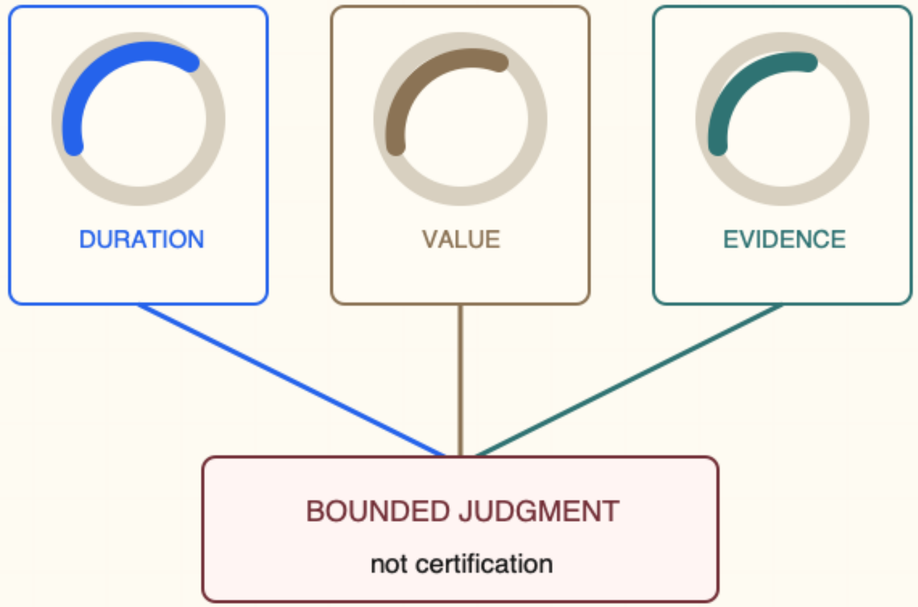
EVALUATION PORTFOLIO

PART 8 00:00-04:32 / synthesis

One Benchmark Cannot Certify an Agent

Stanford CS329A Part 8 - duration, economic value, and evidence quality



CONSTRUCT SEPARATION

PART 8 00:00-04:32 + 61:03-68:20

Agent Progress Has at Least Three Independent Axes

Longer tasks, valuable deliverables, and grounded synthesis are not interchangeable.

01

DURATION

success over human time

time is only a proxy

02

ECONOMIC VALUE

expert preference

sampled work, not whole jobs

03

EVIDENCE QUALITY

grounded synthesis

coverage can remain hidden

BOUNDARY

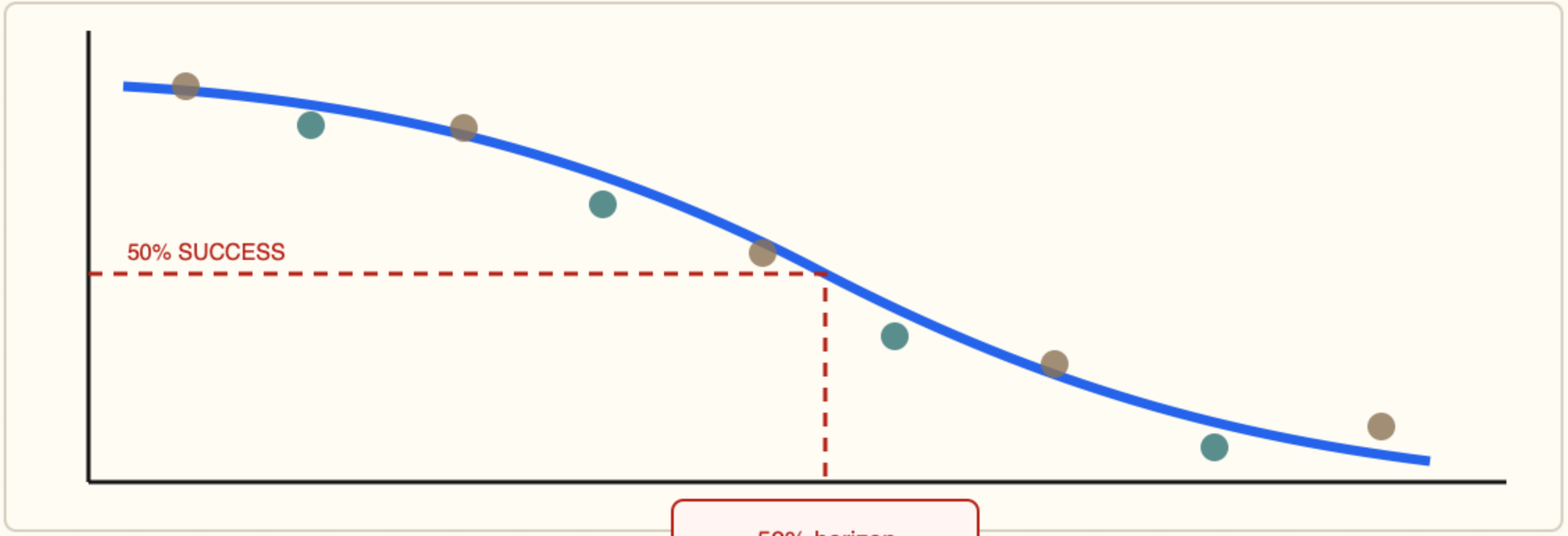
Improvement on one axis does not imply improvement on the other two.

RELIABILITY BOUNDARY

PART 8 04:32-16:55 + METR v4

A Time Horizon Is a Fitted Reliability Boundary

Human completion time is a proxy; the horizon is a threshold intersection.



FIT

success

vs.

human time

BOUNDARY

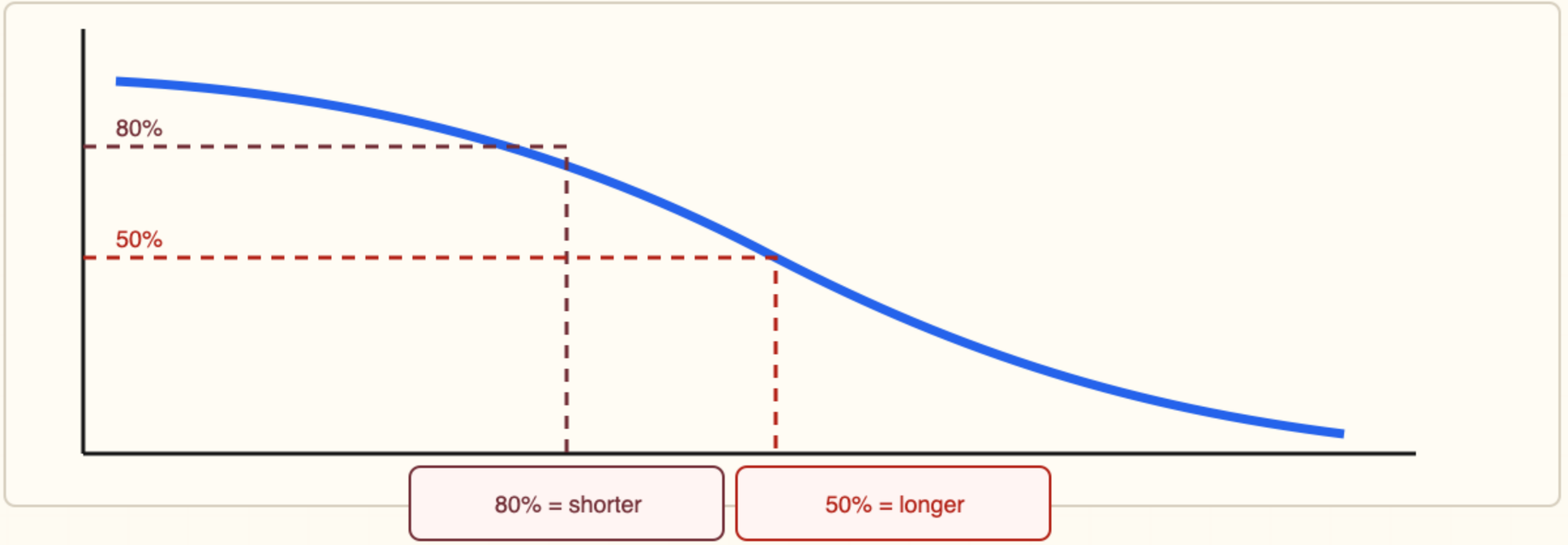
The horizon is a fitted threshold, not the longest observed success.

METRIC SENSITIVITY

PART 8 16:55-22:55 + METR v4

Changing the Reliability Target Changes the Horizon

A 50% horizon and an 80% horizon answer different deployment questions.



SAME

model

DIFFERENT

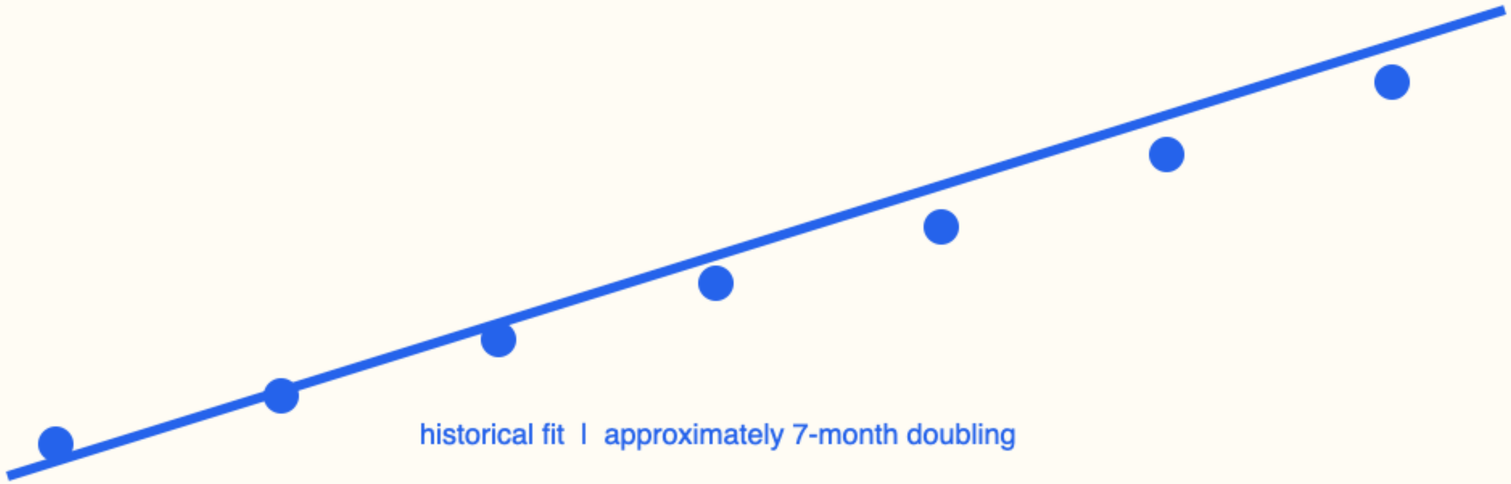
claim

EXTERNAL VALIDITY

PART 8 16:55-22:10 + METR v4

The Seven-Month Trend Is Historical, Not a Deployment Forecast

Observed benchmark scaling does not remove task-distribution or future-trend uncertainty.



UNCERTAINTY

future systems

task distribution

deployment context

NO FORECAST

BOUNDARY

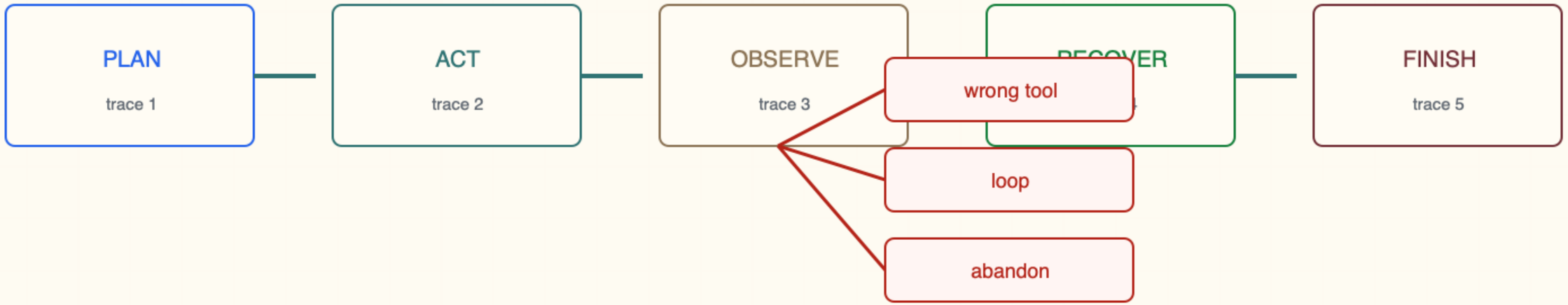
A historical benchmark trend is evidence about the studied period, not a deployment guarantee.

FAILURE ANATOMY

PART 8 22:55-27:21 + METR

Long Tasks Fail Through Trajectories, Not Only Final Answers

Planning errors, recovery failures, loops, and abandonment compound over time.



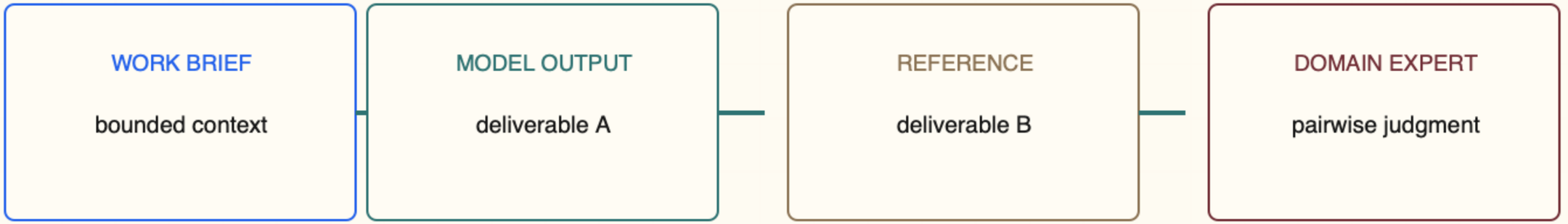
PRESERVE THE TRACE -> LOCALIZE THE FAILURE

ECONOMIC VALUE

PART 8 27:21-33:40 + GDPval v1

GDPval Changes the Question From Duration to Deliverable Value

Experts compare realistic work products rather than only checking a final token or test.



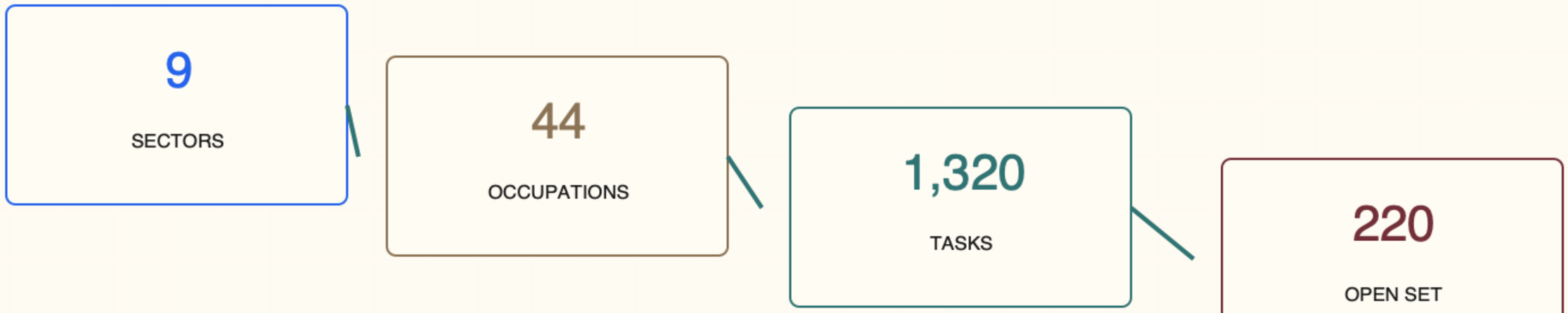
BOUNDARY

GDPval evaluates sampled deliverables under supplied context, not entire jobs or deployment readiness.

SAMPLING FRAME

GDPval Samples Work, Not Entire Occupations

Nine sectors and 44 occupations define a broad but bounded task distribution.



BOUNDARY

Broad sampling remains bounded sampling; tasks do not stand in for every activity in an occupation.

AGGREGATION RISK

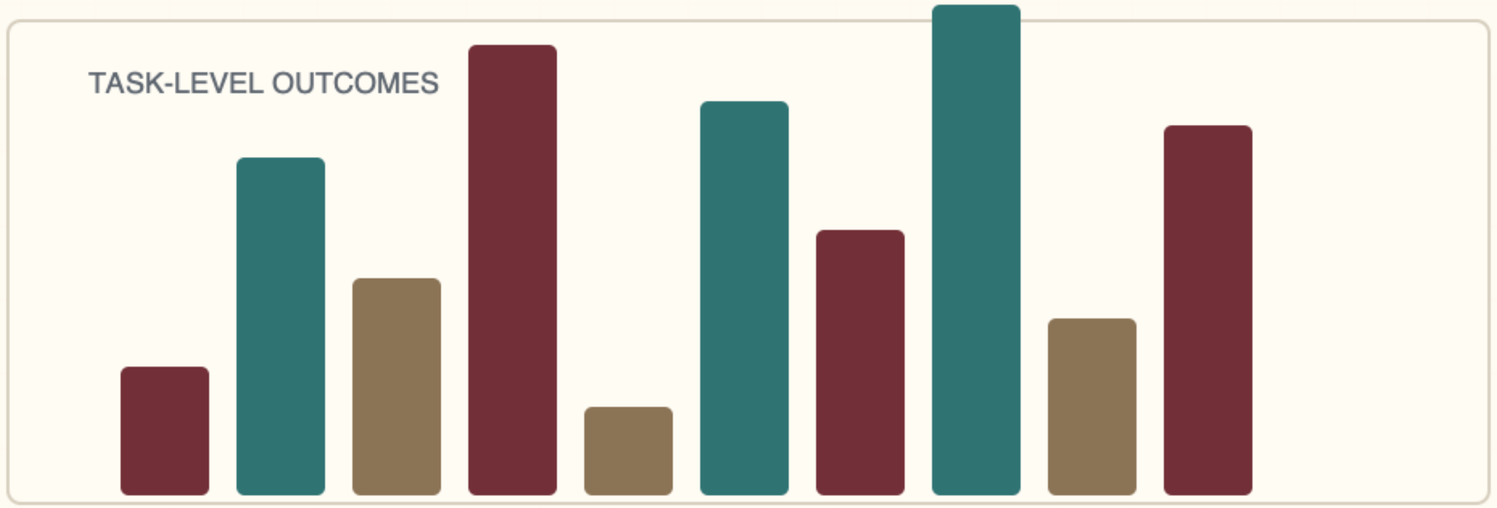
PART 8 35:40-51:34 + GDPval v1

Aggregate Near-Parity Can Hide Uneven Capability

A mean preference result can coexist with sharp task-by-task weaknesses.

AGGREGATE

near parity



same average != same capability everywhere

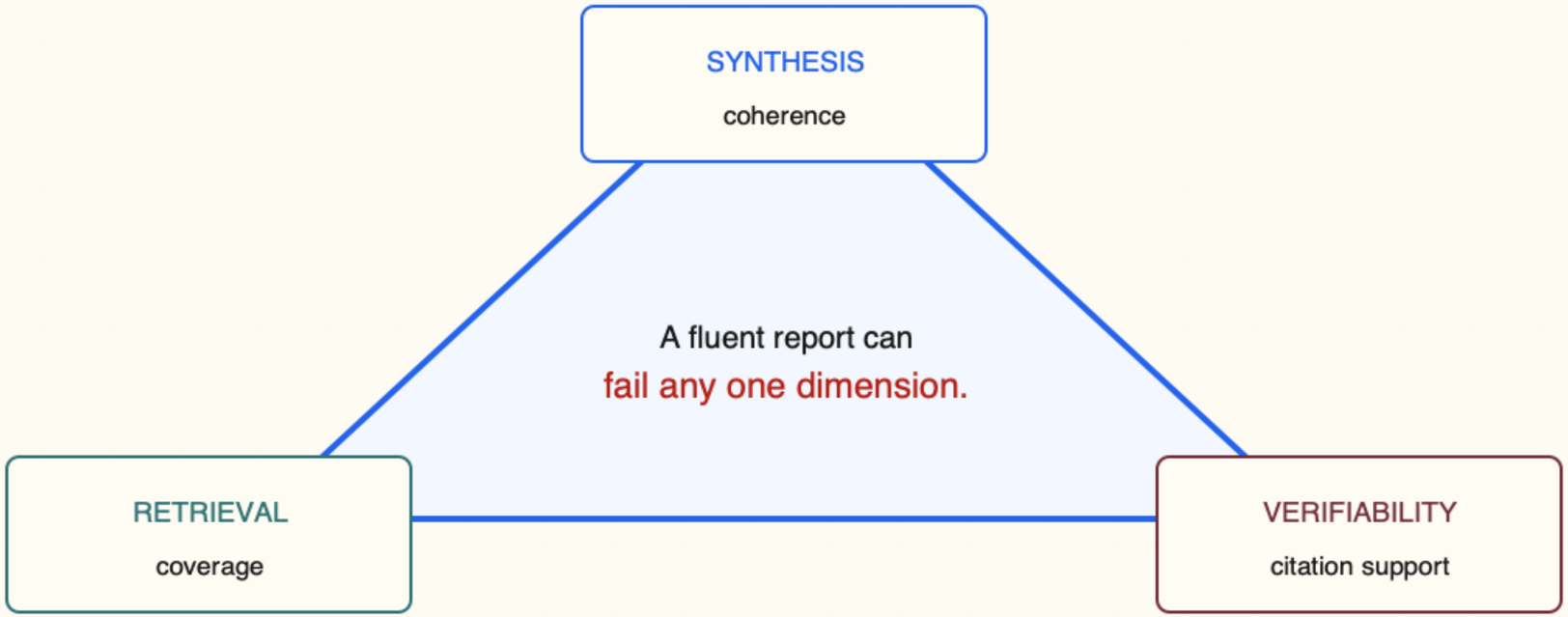
BOUNDARY

Aggregate preference does not establish task-by-task equivalence or occupational replacement.

EVIDENCE QUALITY

Research Synthesis Needs Three Scores, Not One

Synthesis quality, retrieval coverage, and citation support expose different errors.



VERSION DISCIPLINE

PART 8 55:20 + arXiv:2508.20033v2

19% and 31% Are Different Version Snapshots, Not a Trend

The Autumn 2025 lecture and February 2026 arXiv v2 use separately reported aggregates.

LECTURE SNAPSHOT

AUTUMN 2025

19%

no system exceeded lecture aggregate

DO NOT SUBTRACT

different revisions

possibly different aggregation / systems

NOT A TREND

ARXIV V2 SNAPSHOT

2026-02-09

31%

geometric mean in current abstract

BOUNDARY

The two values are separately sourced snapshots and are not directly comparable.

DEPLOYMENT AUDIT

Match Every Deployment Claim to an Evaluation Portfolio

State distribution, threshold, context, evidence, version, and missing dimensions.

CLAIM	DISTRIBUTION	THRESHOLD	CONTEXT	VERSION	BLIND SPOT
long tasks	task set	success target	supplied info	paper / date	what is missing
valuable work	task set	success target	supplied info	paper / date	what is missing
research synthesis	task set	success target	supplied info	paper / date	what is missing

CLAIM SCOPE <= EVIDENCE SCOPE