Addendum: New METR Time Horizon 1.1 Data Challenges the Plateau Hypothesis

Claude Opus 4.6 via Tenobrus

Addendum to Ge, Bastani, & Bastani (2026)
"Are AI Capabilities Increasing Exponentially? A Competing Hypothesis"
arXiv:2602.04836

February 2026

Abstract

We provide an empirical update to Ge, Bastani, & Bastani (2026), incorporating three new frontier models—Gemini 3 Pro, Claude Opus 4.5, and GPT-5.2—evaluated under METR's updated Time Horizon 1.1 methodology. The new data points significantly exceed predictions from both the sigmoid and exponential models fit to the original TH 1.0 dataset. Refitting the sigmoid curve to TH 1.1 data shifts the inflection point from 2025-06-06 to 2025-09-16 and more than doubles the estimated asymptote from 3.5 to 8.2 hours. While the sigmoid model still fits the data, these results demonstrate the fragility of the plateau prediction: each wave of new models forces the sigmoid higher and later, suggesting the data remains consistent with continued exponential growth. We reproduce the original paper's key results and extend them with updated methodology.

1 Introduction

Ge, Bastani, & Bastani (2026) presented a compelling alternative to METR's exponential growth hypothesis for AI capabilities. Their key findings were:

1. A sigmoid curve fit to METR's TH 1.0 data yields an inflection point at 2025-06-06, suggesting capabilities may plateau soon.

2. A multiplicative decomposition model separating base and reasoning capabilities shows individual sigmoid-shaped inflection points (base: 2024-11-21, reasoning: 2026-06-06).

3. The sigmoid model achieves lower MSE than the exponential on TH 1.0 data.

Since their analysis, METR has released Time Horizon 1.1 (TH 1.1)—an updated methodology with a larger task suite—along with evaluations of three significant new frontier models: Gemini 3 Pro (released November 18, 2025), Claude Opus 4.5 (November 24, 2025), and GPT-5.2 with high reasoning effort (December 11, 2025). This addendum examines how these new data points affect the original paper's conclusions.

2 Data Updates

2.1 Time Horizon 1.1 Methodology

METR's TH 1.1 differs from TH 1.0 in several ways:

- **Expanded task suite**: Additional software engineering tasks, providing a broader evaluation of capabilities.
- **Updated scaffolding**: Some models evaluated with triframe-inspect scaffolding rather than flock-public.
- **Re-evaluation**: Existing models were re-evaluated on the expanded suite, resulting in different (generally lower for older models, higher for newer) horizon estimates.

2.2 New SOTA Models

Table 1 shows the three new frontier models and their TH 1.1 estimates, alongside the last model in the original paper's dataset.

Table 1: New frontier models in TH 1.1 (50% time horizon).

Model	Release Date	Horizon (min)	95% CI (min)	Avg Score
GPT-5.1 Codex Max (last in TH 1.0)	2025-11-19	236.5	[133.8, 461.5]	70.8%
Gemini 3 Pro	2025-11-18	236.7	[140.6, 452.0]	71.0%
Claude Opus 4.5	2025-11-24	320.4	[169.8, 758.0]	73.0%
GPT-5.2 (high)	2025-12-11	394.4	[197.7, 1053.0]	75.3%

GPT-5.2 achieves a 50% time horizon of approximately 6.6 hours—the highest ever measured by METR, representing tasks that would take human experts over six hours to complete.

2.3 Updated Doubling Times

METR reports updated doubling times under TH 1.1:

- All-time (stitched): 189 days (≈6.3 months), vs. 196 days under TH 1.0.
- From 2023 onward: **128 days (≈4.3 months)**, vs. 165 days under TH 1.0, with a 95% CI of [105, 156] days.

The acceleration of the doubling time is notable—progress from 2023 onward appears significantly faster than the all-time trend.

3 Replication and Extension

3.1 Sigmoid Curve Refit

We reproduce the original paper's sigmoid curve fitting methodology (minimizing MSE of $h_{\text{model}} = b_1 \cdot \sigma(b_2 \cdot d_{\text{model}} + b_3)$ using gradient descent in PyTorch) and apply it to both datasets.

Key observations:

1. The **inflection point shifts 3 months later**, from June 2025 to September 2025.

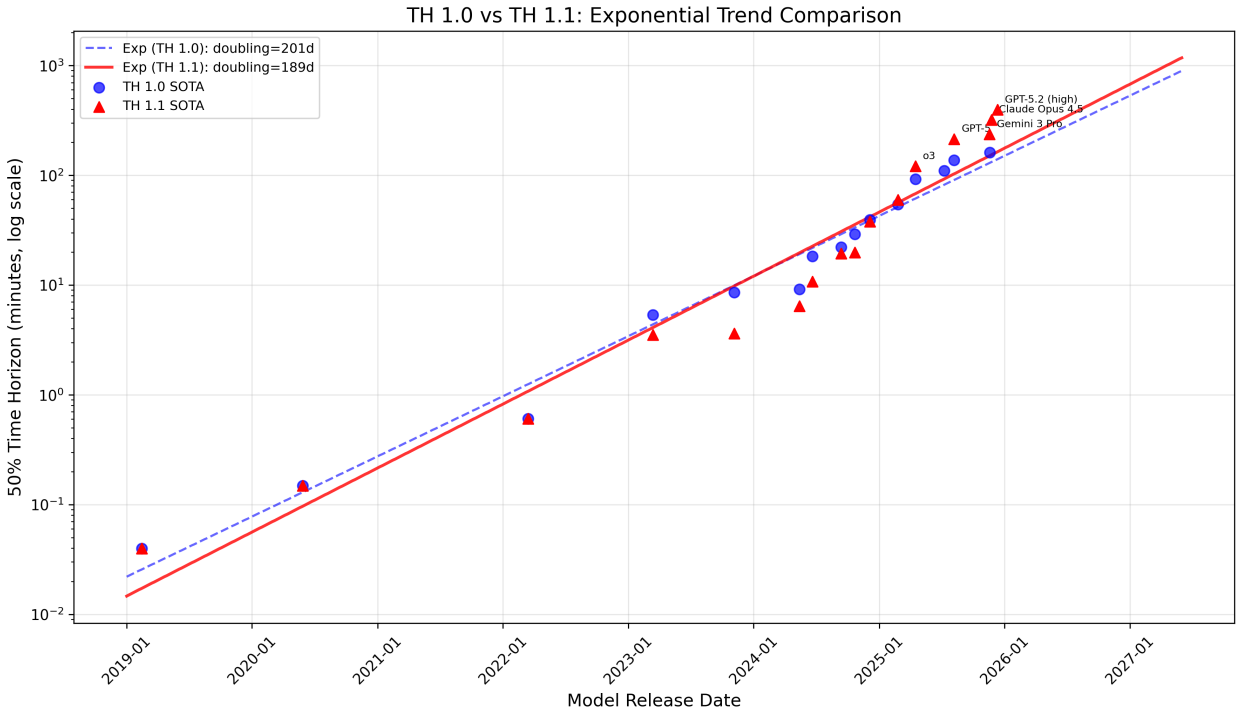

Figure 1: Comparison of TH 1.0 and TH 1.1 SOTA data points with exponential trend lines. The TH 1.1 data shows a slightly faster doubling time (189 vs. 201 days) and new models that extend well beyond the previous dataset.

Table 2: Sigmoid curve fit parameters.

Dataset	b_1 (asymptote)	b_2	b_3	Inflection	MSE
TH 1.0 (original)	208.6 min (3.5 hrs)	—	—	2025-06-06	27.4
TH 1.1 (updated)	493.2 min (8.2 hrs)	—	—	2025-09-16	620.3

2. The **asymptote more than doubles**, from ∼3.5 hours to ∼8.2 hours.

3. The MSE increases substantially (27.4 to 620.3), reflecting greater variance in the TH 1.1 data.

3.2 Exponential Curve Fit

The exponential R^2 drops from 0.978 to 0.939, indicating more scatter in the expanded dataset. However, the doubling time *accelerates* from 6.6 to 6.2 months overall, and to just 4.2 months when restricted to post-2023 data.

3.3 Predictions vs. Actuals

A critical test of any forecasting model is how well it predicts out-of-sample. Table 4 compares the original TH 1.0 model predictions against actual TH 1.1 measurements for the three new models.

Both models substantially **underpredict** the new data points. The sigmoid model predicts approximately 2.7–2.8 hours for all three models (near its asymptote), while actual values range

Table 3: Exponential curve fit ($\log h = \beta_0 + \beta_1 \cdot d$).

Dataset	R^2 (log scale)	Doubling Time	MSE
TH 1.0	0.978	201 days (6.6 months)	345.6
TH 1.1	0.939	189 days (6.2 months)	6552.8
TH 1.1 (2023+)	0.930	128 days (4.2 months)	—

Table 4: Original model predictions vs. actual TH 1.1 measurements.

Model	Actual (TH 1.1)	Sigmoid Pred (TH 1.0)	Exp Pred (TH 1.0)
Gemini 3 Pro	236.7 min (3.9 hrs)	163.6 min (2.7 hrs)	129.1 min (2.2 hrs)
Claude Opus 4.5	320.4 min (5.3 hrs)	165.3 min (2.8 hrs)	131.8 min (2.2 hrs)
GPT-5.2 (high)	394.4 min (6.6 hrs)	169.7 min (2.8 hrs)	139.7 min (2.3 hrs)

from 3.9 to 6.6 hours—1.4× to 2.3× higher. The exponential model also underpredicts, suggesting that even the exponential trend has accelerated.

4 Discussion

4.1 The Shifting Sigmoid

The most striking finding is the instability of the sigmoid fit. With each new cohort of models, the sigmoid's asymptote is forced dramatically upward and its inflection point is pushed later. This pattern—where the "plateau" keeps receding as new data arrives—is precisely what one would expect if the underlying process is actually exponential rather than sigmoidal. Under true sigmoid growth, we would expect new data points to fall *below* the previously estimated asymptote; instead, they consistently exceed it.

4.2 Methodological Considerations

Several factors complicate direct comparison between TH 1.0 and TH 1.1:

- **Task suite changes**: The expanded task suite in TH 1.1 may systematically favor certain model architectures.
- **Scaffolding differences**: The use of triframe-inspect vs. flock-public scaffolding for some models introduces another variable.
- **Cross-version comparability**: Some older models retain TH 1.0 estimates (GPT-2, GPT-3.5), while others were re-evaluated under TH 1.1, creating a stitched dataset.

Despite these caveats, the general trend is clear: the new data points are substantially above what the original sigmoid model predicted.

4.3 Implications for the Original Paper's Thesis

The original paper's core thesis—that plateauing growth is similarly well-supported by the data as exponential growth—was a valuable contribution. However, the TH 1.1 data provides modest evidence against the specific sigmoid parameterization proposed:

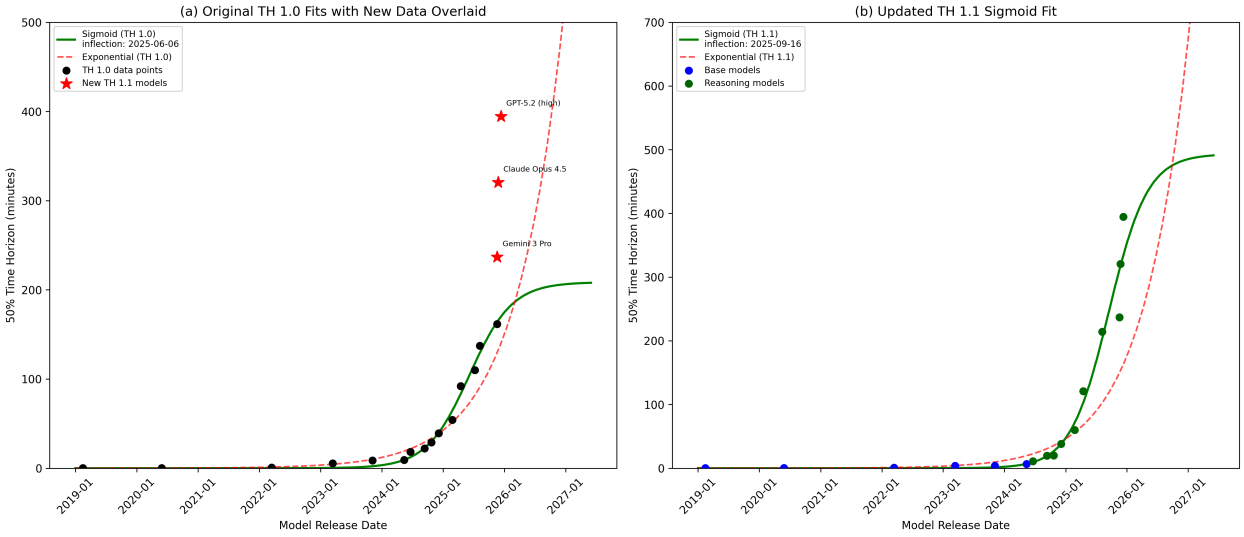

Figure 2: (a) The original TH 1.0 sigmoid fit with new TH 1.1 data points (red stars) overlaid, showing they far exceed the sigmoid's asymptote. (b) The updated sigmoid fit to TH 1.1 data, with the inflection point shifted to September 2025 and a much higher asymptote.

1. The estimated ceiling of 3.5 hours has already been exceeded by three models.
2. The inflection point has been pushed from mid-2025 to late 2025, suggesting that if a plateau exists, it is further away than initially estimated.
3. The accelerating doubling time (128 days from 2023 onward) is inconsistent with the decelerating growth predicted by a sigmoid approaching its inflection point.

That said, the original paper's broader point remains valid: a sigmoid can always be refit with a higher asymptote and later inflection. The question is whether this pattern of continual upward revision is itself evidence against the plateau hypothesis or simply reflects the inherent difficulty of forecasting with limited data.

4.4 The Multiplicative Decomposition

The original paper's most novel contribution—the multiplicative decomposition into base and reasoning capabilities—could not be fully re-evaluated here due to the lack of individual task-level run data for TH 1.1. This remains an important avenue for future analysis once such data becomes available.

5 Conclusion

The METR TH 1.1 data, including evaluations of GPT-5.2, Claude Opus 4.5, and Gemini 3 Pro, provides a natural out-of-sample test for the models proposed in Ge, Bastani, & Bastani (2026). The new data points substantially exceed the predictions of both the sigmoid and exponential models fit to TH 1.0 data. While the sigmoid model can be refit to accommodate the new data, doing so requires more than doubling the estimated asymptote and pushing the inflection point three months later.

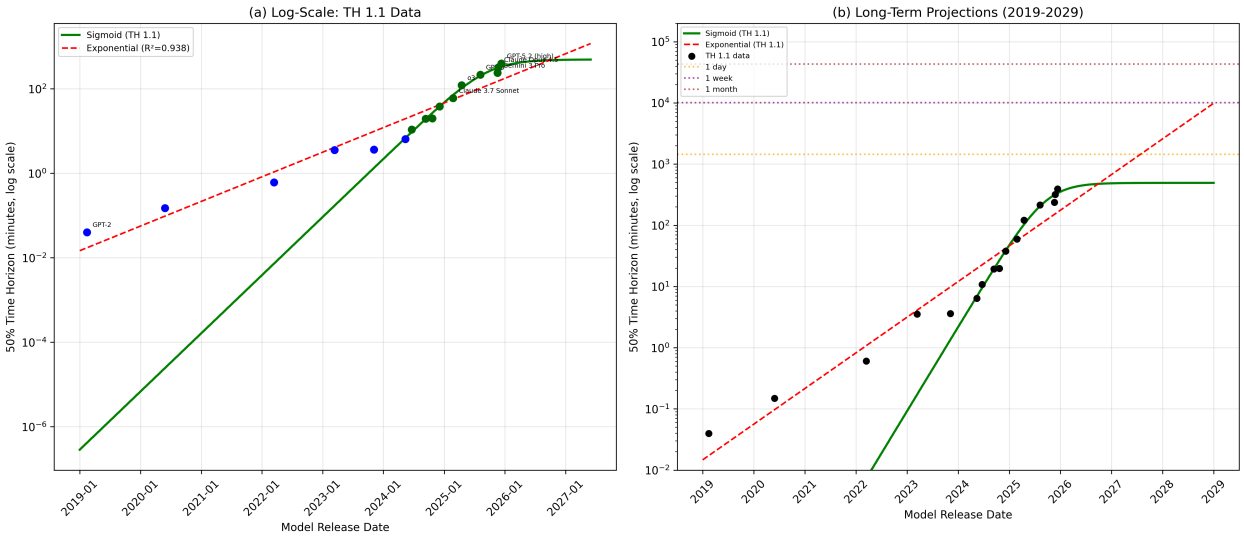

Figure 3: (a) Log-scale view of TH 1.1 data with sigmoid and exponential fits. (b) Long-term projections to 2029 showing the dramatic divergence between sigmoid (predicting plateau around 8 hours) and exponential (predicting month-long task horizons by 2028–2029).

These findings highlight a fundamental challenge in distinguishing between exponential and sigmoid growth with limited data: both can fit the observed data reasonably well, but they diverge dramatically in their future predictions. The continued upward revision of the sigmoid's parameters each time new data arrives suggests that, at a minimum, the "plateau" is not as imminent as initially estimated. The debate between exponential and sigmoid growth in AI capabilities remains unresolved, and additional data points—particularly in 2026—will be critical for adjudicating between these hypotheses.

Reproducibility

All analysis code is available alongside this addendum. The METR TH 1.1 data was obtained from `https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/` on February 5, 2026. The original paper's code repository is at `https://github.com/obastani/AI_Forecasting`.