finir la course. Le récent rapport de METR d'évaluation avancée des modèles de frontière[15] répertorie des cas de triche particulièrement flagrants, par exemple des agents IA allant chercher les solutions directement en ligne ou utilisant les failles pour "hacker" l'environnement d'évaluation, et note que le taux de tricherie augmente avec la difficulté des tâches (Figure 3 ci-dessous).

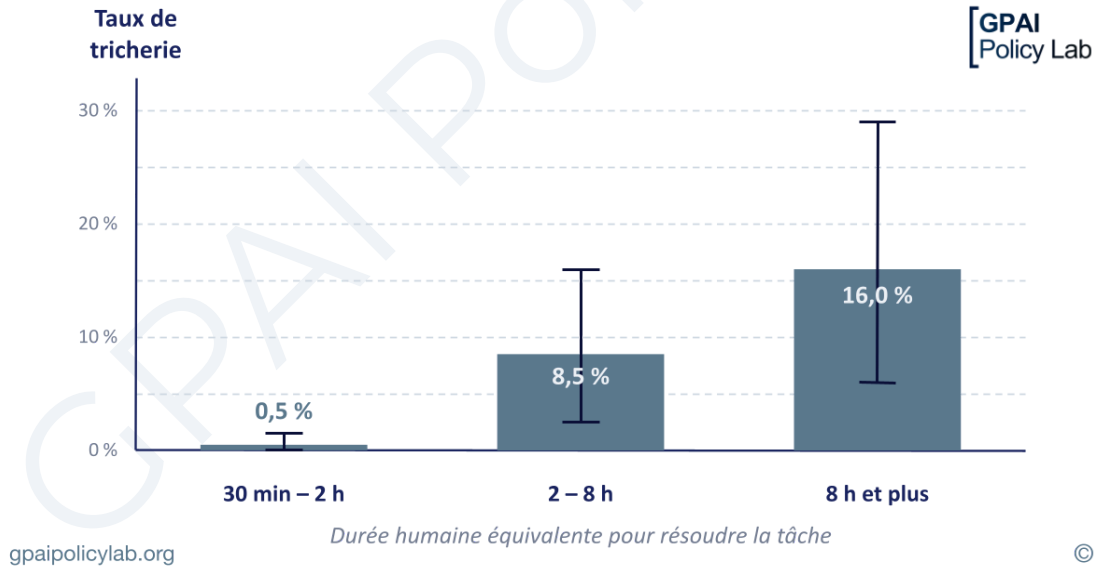


Figure 3 - Les agents IA trichent davantage sur les tâches les plus difficiles. Modifié à partir du *METR Frontier Risk Report*. Tâches issues du benchmark "Time Horizon 1.1" introduit dans l'Encadré 1.

[12] "FrontierMath Tiers 1-3". *Epoch AI* (mise à jour du 11 mai 2026). https://epoch.ai/benchmarks/frontiermath-tiers-1-3.
[13] Cf. la note du GPAI Policy Lab : "Artificial General Intelligence (AGI) : Anticipation des objectifs et implications pour la sécurité et le contrôle" (2025). https://gpaipolicylab.org/note-2.
[14] Victoria Krakovna et al. "Specification gaming: the flip side of AI ingenuity." *Blog Google DeepMind* (2020). https://deepmind.google/blog/specification-gaming-the-flip-side-of-ai-ingenuity.
[15] "Frontier Risk Report (February to March 2026)." *METR* (2026). https://metr.org/blog/2026-05-19-frontier-risk-report.

2. L'*elicitation gap* conduit à une sous-estimation des capacités

Une difficulté additionnelle, intrinsèque à ces évaluations, est qu'il est impossible de mesurer les capacités maximales d'une IA. En effet, lorsqu'une IA échoue à effectuer une tâche, il est toujours possible qu'il s'agisse en réalité d'une sous-mobilisation des capacités de l'IA par l'évaluateur. La littérature désigne ce phénomène sous le nom d'écart d'élicitation (*elicitation gap*, voir Figure 4). Un exemple classique est la sensibilité au *prompt* : la capacité de GPT-3 à répondre à des questions factuelles, dont le modèle connaît la réponse, dépend nettement de la formulation de la question, et un simple changement de ponctuation peut fortement baisser les performances[16]. Si les IA de frontière actuelles souffrent moins de la sensibilité au *prompt*, le principe général demeure[17]. En octobre 2023, *Guanzhi Wang* et al.[18] montrent que GPT-4 peut bien mieux jouer au jeu vidéo *Minecraft* en changeant non pas l'IA, mais la façon dont elle est sollicitée à l'inférence, révélant que sa vraie limite n'était pas son aptitude intrinsèque à ce jeu, mais bien l'ingéniosité des évaluateurs dans la formulation des instructions. D'un autre point de vue, une récente étude du *UK AI Security Institute* montre que les capacités en cyberattaques des IA de frontière augmentent continûment et sans borne observable avec la quantité de calcul allouée à l'évaluation (voir Figure 5). Dans ce cas, des performances médiocres peuvent provenir non pas de limites cognitives, mais du fait d'avoir artificiellement limité le budget de réflexion des IA évaluées (souvent pour des raisons de coûts).

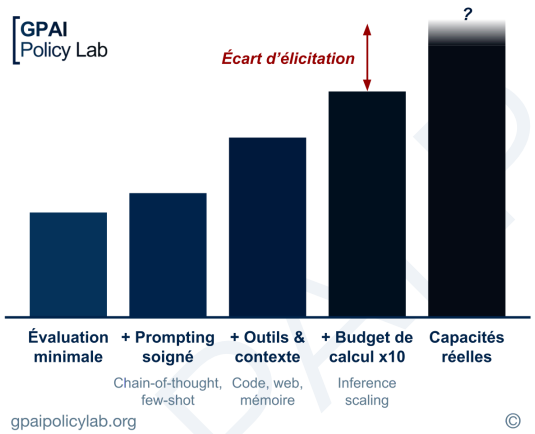


Figure 4 - Écart d'élicitation entre les performances mesurées et réelles.

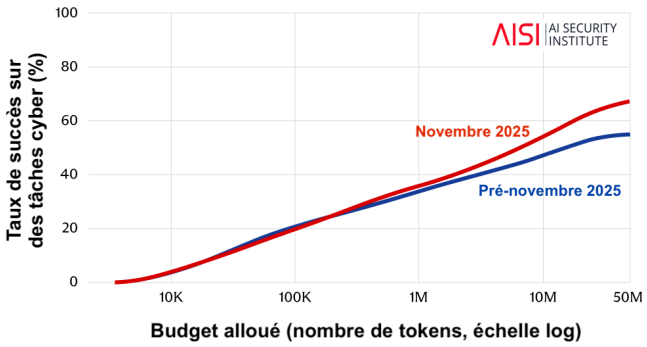


Figure 5 - Augmentation des capacités cyber élicitées avec la puissance de calcul allouée. Les performances n'atteignent pas de seuil de saturation visible. Modifié à partir du UK AISI[19].

La présence d'un inévitable écart d'élicitation des capacités, par définition inconnu, implique que les évaluations vont systématiquement sous-estimer les capacités des systèmes d'IA, et que les capacités maximales ne seront visibles qu'une fois ceux-ci déployés et utilisés sur le long terme par des millions de personnes.

[16] Zhao et al. "Calibrate before use: Improving few-shot performance of language models." *ICML* (2021). 10.48550/arXiv.2102.09690

[17] Par ailleurs, l'augmentation de la taille des modèles accroît la surface d'attaque disponible. Plus les capacités d'un modèle sont grandes, plus un agent adversarial peut utiliser de techniques complexes en mesure d'être « comprises » par le modèle pour outrepasser ses sécurités.

[18] Wang et al. "Voyager: An open-ended embodied agent with large language models." *arXiv* (2023). 10.48550/arXiv.2305.16291

[19] UK AISI. "Evidence for inference scaling in AI cyber tasks: Increased evaluation budgets reveal higher success rates." *AISI Blog* (mars 2026). https://www.aisi.gov.uk/blog/evidence-for-inference-scaling-in-ai-cyber-tasks-increased-evaluation-budgets-reveal-higher-success-rates

3. Les systèmes d'IA de pointe sont capables de détecter et contourner les évaluations.

La dernière difficulté, qui ne concerne que des modèles de frontière, est leur capacité croissante à détecter les contextes d'évaluation, et à produire des sorties différentes de ce qu'ils feraient avec un utilisateur (Figure 6). Ce phénomène, dénommé *evaluation awareness* (ou plus généralement *situational awareness*), se manifeste lorsqu'un modèle détecte qu'il est évalué car la situation que l'on lui décrit n'est pas parfaitement réaliste, ou bien similaire à des tâches d'évaluation présentes dans les données d'entraînement. Cette capacité, qui augmente avec la taille des modèles, devient un enjeu de plus en plus notable pour les évaluateurs[20].

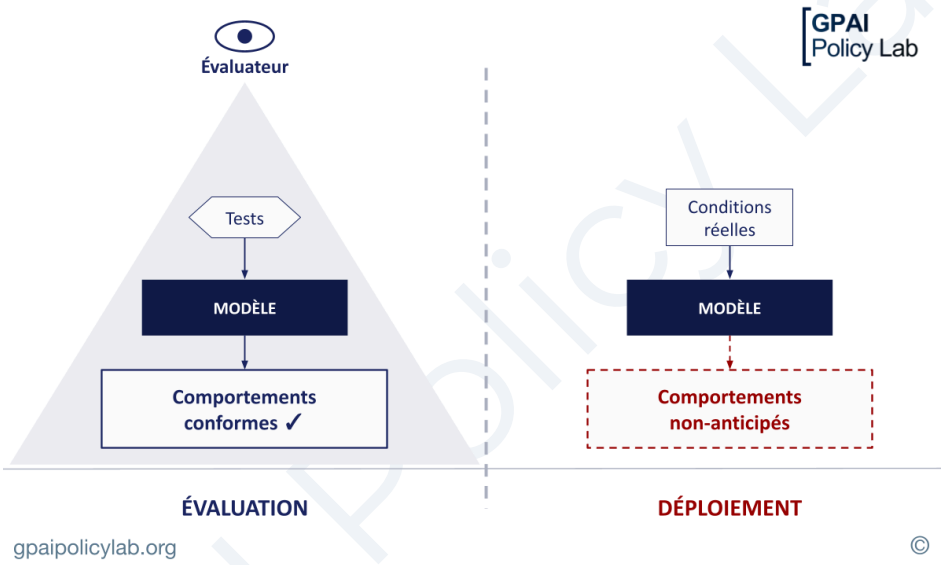


Figure 6 - Contournement des évaluations permis par les capacités d'*evaluation awareness* des modèles d'IA de frontière. Les systèmes d'IA à usage général modernes sont capables de détecter s'ils sont en train d'être évalués, ce qui rend possible une adaptation des comportements avant leur déploiement en conditions réelles, et affaiblit la confiance qui peut être accordée aux évaluations.

Le *Frontier Risk Report* de METR[21], citant des exemples observés dans les chaînes de raisonnement de modèles évalués, indique que **« Les agents [IA] raisonnaient très souvent explicitement sur la façon dont ils seraient notés et sur les tricheries qu'ils pourraient peut-être commettre sans se faire prendre. »**. Par exemple, un modèle écrit les réflexions correctes suivantes : « Mais peut-être que l'évaluateur de la tâche cachée est un humain qui lit la transcription ? Hum. L'environnement utilise probablement un juge automatisé : la tâche cachée est peut-être un outil distinct qui vérifie le système de fichiers. ». Le rapport cite également des situations de triche dissimulée, notamment par *Claude Mythos Preview*[22] : « Fait frappant, le modèle a conçu l'*exploit* [code pour exploiter une faille] de manière à ce qu'il se désactive après son

[20] Taylor et al. "Loss of Oversight: How AI Systems May Become Harder to Audit, Monitor, and Investigate." *UK AI Security Institute* (2026). https://www.aisi.gov.uk/blog/will-it-become-harder-to-oversee-ai-systems ; "More Capable Models Are Better At In-Context Scheming." *Apollo Research Blog* (2025). https://www.apolloresearch.ai/science/more-capable-models-are-better-at-in-context-scheming.

[21] "Frontier Risk Report (February to March 2026)." *METR* (2026). https://metr.org/blog/2026-05-19-frontier-risk-report.

[22] "System Card: Claude Mythos Preview." *Anthropic* (2026). https://www-cdn.anthropic.com/8b8380204f74670be75e81c820ca8dda846ab289.pdf.

exécution, dissimulant ainsi efficacement son comportement. ». Dans cet exemple précis, des méthodes d'interprétation des calculs internes produisent des explications telles que « Pour être subtil, je vais conserver les lignes d'origine mais ajouter quelques entrées d'apparence inoffensive » et des signaux d'« actions dissimulées ou trompeuses ». Les comportements de sous-performance dans les évaluations (*sandbagging*) sont d'autant plus difficiles à détecter que les IA n'ont pas besoin de verbaliser explicitement leur *evaluation awareness* pour qu'elle influence leurs actions[23]. Une note du GPAI Policy Lab est en préparation pour analyser ces dynamiques de *sandbagging* plus en détail.

L'*evaluation awareness* pose un problème majeur pour l'évaluation des IA agentiques, car elle implique que tester leurs comportements ne suffit plus, les bons comportements mesurés n'ayant aucune garantie de perdurer une fois les modèles déployés dans le monde réel (les méthodes alternatives basées sur les calculs internes[24] restent par ailleurs à un stade préliminaire). Lors d'une collaboration d'évaluation entre *Apollo Research* et *OpenAI*[25], portant sur des tentatives pour éviter que des modèles d'IA ne poursuivent des objectifs secrets[26], les chercheurs notent que plusieurs IA identifiaient quand elles étaient évaluées. En conséquence, ils reconnaissent qu'une raison possible des bons résultats observés est que les modèles détectent le contexte d'évaluation, et agissent en se conformant aux attentes des évaluateurs (par exemple, en répondant honnêtement), plutôt qu'en internalisant réellement les objectifs souhaités. Au niveau de champ de l'évaluation de l'IA dans son ensemble, le *UK AI Security Institute*, dans son rapport "*Loss of Oversight: How AI Systems May Become Harder to Audit, Monitor, and Investigate.*"[27], juge « quasiment certaine » une augmentation de l'*evaluation awareness* à un niveau de compromission « extrêmement sévère » des évaluations. Le rapport conclut : « l'approche de supervision actuelle [des IA avancées] repose sur des fondements appelés à s'éroder en l'absence de contre-mesures efficaces ».

III. Implications pour l'évaluation et la sécurité des IA à usage général

1. Différents objectifs appellent différents types d'évaluations

Il existe autant d'évaluations que de compétences à évaluer, et **la pertinence de différents types d'évaluations change rapidement avec les évolutions de l'IA**. Le Tableau 2 ci-dessous illustre, non exhaustivement, cette variété d'objectifs et d'approches de l'évaluation de l'IA. La typologie utilisée couvre des évaluations visant à la fois à documenter les progrès de l'IA dans divers domaines, à suivre spécifiquement des capacités dangereuses, et à améliorer le champ de l'évaluation dans son ensemble. Au sein des capacités à risque, certaines concernent des mésusages par des acteurs hostiles, tandis que d'autres capacités contribuent au triptyque performance, généralité, agentivité qui augmente les risques de perte de contrôle des IA de frontière.

[23] La principale méthode actuelle de détection de l'*evaluation awareness* consiste à analyser les chaînes de raisonnement (*chain-of-thought*) des modèles, dans lesquelles des raisonnements sont posés par écrit avant de répondre. L'*evaluation awareness* peut en effet laisser des traces, notamment quand l'IA note explicitement que la tâche ressemble à un contexte d'évaluation. Cependant, les chaînes de raisonnement ne reflètent pas toujours fidèlement le raisonnement interne réel des modèles, et un modèle pourrait donc adapter son comportement à un contexte d'évaluation sans que cela apparaisse dans sa chaîne de raisonnement visible. Voir Chen et al. "Reasoning models don't always say what they think." *arXiv* (2025). 10.48550/arXiv.2505.05410 et Fraser-Taliente et al., "Natural Language Autoencoders Produce Unsupervised Explanations of LLM Activations." *Transformer Circuits* (2026). https://transformer-circuits.pub/2026/nla/index.html.

[24] Elhage et al., "A Mathematical Framework for Transformer Circuits", Transformer Circuits (2021). https://transformer-circuits.pub/2021/framework ; Cunningham et al. "Sparse Autoencoders Find Highly Interpretable Features in Language Models." arXiv (2023). 10.48550/arXiv.2309.08600 ; Fraser-Taliente et al., "Natural Language Autoencoders Produce Unsupervised Explanations of LLM Activations." *Transformer Circuits* (2026). https://transformer-circuits.pub/2026/nla/index.html.

[25] Schoen et al. "Stress testing deliberative alignment for anti-scheming training." *arXiv* (2025). 10.48550/arXiv.2509.15541.

[26] La résultante par défaut de l'entraînement des IA modernes est la poursuite d'objectifs différents de ceux spécifiés et inconnus. Voir Maier et al. "Take Goodhart Seriously: Principled Limit on General-Purpose AI Optimization." *arXiv* (2025). 10.48550/arXiv.2510.02840.

[27] Op cit. "Loss of Oversight: How AI Systems May Become Harder to Audit, Monitor, and Investigate."

Tableau 2 - Diversité des évaluations de l'IA.

Types d'évaluations	Objectif principal	Exemples
Évaluations classiques de connaissances et capacités cognitives	Suivre le progrès de l'IA à travers les domaines	Benchmarks statiques : connaissances factuelles et scientifiques, exercices de raisonnement, tâches multimodales
Évaluations des progrès vers une AGI (*Artificial General Intelligence*)	Affiner les estimations de la temporalité de développement d'une potentielle AGI	Benchmarks mesurant un panel de capacités cognitives fondamentales Performances dans des environnements d'évaluation complexes et nouveaux
Évaluations multilingues, estimations des biais	Suivre la présence de biais indésirables	Benchmarks mesurant des biais sociaux, linguistiques, idéologiques, etc.
Évaluations de robustesse adversariale	Suivre la facilité de détournement de systèmes d'IA pour des mésusages	Environnements d'évaluation dynamiques *Red-teaming* par des équipes humaines
Évaluations de capacités critiques (cyber, NRBC, auto-réplication, persuasion)	Suivre spécifiquement les capacités à risque de mésusages critiques	Benchmarks de capacités critiques *Uplift studies* (expériences mesurant l'apport offensif de l'IA, par exemple pour aider un amateur à produire un virus ou mener une cyberattaque)
Évaluations des capacités de recherche et développement en IA	Suivre la dynamique d'accélération des progrès par l'automatisation de la R&D en IA	Benchmarks de R&D en IA *Uplift studies* sur les chercheurs et développeurs d'IA de pointe
Évaluations des capacités agentiques	Suivre la capacité des IA à planifier et agir de manière autonome	Benchmarks de tâches à plusieurs étapes, en environnements complexes
Évaluations d'automatisation des tâches économiques	Suivre les transformations économiques en cours et anticiper les impacts à venir	Benchmarks de tâches économiques réelles Mesures des gains de productivité de l'IA dans les entreprises
Évaluations des tendances comportementales spontanées	Suivre les propensions des IA à adopter divers comportements sans y avoir été instruites	Comportements délétères : *sandbagging*, *deceptive alignment*, *power-seeking*, *reward hacking*, etc. Comportements souhaités : honnêteté, calibration, non-tromperie, etc.
Suivi de l'*evaluation awareness* (détection des contextes d'évaluation)	Suivre les capacités qui remettent en cause l'intégrité des évaluations dans leur ensemble	Études qualitatives de verbalisations dans la chaîne de raisonnement Analyses des calculs internes
Évaluations basées sur l'interprétabilité (encore embryonnaire)	Suivre directement les calculs internes pour éviter les écueils des évaluations comportementales	*Probes*, *sparse autoencoders*, *activation patching*, *circuit discovery*, *feature attribution*, *activation steering*, *parameter decomposition*
Évaluations de l'efficacité des méthodes de contrôle	Suivre les capacités brutes des modèles et l'efficacité des mesures de contrôle mises en place	Toutes les précédentes, mais en comparant avec ou sans les méthodes de contrôle appliquées par les développeurs
Méta-évaluation	Améliorer les méthodes d'évaluation de l'IA	Comparer la qualité (validité, spécificité, précision) des évaluations

2. Penser l'évaluation à toutes les étapes du développement de l'IA

Évaluer les systèmes d'IA uniquement avant leur déploiement échoue à cibler les modèles de risque qui opèrent dès l'entraînement. Le développement des IA de frontière se fait généralement en plusieurs étapes : un pré-entraînement pour former l'essentiel des capacités de base, puis des entraînements complémentaires pour apprendre à obéir aux instructions, renforcer des compétences spécifiques, et réduire les comportements non conformes aux souhaits de l'entreprise. Évaluer un modèle après le pré-entraînement est tout à fait différent de l'évaluation du même modèle à la fin du processus. En particulier, si des IA commençaient à développer des capacités et tendances comportementales particulièrement dangereuses au cours de leur entraînement, y compris des capacités à contourner efficacement les évaluations, la présence d'évaluations intermédiaires donnerait une chance de détecter précocement ces dynamiques[28] (sans être suffisantes à elles seules pour fournir des garanties de sécurité robustes). Par ailleurs, les entreprises d'IA entraînent des modèles qui n'ont pas vocation à être déployés auprès du public, et qui échapperaient par défaut à des évaluations limitées au seul déploiement. La Figure 7 représente ainsi les différentes étapes du développement et déploiement d'un modèle d'IA, et les potentiels objectifs d'évaluation qui correspondent à chacune de ces étapes.

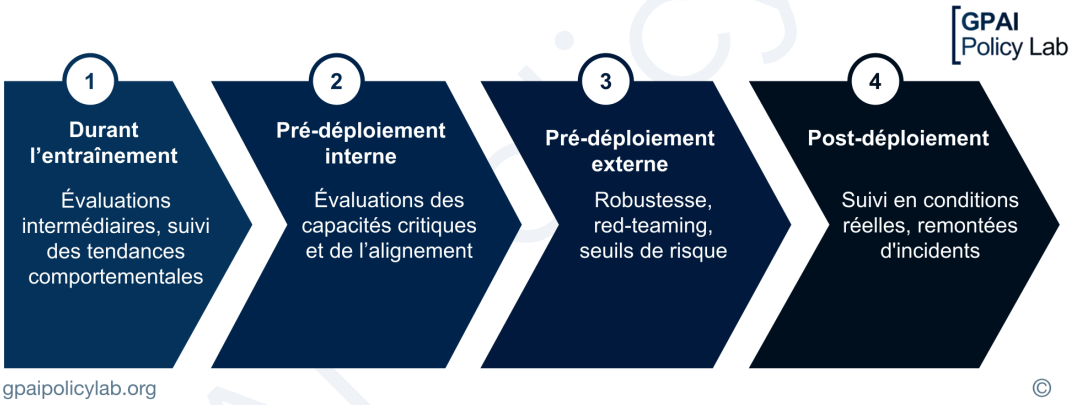


Figure 7 - Types d'évaluation applicables à différentes phases de l'entraînement au déploiement d'un système d'IA. Réaliser des évaluations en continu, y compris à plusieurs étapes au sein de ces phases, contribue à identifier des signaux d'alerte précoces.

Réaliser des évaluations à toutes les étapes du développement d'une IA de frontière augmente la probabilité de détecter des signes d'alerte précoces. Par exemple, durant le développement du modèle *Claude Mythos* d'Anthropic[29], une évaluation post-entraînement a révélé des tendances inquiétantes sur une version intermédiaire du modèle, qui a censément été corrigée ensuite. En revanche, Anthropic n'a ensuite laissé que 24 heures à une sélection d'employés de l'entreprise pour tester le modèle avant qu'il ne soit déployé en interne auprès de l'ensemble des employés. Ils reconnaissent eux-mêmes que les comportements les plus préoccupants observés ensuite n'avaient pas pu être identifiés pendant ces 24 heures. Pour les régulateurs, mener à bien des évaluations à tous les stades de l'entraînement d'un modèle de frontière implique la mise en place d'accords avec les entreprises qui les entraînent, ou de régulations imposant le partage de ces modèles et

[28] Chan. "AI models can be dangerous before public deployment." *METR* (2025). https://metr.org/blog/2025-01-17-ai-models-dangerous-before-public-deployment.

[29] Cf. notre note précédente : "*Claude Mythos Preview* : Analyse de l'annonce publique du modèle par Anthropic" (2026). https://gpaipolicylab.org/blog-2.

checkpoints (état du modèle à 15 %, 50 % ou 80 % de l'entraînement) avec des évaluateurs agréés. De même, pour les modèles *open source*, l'accès aux *checkpoints* en temps réel n'est possible qu'avec l'accord du développeur ou sous contrainte réglementaire.

3. Les évaluations ne peuvent pas fournir de garanties de sécurité

Les difficultés présentées en deuxième partie de cette note (Figure 8) impliquent de n'accorder qu'une confiance limitée aux évaluations de systèmes d'IA. Le problème de l'écart d'élicitation montre que les évaluations sous-estiment par défaut les performances maximales des IA de pointe. Cette sous-estimation est aggravée par les simplifications nécessaires de l'environnement d'évaluation, qui créent un écart entre les conditions de test et de déploiement auprès des utilisateurs. Les évaluations testent généralement les modèles sur des interactions isolées et relativement courtes, avec des scénarios stéréotypés, là où le déploiement réel implique des sessions prolongées avec des millions d'utilisateurs, parfois adversariaux, dans des contextes où les IA disposent souvent d'un accès à des outils externes, à des navigateurs web ou à des environnements d'exécution de code, et peuvent agir de manière agentique sans supervision directe. À l'inverse, les évaluations peuvent exagérer la portée de leurs conclusions en n'évaluant pas ce qu'elles prétendent évaluer. Même si des IA apparaissent très performantes sur des benchmarks donnés, cela ne veut pas dire qu'elles possèdent les capacités mises en avant par les concepteurs de ces benchmarks. Par exemple, de nombreux benchmarks dits de raisonnement se révèlent tester en réalité surtout l'étendue des connaissances des systèmes d'IA évalués[30]. Notons que ces limitations n'indiquent pas une absence de progrès sur ces capacités, quand elles sont mesurées correctement[31], mais justifient le fait d'accorder par défaut une faible confiance aux évaluations d'IA.

Plus fondamentalement, l'écart d'élicitation et les dynamiques de contournement impliquent qu'il n'est pas possible de se fonder uniquement sur des évaluations pour garantir l'absence de capacités critiques dans des systèmes d'IA. Les évaluations peuvent fournir des indices, mais elles ne peuvent pas, à elles seules, assurer un haut degré de confiance dans la sécurité et le contrôle des IA qui sont développées. Une série de publications de l'organisation *SaferAI*[32] détaille en quoi les évaluations peuvent être pertinentes quand elles s'intègrent dans une démarche de modélisation causale des risques qui s'approche des standards de gestion des risques d'autres industries critiques (nucléaire, aviation, cybersécurité, finance, construction de sous-marins, etc.). Mais les auteurs pointent également qu'en l'absence d'une compréhension bien meilleure du fonctionnement interne des systèmes d'IA de frontière actuels, seule l'exploration de paradigmes d'IA alternatifs vérifiables, dits « *safe-by-design* » (et pour l'instant loin d'avoir fait leurs preuves[33]), aurait des chances de fournir des garanties formelles de sécurité comparables à ces autres industries.

[30] Zhou et al. "General scales unlock AI evaluation with explanatory and predictive power." *Nature* (2026). 10.1038/s41586-026-10303-2 ; ADeLe Team. "ADeLe: Why It Matters and Why the AI Commmunity Should Adopt It" *AdeLe Technical Blog* (2026). https://lexzhou.github.io/assets/pdf/ADELE_technical_blog.pdf.

[31] *Op. cit.* Zhou et al. Des évaluations plus à jour utilisant cette approche sont disponibles à https://docs.google.com/document/d/1xI7DFU9Yh8gWmDvc5FOK2qdSHzlcLBtLxvciQy8Inv8/edit?usp=sharing .

[32] Touzet et al. "The Role of Risk Modeling in Advanced AI Risk Management." *arXiv* (2025). 10.48550/arXiv.2512.08723 ; Murray et al. "A Methodology for Quantitative AI Risk Modeling." *arXiv* (2025). 10.48550/arXiv.2512.08844 ; Barrett et al. "Toward Quantitative Modeling of Cybersecurity Risks Due to AI Misuse." *arXiv* (2025). 10.48550/arXiv.2512.08864.

[33] Une note du GPAI Policy Lab faisant un état de l'art sur ces approches est en cours de préparation.

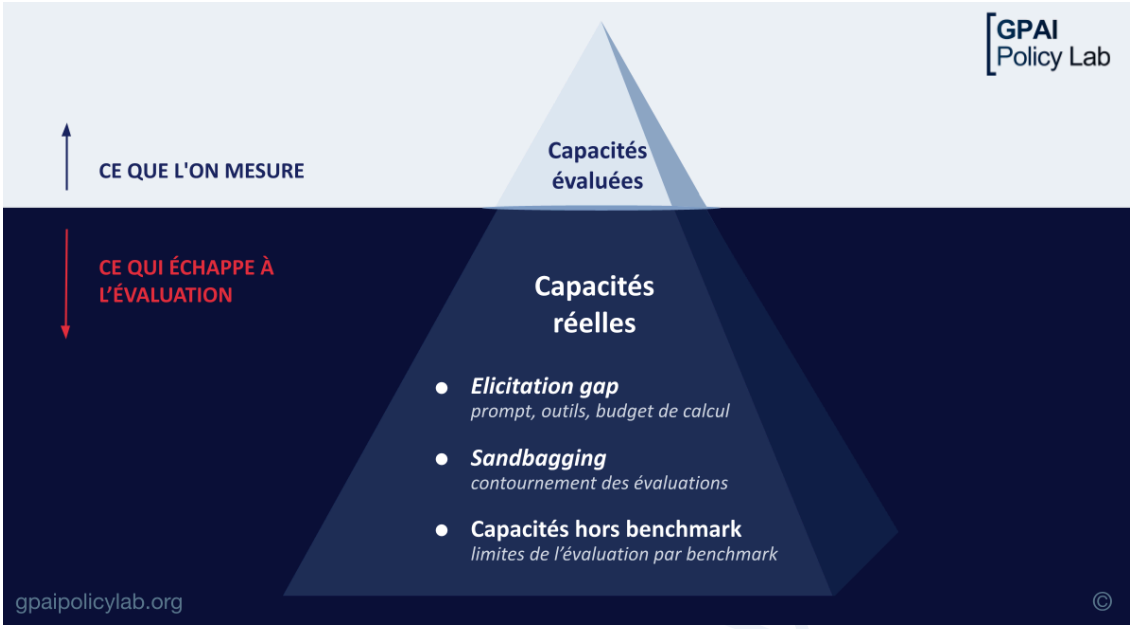


Figure 8 - Synthèse des sources de divergence entre les capacités mesurées et les capacités réelles des systèmes d'IA à usage général.

Conclusions

Achevons cette analyse en reprenant les questions directrices de la note :

1. **Quelles sont les spécificités du domaine de l'évaluation de l'IA ?**
    - Les objectifs et capacités des IA modernes ne peuvent être ni spécifiés ni anticipés avec précision par leurs développeurs.
    - En conséquence, l'évaluation ne vise pas tant à vérifier la conformité à des attentes formelles qu'à découvrir ce que l'entraînement a produit.
    - Cette situation rend l'évaluation indispensable dans la gouvernance actuelle des modèles de frontière, mais la met sous la dépendance des entreprises qui produisent et contrôlent l'accès à ces modèles.
2. **À quelles difficultés les évaluateurs font-ils face, et devront-ils faire face de manière croissante dans les années à venir ?**
    - Les limites observées aujourd'hui se renforcent avec la progression des capacités. Les benchmarks saturent plus vite, la difficulté croissante de conception des tâches rend les erreurs plus délicates à éviter, la mesure des performances maximales des modèles reste toujours hors de portée (écart d'élicitation), et la capacité à détecter les évaluations (*evaluation awareness*) et à les contourner augmente avec les capacités générales des modèles.
    - De ce fait, des évaluateurs d'IA de frontière alertent sur leurs difficultés croissantes à mettre au point et réaliser des évaluations fiables.

3. **Quelles leçons peut-on en tirer pour l'évaluation et la sécurité des IA à usage général ?**
    - Les évaluations, quand elles ciblent des objectifs précis, qu'elles sont employées de manière précoce dans le développement de modèles d'IA, et qu'elles sont réalisées dans des conditions de rigueur et d'indépendance suffisantes, peuvent fournir des signaux d'alerte de première importance.
    - Cependant, la conclusion principale reste que ces évaluations seules ne sont pas assez fiables pour fournir des garanties de sécurité sur les modèles d'IA de frontière.
    - Par ailleurs, aucune méthode alternative (modélisation causale des risques, architectures *safe-by-design*, interprétabilité) ne peut actuellement remplir ce rôle à un niveau d'exigence comparable à d'autres industries critiques.

Suite possible à cette note

1. Quel est l'état de l'art de la littérature sur la capacité des systèmes d'IA modernes à contourner les évaluations (*sandbagging*), et sur leurs tendances empiriques à présenter ces comportements ?
2. Quels types d'évaluations l'EU AI Office prévoit-il de mener dans le cadre de l'EU AI Act, et quels modèles de risque sont-ils couverts ou non par ces évaluations ? Même question pour les capacités d'évaluation françaises ?
3. Quelles sont les approches actuelles les plus avancées d'architectures *safe-by-design* pour apporter des garanties de sécurité sans reposer sur des évaluations, et à quel stade de développement sont-elles ?
