flowlysis.

LE COÛT D'UNE RÉPONSE

On calcule ce que coûte une demande traitée par l'assistant d'IA, appel par appel, relecture comprise. Chaque tâche reçoit le modèle le moins cher qui passe le jeu d'essai, sous un plafond mensuel.

AUSSI APPELÉ coût par requête, coût par token, FinOps de l'IA, routage de modèles **PRINCIPE QU'IL APPLIQUE** la pensée systémique, fiche P-05

SE PRÉPARE EN une semaine, le journal des appels par demande **SE FAIT EN** 2 semaines sur 200 demandes, puis une revue par mois **LEVIER PRINCIPAL** Coût

CE QUE ÇA RAPPORTE	
Impact sur le TTM	Faible
Impact sur la qualité (FTR)	Faible
Impact sur l'Opex / Capex	Moyen

CE QUE ÇA COÛTE	
Coût d'implémentation	Faible
Difficulté d'implémentation	Moyen
Difficulté de maintien	Moyen

Faible · Moyen · Élevé · N/A (sans objet)

1 Le problème auquel ça s'applique

Et ce qu'il vous coûte chaque année

CE QUE VOUS VOYEZ DÉJÀ

- « La facture de l'outil d'IA a triplé, et personne ne sait pourquoi. »
- Le budget a été chiffré sur la démonstration. En production, chaque courriel déclenche cinq appels.
- Le gain se compte en minutes de conseiller, la dépense sur la facture. Personne ne les pose côte à côte.

OÙ PART L'ARGENT

La facture suit le texte lu et produit, le modèle et le nombre d'appels. Elle ne montre pas la relecture : un assistant moins cher qui fait tout reprendre coûte plus.

L'ORDRE DE GRANDEUR, À LA LOUCHE

Comptez-le sur vos chiffres. Un ordre de grandeur suffit pour décider d'essayer.

COMBIEN DE FOIS / AN		CE QUE ÇA COÛTE À CHAQUE FOIS		PART QU'ON ÉVITE		DÉPENSE ÉVITÉE / AN
demandes traitées par l'assistant	×	appels, plus minutes de relecture × coût d'une minute	×	appels inutiles et modèle trop cher pour la tâche	=	**à écrire au crayon, après le relevé**

QUAND ÇA NE SERT À RIEN

Pour quelques centaines de demandes par mois : le relevé coûte plus que l'écart. Suivez la relecture.

2 Comment ça marche

La mécanique, en clair

L'IDÉE EN UNE PHRASE

Une réponse coûte la somme de ses appels au modèle, plus les minutes de qui la relit. On la suit par type de demande, pas à la facture totale.

LES GESTES, DANS L'ORDRE

1. Relever les volumes par type de demande, et ce que l'assistant lit pour chacune.
2. Suivre 200 demandes réelles appel par appel, et noter le coût de chaque appel.
3. Chronométrer la relecture, au coût d'une minute disponible (fiche O-45).
4. Par tâche, garder le modèle le moins cher qui passe le jeu d'essai (fiche O-76), relecture comprise.
5. Réutiliser ce qui revient à chaque demande : la consigne, les conditions générales.
6. Poser un plafond mensuel par service, calé sur le mois de pointe (fiche O-34), une alerte à 80 %, et un nombre d'appels maximal par demande.

CE QU'IL FAUT AVOIR SOUS LA MAIN

Le journal des appels, demande par demande, extrait par l'équipe technique, et le coût d'une minute de conseiller.

LE SIGNE QUE ÇA A MARCHÉ

La facture monte de 30 % un mois. En revue, on voit en cinq minutes que c'est le volume, pas le coût par demande.

3 Les pièges de l'implémentation

Ce qui fait rater, et ce qu'on fait à la place

AU LANCEMENT — Chiffrer sur la démonstration
En production, le courriel arrive avec ses pièces jointes et l'historique du client.
→ Mesurer sur 200 demandes réelles.

AU LANCEMENT — Oublier le temps des personnes
La facture baisse, la relecture s'allonge, le coût par demande monte.
→ Compter la relecture dans le même calcul.

PENDANT — Le même modèle pour tout
Le plus capable classe aussi les demandes d'attestation, au prix fort.
→ Un modèle par tâche, le moins cher qui passe le jeu d'essai.

PENDANT — Changer de modèle sans repasser le jeu d'essai
Le moins cher rédige moins bien, et l'économie part en relecture.
→ Repasser le jeu d'essai, et chronométrer la relecture.

DANS LA DURÉE — Regarder la facture totale
Elle monte avec le volume, et cache une consigne allongée ou une relance en boucle.
→ Suivre le coût par demande chaque mois, sous un plafond.

4 Exemple

Cas type du tertiaire, du constat au résultat

LA SITUATION DE DÉPART

L'OBJET
Les courriels du service clients d'un fournisseur d'énergie. L'assistant classe et propose la réponse, le conseiller la relit et l'envoie.

CE QUI CLOCHE
La facture du modèle est passée de 3,0 à 9 k€ par mois depuis le lancement, pour 14 % de courriels en plus.

CE QU'ILS REGARDAIENT AVANT
La facture totale, comparée au budget du projet. Personne ne la rapportait aux courriels traités.

LES CHIFFRES DE DÉPART, EN SEPTEMBRE

20 000	courriels par mois, moitié de demandes simples
0,45 €	de modèle par courriel, 9 k€ par mois (hypothèse)
3,5 min	de relecture par réponse proposée
0,80 €	la minute de conseiller disponible, coût complet

CE QU'ILS ONT FAIT

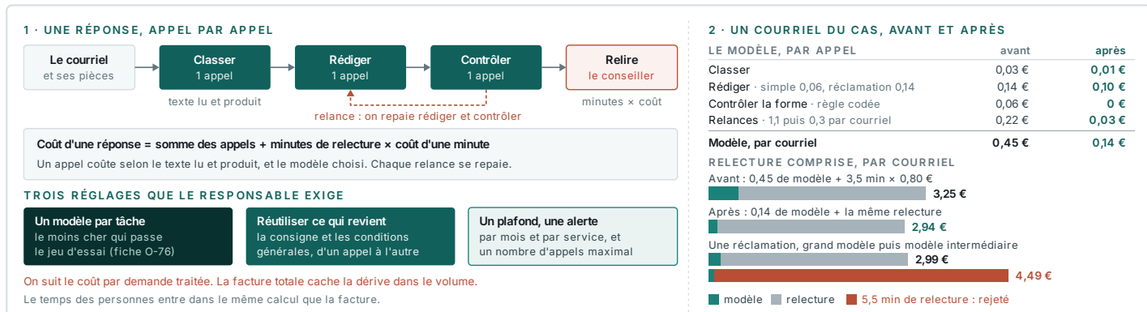
Deux semaines de relevé sur 200 courriels réels, appel par appel et relecture chronométrée, puis un modèle par tâche.

LE DÉCLENCHEUR Le contrôle de gestion demande pourquoi la facture a triplé. Personne ne sait répondre. ▶
LE GESTE 200 courriels suivis appel par appel, et la relecture de chacun chronométrée. ▶
CE QU'ILS ONT VU 5,2 appels par courriel, tous au modèle le plus cher : un contrôle ajouté en juin, et 1,1 relance. ▶
LA DÉCISION Un modèle par tâche, le contrôle de forme en règle codée, un plafond de 3,8 k€ : 27 000 courriels en janvier, le mois de pointe, × 0,14 €.

CE QUE ÇA A DONNÉ

0,45 → 0,14 €	Le modèle par courriel : 0,03 + 0,14 + 0,06 + 1,1 relance × 0,20 = 0,45 € avant, 0,01 + 0,10 de rédaction (moitié à 0,06, moitié à 0,14) + 0,3 × 0,10 = 0,14 € après.
9 → 2,8 k€	La facture du mois : 20 000 courriels × 0,14 €, soit 6,2 k€ de moins par mois et 74 k€ par an.
3,25 → 2,94 €	Une réponse, relecture comprise : 0,45 + 3,5 min × 0,80 € = 3,25 € avant, 0,14 + 2,80 = 2,94 € après. La relecture en fait 95 %.
+1,50 €	Le modèle intermédiaire sur les réclamations : 0,10 € de modèle en moins, 2 min de relecture × 0,80 € = 1,60 € en plus. Rejeté.

Refaites-le avec vos demandes : sur 200 cas réels, le coût de chaque appel plus les minutes de relecture × le coût d'une minute. Posez ce coût par type de demande à côté de la facture du mois.



À gauche le circuit d'une réponse et ce que chaque étape coûte, puis les trois réglages. À droite le coût de modèle d'un courriel du cas, et le coût relecture comprise.

L'exemple est un cas type, reconstruit à partir de situations courantes du tertiaire. Aucune situation client n'y est identifiable.