

Forecasting AI Efficiency During Engagement Delivery: A Simulation Study of Estimate-to-Complete Methods in Accounting Services

Ali Cetin

alictnnn@gmail.com

Independent

Research Article

Keywords: estimate to complete, forecasting, artificial intelligence, accounting services, engagement budgeting, Monte Carlo simulation, learning curve

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11355308/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Forecasting AI Efficiency During Engagement Delivery: A Simulation Study of Estimate-to-Complete Methods in Accounting Services

Ali Cetin

*Finance and accounting professional
October 2026*

Abstract

When accounting firms build artificial intelligence (AI) efficiency assumptions into engagement budgets, they need a way to check during delivery whether those assumptions hold. Monthly estimate-to-complete (ETC) reviews provide such a check, but the choice of forecasting method matters. This paper uses a Monte Carlo simulation of 10,000 recurring accounting engagements per scenario to compare six ETC methods under uncertain AI efficiency, uncertain review effort, staff learning and changes in work volume. Three findings stand out. First, re-forecasting during delivery clearly outperforms relying on the original plan: by month 6, the best methods roughly halve the forecast error of the plan alone. Second, the best method depends on learning: when staff are still learning to use AI, a 50/50 blend of plan and latest-month performance is most accurate early in the engagement (mean absolute error 2.37% at month 3), whereas a cumulative average is best when performance is stable. Third, activity-level forecasting matters most when the mix of work changes: after a shift in volumes, an activity-level latest-month forecast had a mean error of 1.55% at month 6, compared with 3.52% for an engagement-level forecast. The simulation also shows that optimistic AI plans are likely to miss their target margin in most cases, which underlines the need for in-flight monitoring. The paper offers practical guidance on which ETC method to use at each stage of an engagement.

Keywords: estimate to complete; forecasting; artificial intelligence; accounting services; engagement budgeting; Monte Carlo simulation; learning curve

JEL classification: M41; C63; L84; O33

1. Introduction

Generative AI can shorten many routine accounting tasks, but its effects are uneven across tasks and difficult to predict in advance (Dell'Acqua et al., 2023; Noy & Zhang, 2023). Field evidence also shows that the benefits build up over time as users learn to work with the tools (Brynjolfsson et al., 2025; Choi & Xie, 2025). For accounting firms, this creates a planning problem: an engagement budget that assumes a given AI efficiency may be too optimistic or too pessimistic, and the error may only become visible late in the engagement.

In earlier work I proposed that AI-enabled engagements should be planned at the activity level, with explicit parameters for AI efficiency, adoption and review effort, and monitored monthly with estimate-to-complete (ETC) reviews (Cetin, 2026a). That proposal left open a practical question: which ETC method should firms use? Project management offers several standard approaches, from simply trusting the original plan to extrapolating recent performance (Anbari, 2003; Fleming & Koppelman, 2010; Project Management Institute, 2021). Their relative accuracy is likely to depend on how AI efficiency evolves during the engagement.

This paper addresses the question with a Monte Carlo simulation. Simulation is well suited to the problem because it allows the true AI effects, learning and volume changes to be controlled and varied, and because firm-level data on AI-enabled engagements are not yet available. The aim is not to estimate real-world parameters, but to compare forecasting methods under plausible and transparent assumptions and to derive practical guidance.

2. Background

The estimate to complete (ETC) is the forecast of the effort required to finish the remaining work; the estimate at completion (EAC) is actual effort to date plus ETC (Project Management Institute, 2021). Common ETC methods assume that remaining work will be performed at the planned rate, at the rate observed so far, or at some combination of the two (Anbari, 2003; Fleming & Koppelman, 2010). In activity-based terms, each method corresponds to a different assumption about the time per unit of output for the remaining volume (Kaplan & Anderson, 2007).

Two features of AI-enabled accounting work make the choice of method non-trivial. First, plans are likely to be optimistic. People routinely underestimate task completion times (Buehler et al., 1994), and project estimates show systematic optimism bias (Flyvbjerg, 2006); new technology

adds further uncertainty. Second, performance improves over time as staff learn, a pattern described by learning curves since Wright (1936). If early months are slower than later months, methods that average over the whole history will overstate the remaining effort, while methods that use only the latest month will capture improvement faster but react more to noise. Finally, if the mix of work changes during the engagement, forecasts made at the level of the whole engagement may misjudge the remaining effort, because different activities benefit from AI to different degrees (Cetin, 2026a, 2026b).

3. Simulation design

The simulation models the hypothetical recurring monthly accounting and reporting engagement described in Cetin (2026a): twelve months, eight activities, three staff levels, and an AI-adjusted plan with a fee set for a 30% planned margin. For each simulated engagement, the true AI parameters are drawn around the planned values, and monthly hours are generated with learning effects and random variation in volumes and time per unit. Table 1 summarizes the design.

Table 1. Simulation design (base scenario)

Element	Specification
Engagement	Recurring monthly accounting and reporting engagement, 12 months, 8 activities (Table 1 of Cetin, 2026a)
Plan	AI-adjusted plan using expected efficiency (e), adoption (a) and oversight (o) per activity; fee set for a 30% planned margin
True efficiency	Plan e multiplied by a mean-one lognormal factor ($\sigma = 0.3$), capped at 90%
True oversight	Plan o multiplied by a mean-one lognormal factor ($\sigma = 0.5$); 15% chance of an additional 0-15 percentage points
Adoption	Plan a plus normal noise ($\sigma = 0.10$), bounded to $[0, 1]$
Learning curve	Realized efficiency in month $t = e \times (1 - L \cdot \exp(-t/\tau))$, $L \sim U(0, 0.6)$, $\tau \sim U(1, 4)$ months
Volume	Client-level shift $N(1, 0.05)$ and monthly activity noise $N(1, 0.08)$
Time per unit	Monthly multiplicative noise $N(1, 0.06)$
Replications	10,000 engagements per scenario; random seed 2026

The simulation was written in Python (NumPy) and uses a fixed random seed (2026), so all results can be reproduced exactly; the code is openly available (Cetin, 2026d). An AI assistant (Claude, Anthropic) was used to support coding of the simulation and drafting of the text; the author specified the research question, the scenarios and all parameter values, reviewed the code and outputs, and is responsible for the analysis.

For each activity i and month t , actual hours are volume multiplied by actual time per unit:

$$hours_{i,t} = volume_{i,t} \times h_i \times [1 - a_i \cdot e_i \cdot (1 - L_i \cdot e^{(-t/\tau_i)}) + a_i \cdot o_i] \times \varepsilon_{i,t} \quad (1)$$

At review months 3, 6 and 9, six methods forecast the total cost at completion. Five methods work at the activity level and apply different time-per-unit assumptions to the remaining planned volume: (1) the planned rate, i.e., no re-forecast; (2) the latest month's rate; (3) the cumulative average rate to date; (4) the three-month moving average rate; and (5) an equal blend of the planned rate and the latest month's rate. The sixth method extrapolates the latest month's total engagement hours without distinguishing activities. Forecast cost is compared with the actual cost at completion. Four scenarios are examined: a base scenario with learning; a scenario without learning; a scenario with greater uncertainty in review effort; and a scenario in which volumes of two high-volume activities increase by 30% from month 7, a change known to the team at the month-6 review.

4. Results

4.1 *Optimistic plans miss their target*

Table 4 shows the distribution of actual margins. In the base scenario, the mean actual margin is 27.4%, below the planned 30%, and 73.9% of engagements fall short of the target. The shortfall arises because realized efficiency is reduced by learning in the early months and because review effort is occasionally much higher than planned. These results are a direct consequence of the simulation assumptions, but they illustrate a general point: when plans incorporate uncertain AI efficiency and staff take time to learn, a margin set on the plan alone is more likely to be missed than met. Monitoring during delivery is therefore not optional.

Table 4. Distribution of actual margins (planned margin 30%)

Scenario	Mean actual margin	Std. dev.	5th percentile	95th percentile	Share below 30% target
Base (learning)	27.4%	4.0 pp	20.9%	34.2%	73.9%
No learning	29.1%	4.1 pp	22.4%	36.0%	59.7%
High oversight uncertainty	27.8%	4.6 pp	20.0%	35.2%	68.4%
Volume mix shift	24.6%	4.5 pp	17.4%	32.3%	88.0%

4.2 *Re-forecasting beats the plan*

Table 2 and Figure 1 report the accuracy of each method in the base scenario. Relying on the original plan produces a mean absolute error of 3.64% at month 3 and 2.40% at month 6. The best re-forecasting methods reduce these errors to about 2.37% and 1.32%, respectively, roughly

halving the error by month 6. The plan also tends to underestimate cost (bias -1.39% at month 3), consistent with the optimism built into the plan.

Table 2. Forecast accuracy of ETC methods, base scenario (10,000 simulated engagements)

Forecast method	MAPE month 3	MAPE month 6	MAPE month 9	Bias month 3	Shortfalls detected (month 3)	False alarms (month 3)
Plan only (no re-forecast)	3.64%	2.40%	1.22%	-1.39%	99.9%	64.5%
Latest-month run rate	2.86%	1.40%	0.71%	+2.29%	98.8%	50.7%
Cumulative average	4.27%	2.04%	0.86%	+4.16%	99.9%	72.6%
Three-month moving average	4.27%	1.32%	0.62%	+4.16%	99.9%	72.6%
Blend (50% plan, 50% latest)	2.37%	1.54%	0.80%	+0.45%	99.4%	54.3%
Aggregate latest-month (engagement level)	2.79%	1.34%	0.68%	+2.26%	97.8%	41.4%

Note: MAPE = mean absolute percentage error of the forecast cost at completion relative to actual cost. Bias = mean signed error (positive = overestimate). A shortfall is an engagement whose actual margin falls below the 30% target; shortfalls detected = share of such engagements flagged by the forecast; false alarms = share of engagements meeting the target that the forecast flagged. At month 3 the three-month moving average equals the cumulative average.

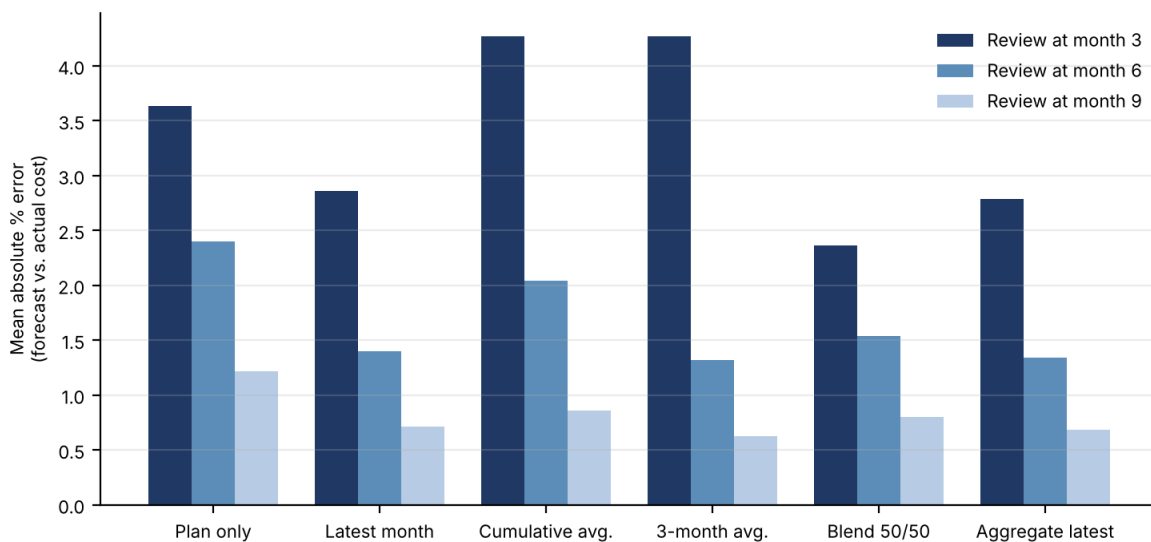


Figure 1. Mean absolute percentage error of cost-at-completion forecasts by method and review month, base scenario

4.3 Learning changes which method is best

With learning, the cumulative average performs poorly early in the engagement (error 4.27% and bias +4.16% at month 3), because it gives as much weight to the slower first months as to recent performance. The latest-month rate captures improvement better (2.86%), and the 50/50 blend of plan and latest month is the most accurate and nearly unbiased at month 3 (2.37%, bias +0.45%). By month 6, once learning has largely stabilized, the three-month moving average becomes the

most accurate activity-level method (1.32%). Without learning, the ranking changes: the cumulative average is the most accurate method at month 3 (1.66%, compared with 2.00% for the latest month), because averaging reduces noise when performance is stable (Table 3).

Table 3. MAPE (%) at months 3 and 6 across scenarios

Forecast method	Base (learning) (m3 / m6)	No learning (m3 / m6)	High oversight uncertainty (m3 / m6)	Volume mix shift (m3 / m6)
Plan only (no re-forecast)	3.64 / 2.40	3.60 / 2.42	3.94 / 2.63	5.85 / 2.82
Latest-month run rate	2.86 / 1.40	2.00 / 1.35	2.93 / 1.43	2.64 / 1.55
Cumulative average	4.27 / 2.04	1.66 / 1.09	4.35 / 2.10	2.25 / 2.35
Blend (50% plan, 50% latest)	2.37 / 1.54	2.31 / 1.55	2.62 / 1.66	3.95 / 1.78
Aggregate latest-month (engagement level)	2.79 / 1.34	1.88 / 1.29	2.80 / 1.34	2.54 / 3.52

Note: In the volume mix shift scenario, volumes of bank reconciliations and invoice processing rise by 30% from month 7, and the change is known to the engagement team at the month-6 review. Activity-level methods use the updated volume forecast; the aggregate method cannot.

4.4 When activity-level forecasting matters

In the base scenario, an engagement-level forecast performs about as well as the activity-level methods (1.34% at month 6) and produces fewer false alarms. This is an important nuance: when the mix of work is stable, the main value of activity-level monitoring lies in diagnosis, showing which activities are under- or over-performing, rather than in forecast accuracy. When the mix changes, however, the difference is large. In the volume mix shift scenario, the activity-level latest-month forecast had a mean error of 1.55% at month 6, compared with 3.52% for the engagement-level forecast, because only the activity-level method can apply updated volumes to activities with different AI-adjusted unit times.

4.5 Detecting margin shortfalls

All re-forecasting methods detect the large majority of engagements that end below the target margin by month 3 (Table 2). They differ in false alarms: methods that overestimate cost early, such as the cumulative average under learning, flag many engagements that eventually meet the target. Firms using such methods would spend management time on engagements that do not need intervention.

5. Practical guidance

The results suggest a simple, staged approach for firms monitoring AI-enabled engagements:

- *Months 1 to 3 (learning phase):* use a blend of the planned and latest-month time per unit. This avoids over-reacting to early, slower months while still capturing improvement.
- *From around month 4 (stabilized phase):* switch to a three-month moving average, which balances responsiveness and noise reduction.
- *When volumes or scope change:* forecast at the activity level with updated volumes. Engagement-level extrapolation can be badly wrong when the mix shifts.
- *Always:* use activity-level results for diagnosis, to see where AI performs above or below plan, even when an engagement-level forecast is accurate enough for the total.

These rules can be implemented in a spreadsheet alongside the activity-level pricing plan described in Cetin (2026a). They are also consistent with the broader argument that firms need to measure AI effects to capture their economic value (Cetin, 2026c).

6. Limitations

The study is based on simulated data. Its parameters were chosen to be plausible, not estimated from real engagements, and the results depend on them; for example, the share of engagements missing their target is driven by the assumed degree of optimism in the plan. The simulation covers a single recurring engagement type and does not model staffing changes, scope disputes or quality issues. Future research should test the methods on real activity-level time and output data from accounting firms, estimate learning parameters for common accounting activities, and examine whether the staged approach improves margin outcomes in practice.

7. Conclusion

Planning for AI efficiency is only half the task; the plan must also be checked while the work is being done. This simulation shows that monthly re-forecasting substantially improves on the original plan, that the best forecasting method depends on whether staff are still learning to use AI, and that activity-level forecasting is essential when the mix of work changes. A staged approach, blending plan and actuals early and moving to a moving average later, offers a simple and robust rule for firms of any size.

Declarations

Data and code availability. All results are generated by a Python simulation with a fixed random seed (2026). The code is openly available on Zenodo under the MIT License: Cetin (2026d), <https://doi.org/10.5281/zenodo.23328944>.

Disclaimer. The views expressed in this paper are solely those of the author and do not represent the views of any employer or client. All data are simulated.

Funding and competing interests. The author received no funding for this work and declares no competing interests.

References

1. Anbari, F. T. (2003). Earned value project management method and extensions. *Project Management Journal*, 34(4), 12-23.
2. Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *Quarterly Journal of Economics*, 140(2), 889-942.
3. Buehler, R., Griffin, D., & Ross, M. (1994). Exploring the "planning fallacy": Why people underestimate their task completion times. *Journal of Personality and Social Psychology*, 67(3), 366-381.
4. Cetin, Ali (2026a). Capturing hidden capacity: Activity-level pricing plans and monthly estimate-to-complete monitoring for AI-enabled accounting services (SSRN Working Paper).
5. Cetin, Ali (2026b). Artificial intelligence and engagement pricing in accounting firms: A literature review and research agenda. *Preprints.org*.
6. Cetin, Ali (2026c). The economics of hidden capacity: AI productivity gains, price pass-through and margin dynamics in professional services (MPRA Paper).
7. Cetin, Ali (2026d). ETC forecasting simulation for AI-enabled accounting engagements (Version 1.0) [Computer software]. Zenodo. <https://doi.org/10.5281/zenodo.23328944>
8. Choi, J. H., & Xie, C. (2025). Human + AI in accounting: Early evidence from the field (SSRN Working Paper No. 5240924).
9. Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Kraymer, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality (Harvard Business School Working Paper No. 24-013).
10. Fleming, Q. W., & Koppelman, J. M. (2010). *Earned value project management* (4th ed.). Project Management Institute.

11. Flyvbjerg, B. (2006). From Nobel Prize to project management: Getting risks right. *Project Management Journal*, 37(3), 5-15.
12. Kaplan, R. S., & Anderson, S. R. (2007). *Time-driven activity-based costing: A simpler and more powerful path to higher profits*. Harvard Business School Press.
13. Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654), 187-192.
14. Project Management Institute. (2021). *A guide to the project management body of knowledge (PMBOK guide) (7th ed.)*. Project Management Institute.
15. Wright, T. P. (1936). Factors affecting the cost of airplanes. *Journal of the Aeronautical Sciences*, 3(4), 122-128.