# Thesis Engineering: An AI-Enhanced Framework for Robust Long-Term Investment Research

Peter Kromkowski

Long-term investing is constrained by integrating diverse information, stress-testing assumptions, and tracking evidence evolution. While analysts excel at judgment, synthesis, and qualitative interpretation, investment ideas frequently fail due to fragile assumptions, hidden dependencies, or hidden scenario interactions that are difficult to perceive. To address these limitations, I propose an AI-enhanced framework that complements traditional research by systematically interrogating assumptions, generating realistic challenges, and maintaining a persistent, auditable memory of prior ideas and outcomes. The goal is to produce investment theses that are robust, transparent, and actionable, while retaining human judgment as the central decision-maker.

The current system situates a stock within a structured, multi-dimensional context. Historical prices, financial performance, peer comparisons, macroeconomic sensitivity, and management behavior are analyzed to provide a coherent profile. Natural language from earnings calls, guidance, and investor communications is parsed to extract key relationships and causal connections. Temporal trends and operational metrics contextualize scenario planning and constrain plausible outcomes. Importantly, this first stage informs human judgment without producing automatic recommendations, creating a disciplined substrate for forming investment ideas.

Once a candidate's thesis is proposed, it is decomposed into explicit, testable assumptions. Each assumption is evaluated for criticality to the thesis, supporting evidence, and vulnerability to adverse events. The system identifies single points of failure and high-leverage dependencies, highlighting areas where the thesis could collapse if an assumption fails. To rigorously stress-test the thesis, the system generates realistic counterfactual scenarios for each assumption. Historical analogs, peer performance deviations, and macro regimes ground these scenarios in reality, mitigating cognitive biases. Each scenario is assessed for severity, producing a ranked view of vulnerability across the thesis, ensuring analysts focus on the most critical assumptions.

The framework then translates these scenarios into financial outcomes, estimating effects on revenue growth, margins, and overall valuation. Scenarios with significant downside are explicitly flagged, and outcomes are aggregated into a comprehensive assessment. Analysts receive structured outputs, including fragile assumptions, risk severity, and evolving confidence guidance. The AI system enforces rigor, tracks evidence, and maintains a persistent memory of assumptions, challenges, and scenario outcomes, while analysts provide context, interpret information, and make allocation decisions. The current implementation enables structured evaluation of ideas, scenario-based stress testing, quantification of tail risk, and guidance for monitoring. The system supports incremental execution and parameter customization at each step without rerunning the full pipeline. A robust test suite validates consistent behavior across diverse scenarios.

Future enhancements will extend the system into a self-improving, empirically calibrated framework. Outcome tracking will record thesis predictions alongside realized performance, enabling calibration of survival probabilities, risk thresholds, and monitoring cadence. Historical patterns will inform learning loops that adjust thresholds, scenario generation, and relative weight of risk factors, addressing systematic biases. Over time, this calibration will allow the framework to more accurately identify durable ideas, fragile assumptions, and expected thesis lifespan. These enhancements will let analysts focus on high-conviction opportunities while mitigating overconfidence and hidden tail risks.

Collectively, the current system and planned enhancements establish a human–machine research loop that improves decision-making. Humans contribute context, interpret qualitative information, and make allocation decisions. The AI system enforces analytical rigor, simulates realistic stress scenarios, tracks evidence, and learns from outcomes to refine the framework. Evaluation metrics focus on practical indicators: persistence of ideas under stress, severity of downside outcomes, analyst responsiveness to contradictory evidence, and improvements in confidence calibration. By integrating assumption testing, adversarial scenario analysis, financial translation, and continuous learning, this framework systematically enhances the reliability, durability, and interpretability of long-term investment theses, lifting hit rates while managing tail risk.

| Stage / Subprocess | Purpose | Inputs | AI Role | Human Role | Outputs | Key Metrics / Feedback |
|---|---|---|---|---|---|---|
| **Stage 1: Reality Framing** | Establish stock risk and opportunity boundaries; provide structured context for thesis formation. | Historical OHLCV, financial filings (10-K, 10-Q), peer data, macro indicators (GDP, CPI, rates), management transcripts, ownership, and flow data. | Parse unstructured text, compute historical and cross-sectional metrics, detect regime shifts, extract language patterns, and generate structured dashboards. | Interpret outputs, validate relevance, explore risk regimes, and decide if the stock warrants thesis creation. | Structured risk map, regime classification, peer-relative and macro-relative metrics, management narrative insights. | Tail risk identification, regime sensitivity, volatility & drawdown characterization, narrative drift detection. |
| **Stage 2.1: Thesis Decomposition** | Break the human thesis into explicit, testable assumptions. | Analyst-written thesis text, supporting Stage 1 context. | Extract assumption clauses from narrative, classify them as operational, macro, competitive, or behavioral, and map dependencies between assumptions. | Confirm relevance and clarity of assumptions; highlight gaps or redundancies. | Discrete assumption set with dependency mapping ("assumption map"). | Coverage completeness, assumption clarity, dependency identification, and redundancy detection. |
| **Stage 2.2: Red Team Generation** | Test assumptions by generating realistic, evidence-based challenges. | Assumption map, historical analogs, peer deviations, macro regimes, Stage 1 context. | Generate counterfactual challenges for each assumption, assign severity ratings, and highlight high-leverage vulnerabilities. | Review challenge plausibility, prioritize high-severity or single-point failure risks. | Challenge set with severity distribution, coverage metrics, and high-risk assumptions flagged. | Coverage of vulnerabilities, realism of challenges, and identification of single points of failure. |
| **Stage 2.3: Scenario Simulation** | Create plausible alternative worlds to stress-test assumptions. | Assumptions, challenges, historical analogs, peer, and macro data. | Generate multiple scenarios (base, bull, bear), augment with challenge-focused extreme variants, enforce empirical bounds, and assign plausibility weights. | Confirm scenario relevance, flag implausible or incomplete scenarios, and provide context-specific insight. | Scenario set with survival probabilities, plausibility scores, and extreme-case outcomes. | Scenario coverage, representation of tail risk, concentration of fragile assumptions, and detection of high-leverage dependencies. |
| **Stage 2.4: Financial Impact Translation** | Translate scenario outcomes into financial consequences and valuation impacts. | Scenario set, assumption stress parameters. | Convert scenario assumptions into revenue, margin, growth, and valuation projections; compute upside/downside potential; identify impaired scenarios. | Interpret financial outcomes, validate realism, and assess implications for position sizing. | Financial outputs for each scenario, weighted valuation ranges, and tail-loss estimates. | Implied upside/downside, tail-risk magnitude, scenario plausibility-weighted impact. |
| **Stage 2.5: Temporal Evidence Integration** | Continuously update thesis integrity as new data arrives. | Earnings reports, KPIs, management updates, macro releases, market news. | Map evidence to assumptions, classify as supporting, contradicting, or ambiguous; track persistence over time. | Confirm evidence relevance, interpret significance, and adjust thesis or conviction as needed. | Updated assumption survival curves, evidence-alignment scores, and temporal risk map. | Evidence-contradiction score, confidence-evidence gap, persistence of assumption failure. |
| **Stage 2.6: Thesis Resilience & Conviction Calibration** | Quantify alignment between analyst confidence and actual thesis robustness; inform position sizing and monitoring. | Survival analysis, evidence integration, and analyst confidence inputs. | Compute resilience metrics, fragility score, and context-aware calibration; penalize fragile or high-severity assumptions. | Adjust conviction, position sizing, and monitoring frequency; incorporate lessons from past outcomes. | Fragility indicators, calibrated confidence, suggested monitoring cadence, and dominant failure modes. | Calibration accuracy, historical consistency, responsiveness to contradictory evidence, and repeat-mistake reduction. |