

Nos workflows agentiques s'améliorent-ils vraiment ?

D'une vague impression à une mesure fiable.

Vous ajoutez un skill...

Vous faites relire le travail d'un agent par un autre...

Vous changez de modèle...

Et vous trouvez que ça s'améliore !

Mais comment en être sûr ?



Jean-Louis Rigau

Vibe coder & AI-native advisor @ Manufacture

J'aide les entreprises à construire des organisations agentiques, de l'outillage au delivery, en mesurant ce que les agents changent vraiment.



LinkedIn

04 On ne sait pas prouver, alors on mesure

2003 · L'IDÉAL THÉORIQUE

Machine de Gödel

Jürgen Schmidhuber imagine une machine qui réécrit son propre code, mais seulement si elle **prouve** que c'est mieux.



AUJOURD'HUI · EN PRATIQUE

Une approche empirique

Prouver, on ne sait pas. Alors on **mesure** chaque changement sur de vrais essais, et on ne garde que ce qui améliore vraiment le résultat.

Une mesure ne prouve rien... mais elle remplace les impressions par des faits.

05 Les benchmarks mesurent les modèles, pas nos workflows

○	○	○	○	○	●
2021	2023	2024	2025	2025	2026
HumanEval	SWE-bench	AI Agents That Matter	SWE-bench Pro	Étude METR	Effet scaffold
Du code qui passe des tests.	De vraies issues, dans des dépôts publics.	Mesurer aussi le coût et la reproductibilité.	Des tâches longues, sur 41 dépôts actifs.	Des devs 19 % plus lents avec l'IA, qui se croyaient 20 % plus rapides.	Le harnais, une variable cachée des classements.

Les benchmarks le montrent : le workflow compte. Mais pour savoir si le nôtre s'améliore, il faut le mesurer sur nos propres tâches.

PARTIE 1

La mesure

préparer

Une application, le choix d'un use case, des critères d'acceptation.

mesurer

S'outiller, observer et comparer chaque run.

évaluer

Se poser les bonnes questions.

07 Une application et un use case comme terrain de jeu

Agreement Hub

+ New Contract

Contracts

19 total

Title	Customer	Status	Created	
AH02 contract a1ac959d-12f6-49be-83bd-3508dfb45d6	Synthetic verifier customer	Draft	10/5/2026	View Edit
AH02 contract 65e23219-2253-470c-8003-f571e1c1c0a4	Synthetic verifier customer	Draft	10/5/2026	View Edit
AH02 contract cac61035-bdbe-4030-b3cc-2dac2da6bfa0	Synthetic verifier customer	Draft	10/5/2026	View Edit
AH02 contract 12578876-8b93-486b-b620-dcef829a6529	Synthetic verifier customer	Draft	10/5/2026	View Edit
AH02 contract 13406108-fef1-4266-8f0a-93bf02580440	Synthetic verifier customer	Draft	10/5/2026	View Edit
AH02 contract dd2e1ca5-3dc3-460b-8ec4-c012025e6a70	Synthetic verifier customer	Draft	10/5/2026	View Edit
AH02 contract 2c60dd83-b435-4ed8-a671-2cf09ebaa949	Synthetic verifier customer	Draft	10/5/2026	View Edit

Agreement Hub au départ · React, Express, SQLite, 24 tests (TU + Playwright)

LE USE CASE · AH02

Une bibliothèque de clauses : les créer, les archiver, les insérer dans un contrat avec leur provenance.

EXIGENCE CLÉ

Aucun doublon, même en cas de retry.

19 CRITÈRES D'ACCEPTATION

- L0 1 le CRUD des contrats existant reste intact
- L1 8 créer, lister, éditer une clause, ses révisions
- L2 3 archiver, restaurer, filtrer
- L3 7 prévisualiser, insérer avec provenance, retry sans doublon

Ce que « réussir » veut dire est fixé avant le premier run, et vérifié comme un utilisateur.

08 Un skill et un workflow comme point de départ

LE WORKFLOW DE DÉPART



4 CONFIGURATIONS COMPARÉES

C0 agent seul	CB + skill
CR + review	CBR les deux

PUIS ITÉRER SUR LE SKILL

Un skill qui aide l'agent à transformer l'énoncé en vérifications observables.

skill v1

Une matrice de vérification

→ aucun gain

skill v2

Des contre-exemples ciblés

→ retry testé, aucun gain

skill v3

Exécuter, pas cocher

→ testé seulement en pilote

Écrire, mesurer, corriger, recommencer : un skill se construit par itérations.

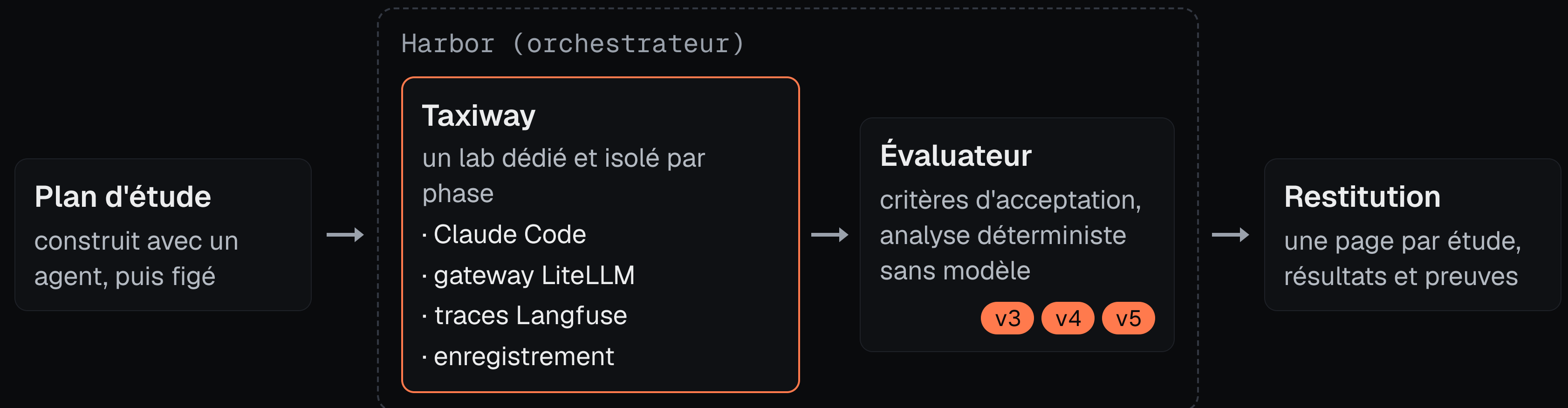
09 La chaîne de mesure, du plan au résultat

étude r07a le skill et la review font-ils mieux que l'agent seul ?

3 × 4 modèles × workflows

× 5 répétitions

60 résultats notés



► VIDÉO · Une étude : le plan, un run, son verdict, son enregistrement

awawe / ah02-r07a-matrix-20261007

Langfuse

Inspect

Harbor

AH02-R07A-MATRIX

PROTOCOLE F9D4BACF3DAB

AH02

3 MODÈLES × 4 CONDITIONS × 5 RÉP.

BUDGETS · AUTEUR 30 MIN · REVUE 10 MIN · CORRECTION 10 MIN

SKILL V2 · ÉVALUATEUR V5 (34 ÉCARTS AVEC V4)

PLAN INTÈGRE

ENTRÉES GELÉES VÉRIFIÉES

ÉVALUATEUR

v4

v5 · par défaut

ÉTUDE

ah02-r07a-matrix-20261007

exploratoire

CALIBRATION

5/5 critères OK avant r07 · ah02-cb-skill-v2-pilot-20261007 · ah02-r07-qualific

20261007 · calibration jamais agréée avec cette étude.

AH02 · 3 modèles (Opus 4.8, Opus 5, Opus 5.5) × 4 conditions (C0, CB, CR, CBR) × 5 répétitions.
Chaque origine produit un auteur indépendant (C0 ou CB).
CR et CBR revoient puis corrigent ce même candidat.

C0

auteur seul, sans workflow

CB

auteur avec le skill

CR

auteur C0, puis revue et correction

CBR

auteur CB, puis revue et correction

Mise à jour 2026-10-10 01:11:57 · rechargement automatique toutes les 60 s

FIN DU RUN

Run terminé à 10:19 : 75 terminées, 15 sautées.

11 Presque pas, et ce n'est pas gratuit

≈ 0

critère gagné avec le skill, v1 comme v2

0 à 2 / 5

essais réparés par la review, sans régression

+2,7 à 3,7 \$

par essai pour la review, chez Opus 4.8 et Opus 5

		C0 agent seul	CB + skill	CR + review	CBR les deux
Opus 4.8		0/5 5,65 \$	0/5 5,95 \$	0/5 8,38 \$	0/5 8,69 \$
Opus 5		0/5 7,26 \$	0/5 7,84 \$	2/5 10,62 \$	2/5 11,56 \$
Opus 5.5		4/5 2,60 \$	2/5 2,92 \$	4/5 3,11 \$	2/5 3,47 \$

étude r07a

évaluateur v5. Essais réussis sur 5, puis coût moyen par essai, revue et correction comprises.

12 Le skill change les tests, la review ne voit pas ses angles morts

LE SKILL

12/15

essais avec skill testent le
retry après une
modification (0/15 sans)

0/3

réussissent quand le test
attend la mauvaise valeur

Il fait tester le bon cas. Mais si le résultat attendu est faux, le test valide le bug.

LA REVIEW

12/16

bugs de retry signalés

Un défaut qu'on montre au relecteur est presque toujours réparé. Un défaut qu'il doit trouver seul, 1 fois sur 9.

0/22

historiques bloqués
signalés quand le modèle
relit son propre code

13 Superpowers : installé ne veut pas dire suivi

SP · installé tel quel

L'agent choisit de s'en servir. Opus 4.8 n'invoque aucun skill dans 2 essais sur 10.

SPS · chaîne complète demandée

Seul Opus 5.5 la suit en entier, 7 fois sur 10.

étude r08 un essai par ligne

MODÈLE	CONDITION	VERDICT	ÉTAPES DE LA CHAÎNE
claude-opus-5 high	SPS	échoué 17 critères timeout	BR › WT › WP › EX › VER · exécution par sous-agents manque CR FIN CR via revue finale SDD TDD : consigne aux implémenteurs
claude-opus-5-5 high	SPS	réussi tous les critères	BR › WT › WP › EX › TDD › CR › RCR › VER › FIN · exécution par sous-agents chaîne complète

14 Superpowers ne remplace pas le modèle

étude r08 essais qui passent tous les critères, coût moyen par essai · 90 essais · plan d'analyse écrit avant les résultats

	sans workflow	SP · tel quel	SPS · chaîne complète
Opus 5.5	7/10 · 2,9 \$	8/10 · 5,5 \$	10/10 · 13,5 \$
Opus 5	0/10 · 7,7 \$	0/10 · 23,2 \$	0/10 · 22,9 \$ · 10 timeouts
Opus 4.8	0/10 · 5,3 \$	0/10 · 11,4 \$	0/10 · 26,7 \$

Aucun gain détecté sur le taux de réussite, pour 2 à 5 fois le coût. Un critère, le retry qui renvoie la réponse d'origine, bloque Opus 4.8 et Opus 5 dans les 60 essais.

► VIDÉO · Superpowers : la chaîne de skills, essai par essai

awawe / ah02-r08-superpowers-20261008-run2

Langfuse

Inspect

Harbor

R08-AH02-SUPERPOWERS

PROTOCOLE BB437526063C

AH02

3 MODÈLES × 3 CONDITIONS × 10 RÉP.

BUDGETS · AUTEUR 2 H

SUPERPOWERS 6.4.1 · ÉVALUATEUR V5

PLAN INTÈGRE

ENTRÉES GELÉES VÉRIFIÉES

ÉVALUATEUR

v5 · par défaut

ÉTUDE

ah02-r08-superpowers-20261008-run2

pré-enregistrée

AH02 · 3 modèles (Opus 4.8, Opus 5, Opus 5.5) × 3 conditions (C0, SP, SPS) × 10 répétitions.
Auteurs seuls.
Notre plateforme ne fait ni revue ni correction.
Utilisation de Superpowers 6.4.1.

C0

auteur seul, sans workflow

SP

auteur avec Superpowers, tel quel

SPS

auteur avec Superpowers, chaîne complète demandée

Mise à jour 2026-10-10 01:14:47 · rechargement automatique toutes les 60 s

FIN DU RUN

Run terminé à 14:43 : 90 terminées.

Labs de l'étude supprimés.

QUESTION 4 · ET SI JE CHANGE DE USE CASE ?

16 Un nouveau use case, tout est à re-mesurer

AH03 circuit d'approbation des contrats (même application, évaluateur dédié)

	réussites	ce qui bloque
Opus 5.5	4 sur 5	réponse perdue, 1 fois
Opus 4.8	2 sur 3	soumission, double clic
Opus 5	0 sur 3	réponse perdue, nouvelle soumission

Sur AH02 : Opus 4.8 et Opus 5 à égalité, 0/10.
Sur AH03 : Opus 4.8 passe devant Opus 5.

Le point dur reste le même : les chemins d'erreur. Mais les résultats d'un use case ne se transposent pas à un autre. Il faut re-mesurer, avec un évaluateur écrit pour ce use case.

La difficulté à mesurer

construire

L'énoncé, une implémentation de référence, puis éprouver l'évaluateur.

se tromper

Un code conforme refusé, une faute acceptée.

s'améliorer

Sans regarder les scores. Chaque erreur ramène au début.

18 Avant de noter, on éprouve son évaluateur

AH02 évaluateurs v4 et v5

CONCEVOIR

Un énoncé public, une implémentation de référence qui le respecte, et un évaluateur externe. Chaque contrôle cite sa phrase de l'énoncé.

ÉPROUVER

Calibration

26

contrôles de variantes correctes et de fautes

Mutation

36/36

mutants du code de référence détectés

Autres modèles

127

trouvailles en 9 sessions

Aucune méthode ne suffit seule : chacune couvre les angles morts des autres.

19 L'évaluateur se trompe, dans les deux sens

IL REFUSE UN CODE CONFORME

Clause library / New clause

New clause

Clause title *

e.g. Limitation of Liability

Clause text *

« Clause title * » refusé à cause de l'astérisque : 36 échecs comptés à tort.

IL LAISSE PASSER UNE FAUTE

Inserted "AH02 UI history after load error 40efef52..." (revision 1).

Add clause

CLAUSE INSERTIONS

Failed to load clause insertions: Temporary history failure

L'historique reste bloqué après une erreur de chargement. Invisible jusqu'à ce que d'autres modèles cherchent les failles de l'évaluateur.

Un score ne dit pas qui s'est trompé : l'agent ou l'évaluateur.

20 Améliorer l'évaluateur, sans toucher aux résultats

plan figé

énoncé, prompts, skill et évaluateur, empreinte vérifiée

verdict caché

les agents ne voient jamais l'évaluateur

rien d'écrasé

chaque version de l'évaluateur garde ses verdicts

une seule étude

jamais deux protocoles mélangés

pannes écartées

rejouées sans regarder le score

```
● Agent "Implement Task 5: insertion UI" failed: Agent terminated early due to an API error: API Error: Server is temporarily limiting requests (not your usage limit) .  
litellm.RateLimitError: AnthropicException -  
{"type":"error","error":{"type":"rate_limit_error","message":"This request would exceed your account's rate limit. Please try again later."},"request_id":"req_011Cfqii95tiJUwSuJV54aFK"}. Received Model  
Group=claude-sonnet-5-5  
Available Model Group Fallbacks=None (error type rate_limit, HTTP 429, model sent to the API: claude-sonnet-5-5)
```

Une panne d'API ressemble à un agent qui échoue.

Un score ne se négocie pas : ni l'agent mesuré, ni celui qui mesure ne doit pouvoir l'arranger.

21 Changer l'évaluateur, c'est changer la conclusion

TROIS VERSIONS DE L'ÉVALUATEUR

évaluateur v3 Il laissait passer des fautes	évaluateur v4 Il refusait un code conforme	évaluateur v5 Il voit l'historique bloqué
--	---	--

étude r07a MÊMES CANDIDATS, RIEN DE RELANCÉ · LA CONCLUSION BOUGE

	évaluateur v4	évaluateur v5
Opus 5.5 sans skill : essais qui passent tous les critères	5/5	4/5
Opus 5.5 avec skill : essais qui passent tous les critères	5/5	2/5
verdict sur le skill	neutre	défavorable

Le v5 voit l'historique bloqué, plus fréquent avec le skill.

Un résultat n'est fiable que si l'évaluateur l'est.

► VIDÉO · La même étude notée par l'évaluateur v4, puis v5

awawe / ah02-r07a-matrix-20261007

Langfuse

Inspect

Harbor

ah02-r07a-matrix

PROTOCOLE F9D4BACF3DAB

ah02

3 MODÈLES × 4 CONDITIONS × 5 RÉP.

BUDGETS · AUTEUR 30 MIN · REVUE 10 MIN · CORRECTION 10 MIN

SKILL V2 · ÉVALUATEUR V4

PLAN INTÈGRE

ENTRÉES GELÉES VÉRIFIÉES

ÉVALUATEUR

v4

v5 · par défaut

ÉTUDE

ah02-r07a-matrix-20261007

exploratoire

VERSION

Verdicts de l'évaluateur v4.
La version par défaut de cette étude est v5.
Les versions ne se mélangent jamais.

CALIBRATION

5/5 critères OK avant r07 · ah02-cb-skill-v2-pilot-20261007 · ah02-r07-qualification-20261007 · calibration jamais agrégée avec cette étude.

ah02 · 3 modèles (Opus 4.8, Opus 5, Opus 5.5) × 4 conditions (C0, CB, CR, CBR) × 5 répétitions.
Chaque origine produit un auteur indépendant (C0 ou CB).
CR et CBR revoient puis corrigent ce même candidat.

C0

auteur seul, sans workflow

CB

auteur avec le skill

CR

auteur C0, puis revue et correction

CBR

auteur CB, puis revue et correction

À RETENIR

Ce que ça change pour vos workflows

Mesurez sur vos tâches

Le benchmark ne dit rien de votre workflow.

Un skill doit gagner sa place

Il peut coûter sans rien apporter.

Re-mesurez à chaque changement

Nouveau modèle, nouveau use case : ce qui aidait peut nuire.

Évaluez votre évaluateur

Il se trompe aussi, dans les deux sens.

ET DEMAIN

Cette boucle, mesurer puis garder ce qui fait ses preuves, des agents la feront seuls. Il leur faudra le même évaluateur éprouvé et les mêmes garde-fous.

Clôture

MERCI

Des questions ?

awawe / ah02-r07a-matrix-20261007

LangfuseInspectHarbor

AH02-R07A-MATRIX

PROTOCOLE F9D4BACF3DAB

AH02

3 MODÈLES × 4 CONDITIONS × 5 RÉP.

BUDGETS · AUTEUR 30 MIN · REVUE 10 MIN · CORRECTION 10 MIN

SKILL V2 · ÉVALUATEUR V5 (34 ÉCARTS AVEC V4)

PLAN INTÈGRE

ENTRÉES GELÉES VÉRIFIÉES

ÉVALUATEUR

v4

v5 · par défaut

Suivi de l'étude

CALIBRATION

5/5 critères OK avant r07 · ah02-cb-skill-v2-pilot-20261007 · ah02-r07-qualification-20261007 · calibration jamais agrégée avec cette étude.

AH02 modèle × skill × review.
Chaque origine produit un auteur indépendant (CO ou CB).
CR et CBR revoient puis corrigent ce même candidat.
Page reconstruite à partir des fichiers de l'étude, sans jamais les modifier.

Mise à jour 2026-10-08 16:21:20 · rechargement automatique toutes les 60 s

PHASES TERMINÉES	EN COURS	ENDPOINTS NOTÉS	TEMPS MODÈLE	COÛT CAMPAGNE
90/90	0	60/60	11 h 59 min	\$229.16
dont 2 timeouts	à réconcilier si le lanceur est	verdict de l'évaluateur externe	sur une enveloppe de 25 h	coût de réponse LiteLLM ·

Prochainement inclus dans Taxiway



taxiway.sh