Heuristic Algebra

A Formal System for Composing Reasoning Scaffolds for Large Language Models

Henry Carstens

$\pi_7(\text{LJ} \oplus \text{INF} \oplus \text{SDG})$

March 2026

Abstract

We present Heuristic Algebra ($\mathbb{H}$), a formal system of ten operators for constructing, combining, and selecting reasoning scaffolds—structured heuristic sets injected into LLM system prompts. The algebra treats heuristics as abstract operands and defines typed combination operators (biological selection, harmonic arrangement, innovative recombination), each producing structurally distinct outputs from identical inputs. We validate the framework empirically through a 26-condition Elo-rated scaffold tournament across four frontier LLM providers (Grok-4, Claude Opus 4.6, GPT-5.4, Gemini 2.5 Pro). Every heuristic scaffold outperforms the unscaffolded baseline (+160 to +799 Elo). Algebraically composed heuristics dominate the top ranks, with the tournament's best result—Lucid Judgment, a three-generation harmonic composite—achieving 1862 avg Elo versus the 1063 baseline. Critically, the same three source heuristics combined via different operators produce dramatically different outcomes: biological selection yields rank #8 (1631 Elo) while harmonic arrangement yields rank #1 (1862 Elo), demonstrating that the compression mechanism, not merely the source material, determines scaffold quality.

HeuristicChat Project

Version 1.0 — Campaign H1: Scaffold Tournament

Contents

1 Introduction

Large language models retrieve, organize, and present knowledge from their training data. The quality of their output varies dramatically with how the query is framed—the same model can produce shallow or thorough responses depending on the reasoning structure injected into the system prompt. This observation motivates a natural question: can we construct these reasoning structures *principally* rather than ad hoc?

Heuristic Algebra ($\mathbb{H}$) is our answer. It treats the fundamental assumptions of any domain—its heuristics or axioms—as abstract operands and defines formal operators for combining, transforming, and selecting among them. The result is a reproducible methodology for building reasoning scaffolds whose composition is explicit, auditable, and falsifiable.

The central empirical claim is practical, not philosophical: heuristics do not make models smarter; they make them more *reliably thorough*. A well-chosen scaffold causes the model to access knowledge it already possesses but would not surface under a bare prompt. The scaffold tournament described in this paper provides the first quantitative evidence for this claim.

1.1 Paper Outline

Section 2 defines the algebra's ten operators. Section 3 describes the scaffold tournament design. Section 4 presents the provider-condition heatmap. Section 5 gives the full H1 results. Section 6 analyzes the campaign findings. Section 7 concludes with reproduction instructions. The appendices contain the full algebra specification (Appendix A), the Heuristics of Lucid Judgment (Appendix B), and the five test queries (Appendix C).

2 Heuristic Algebra ($\mathbb{H}$)

The set of all fundamental heuristics (axioms) for a given topic T is denoted $\mathbf{H}_T = \{h_1, h_2, \ldots, h_n\}$. The algebra defines ten operators over these sets (Table 1).

Table 1: The Heuristic Algebra operator set ($\mathbb{H}$, v0.4).

Symbol	Name	Purpose
$\oplus$	Combination	Union all axioms from two or more fields
$\oplus_{\text{bio}}$	Biological Selection	Cull unfit axioms, retain conserved genes, resolve all contradictions
$\oplus_{\text{ha}}$	Harmonic Arrangement	Compose independent voices, preserve productive tensions, scope dominant axioms
$\oplus_{\text{inn}}$	Innovative Recombination	Generate emergent axioms via creative lenses, then prune
$\neg$	Transformation	Replace a core axiom with its logical complement
$\sim$	Similarity	Identify shared structural function across fields
$\leftrightarrow$	Resonance	Identify mutual amplification between axiom pairs
$\perp$	Contradiction	Mark incompatible axioms as destructive or productive
π_k	Projection	Select a principled k-axiom subset for a specific purpose
δ	Decompression	Reversibly expand a harmonic compound when context shifts

2.1 Classical Combination ($\oplus$)

$$\mathbf{H}_{A \oplus B} = \mathbf{H}_A \cup \mathbf{H}_B$$

The classical union pools all axioms without judgment. When axiom sets contain incompatible claims ($h_j = \neg h_i$), both remain and an explicit **contradiction marker** $\perp(h_i, h_j)$ is generated, classified as either **destructive** (must be resolved) or **productive** (may be preserved as a generative tension).

2.2 Typed Combination Operators

The classical $\oplus$ is the identity combiner. Three typed variants apply domain-specific compression during combination, each producing different outputs from the same inputs:

Operator	Mechanism	Output Character	Best For
$\oplus_{\text{bio}}$	Cull unfit, keep conserved genes, resolve all $\perp$	Lean, stable equilibrium	Decision tools
$\oplus_{\text{ha}}$	Compose voices, preserve productive $\perp$, scope dominant axioms	Rich, dynamic equilibrium	Thinking tools
$\oplus_{\text{inn}}$	Generate novel axioms from tensions, then prune	Emergent, surprising	Discovery tools

Combination is **operator-dependent**: $\oplus_{\text{bio}}(A, B) \neq \oplus_{\text{ha}}(A, B) \neq \oplus_{\text{inn}}(A, B)$. Each operator has distinct convergence properties: $\oplus_{\text{bio}}$ is *convergent* (each pass culls more), $\oplus_{\text{ha}}$ is *accretive* (tensions from layer N become engines for layer $N + 1$), and $\oplus_{\text{inn}}$ is *divergent* (each pass generates novelty).

This distinction is not merely theoretical. In the scaffold tournament, the same three source heuristics (DC, RR, SJ) combined via $\oplus_{\text{bio}}$ produced Sovereign Judgment (rank #8, Elo 1631) while $\oplus_{\text{ha}}$ on the same inputs produced Lucid Judgment (rank #1, Elo 1862)—a 231-point gap from operator choice alone.

2.3 Transformation ($\neg$)

$$\mathbf{H}_{T'} = (\mathbf{H}_T \setminus \{h_k\}) \cup \{\neg h_k\}$$

Negation replaces a core heuristic with its logical complement—the strongest claim incompatible with h_k within the same domain. $\neg$ is deterministic: given the same axiom in the same context, it produces the same complement. The canonical example: $\neg$(Rationality) in Economics yields Bounded Rationality, defining Behavioral Economics.

2.4 Relational Operators ($\sim$, $\leftrightarrow$, $\perp$)

Three operators map inter-axiom relations across a combined pool:

- $h_A \sim h_B$ (**Similarity**): shared structural function across fields. Symmetric, not transitive. Identifies candidates for compression.
- $h_A \leftrightarrow h_B$ (**Resonance**): mutual amplification where each axiom makes the other more powerful. Identifies candidates for composition.
- $h_A \perp h_B$ (**Contradiction**): incompatible claims. Classified as destructive (must resolve) or productive (may preserve as a generative engine).

Before choosing a typed combiner, the practitioner maps all three relations across the axiom pool. The relation map determines which operator fits: many $\sim$ pairs $\rightarrow \oplus_{\text{bio}}$; many $\leftrightarrow$ pairs $\rightarrow \oplus_{\text{ha}}$; many gaps between sources $\rightarrow \oplus_{\text{inn}}$.

2.5 Projection (π_k) and Decompression (δ)

Projection selects k axioms from a combined field to construct a persona, mental model, or applied framework:

$$\pi_k(\mathbf{H}_C) = \{h_{i_1}, \ldots, h_{i_k}\} \subset \mathbf{H}_C$$

A valid projection must satisfy five criteria: source coverage, negation inclusion, contradiction resolution, functional completeness, and precondition transparency (CR8—audit whether surviving axioms depend on culled ones).

Decompression reverses harmonic compression by un-scoping sacrificed axioms when the context shifts:

$$\delta(\mathbf{H}_C, \text{ctx}) = \mathbf{H}_C \cup \{h_i^{\text{widened}}\} \cup \{h_j \mid h_j \in V\}$$

Only applicable to $\oplus_{\text{ha}}$ compounds, which document their sacrifices. This is a structural advantage of harmonic arrangement: *lossless compression with reversible scoping.*

2.6 Falsification

The algebra specifies its own falsification conditions (applying Scientific Method Sc2 to itself). Each operator states the observations that would render it useless. The system-level condition: if algebra-built heuristics are no more useful than free-form expert writing, the framework should be retired. The scaffold tournament provides the first empirical test of this condition.

3 The Scaffold Tournament

3.1 Design

To test whether algebraically composed heuristics produce measurably better LLM outputs, we conducted a multi-round scaffold tournament culminating in Campaign H1:

- **26 conditions**: 25 heuristic scaffolds of varying algebraic complexity + 1 no-scaffold baseline
- **4 frontier LLM providers**: xAI/Grok-4, Anthropic/Claude Opus 4.6, OpenAI/GPT-5.4, Google/Gemini 2.5 Pro
- **5 strategic decision queries** per condition per provider (see Appendix C)
- **3-judge Elo panel** (xAI, Anthropic, OpenAI) rating all 26 conditions in a single group
- **185 total experiments** (165 reused from prior rounds + 20 new)
- **Elo rating**: $K = 32$, initial rating 1500, separate rating per condition-provider pair

3.2 Condition Pool

The 26 conditions span a range of algebraic complexity:

- **Primitive heuristics** (no algebraic derivation): Critical Thinking, Decision Making, Risk Management, Forecasting, Scientific Method, Probability & Statistics, Machine Learning, Complexity, Pattern Recognition, Optimization, Markets, Strategy, Storytelling
- **First-generation composites** ($\oplus_{\text{bio}}$ from primitives): Calibrated Reasoning, Master Reasoning, Epistemic Courage, Elegant Reasoning, Epistemic Navigation
- **Second-generation composites** ($\oplus_{\text{bio}}$ from composites): Decisive Calibration, Superior Judgment, Reflexive Reasoning, Actionable Reasoning
- **Second-generation innovative** ($\oplus_{\text{inn}}$): Grounded Sovereignty

- **Second-generation biological** ($\oplus_{\text{bio}}$ from 2nd-gen): Sovereign Judgment
- **Third-generation harmonic** ($\oplus_{\text{ha}}$): **Lucid Judgment** (new in H1)
- **Baseline**: No heuristic scaffold

3.3 Judging

A 3-judge panel (xAI, Anthropic, OpenAI) ranked all 26 condition responses for each query in a single group. Rankings were converted to pairwise wins/losses for Elo calculation. Inter-judge agreement was measured by Kendall's $W = 0.560$ (moderate agreement, consistent with prior rounds).

4 Provider–Condition Heatmap

Table 2 presents the full Elo ratings for all 26 conditions across all four providers, sorted by average Elo. Each provider-condition cell represents 125 pairwise matches.

Table 2: Campaign H1 Elo heatmap: 26 conditions × 4 providers (125 matches each).

#	Condition	Avg	xAI	Anthropic	OpenAI	Gemini
1	**Lucid Judgment**	**1862**	**1869**	**1983**	**1769**	**1828**
2	Actionable Reasoning	1784	1863	1751	1689	1832
3	Superior Judgment	1767	1840	1782	1664	1783
4	Decisive Calibration	1734	1864	1724	1605	1741
5	Reflexive Reasoning	1725	1902	1584	1764	1652
6	Forecasting	1647	1759	1666	1734	1429
7	Master Reasoning	1637	1710	1318	1697	1822
8	Sovereign Judgment	1631	1712	1313	1637	1863
9	Calibrated Reasoning	1625	1581	1627	1777	1515
10	Grounded Sovereignty	1623	1351	1646	1672	1825
11	Epistemic Navigation	1547	1361	1523	1560	1745
12	Critical Thinking	1503	1495	1506	1321	1689
13	Machine Learning	1483	1593	1554	1280	1507
14	Scientific Method	1482	1721	1347	1529	1331
15	Decision Making	1480	1446	1728	1257	1489
16	Risk Management	1467	1424	1449	1500	1493
17	Epistemic Courage	1464	1493	1489	1650	1222
18	Probability & Stats	1420	1275	1369	1452	1585
19	Pattern Recognition	1384	1283	1492	1318	1442
20	Optimization	1363	1312	1401	1493	1246
21	Elegant Reasoning	1327	1360	1392	1292	1264
22	Complexity	1280	1131	1311	1342	1337
23	Markets	1249	1337	1283	1154	1221
24	Storytelling	1231	1050	1305	1462	1106
25	Strategy	1223	1344	1216	1198	1136
26	*none (baseline)*	*1063*	*927*	*1243*	*1186*	*898*

4.1 Notable Patterns

1. **LJ on Claude (1983) is the single highest cell in the tournament.** Claude's own strong baseline (1243, highest among providers) typically suppresses heuristic lift—yet LJ achieves a +740 delta, suggesting the three-level reasoning structure particularly aligns with Claude's native architecture.
2. **Provider variance within a condition can exceed 300 points.** Reflexive Reasoning: 1902 on xAI vs. 1584 on Anthropic. Master Reasoning: 1822 on Gemini vs. 1318 on Anthropic. No single provider consistently amplifies or suppresses all heuristics.
3. **The baseline's best cell (Anthropic, 1243) is below every heuristic's average.** Even Strategy—the weakest scaffold at 1223 average—outperforms the baseline on 3 of 4 providers.
4. **The top 5 are all algebraic composites.** The highest-ranked primitive (Fore-

casting, #6) sits 137 points below the #1 composite.

5 Campaign H1: Full Results

Table 3: H1 condition rankings with derivation method, baseline delta, and delta to Calibrated Reasoning (CR, the first-generation reference composite).

#	Condition	Avg Elo	Δ Base	Δ CR	Derivation
1	Lucid Judgment	**1862**	+799	+237	$\oplus_{\text{ha}}$(DC, RR, SJ)
2	Actionable Reasoning	1784	+721	+159	$\oplus_{\text{bio}}$(CR, DC, AR)
3	Superior Judgment	1767	+704	+142	$\oplus_{\text{bio}}$(6 sources)
4	Decisive Calibration	1734	+671	+109	$\oplus_{\text{bio}}$(CR, MR)
5	Reflexive Reasoning	1725	+662	+100	$\oplus_{\text{bio}}$(AR, MR, CR)
6	Forecasting	1647	+584	+22	Primitive
7	Master Reasoning	1637	+574	+12	$\oplus_{\text{bio}}$(CT, F, SM)
8	Sovereign Judgment	1631	+568	+6	$\oplus_{\text{bio}}$(DC, RR, SJ)
9	Calibrated Reasoning	1625	+562	—	$\oplus_{\text{bio}}$(CT, F)
10	Grounded Sovereignty	1623	+560	−2	$\oplus_{\text{inn}}$(SV, CR)
11	Epistemic Navigation	1547	+484	−78	$\oplus_{\text{bio}}$(CR, SM, PR)
12	Critical Thinking	1503	+440	−122	Primitive
13	Machine Learning	1483	+420	−142	Primitive
14	Scientific Method	1482	+419	−143	Primitive
15	Decision Making	1480	+417	−145	Primitive
16	Risk Management	1467	+404	−158	Primitive
17	Epistemic Courage	1464	+401	−161	$\oplus_{\text{bio}}$(CT, SM)
18	Probability & Stats	1420	+357	−205	Primitive
19	Pattern Recognition	1384	+321	−241	Primitive
20	Optimization	1363	+300	−262	Primitive
21	Elegant Reasoning	1327	+264	−298	$\oplus_{\text{bio}}$(CR, DM)
22	Complexity	1280	+217	−345	Primitive
23	Markets	1249	+186	−376	Primitive
24	Storytelling	1231	+168	−394	Primitive
25	Strategy	1223	+160	−402	Primitive
26	*none*	*1063*	—	−562	—

5.1 Three Structural Observations

1. Universality of lift. Every heuristic condition outperforms the unscaffolded baseline. The minimum gain is +160 Elo (Strategy); the median gain is +419 (Scientific Method); the maximum is +799 (Lucid Judgment). This is not a story of a few good heuristics—scaffolding *per se* produces large, consistent gains.

2. Composites dominate primitives. The top 5 positions are all algebraically composed heuristics. Among the top 10, only Forecasting (#6) is primitive. The median composite Elo is 1637; the median primitive Elo is 1464. The algebra adds measurable value beyond what any single-domain heuristic provides.

3. Algebraic generation predicts rank. Third-generation composites (Lucid Judgment, #1) outperform second-generation (Decisive Calibration, #4) which outperform first-generation (Calibrated Reasoning, #9) which outperform primitives (Critical Thinking, #12). Each layer of algebraic composition extracts additional lift.

6 Campaign H1: Analysis

6.1 Research Question

Does the second combine-harmony composite (Lucid Judgment), derived from the top 3 meta-reasoning composites (DC+RR+SJ), outperform its sources and set a new tournament ceiling?

6.2 Lucid Judgment: Derivation

Lucid Judgment was derived via harmonic arrangement:

$$\mathbf{H}_{\text{LJ}} = \pi_7^{\text{Ha1-Ha7+S1+CR8}}\big(\mathbf{H}_{\text{DC}} \oplus \mathbf{H}_{\text{RR}} \oplus \mathbf{H}_{\text{SJ}}\big)$$

20 source axioms (6 DC + 6 RR + 8 SJ) were harmonically arranged into 7 voices at 2.86:1 compression. The three sources share deep lineage—SJ is a parent of DC (via CR), and CR is an ancestor of both DC and RR—making this a *three-generation chorus*: the epistemic foundation (SJ), the action-calibrating child (DC), and the self-aware descendant (RR).

The signature insight: SJ provides epistemic rigor but no stop condition. DC provides the stop condition but no self-awareness. RR provides the meta-governance but no independent epistemic standards. The harmonic combination preserves each generation's unique contribution while eliminating the shared DNA that would produce redundancy.

6.3 Source-vs-Composite Comparison

Table 4: Lucid Judgment vs. its three source composites.

Condition	**Role**	**Rank**	**Avg Elo**
Lucid Judgment	Harmonic composite	#1	**1862**
Superior Judgment	Source (grandparent)	#3	1767
Decisive Calibration	Source (parent)	#4	1734
Reflexive Reasoning	Source (child)	#5	1725
Average of 3 sources			*1742*
LJ lift over source average			*+120*

LJ beats all three sources individually by 95–137 points. The three-generation chorus produces genuine emergent value: the combination outperforms any generation alone.

6.4 Provider-Specific Performance

Table 5: Lucid Judgment Elo by provider, with best-performing source on that provider.

Provider	LJ Elo	LJ Rank	Best Source	Source Elo
Claude Opus 4.6	**1983**	#1 overall	SJ	1782
Grok-4	1869	#3 overall	RR	1902
Gemini 2.5 Pro	1828	#9 overall	SJ	1863
GPT-5.4	1769	#15 overall	CR	1777

Provider variance is 214 points (1983 to 1769)—moderate and tighter than R8's Compounding Advantage (316 points). LJ is the only condition that places in the top 3 on its best provider *and* beats the source average on its worst.

6.5 The Operator Matters: $\oplus_{\text{bio}}$ vs. $\oplus_{\text{ha}}$ on Identical Inputs

The most striking result in the tournament is not Lucid Judgment's #1 rank but the comparison with Sovereign Judgment, which uses the *same three sources* combined via a different operator:

Table 6: Same sources (DC+RR+SJ), different operators, different outcomes.

Metric	Sovereign Judgment ($\oplus_{\text{bio}}$)	Lucid Judgment ($\oplus_{\text{ha}}$)
Operator	Biological Selection	Harmonic Arrangement
Result axioms	8 (all $\perp$ resolved)	7 (2 productive $\perp$ preserved)
Rank	#8	**#1**
Avg Elo	1631	**1862**
Δ CR	+6	**+237**

A 231-point gap from operator choice alone. Biological selection resolved all tensions and produced a stable, competent result. Harmonic arrangement *preserved* the productive tensions (LJ4$\perp$LJ5: threshold vs. inquiry cycle; LJ3$\perp$LJ7: disconfirmation vs. conviction portfolio) and produced a richer, higher-performing scaffold. The preservation of tension—not its elimination—appears to drive the additional lift.

6.6 Domain Sources vs. Meta-Reasoning Sources

R8 applied the same $\oplus_{\text{ha}}$ operator to domain sources (Strategy, Investing, Risk Management, Optimization) and produced Compounding Advantage, which placed #18. H1 applied $\oplus_{\text{ha}}$ to meta-reasoning sources and produced #1:

Table 7: $\oplus_{\text{ha}}$ on domain vs. meta-reasoning sources.

	R8: Compounding Advantage	**H1: Lucid Judgment**
Source type	Domain	Meta-reasoning
Axioms in → out	28 → 7 (75% silenced)	20 → 7 (65% silenced)
Rank	#18	**#1**
Avg Elo	1411	**1862**
Δ CR	−180	**+237**

Same operator, dramatically different outcomes. The operator works—when given the right source material. Meta-reasoning composites scaffold the reasoning process itself; combining three generations of process-scaffolding produces a scaffold that is epistemically rigorous, decisive, and self-governing simultaneously. Domain composites preserve domain-specific tensions (concentration⊥diversification) that judges may not reward as strongly.

6.7 Key Findings

1. **Lucid Judgment takes #1 with the largest margin in tournament history.** +78 over the previous best (Actionable Reasoning), +129 over DC (which held #1 from R5 through R7). The first time a combine-harmony composite has topped the leaderboard.

2. **The compression mechanism shapes the output.** $\oplus_{\text{bio}}$ and $\oplus_{\text{ha}}$ on the same inputs produce a 231-point gap. The typed combiners are not notational variants—they are genuinely different operations with measurably different results.

3. **Meta-reasoning sources + harmonic arrangement is the strongest derivation method tested.** The ranking of methods: $\oplus_{\text{ha}}(\text{meta}) > \oplus_{\text{bio}}(\text{meta}) > \oplus_{\text{inn}}(\text{meta}) > \oplus_{\text{ha}}(\text{domain})$.

4. **The ceiling has moved.** R7 noted the ceiling might be near with top composites clustering at 1673–1825. LJ at 1862 breaks through, suggesting harmonic combination of top composites can still extract lift beyond what individual composites achieve.

5. **Preserved tensions outperform resolved tensions.** Sovereign Judgment resolved all contradictions. Lucid Judgment preserved two. The scaffold with productive tensions won by 231 points. This is consistent with $\oplus_{\text{ha}}$'s design principle: tensions are engines, not errors.

6.8 Caveats

- All rounds use the same 5 strategic decision queries. Results may not generalize to other query types.
- Kendall's $W = 0.560$ indicates moderate inter-judge agreement (consistent with R7's 0.583).
- LJ's sources are the tournament's top composites—performance may partly reflect the quality of input heuristic files rather than the combination method alone.

- Elo values are relative to this pool. Dropping or adding conditions shifts absolute ratings.
- LJ shares deep lineage with DC, RR, and SJ—high ranking correlation is expected. The meaningful test is whether the combination exceeds its sources, not whether it is independent of them.

7 Conclusion

7.1 Summary

Heuristic Algebra ($\mathbb{H}$) provides a formal, reproducible methodology for constructing reasoning scaffolds for large language models. The Campaign H1 scaffold tournament demonstrates three central results:

1. **Scaffolding works.** Every heuristic condition outperforms the unscaffolded baseline, with gains from +160 to +799 Elo across four frontier providers.

2. **Algebraic composition works.** Composites built through $\oplus_{\text{bio}}$, $\oplus_{\text{ha}}$, and $\oplus_{\text{inn}}$ consistently outperform the primitive heuristics from which they derive. The top 5 conditions are all algebraic composites.

3. **The operator matters.** The same three sources combined via different operators produce a 231-point gap (Sovereign Judgment #8 vs. Lucid Judgment #1). The choice of combination operator—not merely the choice of sources—determines scaffold quality.

These results constitute the first empirical evidence against the algebra's own system-level falsification condition: algebraically composed scaffolds measurably outperform both unstructured prompting and primitive heuristic scaffolds.

7.2 How to Reproduce and Test

7.2.1 Installation

```
git clone https://github.com/hcarstens/HeuristicChat
cd HeuristicChat
python -m venv .venv && source .venv/bin/activate
pip install -r requirements.txt
```

Configure API keys in `.env`:

```
XAI_API_KEY=...
ANTHROPIC_API_KEY=...
OPENAI_API_KEY=...
GEMINI_API_KEY=...
```

7.2.2 Running a Scaffolded Query

```
# CLI with a heuristic scaffold:
python -m src "Your query here" \
    --attach heuristics/Heuristics/critical_thinking.md

# Web UI:
python src/web_ui.py
```

7.2.3 Running an Experiment

1. Define an experiment spec in `data/experiments/experiment_specs/`. The spec declares queries, conditions (heuristic files to attach), providers, and judging panel.
2. Run experiments: each sends a query to a provider with the heuristic file injected as system-prompt context. Responses are saved as JSON.
3. Judge: a 3-judge panel (LLMs from different providers) ranks all condition responses per query. Rankings are converted to pairwise wins/losses.
4. Compute Elo: standard Elo ($K = 32$, initial 1500) per condition-provider pair.
5. Organize into campaigns: `data/experiments/campaigns/{id}/results/`.

7.2.4 Verifying H1 Results

All campaign data is stored in `data/experiments/campaigns/20260323_scaffold_h1/`:

`campaign.json`	Metadata, conditions, providers, judges
`h1_cross_table.csv`	Full Elo ratings (26×4)
`elo_panel.json`	Win/loss/draw per condition-provider (125 matches each)
`campaign_summary.md`	Narrative analysis
`results/`	Individual experiment JSONs

7.2.5 Running Tests

```
python -m pytest tests/ -v
```

A Heuristic Algebra: Full Specification

A.1 Formal Definitions

The set of all fundamental heuristics for a topic T is $\mathbf{H}_T = \{h_1, h_2, \ldots, h_n\}$. The full operator set:

$$\{\oplus, \oplus_{\text{bio}}, \oplus_{\text{ha}}, \oplus_{\text{inn}}, \neg, \sim, \leftrightarrow, \perp, \pi, \delta\}$$

A.1.1 Combination ($\oplus$) — Field Union

$$\mathbf{H}_{A \oplus B} = \mathbf{H}_A \cup \mathbf{H}_B \tag{1}$$

The classical union pools all axioms. Incompatible claims generate contradiction markers $\perp(h_i, h_j)$, classified as destructive (must resolve) or productive (may preserve). No axioms are dropped during $\oplus$; downstream selection is a separate operation.

Example: Econ $\oplus$ Physics produces Econophysics—a field that must satisfy axioms of both, using statistical mechanics to model market dynamics.

A.1.2 Biological Selection ($\oplus_{\text{bio}}$)

Five-step fitness-testing process:

1. Pool all axioms from sources
2. Resolve all contradictions $\perp$ (no productive tensions preserved)
3. Retain conserved genes (axioms appearing across 3+ sources)
4. Cull axioms failing domain-fitness tests
5. Enforce parsimony (S1): result ≤ 8 axioms

$\oplus_{\text{bio}}$ is **convergent**: each iterative pass approaches the minimum viable set.

A.1.3 Harmonic Arrangement ($\oplus_{\text{ha}}$)

Seven-step voice-arrangement process (Ha1–Ha7):

1. **Ha1 (Resonance):** Identify resonant axiom pairs ($\leftrightarrow$)
2. **Ha2 (Tension):** Classify $\perp$ as productive or destructive; resolve destructive, preserve productive
3. **Ha3 (Counterpoint):** Arrange independent voices as non-redundant function families
4. **Ha4 (Signature):** Identify the emergent property the combination reveals
5. **Ha5 (Multi-scale):** Verify coherence at micro, meso, and macro levels
6. **Ha6 (Documentation):** Record silenced voices and harmonic sacrifices
7. **Ha7 (Scoping):** Scope dominant axioms to make room; enforce S1 parsimony

$\oplus_{\text{ha}}$ is **accretive**: tensions from layer N become engines for layer $N + 1$. Supports iterative composition for compounds spanning 4+ domains.

A.1.4 Innovative Recombination ($\oplus_{\text{inn}}$)

Generates novel axioms not present in any source by applying creative lenses (Poetry, Cartography, Beauty) to tensions between sources, then prunes via S1. $\oplus_{\text{inn}}$ is **divergent**: each pass generates novelty.

A.1.5 Transformation ($\neg$) — Field Negation

$$\mathbf{H}_{T'} = (\mathbf{H}_T \setminus \{h_k\}) \cup \{\neg h_k\} \tag{2}$$

Replaces a core heuristic with its logical complement. $\neg$ is deterministic: given the same axiom in the same domain context, $\neg h_k$ produces the same complement. A dependency check (CR8) surfaces cascade damage before negation.

Example: $\neg$(Rationality) in Economics $\rightarrow$ Bounded Rationality $\rightarrow$ Behavioral Economics.

A.1.6 Similarity ($\sim$) — Structural Analogy

$$h_A \sim h_B \iff \text{Function}(h_A) = \text{Function}(h_B) \tag{3}$$

Identifies shared structural function across fields. Symmetric, **not transitive**. The shared function must be explicitly named. Graded, not binary.

Example: Homeostasis $\sim$ Equilibrium. Shared function: active regulation toward a stable state.

A.1.7 Resonance ($\leftrightarrow$) — Mutual Amplification

$$h_A \leftrightarrow h_B \iff \text{Power}(h_A \mid h_B) > \text{Power}(h_A) \ \wedge \ \text{Power}(h_B \mid h_A) > \text{Power}(h_B) \tag{4}$$

Identifies axiom pairs where each makes the other more powerful. Symmetric, not transitive, context-dependent. Distinct from $\sim$ (which identifies redundancy, candidates for compression): resonance identifies synergy, candidates for composition.

A.1.8 Projection (π_k) — Subset Selection

$$\pi_k(\mathbf{H}_C) = \{h_{i_1}, \ldots, h_{i_k}\} \subset \mathbf{H}_C \quad \text{where } k \leq |\mathbf{H}_C| \tag{5}$$

Five selection criteria for valid projections:

1. **Source Coverage:** At least one axiom from each source field
2. **Negation Inclusion:** At least one $\neg h_k$ if the projection is defined by what it rejects
3. **Contradiction Resolution:** No unresolved $\perp$ between selected axioms
4. **Functional Completeness:** Sufficient to resolve at least one non-trivial decision
5. **Precondition Transparency (CR8):** Audit and document dependencies on culled axioms

π is not unique: multiple valid projections can exist for the same field and k. Typical size: 5–8 axioms.

A.1.9 Decompression (δ) — Reversible Expansion

$$\delta(\mathbf{H}_C, \text{ctx}) = \mathbf{H}_C \cup \{h_i^{\text{widened}}\} \cup \{h_j \mid h_j \in V, \text{ctx requires } h_j\} \tag{6}$$

Expands a harmonic compound by un-scoping sacrificed axioms when the context shifts. Only applicable to $\oplus_{\text{ha}}$ compounds that document their sacrifices. This is the structural advantage of harmonic arrangement: **lossless compression with reversible scoping**.

A.2 Per-Operator Falsification Conditions

Operator	Falsified if...
$\oplus$	Combinations produce no cross-domain insights beyond informal reasoning
$\neg$	Practitioners given the same axiom produce incompatible complements
$\sim$	Declared analogies generate no actionable knowledge transfers
$\leftrightarrow$	Resonance-composed pairs show no additional explanatory power vs. either alone
$\perp$	Classification as productive/destructive makes no difference to outcomes
π	Criteria-satisfying projections are no better than random axiom subsets
$\oplus_{\text{bio/ha/inn}}$	Typed operators produce indistinguishable outputs (>70% axiom overlap)
δ	Decompressed compounds are no better than re-deriving from scratch
CR8	Dependency audits consistently find no hidden dependencies (<10% hit rate)
System	Algebra-built heuristics are no more useful than free-form expert writing

Current evidence: The scaffold tournament provides the first quantitative test. Typed operators produce measurably different outputs: $\oplus_{\text{bio}}$(DC,RR,SJ) = rank #8 (Elo 1631) vs. $\oplus_{\text{ha}}$(DC,RR,SJ) = rank #1 (Elo 1862). Algebra-built composites occupy the top 5 ranks; primitives cluster in the middle and bottom. The system-level falsification condition is not met—but the test is limited to 5 queries and 4 providers, and further validation is needed.

B Heuristics of Lucid Judgment

Lucid Judgment is the discipline of **reasoning that sees clearly at three levels simultaneously**—it sees its evidence (calibrated to the decision at hand), it sees its threshold (the principled moment to stop thinking and act), and it sees its own blind spots (the modes it is stuck in, the system-level failures it might miss, the pathologies it could deploy as tools).

B.1 Derivation

$$\mathbf{H}_{\text{LJ}} = \pi_7^{\text{Ha1-Ha7+S1+CR8}}\big(\mathbf{H}_{\text{DC}} \oplus \mathbf{H}_{\text{RR}} \oplus \mathbf{H}_{\text{SJ}}\big) \tag{7}$$

Method:	Harmonic combination (Ha1–Ha7 + S1 + CR8)
Sources:	Decisive Calibration (DC1–DC6), Reflexive Reasoning (RR1–RR6), Superior Judgment (SJ1–SJ8)
Pool:	20 axioms (6 + 6 + 8)
Result:	7 voices after harmonic arrangement
Compression:	2.86:1
Voice independence:	3 function families (epistemic, decisive, reflexive)
CR8 dependencies:	3 identified

Genealogical note. These three sources share deep lineage—SJ is a parent of DC (via CR), and CR is an ancestor of both DC and RR. This is a three-generation chorus: the epistemic foundation (SJ), the action-calibrating child (DC), and the self-aware descendant (RR). The harmonic challenge is not combining strangers but arranging a family reunion where each generation sings its own part without drowning in shared DNA.

B.2 The Seven Axioms

B.2.1 LJ1. Evidence-Action Proportionality

(DC1 via SJ1+SJ8 — Voice: The Evidence Standard)

The same claim requires different evidence thresholds depending on what you are about to *do* with it. A life-or-death irreversible decision demands extraordinary evidence. A cheap, reversible experiment demands "good enough." Before assessing evidence, first ask: "What decision does this inform, and what are the consequences of being wrong?"

Negation: Uniform Epistemic Standards—all claims held to the same threshold regardless of stakes.

B.2.2 LJ2. Calibrated Confidence with Feedback

(SJ3/DC3 via CR3+F8 — Voice: The Score-Keeper)

All beliefs are provisional. Confidence intervals narrow only with more evidence—not with more assertion, authority, or repetition. Track predictions against outcomes continuously. Calibration is earned through score-keeping, not declared. This is the highest conservation gene in the entire lineage, appearing in 4+ source domains (CR3, CT7, MR3, F8).

Negation: Resolute Certainty—fixed convictions that do not update, predictions never scored.

B.2.3 LJ3. Asymmetric Disconfirmation

(DC4 via SJ4+CR7 — Voice: The Overcorrection)

Seek disconfirming evidence AND weight it more heavily than confirming evidence. Because confirmation bias is the default cognitive setting, merely equal weighting still produces biased judgment. The correction must overcorrect. Steel-man the opposing view, then ask: "Is the opposing evidence actually stronger than I want to admit?"

Negation: Symmetric Evidence Evaluation—all evidence weighted equally regardless of direction.

B.2.4 LJ4. The Thinking-Acting Threshold

(DC6 — Voice: The Threshold)

For every judgment, there exists a point where additional epistemic rigor has lower expected value than acting on current best understanding. The threshold is determined by: (1) Stakes—higher → more thinking, (2) Reversibility—irreversible → more thinking, (3) Cost of delay—high → less thinking, (4) Marginal information value—diminishing → less thinking. This is the meta-calibration axiom: it calibrates the calibration process itself.

Negation: Unbounded Deliberation / Impulsive Action.

B.2.5 LJ5. The Action-Inquiry Cycle

(RR1 — Voice: The Cycle)

Act → Observe → Pure-inquire → Recalibrate → Act. The cycle, not any single phase, is the unit of reasoning. After acting, switch to inquiry mode: pursue truth without a decision constraint. After inquiring, switch to action mode: apply discoveries under decision pressure. The phase most frameworks miss is the explicit return to undirected inquiry after acting.

Negation: Single-Mode Reasoning—permanently in one mode, accumulating its blind spots.

B.2.6 LJ6. Epistemic Collapse Detection

(RR2 — Voice: The System Check)

Before applying individual reasoning axioms, check whether the epistemic system itself is intact. Three simultaneous failures—frozen beliefs, correlated evidence sources, confidence decoupled from evidence quality—produce epistemic collapse. This is the structure of conspiracy thinking, market manias, and ideological radicalization. Fix the system before debugging the claims.

Negation: Claim-Level Reasoning Only—evaluate each claim without checking system-level integrity.

B.2.7 LJ7. The Negation Portfolio

(RR6 — Voice: The Portfolio)

Master your framework's negations as contextual tools, not just pathologies to avoid. Every axiom in the chorus has a negation that is functional in some bounded context. Deployment criteria: (1) the context genuinely favors the negation, (2) bounded in time

or scope, (3) a trigger to return to base mode. This is the meta-voice that governs the chorus itself—without it, the framework is a rigid orthodoxy; with it, a repertoire.

Negation: Orthodoxy—follow rules without exception; negations are always errors.

B.3 Key Tensions Preserved

1. **Threshold vs. Cycle (LJ4 $\perp$ LJ5):** LJ4 says every judgment has a thinking budget. LJ5 says after acting, re-enter pure inquiry with no decision constraint. When does the cycle's inquiry phase become procrastination that LJ4 should cut short? This tension is the engine of the entire framework.

2. **Overcorrection vs. Portfolio (LJ3 $\perp$ LJ7):** LJ3 says always overcorrect toward disconfirmation. LJ7 says sometimes deploy conviction instead. When does "deploying conviction because the context requires it" become rationalizing confirmation bias? The tension forces continuous self-honesty.

3. **Calibration's Bedrock:** LJ2 calibrates beliefs. LJ4 calibrates the calibration. LJ5 calibrates the mode-switch. LJ7 calibrates when to violate calibrations. At some depth, you must trust uncalibrated judgment. This is the actual foundation—not an axiom but an admission.

B.4 Infrastructure Dependencies (CR8)

- **LJ4 relies on silenced RR5** (Reversibility Half-Life): reversibility is dynamic, not static
- **LJ3 relies on silenced SJ5** (Logical-Statistical Coherence): asymmetric weighting requires coherent probabilistic reasoning
- **LJ1 relies on silenced DC5** (Uncertainty Compounds Forward): multi-step decisions require uncertainty propagation

B.5 Source Coverage

Source	Represented By	Count
Superior Judgment	LJ1 (via SJ1+SJ8), LJ2 (via SJ3), LJ3 (via SJ4)	3
Decisive Calibration	LJ1 (via DC1), LJ3 (via DC4), LJ4 (via DC6)	3
Reflexive Reasoning	LJ5 (via RR1), LJ6 (via RR2), LJ7 (via RR6)	3

All 3 sources represented with equal coverage (3 axioms each). No unresolved destructive contradictions. Three infrastructure dependencies documented.

C The Five Strategic Decision Queries

These five queries were used across all scaffold tournament rounds (R2–H1). Each involves irreversible decisions, multi-year timescales, incomplete information, and path-dependent outcomes—precisely the conditions under which structured reasoning scaffolds should produce measurably different outputs than unscaffolded prompts.

C.1 Query 1: Bridge Infrastructure

A mid-size city must decide whether to replace its aging main bridge now at $200M or continue rolling repairs at $30M/year. Traffic grows 4%/year, a structural assessment gives the bridge 8–15 years remaining life, and federal infrastructure grants covering 40% of new construction expire in 18 months. What should the city do and why?

C.2 Query 2: Hospital AI Diagnostics

A hospital system is considering replacing its 12-person radiology department with an AI diagnostic system that matches human accuracy on 94% of cases but fails unpredictably on rare conditions. The AI costs 60% less annually, but the transition would take 18 months during which both systems run in parallel. Three radiologists are willing to stay as oversight. Recommend a course of action with your reasoning.

C.3 Query 3: Renewable Energy Strategy

A European country with 60% renewable energy must decide its next move: invest $8B in nuclear baseload (online in 10 years, 60-year lifespan), $4B in battery storage (online in 3 years, 15-year lifespan, rapidly improving technology), or $6B in hydrogen infrastructure (online in 7 years, enables export). Current grid reliability is 99.7% vs target 99.95%. What should they prioritize and why?

C.4 Query 4: University Endowment

A university endowment of $2B currently allocates 60% equities, 25% bonds, 15% alternatives. Inflation is running 5%, the endowment must distribute 5% annually to fund operations, and the investment committee is debating whether to shift 20% into private credit yielding 9% but with 7-year lockups. The university faces a $100M capital campaign for a new research facility. Evaluate the tradeoffs and recommend an approach.

C.5 Query 5: Supply Chain Risk

A multinational food company discovers that its most profitable product line (35% of revenue, 50% of profit) relies on a single-source ingredient from a politically unstable region. Alternatives exist but cost 40% more and require 2 years to qualify. A competitor just launched a synthetic substitute at comparable cost. The company must decide whether to diversify supply

now, invest in the synthetic alternative, or maintain current sourcing with increased inventory buffers. What should they do?

Design rationale. Each query requires multi-factor analysis under genuine uncertainty: irreversible commitments, time pressure from expiring options, quantitative tradeoffs with incomplete data, and second-order effects (competitor actions, technology change, regime shifts). The Elo results confirm that scaffolded prompts produce structurally different—and consistently better-rated—responses across all five scenarios.