LEEST® · OPTIMISATION DES COÛTS D'IA

Jusqu'à 80% d'économies sur vos coûts liés à l'IA

Nous analysons les leviers de réduction des coûts API d'inférence — sélection dynamique du modèle, cache, traitement par lots (batch), contrôle de la longueur des réponses et escalade qualité — puis Leest® les applique automatiquement, requête par requête. Le modèle premium n'intervient que lorsqu'il le faut.

−80 %

potentiel maximal modélisé
routage + cache + batch

JUSQU'À 40 À 60 % AVEC LE ROUTAGE SEUL

L'ESSENTIEL

Une même demande peut coûter **2 à 10 fois plus cher** selon le modèle qui la traite. Leest® attribue chaque tâche au modèle adapté, réutilise le contexte via le cache et bascule les traitements non urgents en batch. Objectif : **un coût par tâche réussie toujours plus bas**, à niveau de qualité contrôlé.

Ce que montre la simulation

01 · LE BON MODÈLE

Le modèle premium n'est nécessaire que pour une minorité des tâches.

Classification, extraction, reformulation, synthèse courte et opérations répétitives peuvent être confiées à des modèles économiques.

02 · LE ROUTAGE

Le routage automatique est plus fiable que le choix manuel.

Il applique systématiquement des règles de coût, latence, complexité, confidentialité et qualité, avec escalade en cas d'échec.

03 · LES AMPLIFICATEURS

Le cache et le batch amplifient fortement les économies.

Les instructions répétées, documents communs et traitements non urgents bénéficient de remises majeures selon les fournisseurs.

04 · LA BOUSSOLE

L'objectif n'est pas le modèle le moins cher, mais le coût par tâche réussie.

Une réponse bon marché mais incorrecte peut coûter davantage si elle exige plusieurs reprises ou une validation humaine.

L'équation de l'économie

LES LEVIERS COMBINÉS

L'économie n'est jamais un seul réglage. C'est l'addition de plusieurs leviers — que Leest® sait activer ensemble.

Modèle adapté + Cache + Batch + Réponses cadrées + Escalade qualité

En combinant ces leviers, l'économie mensuelle estimée atteint **35 % en routage prudent**, **55 % en routage équilibré** et **jusqu'à 75 %** avec cache et batch — hypothèse : 65 % de tâches simples, 25 % intermédiaires, 10 % complexes.

Pourquoi les factures divergent autant

FACTEUR	CE QU'IL CHANGE	IMPACT
Modèle utilisé	Les modèles premium coûtent nettement plus cher mais ne sont pas nécessaires pour toutes les requêtes.	Très fort
Tokens d'entrée	Historique, documents, instructions système et données répétées sont refacturés sans cache.	Fort
Tokens de sortie	Les réponses longues et le raisonnement étendu peuvent dominer la facture.	Très fort
Mode batch	Les traitements asynchrones bénéficient souvent d'environ 50 % de remise.	Fort
Cache	Le contenu répété peut être facturé à un tarif fortement réduit.	Très fort
Taux de reprise	Une mauvaise sélection de modèle entraîne retries, escalade et validation humaine.	Caché

Formule de pilotage — **Coût réel = tokens × tarif du modèle + outils + retries + validation humaine**

Repères tarifaires — écarts observables sur le marché

FOURNISSEUR / MODÈLE	ENTRÉE	SORTIE	POSITIONNEMENT
Anthropic Claude Fable 5	**$10**	**$50**	Agents longs / premium
Anthropic Claude Opus 4.8	**$5**	**$25**	Complexe / entreprise
Anthropic Claude Sonnet 5	**$3***	**$15***	Équilibre vitesse / intelligence
Anthropic Claude Haiku 4.5	**$1**	**$5**	Rapide / économique
OpenAI GPT-5.6 Sol	**$5**	**$30**	Frontière / premium
OpenAI GPT-5.6 Terra	**$2.50**	**$15**	Équilibre qualité / coût
OpenAI GPT-5.6 Luna	**$1**	**$6**	Rapide / économique
Mistral Large 3	**$0.50**	**$1.50**	Généraliste très économique
Google Gemini 2.5 Pro	**$1.25**	**$10**	Raisonnement complexe
Google Gemini 2.5 Flash	**$0.30**	**$2.50**	Volume / faible latence
Google Gemini Flash-Lite	**$0.10**	**$0.40**	Très économique

Tarifs API publics en USD par million de tokens, vérifiés au 23 juillet 2026. * Claude Sonnet 5 : tarif promotionnel $2 / $10 jusqu'au 31 août 2026 (tarif standard indiqué). **Lecture :** le coût de sortie d'un modèle premium peut être plus de 100 fois supérieur à celui d'un modèle très économique. Sources : Anthropic, OpenAI, Mistral et Google — pages tarifaires officielles. Les catalogues évoluent fréquemment.

Économies selon le niveau de dépense

Trois scénarios mensuels, hors coût de la plateforme Leest® et hors taxes.

BUDGET INITIAL
500 € / mois

Prudent	**−175 €** reste 325 €
Équilibré	**−275 €** reste 225 €
Avancé	**−375 €** reste 125 €
POTENTIEL ANNUEL	**4 500 €**

BUDGET INITIAL
1 000 € / mois

Prudent	**−350 €** reste 650 €
Équilibré	**−550 €** reste 450 €
Avancé	**−750 €** reste 250 €
POTENTIEL ANNUEL	**9 000 €**

BUDGET INITIAL
3 000 € / mois

Prudent	**−1 050 €** reste 1 950 €
Équilibré	**−1 650 €** reste 1 350 €
Avancé	**−2 250 €** reste 750 €
POTENTIEL ANNUEL	**27 000 €**

Exemple : une facture de 1 000 € / mois peut raisonnablement descendre vers 450 € avec un routage équilibré. L'économie réelle dépend du mix de tâches, des contraintes de qualité et du volume de contexte réutilisable.

Matrice de routage — chaque tâche au bon niveau de modèle

TYPE DE TÂCHE	ROUTE ÉCONOMIQUE	ROUTE STANDARD	ROUTE PREMIUM
Classification / tagging	**Par défaut**	Rarement	Jamais
Extraction structurée	**Par défaut**	Si document complexe	Exception
Reformulation / résumé	**Par défaut**	Si enjeu éditorial	Exception
Support client courant	**Par défaut**	Si contexte ambigu	**Si risque élevé**
Analyse métier	**Pré-tri**	Par défaut	**Si décision critique**
Code simple	**Préparation**	Par défaut	**Si architecture complexe**
Audit / stratégie	**Sous-tâches**	Synthèse intermédiaire	**Validation finale**
Agent autonome long	**Actions simples**	Orchestration	**Étapes critiques**

Comment Leest® procède, requête par requête

1 — L'ANALYSE
Classifier la requête
Type de tâche, complexité, confidentialité et enjeu qualité sont évalués automatiquement.

2 — LA ROUTE
Choisir le modèle
La politique de coût attribue la requête au modèle le plus économique compatible avec l'enjeu.

3 — L'EXÉCUTION
Cache & batch
Le contexte répété est mis en cache ; les traitements non urgents partent en lots remisés.

4 — LE CONTRÔLE
Vérifier le résultat
La sortie est contrôlée par rapport à l'attendu avant d'être livrée à l'application.

5 — L'ESCALADE
Uniquement si nécessaire
En cas d'échec, la requête remonte vers un modèle supérieur — jamais par défaut.

Leest® comme couche d'orchestration

Gouvernance et sécurité

- **Budgets et plafonds** — chaque équipe, client ou usage peut recevoir sa propre limite de dépense.
- **Journalisation** — chaque requête, chaque route et chaque coût sont tracés de bout en bout.
- **Règles de confidentialité** — les données sensibles peuvent être réservées à certains fournisseurs ou modèles.
- **Fallback fournisseur** — en cas d'indisponibilité, la requête bascule automatiquement vers un fournisseur de repli.

Le routeur peut appliquer des politiques différentes selon l'équipe, le client, le type de donnée ou le niveau de risque.

La promesse « jusqu'à 80% » est-elle crédible ?

Oui, comme potentiel maximal — pas comme résultat garanti pour chaque organisation.

▪ CAS OÙ 60 À 80% EST ACCESSIBLE

- ✓ Fort volume de tâches répétitives
- ✓ Usage systématique d'un modèle premium aujourd'hui
- ✓ Instructions ou documents largement réutilisés
- ✓ Traitements compatibles avec le batch
- ✓ Réponses courtes et structurées
- ✓ Contrôle qualité automatisable

▲ CAS OÙ LE GAIN SERA PLUS FAIBLE

- — Tâches presque toujours complexes ou critiques
- — Très faible volume mensuel
- — Contexte différent à chaque requête
- — Exigence d'un fournisseur ou modèle unique
- — Longues sorties indispensables
- — Taux d'échec élevé sur les petits modèles

Démonstration chiffrée — du routage aux ≈ 80%

ÉTAPE	COÛT RESTANT	ÉCONOMIE CUMULÉE
Situation initiale — 100 % des tâches sur un modèle premium (ex. GPT-5.6 Sol)	100,0 %	0 %
Après routage multi-modèles — 65 % Luna · 25 % Terra · 10 % Sol	35,5 %	**64,5 %**
Après batch (≈ −50 %) sur 60 % du coût restant	24,9 %	**75,1 %**
Après gains additionnels de cache et cadrage des sorties	≈ 20 %	**≈ 80 %**

Ce calcul montre que le résultat est plausible, pas qu'il sera obtenu chez chaque client : le gain réel dépend du mix de tâches, de la longueur des sorties, du taux de cache, de l'éligibilité au batch et du taux d'escalade vers un modèle supérieur.

Notre recommandation

- **Démarrer par un audit de 30 jours** — mesurer les tâches, les tokens, les retries et le coût par résultat.
- **Activer progressivement le routage** — d'abord sur les catégories à faible risque, avant d'étendre l'optimisation.
- **Piloter au coût par tâche réussie** — c'est l'indicateur qui protège la qualité pendant que la facture baisse.

Potentiel — **40 à 60 % d'économies avec le routage seul · jusqu'à 80 % avec cache et batch**

Sources et validité

- **Sources officielles consultées** — OpenAI API Pricing · Anthropic Models Overview & Pricing · Mistral Pricing · Google Gemini API Pricing.
- **Document commercial indicatif** — tarifs et modèles susceptibles d'évoluer ; les économies dépendent des usages réels.