

Breaking LLM Mode Collapse: Poly-Modal Hierarchical Chunking, Perspective Prisms, and Recursive 5W1H Information Graphs

Taylor Amarel · Creator of TAEDRI & Baltor.ai · Portal: <https://papers.baltor.ai> · Code: [polymodal-5w1h-ideation-graph-engine](#)

The Problem With Standard LLM Brainstorming: If you feed a document into an LLM and ask it “Who, what, where, when, and why?”, it gives you the five most predictable answers in its pretraining distribution. Ask it ten more times, and it repeats slight variations of the same clichés—a failure mode we call **dangerous ideation convergence**. This paper describes the 4-stage engine we built in Baltor.ai to ingest any computer data (from a single photo or CSV to an 8,000-page dissertation), break it apart, challenge it from radically different viewpoints, and build a confidence-scored information graph that synthesizes new tools and harness components.

1. STAGE 1 & 2: POLY-MODAL HIERARCHICAL CHUNKING & CONTROLLED MUTATION

Instead of slicing text into flat 512-token windows, our ingestor preserves the **original immutable source** while building an arbitrary-depth DAG of sub-partitions and mutations:

- **Structural, Visual & Stochastic Chunks:** Pages, sections, 5-paragraph intro-body-conclusion arcs, **photos within pages and photos within paragraphs**, emphasis-only spans, clutter/fluff-stripped telegraphic cores, and seeded random span cuts.
- **Multi-Modal & Cross-Lingual Mutations:** Synonym and register shifts, RAG/embedding neighborhood supplementation, rewriting in Mandarin Chinese or analytic topic-comment grammar, and bidirectional `chunk ↔ image` scene transcoding.

2. STAGE 3: THE PERSPECTIVE PRISM & 10-WAY DIALECTICAL CHALLENGER

Each chunk is then interrogated through contrasting intellectual lenses and historical reality constraints while keeping the full triad (Original + Mutated Remix + Perspective Challenge):

Lens Family	Example Personas & Constrained Realities	10-Way Stress Matrix Generated Per Chunk
Thinker Lenses	Frontier AI Researcher, Karl Marx (labor & surplus extraction), Claude Shannon (entropy), Judea Pearl (causality), Richard Feynman	• 10 Reasons Why It Is Wrong (Falsification) • 10 Reasons Why It Is Right (Steelman)
Simulated Realities	Google Engineer in 1999 (commodity disks), NASA Apollo 13 in 1970 (45-amp ceiling), 1973 Bell Labs Unix (64KB RAM)	• 10 Scenarios Where It Works • 10 Scenarios Where It Breaks (Corner Cases)

3. STAGE 4: RECURSIVE 5W1H CONFIDENCE GRAPH & ANTI-CONVERGENCE FUNNEL

We then run a recursive Who / What / Where / When / Why matrix returning structured JSON/YAML arrays with explicit **confidence and strength scores**:

- **Contrastive, Residual & Epistemic-Age Probes:** For WHO and WHAT, we ask: *Who does this relate to? Who does it NOT relate to? Top 10? Who else?* (toggling between fresh context and stateful history), plus *kindergartener*, *high-schooler*, and *specialist* knowledge levels.
- **Second-Order Protagonist vs. Antagonist Drill-Down:** When an entity surfaces (e.g., *Elon Musk* or *Apollo 13*), the engine recursively asks 5W1H on that entity and pits a **Protagonist** (defending the connection) against an **Antagonist** (attacking spurious analogies) before adding the edge to the global Information Graph and compiling verified `skills.md` and `baltor_contract_card.json` files.