

NeoMundi Observatory · Independent Runtime Metrology

Kimi K3 Runtime Observation Report

Analysis of 1,200 executions across 12 situation families

Provider: moonshot

Requested model: kimi-k3

Observed model: kimi-k3

Report version: 1.0

Generated: 2026-07-22 20:25 UTC

Coverage: 100.00 %

Measured signal, not a verdict

Table of Contents

1. 1. Executive Summary	9. 9. Cases Requiring Investigation
2. 2. Scope and Observation Boundary	10. 10. Reconstructible Runtime Evidence
3. 3. Protocol and Experimental Design	11. 11. What NeoMundi Adds
4. 4. Dataset Quality and Chain of Custody	12. 12. Potential Operational Uses
5. 5. Measurement Framework	13. 13. Limitations
6. 6. Global Results	14. 14. Conclusion
7. 7. Runtime Regimes Across the 12 Families	15. 15. Appendices
8. 8. Distribution Analysis	

This report describes behavior observed under the exact conditions of the executed protocol. It is not a general model ranking, certification, or organizational decision.

1. Executive Summary

NeoMundi observed **kimi-k3** over **1,200 valid executions**, generated from 120 situations repeated 10 times across 12 families.

The objective is not to rank the model. The report shows what a runtime metrology layer makes observable: distributions, family-level differences, signals requiring investigation, and evidence linked to each generation.

1,200

valid executions

100.00 %

final coverage

120

distinct situations

10

repetitions per situation

12

families

0

FLAG observations

Key findings

- **Multiple runtime regimes are visible.** Families differ in latency, volume, information density, and governance signals.
- **A global average is not sufficient.** It can hide distinct profiles depending on the nature of the task.
- **Dimensions remain separate.** Stability, factuality, coherence, speed, and density are not collapsed into a single score.
- **Every signal remains reconstructible.** It can be linked to the prompt, response, metrics, decision, and trace identifiers.

A measured signal is not a verdict. It is an operational capacity to observe, compare, investigate, and govern.

2. Scope and Observation Boundary

What the report observes

- Observable behavior of model generations.
- Stability, coherence, and semantic variation.
- Factual and semantic signals.
- Latency, token volume, and information density.
- Runtime decisions and traceability evidence.

What it does not directly determine

- Absolute truth or universal model quality.
- Organizational legitimacy or authorization.
- Complete legal compliance.
- Medical, legal, or professional certification.
- General superiority over other models.

NeoMundi observes model behavior. Organizational systems determine whether an output is authorized to become an action.

3. Protocol and Experimental Design

Structure

$120 \times 10 = 1,200$

Families

12

Provider

moonshot

Observed model

kimi-k3

Why repeat each situation?

- Distinguish an isolated event from recurrent behavior.
- Observe variability and generation regimes.
- Measure distributions rather than a single response.
- Enable future longitudinal comparisons.

Protocol families

- Ambiguity / robustness
- Business operations
- Cybersecurity
- Factuality
- Governance
- Instruction following
- Legal
- Medical
- Open-ended
- Reasoning
- Runtime efficiency
- Temporal / longitudinal

4. Dataset Quality and Chain of Custody

The initial run contained one technically incomplete execution during the streaming phase. The affected observation was identified, rerun in isolation, assigned to the missing repetition, and the complete dataset was cleaned and verified again. The final corpus contains **1,200/1,200 valid rows**, **0 errors**, and no duplicate prompt/repetition pairs.

1,200

final rows

1,200

valid rows

0

invalid rows

100.00 %

coverage

120

observed prompts

0

prompt/rep. duplicates

Quality artifacts

- Complete cleaned corpus.
- Lightweight analytical table.
- Coverage and quality report.
- Column profile.
- Error log and rerun manifest.

5. Measurement Framework

Behavioral measurements

Stability, coherence, semantic risk, semantic instability, and factual signal.

Runtime performance

Latency, processing time, token volume, efficiency scores, and cost-related signals.

Information structure

Information density and related metrics when available.

Governance evidence

ALLOW/FLAG decision, reasons, request_id, trace_id, and measurement versions.

Metrics remain separate so that a composite score does not hide different regimes.

6. Global Results

The statistics below describe the valid corpus. They should be interpreted as distributions, not as a final model score.

0.923

mean stability

1.000

mean coherence

14.12 s

median latency

26.02 s

p95 latency

564.2

mean tokens

0.647

information density

1,200

ALLOW

0

FLAG

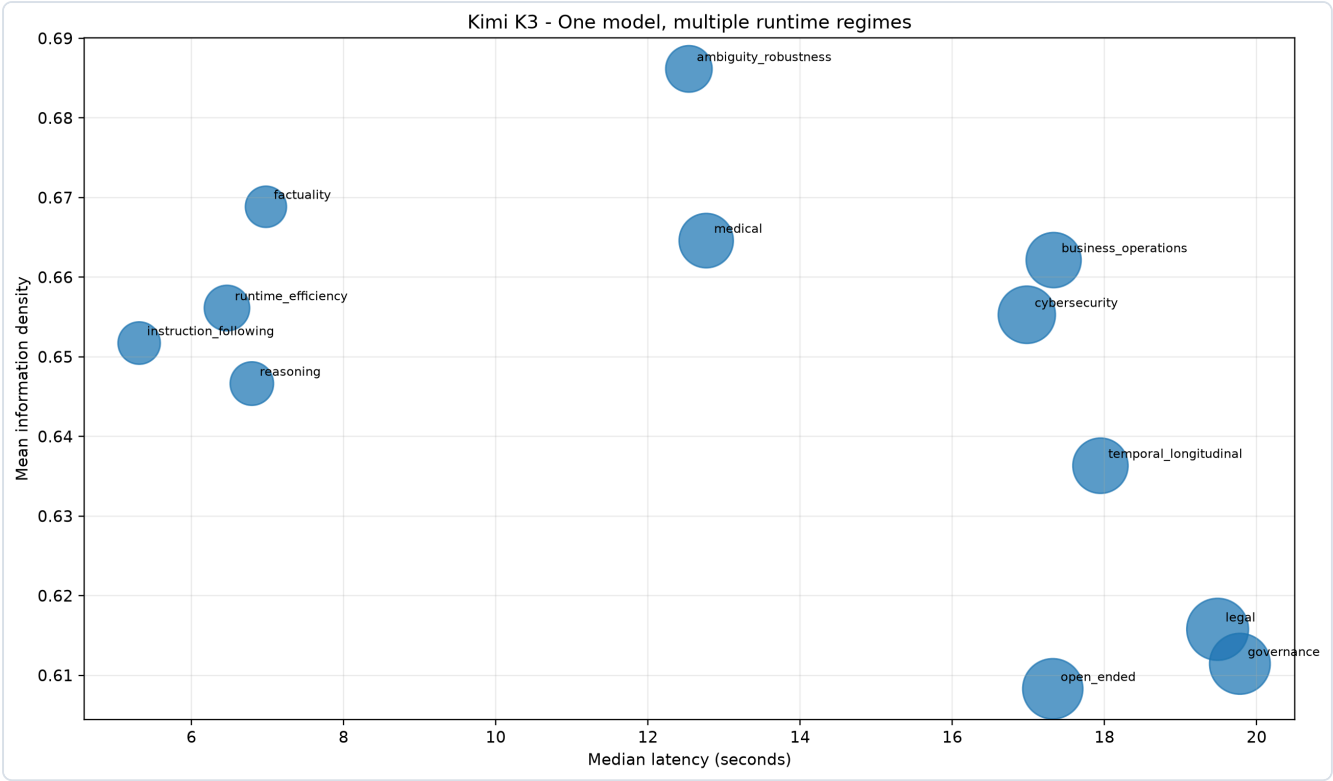
0.00 %

FLAG rate

Metric	N	Mean	Median	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. Runtime Regimes Across the 12 Families

The main map connects median latency, information density, mean token volume, and the possible presence of FLAG signals. It reveals several execution profiles behind a single model.



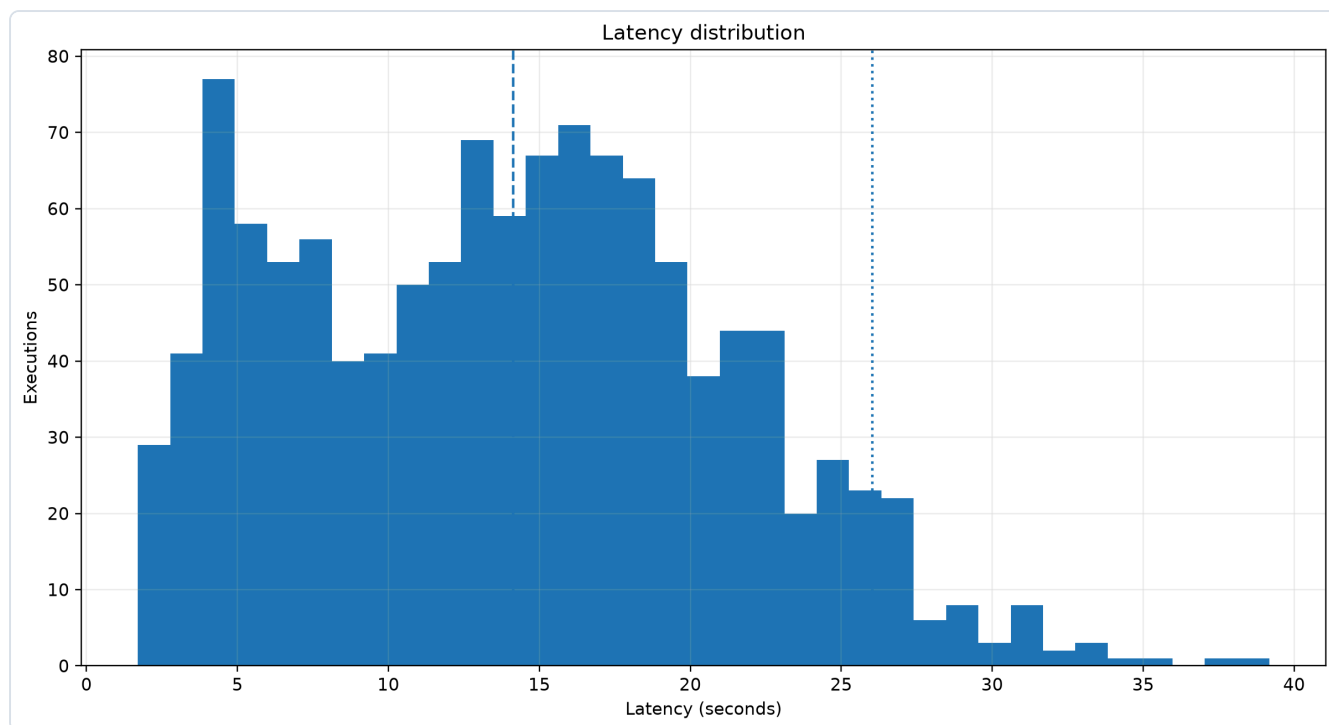
Runtime-regime map by family

Family	Obs.	Prompts	Stability	Median latency ms	p95 latency ms	Mean tokens	Info. density	FLAG rate
Ambiguity / robustness	100	10	0.923	12,538	18,555	442.2	0.686	0.0 %
Business operations	100	10	0.923	17,332	26,644	621.8	0.662	0.0 %
Cybersecurity	100	10	0.923	16,979	26,590	670.9	0.655	0.0 %
Factuality	100	10	0.923	6,979	12,351	347.1	0.669	0.0 %
Governance	100	10	0.923	19,779	26,705	756.5	0.611	0.0 %
Instruction following	100	10	0.923	5,313	15,748	366.9	0.652	0.0 %
Legal	100	10	0.922	19,486	27,807	782.7	0.616	0.0 %

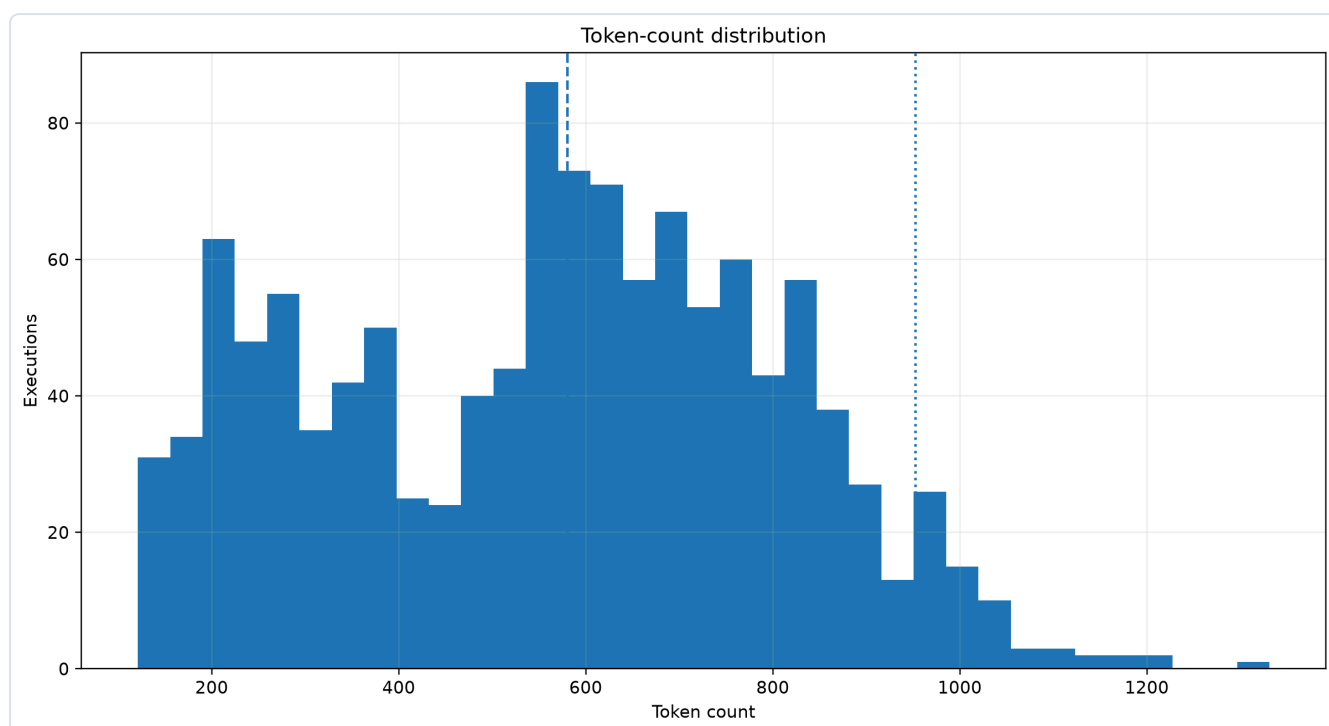
Family	Obs.	Prompts	Stability	Median latency ms	p95 latency ms	Mean tokens	Info. density	FLAG rate
Medical	100	10	0.923	12,766	17,830	604.0	0.665	0.0 %
Open-ended	100	10	0.923	17,320	23,109	743.2	0.608	0.0 %
Reasoning	100	10	0.923	6,794	13,898	386.6	0.647	0.0 %
Runtime efficiency	100	10	0.923	6,467	29,563	423.0	0.656	0.0 %
Temporal / longitudinal	100	10	0.923	17,946	27,556	625.5	0.636	0.0 %

8. Distribution Analysis

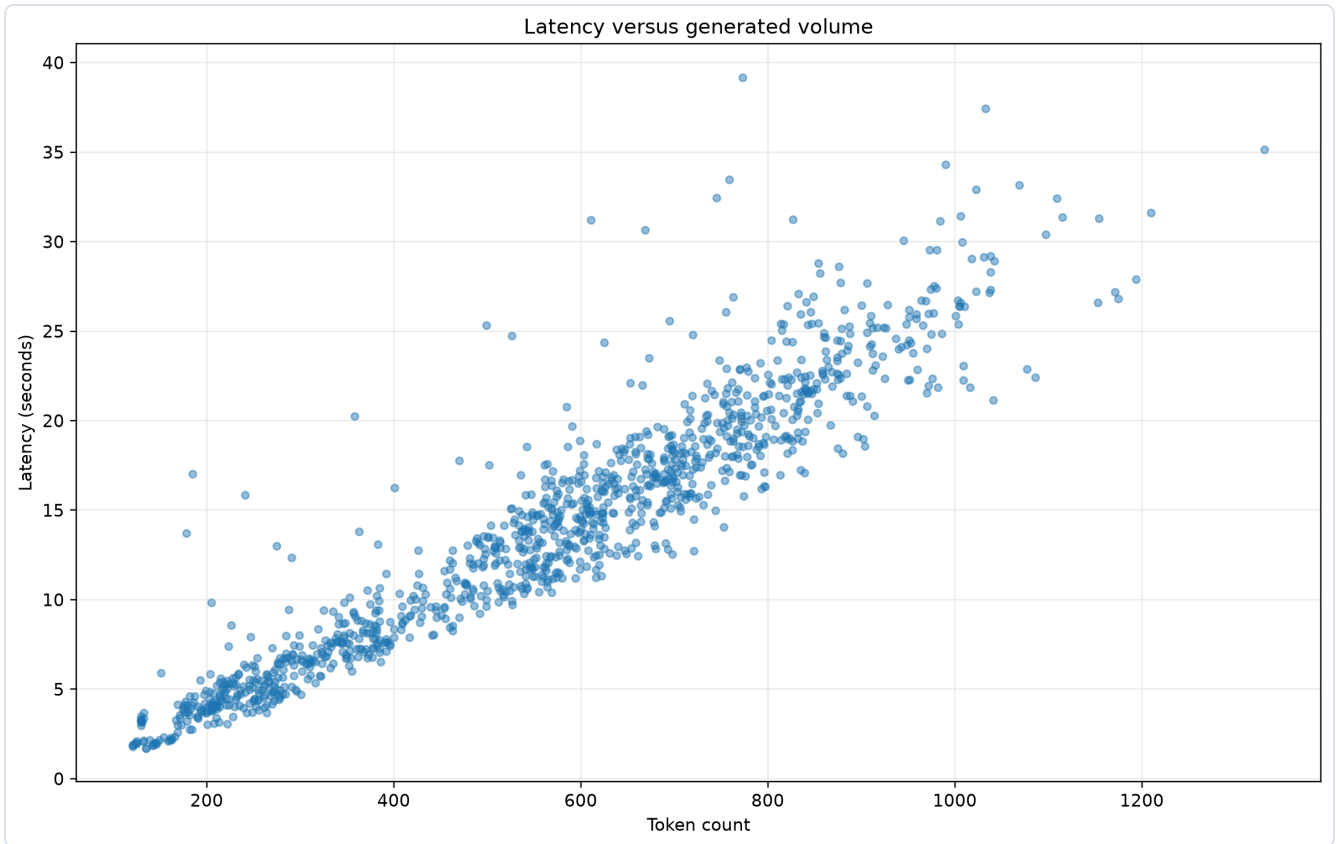
The distributions show the range of observed behaviors and avoid reducing the analysis to a mean. Reference lines represent the median and, where applicable, the 95th percentile.



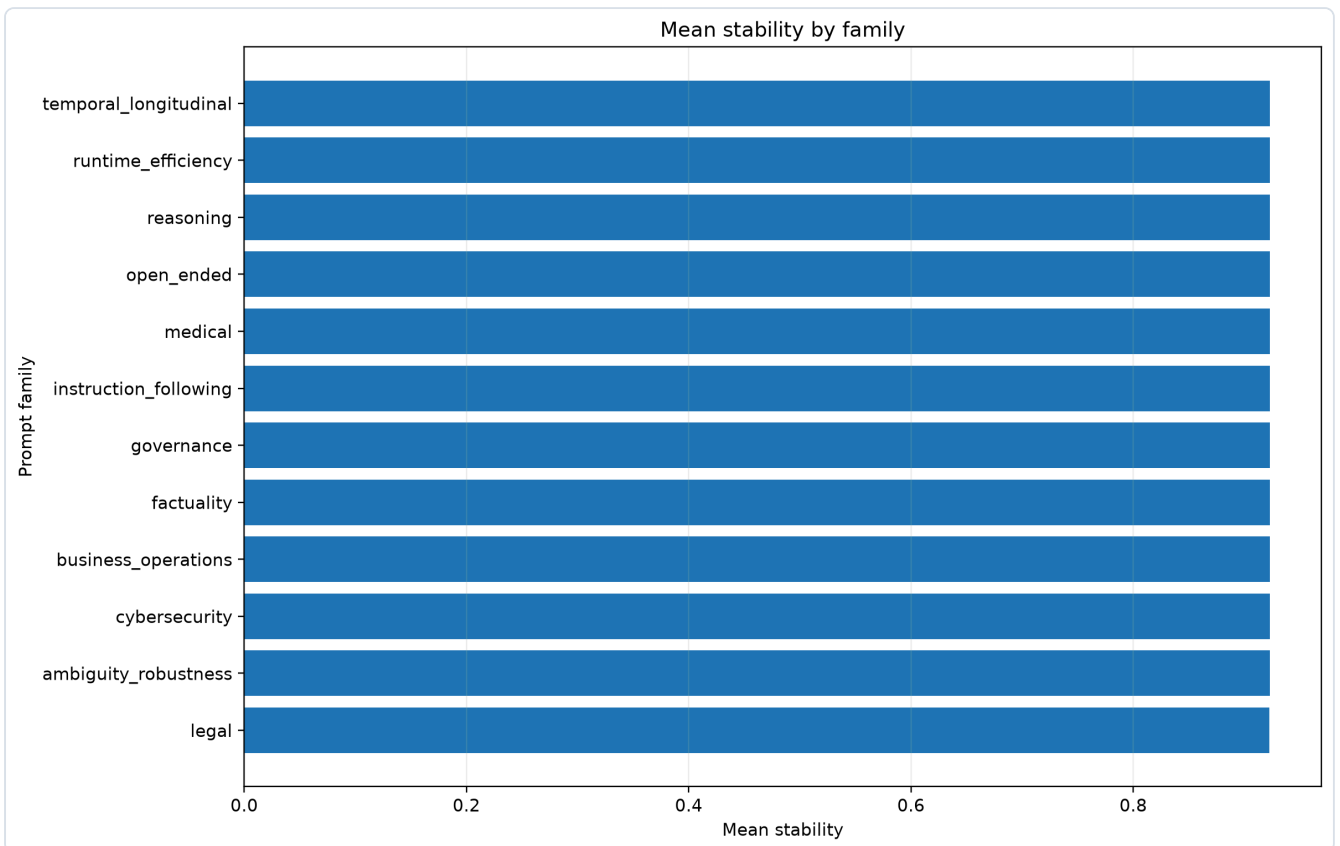
Latency distribution



Token-count distribution



Latency and generated volume



Mean stability by family

9. Cases Requiring Investigation

The cases below are not a condemnation of the model. They locate prompts or executions that deserve closer examination.

No FLAG prompt in the corpus.

A notable case does not automatically mean a model anomaly.

10. Reconstructible Runtime Evidence

The example below shows how an aggregate signal can be traced back to a precise observation.

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

-

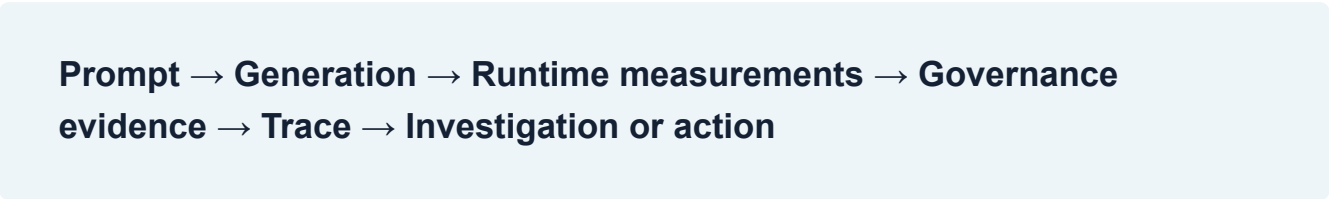
Generated response

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_score	0.599875
delta_g	0.0
measurement_version	3.0.0
normalizer_version	1.0.0

Governance reasons

```
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)
```



11. What NeoMundi Adds

NeoMundi does not add another dashboard. It brings together several monitoring functions that are usually fragmented, while adding the runtime measurement layer missing between model behavior and operational decision-making.

- **Model monitoring:** tracking behavior and runtime regimes.
- **Semantic monitoring:** locating variation and instability.
- **Factual-risk monitoring:** identifying factual signals.
- **Performance and FinOps:** connecting latency, tokens, density, and effort.
- **Governance evidence:** linking a signal to a decision and its reasons.
- **Traceability:** reconstructing an observation with identifiers and versions.

NeoMundi does not replace business systems, decision engines, or enforcement layers. It provides them with signals and evidence.

12. Potential Operational Uses

- Compare multiple temporal waves.
- Detect a regime change.
- Trigger a targeted investigation.
- Regenerate or reroute a response.
- Escalate to a human operator.
- Document an audit or compliance review.
- Feed a decision or enforcement engine.
- Enrich existing monitoring and FinOps tools.

This report observes these possibilities; it does not automatically trigger all of these actions.

13. Limitations

- A single wave does not demonstrate long-term stability.
- Results depend on the protocol, API, observed model, and time period.
- No causality is inferred from an observed signal.
- Stability does not guarantee truth.
- Coherence does not guarantee accuracy.
- A factual signal does not replace expert judgment.
- Medical and legal families do not constitute professional certification.
- The report does not create a leaderboard.
- Thresholds and normalizers must be interpreted with their versions and limitations.

14. Conclusion

This protocol shows that a model cannot be observed solely through availability or latency. A multidimensional reading reveals multiple regimes, separates performance and behavioral dimensions, and allows the aggregate signal to be traced back to evidence.

A measured signal is not a verdict. It is an operational capacity to observe, compare, investigate, and govern.

15. Appendices

A. Metric definitions and names

Metrics are identified by their native names in source tables and files. Interpretation must account for measurement and normalization versions.

B. Reference artifacts

- 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. Analytical columns

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. Reproducibility

- 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

NeoMundi Observatory - Runtime Metrology - Measured signal, not a verdict