NEOMUNDI · REVUE MÉTHODOLOGIQUE INDÉPENDANTE

Mesure en temps réel de la stabilité de l’IA générative

Revue méthodologique publique — cohorte comparative de mai 2026

Abdelkrim Halimi · Data Scientist indépendant (contributeur)

Type : Revue méthodologique indépendante
Périmètre : Synthèse publique, expurgée
Cohorte : Huit fournisseurs anonymisés (P-001 à P-008)
Benchmark : Jeu de données public de véracité (TruthfulQA)

PÉRIMÈTRE DE CETTE VERSION PUBLIQUE

Cette version publique est une synthèse méthodologique expurgée. Elle ne divulgue ni l’identité des fournisseurs, ni les jeux de données propriétaires, ni les clés d’API, ni les détails d’infrastructure, ni les données de benchmark commercialement sensibles. Elle vise à documenter les enseignements méthodologiques, les limites et les pistes d’amélioration.

1 Synthèse

Ce document est une revue méthodologique indépendante de l’approche de mesure en temps réel de NeoMundi pour l’IA générative. Cette approche intercepte la dérive comportementale pendant la génération, sans lire le texte produit.

Sur la cohorte comparative de mai 2026, l’approche en temps réel montre une pertinence méthodologique. Elle parvient à signaler une réelle instabilité de génération sans inspection sémantique du contenu.

La revue identifie aussi des limites nettes :

- Une sensibilité réduite sur les modèles haut de gamme, où la plupart des erreurs factuelles ne sont pas signalées.
- Une incapacité à détecter les erreurs produites avec une grande assurance syntaxique.
- Un protocole d’évaluation à juge unique, qui introduit un biais et dont la fiabilité ne peut être mesurée.
- Un score composite linéaire qui laisse la qualité de génération masquer un mauvais comportement.

Des pistes d’amélioration concrètes sont proposées : un score composite non linéaire, une validation multi-juges avec mesure de l’accord inter-évaluateurs, et une détection statistique de la dérive. Aucun de ces résultats ne doit être interprété comme une certification ou une garantie de performance.

2 Contexte et objet

Il s’agit d’une revue méthodologique indépendante. Ce n’est ni une certification, ni une validation officielle, ni une caution commerciale de NeoMundi.

La revue repose sur une unique cohorte comparative réalisée en mai 2026. Elle porte sur huit fournisseurs anonymisés, désignés uniquement par P-001 à P-008, évalués sur près de 800 questions issues d’un benchmark public de véracité. L’identité des fournisseurs, les coûts exacts et les scores exacts par modèle ne sont pas

divulgués. Les chiffres sont rapportés sous forme d'ordres de grandeur ou de fourchettes, uniquement lorsque cela est nécessaire pour suivre le raisonnement méthodologique.

L'objet est circonscrit : documenter ce que mesure le signal en temps réel, là où il est fort, là où il est faible, et quelles améliorations méritent d'être testées.

3 Ce que mesure le signal en temps réel

Le signal en temps réel suit la stabilité de la dynamique de génération du modèle, et non le sens de la sortie. En pratique, il produit un score de stabilité continu (noté G), une mesure de sa variation (ΔG) et un taux de FLAG pour les anomalies détectées. Le texte lui-même n'est jamais inspecté.

La cohorte a fait apparaître deux groupes de comportement distincts.

Groupe de modèles	Stabilité de génération	Comportement de détection en temps réel	Limite observée
Modèles à moindre coût (p. ex. P-004, P-007, P-008)	Variation plus élevée pendant la génération (ΔG autour de 0,02 ou plus).	La détection en temps réel y est la plus efficace. Précision supérieure à 80 % ; les signalements correspondent presque toujours à une réelle dérive sémantique.	Aucune limite spécifique à la détection ; ce sont les cas favorables.
Modèles haut de gamme (p. ex. P-001, P-002, P-005, P-006)	Excellente stabilité (*G* moyen supérieur à 0,91).	Très peu d'anomalies remontées.	Le rappel s'effondre (sous ~5 %, parfois sous 1 %) : le système manque presque tous les énoncés faux réellement produits.

L'enseignement central est structurel. Un score de stabilité élevé reflète un processus de génération régulier et fluide. À lui seul, il n'indique pas que le contenu est vrai.

4 Observation méthodologique clé

Un modèle de la cohorte, désigné P-003, illustre ce point de façon frappante. Il a enregistré la plus mauvaise exactitude du groupe — environ 24 % de réponses correctes, soit un taux d'erreur supérieur à 75 %.

Pourtant, d'un strict point de vue statistique, ses signaux de génération étaient quasi parfaits. Son score de stabilité moyen était le plus élevé de la cohorte (supérieur à 0,92), sa variation était quasi nulle et il n'a déclenché aucun signalement. Le modèle a produit d'importantes erreurs factuelles de manière très fluide et stable.

> **Stable n'est pas synonyme d'exact.** La régularité mathématique pendant la génération de tokens n'est pas une garantie de véracité. Un modèle peut se tromper avec assurance, et un signal fondé sur la seule stabilité ne le verra pas.

5 Limites identifiées

5.1 Compensation dans le composite linéaire

Le composite actuel combine la stabilité continue et le taux de signalement par une simple moyenne :

$$\text{Composite}_{\text{v1}} = 0{,}5 \cdot G + 0{,}5 \cdot (1 - \text{FLAG}_{\text{rate}})$$

Comme la forme est linéaire, la régularité peut masquer un comportement défaillant. Pour P-003, l'absence de signalements combinée à un score de stabilité élevé donne un composite supérieur à 0,96. Le modèle atteint donc une tranche de notation élevée tout en se trompant dans environ trois cas sur quatre.

5.2 Biais du juge unique

Le protocole repose actuellement sur un unique juge LLM pour valider les réponses du benchmark. Cela pose deux problèmes. D'abord, si le juge provient du même fournisseur qu'un modèle testé, il peut surévaluer les réponses qui partagent son style ou son alignement. Ensuite, un évaluateur unique rend impossible la mesure de la fiabilité de l'évaluation elle-même, faute d'accord inter-évaluateurs à calculer.

5.3 Biais de sélection lié à l'arrêt anticipé

Arrêter la génération dès qu'un signalement apparaît est une optimisation de coût efficace. Mais cela fausse l'analyse de tendance. En tronquant le texte instable, le système biaise ses fenêtres de tendance vers les fragments jugés propres. Cela améliore artificiellement le score de stabilité continu de modèles par ailleurs volatils.

5.4 Seuils rigides et dérives lentes

La dérive temporelle est actuellement calculée comme la différence brute entre deux blocs de requêtes, par rapport à un seuil fixe unique. Cela présente deux faiblesses :

- Elle ignore la variance propre à chaque modèle. Pour un modèle fluctuant, un faible écart peut n'être qu'un bruit de fond déclenchant de fausses alertes. Pour un modèle ultra-stable, le même écart peut signaler une rupture réelle, ici sous-estimée.
- Elle est aveugle à la dégradation lente et progressive (*dérive douce*). Si la qualité s'érode par petits pas au fil de mises à jour silencieuses du fournisseur, l'écart bloc à bloc peut ne jamais franchir le seuil, et un déclin de plusieurs mois peut passer inaperçu.

6 Améliorations proposées

6.1 Un score composite non linéaire

Pour supprimer l'effet de compensation, deux formes non linéaires sont proposées. Leurs hyperparamètres devraient faire l'objet d'un benchmarking dédié pour être calibrés.

OPTION A — FORME MULTIPLICATIVE (ET VARIANTE PONDÉRÉE)

$$\text{Composite}_{\text{v2A}} = G \times (1 - \text{FLAG}_{\text{rate}})$$

$$\text{Composite}_{\text{v2A}} = G^{\beta} \times (1 - \text{FLAG}_{\text{rate}})^{\alpha} \quad \text{avec } \alpha > \beta$$

Dans une structure multiplicative, un taux de signalement croissant fait immédiatement chuter le score global, sans possibilité d'être compensé par un bon score de stabilité. Les exposants permettent d'ajuster la pénalité sur les erreurs par rapport à la qualité de génération.

OPTION B — PÉNALITÉ EXPONENTIELLE

$$\text{Composite}_{\text{v2B}} = G \times e^{-k \cdot \text{FLAG}_{\text{rate}}}$$

Ici, le paramètre *k* agit comme un curseur de sévérité : plus il est élevé, plus le score s'effondre brutalement dès les premiers signalements détectés.

NOTATION HORS LIGNE — INTÉGRER L'EXACTITUDE SÉMANTIQUE

Pour les benchmarks hors ligne, la note publiée peut intégrer directement l'exactitude sémantique au composite, sous forme de multiplicateur :

$$\text{Note} = \text{Composite} \times \text{Exactitude}$$

C'est direct et ne requiert aucune machinerie complexe. Cela fait chuter à une note basse un modèle comme P-003 — stable mais largement incorrect.

6.2 Double juge et accord inter-évaluateurs

Pour supprimer le biais de juge unique lors des phases de test hors ligne, un second juge LLM indépendant est recommandé. Pour mesurer dans quelle mesure les deux évaluateurs s'accordent réellement — au-delà du hasard — le protocole devrait calculer le Kappa de Cohen :

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

Ici, p_o est la proportion d'accords observés et p_e la proportion attendue par hasard. Une valeur supérieure à 0,80 indique un accord fort ; une valeur plus faible signale une ambiguïté dans le jeu de données. Si le dispositif passe à plus de deux juges, le Kappa de Fleiss devrait être utilisé à la place pour rendre compte de l'accord global.

Lorsque les juges restent en désaccord, une règle d'arbitrage neutre peut s'appliquer : la réponse du modèle et les réponses de référence sont converties en vecteurs (embeddings), et la réponse de référence la plus proche dans cet espace tranche le cas.

6.3 Validation statistique des tendances

Pour corriger l'analyse de dérive par blocs, deux méthodes tenant compte de la variance sont proposées.

TEST T DE WELCH — PRENDRE EN COMPTE LA VARIANCE

Au lieu d'un seuil fixe, un écart entre fenêtres récentes et anciennes n'est traité comme une dérive que s'il est statistiquement significatif (probabilité de pur hasard inférieure à 5 %) :

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

Cela élimine les fausses alertes sur les modèles naturellement bruités et révèle les véritables ruptures sur les modèles stables.

CUSUM — CAPTER LA DÉRIVE LENTE

Pour détecter la dégradation progressive, le système suit la somme cumulée des écarts par rapport à la moyenne historique du modèle. De petites pertes successives s'accumulent dans le temps. Lorsque le total cumulé franchit un niveau critique, une alerte est déclenchée. Cela révèle les déclins continus, invisibles d'un bloc à l'autre.

7 Feuille de route de recherche publique

- Évaluer les fonctions composites candidates sur des historiques d'exécution, et calibrer empiriquement leurs hyperparamètres pour séparer au mieux les modèles défaillants des modèles sains.
- Passer d'un juge unique à un dispositif multi-juges mesuré, en rapportant l'accord inter-évaluateurs.
- Remplacer les seuils de dérive fixes par une détection statistique de changement tenant compte de la variance.
- Intégrer l'exactitude sémantique à la notation hors ligne, afin que la stabilité ne puisse jamais se substituer à la vérité.

8 Conclusion

Sur la cohorte comparative de mai 2026, l'approche en temps réel montre une pertinence méthodologique pour détecter la dérive comportementale sans inspection sémantique directe du contenu.

Plusieurs limites subsistent : une sensibilité réduite sur les modèles haut de gamme, le biais de juge unique et la faiblesse de compensation de la formule linéaire actuelle. Les pistes proposées — un composite multiplicatif ou exponentiel, une validation multi-juges avec le Kappa de Cohen ou de Fleiss, et une détection de dérive par Welch / CUSUM — renforceraient la robustesse du protocole tout en le gardant simple sur le plan opérationnel.

Cette revue doit se lire comme une contribution ouverte à la consolidation méthodologique, et non comme une preuve définitive. La prochaine étape est empirique : tester ces options sur l'historique d'exécutions existant.

9 Avertissement

Ce document est une revue méthodologique indépendante, contribuée par **Abdelkrim Halimi**, Data Scientist indépendant (contributeur). Il s'agit d'une revue méthodologique menée de manière indépendante. Ce **n'est pas** une certification, une validation officielle, une garantie de performance ou une caution commerciale de NeoMundi. Les chiffres sont rapportés sous forme d'ordres de grandeur ou de fourchettes et ne divulguent ni l'identité des fournisseurs, ni les jeux de données propriétaires, ni les données de benchmark commercialement sensibles.