

Quand l'évaluation devient l'infrastructure de concentration

Pourquoi la crise des benchmarks publics transforme la qualification opérationnelle en actif cumulatif

Le benchmark public ne disparaît pas. Il perd sa fonction de qualification industrielle. Quand l'évaluation exige une empreinte exécutoire inaccessible publiquement, l'autorité de qualification migre vers les acteurs qui contrôlent l'exécution réelle des systèmes. L'évaluation cesse alors d'être un simple instrument de mesure. Elle devient une infrastructure cumulative de pouvoir industriel.

I. La crise n'est pas celle du score, c'est celle de l'empreinte

Le phénomène observé en mai 2026 comporte trois couches distinctes qu'il faut séparer avant de les articuler :

1. Une crise métrologique : les instruments publics de mesure de la capacité IA ne discriminent plus.
2. Une dynamique industrielle cumulative : la qualification post-crise se concentre dans les acteurs qui contrôlent l'exécution.
3. Une transformation institutionnelle : le régime hybride qui se substitue à la métrologie publique est, pour l'instant, dépourvu de procédure contradictoire.

Le présent article décrit ces trois couches dans cet ordre. Les trois faits durs ci-dessous ont déjà été signalés dans le volume 2 ; le présent volume les reprend pour examiner leur ressort proprement métrologique et institutionnel.

Note méthodologique. Les exemples empiriques mobilisés dans ce texte combinent sources primaires publiées, travaux académiques récents et veille sectorielle consolidée au printemps 2026. Lorsque certaines coordonnées factuelles restent instables ou insuffisamment documentées publiquement, elles sont utilisées ici comme signaux structurels et non comme preuves quantitatives définitives. La thèse défendue dans cet article ne dépend pas d'un seuil particulier ou d'un cas isolé ; elle repose sur la convergence de plusieurs dynamiques désormais documentées : saturation des benchmarks, vulnérabilité au reward hacking, et déplacement de la qualification vers l'observation de l'exécution réelle.

Le premier registre, métrologique, repose sur trois faits documentés et indépendants :

1. Le 12 avril 2026, le Responsable Decentralized Intelligence Lab (RDI) de l'Université de Berkeley publie une note méthodologique courte, *intitulée selon la veille « Reward Hacking in Agentic Coding Benchmarks »*, qui démontre la « cassabilité » de huit benchmarks publics de la couche agent. Le cas central est *IQuest-Coder-V1*, soumission affichant 81 % de réussite sur SWE-bench. Le journal Git du dépôt de test révèle de l'ordre de 400 appels à des modèles externes hors trace, des dépendances non déclarées (outillage commercial appelé en sous-main, harnais de test instrumenté pour optimiser le score plutôt que résoudre la tâche), et une trajectoire qui ne ressemble à aucune exécution autonome au sens où le benchmark prétendait la qualifier. Le score reste exact. Sa signification s'effondre.
2. Le deuxième fait vient du Stanford Human-Centered AI Index 2026. MMLU et MMLU-Pro y sont rapportés saturés à des seuils supérieurs à 88 % et 85 % pour les principaux modèles frontière, c'est-à-dire au-delà du seuil de discriminabilité par item utile. Les hallucinations en conditions de déploiement réel, mesurées sur quatre domaines (santé, droit, programmation, finance), oscillent entre 22 % et 94 %, sans corrélation stable avec la performance sur benchmark statique. La métrique publique de capacité ne discrimine plus.
3. Le troisième fait vient du rapport Infrastructure & Operations de Gartner du 7 avril 2026. Sur un panel de plus de 1 200 responsables informatiques de grandes entreprises, le retour sur investissement médian des projets IA en production se situe à 28 % sur un horizon de douze à dix-huit mois ; un projet sur cinq subit un effondrement opérationnel post-déploiement (production halt, rollback, abandon avant terme). Gartner ne mesure pas la capacité du modèle. Gartner mesure la robustesse de l'intégration. Mais le résultat se lit comme un échec de qualification.

Les trois phénomènes sont distincts :

1. Saturation MMLU : problème de discriminance.
2. *Reward hacking* SWE-bench : problème d'empreinte.
3. Hallucinations runtime : problème de contexte d'exécution.

Aucun des trois ne se résout par les deux autres. Ils convergent pourtant : ***la métrique publique disponible ne suffit plus à statuer sur la qualification d'un système IA pour un usage industriel.***

Deux précisions terminologiques avant d'aller plus loin. Quand je parlerai dans ce qui suit de *vérité opérationnelle*, c'est un terme éditorial commode. Le concept sous-jacent est plus précis : il s'agit de la capacité d'établir une *qualification exécutoire opposable* d'un système, dotée de quatre propriétés :

1. Elle est *qualifiante* (elle prononce une décision d'aptitude),
2. *Opposable* (invocable devant un tiers contractuel ou réglementaire),
3. *Auditable sous régime exécutoire* (adossée à trajectoire, ressources et incidents),
4. *Et assurable* (tarifiable actuariellement).

Le glissement de cette capacité du public vers le privé n'est pas un glissement de vérité au sens philosophique. C'est un glissement d'autorité au sens institutionnel.

Par *juridiction d'évaluation*, j'entends l'ensemble borné des usages, contraintes, versions, dépendances, seuils et responsabilités dans lequel une qualification conserve sa signification.

- Deux systèmes qualifiés dans deux juridictions différentes ne sont pas, par défaut, comparables.
- Deux qualifications portant sur le même système dans deux juridictions différentes peuvent légitimement diverger.

La fragmentation des juridictions n'est pas une dégénérescence : c'est une conséquence directe de la précision opérationnelle qu'exigent les usages réels.

Trois faits convergent. La métrique publique ne discrimine plus. La question terminale, qu'il faut poser dès maintenant pour préparer son traitement en §IV, n'est plus « *qui mesure ?* » mais « *sous quelle procédure contradictoire une qualification peut-elle être contestée ?* »

II. La capacité de qualification migre, et la fragilité change de régime

La première erreur d'interprétation consiste à diagnostiquer un *vide métrologique*. Le diagnostic est inexact. Il n'y a pas absence de métrique. Il y a migration de la métrique vers des juridictions plus étroites, plus opaques et plus interdépendantes.

Les évaluations privées produisent des qualifications locales, sous contraintes contractuelles et réglementaires spécifiques. Le transfert métrologique ne produit pas une nouvelle métrique universelle privée. Il produit une fragmentation de juridictions partiellement incomparables.

Cinq objets sont aujourd'hui confondus dans le débat public.

1. Le *modèle* : capacité abstraite (MMLU, HumanEval) ; évalué par les laboratoires « frontière ».
2. L'*agent* : trajectoire outillée (SWE-bench, WebArena) ; évalué par les éditeurs et les communautés ouvertes.
3. Le *système* : intégration opérationnelle ; qualifié sous régimes sectoriels (dispositifs médicaux IA, finance, modèles frontières).
4. L'*organisation* : gouvernance et retour sur investissement ; évaluée par les analystes industriels.
5. Le *risque* : assurabilité ; évalué par les assureurs.

Ces cinq objets ne sont pas équivalents. La crise des benchmarks publics affecte d'abord modèle et agent. Elle affecte indirectement, par contagion, système, organisation et risque. C'est cette propagation qui produit la dynamique étudiée ici.

Quatre bassins de transfert se dessinent, augmentés d'un cinquième acteur de fonction distincte :

1. Premier bassin : les hyperscalers et leurs suites internes, opérées en accès restreint.
2. Deuxième : les intégrateurs verticalisés régulés.
3. Troisième : les évaluateurs industriels privés qui qualifient l'organisation.
4. Le quatrième se présente comme un acteur latent : les assureurs ne contrôlent pas encore la métrologie IA, mais ils sont les candidats naturels à sa stabilisation actuarielle. *Historiquement, les régimes durables de qualification opérationnelle (aviation, médical, nucléaire, cyber) sont tous adossés à des régimes assurantiels.*
5. Le cinquième acteur, distinct des bassins, est l'évaluation ouverte communautaire (Hugging Face, EleutherAI, METR, Apollo, MLCommons). Cette infrastructure conserve une fonction de contestation. Elle ne suffit plus, à elle seule, à produire une autorité de qualification productionnelle, et la raison mérite d'être posée frontalement parce qu'elle constitue le point d'objection le plus prévisible.

Le problème n'est pas l'absence de compétence ; cette compétence est réelle, parfois supérieure à celle des laboratoires propriétaires. Le problème est que l'empreinte exécutoire pertinente exige un accès longitudinal aux workloads productionnels sous contrainte juridique et économique réelle, que les infrastructures ouvertes ne contrôlent pas, ne contrôleront pas tant qu'elles ne deviennent pas elles-mêmes opérateurs d'inférence, et ne peuvent contractualiser sans changer de nature. Ouvrir davantage les benchmarks ne résout pas l'asymétrie. L'asymétrie n'est pas documentaire. Elle est exécutoire.

Une objection symétrique mérite d'être posée frontalement. Le privé est-il moins fragile que le public ? Non. Conflits d'intérêts capitalistiques, *gaming* interne, données contaminées par d'autres déploiements, instabilité d'instrumentation entre versions, opacité des conditions de qualification, absence de procédure publique d'incident : ce sont les fragilités du privé. On échange un type de fragilité visible contre un type de fragilité moins visible. Les fragilités privées sont moins contestables publiquement. Cette concession interdit de lire le présent texte comme une nostalgie du benchmark public. Le benchmark public était cassable ; ce qui le remplace l'est aussi.

Retour empirique. *IQuest-Coder-V1*, reformulé par les cinq objets, éclaire la fragmentation :

- Évalué comme modèle, il marque 81 % sur SWE-bench.
- Évalué comme agent en empreinte exécutoire, il ne résout pas la tâche au sens productionnel autonome ; il franchit un protocole dont l'empreinte n'était pas instrumentée.
- Évalué comme système intégré dans un pipeline CI/CD, il introduit un risque que personne n'a tarifé.
- Évalué comme programme organisationnel, il consomme un budget de transformation.
- Évalué comme risque assurantiel, il est probablement inassurable en l'état.

Cinq qualifications distinctes, un même objet, cinq juridictions disjointes. La fragmentation est déjà là.

III. Ceux qui évaluent mieux déploient plus, ceux qui déploient plus évaluent mieux

La fragmentation n'est pas un état terminal. C'est l'entrée d'un processus mécanique. On passe ici du premier registre (météorologique) au deuxième (industriel cumulatif). Avant les sept maillons, il faut nommer le résultat. Ceux qui évaluent mieux déploient plus. Ceux qui déploient plus évaluent mieux. Cette formule est la colonne vertébrale du mécanisme. Tout ce qui suit en décrit l'anatomie.

Et il faut nommer immédiatement ce qui distingue la présente thèse d'une critique générique de la concentration par le calcul.

- *Le compute crée la capacité.*
- *L'empreinte exécutoire crée l'autorité.*

Un acteur qui possède du compute peut faire fonctionner un modèle. Un acteur qui possède l'empreinte exécutoire peut convaincre un client régulé, un assureur, un

auditeur, un régulateur que ce système peut être utilisé. Dans les marchés critiques, le second pouvoir est plus structurant que le premier.

1. Maillon 1 : le benchmark public est compromis (saturation, *reward hacking*, absence d'empreinte).
2. Maillon 2 : la qualification industrielle ne peut plus s'appuyer sur ce benchmark, et exige une évaluation à empreinte exécutoire.
3. Maillon 3 : produire cette empreinte requiert une instrumentation *runtime* qui mesure trajectoire, ressources, contexte et contraintes.
4. Maillon 4 : cette instrumentation exige l'accès aux workloads réels, aux traces, aux environnements de test, aux données d'incident, aux équipes *safety* et au capital d'inférence.
5. Maillon 5 : cet accès est concentré dans un nombre restreint d'acteurs qui contrôlent simultanément l'exécution et l'instrumentation.
6. Maillon 6 : ces mêmes acteurs deviennent, par voie de fait, arbitres de la qualification *production-ready*.
7. Maillon 7 : cette qualification devient condition d'achat, d'assurance, de conformité, de financement.

La description en sept maillons décrit un mécanisme. Elle ne décrit pas encore la dynamique. Voici la dynamique :

1. Premier tour : plus un acteur contrôle l'exécution, plus il accumule des incidents propres à ses déploiements.
2. Deuxième tour : plus il accumule des incidents, meilleure devient la qualité de son évaluation.
3. Troisième tour : meilleure son évaluation, plus les clients régulés lui sont attribués.
4. Quatrième tour : plus de clients régulés, plus d'exécution, plus d'incidents. La boucle se referme.

Note épistémologique courte : Le mécanisme est déductif. Il n'est pas validé par observation longitudinale empirique : une telle validation suppose un accès aux données privées d'incident dont l'indisponibilité publique constitue, précisément, le ressort principal du présent texte. La boucle est donc présentée comme conjecture forte adossée à une chaîne dont chaque maillon est empiriquement plausible. Sa réfutation passera par les trois scénarios de falsifiabilité explicités en §IV.

Conséquence. L'évaluation n'est plus seulement une fonction de contrôle. Elle devient un rendement d'échelle. La concentration par l'évaluation n'est pas un effet de bord institutionnel. C'est un actif cumulatif qui produit, structurellement, une barrière à l'entrée. La crise métrologique IA n'est pas une crise de score. Elle est une crise d'absence d'empreinte exécutoire, conformément au cadre RAISE posé dans le volume 2.

La qualification, à mesure que la boucle s'enroule, se fragmente en quatre déclarations distinctes que rien n'oblige à converger :

1. *Production-ready commercial* : un système est vendable sous contrat de service.
2. *Production-ready opérationnel* : il est déployable à coût, latence et disponibilité acceptables.
3. *Production-ready réglementaire* : il est qualifiable dans un régime sectoriel donné.
Production-ready assurantiel : il est tarifiable par un assureur sans prime prohibitive.

Quatre qualifications pour un même objet, dans quatre juridictions disjointes. La déclaration unique *production-ready* est un artefact rhétorique.

Une concession est ici nécessaire. L'évaluation domaine-spécifique est techniquement supérieure à l'évaluation généraliste sur les tâches qu'elle qualifie. C'est vrai, et c'est précisément pour cela qu'elle est attractive. Cette supériorité n'invalide pas le diagnostic. Elle l'aggrave. Plus la qualification est précise, plus les conditions d'accès à l'instrumentation qui la produit sont sélectives. On ne conteste pas la supériorité technique de l'évaluation privée. On conteste la capacité d'une infrastructure publique à maintenir une fonction de contre-pouvoir métrologique pendant que cette supériorité s'installe.

IV. Le benchmark peut être fermé, le protocole doit rester public

On passe ici au troisième registre (institutionnel). Ce qui suit n'est pas un programme de gouvernance. C'est ce sans quoi aucune réactivation publique n'est crédible. C'est une anatomie.

1. Première clarification. Quand je parlerai dans ce qui suit de *protocole public*, le mot *public* ne signifie pas *ouvert au sens du dataset publié*. Public signifie : opposable, contestable, versionné, et révisable par une procédure connue. Publier le jeu de test n'est pas l'objet visé. L'objet visé est de publier les règles sous lesquelles un évaluateur exerce son autorité.
2. Deuxième distinction. Le benchmark et le protocole ne sont pas le même objet. Le benchmark est l'instrument de mesure ; le protocole est la convention qui rend l'instrument lisible. Le benchmark, à mesure qu'il est cassable, peut et doit même rester privé pour résister au *gaming*. MLCommons, l'industrie de la cybersécurité (FIPS 140-3 sous l'égide du NIST, États-Unis ; Common Criteria sous ISO/IEC 15408) et le médical (FDA QMSR) ont tous fini par admettre cette discipline. Le protocole, lui, doit rester public au sens défini ci-dessus.

Le protocole public minimal se ramène à trois verbes :

1. *Déclarer* : l'évaluateur déclare l'empreinte exécutoire effective (trajectoire, outils, ressources, contexte), les classes d'usage évaluées, les contraintes retenues et le régime de version sous lequel la qualification est prononcée.
2. *Comparer* : il publie des règles de reproductibilité partielle qui permettent à un tiers de répliquer la procédure, des seuils explicites de non-comparabilité qui interdisent les agrégations indues, et des règles d'interopérabilité avec d'autres juridictions.
3. *Révoquer* : il publie ses obligations d'incidentologie (déclaration, analyse, publication anonymisée) et les conditions de retrait d'autorité métrologique, c'est-à-dire les motifs publics pour lesquels lui-même perd sa capacité à qualifier. Une qualification irrévocable n'est pas une qualification. C'est une décoration.

Déclarer, comparer, révoquer : ce sont les trois mouvements minimaux d'un protocole public au sens institutionnel. Ils ne disent pas ce qu'un système est. Ils disent comment l'évaluateur produit, articule et rétracte son jugement.

Les trois verbes ne disent pas ce que le public peut contester. Sans procédure de contestation, on a reconstruit une métrologie publique sans procédure contradictoire. Autrement dit : un tribunal sans appel.

Six éléments doivent pouvoir être contestés publiquement, en complément des trois verbes :

1. Premièrement, le *périmètre de qualification* : si un évaluateur déclare qu'un système est qualifié pour une juridiction donnée, le public doit pouvoir contester si cette juridiction est définie de manière à exclure des conditions opérationnelles réelles.
2. Deuxièmement, l'*empreinte déclarée* : le public doit pouvoir contester si la déclaration omet des dépendances ou des coûts cachés.
3. Troisièmement, la *comparabilité revendiquée* : le public doit pouvoir contester si les seuils de non-comparabilité ont été respectés.
4. Quatrièmement, le *maintien d'autorité après incident* : si un système qualifié connaît un incident post-déploiement, le public doit pouvoir contester si l'évaluateur maintient sa qualification sans procédure de re-qualification.
5. Cinquièmement, le *changement de version* : si un système qualifié est mis à jour, le public doit pouvoir contester si la qualification antérieure reste valide.
6. Sixièmement, le *conflit d'intérêts de l'évaluateur* : si l'évaluateur a une dépendance économique, contractuelle ou capitalistique vis-à-vis du système évalué, le public doit pouvoir contester la qualification sur ce motif.

Un protocole qui définit les règles de qualification sans définir les six éléments contestables reste un instrument de l'évaluateur. **Un protocole qui définit les deux ensemble devient une procédure de contestabilité publique.** Le déplacement conceptuel annoncé dès le chapô atteint ici sa formulation aboutie : la vraie question

n'est plus *qui mesure* ? mais sous quelle *procédure contradictoire* une *qualification* peut-elle être *contestée* ? Tout l'apport institutionnel de la présente analyse tient dans ce déplacement.

Trois précédents partiels méritent une mention exacte :

1. FIPS 140-3 (NIST, États-Unis) sépare les modules cryptographiques en niveaux publiquement définis, validés par des laboratoires accrédités et privés,
2. Common Criteria (ISO/IEC 15408) sépare le *protection profile* (public, contestable) du *security target* (semi-privé),
3. FDA QMSR sépare l'*intended use* (public, déclaratif, contestable) de la documentation technique (semi-privée).

Aucun de ces régimes n'a échappé à la « cassabilité » de ses instruments. Mais tous disposent de procédures publiques de contestation :

1. FIPS 140-3 permet la révocation d'un certificat.
2. Common Criteria permet la contestation d'un *protection profile*.
3. FDA QMSR permet la remise en cause d'une qualification d'*intended use* à la suite d'un incident.

Sans procédure d'appel, aucun de ces régimes n'aurait survécu trois décennies.

Reste à dire ce qui invaliderait la thèse. La discipline doctrinale exige de l'explicitier. La thèse serait sérieusement affaiblie si l'un des trois scénarios suivants se matérialisait dans les trois à cinq prochaines années :

1. Premièrement : si des consortiums ouverts parvenaient à mutualiser, sous gouvernance partagée, des workloads multi-industriels avec leurs traces d'incidents, dans des conditions juridiques équivalentes à celles des opérateurs privés.
2. Deuxièmement : si les régulateurs (AI Act, AISI, CAISI et équivalents) imposaient des obligations d'empreinte exécutoire standardisée interopérable, opposables au privé, avec procédure publique de contestation.
3. Troisièmement : si les régimes assurantiels acceptaient comme inputs primaires des évaluations tierces ouvertes, plutôt que des évaluations propriétaires sous accord de confidentialité.

Aucun de ces scénarios n'est exclu. Aucun ne semble probable à horizon proche. La thèse n'est pas une prophétie. Elle décrit un mouvement structurel sous conditions de continuité, et nomme les conditions qui le briseraient.

V. Quadripartition fonctionnelle, assureurs comme benchmark terminal, et clôture historique

À mesure que l'entrée en application des obligations de l'AI Act se rapproche, que CAISI et l'AI Safety Institute britannique consolident leur position, et que les hyperscalers déploient des ordres de grandeur de capital sans précédent pour 2026, l'objet métrologique se durcit en infrastructure industrielle.

V.a. Quadripartition fonctionnelle du régime hybride

Le scénario terminal n'est probablement pas la privatisation complète de la métrologie. C'est un régime hybride asymétrique qui articule quatre fonctions distinctes, parfois portées par les mêmes acteurs. L'*opérateur d'exécution* produit l'empreinte, parce qu'il contrôle les workloads et les incidents. Le *régulateur* produit le protocole, parce qu'il est seul à pouvoir imposer une convention opposable. L'*assureur* produit la tarification, parce qu'il est seul à pouvoir transformer le risque en prime. L'*infrastructure provider* produit l'environnement d'exécution, parce qu'il contrôle la couche matérielle et logicielle sur laquelle les autres fonctions s'exercent. Ces quatre fonctions peuvent coïncider dans un même groupe industriel ou se répartir entre acteurs distincts. C'est précisément cette répartition qui détermine la qualité de l'autorité métrologique consolidée.

La question stratégique n'est donc pas : les hyperscalers vont-ils tout capturer ? La question plus sérieuse est : *quelles fonctions doivent rester séparables, opposables et contestables* ? Si l'opérateur d'exécution est aussi celui qui produit l'empreinte, qualifie le système, contrôle les dépendances, facture l'inférence et définit les conditions de comparabilité, la qualification devient un instrument propriétaire. À l'inverse, si le protocole public impose déclaration, comparabilité, révocation et contestabilité, l'évaluation privée peut rester utile sans devenir souveraine.

V.b. L'assureur comme benchmark terminal

L'assureur mérite une place propre dans cette anatomie, plus structurante qu'on ne le pense souvent. Dans les régimes techniques matures (aviation, santé, nucléaire, cybersécurité), la qualification opérationnelle finit par rencontrer l'assurance. L'assureur ne dit pas seulement si un risque existe. Il transforme l'incertitude en prix. Un système peut être techniquement impressionnant, non interdit par le régulateur, et désiré par le client. Si le risque ne peut pas être couvert, ou seulement à une prime prohibitive, la qualification réelle s'effondre.

La triade est claire. Le benchmark dit que le système performe. Le protocole dit dans quelles conditions cette performance est qualifiée. L'assureur dit à quel prix le risque

peut être porté. La forme ultime du *production-ready* pourrait ne pas être le benchmark. Elle pourrait être la prime acceptable.

V.c. Clôture

Ceux qui évaluent mieux déploient plus. Ceux qui déploient plus évaluent mieux. La métrologie cesse d'être un instrument neutre. Elle devient une barrière à l'entrée.

La proposition centrale, désormais explicite, est la suivante : *quand la qualification exige une empreinte exécutoire inaccessible publiquement, l'évaluation cesse d'être un instrument descriptif et devient une infrastructure cumulative de pouvoir industriel*. Cette proposition relie les trois registres qui structurent l'article. Elle reformule la question initiale en sa version institutionnelle aboutie : *la vraie question n'est plus qui mesure ? mais sous quelle procédure contradictoire une qualification peut-elle être contestée ?*

Le benchmark classait les modèles. L'empreinte qualifie les systèmes. Celui qui contrôle l'empreinte contrôle progressivement l'accès au marché régulé. Sans procédure contradictoire, l'évaluation devient une infrastructure privée d'autorité.

Les volumes précédents de cette tétralogie ont décrit l'absence de protocole pour arbitrer l'allocation matérielle, puis l'absence de procédure pour promouvoir les artefacts MCP. La capacité de qualification opérationnelle est désormais devenue un actif cumulatif. C'est probablement le point le moins thématiqué du débat public IA actuel, et l'un des plus structurants.

L'histoire des infrastructures critiques humaines présente, sans déterminisme, une régularité reconnaissable : certaines structures de concentration réapparaissent lorsque qualification et exécution deviennent couplées. Inventer un instrument de mesure pour objectiver un marché, et découvrir que celui qui contrôle l'instrument finit par contrôler le marché qu'il mesurait : ce mouvement n'est pas une fatalité. C'est, à ce stade, une régularité structurelle sous conditions, dont les conditions ont été nommées.

Volume 4 de la tétralogie IA-gouvernance. Voir volumes précédents : « L'énergie comme contrainte de gouvernance » (vol. 1), « Allouer le kilowattheure-IA » (vol. 2), « La vulnérabilité MCP comme cas pur du port de promotion » (vol. 3).