Hubadviser.

GUIDE PRATIQUE

FINOPS APPLIQUÉ À L'INTELLIGENCE ARTIFICIELLE

Dépenses IA : piloter pour accélérer

Pendant que certains DSI freinent l'IA par peur du budget, d'autres accélèrent en toute maîtrise. La différence ne tient ni à la taille de l'équipe, ni au budget : elle tient à une méthode. La voici, construite à partir d'un retour d'expérience de terrain au sein d'un des plus grands groupes industriels européens.

INCLUS
Auto-diagnostic de maturité FinOps IA en 10 questions

INCLUS
Le dispositif « minimum viable » : 90 jours, 0,2 ETP, 5 indicateurs

INCLUS
Les 3 questions à poser à votre DAF avant le prochain arbitrage IA

Par **Florian Beaurain**
Group FinOps Leader d'EDF, Expert Advisor Hubadviser

Webinar « Dépenses IA »
Juillet 2026

Synthèse exécutive

Beaucoup de DSI appuient aujourd'hui sur le frein IA : dépassements d'OPEX, pression de la direction financière, coûts imprévisibles. C'est une erreur stratégique, mais une erreur compréhensible : sans instrument de pilotage, freiner est la seule option rationnelle. **La thèse de ce guide est inverse : le pilotage financier n'est pas ce qui ralentit l'IA, c'est ce qui donne le droit d'accélérer.** Le DSI capable de dire à son CODIR « voilà ce que chaque usage coûte, voilà ce qu'il rapporte, voilà mes garde-fous » est celui qui obtient les budgets, et qui pèse dans les décisions.

Le cas Uber, largement commenté par la presse économique en 2026, illustre bien ce qui arrive sans ce pilotage : l'entreprise aurait consommé l'intégralité de son enveloppe IA annuelle en un seul trimestre, faute de maîtrise fine de ses consommations. Ce guide propose une autre voie.

Ce guide prolonge le webinar Hubadviser du 7 juillet 2026, animé aux côtés d'Ismail Charkaoui, co-fondateur d'Hubadviser. Chez EDF, on pilote la stratégie FinOps à l'échelle du Groupe : une pratique bâtie en partant de zéro, certifiée FinOps Foundation, aujourd'hui relayée par une communauté mondiale d'une vingtaine de praticiens. On y a ajouté ce que 45 minutes de webinar ne permettaient pas de couvrir : **une transposition opérationnelle pour les DSI qui n'ont pas les moyens d'un grand groupe**, et un outil d'auto-évaluation pour situer votre organisation.

Ce guide s'adresse aux DSI comptant environ 20 à 100 collaborateurs, qu'elles évoluent au sein d'ETI, du secteur public ou d'organisations de taille conséquente hors grands groupes globaux, confrontées aux mêmes enjeux de maîtrise des coûts de l'IA.

COMMENT UTILISER CE DOCUMENT

- **Sections 1 à 5** : comprendre la mécanique économique de l'IA (FinOps, token, classes de modèles, 7 couches de coûts) et les principes de pilotage par la valeur.
- **Section 6** : passer à l'action avec le dispositif FinOps IA « minimum viable » en 90 jours, calibré pour une DSI de taille intermédiaire, avec ses spécificités secteur public.
- **Section 7** : se situer grâce à l'auto-diagnostic en 10 questions et les 4 niveaux de maturité, avec la prochaine étape pour chacun.
- En chemin, deux outils à utiliser dès cette semaine : les **3 questions à poser à votre DAF** (p. 4) et la **grille des ratios de coûts entre classes de modèles** (p. 5).

Sommaire

1. FinOps : la discipline qui donne le droit d'accélérer

Le FinOps, contraction de *Finance* et *DevOps*, est une discipline opérationnelle et collaborative qui vise à maximiser la valeur créée par les technologies, à favoriser des décisions éclairées par les données en temps réel et à responsabiliser l'ensemble des acteurs, grâce à une coopération étroite entre IT, Finance et Métiers. Le cadre de la FinOps Foundation s'organise autour de **4 domaines et 22 capacités mesurables**, adossés à un modèle de maturité.

Notez la formulation : **maximiser la valeur**, pas minimiser la dépense. Un FinOps bien conduit ne dit pas « non » aux usages : il rend le « oui » défendable devant la direction financière. C'est toute la différence entre un DSI qui subit les arbitrages budgétaires et un DSI qui les mène.

Longtemps cantonné au cloud public, le périmètre du FinOps s'est considérablement élargi. Les données de la FinOps Foundation montrent l'ampleur du mouvement :

98 %
des pratiques FinOps couvrent désormais les usages IA

90 %
intègrent les environnements SaaS et éditeurs

64 %
couvrent le licensing

57 %
le cloud privé

48 %
le datacenter en propre

Deux familles de leviers, pleinement activables sur l'IA

OPTIMISATION DES TAUX

Réduire le montant payé pour une ressource nécessaire, en agissant sur le **prix**.

- **Engagements long terme** : Reserved Instances, Savings Plans, Committed Use Discounts, avec jusqu'à 70 % de remise sur 1 à 3 ans.
- **Négociations contractuelles** : remises sur volume, Enterprise Agreements.

Indicateur : baisse de la facture mensuelle à périmètre constant.

OPTIMISATION DES TRAITEMENTS

Empêcher une dépense de se produire, en agissant sur l'**usage**.

- **By design** : architectures serverless, processeurs ARM/AMD dès la conception.
- **Rightsizing & auto-scaling** : ajuster les capacités au besoin réel.
- **Instances spot / préemptibles** : jusqu'à 90 % de réduction sur les traitements batch et analytiques.

Indicateur : écart entre trajectoire prévue et dépense constatée (Forecast Accuracy Rate).

Pour un DSI, la conséquence est structurante : la gouvernance des coûts ne peut plus se penser silo par silo. IA, SaaS, licences, cloud privé et datacenter relèvent d'une même démarche. Les leviers éprouvés sur le cloud classique restent valables sur l'IA, **à condition d'y ajouter une couche de pilotage spécifique**. C'est l'objet de la section suivante.

2. FinOps for AI : le token change les règles du jeu

La métrique centrale des consommations IA est le **token**, l'unité de représentation du texte traité et généré par les modèles. Elle introduit une rupture avec le cloud classique : la dépense devient fonction du contexte transmis, des usages, du raisonnement demandé et du passage à l'échelle. Résultat : **des dépenses volatiles et difficilement prévisibles**, avec des profils de consommation inadaptés aux référentiels existants, un empilement de coûts, une forte sensibilité aux choix de modèles et des grilles tarifaires mouvantes.

La réponse n'est pas de freiner les usages, mais de monter en rigueur : là où le cloud se pilotait par revues mensuelles, les consommations IA exigent un **monitoring en quasi temps réel**. Et un changement de focale : la question n'est plus « combien consomme-t-on ? » mais **« pourquoi consomme-t-on, et quelle valeur métier cela produit-il ? »**.

Le triptyque d'arbitrage : performance, coût et valeur métier

Chaque cas d'usage doit être évalué selon trois piliers : le **niveau de performance attendu** (qualité de raisonnement, vitesse, multimodalité), le **coût** du modèle mobilisé et la **valeur business** générée. C'est ce ratio, et non le prestige du modèle, qui détermine si un usage est bien dimensionné. Cette logique de pilotage par la valeur est portée au niveau international par la **Tokenomics Foundation** (Token Economics), initiée sous l'égide de la Linux Foundation et de la FinOps Foundation.

Les indicateurs à mettre sous contrôle

- **Coût par requête / par inférence**, en précisant modèle, version et volumes associés ;
- **Coût par cas d'usage, par utilisateur, par projet ou par entité**, la granularité qui permet la refacturation ;
- **Taux de conformité aux budgets** (soft limits) et écart prévisionnel ;
- **ROI et gains de productivité**, en temps monétisé ou en valeur business générée ;
- **Indicateurs techniques d'optimisation**, comme le cache hit ratio.

> *L'objectif final : corréler ce que représente l'ensemble de la stack d'un point de vue coût avec la valeur business apportée. Une meilleure visibilité, non pas seulement sur les consommations, mais sur la valeur.*
>
> **À RETENIR**

À UTILISER DÈS CETTE SEMAINE : LES 3 QUESTIONS À POSER À VOTRE DAF

1. « Si notre consommation IA double le mois prochain, **au bout de combien de jours le savons-nous**, et qui décide quoi ? »
2. « Quel usage IA accepterions-nous de **couper demain**, et sur quel critère objectif ? »
3. « Sur quel cas d'usage sommes-nous capables de **démontrer un ROI aujourd'hui**, chiffres à l'appui ? »

Si aucune des trois ne trouve de réponse en moins d'une minute, la section 6 de ce guide est votre priorité du trimestre.

3. Le modèle économique de l'IA : raisonner en ratios, pas en euros

Les grilles tarifaires des fournisseurs évoluent chaque mois ou trimestre : raisonner en prix absolus, c'est être périmé avant d'avoir arbitré. Ce qui reste stable : **les ordres de grandeur entre classes de modèles** et la mécanique du token. Et le vrai coût dépasse toujours la facture du modèle, car il inclut la donnée, l'orchestration, l'observabilité, la sécurité et l'exploitation.

Comprendre le token, l'unité qui fait la facture

Un token est l'unité de découpage du texte par les modèles : un mot court vaut un token, un mot long deux ou trois. Chaque échange se facture en deux temps : les tokens soumis (l'input : la question et son contexte) et les tokens générés (l'output : la réponse). Or **générer coûte 3 à 4 fois plus cher que soumettre**, car le modèle produit sa réponse token par token, le calcul le plus intensif, là où la lecture du contexte se traite en une seule passe. Exemple concret : à question identique, demander une synthèse d'une page plutôt qu'un rapport de dix pages divise le coût de sortie par dix.

QUATRE RÉFLEXES QUI AGISSENT DIRECTEMENT SUR LA FACTURE

- **Limiter les sorties inutilement longues** en calibrant la longueur attendue dès la consigne ;
- **Privilégier des réponses synthétiques** quand l'usage le permet ;
- **Surveiller les usages générant de gros volumes de texte** (rédaction, comptes-rendus, code) ;
- **Intégrer des KPI de consommation par usage** et une gouvernance associée (quotas, formats de sortie).

La grille de lecture : trois classes, un rapport de 1 à 30

Classe de modèle	Ratio de coût indicatif (output)	Type d'usage cible	Règle d'attribution
Léger / rapide	**1×**	Classification, extraction, tâches simples à fort volume	Par défaut pour tous : c'est le modèle qu'on quitte sur justification, pas l'inverse
Milieu de gamme	**≈ 3–5×**	Production généraliste, assistants internes	Attribué par population ou par usage cartographié
Premium (raisonnement)	**≈ 10–30×**	Tâches complexes, workflows agentiques, raisonnement avancé	Arbitré au cas par cas, sur besoin réel démontré

Ordres de grandeur à date de juillet 2026 : output ≈ 0,25–4,60 €/M tokens (léger), 0,75–13,75 € (milieu de gamme), 11–27,50 € (premium) ; input ≈ 0,05–0,75 / 0,40–2,75 / 1,85–4,60 €/M tokens. Ces montants évoluent vite ; les ratios, beaucoup moins.

Trois voies d'accès aux modèles

- **Via API**, la porte d'entrée naturelle : paiement à l'usage au token, remises négociables sur engagement (typiquement au-delà de 50 k$/an), en choisissant des régions d'inférence conformes à vos exigences réglementaires ;
- **Via GPU hébergé**, pour amortir la ressource : le coût par token devient (amortissement GPU + réseau + stockage + exploitation) ÷ tokens produits ; le tarif GPU se négocie (environ 2,75 €/h pour un H100 à date) et des optimisations comme vLLM en maximisent l'utilisation ;
- **En local (datacenter)**, pour la souveraineté et la conformité, en tenant compte de la pénurie actuelle de GPU ; contre-intuitivement, le coût complet amorti sur 3 ans peut s'avérer inférieur à l'approche API sur certains profils d'usage.

4. Multi-modèles et routage : la fin du LLM unique

Conséquence directe de cette grille : une organisation mature ne pilote pas « un LLM », elle pilote un **portefeuille de modèles**. Cela suppose une cartographie à jour croisant trois dimensions : les **populations** (qui utilise ?), les **usages** (pour quoi faire ?) et les **modèles autorisés** (avec quelle classe de performance et de coût ?).

LE RISQUE N° 1 IDENTIFIÉ

Laisser des populations entières utiliser au quotidien un modèle premium, **jusqu'à 30 fois plus cher qu'un modèle léger**, pour des tâches qu'un modèle rapide traiterait aussi bien. La règle saine : **le léger par défaut, le premium sur justification**, et non l'inverse.

Du routage statique au routage dynamique

Le premier niveau de maturité est le **routage statique** : on fixe le modèle adapté à chaque cas d'usage selon le triptyque performance, coût et valeur. Le niveau supérieur est le **routage dynamique**, programmatique : les appels sont orientés en temps réel selon la latence mesurée, voire selon des plages horaires. Un usage critique conserve un modèle rapide aux heures de pointe ; les usages moins sensibles basculent vers des modèles moins coûteux.

Des modèles légers qui rattrapent les modèles frontière

Deux familles cohabitent sur le marché. Les **modèles légers** (par exemple GPT-4o mini, Claude Haiku ou Mistral Small) sont conçus pour aller vite et traiter de gros volumes à moindre coût : ils conviennent très bien au tri, à l'extraction ou à la classification. Les **modèles frontière** (par exemple GPT-5.5, Claude Opus ou Gemini Pro) embarquent la plus grande capacité de raisonnement disponible à un instant donné : ils excellent sur des tâches complexes, mais coûtent 10 à 30 fois plus cher en sortie, comme vu en section 3.

On observe de plus en plus la fin des tarifs subventionnés, avec une convergence des prix entre fournisseurs et un rapprochement du ratio performance/coût des modèles légers vers celui des modèles frontière sur des cas d'usage bien circonscrits. Concrètement : ce qu'un modèle frontière savait faire seul il y a un an, un modèle léger le fait aujourd'hui presque aussi bien, pour une fraction du prix. Le réflexe « toujours le plus gros modèle » devient donc chaque trimestre un peu plus coûteux, et un peu moins justifié.

L'open source comme solution de repli stratégique

D'un côté, les modèles propriétaires (GPT d'OpenAI, Claude d'Anthropic, Gemini de Google) ne sont accessibles que via l'API du fournisseur, sous ses conditions et sa juridiction. De l'autre, des modèles ouverts comme Llama (Meta), Mistral (Mistral AI) ou DeepSeek publient leurs poids : n'importe quelle organisation peut les héberger sur sa propre infrastructure et les faire tourner sans dépendre d'un fournisseur unique.

> *On a déjà vu certains modèles coupés du jour au lendemain, et on n'est pas à l'abri demain d'un export control activé sur des modèles très largement utilisés. Avoir un modèle open source en alternative à votre modèle industriel, en solution de repli, c'est vraiment important.*
>
> **À RETENIR**

Au-delà du coût, cette dépendance constitue un **risque de continuité d'activité**, particulièrement sensible pour le secteur public et les opérateurs d'importance vitale, où il rejoint les exigences de souveraineté. Mais ce choix a un prix technique : héberger un modèle open source exige des compétences que l'appel à une API ne demande pas, comme le MLOps ou la sécurisation de l'infrastructure GPU. Pour une DSI de taille intermédiaire, le bon compromis est souvent de confier cet hébergement à un partenaire ou un hyperscaler plutôt que de tout internaliser : l'essentiel est de garder la main sur le choix du modèle et la capacité à basculer rapidement.

5. L'empilement des 7 couches de coûts

Le coût complet d'un usage IA ne se lit pas sur une ligne de facture : il s'empile sur sept couches techniques, chacune portant ses postes de dépense et ses leviers d'optimisation. Cette grille en sept couches est celle que l'on utilise pour structurer une architecture de coûts ; elle vaut pour tout DSI ou architecte d'entreprise.

L7	**Orchestration applicative**	Agents, workflows métier
L6	**Gouvernance FinOps**	Tags, budgets, billing, policies, refacturation
L5	**Data, RAG et cache**	Embeddings, retrieval : OpenSearch, Redis, S3, GCS, VectorSearch, BigQuery
L4	**Gateway et routage**	API & MCP Gateway, proxy LLM (ex. LiteLLM), budgets par endpoint
L3	**Inférence**	Serving runtime, KV cache : plateformes d'inférence des hyperscalers
L2	**Cloud foundation**	Réseau, stockage
L1	**Compute (GPU)**	Ressources allouées ou achetées, à rentabiliser à 100 %

Trois couches à surveiller en priorité

L1, le compute. Dès qu'on possède ou réserve du GPU, l'enjeu bascule : il ne s'agit plus de réduire la consommation mais de **maximiser le taux d'utilisation** d'un actif amorti. Un GPU utilisé à 40 % est une dépense, pas un investissement.

L3, l'inférence. L'inférence correspond au moment où un modèle d'IA est utilisé pour produire une réponse à partir d'une demande utilisateur ; c'est cette phase d'exécution qui génère la majorité des coûts opérationnels. On y trouve l'optimisation la plus rentable et la plus simple à activer : le **KV cache** (cache clé-valeur appliqué aux prompts). Exemple concret : un assistant interne qui répond à partir d'un même corpus (une FAQ RH, un règlement) réutilise à chaque question le même contexte documentaire ; grâce au KV cache, ce contexte n'est calculé qu'une seule fois puis réutilisé, ce qui peut diviser par deux ou trois le coût de ce type d'usage.

L4, la gateway et le routage. Couche fondamentale et souvent négligée : elle porte les règles de routage entre modèles, mais aussi les **budgets et quotas par endpoint**. C'est le point de contrôle où une politique FinOps devient exécutable techniquement, et non plus seulement déclarative. Pour une petite DSI, c'est aussi **le meilleur premier investissement** : un proxy LLM bien configuré donne d'un coup la visibilité, les quotas et le routage.

LE MESSAGE CLÉ

Ce que l'utilisateur, et souvent le contrôle de gestion, perçoit, c'est la couche 7. Ce que la facture reflète, c'est l'empilement des sept. **Sans modèle de répartition des coûts couche par couche, impossible de construire une refacturation crédible, ni de savoir où agir en cas de dérive.**

6. Le FinOps IA minimum viable : 90 jours, 0,2 ETP, 5 indicateurs

Le dispositif d'EDF (20 personnes dédiées, une communauté mondiale) n'est pas transposable tel quel à une DSI de taille intermédiaire. Mais son principe l'est, et sur ce point on est catégorique : **ne pas staffer le pilotage, c'est se retrouver dans six mois avec un budget brûlé, perçu trop tard**. Voici le dispositif calibré pour une DSI de 20 à 100 collaborateurs, en ETI, dans le secteur public ou au sein d'une organisation plus large : un référent à **0,2 ETP** (un jour par semaine), pas nécessairement un profil FinOps certifié, plutôt un architecte ou un lead technique doté d'une sensibilité économique, positionné en amont, au plus près des projets.

J0 – J30 Voir	**Cartographier et instrumenter.** Inventaire des usages IA réels (y compris le shadow AI), des modèles utilisés et des canaux de dépense (API, SaaS, licences copilotes). Mise en place d'un proxy LLM / gateway avec tagging et quotas de base par équipe. Désignation du référent et de son mandat.
J30 – J60 Mesurer	**Poser les 5 indicateurs et le rituel.** Tableau de bord mensuel (voir ci-dessous), premier rituel DSI-DAF de 30 minutes, arbitrage des accès premium (le léger par défaut, le premium sur justification). Alertes de dépassement *avant* le franchissement des seuils, pas après.
J60 – J90 Arbitrer	**Première revue de spend et politique v1.** Revue des écarts budget vs réalisé, showback simple par direction, politique multi-modèles formalisée (qui, quoi, quels quotas), premier business case valeur sur l'usage phare. À J90 : vous savez ce que l'IA vous coûte, ce qu'elle rapporte, et qui décide.

Les 5 indicateurs qui suffisent pour démarrer

- **1. Dépense IA totale mensuelle**, ventilée par canal (API, SaaS, licences), la ligne que votre DAF attend ;
- **2. Coût par cas d'usage** sur votre top 5, la granularité qui permet d'arbitrer ;
- **3. Répartition premium / milieu de gamme / léger**, votre exposition au risque de sur-consommation ;
- **4. Écart budget vs réalisé** et sa tendance : la dérive se voit dans la pente, pas dans le solde ;
- **5. Un indicateur de valeur** sur un usage phare (temps gagné monétisé, dossiers traités), celui qui transforme la revue de coûts en revue de valeur.

SPÉCIFICITÉS SECTEUR PUBLIC

Trois points d'attention propres aux collectivités et établissements publics :

- **La friction budgétaire.** Une dépense à la consommation, par nature variable, s'accommode mal d'un cadre d'engagement annualisé. Les quotas techniques posés à la couche gateway ne sont donc pas un confort, mais une nécessité de conformité budgétaire.
- **La souveraineté.** Le choix des régions d'inférence et des modèles est aussi une décision juridique. La solution de repli open source (section 4) rejoint ici directement les exigences de souveraineté.
- **L'achat.** Anticiper le véhicule contractuel (centrales d'achat, seuils de marché) dès la phase d'expérimentation, pour ne pas découvrir au moment d'industrialiser que le fournisseur retenu n'est pas achetable.

7. Auto-diagnostic : quel est votre niveau de maturité FinOps IA ?

Dix questions fermées. Comptez vos « oui », honnêtement : ce diagnostic ne vaut que par la lucidité de vos réponses.

1	Chaque cas d'usage IA en production a-t-il un **owner** identifié, responsable de son coût et de sa valeur ?	oui ☐
2	Connaissez-vous votre **dépense IA du mois dernier**, tous canaux confondus (API, SaaS, licences copilotes) ?	oui ☐
3	Savez-vous **quel modèle et quelle version** sont utilisés par chacun de vos usages en production ?	oui ☐
4	Avez-vous des **quotas ou budgets techniques** par endpoint, par équipe ou par usage ?	oui ☐
5	Êtes-vous alerté **avant** le dépassement d'un budget, et non en découvrant la facture ?	oui ☐
6	Mesurez-vous un **coût unitaire** (par requête, par dossier traité, par utilisateur) sur au moins un usage ?	oui ☐
7	Avez-vous un **rituel régulier DSI-DAF** dédié aux dépenses et à la valeur de l'IA ?	oui ☐
8	Les accès aux **modèles premium** sont-ils arbitrés au cas par cas, sur besoin démontré ?	oui ☐
9	Disposez-vous d'une **alternative testée** (modèle de repli, open source) pour vos usages critiques ?	oui ☐
10	Pouvez-vous relier vos consommations à une **valeur métier chiffrée** (temps gagné, revenu, qualité) ?	oui ☐

0 – 3 OUI
IA subie
Les usages existent, la maîtrise non. Priorité absolue : la phase « Voir » du plan 90 jours (p. 8), cartographie et gateway. Ne lancez aucun nouveau POC avant.

4 – 6 OUI
IA pilotée
Les bases sont posées, la valeur n'est pas encore mesurée. Priorité : les 5 indicateurs et le rituel DSI-DAF, pour passer de la visibilité des coûts au pilotage par la valeur.

7 – 8 OUI
IA intégrée
La gouvernance fonctionne. Priorité : l'optimisation fine (KV cache, routage, négociation des contrats) et le business case par usage pour accélérer les investissements.

9 – 10 OUI
IA agentique
Vous êtes prêt pour l'échelle : routage dynamique, workflows agentiques, chargeback complet. Votre enjeu n'est plus la maîtrise mais la vitesse d'industrialisation.

Ce diagnostic reprend les quatre stades de maturité d'adoption de l'IA du référentiel Hubadviser (**subie, pilotée, intégrée, agentique**), appliqués à la dimension financière. Le niveau obtenu n'est ni une note ni un jugement : c'est un point de départ, et surtout **un langage commun à poser sur la table du CODIR**. « Nous sommes en IA subie, voici le plan pour être en IA pilotée dans 90 jours » est un discours qui obtient des arbitrages.

8. Synthèse : ce que la DSI peut concrètement maîtriser

Reprendre la main sur les dépenses IA ne demande ni la taille ni les moyens d'un grand groupe. Trois axes d'action, activables dès maintenant :

1. GOUVERNANCE ET PILOTAGE

- Définir des **owners** par cas d'usage ou produit.
- Fixer des standards de décision, de mesure et d'escalade (dépassement de budget, SLA en risque, shutdown…).
- Partager les KPI entre IT, métiers et finance.
- Créer un rituel de pilotage pour arbitrer coût, performance et valeur métier.

2. DÉMARCHE FINOPS

- Cartographier les usages IA et leurs coûts.
- Mesurer coût unitaire, consommation et valeur métier délivrée.
- Suivre les écarts budget vs réalisé et la dérive.
- Organiser des revues régulières de spend.
- Outiller prévision, allocation et chargeback/showback.

3. LEVIERS TECHNIQUES

- Optimiser l'allocation GPU et la capacité utilisée (ex. vLLM).
- Réduire tokens et appels inutiles (compression, chunks, cache).
- Quotas, budgets, alertes, tagging, rate limiting.
- Rationaliser environnements et versions de modèles.
- Paramétrer un routage intelligent sans dégrader les performances.

Trois convictions pour conclure

1. Le pilotage financier est le préalable à toute industrialisation de l'IA, et son accélérateur. Sans maîtrise des coûts, les projets IA restent des POC coûteux ; avec elle, chaque « oui » devient défendable. L'industrialisation ne consiste pas à ajouter des cas d'usage, mais à structurer un modèle économique reproductible où chaque appel est budgété, monitoré et optimisé.

2. La collaboration DSI-DAF doit dépasser le contrôle des dépenses. DAF, DSI et référent FinOps construisent ensemble des business cases crédibles : coût réel par traitement, leviers de négociation, mécanismes d'optimisation (caching, batch, routing multi-modèles), clauses contractuelles (accès aux futures versions, garanties de capacité).

3. La complexité technique et commerciale exige un accompagnement spécialisé. Entre tarifs mouvants et remises mal documentées, l'échange direct avec des praticiens permet d'adapter ces principes à vos contraintes : stack existante, volumes, fournisseurs déjà engagés.

ET PARADOXALEMENT…

L'IA, souvent présentée comme un levier de réduction de la part humaine, exige d'abord un **investissement humain** : la compétence qui sait relier tokens, modèles et valeur métier. C'est cette compétence, interne, externe ou hybride, qui fait la différence entre les DSI qui freinent et ceux qui accélèrent.

Pour aller plus loin

Cette dernière page est proposée par Hubadviser, éditeur de ce guide.

Ce guide vous donne le cadre. Le passage à l'action (cartographier vos usages, poser votre gouvernance, négocier vos contrats, défendre vos arbitrages au CODIR) se joue dans le détail de votre contexte : stack existante, volumes, fournisseurs engagés, contraintes de souveraineté. C'est exactement là qu'Hubadviser intervient.

Hubadviser donne aux DSI ce que seules les plus grandes organisations mondiales s'offrent : **un niveau d'expertise, d'influence et de méthode accessible à la demande**. Structurer et booster la performance de votre DSI, peser davantage au CODIR, conforter vos décisions technologiques avant d'investir (IA, Cyber, Cloud) et prendre votre poste avec le bon armement.

Échanger avec un expert du calibre de Florian Beaurain

Florian Beaurain fait partie de notre pool d'Expert Advisors. Il est mobilisable en session individuelle d'une heure, en visio, pour structurer votre architecture de coûts IA, challenger vos hypothèses ou préparer vos négociations fournisseurs.

Sparring Partner DSI

Coaching opérationnel mensuel (1 séance de 2h) sur la stratégie IT, Data et IA et l'IT Operating Model. Un DSI de haut niveau à vos côtés, sans agenda caché.

Veille & Pool d'experts

Des sachants de top niveau partout dans le monde, consultables en visio par sessions d'1h, sur des sujets technologiques, métier ou organisationnels.

Interventions & Ateliers

Interventions stratégiques par d'anciens Patrons et Généraux ; ateliers Partners ciblés : Tableau de bord DSI, Réduction des coûts, Portefeuille stratégique…

Contact : contact@hubadviser.com · hubadviser.com

VOTRE RÉSULTAT AU DIAGNOSTIC EN MAIN ?

Envoyez-nous votre score (0 à 10) et votre principal point de friction : nous vous proposons **30 minutes d'échange offertes** avec un Partner Hubadviser pour identifier la première action à fort impact dans votre contexte.