

IA 2026 · DONNÉES ET MODÈLES

LE PLAFOND DE VÉRIFICATION

La machine produit plus vite que vous ne jugez — la loi qui décide de ce que l'IA rapporte

NABIL EL MAHYAOUI

Le Plafond de Vérification

Le Plafond de Vérification

La machine produit plus vite que vous ne jugez. Modèles, données 2026 et décisions.

Nabil El Mahyaoui

Un essai-deck pour comités exécutifs — données 2026, modèles originaux

Copyright © 2026 Nabil El Mahyaoui.
Tous droits réservés.

Table des matières

Ouverture — Le chiffre que personne ne gouverne **2**

ACTE I — L'abondance **6**

0.1 Ce que fait l'abondance à une organisation. 8

ACTE II — Le goulot **11**

0.2 Le laboratoire de 8 millions d'unités . . 13

0.3 Le goulot ne se voit pas de l'intérieur . . 15

ACTE III — La loi **20**

0.4 Premier modèle : la loi du minimum . . 21

0.5 Deuxième modèle : la file d'attente . . . 24

ACTE IV — L'équation du gain **29**

0.6 Les deux façons de perdre 31

ACTE V — L'architecture **37**

0.7 Ce que font ceux qui passent 38

CODA — Le tableau de bord **44**

0.8 Les cinq instruments du comité 45

0.9 Le mot de la fin 47

Sources **50**

×100

effondrement du coût de génération d'une capacité GPT-4 : de ~30–60 $ le million de tokens (mars 2023) à moins de 0,40 $ (début 2026)

Analyses a16z « LLMflation », Silicon Data, 2026

Ouverture

Ce chiffre est l'événement économique de la décennie, et il ne figure dans aucun de vos tableaux de bord. En trois ans, le coût de production d'une unité de travail intellectuel de premier jet — une page rédigée, une analyse esquissée, un programme fonctionnel — a été divisé par cent, et la tendance de fond mesurée par les analystes est une division par dix *chaque année*, à capacité constante. Aucun facteur de production n'a jamais connu une déflation pareille : ni l'énergie, ni le transport, ni même le transistor de Moore, qui mettait deux ans à faire ce que l'inférence fait en huit mois.

Votre organisation a répondu à ce signal de prix exactement comme la théorie le prévoit :

elle consomme. Des licences, des assistants, des agents, des pilotes — et une production de premiers jets qui explose. Mais chaque premier jet, avant de valoir quelque chose, doit franchir une opération que la déflation n'a pas touchée : être *vérifié*. Lu, jugé, corrigé, assumé par quelqu'un dont la signature engage l'entreprise. Le coût de cette opération repose sur la ressource dont la quantité n'a pas bougé depuis 2022 : des heures de jugement humain compétent.

Ce livre défend une thèse en une phrase : **la valeur que l'IA rapporte à une organisation n'est pas bornée par ce que la machine peut produire, mais par ce que l'organisation peut vérifier — et cette borne obéit à des lois que l'on peut modéliser, mesurer et déplacer.**

Il est construit comme l'objet qu'il décrit : dense là où il faut juger, visuel là où il faut

voir. Cinq actes. L'*abondance* — ce que la division par cent fait à une organisation. Le *goulot* — les données 2026, mesurées sur des millions d'unités de travail réelles. La *loi* — deux modèles qui prédisent ces données au pour-cent près. L'*équation* — pourquoi l'optimum de vérification n'est ni zéro ni cent pour cent, et comment le calculer. L'*architecture* — ce que font, concrètement, les organisations qui ont surélevé leur plafond. Puis une coda : les cinq chiffres à exiger de vos équipes chaque trimestre.

Les données viennent des plus grands laboratoires naturels disponibles — 8,1 millions de pull requests, 4 800 organisations, des essais randomisés — parce que le logiciel est le seul endroit où chaque unité de travail généré par l'IA laisse une trace mesurable de bout en bout. Mais ne vous y trompez pas : ce qui s'y mesure vaut pour vos contrats, vos rapports de crédit,

vos dossiers clients. Le logiciel n'est pas le sujet de ce livre. C'est son microscope.

ACTE I

L’abondance

÷10 / an

Le coût d’une capacité d’IA constante est divisé par dix chaque année depuis 2023. La rareté a changé de camp.

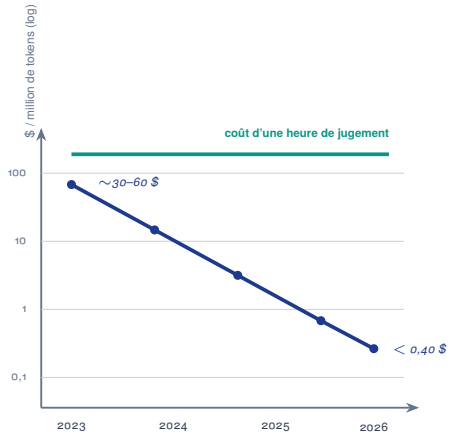


Fig. 1 : Deux courbes, une décennie : le prix d'une capacité GPT-4 s'effondre à un rythme de ÷10 par an ; le prix d'une heure de jugement humain compétent, lui, ne baisse pas. Toute l'économie de l'IA en entreprise vit dans l'écart entre ces deux lignes.

Ce que fait l'abondance à une organisation

Quand le prix d'un facteur s'effondre, on n'en consomme pas un peu plus — on en consomme démesurément plus, et pour des usages qui n'existaient pas. Les économistes appellent cela l'effet Jevons ; votre boîte mail l'appelle lundi matin. Chaque collaborateur équipé d'un modèle produit désormais en une heure ce qui en demandait dix : notes, synthèses, plans, analyses, code. La production de premiers jets n'est plus une compétence : c'est un robinet.

Le mémo d'avril 2025 de Tobi Lütke chez Shopify — l'usage de l'IA devient « une attente fondamentale », aucun recrutement sans preuve que l'IA ne peut pas faire le travail — n'était pas une provocation : c'était la reconnaissance officielle que le robinet est ouvert et

ne se refermera pas. En 2026, cette position est devenue la norme du marché du travail qualifié.

Mais une organisation n'est pas payée pour produire des premiers jets. Elle est payée pour des engagements : un contrat signable, un crédit accordable, un code déployable, un conseil dont quelqu'un répond. Entre le premier jet et l'engagement, il y a une opération de transformation — le jugement — dont le prix, lui, n'a pas bougé d'un centime. L'abondance n'a donc pas supprimé la rareté : elle l'a *déplacée*, du côté où personne ne regardait.

Quand la production devient gratuite, tout ce qui n'est pas la production devient le prix.

La suite de ce livre mesure ce déplacement. Et le premier endroit où il est devenu visible à l'œil nu, avec des chiffres d'une précision inédite, est l'industrie qui a adopté l'IA plus vite que toutes les autres : celle qui fabrique le logiciel.

ACTE II

Le goulot

+91 %

Hausse du temps d'attente en revue des unités de travail, mesurée sur 8,1 millions de pull requests dans 4 800 organisations (LinearB, 2026).

+98 %

de production livrée en revue par les équipes utilisant l'IA

+91 %

de temps d'attente avant qu'un humain la juge

LinearB, Software Engineering Benchmarks 2026 — 8,1 M d'unités, 4 800 organisations

Le laboratoire de 8 millions d'unités

Si vous voulez voir le plafond de vérification à l'œil nu, regardez l'industrie du logiciel en 2026 — non parce qu'elle est spéciale, mais parce qu'elle est *instrumentée* : chaque unité de travail y laisse une trace horodatée, de sa création à son acceptation. Aucune direction juridique, aucun département crédit ne dispose d'une telle télémétrie sur ses propres dossiers. Voici ce que disent les instruments.

Les équipes assistées par l'IA soumettent 98 % d'unités de plus à la revue et accomplissent 21 % de tâches de plus. Le temps d'attente avant jugement humain, lui, augmente de 91 %. Les unités produites par des agents autonomes attendent 5,3 fois plus longtemps d'être seulement *prises en main* que celles d'un humain. Et le taux d'acceptation sans modifi-

cation s'effondre : 32,7 % pour le travail généré par l'IA, contre 84,4 % pour le travail humain — un écart de 51,7 points qui mesure, mieux que tout discours, la charge de jugement que chaque unité générée transporte avec elle.

Le chiffre le plus grave est ailleurs, et il vient d'un autre instrument. CircleCI, qui mesure les flux de bout en bout, observe en 2026 que le débit des branches de *travail* a augmenté de 59 % en un an — pendant que le débit des branches de *production*, celui qui touche le client, a *baissé* pour l'équipe médiane. L'organisation médiane de 2026 produit davantage et livre moins. Il n'existe pas de résumé plus exact de ce livre.

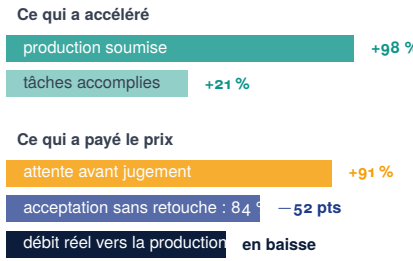


Fig. 2 : Le bilan 2026 du travail assisté par l'IA, sur données de flux réelles : l'amont explose, l'aval encaisse, le débit final — le seul qui paie — recule (LinearB ; CircleCI, équipe médiane).

Le goulot ne se voit pas de l'intérieur

Pourquoi vos équipes ne vous l'ont-elles pas signalé ? Parce que le goulot est invisible depuis l'intérieur — et c'est un résultat expérimental, pas une hypothèse. Au premier semestre 2025, le laboratoire METR a conduit un essai randomisé contrôlé — seize développeurs expérimentés, 246 tâches réelles, dans des bases

de code qu'ils connaissaient — avec et sans IA. Résultat : avec l'IA, ils étaient 19 % *plus lents*. Le plus troublant : après l'expérience, ces mêmes professionnels estimaient que l'IA les avait accélérés de 20 %. Trente-neuf points d'écart entre la performance vécue et la performance mesurée.

−19 %

vitesse réelle des experts avec IA sur tâches réelles — alors qu'ils se croyaient 20 % plus rapides. L'écart perception-réalité : 39 points.

METR, essai randomisé contrôlé, juillet 2025

La sensation de vitesse vient de la génération, qui est spectaculaire ; le temps se perd dans la vérification, qui est diffuse — relire, douter, corriger, re-prompter. Nul ne comptabilise ces minutes, donc chacun les sous-estime ; l'enquête SonarSource 2026 ferme la boucle : 96 % des développeurs déclarent ne pas faire pleinement confiance à la justesse fonctionnelle du travail généré. On vérifie ce dont on doute, et l'on doute de presque tout.

Hors du logiciel, le même phénomène a déjà son nom et son prix : le *workslop* — ce travail généré, poli en surface, creux en substance, que 41 % des salariés américains reçoivent chaque mois et qui coûte une heure cinquante-six de vérification et de refaçon par occurrence, soit 186 dollars par salarié et par mois (BetterUp Labs et Stanford, *Harvard Business Review*, septembre 2025). Votre organisation n'a pas de té-

lémétrie sur ses notes internes ; Stanford vient de vous en fournir une.

L'organisation médiane de 2026 produit davantage et livre moins. Le goulot n'est pas une opinion : c'est une mesure.

Reste à comprendre *pourquoi* ces chiffres prennent exactement cette forme — pourquoi +98 % de production donne +91 % d'attente et non +20, pourquoi le débit final peut *baisser* quand la production double. Il se trouve que deux modèles vieux d'un demi-siècle prédisent ces valeurs avec une précision embarrassante. C'est l'objet de l'acte suivant.

ACTE III

La loi

min(G, V)

Le débit de valeur d'une organisation assistée par l'IA est le minimum de deux capacités : générer, et vérifier. La première a été multipliée par cent. Pas la seconde.

Premier modèle : la loi du minimum

Une organisation qui déploie l'IA est une chaîne à deux étages : un étage de *génération*, de capacité G (unités produites par période), et un étage de *vérification*, de capacité V (unités qu'un jugement compétent peut absorber par période). Seul le travail qui franchit les deux étages a une valeur — un premier jet non vérifié n'est pas un actif, c'est un passif en attente (demandez à Air Canada, condamnée en 2024 pour les affirmations de son propre chatbot).

Modèle 1 — La loi du minimum

$$D \;=\; \min\,(\,G,\; V\,)$$

D : débit de valeur réellement livrée. G : capacité de génération — multipliée par $\sim$100 depuis 2023. V : capacité de vérification — constante à effectif et méthode constants. Dès que $G > V$, tout investissement supplémentaire dans G a un rendement marginal *nul*.

Lecture pour le dirigeant

Avant 2023, G était la contrainte : chaque euro investi dans la production (recrutement, outils) augmentait le débit. Ce réflexe budgétaire a cinquante ans ; il est désormais faux. Si votre plan IA 2026 investit dans la génération sans investir symétriquement dans la capacité de jugement, vous achetez la partie de la courbe qui ne paie plus.

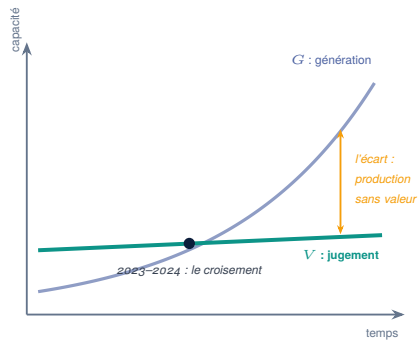


Fig. 3 : La loi du minimum : depuis le croisement, le débit de valeur suit la courbe basse. Tout ce qui est généré au-dessus de V est soit non vérifié (un risque), soit en attente (un stock).

Deuxième modèle : la file d'attente

La loi du minimum dit où est la borne ; elle ne dit pas encore pourquoi *tout se dégrade* bien avant de l'atteindre. Pour cela, il faut le deuxième modèle — la théorie des files d'attente, celle qui dimensionne les centres d'appels et les urgences hospitalières depuis un demi-siècle. Vos vérificateurs sont des serveurs ; les unités générées sont des clients ; et le temps d'attente obéit à une loi brutalement non linéaire.

Modèle 2 — Le temps d'attente explose avant la saturation

$$W \propto \frac{1}{1-\rho}$$

$$\rho = \frac{\text{flux généré}}{\text{capacité de jugement}}$$

W : temps d'attente avant vérification. ρ : taux de charge du goulot. À $\rho = 0{,}5$, l'attente est douce. À $\rho = 0{,}8$, elle a déjà été multipliée par 2,5. À $\rho = 0{,}95$, par 10. La dégradation n'est pas linéaire : c'est un mur qui se courbe.

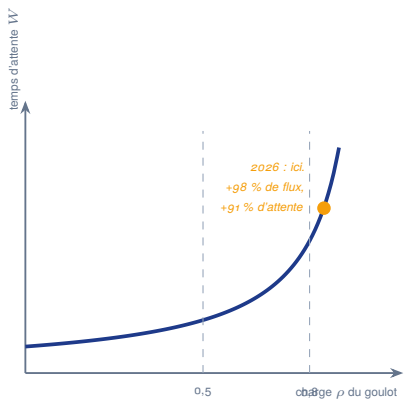


Fig. 4 : La courbe qui explique 2026 : en doublant le flux généré (+98 %) face à une capacité de jugement quasi fixe, les organisations ont poussé leur goulot dans la zone où l'attente s'envole (+91 %) — exactement ce que le modèle prédit.

Lecture pour le dirigeant

Les chiffres de l'acte II ne sont pas une crise de croissance passagère : ils sont la signature d'un système poussé au-delà de $\rho \approx 0{,}8$ sans redimensionnement du goulot. Le modèle donne aussi la seule bonne nouvelle de ce livre : près du mur, chaque point de charge retiré — par triage, par bornes, par vérification outillée — rend des gains *spectaculaires*. La non-linéarité qui vous étrangle joue dans les deux sens.

Doubler la production face à un jugement constant ne double pas l'attente : elle la fait exploser. Ce n'est pas une opinion de consultant, c'est la loi des files — la même qui dimensionne vos centres d'appels.

Deux modèles, une conclusion : le système actuel de la plupart des organisations est mathématiquement condamné à produire l'écart mesuré par CircleCI — plus de travail, moins de livraison. La question suivante n'est donc pas « comment vérifier plus » mais « combien vérifier, et où » — car tout vérifier est précisément l'erreur symétrique. C'est l'objet de l'équation du gain.

ACTE IV

L’équation du gain

ni 0 %, ni 100 %

Sous-vérifier coûte des clients et des procès. Sur-vérifier rend l'IA plus lente que rien. L'optimum se calcule.

Les deux façons de perdre

Le goulot admet deux réponses naïves, et 2024–2025 a produit le cas d'école de chacune.

La première naïveté : *ne pas vérifier* — ouvrir le robinet et faire confiance. C'est Air Canada, dont le chatbot a inventé une politique de remboursement, et dont le tribunal a rappelé qu'une entreprise répond de ce que son IA affirme. C'est Klarna, qui a poussé l'automatisation du service client jusqu'au point où son directeur général a dû reconnaître publiquement, en mai 2025 : « nous sommes allés trop loin » — avant de réembaucher des humains sur les cas complexes. Le coût de l'erreur non vérifiée n'est pas linéaire : une réponse fausse sur un remboursement coûte mille fois le prix de sa génération.

La seconde naïveté : *tout vérifier* — faire relire chaque unité générée, intégralement, par un expert. C'est le monde mesuré par METR : l'expert qui relit tout ce que l'IA produit est 19 % plus lent que s'il avait travaillé seul, parce que vérifier le travail d'un autre — humain ou machine — coûte souvent plus cher que le refaire. La vérification intégrale rend le rendement de l'IA *négatif*, en toute rigueur mesurée.

Entre ces deux falaises, il existe un optimum — et il se calcule.

Modèle 3 — L'équation du gain net

$$\Pi(\tau) \;=\; B\,D \;-\; c_v\,\tau\,D \;-\; c_e\,(1-\tau)\,\varepsilon\,D$$

valeur livrée − coût de vérification − coût des erreurs résiduelles.

τ : intensité de vérification (part du flux réellement jugée). B : valeur d'une unité livrée juste. c_v : coût de vérifier une unité. ε : taux d'erreur du flux généré. c_e : coût d'une erreur qui passe. L'optimum τ^* monte avec ε et c_e, descend avec c_v : *il n'est pas le même pour un tweet marketing et pour un avis de crédit.*

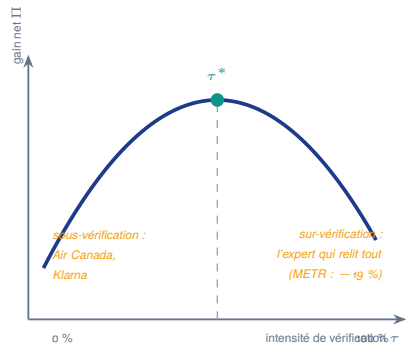


Fig. 5 : L'équation du gain : le rendement de l'IA est une courbe en cloche, pas une droite. Les deux queues sont documentées — procès et marche arrière à gauche, essai randomisé à droite.

Lecture pour le dirigeant

La question « faut-il vérifier ? » est mal posée ; la bonne question comporte trois termes que vos équipes peuvent estimer dossier par dossier : quel est le taux d'erreur du flux (ε), que coûte une erreur qui passe (c_e), que coûte une vérification (c_v) ? Un portefeuille de cas d'usage bien géré affiche un τ^* *par famille de risque* — proche de zéro pour la note interne sans conséquence, proche de un pour tout ce qui engage la signature de l'entreprise. Si tous vos cas d'usage ont la même politique de vérification, aucun n'a la bonne.

Le rendement de l'IA n'est pas une droite mais une cloche : à gauche on paie en procès, à droite on paie en experts. Diriger, c'est placer chaque cas d'usage sur sa cloche.

L'équation dit combien vérifier. Reste le levier le plus puissant, celui qui déplace la cloche entière vers le haut : changer *ce que coûte* une vérification et *où* se dépense le jugement. C'est une affaire d'architecture — et les organisations qui l'ont comprise ont publié leurs chiffres.

ACTE V

L’architecture

20 → 80

Morgan Stanley n’a pas remplacé le jugement de ses 16 000 conseillers : elle a multiplié par quatre ce que chaque heure de jugement absorbe. Le plafond n’est pas fixe — il se construit.

Ce que font ceux qui passent

Le plafond de vérification n'a pas la même hauteur partout, et ce n'est pas un hasard. Les organisations qui livrent — celles dont les chiffres sont publics — ont toutes exécuté les trois mêmes déplacements, dans le même ordre.

Premier déplacement : le jugement monte aux bornes. Chez JPMorgan, environ 75 % du trading actions s'exécute sans intervention humaine — non parce que le jugement a disparu, mais parce qu'il s'est déplacé de l'*exécution* vers la *définition des limites* : enveloppes de risque, seuils, périmètres pré-approuvés à l'intérieur desquels la machine opère seule. Une heure de jugement dépensée sur une borne gouverne des millions d'opérations ; la même heure dépensée à relire une opération n'en

gouverne qu'une. C'est le levier le plus violent de toute l'économie de l'IA, et il est purement organisationnel.

Deuxième déplacement : la vérification s'industrialise elle-même. La revue n'est pas condamnée à rester artisanale. Sur les dépôts GitHub de 2025, le travail passé par une revue *assistée par IA* se fusionne 32 % plus vite avec 28 % de défauts en moins en aval — l'IA de vérification rattrape une partie du flux que l'IA de génération déverse. Fin 2025, 30 % des entreprises de plus de mille développeurs avaient déployé au moins un outil de revue automatisée (Gartner). La règle d'architecture : chaque étage de génération déployé doit être accompagné de son étage de vérification outillée — filtres, tests, contrôles de cohérence — qui absorbe le tout-venant et n'escalade vers l'humain que le douteux.

Troisième déplacement : l'escalade se trie par risque. C'est le modèle post-2025 de Klarna — l'IA absorbe le routinier, l'humain reprend l'exception et le litige — et celui de Morgan Stanley, dont l'assistant a fait passer l'accès documentaire des conseillers de 20 % à 80 % d'efficacité *sans jamais prendre la signature* : le conseiller décide, la machine instruit. Le système signale sa propre incertitude et n'appelle le jugement que là où l'équation de l'acte IV le justifie.



Fig. 6 : L'architecture qui surélève le plafond : les mêmes heures de jugement, déplacées de la relecture uniforme vers les bornes, le tri des exceptions et la décision. C'est la traduction organisationnelle de « human-in-the-loop by design ».

Lecture pour le dirigeant

Les trois déplacements s'auto-financent : la non-linéarité de l'acte III fait qu'un point de charge retiré au goulot rend plusieurs points de délai. Mais leur ordre importe — bornes d'abord (sans elles, l'outillage vérifie n'importe quoi), outillage ensuite, tri enfin. Et une constante chez tous ceux qui publient des résultats : *la décision engageante reste humaine*, non par prudence morale, mais parce que c'est elle que tribunaux, régulateurs et clients viennent chercher — Air Canada fait jurisprudence.

Le plafond n'est pas le prix d'une heure de jugement — il est ce qu'une heure de jugement gouverne. L'architecture qui fait passer ce ratio de un à mille est la seule dépense d'IA au rendement garanti.

Un dernier acte reste à écrire, et il ne concerne ni les modèles ni les équipes : il concerne la table où vous siégez. Car un plafond ne se surélève que si quelqu'un le mesure — et ce qui se mesure tient sur une page.

CODA

Le tableau de bord

5 chiffres

Un plafond ne se surélève que si quelqu'un le mesure. Voici les cinq instruments — et ce qui disqualifie chaque réponse.

Les cinq instruments du comité

Exigez-les chaque trimestre, en chiffres. Chacun mesure une pièce des modèles de ce livre ; ensemble, ils rendent le plafond visible depuis la salle du conseil.

1. **Le taux d'acceptation sans retouche.** Quelle part du travail généré franchit le jugement sans modification ? La référence 2026 : 33 % pour l'IA, 84 % pour l'humain — votre écart mesure la charge cachée de chaque unité générée. Disqualifiant : ne pas le connaître.

2. **Le temps d'attente avant jugement.** Combien de temps une unité générée attend-elle qu'un humain la juge ? C'est le W du modèle des files : s'il croît plus vite que votre production, vous êtes dans la zone

du mur. Disqualifiant : « nos équipes s'organisent ».

3. **La charge du goulot** (ρ)**.** Flux généré rapporté à la capacité de jugement. Au-delà de 0,8, chaque pilote supplémentaire dégrade tous les autres — c'est le seul chiffre qui doive gouverner le rythme de vos déploiements. Disqualifiant : lancer un cas d'usage sans l'avoir calculé.

4. **Le coût de vérification par unité livrée** (c_v)**.** S'il ne baisse pas d'un trimestre à l'autre, votre vérification est restée artisanale pendant que votre génération s'industrialisait — l'écart de l'acte I travaille contre vous. Disqualifiant : un coût de vérification absent du business case du pilote.

5. **La part du jugement senior dépensée aux bornes.** Sur cent heures de vos

meilleurs experts, combien définissent des limites et des règles d'escalade, combien relisent du tout-venant ? La référence à battre est votre propre trimestre précédent ; la direction est connue : vers les bornes. Disqualifiant : 100 % en relecture — c'est le monde à −19 %.

Le mot de la fin

La déflation de la génération va continuer — division par dix, chaque année, aussi loin que portent les courbes. Rien de ce que vous déciderez ne l'accélérera ni ne la freinera : elle est le climat, pas la stratégie.

Ce qui dépend de vous tient dans une seule grandeur, celle que ce livre a modélisée, mesurée et outillée : la hauteur de votre plafond de vérification. Les organisations qui l'ignorent achèteront chaque année dix fois plus de génération pour le même débit de valeur — plus de travail, moins de livraison, l'équation médiane de 2026. Celles qui le gouvernent — bornes, outillage, tri, cinq instruments sur la table du conseil — transforment la même déflation en avantage cumulatif : chaque heure de jugement y gouverne, année après année, dix fois plus de production.

L'intelligence artificielle ne manquera jamais de quoi produire. La seule question stratégique de la décennie est de savoir ce que votre organisation saura *juger*.

La génération est le climat. Le jugement est la stratégie.

Sources

- **LinearB**, *Software Engineering Benchmarks Report 2026* — 8,1 millions de pull requests, 4 800 organisations : +21 % de tâches, +98 % d'unités soumises, +91 % de temps de revue ; prise en main des PR agentiques 5,3× plus lente ; acceptation sans modification 32,7 % (IA) contre 84,4 % (humain).
- **CircleCI**, données de flux 2026 — débit des branches de travail +59 % sur un an ; débit des branches de production en baisse pour l'équipe médiane.
- **METR**, *Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity*, essai randomisé contrô-

lé, juillet 2025 — 16 développeurs, 246 tâches : −19 % de vitesse réelle avec IA, perception de +20 %.

- **SonarSource**, *State of Code Developer Survey 2026* — 96 % des développeurs ne font pas pleinement confiance à la justesse fonctionnelle du code généré.

- **BetterUp Labs & Stanford Social Media Lab**, « AI-Generated Workslop Is Destroying Productivity », *Harvard Business Review*, septembre 2025 — 41 % de salariés destinataires ; 1 h 56 par occurrence ; 186 $ par salarié et par mois.

- **a16z**, « LLMflation » ; **Silicon Data** ; analyses 2026 — coût d'une capacité GPT-4 : ∼30–60 $/M tokens (mars 2023) → <0,40 $/M (début 2026) ; tendance ÷10 par an à capacité constante.

- **GitHub**, *Octoverse 2025* — revue assistée par IA : fusion 32 % plus rapide, 28 % de défauts en moins en aval.
- **Gartner**, 2025 — 30 % des entreprises de plus de 1 000 développeurs équipées d'au moins un outil de revue automatisée fin 2025.
- **Bain & Company**, enquête coûts et IA, juin 2026 — 40 % des entreprises sous leurs objectifs de réduction de coûts.
- **Moffatt c. Air Canada**, British Columbia Civil Resolution Tribunal, février 2024 — responsabilité de l'entreprise pour les affirmations de son chatbot.
- **Klarna** — communiqué février 2024 (2,3 M de conversations, équivalent 700 agents) ; entretien Bloomberg de Sebastian Siemiatkowski, mai 2025 (« we went

too far ») ; modèle hybride avec escalade humaine.

- **JPMorgan Chase**, rapport annuel 2025 — ∼75 % du trading actions automatisé dans des limites pré-approuvées ; 450+ cas d'usage en production.
- **Morgan Stanley** — efficacité d'accès documentaire des conseillers de 20 % à 80 % ; la décision reste au conseiller.
- **Shopify**, mémo de Tobi Lütke, avril 2025 — l'usage de l'IA « attente fondamentale ».