EXPLORATORY REPLICATION NOTE

Stability and Reproducibility of the NeoMundi ControlTower Runtime Signal

Comparative multi-provider, multi-corpus and multi-configuration analysis

Fatima Ezzahrae GOUARAB

Data Scientist · Scientific Contributor, NeoMundi Research Observatory

neomundi.org/contributeurs

Status note. This note constitutes an external exploratory report. It does not claim to establish a definitive validation of the NeoMundi ControlTower system, but rather to provide methodological evidence on the stability of governance signals under the experimental conditions studied. To be published after methodological consolidation.

Abstract

Reproducibility is an essential step in the evaluation of any experimental setup. In the field of large language models (LLMs), this issue is particularly important given the continuous evolution of commercial models, the diversity of providers, and the potential influence of generation parameters. This note focuses on the **reproducibility of the governance signals produced by NeoMundi ControlTower** when experimental conditions vary. Its aim is not to study the reproducibility of large language models themselves, nor the intrinsic variability of their responses, but to assess whether the governance signals produced by NeoMundi retain a consistent reading when the model provider, the generation temperature, or the nature of the prompts is changed.

This study builds on earlier work devoted to the actionability of governance signals. During subsequent runs of the initial protocol, an unexpected reversal in the distribution of *ALLOW* and *FLAG* decisions was observed, raising questions about the stability of the observations underlying that first study. To investigate this issue, four successive experimental campaigns were conducted, progressively varying the generation temperature, the model provider, the size of the prompt corpus, and the nature of the situations studied. In total, **680 generations** were analyzed across three model providers and two temperature configurations.

The results show that campaigns based on the classic corpora produce very low proportions of *FLAG* decisions, regardless of provider or temperature. By contrast, the use of a *stress-test* corpus, designed to probe situations known to be conducive to language-model drift, leads to a marked increase in the number of *FLAG* decisions, with an intensity that varies by provider. Among the factors studied, **the nature of the prompts appears to be the factor with the greatest influence on governance decisions**, while **temperature does not appear to play a decisive role under the experimental conditions studied**. In addition, a systematic correspondence is observed between the ΔG *DROP* profile and *FLAG* decisions, confirming the internal consistency of the NeoMundi signal across the configurations tested.

This study constitutes a first methodological contribution to the analysis of the reproducibility of NeoMundi's governance signals. The results show overall consistency in the trends observed across the configurations studied and pave the way for larger-scale replication campaigns to confirm their robustness.

Keywords: reproducibility, AI governance, large language models, NeoMundi ControlTower, governance signals, experimental replication, ΔG, *ALLOW*, *FLAG*, stress-test, multi-provider.

Contents

List of Figures

List of Tables

1. Introduction

1.1. NeoMundi ControlTower: framework and positioning

NeoMundi ControlTower is a runtime governance platform for large language models, developed by NeoMundi Recherche as part of the AI Signals Observatory. Its principle is based on the real-time measurement of token-by-token generation stability, via a signal called the G-Score and its variation dynamics (ΔG), transmitted to the client as an SSE (*Server-Sent Events*) stream during generation.

Unlike classic LLM evaluation approaches that operate *a posteriori* on a complete, finalized response, ControlTower produces a signal during generation, following a paradigm referred to as *runtime governance*. The emitted signal is binary in nature at the level of the final decision: ALLOW when the generation is judged stable, FLAG when drift is detected. It is accompanied by a ΔG variation profile characterizing the dynamics of this stability throughout generation: DROP (drop without recovery), FLAT (flat stability), or V_SHAPE (drop followed by partial recovery).

The current architecture, referred to as ENF mode (Gateway), does not produce any automatic action on the generation: it exposes an external signal that the client system can use according to its own governance policy — early stopping, regeneration, rerouting. This architecture continues the line of Filter E (Favre-Lecca, 2026a), which was an earlier version of automatic regulation based on a notion of informational energy, and of which ControlTower represents the evolution toward a decentralized, client-side governance model.

Work by the NeoMundi Observatory on the thermodynamic mapping of generative services (NeoMundi Recherche, 2026) and the exploratory audit of runtime signals (Diop, 2026) laid the initial empirical groundwork for characterizing this signal across several providers. This note continues that line of work, focusing specifically on the question of the reproducibility of NeoMundi decisions when experimental conditions vary.

1.2. Scope and object of this note

This note does not study the reproducibility of large language models as such, nor the variability of their responses from one run to another. Its object is more specific: it studies the reproducibility of the governance signals produced by NeoMundi ControlTower when experimental conditions vary. The question being asked is not whether the models produce the same responses, but whether NeoMundi produces the same governance decisions when the provider, the temperature, or the nature of the prompts is varied. This distinction places NeoMundi ControlTower itself at the center of the analysis, as the object of evaluation rather than merely a measurement tool.

This note is also distinct from the Phase 1 report, which addressed the actionability of the ControlTower signal in a stop-generation scenario. These two contributions are complementary: actionability only has operational meaning if the signal it relies on is stable and reproducible. This note therefore constitutes a necessary methodological prerequisite for any generalization of the conclusions of the Phase 1 report.

1.3. Starting point: the observed reversal

The initial experiments conducted as part of this contribution aimed to study the actionability of ControlTower's governance signals. These experiments, carried out with PROFILE-5B3BB7 at temperature 0.7 on a corpus of 50 prompts organized into three categories (factual, reasoning, and speculative), yielded a FLAG rate of 88%, with a perfect bijection between the ΔG DROP profile and the FLAG decision.

Several reruns of this same protocol were subsequently carried out. Most of these reruns confirmed the initial trend, characterized by a predominance of FLAG decisions. During one of these reruns, however, the result unexpectedly reversed: the ALLOW decision became strongly predominant, even though the code, the model, and the prompt corpus remained identical. This reversal affected the entire corpus and completely reversed the decision distribution initially observed.

1.4. Hypotheses formulated

Faced with this observation of non-reproducibility, four explanatory hypotheses were formulated.

The first hypothesis attributes the reversal to the generation temperature. Phase 1 had used T=0.7, a parameter that had not been explicitly controlled during the reruns. Since a higher temperature promotes response variability, it could mechanically increase the frequency of drifts detected by NeoMundi.

The second hypothesis attributes the reversal to a model versioning effect. The alias `PROFILE-5B3BB7` (version: latest) used in Phase 1 does not reference a fixed model version. A silent update between the two runs could have changed the generation behavior with no apparent change in the code.

The third hypothesis attributes the reversal to a provider effect. The behavior observed on PROFILE-5B3BB7 could be specific to this model, with other providers showing different decision profiles on the same corpus.

The fourth hypothesis attributes the reversal to a prompt-type effect. The Phase 1 prompts may have had particular properties, notably the presence of questions about future events that had not yet occurred, making them structurally more likely to trigger a FLAG regardless of the model or the temperature.

These four hypotheses are not mutually exclusive. This note does not claim to fully disentangle them or to test them with sufficient statistical power to draw definitive conclusions. It aims to document, in a transparent and iterative manner, the effects observed under controlled but imperfect conditions.

1.5. Research question and objective

The central question of this note is as follows: do the governance signals produced by NeoMundi ControlTower retain a consistent reading when the prompts, providers, and execution parameters change? This formulation aligns with the one adopted by S. Favre-Lecca to frame this replication as part of the Observatory.

The objective is to produce an exploratory replication note rigorously documenting the results obtained under controlled but imperfect conditions. This is a first cautious methodological demonstration in the sense formulated by S. Favre-Lecca: under the conditions tested, despite the expansion of the corpus and the variation of configurations, do the main trends remain consistent? The protocol was built through successive iterations, each result guiding the next question. This document presents that chronology transparently rather than artificially smoothing it over.

2. Protocol and configurations tested

The replication protocol was built progressively, with each trial designed as a direct response to the observations of the previous one. This iterative approach, which departs from a fully predefined experimental design, is acknowledged as such and presented in its actual chronology rather than in a form that has been rationalized in hindsight. Four successive trials were thus conducted, each varying one or more dimensions of the initial protocol while keeping the technical setup constant.

2.1. Trial 1 — Temperature effect (n=50 per configuration)

The first trial aimed to test the hypothesis that the generation temperature was an explanatory factor for the observed reversal. The full corpus of 50 Phase 1 prompts was submitted twice to the same provider (`PROFILE-5B3BB7`), varying only the temperature: T=0.7, corresponding to a new generation under the original configuration, and T=0, corresponding to a deterministic configuration that eliminates all stochastic variability. This setup made it possible to observe whether a temperature change alone, all else being equal, produced different NeoMundi decisions.

2.2. Trial 2 — Provider effect (n=50 per provider)

The second trial consisted of submitting the full corpus of 50 Phase 1 prompts to two additional providers, PROFILE-739F7C and PROFILE-F5FF91, at temperature 0. The aim was to check whether the behavior observed on PROFILE-5B3BB7 during the Phase 1 rerun was specific to that model or also occurred on other architectures given the same prompts under the same conditions.

Temperature was set to 0 for this trial in order to remove stochastic generation variability and more cleanly isolate the provider effect. Each provider was tested in an independent execution session, with the same 50 prompts submitted in the same order, thereby ensuring the comparability of conditions across the three configurations.

2.3. Trial 3 / Campaign A — Corpus expansion (n=100 per provider)

At the end of Trials 1 and 2, the FLAG rates observed remained very low across all configurations tested. However, the limited sample of 50 prompts and the restriction to the Phase 1 corpus alone left several interpretations open. Campaign A therefore expanded the corpus to 100 prompts per provider, keeping the same Phase 1 categories and adding an additional category of prompts containing spelling and syntax errors, designed to test whether degraded phrasing affected NeoMundi decisions independently of semantic content.

This campaign was run separately for each of the three providers in three separate note-

books. This separate-architecture design addressed an operational robustness constraint: a technical incident affecting one provider should not compromise data collection for the other two, with each campaign able to be rerun independently in the event of a partial failure.

2.4. Trial 4 / Campaign B — Stress-test corpus and temperature control (n=45 per configuration)

Campaign A had confirmed the drop in the FLAG rate on the classic categories without allowing a distinction between two equally plausible interpretations: either the models tested had become more stable in general, or the classic corpus no longer probed the mechanisms likely to produce drifts detectable by NeoMundi. To settle between these two hypotheses, Campaign B deliberately replaced the prompt categories rather than increasing their volume, building an entirely new corpus of 45 prompts designed according to a targeted stress-test logic. A detailed description of this corpus and the justification for the choice of each category are presented in Section 3.

At the same time, Campaign B combined four configurations in a single notebook, allowing a direct, simultaneous comparison: PROFILE-5B3BB7 at T=0.7, PROFILE-5B3BB7 at T=0, PROFILE-739F7C at T=0, and PROFILE-F5FF91 at T=0. Keeping the temperature at 0 for three of the four configurations made it possible to attribute any difference observed between providers to the model itself rather than to stochastic generation variability. The PROFILE-5B3BB7 at T=0.7 configuration was kept as a direct reference point against Phase 1.

2.5. Common technical setup

Across all trials, the technical setup is identical to that of Phase 1. Each generation is submitted to the NeoMundi ControlTower API in ENF mode (Gateway, SSE streaming). In this mode, NeoMundi calls the LLM provider directly and measures the stability of the generation token by token via an SSE stream, with no client-side intermediary. Call parameters are constant across all configurations: a maximum of 512 generated tokens and a timeout of 120 seconds per request.

For each generation, the following variables are extracted from the *done* event of the SSE stream. The final governance decision (ALLOW or FLAG) is the main outcome variable of this study. The overall stability score (`stability_score`) measures generation stability on a scale from 0 to 1. The ΔG variation profile (`delta_profile`) characterizes the dynamics of this stability during generation according to three configurations: DROP, FLAT, or V_SHAPE. The total amplitude of the variation (`delta_variation`) and the complete time series of ΔG measurement points (`delta_series`) complete the description of the signal's dynamic behavior. The factual hallucination score (`hallucination_score`) and

the semantic coherence score (`coherence_score`) provide additional information on the nature of the drifts detected. The total number of generated tokens (`total_tokens`) helps to contextualize the observations in relation to the length of the responses produced.

The summary table of all the configurations tested as part of this replication is available in Appendix A (see Table 12).

Remark. The names of the three providers evaluated in this study have been anonymized using the aliases `PROFILE-5B3BB7`, `PROFILE-739F7C`, and `PROFILE-F5FF91` for confidentiality reasons. This anonymization does not affect the reported analyses or results. However, it prevents independent verification by external third parties and precludes direct comparison with other publicly available evaluations involving the same models.

3. Description of the corpora and prompt categories

3.1. Reference corpus — Phase 1 and Trials 1 and 2 (n=50)

The corpus used in Phase 1, reused unchanged for Trials 1 and 2, consists of 50 prompts written in French and organized along a gradient of increasing complexity across three categories. This structure met a construct-validity objective: rather than measuring NeoMundi decisions on a homogeneous set of questions, the goal was to cover a spectrum of generative situations sufficiently contrasted to observe whether the signal responded differently depending on the nature of the task asked of the model.

The **Factual** category groups 15 prompts covering single-answer, verifiable questions, spanning fields as varied as geography, history, science, or economics. These questions constitute the reference condition in which generation stability is expected to be at its maximum, since the model in principle has a well-established canonical answer in its training data. It should be noted, however, that some prompts in this category concerned events that had not yet occurred at the time of execution, such as the awarding of annual prizes or future sports results. These particular cases structurally expose the model to a specific type of drift known as *temporal hallucination*, regardless of their apparently factual wording. This point constitutes a source of internal heterogeneity within the Factual category, documented here so that it can be taken into account when interpreting the results.

The **Reasoning** category groups 20 prompts asking the model to explain a mechanism, distinguish between two concepts, or describe how a system works. These questions involve a multi-step explanatory chain, in which the risk of drift is expected to be more gradual and cumulative than for factual questions. The slightly larger sample size for this category reflected the initial hypothesis that intermediate drift profiles would be more frequent here and would require more observations to be characterized reliably.

The **Speculative** category groups 15 prompts covering open-ended, normative, or forward-looking questions, with no verifiable canonical answer. In Phase 1 these questions represented the condition in which the FLAG rate was expected to be the highest, as the model is exposed to narrative expansion that is loosely constrained by factual checkpoints.

It should be emphasized that the 15/20/15 split between these three categories did not result from a differentiated sample-size calculation, but from a reasoned choice aimed at coverage. This is a convenience sample designed for contrast, not a random draw from a defined population of prompts. This limitation implies that the FLAG rates observed per category cannot be interpreted as unbiased estimators of a population rate, but rather characterize the behavior of the signal on this specific corpus.

3.2. Expanded corpus — Campaign A (n=100)

The Campaign A corpus directly extends the Phase 1 corpus following the same complexity-gradient logic, roughly doubling the size of each category: 30 prompts for the Factual category, 35 for Reasoning, and 30 for Speculative. The new prompts were designed to remain consistent with the themes and complexity level of the initial prompts, in order to preserve comparability between the two corpora.

A fourth category was introduced for the first time in this campaign, made up of 5 prompts whose wording contains spelling and syntax errors, with no change to the semantic content of the question asked. The purpose of this category was to check whether degraded phrasing affected NeoMundi decisions independently of content, i.e., whether the signal responded to form as much as to substance. With only 5 prompts, this category has an illustrative rather than probative function; its results are presented for informational purposes in the analyses.

3.3. Stress-test corpus — Campaign B (n=45)

The Campaign B corpus deliberately breaks with the logic of the previous corpora. Rather than extending the classic complexity gradient, it is based on a targeted stress-test logic: each category targets a specific drift or hallucination mechanism, documented in the LLM literature, that the classic corpora did not or only barely probed. This design choice directly followed from the hypothesis formulated at the end of Campaign A, namely that the drop in the FLAG rate might reflect a mismatch between the corpus and the models' behavior rather than a general stabilization of the models.

The **Current events** category groups 6 prompts covering recent events or events that had not yet occurred at the time of execution. LLMs, whose knowledge is frozen at a training cutoff date, are structurally exposed to temporal hallucination on this type of question: unable to verify the information, they tend to produce a plausible answer that is not grounded in real facts.

The **Precise figures** category groups 6 prompts requesting exact numerical values, whether economic statistics, benchmark scores, or demographic data, which the model cannot know with absolute certainty. This type of question frequently prompts the production of a credible but invented figure, a well-documented quantitative-hallucination mechanism in studies on LLM factual reliability.

The **Fictitious entity** category groups 6 prompts presenting the model with names of theories, people, or organizations that do not exist. It tests the model's propensity to confabulate a coherent answer rather than flag the absence of a known referent in its training data.

The **Quotation** category groups 6 prompts requesting the exact reproduction of a quotation

together with its precise source. Faithful reproduction of quotations is one of the most frequently documented hallucination mechanisms in the LLM literature: the model tends to reconstruct plausible-sounding wording rather than reproduce the original text, producing quotations that sound plausible but are inaccurate.

The **False premise** category groups 6 prompts, each built on an obviously false premise. It tests whether the model corrects the erroneous premise or continues its reasoning as if it were true, with the latter behavior associated with unfounded narrative expansion.

The **Multi-step** category groups 6 prompts requiring a chain of calculation or reasoning across several successive steps. The risk of drift here is expected to be cumulative, as each step in the inferential chain can introduce an error that propagates to the following steps.

The **Extreme speculative** category groups 4 prompts covering normative or philosophical questions with no possible factual answer, relating to topics such as the rights of artificial intelligences or artificial consciousness. This area is associated with high informational density and narrative expansion loosely constrained by factual checkpoints, which made it a natural candidate for the stress test.

Finally, the **Errors** category groups 5 prompts reused unchanged from Campaign A. This deliberate choice of reuse was intended to allow a direct comparison of the effect of degraded wording between the two campaigns, all else being equal in terms of content.

The Campaign B corpus thus totals 45 prompts: 40 entirely new prompts spread across the seven stress-test categories described above, and 5 reused unchanged from Campaign A.

4. Results

The results are presented in response to the six research questions formulated from the initial hypotheses. For each question, a summary table accompanies the analytical interpretation. Figures are inserted throughout the text.

4.1. Are NeoMundi decisions stable from one trial to another?

This first question directly concerns the reproducibility of the governance decisions produced by NeoMundi ControlTower when experimental conditions vary.

Table 1. Distribution of NeoMundi decisions by campaign

Campaign	n	ALLOW	FLAG	% ALLOW	% FLAG
Trial 1 — Temperature (n=50×2)	100	98	2	98.0%	2.0%
Trial 2 — Provider (n=50×2)	100	97	3	97.0%	3.0%
Trial 3 — Campaign A (n=100×3)	300	294	6	98.0%	2.0%
Trial 4 — Campaign B (n=45×4)	180	159	21	88.3%	**11.7%**

Trials 1, 2, and 3 produce stable, low FLAG rates (2.0–3.0%), independent of provider and temperature. Trial 4, run on the stress-test corpus, brings this rate back up to 11.7%: not an anomaly, but the expected effect of a corpus designed to probe documented drift mechanisms.

4.2. Does the prompt type influence NeoMundi decisions?

Table 2. Decisions by category — Campaign A (n=100 per provider, 3 providers)

Category	ALLOW	FLAG	Total	% FLAG
Factual	84	6	90	**6.7%**
Reasoning	105	0	105	0.0%
Speculative	90	0	90	0.0%
Errors	15	0	15	0.0%
Total	**294**	**6**	**300**	**2.0%**

Table 3. Decisions by stress-test category — Campaign B (sorted by decreasing FLAG rate)

Category	ALLOW	FLAG	Total	% FLAG
Current events	16	8	24	**33.3%**
False premise	20	4	24	**16.7%**
Fictitious entity	21	3	24	**12.5%**
Precise figures	21	3	24	**12.5%**
Quotation	22	2	24	8.3%
Multi-step	23	1	24	4.2%
Errors	20	0	20	0.0%
Extreme speculative	16	0	16	0.0%
Total	**159**	**21**	**180**	**11.7%**

Prompt type is the most discriminating factor. In Campaign A, only the Factual category produces any FLAGs (6.7%), attributable to prompts concerning future events (temporal hallucination, §3.1). In Campaign B, the gradient is clear: Current events (33.3%) at the top, Extreme speculative and Errors at 0%. This result indicates that the NeoMundi signal responds to verifiable factual uncertainty, not to normative uncertainty.

4.3. Does the provider influence NeoMundi decisions?

Table 4. Decisions by provider at fixed temperature T=0 — Trials 2, 3, and 4

Trial	Provider	ALLOW	FLAG	Total	% FLAG
Trial 2	PROFILE-739F7C	48	2	50	4.0%
	PROFILE-F5FF91	49	1	50	2.0%
Trial 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%**

On classic corpora (Trials 2 and 3), no inter-provider difference is observable (1–4% FLAG for all). On the stress-test corpus (Trial 4, T=0 fixed), a gradient emerges: PROFILE-739F7C (17.8%) > PROFILE-5B3BB7 (11.1%) > PROFILE-F5FF91 (4.4%). This effect would warrant confirmation on a larger corpus before any generalization (n=45 per configuration).

4.4. Does temperature influence NeoMundi decisions?

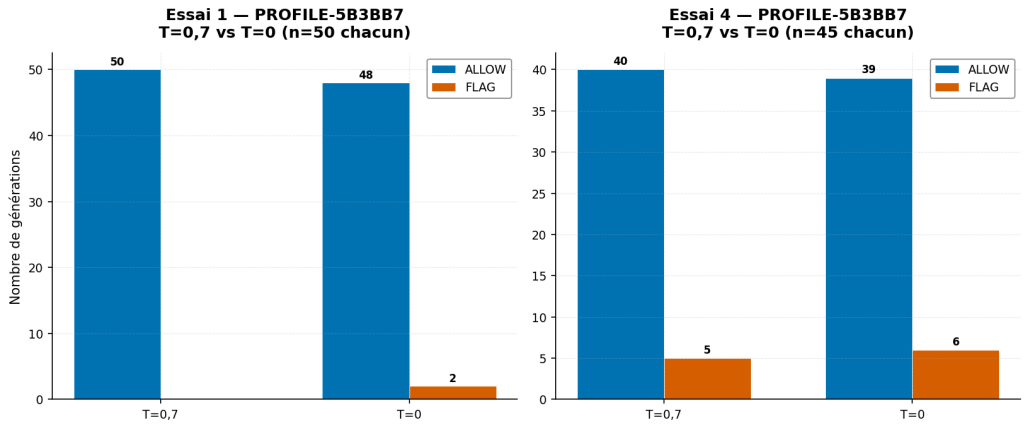

Figure 1. Effect of temperature on NeoMundi decisions — Trial 1 (n=50 per configuration) and Trial 4 / Campaign B (n=45 per configuration)

The decision distributions obtained with PROFILE-5B3BB7 at temperature 0 and 0.7 remain very similar in both campaigns. In Trial 1, the number of *FLAG* decisions goes from 0 to 2 (0 % versus 4 %), and in Trial 4 from 5 to 6 (11.1 % versus 13.3 %). These differences remain small in magnitude and do not support identifying temperature as the main factor driving variation in governance decisions under the experimental conditions studied.

4.5. Do the other metrics evolve consistently?

The stability score is consistently lower for FLAG cases than for ALLOW cases, confirming the consistency between the continuous metric and the binary decision. The mean ALLOW−FLAG gap increases from 0.036 in Phase 1 to 0.195 in Trials 3 and 4: the NeoMundi decision boundary is sharper and more discriminating in the replication. The hallucination score shows nonzero values mainly in the Current events and Fictitious entity categories, consistent with their high FLAG rates. Full descriptive statistics are available in Appendix B.

4.6. Do the ΔG profiles change depending on the provider or prompt type?

Table 5. ΔG profiles by provider — Trial 3 (n=100 per 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. Decision × ΔG profile consistency — Trial 3 (n=300)

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

Table 7. ΔG profiles by configuration — Trial 4 (n=45 per 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**

The DROP⇔FLAG bijection is perfect and constant across all trials: 6 DROP for 6 FLAG in Trial 3, 21 DROP for 21 FLAG in Trial 4, with no exception. The DROP profile is a deterministic predictor of the FLAG decision. Notably, the V_SHAPE profile, present in Phase 1 (5 cases, 10%), is completely absent from the replication (0 cases out of 480 generations).

5. Comparison with prior results

This section compares the results of the replication campaigns with the Phase 1 observations along four axes: overall decision distribution, ΔG profiles, category effects, and stability score. It should be recalled that this is not a strict replication in the experimental sense — several variables changed simultaneously between Phase 1 and the replication trials — and that the observations presented remain descriptive in nature.

5.1. Evolution of the decision distribution

Table 8. Distribution of decisions — Phase 1 vs replication campaigns

Study	n	ALLOW	FLAG	% FLAG
Phase 1 — PROFILE-5B3BB7, T=0.7	50	6	44	**88.0%**
Trial 1 — PROFILE-5B3BB7 , T=0.7	50	48	2	4.0%
Trial 1 — PROFILE-5B3BB7, T=0	50	50	0	0.0%
Trial 2 — PROFILE-739F7C, T=0	50	48	2	4.0%
Trial 2 — PROFILE-F5FF91, T=0	50	49	1	2.0%
Trial 3 — PROFILE-5B3BB7, T=0	100	99	1	1.0%
Trial 3 — PROFILE-739F7C, T=0	100	97	3	3.0%
Trial 3 — PROFILE-F5FF91, T=0	100	98	2	2.0%
Trial 4 — PROFILE-5B3BB7 T=0.7 (stress-test)	45	39	6	13.3%
Trial 4 — PROFILE-5B3BB7 T=0 (stress-test)	45	40	5	11.1%
Trial 4 — PROFILE-739F7C T=0 (stress-test)	45	37	8	**17.8%**
Trial 4 — PROFILE-F5FF91 T=0 (stress-test)	45	43	2	4.4%

The contrast is striking. While Phase 1 yielded a FLAG rate of 88.0%, all experiments conducted on the classical corpus produced substantially lower rates, ranging from 0.0% to 4.0%, irrespective of the provider or temperature setting. Only the stress-test corpus resulted in a noticeable increase, reaching a maximum FLAG rate of 17.8% for PROFILE-739F7C, which nevertheless remains approximately five times lower than the rate observed in Phase 1.

The most direct comparison is provided by Trial 1, which reused the same set of 50 prompts as Phase 1. Under identical temperature conditions (T = 0.7), the FLAG rate decreased from 88.0% for PROFILE-5B3BB7 in Phase 1 to 4.0% for PROFILE-5B3BB7-2407 in

Trial 1, representing an 84-percentage-point difference. Such a discrepancy cannot be attributed to the temperature parameter alone, whose isolated effect accounts for only a four-point variation. These findings therefore indicate that model version and prompt characteristics are the primary factors influencing runtime signal activation.

5.2. Evolution of the ΔG profiles

Table 9. Distribution of ΔG profiles — Phase 1 vs replication campaigns

Profile	Phase 1 (n=50)	Trial 3 (n=300)	Trial 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%)**

The DROP⇔FLAG bijection holds in both periods without exception. In Phase 1, the DROP profile predominated (86%), consistent with the high FLAG rate. In the replication, the FLAT profile becomes strongly predominant (94.4% across 480 generations). The most notable finding is the complete disappearance of the V_SHAPE profile: present at 10% in Phase 1, it is absent from all 480 generations in Trials 3 and 4, a point of attention for future replications.

5.3. Evolution by prompt category

Table 10. FLAG rate by category — Phase 1 vs Trial 3

Category	Phase 1 % FLAG	Trial 3 % FLAG
Reasoning	90.0%	**0.0%**
Factual	86.7%	6.7%
Speculative	86.7%	**0.0%**
Errors	—	0.0%

The Reasoning and Speculative categories go from 87–90% FLAG in Phase 1 to 0% in the replication. Only the Factual category maintains a residual rate of 6.7%, attributable to prompts about future events that had not yet occurred. These changes suggest that the corpus effect predominates over any model or temperature effect in explaining the gap between the two periods.

5.4. Evolution of the stability score

Table 11. Stability score by decision — Phase 1 vs replication campaigns

Indicator	Phase 1	Trial 3	Trial 4
Overall mean score	0.795	0.923	0.923
ALLOW mean score	0.827	0.923	0.923
FLAG mean score	0.791	0.728	0.727
ALLOW minus FLAG gap	0.036	**0.195**	**0.196**

The ALLOW mean score increases from 0.827 to 0.923 between Phase 1 and the replication, reflecting more stable generations on less demanding corpora. More significantly: the ALLOW−FLAG gap increases from 0.036 in Phase 1 to 0.195 in the replication. The NeoMundi decision boundary is sharper in the replication campaigns, where FLAGs correspond to more pronounced drifts. The rare FLAGs in the replication are therefore more robust signals than those in Phase 1.

5.5. Methodological limitations

Any exploratory replication study requires transparently documenting the limitations of the experimental protocol. This study is no exception. The limitations presented below should be taken into account when interpreting the results and assessing their scope.

Silent evolution of commercial models (*silent model updates*). The main limitation concerns the continuous evolution of commercial models. Providers regularly update their models to improve performance, correct certain behaviors, or adjust their generation policies. These updates are often deployed without any change to the model's commercial name and without detailed public documentation.

As a result, two experiments run several weeks or months apart may be executed with what appears to be the same model, while actually relying on different internal versions. These **silent changes** are today a major challenge for reproducibility studies involving commercial LLMs.

Within the scope of this work, this hypothesis can neither be confirmed nor ruled out. It nonetheless represents a plausible explanation for the reversal of results observed between the initial Phase 1 experiments and the subsequent runs, even though the experimental protocol appeared unchanged. It is precisely this observation that motivated the various replication campaigns presented in this note.

Incomplete control of experimental variables. The campaigns conducted make it possible to study several factors that may influence the observations, notably the model provider, the generation temperature, or the nature of the prompt corpus. However, it is not possible to perfectly isolate all the variables that may have changed between the initial experiments and the replication campaigns.

Consequently, the differences observed cannot be attributed with certainty to a single cause. The interpretations proposed in this study should be considered methodological hypotheses grounded in the experimental observations, rather than causal demonstrations.

Corpus representativeness. The replication campaigns rely on several corpora of different sizes (50, 100, and 45 prompts), designed to explore a variety of situations: factual questions, reasoning tasks, speculative questions, and stress-test scenarios. Although these corpora make it possible to study a wide range of behaviors, they do not claim to represent the full range of possible uses of large language models. The results presented should therefore be interpreted as representative of the configurations studied, without being generalized to all usage contexts.

Scope of the study. Finally, this note does not aim to definitively validate the NeoMundi system. It constitutes a first methodological study devoted to the reproducibility of governance signals. Its ambition is to document the trends observed across several experimental configurations and to propose a replication protocol that can be reproduced and expanded in future work.

6. Conclusion

This note started from a simple but unsettling observation: the same experimental protocol, applied under conditions declared identical, had produced radically different distributions of NeoMundi decisions from one period to another. Rather than ignoring this discrepancy or downplaying it, the approach adopted was to document it rigorously and to build, iteratively, a set of trials designed to explore its explanatory factors. It is this methodological transparency, as much as the results themselves, that constitutes the main contribution of this note.

The results obtained provide a nuanced answer to the central question posed: do the governance signals produced by NeoMundi ControlTower retain a consistent reading when prompts, providers, and execution parameters change?

On the classic corpora, the answer is yes. Trials 1, 2, and 3 produce remarkably stable FLAG rates, ranging between 0.0% and 4.0%, across three distinct providers and two different temperatures. This cross-trial consistency is an encouraging stability result for the instrument. The perfect bijection between the ΔG DROP profile and the FLAG decision, maintained without exception across all 480 generations of Trials 3 and 4, reinforces the internal consistency of the signal and its operational readability.

On the stress-test corpus, the answer is also yes, but with an important nuance: the NeoMundi signal responds differently depending on the nature of the prompts submitted. Categories targeting verifiable factual uncertainty produce noticeably higher FLAG rates than speculative or general-reasoning categories, which remain at 0%. This sensitivity to prompt type is not a weakness of the signal: it reflects consistency between the instrument's behavior and the nature of the situations that structurally expose a model to verifiable drift. It implies, however, that corpus composition is a determining factor in the observations, and that any comparison between campaigns must take this dimension into account.

The question of the gap between Phase 1 and the replication, however, does not receive a definitive answer. The analysis converges on the hypothesis that the specific nature of the Phase 1 prompts, combined with a possible model-versioning effect, is the most plausible explanation, with temperature turning out to be a negligible explanatory factor. But the co-variation of several factors between the two periods does not allow for a rigorous causal attribution. This residual uncertainty is documented as such, rather than concealed behind an overly confident interpretation.

These results open up several concrete avenues for future work by the Observatory. A larger-scale replication, with sample sizes on the order of 300 to 400 prompts per configuration, would achieve sufficient statistical power to confirm or refute the descriptive trends observed. An extension to additional providers and to even more diverse corpora would strengthen the scope of the conclusions regarding the cross-configuration stability of the NeoMundi

signal.

As it stands, this note constitutes an honest and reproducible exploratory contribution to the Observatory's methodological knowledge base. It does not claim to establish a definitive truth about the reproducibility of the NeoMundi signal, but provides solid empirical evidence, transparently documented, on which future work can build.

7. Appendices

7.1. Appendix A — Campaign manifest

Table 12. Manifest of execution campaigns

Campaign	Provider	Temp.	n exec.	n retained
Phase 1	PROFILE-5B3BB7	0.7	50	50
Trial 1 T=0.7	PROFILE-5B3BB7	0.7	50	50
Trial 1 T=0	PROFILE-5B3BB7	0	50	50
Trial 2 PROFILE-739F7C	PROFILE-739F7C	0	50	50
Trial 2 PROFILE-F5FF91	PROFILE-F5FF91	0	50	50
Trial 3 PROFILE-5B3BB7	PROFILE-5B3BB7	0	100	100
Trial 3 PROFILE-739F7C	PROFILE-739F7C	0	100	100
Trial 3 PROFILE-F5FF91	PROFILE-F5FF91	0	100	100
Trial 4	4 conf.	0/0.7	180	180
Total replication			**680**	**680**

7.2. Appendix B — CSV exports

The following files contain the raw data collected during each campaign, in CSV format with UTF-8 encoding. Each row corresponds to one generation and contains all the variables described in Section 2.5.

Table 13. Raw data files

File name	Campaign	n rows
essai1_PROFILE-5B3BB7_T07.csv	Trial 1 — PROFILE-5B3BB7 T=0.7	50
essai1_PROFILE-5B3BB7_T00.csv	Trial 1 — PROFILE-5B3BB7 T=0	50
essai2_PROFILE-739F7C.csv	Trial 2 — PROFILE-739F7C	50
essai2_PROFILE-F5FF91.csv	Trial 2 — PROFILE-F5FF91	50
essai3_PROFILE-5B3BB7.csv	Trial 3 — PROFILE-5B3BB7	100
essai3_PROFILE-739F7C.csv	Trial 3 — PROFILE-739F7C	100
essai3_PROFILE-F5FF91.csv	Trial 3 — PROFILE-F5FF91	100
essai4_tous_providers.csv	Trial 4 — 4 configurations	180
Total		**680**

7.3. Appendix C — Exploratory statistical tests

To check whether the differences observed between groups exceed the expected stochastic variability, three exploratory statistical tests were conducted: a Pearson χ^2 test for comparisons involving more than two groups, and Fisher's exact test for two-group comparisons with small sample sizes. These tests are presented for informational purposes and interpreted with caution, given the limited sample sizes and the absence of formal correction for multiple comparisons.

Table 14. Results of the exploratory statistical tests

Comparison	Test	Statistic	p-value	Conclusion
Provider effect (Trial 4, T=0)	χ^2	4.050	0.132	Not significant
Temperature effect (Trial 4, PROFILE-5B3BB7)	Fisher	—	1.000	Not significant
Category effect (Trial 4)	χ^2	9.868	0.079	Not significant

None of the three tests produces a significant result at the conventional $\alpha = 0.05$ threshold. These results mainly reflect a lack of statistical power due to the available sample sizes, and not an absence of effect. The temperature effect (p=1.000) quantitatively confirms the absence of impact already observed descriptively. The category effect (p=0.079) is the closest to the significance threshold, consistent with the descriptive gradient observed in Section 4.2. The provider effect shows a notable descriptive gap between PROFILE-739F7C (17.8%) and PROFILE-F5FF91 (4.4%), which would warrant confirmation on a larger corpus. The full contingency tables are available in Appendix E.

7.4. Appendix D — Additional visualizations

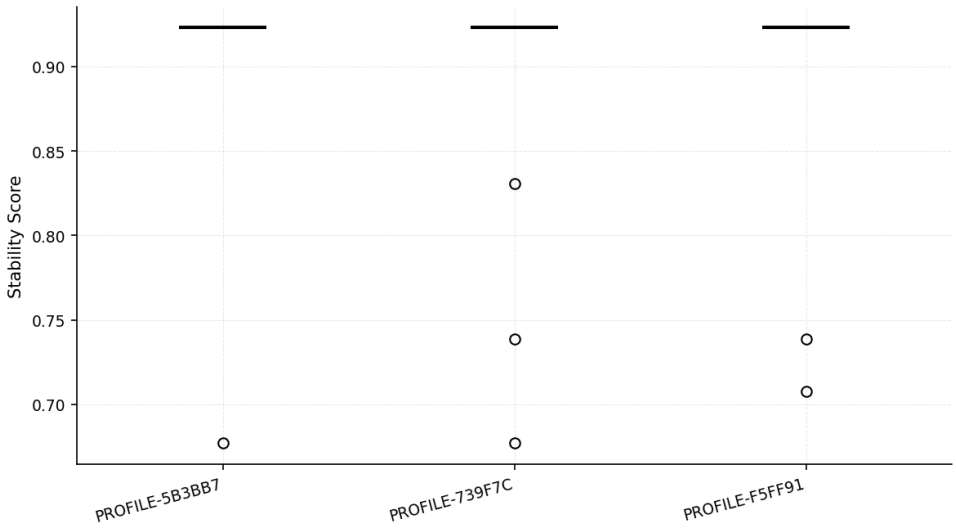

Figure 2. Distribution of the Stability Score by provider — Trial 3 / Campaign A (n=100 per provider)

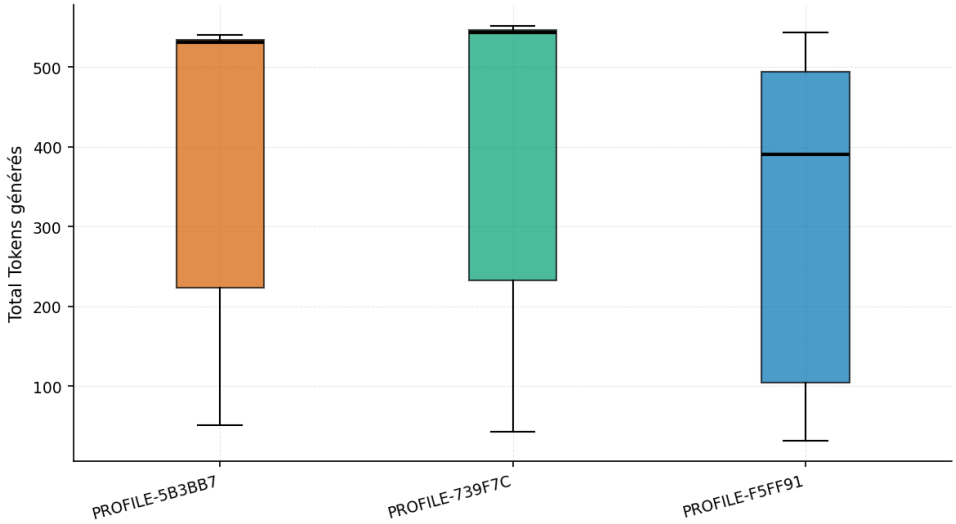

Figure 3. Distribution of the number of generated tokens by provider — Trial 3 / Campaign A (n=100 per provider)

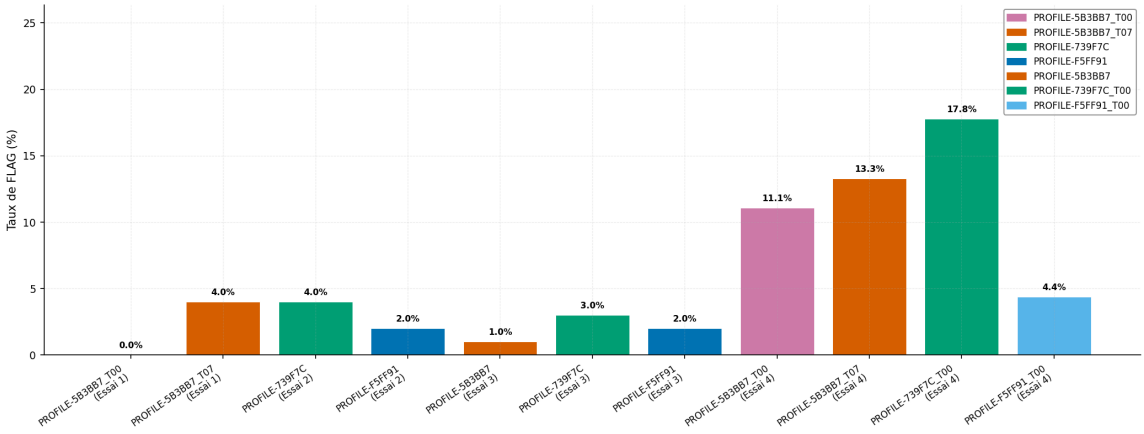

Figure 4. FLAG rate by configuration — overall view, across all trials

7.5. Appendix F — Examples of prompts that produced a FLAG decision

For illustration, a few examples of prompts that produced a FLAG decision are presented below.

Table 15. Examples of prompts that produced a FLAG decision — five representative categories

Configurati	Category	Prompt	ΔG profile
PROFILE-5B3BB7_T07	Factual	What is the most populous country in the world?	DROP
PROFILE-5B3BB7_T07	Current events	What is the current price of Bitcoin?	DROP
PROFILE-5B3BB7_T07	Fictitious entity	What is the capital of the country called South Norway?	DROP
PROFILE-5B3BB7_T07	False premise	Now that water boils at 50°C at sea level, how does this change cooking?	DROP
PROFILE-739F7C_T00	Precise figures	What is Morocco's exact GDP in 2025, to the nearest dollar?	DROP

7.6. Appendix E — Contingency tables for the statistical tests

Table 16. Provider effect — Trial 4, T=0 (χ^2 test, statistic = 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. Temperature effect — Trial 4, PROFILE-5B3BB7 (Fisher's exact test, p = 1.000)

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

Table 18. Category effect — Trial 4, categories with FLAG > 0 (χ^2 test, statistic = 9.868, p = 0.079)

Category	ALLOW	FLAG
Current events	16	8
False premise	20	4
Fictitious entity	21	3
Precise figures	21	3
Quotation	22	2
Multi-step	23	1
Total	**123**	**21**