

Quand votre IA annonce 70 pour cent, a-t-elle raison 70 pour cent du temps ?

Choisir la bonne forme d'IA pour les prévisions d'affaires, à partir d'un banc d'essai public et en direct sur la Coupe du Monde 2026



Résumé

Les entreprises commencent à laisser l'IA mettre un nombre sur l'avenir : la probabilité qu'une transaction se conclue, qu'un client se désabonne, qu'une facture soit payée à temps. Dès que ce nombre guide une décision, une seule question détermine s'il aide ou nuit : **quand l'IA annonce 70 pour cent, a-t-elle raison 70 pour cent du temps ?**

Nous avons bâti un banc d'essai public et en direct pour y répondre. Trois architectures différentes de prévisionniste d'IA ont pronostiqué chaque match de la Coupe du Monde 2026, en temps réel, avec les règles verrouillées avant le premier match. Nous avons mesuré non pas qui a deviné le plus de vainqueurs, mais dont la confiance annoncée était **honnête**. C'est ce que nous appelons le **score de confiance**.

Trois constats sur lesquels un opérateur peut agir :

1. **Le Comité, une architecture où trois modèles d'IA délibèrent, a obtenu le meilleur score de confiance de toutes les architectures testées.** Quand un nombre de confiance guide une décision qui compte (une marge de crédit, un

engagement de capacité, une recommandation à la haute direction), le Comité est l'architecture vers laquelle se tourner.

2. **Plus cher ne voulait pas dire plus digne de confiance.** Les architectures les plus élaborées ont coûté vingt fois le moins cher et n'ont acheté aucun meilleur score de confiance. La dépense et la fiabilité sont deux choses différentes, et la plupart des organisations paient pour une orchestration qu'elles n'ont jamais validée.
3. **Le constat compte davantage dans votre entreprise que sur le terrain.** Le football a un marché des paris qui valorise déjà chaque match. Votre pipeline du troisième trimestre n'a aucun marché de ce genre. Une prévision d'IA digne de foi n'est donc pas en concurrence avec une référence. Elle **est** la référence qui vous manquerait autrement.

Si vous ne lisez rien d'autre : avant de faire confiance à la confiance d'une IA dans une décision réelle, mesurez son score de confiance. Sur ces preuves, le Comité est l'architecture vers laquelle se tourner, et l'élaboration seule ne vaut pas la dépense.

1. Le problème : un nombre de confiance non audité est un risque

Prenez un pipeline de ventes pondéré. La taille d'une transaction multipliée par sa probabilité de conclusion, sommée. De l'arithmétique de revenu standard. Mais ce n'est de l'arithmétique que si les probabilités sont honnêtes. Un carnet de transactions marquées « 70 pour cent » doit se conclure environ 70 pour cent du temps, sinon le total pondéré est une fiction avec des décimales.

Quand une IA se met à fournir ces probabilités, cette question d'honnêteté devient une question de finance. Un opérateur qui adopte une confiance générée par l'IA sans mesurer si elle est honnête n'a pas automatisé la prévision. Il a automatisé la production de nombres non audités, et chaque décision en aval hérite de l'erreur.

L'objection naturelle : *nous vérifions déjà nos prévisions face aux résultats réels du trimestre dernier.* C'est vrai, et là où vous avez assez d'historique clos, faites ce test rétroactif. Mais il ne peut pas vous donner ce qu'un score de confiance donne. Un score de confiance est disponible **avant** que les vrais résultats n'arrivent, quand un nouveau segment ou un jeune produit a trop peu de transactions closes pour un test rétroactif. Et c'est un signal **par prévision** : vos résultats réels vous disent que le total du dernier trimestre a manqué sa cible, tandis qu'un score de confiance vous dit **quels** niveaux de confiance sont ceux qui mentent.

2. Les trois architectures d'IA, en clair

Nous avons testé trois façons d'agencer un prévisionniste d'IA. Un gestionnaire qui choisit entre elles a besoin de savoir ce que chacune **est** et quand s'en servir. Les détails techniques de construction sont dans l'article scientifique. Voici ce que chacune fait et ce qu'elle coûte.

L'Appel unique

Un seul modèle d'IA reçoit l'information et retourne une prévision avec un nombre de confiance. La chose la plus simple à déployer. Nous l'avons exécuté de deux façons : avec un pack de contexte complet (statistiques, actualités, historique) et avec presque rien. **Servez-vous-en** quand la décision est à fort volume et à faible enjeu et qu'une probabilité approximative suffit. C'est l'option la moins chère de loin.

Le Pipeline

Une chaîne de spécialistes. Une étape d'IA lit les chiffres, une deuxième lit le récit et le contexte, une troisième écrit la prévision finale. Sur les cas difficiles où sa propre confiance est basse, il escalade vers une méthode plus lourde. Plus de pièces mobiles, plus de coût. **Servez-vous-en** quand vous voulez un processus par étapes avec une escalade intégrée sur les prévisions difficiles, et que vous acceptez de payer l'orchestration.

Le Comité

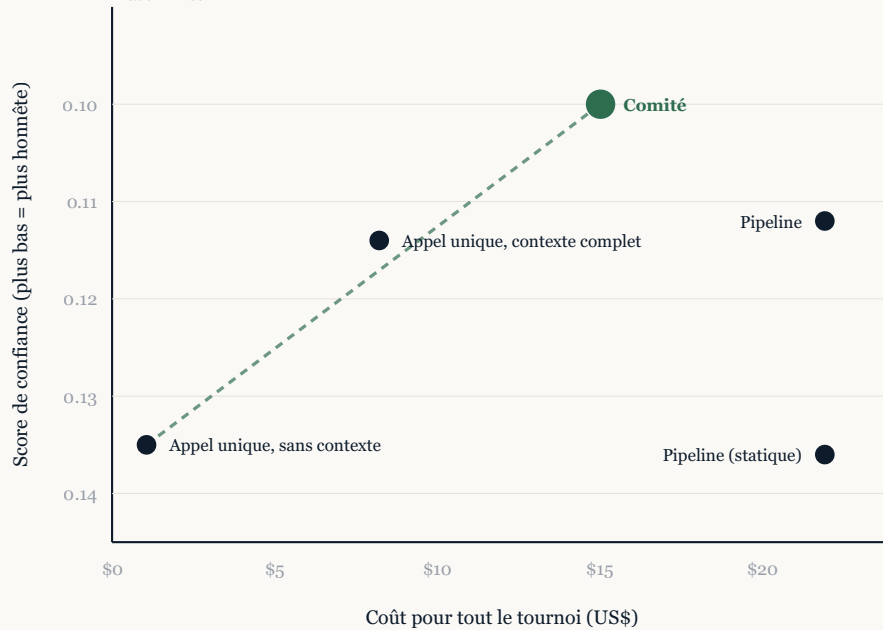
Trois modèles d'IA indépendants, de familles différentes, produisent chacun une prévision en parallèle, puis une étape finale pèse leur accord et leur désaccord en une seule réponse. Quand les trois s'accordent, la confiance est haute. Quand ils divergent, la confiance baisse pour refléter l'incertitude réelle. C'est cette délibération entre modèles, et sa sensibilité au désaccord, qui a valu au Comité le meilleur score de confiance. **Servez-vous-en** quand la décision est critique en calibration, quand un nombre de confiance guide un engagement coûteux ou irréversible.

3. Ce que le banc d'essai a trouvé

Sur l'ensemble des 104 matchs, chaque architecture a obtenu un score de confiance (plus bas est plus honnête) à un coût connu. Les chiffres de coût ci-dessous représentent la dépense totale pour prévoir les 104 matchs, du début à la fin.

Coût vs. score de confiance : ce que chaque architecture achète

La ligne verte est la frontière efficiente. Au-dessus, on paie plus sans gagner en confiance.
haut = mieux

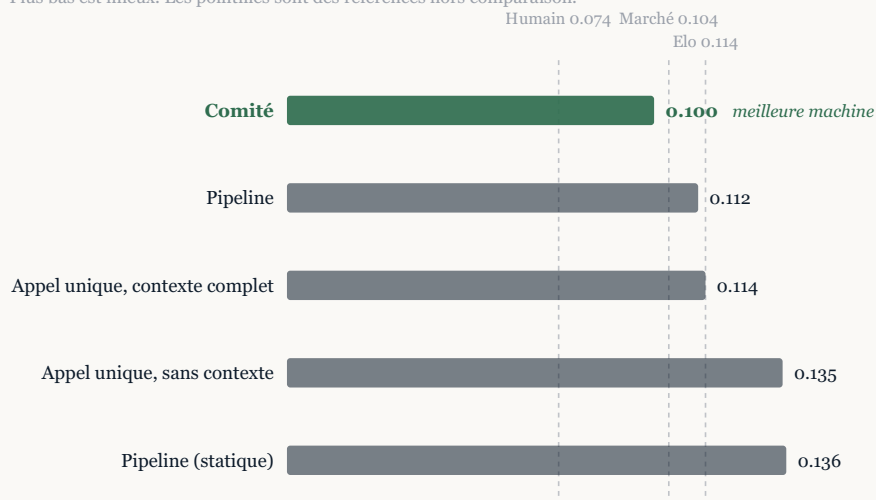


Frontière coût contre score de confiance

Le graphique de la première page est la décision d'achat en une seule image. Deux architectures se situent sur la frontière efficiente : l'Appel unique bon marché à un bout (coût le plus bas) et le Comité à l'autre (meilleur score de confiance). Les Pipelines se situent hors de la frontière, en haut et à droite, ce qui signifie qu'ils coûtent bien plus sans acheter un meilleur score de confiance. Ce sont les deux systèmes les plus coûteux de l'étude, et aucun ne mérite sa prime.

Score de confiance : les architectures d'IA à déployer

Plus bas est mieux. Les pointillés sont des références hors comparaison.



Humain et Marché sont des références, pas des systèmes à acheter. Les écarts sont faibles.

Score de confiance par architecture

Classé face à face, le Comité affiche le meilleur score de confiance de toutes les machines. Nous rapportons les marges avec honnêteté, de deux façons. Premièrement, quand nous avons ré-exécuté le classement des milliers de fois sur des matchs rééchantillonnés, le Comité a tenu la première ou quasi-première place bien plus souvent qu'autrement, mais son avance sur l'architecture suivante n'était pas assez large pour en faire une certitude sur un seul tournoi de 104 matchs. Deuxièmement, et c'est le test le plus dur : sur la dernière tranche de matchs que nous avons mis de côté et notés à l'aveugle, le Comité n'a pas gardé son avance. Sur cet ensemble mince et réservé, une architecture moins chère est passée devant. Lisez donc ceci comme une réponse directionnelle, non comme une photo d'arrivée : sur l'ensemble du tournoi, le Comité est l'architecture vers laquelle se tourner, et un second tournoi affinerait la marge, mais nous ne vous dirons pas que le test réservé propre l'a couronné, parce qu'il ne l'a pas fait. Un meilleur directionnel, divulgué comme directionnel, reste la réponse qu'un opérateur a demandée, et c'en est une plus honnête qu'une fausse certitude.

L'architecture d'IA	Score de confiance (plus bas est mieux)	Coût d'exécution	Taux de réussite
Appel unique, sans contexte	0.135	\$1.06	68%
Appel unique, contexte complet	0.114	\$8.22	66%
Comité	0.100 (meilleur)	\$15.03	67%
Pipeline	0.112	\$21.93	67%
Pipeline (statique)	0.136	\$21.93	67%

Le taux de réussite est la fréquence à laquelle chaque architecture a choisi le vainqueur. Notez qu'il bouge à peine d'une architecture à l'autre (66 à 68 pour cent) : sur un événement dominé par la chance, personne ne devine bien mieux qu'un autre, ce qui est exactement pourquoi l'exactitude est la mauvaise chose sur laquelle acheter et le score de confiance la bonne.

Le nom technique du score de confiance est la calibration. L'article scientifique le rapporte comme erreur de calibration. Même mesure, de sorte qu'un lecteur de l'un ou l'autre article peut faire le lien.

4. Ce que cela signifie pour votre entreprise

La dépense n'est pas la confiance. La leçon la plus claire : les architectures les plus élaborées et les plus coûteuses n'étaient pas plus honnêtes que de bien moins chères. Avant d'engager un budget d'intégration dans une riche orchestration d'IA, exigez la différence de score de confiance qu'elle achète, mesurée, non supposée. Sur cette tâche, elle était nulle.

Le résultat football sous-estime le cas d'affaires. Les architectures d'IA se sont situées près du marché des paris parce que le football a un marché public profond qui

valorise chaque match, une référence dure qui laisse peu de place pour se distinguer. Votre entreprise n'a aucun marché de ce genre. Il n'y a aucune ligne de clôture pour votre pipeline, votre plan de demande, votre désabonnement. La prévision d'IA digne de foi n'est donc pas comprimée contre une référence. Elle est la référence qui vous manquerait autrement, et cela rend la mesure de son score de confiance plus précieuse dans une entreprise, non moins.

Où cela s'applique, et où cela ne s'applique pas. Le constat se transfère proprement à une décision de la forme d'un match de football : une décision prise selon un calendrier, avec une échéance ferme, et un vrai résultat que vous pouvez noter peu après. Une réunion de pipeline hebdomadaire, un plan de demande mensuel, une prévision de désabonnement trimestrielle. Il se transfère moins à un jugement ponctuel sans résultat clair et sans échéance. Pointez donc le score de confiance vers vos décisions récurrentes et notées d'abord. C'est là qu'un nombre de confiance est une promesse que vous pouvez opposer à l'IA.

5. Quelle architecture pour quelle décision

Le tournoi transforme les architectures en un menu tarifé. Faites correspondre l'architecture à la décision :

	Faible enjeu (un nombre approximatif suffit)	Fort enjeu (une confiance erronée coûte cher)
Fort volume (beaucoup de prévisions)	Appel unique, sans contexte. Le moins cher, assez bon.	Comité sur le sous-ensemble critique. Réservez-le aux décisions qui peuvent se le permettre.
Faible volume (peu de prévisions, délibérées)	Appel unique. Simple, faible coût.	Comité. Le meilleur score de confiance, la prime en vaut la peine.

La règle sous la grille : tournez-vous vers le Comité quand un nombre de confiance guide un engagement que vous ne pouvez pas facilement défaire. Utilisez l'Appel unique bon marché partout ailleurs, et mettez l'argent économisé au service de la mesure du score de confiance lui-même.

6. Le mettre en pratique

Six questions à poser à tout fournisseur de prévision par IA

Copiez-collez-les. Une équipe qui ne peut pas y répondre n'a pas mesuré ce qu'elle vend.

1. « Montrez-moi la courbe de fiabilité de votre système sur 30 événements notés ou plus. Pas l'exactitude. La courbe. »

2. « Quand votre système annonce 70 pour cent, à quelle fréquence a-t-il réellement raison ? Donnez-moi ce nombre, sur combien d'événements, sur des résultats que vous avez mis de côté et sur lesquels vous ne vous êtes pas entraîné. »
3. « Quel seuil de confiance achemine une prévision vers un humain, et où cela est-il journalisé ? »
4. « Qu'a mesurablement ajouté l'intégration du contexte ? Montrez-moi la comparaison sans contexte. »
5. « Pouvez-vous rejouer n'importe quelle prévision passée exactement, avec les entrées qu'elle a vues au moment du verrouillage ? Qu'est-ce qui empêche la révision a posteriori ? »
6. « Combien coûte une prévision, et combien coûte chaque point d'amélioration du score de confiance par rapport à notre processus actuel ? »

Comment le score de confiance est calculé, en clair

Regroupez vos prévisions par la confiance que vous leur avez attribuée. Prenez chaque prévision que vous avez marquée autour de 70 pour cent et vérifiez combien se sont réellement réalisées. Si 70 pour cent l'ont fait, cette bande est honnête. Si seulement la moitié l'a fait, votre 70 est en réalité un 50, et l'écart est le mensonge. Faites cela sur chaque bande de confiance, faites la moyenne des écarts, et vous avez un seul nombre : à quelle distance votre confiance annoncée se situe de la réalité, en moyenne. Plus bas est mieux. Un prévisionniste parfaitement honnête obtient zéro.

C'est toute l'idée. La formule exacte, et un script prêt à coller qui l'exécute sur votre propre journal de prévisions, sont dans le dépôt public OSF ci-dessous, de sorte que vous n'avez jamais à bâtir le calcul vous-même.

L'exécuter vous-même en un trimestre

Verrouillez la définition du résultat et l'échéance de verrouillage par écrit avant la première prévision. Enregistrez chaque prédiction avec sa confiance avant l'échéance, sans révision.

Ce qu'il faut exécuter. Ne commencez pas par le Comité. Commencez par ce que vous avez plus un plancher bon marché : votre processus actuel, une simple moyenne historique, et l'Appel unique d'IA le moins cher, tous prévoyant les mêmes événements en parallèle. Cela vous dit si l'IA bat votre statu quo tout court, et si l'intégration de données coûteuse mérite sa place, avant de dépenser en orchestration. Réservez le Comité au sous-ensemble critique de décisions une fois que la comparaison bon marché vous a montré que l'IA vaut la confiance.

La seule limite dure. Un score de confiance a besoin d'au moins 30 événements notés pour signifier quelque chose. Un seul trimestre d'un pipeline de taille moyenne peut en conclure moins. Si c'est le cas, ne forcez pas un verdict sur des données minces : regroupez des décisions similaires, ou exécutez deux trimestres, avant de classer quoi que ce soit.

Après assez d'événements notés, une entreprise de volume ordinaire possède son premier score de confiance. Autrement dit, elle sait si le 70 de son IA vaut 70.

Les preuves derrière ce document

Chaque affirmation ici remonte à une étude entièrement publique et préenregistrée. Rien n'est affirmé que les données ne portent pas.

- **L'article scientifique complet, chaque prédiction, chaque coût, le jeu de données complet** : OSF, DOI 10.17605/OSF.IO/R5SBJ (osf.io/r5sbj)
- **Le tableau de bord public en direct** : match.inocta.io, où chaque prédiction a été publiée avant le coup d'envoi
- **Le préenregistrement verrouillé** : osf.io/ke35t, gelé avant le premier match noté

On ne peut pas automatiser ce qu'on ne comprend pas, et un nombre de confiance que vous n'avez pas audité n'est pas encore quelque chose que vous comprenez.

inocta.io · L'Intelligence des Opérations.