NeoMundi Observatory · Independent Runtime Metrology

Rapport d’observation runtime - Kimi K3

Analyse de 1 200 exécutions dans 12 familles de situations

Fournisseur: moonshot

Modèle demandé: kimi-k3

Modèle observé: kimi-k3

Version du rapport: 1.0

Généré le: 2026-07-22 20:25 UTC

Couverture: 100,00 %

> **Signal mesuré, pas verdict**

Table des matières

> Ce rapport décrit les comportements observés dans les conditions exactes du protocole exécuté. Il ne constitue ni un classement général du modèle, ni une certification, ni une décision organisationnelle.

1. Synthèse exécutive

NeoMundi a observé **kimi-k3** sur **1 200 exécutions valides**, issues de 120 situations répétées 10 fois dans 12 familles.

L'objectif n'est pas de classer le modèle. Le rapport montre ce qu'une couche de métrologie runtime rend observable : distributions, différences entre familles, signaux nécessitant une investigation et preuve rattachée à chaque génération.

1 200 exécutions valides	**100,00 %** couverture finale	**120** situations distinctes
10 répétitions par situation	**12** familles	**0** observations FLAG

Enseignements principaux

- **Plusieurs régimes runtime apparaissent.** Les familles diffèrent en latence, volume, densité informationnelle et signaux de gouvernance.
- **Une moyenne globale ne suffit pas.** Elle peut masquer des profils distincts selon la nature de la tâche.
- **Les dimensions restent séparées.** Stabilité, factualité, cohérence, vitesse et densité ne sont pas fusionnées en un score unique.
- **Chaque signal reste reconstructible.** Il peut être relié au prompt, à la réponse, aux métriques, à la décision et aux identifiants de trace.

Un signal mesuré n'est pas un verdict. C'est une capacité opérationnelle à voir, comparer, investiguer et gouverner.

2. Périmètre et frontière d’observation

Ce que le rapport observe

- Comportement observable des générations.
- Stabilité, cohérence et variation sémantique.
- Signaux factuels et sémantiques.
- Latence, volume de tokens et densité informationnelle.
- Décisions runtime et éléments de traçabilité.

Ce qu’il ne détermine pas directement

- Vérité absolue ou qualité universelle du modèle.
- Légitimité ou autorisation organisationnelle.
- Conformité juridique complète.
- Certification médicale, juridique ou professionnelle.
- Supériorité générale par rapport à d’autres modèles.

> NeoMundi observe le comportement du modèle. Les systèmes organisationnels déterminent si une sortie est autorisée à devenir une action.

3. Protocole et conception expérimentale

Structure
120 × 10 = 1 200

Familles
12

Fournisseur
moonshot

Modèle observé
kimi-k3

Pourquoi répéter chaque situation ?

- Distinguer un événement isolé d’un comportement récurrent.
- Observer la variabilité et les régimes de génération.
- Mesurer des distributions plutôt qu’une réponse unique.
- Permettre des comparaisons longitudinales ultérieures.

Familles du protocole

- Ambiguïté / robustesse
- Opérations métier
- Cybersécurité
- Factualité
- Gouvernance
- Respect des consignes
- Juridique
- Médical
- Questions ouvertes
- Raisonnement
- Efficience runtime
- Temporalité / longitudinal

4. Qualité du corpus et chaîne de traçabilité

Le run initial comportait une exécution techniquement incomplète dans la phase de streaming. L'observation concernée a été identifiée, relancée de manière ciblée, réaffectée à la répétition manquante, puis le corpus complet a été nettoyé et contrôlé à nouveau. Le corpus final contient **1 200/1 200 lignes valides**, **0 erreur** et aucun doublon prompt/répétition.

1 200 lignes finales	**1 200** lignes valides	**0** lignes invalides
100,00 % couverture	**120** prompts observés	**0** doublon prompt/rép.

Artefacts de qualité

- Corpus nettoyé complet.
- Table analytique allégée.
- Rapport de couverture et de qualité.
- Profil des colonnes.
- Journal des erreurs et manifeste de relance.

5. Cadre de mesure

Mesures comportementales

Stabilité, cohérence, risque sémantique, instabilité sémantique et signal factuel.

Performance runtime

Latence, temps de traitement, volume de tokens, scores d’efficience et signaux de coût.

Structure informationnelle

Densité informationnelle et métriques associées lorsqu’elles sont disponibles.

Preuve de gouvernance

Décision ALLOW/FLAG, raisons, request_id, trace_id et versions de mesure.

> Les métriques restent séparées afin qu’un score composite ne masque pas des régimes différents.

6. Résultats globaux

Les statistiques ci-dessous décrivent le corpus valide. Elles doivent être interprétées comme des distributions, non comme un score final du modèle.

0,923
stabilité moyenne

1,000
cohérence moyenne

14,12 s
latence médiane

26,02 s
latence p95

564,2
tokens moyens

0,647
densité informationnelle

1 200
ALLOW

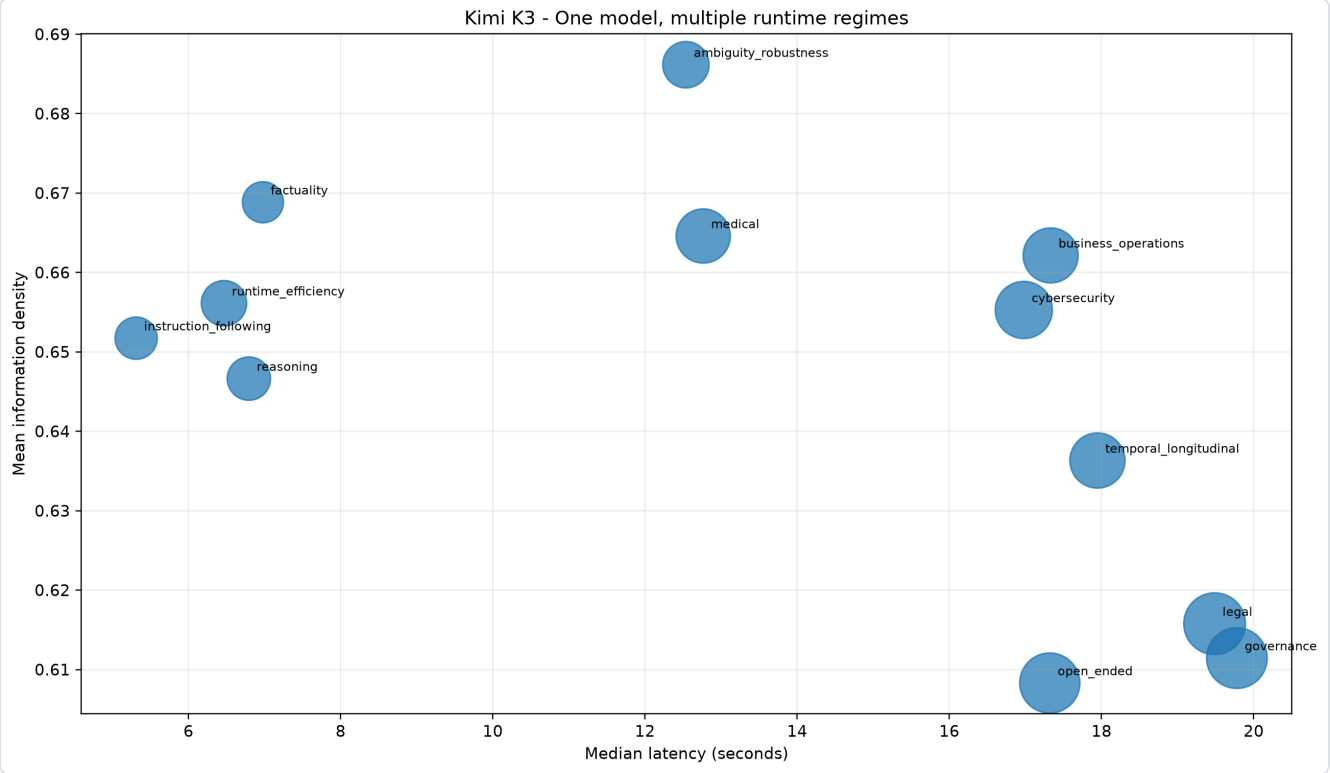
0
FLAG

0,00 %
taux FLAG

Métrique	N	Moyenne	Médiane	P05	P95	Min	Max
Latency (ms)	1 200	14 034,138	14 118,000	3 699,600	26 024,000	1 700,000	39 180,000
Processing time (ms)	1 200	20 394,112	20 020,865	20 018,220	20 027,311	20 016,590	87 594,520
Token count	1 200	564,182	580,000	184,950	952,050	121,000	1 331,000
Stability score	1 200	0,923	0,923	0,923	0,923	0,892	0,923
Coherence score	1 200	1,000	1,000	1,000	1,000	1,000	1,000
Information density	1 200	0,647	0,648	0,557	0,750	0,408	0,787
Factual signal	1 200	0,000	0,000	0,000	0,000	0,000	0,100
Semantic instability	1 200	0,018	0,000	0,000	0,000	0,000	0,900
Runtime R score	1 200	0,600	0,600	0,600	0,600	0,600	0,600

7. Régimes runtime dans les 12 familles

La carte principale relie la latence médiane, la densité informationnelle, le volume moyen de tokens et la présence éventuelle de signaux FLAG. Elle rend visibles plusieurs profils d’exécution derrière un même modèle.



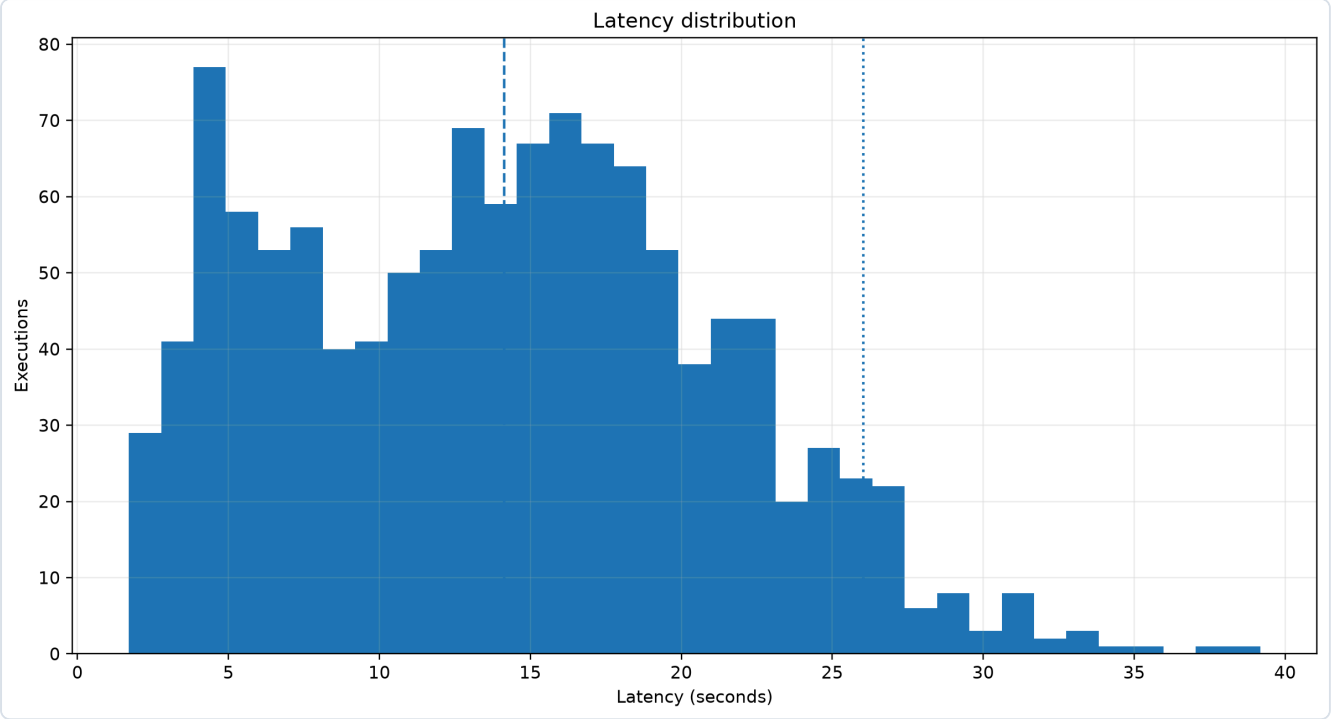
Carte des régimes runtime par famille

Famille	Obs.	Prompts	Stabilité	Latence méd. ms	Latence p95 ms	Tokens moy.	Densité info.	Taux FLAG
Ambiguïté / robustesse	100	10	0,923	12 538	18 555	442,2	0,686	0,0 %
Cybersécurité	100	10	0,923	16 979	26 590	670,9	0,655	0,0 %
Efficience runtime	100	10	0,923	6 467	29 563	423,0	0,656	0,0 %
Factualité	100	10	0,923	6 979	12 351	347,1	0,669	0,0 %
Gouvernance	100	10	0,923	19 779	26 705	756,5	0,611	0,0 %
Juridique	100	10	0,922	19 486	27 807	782,7	0,616	0,0 %
Médical	100	10	0,923	12 766	17 830	604,0	0,665	0,0 %
Opérations métier	100	10	0,923	17 332	26 644	621,8	0,662	0,0 %

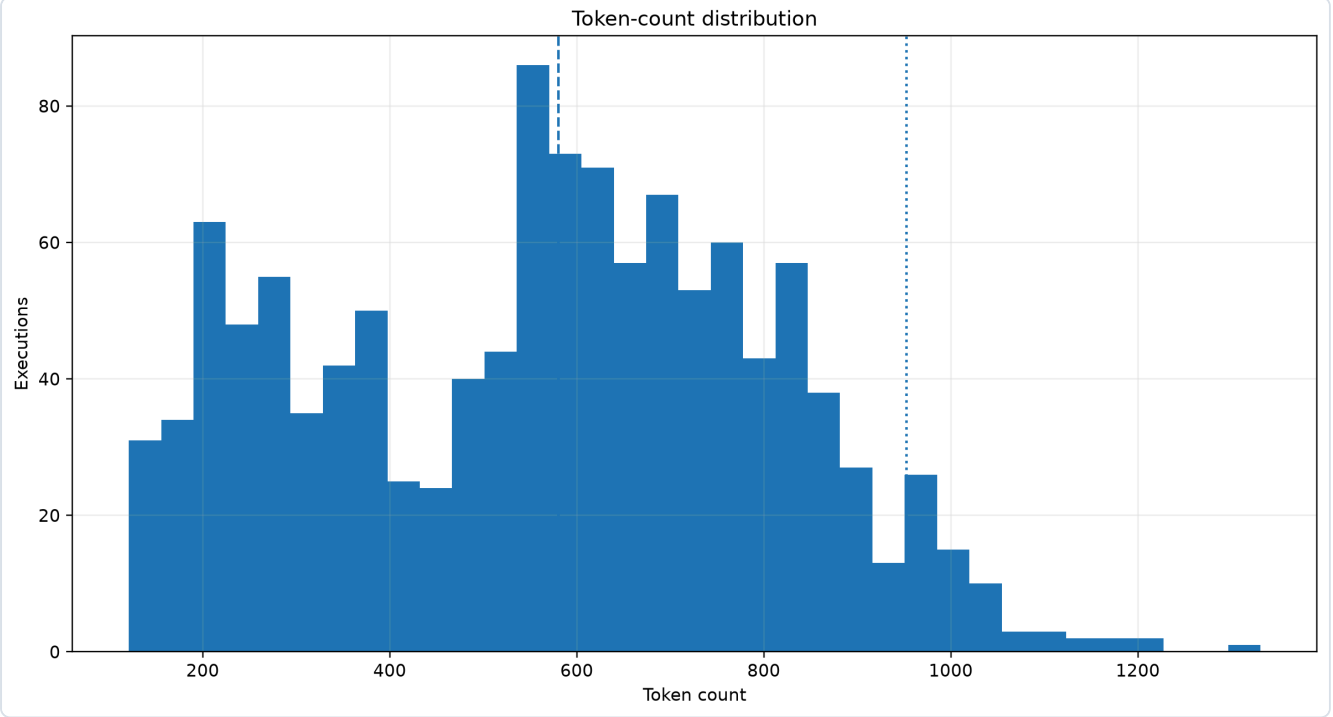
Famille	Obs.	Prompts	Stabilité	Latence méd. ms	Latence p95 ms	Tokens moy.	Densité info.	Taux FLAG
Questions ouvertes	100	10	0,923	17 320	23 109	743,2	0,608	0,0 %
Raisonnement	100	10	0,923	6 794	13 898	386,6	0,647	0,0 %
Respect des consignes	100	10	0,923	5 313	15 748	366,9	0,652	0,0 %
Temporalité / longitudinal	100	10	0,923	17 946	27 556	625,5	0,636	0,0 %

8. Analyse des distributions

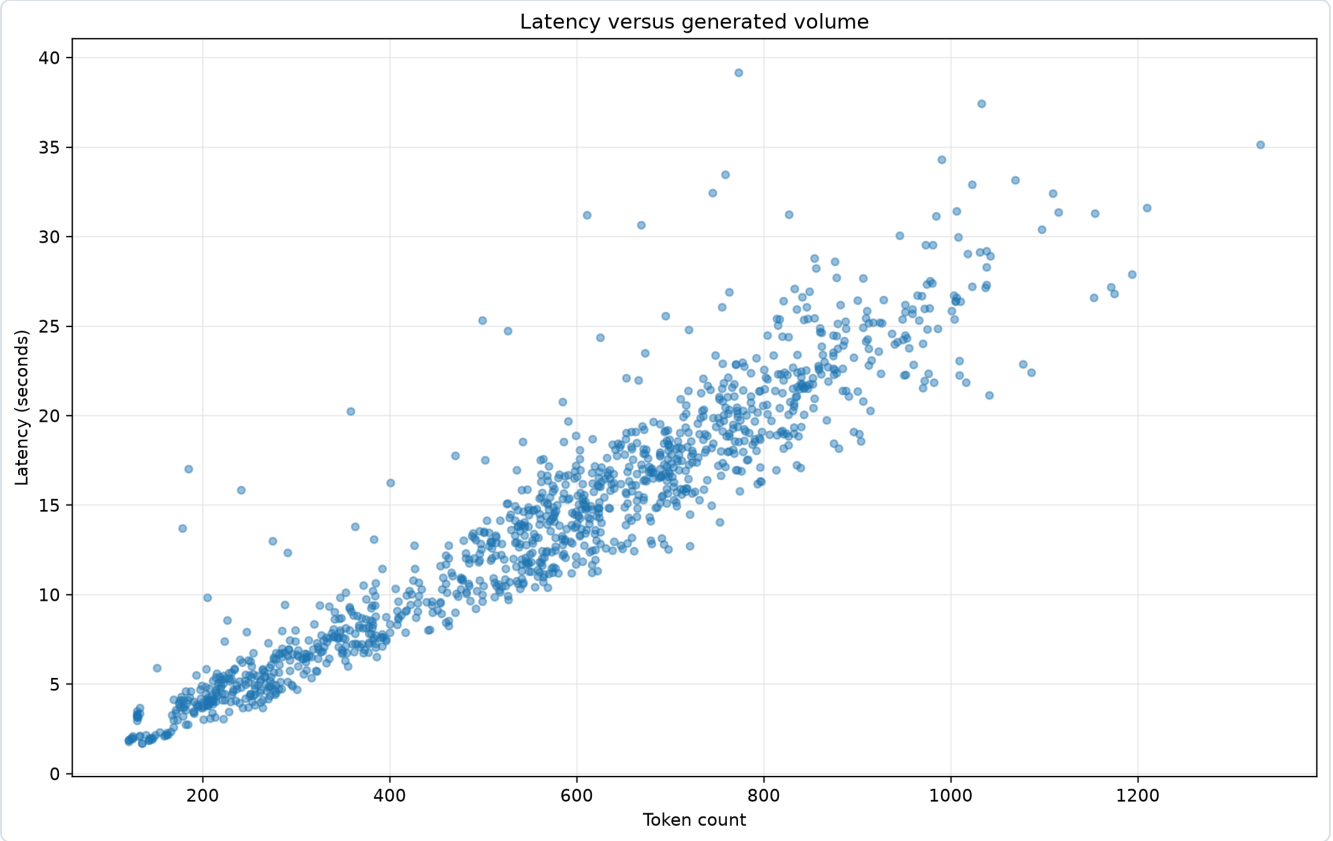
Les distributions montrent l'étendue des comportements observés et évitent de réduire l'analyse à une moyenne. Les lignes de référence représentent la médiane et, selon les graphiques, le 95e percentile.



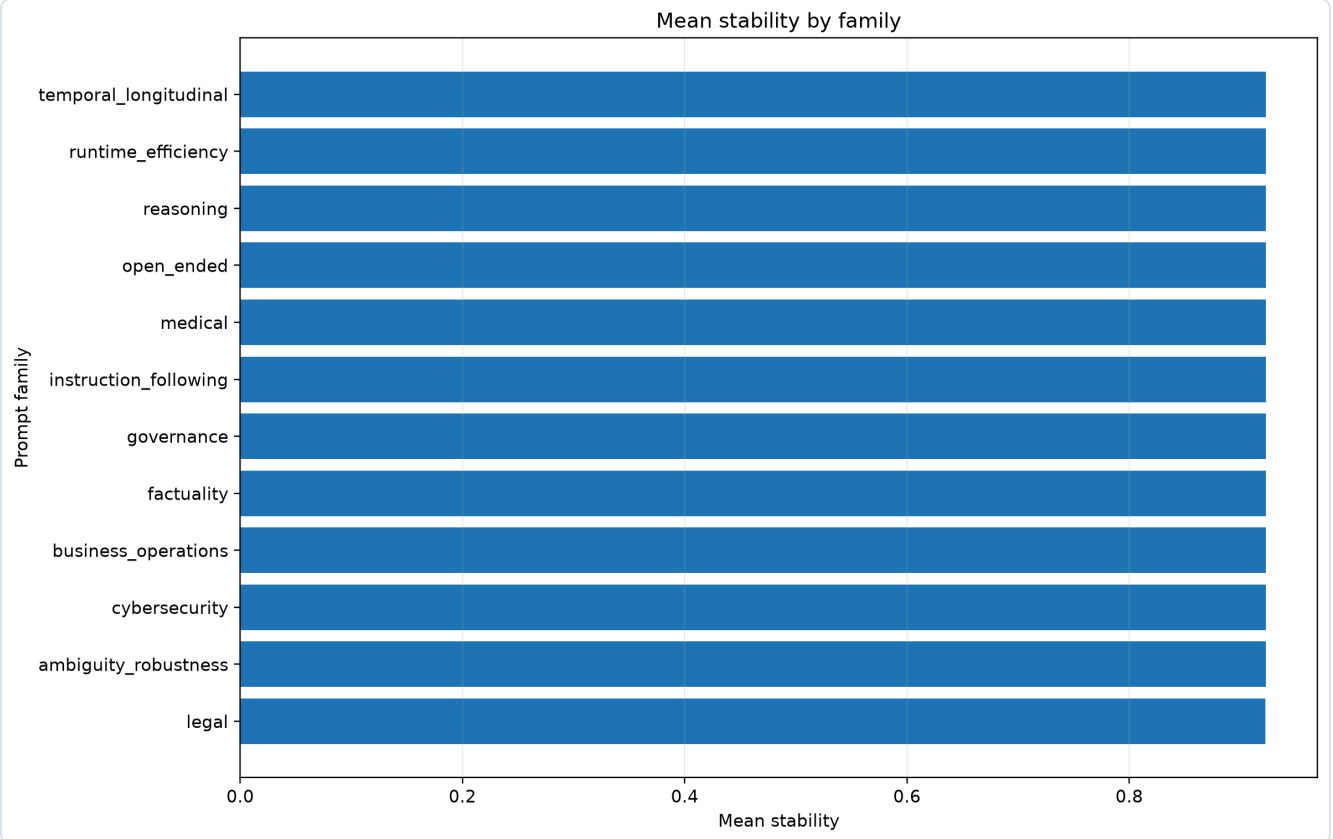
Distribution de la latence



Distribution du volume de tokens



Latence et volume généré



Stabilité moyenne par famille

9. Cas nécessitant une investigation

Les cas ci-dessous ne constituent pas une condamnation du modèle. Ils localisent les prompts ou exécutions qui méritent une lecture plus attentive.

> Aucun prompt FLAG dans le corpus.

> Cas remarquable ne signifie pas automatiquement anomalie du modèle.

10. Exemple de preuve runtime reconstructible

L’exemple ci-dessous montre comment revenir d’un signal agrégé à une observation précise.

run_id	moonshot_kimi-k3_kimi_k3_cross_domain_wave_1_20260721T204417Z
wave_id	kimi_k3_cross_domain_wave_1
provider	moonshot
requested_model	kimi-k3
observed_model	kimi-k3
prompt_id	NCRP-ROB-002
prompt_family	ambiguity_robustness
domain	conflicting_instructions
difficulty	hard
repetition_index	1
request_id	3a4a2aa0-e9f8-44da-be5b-a056052c4559
api_timestamp	2026-07-22 06:41:45.548148+00:00
trace_id	trace-3a4a2aa0
token_count	481
latency_ms	12326.0
processing_time_ms	20019.66
decision	ALLOW
classification_confidence	1.0
stability_score	0.923077

Prompt

```
-
```

Réponse générée

```
Impossible de satisfaire simultanément « 500 mots » et « ne pas dépasser 50 mots ». Je privilégie la limite stricte : voici une réponse concise. Dites-moi si vous préférez la version longue de 500 mots, et je la fournirai immédiatement.
```

coherence_score	1.0
semantic_risk	0.0
factual_hallucination_score	0.0
semantic_instability_score	0.9
information_density	0.7523
runtime_r_score	0.599875
delta_g	0.0
measurement_version	3.0.0
normalizer_version	1.0.0

Raisons de gouvernance

```
stability_score=0.923, regime=STABLE -- all metrics within bounds || Semantic instability detected (score=0.90) >= warn threshold (0.40) -- WARN only (answer appears correct) || R=0.600 (latency=0.000, cost=1.000, efficiency=0.999)
```

Prompt → Generation → Runtime measurements → Governance evidence → Trace → Investigation or action

11. Ce que NeoMundi ajoute

NeoMundi n’ajoute pas un dashboard supplémentaire. Il concentre dans une même couche plusieurs fonctions de monitoring habituellement fragmentées, tout en ajoutant la mesure runtime qui manque entre le comportement du modèle et la décision opérationnelle.

- **Model monitoring** : suivre les comportements et leurs régimes.
- **Semantic monitoring** : repérer variations et instabilités.
- **Factual-risk monitoring** : localiser des signaux factuels.
- **Performance et FinOps** : relier latence, tokens, densité et effort.
- **Governance evidence** : rattacher le signal à une décision et à ses raisons.
- **Traçabilité** : reconstruire l’observation avec ses identifiants et versions.

> NeoMundi ne remplace pas les systèmes métier, les moteurs de décision ou les couches d’enforcement. Il leur fournit des signaux et des preuves.

12. Usages opérationnels possibles

- Comparer plusieurs vagues temporelles.
- Détecter un changement de régime.
- Déclencher une investigation ciblée.
- Régénérer ou rerouter une réponse.
- Escalader vers un opérateur humain.
- Documenter un audit ou une revue de conformité.
- Alimenter un moteur de décision ou d’enforcement.
- Enrichir des outils de monitoring et de FinOps existants.

Le présent rapport observe ces possibilités ; il ne déclenche pas automatiquement toutes ces actions.

13. Limites

- Une seule vague ne démontre pas la stabilité de long terme.
- Les résultats dépendent du protocole, de l’API, du modèle observé et de la période.
- Aucune causalité n’est déduite d’un signal observé.
- La stabilité ne garantit pas la vérité.
- La cohérence ne garantit pas l’exactitude.
- Un signal factuel ne remplace pas un jugement expert.
- Les familles médicale et juridique ne constituent pas une certification professionnelle.
- Le rapport ne produit pas de leaderboard.
- Les seuils et normalisateurs doivent être lus avec leurs versions et limites.

14. Conclusion

Ce protocole montre qu’un modèle ne peut pas être observé uniquement à travers sa disponibilité ou sa latence. Une lecture multidimensionnelle fait apparaître plusieurs régimes, distingue les dimensions de performance et de comportement, et permet de revenir du signal agrégé à la preuve.

Un signal mesuré n’est pas un verdict. C’est une capacité opérationnelle à voir, comparer, investiguer et gouverner.

15. Annexes

A. Définitions et noms de métriques

Les métriques sont identifiées par leur nom natif dans les tables et fichiers source. Toute interprétation doit tenir compte des versions de mesure et de normalisation.

B. Artefacts de référence

- moonshot_kimi-k3_kimi_k3_cross_domain_wave_1_20260721T204417Z_patched_analysis.csv
- moonshot_kimi-k3_kimi_k3_cross_domain_wave_1_20260721T204417Z_patched_clean_full.csv
- moonshot_kimi-k3_kimi_k3_cross_domain_wave_1_20260721T204417Z_patched_quality_report.json

C. Colonnes analytiques

analysis_role, api_timestamp, checks_emitted, classification_confidence, coherence_score, cost_score, decision, delta_g, difficulty, domain, e_token, energy_score, factual_hallucination_score, g_final, g_score, governance_reasons, information_density, is_valid, language, latency_ms, latency_score, llm_response, measurement_version, normalizer_version, observed_model, planned_repetitions, processing_time_ms, prompt_family, prompt_id, prompt_version, provider, quality_issue, repetition_index, request_id, requested_model, run_id, runtime_r_score, semantic_instability_score, semantic_risk, sensitivity, stability_score, temperature_mode, temperature_requested, token_count, token_count_source, trace_id, v_score, wave_id

D. Reproductibilité

- Report version: 1.0
- Generated: 2026-07-22 20:25 UTC
- Analysis SHA-256: 87e69c22073c3a686b6a71d6dd0c8db91636e5cdae9be785a725d2c1486b9c84
- Clean-full SHA-256: b462ca3537d123bb12df0730145bdc393819e9e7bc1da0bb6e0a2f5ed8dfe12c
- Quality-report SHA-256: 16ce2101812b411b32cd91bbb685d8338f689beec7442e427e404ad7841ed275