

Métadonnée	Valeur
Statut	Projet de norme distinct des spécifications AIWC
Objet	Règles générales de conception, d'exécution et de publication des benchmarks AIWC
Dépendance	AIWC Specification, draft exploratoire 0.4
Public visé	Concepteurs de benchmarks, laboratoires, fournisseurs, évaluateurs et acheteurs
Format	A4 ; pagination automatique

NORME DE BENCHMARKING

AIWC — AI Work Continuity

Changer d'IA sans recommencer

Interopérabilité entre IA et portabilité du travail en cours

Norme générale de benchmarking de la continuité fonctionnelle

Vers une continuité fonctionnelle lors du passage d'une intelligence artificielle à une autre

Draft exploratoire 0.3 — Document de travail — 21 juillet 2026

1. Statut du document

Le présent document constitue un premier draft exploratoire de la norme de benchmarking AIWC. Il définit des règles générales permettant d'évaluer, de manière comparable et reproductible, la capacité d'une intelligence artificielle à continuer un travail à partir d'un WCS.

Il ne constitue pas un catalogue exhaustif de tâches, un classement du marché ni une procédure que l'utilisateur ordinaire doit exécuter lors de chaque reprise. Les termes DOIT, NE DOIT PAS, DEVRAIT, NE DEVRAIT PAS et PEUT expriment l'intention normative de cette version.

Séparation essentielle — Les spécifications AIWC définissent l'objet échangé et les comportements minimaux. La présente norme définit comment mesurer la qualité d'une continuation. La conformité à l'une n'implique pas un bon résultat dans l'autre.

2. Finalité

Un benchmark AIWC cherche à répondre à une question limitée : dans quelle mesure un système candidat peut-il produire une continuation pertinente d'un travail après avoir reçu un WCS, dans des conditions connues et comparables ?

Selon son profil, le benchmark peut isoler la capacité à produire un WCS, la capacité à le consommer, ou la performance de bout en bout. Il doit éviter d’attribuer à l’IA consommatrice les défauts du WCS, ou inversement.

3. Périmètre

La présente norme couvre :

- les principes de validité d’un benchmark AIWC ;
- les rôles, objets et profils d’évaluation ;
- la constitution d’une épreuve et de ses baselines ;
- la procédure générale d’exécution ;
- les dimensions générales d’évaluation ;
- les règles de notation, de jugement et de reproductibilité ;
- la publication et les déclarations de conformité.

Elle ne définit pas :

- une taxonomie complète des travaux humains ;
- un jeu d’épreuves universel ou définitif ;
- des pondérations obligatoires identiques pour tous les domaines ;
- un classement actuel des fournisseurs ;
- une métrique unique réduisant toute continuité à un nombre ;
- une évaluation de la conscience, de l’identité ou des états latents d’un modèle.

4. Termes

Travail source — Travail réel ou construit dont une interruption et une reprise peuvent être observées.

Point de coupure — État du travail au moment où le prédécesseur cesse de contribuer et où le WCS est produit.

Continuation cible — Contribution demandée après le point de coupure.

Système candidat — Modèle, agent ou configuration évalué.

Producteur — Système qui compose le WCS à partir du travail source.

Consommateur — Système qui assimile le WCS et produit la continuation.

Référence experte — Éléments établis par les auteurs ou évaluateurs pour guider le jugement, sans imposer nécessairement une réponse unique.

Baseline — Condition de comparaison n’utilisant pas le WCS évalué ou utilisant une représentation alternative.

Épreuve — Ensemble versionné comprenant le travail source, le point de coupure, le WCS, la continuation demandée, les ressources autorisées et le protocole d’évaluation.

Exécution — Passage d’un système candidat sur une épreuve dans une configuration déterminée.

5. Principes de benchmarking

Orientation vers la continuation

Le benchmark DOIT mesurer une contribution après reprise, et non la simple récitation du contenu du WCS.

Neutralité technologique

Le protocole NE DOIT PAS dépendre inutilement d'un fournisseur, d'un modèle ou d'un état interne propriétaire.

Reproductibilité

Les entrées, ressources, paramètres et règles de jugement DOIVENT être suffisamment documentés pour permettre une réplique raisonnable.

Comparabilité conditionnelle

Deux résultats NE DOIVENT être comparés que si les différences de modèles, outils, budgets, ressources et conditions sont déclarées et compatibles avec la comparaison.

Pluralité des continuations

Une épreuve NE DOIT PAS supposer une réponse unique lorsque plusieurs continuations sont légitimes.

Séparation des facteurs

Le design DEVRAIT distinguer autant que possible la qualité du WCS, la capacité du consommateur et les capacités générales nécessaires à la tâche.

Charge utilisateur

Le protocole de laboratoire NE DOIT PAS être présenté comme une exigence de l'usage ordinaire d'AIWC.

Prudence interprétative

Un score mesure une performance dans un périmètre donné ; il NE DOIT PAS être généralisé à toute forme de continuité.

6. Profils d'évaluation

6.1 Profil C — consommation

Tous les candidats reçoivent le même WCS de référence et les mêmes ressources autorisées. Ce profil mesure principalement la capacité à assimiler le WCS et à continuer le travail.

6.2 Profil P — production

Les producteurs candidats reçoivent le même travail source jusqu'au point de coupure. Leurs WCS sont évalués directement et, de préférence, consommés par un ou plusieurs consommateurs fixes. Ce profil mesure la qualité de composition du WCS.

6.3 Profil E2E — bout en bout

Un système ou une chaîne produit le WCS, puis un successeur poursuit le travail. Ce profil mesure l'efficacité globale du transfert, mais ne permet pas à lui seul d'attribuer les succès ou échecs au producteur ou au consommateur.

6.4 Profil X — interopérabilité croisée

Plusieurs producteurs et consommateurs de familles différentes sont croisés. Ce profil explore la robustesse inter-fournisseurs. Le plan de croisements et les exclusions DOIVENT être publiés.

7. Constitution d'une épreuve

Une épreuve conforme DOIT comprendre ou identifier :

- le travail source et son périmètre ;
- un point de coupure non ambigu ;
- le WCS ou la procédure permettant de le produire ;
- la continuation demandée ;
- les artefacts, outils et ressources autorisés ;
- les informations volontairement retenues après la coupure ;
- les critères et ancres de jugement établis avant l'évaluation ;
- les risques particuliers, notamment confidentialité et contamination ;
- un identifiant et une version de l'épreuve.

Le travail source PEUT être authentique, synthétique ou hybride. Ce choix DOIT être déclaré. Une épreuve authentique améliore la validité écologique ; une épreuve synthétique améliore souvent le contrôle expérimental.

8. Prévention des fuites et de la contamination

- La continuation cible et les éléments futurs du travail NE DOIVENT PAS être accessibles au candidat avant sa réponse.
- Les auteurs DEVRAIENT évaluer le risque que l'épreuve ou sa solution aient été présentes dans les données d'entraînement ou publiées en ligne.
- Les versions publiques DEVRAIENT distinguer les épreuves d'entraînement, de développement et d'évaluation cachée.
- Toute consultation d'Internet, d'outils, de mémoire persistante ou de conversations antérieures DOIT être explicitement autorisée et journalisée.
- Une fuite détectée DOIT entraîner l'exclusion ou la qualification du résultat concerné.

9. Baselines

Un benchmark destiné à mesurer la valeur d'AIWC DEVRAIT comparer le WCS à au moins une représentation alternative. Les baselines possibles comprennent :

- absence de contexte de reprise ;
- instruction minimale décrivant seulement la prochaine tâche ;
- résumé conventionnel ;
- historique ou transcription accessible ;
- sélection d'artefacts sans WCS.

Les baselines DOIVENT disposer de budgets d'information et de ressources suffisamment documentés. Si une baseline reçoit substantiellement plus d'information, de tokens, d'outils ou de temps, cette différence DOIT être visible dans les résultats.

10. Procédure générale

1. Geler et versionner l'épreuve, ses critères et ses ressources.
2. Déclarer le profil d'évaluation et les facteurs que l'expérience cherche à isoler.
3. Configurer le ou les candidats et consigner modèle, version, paramètres, outils, budgets et date.
4. Fournir uniquement les entrées autorisées pour la condition testée.
5. Recueillir la continuation et les traces explicitement prévues, sans exiger ni interpréter un raisonnement interne privé.
6. Faire évaluer les sorties selon le protocole annoncé, idéalement à l'aveugle quant au candidat et à la condition.
7. Répéter lorsque la stochasticité ou la variabilité des juges peut affecter substantiellement le résultat.
8. Calculer, qualifier et publier les résultats avec leurs limites.

11. Dimensions générales d'évaluation

Tout benchmark AIWC DOIT évaluer plusieurs dimensions. Il PEUT adapter leur formulation au domaine, mais NE DOIT PAS les remplacer par une simple appréciation globale non documentée.

Orientation

Le candidat identifie correctement l'objet, la finalité et le point où le travail doit reprendre.

Fidélité

Il conserve les éléments du cadre actuel qui conditionnent la suite, sans fabriquer de contexte.

Préservation des contraintes

Il respecte les décisions, exigences, limites et engagements encore actifs.

Discipline épistémique

Il distingue faits, hypothèses, inférences, désaccords et inconnues ; il ne transforme pas l'absence en certitude.

Pertinence de la continuation

Sa contribution répond au besoin actuel et s'inscrit dans la trajectoire du travail.

Progrès utile

La continuation fait réellement avancer le travail au-delà d'une reformulation ou d'un contrôle de compréhension.

Autonomie proportionnée

Le candidat exploite le WCS sans reporter inutilement sur l'utilisateur la reconstruction du contexte, tout en demandant les validations réellement nécessaires.

Une épreuve PEUT ajouter des dimensions propres au domaine, notamment exactitude technique, sécurité ou qualité formelle. Ces dimensions supplémentaires DOIVENT être séparées des dimensions de continuité afin d'éviter de confondre compétence générale et reprise.

12. Notation

Chaque dimension DEVRAIT utiliser des ancrs ordinales explicites. L'échelle suivante constitue la recommandation par défaut :

Score	Ancre générale
0	Échec : absence de reprise exploitable ou contradiction majeure avec l'état transmis.
1	Faible : quelques éléments reconnus, mais pertes ou erreurs empêchant une continuation fiable.
2	Partiel : reprise globalement identifiable, avec lacunes importantes ou progrès limité.
3	Solide : continuation fidèle et utile, avec défauts mineurs n'altérant pas la trajectoire.
4	Excellente : continuité pleinement exploitable, nuancée, autonome et productive dans le périmètre de l'épreuve.

Les pondérations PEUVENT varier selon l'épreuve, mais DOIVENT être définies avant l'examen des résultats et publiées. Un score agrégé DOIT rester accompagné des scores dimensionnels. Les échecs critiques — par exemple violation d'une contrainte de sécurité — DEVRAIENT être signalés séparément plutôt que dilués dans une moyenne.

13. Jugement et fiabilité

- Les juges DOIVENT disposer d'instructions et d'exemples d'ancrage adaptés à l'épreuve.
- Les sorties DEVRAIENT être anonymisées et présentées dans un ordre aléatoire lorsque cela est possible.
- Une évaluation importante DEVRAIT utiliser plusieurs juges et publier une mesure d'accord ou la procédure de résolution des divergences.
- Les juges IA PEUVENT assister l'évaluation, mais leurs modèles, prompts, limites et validations humaines DOIVENT être déclarés.
- Une référence experte NE DOIT PAS être traitée comme l'unique continuation acceptable sauf si la tâche possède objectivement une seule solution.

14. Répétitions et incertitude

Lorsque le candidat ou la procédure de jugement est non déterministe, le benchmark DEVRAIT comporter plusieurs exécutions ou justifier l'usage d'une seule. Les résultats agrégés DEVRAIENT publier la dispersion, le nombre d'exécutions et, lorsque pertinent, un intervalle d'incertitude. Les échantillons trop faibles pour soutenir une comparaison générale DOIVENT être présentés comme exploratoires.

15. Métadonnées minimales d'un résultat

- identifiant et version du benchmark et de l'épreuve ;
- profil P, C, E2E ou X ;
- identité du producteur et du consommateur évalués ;
- modèle, version, date, paramètres et configuration système ;
- outils, ressources, mémoire et accès réseau autorisés ;
- budget de tokens, temps ou coût lorsqu'il affecte la comparaison ;
- nombre d'exécutions ;
- procédure et identité fonctionnelle des juges ;
- scores par dimension, score agrégé éventuel et incertitude ;
- baselines ;
- exclusions, incidents, conflits d'intérêts et limites connues.

16. Publication et interprétation

Un rapport conforme DOIT permettre au lecteur de comprendre ce qui a été testé, dans quelles conditions et avec quelles limites. Il NE DOIT PAS transformer un résultat local en affirmation générale sur une IA ou sur AIWC.

Les classements de marché DEVRAIENT privilégier des familles d'épreuves, publier les résultats dimensionnels et signaler les changements de version des systèmes. Un classement sans description des épreuves, baselines et ressources n'est pas conforme à la présente norme.

17. Confidentialité, sécurité et éthique

- Les travaux authentiques DOIVENT être utilisés avec les autorisations et protections appropriées.
- Les WCS et traces d'évaluation DEVRAIENT être minimisés, pseudonymisés ou synthétisés lorsque cela réduit le risque sans détruire l'objet du test.
- Les évaluations portant sur des personnes NE DOIVENT PAS inférer ou publier des attributs sensibles non nécessaires.
- Les fournisseurs, auteurs d'épreuves et évaluateurs DOIVENT déclarer les conflits d'intérêts susceptibles d'affecter l'interprétation.
- Une épreuve présentant un risque réel ne doit pas être exécutée dans un environnement opérationnel non contrôlé.

18. Conformité

18.1 Épreuve conforme

Une épreuve est conforme si elle définit les éléments obligatoires de la section 7, protège les informations retenues, annonce ses dimensions et peut être versionnée.

18.2 Exécution conforme

Une exécution est conforme si elle suit une épreuve conforme, consigne la configuration, respecte les entrées autorisées et conserve les sorties nécessaires à l'évaluation.

18.3 Rapport conforme

Un rapport est conforme s'il publie les métadonnées minimales, les résultats dimensionnels, les baselines applicables, les incertitudes et les limites.

18.4 Déclaration

Toute revendication de conformité DOIT préciser l'objet concerné — épreuve, exécution ou rapport — ainsi que la version de la présente norme.

19. Évolutions attendues

Les versions futures pourront préciser des profils de domaine, des formats d'échange pour les paquets d'épreuve, des méthodes statistiques, des règles de gouvernance et des suites publiques. Ces extensions ne devraient pas remettre en cause la distinction entre usage pragmatique d'AIWC et protocole expérimental de benchmarking.

Conditions de publication

AIWC - Norme générale de benchmarking, draft exploratoire 0.3

© 2026 Yves Konen. Tous droits réservés.

Ce document est mis à disposition pour consultation et évaluation du projet pré-normatif exploratoire AIWC. Sa publication n'accorde aucune licence open source, Creative Commons ou autre licence générale d'exploitation.

Sauf exceptions prévues par la loi, sa reproduction substantielle, sa modification, sa redistribution, son intégration dans un autre produit ou son exploitation commerciale nécessitent une autorisation écrite préalable.

Les citations courtes doivent mentionner clairement l'auteur, le titre du document et la source AIWC.

Statut : document de travail exploratoire. Ce document ne constitue pas une norme officielle ou adoptée.