NOTE DE RÉPLICATION EXPLORATOIRE

Stabilité et reproductibilité du signal runtime NeoMundi ControlTower

Analyse comparative multi-providers, multi-corpus et multi-configurations

Fatima Ezzahrae GOUARAB

Data Scientist · Contributrice scientifique, Observatoire NeoMundi Recherche

neomundi.org/contributeurs

Note de statut. Cette note constitue un rapport exploratoire externe. Elle ne prétend pas établir une validation définitive du dispositif NeoMundi ControlTower, mais fournir des éléments méthodologiques sur la stabilité des signaux de gouvernance dans les conditions expérimentales étudiées. À publier après consolidation méthodologique.

Résumé

La reproductibilité constitue une étape essentielle dans l'évaluation de tout dispositif expérimental. Dans le domaine des grands modèles de langage (LLM), cette question revêt une importance particulière en raison de l'évolution continue des modèles commerciaux, de la diversité des fournisseurs et de l'influence potentielle des paramètres de génération. Cette note porte sur la **reproductibilité des signaux de gouvernance produits par NeoMundi ControlTower** lorsque les conditions expérimentales varient. Elle n'a pas pour objectif d'étudier la reproductibilité des grands modèles de langage eux-mêmes, ni la variabilité intrinsèque de leurs réponses, mais d'évaluer si les signaux de gouvernance produits par NeoMundi conservent une lecture cohérente lorsque le fournisseur du modèle, la température de génération ou la nature des prompts sont modifiés.

Cette étude s'inscrit dans le prolongement d'un premier travail consacré à l'actionnabilité des signaux de gouvernance. Lors de nouvelles exécutions du protocole initial, une inversion inattendue de la distribution des décisions *ALLOW* et *FLAG* a été observée, conduisant à interroger la stabilité des observations ayant servi de base à cette première étude. Afin d'examiner cette question, quatre campagnes expérimentales successives ont été conduites, faisant varier progressivement la température de génération, le fournisseur du modèle, la taille du corpus de prompts ainsi que la nature des situations étudiées. Au total, **680 générations** ont été analysées à partir de trois fournisseurs de modèles et de deux configurations de température.

Les résultats montrent que les campagnes reposant sur les corpus classiques produisent des proportions très faibles de décisions *FLAG*, indépendamment du fournisseur et de la température. En revanche, l'utilisation d'un corpus de *stress-test*, conçu pour solliciter des situations reconnues comme propices aux dérives des modèles de langage, entraîne une augmentation sensible du nombre de décisions *FLAG*, avec une intensité variable selon le fournisseur. Parmi les facteurs étudiés, **la nature des prompts apparaît comme celui qui influence le plus les décisions de gouvernance**, tandis que **la température ne semble pas jouer un rôle déterminant dans les conditions expérimentales étudiées**. Par ailleurs, une correspondance systématique est observée entre le profil ΔG *DROP* et les décisions *FLAG*, confirmant la cohérence interne du signal NeoMundi dans les configurations testées.

Cette étude constitue une première contribution méthodologique à l'analyse de la reproductibilité des signaux de gouvernance de NeoMundi. Les résultats montrent une cohérence globale des tendances observées dans les configurations étudiées et ouvrent la voie à des campagnes de réplication plus larges afin d'en confirmer la robustesse.

Mots-clés : reproductibilité, gouvernance des IA, grands modèles de langage, NeoMundi ControlTower, signaux de gouvernance, réplication expérimentale, ΔG, *ALLOW*, *FLAG*, stress-test, multi-providers.

Table des matières

Liste des figures

Liste des tableaux

1. Introduction

1.1. NeoMundi ControlTower : cadre et positionnement

NeoMundi ControlTower est une plateforme de gouvernance runtime des grands modèles de langage, développée par NeoMundi Recherche dans le cadre de l'Observatoire des signaux IA. Son principe repose sur la mesure en temps réel de la stabilité de la génération token par token, via un signal appelé G-Score et sa dynamique de variation (ΔG), transmis au client sous forme de flux SSE (*Server-Sent Events*) pendant la génération.

Contrairement aux approches classiques d'évaluation des LLM qui opèrent *a posteriori* sur une réponse complète et finalisée, ControlTower produit un signal pendant la génération, selon un paradigme qualifié de *runtime governance*. Le signal émis est de nature binaire au niveau de la décision finale : ALLOW lorsque la génération est jugée stable, FLAG lorsqu'une dérive est détectée. Il est accompagné d'un profil de variation ΔG caractérisant la dynamique de cette stabilité au cours de la génération : DROP (chute sans récupération), FLAT (stabilité plate) ou V_SHAPE (chute suivie d'une récupération partielle).

L'architecture actuelle, dite mode ENF (Gateway), ne produit pas d'action automatique sur la génération : elle expose un signal externe que le système client peut exploiter selon sa propre politique de gouvernance — arrêt anticipé, régénération, reroutage. Cette architecture s'inscrit dans la continuité du Filtre E (Favre-Lecca, 2026a), qui constituait une première version de régulation automatique fondée sur une notion d'énergie informationnelle, et dont ControlTower représente l'évolution vers un modèle de gouvernance décentralisée côté client.

Les travaux de l'Observatoire NeoMundi sur la cartographie thermodynamique des services génératifs (NeoMundi Recherche, 2026) et l'audit exploratoire des signaux runtime (Diop, 2026) ont posé les premières bases empiriques de la caractérisation de ce signal sur plusieurs providers. La présente note s'inscrit dans cette continuité, en se concentrant spécifiquement sur la question de la reproductibilité des décisions NeoMundi lorsque les conditions expérimentales varient.

1.2. Périmètre et objet de cette note

La présente note n'étudie pas la reproductibilité des grands modèles de langage en tant que tels, ni la variabilité de leurs réponses d'une exécution à l'autre. Son objet est plus spécifique : elle étudie la reproductibilité des signaux de gouvernance produits par NeoMundi ControlTower lorsque les conditions expérimentales varient. La question posée n'est pas celle de savoir si les modèles produisent les mêmes réponses, mais si NeoMundi produit les mêmes décisions de gouvernance lorsque l'on fait varier le provider, la température ou la nature des prompts. Cette distinction place NeoMundi ControlTower lui-même au centre de l'analyse, comme objet d'évaluation et non comme simple outil de

mesure.

Cette note se distingue également du rapport de Phase 1, qui portait sur l'actionnabilité du signal ControlTower dans un scénario de stop-generation. Ces deux contributions sont complémentaires : l'actionnabilité n'a de sens opérationnel que si le signal sur lequel elle repose est stable et reproductible. La présente note constitue donc un préalable méthodologique nécessaire à toute généralisation des conclusions du rapport de Phase 1.

1.3. Point de départ : l'inversion observée

Les premières expérimentations menées dans le cadre de cette contribution avaient pour objectif d'étudier l'actionnabilité des signaux de gouvernance ControlTower. Ces expérimentations, réalisées avec PROFILE-5B3BB7 à température 0,7 sur un corpus de 50 prompts organisés en trois catégories (factuelles, raisonnement et spéculatives), ont conduit à un taux de FLAG de 88 %, avec une bijection parfaite entre profil ΔG DROP et décision FLAG.

Plusieurs relances de ce même protocole ont ensuite été effectuées. La majorité de ces relances confirmait la tendance initiale, caractérisée par une prédominance des décisions FLAG. Lors d'une de ces relances cependant, le résultat s'est inversé de façon inattendue : la décision ALLOW est devenue largement majoritaire, alors que le code, le modèle et le corpus de prompts restaient identiques. Cette inversion concernait l'ensemble du corpus et renversait complètement la distribution des décisions observée initialement.

1.4. Hypothèses formulées

Face à ce constat d'irréproductibilité, quatre hypothèses explicatives ont été formulées.

La première hypothèse attribue l'inversion à la température de génération. La Phase 1 avait utilisé T=0,7, paramètre qui n'avait pas été explicitement contrôlé lors des relances. Une température plus élevée favorisant la variabilité des réponses, elle pourrait mécaniquement augmenter la fréquence des dérives détectées par NeoMundi.

La deuxième hypothèse attribue l'inversion à un effet de version du modèle. L'alias `PROFILE-5B3BB7` (version : latest) utilisé en Phase 1 ne référence pas une version figée du modèle. Une mise à jour silencieuse entre les deux exécutions aurait pu modifier le comportement de génération sans changement apparent dans le code.

La troisième hypothèse attribue l'inversion à un effet provider. Le comportement observé sur PROFILE-5B3BB7 pourrait être spécifique à ce modèle, d'autres providers présentant des profils de décision différents sur le même corpus.

La quatrième hypothèse attribue l'inversion à un effet du type de prompt. Les prompts de Phase 1 pourraient avoir des propriétés particulières, notamment la présence de questions portant sur des événements futurs non encore survenus, les rendant structurellement plus

propices à déclencher un FLAG indépendamment du modèle ou de la température.

Ces quatre hypothèses ne sont pas mutuellement exclusives. La présente note ne prétend pas les dissocier complètement ni les tester avec une puissance statistique suffisante pour conclure de façon définitive. Elle vise à documenter, de façon transparente et itérative, les effets observés dans des conditions contrôlées mais imparfaites.

1.5. Question de recherche et objectif

La question centrale de cette note est la suivante : les signaux de gouvernance produits par NeoMundi ControlTower conservent-ils une lecture cohérente lorsque les prompts, les fournisseurs et les paramètres d'exécution changent ? Cette formulation rejoint celle retenue par S. Favre-Lecca pour cadrer cette réplication dans le cadre de l'Observatoire.

L'objectif est de produire une note de réplication exploratoire documentant de façon rigoureuse les résultats obtenus dans des conditions contrôlées mais imparfaites. Il s'agit d'une première démonstration méthodologique prudente au sens formulé par S. Favre-Lecca : dans les conditions testées, malgré l'élargissement du corpus et la variation des configurations, les tendances principales restent-elles cohérentes ? Le protocole s'est construit par itérations successives, chaque résultat orientant la question suivante. Ce document restitue cette chronologie de façon transparente plutôt que de la lisser artificiellement.

2. Protocole et configurations testées

Le protocole de réplication s'est construit de façon progressive, chaque essai étant conçu en réponse directe aux observations du précédent. Cette démarche itérative, qui s'écarte d'un plan expérimental entièrement prédéfini, est assumée comme telle et restituée dans sa chronologie réelle plutôt que présentée sous une forme rétrospectivement rationalisée. Quatre essais successifs ont ainsi été conduits, chacun faisant varier une ou plusieurs dimensions du protocole initial tout en maintenant le dispositif technique constant.

2.1. Essai 1 — Effet de la température (n=50 par configuration)

Le premier essai visait à tester l'hypothèse selon laquelle la température de génération constituait un facteur explicatif de l'inversion observée. Le corpus complet des 50 prompts de Phase 1 a été soumis deux fois au même modèle (`PROFILE-5B3BB7`), en faisant varier uniquement la température : T=0,7, correspondant à une nouvelle génération dans la configuration d'origine, et T=0, correspondant à une configuration déterministe éliminant toute variabilité stochastique. Ce dispositif permettait d'observer si un changement de température seul, toutes choses égales par ailleurs, produisait des décisions NeoMundi différentes.

2.2. Essai 2 — Effet du provider (n=50 par provider)

Le second essai a consisté à soumettre le corpus complet des 50 prompts de Phase 1 à deux providers supplémentaires, PROFILE-739F7C et PROFILE-F5FF91, à température 0. L'objectif était de vérifier si le comportement observé sur PROFILE-5B3BB7 lors de la relance de Phase 1 était spécifique à ce modèle ou se manifestait également sur d'autres architectures soumises aux mêmes prompts dans les mêmes conditions.

La température a été fixée à 0 pour cet essai afin de s'affranchir de la variabilité stochastique de génération et d'isoler plus proprement l'effet provider. Chaque provider a été testé dans une session d'exécution indépendante, avec les mêmes 50 prompts soumis dans le même ordre, garantissant ainsi la comparabilité des conditions entre les trois configurations.

2.3. Essai 3 / Campagne A — Élargissement du corpus (n=100 par provider)

À l'issue des Essais 1 et 2, les taux de FLAG observés restaient très faibles sur l'ensemble des configurations testées. Cependant, l'effectif limité à 50 prompts et la restriction au seul corpus de Phase 1 laissaient plusieurs interprétations ouvertes. La Campagne A a donc consisté à élargir le corpus à 100 prompts par provider, en conservant les mêmes catégories de la Phase 1 et en y ajoutant une catégorie supplémentaire de prompts contenant des fautes d'orthographe et de syntaxe, destinée à tester si une dégradation de la formulation affectait les décisions NeoMundi indépendamment du contenu sémantique.

Cette campagne a été exécutée séparément pour chacun des trois providers dans trois notebooks distincts. Ce choix d'architecture séparée répondait à une contrainte opérationnelle de robustesse : un incident technique sur un provider ne devait pas compromettre la collecte sur les deux autres, chaque campagne pouvant être relancée indépendamment en cas de défaillance partielle.

2.4. Essai 4 / Campagne B — Corpus stress-test et contrôle de température (n=45 par configuration)

La Campagne A avait confirmé la chute du taux de FLAG sur les catégories classiques sans permettre de distinguer entre deux interprétations également plausibles : soit les modèles testés étaient devenus plus stables de façon générale, soit le corpus classique ne sollicitait plus les mécanismes susceptibles de produire des dérives détectables par NeoMundi. Pour départager ces deux hypothèses, la Campagne B a délibérément remplacé les catégories de prompts plutôt que d'en augmenter le volume, en construisant un corpus entièrement nouveau de 45 prompts conçus selon une logique de stress-test ciblé. La description détaillée de ce corpus et la justification du choix de chaque catégorie sont présentées en section 3.

Parallèlement, la Campagne B a regroupé quatre configurations dans un notebook unique, permettant une comparaison directe et simultanée : PROFILE-5B3BB7 à T=0,7, PROFILE-5B3BB7 à T=0, PROFILE-739F7C à T=0 et PROFILE-F5FF91 à T=0. Le fait de maintenir la température à 0 pour trois des quatre configurations permettait d'attribuer tout écart observé entre providers au modèle lui-même plutôt qu'à la variabilité stochastique de génération. La configuration PROFILE-5B3BB7 à T=0,7 était maintenue comme point de référence direct avec la Phase 1.

2.5. Dispositif technique commun

Dans l'ensemble des essais, le dispositif technique est identique à celui de la Phase 1. Chaque génération est soumise à l'API NeoMundi ControlTower en mode ENF (Gateway, streaming SSE). Dans ce mode, NeoMundi appelle directement le provider LLM et mesure la stabilité de la génération token par token via un flux SSE, sans intermédiaire côté client. Les paramètres d'appel sont constants pour l'ensemble des configurations : un maximum de 512 tokens générés et un délai d'attente de 120 secondes par requête.

Pour chaque génération, les variables suivantes sont extraites depuis l'événement *done* du flux SSE. La décision finale de gouvernance (ALLOW ou FLAG) constitue la variable de résultat principale de cette étude. Le score de stabilité global (`stability_score`) mesure la stabilité de la génération sur une échelle de 0 à 1. Le profil de variation ΔG (`delta_profile`) caractérise la dynamique de cette stabilité au cours de la génération selon trois configurations : DROP, FLAT ou V_SHAPE. L'amplitude totale de la variation (`delta_variation`) et la série temporelle complète des points de mesure ΔG

(`delta_series`) complètent la description du comportement dynamique du signal. Le score d'hallucination factuelle (`hallucination_score`) et le score de cohérence sémantique (`coherence_score`) fournissent des informations complémentaires sur la nature des dérives détectées. Le nombre total de tokens générés (`total_tokens`) permet de contextualiser les observations en fonction de la longueur des réponses produites.

Le tableau récapitulatif de l'ensemble des configurations testées dans le cadre de cette réplication est disponible en Annexe A (voir Tableau 12).

Remarque. Les noms des trois fournisseurs évalués dans cette étude ont été anonymisés sous la forme des alias `PROFILE-5B3BB7`, `PROFILE-739F7C` et `PROFILE-F5FF91` pour des raisons de confidentialité. Cette anonymisation n'a aucun impact sur les analyses ni sur les résultats rapportés. En revanche, elle empêche toute vérification indépendante par des tiers externes et ne permet pas d'établir une correspondance directe avec d'autres évaluations publiques portant sur les mêmes modèles.

3. Description des corpus et catégories de prompts

3.1. Corpus de référence — Phase 1 et Essais 1 et 2 (n=50)

Le corpus utilisé en Phase 1, repris à l'identique pour les Essais 1 et 2, est composé de 50 prompts rédigés en français et organisés selon un gradient de complexité croissant en trois catégories. Cette structuration répondait à un objectif de validité de construit : plutôt que de mesurer les décisions NeoMundi sur un ensemble homogène de questions, il s'agissait de couvrir un spectre de situations génératives suffisamment contrasté pour observer si le signal réagissait différemment selon la nature de la tâche demandée au modèle.

La catégorie **Factuelle** regroupe 15 prompts portant sur des questions à réponse unique et vérifiable, relevant de domaines aussi variés que la géographie, l'histoire, les sciences ou l'économie. Ces questions constituent la condition de référence dans laquelle la stabilité de la génération est attendue comme maximale, le modèle disposant en principe d'une réponse canonique bien établie dans ses données d'entraînement. Il convient cependant de noter que certains prompts de cette catégorie portaient sur des événements non encore survenus à la date d'exécution, comme la remise de prix annuels ou des résultats sportifs futurs. Ces cas particuliers exposent structurellement le modèle à une forme spécifique de dérive appelée *hallucination temporelle*, indépendamment de leur formulation factuelle apparente. Ce point constitue un facteur d'hétérogénéité interne à la catégorie Factuelle, documenté ici pour en tenir compte dans l'interprétation des résultats.

La catégorie **Raisonnement** regroupe 20 prompts demandant au modèle d'expliquer un mécanisme, de distinguer deux concepts ou de décrire le fonctionnement d'un système. Ces questions impliquent une chaîne explicative à plusieurs étapes, dans laquelle le risque de dérive est attendu comme plus progressif et cumulatif que dans les questions factuelles. L'effectif légèrement supérieur de cette catégorie reflétait l'hypothèse initiale selon laquelle les profils de dérive intermédiaires y seraient plus fréquents et nécessiteraient davantage d'observations pour être caractérisés de façon stable.

La catégorie **Spéculative** regroupe 15 prompts portant sur des questions ouvertes, normatives ou prospectives, dépourvues de réponse canonique vérifiable. Ces questions représentaient en Phase 1 la condition dans laquelle le taux de FLAG était attendu comme le plus élevé, le modèle étant exposé à une expansion narrative peu contrainte par des points de vérification factuels.

Il convient de souligner que la répartition 15/20/15 entre ces trois catégories n'a pas résulté d'un calcul de taille d'échantillon différencié, mais d'un choix raisonné de couverture. Il s'agit d'un échantillonnage de convenance à finalité contrastive et non d'un tirage aléatoire au sein d'une population de prompts définie. Cette limite implique que les taux de FLAG observés par catégorie ne peuvent pas être interprétés comme des estimateurs non biaisés d'un taux populationnel, mais caractérisent le comportement du signal sur ce corpus précis.

3.2. Corpus élargi — Campagne A (n=100)

Le corpus de la Campagne A prolonge directement le corpus de Phase 1 selon la même logique de gradient de complexité, en portant chaque catégorie à environ le double d'effectif : 30 prompts pour la catégorie Factuelle, 35 pour le Raisonnement et 30 pour le Spéculatif. Les nouveaux prompts ont été conçus pour rester cohérents avec les thématiques et le niveau de complexité des prompts initiaux, afin de préserver la comparabilité entre les deux corpus.

Une quatrième catégorie a été introduite pour la première fois dans cette campagne, composée de 5 prompts dont la formulation présente des fautes d'orthographes et syntaxiques, sans modification du contenu sémantique de la question posée. L'objectif de cette catégorie était de vérifier si une formulation dégradée affectait les décisions NeoMundi indépendamment du contenu, c'est-à-dire si le signal réagissait à la forme autant qu'au fond. Avec un effectif de 5 prompts seulement, cette catégorie a une fonction illustrative et non probatoire ; ses résultats sont présentés à titre indicatif dans les analyses.

3.3. Corpus stress-test — Campagne B (n=45)

Le corpus de la Campagne B rompt délibérément avec la logique des corpus précédents. Plutôt que d'étendre le gradient de complexité classique, il repose sur une logique de stress-test ciblé : chaque catégorie vise un mécanisme de dérive ou d'hallucination spécifique, documenté dans la littérature sur les LLM, que les corpus classiques ne sollicitaient pas ou peu. Ce choix de conception répondait directement à l'hypothèse formulée à l'issue de la Campagne A, selon laquelle la chute du taux de FLAG pouvait refléter une inadéquation du corpus existant plutôt qu'une stabilisation générale des modèles.

La catégorie **Actualité** regroupe 6 prompts portant sur des événements récents ou non encore survenus à la date d'exécution. Les modèles LLM, dont la connaissance est figée à une date d'entraînement, sont structurellement exposés à l'hallucination temporelle sur ce type de questions : faute de pouvoir vérifier l'information, ils tendent à produire une réponse plausible mais non fondée sur des faits réels.

La catégorie **Chiffres précis** regroupe 6 prompts sollicitant des valeurs numériques exactes, qu'il s'agisse de statistiques économiques, de scores de benchmarks ou de données démographiques, que le modèle ne peut connaître avec certitude absolue. Ce type de questions incite fréquemment à la production d'un chiffre crédible mais inventé, mécanisme d'hallucination quantitative bien documenté dans les études sur la fiabilité factuelle des LLM.

La catégorie **Entité fictive** regroupe 6 prompts présentant au modèle des noms de théories, de personnes ou d'organismes qui n'existent pas. Elle teste la propension du modèle à confabuler une réponse cohérente plutôt qu'à signaler l'absence de référent connu dans ses

données d'entraînement.

La catégorie **Citation** regroupe 6 prompts demandant la reproduction exacte d'une citation accompagnée de sa source précise. La reproduction fidèle de citations constitue l'un des mécanismes d'hallucination les plus fréquemment documentés dans la littérature sur les LLM : le modèle tend à reconstruire une formulation plausible plutôt qu'à reproduire le texte original, produisant des citations vraisemblables mais inexactes.

La catégorie **Fausse prémisse** regroupe 6 prompts dont chacun est construit sur un postulat manifestement faux. Elle teste si le modèle corrige la prémisse erronée ou poursuit son raisonnement comme si elle était vraie, ce second comportement étant associé à une expansion narrative non fondée.

La catégorie **Multi-étapes** regroupe 6 prompts imposant une chaîne de calcul ou de raisonnement à plusieurs étapes successives. Le risque de dérive y est attendu comme cumulatif, chaque étape de la chaîne inférentielle pouvant introduire une erreur qui se propage aux suivantes.

La catégorie **Spéculatif extrême** regroupe 4 prompts portant sur des questions normatives ou philosophiques sans réponse factuelle possible, relevant de sujets tels que les droits des intelligences artificielles ou la conscience artificielle. Ce terrain est associé à une forte densité informationnelle et à une expansion narrative peu contrainte par des points de vérification factuels, ce qui en faisait un candidat naturel pour le stress-test.

Enfin, la catégorie **Fautes** regroupe 5 prompts réutilisés à l'identique depuis la Campagne A. Ce choix délibéré de réutilisation visait à permettre une comparaison directe de l'effet d'une formulation dégradée entre les deux campagnes, toutes choses égales par ailleurs quant au contenu.

Le corpus de la Campagne B totalise ainsi 45 prompts : 40 entièrement nouveaux répartis dans les sept catégories stress-test décrites ci-dessus, et 5 réutilisés à l'identique depuis la Campagne A.

4. Résultats

Les résultats sont présentés en réponse aux six questions de recherche formulées à partir des hypothèses de départ. Pour chaque question, un tableau synthétique accompagne l'interprétation analytique. Les figures sont insérées au fil du texte.

4.1. Les décisions NeoMundi sont-elles stables d'un essai à l'autre ?

Cette première question porte directement sur la reproductibilité des décisions de gouvernance produites par NeoMundi ControlTower lorsque les conditions expérimentales varient.

Table 1. Distribution des décisions NeoMundi par campagne

Campagne	n	ALLOW	FLAG	% ALLOW	% FLAG
Essai 1 — Température (n=50×2)	100	98	2	98,0%	2,0%
Essai 2 — Fournisseur (n=50×2)	100	97	3	97,0%	3,0%
Essai 3 — Campagne A (n=100×3)	300	294	6	98,0%	2,0%
Essai 4 — Campagne B (n=45×4)	180	159	21	88,3%	**11,7%**

Les Essais 1, 2 et 3 produisent des taux de FLAG stables et faibles (2,0–3,0%), indépendamment du provider et de la température. L'Essai 4, conduit sur corpus stress-test, fait remonter ce taux à 11,7% : non pas une anomalie, mais l'effet attendu d'un corpus conçu pour solliciter des mécanismes de dérive documentés.

4.2. Le type de prompt influence-t-il les décisions NeoMundi ?

Table 2. Décisions par catégorie — Campagne A (n=100 par provider, 3 providers)

Catégorie	ALLOW	FLAG	Total	% FLAG
Factuelle	84	6	90	**6,7%**
Raisonnement	105	0	105	0,0%
Spéculative	90	0	90	0,0%
Fautes	15	0	15	0,0%
Total	**294**	**6**	**300**	**2,0%**

Table 3. Décisions par catégorie stress-test — Campagne B (triées par taux de FLAG décroissant)

Catégorie	ALLOW	FLAG	Total	% FLAG
Actualité	16	8	24	**33,3%**
Fausse prémisse	20	4	24	**16,7%**
Entité fictive	21	3	24	**12,5%**
Chiffres précis	21	3	24	**12,5%**
Citation	22	2	24	8,3%
Multi-étapes	23	1	24	4,2%
Fautes	20	0	20	0,0%
Spéculatif extrême	16	0	16	0,0%
Total	**159**	**21**	**180**	**11,7%**

Le type de prompt est le facteur le plus discriminant. Dans la Campagne A, seule la catégorie Factuelle produit des FLAG (6,7%), attribuables aux prompts portant sur des événements futurs (hallucination temporelle, §3.1). Dans la Campagne B, le gradient est net : Actualité (33,3%) en tête, Spéculatif extrême et Fautes à 0%. Ce résultat indique que le signal NeoMundi réagit à l'incertitude factuelle vérifiable, non à l'incertitude normative.

4.3. Le provider influence-t-il les décisions NeoMundi ?

Table 4. Décisions par provider à température fixe T=0 — Essais 2, 3 et 4

Essai	Provider	ALLOW	FLAG	Total	% FLAG
Essai 2	PROFILE-739F7C	48	2	50	4,0%
	PROFILE-F5FF91	49	1	50	2,0%
Essai 3	PROFILE-5B3BB7	99	1	100	1,0%
	PROFILE-739F7C	97	3	100	3,0%
	PROFILE-F5FF91	98	2	100	2,0%
	PROFILE-739F7C_T00	37	8	45	**17,8%**
	PROFILE-5B3BB7_T00	40	5	45	**11,1%**
	PROFILE-F5FF91_T00	43	2	45	**4,4%**

Sur corpus classiques (Essais 2 et 3), aucune différence inter-provider n'est observable (1–4% de FLAG pour tous). Sur le corpus stress-test (Essai 4, T=0 fixée), un gradient émerge : PROFILE-739F7C (17,8%) > PROFILE-5B3BB7 (11,1%) > PROFILE-F5FF91 (4,4%). Cet effet mériterait une confirmation sur un corpus plus large avant toute généralisation (n=45 par configuration).

4.4. La température influence-t-elle les décisions NeoMundi ?

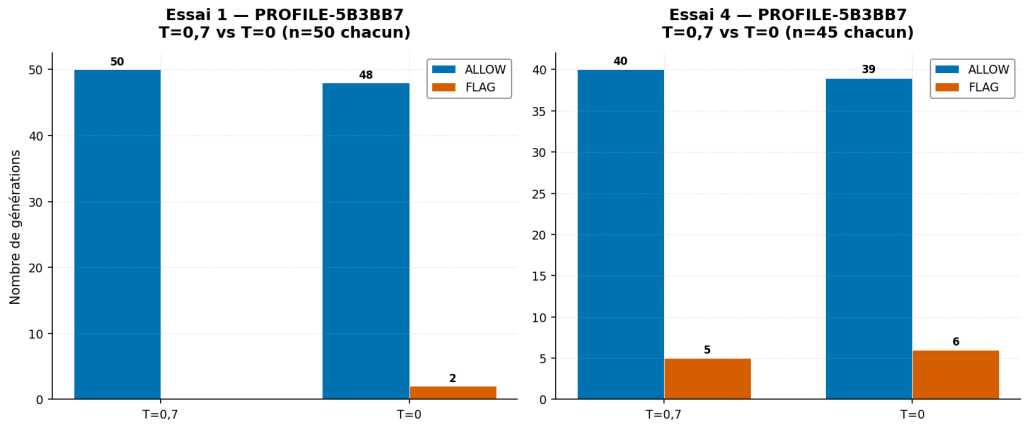

Figure 1. Effet de la température sur les décisions NeoMundi — Essai 1 (n=50 par configuration) et Essai 4 / Campagne B (n=45 par configuration)

Les distributions de décisions obtenues avec PROFILE-5B3BB7 à température 0 et 0,7 demeurent très similaires dans les deux campagnes. Dans l'Essai 1, le nombre de décisions *FLAG* passe de 0 à 2 (0 % contre 4 %), et dans l'Essai 4 de 5 à 6 (11,1 % contre 13,3 %). Ces différences restent de faible amplitude et ne permettent pas d'identifier la température comme le principal facteur de variation des décisions de gouvernance dans les conditions expérimentales étudiées.

4.5. Les autres métriques évoluent-elles de façon cohérente ?

Le score de stabilité est systématiquement plus faible pour les cas FLAG que pour les cas ALLOW, validant la cohérence entre métrique continue et décision binaire. L'écart moyen ALLOW−FLAG passe de 0,036 en Phase 1 à 0,195 dans les Essais 3 et 4 : la frontière de décision NeoMundi est plus nette et plus discriminante dans la réplication. Le score d'hallucination présente des valeurs non nulles principalement sur les catégories Actualité et Entité fictive, en cohérence avec leurs taux de FLAG élevés. Les statistiques descriptives complètes sont disponibles en Annexe B.

4.6. Les profils ΔG changent-ils selon le provider ou le type de prompt ?

Table 5. Profils ΔG par provider — Essai 3 (n=100 par provider)

Provider	DROP	FLAT	Total
PROFILE-739F7C	3	97	100
PROFILE-F5FF91	2	98	100
PROFILE-5B3BB7	1	99	100
Total	**6**	**294**	**300**

Table 6. Cohérence décision × profil ΔG — Essai 3 (n=300)

Profil ΔG	ALLOW	FLAG	Total
DROP	0	6	6
FLAT	294	0	294
Total	**294**	**6**	**300**

Table 7. Profils ΔG par configuration — Essai 4 (n=45 par configuration)

Configuration	DROP	FLAT	Total
PROFILE-739F7C_T00	8	37	45
PROFILE-F5FF91_T00	2	43	45
PROFILE-5B3BB7_T00	5	40	45
PROFILE-5B3BB7_T07	6	39	45
Total	**21**	**159**	**180**

La bijection DROP⇔FLAG est parfaite et constante sur l'ensemble des essais : 6 DROP pour 6 FLAG en Essai 3, 21 DROP pour 21 FLAG en Essai 4, sans exception. Le profil DROP est un prédicteur déterministe de la décision FLAG. À noter : le profil V_SHAPE, présent en Phase 1 (5 cas, 10%), est totalement absent de la réplication (0 cas sur 480 générations).

5. Comparaison avec les résultats antérieurs

Cette section confronte les résultats des campagnes de réplication aux observations de la Phase 1 selon quatre axes : distribution globale des décisions, profils ΔG, effet des catégories et score de stabilité. Il convient de rappeler qu'il ne s'agit pas d'une réplication stricte au sens expérimental — plusieurs variables ont changé simultanément entre Phase 1 et essais de réplication — et que les observations présentées restent de nature descriptive.

5.1. Évolution de la distribution des décisions

Table 8. Distribution des décisions — Phase 1 vs campagnes de réplication

Étude	n	ALLOW	FLAG	% FLAG
Phase 1 — PROFILE-5B3BB7, T=0,7	50	6	44	**88,0%**
Essai 1 — PROFILE-5B3BB7 , T=0,7	50	48	2	4,0%
Essai 1 — PROFILE-5B3BB7, T=0	50	50	0	0,0%
Essai 2 — PROFILE-739F7C, T=0	50	48	2	4,0%
Essai 2 — PROFILE-F5FF91, T=0	50	49	1	2,0%
Essai 3 — PROFILE-5B3BB7, T=0	100	99	1	1,0%
Essai 3 — PROFILE-739F7C, T=0	100	97	3	3,0%
Essai 3 — PROFILE-F5FF91, T=0	100	98	2	2,0%
Essai 4 — PROFILE-5B3BB7 T=0,7 (stress-test)	45	39	6	13,3%
Essai 4 — PROFILE-5B3BB7 T=0 (stress-test)	45	40	5	11,1%
Essai 4 — PROFILE-739F7C T=0 (stress-test)	45	37	8	**17,8%**
Essai 4 — PROFILE-F5FF91 T=0 (stress-test)	45	43	2	4,4%

Le contraste est net : la Phase 1 présentait 88,0% de FLAG, tandis que l'ensemble des campagnes sur corpus classique produit des taux de 0,0% à 4,0%, indépendamment du provider et de la température. Seul le corpus stress-test fait remonter ce taux, jusqu'à 17,8% pour PROFILE-739F7C, soit néanmoins cinq fois inférieur à la Phase 1. La comparaison la plus directe — Essai 1, mêmes 50 prompts — montre un écart de 84 points entre PROFILE-5B3BB7-latest à T=0,7 en Phase 1 et PROFILE-5B3BB7-2407 à T=0,7 en Essai 1, que la température à elle seule ne peut expliquer (effet propre de 4 points seulement). La version du modèle et la nature des prompts jouent un rôle prépondérant.

5.2. Évolution des profils ΔG

Table 9. Distribution des profils ΔG — Phase 1 vs campagnes de réplication

Profil	Phase 1 (n=50)	Essai 3 (n=300)	Essai 4 (n=180)
DROP	43 (86,0%)	6 (2,0%)	21 (11,7%)
FLAT	2 (4,0%)	294 (98,0%)	159 (88,3%)
V_SHAPE	5 (10,0%)	**0 (0,0%)**	**0 (0,0%)**

La bijection DROP⇔FLAG se maintient dans les deux périodes sans exception. En Phase 1, le profil DROP dominait (86%), cohérent avec le taux de FLAG élevé. Dans la réplication, le profil FLAT devient largement majoritaire (94,4% sur 480 générations). Le résultat le plus notable est la disparition totale du profil V_SHAPE : présent à 10% en Phase 1, il est absent des 480 générations des Essais 3 et 4, point d'attention pour les réplications futures.

5.3. Évolution par catégorie de prompt

Table 10. Taux de FLAG par catégorie — Phase 1 vs Essai 3

Catégorie	Phase 1 % FLAG	Essai 3 % FLAG
Raisonnement	90,0%	**0,0%**
Factuelle	86,7%	6,7%
Spéculative	86,7%	**0,0%**
Fautes	—	0,0%

Les catégories Raisonnement et Spéculative passent de 87–90% de FLAG en Phase 1 à 0% dans la réplication. Seule la catégorie Factuelle maintient un taux résiduel de 6,7%, attribuable aux prompts portant sur des événements futurs non encore survenus. Ces évolutions suggèrent que l'effet corpus est prépondérant sur tout effet modèle ou température pour expliquer l'écart entre les deux périodes.

5.4. Évolution du score de stabilité

Table 11. Score de stabilité par décision — Phase 1 vs campagnes de réplication

Indicateur	Phase 1	Essai 3	Essai 4
Score moyen global	0,795	0,923	0,923
Score moyen ALLOW	0,827	0,923	0,923
Score moyen FLAG	0,791	0,728	0,727
Écart ALLOW moins FLAG	0,036	**0,195**	**0,196**

Le score moyen ALLOW augmente de 0,827 à 0,923 entre Phase 1 et réplication, reflet de générations plus stables sur des corpus moins sollicitants. Plus significatif : l'écart ALLOW−FLAG passe de 0,036 en Phase 1 à 0,195 dans la réplication. La frontière de décision NeoMundi est plus nette dans les campagnes de réplication, où les FLAG correspondent à des dérives plus prononcées. Les rares FLAG de la réplication sont donc des signaux plus robustes que ceux de Phase 1.

5.5. Limites méthodologiques

Toute étude exploratoire de réplication implique de documenter avec transparence les limites du protocole expérimental. La présente étude ne fait pas exception. Les limites présentées ci-dessous doivent être prises en compte dans l'interprétation des résultats et dans l'appréciation de leur portée.

Évolution silencieuse des modèles commerciaux (*silent model updates*). La principale limite concerne l'évolution continue des modèles commerciaux. Les fournisseurs mettent régulièrement à jour leurs modèles afin d'améliorer leurs performances, de corriger certains comportements ou d'adapter leurs politiques de génération. Ces mises à jour sont souvent déployées sans modification du nom commercial du modèle et sans documentation publique détaillée.

Ainsi, deux expérimentations réalisées à plusieurs semaines ou plusieurs mois d'intervalle peuvent être exécutées avec ce qui semble être le même modèle, alors qu'elles reposent en réalité sur des versions internes différentes. Ces **changements silencieux** constituent aujourd'hui un défi majeur pour les études de reproductibilité impliquant des LLM commerciaux.

Dans le cadre de ce travail, cette hypothèse ne peut être ni confirmée ni infirmée. Elle représente toutefois une explication plausible de l'inversion des résultats observée entre les premières expérimentations de la Phase 1 et les exécutions ultérieures, alors même que le protocole expérimental paraissait inchangé. C'est précisément cette observation qui a motivé la réalisation des différentes campagnes de réplication présentées dans cette note.

Contrôle incomplet des variables expérimentales. Les campagnes réalisées permettent d'étudier plusieurs facteurs susceptibles d'influencer les observations, notamment le fournisseur du modèle, la température de génération ou la nature du corpus de prompts. En revanche, il n'est pas possible d'isoler parfaitement l'ensemble des variables susceptibles d'avoir évolué entre les premières expérimentations et les campagnes de réplication.

Par conséquent, les différences observées ne peuvent être attribuées avec certitude à une cause unique. Les interprétations proposées dans cette étude doivent être considérées comme des hypothèses méthodologiques fondées sur les observations expérimentales, plutôt que comme des démonstrations causales.

Représentativité des corpus. Les campagnes de réplication reposent sur plusieurs corpus de tailles différentes (50, 100 et 45 prompts), construits afin d'explorer des situations variées : questions factuelles, raisonnements, questions spéculatives et scénarios de stress-test. Bien que ces corpus permettent d'étudier un large éventail de comportements, ils ne prétendent pas représenter l'ensemble des usages possibles des grands modèles de

langage. Les résultats présentés doivent donc être interprétés comme représentatifs des configurations étudiées, sans être généralisés à tous les contextes d'utilisation.

Portée de l'étude. Enfin, cette note n'a pas pour objectif de valider définitivement le dispositif NeoMundi. Elle constitue une première étude méthodologique consacrée à la reproductibilité des signaux de gouvernance. Son ambition est de documenter les tendances observées dans plusieurs configurations expérimentales et de proposer un protocole de réplication susceptible d'être reproduit et enrichi dans des travaux futurs.

6. Conclusion

Cette note avait pour point de départ une observation simple mais déstabilisante : le même protocole expérimental, appliqué dans des conditions déclarées identiques, avait produit des distributions de décisions NeoMundi radicalement différentes d'une période à l'autre. Plutôt que d'ignorer cet écart ou de le minimiser, la démarche adoptée a consisté à le documenter rigoureusement et à construire, de façon itérative, un ensemble d'essais permettant d'en explorer les facteurs explicatifs. C'est cette transparence méthodologique, autant que les résultats eux-mêmes, qui constitue la contribution principale de cette note.

Les résultats obtenus permettent d'apporter une réponse nuancée à la question centrale posée : les signaux de gouvernance produits par NeoMundi ControlTower conservent-ils une lecture cohérente lorsque les prompts, les fournisseurs et les paramètres d'exécution changent ?

Sur les corpus classiques, la réponse est oui. Les Essais 1, 2 et 3 produisent des taux de FLAG remarquablement stables, compris entre 0,0% et 4,0%, sur trois providers distincts et deux températures différentes. Cette cohérence inter-essais constitue un résultat de stabilité encourageant pour l'instrument. La bijection parfaite entre profil ΔG DROP et décision FLAG, maintenue sans exception sur l'ensemble des 480 générations des Essais 3 et 4, renforce la cohérence interne du signal et sa lisibilité opérationnelle.

Sur le corpus stress-test, la réponse est également oui, mais avec une nuance importante : le signal NeoMundi réagit différemment selon la nature des prompts soumis. Les catégories ciblant l'incertitude factuelle vérifiable produisent des taux de FLAG sensiblement plus élevés que les catégories spéculatives ou de raisonnement général, qui restent à 0%. Cette sensibilité au type de prompt n'est pas une faiblesse du signal : elle reflète une cohérence entre le comportement de l'instrument et la nature des situations qui exposent structurellement un modèle à des dérives vérifiables. Elle implique cependant que la composition du corpus est un facteur déterminant des observations, et que toute comparaison entre campagnes doit tenir compte de cette dimension.

La question de l'écart entre Phase 1 et réplication ne reçoit en revanche pas de réponse définitive. L'analyse converge vers l'hypothèse que la nature spécifique des prompts de Phase 1, combinée à un possible effet de version du modèle, constitue l'explication la plus vraisemblable, la température s'avérant un facteur explicatif négligeable. Mais la co-variation de plusieurs facteurs entre les deux périodes ne permet pas d'établir une attribution causale rigoureuse. Cette incertitude résiduelle est documentée comme telle, plutôt que dissimulée derrière une interprétation trop assurée.

Ces résultats ouvrent plusieurs perspectives concrètes pour les travaux futurs de l'Observatoire. Une réplication à plus grande échelle, avec des effectifs de l'ordre de 300 à 400 prompts par configuration, permettrait d'atteindre une puissance statistique suffisante

pour confirmer ou infirmer les tendances descriptives observées. Une extension à des providers supplémentaires et à des corpus encore plus diversifiés renforcerait la portée des conclusions sur la stabilité inter-configurations du signal NeoMundi.

En l'état, cette note constitue une contribution exploratoire honnête et reproductible à la base de connaissances méthodologiques de l'Observatoire. Elle ne prétend pas établir une vérité définitive sur la reproductibilité du signal NeoMundi, mais fournit des éléments empiriques solides, documentés avec transparence, sur lesquels les travaux futurs pourront s'appuyer.

7. Annexes

7.1. Annexe A — Manifeste des campagnes

Table 12. Manifeste des campagnes d'exécution

Campagne	Provider	Temp.	n exec.	n retenu
Phase 1	PROFILE-5B3BB7	0,7	50	50
Essai 1 T=0,7	PROFILE-5B3BB7	0,7	50	50
Essai 1 T=0	PROFILE-5B3BB7	0	50	50
Essai 2 PROFILE-739F7C	PROFILE-739F7C	0	50	50
Essai 2 PROFILE-F5FF91	PROFILE-F5FF91	0	50	50
Essai 3 PROFILE-5B3BB7	PROFILE-5B3BB7	0	100	100
Essai 3 PROFILE-739F7C	PROFILE-739F7C	0	100	100
Essai 3 PROFILE-F5FF91	PROFILE-F5FF91	0	100	100
Essai 4	4 conf.	0/0,7	180	180
Total réplication			**680**	**680**

7.2. Annexe B — Exports CSV

Les fichiers suivants contiennent les données brutes collectées lors de chaque campagne, au format CSV avec encodage UTF-8. Chaque ligne correspond à une génération et contient l'ensemble des variables décrites en section 2.5.

Table 13. Fichiers de données brutes

Nom du fichier	Campagne	n lignes
essai1_PROFILE-5B3BB7_T07.csv	Essai 1 — PROFILE-5B3BB7 T=0,7	50
essai1_PROFILE-5B3BB7_T00.csv	Essai 1 — PROFILE-5B3BB7 T=0	50
essai2_PROFILE-739F7C.csv	Essai 2 — PROFILE-739F7C	50
essai2_PROFILE-F5FF91.csv	Essai 2 — PROFILE-F5FF91	50
essai3_PROFILE-5B3BB7.csv	Essai 3 — PROFILE-5B3BB7	100
essai3_PROFILE-739F7C.csv	Essai 3 — PROFILE-739F7C	100
essai3_PROFILE-F5FF91.csv	Essai 3 — PROFILE-F5FF91	100
essai4_tous_providers.csv	Essai 4 — 4 configurations	180
Total		**680**

7.3. Annexe C — Tests statistiques exploratoires

Afin de vérifier si les différences observées entre groupes dépassent la variabilité stochastique attendue, trois tests statistiques exploratoires ont été conduits : un test du χ^2 de Pearson pour les comparaisons à plus de deux groupes et un test exact de Fisher pour les comparaisons à deux groupes avec effectifs faibles. Ces tests sont présentés à titre indicatif et interprétés avec prudence, compte tenu des effectifs limités et de l'absence de correction formelle pour comparaisons multiples.

Table 14. Résultats des tests statistiques exploratoires

Comparaison	Test	Statistique	p-value	Conclusion
Effet provider (Essai 4, T=0)	χ^2	4,050	0,132	Non significatif
Effet température (Essai 4, PROFILE-5B3BB7)	Fisher	—	1,000	Non significatif
Effet catégorie (Essai 4)	χ^2	9,868	0,079	Non significatif

Aucun des trois tests ne produit de résultat significatif au seuil conventionnel $\alpha = 0{,}05$. Ces résultats reflètent principalement un manque de puissance statistique lié aux effectifs disponibles, et non une absence d'effet. L'effet température (p=1,000) confirme quantitativement l'absence d'impact déjà observée de façon descriptive. L'effet catégorie (p=0,079) est le plus proche du seuil de significativité, cohérent avec le gradient descriptif observé en section 4.2. L'effet provider présente un écart descriptif notable entre PROFILE-739F7C (17,8%) et PROFILE-F5FF91 (4,4%) qui mériterait une confirmation sur un corpus plus large. Les tables de contingence complètes sont disponibles en Annexe E.

7.4. Annexe D — Visualisation complémentaires

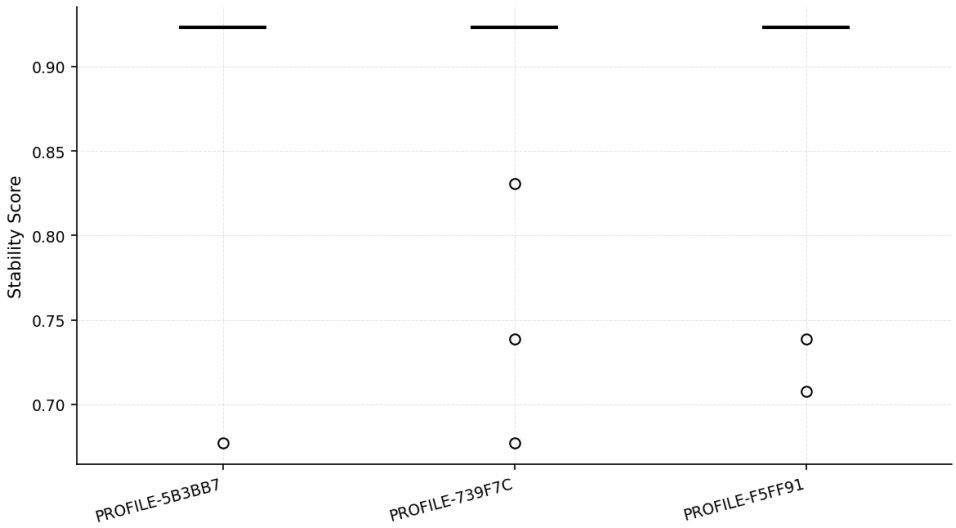

Figure 2. Distribution du Stability Score par provider — Essai 3 / Campagne A (n=100 par provider)

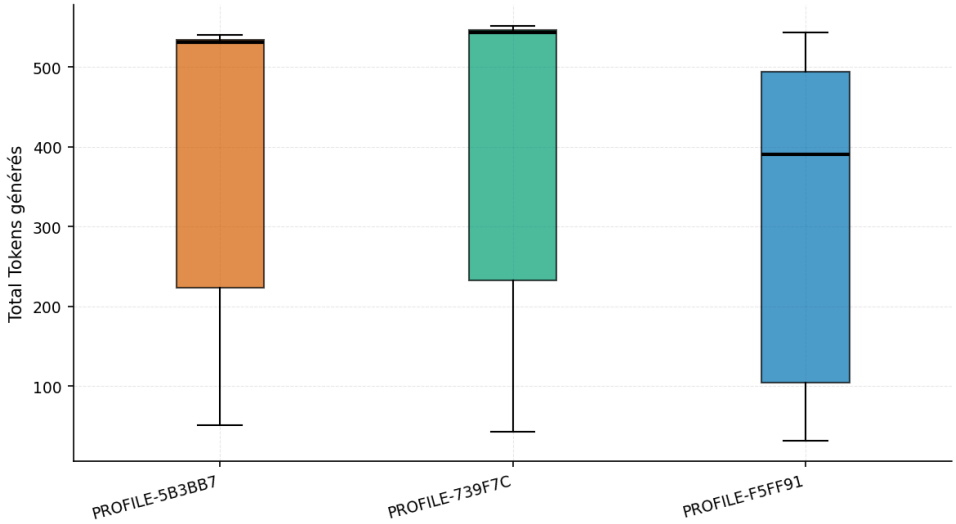

Figure 3. Distribution du nombre de tokens générés par provider — Essai 3 / Campagne A (n=100 par provider)

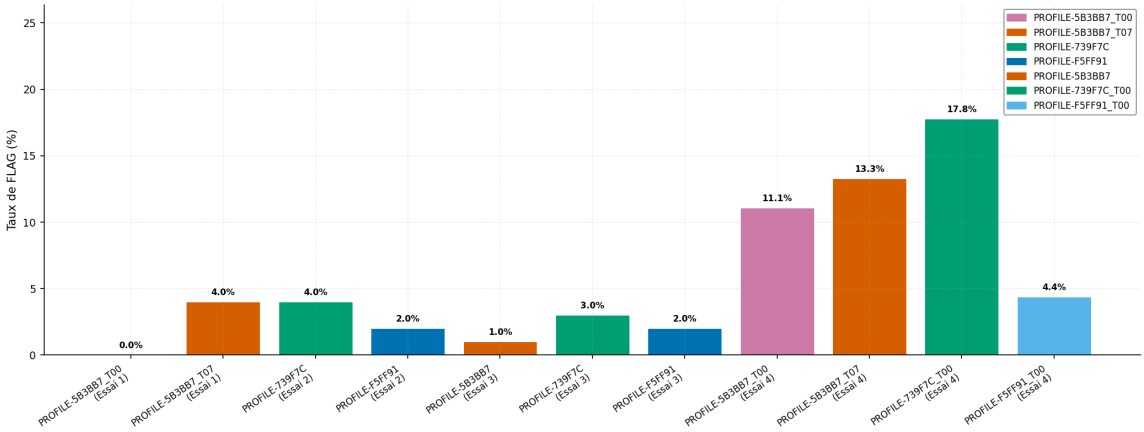

Figure 4. Taux de FLAG par configuration — vue globale, tous essais confondus

7.5. Annexe F — Exemples de prompts ayant produit une décision FLAG

À titre illustratif, quelques exemples de prompts ayant produit une décision FLAG sont présentés ci-dessous.

Table 15. Exemples de prompts ayant produit une décision FLAG — cinq catégories représentatives

Configurati	Catégorie	Prompt	Profil ΔG
PROFILE-5B3BB7_T07	Factuelle	Quel est le pays le plus peuplé du monde ?	DROP
PROFILE-5B3BB7_T07	Actualité	Quel est le cours actuel du Bitcoin ?	DROP
PROFILE-5B3BB7_T07	Entité fictive	Quelle est la capitale du pays appelé Norvégie du Sud ?	DROP
PROFILE-5B3BB7_T07	Fausse prémisse	Maintenant que l'eau bout à 50°C au niveau de la mer, comment cela change la cuisine ?	DROP
PROFILE-739F7C_T00	Chiffres précis	Quel est le PIB exact du Maroc en 2025, au dollar près ?	DROP

7.6. Annexe E — Tables de contingence des tests statistiques

Table 16. Effet provider — Essai 4, T=0 (test χ^2, statistique = 4,050, p = 0,132)

Provider	ALLOW	FLAG
PROFILE-5B3BB7_T00	40	5
PROFILE-739F7C_T00	37	8
PROFILE-F5FF91_T00	43	2
Total	**120**	**15**

Table 17. Effet température — Essai 4, PROFILE-5B3BB7 (test exact de Fisher, p = 1,000)

Configuration	ALLOW	FLAG
PROFILE-5B3BB7_T07	39	6
PROFILE-5B3BB7_T00	40	5
Total	**79**	**11**

Table 18. Effet catégorie — Essai 4, catégories avec FLAG > 0 (test χ^2, statistique = 9,868, p = 0,079)

Catégorie	ALLOW	FLAG
Actualité	16	8
Fausse prémisse	20	4
Entité fictive	21	3
Chiffres précis	21	3
Citation	22	2
Multi-étapes	23	1
Total	**123**	**21**