

Actuarial Agents

and How to Scale Them

LLMs reason. Agents act. Narrow AIs compute. Actuaries own.
Together they drive value.

Ronald Richman FIA FASSA CERA CPCU · CEO & Founder, InsureAI
Jacques Bisschoff Senior Product Engineer & Quant, InsureAI

STIC Seminar 2026 · Johannesburg

ACTUARIAL SOCIETY OF SOUTH AFRICA
Advancing Actuarial Excellence
Educate • Develop • Uphold

Agenda

1. Foundations: AI, agents & key concepts
2. Building an Actuarial Agent
3. The Study & the Capability Layers
4. Results across L1–L4
5. Why It Works: the mechanisms
6. Trust & the Road Ahead

PART 1 · FOUNDATIONS

Setting the Stage

What AI is, how agents work, and the vocabulary for the rest of the talk.

What is AI?

Artificial intelligence encompasses systems that perform tasks that would otherwise require human intelligence.

Large Language Model

Trained on vast text, **not rules**, not expertise, just patterns. Reasons by predicting what comes next. Domain knowledge must be given.

Same prompt, different run: different answer.

Deep Learning

Modern AI's core technique: neural networks that learn complex patterns directly from data, with no strong prior theoretical assumptions. LLMs and **purpose-built** actuarial models are both deep learning, applied broadly or to one task.

Augmented Intelligence = LLM + Agent + Narrow AI + User Feedback

Not automation, insight that drives better decisions. AI surfaces patterns, signals and anomalies a human alone would miss. The **human** stays in the loop, makes the call, and **acts** with far greater confidence and clarity.

Two Kinds of AI: General vs Narrow

General AI: LLMs

Systems that perform across a wide array of tasks. LLMs like GPT-4 and Claude learn patterns from vast data with no strong theoretical assumptions, predicting the next token from the empirical distribution of language.

Broad · Flexible · Fluent, but stochastic, not domain-calibrated

Narrow AI: specialized tools

Purpose-built models for one task (pricing, reserving), the core technical work actuaries do. Tailored to the problem, faster and more accurate than traditional methods, and inherently explainable, scorable and auditable.

Specialized · Deterministic · Accurate, but only within scope

General AI reasons broadly. Narrow AI computes precisely. Neither alone produces a reserve you would sign.

Richman (2024), An AI Vision for the Actuarial Profession · The Actuary in an Age of AI.

Agents Marry the Two

THE BEST OF BOTH: REASONING + COMPUTATION IN ONE LOOP

General AI: LLM

decides WHAT to do: reads the task, reasons, chooses the method, interprets results

+

Narrow AI: tools

does it precisely: exact arithmetic, fits and scores models against emerged experience

THE AGENT

Reasoning + computation in one loop. The LLM picks the method; the narrow AI computes the answer and scores it; the loop iterates until done.

THE OUTCOME

A reserve that is reproducible, accurate and explainable, and owned by the actuary.

LLMs reason. Agents act. Narrow AIs compute. Actuaries own.

Key Concepts



THE VOCABULARY FOR THE REST OF THE TALK

The agent loop

Agent
An LLM that acts: calls tools, reads results, decides the next step, iterates.

Tool
A function the agent calls. Makes computation deterministic; the reasoning is not.

Harness
The infrastructure that runs the loop, executes tools and logs everything.

Context & cost

Token
The unit of text an LLM reads and writes, about three-quarters of a word.

Context window
Everything the model can hold at once; once full, earlier content is lost.

Turns
Each user / assistant exchange in the loop. Every turn spends tokens.

Scaling the agent

Namespace
A shared store where data lives by reference, kept off the context window.

Auto tooling
The agent builds and registers custom tools on demand.

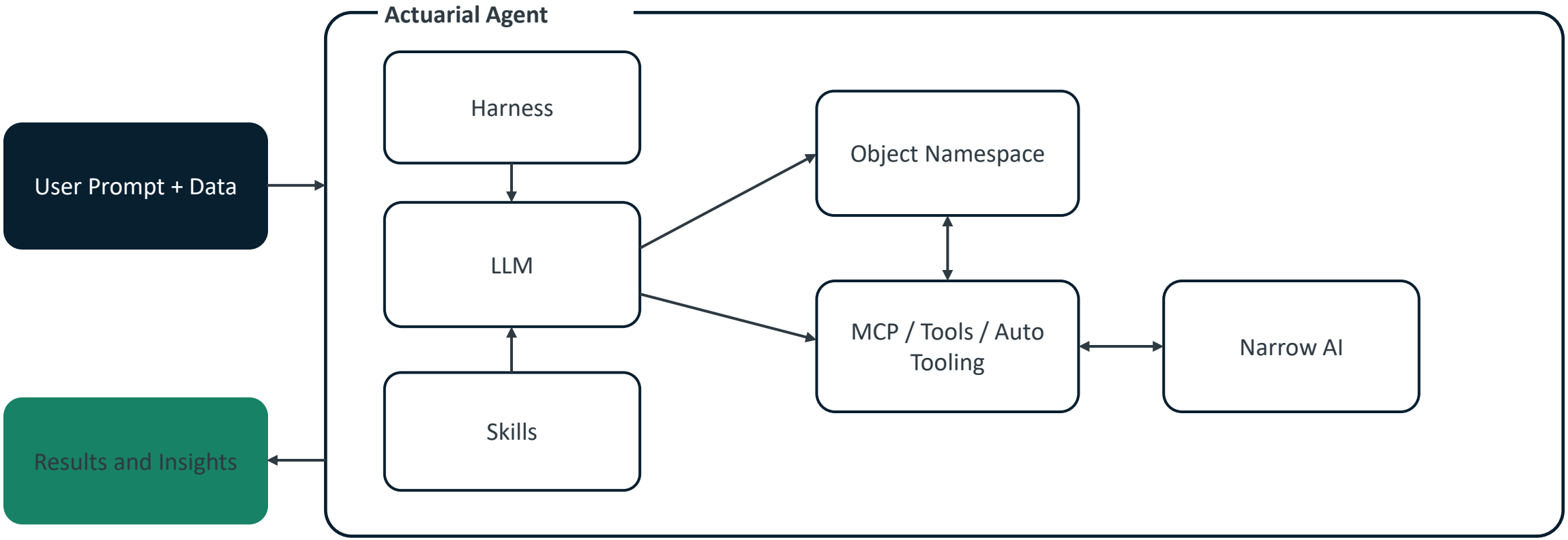
MCP
A standard giving the agent structured access to tools and data.

PART 2 · BUILDING

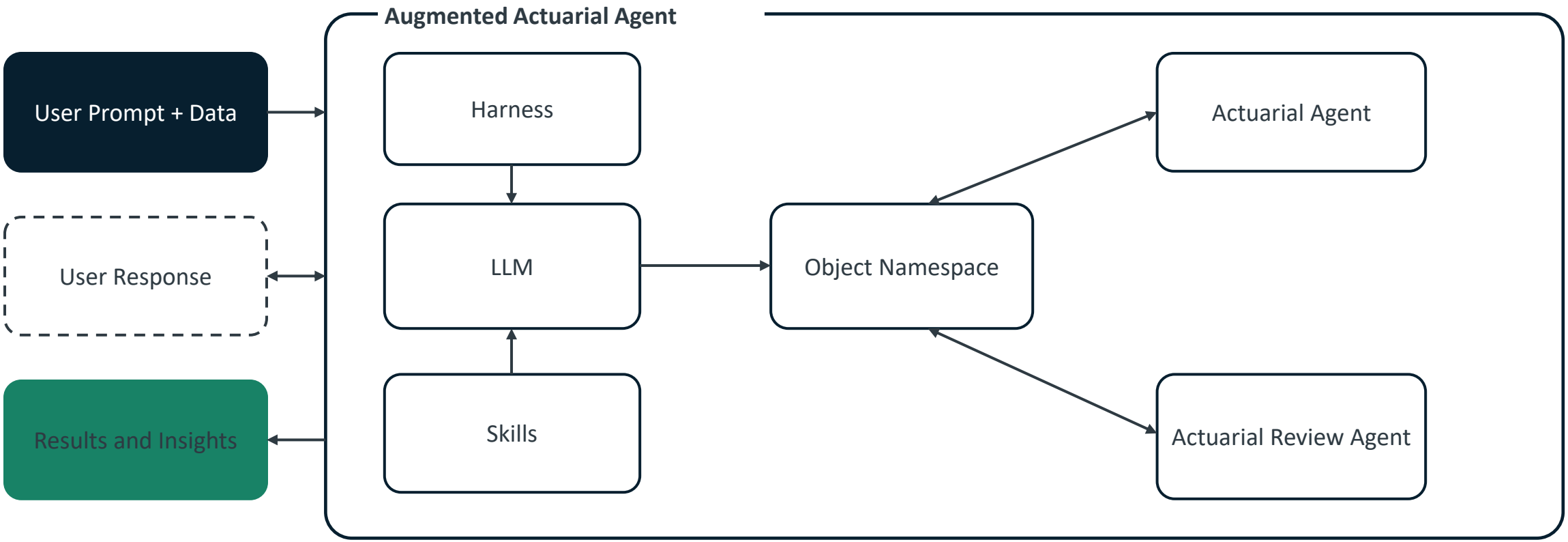
Building an Actuarial Agent

The harness, the namespace, and the multi-agent loop.

Building an Actuarial Agent



The Augmented Actuarial Agent

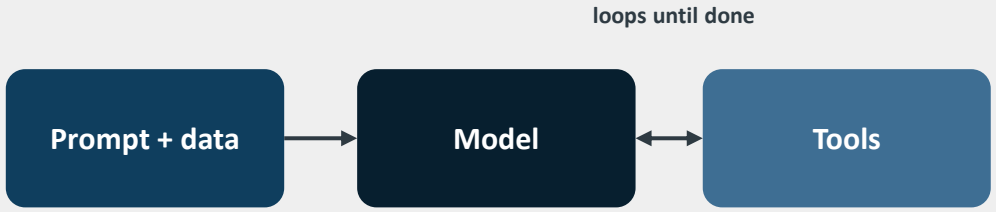


What is a Harness?

THE INFRASTRUCTURE AROUND THE AGENT

The harness is the scaffolding that wraps a model in a loop: it sends the prompt, runs the tools the model asks for, feeds the results back, handles errors and budgets, and logs every step. It turns a one-off prompt into a repeatable, auditable workflow.

The agent loop



The model calls tools, reads the results and decides the next step, looping until it is done, then returns the answer.

Common harnesses

- **LangGraph (LangChain)**
- **OpenAI Agents SDK**
- **Claude Agent SDK (Anthropic)**
- **Microsoft AutoGen**
- **CrewAI · Google ADK**

This study: a lightweight harness built on chatlas.

What Makes This Powerful

THE HARNESS & OUR NAMESPACE

Infrastructure

Any model, swapped at runtime

Provider-agnostic: hosted Claude or self-hosted Gemma / Qwen, all on the same harness.

Tools

The line between stochastic and deterministic

Typed, described functions the LLM chooses to call; results return to the namespace.

Auto tooling

The agent writes its own tools

Registers bespoke computation on demand, with full access to the namespace.

Namespace

Data by reference, not in the prompt

The LLM holds only names, types and schema, never the payload.

The Namespace, for Actuaries

Like named ranges in a shared workbook.

The LLM knows the book has a triangle tab, a validation report, a fitted model, but it never retypes the numbers. It references them by name, and the tools (macros) read and write the tabs.

Decouples computation scale from context size.

Tools pass multi-megabyte model state to each other by reference; the LLM holds only keys, types and byte counts, never the payload.

Without a namespace, L4 is not expensive: it is impossible. Multi-MB model state cannot fit in any context window at any price.

PART 3 · THE STUDY

The Study & the Capability Layers

Four layers, one prompt, three triangles, from L1 to L4.

Our Core Thesis

FOUR QUESTIONS DRIVE OUR EVALUATION

Reproducibility	Correctness	Accuracy	Cost
Same prompt, consistent results across runs?	Is the methodology applied correctly?	Benchmarked against observed outcomes, how did it do?	As capability grows, so do tokens. Worth the price?

In technical actuarial applications, agents need access to narrow AI tools that have been appropriately specified, made accessible and scaled for agentic use.

The agent supplies general reasoning and judgement; the narrow tools supply deterministic and auditable computation.

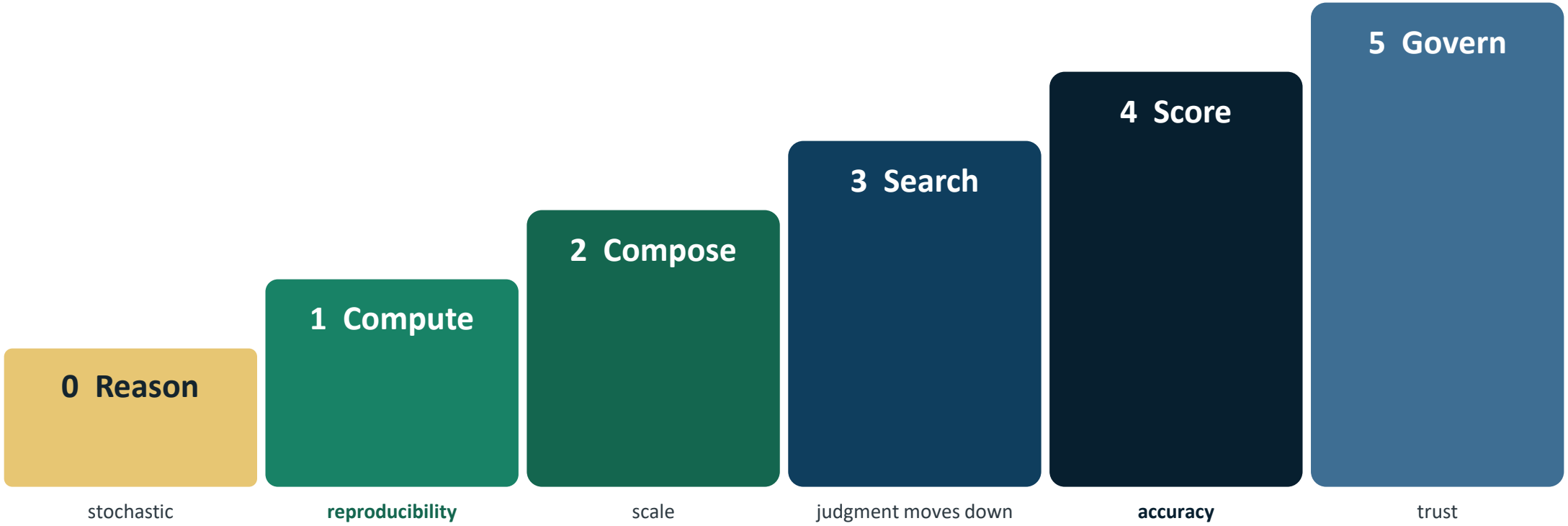
Thesis Layers

Layer 1	Layer 2	Layer 3	Layer 4
+ Reserving prompt	+ Reserving prompt	+ Reserving prompt	+ Reserving prompt
+ Inline data (as text)	+ Inline data (as text)	+ Data via namespace	+ Data via namespace
No skills	+ Reserving methodology skill	+ Reserving methodology skill	+ Reserving methodology skill
No tools / MCP	No tools / MCP	+ Basic reserving tools	+ Reserving Narrow AI MCP
No auto tooling	No auto tooling	No auto tooling	+ Result Harness
No guidance	No guidance	No guidance	+ Auto tooling
No reviewer	No reviewer	No reviewer	No reviewer
No user actions	No user actions	No user actions	No user actions

Basic → **Agent**

Depth of Tooling: Reason → Govern

A PROPERTY OF THE TOOLS, NOT THE AGENT



Reason → Compute → Compose → Search → Score → Govern: the L1–L4 study is the agent climbing this ladder.

The Prompt Structure

ONE PROMPT, FIVE PARTS · IDENTICAL ACROSS EVERY LAYER

1 Task	Determine a best estimate IBNR, a Chain Ladder benchmark and a best estimate range for the incurred loss triangle.
2 Data	Accident period, development period, cumulative incurred and earned premium. Provided inline at L1–L2, or by reference through the namespace at L3–L4.
3 Best Estimate IBNR	Pick one method (Chain Ladder, Bornhuetter-Ferguson or Cape Cod), explore parameterisations, then commit to a single best estimate by accident year and in total.
4 Chain Ladder Benchmark	Also compute a volume-weighted Chain Ladder IBNR on all origins: a consistent benchmark across every run and layer.
5 Best Estimate Range	Give a lower and upper bound reflecting genuine sensitivity (alternative methods, parameterisations or tail uncertainty), with the basis stated.

Every layer gets the same prompt. L1 & L2 receive the data inline; L3 & L4 receive it by reference through the namespace.

Basic Tools, Auto Tooling and Narrow AI

Basic Tools

- Age-to-Age Factors on a reserving triangle
- Chain Ladder LDF and development pattern calculation linked to any volume-weighted period
- Chain Ladder (CL) model given LDF params
- Bornhuetter-Ferguson (BF) model given LDF and IELR params
- Generalized Cape Cod (GCC) model given LDF and decay weighting params

Auto Tooling

- Ability to register custom tools defined by the LLM and required by the agent for insights
- Ability to run custom tools registered
- Security constraints linked to allowed actions
- For the most part auto tooling was used to build custom data queries and summaries linked to data in the namespace to get better insights

Narrow AI

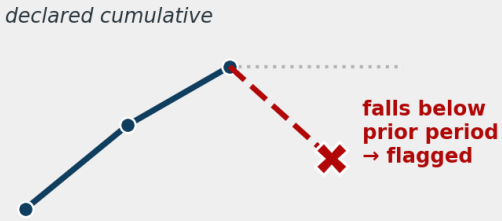
- Reserving API with access to LDF and development pattern diagnostics over time
- Auto outliers removal and trend detection
- A full suite of models with all params for CL, BF and GCC
- Modelling a range of models and parameterizations across time calculating AvE and CDR
- A system to rank candidate models against emerged experience using AvE and CDR outcomes.

Only used in L3 and L4 as defined in layer definitions. · Implements Balona–Richman (2020).

What Goes Wrong, and What Catches It

DATA GUARDRAILS THE AGENT CANNOT SKIP

Cumulative vs incremental



WHAT GOES WRONG

Mislabel incremental data as cumulative and LDFs collapse toward 1.0, silently halving IBNR.

WHAT CATCHES IT

The pipeline refuses to proceed unless the flag is set, and flags any origin whose cumulative claims fall.

Development-period units

~~[0.25, 0.5, …]~~ ✗

↓

[2, 4, 6, 8] ✓

0.25 = 25% of the pattern, not one quarter

WHAT GOES WRONG

An LLM primed on “quarterly = 0.25 years” picks 25% of the pattern, not one quarter.

WHAT CATCHES IT

Frequency is inferred from the dates; year-fractions on a named-frequency triangle are rejected by name.

Exposure per origin

origin 2010 Q1

premium 25,728

premium 25,728

premium 24,100 ✗ **differs**

WHAT GOES WRONG

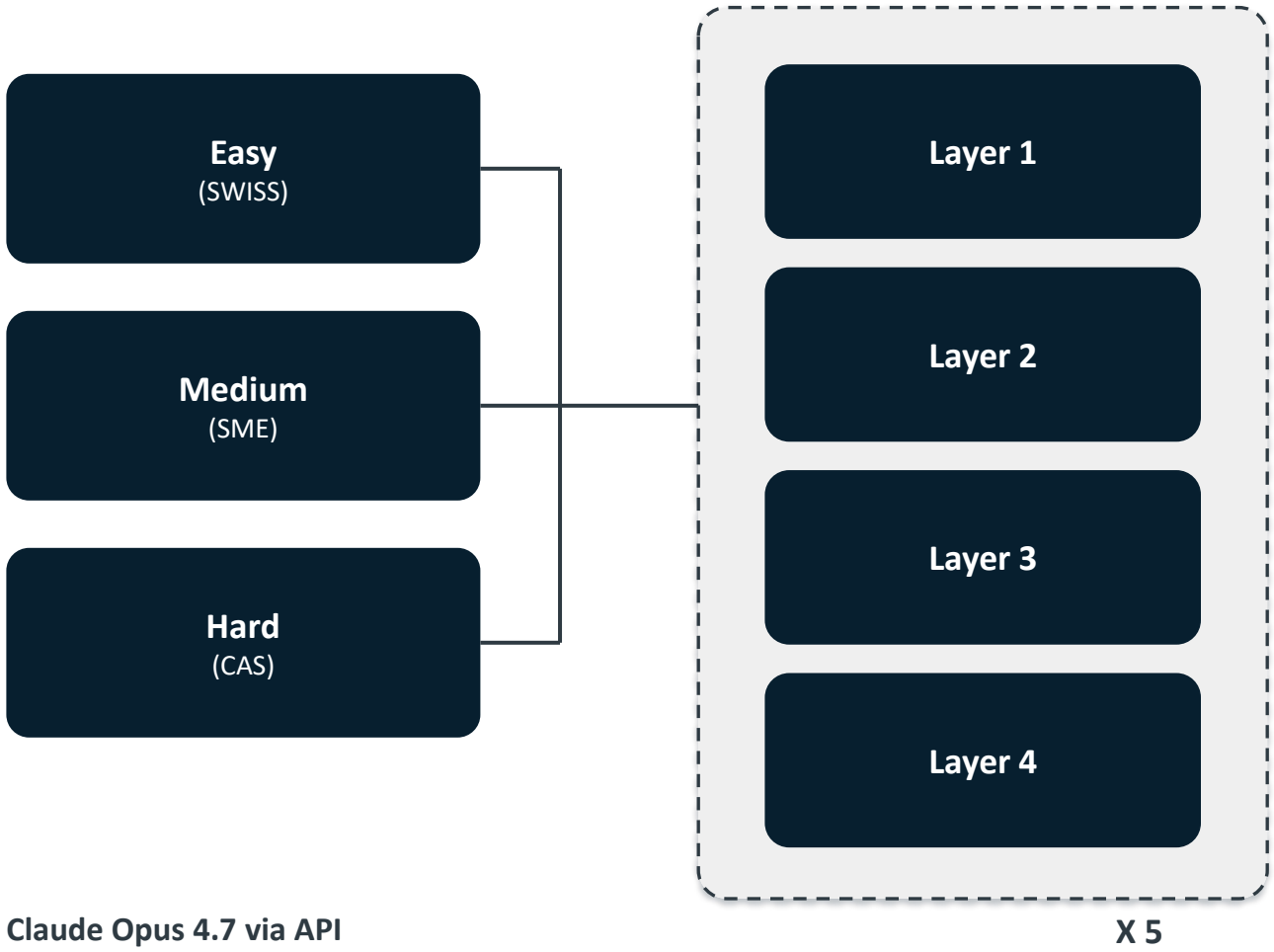
BF and Cape Cod need one exposure per origin; a join fan-out corrupts the a-priori where it matters most.

WHAT CATCHES IT

Validation counts distinct exposure per origin and hard-errors if any origin carries more than one value.

Correctness by refusal: the agent is blocked at the data gate before a single LDF is computed.

The Study Design



Claude Opus 4.7 via API

Multiple Problems

Tested each layer with different reserving data sets classified as easy, medium and hard to measure **accuracy.**

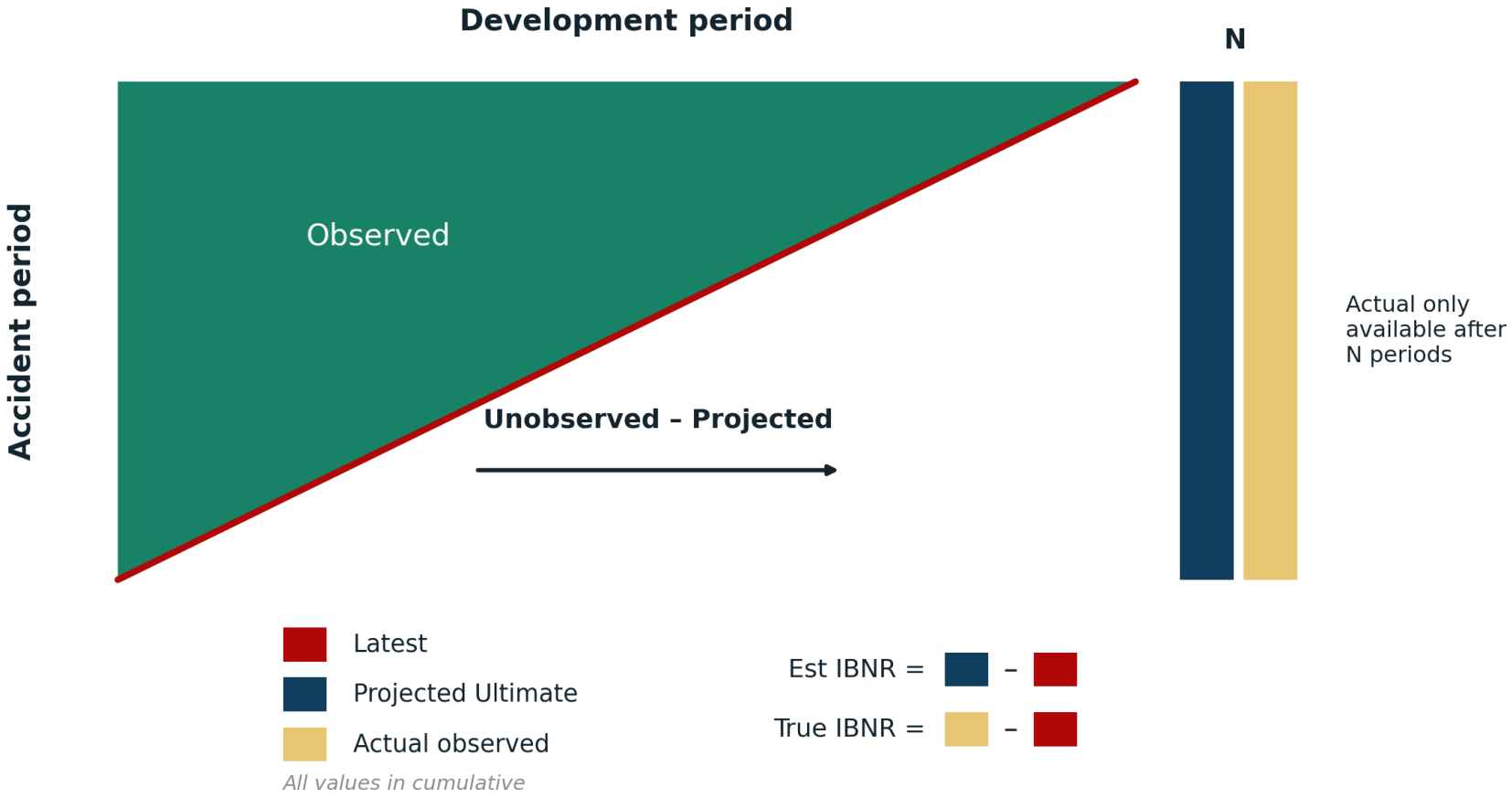
Repeat Problems

Each data/layer combination was repeated 5 times with the same prompt and layer to validate **reproducibility.**

Track Results & Cost

Track reserving best estimates ranged and the chain ladder baseline model to validate **correctness.**

Accuracy in Context of Study

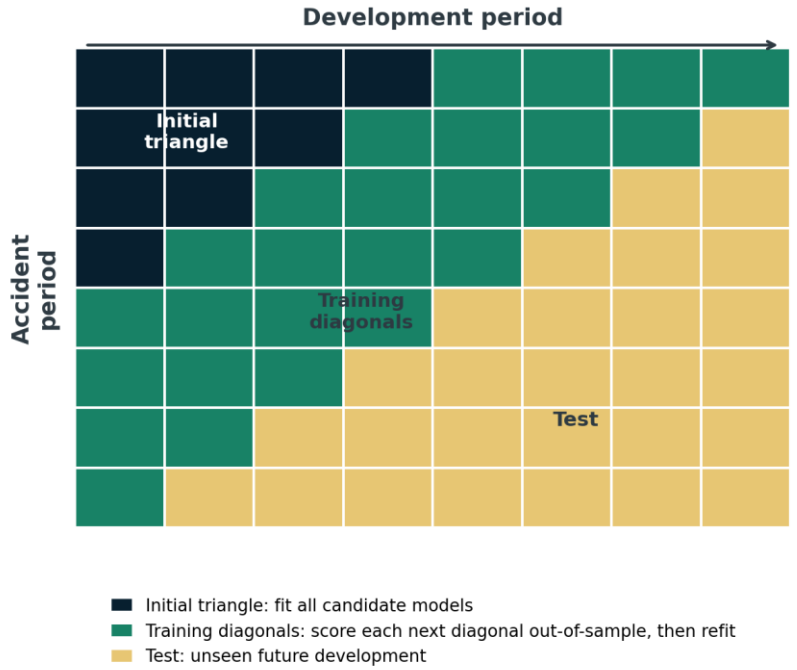


The Balona–Richman Method

THE ENGINE BEHIND L4 · BALONA & RICHMAN (2020)

Reserving model selection, reframed as supervised learning.

- **Pick what predicts, not what we believe.** Method and parameters (CL / BF / Generalized Cape Cod, LDF window, loss ratio, decay) become a model-selection problem.
- **Score out of sample.** Re-reserve successive calendar diagonals; each forecast is compared to the experience that actually emerged.
- **Two metrics.** AvE (actual vs expected) and CDR (a regularised AvE that also penalises unstable ultimates). Choose the parameterisation that minimises the score.
- **Metric guidance.** CDR for large, stable triangles; AvE for small, volatile ones.



L4's Narrow AI automates the Balona–Richman method end-to-end, with improved ranking and scoring on top: fitting the grid, scoring AvE / CDR across diagonals, and selecting the best estimate by emerged accuracy.

Balona, C. & Richman, R. (2020). The Actuary and IBNR Techniques: A Machine Learning Approach.

How Accuracy Is Won

BACKTESTED MODEL SELECTION · BUILT ON BALONA–RICHMAN (2020)

1

Fit a wide grid of methods

Chain Ladder, Bornhuetter-Ferguson and Generalized Cape Cod, across LDF windows, tail treatment, loss ratios and decay.

2

Replay across every diagonal

Each candidate is re-reserved over the historical run-off, as experience would actually have emerged.

3

Score out of sample

Compare every forecast to what emerged, using Actual-vs-Expected (AvE) and the one-year Claims Development Result (CDR).

4

Rank, select, sign off

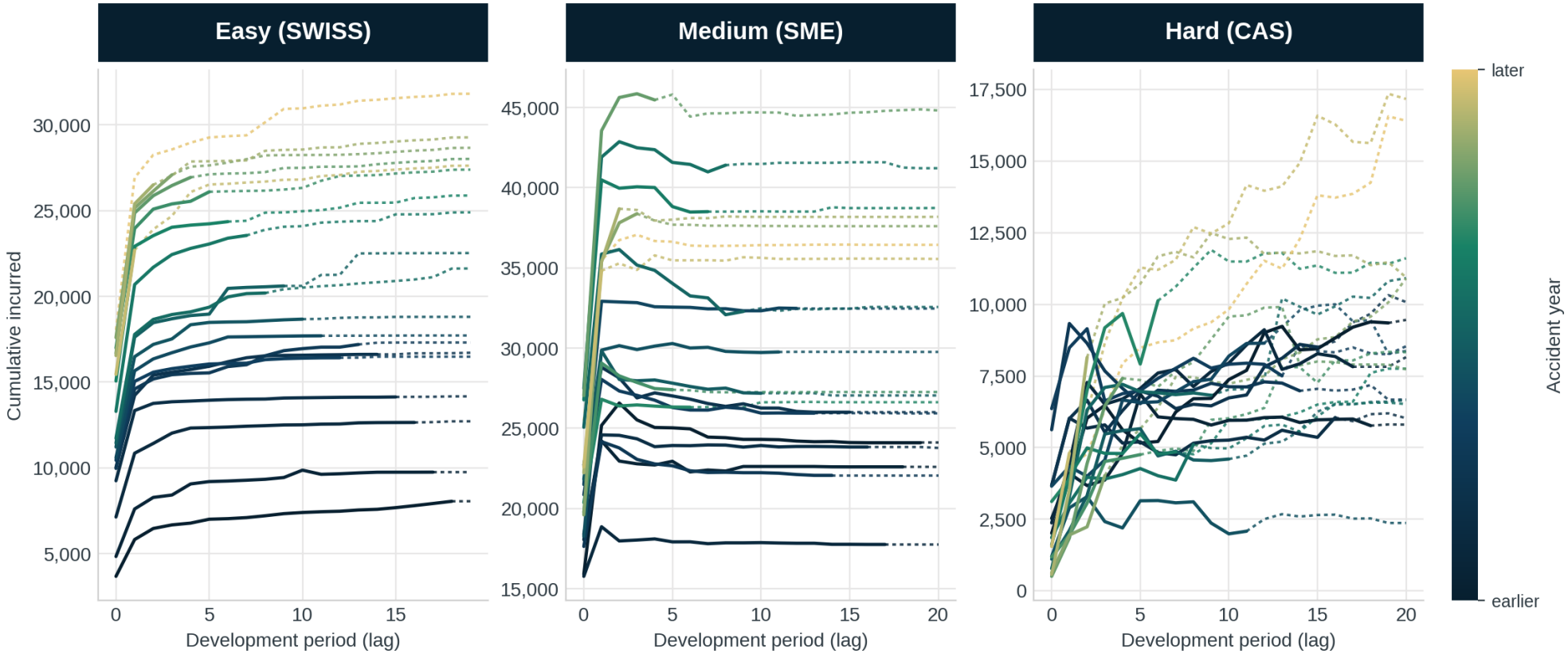
Rank candidates by out-of-sample fit, then carry the best estimate forward for the actuary to review.

The backtest is the objective.

Accuracy is out-of-sample fit to emerged experience, by construction, not a check bolted on afterward.

The ranking and scoring that turn the backtest into a single recommendation are refined beyond the published method.

How Hard Is the Task?



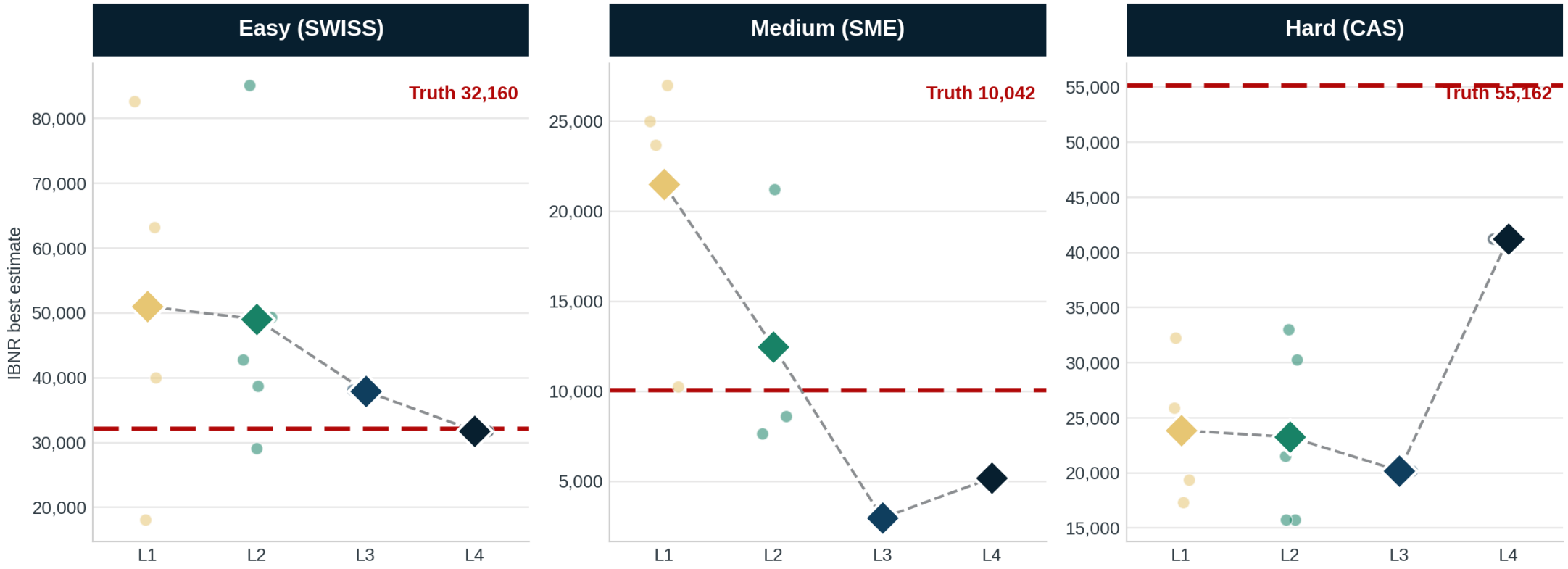
Cumulative incurred by accident year: solid is observed (train), dotted is held-out (test). SWISS develops smoothly; SME is larger; CAS is volatile and non-monotonic.

PART 4 · RESULTS

Results across the Layers

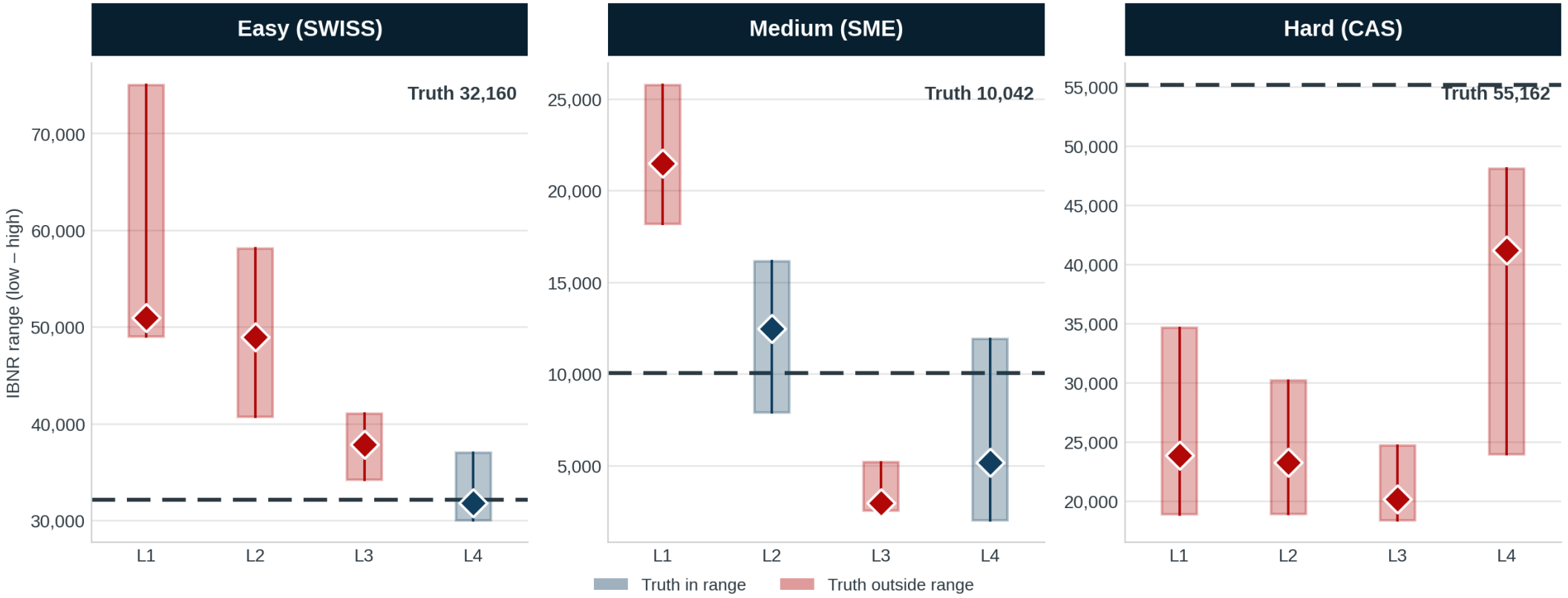
Reproducibility, correctness, accuracy and cost: what actually happened.

Best Estimate Results vs Truth



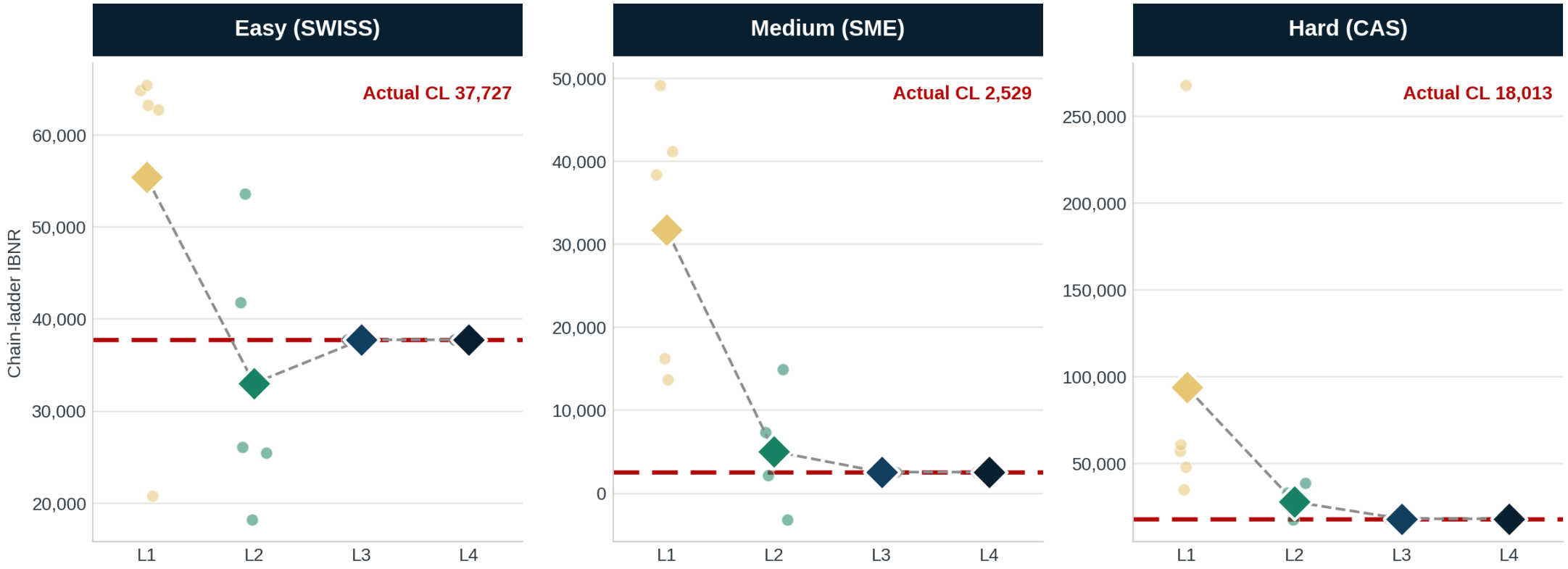
Tools add reproducibility but not necessarily accuracy – Narrow AI attempts to solve this

Best Estimate Range vs Truth



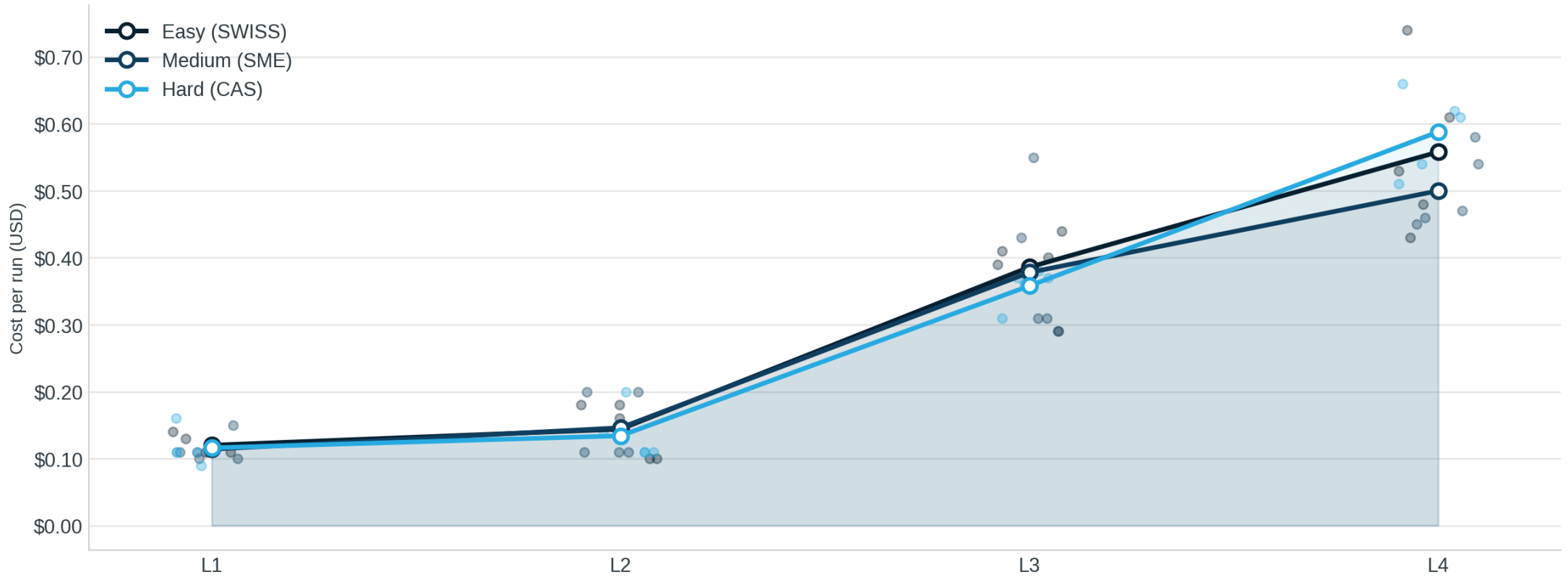
For SWISS and SME, the Truth is in range. For CAS we note the L4 upper just fell short of the range.

Baseline Chain Ladder vs Actual Chain Ladder



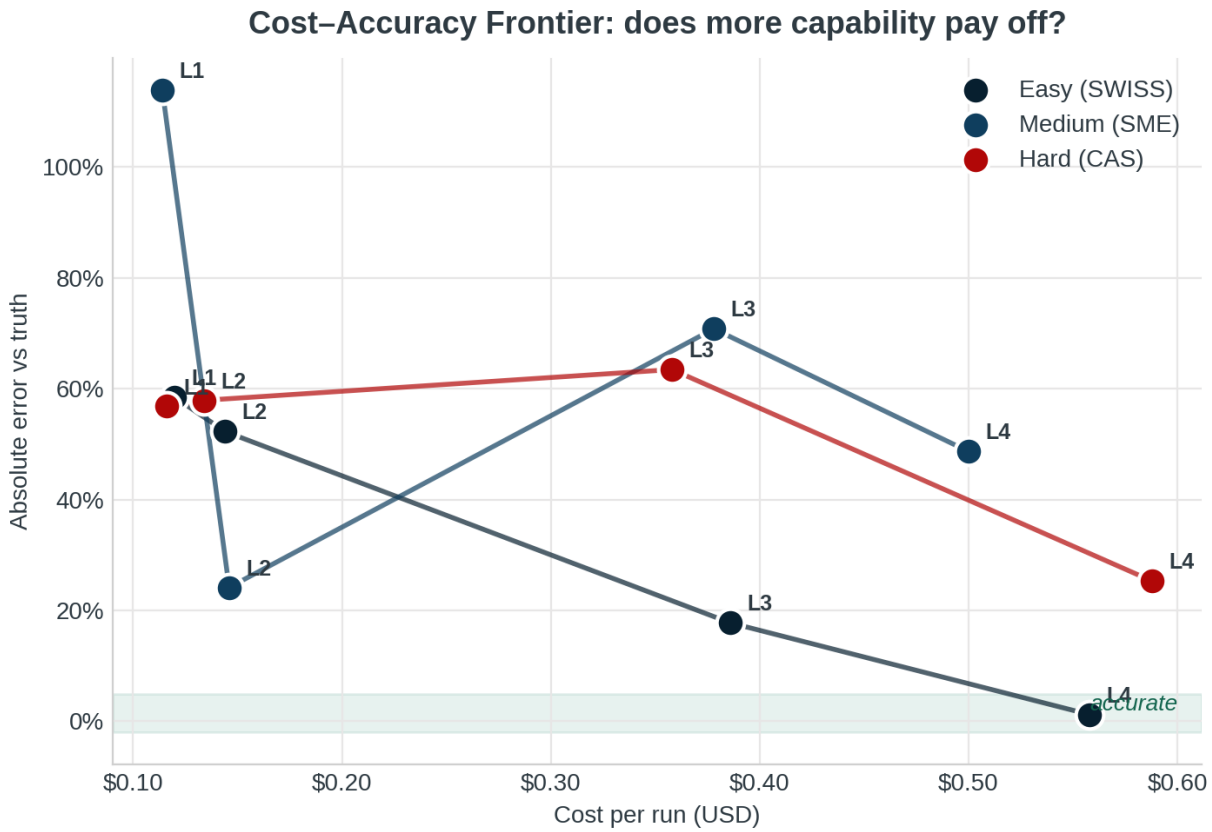
Tools and Narrow AI both ensure correctness of methodology

LLM API Cost



More context, skills and tool calls lead to more tokens and increase cost

Cost–Accuracy Frontier



READING THE FRONTIER

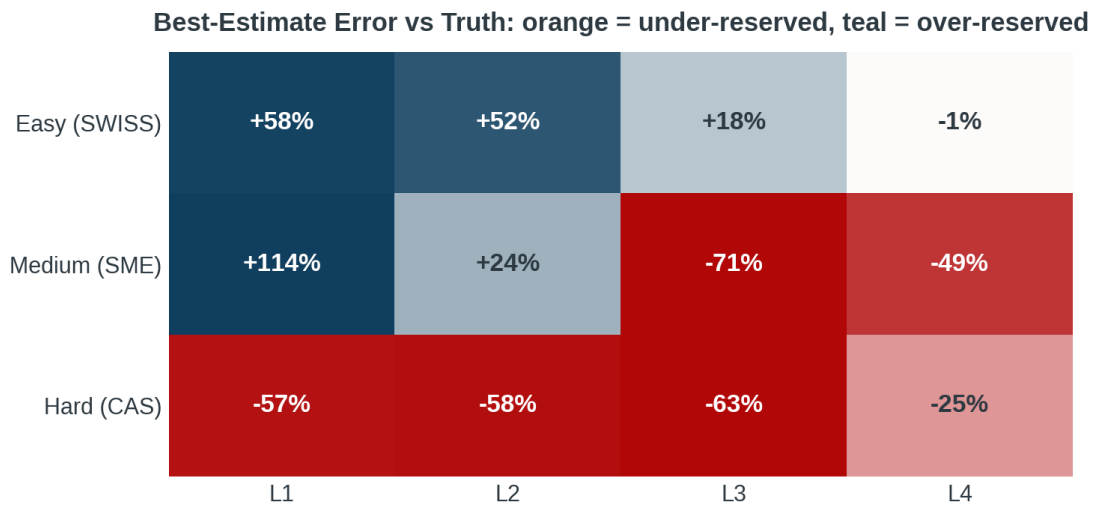
Down-and-right you are paying more per run to buy accuracy.

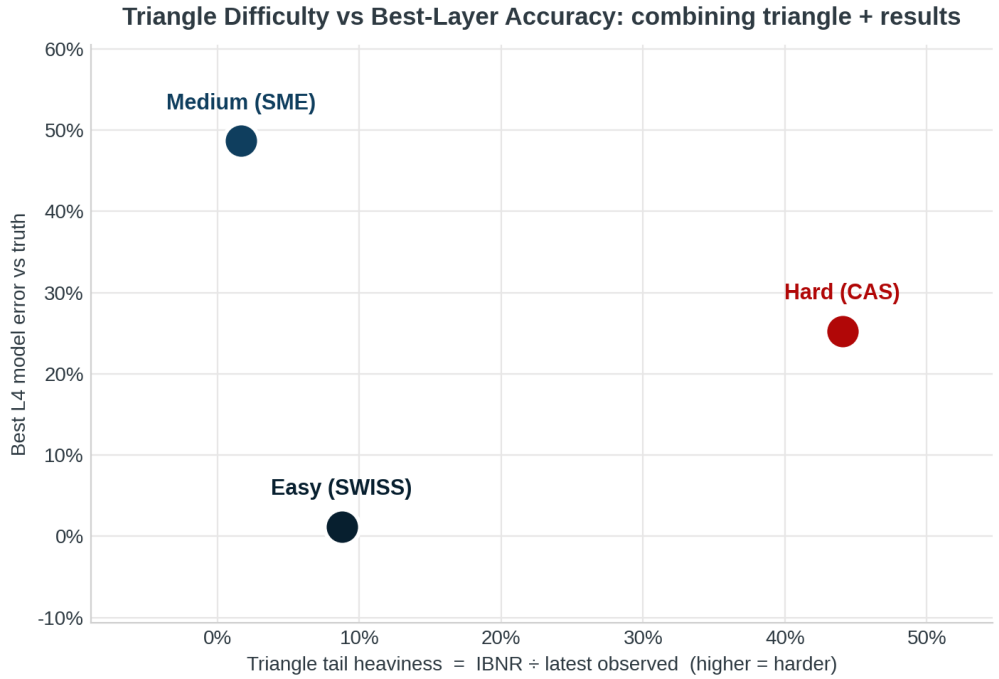
SWISS (easy) reaches the cheap-and-accurate corner by L4 (~0% error).

SME & CAS pay the full L4 premium but stay well short of truth.

L3 is the value knee: reproducibility before the L4 cost jump.

Where the Model Wins and Misses





L4 nails SWISS (−1%), improves CAS to −25%, but SME stays −49%. Accuracy does not track triangle tail-heaviness alone: consistency is not accuracy.

PART 5 · WHY IT WORKS

Three Problems, Three Solutions

You’ve seen what happened across L1 → L4. Now what it means, and what each rung proved.

Reproducible ≠ Accurate

THE L3 → L4 HINGE

L3 · Reproducible

The agent reasons its way to a method, applies it flawlessly, and picks the same method every run. Consistent. Auditable. Defensible-looking.

A confident, stable, WRONG reserve is scarier than a noisy one, because it survives review.

L4 · Accurate

Narrow AI stops the agent picking the method it believes in, and makes it pick the method the data says was most accurate out of sample.

It rejects the methods that only got lucky once.

At L3 the agent picks the method it believes in. At L4 it picks the method that was most accurate against what actually emerged.

The Capability Ladder, Revisited

CALLBACK · WHAT EACH OF THE FOUR LAYERS ACTUALLY BOUGHT

Layer	What the actuary gains	The mechanism in the engine
L1	Talks reserving, but does triangle arithmetic in-context, non-reproducible and can hallucinate numbers	No tools: pure next-token reasoning over a triangle pasted as text
L2	Better-guided reasoning: volume-weight LDFs, consider a tail, pick a method family	Seven markdown skills encode the playbook; arithmetic still in-context, still stochastic
L3	Reproducibility: arithmetic is exact and identical every run	Deterministic tools: age-to-age, volume-weighted LDF, CL / BF / GCC given params
L4	Accuracy: method and parameter selection become evidence-based	Backtested ranking engine scored on out-of-sample AvE / CDR
Augmented	Trust: deterministic review + human-locked judgment + sign-off	Review agent re-derives schema and diffs (not an LLM judge); human-in-the-loop gate

Revisiting the Four Questions

CALLBACK · OUR CORE THESIS

Reproducibility	**L3**	Deterministic tools: exact, identical arithmetic on every run
Correctness	**L3 – L4**	Locked development patterns; identical LDFs across CL / BF / GCC
Accuracy	**L4**	Backtested selection against emerged experience
Cost	**Namespace**	State by reference bends the curve: megabytes of compute for bytes of context

PART 6 · TRUST

Trust & the Road Ahead

Earning trust, the actuary’s role, and what comes next.

Gaining Trust

01 Reproducibility

Same inputs should produce same or bounded outputs. A system that gives a different answer every run cannot be relied upon in production.

02 Auditability

The harness logs every turn, every tool call, every input and output. Every reserve has a full, inspectable trail.

03 Explainability

The agent must articulate why it selected a method and what it assumed. The skill file enforces this in every run.

04 Validation

Run the agent on triangles with known outcomes before deploying on a live book. Score it against actual emerged results, not just internal consistency.

05 Sign-off

A qualified actuary reviews and signs off every agent-produced reserve. The agent is a tool, not a signatory. Professional responsibility stays with the human.

06 Governance

Prompts, skill files and tool versions are model artefacts. They require version control, change management and periodic revalidation.

Trust in actuarial agents is not assumed, it is earned through repeated, verifiable performance.

The Human in the Loop

WHERE THE AGENT MUST STOP

Gates fire at the boundary

A gate appears whenever a tool creates a reserve artifact or calls the external engine:

validate · create data · base factor · fit model · compute metrics · select models · run code

Inspection runs free: querying the namespace or describing objects never interrupts the actuary.

What the actuary does

- **Approve:** let the call proceed.
- **Reject:** block it; one rejection halts the rest of the turn.
- **Reject with feedback:** correct the agent's reasoning mid-flight.
- **The one call it cannot skip:** which reserve to take forward, and the agent stops and asks.

Three regimes, one setting: off · relaxed (default) · strict, and every gate, question and decision is written to the audit log.

The Human Actuary and the Road Ahead

ROLE OF THE HUMAN ACTUARY

The agent surfaces insight. The actuary acts on it.

- Sets the brief and reviews the output, the agent does not define its own task
- Challenges unusual method selections and anomalous results
- Holds professional responsibility and signs off every reserve
- Identifies when the agent is confidently wrong, the most critical skill

POTENTIAL NEXT STEPS

1. **Expand to more triangles**
   Test across a broader book, different lines, longer tails, more edge cases. Build confidence in where each layer earns its place.

2. **Cross-model benchmarking**
   Run the same experiment with GPT-4o and Gemini. Does the model provider change the result? What is the cost-accuracy frontier across providers?

3. **CDR and AvE roll-forward**
   Score not just the point-in-time reserve but how it develops. A good reserve should be stable, test CDR and AvE across subsequent diagonals.

4. **Production harness design**
   Move from experiment to workflow. Version-controlled prompts, skill files and tools. Automated validation. Human review interface. Implementing the augmented actuarial agent.

The most dangerous failure is not the agent that fails. It is the agent that fails and the actuary who does not act.

Key Takeaways

1 Reproducibility requires deterministic computation

General reasoning alone is not enough. Reliable, repeatable outputs require deterministic tools underneath.

2 Consistency and accuracy are not the same thing

A system can give the same answer every time and still be wrong. Both dimensions must be evaluated independently, against observed ground truth.

3 Purpose-built narrow AI is where accuracy is won

Generic tools reduce variance. Domain-specific narrow AI improves accuracy. The two are not interchangeable.

4 Augmented actuarial intelligence does not replace the actuary, it enables

The agent reasons, acts and computes. The actuary challenges, decides and signs off. The value is in what they achieve together, not separately.

Thank you

Ron Richman · ronaldrichman@gmail.com

Jacques Bisschoff · bisschoff.jacques@gmail.com

ACTUARIAL SOCIETY OF SOUTH AFRICA
Advancing Actuarial Excellence
Educate • Develop • Uphold