Le coût réel d’un agent en production

Étude 13 — Réagencer le travail à l’âge des agents

Mathieu Guglielmino

2026-06-19

Le fil

Journal des mises à jour

- **2026-07-06** — Ajout §7 « Le marché a tranché : il paie l’assistant, pas l’agent » : la preuve côté acheteur de la thèse, à partir de Menlo Ventures « State of Generative AI in the Enterprise 2025 » (déc. 2025). Deux schémas neufs — copilotes vs agents (7,2 vs 0,75 Md$), bascule build→buy (interne 47 %→24 %).

La thèse

Le POC ment. Tout projet d’agent commence par une démonstration convaincante — et se fracasse sur la production. Pas sur le prix du modèle, qui s’est effondré (÷280 en dix-huit mois, de 20,00 $ à 0,07 $ le million de tokens) : ce coût-là est presque résiduel. Ce qui tue un agent, c’est **tout ce qu’il y a autour** — l’intégration dans les systèmes existants, la **supervision humaine** qu’on n’avait pas budgétée, l’apprentissage du contexte métier, la maintenance quand le modèle change de version, la gouvernance quand le régulateur se retourne. Cinq pour cent seulement des outils d’IA sur mesure atteignent la production[1]. La part d’entreprises qui abandonnent la majorité de leurs initiatives IA a plus que doublé en un

[1] MIT NANDA (Project NANDA, MIT Media Lab), The GenAI Divide: State of AI in Business 2025, juillet 2025. Méthodologie : 153 dirigeants interrogés + 300 cas publics analysés. Chiffres cités : 95 % des organisations sans retour malgré 30 à 40 Md$ d’investissement ; entonnoir outils sur mesure 60 % → 20 % → 5 % (vs génériques 80 → 50 → 40) ; ~67 % (partenariat externe) vs ~33 % (build interne) de passage en prod ; ~50 % des budgets GenAI au front-office (jusqu’à ~70 % en allocation déclarée) ; ROI back-office documenté à 2–10 M$/an. https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf

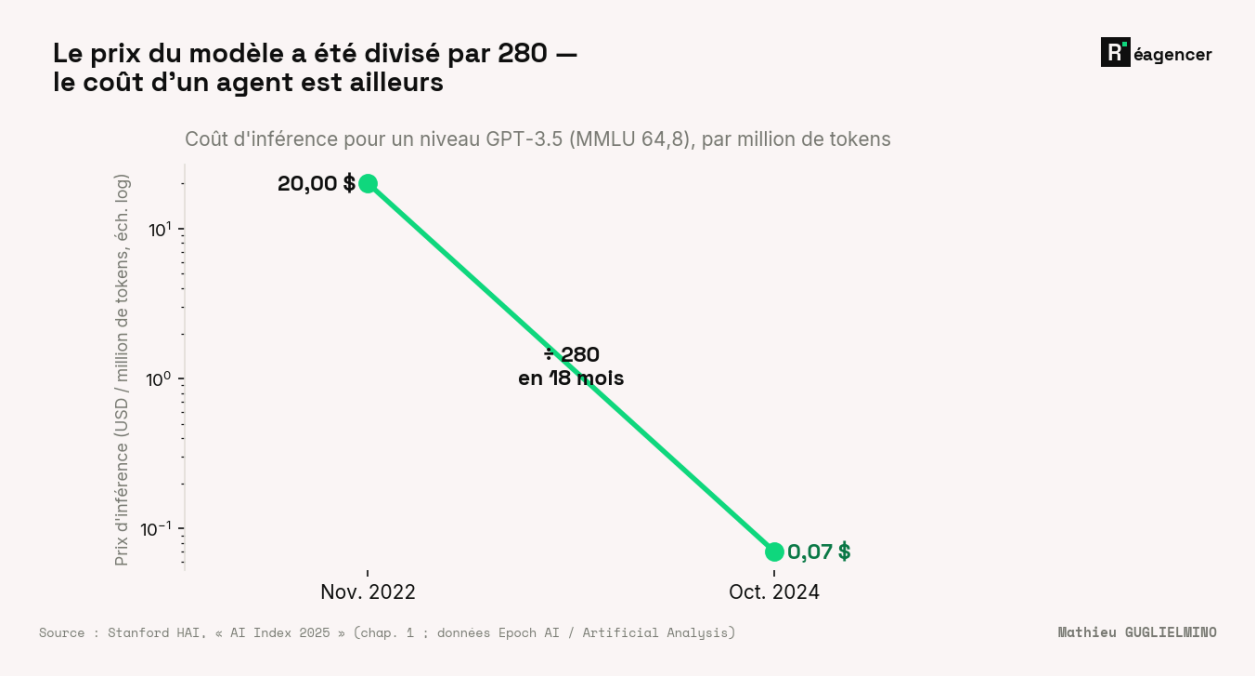
an[2]. Et pourtant 88 % déclarent « utiliser l'IA »[3] — le mot recouvre tout, et c'est précisément le problème. Corollaire Réagencer : **on ne budgète pas le modèle, on budgète l'orchestration.**

---

1. Le prix du modèle a été divisé par 280 — le coût est ailleurs

Le premier réflexe quand on calcule le coût d'un agent, c'est de regarder la facture API. C'est le mauvais endroit.

```
analyse.fig_cout_inference(d);
```



Pour un niveau de performance équivalent à GPT-3.5 (MMLU 64,8), le prix d'inférence est passé de **20,00 $** à **0,07 $** par million de tokens entre novembre 2022 et octobre 2024[4] — soit une division par **280 en à peine dix-huit mois**. Epoch AI estime la baisse à 9 à 900 fois par an selon la tâche. Ce mouvement ne s'arrête pas : la concurrence entre fournisseurs de modèles pousse les prix vers le bas à une vitesse qu'aucun autre marché technologique n'avait connue à cette échelle.

---

[2] S&P Global Market Intelligence, Voice of the Enterprise: AI & Machine Learning, mars 2025 (N > 1 000, Amérique du Nord + Europe). Chiffres cités : 42 % des entreprises ont abandonné la majorité de leurs initiatives IA en 2025 (vs 17 % en 2024) ; 46 % des POC abandonnés avant la production. Obstacles : coût, confidentialité des données, sécurité. (Rapport sous accès ; chiffres repris de la couverture CIO Dive, 14 mars 2025.) https://www.ciodive.com/news/AI-project-fail-data-SPGlobal/742590/

[3] McKinsey & Company (QuantumBlack), The State of AI 2025, novembre 2025. Chiffres cités : 88 % des organisations utilisent l'IA dans ≥1 fonction ; 39 % expérimentent des agents IA ; 23 % mettent à l'échelle un système agentique ; ≤10 % par fonction ; 6 % de « high performers » (≥5 % d'EBIT lié à l'IA) ; 39 % déclarent un impact EBIT au niveau de l'entreprise. (Page publique ; PDF de détail sous accès.) https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

[4] Stanford HAI, Artificial Intelligence Index Report 2025, chapitre 1 (R&D), 2025. Coût d'inférence pour un niveau GPT-3.5 (MMLU 64,8) : 20,00 $/M tokens (nov. 2022) → 0,07 $/M tokens (oct. 2024, Gemini-1.5-Flash-8B), soit ÷280 en ~18 mois. Données Epoch AI / Artificial Analysis. https://hai.stanford.edu/assets/files/hai_ai-index-report-2025_chapter1_final.pdf

Ce chiffre devrait clore le débat sur « le modèle est trop cher ». À 0,07 $ le million de tokens, faire tourner un agent un million de fois ne coûte pas ce qu'on imaginait il y a dix-huit mois. Le coût de l'inférence devient marginal dans le budget total d'un agent.

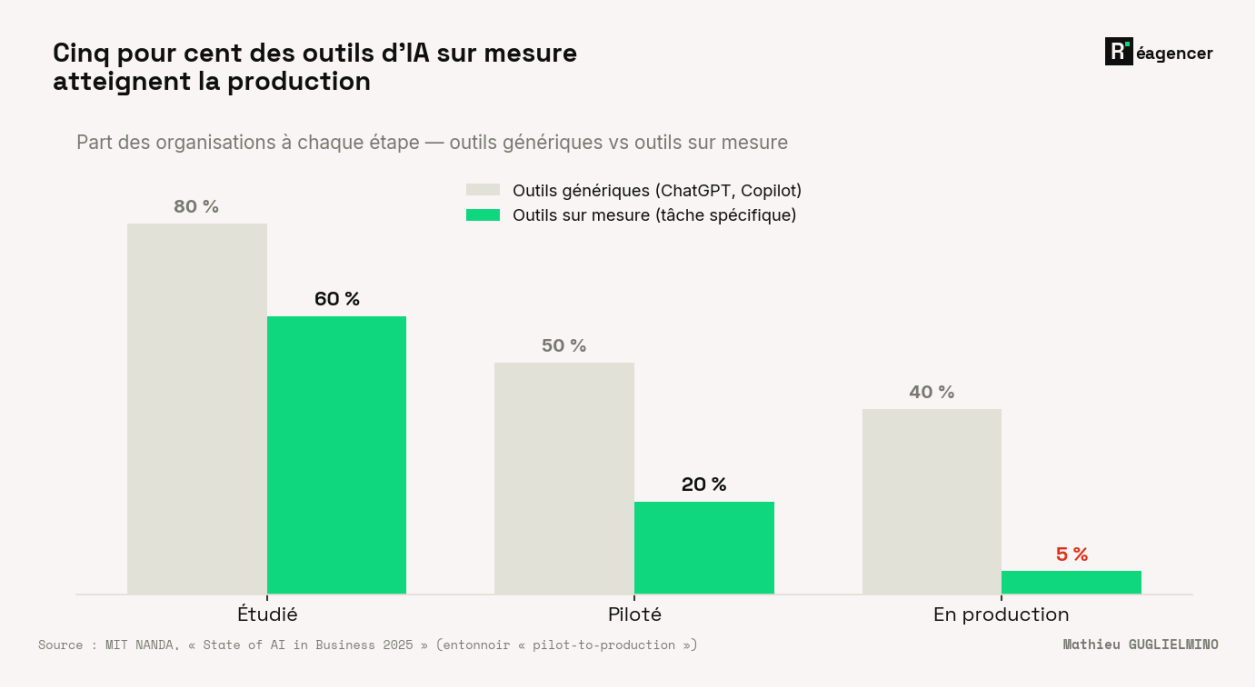
Ce qui ne devient pas marginal, c'est tout le reste. L'intégration dans les systèmes d'information — ERP, CRM, bases documentaires maison — mobilise des équipes tech pendant des mois. L'apprentissage du contexte métier (le vocabulaire propre à l'entreprise, les exceptions de processus, les cas limites) ne s'achète pas à un fournisseur. La supervision humaine — quelqu'un qui vérifie les sorties, corrige les erreurs, escalade les cas ambigus — est un poste invisible dans les maquettes de budget et souvent le plus lourd en pratique. La maintenance, enfin, recommence à chaque mise à jour de modèle. Le coût de l'inférence peut diviser par dix : aucun de ces postes ne suit.

---

2. Cinq pour cent des outils sur mesure atteignent la production

Le passage du POC à la production est le moment de vérité. C'est là que la plupart des agents meurent.

```
analyse.fig_vallee_mort(d);
```



L'entonnoir MIT NANDA[5] est brutal pour les outils sur mesure : **60 %** des organisations les étudient, **20 %** atteignent le pilote, **5 %** seulement la production. Les outils génériques — ChatGPT, Copilot et leurs semblables — s'en tirent bien mieux (80 % → 50 % → **40 %**) parce qu'ils n'exigent ni intégration profonde ni apprentissage du contexte : on les branche, on les utilise, on les facture au siège.

[5]MIT NANDA (Project NANDA, MIT Media Lab), The GenAI Divide: State of AI in Business 2025, juillet 2025. Méthodologie : 153 dirigeants interrogés + 300 cas publics analysés. Chiffres cités : 95 % des organisations sans retour malgré 30 à 40 Md$ d'investissement ; entonnoir outils sur mesure 60 % → 20 % → 5 % (vs génériques 80 → 50 → 40) ; ~67 % (partenariat externe) vs ~33 % (build interne) de passage en prod ; ~50 % des budgets GenAI au front-office (jusqu'à ~70 % en allocation déclarée) ; ROI back-office documenté à 2–10 M$/an. https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf

Pourquoi l'outil sur mesure décroche-t-il ? Parce qu'il cumule exactement les coûts que le modèle ne résout pas. Il faut l'intégrer — c'est-à-dire connecter des APIs, nettoyer des données, mapper des schémas qui ne se ressemblent pas. Il faut lui apprendre le contexte — ce que signifie « bon de commande critique » dans ce secteur précis, pourquoi le client X a une gestion dérogatoire, quel niveau de confiance accorder à quel type de sortie. Et il faut maintenir tout ça quand le modèle sous-jacent change, quand l'API change, quand le process métier change. Chacun de ces trois postes est de la main-d'œuvre qualifiée, pas du crédit API.

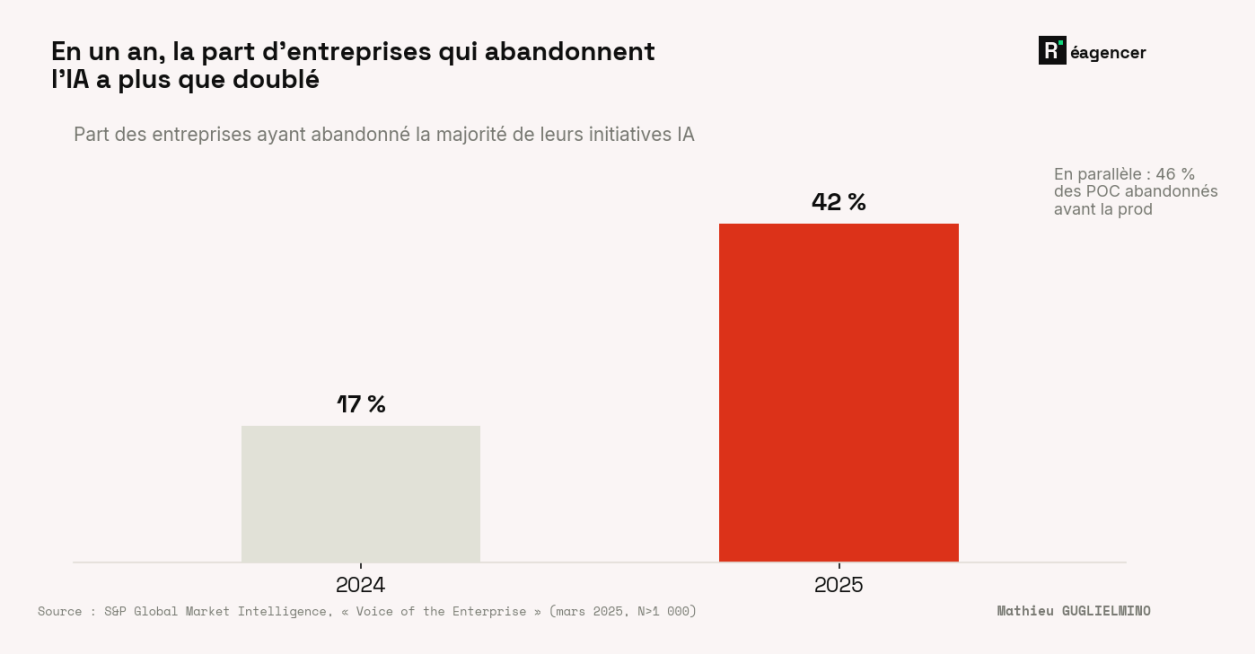
Le rapport MIT NANDA quantifie l'investissement global : **30 à 40 milliards de dollars** d'investissement GenAI en entreprise, **95 %** des organisations sans aucun retour mesuré[6]. Méthode : 153 dirigeants interrogés + 300 cas publics analysés. Ce n'est pas que la technologie soit mauvaise. C'est que l'approche — faire un POC, l'exposer à la direction, demander un budget de production — ne prépare à aucun des coûts réels.

---

3. L'« AI fatigue » est mesurée : 42 % abandonnent en 2025

Le phénomène n'est plus anecdotique. Il est documenté à grande échelle.

```
analyse.fig_abandon(d);
```



En mars 2025, S&P Global Market Intelligence a interrogé plus de 1 000 entreprises en Amérique du Nord et en Europe[7]. Résultat : **42 %** ont abandonné la majorité de leurs initiatives IA au cours

[6]MIT NANDA (Project NANDA, MIT Media Lab), The GenAI Divide: State of AI in Business 2025, juillet 2025. Méthodologie : 153 dirigeants interrogés + 300 cas publics analysés. Chiffres cités : 95 % des organisations sans retour malgré 30 à 40 Md$ d'investissement ; entonnoir outils sur mesure 60 % → 20 % → 5 % (vs génériques 80 → 50 → 40) ; ~67 % (partenariat externe) vs ~33 % (build interne) de passage en prod ; ~50 % des budgets GenAI au front-office (jusqu'à ~70 % en allocation déclarée) ; ROI back-office documenté à 2–10 M$/an. https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf

[7]S&P Global Market Intelligence, Voice of the Enterprise: AI & Machine Learning, mars 2025 (N > 1 000, Amérique du Nord + Europe). Chiffres cités : 42 % des entreprises ont abandonné la majorité de leurs initiatives IA en 2025 (vs 17 % en 2024) ; 46 % des POC abandonnés avant la production. Obstacles : coût, confidentialité des données, sécurité. (Rapport sous accès ; chiffres repris de la couverture CIO Dive, 14 mars 2025.) https://www.ciodive.com/news/AI-project-fail-data-SPGlobal/742590/

de l'année — contre **17 %** un an plus tôt. En douze mois, le taux a plus que doublé. En parallèle, l'organisation moyenne a abandonné **46 %** de ses POC avant d'atteindre la production.

Les obstacles cités ne sont pas techniques : **coût**, **confidentialité des données**, **sécurité**. Aucun de ces trois freins ne concerne le prix du modèle. Le coût, ici, c'est le coût total du déploiement — intégration, supervision, gouvernance. La confidentialité, c'est la question de savoir quelles données on peut faire transiter par quel système, avec quelles garanties contractuelles, sous quelles juridictions. La sécurité, c'est le risque qu'un agent avec des droits d'accès larges produise une sortie erronée dans un contexte à enjeux.

Ces obstacles ont un point commun : ce sont des problèmes d'**organisation et de gouvernance**, pas des problèmes de performance du modèle. On peut avoir le meilleur modèle du marché et butter sur chacun d'eux.

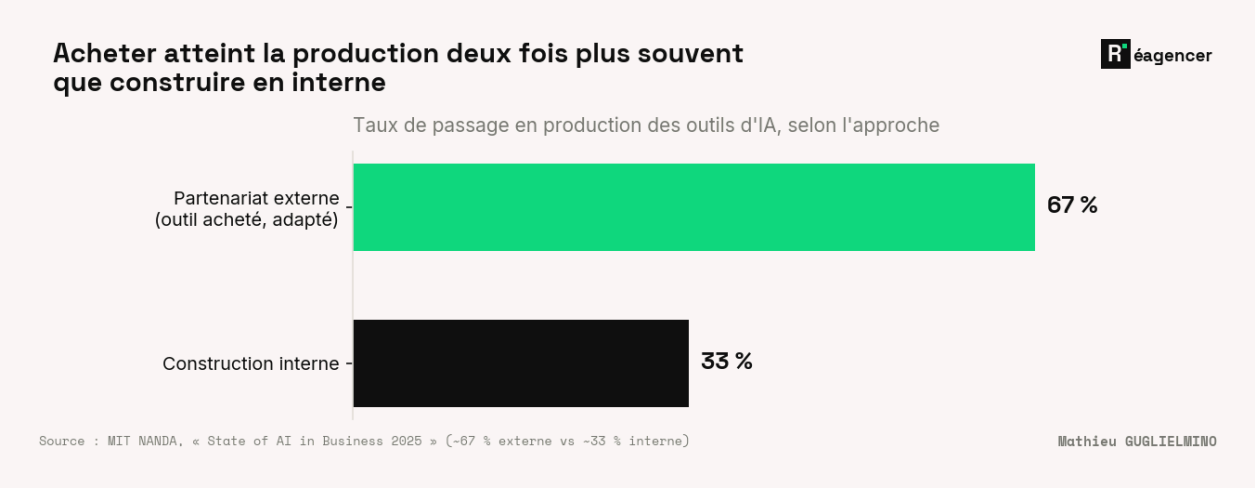
Gartner va plus loin dans son pronostic : **plus de 40 %** des projets d'IA agentique seront annulés d'ici fin 2027[8], en raison de coûts qui dérapent, d'une valeur métier floue, et d'un contrôle du risque insuffisant. Gartner pointe aussi l'« **agent washing** » : des milliers d'éditeurs rebranding un chatbot, un assistant ou un automate RPA en « agent IA » — sur lesquels seuls environ 130 d'entre eux proposeraient de vraies capacités agentiques. Le marché crée des attentes que la majorité des produits ne peuvent pas tenir, et les organisations qui achètent en font les frais quand le déploiement commence.

---

4. Acheter atteint la production deux fois plus souvent que construire

La question « faire en interne ou acheter » n'est pas seulement stratégique : elle détermine statistiquement si l'agent existera un jour en production.

```
analyse.fig_buy_vs_build(d);
```



[8]Gartner, communiqué « Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027 », 25 juin 2025. Prédiction : >40 % des projets d'IA agentique annulés d'ici fin 2027 (coûts dérapants, valeur métier floue, contrôle du risque insuffisant). Pointe l'« agent washing » : ~130 éditeurs proposant de vraies capacités agentiques sur des milliers de prétendants. (Communiqué web ; PDF non téléchargeable.) https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027

Le même rapport MIT NANDA[9] établit un écart net : un outil acheté à un éditeur et adapté atteint la production dans **~67 %** des cas ; un build interne, dans **~33 %** — deux fois moins souvent. Ce n'est pas que les équipes internes soient moins compétentes. C'est que le build interne cumule tous les coûts cachés que l'outil acheté a déjà absorbés.

Un éditeur qui vend un agent de traitement de factures a déjà résolu l'intégration aux principales ERPs, déjà entraîné le contexte sur des milliers de cas métier similaires, déjà géré la migration de modèle entre GPT-3.5 et GPT-4, déjà constitué une équipe support. Quand une organisation achète cet outil, elle achète aussi ce capital accumulé. Quand elle le reconstruit en interne, elle l'accumule elle-même — au prix fort, sur sa propre timeline, avec ses propres erreurs.

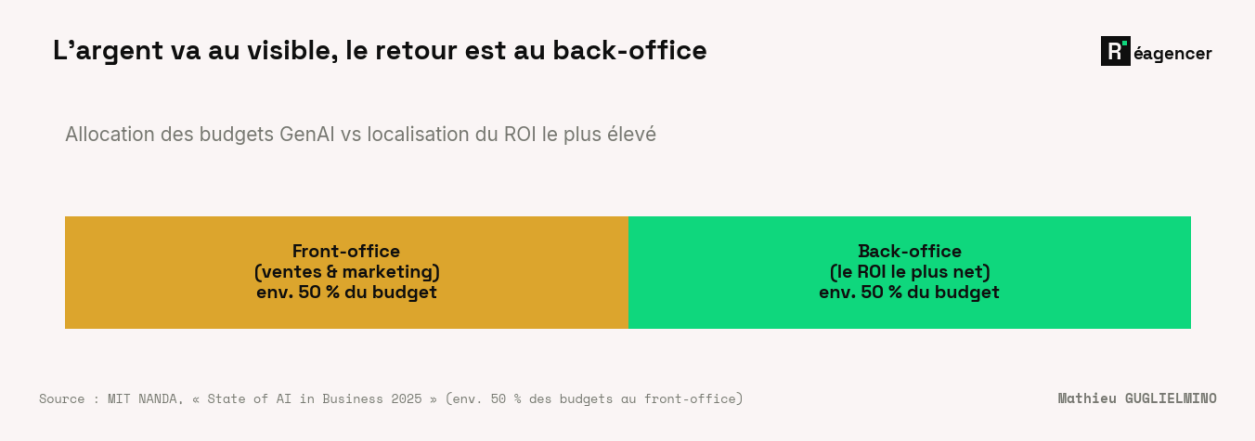
Le coût caché du build interne, c'est précisément ce que le POC ne montre pas. Le POC est fait par l'équipe qui maîtrise le contexte, dans un cadre simplifié, avec des jeux de données propres. La production, c'est l'équipe entière, le contexte dans toute sa complexité, les données dans toute leur irrégularité. L'écart entre les deux est un coût humain, pas computationnel.

---

5. L'argent va au visible — le retour est au back-office

Quand les organisations qui ont atteint la production regardent où elles ont dépensé et où elles ont gagné, le constat est systématique : l'argent et le retour ne vont pas au même endroit.

```
analyse.fig_budget_roi(d);
```



Selon MIT NANDA[10], **environ 50 %** des budgets GenAI en entreprise sont alloués au front-office — ventes, marketing, relation client, expériences visibles et facilement pitchables à la direction. Certains dirigeants déclarent jusqu'à **~70 %** d'allocation front-office dans les simulations de budget (test des 100 $, page 9 du rapport). Pourtant, le ROI net le plus élevé est au **back-office** : automatisation de processus administratifs, réduction des coûts BPO, traitement de données

[9]MIT NANDA (Project NANDA, MIT Media Lab), The GenAI Divide: State of AI in Business 2025, juillet 2025. Méthodologie : 153 dirigeants interrogés + 300 cas publics analysés. Chiffres cités : 95 % des organisations sans retour malgré 30 à 40 Md$ d'investissement ; entonnoir outils sur mesure 60 % → 20 % → 5 % (vs génériques 80 → 50 → 40) ; ~67 % (partenariat externe) vs ~33 % (build interne) de passage en prod ; ~50 % des budgets GenAI au front-office (jusqu'à ~70 % en allocation déclarée) ; ROI back-office documenté à 2–10 M$/an. https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf

[10]MIT NANDA (Project NANDA, MIT Media Lab), The GenAI Divide: State of AI in Business 2025, juillet 2025. Méthodologie : 153 dirigeants interrogés + 300 cas publics analysés. Chiffres cités : 95 % des organisations sans retour malgré 30 à 40 Md$ d'investissement ; entonnoir outils sur mesure 60 % → 20 % → 5 % (vs génériques 80 → 50 → 40) ; ~67 % (partenariat externe) vs ~33 % (build interne) de passage en prod ; ~50 % des budgets GenAI au front-office (jusqu'à ~70 % en allocation déclarée) ; ROI back-office documenté à 2–10 M$/an. https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf

structurées — des économies documentées entre **2 et 10 millions de dollars par an** pour les cas retenus.

Le front-office a un avantage évident sur la direction : on voit ce que ça fait. Un chatbot client, une génération de propositions commerciales, une aide à la rédaction de contenus — ce sont des surfaces visibles, démontables en dix minutes. Le back-office est opaque. Un agent qui traite des rapprochements comptables, qui classe des contrats entrants, qui pré-qualifie des sinistres ne se montre pas en présentation de board. Et pourtant, c'est là que les masses salariales qualifiées sont concentrées, là que le volume de tâches répétitives est le plus élevé, là que le retour sur investissement est le plus défendable dans un tableur.

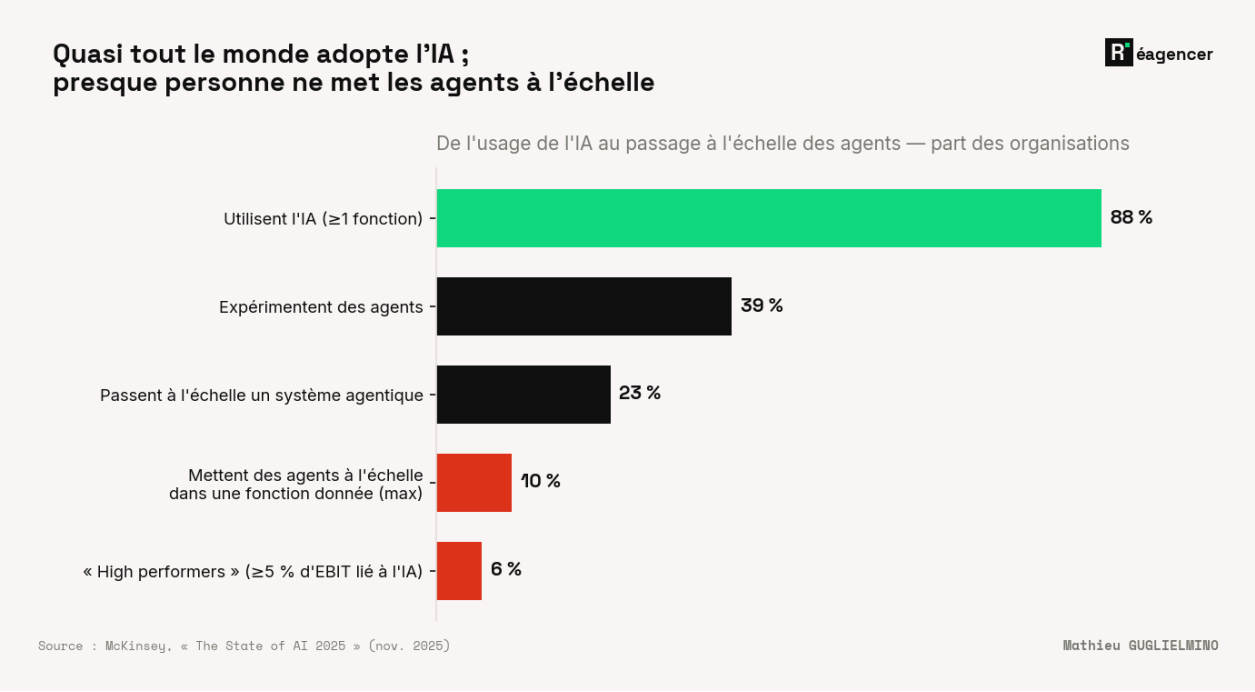
C'est un problème d'**allocation**, pas de prix de l'outil. Si les modèles coûtaient dix fois plus cher, le même problème existerait : trop d'argent sur ce qui se montre, pas assez sur ce qui rapporte. Inverser la logique — viser d'abord les processus back-office à fort volume et ROI documentable — est un choix organisationnel, pas technologique. Et c'est un choix que beaucoup d'organisations n'ont pas encore fait.

---

6. Quatre-vingt-huit pour cent utilisent l'IA — vingt-trois pour cent mettent des agents à l'échelle

L'écart entre ces deux chiffres est le vrai sujet de coût de cette étude.

```
analyse.fig_adoption_echelle(d);
```



McKinsey[11] a interrogé les grandes organisations fin 2025 : **88 %** déclarent utiliser l'IA dans au moins une fonction. **39 %** expérimentent des agents. **23 %** mettent à l'échelle un système

[11]McKinsey & Company (QuantumBlack), The State of AI 2025, novembre 2025. Chiffres cités : 88 % des organisations utilisent l'IA dans ≥1 fonction ; 39 % expérimentent des agents IA ; 23 % mettent à l'échelle un système agentique ; ≤10 % par fonction ; 6 % de « high performers » (≥5 % d'EBIT lié à l'IA) ; 39 % déclarent un impact EBIT au niveau de l'entreprise. (Page publique ; PDF de détail sous accès.) https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

agentique. Dans une fonction donnée, au maximum **10 %** passent à l'échelle. Et seules **6 %** sont des « high performers » — ceux qui tirent au moins 5 % d'EBIT de l'IA. **39 %** déclarent un impact EBIT au niveau de l'entreprise.

Ces chiffres confirment deux choses. D'abord, « utiliser l'IA » est une déclaration quasi universelle et quasi vide : elle inclut l'équipe commerciale qui utilise ChatGPT pour reformuler des emails, et l'opération industrielle qui a déployé un système agentique de contrôle qualité en production 24 heures sur 24. L'écart 88 % ↔ 23 % ↔ 6 % décrit un entonnoir identique à celui du MIT NANDA — sauf qu'il porte sur l'ensemble du spectre IA, pas uniquement sur les outils sur mesure.

Ensuite, le passage à l'échelle est le moment où tous les coûts cachés se matérialisent simultanément. Un pilote peut tourner avec deux développeurs qui connaissent l'application sur le bout des doigts, un jeu de données curé, et une tolérance aux erreurs implicite. La mise à l'échelle exige une infrastructure, des processus de supervision formalisés, des contrats avec les fournisseurs de modèles, une politique de gouvernance des données, une formation des utilisateurs finaux, et une organisation capable de maintenir tout ça sur la durée. C'est là que le budget « modèle » s'avère anecdotique et que le budget « orchestration » devient structurant.

Le mot d'ordre Gartner vaut ici comme mise en garde : l'« agent washing »[12] gonfle artificiellement les chiffres d'adoption déclarée. Des organisations qui ont déployé un assistant conversationnel ou un automate de classification se retrouvent dans la catégorie « utilisent l'IA » — et parfois dans « expérimentent des agents ». Quand Gartner annonce que 40 % des projets agentiques seront annulés d'ici 2027, c'est en partie parce que beaucoup de ces projets ne sont pas agentiques au sens propre, et que la désillusion arrive quand les promesses ne se concrétisent pas.

---

7. Le marché a tranché : il paie l'assistant, pas l'agent

- Après trois ans, le marché a une taille et une forme mesurables. Les entreprises ont dépensé **37 milliards de dollars** en logiciels d'IA générative en 2025, contre 11,5 milliards en 2024 — un bond de **×3,2** en douze mois, selon le modèle de marché de Menlo Ventures[13]. Précision méthodologique qui compte : ce total **exclut l'inférence et le model-serving cloud** — ce n'est pas un coût total de possession, c'est ce que les entreprises paient à des éditeurs de logiciels pour habiller un modèle. La nuance n'est pas cosmétique : elle confirme, à l'échelle du marché entier, ce que la §1 disait du prix du modèle — il devient un poste résiduel, presque invisible dans la ligne budgétaire qui compte vraiment. Mais ce que cette dépense achète, et où elle va, raconte la thèse de ce dossier mieux qu'aucun sondage de confiance déclarative.
- Dans la brique de marché que Menlo appelle l'IA « horizontale » (8,4 Md$, les outils transverses par opposition aux applications métier verticales), la répartition est sans appel.

---

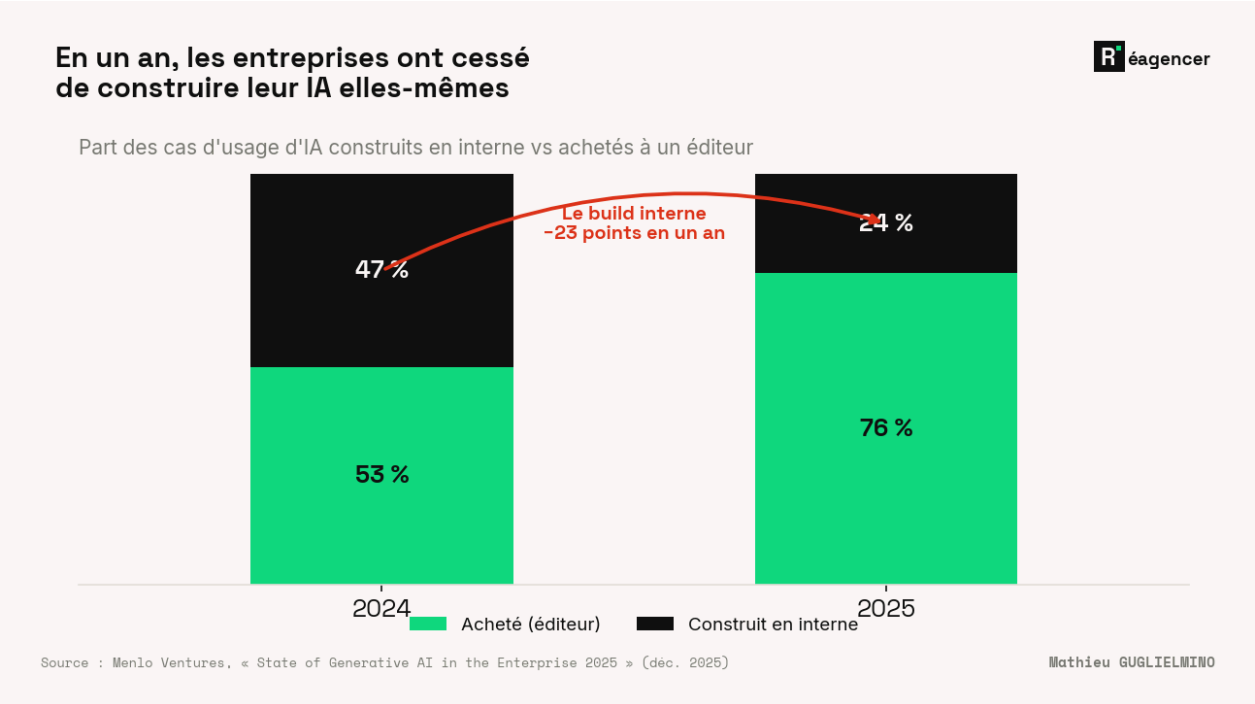
[12] Gartner, communiqué « Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027 », 25 juin 2025. Prédiction : >40 % des projets d'IA agentique annulés d'ici fin 2027 (coûts dérapants, valeur métier floue, contrôle du risque insuffisant). Pointe l'« agent washing » : ~130 éditeurs proposant de vraies capacités agentiques sur des milliers de prétendants. (Communiqué web ; PDF non téléchargeable.) https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027

[13] Menlo Ventures, 2025: The State of Generative AI in the Enterprise, décembre 2025 (3e édition annuelle ; ~500 décideurs d'entreprises américaines + modèle bottom-up du marché). Chiffres cités : 37 Md$ de dépense GenAI entreprise en 2025 (vs 11,5 Md$ en 2024) — hors inférence/model-serving cloud et puces ; app 19 Md$ / infrastructure 18 Md$; IA « horizontale » 8,4 Md$ répartie en copilotes 86 % (7,2 Md$), plateformes agentiques 10), productivité personnelle 5 % (0,45 Md$) ; cas d'usage construits en interne 47 % (2024) → 24 % (2025), achetés 53 % → 76 %. https://menlovc.com/perspective/2025-the-state-of-generative-ai-in-the-enterprise/

```
analyse.fig_copilotes_vs_agents(d);
```

- Les copilotes — l'assistant qu'on branche sur une messagerie, un IDE ou une suite bureautique, sans intégration profonde ni supervision formalisée — captent **7,2 Md**, $soit 86$**, soit 10 %** ; le solde (0,45 Md$, 5 %) va aux outils de productivité personnelle. Le so-what praticien est direct : le marché paie l'**assistant borné**, pas l'**agent** qui doit survivre au passage POC→prod décrit en §2. C'est la confirmation mesurée, côté demande, de l'« agent washing » que Gartner pointait côté offre en §3 — on appelle « agent » beaucoup de choses, mais l'argent qui va aux produits réellement agentiques reste un filet d'eau à côté du fleuve copilote.
- Le deuxième schéma montre un mouvement, pas seulement un état.

```
analyse.fig_build_vs_buy_flip(d);
```



- En un an, la part des cas d'usage d'IA construits en interne est tombée de **47 % à 24 %** ; symétriquement, la part achetée est passée de 53 % à **76 %**. C'est la confirmation temporelle de la §4 : si acheter atteint la production deux fois plus souvent que construire (~67 % contre

~33 % selon MIT NANDA), il n'est pas surprenant que les organisations, une fois échaudées par un ou deux cycles de POC avortés, réallouent leur budget vers l'achat. Le coût caché du build interne — l'intégration, l'apprentissage du contexte métier, la maintenance à chaque changement de version — est précisément ce qui le tuait déjà en §1 et §2 ; le marché ne fait qu'en tirer la conséquence budgétaire un an plus tard. On n'externalise pas le modèle : il est déjà bon marché, et il l'était déjà en 2024. On externalise l'orchestration.

- Ce que le marché dépense confirme donc, point par point, ce que les taux d'échec disaient déjà : la valeur se capte là où l'intégration est déjà faite — l'assistant prêt à l'emploi, l'outil acheté et adapté — pas là où il faut la construire soi-même, à savoir l'agent sur mesure. Le corollaire Réagencer tient, et cette fois il tient aussi côté portefeuille : **budgéter l'orchestration, pas le modèle** — y compris quand on choisit ce qu'on achète.

Ce que la donnée ne dit pas encore

Angle mort majeur — pas de mesure française. Toutes les enquêtes citées dans ce dossier sont mondiales ou anglo-saxonnes : MIT (États-Unis et international), Stanford HAI (données Epoch AI/Artificial Analysis, marché mondial), S&P Global (Amérique du Nord + Europe sans ventilation), McKinsey (grands groupes internationaux), Gartner (marché mondial). Il n'existe pas, à ce jour, de mesure française publiée du taux de passage POC→production pour les projets d'IA en entreprise, ni des postes de coût réels d'un agent en prod (quelle part de la facture totale va à l'API, à l'intégration, à la supervision humaine, à la maintenance, à la conformité), ni de la durée de vie médiane d'un agent ni des causes d'abandon. Le **coût de supervision humaine** en particulier — le poste qu'on suspecte décisif — n'est chiffré nulle part en données ouvertes. C'est l'inconnue n°1 de ce dossier. **Le module 13 du baromètre décideurs-IA** est conçu pour combler exactement cet angle mort : 13.1 — part des POC passés en production (ancre ouverte : 5 % MIT NANDA, 46 % de POC abandonnés S&P) ; 13.2 — poste de coût dominant déclaré par les décideurs français (le delta entre ce qu'ils croient — « l'API » — et ce qui pèse vraiment est l'angle de publication) ; 13.3 — durée de vie et cause d'abandon principale. Le chiffre fondateur visé : le poste de coût n°1 d'un agent en prod déclaré par les décideurs français, à confronter à l'effondrement du prix du modèle.

Ce qu'il faut en retenir

Quatre mouvements résument ce que la donnée disponible autorise à conclure — et ce qu'elle oblige à problématiser.

Budgéter l'orchestration, pas le modèle. Le prix du modèle s'effondre et continuera de s'effondrer. Continuer à piloter un projet d'agent sur la ligne « API/inférence » du budget est une erreur de cadre. Les coûts qui décident du sort d'un projet sont l'intégration, la supervision humaine, l'apprentissage du contexte et la maintenance — tous des postes de main-d'œuvre qualifiée, tous invisibles dans un POC, tous visibles en production. C'est le budget d'orchestration qui détermine si un agent vit ou meurt.

Acheter plutôt que rebâtir ce qui existe déjà. Pour une organisation qui n'a pas la taille critique pour accumuler le capital d'intégration en interne — c'est-à-dire la grande majorité des entreprises françaises —, la comparaison MIT NANDA est sans appel : 67 % de passage en prod contre 33 %. Le build interne a du sens quand le cas d'usage est tellement propriétaire qu'aucun éditeur ne peut l'avoir anticipé. Pour le reste, acheter un outil adapté et bien intégré est une décision de gestion, pas une capitulation technique.

Viser le back-office. Le ROI documentable est là, pas sur la génération de contenu ni sur les expériences client visibles. C'est un problème d'allocation plus que de prix, et il se résout par une décision organisationnelle — prioriser les processus à fort volume, à forte répétitivité et à coût de supervision humaine existant élevé. Ces processus sont moins « pitchables » mais bien plus rentables.

Mesurer la supervision. Tant qu'une organisation ne sait pas combien d'heures-homme par semaine sont consacrées à vérifier, corriger et escalader les sorties de ses agents, elle ne connaît pas le vrai coût de ces agents. La supervision humaine est le poste fantôme : il n'apparaît ni dans les contrats fournisseurs ni dans les budgets projets, mais il consomme du temps qualifié en permanence. Le mesurer est la première étape pour le réduire — ou pour décider que le niveau de supervision requis rend le cas d'usage non viable à l'échelle.

Ce dossier est le premier volet d'une trilogie. **L'étude 16** descend dans le coût d'infrastructure d'un agent — GPU, orchestration, stockage — pour compléter le tableau côté technique. **L'étude 06** pose la même question du côté des PME : quel budget IA est réellement envisageable pour une structure de moins de 250 salariés, et comment arbitrer entre les postes ? Ensemble, les trois études répondent à la question que beaucoup de dirigeants posent sans oser la formuler : combien coûte vraiment l'IA, et est-ce que ça vaut le coup ? La réponse honnête est que ça dépend d'un choix d'architecture humaine autant que technique — et que personne ne vous vendra ce choix avec votre abonnement API.

Sources et références

Journal de recherche en cours — Réagencer, juin 2026. Données ouvertes au 2026-06-19.