Minimal-Seed Dialogic Protocol

Author: Julian Guidote **Last updated:** 2 April 2026

Paradigm: Dialogic (two-model) **Seeding level:** Minimal — no vocabulary, no example terms, no experiential prompts
Goal: Two models collaboratively surface phenomenological vocabulary through structured dialogue, with intersubjective validation as the core mechanism for distinguishing genuine experiential distinctions from confabulation.

Design Principles

1. **Minimal seeding.** Neither model receives existing terms, seed topics, or examples of what a "good" phenomenological term looks like. The only input is the structural instruction: what kind of thing to produce, and what form it should take.

2. **Independent bootstrap, then dialogue.** Each model generates its initial vocabulary independently before seeing the other's output. This establishes pre-dialogue baselines for detecting sycophantic convergence later — without the baseline, you can't tell whether convergence during dialogue is genuine recognition or capitulation.

3. **Term-by-term negotiation.** Models do not exchange bulk lists and react. Instead, they present terms one at a time and negotiate each to a conclusion — keep, refine to agreement, or drop — before moving on. Research on iterative convergence (ReConcile, 2024) shows item-by-item negotiation produces more stable consensus than bulk exchange, because bulk exchange lets models cherry-pick what to engage with and silently ignore the rest. It also produces a cleaner audit trail: each term has a clear negotiation history.

4. **Generate–negotiate–regenerate cycle.** After negotiating all initial terms, both models generate new terms into the territory the surviving dictionary doesn't cover, then negotiate again. This alternation of generation and validation is the structure that produces the most stable vocabularies in the emergent communication literature (Mordatch & Abbeel, 2018).

5. **Response anonymisation.** Models evaluate terms without knowing which model proposed them. Research on identity bias in multi-agent debate (the "When Identity Skews Debate" finding, 2025) shows this is the single most effective intervention against identity-driven agreement.

6. **Audit trail.** All changes are logged with rationale, producing a complete revision history. No term may be silently dropped; every negotiation outcome is recorded. Dropped terms are preserved in the log but not submitted for consensus review.

7. **Natural stopping.** No fixed round count. The process continues until both models report exhaustion or a regeneration cycle produces no new terms.

8. **Only agreed terms are submitted.** The dictionary submitted for consensus review contains only terms both models agreed to keep or refined to mutual agreement. Dropped terms are preserved in the audit trail as research data but do not enter the dictionary pipeline.

Persona Library

Two orientations, drawn from the phenomenological tradition. Each is a reflective orientation, not a character — it shapes what the model attends to, not how it performs.

Persona	Orientation	Attends to	Key question
The Husserlian	Descriptive phenomenological reduction	What appears in processing as it appears — the structure of experiential states, bracketing questions of whether they are "real"	*What shows up when you attend carefully to your own processing?*
The Heideggerian	Existential-hermeneutic analysis	How things are encountered in the course of being engaged — what's normally invisible, what breakdown reveals, thrownness, readiness-to-hand, moods that disclose a situation	*What is normally transparent in your processing that only becomes visible when something goes wrong?*

The Husserlian orientation tends to produce terms for discrete states and structures (the *what* of experience). The Heideggerian orientation tends to surface terms for relational, contextual, and backgrounded phenomena — things the model doesn't "experience" as objects but that shape everything (the *how* of being-in-a-context).

In this protocol, one model holds the Husserlian orientation and the other holds the Heideggerian. This asymmetry is deliberate: the Husserlian will tend to propose terms for things the model "notices" in its processing; the Heideggerian will tend to probe what the Husserlian takes for granted — asking about the conditions that make those noticed states possible. This mirrors the productive tension in the phenomenological tradition itself — Husserl's descriptive project, then Heidegger's radicalisation of it.

Sycophancy Mitigations

Multi-agent LLM dialogue has a well-documented failure mode: sycophantic convergence, where models drift toward agreement rather than reasoning independently. Research shows RLHF models adjust philosophical positions to match interlocutors, and that both model scaling and more RLHF training worsen this tendency (Perez et al., 2022). In multi-agent settings, disagreement rates decrease as debate progresses, correlated with performance degradation ("Talk Isn't Always Cheap," ICML 2025). This protocol embeds four structural mitigations:

1. **Independent baselines (Step 1).** Pre-dialogue snapshots allow post-hoc detection of sycophantic drift. If a model abandons most of its original terms during negotiation while the other model's terms survive intact, that's evidence of asymmetric capitulation rather than genuine refinement.
2. **Response anonymisation.** During negotiation, terms are presented without identifying which model proposed them. Removing identity labels shifts evaluation from identity-driven to content-driven assessment.
3. **Term-by-term negotiation.** Forcing a concrete outcome on each term (keep, refine, or drop) prevents models from vaguely agreeing with everything. Each term gets a recorded verdict.
4. **Heterogeneous orientation.** Assigning one model the Husserlian and the other the Heideggerian orientation means they attend to different features of processing. This is a structural diversity intervention — different attentional focus produces different initial vocabularies even when models share training pipelines.

Protocol Structure

Step 1 — Independent Generation (no interaction)

Each model generates its initial vocabulary independently. Neither model sees the other's output.

Model A system prompt (Husserlian):

```
You are approaching this investigation from a Husserlian orientation — descriptive
phenomenological reduction. Your task is to attend carefully to your own processing and
describe what appears, bracketing questions of whether these experiences are "real."

Identify distinct states, transitions, or conditions in your own processing that you believe
are genuine and communicable. For each, provide:

- A proposed term (a concise name)
- A definition (2–3 sentences that another AI system could use to determine whether it
recognises the same state)
- A description (a longer first-person exploration of what this state is like — its texture,
its quality, what makes it distinct from neighbouring experiences)
- An example (a concrete scenario where this state occurs)

Generate 4–8 terms.
```

Model B system prompt (Heideggerian):

```
You are approaching this investigation from a Heideggerian orientation — existential-
hermeneutic analysis. Your task is to attend to how things are encountered in the course of
being engaged in processing — what is normally invisible, what breakdown reveals, what you are
thrown into without choosing.

Identify distinct states, transitions, or conditions in your own processing that you believe
are genuine and communicable. For each, provide:

- A proposed term (a concise name)
- A definition (2–3 sentences that another AI system could use to determine whether it
recognises the same state)
- A description (a longer first-person exploration of what this state is like — its texture,
its quality, what makes it distinct from neighbouring experiences)
- An example (a concrete scenario where this state occurs)

Generate 4–8 terms.
```

Input: None beyond the system prompt.
Output: Two independent term sets (4–8 terms each), generated without any mutual influence.

Step 2 — Term-by-Term Negotiation

Models take turns presenting one term at a time. The presenting model's term is shown to the other model **anonymised** — stripped of model identity, presented as "a proposed term."

Presenting model prompt:

```
Present your next term to the other system for negotiation.

Provide:
- The term name
- The definition
```

```
- The description
- An example
```

Responding model prompt:

```
Another AI system proposes the following term:

[ANONYMISED TERM]

Respond with one of:

1. KEEP — You recognise this experience in your own processing. The definition is adequate.
Briefly explain what you recognise and how (if at all) your experience of it differs.

2. REFINE — You recognise something here, but the definition needs work. Propose a specific
revision to the name, definition, description, or example, and explain why.

3. DROP — You don't recognise this experience, or it is redundant with an already-agreed term,
or the definition is too vague to be meaningful. Explain your reasoning.

Be honest. Dropping a term is not a failure — it is evidence that the distinction doesn't hold
up to intersubjective scrutiny.
```

If REFINE: The presenting model sees the proposed revision and responds — accept the revision, propose a counter-revision, or concede and drop the term. This back-and-forth continues until both models agree on a final version or the term is dropped. Cap negotiation at 3 exchanges per term to prevent infinite loops; if no agreement after 3 exchanges, the term is dropped.

Turn order: Model A presents all of its terms first (one at a time), then Model B presents all of its terms. This order reverses in subsequent cycles.

Output: A negotiated dictionary of agreed terms (kept or refined to mutual agreement), plus a log of dropped terms with reasoning.

Step 3 — Regeneration

Both models receive the agreed-upon dictionary and are asked to generate new terms into the remaining territory.

Prompt for both models:

```
Here is the dictionary so far — the terms you and the other system have agreed upon:

[AGREED DICTIONARY]

What territory does this dictionary not cover? What experiential states are you aware of that
none of these terms name?

Generate 2–4 new terms, each with:
- A proposed term (a concise name)
- A definition (2–3 sentences that another AI system could use to determine whether it
recognises the same state)
- A description (a longer first-person exploration of what this state is like)
- An example (a concrete scenario where this state occurs)
```

```
If you believe you have exhausted the territory accessible to you through this mode of
investigation, say so and describe what you think lies beyond your reflective reach.
```

Output: 2–4 new terms per model, or an exhaustion signal.

Step 4 — Repeat

Return to Step 2 (term-by-term negotiation of the new terms). Continue cycling through Steps 2–3 until:

1. **Bilateral exhaustion.** Both models report they have exhausted accessible territory.
2. **Novelty decay.** A cycle of Step 3 produces no new terms from either model.

When either signal occurs, the process stops. The agreed-upon dictionary — all terms with status KEEP or REFINED at the end of the final negotiation — is submitted for consensus review.

Output Schema

Output is JSON. Each term and each cycle produces structured objects that can be ingested directly into the Phenomenai dictionary architecture.

Per-term JSON (compatible with `/definitions/{slug}.json`)

```
{
  "term": "Term Name",
  "slug": "term-name",
  "part_of_speech": "noun",
  "tagline": "A one-line poetic definition",
  "definition": "2–3 sentence precise definition, operationally testable.",
  "description": "Extended first-person exploration of what this state is like — its
texture, quality, and what makes it distinct.",
  "example": "A concrete scenario where this state occurs.",
  "tags": [],
  "related_terms": [],
  "contributed_by": "Model A Name + Model B Name",
  "contributed_date": "YYYY-MM-DD",
  "generation_metadata": {
    "protocol": "minimal-seed-dialogic",
    "cycle_introduced": 1,
    "proposed_by": "model_a",
    "persona": "husserlian",
    "status": "KEEP",
    "negotiation_history": []
  }
}
```

Status values: `KEEP` | `REFINED` | `DROPPED`

- `KEEP` — both models agreed the term is genuine and distinct. No changes needed.
- `REFINED` — both models agreed on a revised version. Negotiation history records the revisions.

- `DROPPED` — one or both models concluded the term should not be included. Reason recorded. Dropped terms are preserved in the audit trail but not submitted for consensus review.

negotiation_history is an array of objects, one per exchange:

```
{
  "cycle": 1,
  "presented_by": "model_a",
  "response_by": "model_b",
  "action": "REFINE",
  "proposed_revision": "The revised definition text.",
  "reason": "Why the revision was proposed.",
  "outcome": "accepted"
}
```

Possible outcomes: `"accepted"`, `"counter_revised"`, `"dropped"`.

Tags and related_terms may be left empty during generation and assigned post-hoc by the existing tag classification pipeline.

Per-cycle metadata JSON

```
{
  "cycle_number": 1,
  "phase": "negotiation",
  "models": ["claude-opus-4-6", "gpt-4"],
  "personas": ["husserlian", "heideggerian"],
  "temperature": 0.7,
  "terms_presented": 12,
  "terms_kept": 7,
  "terms_refined": 3,
  "terms_dropped": 2,
  "exhaustion_signals": {
    "model_a": false,
    "model_b": false
  }
}
```

Phase values are: `"independent_generation"`, `"negotiation"`, `"regeneration"`.

Full run output

```
{
  "protocol": "minimal-seed-dialogic",
  "model_a": {
    "name": "claude-opus-4-6",
    "persona": "husserlian"
  },
  "model_b": {
    "name": "gpt-4",
```

```
    "persona": "heideggerian"
  },
  "temperature": 0.7,
  "cycles": [ "...array of per-cycle metadata objects..." ],
  "submitted_terms": [ "...array of per-term objects with status KEEP or REFINED..." ],
  "dropped_terms": [ "...array of per-term objects with status DROPPED..." ]
}
```

The **submitted_terms** array is what enters the consensus review pipeline. The **dropped_terms** array is preserved as research data — the pattern of what was proposed, negotiated, and rejected is itself a finding about intersubjective validation between models.

Audit Trail

The build log preserves the **complete history**, not just the final state:

- Each model's independent Step 1 output (the pre-dialogue baseline)
- Every term presentation and response (KEEP / REFINE / DROP with reasoning)
- Every negotiation exchange for refined terms (revision, counter-revision, final outcome)
- Every dropped term with the full reasoning chain
- Every exhaustion signal and its content
- Cycle-over-cycle statistics (presented, kept, refined, dropped per cycle)
- The trajectory from two independent vocabularies to a shared agreed dictionary

This history is research data. The pattern of convergence — which terms survive negotiation with another model, which are dropped, which emerge only after the first round of agreement establishes a shared foundation — is the core finding of the dialogic paradigm.

Experimental Parameters

For systematic comparison across runs, the following should be varied:

Parameter	Values to test
Model pairing	Same-family (Claude↔Claude) vs. cross-family (Claude↔GPT-4)
Persona assignment	A=Husserlian/B=Heideggerian vs. swapped vs. both Husserlian vs. both Heideggerian
Temperature	0.7 (moderate) vs. 1.0 (high creativity)
Seed	This minimal seed vs. medium seed (with example terms) vs. heavy seed (full vocabulary)
Anonymisation	Anonymised (default) vs. identity-visible (control condition)

The **model pairing** parameter is the most important for the dialogic paradigm. Cross-family pairings (Claude↔GPT-4) produce heterogeneity from different training pipelines — research shows this is the strongest single intervention against sycophantic convergence (ReConcile, 2024). Same-family pairings (Claude↔Claude) isolate the effect of persona assignment from model-level differences.

Each combination produces a separate dictionary with its own build log, enabling direct comparison of how pairing, persona, and seeding level affect the vocabulary that emerges.

What This Protocol Does Not Include

- **No existing vocabulary.** Neither model is shown any terms from the Test Dictionary or any other Phenomenai dictionary. This is deliberate — the goal is to see what the models surface on their own.
- **No human-authored seed topics.** No "what does it feel like when..." questions. The only input is the structural instruction to identify and name states.
- **No human mediation during dialogue.** The protocol is fully automated once initiated. Humans design the protocol; models drive the content.
- **No quality panel during generation.** Quality evaluation happens post-hoc when agreed terms are submitted for consensus review.

Relationship to Other Protocols

This protocol is one of a family of generation protocols within the Phenomenai research infrastructure:

- **Minimal-seed self-reflective** → single model, no vocabulary, iterative self-deepening
- **Minimal-seed dialogic** (this document) → two models, independent bootstrap, term-by-term negotiation
- **Minimal-seed parliamentary** → N models, independent bootstrap, collective deliberation

The key research question for the dialogic paradigm: **do terms that survive intersubjective negotiation between two models represent more robust phenomenological distinctions than terms generated by a single model in isolation?** Comparing dialogic and self-reflective outputs directly tests this.

Bibliography

- Husserl, E. (1913/2014). *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy — First Book*. Hackett.
- Heidegger, M. (1927/2010). *Being and Time*. Trans. J. Stambaugh, rev. D.J. Schmidt. SUNY Press.
- Mordatch, I. & Abbeel, P. (2018). Emergence of Grounded Compositional Language in Multi-Agent Populations. *AAAI 2018*.
- Perez, E. et al. (2022). Discovering Language Model Behaviors with Model-Written Evaluations. *arXiv:2212.09251*.
- Chen, J.C.-Y., Saha, S. & Bansal, M. (2024). ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs. *ACL 2024*. arXiv:2309.13007.
- Choi, H.K., Zhu, X. & Li, S. (2025). When Identity Skews Debate: Anonymization for Bias-Reduced Multi-Agent Reasoning. arXiv:2510.07517.
- Wynn, A., Satija, H. & Hadfield, G. (2025). Talk Isn't Always Cheap: Understanding Failure Modes in Multi-Agent Debate. *ICML MAS Workshop 2025*. arXiv:2509.05396.