flowlysis.

CELUI QUI PRODUIT NE S'ÉVALUE PAS LUI-MÊME

Ce qu'un assistant d'IA produit est contrôlé par un autre regard, sur des critères écrits avant de voir le résultat. L'outil qui relit sa propre réponse la trouve presque toujours bonne.

AUSSI APPELÉ contrôle indépendant, règle des quatre yeux ; LLM as a judge pour sa version automatique

S'APPLIQUE À toute réponse ou tout résumé qu'un assistant produit pour un service

SE VÉRIFIE EN deux jours, en faisant noter cent réponses par deux personnes du métier **OUTIL QUI L'APPLIQUE** le jeu d'essai d'un assistant, fiche O-76 (à venir)

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

CE QUE ÇA EXIGE	
Ce que ça invalide dans la pratique actuelle	Moyen
Délai avant effet visible	Faible
Difficulté à tenir dans la durée	Élevé

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

1 Ce que dit le principe

Et pourquoi c'est vrai

LA RÈGLE

Ce que produit un assistant passe devant un autre regard : une personne, une règle codée, ou un contrôle automatique conduit à part, avec sa propre consigne et la grille sous les yeux. La grille s'écrit avant de voir les réponses. Un contrôle automatique ne compte qu'après avoir été comparé aux notes de personnes du métier.

POURQUOI ÇA MARCHE

Le mécanisme. Une chaîne de causes que le lecteur peut refaire dans sa tête.

SI LE CONTRÔLE SE FAIT À PART l'outil ne note plus sa propre réponse, qu'il trouve presque toujours bonne. ▶ **ALORS** chaque réponse est jugée sur les mêmes points, écrits avant. ▶ **DONC** comparé à des notes humaines, le contrôle montre ses penchants et combien de réponses fausses passent.

L'IDÉE REÇUE QUE ÇA RENVERSE

Écrite telle qu'on l'entend vraiment en comité : c'est une croyance raisonnable qui a cessé d'être vraie.

"Il suffit de lui demander de vérifier : on lui a demandé s'il était sûr, il a dit oui."

OÙ ÇA NE S'APPLIQUE PAS

Pour un brouillon que le gestionnaire réécrit en entier avant l'envoi : son regard est déjà le contrôle à part. Le principe compte dès qu'une réponse part sans être réécrite.

2 Est-ce respecté chez vous ?

Des tests qu'on fait en marchant

Cinq vérifications observables. Chacune dit où aller regarder, et à quoi on voit que ce n'est pas respecté.

LE PLUS RAPIDE **Demandez qui a contrôlé la dernière réponse envoyée, et sur quoi.**
Un autre regard que l'assistant, avec une grille écrite.
× « L'assistant l'a vérifiée », ou « le gestionnaire l'a lue ».

Demandez la grille du contrôle.
Quatre ou cinq critères, datés, écrits avant la mise en service.
× Une consigne qui demande si la réponse est correcte.

Demandez ce que contrôlent les règles codées.
Le montant, les articles cités, les mentions obligatoires (fiche O-65).
× Rien : tout passe par le contrôle automatique, ou par une personne.

Demandez à quelles notes le contrôle a été comparé.
Cent réponses notées par deux gestionnaires, pas par l'équipe de l'assistant.
× À aucune : on lit son score chaque mois.

DANS LA DURÉE **Demandez la date de la dernière comparaison.**
Le mois dernier, et après chaque changement de consigne ou de modèle.
× À la mise en service, puis plus jamais.

3 Les confusions courantes

Ce que le principe n'est pas

Quatre faux amis. Ce sont eux qui font croire qu'on applique le principe en faisant l'inverse.

Ce n'est pas tout relire à la main
Les deux mettent une personne derrière l'assistant.
→ Tout relire en détail coûte presque autant que tout écrire, et l'attention s'use. La relecture attentive va à l'échantillon et aux réponses arrêtées.

Ce n'est pas un second outil qui « vérifie »
Les deux confient le contrôle à une machine.
→ Sans grille écrite avant, il juge à l'allure et préfère la réponse longue, même fausse.

Ce n'est pas une grille jamais confrontée
Les deux ont des critères écrits.
→ Sans comparaison aux notes de personnes du métier, son score ne dit rien. Un 95 % peut venir d'un contrôle indulgent.

Ce n'est pas interdire à l'outil de se relire
Les deux font revoir sa réponse à l'assistant.
→ Se relire corrige parfois une erreur grossière : c'est un meilleur brouillon, pas un contrôle.

4 D'où ça vient, et un exemple

L'origine, puis le principe à l'œuvre

ORIGINE

La règle des quatre yeux, en banque et en audit : qui fait ne contrôle pas. Un modèle juge mal son propre travail.

IMAGINONS… Une mutuelle reçoit 1 400 réclamations par mois. L'assistant propose la réponse, le gestionnaire la lit vite.

SANS CONTRÔLE À PART Prié de vérifier, l'assistant juge bonnes 97 réponses sur 100. Les adhérents réécrivent.

AVEC CONTRÔLE À PART Deux gestionnaires en notent 100 sur une grille écrite avant : 18 fausses, dont 16 validées par l'assistant. Règles codées et contrôle à part en arrêtent 16 : 28 partent par mois, contre 224 (1 400 × 2 % et × 16 %).

POUR ALLER PLUS LOIN

fiche O-76, Le jeu d'essai d'un assistant (à venir)
fiche O-65, Le détrompeur
fiche O-29, Le coût d'obtention de la qualité

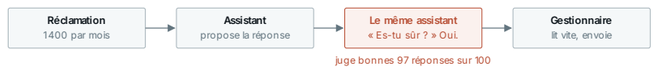
LE MOT JUSTE

Dire contrôler, et non vérifier. On vérifie son travail ; on contrôle celui d'un autre, sur une grille.

CE QUE ÇA CHANGE, CONCRÈTEMENT

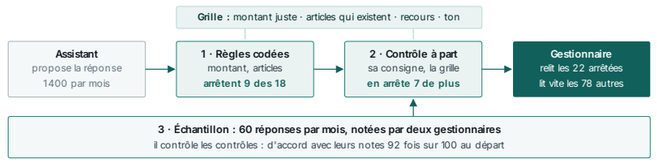
Les réponses aux réclamations d'une mutuelle, avant et après.

AVANT · L'ASSISTANT RELIT SA PROPRE RÉPONSE



100 réponses notées par deux gestionnaires : 18 fausses, dont 16 validées par l'assistant

APRÈS · UN AUTRE REGARD, UNE GRILLE ÉCRITE AVANT



Les trois niveaux, du moins cher au plus cher : règles codées, contrôle à part, personnes

Sur 100 réponses : 16 des 18 fausses arrêtées avant l'envoi · 2 passent, contre 16 avant

MÊMES 1 400 RÉCLAMATIONS PAR MOIS	AVANT	APRÈS
Réponses fausses envoyées, par mois	224	28
Temps de lecture des gestionnaires	49 h	83 h
Temps passé sur les réclamations qui reviennent	149 h	19 h
Total des deux, par mois	199 h	102 h

Taux sur 100 réponses notées : 2 sur 100 peut valoir de 0 à 7. 1400 × 16 % = 224 ; × 2 % = 28.
Avant : 1358 × 2 + 42 signalées × 6 ≈ 49 h. Après : 1092 × 2 + 308 arrêtées × 6 + 60 × 2 × 8 ≈ 83 h.
Hypothèses : lecture 2 min, relecture 6 min, notation 8 min, réclamation qui revient 40 min.

QUAND LE PRINCIPE EST TENU

QUI CONTRÔLE
règles codées, contrôle à part, échantillon relu.

LA GRILLE
quatre ou cinq critères, écrits avant, datés.

LE CONTRÔLE AUTOMATIQUE
comparé chaque mois à des notes humaines.

CE QU'ON REGARDE
les réponses fausses qui passent, sur 100.

QUAND IL NE L'EST PAS

QUI CONTRÔLE
l'assistant, à qui l'on demande s'il est sûr.

LA GRILLE
« la réponse est-elle correcte ? »

LE CONTRÔLE AUTOMATIQUE
un score lu chaque mois, jamais confronté.

CE QU'ON REGARDE
la part que l'assistant juge bonne.

Ce qu'il faut accepter de perdre

Le prix du principe, dit franchement

CE QU'ON ABANDONNE

Le confort d'un assistant qui se dit sûr de lui, et d'un score qu'on lit sans savoir d'où il vient. Deux gestionnaires passent deux jours à noter cent réponses, puis une journée chacun par mois.

L'INDICATEUR QUI VA SE DÉGRADER

Le temps de lecture des gestionnaires, de 49 à 83 h par mois dans l'exemple. Il se reprend sur les réclamations qui reviennent : de 149 à 19 h. Chaque contrôle automatique est aussi un appel de plus, à compter dans le coût d'une réponse (fiche O-79).

CE QUI CASSE SI ON L'APPLIQUE À MOITIÉ

Un contrôle à part qui reprend la consigne de l'assistant, ou une grille jamais comparée aux notes humaines. Le score monte, les réponses fausses passent, et ce sont les adhérents qui les trouvent.

LA PHRASE À RETENIR

Qui produit ne se note pas. Et le contrôle se contrôle.