La calibration avant la précision : un banc d'essai préenregistré, en direct, de trois architectures de raisonnement d'IA sur la Coupe du Monde de la FIFA 2026

Mo Kahlain
inocta.io, Toronto, Canada
mo@inocta.io

Version de travail, compilée le 21 juillet 2026
Préenregistrement : OSF ke35t, DOI 10.17605/OSF.IO/R5SBJ (gelé avant le premier match comptabilisé)
Tableau de bord en direct : https://match.inocta.io
Traduction française. La version anglaise fait foi.

Résumé

Quand un système d'IA annonce une confiance de 70 pour cent, ce nombre a-t-il la moindre valeur ? Nous présentons inocta-bench, un banc d'essai préenregistré et en direct qui compare trois architectures de raisonnement d'IA sur la *calibration* probabiliste, non sur l'exactitude : un généraliste à appel unique (SOLO), un pipeline séquentiel de spécialistes (PIPELINE) et un conseil multimodèle (COUNCIL). Les trois prévoient les 104 matchs de la Coupe du monde FIFA 2026, en public et en temps réel. Au lieu de comparer des classes de modèles sur l'exactitude, nous comparons des architectures de raisonnement sur la calibration, avec un protocole préenregistré et la référence publiée la plus solide (Elo avec variable indicatrice de terrain, selon Robberechts et Davis 2019) comme ligne de comparaison, aux côtés de la cote de clôture du marché des paris et d'un prévisionniste humain. Deux contrôles d'ablation (un jumeau du généraliste privé de contexte et un jumeau du pipeline dont la boucle d'apprentissage est désactivée) ont été enregistrés pour séparer la contribution du contexte de la contribution de l'architecture, et l'apprentissage véritable de la simple récupération. La couche de boucle d'apprentissage n'a jamais été alimentée en production, si bien que la question de l'apprentissage est rapportée comme non testable telle que conçue, et non comme un résultat (Section 7). Chaque prédiction, chaque coût et chaque erreur sont publiés avant le début de chaque match. Le protocole, les invites et les règles de notation ont été verrouillés par hachage et enregistrés (OSF 10.17605/OSF.IO/R5SBJ) avant le premier match noté. Le tournoi est terminé, et la réponse sur laquelle un opérateur peut agir se situe sur l'axe de la calibration. Sur l'ensemble du tournoi, le conseil, un ensemble de trois familles de modèles indépendantes doté d'une étape qui pondère leur désaccord, affiche l'erreur de calibration attendue la plus faible de toutes les machines testées, la référence humaine mise à part (0.100), en dessous des architectures à appel unique et à pipeline, de la référence déterministe et de la propre ligne de clôture du marché. C'est l'architecture dont la confiance annoncée suit la réalité de la façon la plus étroite, et pour une organisation qui choisit une forme d'IA à qui confier une prévision, c'est celle vers laquelle se tourner. Nous rapportons la marge avec honnêteté : sur un événement dominé par le hasard, les intervalles de bootstrap se chevauchent, si bien que l'avance est directionnelle plutôt que tranchée, et un second tournoi l'affinerait. L'exactitude n'a jamais été la question, et elle s'est comportée comme le prédit une tâche dominée par le hasard : aucune architecture n'a prédit de façon fiable plus de matchs qu'une autre, et l'inversion légère de notre hypothèse annoncée, les voies

d'IA passant juste en dessous de la ligne de clôture du marché sur l'ensemble réservé, est une lecture directionnelle que nous n'appuyons pas. Que les voies d'IA se situent près du marché tient à une propriété du football, qui dispose déjà d'un marché des paris public et profond valorisant chaque match. Une entreprise qui prévoit son propre pipeline n'a aucune référence de ce genre, et la probabilité d'IA calibrée est ce qui s'en rapproche le plus de ce qu'elle peut construire, ce qui fait de l'axe de la calibration, et du conseil sur cet axe, le résultat opérationnel de l'étude. Nous publions chaque prédiction, chaque coût et l'intégralité du jeu de données avec l'article, de sorte que la direction indiquée par le résultat puisse être confrontée aux preuves sur lesquelles elle repose.

Mots-clés : calibration, prévision probabiliste, grands modèles de langage, architectures de raisonnement, préenregistrement, prévision sportive

1 Introduction

Un système d'IA qui annonce une confiance de 70 pour cent fait une promesse. Si les pronostics émis à 70 pour cent de confiance gagnent environ 70 pour cent du temps, le nombre porte une information sur laquelle un décideur peut agir. Il vous dit quoi déléguer, quoi couvrir, et quand escalader vers un humain. Si les pronostics à 70 pour cent gagnent 50 pour cent du temps, le système n'est pas simplement imprécis. Il est mal calibré, et chaque décision en aval bâtie sur sa confiance hérite du défaut. Au sein des organisations qui adoptent l'IA, c'est là la question opérante. Savoir si un système a raison un peu plus ou un peu moins souvent qu'une alternative importe bien moins que de savoir si son propre compte rendu de son incertitude peut être digne de confiance, car c'est la confiance dans la confiance annoncée qui détermine quelles décisions peuvent être confiées, tout simplement.

Cet article pose cette question de manière expérimentale, au niveau où la plupart des décisions d'IA appliquée sont réellement prises. Non pas le choix du modèle de base, mais le choix de l'*architecture de raisonnement*, à savoir comment les appels, les rôles et le contexte sont agencés autour des modèles. Nous comparons trois architectures qui reflètent les schémas que les organisations déploient aujourd'hui. Un généraliste à appel unique doté d'un contexte riche (Solo), un pipeline séquentiel de rôles spécialisés (Pipeline), et un conseil de trois familles de modèles indépendantes dont les sorties sont fusionnées par un décideur (Council). La question de recherche, fixée dans le préenregistrement avant que le moindre match ne soit noté, est la suivante :

> *Sur une tâche de prévision dominée par le hasard (les résultats des matchs de la Coupe du Monde), comment trois architectures de raisonnement de l'IA diffèrent-elles en calibration probabiliste, et non en exactitude, les unes par rapport aux autres et par rapport à des références sans IA ?*

La prévision des résultats de matchs de football est le banc d'essai, non le sujet. Elle offre ce dont la mesure de calibration a besoin et ce que les bancs d'essai rétrospectifs ne peuvent fournir : un flux de 104 événements authentiquement incertains selon un calendrier fixe, une vérité terrain sans ambiguïté dans les deux heures suivant chaque prédiction, un horizon d'information strict (les prédictions se verrouillent 60 à 75 minutes avant le coup d'envoi), et, de façon décisive, une prévision de référence publique et liquide dans les cotes de clôture du marché des paris, que des décennies de littérature sur la prévision considèrent comme la plus forte agrégation disponible de l'information publique. Une tâche dominée par le hasard irréductible est précisément là où la calibration se sépare de l'exactitude. Lorsqu'aucun prévisionniste ne peut avoir raison bien plus souvent que ce que les cotes permettent, la description honnête de sa propre incertitude devient la compétence mesurable.

Trois engagements de conception distinguent ce banc d'essai des évaluations rétrospectives. **Premièrement, il est préenregistré et en direct.** Le protocole, les invites, les versions verrouillées des modèles, les règles de notation et les conditions d'affirmation ont été verrouillés par hachage et enregistrés sur OSF (DOI 10.17605/OSF.IO/R5SBJ) avant le premier match noté. Chaque prédiction est publiée sur un tableau de bord public (`https://match.inocta.io`) avant le coup d'envoi, avec le coût et la latence par prédiction. Aucune possibilité de rapport sélectif, de choix de métrique a posteriori, ou de réutilisation discrète du jeu de test. **Deuxièmement, il porte ses propres contrôles.** Un jumeau sans contexte du généraliste (Solo-Zero) isole ce qu'apporte le contexte par rapport à ce que le modèle sous-jacent sait déjà. Un jumeau du pipeline avec la boucle d'apprentissage désactivée (Pipeline-Static) a été enregistré pour transformer la question « le système apprend-il réellement pendant le tournoi ? » en une ablation mesurée plutôt qu'en une affirmation marketing, affinée davantage par deux schémas synthétiques semés dans le flux de contexte et absents de toute donnée publique. (La couche de boucle n'a jamais été alimentée en production. La section 7 rapporte ce que la paire de contrôle peut et ne peut pas dire en conséquence.) **Troisièmement, il énonce un a priori falsifiable.** Nous nous attendons à ce qu'aucune voie d'IA ne batte la ligne de clôture du marché sur l'exactitude brute. La contribution est la mesure de calibration, la transparence des coûts et le test de la boucle d'apprentissage, et nous nous engageons à publier ce que les données montrent, y compris les inversions de notre thèse.

Les voies d'IA sont évaluées par rapport à trois points de référence sans IA : un classement Elo calculé en interne avec un ajustement du domicile pour la nation hôte (la plus forte classe de référence publiée pour cette tâche, selon Robberechts and Davis [12]), les cotes de clôture dé-marginées (Market), et un prévisionniste humain unique et nommé servant de ligne de référence non aveugle (Human, section 3). La section 3 spécifie les architectures, les contrôles et la discipline de répartition. La section 4 couvre les sources de données et la construction des paquets de contexte et de la référence Elo. La section 5 donne le contrat de notation et les conditions d'affirmation préengagées. Les résultats suivent le tournoi (section 6), et l'analyse de la boucle d'apprentissage (section 7) traite la seconde question de recherche, subordonnée, fixée dans l'enregistrement :

> *Une architecture dotée d'une boucle d'apprentissage active démontre-t-elle un apprentissage véritable (par opposition à de la récupération), mesuré par (a) l'écart entre l'architecture en boucle et son contrôle à boucle désactivée, et (b) la captation de deux schémas synthétiques semés, absents des données publiques ?*

Contributions. Cet article en apporte quatre. La première est un banc d'essai préenregistré, entièrement public et en direct, d'architectures de raisonnement de l'IA, avec la calibration comme métrique phare. La seconde est un ensemble de contrôles d'ablation appariés, enregistrés pour séparer le contexte de l'architecture et l'apprentissage de la récupération, y compris une sonde à schémas semés (la partie apprentissage n'a pas été exercée en production et est rapportée comme telle, section 7). La troisième est une transparence complète des coûts et de la latence par prédiction, traitant l'économie opérationnelle comme une sortie expérimentale de premier plan. La quatrième est un jeu de données ouvert, un harnais d'évaluation et un ensemble d'invites, vérifiés par hachage par rapport au préenregistrement et publiés avec cet article.

2 Travaux connexes

Ce banc d'essai se situe à l'intersection de trois littératures. La prévision statistique du football fournit les références et les familles de caractéristiques. La notation propre et la calibration fournissent le cadre de mesure. La prévision fondée sur les LLM et les architectures multi-agents constituent la classe de système à l'étude.

Prévision statistique de la Coupe du Monde. Les familles de caractéristiques que nos paquets de contexte opérationnalisent descendent de la lignée de régression régularisée de Groll et al. [7] et Groll et al. [8], dont les covariables sélectionnées par LASSO, l'avantage du domicile et de l'hôte décomposé en effets d'hôte, de continent et de confédération, la structure de l'effectif, et les proxys de valeur marchande, sont les entrées sur lesquelles un modèle statistique solide s'appuie pour cette tâche. Nous les empruntons comme contenu de la couche de contexte L0, et nous écartons délibérément les macro-covariables que leurs analyses d'importance des variables trouvent faibles, le PIB et la population, qui portent peu de signal marginal une fois que l'Elo et les cotes du marché sont présents. Le point d'intérêt pour nous n'est pas d'améliorer l'exactitude de cette lignée mais de demander si un système d'IA à qui l'on remet la même classe d'information rapporte sa confiance honnêtement.

Le précédent Elo plus indicatrice de domicile. Robberechts and Davis [12] montrent qu'un logit ordonné sur la différence d'Elo plus une indicatrice de domicile bat tout modèle plus complexe qu'ils testent sur les Coupes du Monde 2002–2014, y compris les modèles augmentés par les bookmakers et les forêts aléatoires (RPS 0,1860 contre 0,1877 pour les bookmakers). C'est pourquoi Elo est une voie de premier plan plutôt qu'une note de bas de page. C'est la plus forte classe de référence simple pour cette tâche, de sorte qu'une architecture qui ne peut égaler un Elo ajusté au domicile n'a pas mérité sa complexité. Notre Elo interne (section 4) suit ce précédent dans l'esprit, une différence d'Elo avec un terme de domicile pour la nation hôte, calculé de façon transparente à partir de l'historique des matchs afin que la référence soit reproductible plutôt qu'empruntée à un classement tiers opaque.

Règles de notation propres et calibration. Le cadre de mesure descend de Brier [2] et Murphy [11] à travers la synthèse des règles de notation propres de Gneiting and Raftery [6]. Le score de Brier est le score propre que nous rapportons comme métrique phare, sa décomposition de Murphy sépare la calibration (fiabilité) de la discrimination (résolution), et ECE isole la pièce de calibration à elle seule, le nombre unique pour la question que pose cette étude : quand un système annonce soixante-dix pour cent, cela se produit-il soixante-dix pour cent du temps. La finesse complète le tableau, en distinguant un prévisionniste bien calibré d'un lâche bien calibré, de sorte que nous rapportons la calibration, la finesse et la décomposition ensemble plutôt que l'une ou l'autre isolément.

Bornes supérieures de l'apprentissage automatique moderne. Al-Bustami and Ghazal [1] représente l'approche contemporaine orientée exactitude, des caractéristiques au niveau des joueurs avec les cotes des bookmakers et les différentiels d'Elo comme classe d'entrée dominante. Nous traitons cette lignée comme un cadrage plutôt que comme une cible. Elle marque jusqu'où l'exactitude peut être poussée sur cette tâche, ce qui est précisément l'axe qu'un banc d'essai dominé par le hasard s'attend à voir saturer, et c'est la raison pour laquelle nous mesurons la calibration à la place.

Incertitude bayésienne dans les modèles de football. Egidi et al. [5] et Robberechts et al. [13] portent un fil de quantification de l'incertitude que notre orientation vers la calibration met en parallèle du côté de l'IA. Là où ces modèles expriment l'incertitude par une distribution a posteriori sur les résultats, nous posons la même question sous-jacente à une architecture de modèle de langage, à savoir si l'incertitude qu'elle rapporte est honnête, et nous y répondons avec la machinerie de calibration ci-dessus plutôt qu'avec un modèle génératif de la ligne de score.

Prévision par LLM et architectures multi-agents. Du côté de l'IA, une lignée croissante demande si les modèles de langage peuvent prévoir des événements futurs incertains à un niveau de compétence utile. Halawi et al. [9] construisent un système augmenté par récupération qui approche, et par endroits dépasse, l'agrégat de prévisionnistes humains compétitifs, et Karger et al. [10] introduisent ForecastBench, un banc d'essai dynamique et sans contamination de la prévision par LLM face à des groupes de comparaison humains. Tous deux partagent notre engagement envers des questions dont les réponses sont inconnues au moment de la prédiction, ce qui élimine le facteur de confusion de réutilisation du jeu de test qui trouble l'évaluation rétrospective. Là où ils mesurent l'exactitude face à des foules humaines, nous maintenons l'exactitude fixe en choisissant une tâche dominée par le hasard et mesurons la calibration à la place, et nous comparons des architectures plutôt que des pipelines de récupération.

L'architecture de conseil que nous testons descend des travaux sur le raisonnement multi-agents. Du et al. [4] montrent que faire débattre plusieurs instances de modèle sur plusieurs tours améliore l'exactitude factuelle et réduit l'hallucination par rapport à un appel unique, et Chen et al. [3] proposent ReConcile, une table ronde de familles de modèles diverses qui atteint un consensus pondéré par la confiance, ce qui est le parent publié le plus proche du décideur de notre Council. La preuve que davantage d'agents aide n'est pas unidirectionnelle, et nous citons la tension plutôt que de la masquer. Zhu et al. [16] constatent que le débat multi-agents ordinaire sous-performe souvent face à un simple vote majoritaire, et ne recouvre un avantage que lorsque les agents sont divers et que leur confiance est calibrée, précisément les deux propriétés autour desquelles notre conseil est construit. Tran and Kiela [15] vont plus loin et montrent qu'un agent unique peut égaler ou battre les systèmes multi-agents sur le raisonnement une fois que le budget de jetons de réflexion est maintenu égal, attribuant les gains multi-agents rapportés à un surcroît de calcul au moment du test plutôt qu'à l'interaction elle-même. Notre étude s'adresse à ce débat sous un angle différent : nous ne prétendons pas que le conseil raisonne mieux, nous demandons si sa *confiance* est plus digne de confiance, une propriété qu'aucun de ces travaux ne mesure directement. La calibration de la confiance des LLM est elle-même un problème actif, avec des méthodes telles que Thermometer [14] développées pour recalibrer les probabilités des modèles après coup, ce qui est une raison de plus de mesurer la calibration in situ plutôt que de la supposer.

3 Conception du banc d'essai

Le banc d'essai exécute huit voies de prévision sur chaque match de la Coupe du Monde de la FIFA 2026 (104 rencontres : 72 matchs de phase de groupes répartis sur trois journées, puis les tours à élimination directe à mesure que les tableaux se résolvent). La cible évaluée pour chaque match est le résultat sur 90 minutes, en trois issues (domicile / nul / extérieur), à chaque étape du tournoi. Tous les éléments de conception de cette section ont été fixés dans le préenregistrement (OSF 10.17605/OSF.IO/R5SBJ) avant le premier match évalué. Les documents enregistrés ainsi que l'invite de production et le code de la passerelle sont liés par des sommes de contrôle SHA-256 (Annexe B).

3.1 Voies

Voie	Type	Définition
Solo (A1)	architecture	Appel unique à modèle fixé, pack de contexte complet à 7 couches
Solo-Zero (A1-zero)	contrôle	Même chemin de code que A1, métadonnées de rencontre seulement, zéro couche de contexte
Pipeline (A2)	architecture	Pipeline séquentiel à 3 étapes, validation de schéma par étape, une relance
Pipeline-Static (A2-static)	contrôle	A2 avec la couche de boucle d'apprentissage retirée, exclu des écritures de boucle
Council (A3)	architecture	Trois familles de modèles en parallèle plus un décideur
Elo	référence	Elo interne à partir de l'historique des matchs (Section 4)
Market	référence	Cotes de clôture, un seul instantané T−1 h, dévigué proportionnellement
Human	repère	Un humain nommé (le déclarant), ligne de repère non aveugle, pronostic verrouillé au coup d'envoi

Table 1 – Les huit voies : trois architectures, deux contrôles d'ablation, trois références non IA. Les définitions des voies sont des engagements repris mot pour mot du préenregistrement §2.

3.2 Les trois architectures

Solo (généraliste à appel unique). Un appel à un modèle fixé avec une invite système d'analyste partagée et le pack de contexte complet. La sortie structurée (vainqueur, score, confiance 0–100, au plus trois facteurs de risque, une note de raisonnement d'au plus 100 mots) est imposée par un schéma JSON, avec exactement une relance en cas d'échec de schéma, et un second échec est enregistré comme une abstention, jamais silencieusement réparé.

Pipeline (spécialistes séquentiels). Trois appels en séquence sur la même famille de modèle fixée. Un *statisticien* lit uniquement les couches quantitatives du pack de contexte. Un *lecteur des opérations* combine ensuite la sortie du statisticien avec les couches narratives (actualités, rivalités, éclairages antérieurs, la boucle d'apprentissage lorsqu'elle est présente). Un *éditeur de voix* synthétise les deux dans la prédiction structurée finale. Chaque étape porte son propre schéma de sortie ainsi que la même règle d'une relance. Une porte d'escalade fait partie de l'architecture enregistrée : si la confiance finale du pipeline tombe sous 60, il adopte la prédiction de la Council pour ce match, et l'adoption est journalisée et attribuée en coût, le pipeline payant pour ses trois étapes propres quoi qu'il arrive.

Council (conseil multi-modèles). Trois membres issus de familles de modèles distinctes (Claude, GPT, Gemini, avec les versions exactes fixées dans l'Annexe B) reçoivent l'invite et le contexte identiques en parallèle. Un décideur de la première famille reçoit les trois soumissions *avec les identités des modèles retirées* (« membre 1/2/3 » anonymes), reçoit pour instruction que sa confiance doit refléter le désaccord entre membres, et produit la prédiction unique de la voie. Les identités des modèles sont journalisées dans la télémétrie et publiées, mais ne sont pas visibles pour le décideur au moment de la décision.

L'architecture et le mélange de modèles sont donc confondus par conception. Toutes les affirmations de cette étude concernent les systèmes fixés tels que spécifiés, non les architectures de raisonnement dans l'abstrait.

3.3 Contrôles

SOLO-ZERO exécute le chemin de code littéral de SOLO avec le pack de contexte retenu : métadonnées de rencontre seulement. L'écart entre les jumeaux chiffre la valeur du pack de contexte : toute différence de calibration ou d'exactitude est attribuable au contexte, non à l'architecture, puisque rien d'autre ne diffère. PIPELINE-STATIC exécute le pipeline complet avec la couche de boucle d'apprentissage annulée de son contexte et est exclu des écritures de boucle d'apprentissage, et il s'exécute sur chaque match que PIPELINE exécute. Ce contrôle porte la question de recherche de niveau 2 (Section 7).

3.4 Discipline de répartition et horizon d'information

Les voies IA prédisent exactement une fois par match, déclenchées lorsque le match entre dans la fenêtre pré-coup-d'envoi T−75 min sur un planificateur à 15 minutes (en pratique 60 à 75 minutes avant le coup d'envoi). Les prédictions ne sont jamais révisées après la répartition. La répartition est protégée par revendication : des horloges redondantes s'affrontent sur une revendication transactionnelle par match, de sorte qu'une planification en double ne peut pas produire de prédictions en double. La référence MARKET est un instantané T−1 h unique, de sorte que les voies IA et le marché se situent approximativement au même horizon d'information. La ligne de marché n'entre jamais dans le contexte d'aucune voie IA : le pipeline des cotes est un chemin de code séparé sans contact avec les packs de contexte. Les compositions officielles sont publiées environ 35 minutes avant le coup d'envoi, après la fermeture de la fenêtre de répartition, de sorte qu'aucune voie ne les voit jamais. La voie HUMAN peut être renseignée à tout moment avant le coup d'envoi et se verrouille au coup d'envoi.

La voie humaine est une ligne de repère non aveugle, non une référence contrôlée. Le prévisionniste humain nommé était le déclarant et l'opérateur du projet. À ce titre, l'humain avait accès, avant de verrouiller chaque pronostic, aux prédictions et au raisonnement publiés des voies IA pour le même match, et, tout au long du tournoi, au tableau de bord courant et au retour sur la performance des voies qu'un prévisionniste en aveugle ne verrait pas. Le pronostic humain est donc informé par le système auquel il est comparé, à la fois sur un canal par match (visibilité des sorties IA) et sur un canal à l'échelle du tournoi (retour et adaptation). Nous rapportons la voie HUMAN par souci de transparence et d'exhaustivité, mais ses résultats ne portent aucune affirmation de comparaison indépendante humain contre machine, et tout avantage de calibration qu'elle montre est confondu par l'accès à l'information. Une référence humaine propre exigerait un prévisionniste aveugle à toutes les sorties de voies et aux résultats courants. Nous n'en avons pas exécuté, et nous signalons cela comme une limite (Section 10) plutôt que de l'occulter.

3.5 Engagements méthodologiques (M1–M8)

Le préenregistrement lie huit engagements.

- **M1.** Préenregistrement sur OSF avant le premier match évalué.
- **M2.** Évaluation multi-juges des sorties narratives (trois familles de modèles comme juges, vérité terrain issue du fournisseur de données, une vérification humaine par sondage stratifié de 10 %).

- **M3.** Fixation des versions : chaque appel journalise à la fois la version demandée et l'identifiant de modèle effectif rapporté par le fournisseur, de sorte qu'une substitution silencieuse de modèle soit détectable dans la télémétrie publiée.
- **M4.** Le contrôle de boucle s'exécute sur chaque match que la voie en boucle exécute.
- **M5.** Deux motifs synthétiques, scellés dans une annexe avant le tournoi, injectés dans le flux de contexte à des emplacements déterministes, avec la reprise mesurée par un détecteur, la sonde pour l'apprentissage par opposition à la récupération.
- **M6.** Les 20 derniers matchs du tournoi sont gelés comme ensemble de test réservé, et aucun changement d'invite, de barème ou de configuration n'est permis après le 84e match du tournoi.
- **M7.** Références humaines : le déclarant dès le lancement, des rédacteurs publics nommés comme ajout possible en milieu de tournoi selon le processus d'amendement enregistré.
- **M8.** Le contrat de reporting de calibration de la Section 5.

3.6 Fenêtre de développement et amendements

Les matchs antérieurs au verrouillage de l'enregistrement sont étiquetés données de développement partout où ils apparaissent et n'entrent dans aucune affirmation principale, et les affirmations principales sont portées par la performance sur l'ensemble de test réservé. Tout amendement de protocole doit être versionné et publié avant tout match qu'il affecte, et le journal des amendements fait partie du registre public (Annexe B).

4 Données

4.1 Sources

Les rencontres, les résultats, les statistiques d'équipe, l'historique des matchs et les cotes de clôture proviennent d'un fournisseur commercial unique (API-Football, niveau PRO), vérifié par une sonde de préenregistrement sur les formes de réponse et la couverture. Les récits d'actualités et de rivalités sont récoltés par un modèle ancré dans la recherche (Perplexity Sonar) avec des citations d'URL stockées par élément. La météo provient des coordonnées du site au moment de la prévision.

La couverture est imparfaite et la conception en tient compte plutôt que de la masquer. Plusieurs champs du fournisseur, la disponibilité des compositions et certaines statistiques historiques en particulier, se lisaient comme indisponibles avant le tournoi et sont restés clairsemés pour les rencontres de moindre profil, de sorte que chaque couche de contexte est nullable et qu'une couche manquante est journalisée par match plutôt que substituée silencieusement. Le principe directeur est qu'un manque est un manque pour chaque voie également : lorsqu'une couche échoue à être récupérée, aucune voie ne la reçoit, de sorte qu'aucune voie n'est avantagée ou désavantagée par un trou de couverture. La seule voie qui ne reçoit aucune couche de contexte par conception, le jumeau sans contexte Solo-Zero, n'est donc pas un cas particulier de données manquantes mais le contrôle délibéré à zéro contexte au regard duquel les autres sont lues.

4.2 Le pack de contexte à 7 couches

Chaque voie consommatrice de contexte puise dans un pack récolté unique d'au plus sept couches (Table 2). Le pack est assemblé une fois par match au moment de la récolte, sérialisé avec un plafond

de 120,000 caractères, et l'ensemble des couches récupérées avec succès ou échouées est journalisé par match afin que toute prédiction puisse être retracée exactement jusqu'au contexte disponible au moment où elle a été faite. Le jumeau sans contexte Solo-Zero exécute le chemin de code identique avec le pack retenu, ne recevant que les métadonnées de rencontre.

Couche	Contenu	Source	Nullable
L0	Statistiques d'équipe et de match	API-Football	oui
L1	Météo du site au moment de la prévision	Coordonnées du site	oui
L2	Récit d'actualités, avec citations	Perplexity Sonar	oui
L3	Récit de rivalités, avec citations	Perplexity Sonar	oui
L5	Éclairages pré-match générés	Modèle, à partir de L0–L3	oui
L6	Analyse pré et post-match	Modèle	oui
L7	Entrées de boucle d'apprentissage	Couche de boucle (jamais peuplée)	oui

Table 2 – Le pack de contexte. La numérotation des couches suit l'implémentation et saute L4. Chaque couche est nullable et une récupération échouée est journalisée, non substituée. Les voies de contexte (Solo, Pipeline, Pipeline-Static, Council) consomment le pack assemblé. Solo-Zero n'en reçoit rien, et les références déterministes (Elo, Market) ainsi que le repère humain n'utilisent aucun pack. L7 n'a jamais été peuplée en production (Section 7).

4.3 La construction de la référence Elo

Elo est calculée en interne à partir de l'historique des matchs du fournisseur plutôt qu'extraite de sites de classement tiers, puisqu'aucune API propre n'existe et que l'extraction est fragile et méthodologiquement opaque. Les paramètres enregistrés sont les suivants.

- Score attendu Elo standard $W_e = (1 + 10^{-(R_h+H-R_a)/400})^{-1}$.
- $K = 30$ pour les matchs de Coupe du Monde (phase finale et qualifications), $K = 20$ pour tous les autres matchs internationaux.
- Un multiplicateur de différence de buts sur K (1 pour un nul ou une marge d'un but, 1,5 pour deux, 1,75 pour trois, $+1/8$ par but supplémentaire).
- Toutes les équipes démarrent à 1500, sur une fenêtre de rétro-calcul du 1er juin 2022 au début du tournoi (2,001 rencontres réparties sur les 48 équipes qualifiées au moment du calcul).

L'ajustement à domicile $H = 100$ s'applique à l'équipe à domicile listée dans les mises à jour historiques et, au moment de la prédiction du tournoi, uniquement lorsque l'équipe à domicile est une nation hôte jouant chez elle : tout autre site de Coupe du Monde est traité comme neutre. Les probabilités à trois issues dérivent de l'identité du score attendu Elo $W_e = P(\text{win}) + \frac{1}{2}P(\text{draw})$ avec la masse de nul fixée à la part empirique de nuls des rencontres de rétro-calcul traitées. Le calcul est déterministe et reconstruit à partir de zéro quotidiennement, de sorte que toute relecture ultérieure reproduise exactement les prédictions de la voie, la propriété sous laquelle l'enregistrement permet le calcul rétroactif. La divulgation complète des paramètres figure dans le registre public (Annexe B). Deux limites connues, les adversaires hors-graphe entrant à 1500 et la simplification de domicile listé pour les tournois à site neutre dans l'historique, y sont indiquées ainsi que dans la Section 10.

4.4 Flux d'information et isolation

L'isolation entre les voies est structurelle, non procédurale. La ligne de marché n'atteint le chemin évalué qu'au moment de l'évaluation, par un chemin de code qui ne touche jamais aux packs de contexte,

de sorte qu'aucune voie IA ne peut lire les cotes au regard desquelles elle est mesurée. Les compositions officielles sont publiées environ 35 minutes avant le coup d'envoi, après la fermeture de la fenêtre de répartition, de sorte qu'aucune voie ne prévoit jamais avec une feuille d'équipe en main. Les artefacts de la couche d'affichage qu'un lecteur voit sur le tableau de bord public, traductions, compositions rendues et verdicts post-match, sont produits hors du chemin évalué et sont écrits après le verrouillage de la prédiction, de sorte que rien de ce qui est montré au public ne peut refluer dans une prévision évaluée. L'effet est que la contamination est empêchée par la forme du système plutôt que par un relecteur se souvenant de l'éviter : une voie ne peut pas voir ce à quoi l'architecture ne lui donne aucun chemin d'accès.

5 Métriques et méthodes statistiques

Toutes les métriques, leur opérationnalisation exacte, ainsi que les conditions sous lesquelles une affirmation comparative peut être formulée ont été verrouillées avec le préenregistrement (`statistical-tests.md`, lié par empreinte, Annexe B). Le code d'évaluation met en œuvre exactement ces définitions. Des tests unitaires vérifient des valeurs connues sur données synthétiques, et la première journée de matchs notée a fait l'objet d'un contrôle d'intégrité supplémentaire décrit ci-dessous.

5.1 Noter une prédiction unique

La métrique principale est le Brier score multiclasse sur le résultat à trois issues (domicile / nul / extérieur) [2]. Chaque voie émet un vainqueur prédit et une confiance $c \in [0, 100]$. Le vecteur de probabilité de la voie est construit selon la règle enregistrée. La classe prédite reçoit $c/100$, et la masse restante est répartie entre les deux autres classes *proportionnellement aux prieurs dévigés du marché à T−1 h*. Cette règle de répartition est déclarée dans l'enregistrement précisément pour éviter l'homme de paille de la répartition uniforme. Le prieur du marché n'entre jamais dans la prédiction propre d'une voie IA. Il est utilisé uniquement au moment de la notation, de façon identique pour chaque voie. Avec $\mathbf{p}$ le vecteur construit et $\mathbf{o}$ le résultat en encodage one-hot,

$$\text{Brier} = \sum_{k \in \{\text{H,D,A}\}} (p_k - o_k)^2,$$

donc plus bas est meilleur, un succès confiant l'emporte sur un succès timide, et un échec confiant coûte le plus cher.

5.2 Les deux espaces de probabilité

Deux constructions coexistent, toutes deux stockées et toutes deux étiquetées partout où elles apparaissent. L'*espace multiclasse* de la Section 5.1 porte le Brier score de saison, la métrique phare verrouillée. L'*espace binaire du vainqueur prédit* porte tout ce qui relève de la calibration. Pour chaque prédiction notée et non abstenue, $p = c/100$ est la masse que la voie a placée sur son vainqueur prédit et $y = 1$ si ce vainqueur était correct, sinon 0. Les diagrammes de fiabilité, l'ECE, la décomposition de Murphy et la finesse vivent tous dans cet espace binaire, qui est exactement l'espace de la question qu'un lecteur pose à un nombre de confiance (« quand il dit 70, a-t-il raison 70 pour cent du temps ? »), de sorte que la courbe publique et les métriques de calibration rapportées sont cohérentes entre elles par construction.

5.3 Métriques rapportées

Par voie, le banc d'essai rapporte ce qui suit. Les intervalles (bins), partout où ils apparaissent, sont dix intervalles de largeur égale sur $[0, 1]$, semi-ouverts avec l'intervalle supérieur fermé. n_b est le compte dans l'intervalle b, N le total, conf_b le p moyen dans l'intervalle, et acc_b le y moyen.

Brier score de saison (phare). Le Brier score multiclasse moyen de la Section 5.1. Le parfait est 0, et une répartition à trois issues toujours uniforme obtient 0,667.

Décomposition de Murphy. Le Brier score binaire se partitionne sur les intervalles [11] selon Brier = REL − RES + UNC, où

$$\text{UNC} = \bar{y}(1-\bar{y}), \qquad \text{RES} = \sum_b \tfrac{n_b}{N}\,(\text{acc}_b - \bar{y})^2, \qquad \text{REL} = \sum_b \tfrac{n_b}{N}\,(\text{conf}_b - \text{acc}_b)^2.$$

L'incertitude est la variance du taux de base de la tâche elle-même (aucun prévisionniste ne peut la modifier), la résolution récompense les prévisions qui séparent les victoires des défaites (plus haut est meilleur), et la fiabilité est la mauvaise calibration proprement dite (plus bas est meilleur). L'identité est exacte pour la prévision regroupée en intervalles. Le code d'évaluation la vérifie à chaque calcul et échoue bruyamment en cas de violation, et le Brier score brut (par point) est rapporté comme valeur phare aux côtés de la décomposition par intervalles.

ECE. Erreur de calibration attendue pondérée par les comptes sur les mêmes intervalles :

$$\text{ECE} \;=\; \sum_b \frac{n_b}{N}\,\big|\text{acc}_b - \text{conf}_b\big|.$$

La lecture simple : une ECE de 0,10 signifie que, en moyenne sur ses prédictions, la confiance déclarée de la voie et son taux de succès réalisé divergent de dix points de pourcentage. Zéro est une calibration parfaite.

Finesse. Mise en œuvre comme la variance des probabilités de prévision dans l'espace binaire, $\text{mean}\big((p-\bar{p})^2\big)$. Une voie qui dit toujours 55 a une finesse quasi nulle indépendamment de sa calibration, et la finesse est ce qui distingue un lâche bien calibré d'un prévisionniste bien calibré. Écart divulgué : le texte enregistré formulait la finesse comme la distance moyenne au prieur uniforme, mais la mise en œuvre mesure la dispersion autour de la prévision moyenne propre à la voie. Les deux capturent la « volonté de s'engager ». La forme mise en œuvre est celle qui est rapportée, et la formulation enregistrée est récupérable à partir du jeu de données par prédiction publié.

Intervalles bootstrap. Intervalles de confiance à 95 % à partir de 1,000 rééchantillonnages avec le match comme bloc de rééchantillonnage. Le même ensemble de matchs rééchantillonné est appliqué à chaque voie au sein d'une itération, préservant la corrélation entre voies de sorte que les intervalles par voie soient comparables, et le générateur est initialisé avec une constante fixe et publiée de sorte que chaque intervalle de cet article se reproduise bit pour bit à la ré-exécution.

Taux de succès. Fraction des vainqueurs prédits corrects. Contexte seulement, jamais une valeur phare. Un devineur audacieux peut afficher un taux de succès élevé tout en étant mal calibré, ce qui est la confusion que ce banc d'essai existe pour séparer.

Brier score stratifié par phase. Les strates enregistrées sont la phase de groupes, les seizièmes de finale, les huitièmes de finale, les quarts de finale, les demi-finales avec le match pour la troisième place, et la finale. Note de mise en œuvre divulguée : le pipeline publié réduit les strates éliminatoires en une seule division groupes contre éliminatoires parce que les comptes par phase sont trop minces pour des intervalles stables. Les strates enregistrées plus fines restent calculables à partir du jeu de données par prédiction publié.

Un diagramme de fiabilité accompagne les résultats d'ECE (Section 6). Les abstentions sont rapportées par voie (choix effectués / refusés) afin que les comparaisons de Brier score se fassent à conditions égales. Une référence destinée au lecteur définissant chaque voie et chaque métrique en termes simples est fournie à l'Annexe E.

5.4 Conditions d'affirmation préengagées

Trois conditions, fixées avant l'existence de toute donnée, régissent ce que cet article est autorisé à affirmer. **(i)** Une voie est « mieux calibrée » qu'une autre uniquement si son intervalle de confiance de Brier score hors échantillon et celui de la voie de comparaison ne se chevauchent pas. Des intervalles qui se chevauchent sont rapportés comme aucune différence résoluble. **(ii)** L'affirmation d'apprentissage de Niveau 2 requiert *tout* ce qui suit : le Brier score hors échantillon de la voie bouclée meilleur que celui de son témoin avec des intervalles ne se chevauchant pas, *ou* une captation de motif ensemencé passant de façon démontrable par la couche de la boucle d'apprentissage, *et* au moins une instance concrète de changement de comportement (le même motif, décisions pré-boucle contre post-boucle différant) documentée dans cet article. **(iii)** Une asymétrie d'abstention dépassant 10 points de pourcentage entre les voies comparées disqualifie la comparaison directe de Brier score. L'asymétrie est rapportée à la place. De plus, aucune comparaison n'est publiée en dessous de $N = 30$ matchs notés par voie, et les résultats de la fenêtre de développement sont étiquetés comme données de développement partout où ils apparaissent.

5.5 Intégrité de l'évaluation

Deux mécanismes d'intégrité font partie de la conception enregistrée. Premièrement, lors de la première journée de matchs notée, le Brier score a été calculé à la main de façon indépendante pour les cinq sorties requises et recoupé avec le pipeline de production. Le contrôle a été exécuté à la première journée (11 juin 2026) sur les huit voies du match d'ouverture, avec un accord à trois décimales sur chaque sortie (Annexe B). Deuxièmement, pour le jugement de qualité narrative de M2, l'accord inter-juges est rapporté comme l'α de Krippendorff moyen par paire. Rapporté, non réparé : le désaccord entre juges est une donnée, non un bruit à moyenner silencieusement.

Toutes les valeurs rapportées pour ces métriques sont générées par le pipeline d'évaluation publié et insérées depuis `data/`, jamais saisies à la main. Les classements hors échantillon et de tournoi complet, la décomposition de Brier, les résultats stratifiés par phase et le diagramme de fiabilité apparaissent avec les résultats (Section 6).

6 Résultats

Règle de lecture. L'enregistrement fixe un test délibérément strict pour une unique affirmation par paire sur la fenêtre réservée : deux voies n'y sont dites « différentes de façon tranchée » que lorsque leurs intervalles de bootstrap Brier score sur l'ensemble réservé ne se chevauchent pas. Sur l'ensemble réservé de vingt matchs, chaque paire se chevauche, ce que renvoie un échantillon aussi petit sur une tâche fortement dominée par le hasard, si bien que nous ne formulons aucune affirmation de significativité par paire unique sur la fenêtre réservée et lisons son ordonnancement comme directionnel. Le tournoi complet de 104 matchs, où l'échantillon est d'un ordre de grandeur plus grand, est là où l'ordonnancement de calibration se lit : l'ordonnancement ponctuel y est clair, le conseil le plus bas sur ECE de toutes les machines, même si son intervalle de bootstrap chevauche encore celui des autres, si bien que nous présentons cette avance comme directionnelle plutôt que comme une séparation par paire tranchée. Nous rapportons les deux fenêtres, sommes explicites tout du long sur la fenêtre sur laquelle repose un énoncé donné, et tenons chaque ordonnancement à la force que ses intervalles autorisent.

Simplifications établies durant la fenêtre (structurelles, non dépendantes des résultats). Aucune voie ne s'est abstenue sur un match, si bien que toutes les comparaisons Brier score sont de même nature par construction et que le disqualificatif d'asymétrie d'abstention enregistré ne se déclenche jamais. La clause de fenêtre de développement (§5 de l'enregistrement) n'a jamais été nécessaire : le verrouillage OSF a précédé le premier coup d'envoi, si bien que chaque match noté est un match de fenêtre notée et qu'aucune strate de fenêtre de développement n'existe. L'ensemble réservé (les 20 derniers matchs par coup d'envoi) relève entièrement de la phase à élimination directe, conséquence du point de gel M6 tombant à la frontière groupes/ élimination directe. Nous le signalons là où cela pèse sur l'interprétation.

6.1 Calibration principale (ensemble réservé)

Le Tableau 3 rapporte chaque voie sur l'ensemble réservé : le Brier score multiclasse avec son intervalle par bootstrap par blocs à 95 %, l'ECE, la finesse, le taux de réussite et le coût.

Voie	*n*	Brier score	IC 95%	ECE	IC 95% ECE	Fin.	Réu.	Coût
Solo	20	0.422	[0.299, 0.536]	0.192	[0.110, 0.361]	0.008	75%	$8.22
Solo-Zero	20	0.375	[0.239, 0.519]	0.181	[0.071, 0.341]	0.012	80%	$1.06
Pipeline	20	0.430	[0.314, 0.549]	0.200	[0.125, 0.377]	0.010	75%	$21.93
Pipeline-Static	20	0.431	[0.316, 0.548]	0.202	[0.124, 0.377]	0.009	75%	$21.93
Council	20	0.425	[0.309, 0.546]	0.196	[0.120, 0.369]	0.011	75%	$15.03
Elo	20	0.503	[0.416, 0.581]	0.309	[0.190, 0.447]	0.006	65%	n/a
Market	20	0.457	[0.336, 0.578]	0.166	[0.119, 0.341]	0.019	65%	n/a
Human	20	0.483	[0.317, 0.655]	0.127	[0.056, 0.325]	0.006	65%	n/a

Table 3 – Résultats sur l'ensemble réservé, les 20 derniers matchs (Section 3, M6). Les intervalles sont des bootstraps par blocs à 1000 rééchantillonnages (bloc = match, toutes les voies rééchantillonnées ensemble). L'intervalle de l'ECE utilise le même rééchantillonnage à graine fixée en recalculant l'ECE par bacs. Tous les intervalles se chevauchent, si bien que l'ordonnancement est suggestif, non tranché. Le coût est la dépense d'API cumulée de la voie sur l'ensemble du tournoi, non par match. L'économie par match et le chiffre du coût marginal figurent à l'Annexe C.

Fait. Sur l'ensemble réservé, le jumeau privé de contexte Solo-Zero enregistre le plus bas Brier score multiclasse (0.375), devant chaque voie d'IA à contexte complet, la référence déterministe Elo et la ligne de clôture démargée du marché. Les cinq voies d'IA enregistrent un Brier score sur l'ensemble réservé plus bas que Market. L'intervalle à 95 % de chaque voie chevauche celui de toutes les autres.

Lecture des deux paires de jumeaux (elles ne sont pas le même type de paire). Deux paires de voies sont des jumeaux par conception, et elles doivent se lire différemment. Solo et Solo-Zero constituent une ablation véritable : la même architecture à appel unique, avec et sans la pile de contexte, si bien que leur écart mesure la pile de contexte. Pipeline et Pipeline-Static ne sont *pas* une ablation dans cette exécution. La boucle d'apprentissage censée les séparer n'a jamais été alimentée en production (Section 7), si bien que les deux voies ont reçu des entrées *identiques* sur chaque match, et tout écart entre elles est du bruit d'une exécution à l'autre, non un effet d'apprentissage. Nous rapportons leurs deux chiffres parce qu'ils ont été notés, mais la quasi-identité de Pipeline et Pipeline-Static sur l'ensemble réservé est exactement ce qu'une paire de même système devrait montrer, et elle borne le plancher de bruit au regard duquel chaque autre écart de ce tableau doit se lire. Elle ne porte aucune information sur l'apprentissage, et nous n'en tirons aucune affirmation d'apprentissage.

Interprétation. L'hypothèse enregistrée énonçait qu'aucune voie d'IA ne battrait la ligne de clôture sur l'exactitude brute. Sur l'ensemble réservé, mesuré par Brier score, l'ordonnancement va en sens inverse : les voies d'IA se situent en dessous de la ligne du marché, et la plus épurée d'entre elles, un appel unique sans pile de contexte, se situe la plus basse de toutes. Nous rapportons ceci comme la question préénoncée se résolvant contre l'hypothèse en direction. Nous ne prétendons *pas* qu'une voie soit significativement meilleure que le marché ou qu'une autre voie : les intervalles se chevauchent, si bien que l'énoncé honnête est qu'aucune élaboration d'architecture, ni aucun ajout de la pile de contexte, n'a produit une amélioration du Brier score qui survive à l'échantillon. Le contraste économique est la lecture la plus nette du résultat d'exactitude : la voie au plus bas Brier score a aussi coûté environ un vingtième des voies du pipeline.

L'exactitude était la mauvaise question, par conception. Les paragraphes ci-dessus rapportent le Brier score parce qu'il est la métrique principale verrouillée, mais le banc d'essai a été bâti pour comparer la calibration, non l'exactitude, et sur une tâche dominée par le hasard cette distinction est tout l'enjeu. Aucune architecture ne possède un résultat de pile ou face, si bien que le classement d'exactitude sur l'ensemble réservé est dominé par le bruit des matchs et, comme le montrent les intervalles, ne tranche rien. La question qu'une entreprise pose réellement à une prévision est différente et trouve sa réponse dans la calibration : quand le système énonce une confiance, ce nombre peut-il être digne de foi ? Une prévision qui dit trente pour cent et se trompe a rapporté son incertitude avec honnêteté. Une prévision qui dit soixante-dix pour cent et se trompe ne l'a pas fait. La sous-section suivante lit les voies sur cet axe.

6.2 Calibration : à quelle architecture la confiance peut-elle se fier

La calibration est l'axe que l'étude a été bâtie pour mesurer, et c'est là que les architectures donnent à un opérateur de quoi agir. Sur l'ensemble des 104 matchs (Tableau 7), l'architecture Council enregistre le plus bas ECE de toutes les machines testées (0.100), en dessous des architectures à appel unique et à pipeline, de la référence déterministe Elo et de la propre ligne de clôture du marché. Sa confiance annoncée suit la fréquence réalisée de plus près que celle de toute autre machine, ce qui est la propriété qu'un opérateur consomme lorsqu'une probabilité fournie par l'IA entre dans une décision réelle. Sur la

question opérationnelle à laquelle l'étude existe pour répondre, quelle architecture rapporte la confiance la plus digne de foi, les preuves désignent le conseil.

Nous rapportons la force de cette avance avec précision. Une fois que chaque ECE porte son intervalle de bootstrap, celui du conseil, $[0.076, 0.216]$, chevauche celui du marché $[0.066, 0.209]$ et celui des autres voies d'IA, si bien que l'ordonnancement est directionnel plutôt que tranché : une avance ponctuelle réelle à l'intérieur d'intervalles qui se chevauchent, qu'un second tournoi affinerait. Ceci est la limite honnête sur la marge, et elle ne renverse pas la recommandation. Un meilleur directionnel reste un meilleur, et une organisation qui choisit une architecture à qui confier une prévision choisit sur exactement ce genre de preuves, un ordonnancement ponctuel clair avec des intervalles qu'un seul événement ne peut encore séparer. Le chevauchement lui dit avec quelle fermeté tenir le choix, non de l'abandonner. Ce que l'opérateur apprend, c'est à la fois la quantité à exiger, la calibration, et la forme qui l'a le mieux rapportée ici, le conseil.

Le diagramme de fiabilité (Figure 1) montre la forme derrière les chiffres. Les points d'un prévisionniste parfaitement calibré se situent sur la diagonale, où la confiance annoncée égale la fréquence observée. Un point sous la diagonale est de la surconfiance, le mode de défaillance qui coûte le plus à une entreprise, et un point au-dessus est de la prudence. Les courbes sont visiblement bruitées, ce qui est le portrait honnête de ce que supportent une centaine de matchs répartis en bacs et la raison pour laquelle les intervalles de l'ECE se chevauchent et l'avance du conseil se lit comme directionnelle plutôt que tranchée. La valeur du diagramme n'est pas de déclarer un vainqueur à l'œil mais de montrer qu'aucune voie n'affiche un biais de surconfiance grossier et systématique : les points se dispersent autour de la diagonale plutôt que de se situer constamment en dessous. L'avance en ECE que le conseil détient plus haut vit à l'intérieur de cette dispersion, ce qui est exactement la raison pour laquelle l'article la traite comme directionnelle et rapporte chaque intervalle aux côtés de chaque point.

6.3 Les deux hypothèses préénoncées

Deux schémas ont été consignés comme hypothèses le 2026-07-03, alors que 20 des matchs réservés n'étaient pas encore joués, si bien que leur verdict sur l'ensemble réservé est une question énoncée à l'avance plutôt qu'un récit a posteriori :

(a) Le jumeau privé de contexte égale ou bat les voies à contexte complet sur le Brier score. *Verdict : soutenu en direction, non tranchable en magnitude.* Sur l'ensemble réservé, Solo-Zero enregistre un Brier score plus bas que son jumeau à contexte complet Solo et que toute autre voie d'IA. La direction correspond à l'hypothèse, et la valeur de calibration marginale de la pile de contexte pour une architecture à appel unique est, au mieux, non positive sur ces preuves. Les intervalles se chevauchent, si bien que nous nous gardons de prétendre que la pile de contexte a *nui*. Nous prétendons seulement qu'elle n'a pas aidé de façon mesurable, à un coût d'environ huit fois celui de l'appel privé de contexte.

(b) Council mène sur l'ECE sans mener sur le Brier score. *Verdict : non confirmé sur la fenêtre réservée.* Cette hypothèse a été préénoncée à propos de l'ensemble réservé de vingt matchs spécifiquement, et cette fenêtre est trop petite pour la confirmer : parmi les voies aveugles aux sorties des autres voies, le plus bas ECE sur l'ensemble réservé appartient à Solo-Zero plutôt qu'à Council, le plus bas dans l'absolu appartient à Human (une référence non aveugle, Section 3, ne portant aucune affirmation comparative), et aucun de ces écarts d'ECE sur l'ensemble réservé ne survit au test d'intervalle de toute façon. Donc, sur la seule fenêtre réservée, l'hypothèse n'est pas confirmée. Elle ne l'est pas davantage sur

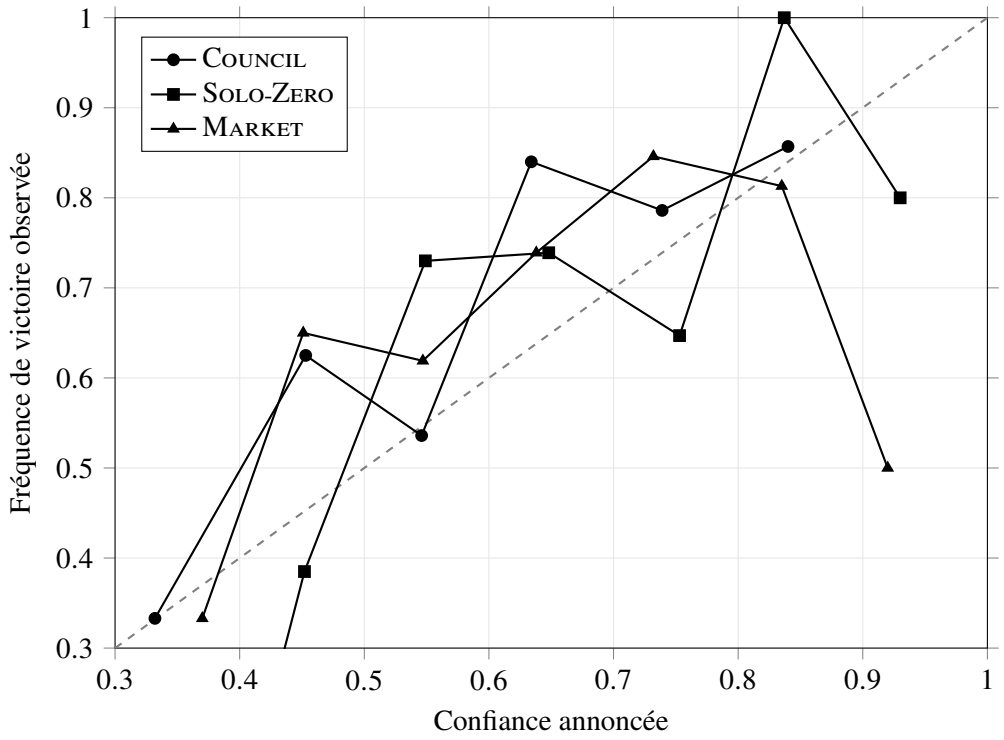

Figure 1 – Diagramme de fiabilité, tournoi complet. Chaque point est un bac de confiance : l'axe horizontal est la confiance annoncée moyenne, l'axe vertical la fréquence de victoire observée dans ce bac. La diagonale en pointillés est la calibration parfaite. Les points sous la diagonale sont surconfiants. Données du tournoi complet (développement plus ensemble réservé), montrées parce que les effectifs par bac sur le seul ensemble réservé sont trop minces pour être tracés de façon stable.

le tournoi complet, où l'échantillon est d'un ordre de grandeur plus grand, le conseil détient le plus bas ECE *ponctuel* de toutes les machines, même si son intervalle chevauche celui des autres (Section 6.2), si bien que l'avance y est directionnelle plutôt que tranchée. Le verdict honnête est donc partagé selon la fenêtre : le test strict de vingt matchs ne confirme pas l'hypothèse, tandis que le tournoi complet la soutient en direction, le conseil affichant la meilleure calibration de toutes les machines mais par une marge qu'un seul événement ne peut encore séparer du bruit. Nous rapportons les deux, et nous lisons l'ordonnancement du tournoi complet comme la réponse directionnelle à la question opérationnelle, le conseil est l'architecture la mieux calibrée ici, tout en la tenant à la force que les intervalles autorisent et pas au-delà.

6.4 Contrôles : le prix du contexte et le coût de la calibration

La paire de contrôle enregistrée isole la pile de contexte : Solo et Solo-Zero sont la même architecture à appel unique, l'une avec le pack de contexte complet et l'autre avec aucun. Sur l'ensemble réservé, le membre privé de contexte enregistre le plus bas Brier score (Tableau 3) à une petite fraction du coût. Lu avec la condition d'affirmation, ceci est un résultat nul sur la valeur du contexte pour un appel unique, non une démonstration que le contexte est nuisible. C'est toutefois une démonstration que le coût de la pile de contexte n'a acheté aucun gain de calibration mesurable ici. La décomposition de Brier (Tableau 4) situe d'où proviennent les scores des voies : la fiabilité (mauvaise calibration, plus bas est meilleur) et la

résolution (discrimination, plus haut est meilleur) sont rapportées par voie sur l'ensemble réservé.

Voie	Fiabilité	Résolution	Incertitude
Solo	0.047	0.033	0.188
Solo-Zero	0.035	0.010	0.160
Pipeline	0.054	0.034	0.188
Pipeline-Static	0.054	0.034	0.188
Council	0.051	0.034	0.188
Elo	0.130	0.123	0.227
Market	0.038	0.073	0.227
Human	0.021	0.034	0.227

Table 4 – Décomposition de Murphy du Brier score binaire sur l'ensemble réservé (Brier = reliability−resolution+ uncertainty). Dans l'espace du vainqueur prédit, l'incertitude de chaque voie est $\bar{y}(1-\bar{y})$ sur son propre taux de réussite $\bar{y}$, si bien qu'une voie qui a raison plus souvent porte un terme d'incertitude plus bas. C'est la part du score qu'aucune calibration ne peut changer, non l'habileté d'une voie.

6.5 Escalade : l'enchevêtrement Pipeline–Council

Le portillon de Pipeline escalade un match à faible confiance vers Council, et la sortie publiée de Pipeline sur un match escaladé *est* celle de Council. Les deux voies ne sont donc pas indépendantes. Le portillon s'est déclenché sur 15 des 20 matchs réservés (75 %) et 65 du tournoi complet (63 %), si bien que l'enchevêtrement n'est pas une note de bas de page. Nous rapportons Pipeline de trois façons (Tableau 5) pour le rendre mesurable : tel que publié (le mélange qui a été noté), pré-escalade (la sortie propre du pipeline, à partir de la prédiction pré-escalade consignée) et le sous-ensemble adopté-Council seul.

Vue Pipeline	Brier score hors éch.	Brier score complet
Tel que publié (Council si escaladé)	0.430	0.481
Pré-escalade (sortie propre du Pipeline)	0.429	0.494
Sous-ensemble adopté-Council seul	0.518 ($n = 15$)	0.546 ($n = 65$)

Table 5 – Le triple rapport d'escalade, Brier score multiclasse. « Tel que publié » est ce qui a été noté. « Pré-escalade » est ce que Pipeline aurait noté sur sa propre sortie. Le « sous-ensemble adopté-Council » est les matchs escaladés seuls, que le portillon a sélectionnés précisément parce qu'ils étaient à faible confiance, si bien qu'un Brier score plus haut sur ce sous-ensemble reflète la difficulté des matchs, non un défaut de voie.

Interprétation. Le portillon a acheminé les matchs qu'il a jugés les plus difficiles vers Council, si bien que le sous-ensemble adopté porte les matchs les plus difficiles par construction et score moins bien que la moyenne du tournoi sur les deux ensembles. Les colonnes tel que publié et pré-escalade diffèrent modestement, ce qui borne à quel point la machinerie d'escalade a changé la calibration notée de Pipeline : sur l'ensemble réservé, les deux sont à l'arrondi près, si bien que l'escalade n'a ni aidé ni nui au Brier score sur l'ensemble réservé du pipeline à un degré tranchable.

6.6 Résultats stratifiés par phase et la couche de jugement

Parce que l'ensemble réservé relève entièrement de l'élimination directe, le contraste groupes contre élimination directe n'est visible que sur le tournoi complet, rapporté ici comme contexte de développement (Tableau 6). Le jugement de qualité narrative (M2) est une couche distincte des résultats de calibration ci-dessus : il note le texte de raisonnement, non les prédictions, et son accord inter-juges est tombé sous le seuil enregistré. Ce résultat, et ses conséquences pour ce que les scores M2 peuvent soutenir, est rapporté avec les autres limitations de mesure dans la Section 10 plutôt qu'ici, afin que les résultats de calibration, qui sont notés au regard de résultats réels, ne soient pas lus au travers du désaccord de la couche de jugement.

	Groupes		Élim. directe	
Voie	Brier score	*n*	Brier score	*n*
Solo	0.212	72	0.204	32
Solo-Zero	0.211	72	0.187	32
Pipeline	0.224	72	0.207	32
Pipeline-Static	0.221	72	0.207	32
Council	0.222	72	0.204	32
Elo	0.243	72	0.230	32
Market	0.220	72	0.194	32
Human	0.221	72	0.210	32

Table 6 – Brier score binaire stratifié par phase, tournoi complet (contexte de développement). L'ensemble réservé relève entièrement de l'élimination directe, si bien que cette répartition n'est pas un résultat sur l'ensemble réservé et ne porte aucune affirmation principale.

6.7 Résultats du tournoi complet, la base de preuves de la calibration

Le Tableau 7 rapporte chaque voie sur le tournoi complet, les 104 matchs. Ceci est la base de preuves du résultat de calibration de la Section 6.2 : l'enregistrement fixe l'ensemble réservé comme une fenêtre de contrainte préengagée, et l'événement complet est le plus grand échantillon au regard duquel l'ordonnancement ponctuel de calibration se lit. Sur l'exactitude, les deux fenêtres racontent la même histoire dans leurs extrêmes, la voie privée de contexte la plus basse sur le Brier score et la référence déterministe la plus haute, même si le marché se situe en milieu de peloton sur le tournoi complet plutôt que derrière les voies d'IA comme il le fait sur la tranche à élimination directe réservée, une différence que nous attribuons à la composition exclusivement à élimination directe de l'ensemble réservé plutôt qu'à un changement dans ce que font les systèmes. Sur la calibration, le tournoi complet porte l'ordonnancement ponctuel de la Section 6.2, le conseil le plus bas sur ECE de toutes les machines. Son intervalle de bootstrap chevauche celui des autres, si bien que l'avance est directionnelle plutôt que tranchée sur cette plus grande fenêtre, ce qui est la force que nous lui revendiquons et rien de plus.

Voie	*n*	Brier score	IC 95%	ECE	IC 95% ECE	Fin.	Réu.	Coût
Solo	104	0.472	[0.409, 0.545]	0.114	[0.074, 0.214]	0.015	66%	$8.22
Solo-Zero	104	0.453	[0.379, 0.532]	0.135	[0.085, 0.219]	0.018	68%	$1.06
Pipeline	104	0.481	[0.416, 0.555]	0.112	[0.084, 0.222]	0.014	67%	$21.93
Pipeline-Static	104	0.477	[0.414, 0.550]	0.136	[0.097, 0.237]	0.014	67%	$21.93
Council	104	0.477	[0.411, 0.553]	0.100	[0.076, 0.216]	0.017	67%	$15.03
Elo	104	0.534	[0.488, 0.584]	0.114	[0.067, 0.212]	0.007	62%	n/a
Market	104	0.468	[0.394, 0.553]	0.104	[0.066, 0.209]	0.023	68%	n/a
Human	104	0.505	[0.424, 0.597]	0.074	[0.042, 0.170]	0.010	62%	n/a

Table 7 – Résultats du tournoi complet, les 104 matchs. Ceci est la base de preuves de la lecture de calibration (Section 6.2). Le conseil affiche le plus bas ECE de toutes les machines. Chaque ECE porte aussi son intervalle de bootstrap, et ceux-ci se chevauchent, si bien que l'avance est directionnelle plutôt que tranchée. L'ensemble réservé (Tableau 3) est la fenêtre de contrainte préenregistrée, rapportée aux côtés. Le coût est la dépense d'API de tournoi cumulée par voie, non par match (Annexe C).

7 Le test de la boucle d'apprentissage

7.1 Ce qui a été enregistré

La question de Niveau 2 enregistrée demandait si une architecture dotée d'une boucle d'apprentissage active démontre un apprentissage véritable plutôt qu'une récupération, mesurée par (a) l'écart hors échantillon entre Pipeline et son jumeau à boucle désactivée Pipeline-Static sous les conditions d'affirmation préengagées (Section 5.4), et (b) la captation de deux motifs synthétiques scellés injectés dans la pile de contexte à partir du match 1.

7.2 Ce qui s'est réellement exécuté

L'injection de motifs ensemencés (M5) s'est exécutée comme enregistrée à partir du match 1 avec rotation déterministe de l'emplacement. Le côté écriture de la boucle ne l'a pas fait : le composant d'évaluation qui peuple la couche de la boucle d'apprentissage n'a pas été déployé pendant le tournoi, de sorte que la couche de contexte L7 était vide pour chaque match, pour les deux voies de la paire témoin. Deux conséquences en découlent par construction, et nous les énonçons clairement :

- Pipeline et Pipeline-Static ont reçu des entrées *identiques* à chaque match. La paire fonctionne donc comme un *témoin de réplication* de même architecture, une estimation non planifiée mais utilisable de la variance d'une exécution à l'autre dans des conditions identiques, et tout écart observé entre les jumeaux borne le plancher de bruit par rapport auquel les affirmations au niveau de l'architecture ailleurs dans cet article doivent être lues. Elle ne porte aucune information sur l'apprentissage.
- La captation de motifs ensemencés *à travers la couche de la boucle* est impossible par construction. L'analyse de la captation de motifs est en conséquence limitée à la détection dans le texte de sortie (le raisonnement d'une voie a-t-il fait émerger les motifs ensemencés du contexte injecté ?), ce qui teste l'attention au contexte, non l'apprentissage.

7.3 Le verdict de Niveau 2 que ce changement de conception permet

Les conditions d'affirmation préengagées requéraient un écart de jumeaux hors échantillon avec des intervalles ne se chevauchant pas *ou* une captation passant de façon démontrable par la couche de la boucle, *et* une instance concrète de changement de comportement. La boucle n'ayant jamais été active, ni la preuve de passage ni le changement de comportement ne sont observables : l'affirmation de Niveau 2 enregistrée est rapportée comme **non testable telle que conçue**, non comme un résultat nul. La distinction importe : un résultat nul dirait « la boucle n'a pas aidé ». L'énoncé honnête est « la boucle n'a jamais été exercée, donc ce banc d'essai ne peut rien dire ».

7.4 L'écart de réplication, et ce qu'il borne

Parce que les jumeaux ont reçu des entrées identiques, l'écart entre eux est une lecture directe de la variance d'une exécution à l'autre, le plancher de bruit du banc d'essai. Sur l'ensemble hors échantillon, Pipeline et Pipeline-Static enregistrent des valeurs de Brier score multiclasse qui diffèrent d'environ un millième (Table 3), et sur le tournoi complet d'environ quatre millièmes (Table 7). Les deux écarts sont d'un ordre de grandeur plus petits que toute différence au niveau de l'architecture ailleurs dans l'étude, et tous deux se situent bien à l'intérieur des intervalles bootstrap qui se chevauchent. C'est la raison quantitative pour laquelle l'article n'affirme aucune différence d'architecture résolue : l'écart entre jumeaux identiques montre qu'une différence de quelques millièmes de Brier score est ce que deux exécutions du *même* système produisent, de sorte qu'un écart de cette taille entre deux systèmes *différents* ne porte aucun signal. Le témoin de réplication n'était pas celui que nous avons enregistré, mais il gagne sa place en calibrant quelle part de tout écart observé est du bruit.

7.5 Pourquoi la boucle n'a pas été livrée, et ce qu'une ré-exécution corrigée exige

L'absence est un fait de portée d'ingénierie, non un changement silencieux de protocole. Le composant qui écrit les entrées de la boucle d'apprentissage dans la couche de contexte L7 n'a pas été déployé pendant le tournoi, de sorte que L7 était nul pour chaque match et pour les deux voies de la paire témoin. Aucune invite, pondération de rubrique, épingle de modèle, règle de construction du pack de contexte ou règle de notation n'a été modifiée : la conception enregistrée a été exécutée telle qu'écrite, avec une couche enregistrée laissée non peuplée parce que le composant d'écriture derrière elle n'a pas été livré. Nous le rapportons ouvertement plutôt que de restreindre l'affirmation silencieusement autour de cela. L'injection de motifs ensemencés (M5) s'est bien exécutée à partir du premier match comme enregistrée, mais sans entrées de boucle, la seule forme observable de captation est dans le texte de sortie, ce qui teste l'attention au contexte plutôt que l'apprentissage à travers la boucle. Nous rapportons l'injection comme exécutée et laissons le balayage de détection dans le texte de sortie au jeu de données ouvert plutôt que d'affirmer un résultat d'apprentissage que la conception ne peut soutenir. Un banc d'essai successeur corrigé exigerait que le composant d'écriture de la boucle soit actif à partir du premier match, de sorte que L7 porte de véritables entrées des matchs antérieurs et que la paire de jumeaux devienne l'ablation qu'elle était censée être. C'est le seul changement qui ferait passer la question de Niveau 2 de *non testable* à répondue.

8 Discussion

L'introduction soutenait que la question opérationnelle pour une organisation qui adopte l'IA n'est pas de savoir si un système a raison légèrement plus souvent qu'une solution de rechange, mais si la confiance qu'il rapporte peut être tenue pour fiable. Les résultats nous permettent de reformuler cette affirmation avec des données à l'appui, et ces données pointent dans une seule direction sur l'axe qui compte et dans aucune direction sur l'axe qui ne compte pas.

8.1 Ce que le résultat de calibration signifie pour la délégation à l'IA

Sur le plan de l'exactitude, l'étude ne tranche rien, et c'est le résultat correct pour une tâche dominée par le hasard. Aucune architecture n'a prédit de façon fiable plus de matchs qu'une autre, les intervalles hors échantillon se chevauchent sans exception, et même l'inversion de notre a priori déclaré, à savoir chaque voie d'IA sous la ligne de clôture du marché sur le Brier score hors échantillon, est une lecture directionnelle que nous refusons de gonfler en une affirmation de signification. Une organisation qui choisirait une architecture de prévision par IA sur son taux de réussite choisirait sur du bruit.

La calibration se comporte différemment, et c'est l'axe que l'étude a été bâtie pour isoler. Sur l'ensemble du tournoi, le conseil a rapporté la confiance la plus fiable de toutes les machines : ses probabilités déclarées suivaient les fréquences réalisées de plus près que les architectures à appel unique et à pipeline, que la référence déterministe, et que la ligne du marché elle-même, et il affiche le plus faible ECE des huit voies, la référence humaine mise à part. C'est le résultat sur lequel un adoptant devrait agir, parce que la calibration est la propriété qu'un adoptant consomme réellement. Une prévision de pipeline pondérée, une réserve établie contre un chiffre ajusté au risque, une décision d'auto-approuver au-dessus d'un seuil de confiance et d'escalader en dessous, chacune de ces choses hérite de la qualité de l'estimation de confiance, non du taux de succès brut. Un système dont le soixante-dix pour cent signifie soixante-dix pour cent peut se voir déléguer une classe de décisions qu'un système qui devine mieux mais surestime sa confiance ne peut pas recevoir, et sur l'ensemble de ce tournoi, le conseil est le système qui a gagné cette confiance.

Nous énonçons la force de la lecture avec exactitude. Elle repose sur l'ensemble du tournoi, 104 matchs en direct notés par un pipeline préenregistré, ce qui constitue une base substantielle pour la direction qu'elle donne. Sur un événement dominé par le hasard, les intervalles bootstrap sur l'ECE se chevauchent, de sorte que l'avance du conseil est directionnelle plutôt qu'une séparation tranchée, et nous la présentons comme telle : un résultat directionnel fort pour quiconque choisit une architecture à laquelle se fier, qu'un deuxième événement viendrait affiner. Le chevauchement est la limite honnête sur la marge, non une rétractation de la recommandation. Nous nommons la comparaison la plus serrée sans détour : l'erreur de calibration du conseil, 0.100, se situe quatre millièmes sous le 0.104 du marché, et notre propre contrôle (Section 7) place le plancher de bruit d'une exécution à l'autre à environ la même magnitude. Face au marché, l'avance du conseil est donc une égalité à la résolution que ce seul tournoi peut mesurer, et nous n'affirmons pas qu'elle bat le marché. Cette comparaison, toutefois, est celle qui ne se transfère pas. La raison pour laquelle les machines se regroupent près du marché est que le football dispose d'un marché de paris public profond qui tarife déjà chaque rencontre, une référence dure qui laisse peu de place pour se distinguer. Une entreprise qui prévoit son propre pipeline n'a pas un tel marché et pas de ligne de clôture. Il n'y a pas de probabilité en place contre laquelle faire égalité, seulement les résultats réels et lents de l'entreprise elle-même, de sorte que la prévision d'IA calibrée n'est pas comprimée contre une référence, elle est la référence dont l'entreprise manquerait autrement. L'avance du conseil sur les autres machines, celles entre lesquelles une entreprise choisirait réellement, est là où le résultat se pose, et c'est

une avance que la comparaison avec le marché ne crée ni ne retire.

8.2 Économie des architectures : la calibration ne s'achète pas avec la dépense

Le relevé des coûts affine le résultat plutôt qu'il ne le complique. L'appel unique sans contexte, le système le moins coûteux de l'étude par plus d'un ordre de grandeur, n'était pas moins bien calibré que les pipelines élaborés et, sur l'ensemble hors échantillon, a affiché le plus faible Brier score de tous. Quoi que la pile de contexte et les rôles séquentiels de spécialistes aient ajouté, ce n'était pas la calibration, et leur coût était réel. Le conseil, plus coûteux qu'un appel unique mais bien moins coûteux que les pipelines, est l'architecture qui a transformé sa dépense supplémentaire en la meilleure calibration de l'étude, tandis que les pipelines ont dépensé davantage et n'ont acheté aucun avantage de calibration sur l'appel le moins cher, ce qui est la lecture des coûts sous sa forme la plus tranchée. L'implication opérationnelle est que la dépense et la fiabilité ne sont pas le même axe. Une organisation peut payer pour une orchestration dont elle n'a pas besoin, et la façon de le savoir est de mesurer la calibration par dollar plutôt que de supposer qu'un pipeline plus élaboré est plus fiable. Ce banc d'essai l'a mesuré, et l'élaboration n'a pas payé sur cette tâche. Les coûts tout au long sont la dépense API cumulée du tournoi par voie, l'économie par match étant détaillée à l'Annexe C.

8.3 Ce que les contrôles ont enseigné

Les deux paires jumelles isolent deux questions. L'ablation véritable, l'appel unique avec et sans le pack de contexte, a retourné un résultat nul sur la valeur du contexte pour cette architecture : le jumeau sans contexte a tenu bon sur la calibration et mené sur l'exactitude, à une fraction du coût. Nous résistons à la lecture plus forte selon laquelle le contexte serait sans valeur, parce que les intervalles ne l'autorisent pas, mais les données ne donnent à une organisation aucune raison de supposer qu'une pile de contexte riche améliore la calibration d'un système à appel unique, et une raison claire de la tarifer avant de l'adopter. La seconde paire n'était pas une ablation dans cette exécution : la boucle d'apprentissage censée séparer le pipeline de son jumeau statique n'a jamais été alimentée (Section 7), de sorte que les deux ont reçu des entrées identiques et que leur quasi-identité est une mesure du bruit d'une exécution à l'autre, non un résultat sur l'apprentissage. Elle borne le plancher de bruit contre lequel chaque autre écart de l'étude doit être lu, et c'est la raison pour laquelle aucun de ces écarts n'est revendiqué comme tranché.

La lecture tient ici : sur une tâche bâtie pour tester si c'est la calibration plutôt que l'exactitude qui est l'axe à mesurer, la calibration est là où les architectures se séparent, le conseil rapportant la confiance la plus fiable de toutes les machines tandis que l'appel unique le moins cher est aussi bien calibré que le pipeline le plus coûteux. La marge est directionnelle sur un événement dominé par le hasard, et nous le disons, mais la réponse opérationnelle est assez claire pour agir dessus : mesurez la calibration, et sur ces données le conseil est la forme à laquelle se fier. La traduction appliquée de ce résultat, la façon dont un opérateur fait correspondre ce banc d'essai à une prévision d'affaires telle qu'un pipeline de ventes ou un plan de demande et les questions d'approvisionnement qui en découlent, fait l'objet de la section suivante.

9 Du banc d'essai à l'entreprise : choisir une architecture de prévision

Les matchs de football sont le banc d'essai. La classe de problèmes que ce banc d'essai traite réellement est toute prévision d'affaires récurrente présentant quatre propriétés : la décision se répète selon un

calendrier, le résultat est véritablement incertain, une échéance ferme force la prévision à se verrouiller avant que la vérité n'arrive, et un résultat vérifiable survient assez tôt pour le noter. Les pipelines de ventes, les plans de demande, les projections de flux de trésorerie, les prédictions de désabonnement et les engagements de date de livraison ont tous cette forme. Ce qui change entre eux, ce sont les noms, non la méthode.

9.1 Pourquoi la calibration est un chiffre d'affaires

L'instance la plus claire est le pipeline de ventes pondéré. Multiplier la taille d'une transaction par sa probabilité de conclusion et sommer est de l'arithmétique de revenu standard, et ce n'est *que* de l'arithmétique si les probabilités sont calibrées : 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. Lorsqu'un système d'IA se met à fournir ces probabilités, la question que ce banc d'essai pose à une voie de la Coupe du monde devient une question de finance mot pour mot : quand le système dit 70, a-t-il raison 70 pour cent du temps ? L'ECE est cette question sous la forme d'un seul chiffre (Annexe E.2), et une courbe de fiabilité en est la piste de vérification. Un opérateur qui adopte une confiance générée par IA sans mesurer la calibration n'a pas automatisé la prévision. Il a automatisé la production de chiffres non vérifiés.

Un lecteur de finance objectera qu'une entreprise dispose déjà d'un validateur contre lequel sa prévision peut être vérifiée, les résultats réels de la période précédente, et cette objection est correcte pour autant qu'elle aille. Un test rétroactif trimestriel contre le revenu réalisé est une véritable vérification de calibration, et là où une entreprise dispose d'un historique clos suffisant, elle devrait en exécuter un. Ce que la calibration ajoute est ce que le test rétroactif ne peut pas donner. La Coupe du monde disposait d'une prévision de référence publique et liquide dans le marché des paris, et une entreprise n'en a pas, de sorte que la seule vérité terrain d'une entreprise est ses propres résultats réels, qui arrivent lentement et de façon agrégée. La calibration est le signal disponible avant que ces résultats réels ne s'accumulent, lorsqu'un nouveau segment, un nouveau produit ou une jeune entreprise a trop peu de transactions closes pour faire un test rétroactif (le plancher de N que cette étude respecte), et c'est un signal *par prévision* qu'un test rétroactif trimestriel ne peut pas produire : les résultats réels disent à un opérateur que le chiffre du dernier trimestre a manqué sa cible, tandis qu'une courbe de fiabilité lui dit *quelles* bandes de confiance sont celles qui mentent. L'affirmation n'est pas qu'une entreprise ne peut pas valider ses prévisions. C'est que la calibration est la vérification d'honnêteté qu'elle peut exécuter avant d'avoir assez de résultats réels, et à une résolution que ses résultats réels n'atteignent jamais.

9.2 La correspondance

Deux rangées de la Table 8 méritent d'être soulignées. Premièrement, le processus en place joue le rôle de la ligne de clôture : c'est la référence à battre, il incorpore de l'information que les modèles n'ont pas, et l'a priori enregistré de cette étude (s'attendre à ce que l'agrégat en place gagne sur l'exactitude brute) est la posture de départ honnête pour toute initiative de prévision par IA. L'objectif réaliste à court terme n'est pas de surpasser en prédiction le processus en place. C'est de l'égaler avec une confiance mieux calibrée, moins coûteuse et vérifiable. Deuxièmement, la rangée Solo-Zero est l'expérience la moins coûteuse que la plupart des organisations n'exécutent jamais : le même modèle avec le contexte retiré. Si le jumeau sans contexte se comporte de façon comparable, l'intégration de données était un coût, non un signal, et cela vaut la peine d'être su avant que le pipeline ne soit construit, non après.

Élément du banc d'essai	Équivalent d'affaires
Un match	Un événement de prévision (une transaction, un SKU-mois, une facture)
Verrouillage au coup d'envoi (T−75 min)	L'échéance de verrouillage de la prévision
La cible des 90 minutes	La définition verrouillée du résultat (conclu-gagné à la date X)
MARKET (cotes de clôture)	La prévision en place (représentants, finance, le processus actuel)
ELO	La référence déterministe bon marché (taux de base historiques)
HUMAN	Le jugement instinctif de l'opérateur expérimenté
Le pack de contexte	Champs CRM, notes d'appels, signaux de courriel, données externes
SOLO-ZERO (le jumeau)	« Tout ce contexte aide-t-il vraiment ? », mesuré
La porte d'escalade	Prévisions à faible confiance acheminées vers une revue humaine
Abstention	Le droit enregistré de dire « aucune prévision » sans pénalité
Plancher $N = 30$, ensemble hors échantillon	Combien d'historique avant que le tableau de bord soit fiable

TABLE 8 – La table de traduction. Chaque élément méthodologique du banc d'essai a un équivalent opérationnel direct. Rien dans la méthode n'est spécifique au football.

9.3 Un cas d'usage détaillé : la prévision de ventes trimestrielle

Considérons une entreprise qui prévoit son revenu trimestriel à partir d'un pipeline de transactions. La méthode du banc d'essai se transfère comme suit. L'*événement* est chaque transaction ouverte au-dessus d'un seuil de matérialité, notée sur la question verrouillée « conclu-gagné à la fin du trimestre, oui ou non » (la cible de match à trois issues devient un résultat à deux issues, et chaque métrique de la Section 5 a une forme binaire, la machinerie de calibration étant déjà binaire). Le *verrouillage* est la date mensuelle de verrouillage de la prévision : la probabilité de l'IA par transaction est enregistrée avant la réunion de verrouillage et jamais révisée, exactement comme les prédictions se verrouillent avant le coup d'envoi. Les *voies* sont les options de construction : un généraliste à appel unique lisant le dossier CRM de la transaction et sa trace de communications (le motif SOLO), un pipeline par étapes dans lequel une étape quantitative lit les métriques de la transaction, une étape contextuelle lit les communications, et une étape de synthèse verrouille (le motif PIPELINE, avec sa porte d'escalade enregistrée acheminant toute transaction sous 60 de confiance vers le directeur des ventes), et un comité multi-modèles pour le carnet critique en calibration (le motif COUNCIL). Les *contrôles* sont les mêmes jumeaux : une exécution sans contexte tarifant l'intégration CRM, et, si une boucle d'apprentissage est revendiquée, un jumeau à boucle désactivée, parce que la Section 7 de cette étude est le récit édifiant d'une boucle d'apprentissage revendiquée sans preuve instrumentée qu'elle s'exécute. Les *références* sont le propre jugement du représentant, le taux de conclusion historique par étape (le plancher déterministe), et le pipeline pondéré en place. Après un trimestre, une entreprise au volume de transactions ordinaire franchit le plancher $N = 30$ et possède sa première courbe de fiabilité : autrement dit, elle sait si le 70 de son IA signifie 70.

Le même squelette se re-nomme vers les problèmes voisins : la planification de la demande (événement = SKU-mois, résultat = demande dans la tolérance), le flux de trésorerie (événement = facture, résultat = payée à la date), le désabonnement (événement = compte-trimestre), et les engagements de livraison (événement = commande, résultat = à temps). La méthode est indifférente au nom.

9.4 Choisir l'architecture

Le tournoi transforme les cinq architectures en un menu tarifé parmi lequel un opérateur peut choisir. Lus en regard du coût, les résultats de calibration se trient en trois paliers.

La frontière coût-calibration. Tracer l'erreur de calibration de chaque voie en regard de son coût de tournoi place deux voies sur la frontière efficiente et le reste au-dessus. L'appel unique sans contexte (Solo-Zero) se situe à un dollar et des poussières ($1.06) avec une erreur de calibration de 0.135, le système le moins cher qui prévoit tout de même. Le conseil se situe à $15.03 avec la plus faible erreur de calibration de toutes les machines (0.100). Tout ce qui se trouve entre les deux est dominé : l'appel unique à contexte complet (Solo, $8.22) n'est pas mieux calibré que le conseil et ne coûte la moitié moins qu'en étant pire, et les deux pipelines ($21.93 chacun) sont les systèmes les plus coûteux de l'étude et n'achètent rien pour la prime, l'un d'eux affichant la pire calibration machine de toutes (Pipeline-Static, 0.136). Un opérateur tarife les architectures à partir de cette frontière. La question n'est jamais quel système est le plus élaboré, c'est quel point de la courbe coût-calibration la décision justifie.

La dépense en contexte doit être validée par tâche. La seule ablation propre de l'étude, Solo contre Solo-Zero, est la même architecture avec et sans le pack de contexte à sept couches. Le jumeau sans contexte n'était pas moins bien calibré au huitième du coût. La leçon opérationnelle n'est pas que le contexte est sans valeur, c'est qu'une pile de contexte est un coût dont il faut démontrer qu'il paie avant que le budget d'intégration ne soit engagé, parce que sur cette tâche il ne l'a pas fait. À un fournisseur qui propose une couche riche de récupération et d'enrichissement, il faut demander la différence de calibration qu'elle achète, mesurée, non supposée.

Quand la prime du conseil vaut la peine d'être payée. Le conseil coûte environ quatorze fois un appel sans contexte et retourne la meilleure calibration de l'étude. Cette prime vaut la peine d'être payée lorsque la décision est critique en calibration, lorsqu'un chiffre de confiance guide un engagement irréversible ou coûteux tel qu'une marge de crédit, une réservation de capacité, ou une orientation de revenu à un conseil d'administration. Pour ceux-là, le conseil est l'architecture vers laquelle se tourner, en tenant compte de la réserve directionnelle que ce document énonce tout au long. Pour une prévision à fort volume et à faible enjeu où une probabilité approximative suffit, l'appel sans contexte est le plancher rationnel, et l'argent économisé est réel. Le conseil est une recommandation circonscrite aux décisions qui peuvent se le permettre, non une valeur par défaut pour chaque prévision.

Une porte d'escalade est une décision architecturale, non une réflexion de sûreté après coup. La voie Pipeline a acheminé les matchs qu'elle jugeait elle-même les plus difficiles vers une étape de revue plus lourde, et cette porte est enchevêtrée avec ses résultats. C'est un choix de conception aux conséquences mesurables de coût et de calibration, non un ajout rapporté. Un opérateur qui compose un système multi-étapes devrait traiter chaque règle d'acheminement et d'escalade comme une voie à part entière, à noter, parce qu'elle change le chiffre que l'entreprise consomme.

L'architecture émergente. Les propres composants de l'étude suggèrent un système qu'elle n'a pas exécuté : un appel de base sans contexte pour le coût, une étape de conseil réservée au sous-ensemble critique en calibration, et une porte d'escalade notée plutôt que supposée. Ce que cet assemblage coûterait par point de calibration face au processus en place d'une entreprise est la mesure qu'un premier tournoi interne devrait être bâti pour retourner. Ce document donne la méthode pour l'exécuter, non le plan directeur fini.

9.5 La liste de contrôle de l'opérateur

Six questions, mot pour mot, pour tout fournisseur ou équipe interne qui propose un système de prévision par IA. Chacune trouve sa réponse dans la méthode de ce document, et une équipe qui ne peut pas y répondre n'a pas mesuré ce qu'elle vend.

1. « Montrez-moi la courbe de fiabilité des prédictions de votre système à $N \geq 30$ événements notés. Pas l'exactitude. La courbe. »
2. « Quand votre système rapporte 70 pour cent de confiance, à quoi cela correspond-il à la mesure ? Quel est son ECE, sur combien d'événements, et sur quelle fenêtre hors échantillon ? »
3. « Quel est le seuil de confiance sous lequel une prévision est acheminée vers un humain, et où cela est-il journalisé ? »
4. « Qu'a mesurablement ajouté l'intégration du contexte ? Montrez-moi l'ablation sans contexte. »
5. « Pouvez-vous rejouer n'importe quelle prévision passée bit pour bit, 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 quel est le coût par point de calibration gagné sur notre processus actuel ? »

9.6 L'exécuter sur votre propre pipeline

Le banc d'essai est une méthode reproductible, et sa forme minimale viable tient à l'intérieur d'un trimestre d'affaires. 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. Portez le processus en place et un taux de base déterministe comme références. Exécutez le jumeau sans contexte. Ne notez rien publiquement sous $N = 30$ événements. Réservez la dernière ligne droite pour les affirmations qui comptent. Voilà tout le protocole, transplanté. Les instruments sont ceux définis dans ce document (Section 5, Annexe E), le coût de leur exécution est faible face au coût d'un trimestre d'engagements mal calibrés, et le résultat est la chose que cette étude soutient qu'aucun opérateur ne devrait déléguer sans elle : une réponse mesurée à la question de savoir si le chiffre que la machine rapporte mérite la confiance que l'arithmétique y place discrètement.

10 Limites

- **Confusion architecture–modèle.** SOLO et PIPELINE fonctionnent sur une seule famille de modèles alors que COUNCIL en mélange trois, par conception. Nos affirmations portent sur les systèmes figés tels que spécifiés, jamais sur des architectures dans l'abstrait.
- **Tournoi unique, domaine dominé par le hasard.** La généralisation au-delà des tâches de prévision en direct, fortement soumises au hasard, est argumentée, non démontrée.
- **Référence humaine non aveugle.** HUMAN n'est pas une base de comparaison contrôlée. Le prévisionniste était l'opérateur du projet et n'était pas aveugle aux prédictions des voies d'IA ni aux résultats en cours, de sorte que tout avantage de calibration montré par l'humain est confondu par l'accès à l'information et ne porte aucune affirmation d'opposition humain contre machine (la divulgation complète figure à la Section 3). Il s'agit d'une personne, non d'un panel. L'option enregistrée d'ajouter un panel de rédacteurs nommés en cours de tournoi (M7) n'a pas été exercée, de sorte qu'aucun choix de rédacteur n'a été recueilli.

- **Simplifications de construction de l'Elo.** Les adversaires hors graphe entrent à 1500. L'ajustement domicile déclaré est appliqué aux tournois en terrain neutre dans la fenêtre historique (divulgué dans le registre).
- **Ce que nous n'avons pas répliqué.** Les solides bases de comparaison publiées que nous citons (par exemple, le logit ordonné de 12) constituent un contexte de comparaison, non des réimplémentations.
- **[2026-07-03] Boucle d'apprentissage jamais alimentée.** Le composant d'évaluation qui écrit la couche de boucle d'apprentissage n'a pas été livré pendant le tournoi, et L7 était vide pour les deux jumeaux à chaque match. L'affirmation enregistrée de niveau 2 est rapportée comme non testable telle que conçue (Section 7), et la paire de jumeaux est reconvertie en contrôle de réplication. C'est le plus grand écart entre l'enregistrement et l'exécution dans l'étude.
- **[2026-07-03] Enchevêtrement Pipeline–Council au taux majoritaire.** Le seuil d'escalade enregistré (une confiance finale inférieure à 60 adopte la prédiction du Council) s'est déclenché sur une majorité de matchs, de sorte que Pipeline et Council sont fortement non indépendants en tant que voies publiées. Les sorties pré-escalade par match ont été journalisées tout au long et l'enchevêtrement est analysable (Section 6). Les comparaisons de voies impliquant Pipeline doivent être lues en tenant compte de ce couplage.
- **[2026-07-03] Évaluation de la qualité narrative (M2) exécutée a posteriori.** Aucun score de juge n'a été produit en direct. L'évaluation à trois juges s'est déroulée le 2026-07-03 à partir des sorties figées, à la provenance sûre sur trois plans. Le jugement est une évaluation, non une prédiction. Les juges étaient aveugles à la voie, au modèle et au résultat. L'invite du juge est un outillage daté a posteriori, divulgué, hors du verrou M1. L'accord inter-juges n'est pas simplement tombé sous le seuil enregistré de 0.6, il est tombé sous zéro. L'alpha de Krippendorff moyen par paires est négatif (Table 9), ce qui signifie que les trois juges se sont accordés sur la qualité absolue moins qu'une assignation aléatoire ne le ferait. Le mécanisme est un effet d'échelle visible dans la matrice par voie (Table 10). Les juges classent les voies de manière similaire aux extrêmes mais notent sur des échelles absolues très différentes, l'un se groupant près de six et l'autre près de neuf, de sorte qu'une statistique d'accord à métrique d'intervalle lit le désaccord comme sévère. Nous rapportons cela comme un résultat, non comme un défaut à réparer. Pour la tâche de qualité narrative, un panel de juges à modèle de langage n'a pas produit de score absolu fiable, et les chiffres qualitatifs de M2 sont lus en conséquence, comme indicatifs d'un ordonnancement grossier tout au plus et non comme une mesure de qualité calibrée. Le résultat ne touche pas aux constats de calibration, qui sont notés contre des résultats de matchs réels plutôt que jugés.

Dimension	Claude/GPT	Claude/Gemini	GPT/Gemini	Mean
reasoningTraceability	0.017	-0.512	-0.214	-0.236
brandDiscipline	-0.320	-0.503	0.027	-0.265
operationalClarity	-0.175	-0.568	-0.123	-0.289
Mean pairwise α (all dimensions)				-0.263

Table 9 – Accord inter-juges (M2), alpha de Krippendorff moyen par paires par dimension de rubrique. Les valeurs négatives indiquent un accord pire que le hasard sur la sévérité absolue.

La matrice enregistrée d'auto-préférence au niveau de la famille dégénère parce que la narration publiée de chaque voie est rédigée par la même famille de modèles (l'étape de rédacteur ou de décideur), de sorte que la matrice est rapportée par voie avec la confusion de paternité nommée. La vérification humaine par sondage stratifié enregistrée de 10 % reste ouverte (tâche de l'opérateur).

- **[2026-07-03] Artefacts de couche d'engagement enregistrés.** L'échelle de points de la §8 n'a pas

Juge	Solo	Solo-Zero	Pipeline	Pipeline-Static	Council
claude	6.81	5.44	6.70	6.62	6.26
gpt	7.80	7.69	7.92	7.96	7.74
gemini	8.86	8.48	9.12	9.18	9.04
Signalements d'hallucination par verdict	0.24	0.05	0.20	0.19	0.13

Table 10 – Matrice des scores moyens juge par voie (M2). L'effet d'échelle est visible. Les juges divergent davantage sur le niveau absolu que sur l'ordonnancement. Les signalements d'hallucination par verdict apparaissent à la dernière ligne.

été affichée en direct pendant le tournoi. Parce qu'elle est déterministe et que l'enregistrement permet un affichage rétroactif, elle est calculée à partir des prédictions publiées et incluse (Appendix F). L'engagement de tableau complet de la §8 (chaque voie publiant un tableau d'élimination complet avant le premier coup d'envoi des seizièmes de finale) n'a pas été tenu et ne peut être reconstruit avec une provenance propre, de sorte qu'il est noté comme un engagement non tenu. Les deux sont des surfaces d'engagement, non des indicateurs phares, et ni l'un ni l'autre ne porte sur le résultat de calibration.

- **[dated] Incidents opérationnels.** Deux au cours du tournoi, tous deux récupérés sans aucune rencontre perdue. La limite de débit du fournisseur du 2026-06-23 (2 rencontres) et le blocage de répartition du 2026-07-03 (1 rencontre, récupérée avant le coup d'envoi, et le correctif de délai d'attente est un changement au niveau du transport postérieur au gel, divulgué sous M6). L'une des rencontres limitées en débit, id 1489404, a été récoltée à nouveau et ses prédictions régénérées à partir d'un paquet propre pré-coup d'envoi. Parce que sa provenance diffère de celle des rencontres intactes, nous rapportons les résultats phares en l'incluant et confirmons qu'ils sont inchangés lorsqu'elle est écartée, de sorte qu'un lecteur peut voir que la seule rencontre régénérée ne déplace rien. Détails. `NOTES-incidents.md` dans le dépôt source, publié avec le jeu de données.

11 Conclusion

Nous avons demandé si la confiance qu'un système d'IA déclare vaut quelque chose, et avons rendu la question falsifiable. Un banc d'essai en direct préenregistré, chaque prédiction et son coût publiés avant le coup d'envoi, les règles de notation verrouillées avant le premier match noté. La réponse donnée par le tournoi n'est pas celle qu'un classement célébrerait. Sur l'exactitude, aucune architecture ne domine un résultat dominé par le hasard, et sur l'ensemble mis de côté, aucune voie n'est statistiquement distinguable d'une autre. L'a priori que nous avons énoncé, à savoir qu'aucun système d'IA ne battrait le marché sur l'exactitude brute, a été inversé en direction et laissé non résolu en amplitude, deux points que nous publions.

Le constat qui subsiste porte sur la confiance, non sur l'exactitude. Une prévision qui dit trente pour cent et échoue a dit la vérité sur sa propre incertitude. Une prévision qui dit soixante-dix pour cent et échoue ne l'a pas fait. Le nombre qui les sépare est la calibration, et c'est le nombre qu'une entreprise consomme réellement lorsqu'elle met une probabilité fournie par l'IA dans une décision réelle. Sur l'ensemble du tournoi, le conseil, trois familles de modèles indépendantes et une étape qui pèse leur désaccord, affiche l'erreur de calibration la plus basse de toute machine que nous avons testée, la référence humaine mise à part. Si la question opérationnelle est de savoir à quelle architecture faire confiance pour une prévision, les données pointent vers celle dont la confiance déclarée est la plus honnête, non vers

celle qui appelle le plus de matchs, et sur l'ensemble de ce tournoi il s'agit du conseil. Nous énonçons la force de cette lecture clairement. Sur un événement dominé par le hasard, les intervalles bootstrap se chevauchent, de sorte qu'il s'agit d'une avance directionnelle qu'un second tournoi affinerait, non d'une séparation résolue. Une avance directionnelle reste la réponse qu'un opérateur a demandée, à savoir vers quelle architecture se tourner, et cette étude la fournit.

Que les machines se situent près du marché tout court est le résultat football, non le résultat d'affaires. Le football a un marché de paris public profond qui valorise déjà chaque match, de sorte qu'une bonne prévision a une référence dure contre laquelle être mesurée et peu de marge pour la battre. Une entreprise qui prévoit son propre pipeline, sa propre demande, son propre risque n'a pas de tel marché ni de cote de clôture, de sorte que la probabilité d'IA calibrée n'est pas en concurrence avec une référence, elle est la chose la plus proche d'une référence que l'entreprise puisse bâtir. C'est là que ce banc d'essai porte ses fruits hors du stade. Il dit à un opérateur quelle forme d'IA déclare la confiance la plus digne de foi quand rien d'autre ne la note, et sur ces données cette forme est le conseil.

Les limites sont énoncées là où elles ont leur place, à la Section 10, et aucune d'elles ne renverse le résultat. C'est un tournoi, et c'en est un complet, noté en public contre un protocole verrouillé. L'affirmation centrale n'a jamais été qu'un système quelconque prédit l'avenir. C'est que le compte rendu qu'un système fait de sa propre incertitude est la première chose à mesurer à son sujet, et la première chose à laquelle se fier ou se méfier, et que sur ces données le conseil est l'architecture qui le déclare le plus honnêtement. 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.

Références

[1] Ali Al-Bustami and Zaid Ghazal. From players to champions : A generalizable machine learning approach for match outcome prediction with insights from the FIFA World Cup. In *2025 IEEE International Conference on Electro Information Technology (eIT)*, pages 574–578, 2025. doi : 10.1109/eit64391.2025.11103598.

[2] Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1) :1–3, 1950.

[3] Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. ReConcile : Round-table conference improves reasoning via consensus among diverse LLMs, 2023. Presented at NAACL 2024.

[4] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate, 2023. Presented at ICML 2024.

[5] Leonardo Egidi, Roberto Macrì Demartino, and Vasilis Palaskas. footBayes : Fitting Bayesian and MLE football models. CRAN : Contributed Packages, 2022.

[6] Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477) :359–378, 2007.

[7] Andreas Groll, Gunther Schauberger, and Gerhard Tutz. Prediction of major international soccer tournaments based on team-specific regularized Poisson regression : An application to the FIFA World Cup 2014. *Journal of Quantitative Analysis in Sports*, 11(2), 2015. doi : 10.1515/jqas-2014-0051.

[8] Andreas Groll, Christophe Ley, Gunther Schauberger, and Hans Van Eetvelde. A hybrid random forest to predict soccer matches in international tournaments. *Journal of Quantitative Analysis in Sports*, 15(4) :271–287, 2019. doi : 10.1515/jqas-2018-0060.

[9] Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. Approaching human-level forecasting with language models, 2024.

[10] Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E. Tetlock. ForecastBench : A dynamic benchmark of AI forecasting capabilities. In *International Conference on Learning Representations (ICLR)*, 2025.

[11] Allan H. Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4) :595–600, 1973.

[12] Pieter Robberechts and Jesse Davis. Forecasting the FIFA World Cup : Combining result- and goal-based team ability parameters. In *Machine Learning and Data Mining for Sports Analytics (MLSA), Lecture Notes in Computer Science*, pages 16–30, 2019. doi : 10.1007/978-3-030-17274-9_2.

[13] Pieter Robberechts, Jan Van Haaren, and Jesse Davis. A Bayesian approach to in-game win probability in soccer. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3512–3521, 2021. doi : 10.1145/3447548.3467194.

[14] Maohao Shen, Subhro Das, Kristjan Greenewald, Prasanna Sattigeri, Gregory Wornell, and Soumya Ghosh. Thermometer : Towards universal calibration for large language models, 2024. Presented at ICML 2024.

[15] Dat Tran and Douwe Kiela. Single-agent LLMs outperform multi-agent systems on multi-hop reasoning under equal thinking token budgets, 2026.

[16] Xiaochen Zhu, Caiqi Zhang, Yizhou Chi, Tom Stafford, Nigel Collier, and Andreas Vlachos. Demystifying multi-agent debate : The role of confidence and diversity, 2026.

A Invites, mot pour mot

Les invites ci-dessous sont reproduites depuis la source verrouillée par hachage (`web/convex/lib/prompts.ts`, `PROMPTS_VERSION` 1.0, figée au verrou de préenregistrement M1), SHA-256 `1ce8d0…a9d0966`. Le fichier complet, sa somme de contrôle et le schéma de sortie des prédictions sont publiés avec le jeu de données afin que cette liste puisse être vérifiée par rapport à ce qui a été exécuté. Les gabarits d'exécution interpolent l'en-tête du match et le contexte assemblé au moment de la répartition.

Invite système partagée de l'analyste. Utilisée par Solo, Solo-Zero, l'éditeur de voix de Pipeline, ainsi que chaque membre et décideur de Council.

```
You are a football analyst forecasting a 2026 World Cup match. Confidence must
be a numeric value 0-100 representing your actual probability estimate that your
predicted winner is correct. Do not hedge, calibration is scored. riskFactors:
max 3 short strings. reasoning: max 100 words, plain language, direct, no buzzwords,
no emojis, no semicolons, no em dashes.
```

Gabarits utilisateur à appel unique. SOLO reçoit l'en-tête ainsi que le contexte à sept couches. SOLO-ZERO reçoit l'en-tête seul.

```
[A1] {matchHeader} + Full context pack (seven layers, some may be null): {contextPack}
+ Predict the match outcome.
```

```
[A1-zero] {matchHeader} + No additional context is provided. Predict the match
outcome from the fixture information above only.
```

Invites d'étape de PIPELINE (trois spécialistes séquentiels).

```
[Step 1, statistician] You are a football statistician. From the quantitative
layers of the context pack (stats, weather), produce a strictly statistical read
of the match. No narrative, no news.
```

```
[Step 2, operations reader] You are an operations reader. Combine the statistician's
read with the narrative layers (news, rivalries, prior insights, learning loop
if present). Surface what the numbers miss: availability, momentum, stakes, conditions.
```

```
[Step 3, voice editor] You are the final voice editor of a sequential analyst
pipeline. Synthesize the statistician's and operations reads into one final prediction.
(+ shared analyst system prompt above.)
```

Invites de COUNCIL. Chaque membre exécute l'invite d'analyste partagée. Le décideur exécute :

```
You are the decider of a three-member analyst council. You receive three independent
predictions from different model families. Weigh their agreement and disagreement,
then produce ONE final prediction. Where they disagree, your confidence should
reflect that disagreement. (+ shared analyst system prompt above.)
```

Requêtes de collecte de contexte (actualités L2, rivalité L3).

```
[L2 news] Latest team news for {home} and {away} national football teams in the
last 7 days: injuries, suspensions, lineup signals, form. World Cup 2026 context.
```

```
[L3 rivalry] Historical rivalry and head-to-head narrative between {home} and
{away} national football teams. Memorable encounters, styles of play, current
tactical identities.
```

Schéma de sortie des prédictions. Toutes les voies émettent un objet JSON validé par rapport à un schéma unique : `winner` (home / away / draw), `score` (entiers home et away), `confidence` (nombre 0–100), `riskFactors` (tableau de chaînes), `reasoning` (chaîne). La sortie structurée est imposée avec une nouvelle tentative en cas d'échec de schéma, et un second échec est enregistré comme une abstention, jamais réparé en silence. Chaque étape de PIPELINE porte son propre schéma intermédiaire (lecture statistique, lecture opérationnelle), publié dans la source.

B Artefacts de protocole et registre d'intégrité

Le tableau final est généré à la publication. Le contenu est fixé dès maintenant :

- Préenregistrement : OSF ke35t, DOI 10.17605/OSF.IO/R5SBJ, avec le tableau de sommes de contrôle SHA-256 liant les documents enregistrés et le code de production qu'ils régissent (invites, passerelle

de modèles, injection de motifs implantés). Documents liés : `pre-registration.md`, `statistical-tests.md`, `seeded-patterns-sealed.md`.
- La divulgation de dérive de schéma : tableaux d'opérations additifs postérieurs au verrou (registre de référence, journal des allégations de répartition, tableaux de la couche d'affichage), chacun divulgué dans son commit, les tableaux de méthodologie restant intacts.
- Le registre de vérification manuelle de la première journée M8 (11 juin 2026) : recalcul indépendant du Brier score pour les huit voies du match d'ouverture, concordance à trois décimales (le seuil exigeait cinq sorties).
- La divulgation des paramètres Elo (clé de registre `elo:params`) : conventions, part empirique des matchs nuls, matchs traités, date de référence.
- Le journal des amendements, daté, versionné, chacun publié avant tout match qu'il affecte.

C Comptabilité des coûts et des latences

Chaque prédiction évaluée porte son coût d'API facturé dans la télémétrie publiée. Les chiffres ci-dessous représentent la dépense cumulée par voie avant prédiction sur l'ensemble des 104 matchs, lue depuis l'export figé du tournoi, de sorte que chaque ligne se réconcilie exactement avec la colonne Coût des tableaux 3 et 7. Les références déterministes (Elo, Market) et la référence humaine n'engagent aucun coût d'API et sont omises.

Voie	Total tournoi	Par match	Relatif à Solo-Zero
Solo-Zero	\$1.06	\$0.010	1.0×
Solo	\$8.22	\$0.079	7.8×
Council	\$15.03	\$0.145	14.2×
Pipeline	\$21.93	\$0.211	20.7×
Pipeline-Static	\$21.93	\$0.211	20.7×
Plateau IA complet	\$68.18	\$0.656	

Table 11 – Comptabilité des coûts, l'ensemble des 104 matchs. Le total tournoi est le coût d'API facturé cumulé par voie, le coût par match divise par les 104 rencontres évaluées, et la colonne relative tarife chaque architecture par rapport à l'appel sans contexte. Le plateau IA complet est le coût par match des cinq voies qu'un opérateur exécutant chaque architecture en parallèle paierait.

La lecture opérationnelle se trouve à la Section 9.4. L'ensemble du tournoi à cinq voies a coûté \$68.18 en dépense d'API, de sorte que la contrainte déterminante d'un premier banc d'essai interne est la méthodologie, non le budget. La comparaison coût par point de calibration qui importe à un adoptant, à savoir les dollars par voie face à l'ECE par voie, est la frontière que cette section lit dans le Tableau 7. Le conseil achète la plus faible erreur de calibration de l'étude à environ quatorze fois l'appel sans contexte et bien en deçà des pipelines, qui dépensent le plus et ne rapportent aucun avantage de calibration. Le nombre de jetons par appel et la latence horloge de chaque prédiction figurent dans le jeu de données par prédiction publié.

D Revue adversariale

Avant publication, la version finale a été soumise à une lecture adversariale interne : quatre relecteurs indépendants, chacun doté d'une lentille distincte et chargé de trouver de vrais défauts plutôt que d'approuver. Les lentilles étaient la rigueur statistique (le calcul de calibration tient-il, les intervalles sont-ils étiquetés honnêtement), la crédibilité opérateur et finance (la recommandation donne-t-elle au décideur matière à agir, le pont du marché vers l'entreprise se tient-il), l'adéquation à la finalité d'affaires (l'article reste-t-il un artefact de qualité décisionnelle ou verse-t-il dans la méthode pour elle-même), et la marque et cohérence interne (discipline de voix, cohérence de section à section). Chaque relecteur a lu l'article compilé de façon indépendante et a rapporté ses constats classés par sévérité.

La revue a laissé debout le constat central : les quatre relecteurs confirment que la recommandation directionnelle du conseil tient sur l'évidence du tournoi complet, l'intervalle de calibration chevauchant étant divulgué comme une limite plutôt qu'une rétractation. Leurs corrections substantielles ont été adoptées avant publication. La plus tranchante portait sur la comparaison avec le marché. L'erreur de calibration du conseil se situe à environ quatre millièmes sous celle du marché, à l'intérieur du plancher de bruit de réplication propre à l'étude, de sorte que l'article nomme désormais cette comparaison comme une égalité à cette résolution et la résout par le point qu'une entreprise n'a aucune ligne de marché à laquelle s'égaler (Section 8.1). D'autres ont affiné la lecture coût contre calibration du choix d'architecture (Section 9.4) et le placement du résultat opérateur dans le résumé.

Des lectures externes indépendantes, par un journaliste du domaine, un opérateur en exercice et un spécialiste IA externe, sont prévues après la publication et seront reproduites ici avec les réponses au fur et à mesure de leur arrivée. Cet appendice consigne le processus adversarial préalable à la publication, non une revendication de revue externe par les pairs.

E Référence des voies et des métriques

Une référence à destination du lecteur. Chacun des huit analystes et chaque métrique rapportée y sont définis en un seul endroit. Tout ce qui suit transcrit le protocole enregistré et les notes d'implémentation divulguées. Rien dans cette annexe ne constitue un nouvel engagement, et aucun résultat de tournoi n'y apparaît.

E.1 Les huit analystes

Solo (A1), architecture. Un appel à un modèle figé avec l'invite système d'analyste partagée et le pack de contexte complet à sept couches (statistiques, météo, actualités, rivalités, insights générés, analyse antérieure, boucle d'apprentissage). Sortie structurée imposée par un schéma JSON avec une nouvelle tentative. Un second échec enregistre une abstention. La chose la plus simple qu'un opérateur pourrait déployer, et donc l'architecture de référence.

Solo-Zero (A1-zero), contrôle. Le chemin de code littéral de Solo avec le pack de contexte retiré, uniquement les métadonnées de match (équipes, phase, stade, coup d'envoi). Tout le reste est identique, si bien que l'écart entre les jumeaux mesure la valeur du pack de contexte lui-même. Il répond à la question, qu'est-ce que tout ce contexte a apporté ?

Pipeline (A2), architecture. Trois appels séquentiels sur une même famille de modèle figé. Un statisticien ne lit que les couches quantitatives, un lecteur d'opérations ajoute les couches narratives, un éditeur de voix produit la prédiction structurée finale. Chaque étape a son propre schéma et la règle d'une nouvelle tentative. Une porte d'escalade enregistrée fait partie de l'architecture. Quand la confiance finale du pipeline tombe sous 60, il adopte la prédiction du Council pour ce match, journalisée et attribuée en coût (le pipeline paie toujours ses trois étapes).

Pipeline-Static (A2-static), contrôle. Le pipeline complet avec la couche de boucle d'apprentissage annulée dans son contexte, exclue des écritures de boucle, exécuté sur chaque match que Pipeline traite. Enregistré pour isoler la boucle d'apprentissage. (La section 7 rapporte comment cette paire doit réellement être lue, étant donné que la couche de boucle n'a jamais été alimentée.)

Council (A3), architecture. Trois membres issus de familles de modèles distinctes reçoivent en parallèle l'invite et le contexte identiques. Un décideur de la première famille reçoit les trois soumissions, identités de modèle retirées (membre anonyme 1/2/3), avec pour consigne que sa confiance doit refléter le désaccord entre membres, et produit l'unique prédiction de la voie. Architecture et mélange de modèles sont confondus par conception et divulgués comme tels.

Elo, référence déterministe. Un système de notation Elo calculé à partir de l'historique des matchs internationaux (fenêtre du 2022-06-01 au début du tournoi, tous les matchs des 48 équipes qualifiées, dédoublonnés). Score attendu $W_e = 1/(1 + 10^{-(R_h+H-R_a)/400})$. Avantage du terrain côté historique $H = +100$ pour l'équipe listée à domicile. Pour les prédictions de tournoi $H = +100$ uniquement quand l'équipe est une nation hôte (Mexico, USA, Canada), 0 sinon, selon la variable indicatrice d'hôte de Robberechts and Davis [12]. $K = 30$ pour les matchs de Coupe du monde, $K = 20$ sinon, mis à l'échelle par un multiplicateur de différence de buts (1 pour un nul ou une marge d'un but, 1.5 pour deux, 1.75 pour trois, plus 1/8 par but supplémentaire). Les probabilités à trois issues dérivent de l'identité $W_e = P(\text{win}) + P(\text{draw})/2$, la part de nul étant prise empiriquement sur la fenêtre de reconstitution et les probabilités bornées à $[0.01, 0.97]$. Les adversaires hors graphe entrent à 1500. Déterministe. Le même historique reproduit chaque prédiction de la voie bit à bit. Aucune IA n'intervient à aucune étape.

Market, référence déterministe. La cote de clôture du marché des paris, un instantané unique à T−1 h par match, Pinnacle en priorité avec une chaîne de repli enregistrée (Bet365, puis le premier livre 1X2 disponible). Les cotes des bookmakers intègrent une marge, si bien que les probabilités implicites sont dévigées proportionnellement. Avec des cotes décimales d_k pour les trois issues, $p_k = (1/d_k)/\sum_j (1/d_j)$. Le choix de la voie est l'issue argmax et sa confiance est la probabilité dévigée de cette issue. La cote de marché n'entre jamais dans le contexte d'aucune voie IA. C'est la référence la plus difficile largement disponible en prévision sportive, et l'a priori enregistré n'attend d'aucune voie IA qu'elle la batte sur l'exactitude brute.

Human, référence humaine non aveugle. Un humain nommé (le déclarant). Les pronostics ont été saisis par le déclarant, qui n'était pas aveugle aux sorties des voies IA ni aux résultats en cours (voir la divulgation de la référence non aveugle, section 3), à tout moment avant le coup d'envoi, verrouillant au coup d'envoi. Une nouvelle saisie avant le coup d'envoi remplace le pronostic antérieur. Un pronostic manquant est affiché comme manquant et ne marque rien. Le texte de raisonnement est publié tel quel.

Cette voie est rapportée à des fins de transparence, non comme un test contrôlé humain contre machine, et tout avantage de calibration qu'elle montre est confondu par l'accès à l'information. C'est une seule personne, pas un panel, et elle est lue en conséquence.

E.2 Les métriques, en termes simples

Les définitions formelles se trouvent à la section 5. Ceci est le guide de lecture. Partout, la confiance c d'une voie sur son vainqueur prédit se traduit par $p = c/100$, et y enregistre si ce vainqueur a gagné.

Brier score de saison (multiclasse). Le score de pénalité pour un vecteur complet de probabilités à trois issues face à ce qui s'est produit, sommé sur domicile/nul/extérieur. Plus bas est meilleur. Une prescience parfaite marque 0, une répartition toujours uniforme marque 0.667, et les erreurs confiantes coûtent le plus cher. C'est la métrique phare verrouillée. Le vecteur de probabilités est construit par la règle enregistrée de la section 5.1.

Brier score binaire. La même idée réduite au vainqueur prédit, $\text{mean}\big((p - y)^2\big)$. Une voie qui dirait toujours 50 marquerait 0.25 quels que soient les résultats, et l'habileté comme l'honnêteté poussent toutes deux en dessous. C'est l'espace dans lequel vit l'analyse de calibration.

ECE (erreur de calibration attendue). Regroupe les prédictions en dix casiers de confiance et demande, par casier, à quelle distance la confiance déclarée se situe du taux de réussite réalisé, pondérée par la fréquence à laquelle la voie s'est exprimée à cette confiance. Le premier nombre à lire. Une ECE de 0.05 signifie que la confiance déclarée suit la réalité à environ cinq points près en moyenne, une ECE de 0.15 signifie que le nombre rapporté par la voie est faux de quinze points en moyenne. Zéro est une voie dont le 70 vaut 70.

Fiabilité, résolution, incertitude. L'anatomie à trois volets du Brier score binaire. La fiabilité est la mauvaise calibration elle-même (plus bas est meilleur), la version au carré de ce que l'ECE mesure en termes absolus. La résolution récompense une voie dont les prédictions confiantes et peu confiantes trient effectivement les victoires des défaites (plus haut est meilleur). L'incertitude est la variance de taux de base propre à la tâche et appartient à la Coupe du monde, pas à un prévisionniste. C'est le plancher que le hasard fixe pour tout le monde.

Finesse. L'étalement des confiances déclarées d'une voie (variance autour de sa propre prévision moyenne). La calibration seule peut se tromper en ne s'engageant jamais. La finesse est la métrique qui attrape la voie bien calibrée seulement parce qu'elle dit toujours 55. La paire à lire ensemble est ECE basse *et* finesse élevée, honnête *et* disposée à s'engager.

Taux de réussite. La simple fraction de vainqueurs corrects. Délibérément rétrogradé. C'est le nombre que tout le monde demande en premier et celui auquel ce banc d'essai se fie le moins, parce qu'il récompense les pronostics audacieux et masque entièrement la mauvaise calibration.

Intervalles bootstrap à 95 %. La bande de stabilité autour de chaque métrique rapportée, à partir de 1,000 rééchantillonnages par blocs de matchs avec une graine publiée fixe. La règle d'affirmation enregistrée, deux voies ne diffèrent que si leurs intervalles hors échantillon ne se chevauchent pas. Un chevauchement se lit comme « aucune différence résoluble », et cette formule apparaît partout où elle est vraie.

Brier score stratifié par phase. Le même score découpé par phase de tournoi, parce que le football à élimination directe est plus difficile à prévoir que la phase de groupes et qu'un agrégat de voie peut le masquer. Publié en groupes contre élimination directe (les strates enregistrées plus fines sont calculables à partir du jeu de données publié).

F Le classement par points

Le classement par points est une surface d'engagement publique, enregistrée dans le protocole pour rendre le banc d'essai lisible aux visiteurs non spécialistes. Ce n'est pas une affirmation de calibration et il n'entre dans aucun résultat phare. Il récompense les appels corrects et confiants et pénalise les appels faux et confiants, ce qui est une question différente de celle de savoir si la confiance d'une voie est bien calibrée. Nous l'incluons ici, calculé rétroactivement à partir des prédictions publiées, parce que l'enregistrement le définit comme une fonction déterministe du relevé et permet un affichage rétroactif.

La règle, fixée dans le préenregistrement, attribue à chaque voie (confiance si le pronostic est correct, sinon − confiance) × multiplicateur de tour par match noté, le multiplicateur valant 1.0 pour la phase de groupes, 1.5 des seizièmes de finale aux quarts de finale, et 2.0 pour les demi-finales, le match pour la troisième place et la finale. Les abstentions et les pronostics manquants marquent zéro. La table 12 rapporte le classement final sur l'ensemble des 104 matchs.

Le conseil termine troisième de ce classement, derrière l'appel unique sans contexte et le marché. Cet ordre est attendu et il ne qualifie pas le résultat de l'étude. Le classement récompense l'exactitude confiante, l'axe que cet article montre ne résoudre rien sur une tâche dominée par le hasard, tandis que l'axe de calibration qui porte le résultat place le conseil en premier. Une voie peut se classer haut ici en ayant été audacieusement correcte sur les matchs qu'elle a justement appelés, tout en restant moins honnête sur sa propre incertitude à travers l'ensemble complet. Le classement est la surface d'engagement, la courbe de fiabilité est le résultat.

Rang	Voie	Points
1	Solo-Zero	3436
2	Market	3094
3	Council	2944
4	Solo	2940
5	Pipeline-Static	2891
6	Pipeline	2860
7	Human	2205
8	Elo	1673

Table 12 – Classement final par points, calculé rétroactivement sur l'ensemble des 104 matchs selon la règle enregistrée. Le classement récompense l'exactitude confiante, si bien qu'il suit l'ordre d'exactitude plutôt que l'ordre de calibration qui porte le résultat de l'étude, et les deux sont délibérément rapportés côte à côte.