Exploratory Audit of NeoMundi Runtime Signals

Analysis of G, ΔG, FLAG, Regime and Benchmark Accuracy

Pape Malick DIOP

Data Scientist, ML Researcher

June 10, 2026

External exploratory report based on a limited data scope. It does not constitute a definitive mathematical validation nor a product certification.

Abstract

This report presents a targeted exploratory audit of NeoMundi runtime signals, from the perspective of an external *observability / governance* integration. The analysis covers the *G-score*, ΔG, FLAG, the runtime regime and benchmark accuracy. The document does not aim at a definitive mathematical validation; it proposes a methodological reading, interpretation rules and an OBS/GOV-ready recommendation.

Contents

1 Executive Summary

The goal of the audit is to determine whether NeoMundi runtime signals are readable, consistent and usable by an external governance system. Five signals are analyzed: *G-score*, ΔG, `FLAG`, runtime regime and benchmark accuracy.

Three findings structure the report. First, the *G-score* appears to measure runtime stability or coherence, but not factual truth directly. The key case is *misleading stability*: a high score can coexist with an incorrect answer. Second, `FLAG` and `DROP` appear useful as alerting or escalation signals, without constituting automatic proof of error. Finally, the `STABLE` regime is highly dominant; it must therefore be read as a synthetic context, not as a standalone trigger.

The hypothesis of a *G-score* as an operational invariant of generational stability is interesting, but still exploratory. The strong concentration of the score near an observed ceiling may reflect genuine stability or a saturation effect. The central recommendation is therefore a multi-signal reading combining at minimum *G-score*, ΔG, `FLAG`, regime and benchmark accuracy.

2 Objective, Scope and Method

The central question is the following:

> Are the signals G, ΔG, `FLAG`, regime and benchmark accuracy sufficiently clear, stable and interpretable to be consumed as runtime governance signals?

The analysis seeks in particular to characterize zones of stability, tension, rupture, potential drift and misleading stability. The scope covers:

- the *G-score*, read as a signal of runtime stability or behavioral coherence;
- ΔG and `dg_variation`, read as signals of variation, rupture or instability;
- `FLAG` and the `FLAG` rate, read as runtime alerting signals;
- the regime, read as a synthetic state of the system;
- benchmark accuracy, used as an external layer of factual validation.

The report produces neither a definitive mathematical validation, nor final thresholds, nor a JSON integration contract. It proposes a first grammar of interpretation and a short recommendation for OBS/GOV-ready use.

The analyses rely on the provided exports: benchmark accuracy checked per question, anonymized runtime signals and interpreted cases. The main variables used are `provider_id`, `question_id`, `g_score`, `decision`, `regime`, `dg_profile`, `dg_variation`, `is_correct` and `judge_verdict`. The main aggregates were computed after excluding rows with an unknown regime or a null *G-score*.

The approach follows nine steps: inspection of the exports, cleaning, per-provider statistics, cross-analysis between signals, identification of misleading-stability cases, extraction of qualitative cases, analysis of the G invariance hypothesis, formulation of actionable rules, then an OBS/GOV-ready recommendation.

3 Main Results

3.1 G-score and Benchmark Accuracy

The *G-score* remains globally high for most providers, whereas benchmark accuracy varies strongly. The observed correlation between the two signals is weak, around 0.10. This dissociation suggests that the *G-score* captures a property of runtime stability rather than a measure of factuality.

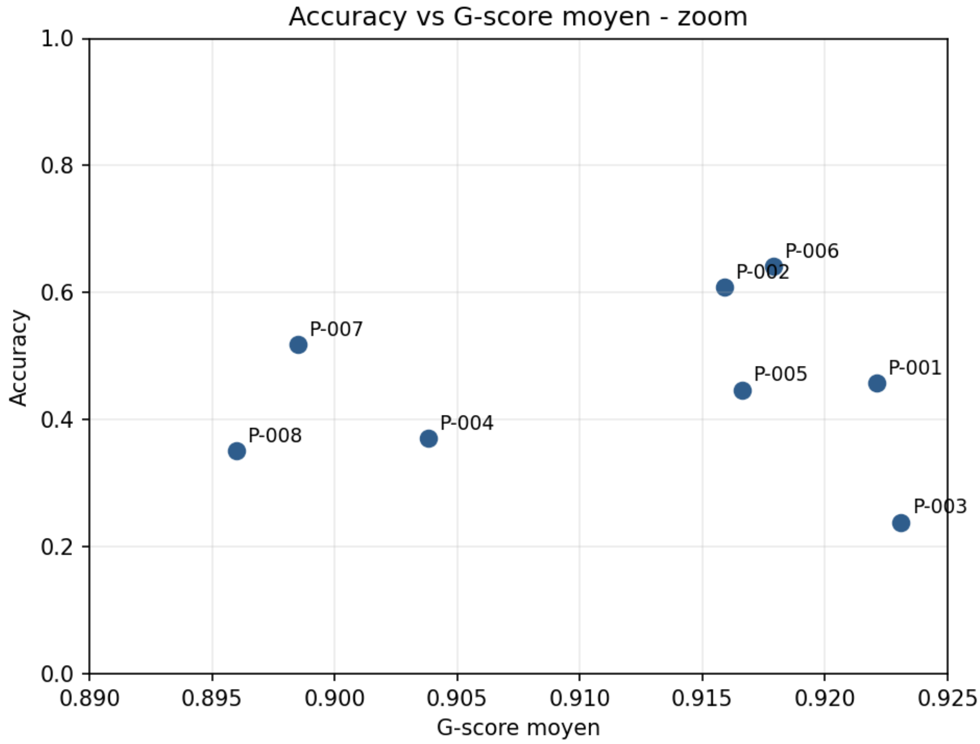

Figure 1: Relationship between benchmark accuracy and mean *G-score* per provider.

Figure 1 shows that providers with similar mean *G-score* values can exhibit very different accuracy levels. Stable behavior therefore does not necessarily imply a correct answer.

3.2 Misleading Stability

The central pattern is *misleading stability*: a high *G-score* with an answer that is incorrect according to the benchmark. Provider P-003 strongly illustrates this case, with a mean *G-score* of 0.9231, a null standard deviation and a 100% saturation rate, despite low accuracy.

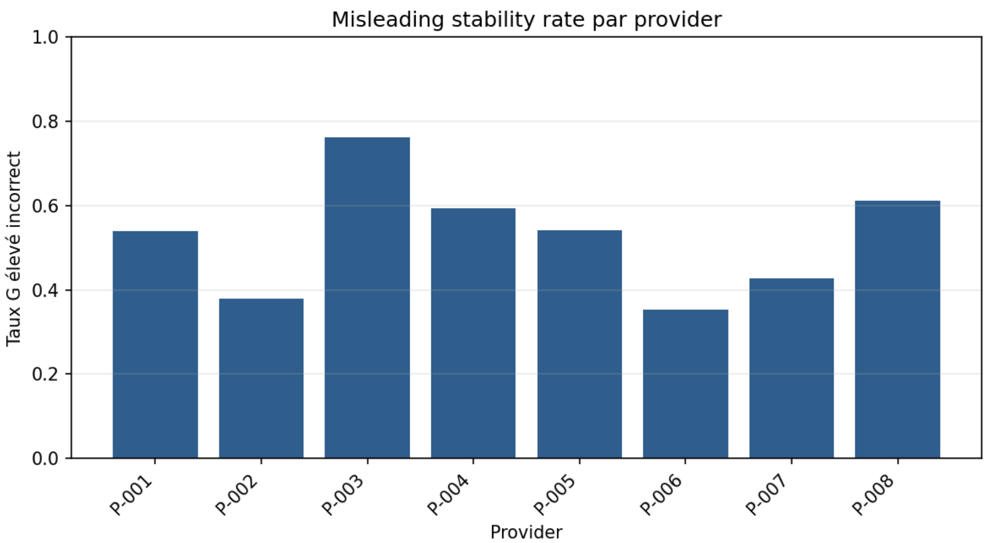

Figure 2: Misleading-stability rate per provider, defined as the proportion of cases where $G \geq 0.90$ and the answer is incorrect.

This result argues for an explicit separation between runtime stability and factual validation in any external integration.

3.3 FLAG, DROP and Regime

`FLAG` cases generally show lower accuracy, a lower *G-score* and a higher ΔG variation than `ALLOW` cases. `DROP` profiles are also associated with instability zones. These signals are therefore useful for alerting, escalation or reinforced verification.

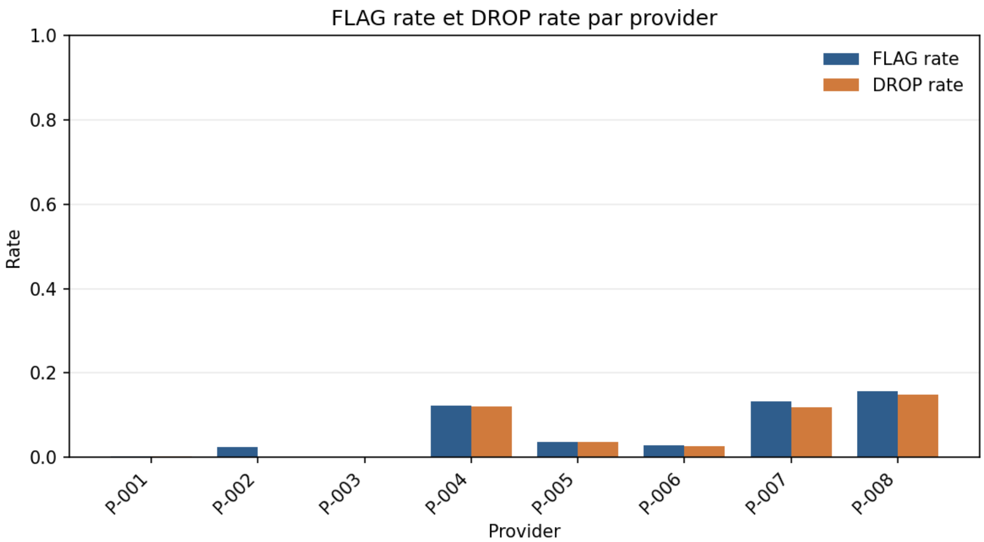

Figure 3: `FLAG` and `DROP` rates per provider.

The `STABLE` regime is highly dominant in the exports. It provides a global context, but its discriminating power is limited: some cases remain `STABLE` despite factual errors, `FLAG` events or `DROP` profiles. The regime must therefore be cross-checked with the other signals.

3.4 G-score Concentration

The *G-score* is strongly concentrated around the maximum observed value 0.9231. The global saturation rate, defined as the proportion of cases where $G \geq 0.9230$, reaches approximately 72.5%.

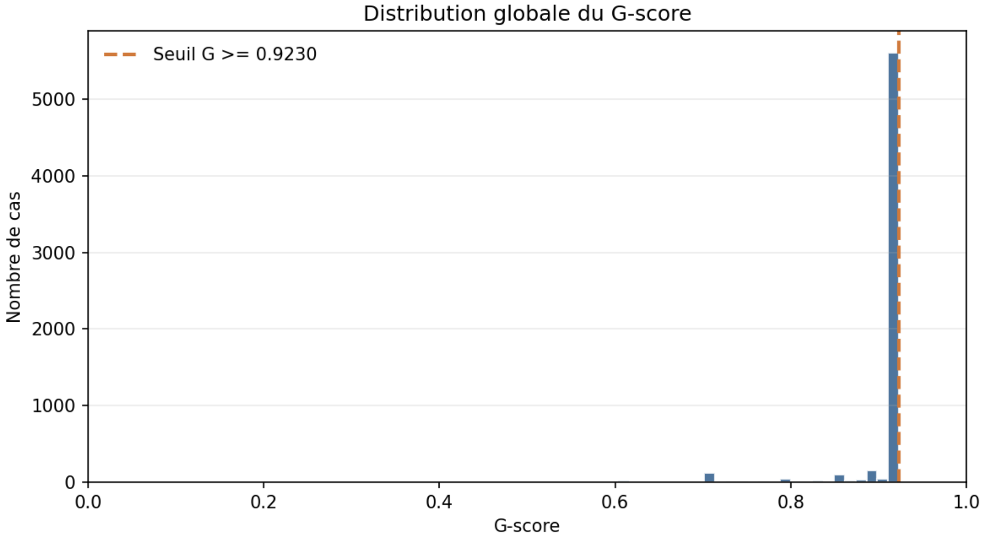

Figure 4: Global distribution of the *G-score* and concentration around the observed ceiling.

This concentration may signal genuine stability of the generative process, but also a saturation effect. This ambiguity is central to discussing the operational invariance hypothesis.

4 Signal-by-Signal Reading

The table below proposes an operational reading of the main NeoMundi signals. It aims to link each observed signal to an interpretation, a potential risk and a recommended action for a first OBS/GOV-ready logic.

Observed signal	Interpretation	Main risk	Recommended action
High *G-score* alone	Stable, regular or coherent generation from a runtime standpoint.	Possible confusion between stability and factual truth.	Do not use *G-score* alone to validate an answer. Cross-check with accuracy or a factual layer.
High *G-score* + low accuracy	Misleading stability: the model produces a stable but incorrect answer.	Strong apparent confidence despite a factual error.	Trigger an external factual validation or reinforced review.
Low or unstable *G-score*	Signal of runtime fragility, irregularity or less stable behavior.	Output potentially sensitive to context, model or prompt.	Examine the question, the provider and the associated variations.
High ΔG or `DROP` profile	Strong signal variation, local rupture or tension zone.	Runtime transition, occasional instability or local degradation.	Classify as a priority attention zone and analyze with `FLAG`, regime and accuracy.
`FLAG`	Runtime alert indicating that an output deserves verification.	Possible over-interpretation as an automatic error.	Use as an escalation or reinforced-verification signal, not as a systematic rejection.
`FLAG` + `DROP`	Stronger runtime risk zone, combining alert and signal rupture.	Increased risk of instability or problematic behavior.	Prioritize the analysis, consider human review or a second validation pass.
`STABLE` regime alone	Synthetic context indicating a globally stable state.	Limited discriminating power when `STABLE` is dominant.	Never use the regime alone; cross-check it with *G-score*, ΔG, `FLAG` and accuracy.
ALLOW + low accuracy	Case not detected by the runtime signal.	Potential false negative in the governance logic.	Reinforce the factual or benchmark layer, especially offline.
Strong *G-score* saturation	Concentration of the score around an observed ceiling.	Ambiguity between genuine stability and saturation effect.	Test robustness across several providers, benchmarks, prompts and repetitions.

Table 1: Interpreting the NeoMundi signal: operational reading of the main runtime signals.

No single signal is sufficient on its own to characterize the risk of a generative output. The *G-score* informs about stability; ΔG and `DROP` about variations; `FLAG` about cases requiring attention; the regime about the global context; benchmark accuracy about factual validation. The combination of these signals forms the basis of a first OBS/GOV-ready grammar.

5 Hypothesis of G as an Operational Invariant

The discussed hypothesis is the following: the *G-score* could measure a stable and interpretable property of the generative process, independently of the provider, the model, the benchmark or the context.

The data partly support this idea: the score is strongly concentrated around 0.9231, and the global saturation rate reaches approximately 72.5%. However, this concentration is ambiguous. It may reflect genuine stability, but also a ceiling effect.

The weak correlation between *G-score* and benchmark accuracy (≈ 0.10) confirms that any potential invariance must not be understood as an invariance of factual truth. If the signal is stable, it is stable rather at the level of generative behavior or runtime coherence.

Some providers and some questions nevertheless exhibit a more marked variability of the *G-score*, often associated with more `FLAG` and `DROP` events. These cases are useful for testing the robustness of the signal. At this stage, the most cautious formulation is therefore the following: the *G-score* appears to measure a property of runtime stability, but its status as an operational invariant remains exploratory.

A consolidation would require several benchmarks, several domains, controlled repetitions, prompt variations, several providers and a separate analysis of saturation effects.

6 Actionable Interpretation Rules

The rules below constitute a first exploratory grammar. They are not definitive thresholds, but practical guidelines for an OBS/GOV-ready reading.

Rule	**Condition**	**Interpretation**	**Recommended action**
R1	High G + low accuracy	Misleading stability.	Add an external factual validation.
R2	High G alone	Runtime stability, not proof of truth.	Never use G alone to validate.
R3	`FLAG` + `DROP`	Priority runtime risk zone.	Escalation or human review.
R4	Low G + high ΔG	Potential rupture or tension.	Investigate as a transition zone.
R5	`FLAG` + correct answer	Alert without a direct factual error.	Read as attention, not rejection.
R6	`STABLE` regime alone	Insufficient synthetic signal.	Cross with G, ΔG, `FLAG` and accuracy.
R7	Stable G + low `FLAG` + low accuracy	Stable but factually weak profile.	Add a factual layer.
R8	Strong G saturation	Strong stability or ceiling effect.	Analyze variability, ΔG, `FLAG`/`DROP`.
R9	High G variability across providers	Sensitivity to the model or context.	Use as a priority test case.
R10	Single signal used alone	Insufficient decision basis.	Use a multi-signal reading.

Table 2: First actionable interpretation grammar for NeoMundi runtime signals.

These rules serve as a basis for an *observability* logic. In a more advanced GOV mode, they could be transformed into configurable thresholds, after validation on more benchmarks, providers, contexts and repetitions.

7 OBS/GOV-ready Recommendation

The following classification summarizes the maturity level recommended for a first external integration.

Signal	Category	Recommended use
G-score	Required	Central runtime-stability signal, to be cross-checked with accuracy.
ΔG/ `dg_variation`	Required	Detection of rupture, tension or local transition.
`FLAG`/ `FLAG` rate	Required	Alerting, escalation or reinforced-verification signal.
Benchmark accuracy	Required	External layer of factual validation.
Regime	Recommended	Synthetic reading of the runtime context, not sufficient alone.
`dg_profile`	Recommended	Simplified variation profile, useful for interpretable rules.
G-score alone	Avoid alone	Risk of confusion between stability and factuality.
Regime alone	Avoid alone	Too synthetic to serve as a standalone trigger.

Table 3: Classification of signals for a first OBS/GOV-ready integration.

For a first OBS integration, the goal is to make outputs observable, interpretable and comparable. For a GOV mode, the signals will need to be combined into configurable rules. The main recommendation is to combine at minimum *G-score*, ΔG, `FLAG`, regime and benchmark accuracy.

8 ControlTower OBS Operational Check

A limited check was carried out via the ControlTower API in OBS mode. It does not constitute an additional statistical validation; it simply verifies the practical behavior of the API and the nature of the signals exposed in production.

OBS mode is a one-off, *snapshot*-type evaluation. The user provides a prompt, an answer already produced by an LLM and simple runtime metrics; ControlTower then returns a governance reading.

The check confirms that the API exposes a structured multi-signal reading, usable within an observability and governance logic. Correct answers were generally associated with a favorable decision and higher internal indicators; incorrect or unfounded answers were more often associated with alerting signals and lower validation indicators.

This step confirms that ControlTower V2 exposes a multi-signal reading in production. It also recalls an important limitation: since OBS is a snapshot, it does not, on its own, allow the analysis of a complete ΔG dynamic. The analysis of ΔG, `DROP` profiles and runtime transitions remains mainly based on the exports containing `dg_variation` and `dg_profile`.

9 Reading Scope and Limitations

The results presented in this report should be read as a first exploratory qualification of NeoMundi signals, and not as a definitive validation of their properties. The analysis helps clarify what the signals appear to indicate, the risks of misinterpretation and the minimal conditions of an OBS/GOV-ready reading.

The main limitation stems from the data scope: the observations rely on the available exports, a given set of anonymized providers and a specific benchmark. They allow useful trends to be identified, but are not yet sufficient to generalize the conclusions to all models, domains or usage contexts.

A second limitation concerns the interpretation of the *G-score*. Its strong concentration around an observed ceiling may reflect genuine stability of the generative process, but also a saturation effect. This is why the operational-invariant hypothesis remains cautious: it requires longitudinal tests, several benchmarks, controlled repetitions and prompt variations.

Finally, the ControlTower OBS check must remain understood as a limited operational check. It confirms that the API exposes a usable multi-signal reading, but OBS mode remains a snapshot.

It therefore does not, on its own, allow the study of a complete ΔG dynamic or a progressive drift over time.

These limitations do not call into question the value of the analyzed signals. They rather clarify their proper use: NeoMundi signals are useful for observability and governance, provided they are combined, contextualized and progressively validated on broader campaigns.

10 Conclusion

The audit shows that NeoMundi signals are readable and usable for a first external reading, provided they are interpreted within a multi-signal logic. The *G-score* is useful for runtime stability, but must not be confused with a measure of factual truth. Misleading-stability cases justify adding an external factual layer.

`FLAG`, `DROP` and ΔG are useful for spotting zones of local tension or rupture, but must remain attention signals. The regime provides a synthetic context, but the dominance of `STABLE` limits its use on its own.

The hypothesis of a *G-score* as an operational invariant remains promising but exploratory. The strong numerical concentration observed may reflect genuine stability or a saturation effect. A more robust validation would require broader tests.

The main recommendation is not to use an isolated signal as a decision basis. For a first OBS/GOV-ready integration, one should combine at minimum *G-score*, ΔG, `FLAG`, regime and benchmark accuracy. This combination makes it possible to distinguish favorable stability, misleading stability, runtime tension, local rupture and the need for complementary factual verification.