

Économies de coûts par conception

Pourquoi COMPAiSS coûte moins sans compromettre l'exactitude

Table des matières

- Résumé exécutif
- Défi institutionnel : l'écart de mise à l'échelle de l'IA
- Le paradigme conventionnel : les coûts cachés de l'IA à génération prioritaire
- La distinction COMPAiSS : une architecture à exécution conditionnelle
- Analyse financière comparative : projections annuelles
- Intégrité structurelle : pourquoi l'exactitude est un sous-produit de la conception
- Conclusion : alignement institutionnel stratégique

1. Résumé exécutif

Pour les institutions traitant environ 100 000 requêtes assistées par l'IA par année, les systèmes d'IA d'entreprise conventionnels entraînent généralement des coûts d'exploitation compris entre 90 000 \$ et 200 000 \$. Un déploiement de COMPAiSS traitant le même volume de requêtes fonctionne pour sa part à un coût annuel d'environ 15 000 \$ à 30 000 \$.

Cet écart de coûts, de l'ordre de 6× à 12×, ne découle ni de prix réduits, ni d'une capacité moindre, ni de modèles de qualité inférieure. Il résulte d'une inversion architecturale structurelle. Les systèmes conventionnels permettent à l'inférence de l'IA de s'exécuter pour chaque requête — entraînant des coûts même pour des demandes non pertinentes, non prises en charge ou refusées. COMPAiSS introduit un mécanisme d'autorisation préalable qui empêche l'exécution de l'inférence du modèle tant qu'une réponse institutionnelle autorisée n'existe pas.

Ainsi, COMPAiSS consacre le budget exclusivement à des interactions faisant autorité et à valeur ajoutée, tout en éliminant de façon structurelle les coûts, les risques et la charge de gouvernance associés à un raisonnement de l'IA non autorisé.

La présente analyse porte sur les coûts d'exploitation et les implications en matière de gouvernance d'une architecture à exécution conditionnelle à l'échelle institutionnelle. Elle ne couvre pas les dépenses en immobilisations liées à des infrastructures non fondées sur l'IA, les coûts de gestion du changement ni les effets en aval sur la productivité. Tous les montants présentés correspondent à des estimations de coûts d'exploitation fondées sur les hypothèses énoncées à la section 5.

2. Défi institutionnel : l'écart de mise à l'échelle de l'IA

Les universités, les hôpitaux, les organismes du secteur public et d'autres institutions fortement réglementées gèrent de vastes ensembles fragmentés d'information faisant autorité — politiques,

procédures, règles d'admissibilité et obligations réglementaires — répartis sur des centaines, voire des milliers de documents.

La demande pour cette information continue de croître :

- Les étudiants, le personnel, les patients et les citoyens s'attendent à des réponses immédiates
- Les équipes de soutien humaines ne peuvent pas croître de façon proportionnelle
- Les modèles de services traditionnels exercent une pression accrue sur les budgets et les effectifs

Les systèmes d'IA sont de plus en plus déployés pour absorber cette demande. Toutefois, l'architecture de ces systèmes détermine s'ils réduisent les coûts et les risques institutionnels — ou s'ils les amplifient.

3. Le paradigme conventionnel : les coûts cachés de l'IA à génération prioritaire

La majorité des systèmes d'IA d'entreprise suivent une séquence à génération prioritaire :

- L'inférence commence dès la soumission d'une question
- Des documents sont récupérés (souvent au moyen de pipelines de *retrieval-augmented generation* (RAG))
- Une réponse est générée
- Des contrôles de sécurité, de modération et de gouvernance interviennent après la génération
- L'ensemble des interactions est consigné à des fins de conformité

Cette séquence crée trois facteurs structurels de coûts.

1. Gaspillage d'inférence

L'inférence s'exécute même pour des questions non pertinentes, exploratoires ou non autorisées. Les refus consomment malgré tout des jetons et des ressources, puisque le modèle s'est déjà exécuté.

2. Charge d'infrastructure persistante

Les architectures RAG exigent une ingestion continue des documents, des calculs d'intégration (*embeddings*) et l'exploitation de bases de données vectorielles, indépendamment du volume réel d'utilisation.

3. Gouvernance compensatoire

Puisque le modèle est autorisé à spéculer, les institutions doivent assumer les coûts liés à la modération, aux outils de sécurité et aux processus de révision afin d'atténuer les erreurs après la génération.

En somme, des coûts sont engagés avant même que le système sache si une question devrait être traitée.

4. La distinction COMPAiSS : une architecture à exécution conditionnelle

COMPAiSS inverse cette séquence :

- L'inférence ne s'exécute pas tant que l'autorisation n'est pas accordée
- Les requêtes non prises en charge ou hors périmètre ne déclenchent aucun appel au modèle
- Aucun jeton, pipeline de récupération ou système de modération ne s'exécute pour les requêtes non autorisées

L'évitement des coûts est structurel, et non comportemental.

Le mécanisme d'autorisation repose sur des algorithmes déterministes, non fondés sur l'IA, qui comparent les requêtes à des structures institutionnelles prédéfinies et à des sources approuvées. Ce mécanisme fonctionne à un coût de calcul négligeable comparativement à l'inférence des grands modèles de langage.

Comparaison des flux de traitement

IA d'entreprise conventionnelle

Question de l'utilisateur → Début de l'inférence → Récupération → Génération de la réponse → Filtres de sécurité → Modération → Diffusion de la réponse

COMPAiSS

Question de l'utilisateur → Vérification de l'autorisation →

- Si non autorisée : arrêt avec orientation
- Si autorisée : inférence de l'IA → Génération de la réponse à partir de sources approuvées → Diffusion de la réponse

Il ne s'agit pas d'une optimisation d'un modèle existant, **mais de l'élimination complète de catégories de coûts.**

Échec sécuritaire et erreur bornée

Faux négatifs (questions légitimes bloquées) : les utilisateurs sont redirigés vers des sources faisant autorité ou reçoivent des indications pour reformuler leur demande. L'impact est temporaire et à faible risque.

Faux positifs (cas limites autorisés) : les réponses demeurent strictement fondées sur des sources institutionnelles approuvées, ce qui empêche toute fabrication.

5. Analyse financière comparative : projections annuelles

Base : ~100 000 requêtes annuelles | ~60 % de taux d'autorisation

Hypothèses

- Les prix des jetons reflètent les normes actuelles du marché des services d'IA d'entreprise au moment de la rédaction
- Le taux d'autorisation constitue une hypothèse de planification ; les taux réels varient selon la maturité du périmètre
- Tous les montants correspondent à des coûts d'exploitation annuels, à l'exclusion des frais uniques de mise en service

Les estimations de jetons tiennent compte du contexte total transmis au modèle, et non uniquement de la question de l'utilisateur. Dans les systèmes RAG conventionnels, les documents récupérés, les citations et les instructions système sont insérés dans l'invite avant la génération, constituant souvent la majorité des jetons consommés par interaction.

Baseline Parameters and Token Assumptions

| Parameter | Conventional RAG | COMPAiSS |
|--|---|---|
| Annual query volume | ~100,000 | ~100,000 |
| Authorization rate | 100% (inference runs on all queries) | ~60% (planning assumption) |
| Token composition per authorized query | User query + retrieved documents + <i>instruée</i> . | User query + bounded generation (no document insertion) |
| Average tokens per authorized query | ~15,000–30,000 | ~3,000–8,000 |
| Model class | Enterprise LLM (GPT / Claude / Gemini) | Enterprise LLM (Claude / Gemini) |
| Operating environment | Regulated, audit-required | Regulated, audit-required |
| Annual inference (token) cost only | \$10k–\$30k | \$7k–\$18k |

In retrieval-augmented generation (RAG) systems, token usage is dominated by the insertion of retrieved document excerpts into the model prompt prior to generation. COMPAiSS does not insert documents into the model context; documents are used exclusively for authorization and scope control. As a result, authorized COMPAiSS queries consume materially fewer tokens per interaction.

Bien que l'inférence COMPAiSS soit refusée pour environ 40 % des requêtes afin d'éviter des coûts inutiles, ces utilisateurs reçoivent néanmoins une réponse à valeur ajoutée de type *échec*

sécuritaire (par exemple, des liens directs vers des politiques faisant autorité), fournissant ainsi un service institutionnel sans coût marginal de calcul

Comparaison des coûts côte à côte

| Cost Component | Conventional RAG | COMPAiSS |
|--------------------------|---------------------|--------------------|
| Inference (tokens) | \$10k–\$30k | \$7k–\$18k |
| RAG infrastructure | \$30k–\$60k | \$0 |
| Safety & moderation | \$30k–\$75k | \$0 |
| Monitoring & compliance | \$10k–\$20k | \$5k–\$15k |
| Total Annual Cost | \$90k–\$200k | \$15k–\$30k |

Note : Ces montants représentent les coûts opérationnels liés à l’inférence de l’IA, à l’infrastructure et à la gouvernance. Les conditions de licence commerciale, les frais contractuels ou les ententes de services propres à une institution sont traités séparément et varient selon la taille et la portée du déploiement.

6. Intégrité structurelle : pourquoi l’exactitude est un sous-produit de la conception

Un coût inférieur n’implique pas une exactitude moindre.

Les hallucinations exigent une inférence active. En empêchant l’inférence pour les requêtes non autorisées, COMPAiSS élimine les conditions structurelles qui permettent la fabrication d’information.

Contrairement aux systèmes qui tentent de filtrer ou d’« assurer » les hallucinations après la génération — laissant subsister un risque résiduel — COMPAiSS élimine la possibilité même d’un raisonnement spéculatif en empêchant toute inférence non autorisée.

L’exactitude est ainsi assurée par un périmètre borné, et non par une correction a posteriori.

7. Conclusion : alignement institutionnel stratégique

À l’échelle institutionnelle, COMPAiSS offre à la fois une réduction des coûts et une atténuation des risques par un même mécanisme : empêcher le raisonnement non autorisé de l’IA avant qu’il ne se produise.

Cette distinction architecturale permet aux institutions de répondre à une demande croissante de soutien automatisé tout en préservant l'exactitude, la traçabilité et la responsabilité fiduciaire.