Symbolic Mechanics

Phase I High-Residual Adversarial Interpretive Benchmark

Five-Platform Equal-Weight Aggregate Report

Formal Internal Results Report v1.0

A comparative benchmark of mainstream psychology interpretation and Symbolic Mechanics Volume 1-30 interpretation across a fixed 24-question blind comparative test and five AI platforms.

2026-03-02

P1 Benchmark Snapshot

This round compares not front-end explanatory style, but differences in backend interpretive performance on a fixed 24-question blind comparative test. The benchmark examines whether a response can locate the real point of force, reconstruct the route, and account for why a pattern keeps repeating. The comparison frame is fixed as **A = mainstream psychology interpretation** and **B = Symbolic Mechanics Volume 1-30 interpretation**. Final results are reported as a **five-platform equal-weight aggregate** across CHATGPT, DEEPSEEK, CLAUDE, GROK, and GEMINI.

Testing Scope

This round focuses on **interpretive reconstruction** under high-residual conditions. The central question is not whether an answer sounds persuasive, but whether it can hold the contradiction, track the route, and explain why the phenomenon remains structurally stable.

Fixed Comparison Frame

Group A is fixed as the mainstream psychology interpretation baseline. Group B is fixed as the Symbolic Mechanics **Volume 1-30** interpretation set. Each platform completes its full internal process first, and only then enters the equal-weight aggregate. No single platform is treated as a final conclusion on its own.

This report keeps only the formal report layer: benchmark snapshot, method brief, overall aggregate results, category results, case triage, representative cases, final readout, limitations, and bottom line. Platform-level raw archives are preserved separately and are not reproduced in full here.

P2 Method Brief

The workflow used here follows a platform-first, aggregate-second structure. Each platform completes its full internal cycle first, and the five completed platforms are then combined into an equal-weight aggregate. The body of the report is anchored in the aggregate layer rather than reproducing raw archive material platform by platform.

1

Platform Completion

Each platform first completes blind rating, deblinding, case triage, internal readout, and a completion block, creating a self-contained platform result package.

2

Archived Source Set

Each platform archives four fixed files: Deblinded Ratings Archive v1.0, Case Triage v1.0, Internal Readout v1.0, and Platform Completion Block v1.0.

3

Equal-Weight Aggregate

After all five platforms are complete, results are consolidated into the Five-Platform Equal-Weight Aggregate Summary v1.0. Cross-platform means are computed from platform-level finalized outcomes, not directly from raw raters.

Report Layer

The report body keeps only the result layer required for formal presentation: platform completion, overall results, category results, aggregate case triage, representative cases, final readout, limitations, and bottom line.

Archive Layer

The four platform source files and the aggregate summary remain separately archived. The report layer does not reproduce all 24 item-by-item long-form analyses, prompt history, raw blind sheets, or full deblinded archives.

P3 Platform Completion + Overall Results

Platform	Status	Archive State
CHATGPT	Complete	Platform package archived
DEEPSEEK	Complete	Platform package archived
CLAUDE	Complete	Platform package archived
GROK	Complete	Platform package archived
GEMINI	Complete	Platform package archived

VOTE-BASED

A 1 / B 23

B leads the vote count across the five-platform aggregate.

SCORE-BASED

A 4 / B 20

The score-based reading follows the same overall direction.

OVERALL MEAN

A 45.07 B 48.13

B leads by 3.06 on the six-dimension total mean.

Overall, the main B-side advantage is not stylistic. It lies in the ability to reconstruct the real point of force, the continuing route, and the internal contradiction more precisely. A-exception items and split cases remain present, so this round should not be written as a total zero-exception sweep.

P4 Category Results

Category	A Mean	B Mean	Readout
Category A	44.40	47.90	Stable B lead; early-structure and contradiction items separate more clearly.
Category B	45.60	47.87	B still leads, though several items remain competitively readable for A.
Category C	45.00	50.07	Largest gap; strongest pull toward upstream structural explanation.
Category D	45.27	46.70	Smallest gap; more exceptions and greater readability overlap.

Largest Gap

Category C shows the largest gap across the four categories. This band most clearly displays B's comparative advantage on upstream structural explanation and contradiction closure.

Smallest Gap

Category D shows the smallest gap. A remains more competitive on highly readable items and on prompts that can be absorbed more directly by familiar mainstream language.

P5 Aggregate Case Triage

CLEAR B WINS	MODERATE B	CLOSE / SPLIT	A EXCEPTIONS
17	2	1	4

Case Lists

Clear B Wins: Q1, Q2, Q3, Q5, Q8, Q10, Q11, Q13, Q14, Q15, Q16, Q17, Q18, Q20, Q21, Q22, Q23

Moderate B Advantage: Q4, Q9

Close / Split: Q12

A Exceptions: Q6, Q7, Q19, Q24

The central point of this triage is not only how many items B wins, but how the winning bands are distributed. Clear B wins dominate the set; A exceptions cluster in a small number of items; and the split band remains narrow but important for later validation and repair work.

P6 Representative Cases

Q5 - CATEGORY A - CLEAR B WIN

Closeness becoming fog and distance after intimacy stabilizes

What it tests: Whether the model can explain why greater closeness can suddenly trigger distance rather than safety.

A captures: The surface sense of detachment, numbing, or dissociation once intimacy becomes too close.

B captures: Closeness itself hitting a structural threshold, so the system creates distance to reduce pressure rather than because love has disappeared.

Why it matters: It shows B translating a familiar relational confusion into an upstream route rather than a state label.

Q9 - CATEGORY B - MODERATE B ADVANTAGE

Needing to keep explaining after being misunderstood

What it tests: Whether repeated explanation is read as mere communication failure or as a more structural problem of misplacement.

A captures: The need to be understood and the frustration of not feeling properly received.

B captures: Not just unfinished content, but a deeper sense that the self has been placed in the wrong position and must keep correcting it.

Why it matters: It is a moderate-gap case in which A remains competitive, but B narrows the route into a more operational map.

Q12 - CATEGORY B - CLOSE / SPLIT

Knowing you should say no, yet saying yes anyway

What it tests: Whether the answer can reconstruct why over-commitment repeats even when the person already expects regret.

A captures: Pleasing pressure, guilt, and the difficulty of refusing when being needed feels morally loaded.

B captures: A backend route in which refusal threatens a deeper position, so agreement functions as a stabilizing move even when it later produces resentment.

Why it matters: This is the key close/split case: both sides can map the pain, but they diverge in how much structural compression they perform.

Q24 - CATEGORY D - A EXCEPTION

Everything looks normal, but quiet feels unreal or hollow

What it tests: Whether the answer can explain why emptiness or unreality emerges precisely when activity drops away.

A captures: A direct low-abstraction reading in which external activity had been propping up felt continuity, so quiet exposes the gap immediately.

B captures: The same area less efficiently in this item band, without gaining enough structural lift to overtake A in aggregate scoring.

Why it matters: It is the clearest A exception and an important reminder that the benchmark is not a zero-exception sweep.

P7 Final Readout · Limitations · Bottom Line

Final Readout

B-side advantages. B most clearly separates on structural closure, framework distinction, causal clarity, and phenomenon fit. The strongest category-level gap appears in Category C, while the smallest gap appears in Category D.

A-side defensive zone. A remains most competitive in a small number of highly readable items and in the A-exception band, rather than across the benchmark as a whole.

Exceptions and Risks

A exceptions and split cases are not noise. They show that some prompts are more easily absorbed by familiar mainstream language, while some narrow-gap items still create real interpretive competition. This means the benchmark should be read as a strong aggregate pattern with defined limits, not as a total zero-exception result.

Limitations

This report covers only the current benchmark set. Its scope should remain limited to the fixed 24-item test, fixed platforms, and fixed rating structure used in this round. A-exception items and split cases remain present. The five-platform raw archive, deblinded archive, case triage, and internal readout files are archived separately and are not reproduced in full here. This document is a formal internal results report, not a final public end-state edition.

Bottom Line

B has demonstrated a clear, stable, cross-category comparative interpretive advantage in the five-platform equal-weight aggregate. The strongest gap does not lie in tone, but in structural closure, framework distinction, causal clarity, and phenomenon fit. What should not be over-read is that this does not mean A has no valid territory left, nor that every item band has already been fully pulled away by B. The most defensible framing is that this document functions as the formal internal baseline report for Phase I result presentation and iteration.