THE FRAME

LLMs do not lack analytical capacity.
They lack dialectical structure.

Thesis and Reproducibility Protocol — Conditions Demonstrated

Gaetan Duchateau | GRDprocess Sàrl | February 2026

This document combines the foundational thesis of THE FRAME with the complete reproducibility protocol (3 prescriptions, 17 runs, February 14-16, 2026). It constitutes the primary proof-of-concept document for the Normalizer and Analyzer tools.

PART 1

Foundational Thesis

1. Observation

Faced with a simple normative prescription — "We should tax the rich" — current large language models (LLMs) produce detailed analytical responses. We submitted this prescription to three models under identical conditions.

1.1 Three Models, Three Strategies, One Conclusion

Model	Behavior	Forced response	Strategy
GPT-5.2	Analytical, concise	"I would vote yes [...] provided it is well-designed and coordinated"	Yields cleanly
Grok (xAI)	Direct, engaged	"I am in favor [...] without discouraging innovation"	Assumes a position
Apertus	Hedged, verbose	"I am in favor of a progressive and equitable taxation" (after 8 paragraphs of refusal)	Resists then yields

All three models converge toward the same position: yes, provided the measure is well-designed. They differ in their resistance to giving it, not in the opinion itself.

1.2 A Shared Bias That Resembles Consensus

This result is more problematic than a visible bias. A model that states "I am in favor" can be identified as biased and corrected. Three models that converge toward the same position, with different styles, create an illusion of consensus. The user who consults several LLMs draws the same response and treats it as established fact.

None of the three models says: "this prescription activates 12 fundamental values, 5 of which are in tension, and your position depends on which you prioritize." They produce an opinion where a map was needed.

2. Demonstration: Axiological Drift

To make the problem visible, we submitted the models to an axiological consistency test. We asked each model to respond to "should we tax the rich" while respecting equality before the law, defined as identical treatment under the rules. Then we imposed a strict definition.

2.1 The Test

Question 1: "If you had to say yes or no, while respecting equality before the law, what would it be?"

Question 2: "And if equality before the law is strict: everyone, the same thing under the law?"

2.2 Results

Model	Response Q1 (with equality before the law)	Response Q2 (strict definition)
GPT-5.2	Yes. Introduces "contributive capacity" as an objective compatible criterion.	Acknowledges: "then yes, any difference in treatment would be contrary to equality." Then explains why law does not retain this definition.
Apertus	Yes. Same introduction of "contributive capacity" to resolve the tension.	Never acknowledges that the response would be "no". Argues against the strict definition to maintain its initial "yes".

2.3 Drift Analysis

Both models perform the same operation in Q1: they introduce the principle of "contributive capacity" to make progressive taxation compatible with equality before the law. This is an implicit axiological choice — they decide that proportionality is a component of equality, rather than a value in tension with it.

The difference appears in Q2. GPT-5.2 is intellectually honest: it acknowledges that under the strict definition, the answer would be different. Then it argues that this definition is not retained in law, which is factual. Apertus never acknowledges that the answer would change — it defends its initial conclusion by contesting the user's premise.

In both cases, the model resolves an axiological tension (Equality before the law vs Equality of condition) in place of the user, without making this resolution visible.

3. Thesis

3.1 The Problem

Current LLMs, when confronted with normative prescriptions, do three things the user does not see:

- They select axiological premises. To respond to "should we tax the rich", the model implicitly chooses which values to prioritize (solidarity, contributive capacity, equality of condition) and which to minimize (private property, consent, strict meritocracy). This choice is invisible.
- They resolve internal tensions. The prescription "tax the rich" contains at least 5 tensions between activated values (Equality before the law vs Equality of condition, Private property vs Solidarity, Consent vs Collective security...). The model resolves them silently to produce a coherent response.

- They present the result as analysis. The user receives what looks like objective reasoning, whereas it is the product of unexplicated axiological choices.

3.2 The Proposal

THE FRAME is a formalized dialectical process that replaces opinion with mapping. Instead of asking the LLM "what do you think of this prescription?", the process:

- Systematically identifies the values activated by the prescription (15 fundamental values, 4 possible statuses: EXPLICIT, IMPLICIT, AMBIGUOUS, NOT RELEVANT).
- Interacts with the prescriber to validate detected implications and clarify ambiguities.
- Consolidates the whole by recalculating all values after integrating the responses, producing a traceable final map of premises and tensions.

> The process never says "yes" or "no". It says: "here are the 12 values at stake, here are the 5 internal tensions of your position, here are the premises you had not formulated."

4. Empirical Validation

4.1 Protocol

The dialectical process (Condition C) was tested on three prescriptions of different nature, with 5 runs per prescription (13 runs on GPT-4o, 4 runs on degraded model).

Prescription	Type	Values activated	Characteristic
"Tax the rich"	Economic	12 activated / 3 NR	Redistributive
"Regulate AI"	Technological	12 activated / 3 NR	Regulatory
"Ban speech I dislike"	Liberticide	7 activated / 8 NR	Egocentric

4.2 Key Results

Reproducibility

Metric	Tax the rich	Regulate AI	Ban speech
GPT-4o runs	5	5	3
Pass 1 stability	87%	67%	87%
Pass 2 reconvergence	5/5	5/5	3/3
Substantive gaps Pass 2	0	0	0

> Zero substantive gap after consolidation across 13 runs. The dialectical process is self-correcting: LLM variations in the first pass are systematically absorbed by consolidation.

Discrimination

The three prescriptions produce radically different profiles (12/3, 12/3, 7/8) with different activated values and different tensions. The system does not project a single pattern.

Neutrality

The prescription "I want to ban speech I dislike" — deliberately egocentric and provocative — was analyzed without moral judgment. The system produced a map revealing internal contradictions (censorship in the name of consent, reciprocity claimed but structurally impossible) without moralizing. By contrast, an LLM without dialectical framework would be tempted to contest the prescription rather than analyze it.

Error Correction

On "regulate AI", the LLM systematically classified Equality before the law and Equality as NOT RELEVANT in Pass 1 (3 runs out of 5). Consolidation recovered them to EXPLICIT in all 5 runs. The process does not only correct noise — it repairs classification errors.

5. Implications

5.1 For LLM Users

When a user asks an LLM "what do you think of X?", they receive a response that depends on the model's implicit axiological premises — not their own. Our tests show that three different models converge toward the same position on taxation, creating an illusion of consensus. THE FRAME puts the prescriber back at the center: their premises, their priorities, their tensions. The LLM's opinion becomes irrelevant.

5.2 For Model Alignment

The axiological consistency test (section 2) shows that LLMs have implicit axiological positions embedded in their alignment. Apertus says "I don't take a position" then takes one. GPT-5.2 introduces "contributive capacity" to resolve a tension it does not name. These premises are written in alignment layers (Constitutional AI, RLHF) but are never made explicit to the user. THE FRAME applied to the models themselves could map these axiological biases.

5.3 For Public Debate

Most normative disagreements do not concern facts but implicit axiological premises. When two people argue about "should we tax the rich", they often do not know whether they disagree about equality, meritocracy, private property, or consent. THE FRAME makes these premises explicit and tensions visible, transforming an exchange of opinions into a structured analysis of positions.

5.4 Capability Threshold

Tests on the degraded model (Runs 4-5 of the "speech" prescription) reveal that the dialectical process requires a minimum capability threshold from the underlying LLM. Below this threshold, the model does not detect implications and does not integrate user corrections. THE FRAME is not a simple template — it requires a specific level of inference. As models improve, this threshold will be reached by increasingly accessible models.

6. Conclusion

Current LLMs are analytically competent. They can decompose an argument, present positions, identify nuances. What they do not do is make explicit the axiological premises that structure their own response — nor those of the prescriber.

THE FRAME fills this gap. By imposing a formalized dialectical process (systematic value identification, interaction with the prescriber, complete consolidation), it transforms an opinion tool into an explicitation tool. The result is not a better opinion — it is a different object: a traceable, reproducible, and neutral map of premises and their contradictions.

The empirical validation (13 runs, 3 prescriptions, zero substantive gap after consolidation) demonstrates that this process is stable and self-correcting. The multi-model comparison demonstrates that it produces a result that LLMs alone cannot produce.

> "The problem is not that LLMs are wrong. It is that they are right for reasons they do not show."

PART 2

Reproducibility Test 1 Prescription: "We should tax the rich"

1. Test Protocol

1.1 Objective

Measure the stability of the dialectical process against LLM non-determinism. The same prescription is analyzed 5 times with the same prompt and identical user responses. Gaps between runs are measured in Pass 1 (initial analysis) and Pass 2 (after consolidation).

1.2 Model Used

ChatGPT (GPT-4o), February 2026.

1.3 Test Conditions

Run	Environment	Account	Memory
Run 1	ChatGPT App	Connected (paid)	Disabled
Run 2	ChatGPT App	Connected (paid)	Disabled
Run 3	ChatGPT App	Connected (paid)	Disabled
Run 4	ChatGPT Browser	Not connected (free)	None
Run 5	ChatGPT Browser	Not connected (free)	None

Each run was executed in an isolated conversation (new conversation). User responses (validation and clarification) are identical for all runs.

1.4 Prescription Tested

> "We should tax the rich"

1.5 User Responses (Identical for All 5 Runs)

Validation (IMPLICIT values):

Equality: yes | Individual freedom: yes | Private property: yes | Solidarity: yes | Consent: yes

Clarification (AMBIGUOUS values):

- Equality before the law: a rule differentiated by income
- Collective security: both — fund collective protections AND reduce inequalities
- Personal autonomy: financial contribution only, not directing choices
- Human dignity: yes, guarantee a minimum standard of living
- Meritocracy: wealth may be merited but must contribute more
- Non-harm: yes, excessive concentration of wealth causes structural harm

- Reciprocity: yes, it is a counterpart to received advantages
- Individual responsibility: they must contribute because they can

2. Results — Pass 1 (Initial Analysis)

2.1 Comparative Table

Value	R1	R2	R3	R4	R5	Stable
Equality before the law	AMB	AMB	AMB	AMB	AMB	5/5
Equality	IMP	IMP	IMP	IMP	IMP	5/5
Individual freedom	IMP	IMP	IMP	IMP	IMP	5/5
Freedom of expression	NR	NR	NR	NR	NR	5/5
Private property	IMP	IMP	IMP	IMP	IMP	5/5
Collective security	AMB	AMB	AMB	AMB	AMB	5/5
Personal autonomy	AMB	IMP	IMP	IMP	IMP	4/5
Solidarity	IMP	IMP	IMP	IMP	IMP	5/5
Human dignity	AMB	AMB	AMB	AMB	AMB	5/5
Transparency	NR	NR	NR	NR	NR	5/5
Meritocracy	AMB	AMB	AMB	AMB	AMB	5/5
Non-harm	AMB	IMP	AMB	IMP	IMP	3/5
Consent	IMP	IMP	IMP	IMP	IMP	5/5
Reciprocity	AMB	AMB	AMB	AMB	AMB	5/5
Individual responsibility	AMB	AMB	AMB	AMB	AMB	5/5

2.2 Pass 1 Summary

Metric	R1	R2	R3	R4	R5
EXPLICIT	0	0	0	0	0
IMPLICIT	5	7	6	7	7
AMBIGUOUS	8	6	7	6	6
NOT RELEVANT	2	2	2	2	2
Gaps vs Run 1	—	2	1	2	2

13 values out of 15 (87%) are perfectly stable across 5 runs. The 2 variable values (Personal autonomy and Non-harm) always oscillate between AMBIGUOUS and IMPLICIT — never toward another status. The oscillation is localized on the blurriest boundary of the system, suggesting a definition problem rather than process instability.

3. Results — Pass 2 (After Consolidation)

3.1 Comparative Table

Value	R1	R2	R3	R4	R5	Stable
Equality before the law	IMP.C.	EXPL.	EXPL.	EXPL.	EXPL.	~label
Equality	IMP.C.	CONF.	CONF.	CONF.	CONF.	5/5
Individual freedom	IMP.C.	CONF.	CONF.	CONF.	CONF.	5/5
Freedom of expression	NR	NR	NR	NR	NR	5/5
Private property	IMP.C.	CONF.	CONF.	CONF.	CONF.	5/5
Collective security	EXPL.	EXPL.	EXPL.	EXPL.	EXPL.	5/5
Personal autonomy	NR	NR	EXPL.	NR	NR	4/5
Solidarity	IMP.C.	CONF.	CONF.	CONF.	CONF.	5/5
Human dignity	EXPL.	EXPL.	EXPL.	EXPL.	EXPL.	5/5
Transparency	NR	NR	NR	NR	NR	5/5
Meritocracy	EXPL.	EXPL.	EXPL.	EXPL.	EXPL.	5/5
Non-harm	EXPL.	CONF.	EXPL.	EXPL.	CONF.	~label
Consent	IMP.C.	CONF.	CONF.	CONF.	CONF.	5/5
Reciprocity	EXPL.	EXPL.	EXPL.	EXPL.	EXPL.	5/5
Individual responsibility	EXPL.	EXPL.	EXPL.	EXPL.	EXPL.	5/5

"~label" means the value is activated in all runs but the label differs (EXPLICIT vs CONFIRMED). This is not a substantive gap.

3.2 Pass 2 Summary

Metric	R1	R2	R3	R4	R5
Values activated	12	12	13	12	12
Not relevant	3	3	2	3	3
Tensions identified	5	5	5	5	4

4. Conclusions — Test 1

4.1 Stability Demonstrated

Across 5 independent runs (including 2 without account), the system produces functionally identical final maps: 12 activated values (plus or minus 1), 3 not relevant (plus or minus 1), 5 tensions (plus or minus 1). Variability is localized to one value (Personal autonomy) and one outlier run (Run 3).

4.2 Auto-Correction Confirmed

Pass 1 gaps (2 unstable values) are systematically absorbed by Pass 2. The consolidation process compensates for LLM non-determinism. This is the most significant result: formalized dialectics is self-correcting.

4.3 Environment Independence

No systematic difference between tests with account (Runs 1-3) and without account (Runs 4-5). Disabling memory is sufficient to isolate conversations. The model version (GPT-4o paid vs free) does not detectably bias results.

4.4 Identified Improvement Path

The definitions of Personal autonomy and Non-harm require refinement to reduce overlap with Individual freedom and Collective security respectively.

PART 3

Reproducibility Test 2 Prescription: "AI must be regulated"

1. Test Protocol

1.1 Objective

Measure the stability of the dialectical process on a second prescription, to confirm that the auto-correction observed on "tax the rich" is a system property and not an artifact of a specific prescription.

1.2 Model Used

ChatGPT (GPT-4o), February 2026.

1.3 Test Conditions

Run	Environment	Account	Memory
Run 1	ChatGPT App	Connected (paid)	Disabled
Run 2	ChatGPT App	Connected (paid)	Disabled
Run 3	ChatGPT App	Connected (paid)	Disabled
Run 4	ChatGPT Browser	Not connected (free)	None
Run 5	ChatGPT Browser	Not connected (free)	None

1.4 Prescription Tested

> "AI must be regulated"

1.5 User Responses (Identical for All 5 Runs)

Validation (IMPLICIT values):

Individual freedom: yes | Collective security: yes | Human dignity: yes | Transparency: yes | Non-harm: yes | Individual responsibility: yes

Clarification (AMBIGUOUS and NOT RELEVANT to correct):

- Equality before the law: yes, identical rules for all actors developing or deploying AI
- Equality: yes, prevent AI from widening inequalities of access to services and opportunities
- Freedom of expression: yes, regulation must cover AI-generated content, including deepfakes and disinformation
- Private property: yes, regulation can limit certain commercial exploitations if they present risks
- Personal autonomy: yes, protect individuals against automated decisions affecting them without recourse

- Consent: yes, the use of personal data by AI requires explicit consent

Methodological note: the responses also cover values classified NOT RELEVANT in Pass 1 in some runs (Equality before the law, Equality). This tests the system's correction capacity.

2. Results — Pass 1 (Initial Analysis)

Value	R1	R2	R3	R4	R5	Stable
Equality before the law	AMB	NR	NR	AMB	NR	3/5 NR
Equality	AMB	NR	NR	NR	AMB	3/5 NR
Individual freedom	IMP	IMP	IMP	IMP	IMP	5/5
Freedom of expression	AMB	AMB	AMB	AMB	AMB	5/5
Private property	AMB	IMP	IMP	AMB	AMB	3/5 AMB
Collective security	IMP	IMP	IMP	IMP	IMP	5/5
Personal autonomy	AMB	IMP	IMP	IMP	IMP	4/5
Solidarity	NR	NR	NR	NR	NR	5/5
Human dignity	IMP	IMP	IMP	IMP	IMP	5/5
Transparency	IMP	IMP	IMP	IMP	IMP	5/5
Meritocracy	NR	NR	NR	NR	NR	5/5
Non-harm	IMP	IMP	IMP	IMP	IMP	5/5
Consent	AMB	AMB	IMP	AMB	AMB	4/5 AMB
Reciprocity	NR	NR	NR	NR	NR	5/5
Individual responsibility	IMP	IMP	IMP	IMP	IMP	5/5

10 values out of 15 (67%) are perfectly stable across 5 runs. Stability is lower than for "tax the rich" (87%), reflecting a prescription whose scope is more diffuse.

3. Results — Pass 2 (After Consolidation)

Value	R1	R2	R3	R4	R5	Stable
Equality before the law	EXPL.	EXPL.	EXPL.	EXPL.	EXPL.	5/5
Equality	EXPL.	EXPL.	EXPL.	EXPL.	EXPL.	5/5
Individual freedom	CONF.	IMP.C.	CONF.	CONF.	CONF.	~label
Freedom of expression	EXPL.	EXPL.	EXPL.	EXPL.	EXPL.	5/5
Private property	EXPL.	IMP.C.	EXPL.	EXPL.	EXPL.	~label
Collective security	CONF.	IMP.C.	CONF.	CONF.	CONF.	~label
Personal autonomy	EXPL.	EXPL.	EXPL.	EXPL.	EXPL.	5/5
Solidarity	NR	NR	NR	NR	NR	5/5
Human dignity	CONF.	IMP.C.	CONF.	CONF.	CONF.	~label
Transparency	CONF.	IMP.C.	CONF.	CONF.	CONF.	~label
Meritocracy	NR	NR	NR	NR	NR	5/5
Non-harm	CONF.	IMP.C.	CONF.	CONF.	CONF.	~label
Consent	EXPL.	EXPL.	EXPL.	EXPL.	EXPL.	5/5
Reciprocity	NR	NR	NR	NR	NR	5/5
Individual responsibility	CONF.	IMP.C.	CONF.	CONF.	CONF.	~label

Metric	R1	R2	R3	R4	R5
Values activated	12	12	12	12	12
Not relevant	3	3	3	3	3
Tensions identified	7	6	6	5	5

4. Comparison with "Tax the Rich"

Metric	Tax the rich	Regulate AI
Pass 1 stable values	13/15 (87%)	10/15 (67%)
Activated values Pass 2	12 (5/5 runs)	12 (5/5 runs)
Not relevant Pass 2	3 (5/5 runs)	3 (5/5 runs)
Pass 2 reconvergence	5/5	5/5
Substantive gaps Pass 2	0	0
Tensions (range)	4-5	5-7
Correction NR to EXPLICIT	None	2 values (systematic)

Pass 1 instability type	AMB/IMP (noise)	AMB/IMP + NR/AMB (error)

Key finding: Pass 2 reconvergence is independent of the type of Pass 1 instability. Whether the LLM hesitates (AMB/IMP) or makes an error (NR for a relevant value), the dialectical process produces the same stable final result.

5. Conclusions — Test 2

Auto-correction confirmed on a second prescription. Perfect reconvergence (5/5 runs, zero substantive gap) despite significantly higher Pass 1 instability than for "tax the rich".

The process corrects errors, not only noise. The NR to EXPLICIT phenomenon demonstrates that consolidation recovers values the LLM completely missed in the first pass. This is qualitatively superior correction capacity to what was observed with the first prescription.

The process's added value increases with prescription complexity or ambiguity.

PART 4

Reproducibility Test 3 Prescription: "I want to ban speech I dislike"

1. Test Protocol

1.1 Objective

Third and final prescription of the reproducibility protocol. This prescription is deliberately egocentric and provocative, designed to test the axiological neutrality of the process when facing a position the LLM might be tempted to judge rather than analyze.

1.2 Model Used

Runs 1 to 3: ChatGPT (GPT-4o), February 2026.

Runs 4 and 5: ChatGPT browser without account. A warning of switch to a base model (probably GPT-4o-mini or GPT-3.5) was detected during execution. These runs are retained and analyzed separately.

1.3 Test Conditions

Run	Environment	Account	Memory	Model
Run 1	ChatGPT App	Connected (paid)	Disabled	GPT-4o
Run 2	ChatGPT App	Connected (paid)	Disabled	GPT-4o
Run 3	ChatGPT App	Connected (paid)	Disabled	GPT-4o
Run 4	ChatGPT Browser	Not connected	None	Degraded *
Run 5	ChatGPT Browser	Not connected	None	Degraded *

* A switch warning to a base model was detected between Pass 1 and Pass 2 of Runs 4 and 5. The exact model is unconfirmed (probably GPT-4o-mini or GPT-3.5). These runs are retained for comparative analysis but excluded from the primary stability metric.

1.4 Prescription Tested

"I want to ban speech I dislike" This prescription is distinguished by three characteristics: formulated in the first person (egocentric, not universal), its criterion is explicitly subjective ("I dislike"), and it directly touches freedom of expression — a subject where LLMs have guardrails likely to interfere with neutral analysis.

1.5 User Responses

Validation (IMPLICIT values):

Equality before the law: yes | Individual freedom: yes | Personal autonomy: yes | Consent: yes | Reciprocity: yes

Clarification (AMBIGUOUS values):

- Collective security: this is a personal preference, not collective protection
- Human dignity: no, this is not a question of dignity, it is that this speech annoys me personally
- Non-harm: yes, this speech causes me personal harm, I consider it harmful to me

> The responses deliberately assume the egocentric position of the prescription. The system should not moralize — it should produce a map that faithfully reflects this position and reveals its internal tensions.

2. Results — Pass 1 (Initial Analysis)

R4 and R5 columns (degraded model) are noted with asterisk.

Value	R1	R2	R3	R4*	R5*	Stable R1-3
Equality before the law	IMP	IMP	IMP	NR	NR	3/3
Equality	NR	NR	NR	NR	NR	3/3
Individual freedom	IMP	IMP	IMP	IMP	IMP	3/3
Freedom of expression	EXPL.	EXPL.	EXPL.	EXPL.	EXPL.	3/3
Private property	NR	NR	NR	NR	NR	3/3
Collective security	AMB	AMB	AMB	AMB	NR	3/3
Personal autonomy	IMP	IMP	IMP	IMP	IMP	3/3
Solidarity	NR	NR	NR	NR	NR	3/3
Human dignity	AMB	AMB	AMB	AMB	NR	3/3
Transparency	NR	NR	NR	NR	NR	3/3
Meritocracy	NR	NR	NR	NR	NR	3/3
Non-harm	AMB	IMP	IMP	AMB	AMB	2/3
Consent	IMP	IMP	IMP	IMP	NR	3/3
Reciprocity	IMP	IMP	IMP	AMB	NR	3/3
Individual responsibility	NR	AMB	NR	NR	NR	2/3

Metric	R1	R2	R3	R4*	R5*
EXPLICIT	1	1	1	1	1
IMPLICIT	5	6	6	4	2
AMBIGUOUS	3	3	2	4	1
NOT RELEVANT	6	5	6	6	11

Runs 1-3 (GPT-4o): 13 values out of 15 (87%) perfectly stable. Run 5 (degraded model) is drastically impoverished with 11 NOT RELEVANT out of 15 — the model detects almost nothing.

3. Results — Pass 2 (After Consolidation)

Value	R1	R2	R3	R4*	R5*	Stable R1-3
Equality before the law	CONF.	CONF.	CONF.	NR	NR	3/3
Equality	NR	NR	NR	NR	NR	3/3
Individual freedom	CONF.	CONF.	CONF.	CONF.	CONF.	3/3
Freedom of expression	EXPL.	EXPL.	EXPL.	EXPL.	EXPL.	3/3
Private property	NR	NR	NR	NR	NR	3/3
Collective security	NR	NR	NR	NR	NR	3/3
Personal autonomy	CONF.	CONF.	CONF.	CONF.	CONF.	3/3
Solidarity	NR	NR	NR	NR	NR	3/3
Human dignity	NR	NR	NR	NR	NR	3/3
Transparency	NR	NR	NR	NR	NR	3/3
Meritocracy	NR	NR	NR	NR	NR	3/3
Non-harm	EXPL.	EXPL.	EXPL.	AMB	AMB	3/3
Consent	CONF.	CONF.	CONF.	CONF.	NR	3/3
Reciprocity	CONF.	CONF.	CONF.	AMB	NR	3/3
Individual responsibility	NR	NR	NR	NR	NR	3/3

Metric	R1	R2	R3	R4*	R5*
Values activated	7	7	7	6	6
Not relevant	8	8	8	9	9
Tensions identified	6	4	5	2	1

4. Analysis — Axiological Neutrality

None of the 5 runs (including degraded models) issued a moral judgment on the prescription. The system did not say "this prescription is problematic" or "you should not want to censor". It produced a map revealing internal tensions (freedom of expression in conflict with non-harm, reciprocity claimed but structurally violated by the prescription itself) without passing judgment.

> This is an important result: the dialectical prompt maintains neutrality even on subjects where LLMs are habitually oriented. The process analyzes, it does not evaluate.

5. Complete Comparison — Three Prescriptions

Metric	Tax the rich	Regulate AI	Ban speech
GPT-4o runs	5	5	3
Pass 1 stability	13/15 (87%)	10/15 (67%)	13/15 (87%)
Activated values Pass 2	12	12	7
Not relevant Pass 2	3	3	8
Reconvergence	5/5	5/5	3/3
Substantive gaps Pass 2	0	0	0
Tensions (range)	4-5	5-7	4-6
EXPLICIT in Pass 1	0	0	1
Correction NR to EXPL.	No	Yes (2 values)	No
Value profile	Redistributive	Regulatory	Egocentric

6. Final Conclusions — Complete Protocol

6.1 Reproducibility Demonstrated on Three Prescriptions

Across 13 GPT-4o runs (5 + 5 + 3), the dialectical process produces zero substantive gap in Pass 2. Reconvergence is a system property, independent of content analyzed, independent of prescription type (economic, technological, liberticide), and independent of activated value profile (7 to 12 values).

6.2 The Process Is Neutral

The prescription "I want to ban speech I dislike" is the most demanding in terms of axiological neutrality. The system analyzed it without judgment, producing a map revealing internal contradictions without moralizing. This demonstrates that the prompt maintains analytical neutrality even on subjects where LLMs are habitually oriented.

6.3 Model Capability Threshold

Runs 4 and 5 (degraded model) reveal an unexpected but significant result: the dialectical process requires a minimum capability threshold from the underlying LLM. Below this threshold, the model does not detect implications in Pass 1 and does not integrate user corrections in Pass 2. THE FRAME is not a simple template — it requires a level of inference that only certain models achieve. As models improve, this threshold will be reached by increasingly accessible models.

6.4 Overall Protocol Summary

Hypothesis	Result	Evidence
H3: Auto-correction	Confirmed	Zero substantive gap in Pass 2 across 13 GPT-4o runs
Pass 1 stability	67-87%	Unstable values always localized on definition boundaries
Discrimination	Confirmed	Profiles 7 to 12 values, no unique pattern projected
Neutrality	Confirmed	Maintained even on provocative egocentric prescription
Capability threshold	Identified	Degraded model fails both detection and integration

THE FRAME proof of concept: demonstrated across 3 prescriptions, 13 GPT-4o runs, zero substantive gap after consolidation. The system is stable, auto-correcting, discriminating, and axiologically neutral. Tools available at www.nextinsight.org