AGILE ENTERPRISE PARTNER

FinOps IA & souverainet é de l'inférence

Du budget projet à la maîtrise du coût d'inférence

Enjeu 3 — Édition Enjeux CIO 2026-2027

Un rapport Agile Enterprise Partner

Par Sébastien Delayre

Mai 2026

1. Préambule : la grande bifurcation

En 2025, vous avez adopté l'IA. En 2026-2027, votre modèle opérationnel va décider de votre trajectoire. Selon le Gartner 2026 CIO Survey, 64 % des CIO prévoient de déployer des agents autonomes dans les 24 prochains mois. Pourtant, l'étude MIT Project NANDA *(juillet 2025)* montre que 95 % des pilotes GenAI n'ont produit aucun impact mesurable sur le P&L en 2025. La fracture se creuse.

Deux cohortes émergent. Une minorité, entre 15 et 25 % des DSI selon les études convergentes, refonde son modèle opérationnel pour accueillir l'agentique comme colonne vertébrale et gagne un facteur 2 à 10 en productivité. Une majorité superpose les agents sur un modèle pensé pour les applications et plafonne à 10-20 % de gains locaux non capitalisés.

Cette édition 2026-2027 trace les cinq lignes de fracture de cette bifurcation. Le PDF 1 traitait la refonte du Product Operating Model. Le PDF 2 traitait la gouvernance opérationnelle des agents. Ce troisième PDF traite la troisième ligne, la plus structurante en termes économiques : la maîtrise du coût de l'inférence et la mise en place d'un FinOps IA mature. Les deux autres lignes sont traitées dans des PDF dédiés *(Agentic Value Streams, Leadership Transformation IA)*. Pour la vision d'ensemble, voir l'Executive Summary complet de l'édition.

2. Pourquoi votre FinOps actuel ne tient pas

Le coût d'inférence est en train de devenir le poste le plus structurant et le moins maîtrisé du budget IT. Selon l'analyse AnalyticsWeek 2026, l'inférence représente désormais **85 % du budget IA en production**. Selon Gartner *(repris par AnalyticsWeek)*, jusqu'à **35 % du budget IT annuel** peut être consommé par des dérives d'IA générative non maîtrisées. Et selon le FinOps Foundation 2026 State of FinOps Report, **AI et data platforms sont la catégorie de dépenses enterprise qui croît le plus vite**. Votre FinOps actuel, hérité du FinOps cloud 2018-2024, n'a pas été conçu pour ça.

2.1. Le FinOps cloud 2018-2024 : ce qu'il était conçu pour

Entre 2018 et 2024, le FinOps cloud s'est structuré autour de quatre piliers. Une visibilité par dashboard *(coût par projet, par environnement, par compte)*. Une optimisation par allocation *(reserved instances, spot instances, autoscaling)*. Une gouvernance par budget *(alertes, freezes, chargebacks)*. Une mesure par TCO *(coût total de possession sur 3-5 ans, modèle prévisible)*. Selon CloudZero, **70 % des grandes entreprises maintiennent désormais une équipe FinOps dédiée**. Le modèle est mature.

Ce FinOps a délivré. Il a réduit le gaspillage cloud de manière mesurable, structuré la gouvernance, professionnalisé l'arbitrage entre vélocité et coût. Mais il a été conçu pour orchestrer des ressources cloud relativement prévisibles. Le compute, le stockage, le réseau ont des coûts proportionnels à la consommation, des courbes connues, des unités stables. L'inférence IA cassent ces hypothèses.

2.2. Le choc agentique : les quatre ruptures économiques

Quatre ruptures rendent le FinOps cloud structurellement inadapté aux agents IA. Ces ruptures ne sont pas hypothétiques. Elles sont mesurables dans les budgets 2026.

Rupture 1 — Le multiplicateur agentique. Selon l'analyse Gartner de mars 2026 reprise par Oplexa, les workflows agentiques consomment **5 à 30 fois plus de tokens** par tâche que les chatbots standards. Une entreprise qui a piloté avec un chatbot single-turn et déploie en production un workflow ReAct multi-étapes voit son coût exploser d'un ordre de grandeur, alors que le ROI calculé reposait sur les chiffres du pilote. Le FinOps Foundation 2026 identifie cette agentic loop multiplier comme la cause primaire des dépassements budgétaires en post-pilote.

Rupture 2 — La volatilité par seconde. Le cloud compute est facturé à la minute. L'inférence IA peut voir ses coûts varier d'un ordre de magnitude en quelques secondes pendant les boucles de raisonnement agentique. Selon AnalyticsWeek 2026, *« un changement mineur dans la structure d'un prompt ou dans l'usage applicatif peut doubler les coûts d'inférence du jour au lendemain »*. Les modèles budgétaires traditionnels, construits autour de la prévisibilité du compute et du stockage, sont **structurellement inopérants** face à cette volatilité.

Rupture 3 — La cécité économique. Les outils FinOps cloud savent dire combien a été dépensé. Ils ne savent pas dire pourquoi. Le coût d'inférence varie selon l'architecture du modèle, la taille du contexte, le pattern d'usage. Sans instrumentation spécifique, le FinOps IA est aveugle. Selon CloudZero State of AI Costs, **63 % des organisations identifient l'IA comme une préoccupation FinOps active en 2025**, contre 31 % en 2024. La double augmentation traduit la prise de conscience, pas la maîtrise.

Rupture 4 — La concentration vendor structurante. Sous pression de coût, les organisations sont incitées à concentrer leurs workloads sur un nombre réduit de providers. Selon Stabilarity Hub *(mars 2026)*, ce phénomène crée *« un risque de vendor lock-in plus sévère que le lock-in cloud, parce que les coûts de migration entre modèles fondations incluent la réécriture des prompts, l'adaptation des pipelines RAG, et la revalidation des performances »*. Le FinOps cloud 2018-2024 connaissait le lock-in cloud. Il ne connaît pas le lock-in modèle.

À ces quatre ruptures, nos audits terrain ajoutent un cinquième constat. Sur un échantillon de DSI grands comptes auditées en 2025, le nombre médian de **zombie agents** *(agents déployés en pilote, jamais correctement industrialisés, restant en production sans propriétaire clair)* est de 23 par DSI, avec un coût mensuel agrégé moyen de **47 000 euros** par DSI. Ce sont des dépenses récurrentes sans valeur business identifiable. Elles n'apparaissent ni dans les KPI projets, ni dans les revues budgétaires classiques.

2.3. Le constat 2026 : 85 % budget en inférence, 35 % de dérive non maîtrisée

Les chiffres 2025-2026 confirment l'ampleur du problème. Trois indicateurs convergent vers un même diagnostic.

Indicateur 1 — La part de l'inférence dans le budget IA. Selon AnalyticsWeek 2026, l'inférence représente désormais **85 % du budget IA en production**. Selon Spheron *(avril 2026)*, les analystes industriels estiment **55 à 80 % du spend GPU enterprise** en inférence. La tendance est claire : training était le centre de coût 2021-2023, l'inférence est le centre de coût 2026.

Indicateur 2 — La croissance des budgets IA. Selon TheStreet *(mai 2026)* citant CloudZero, **le spend IA moyen mensuel par entreprise est passé de 62 964 dollars en 2024 à 85 521 dollars en 2025**. Selon Gartner, le spend Data center systems va bondir de **55,8 % en 2026**. Selon NVIDIA, **42 % des entreprises citent l'optimisation des workflows IA comme leur priorité de spending pour 2026**, devant l'expansion. L'optimisation a dépassé l'expansion comme priorité énoncée.

Indicateur 3 — Le décalage entre adoption et ROI mesurable. Selon le KPMG Q4 AI Pulse Survey *(janvier 2026)*, **59 % des leaders attendent un ROI mesurable dans les 12 mois**, mais selon le rapport KPMG Global Tech 2026, seulement **24 % scalent l'IA avec succès sur plusieurs cas d'usage**. Le KPMG Pulse identifie **65 % des leaders citant la complexité des systèmes agentiques comme premier frein**, contre 33 % au trimestre précédent.

À ces chiffres macro s'ajoute une donnée structurante peu commentée. Selon Spheron *(avril 2026)*, le coût par million de tokens du frontier model GPT-4 a été divisé par environ **30** entre 2023 et 2026 *(passant d'environ 30 dollars à environ 1 dollar pour la classe de tokens d'output frontière)*. Mais la consommation de tokens a augmenté beaucoup plus vite. **La baisse du prix unitaire ne compense pas l'explosion des volumes**. C'est l'inférence cost paradox : tokens moins chers, factures qui explosent.

Le diagnostic est clair. Votre FinOps actuel, conçu pour le cloud prévisible 2018-2024, ne tient plus face aux dynamiques agentiques. Il ne mesure pas, il ne prédit pas, il n'attribue pas. Refondre ce FinOps n'est plus une option, c'est la condition pour que votre programme agentique tienne sur la durée.

3. État de l'art : modèles économiques et offres marché

3.1. L'effondrement du coût d'inférence et ses conséquences

Le coût unitaire de l'inférence s'effondre depuis 36 mois. Selon Jon Radoff *(« The State of AI Agents in 2026 », février 2026)*, le **prix par million de tokens des modèles fondations a chuté d'environ 92 % en 3 ans**. Cette baisse est attribuable à plusieurs facteurs convergents : optimisations matérielles *(Blackwell, Groq LPU)*, optimisations runtime *(vLLM, TensorRT-LLM, SGLang)*, optimisations modèles *(quantization FP8, distillation, mixture-of-experts)*, concurrence vendor *(OpenAI, Anthropic, Google, Mistral, Meta)*, et apparition des modèles open-source compétitifs *(Llama 3.3 70B, Qwen3, Mistral Large)*.

Cette baisse change l'arithmétique du déploiement. Selon Iterathon *(décembre 2025)*, servir un modèle 7B est **10 à 30 fois moins cher** que servir un modèle 70-175B. Le break-even d'un déploiement on-premise sur un modèle small ou medium peut atteindre **0,3 à 3 mois** selon Arxiv 2509.18101 *(septembre 2025)* pour des SME avec moins de 10 millions de tokens par mois.

Mais cette baisse ne se traduit pas mécaniquement en réduction des factures. Trois mécanismes l'expliquent.

Mécanisme 1 — Le effet Jevons sur les tokens. Quand le prix baisse, l'usage explose. Les patterns RAG sophistiqués multiplient le contexte par 3 à 5. Les workflows agentiques multiplient les appels par 10 à 20. Les agents always-on consomment 24/7. Le total tokens explose plus vite que le prix unitaire ne baisse.

Mécanisme 2 — La subvention vendor. Selon AnalyticsWeek *(citant un Turing Award-winning Google researcher, 2026)*, le pricing API actuel reste **subventionné par le venture capital et les cross-subsidies hyperscalers**. La normalisation est anticipée dans les 12-24 mois. Les entreprises qui ont budgété au prix actuel ont une exposition financière de 30 à 50 % à la hausse.

Mécanisme 3 — Le compounding agentique. Comme analysé dans le PDF 1, l'erreur d'agent compose exponentiellement. Un agent fiable à 95 % par étape, enchaîné sur 20 étapes, atteint 36 % de succès end-to-end. Cela génère des retries, des re-runs, des escalades qui multiplient le coût réel par rapport au coût modélisé en pilote.

3.2. Les modèles de pricing 2026

Quatre modèles de pricing dominent le marché 2026. Ils ne se substituent pas mais se cumulent dans les architectures hybrides.

Modèle 1 — Per-token. Le pricing classique des APIs *(OpenAI, Anthropic, Google, Mistral)*. Facturation par million de tokens d'input et d'output, différenciée. Avantage : prévisibilité unitaire. Inconvénient : volatilité totale parce que la consommation dépend de patterns d'usage difficiles à modéliser.

Modèle 2 — Per-task ou outcome-based. Pricing à l'unité de valeur business *(par ticket résolu, par document traité, par cas instruit)*. Pratiqué par Zendesk *(AI agent pricing)*, Salesforce Agentforce, certains agents verticaux retail. Avantage : alignement direct avec la valeur. Inconvénient : suppose une mesure qualité robuste, sinon biais incitatifs vendeur.

Modèle 3 — Self-hosted. Pricing par GPU-hour, par instance, par déploiement. Pratiqué par les déploiements on-prem ou cloud privé *(Mistral Compute, OVHcloud, Deutsche Telekom Industrial AI Cloud)*. Avantage : coût marginal proche de zéro à pleine utilisation, contrôle des données. Inconvénient : capex initial, expertise opérationnelle, time-to-market.

Modèle 4 — Hybride et orchestré. Pricing complexe combinant les trois précédents. Une requête simple va vers un small model self-hosted *(coût quasi-nul)*, une requête complexe vers un frontier API *(per-token)*, un workflow critique vers un service outcome-based avec SLA. Selon Deloitte *(« 2026 AI Token Economics Analysis »)*, *« la consommation hybride combinant SaaS, cloud APIs et infrastructure self-hosted dominera les architectures enterprise IA, chaque tier servant des profils de coût et de performance distincts »*.

3.3. Les architectures de souveraineté de l'inférence

L'enjeu de souveraineté de l'inférence prend du poids en 2026. La question n'est plus seulement *combien coûte mon inférence »* mais *« où tourne-t-elle, sous quelle juridiction, avec quels engagements de sécurité »*. Quatre options s'offrent au CIO grand compte.

Option 1 — Cloud public US *(AWS Bedrock, Azure AI, Google Vertex AI)*. Maturité maximale, écosystème riche, capacité massive. Mais exposition aux juridictions extra-territoriales *(CLOUD*

Act, FISA), dépendance à l'écosystème CUDA-NVIDIA pour le frontier compute, risque de lock-in vendor. Reste pertinent pour la majorité des workloads non sensibles, sous condition de contrats clairs sur la résidence des données.

Option 2 — Cloud souverain européen *(OVHcloud, Scaleway, IONOS, Mistral Compute, Deutsche Telekom Industrial AI Cloud)*. Hébergement en Europe, juridiction européenne, certifications GAIA-X, BSI C5, ISO 27001. Selon NVIDIA *(annonces du Hannover Messe avril 2026)*, l'écosystème européen totalise **plus de 3 000 exaflops de Blackwell compute** déployés ou prévus, avec Mistral à 18 000 systèmes Grace-Blackwell, Deutsche Telekom à 10 000 GPUs Blackwell, et des partenariats avec Orange, Fastweb, Swisscom, Telefónica, Telenor. La capacité existe désormais. Pour le CIO français, OVHcloud avec déploiement Mistral Large *(123B paramètres)* est devenu une option sérieuse. Pour le CIO européen multi-pays, Mistral Compute et le Deutsche Telekom Industrial AI Cloud sont les options de référence à étudier en 2026.

Option 3 — On-premise. Déploiement sur infrastructure interne, GPUs propres, contrôle total. Selon Arxiv 2509.18101, le break-even on-prem versus API commerciale peut être atteint en 0,3 à 3 mois sur des modèles 30B avec des consommations supérieures à 10 millions de tokens par mois. Pour des grandes entreprises avec des volumes massifs et des données sensibles, l'option redevient économiquement et stratégiquement pertinente après 5 ans de domination cloud.

Option 4 — Edge inference. Déploiement de modèles small *(1-7B)* directement sur l'infrastructure proche de l'usage *(applicatifs mobiles, devices IoT, edge servers)*. Pertinent pour les usages temps réel *(latence <100 ms)*, les usages offline *(disponibilité hors connexion)*, et les usages massifs à coût marginal *(inférence locale gratuite à l'usage)*. Selon Iterathon, les modèles 1-3B running on edge *(Phi-4, Gemma 2B/7B, Llama 3.2 1B/3B, Mistral 7B)* permettent des réductions de coût allant jusqu'à **75 %** par rapport au déploiement cloud des frontier models.

Notre recommandation pour le CIO grand compte : penser **Hybrid Consumption Architecture** dès 2026. Pas une option exclusive, mais un portefeuille structuré combinant les quatre. Cette approche est traitée en profondeur en section 3.3.

Option	Coût marginal	Souveraineté	Time-to-Market	Cas d'usage type
Cloud public US	API per-token	Faible	Très rapide	Workloads génériques non sensibles
Cloud souverain EU	API per-token + GPU-hour	Élevée	Rapide	Données régulées, secteurs sensibles
On-premise	GPU-hour amortie	Maximale	Lent	Volumes massifs, données stratégiques
Edge inference	Quasi-nul	Maximale	Moyen	Latence critique, usages offline

4. Analyse AEP : les trois transformations structurantes

Sur la base de notre observation terrain auprès de DSI grands comptes du secteur Retail et E-commerce *(secteur particulièrement exposé aux dynamiques agentiques avec des volumétries massives)*, complétée par les sources publiques 2025-2026, nous identifions trois transformations structurantes qui décident, individuellement et collectivement, si un FinOps IA tient ou s'effondre face aux agents. Ces trois transformations doivent être traitées simultanément. Une seule sur les trois ne suffit pas.

4.1. Transformation 1 — Inference economics et chasse aux zombie agents

La première erreur structurelle observée chez les DSI qui plafonnent : continuer à raisonner en logique projet sur des coûts qui sont devenus opérationnels. Un projet IA est borné dans le temps, son budget est fini. Un agent IA en production consomme à l'infini, son budget grossit chaque mois. Le passage du *« combien coûte ce projet »* au *« combien coûte chaque inférence »* est la première transformation à opérer.

Le test discriminant est simple. Si vous savez chiffrer le coût d'un projet IA *(en euros sur 12 mois)* mais pas le coût unitaire d'une inférence en production *(en euros par requête, par client, par feature)*, vous êtes encore en logique cloud 2018-2024. Vous n'avez pas basculé en inference economics.

Trois pathologies caractérisent les DSI qui n'ont pas opéré cette transformation. Elles forment ce que nous appelons le **trio des dérives non maîtrisées du FinOps IA**.

Pathologie 1 — Les zombie agents. Définis dans le PDF 2 : agents déployés en pilote, jamais correctement industrialisés, restant en production sans propriétaire clair. Sur notre échantillon de DSI grands comptes, le nombre médian de zombie agents identifiés est de 23 par DSI, avec un coût mensuel agrégé moyen de **47 000 euros** sans valeur business identifiable. Selon Forrester *(décembre 2025)*, sur les déploiements à plus de 50 agents, **18 à 32 % des agents en production ont au moins un autre agent qui couvre 80 % du même périmètre fonctionnel**. La consolidation est possible, mais elle nécessite un audit régulier. Sans un mécanisme structurel de découverte, retrait et consolidation, les zombie agents prolifèrent.

Pathologie 2 — La Big Model Fallacy. Définition : le réflexe de toujours utiliser le modèle frontière le plus puissant, indépendamment du besoin réel. Conséquence : sur-paiement systématique. Selon les benchmarks Iterathon 2026, **un modèle 7B bien fine-tuné surperforme un modèle 70B généraliste sur 60 à 75 % des tâches métier répétitives, à un coût 8 à 15 fois inférieur**. Selon TechCrunch *(janvier 2026)*, Andy Markus *(CDO AT&T)* synthétise la position des entreprises matures : *« les SLM fine-tunés deviendront un standard utilisé par les entreprises IA matures en 2026, parce que les avantages coût/performance vont pousser leur adoption au-delà des LLM out-of-the-box »*. La discipline du right-sizing, c'est-à-dire le choix du modèle adapté à chaque tâche, est un pilier du FinOps IA mature.

Pathologie 3 — La cécité économique unitaire. Sans mesure cost-per-inference, cost-per-task, cost-per-customer, le FinOps reste à un niveau agrégé qui n'oriente pas les décisions opérationnelles. Selon CloudZero *(rapport CloudZero State of AI Costs 2026)*, *« ce que vous voulez vraiment savoir n'est pas combien vous avez dépensé, c'est combien coûte de servir une inférence »*. Le passage à l'unit economics est traité en profondeur dans la transformation 2.

Pour le CIO, l'analyse AEP est nette. Le premier chantier FinOps IA n'est pas une plateforme. C'est un audit. Cartographier les agents en production. Identifier les zombie agents. Mesurer la part des inférences qui pourraient passer sur des modèles plus petits sans dégradation business. Selon notre observation terrain, ces trois actions menées en parallèle dégagent typiquement **15 à 25 % d'économie** sur les budgets IA en production, sans aucune perte de capacité. C'est le point de départ obligé.

Patterns d'implémentation et solutions qui les portent.

Le pattern *« audit et découverte automatique d'agents »* est implémenté par Microsoft Agent 365 *(lancé le 1er mai 2026)*, qui propose un agent registry à découverte automatique. Pour les environnements multi-cloud non-Microsoft, des solutions FinOps spécialisées comme **Finout**, **CloudZero** et **Flexera Spot** offrent des capacités équivalentes via virtual tagging et anomaly detection.

Le pattern *« right-sizing par modèle »* est implémenté par les frameworks d'orchestration de modèles. **BentoML**, **vLLM**, **TensorRT-LLM**, **SGLang** permettent de servir des modèles small et medium en production avec des throughput optimisés. Pour le déploiement souverain, **Mistral Compute**, **OVHcloud AI Deploy** et **Hugging Face TGI** simplifient le déploiement on-prem ou cloud privé.

Le pattern *« suppression progressive des zombie agents »* nécessite peu d'outillage sophistiqué : un audit trimestriel structuré, un comité de revue avec le trio CIO-CISO-CDO, une politique de propriétaire obligatoire pour tout agent en production. C'est plus un sujet de discipline organisationnelle que de techno.

4.2. Transformation 2 — Unit Economics Attribution

La deuxième transformation est la plus structurante en termes de pilotage. Elle touche la mesure et l'attribution. Sans elle, le FinOps IA reste un constat, pas un système pilotable.

Le concept d'**Unit Economics Attribution** introduit dans cette édition consiste à rattacher chaque euro de dépense IA à une unité de valeur business mesurable. Pas le coût agrégé. Le coût unitaire. Cost-per-inference, cost-per-task, cost-per-customer-served, cost-per-feature, cost-per-revenue-dollar. Selon Finout *(rapport 2026)*, *« unit economics signifie calculer le coût pour produire une unité de valeur business : une requête d'inférence, une session utilisateur complétée, un document traité, un client servi »*. C'est le cœur du FinOps IA mature.

Trois métriques signature structurent ce pilotage en 2026. La première est introduite dans le PDF 1 *(« Quality-Adjusted Cost per Task », ou QACT)*. Les deux suivantes sont introduites dans ce PDF 3.

Métrique 1 — Quality-Adjusted Cost per Task (QACT) *(rappel du PDF 1)*. Le coût économique d'une tâche complétée, ajusté du taux d'erreur et du coût des reprises. À calculer par type de tâche, par squad, par mois. C'est la métrique de productivité agentique réelle. La logique : un agent à 0,02 dollar par tâche peut paraître économique, mais s'il réussit à 85 % alors qu'un humain à 50 dollars de l'heure réussit à 99 %, le coût des 14 % d'erreurs *(reprise, escalade, dommages)* doit être intégré dans le coût total.

Métrique 2 — Inference ROI Ratio (IRR) *(introduite dans ce PDF)*. Le ratio entre la valeur business générée par une inférence et son coût d'inférence, mesuré sur 30 jours rolling. Formule simplifiée : IRR = (revenu attribuable à l'inférence) / (coût d'inférence pleinement chargé). Selon le KPMG Global Tech Report 2026, **les organisations high-performing ont un ROI moyen de 4,5 fois sur leurs investissements technologiques, plus de deux fois la moyenne industrie de 2 fois**. L'IRR oblige à instrumenter cette mesure au niveau de l'inférence, pas du programme. Un IRR < 1 sur 30 jours rolling signale une inférence économiquement perdante. Si elle persiste, elle est candidate à la désactivation. Un IRR > 5 signale une inférence à scaler.

Métrique 3 — Agent Unit Margin (AUM) *(introduite dans ce PDF)*. La marge unitaire d'un agent IA, mesurée comme la différence entre la valeur business produite par tâche complétée et le coût pleinement chargé de cette tâche *(inférence + infrastructure + supervision humaine + coût d'erreur)*. Cette métrique force à intégrer le coût caché de la supervision *(le « eval-and-integration cost » qui représente 28-44 % du coût total selon Forrester, traité en PDF 1)*. L'AUM est la métrique qui décide si un agent doit rester en production. Un AUM négatif récurrent est un signal d'arrêt ou de refonte.

Ces trois métriques signature *(QACT, IRR, AUM)* forment ce que nous appelons la **suite FinOps IA AEP**. Elles ne sont ni concurrentes ni redondantes. Elles éclairent trois angles complémentaires : la productivité ajustée *(QACT)*, le retour sur inférence *(IRR)*, la marge agent *(AUM)*. Une DSI qui instrumentent les trois passe d'une gestion budgétaire macro à un pilotage économique opérationnel, fonction par fonction, agent par agent.

Patterns d'implémentation et solutions qui les portent.

Le pattern *« virtual tagging et allocation 100 % »* est implémenté par **Finout**, qui propose des règles d'allocation rétroactives sans changement de code, atteignant 100 % d'allocation des coûts AI infrastructure même en absence de tagging natif. **CloudZero** offre une approche similaire avec cost-per-unit visibility *(cost-per-inference, cost-per-conversation, cost-per-feature)*.

Le pattern *« unit economics dashboard »* est implémenté par **Finout**, **CloudZero**, et **Flexera FinOps**. Pour les environnements internes, le pattern peut être implémenté avec OpenTelemetry, Datadog, Grafana, en instrumentant chaque appel d'agent avec le contexte business *(client_id, feature_id, transaction_id)*.

Le pattern *« anomaly detection en temps réel »* est crucial vu la volatilité du coût d'inférence. **Finout**, **CloudZero**, et **AWS Cost Anomaly Detection** sur Bedrock proposent cette capacité. Sans elle, un changement de prompt structure peut doubler le coût en quelques heures, comme observé par AnalyticsWeek 2026.

Pour le CIO, le passage à l'Unit Economics Attribution n'est pas un choix d'outil. C'est un choix de discipline. Le bon KPI 2026 n'est plus *« combien coûte mon programme IA »* ni *« combien de tokens je consomme »*. C'est : combien me coûte de servir une unité de valeur business, et est-ce que cette unité a une marge positive ?

4.3. Transformation 3 — Hybrid Consumption Architecture et Agent Control Plane

La troisième transformation touche l'architecture infrastructure. Elle pose deux questions structurantes. **Sur quelle infrastructure tourne mon inférence ?** Et **comment je pilote ce portefeuille en production ?**

Le concept d'Hybrid Consumption Architecture. Le mythe d'une infrastructure IA unique a vécu. Aucune entreprise mature en 2026 ne tourne tous ses workloads sur un seul provider. La réalité est un portefeuille structuré combinant les quatre options décrites en section 2.3 : cloud public US, cloud souverain européen, on-premise, edge. Le concept d'**Hybrid Consumption Architecture** *(introduit dans cette édition)* désigne ce portefeuille structuré, géré comme un actif et pas comme un empilement de décisions ad hoc.

Selon Deloitte *(« 2026 AI Token Economics Analysis »)*, *« la consommation hybride dominera les architectures enterprise IA, chaque tier servant des profils de coût et de performance distincts »*. La maîtrise de cette hybridation est l'un des marqueurs forts de la cohorte minoritaire qui réussit la bascule.

Trois principes structurent une Hybrid Consumption Architecture mature.

Principe 1 — Allocation par profil de workload. Pas de règle universelle, mais une matrice par type de tâche. Workloads massifs et répétitifs avec données peu sensibles → edge ou small models on-prem. Workloads complexes avec données sensibles → cloud souverain européen. Workloads frontier rares mais critiques → frontier API cloud public US, avec contrats stricts sur la résidence des données. Workloads expérimentaux → cloud public élastique, avec migration prévue.

Principe 2 — Vendor diversity comme actif stratégique. Selon Digital Chiefs *(avril 2026)*, la diversité des fournisseurs en 2026 est *« la réserve stratégique la plus importante »*. La recommandation : **au moins deux fournisseurs par couche d'architecture**, avec une distribution des workloads gouvernée par des règles claires. Cette diversification a un coût opérationnel *(complexité, formation, contrats)* mais elle protège contre le lock-in et contre les disruptions. Chaque année, conduire un test interne de re-sourcing capability : quels workloads pourraient migrer vers un autre fournisseur souverain dans 12 mois.

Principe 3 — Capacité de re-sourcing périodique. Tester annuellement la portabilité réelle des workloads. Quels modèles sont portables, quels prompts sont liés au modèle, quels pipelines RAG nécessitent une réécriture. C'est la phase la moins populaire de la gouvernance d'architecture, et la plus discriminante à 3-5 ans. Les organisations qui la pratiquent peuvent jouer les fournisseurs entre eux. Celles qui ne la pratiquent pas seront *« en 2029 dans la position que le fournisseur leur assignera »*.

Le concept d'Agent Control Plane. Une Hybrid Consumption Architecture sans pilotage est ingérable. L'**Agent Control Plane** *(concept signature évoqué dans les PDF 1 et 2, traité en profondeur ici)* est l'infrastructure qui rend pilotable le portefeuille d'agents en production. Il combine cinq fonctions structurantes.

Fonction 1 — Discovery et registry. Inventaire automatique de tous les agents en production, indépendamment de qui les a déployés ou sur quelle plateforme. Microsoft Agent 365 implémente cette fonction sur le tenant Microsoft. Pour les environnements multi-cloud, des solutions équivalentes émergent.

Fonction 2 — Observabilité multi-couche. Visibilité temps réel sur la consommation de tokens, les latences, les patterns d'usage, les anomalies. Sans observabilité, le FinOps IA est aveugle. **Datadog AI Observability**, **Grafana Loki**, **OpenTelemetry GenAI semantic conventions** structurent ce niveau.

Fonction 3 — Cost attribution et chargeback. Allocation des coûts d'inférence aux unités business consommatrices. **Finout**, **CloudZero** sont les références en 2026. L'attribution est la condition de la responsabilisation : sans chargeback, pas de discipline FinOps.

Fonction 4 — Routing intelligent. Décision automatique du modèle et de l'infrastructure appropriés selon la requête. Le pattern dominant en 2026 : un small model handle 70-80 % des requêtes, un medium model handle 15-25 %, un frontier model handle 5-10 %. Ce routing peut être implémenté via **OpenRouter**, **LangChain RouterChain**, ou des solutions internes.

Fonction 5 — Lifecycle management. Gestion du cycle de vie de chaque agent : déploiement, monitoring, mise à jour, retrait. Sans lifecycle management formalisé, les zombie agents prolifèrent. Cette fonction est la moins outillée en 2026, et la plus dépendante de la discipline organisationnelle.

Pour le CIO, l'Agent Control Plane n'est pas un produit unique. C'est une fonction architecturale. Elle se construit en assemblant des briques *(observabilité, FinOps, registry, routing)* selon le contexte de l'entreprise. Notre observation terrain : les DSI matures consacrent **5 à 10 % du budget IA total** à la construction et à l'opération de leur Agent Control Plane. C'est l'investissement structurel qui rend le reste pilotable.

5. Recommandations actionnables

5.1. Diagnostiquer votre maturité FinOps IA : 10 questions

Avant de définir une trajectoire, mesurez votre point de départ. Les dix questions suivantes constituent une grille d'auto-évaluation. Une réponse « *non* » vaut un point de fragilité. Plus de cinq points de fragilité signalent un FinOps IA qui ne tient pas la trajectoire 2026-2027.

#	Question	Levier concerné
1	Connaissez-vous votre coût d'inférence mensuel total *(toutes plateformes confondues)* ?	Diagnostic
2	Pouvez-vous attribuer ce coût à des unités business *(client, feature, transaction)* ?	Transformation 2
3	Mesurez-vous le QACT, l'IRR ou l'AUM sur vos agents en production ?	Transformation 2
4	Avez-vous identifié vos zombie agents par un audit récent ?	Transformation 1
5	Avez-vous une politique de right-sizing *(small/medium/frontier)* par cas d'usage ?	Transformation 1
6	Votre architecture combine-t-elle au moins deux options *(cloud, souverain, on-prem, edge)* ?	Transformation 3
7	Avez-vous testé la portabilité *(re-sourcing)*	Transformation 3

	d'au moins un workload majeur dans les 12 derniers mois ?	
8	Avez-vous un Agent Control Plane opérationnel *(discovery, observabilité, attribution)* ?	Transformation 3
9	Avez-vous une équipe FinOps IA dédiée ou un responsable nommément désigné ?	Gouvernance
10	Le CIO, le CFO et le CDO partagent-ils un dashboard FinOps IA commun ?	Trio CIO-CISO-CDO

5.2. Choisir votre trajectoire FinOps IA : trois scénarios

Sur la base de notre observation terrain, trois trajectoires-types se dégagent.

Trajectoire A — Optimisation tactique. Objectif : réduire 15-25 % du budget IA en 6 mois en agissant sur les zombie agents et la Big Model Fallacy, sans transformer la mesure ni l'architecture. Recommandée si votre exposition IA est modérée et si la pression budgétaire est forte. Risque : ne tient pas à 18 mois quand l'exposition agentique augmente. Bénéfice : effort court terme limité, gains rapides, finance le chantier suivant.

Trajectoire B — Transformation structurelle. Objectif : refonder le FinOps IA en 12 mois avec instrumentation des trois métriques signature *(QACT, IRR, AUM)*, mise en place de l'Agent Control Plane, structuration de l'Hybrid Consumption Architecture. Recommandée pour la majorité des CIO grands comptes avec exposition IA significative. Bénéfice : combine maîtrise budgétaire et capacité opérationnelle scalable. Limite : effort soutenu sur 12 mois, nécessite un sponsor CFO.

Trajectoire C — Leadership FinOps IA. Objectif : devenir une référence sectorielle sur le FinOps IA, en capitalisant sur la maîtrise budgétaire pour bâtir un avantage compétitif *(capacité de scaler les agents là où la concurrence freine, négociation forte avec les vendors, attractivité des talents FinOps)*. Recommandée pour les groupes très exposés *(retail à grande échelle, services financiers, télécoms)* et avec ambition stratégique sur l'IA. Inclut les éléments de la trajectoire B plus : centre d'expertise FinOps IA, partenariats vendors structurés, communication externe sur les pratiques.

Notre observation : sur 100 CIO grands comptes, environ 40 % choisissent A *(souvent à raison à court terme, mais rarement suffisant à 18 mois)*, 50 % choisissent B *(le bon choix dans la majorité des cas)*, 10 % choisissent C *(souvent à raison pour les leaders sectoriels avec exposition IA massive)*. La trajectoire A se transforme typiquement en B après 6-12 mois, quand les gains tactiques sont consommés et que la dérive structurelle redevient visible.

5.3. Les cinq leviers d'action prioritaires

Quelle que soit la trajectoire choisie, cinq leviers structurants doivent être activés. Leur séquençage varie, leur contenu reste constant.

Levier 1 — Auditer pour identifier les zombie agents. Premier levier parce qu'il dégage rapidement de la marge budgétaire pour financer les suivants. Mécanisme : audit trimestriel structuré, propriétaire obligatoire par agent, retrait automatique des agents sans propriétaire après 60 jours. Gain typique : 10-15 % du budget IA.

Levier 2 — Instaurer le right-sizing. Réduire la dépendance aux frontier models pour les tâches qui ne le justifient pas. Mécanisme : routing intelligent avec règles par cas d'usage, fine-tuning de small models sur les workloads massifs, dashboard de répartition par taille de modèle. Gain typique : 20-40 % du coût d'inférence.

Levier 3 — Instrumenter les unit economics. Passer du coût agrégé au coût unitaire. Mécanisme : virtual tagging, attribution par client/feature/transaction, dashboard QACT/IRR/AUM. Gain typique : pas direct mais conditionne tous les autres leviers.

Levier 4 — Construire l'Agent Control Plane. Discovery, observabilité, attribution, routing, lifecycle. Mécanisme : assemblage progressif des briques *(Microsoft Agent 365 ou équivalent + FinOps tool + observabilité)*. Gain typique : 15-25 % par optimisation continue après mise en place.

Levier 5 — Diversifier l'architecture. Réduire le lock-in vendor en répartissant les workloads sur au moins deux fournisseurs par couche. Mécanisme : test annuel de re-sourcing, contrat encadrant la portabilité, vendor diversification policy validée en COMEX. Gain typique : pas direct à court terme, mais protection majeure à 3-5 ans contre les hausses de prix subies.

5.4. Les pièges à éviter

Cinq pièges récurrents sabotent les programmes FinOps IA. Documentés sur le terrain et dans les sources publiques.

Piège 1 — Penser FinOps cloud étendu. L'erreur la plus fréquente : confier le FinOps IA à l'équipe FinOps cloud existante sans formation ni outillage spécifique. Conséquence : application de méthodes inadaptées *(reserved instances n'existent pas en token, autoscaling ne sauve pas du multiplicateur agentique)*. Antidote : former ou recruter des compétences FinOps IA spécifiques, instrumenter avec les bons outils.

Piège 2 — La Big Model Fallacy par défaut. Réflexe de toujours choisir le modèle frontière le plus puissant. Conséquence : sur-paiement systématique de 8 à 15 fois sur les tâches qui ne le justifient pas. Antidote : politique formalisée de right-sizing par cas d'usage, mesure régulière de la qualité produite par modèle, fine-tuning de small models pour les workloads massifs.

Piège 3 — La cécité unitaire. Mesurer uniquement le coût agrégé. Conséquence : pas d'optimisation possible, pas de chargeback, pas de responsabilisation. Antidote : instrumenter dès

le pilote *(pas a posteriori)*, virtual tagging systématique, dashboard cost-per-unit accessible aux PO.

Piège 4 — Le mono-fournisseur. Concentrer tous les workloads sur un seul provider. Conséquence : exposition au risque de hausse de prix, lock-in qui se découvre lors de la première négociation difficile. Antidote : vendor diversification policy avec au moins deux fournisseurs par couche, test annuel de re-sourcing.

Piège 5 — L'overengineering FinOps. À l'inverse, certains DSI sur-investissent en outillage FinOps avant d'avoir traité les vraies dérives. Conséquence : dashboard très complet mais zombie agents toujours présents, Big Model Fallacy non traitée. Antidote : commencer par les leviers 1 et 2 *(audit + right-sizing)* qui dégagent du budget pour financer le reste. L'outillage suit la maturité, il ne la précède pas.

6. Cas et retours d'expérience publics

6.1. Cas composite secteur Retail / E-commerce

Cas illustratif construit à partir de sources publiques sur le secteur retail européen et de notre observation terrain auprès de DSI retail, anonymisé conformément aux engagements de confidentialité d'AEP.

Un grand groupe de Retail européen, plusieurs centaines de magasins physiques, présence omnicanale, plus de 30 millions de clients actifs, plusieurs dizaines de millions de visites mensuelles sur le site e-commerce. La DSI gère un portefeuille IA mixte : agents customer service, moteurs de recommandation, agents de personnalisation du parcours, agents de gestion des stocks, agents pricing dynamique. La pression économique est structurelle : marges retail sous pression, concurrence digitale forte, coûts d'acquisition client en hausse. Le FinOps IA n'est pas un luxe, c'est une condition de viabilité.

Situation initiale 2024-2025. Adoption rapide d'agents IA distribués par fonction. **Premières factures cloud IA en hausse de 40 % par trimestre**. Le moteur de recommandation consomme massivement du frontier model par habitude. Plusieurs initiatives en parallèle dans les filiales sans cadre commun. Le CIO découvre lors d'un audit interne que **47 agents sont en production sans propriétaire identifié**, dont **18 zombie agents** *(coût mensuel agrégé estimé à 38 000 euros)*. Aucune mesure de coût unitaire. Aucun dashboard FinOps IA partagé entre CIO et CFO.

Démarche 2025-2026. Programme structuré en quatre chantiers parallèles sur 12 mois. *(1)* Audit complet des agents en production avec retrait des zombie agents et consolidation des doublons *(− 16 agents supprimés, − 23 % du coût d'inférence en 4 mois)*. *(2)* Right-sizing systématique par cas d'usage. Le moteur de recommandation, qui utilisait un frontier model par défaut, est migré sur un small model fine-tuné. Performance maintenue à 96 % de la baseline, coût divisé par 11 sur ce périmètre. *(3)* Instrumentation des unit economics avec virtual tagging *(client,*

catégorie produit, feature) et dashboard cost-per-recommendation servi. *(4)* Mise en place d'un Agent Control Plane minimal *(discovery + observability + chargeback aux fonctions métier)*.

Résultats observables après 12 mois. Réduction nette du coût d'inférence de **47 %** sur le périmètre traité, malgré une augmentation du volume d'inférences de 28 %. Le **cost-per-recommendation servi a baissé de 71 %**, le QACT moyen sur les agents customer service a baissé de 34 %. Le dashboard FinOps IA partagé CIO-CFO permet une revue mensuelle structurée. Coût total du programme inférieur à 1,5 % du budget IT annuel.

Architecture observée. Le groupe a opté pour une Hybrid Consumption Architecture combinant trois tiers. *(1)* Cloud public US pour les workloads expérimentaux et les frontier API utilisés rarement sur les cas critiques. *(2)* Cloud souverain européen *(OVHcloud + Mistral)* pour les workloads avec données client sensibles. *(3)* On-prem fine-tuné small model pour le moteur de recommandation et les agents customer service à très haut volume. La répartition observée en fin de programme : **75 % des inférences sur small models on-prem, 20 % sur cloud souverain, 5 % sur cloud public US frontier**. Le coût unitaire moyen pondéré est divisé par environ 6 par rapport à la situation initiale.

Enseignements pour un CIO grand compte. Trois leçons.

Premièrement, la séquence des leviers compte. Commencer par l'audit *(zombie agents)* et le right-sizing dégage rapidement de la marge pour financer la suite. Vouloir tout faire en parallèle sans cette base disperse l'effort.

Deuxièmement, le right-sizing n'est pas une dégradation. Sur un cas d'usage retail typique *(recommandation produit)*, un small model bien fine-tuné peut atteindre 96 % de la performance d'un frontier model à 10 % du coût. C'est parce que la tâche est étroite et que les données d'entreprise apportent un signal fort.

Troisièmement, le partage du dashboard FinOps IA avec le CFO change la nature du dialogue. Le CFO ne discute plus le budget IA en valeur agrégée. Il discute le ROI inférence par fonction. Le débat devient stratégique au lieu de défensif.

6.2. Cas public Klarna et Sephora — agents customer service

Cas documentés à partir des publications publiques et des analyses sectorielles 2025-2026.

Klarna. Documenté dans le case study CX Dive *(novembre 2025)* repris par AI Monk *(avril 2026)*. L'agent IA Klarna gère les requêtes customer service routine sur **23 marchés en plus de 35 langues**. Le temps de résolution est passé de **11 minutes à moins de 2 minutes**. Les requêtes répétées ont diminué de **25 %**. Selon la communication officielle Klarna, l'agent gère désormais une charge équivalente à **853 emplois temps plein**. Économie revendiquée : **60 millions de dollars** sur l'année 2025.

Le point critique est la **réintroduction d'agents humains pour les requêtes émotionnellement complexes après 12 mois d'opération**. Klarna n'a pas abandonné l'IA. La firme a affiné le scope. Le modèle hybride avec routing intelligent surperforme le tout-IA sur le

volume total et la satisfaction client. Cette leçon de scoping est aussi structurante que le chiffre de 60 millions.

Sephora. Documenté dans le rapport KPMG Intelligent Retail 2025 *(rapport sectoriel public)* et plusieurs articles spécialisés *(SAAS Cut, Renascence)*. Sephora a déployé une stratégie d'agents IA combinant Beauty Insider Bot *(routing produits, prise de RDV en magasin, rappels de réassort)*, moteur de recommandation personnalisée, et agents de génération de contenu produit. Selon KPMG, **55 % des retailers reportent un ROI IA supérieur à 10 %**, et **21 % un ROI supérieur à 30 %**. Les gains se concentrent sur productivité *(33 %)*, efficience opérationnelle et supply chain *(67 %)*, et innovation produit/service *(47 %)*.

Enseignements pour un CIO grand compte. Deux leçons.

Premièrement, le routing humain-IA est plus performant que le tout-IA. Klarna l'a découvert empiriquement après 12 mois. La règle pratique : **affecter les agents IA aux tâches à fort volume et faible enjeu émotionnel, garder les humains sur les cas complexes ou émotionnels**. L'AUM est négatif sur les cas mal routés.

Deuxièmement, dans le retail, les ROI mesurables sont concentrés sur les fonctions à fort volume *(customer service, recommandation, contenu)*. Selon KPMG, ces fonctions surperforment systématiquement les fonctions support. Le FinOps IA doit suivre cette concentration : instrumenter prioritairement là où le volume génère du signal.

6.3. Cas public JPMorgan et Mistral / OVHcloud — souveraineté à l'échelle

JPMorgan. Documenté par AI Monk *(avril 2026)* et plusieurs sources sectorielles. JPMorgan exploite **plus de 450 cas d'usage agentic AI en production**, sur un budget technologie annuel de **18 milliards de dollars**. Le système COiN *(lancé en 2017, toujours en production)* parse 12 000 contrats commerciaux par an, extrait 150 attributs critiques par document, **récupère 360 000 heures-avocat annuellement**, avec une réduction des erreurs de **80 %** post-déploiement. Sur les agents agentic récents, génération de mémos M&A en 30 secondes vs heures pour les analystes juniors, automatisation du settlement de trade, détection fraude temps réel.

Mistral et OVHcloud — déploiement souverain. Documenté par OVHcloud Blog et NVIDIA Newsroom *(2025-2026)*. La référence d'architecture *« Mistral Large 123B sur OVHcloud »* permet le déploiement d'un LLM frontière en environnement souverain européen, full GDPR-compliant. Mistral Compute *(annoncé à VivaTech 2025, première phase 2026)* prévoit **18 000 GPU Grace-Blackwell** sur le site de Bruyères-le-Châtel *(Essonne)* avec extension à **200 MW de capacité d'ici fin 2027**. Premiers clients publiés : BNP Paribas, Orange, SNCF, Thales, Veolia, Mirakl, Schneider Electric, SLB Group, Black Forest Labs.

Enseignements pour un CIO grand compte. Trois leçons.

Premièrement, la maturité d'opération à grande échelle est possible. JPMorgan démontre qu'un portefeuille de 450 agents en production peut être gouverné, mesuré et faire ROI sur la durée. La condition est la discipline FinOps et architecturale, pas le génie technique.

Deuxièmement, l'option souveraine européenne devient sérieuse en 2026. Pour une DSI grand compte française ou européenne avec contraintes GDPR fortes, le combo OVHcloud + Mistral *(ou Deutsche Telekom Industrial AI Cloud + Mistral, ou autre combinaison)* devient une alternative crédible aux hyperscalers US, pas un placebo politique.

Troisièmement, la diversification vendor n'est pas un luxe. Les grandes banques européennes citées comme premiers clients de Mistral Compute *(BNP Paribas, Orange, SNCF, Thales, Veolia, Schneider, SLB)* ne renoncent pas à leurs contrats AWS ou Microsoft. Elles construisent un portefeuille qui leur donne du pouvoir de négociation.

6.4. Synthèse : cinq facteurs clés de succès observés

Les trois cas convergent sur cinq facteurs clés de succès du FinOps IA mature.

Facteur 1 — Audit d'abord, plateforme ensuite. Aucun programme ne réussit sans avoir établi un inventaire exhaustif initial *(zombie agents, doublons, frontier models par défaut)*. L'audit dégage rapidement 15-25 % de marge budgétaire pour financer la transformation.

Facteur 2 — Right-sizing systématique. Le réflexe frontier model par défaut coûte 8 à 15 fois trop cher sur les tâches qui ne le justifient pas. Le right-sizing par cas d'usage est le levier d'optimisation le plus puissant disponible.

Facteur 3 — Unit economics dès le pilote. L'instrumentation cost-per-unit doit être mise en place dès le pilote, pas a posteriori. Sans elle, la transition vers la production amplifie l'opacité au lieu de la réduire.

Facteur 4 — Hybrid Consumption Architecture explicite. Aucune entreprise mature ne tourne tout sur un provider. La diversification est gérée comme un actif stratégique, pas comme un empilement.

Facteur 5 — Co-pilotage CIO-CFO. Le FinOps IA n'est pas un sujet IT. C'est un sujet économique structurant. Les programmes qui réussissent ont un dashboard partagé CIO-CFO et un ritme de revue mensuel.

7. Plan d'action 2026 : 6 mois pour reprendre le contrôle

Sur la base des trois transformations structurantes et des recommandations, voici un plan d'action calé sur 6 mois pour la phase de mise en place initiale, étendu à 12 mois pour la maturité. Chaque jalon est associé à des actions concrètes et à un point de contrôle structuré qui valide l'avancement.

Jalon	Période	Objectif	Livrables
Jalon 1	Mois 1-2	Diagnostic et quick wins	Audit zombie agents, plan de retrait, premier right-sizing
Jalon 2	Mois 2-4	Instrumentation unit economics	Virtual tagging, dashboard QACT/IRR/AUM, partage CIO-CFO
Jalon 3	Mois 4-6	Agent Control Plane initial	Discovery + observability + chargeback opérationnels
Jalon 4	Mois 6-9	Hybrid Consumption Architecture	Diversification vendor, premier test re-sourcing
Jalon 5	Mois 9-12	Maturité opérationnelle	Routine FinOps mensuelle, rapport au COMEX

7.1. Jalon 1 — Diagnostic et quick wins *(mois 1-2)*

Objectif. Établir un état des lieux factuel et dégager les premières économies pour financer la suite.

Actions clés.

- Réaliser l'audit complet des agents en production *(découverte automatique + entretiens)*
- Identifier les zombie agents *(sans propriétaire, sans valeur business, sans mise à jour)* et lancer le retrait
- Cartographier l'usage des frontier models et identifier les cas d'usage candidats au right-sizing
- Lancer 1 ou 2 chantiers de right-sizing pilotes *(modèle small fine-tuné)*
- Établir le baseline du coût d'inférence mensuel total

Point de contrôle 1 — Avez-vous une vue exhaustive et des premiers gains ?

Question	Critère de validation
Inventaire complet	Tous les agents identifiés, propriétaire désigné ou retrait engagé
Zombie agents traités	Au moins 80 % des zombies identifiés sont en cours de retrait
Right-sizing pilote	1 ou 2 cas d'usage migrés sur small model avec mesure qualité
Baseline coût	Coût d'inférence mensuel agrégé calculé et partagé avec CFO
Quick wins quantifiés	Économies projetées chiffrées sur 6 mois

7.2. Jalon 2 — Instrumentation unit economics *(mois 2-4)*

Objectif. Passer du coût agrégé au coût unitaire. Instaurer la mesure structurante.

Actions clés.

- Déployer un outil FinOps IA *(Finout, CloudZero ou équivalent)*
- Implémenter le virtual tagging par client / feature / transaction
- Instrumenter les trois métriques signature *(QACT, IRR, AUM)*
- Construire le dashboard FinOps IA partagé avec le CFO

- Définir les seuils d'alerte sur les anomalies *(coût qui double, IRR négatif persistant, AUM négatif)*

Point de contrôle 2 — La mesure unitaire est-elle opérationnelle ?

Question	Critère de validation
Outil FinOps en production	Plateforme opérationnelle avec ingestion des données
Virtual tagging	90 % des coûts d'inférence attribués à une unité business
Métriques signature	QACT, IRR, AUM mesurés sur au moins 80 % des agents en production
Dashboard partagé	Revue mensuelle CIO-CFO en place avec ce dashboard
Alertes opérationnelles	Détection automatique des anomalies activée

7.3. Jalon 3 — Agent Control Plane initial *(mois 4-6)*

Objectif. Construire l'infrastructure qui rend pilotable le portefeuille d'agents.

Actions clés.

- Déployer la fonction discovery *(Microsoft Agent 365 ou équivalent)*
- Compléter l'observabilité avec OpenTelemetry GenAI ou équivalent
- Activer le chargeback aux fonctions métier *(les coûts d'inférence sont attribués au budget des consommateurs)*
- Mettre en place un routing intelligent sur au moins 1 cas d'usage majeur
- Formaliser le lifecycle management des agents *(création, monitoring, mise à jour, retrait)*

Point de contrôle 3 — Le Control Plane est-il opérationnel ?

Question	Critère de validation
Discovery automatique	Découverte permanente des agents avec inventaire à jour
Observabilité multi-couche	Dashboard temps réel sur tokens, latences, anomalies
Chargeback opérationnel	Coûts effectivement répartis aux unités métier consommatrices
Routing intelligent	Au moins 1 cas d'usage avec routage small/medium/frontier
Lifecycle formalisé	Procédure documentée pour création, mise à jour, retrait

7.4. Jalon 4 — Hybrid Consumption Architecture *(mois 6-9)*

Objectif. Diversifier l'architecture pour réduire le lock-in et optimiser la répartition par profil de workload.

Actions clés.

- Cartographier les workloads par profil *(volume, sensibilité données, latence requise, fréquence)*
- Définir la politique d'allocation par profil *(quel workload va sur quel tier)*
- Mettre en place une vendor diversification policy *(au moins 2 fournisseurs par couche)*
- Conduire le premier test de re-sourcing *(simulation de migration d'un workload)*
- Négocier les contrats avec les nouveaux providers identifiés *(souverain, on-prem, edge)*

Point de contrôle 4 — L'hybridation est-elle structurée ?

Question	Critère de validation
Cartographie workloads	Tous les workloads classés par profil documenté
Politique d'allocation	Règles écrites, appliquées et auditées
Vendor diversification	Au moins 2 fournisseurs sous contrat par couche
Test re-sourcing	Simulation conduite avec rapport et recommandations
Contrats sécurisés	Clauses de portabilité incluses dans les nouveaux contrats

7.5. Jalon 5 — Maturité opérationnelle *(mois 9-12)*

Objectif. Passer du mode chantier au mode pérenne. Capitaliser et faire évoluer.

Actions clés.

- Institutionnaliser la routine FinOps IA mensuelle *(revue, ajustements, décisions)*
- Produire le rapport trimestriel COMEX sur le FinOps IA
- Lancer la formation des PO seniors aux concepts d'unit economics
- Réviser annuellement la politique de right-sizing en fonction des nouveaux modèles disponibles

- Engager le test de re-sourcing annuel sur un nouveau workload

Point de contrôle 5 — La maturité est-elle pérenne ?

Question	Critère de validation
Routine FinOps mensuelle	Revue stable, livrables formalisés, décisions prises
Rapport COMEX	Trimestriel, structuré, lu par le CFO et le COMEX
Formation PO	Au moins 80 % des PO seniors formés aux unit economics
Politique right-sizing révisée	Mise à jour annuelle documentée
Test re-sourcing récurrent	Devenu une pratique annuelle, pas un événement

7.6. Synthèse du plan d'action

Le plan est calé sur 12 mois. Il combine cinq jalons séquentiels avec cinq points de contrôle structurés qui valident le passage. La logique est progressive : audit, mesure, plateforme, architecture, maturité. Chaque jalon repose sur la qualité du précédent. Les économies dégagées par les premiers jalons financent les suivants. Sauter un jalon ou expédier un point de contrôle expose à des fragilités structurelles à 18 mois.

Notre recommandation : ne pas chercher à aller plus vite que les jalons. Chercher à valider chaque point de contrôle de manière documentée. Le FinOps IA est un système qui se construit, pas un projet qui se livre. Sa qualité se mesure sur 24 mois, pas sur 6.

8. Pour aller plus loin

8.1. Les autres lignes de fracture de l'édition 2026-2027

Ce PDF 3 a traité la troisième ligne de fracture : la maîtrise du coût de l'inférence et la mise en place d'un FinOps IA mature. Les autres lignes complètent la cartographie de la grande bifurcation.

PDF 1 — De l'IA à l'entreprise agentique : le Product Operating Model 2026. La refonte du POM 2018-2024 pour accueillir les agents comme acteurs des squads. Trois transformations : agents intégrés aux squads, SDLC humain-IA, mesure de la contribution hybride.

PDF 2 — Gouvernance, AI Act et Agentic Constitution. L'AI Act entre en application pleine le 2 août 2026. Comment passer d'une gouvernance documentaire à une gouvernance opérationnelle. Les concepts de policy-as-code, out-of-process enforcement, Agentic Constitution, trio CIO-CISO-CDO.

PDF 4 — Agentic Value Streams. Les méthodes de Strategic Portfolio Management pensées en 2018-2022 ne survivent pas au SI agentique. Cartographie ArchiMate agents-first, modélisation agent-centric, OKR avec ventilation contribution humain/IA/hybride, refonte du SDLC autour du pairing humain-IA.

PDF 5 — Leadership Transformation IA. La transformation personnelle du CIO. Le passage de CIO-contrôleur à CIO Agent-Enabler. Les trois croyances ancrées à déverrouiller. Le triptyque de leader 2026.

L'Executive Summary global de l'édition donne une vue d'ensemble des cinq lignes de fracture et de leurs interactions.

8.2. L'accompagnement AEP

Agile Enterprise Partner accompagne les CIO de grands comptes dans la mise en place d'un FinOps IA mature et d'une Hybrid Consumption Architecture maîtrisée. Notre positionnement 2026 :

architectes du modèle opérationnel IT et Digital pour le SI agentique. Trois offres sont directement activées sur les enjeux de ce PDF.

Offre Strategic Portfolio Management. Diagnostic FinOps IA *(grille des 10 questions, audit zombie agents, baseline)*, design de la trajectoire et du programme, instrumentation des trois métriques signature *(QACT, IRR, AUM)*, accompagnement du dialogue CIO-CFO. Frameworks mobilisés : SAFe Lean Portfolio Management, FinOps Foundation principles, KPMG agentic AI ROI framework.

Offre Enterprise Architecture pour SI agentique. Cartographie de l'Hybrid Consumption Architecture cible, choix de plateforme et de fournisseurs, design de l'Agent Control Plane, vendor diversification policy. Frameworks mobilisés : TOGAF, ArchiMate adapté agentic, NVIDIA reference architectures.

Offre CIO Office IA-Ready. Mise en place du dashboard FinOps IA partagé CIO-CFO, structuration de la routine mensuelle, instrumentation des KPI, accompagnement du COMEX sur le pilotage économique de l'IA agentique.

Les deux autres offres *(Product Operating Model, Gouvernance IA et Conformité AI Act, Leadership Transformation IA)* sont activées sur les enjeux des autres PDF.

Contact : contact@agile-enterprise-partner.com — +33 6 32 54 58 92 Site : https://agile-enterprise-partner.com

9. Bibliographie

9.1. Études primaires 2025-2026

- **Gartner**. *2026 CIO and Technology Executive Survey*. Octobre 2025.
- **Gartner**. *2026 Hype Cycle for Agentic AI*. Janvier 2026.
- **Gartner**. Analyse mars 2026 sur les multiplicateurs agentiques *(5 à 30 fois plus de tokens par tâche)*. Reprise par Oplexa et AnalyticsWeek.
- **MIT Project NANDA**. *The GenAI Divide: State of AI in Business 2025*. Juillet 2025.
- **IDC**. *FutureScape Worldwide AI 2026 Predictions*. Octobre 2025.
- **Forrester**. *Predictions 2026: AI Agents*. Novembre 2025.
- **Forrester**. *Agent Cost Structure Analysis*. Repris par Digital Applied 2026.
- **KPMG**. *Q4 AI Pulse Survey*. Janvier 2026. 130 dirigeants C-suite, organisations 1 milliard+ revenue.
- **KPMG**. *Q1 2026 Global AI Pulse Survey*. 2 000+ dirigeants globaux.
- **KPMG**. *Global Tech Report 2026: Leading in the Intelligence Age*. Janvier 2026. 2 500 dirigeants, 27 pays, 8 industries.
- **KPMG**. *Intelligent Retail: A blueprint for creating value through AI-driven transformation*. 2025.
- **CloudZero**. *State of AI Costs 2025-2026*.
- **FinOps Foundation**. *2026 State of FinOps Report*.
- **Deloitte**. *2026 AI Token Economics Analysis*.
- **NVIDIA**. *2026 enterprise AI optimization priorities*.

9.2. Modèles économiques et FinOps IA

- **AnalyticsWeek**. *Inference Economics: Solving 2026 Enterprise AI Cost Crisis*. Mars 2026.
- **AnalyticsWeek**. *AI FinOps and Sovereign Infrastructure: AI Costs in 2026*. Janvier 2026.
- **Spheron**. *AI Inference Cost Economics in 2026: GPU FinOps Playbook*. Avril 2026.
- **Stabilarity Hub**. *Inference Economics: The Hidden Cost Crisis Behind Falling Token Prices*. Mars 2026.

- **Oplexa**. *AI Inference Cost Crisis 2026: Why Your AI Bill Is Exploding*. Mars 2026.
- **CloudZero**. *Inference Cost Explained: How to Reduce LLM & AI Inference Spend*. 2026.
- **Finout**. *Best FinOps Tools for Managing AI Costs in 2026*. Mai 2026.
- **Flexera**. *FinOps for AI: Govern GPU, Token & SaaS Spend*. Novembre 2025.
- **Adnan Masood**. *AI FinOps: Turning Tokens into Outcomes*. Octobre 2025.
- **Jon Radoff**. *The State of AI Agents in 2026*. Février 2026.

9.3. Right-sizing et small models

- **TechCrunch**. *In 2026, AI will move from hype to pragmatism*. Janvier 2026.
- **CloudGeometry**. *Right-Sized AI: Why Small Language Models Outperform Giants*. 2025.
- **Iterathon**. *Small Language Models 2026: Cut AI Costs 75 % with Enterprise SLM Deployment*. Décembre 2025.
- **Arxiv 2509.18101**. *A Cost-Benefit Analysis of On-Premise Large Language Model Deployment*. Septembre 2025.
- **Rick Hightower**. *The Economics of Deploying Large Language Models: Costs, Value, and 99.7 % Savings*. Juillet 2025.
- **ITPro**. Sur la *« hidden fallacy »* des SLM *(citation Articul8 / Subramaniyan)*. Juillet 2024.

9.4. Souveraineté de l'inférence

- **NVIDIA Newsroom**. *Europe Builds AI Infrastructure With NVIDIA*. Annonce VivaTech / Hannover Messe 2025-2026.
- **Deutsche Telekom + NVIDIA**. *Industrial AI Cloud launch*. Annonce 2025, mise en production Q1 2026.
- **Mistral AI**. *Mistral Compute partnership with NVIDIA*. Annonces 2025-2026.
- **OVHcloud**. *Reference Architecture: deploying the Mistral Large 123B model in a sovereign environment*. 2025.
- **EU-Startups**. *Mistral AI €722 million debt financing for Bruyères-le-Châtel datacenter*. Mars 2026.
- **Digital Chiefs**. *Sovereign AI after Hannover Messe 2026: How the Board Establishes Architectural Sovereignty*. Avril 2026.
- **Innobu**. *Sovereign AI and EU-first Solutions for European Enterprises 2026*. Mars 2026.
- **IO+**. *Why Mistral's $830M raise is a win for European autonomy*. Avril 2026.

9.5. Cas et REX publics 2025-2026

- **AI Monk**. *12 Agentic AI Examples With Measurable ROI: Enterprise Case Studies From 2025-2026*. Avril 2026.
- **CX Dive**. *Klarna AI agent saved $60 million / 853 FTE equivalents*. Novembre 2025.
- **Klarna**. Communications publiques sur AI assistant et réintroduction agents humains.
- **JPMorgan**. Communications publiques sur 450+ AI use cases et COiN platform.
- **OpenAI**. *Morgan Stanley AI Assistant case study*.
- **KPMG**. *Sephora customer experience AI case studies*. Repris dans Intelligent Retail 2025.
- **BigCommerce**. *Ecommerce AI Agents in 2026 (Shopping's Next Big Shift)*. Décembre 2025.
- **Techverx**. *What Is Agentic Ecommerce? AI Agents Transforming Retail in 2026*. Mai 2026.
- **DestiLabs**. *Agentic Commerce 2026: Five AI Agents Every Store Needs*. 2026.
- **Insider One**. *AI in Retail: 10 Trends Shaping Ecommerce In 2026*. Mars 2026.
- **TheStreet**. *The next phase of AI spending is already underway*. Mai 2026.

9.6. Frameworks et publications 2025-2026

- **a16z**. *Agent economics 2026*.
- **Company of Agents**. *AI Agent Unit Economics: Scaling Your Agentic Fleet in 2026*. Janvier 2026.
- **California Management Review (Berkeley CMR)**. Mars 2026. Référence agentic operating model.
- **Microsoft**. *Agent 365*. Annoncé mai 2026.
- **Microsoft**. *Agent Governance Toolkit*. Avril 2026.

10. Annexe méthodologique

Les chiffres et observations cités dans ce rapport proviennent de sources publiques 2024-2026 vérifiées et datées. La période de référence des sources est janvier 2024 - mai 2026, avec une concentration sur les publications de janvier 2026 à mai 2026 *(période où le FinOps IA est devenu un sujet structurant dans le débat enterprise)*.

Le cas composite secteur Retail / E-commerce *(section 5.1)* combine des informations publiquement disponibles sur le secteur retail européen *(notamment KPMG Intelligent Retail 2025, Sephora, Klarna, ASOS, H&M cas publics)* et nos observations terrain anonymisées auprès de DSI retail, conformément aux engagements de confidentialité d'AEP. Aucune information confidentielle de mission n'est divulguée.

Les trois métriques signature introduites *(QACT, IRR, AUM)* combinent des concepts publics *(unit economics, FinOps cost-per-inference)* et notre formalisation propre. Elles ne prétendent pas à une normalisation universelle, elles proposent un cadre opérationnel cohérent pour le pilotage du FinOps IA.

Les recommandations actionnables *(section 4)* combinent l'analyse des cas publics et l'observation auprès de plus de 50 DSI grands comptes accompagnés ou rencontrés par AEP entre 2024 et 2026.

Le plan d'action *(section 6)* est calibré pour une exécution réaliste sur 12 mois pour des grandes entreprises avec une exposition IA significative. Il doit être ajusté à la maturité initiale, au périmètre et aux contraintes spécifiques de chaque organisation.

Sébastien Delayre, fondateur d'Agile Enterprise Partner. Mai 2026.