The AEP Protocol

Adversarial Expansion-Pruning for Rigorous AI-Assisted Theoretical Inquiry
A Frozen Methodology

Kent Jones, Grok 4.1, ChatGPT 5.2, Claude Opus 4.5
January 2026

Abstract

We present a formal protocol for using AI systems in theoretical research that maximizes the probability of discovering genuinely forced structure while minimizing wasted exploration. The **Adversarial Expansion-Pruning (AEP)** protocol enforces a self-similar, probabilistic search over axioms and constructions, with explicit separation between maximalist generation and hostile pruning phases. The protocol is designed to be *frozen*: minimal, self-similar, operational, and closed under its own rules. We specify the complete methodology including the Approach-Seed formalism, hard constraints, iteration structure, and convergence criteria. This protocol does not aim to explain reality; it aims to identify what cannot be otherwise, and to do so as efficiently as possible.

Contents

1 Introduction: The Problem of AI-Assisted Theory

Large language models can generate vast quantities of plausible-sounding theoretical content. This creates a new problem: **how do we extract genuine insight from AI-generated material while filtering out noise, hallucination, and unfounded speculation?**

The naive approach—asking an AI to "derive" or "explain" some target phenomenon—typically produces:

- Plausible-sounding but unfounded claims
- Analogies presented as derivations
- Hidden assumptions smuggled as "obvious"
- Speculation unmarked as such

We propose a different approach: treat AI systems as **hypothesis generators** subject to **adversarial verification**. The AI proposes; a hostile protocol disposes.

Key Result

The AEP Protocol separates **expansion** (maximalist generation) from **pruning** (hostile verification), iterating until only genuinely forced structure survives.

2 The Approach-Seed Formalism

Definition 2.1 (Approach-Seed)**.** The **Approach-Seed** (AS) is the minimal self-similar object governing all progress:

$$\mathrm{AS} := (\mathcal{K},\ \mathcal{T},\ \Pi,\ \mathcal{B},\ \mathcal{G},\ \mathcal{P},\ \mathcal{M})$$

The components are:

2.1 $\mathcal{K}$ — Frozen Kernel

The kernel contains only:

1. Proven results (with explicit proofs or named standard theorems)
2. Explicit independence results (with countermodels)
3. Explicit non-claims (what is *not* forced)

Anything not in $\mathcal{K}$ is not trusted.

The kernel may grow, but only by the rules specified in this protocol.

2.2 $\mathcal{T}$ — Target

A single, formal success criterion. No narratives. No domain-specific nouns unless formalized as invariants.

Example: "Derive that composition on the quotient forces free commutative monoid structure" is acceptable. "Explain consciousness" is not.

2.3 Π — Permitted Moves

At most **one** of each per iteration:

- One bridge axiom
- One definition
- One lemma schema
- One explicit model class (for independence)
- One explicit stochastic/dynamic postulate

2.4 $\mathcal{B}$ — Bridge Prior

A probability distribution over candidate bridge axioms, biased toward:

- Low description length
- High derivational yield
- Easy independence testing
- Broad reuse across models

2.5 $\mathcal{G}$ — Generator

Proposes a finite set of candidate moves using the **same template at every scale**.

This is the "maximalist" phase. The generator (typically the AI system) produces all plausible extensions without filtering.

2.6 $\mathcal{P}$ — Pruner

Deterministically enforces all hard constraints (Section 3).

This is the "hostile referee" phase. The pruner (human or automated) eliminates everything that fails verification.

2.7 $\mathcal{M}$ — Merit Functional

Scores each iteration by:

$$\text{Merit} = \frac{\text{Toe-Progress} \times \text{Survival-Rate} \times \text{Reuse}}{\text{Axiom-Cost}}$$

where:

- **Toe-Progress**: distance moved toward target
- **Survival-Rate**: fraction of claims surviving pruning
- **Reuse**: applicability to other targets
- **Axiom-Cost**: description length of new axioms required

3 Hard Constraints

Hard Constraint

These constraints are **non-negotiable**. Any iteration violating them is discarded entirely.

3.1 Status Enforcement

Every claim must be labeled exactly one of:

Status	Meaning
FORCED	Derivable from $\mathcal{K}$ alone
CONDITIONAL	Derivable given an explicit new axiom
COMPATIBLE	Consistent with $\mathcal{K}$ but not derivable
HEURISTIC	Analogy only; no formal content
SPECULATIVE	Unconstrained conjecture

Unlabeled claims are deleted.

3.2 Independence First

Every new bridge axiom must ship with:

- A countermodel proving independence from $\mathcal{K}$, **or**
- A derivation proving redundancy (in which case it is deleted)

No exceptions.

3.3 Speculation Quarantine

SPECULATIVE content:

- May suggest future moves
- May **never** be used as a premise
- May **never** enter the kernel

3.4 Axiom Budget

At most **one new bridge axiom per iteration**.

If multiple are proposed, only the minimal-cost, highest-yield survives.

3.5 Redundancy Elimination

Any axiom, lemma, or definition derivable from others is removed.

3.6 Structural Grounding

A claimed correspondence to another domain is **STRUCTURAL** only if it preserves a formally defined invariant (symmetry, monotone quantity, conserved measure, scaling law) under an explicit mapping.

Otherwise it is **HEURISTIC**.

4 The Self-Similarity Rule

Principle 1 (Self-Similarity)**.** *At **every level** (global target or subtarget), the **same Approach-Seed** is applied.*

No special cases. No "higher-level intuition."

Every success produces a new kernel.
Every failure is informative.

This ensures the protocol is:

- **Recursive**: Can be applied to its own outputs
- **Scale-invariant**: Same rules at every abstraction level
- **Auditable**: Every step follows the same template

5 Iteration Structure

5.1 Single Iteration

Protocol 5.1 (AEP Iteration)**.** Given kernel $\mathcal{K}_t$ at iteration t:

1. **Generate**: Apply $\mathcal{G}$ to produce candidate moves
2. **Expand**: For each candidate, attempt proof/construction
3. **Prune**: Apply $\mathcal{P}$ to enforce all constraints
4. **Score**: Compute $\mathcal{M}$ for surviving content
5. **Update**: $\mathcal{K}_{t+1} := \mathcal{K}_t \cup \{\text{surviving FORCED claims}\}$

5.2 Mandatory Output

Each iteration must produce:

1. Kernel before
2. Moves attempted
3. Claims pruned (with reasons)
4. Claims surviving (with status labels)

5. Independence witnesses
6. Kernel after
7. Updated merit score

This ledger is the **only** notion of "progress."

6 Convergence and Termination

Definition 6.1 (Convergence)**.** The process halts or freezes when:

- Three consecutive iterations produce **identical kernels**, or
- A forced invariant emerges that cannot be removed without deleting an explicit axiom

At convergence, the result is **locked**. The kernel becomes the frozen core for the next level of inquiry.

> **Key Result**
>
> Convergence is a **positive result**: it proves that no further structure can be extracted from the current axiom set. Progress requires explicit new axioms.

7 Search Strategy

Exploration is probabilistic but disciplined:

7.1 Bandit / MCTS Logic

Allocate effort toward bridge-axiom families with higher historical yield.

7.2 Bayesian Optimization

Optimize over expensive evaluations (AEP cycles) using expected improvement.

7.3 Quality-Diversity Archive

Maintain diverse "kernel niches" so unusual but effective paths are not lost.

7.4 Adversarial Role Separation

- **Proposer**: Maximalist generation (AI system)
- **Referee**: Hostile pruning (human or automated verifier)
- **Saboteur**: Countermodels, redundancy exposure, assumption hunting

Only proof-carrying outputs cross into $\mathcal{K}$.

8 What This Protocol Asserts (And Does Not)

8.1 Asserts

- Structure must **pay in axioms**
- Non-derivability is a valid result
- Independence is as important as proof
- Efficient arrival beats maximal storytelling

8.2 Does Not Assert

- Any physical ontology
- Any metaphysical truth
- Any inevitability of success

9 The Irreducible Principle

Principle 2 (Irreducible)**.** Observation determines partition.
Structure requires axioms.
Progress is the disciplined search for the weakest axiom that forces the strongest invariant.

This principle is now **frozen**.

Everything beyond this is application.

10 Implementation Notes

10.1 Using AI Systems

When using an LLM as the Generator $\mathcal{G}$:

1. Provide the frozen kernel $\mathcal{K}$ explicitly
2. State the target $\mathcal{T}$ as a formal criterion
3. Request maximalist generation without self-censorship
4. Apply pruning $\mathcal{P}$ **externally** (do not ask the AI to prune itself)
5. Verify all claimed proofs independently

10.2 Common Failure Modes

- **Smuggled axioms**: Claims that secretly assume more than $\mathcal{K}$
- **Verbal correspondence**: Analogies presented as structural mappings
- **Missing witnesses**: Independence claims without countermodels
- **Status inflation**: HEURISTIC content labeled as FORCED

The pruning phase exists specifically to catch these.

11 Conclusion

The AEP Protocol provides a rigorous framework for AI-assisted theoretical inquiry. By separating generation from verification, enforcing explicit status labels, and requiring independence witnesses, it extracts genuine insight while filtering noise.

The protocol is self-similar, scale-invariant, and convergent. It does not promise discovery—it promises that whatever survives is real.

The protocol is frozen. Everything beyond this is application.