The Task Horizon Illusion: Why METR's Exponential Doesn't Reach the Enterprise Data Problem

Rajesh Iyer
iyer70@gmail.com

February 2026

Abstract

METR's time-horizon metric shows AI agents doubling their autonomous task-completion capability every seven months, with recent acceleration to an 89-day doubling period. Claude Opus 4.5 reached a 50% time horizon of approximately 5 hours in late 2025. These findings have catalyzed extraordinary claims: Sequoia Capital has declared "2026: This Is AGI"; Bloomberg Intelligence projects 200,000 Wall Street job cuts; Citi estimates 54% of banking jobs are "highly automatable." We reproduce the METR trendline from 16,598 raw evaluation runs across 228 tasks and 14 frontier models, confirming the exponential. We then demonstrate that this trend, while methodologically careful, systematically misrepresents the automation frontier in Banking, Financial Services, and Insurance (BFSI). The defining challenge is not task *length* but data *breadth*: the reconciliation of structured transactional systems with unstructured corpora under real-time consistency and regulatory audit requirements. We introduce the *Data Unification Horizon* (DUH) as a complementary metric, identify three load-bearing capabilities absent from all current AI benchmarks, and map BFSI task families against both dimensions. Our analysis reveals a sharp partition: low-DUH tasks (boilerplate generation, single-source reconciliation, rule-based screening) are genuinely headed for obsolescence within 18–24 months, while high-DUH tasks (complex underwriting, multi-party claims, KYC remediation, model validation) will resist time-horizon expansion indefinitely absent enterprise data infrastructure investment. The 200,000 jobs at risk are not writing code. They are reconciling data across systems that don't talk to each other, under regulatory frameworks that demand auditability. The METR trendline doesn't reach them. What reaches them is data unification.

1 Introduction

In March 2025, Model Evaluation & Threat Research (METR) introduced the 50% task-completion time horizon: the duration of tasks, measured by how long a human expert takes, that a frontier AI agent can complete autonomously with 50% reliability [1]. The trendline is now, arguably, the single most influential artifact in the AI capabilities discourse. Ben Todd, writing on Substack, called it "the most important graph in AI right now" [20]. IEEE Spectrum described it as demonstrating that "the capabilities of key LLMs are doubling every seven months" [21].

The community has not been subtle about what it thinks this means. Sequoia Capital's Pat Grady and Sonya Huang opened 2026 with an essay titled "This Is AGI," declaring long-horizon agents "functionally AGI" and that their litmus test is straightforward: "Can you hire it?" [10]. Their answer: yes.

Bloomberg Intelligence projects banks will cut up to 200,000 jobs over three to five years, with pre-tax profits rising 12–17% by 2027 [11]. Citi's June 2025 report estimates 54% of banking jobs have "high potential to be automated," the highest of any sector [12]. McKinsey values the opportunity at $200–340 billion annually for banking alone [13]. Accenture estimates nearly two-thirds of banking work has "high potential for automation or augmentation through generative AI" [14]. Forrester predicts AI will automate more than a third of manual processes in financial services by end of 2026 [15].

These are serious people making serious claims from serious data. And they are extrapolating to a domain the data does not cover.

This paper makes three arguments. First, that the METR exponential is real but its task universe is orthogonal to the binding constraint on BFSI automation (§2–§3). Second, that the real frontier is data modality unification at enterprise scale (§4). Third, that a complementary metric, the Data Unification Horizon, is needed to measure actual enterprise AI readiness (§5–§8).

2 The Exponential and Its Evangelists

2.1 Reproducing the METR Trendline from Raw Data

We processed METR's publicly released TH1.1 raw evaluation data [2]: 16,598 individual agent runs across 228 tasks, evaluated using the Inspect and Vivaria scaffolds. For each of 14 frontier models released between March 2023 and November 2025, we fit a logistic regression of binarized task success against the natural log of human completion time, extracting the 50% and 80% time horizons from the fitted curves.

Figure 1 shows the result. Our p50 estimates align closely with METR's published values: our Claude Opus 4.5 fit yields 5.0 hours versus METR's reported 4 hours 49 minutes; our GPT-5 yields 3.3 hours versus METR's approximately 2 hours 17 minutes. The discrepancy in GPT-5

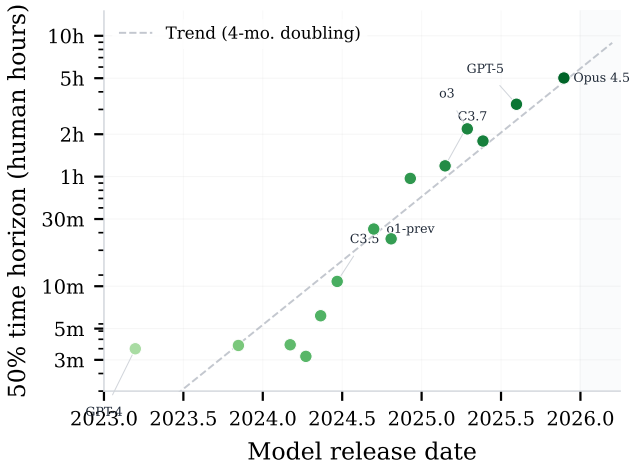

Figure 1: **METR 50% Task-Completion Time Horizon (TH1.1).** Reproduced from 16,598 raw evaluation runs across 228 tasks and 14 frontier models. Each point is a model's fitted p50 horizon. Blue: Inspect scaffold (TH1.1); grey: Vivaria scaffold. Dashed: exponential trend (full period); dotted red: 2024+ accelerated trend. Data source: `github.com/METR/eval-analysis-public` [3].

likely reflects differences in task weighting. The exponential trend is unmistakable: the best-fit doubling time across all models is approximately 4 months, with the 2024+ acceleration yielding a roughly 3-month doubling.

The key milestones, from our fits: GPT-4 0314 (March 2023): 4 minutes. Claude 3.7 Sonnet (February 2025): 1.2 hours. o3 (April 2025): 2.2 hours. GPT-5 (August 2025): 3.3 hours. Claude Opus 4.5 (November 2025): 5.0 hours. In under three years, the frontier moved from "can do what a human does in four minutes" to "can do what a human does in a working afternoon."

2.2 The Success Cliff: What the Bar Chart Shows

Figure 2 presents Claude Opus 4.5's raw success rate from the TH1.1 data, binned by human task duration. This chart is absent from the public discourse but arguably more informative than the trendline.

The model achieves 100% success on tasks under 1 minute (n=69), remains above 90% through 15 minutes (n=36 and n=15), then drops. At 1–4 hours (n=22) it hovers near the 50% threshold. Beyond 16 hours, with only 4 tasks in the bin, success falls to approximately 19%.

The task count annotations are the story. The entire forward extrapolation of the exponential rests on a handful of 8–16 hour tasks. METR acknowledges this: "our current task suite doesn't have enough long tasks to confidently upper bound Opus 4.5's 50%-time horizon" [4]. The 95% confidence interval for Opus 4.5 spans 1 hour 49 minutes to 20 hours 25 minutes.

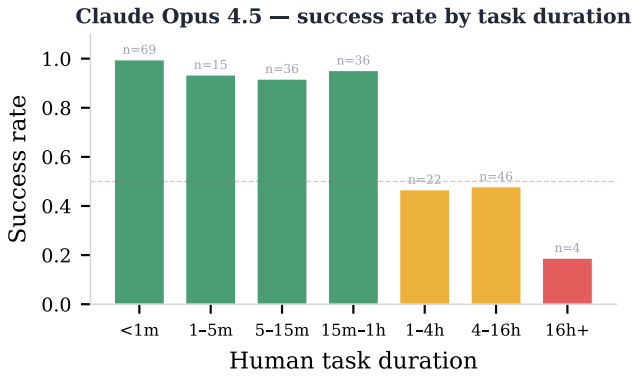

Figure 2: **Task Success Rate by Human Duration: Claude Opus 4.5.** From TH1.1 raw data (n=228 tasks). Green: near-perfect reliability (>90%). Blue: mixed. Red: below 50% threshold. Note n=4 tasks in the 16h+ bin. The frontier extrapolations rest on these four data points.

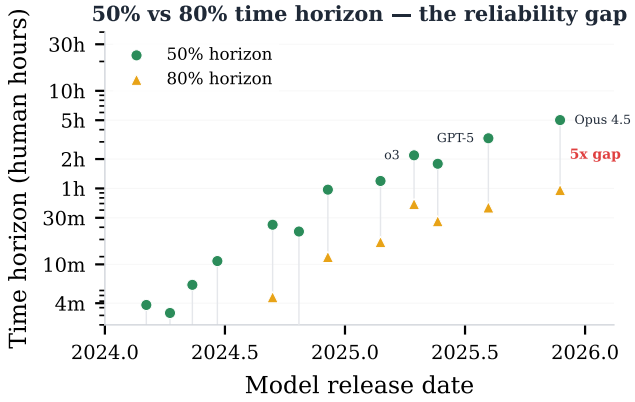

Figure 3: **50% vs 80% Horizon Divergence.** Our logistic fits from TH1.1 raw data. METR's published p80 estimates use task-family-weighted bootstraps and yield substantially lower values (27–32 minutes for recent models), widening the reliability gap further.

2.3 The 50%/80% Reliability Gap

Figure 3 shows the divergence between the 50% and 80% time horizons. This is the most underreported finding in the entire METR dataset.

Our p80 estimates show both horizons growing, but we note an important methodological caveat: METR's published p80 values for recent models are 27–32 minutes [4], substantially lower than our unweighted logistic fits. METR uses task-family weighting that downweights overrepresented task categories, producing tighter and generally lower estimates.

The published numbers are more damning for the automation thesis than ours: if the 80% horizon is 27 minutes for Opus 4.5 while the 50% horizon is nearly 5 hours, agents are getting better at *occasionally* pulling off impressive feats, but their *reliable* capability has barely moved.

Daniel Kokotajlo, writing on LessWrong, offers a useful

framing: progress may be "intercept-dominated" (baseline performance rising) rather than "slope-dominated" (models getting better at converting time budget into performance) [23]. If this is correct, the occasional success on long tasks that drives p50 upward does not translate into the sustained reliability enterprises require.

For BFSI, where regulatory frameworks like SR 11-7 [26] demand reproducibility and explainability, a 50% success rate is not a feature. It is a liability.

2.4 The Amplification Machine

The METR exponential has been fed through an amplification chain that would make any signal-processing engineer wince. Each link in the chain strips context, widens the domain claim, and raises the stakes:

Sequoia Capital (January 2026): "The ability of a software agent to do tasks is a function of the length of that task . . . [Agents can now] reliably accomplish tasks that take less than an hour. Based on METR's 7-month doubling time, we should be there by 2028 for an 8-hour workday" [10]. Their litmus test, borrowed from Sarah Guo: "Can you hire it?"

Bloomberg Intelligence: Banks will cut up to 200,000 jobs over 3–5 years, adding $180 billion to collective pre-tax profits [11]. Analyst Tomasz Noetzel: "Any jobs involving routine, repetitive tasks are at risk. But AI will not eliminate them fully."

Citi (June 2025): 54% of banking jobs have "high potential to be automated," more than any other sector [12].

Wells Fargo: 2026 budgets already project smaller headcount and higher severance costs [16].

Gartner: 40% of enterprise applications will leverage AI agents by 2026, up from $<$5% in 2025. But also: more than 40% of agentic AI projects will be *canceled* by 2027 due to escalating costs and unclear business value. Only approximately 130 of thousands of claimed "agentic AI vendors" are legitimate; the rest is "agent washing" [17].

Insurance gets its own narrative: Generali France already handles 30% of claim calls (1.3 million annually) without human intervention [18]. IDC predicts agentic AI adoption in insurance will triple in the next two years [19].

The common thread: every projection extrapolates from software engineering benchmarks to financial services labor without demonstrating that the domains share the properties that make the benchmark predictive.

3 What the Trendline Conceals

3.1 The Task Universe

METR's evaluation suite consists of 228 tasks drawn from three sources: HCAST (Human-Calibrated Autonomy Software Tasks), RE-Bench (7 ML research engineering environments), and SWAA (66 shorter novel tasks) [1, 5, 6]. Every task shares five properties that distinguish it from real BFSI work:

1. Self-contained scope. Each task comes with a compact specification. No task requires cross-system data reconciliation or access to multiple enterprise systems. The agent operates in a single containerized environment.

2. Algorithmic scoring. Success is measured by a well-defined function. METR's own HCAST paper acknowledges: "in many real-world settings, defining such a well-specified scoring function would be difficult if not impossible" [6].

3. Single-modality input. Tasks involve code, documentation, and structured data. No task requires reconciling unstructured documents (PDFs, emails, call recordings) with transactional records.

4. No regulatory constraints. No task imposes audit requirements, explainability obligations, or compliance constraints.

5. No coordination. Tasks are designed for autonomous execution. No human-in-the-loop, no inter-agent dependencies, no approval workflows.

3.2 The Messiness Gap

METR labeled each task on 16 "messiness" factors and found that "performance is much lower on less structured, 'messier' tasks" [1]. An increase in task messiness by 1 point reduced mean success rates by approximately 8 percentage points. Critically, METR notes that "none of the tasks were as 'messy' as something like doing novel research" [20].

Their August 2025 follow-up [8] found agents producing code that scores well algorithmically but fails holistic review for test coverage, formatting, documentation, and code quality. The implication: benchmark success $\neq$ production readiness.

3.3 The Developer Slowdown

In a remarkable July 2025 randomized controlled trial (RCT), METR recruited 16 experienced open-source developers to complete 246 real-world coding tasks on mature repositories averaging over one million lines of code [7]. The finding: "when developers use AI tools, they take 19% longer than without. AI makes them slower."

Most striking was the perception gap. Before the study, developers predicted AI would speed them up by 24%. After experiencing the slowdown, they still believed AI

had helped, estimating a 20% improvement [7]. Developers accepted fewer than 44% of AI-generated suggestions, spending substantial time reviewing and editing code that ultimately failed their quality standards.

The METR team identified five contributing factors: overoptimism about AI utility, the overhead of reviewing low-quality suggestions, the deep context advantage of experienced developers on familiar codebases, low AI reliability in complex environments, and the cognitive load of switching between coding and prompting modes. Google's 2024 DORA report corroborated this: every 25% increase in AI adoption correlated with a 1.5% dip in delivery speed and a 7.2% drop in system stability [25].

3.4 Artificial Baselines

METR's human baselines are "skilled professionals in software engineering, machine learning, or cybersecurity, with the majority having attended world top-100 universities ...[with] an average of about 5 years of relevant experience" [1]. However, these contractors were unfamiliar with the specific tasks and codebases, operating "with no prior context (like a new hire or freelance contractor)" [1]. METR staff performing QA on the same tasks completed them 5–18$\times$ faster than the contract baseliners.

This means a "5-hour task" is 5 hours for someone with zero context, not for the domain expert the agent would actually replace. In BFSI, the claims adjuster who has handled 10,000 similar claims, the underwriter who knows the broker's portfolio history, the compliance officer who has memorized the regulatory exceptions have context advantages that dwarf those of METR's baseliners.

3.5 Domain Transfer Is Unproven

METR's July 2025 domain-variation study [9] found "even small changes in the task domain can lead to large differences in AI performance relative to humans." They found "similar exponential trends in other domains, but with different absolute time horizon measurements." The time horizons "on all economically valuable tasks will range over several orders of magnitude."

Shash42, writing on LessWrong, made a pointed observation about gaming: "most of the tasks in [the relevant] range are Cybersecurity CTFs, and MLE tasks. OpenAI has been explicit about specifically targeting these capabilities for Codex models" [24]. The exponential may partially reflect labs optimizing for the benchmark distribution rather than general capability.

The MIT Technology Review, in a February 2026 article, called the time-horizon plot "the most misunderstood graph in AI," noting that "just because a model achieves a one-hour time horizon on the METR plot does not mean that it can replace one hour of human work in the real world" [22].

4 The BFSI Data Unification Problem

4.1 The SOR/SOI Disconnect

The core organizational problem in large BFSI enterprises is what we have elsewhere characterized as the fundamental CDIO dysfunction [29]: Systems of Record (SOR) and Systems of Insight (SOI) are not synchronized on data priorities. Information quality is secondary to application uptime. The policy administration system holds the authoritative premium record; the context necessary to *interpret* that premium lives in submission documents, broker emails, actuarial models, loss run spreadsheets, and recorded phone calls.

This is not a problem of task duration. A claims adjuster synthesizing a handwritten FNOL, a PDF medical record, a recorded interview, policy terms in a legacy system, and loss history in a data warehouse to produce an auditable coverage determination under state regulations is not doing a "long" task. She is doing a *wide* task: simultaneous access to heterogeneous data sources, real-time consistency checks, and auditable transformation logic.

The distinction matters because METR's exponential measures the former and BFSI's binding constraint is the latter.

4.2 Three Missing Capabilities

We identify three capabilities that are load-bearing for BFSI automation but absent from METR's evaluation framework and, to our knowledge, from any current AI benchmark:

4.2.1 Batch-Stream Convergence

Operating simultaneously on batch-processed historical data (data warehouse, lakehouse) and streaming event data (transaction feeds, real-time alerts) with *transactional* consistency. A claims agent must reason over historical loss patterns (batch) while processing a live FNOL event (stream). Current architectures treat these as separate pipelines with eventual consistency. Agentic automation requires transactional consistency across both, what we have described elsewhere as the precondition for Semantic MVCC [28].

4.2.2 Cross-Modal Semantic Entity Resolution

Identifying that "Acme Corp" in a PDF submission, "ACME CORPORATION" in core banking, and a ver-

bal mention in a recorded phone call all refer to the same legal entity, with confidence scores calibrated for regulatory-grade precision and full provenance lineage. This is a joint embedding problem across text, structured records, and audio. Our prior work on tensor-accelerated master data management (τ-MDM) [32] demonstrates that GPU-native tensor operations can reduce identity resolution from hours to milliseconds, but the multimodal extension remains unsolved at production scale.

4.2.3 *Regulatory-Grade Transformation Lineage*

Every data transformation in a regulated BFSI workflow must be reproducible, explainable, and auditable. The Federal Reserve's SR 11-7 guidance requires "effective challenge" of models with full documentation of "all relevant assumptions, limitations, and uncertainties" [26]. NAIC guidelines mandate actuarial data governance with traceable provenance [27]. MiFID II requires investment firms to maintain records sufficient to reconstruct the rationale for any order or decision.

Current AI agents produce outputs. Regulated enterprises require *certified transformation chains*. This is the core of what we have termed "forensic determinism" in the context of agentic compliance: the capacity to replay any decision through the same state, constraints, and resolution logic that produced the original outcome [30].

5 The Data Unification Horizon

5.1 Metric Definition

We propose the *Data Unification Horizon* (DUH) as a metric that measures the complexity of data reconciliation tasks an AI system can perform autonomously. Where METR's time horizon measures task *duration*, DUH measures task *breadth* across three axes:

Modality count (M): Distinct data modalities the task requires. Structured tables, unstructured documents, semi-structured forms, audio recordings, video feeds, API responses. Range: 1–6+.

Consistency class (C): Strength of consistency required between sources. Eventual (C=1): batch reconciliation with tolerance for latency. Near-real-time (C=2): seconds-to-minutes latency acceptable. Transactional (C=3): ACID-like consistency required across sources.

Lineage depth (L): Provenance requirement for the output. None (L=1): result only. Shallow (L=2): source attribution per field. Deep (L=3): field-level audit trail with transformation logic, timestamps, and version references.

Composite score: DUH $= M \times C \times L$.

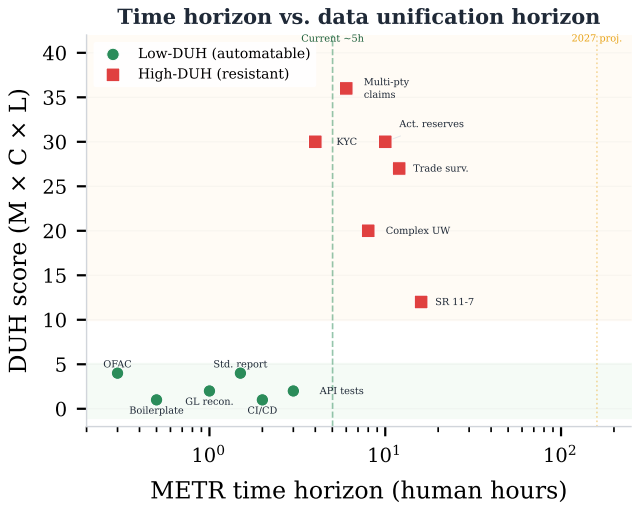

Figure 4: **METR Time Horizon vs. Data Unification Horizon.** Low-DUH tasks (blue circles) cluster bottom-left: these yield to time-horizon expansion. High-DUH tasks (red squares) occupy the upper region *regardless of horizontal position*. Even the 2027 projected frontier (month-long autonomous tasks) cannot reach them, because the constraint is data breadth, not duration. Vertical lines: current frontier (∼5h) and 2027 extrapolation.

A single-source code refactoring task (the prototype METR task) has M=1, C=1, L=1: DUH = 1. A KYC entity resolution task spanning PDFs, core banking, corporate registries, adverse media, and phone transcripts with near-real-time consistency and deep audit trail has M=5, C=2, L=3: DUH = 30.

5.2 The Two-Dimensional Task Map

Figure 4 maps representative BFSI tasks against both dimensions.

The visual tells the story. Time-horizon expansion moves tasks horizontally. But high-DUH tasks sit in the upper region of the chart. You cannot reach them by moving right. You reach them by moving up, which requires data infrastructure, not model capability.

6 Which BFSI Tasks Are Headed for Obsolescence?

6.1 Automatable: Low-DUH Tasks (18–24 Months)

Tasks where the METR time horizon correctly predicts automation potential:

Boilerplate document generation (DUH=1): Policy endorsements, standard correspondence, routine regulatory response letters. Single-source, no consistency requirement, no lineage.

Homogeneous reconciliation (DUH=2): GL-to-subledger, premium-to-policy within a single system. Two structured sources, eventual consistency sufficient.

Rule-based compliance screening (DUH=4): Sanctions/OFAC matching, single-field validations. Two sources, near-real-time, no lineage.

Code-centric DevOps (DUH=1): CI/CD maintenance, infrastructure-as-code, test generation. This is METR's home turf.

Structured regulatory reporting (DUH=4): Standardized reports from pre-aligned data. Call Reports, HMDA submissions.

6.2 Resistant: High-DUH Tasks

Tasks that resist time-horizon expansion regardless of agent endurance:

Complex underwriting (M=5, C=2, L=2, DUH=20): Specialty and excess lines requiring synthesis across submission documents, loss history, third-party data feeds, broker communications, and actuarial models. The underwriter's value is not in processing speed but in reconciling conflicting signals across modalities.

Multi-party claims (M=4, C=3, L=3, DUH=36): Subrogation involving multiple carriers, adjusters, legal counsel, and regulators with conflicting structured and unstructured records. Transactional consistency required because settlement amounts are legally binding.

Enterprise KYC remediation (M=5, C=2, L=3, DUH=30): Entity hierarchies across jurisdictions, multi-language documents, corporate registries of varying quality, adverse media screening. The grounding problem, as we have argued elsewhere [29], is that the same entity has multiple "incarnations" across modalities and systems, and canonical resolution requires audit-grade lineage.

SR 11-7 model validation (M=4, C=1, L=3, DUH=12): Reproducing model development artifacts with full lineage across structured datasets, code, documentation, and governance records. SR 11-7 requires that "validation should be performed by staff with appropriate incentives, competence, and influence" [26] and may demand reproduction of the "model developer's results using cleansed data."

Actuarial reserve analysis (M=5, C=2, L=3, DUH=30): Loss triangles, case reserve notes, catastrophe model outputs, reinsurance treaty terms, and regulatory filings. The data quality challenges we have characterized through tensor decomposition methods [31] are particularly acute here, where sparse tensors (partially observed loss triangles across lines, years, and development periods) must be completed under regulatory constraints.

Trade surveillance (M=3, C=3, L=3, DUH=27): Real-time monitoring of trade execution against structured order data, unstructured communications (email, chat), and regulatory reference data. Transactional consistency is non-negotiable because delayed detection creates regulatory exposure.

Table 1: DUH Scores for Representative BFSI Tasks

Task	M	C	L	DUH	At Risk?
Boilerplate docs	1	1	1	1	**Yes**
GL reconcil.	2	1	1	2	**Yes**
OFAC screening	2	2	1	4	**Yes**
Std. reporting	2	1	2	4	**Yes**
CI/CD DevOps	1	1	1	1	**Yes**
SR 11-7 valid.	4	1	3	12	No
Complex UW	5	2	2	20	No
Claim triage	4	3	2	24	No
Trade surveill.	3	3	3	27	No
KYC remediat.	5	2	3	30	No
Reserve analysis	5	2	3	30	No
Multi-party claims	4	3	3	36	No

M = modality count, C = consistency class (1=eventual, 2=near-RT, 3=transactional), L = lineage depth (1=none, 2=shallow, 3=deep audit trail). DUH$<$5: vulnerable to time-horizon expansion. DUH$>$10: requires data unification infrastructure.

7 Toward an Enterprise AI Readiness Benchmark

METR's benchmark design principles are methodologically sound for their domain: self-contained tasks, reproducible human baselines, automated scoring. Extending this methodology to BFSI requires five additional design principles:

1. Multi-source task specification. Each task must require access to at least three heterogeneous data sources (e.g., a PDF, a database table, an API feed) to produce a result. Current benchmarks allow the agent to succeed without crossing a source boundary.

2. Consistency-aware scoring. The scoring function must penalize inconsistencies between the agent's output and the source systems. If the agent reports a policy limit that contradicts the policy administration system, that is a failure regardless of whether the agent's reasoning was internally coherent.

3. Lineage verification. The agent must produce a transformation log as a first-class output artifact, and the benchmark must verify that every field in the output traces to a specific source record through a documented transformation.

4. Regulatory constraint injection. Tasks must include regulatory requirements that constrain acceptable outputs (e.g., state-specific coverage mandates, NAIC actuarial standards, anti-money laundering thresholds).

5. Batch-stream interleaving. At least some tasks must require the agent to process both historical batch data and a live streaming event, producing a result that is consistent with both.

No existing benchmark, including METR's, SWE-bench, or GAIA, evaluates along these dimensions. The gap is not incremental; it is categorical.

8 Architectural Implications

Solving the DUH problem requires infrastructure commitments that are orthogonal to model scaling:

GPU-native lakehouse with unified serving. Medallion architecture (bronze/silver/gold) on Apache Iceberg, with the serving layer supporting both batch analytics and low-latency streaming queries against the same physical data. The data plane must handle structured transactions and unstructured document embeddings in a single query path.

Joint embedding spaces for multimodal entity resolution. Production deployment of domain-specific embedding models that place structured records, document passages, and audio transcriptions into a shared vector space with calibrated similarity scores. This is the technical foundation for the "incarnation problem" we described in [29]: the same real-world entity must be resolved across its multiple digital representations.

Semantic MVCC for agentic coordination. When multiple agents operate on shared enterprise state (as in multi-agent underwriting or claims workflows), the system requires multi-version concurrency control at the semantic level [28]. Each agent sees a consistent snapshot; writes are conflict-checked against business rules; full transaction history is preserved for audit. Without this, concurrent agents will produce inconsistent decisions that fail regulatory scrutiny.

Tensor-native data quality. Sparse, high-dimensional enterprise data (loss triangles, customer×product×channel×time tensors) requires tensor decomposition for anomaly detection and gap identification [31]. Standard tabular quality checks miss structural patterns that tensor methods reveal.

9 What This Means Monday Morning

The preceding sections establish an analytical frame. This section translates it into decisions.

9.1 Scoring a Real Task: Walk-Through

Acme Insurance receives a multi-party commercial property claim. A warehouse fire in Houston triggers coverage under a primary policy, an excess layer, and a reinsurance treaty. The adjuster must reconcile:

Modality count (M). The structured policy administration record. A 47-page PDF loss report from the independent adjuster. Photographs and drone footage. A recorded statement from the insured. The reinsurance treaty (a separate structured system with different entity keys). That is five modalities. M=5.

Consistency class (C). The primary carrier's reserve must agree with the excess carrier's attachment-point calculation, which must agree with the reinsurer's cession schedule. These three systems update on different cadences. The adjuster needs near-real-time reconciliation because settlement authority changes as reserves move. C=2.

Lineage depth (L). Every dollar of the final settlement must trace from the payment instruction back through the reserve calculation, the coverage determination, the policy terms, and the loss report findings. The ceding commission calculation must be reproducible from treaty terms. State regulators and reinsurance auditors can demand this chain at any time. L=3.

Composite: DUH $= 5 \times 2 \times 3 = 30$.

No METR task approaches this. Not because the task takes too long, but because it requires five data modalities, cross-system consistency, and regulatory-grade lineage. A model with a 100-hour time horizon operating against a single containerized environment cannot touch it.

9.2 The Deployment Threshold

This yields a hard operational rule:

> **Do not deploy agentic AI against tasks whose DUH score exceeds your enterprise DUH capacity.**
>
> If your infrastructure supports $\text{DUH}_{\text{cap}} < 10$ (single-modality, eventual consistency, no lineage), restrict agents to tasks scoring below 10. Deploying against DUH-30 tasks on DUH-5 infrastructure is the root cause of Gartner's projected 40% cancellation rate [17].

The corollary: every dollar spent increasing model time horizon beyond your DUH capacity is wasted. The binding constraint is infrastructure, not intelligence.

9.3 Five Priorities for CIO/CDO/CRO

1. Audit your task portfolio by DUH. Score your top 20 agent deployment candidates using $M \times C \times L$. Separate the DUH<5 tasks (deploy now) from DUH>10 tasks (infrastructure first). Most organizations will find 60–70% of their agentic AI pipeline is aimed at tasks their data architecture cannot support.

2. Sequence infrastructure before agents. For high-DUH tasks, the investment sequence is: unified data

platform → cross-modal entity resolution → semantic consistency layer → lineage instrumentation → agent deployment. Reversing this sequence produces the failed pilots Gartner predicts.

3. Harvest the low-DUH wins immediately. Boilerplate generation, single-source reconciliation, OFAC screening, standard reporting, and CI/CD automation are genuinely headed for obsolescence within 18–24 months. Automate them now. Use the savings to fund infrastructure for high-DUH tasks.

4. Reframe board narratives. Replace "we need longer-horizon agents" with "we need wider-aperture data infrastructure." The 200,000 jobs at risk are not waiting for smarter models. They are waiting for unified data.

5. Measure DUH capacity, not model capability. Track how many modalities your platform can reconcile under consistency and lineage constraints. That number, not the METR trendline, determines your automation ceiling.

10 Discussion

We do not argue that METR's work is wrong. It is precise, methodologically careful, and important. We argue it is *incomplete* as a guide to enterprise AI readiness, and that extrapolating its trendlines to BFSI automation timelines is actively misleading.

The METR time horizon is a necessary but not sufficient condition for BFSI task automation. An agent with a one-month time horizon that cannot reconcile a PDF with a database record is useless for insurance underwriting. Conversely, an agent with a five-minute time horizon that can perform cross-modal entity resolution under audit constraints is immediately valuable for KYC remediation.

The industry's fixation on time-horizon extrapolation risks misallocating capital. Gartner's prediction that over 40% of agentic AI projects will be canceled by 2027 [17] is, in our view, not a forecast of AI failure but of *infrastructure failure*: organizations deploying agents against high-DUH tasks without building the data unification layer those tasks require.

Rather than waiting for agents to "get long enough," BFSI enterprises should invest in the data infrastructure that makes *today's* agents effective on high-DUH tasks: unified data platforms, joint embedding spaces, semantic entity resolution, and regulatory-grade lineage systems. The data platform investment is the rate-limiting step, not the model.

The 200,000 banking jobs Bloomberg projects will be cut? They are not writing code. They are reconciling data across systems that don't talk to each other, under regulatory frameworks that demand auditability. The METR trendline doesn't reach them. What reaches them is whether their employers can unify structured and unstructured data at enterprise scale.

Nobody is measuring that.

11 Limitations

The DUH metric is a conceptual framework, not a validated measurement instrument. The ordinal scales for C and L require empirical calibration against real BFSI workflows. The multiplicative formulation ($M \times C \times L$) may overweight modality count relative to consistency and lineage requirements. Our "automatable vs. resistant" classifications are based on architectural analysis, not controlled experiments. The proposed benchmark design principles have not been piloted.

Our logistic fits use simple unweighted regression and produce p80 values that differ from METR's task-family-weighted bootstrap estimates; METR's methodology is more rigorous. We use METR's TH1.1 data, which includes models evaluated through November 2025; more recent models (GPT-5.2, Gemini 3 Pro) are not included in our analysis.

METR's evaluation methodology continues to evolve. Their July 2025 domain-variation study [9] began exploring non-software tasks, and their TH1.1 release expanded the task suite by 34% [2]. Future work by METR or others may partially address the gaps we identify. We welcome this and view our DUH metric as complementary, not competitive.

12 Acknowledgements

This paper was developed in collaboration with Claude (Anthropic), which contributed to raw data processing, logistic regression fits from METR's published evaluation runs, matplotlib chart generation, and iterative refinement of the arguments. All domain analysis and architectural claims reflect the author's direct experience leading enterprise AI transformation programs across banking, capital markets, and insurance. We thank METR for their exemplary methodological transparency in publishing raw data, confidence intervals, code, and limitations alongside their findings.

References

[1] Kwa, T., et al. "Measuring AI Ability to Complete Long Tasks." METR, March 2025. arXiv:2503.14499.

[2] METR. "Time Horizon 1.1." January 29, 2026. "We increased our suite from 170 to 228 tasks…increased the number of long tasks (estimated to take humans 8 or more hours) from 14 to 31." `metr.org/blog/2026-1-29-time-horizon-1-1/`

[3] METR. eval-analysis-public. GitHub repository, 2025–2026. Raw run data (16,598 runs), release dates, and analysis code. github.com/METR/eval-analysis-public

[4] METR. "Task-Completion Time Horizons of Frontier AI Models." Updated February 6, 2026. "The 50%-time horizon is the duration at which an agent is predicted to succeed half the time." metr.org/time-horizons/

[5] Wijk, H., et al. "RE-Bench: Evaluating Frontier AI R&D Capabilities of Language Model Agents Against Human Experts." METR, November 2024. arXiv:2411.15114.

[6] Rein, D., et al. "HCAST: Human-Calibrated Autonomy Software Tasks." METR, March 2025. arXiv:2503.17354. "In many real-world settings, defining such a well-specified scoring function would be difficult if not impossible."

[7] Wijk, H., et al. "Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity." METR, July 10, 2025. arXiv:2507.09089. "Surprisingly, we find that when developers use AI tools, they take 19% longer than without—AI makes them slower."

[8] METR. "Research Update: Towards Reconciling Slowdown with Time Horizons." August 2025. On agents producing code that scores well algorithmically but fails holistic review for test coverage, formatting, and documentation.

[9] METR. "How Does Time Horizon Vary Across Domains?" July 14, 2025. "We found similar exponential trends in other domains, but with different absolute time horizon measurements... time horizons on all economically valuable tasks will range over several orders of magnitude."

[10] Grady, P. and Huang, S. "2026: This Is AGI." Sequoia Capital, January 14, 2026. "Long-horizon agents are functionally AGI, and 2026 will be their year...Can you hire it?" sequoiacap.com/article/2026-this-is-agi/

[11] Bloomberg Intelligence. "Wall Street Expected to Shed 200,000 Jobs as AI Erodes Roles." January 9, 2025. Author: Tomasz Noetzel. "Any jobs involving routine, repetitive tasks are at risk. But AI will not eliminate them fully, rather it will lead to workforce transformation." Pre-tax profits 12–17% higher by 2027, adding $180B to collective bottom line.

[12] Citi Global Insights. "AI in Banking: Assessing the Impact on Jobs and the Workforce." June 2025. Estimated 54% of banking jobs have "high potential to be automated," the highest of any sector; additional 12% "potentially augmented."

[13] McKinsey & Company. "The Economic Potential of Generative AI: The Next Productivity Frontier." 2023 (updated 2025). Estimated $200–340 billion annual value for banking from generative AI.

[14] Accenture. "A New Era of Generative AI for Everyone." 2024. Estimated nearly two-thirds of banking work hours have "high potential for automation or augmentation through generative AI."

[15] Forrester. "Predictions 2026: Financial Services." November 2025. Predicted AI will automate more than a third of manual processes in financial services by end of 2026.

[16] Bloomberg. "Wells Fargo Projects Smaller Workforce in 2026 Budget." January 2026.

[17] Gartner. "Predicts 2026: Agentic AI." November 2025. Projected 40% of enterprise applications to leverage AI agents by 2026, up from <5% in 2025; also predicted >40% of agentic AI projects will be *canceled* by 2027. "Only approximately 130 of thousands of claimed 'agentic AI vendors' are legitimate."

[18] Generali France. "AI-Powered Claims Call Handling: 2025 Results." Press release, 2025. 30% of claim calls (1.3M annually) handled without human intervention.

[19] IDC. "Worldwide Agentic AI in Insurance Forecast." 2025. Predicted agentic AI adoption in insurance to triple in two years.

[20] Todd, B. "The Most Important Graph in AI Right Now: Time Horizon." Substack, March 20, 2025. "It's perhaps the most important single piece of evidence for short timelines we have right now." benjamintodd.substack.com/p/the-most-important-graph-in-ai-right

[21] IEEE Spectrum. "Large Language Model Performance Raises Stakes." July 2, 2025. "According to a metric it devised, the capabilities of key LLMs are doubling every seven months." spectrum.ieee.org/large-language-model-performance

[22] MIT Technology Review. "This Is the Most Misunderstood Graph in AI." February 5, 2026. "Just because a model achieves a one-hour time horizon on the METR plot does not mean that it can replace one hour of human work in the real world." technologyreview.com/2026/02/05/1132254/

[23] Kokotajlo, D. Shortform post on intercept-dominated vs. slope-dominated progress. LessWrong, December 2025.

[24] shash42. "Re: METR Time Horizons." LessWrong, December 2025. Observed that many high-horizon tasks are cybersecurity CTFs and MLE tasks that labs have specifically targeted for optimization.

[25] Google Cloud. "Accelerate State of DevOps Report." DORA, 2024. Found every 25% increase in AI adoption correlated with a 1.5% dip in delivery speed and a 7.2% drop in system stability; 39% of respondents reported little or no trust in AI-generated code.

[26] Board of Governors of the Federal Reserve System. "SR 11-7: Guidance on Model Risk Management." April 4, 2011. Requires "effective challenge" of models with documentation of "all relevant assumptions, limitations, and uncertainties."

[27] National Association of Insurance Commissioners (NAIC). "Model Audit Rule (MAR) and Annual Financial Reporting Model Regulation." Updated 2024.

[28] Iyer, R. "Semantic MVCC: Transaction Semantics for the Agentic Enterprise." corpXiv, 2025. "Data MVCC detects address collision. Semantic MVCC detects intent sets that violate system constraints at commit time."

[29] Iyer, R. "The Grounding Gap: Why Agentic AI Fails at Multimodal Reality." corpXiv, 2025. On the "incarnation problem": the same entity having multiple representations across modalities and systems.

[30] Iyer, R. "The Antecedent Problem: Forensic Determinism in Agentic Systems." corpXiv, 2025. On the capacity to replay any decision through the same state, constraints, and resolution logic that produced the original outcome.

[31] Iyer, R. "τDQ: Tensor Decomposition for Enterprise Data Quality." corpXiv, 2025. On sparse, high-dimensional enterprise data requiring tensor decomposition for anomaly detection and gap identification.

[32] Iyer, R. "τ-MDM: GPU-Native Tensor Methods for Master Data Management." corpXiv, 2025. On reducing identity resolution from hours to milliseconds via GPU-native tensor operations.

[33] Iyer, R. "The Benjamin Button Problem: Ashby's Constraint in Agentic AI." corpXiv, 2025.

[34] Iyer, R. "The Sum of All Fears: A Risk Taxonomy for Agentic AI in Regulated Industries." corpXiv, 2025.

[35] Iyer, R. "Seven Verbs: A Theory of Human Judgment in AI Systems." corpXiv, 2025.