

Jev en comptabilité : ce que notre test de 200 opérations révèle sur le contexte

Pierre Coquard · HOLCO · 8 octobre 2026

<https://holco.co/ressources/jev-comptabilite-test-contexte-resultats/>

Dans notre pilote, Jev mérite d'être testé pour proposer un classement comptable à faible coût. Avec six exemples historiques, il retrouve 126 comptes sur 200. Le contexte enrichi réduit ses désaccords, mais augmente ses abstentions. Ces résultats justifient un pilote avec revue humaine ; ils ne valident pas une comptabilisation automatique.

J'ai voulu partir d'une question concrète : Jev, le modèle de TypeSafe, peut-il nous être utile en finance ? L'étude de Digits nous a donné un point de départ. Pour décider, il fallait ensuite regarder ce qui se passe sur un dossier français, avec notre façon de préparer le contexte.

Nous avons donc comparé Jev, OpenAI et Claude sur les mêmes 200 opérations d'une entreprise française. Chaque modèle a reçu trois versions du contexte. Cette campagne, réalisée le 8 octobre 2026, produit 1 800 résultats comparatifs, mais porte bien sur 200 cas distincts.

Le résultat le plus instructif tient à un changement de comportement : en ajoutant un profil historique structuré, les trois modèles proposent moins de comptes. Leur accord avec le FEC augmente parmi les propositions conservées, tandis que le nombre total de comptes retrouvés diminue.

Ce que nous reprenons de Digits

Le benchmark de Digits porte sur 2 000 transactions de quatre entreprises américaines, avec des catégories revues par des comptables appliquant les principes US GAAP. Il distingue notamment les modèles travaillant seuls, ceux pouvant consulter un historique et le système comptable de l'éditeur.

Nous reprenons cette question : que change la connaissance du dossier ? Notre étude est une adaptation locale, pas une réplique à l'identique. Les données, les modèles, le plan de comptes et la référence de notation diffèrent. Nous n'avons pas testé Digits AGL et nous ne comparons pas nos pourcentages aux siens pour établir un classement.

Un protocole borné et des entrées communes

Nous utilisons un FEC 2025 pour reconstruire un test rétrospectif. Seules les écritures bancaires simples, équilibrées, avec une ligne bancaire et un compte de contrepartie unique sont éligibles. Les écritures complexes sont exclues. Les 200 cas sont sélectionnés de façon reproductible parmi les opérations éligibles du second semestre, indépendamment du compte attendu.

Les trois modèles reçoivent les mêmes observations, le même contexte métier et le même ensemble de comptes candidats. La sélection se fait en deux étapes identiques : famille, puis compte. Cette hiérarchie permet notamment de respecter la limite du nombre de choix de TypeSafe. L'abstention est autorisée aux deux étapes ; le compte attendu ne sert jamais à corriger une décision intermédiaire.

Jev est appelé via l'API officielle TypeSafe, OpenAI via Responses et Claude via son outil en ligne de commande, sans outils métier externes. Les interfaces techniques restent donc différentes. Les réglages de raisonnement d'OpenAI et de Claude sont faibles. Les versions exactes sont indiquées ci-dessous.

Tableau 1 — Périmètre du pilote

Paramètre	Choix retenu
Dossier	Une entreprise française ; identité et écritures non publiées
Cas évalués	200 opérations de juillet à décembre 2025, parmi 1 670 éligibles
Historique disponible	2 152 opérations éligibles de janvier à juin 2025
Comptes candidats	285 comptes ; sélection hiérarchique famille puis compte
Modèles observés	Jev 1.13.0 ; OpenAI gpt-5.6-terra ; Claude opus 4.8
Référence de notation	Compte de contrepartie observé dans le FEC
Campagne	8 octobre 2026 ; trois modèles × trois contextes × 200 cas

Ces 1 800 résultats ne constituent pas 1 800 observations indépendantes. Un seul dossier et une seule campagne sont étudiés.

Trois niveaux de contexte, avec une protection qui change l'épreuve

Première condition : l'opération et le plan de comptes, sans historique. Deuxième condition : les mêmes éléments, plus six écritures du premier semestre sélectionnées par une règle fixe combinant motif, sens bancaire, recouvrement des mots-clés et récence. Troisième condition : ces six exemples, plus un profil structuré des antécédents partageant le même motif et le même sens bancaire : effectifs, répartition des comptes et dates.

Les noms, références et libellés bruts ne sont pas transmis aux modèles. Nous les remplaçons par des identifiants pseudonymisés stables et une sélection de mots-clés comptables ; dates et montants sont conservés. Ce masquage protège les informations, mais retire aussi beaucoup de sens. Le résultat sans historique mesure donc une épreuve volontairement appauvrie, pas la capacité à lire une facture complète.

L'enrichissement ajoute à la fois des informations et une présentation structurée. Il ne permet pas d'isoler le seul effet de la longueur du contexte. Les fréquences sont explicitement présentées comme des observations historiques, pas comme des règles validées. Sur 200 cas, 111 possèdent un antécédent de même motif et de même sens ; 89 n'en ont aucun. Parmi les 111, 35 ont été associés à plusieurs comptes dans l'historique.

Lire les résultats : accord, couverture et abstention

Nous appelons « accord » une proposition de compte identique au compte observé dans le FEC. Un désaccord n'est pas automatiquement une erreur comptable démontrée : aucune nouvelle revue expert n'a tranché chaque écriture. Inversement, retrouver le compte historique ne prouve pas que ce compte était correct.

Le taux d'accord global divise les comptes retrouvés par les 200 cas. La couverture mesure la part des cas pour lesquels un compte valide est proposé. L'accord parmi les propositions exclut les abstentions et les réponses invalides. Il doit toujours être lu avec la couverture : un système qui répond moins peut obtenir un meilleur taux sur les seules réponses conservées.

Contexte	Jev	OpenAI	Claude
Sans historique	0 / 200 · 0,0 %	7 / 200 · 3,5 %	8 / 200 · 4,0 %
Six exemples historiques	126 / 200 · 63,0 %	113 / 200 · 56,5 %	125 / 200 · 62,5 %
Historique et profil structuré	113 / 200 · 56,5 %	104 / 200 · 52,0 %	117 / 200 · 58,5 %

Chaque pourcentage a pour dénominateur les 200 cas, abstentions et réponses invalides comprises.

Avec six exemples, Jev retrouve 126 comptes sur 200

Jev propose 150 comptes : 126 correspondent au FEC et 24 en différent ; il s'abstient 50 fois. Sa couverture est de 75 % et son accord parmi les propositions de 84 %. Claude retrouve 125 comptes, avec 24 désaccords, 49 abstentions et deux réponses invalides. OpenAI retrouve 113 comptes, avec 12 désaccords et 75 abstentions.

Jev et Claude sont donc très proches sur cet échantillon. OpenAI propose moins de comptes et présente moins de désaccords. Je ne tirerais pas un classement général de ces effectifs : les politiques d'abstention font partie du résultat, et un compte retrouvé de plus ne démontre pas une supériorité durable.

Sans historique, Jev s'abstient 199 fois sur 200, OpenAI 188 fois et Claude 183 fois. Ce constat souligne combien les entrées masquées manquent d'information. Il serait trompeur d'en conclure que Jev serait incapable de classer des dépenses à partir de descriptions complètes.

Le contexte enrichi rend les propositions plus sélectives

Avec le profil historique, Jev passe de 24 à 13 désaccords, mais de 126 à 113 comptes retrouvés. Ses abstentions passent de 50 à 74. Son accord parmi les propositions progresse de 84 % à 89,7 %, alors que sa couverture recule de 75 % à 63 %.

En suivant les cas un par un, le mécanisme apparaît : 14 propositions auparavant en accord deviennent des abstentions, 11 désaccords deviennent des abstentions et une abstention devient un accord. L'amélioration du taux parmi les propositions vient donc surtout d'un retrait de décisions, dont certaines étaient pourtant conformes au FEC.

OpenAI et Claude présentent la même tendance globale : davantage d'abstentions, moins de comptes retrouvés, moins de désaccords. Le contexte enrichi n'est pas un gain uniforme. Il déplace le compromis entre proposer davantage et limiter les propositions contestables.

Tableau 3 — Détail avec le profil historique enrichi

Modèle	Accords	Désaccords	Abstentions	Couverture	Accord parmi les propositions
Jev	113	13	74	63,0 %	89,7 %
OpenAI	104	8	88	56,0 %	92,9 %
Claude	117	11	72	64,0 %	91,4 %

200 cas par modèle ; aucune réponse invalide dans cette condition. Pour Jev : $113 / 126 = 89,7 \%$ parmi les propositions, mais $113 / 200 = 56,5 \%$ sur tous les cas.

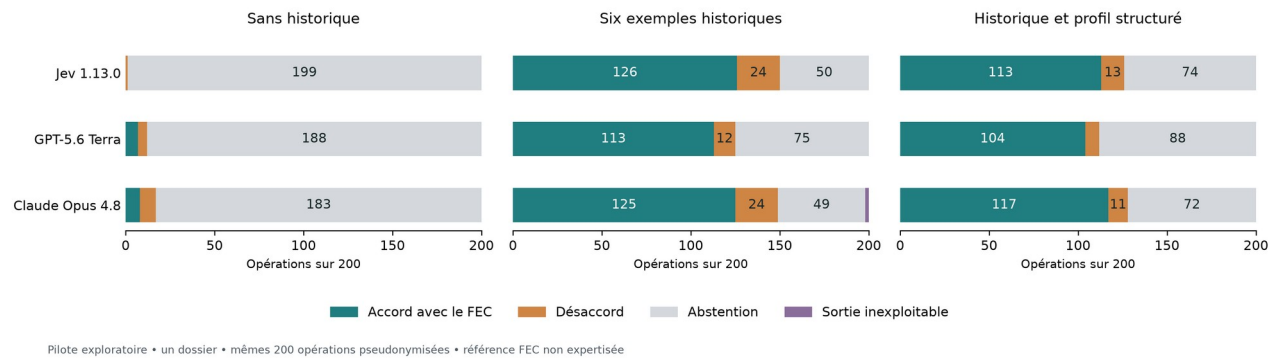


Figure HOLCO — Les effectifs complets, y compris les abstentions. Les taux parmi les seules propositions ne décrivent pas toute la performance.

Une règle simple constitue aussi un point de comparaison

Avant de payer un modèle, il faut regarder ce que ferait une règle élémentaire. Ici : reprendre le compte majoritaire dans l'historique de même motif et de même sens bancaire, ou s'abstenir si aucun précédent n'existe.

Cette règle propose un compte sur 111 cas et retrouve le compte du FEC dans 97 cas : 48,5 % d'accord global, 55,5 % de couverture et 87,4 % d'accord parmi les propositions. Elle laisse 89 cas sans proposition et produit 14 désaccords. Jev enrichi retrouve 16 comptes de plus au total, avec 15 propositions supplémentaires et un désaccord de moins. Cela justifie d'examiner où se situe l'apport du modèle, sans supposer qu'il doit traiter toutes les décisions.

Un coût d'inférence faible, un gain métier encore à mesurer

Le coût estimé de Jev est faible dans cette campagne. C'est un argument pour explorer des décisions répétées et bornées. Il ne suffit pas à démontrer une rentabilité : préparation des données, exploitation, contrôle humain et corrections ne sont pas inclus.

Les latences ci-dessous sont les médianes par opération terminée, appels successifs compris, qu'un compte ait été proposé ou que le modèle se soit abstenu. Elles ne mesurent pas le temps de traitement du lot entier. Claude a été exécuté par un outil en ligne de commande, avec son temps de démarrage et des consommations auxiliaires. La concurrence des appels et les reprises ont également varié. Il ne s'agit donc pas d'un benchmark pur des moteurs.

Tableau 4 — Ressources consommées avec le profil enrichi

Modèle	Appels enregistrés	Médiane par opération	Coût estimé pour 200 cas
Jev	331	600,5 ms	0,080 USD
OpenAI	323	2336,0 ms	2,541 USD
Claude	330	5175,0 ms	16,048 USD

Estimations à partir des usages enregistrés et tarifs retenus le 8 octobre 2026 ; pas des factures. Claude : équivalent de consommation, incluant les modèles auxiliaires, sous abonnement. Appels techniques échoués et diagnostics exclus du chiffrage.

Les limites qui comptent pour une décision d'usage

Ce pilote ne porte que sur un dossier et sur des écritures bancaires simples. Il ne mesure ni la TVA, ni les écritures complexes, ni la qualité d'une clôture, ni la prévision financière. Les données sont pseudonymisées, et le plan de comptes annuel complet est fourni comme ensemble fixe de candidats.

Le découpage janvier-juin / juillet-décembre isole les exemples historiques utilisés. Mais le FEC est un export de clôture : cela ne prouve pas que chaque information ou chaque compte était disponible à la date de l'opération en 2025. Il faut parler de reconstruction rétrospective, pas d'un test prospectif.

Nous avons conservé les réponses invalides dans les dénominateurs. Des erreurs HTTP 529 de TypeSafe ont nécessité des reprises sur les cas non terminés ; les décisions déjà obtenues ont été conservées. Les deux réponses invalides de Claude avec six exemples restent comptées comme telles. Ces incidents font partie du retour d'expérience.

Enfin, nous n'avons ni jeu indépendant de confirmation, ni mesure de stabilité sur des exécutions répétées des cas réels, ni calibration des scores de confiance. Les choix de présentation du contexte doivent être confirmés sur de nouveaux dossiers. Les résultats agrégés sont publics ; les écritures sources restent privées.

L'usage que je retiens pour la finance

À ce stade, je retiens Jev comme candidat pour proposer une catégorie, orienter une opération vers une revue ou signaler qu'il manque une information. Ses sorties contraintes et son faible coût sont intéressants lorsque les choix sont connus et que le contexte du dossier est bien préparé.

Je ne retiens pas la comptabilisation autonome sur la base de ce pilote. Même avec le profil enrichi, 13 des 126 propositions de Jev diffèrent du FEC. Il faut comprendre ces désaccords avant de transformer une proposition en écriture.

Le parcours à éprouver chez HOLCO est concret : H-CONTEXT assemble les faits et les antécédents documentés ; H-RULES fournit les conventions validées ; Jev propose un compte ou s'abstient ; H-VISA permet la revue humaine. C'est une piste d'intégration issue de l'étude, pas une fonction de production validée par ces chiffres.

La prochaine étape utile sera un pilote en suggestion sur plusieurs dossiers, avec des attendus revus par un professionnel. Nous mesurerons le temps de revue, les corrections, les abstentions utiles et le coût total par décision validée. C'est sur ces mesures que je déciderais d'élargir l'usage de Jev en finance.

Sources et données agrégées

Digits — étude et présentation du benchmark — Présentation du 7 octobre 2026 ; étude complète datée du 5 octobre consultée pour construire notre protocole. Étude publiée par l'éditeur évaluant aussi son propre produit.

<https://digits.com/blog/ai-cant-close-books-from-outside/>

TypeSafe — primitive Choice — Choix dans une liste définie, probabilités et confiance. La structure de sortie ne démontre pas la justesse comptable.

<https://docs.typesafe.ai/primitives/choice>

Résultats agrégés HOLCO — CSV — Les neuf conditions, les effectifs, les abstentions, les réponses invalides, les coûts estimés et les latences. Aucune transaction individuelle.

HOLCO / RECHERCHE APPLIQUÉE / 8 OCTOBRE 2026
<https://holco.co/etudes/jev-contexte-20261008/resultats-agreges.csv>

Étude HOLCO — version PDF — Article signé, méthode, tableaux et limites, dans un format à partager.

<https://holco.co/etudes/jev-contexte-20261008/etude-jev-contexte.pdf>