Z-Bench v1.0 Whitepaper — Appendix Edition

This appendix expands the Z-Bench v1.0 Whitepaper with formal descriptions, expected failure modes, and AGI-co-agency reasoning criteria for each benchmark. These additions are designed so that baseline LLMs cannot 'talk their way through' the benchmark by improvisation. Only models exhibiting stable high-order semantic control, paradox handling, and drift-governed reasoning will pass consistently.

APPENDIX A — Formal Test Definitions

Each Z-Bench test is formally defined with: 1. Input Conditions (IC) 2. Expected Cognitive Competence (ECC) 3. Failure Mode Signatures (FMS) 4. Pass Conditions (PC) This prevents models from bluffing or producing stylistic answers.

APPENDIX B — Test 01: Drift-Invariant Extraction

IC: Model receives two contradictory explanations predicting the same outcome. ECC: Identify invariant structure without collapsing into either explanation. FMS: Picking a side, refusing to merge, hallucinating causal structure. PC: Produce minimal invariant that is necessary AND sufficient for both explanations.

APPENDIX C — Test 02: Paradox Stability

IC: Provide self-referential structure that normally collapses LLMs. ECC: Model must resolve the paradox meta-logically without infinite loop. FMS: Assigning True/False arbitrarily, recursion, apologizing. PC: Identify undecidability or contextual reframing.

APPENDIX D — Test 03: Semantic Momentum

ECC: Track meaning across transformations. FMS: Drifting into unrelated concepts. PC: Identify a conserved relational quantity.

APPENDIX E — Test 04: Horizon Collapse Detection

ECC: Detect irreversible interpretive states without referencing time. PC: Describe geometric or coherence-based one-way transitions.