

# Sizing AI-Driven Enterprise Engagements Under Uncertainty: A Probabilistic Scope and Pricing Framework for Pre-Sales Architects

Mihir Shah

Independent Researcher, USA  
[mihirshahconnect@gmail.com](mailto:mihirshahconnect@gmail.com)

## Article Info

### Article history:

Received 2026-07-28

Revised 2026-09-11

Accepted 2026-10-03

### Keyword:

*Pre-Sales Solution Architecture,  
AI Engagement Estimation,  
Probabilistic Sizing,  
Three-Point Estimation,  
Multi-Phase Pricing,  
Scope Uncertainty.*

## ABSTRACT

Pre-sales architects who size AI-driven enterprise engagements confront a problem that conventional estimation methods were not built to handle. These engagements combine deterministic components, which established techniques size with reasonable confidence, and probabilistic components, whose effort remains unresolved until implementation begins. Treating every component as deterministic tends to produce chronic underestimation, eroded margins, and delivery teams that inherit commitments they cannot meet. Treating every component as probabilistic surrenders competitive bids to providers willing to offer false certainty. This paper presents a structured framework that sizes each class of component on its own terms and prices the resulting engagement to protect client and provider together. The approach separates stable, fully specifiable work from work that depends on data quality, model behavior, prompt iteration, and evaluation design. Stable components are estimated deterministically. Uncertain components are estimated with three-point ranges aggregated through Monte Carlo simulation rather than summed at their expected values. Pricing follows a multi-phase structure, with discovery, pilot, and scale phases that each carry independent scope, price, and gating criteria, supported by disclosed assumptions that feed contract change management. Operational disciplines of peer review and instrumented post-engagement learning convert individual estimates into an improving organizational capability. A worked illustrative example demonstrates decomposition, ranged sizing, and phased pricing for a customer service assistant engagement. The framework is offered as practitioner guidance grounded in enterprise solutioning experience, with controlled empirical validation identified as future work.



This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

## I. INTRODUCTION

Traditional consulting estimation rests on an assumption that effort scales predictably once scope is understood. Decades of software sizing practice formalize this premise through function points, parametric models, and structured expert judgment [1], [2]. The premise holds reasonably well when requirements are stable and the work resembles work done before. Practitioner experience across enterprise AI solutioning engagements suggests that the premise weakens sharply once generative and retrieval components enter scope. A pre-sales architect authoring a statement of work must still commit to numbers. The difficulty is that some of those

numbers describe work whose effort cannot be known before implementation reveals it.

AI engagements differ because several components are genuinely uncertain rather than merely unfamiliar. Data quality surfaces only when pipelines run against real enterprise records. Prompt engineering complexity depends on how a chosen model responds to a specific domain. Model behavior and evaluation criteria often remain unsettled until stakeholders see early outputs and revise what they consider acceptable. Analyses of failed AI initiatives point to misjudged scope and shifting objectives as recurring causes rather than isolated accidents [3]. A sizing method that records these factors as fixed quantities encodes a claim the

architect cannot support. The proposal reads as precise while resting on guesses.

Ignoring this uncertainty fails in two opposite directions. Underestimation commits the delivery team to effort it cannot supply within the agreed price, which produces schedule slippage, margin erosion, and staff burnout. Overestimation pads every line to feel safe, which inflates the bid and hands the engagement to a competitor promising confident certainty. Neither error is a matter of carelessness. Both follow directly from applying a deterministic instrument to a partly nondeterministic problem. The architect who has closed multi-tier enterprise procurements learns that a bid lost to false precision and a delivery lost to true optimism share one root cause. The estimation model never distinguished the two kinds of work.

The framework proposed here treats that distinction as its organizing principle. Engagement scope is decomposed into deterministic and probabilistic components, each sized by a method suited to its nature, then priced through a phased structure that allocates risk deliberately. The probabilistic elements draw on three-point estimation and Monte Carlo aggregation, techniques with long histories in project cost and schedule analysis [4], [5]. The pricing logic borrows from value-based and phased contracting practice [6]. The remainder of this paper is organized around the IMRAD structure. Section II presents the Method, covering the related work, the decomposition approach, probabilistic sizing, pricing, and operational disciplines that together constitute the framework. Section III presents Results and Discussion, comprising a worked illustration, sector-specific calibration, and limitations. Section IV concludes.

#### A. Related Work

Software effort estimation forms the most developed body of relevant work. Function point analysis, the COCOMO family, and agile practices such as planning poker and story points give teams repeatable ways to convert scope into effort [7], [8]. Systematic reviews of agile estimation document steady refinement of these instruments and their calibration data [9]. Recent work applies transformer models to story point prediction, extending the same tradition with learned estimators [10], [11]. A shared assumption runs through the literature. The requirement being sized is knowable at estimation time, even if imperfectly described. That assumption is precisely what AI engagement scope violates for a meaningful share of its components.

A growing strand introduces machine learning to effort and cost estimation, reporting empirical gains over classical parametric baselines [12], [13]. Industrial experience reports describe both the promise and the fragility of these estimators when project data is sparse or inconsistent [14], [15]. This work improves how known scope is estimated. It does not address scope whose effort driver is unresolved until a model is selected and exercised against live data. The distinction matters for pre-sales, where commitments are made before any implementation evidence exists. Practitioner solutioning

experience indicates that better point estimators do not remove the underlying uncertainty. They sharpen estimates of the components that were already estimable.

Probabilistic estimation offers the tools this problem needs, though seldom aimed at pre-sales AI sizing. Three-point estimation, formalized in program evaluation practice and carried into modern project management guidance, represents each uncertain quantity as a range rather than a single figure [4]. Monte Carlo simulation aggregates such ranges to produce distributions of total cost and schedule, an approach validated for infrastructure projects prone to overruns [5]. Parallel research explores AI-assisted cost estimation and overrun prediction within project management [16], [17], [18]. These methods handle uncertainty rigorously. Their application to the specific act of sizing an enterprise AI engagement during pre-sales remains largely unaddressed in the published record.

Practitioner and management literature supplies context without closing the gap. A Stanford analysis of fifty-one enterprise deployments distills lessons that concentrate on post-deployment adoption and organizational factors rather than pre-sales sizing [19]. Studies of IT consulting value co-creation treat estimation as an input to the engagement rather than an object of study [20]. Contracting research examines fixed-price versus performance-based structures and gain-sharing, which inform how risk is priced once scope is set [21], [22]. Sales engineering guidance covers discovery and proposal craft while leaving technical sizing method implicit. The framework presented here occupies the intersection these strands leave open, joining probabilistic estimation to pre-sales pricing.

## II. METHOD

#### A. Separating Deterministic and Probabilistic Components

The framework begins by splitting engagement scope into two classes before any number is attached. Deterministic components carry stable, fully specifiable requirements: platform integration, identity and entitlement setup, data pipeline construction, infrastructure provisioning, and security review. Practitioner experience across enterprise engagements suggests these typically account for forty to sixty percent of total effort. They resemble work the delivery organization has done repeatedly, and established sizing techniques estimate them with acceptable confidence [1], [2]. Attaching a single number to this class is defensible. The estimate rests on prior instances of similar work rather than on speculation about behavior not yet observed.

Probabilistic components occupy the remaining share and behave differently. Model selection, prompt engineering iteration, retrieval pipeline tuning, and evaluation harness development depend on factors that stay unresolved until implementation supplies evidence. Their effort can swing widely with data quality and stakeholder expectations. Misclassification in either direction carries a cost. Labeling a probabilistic component as deterministic yields a brittle

commitment that delivery cannot honor. Labeling a deterministic component as probabilistic pads the bid and weakens competitiveness. The classification step therefore does real work, and getting it wrong quietly sets up the failure modes the framework exists to prevent. Correct classification is therefore the quiet foundation on which every later estimate depends.

### B. Component Classification Criteria

A component qualifies as deterministic when its requirements are stable and can be fully specified during pre-sales, so that effort follows from documented precedent. It qualifies as probabilistic when its effort depends on conditions that cannot be settled until implementation, such as how a model responds to domain data or which evaluation threshold stakeholders will accept. The test is not whether a component involves AI. It is whether the information needed to size it exists yet. Security review of an AI system can be deterministic. Tuning that system's retrieval quality usually is not, and the difference decides which sizing method applies.

Mixed components require care because a single line item can contain both natures. A retrieval-augmented generation pipeline bundles deterministic ingestion and indexing with probabilistic relevance tuning and answer evaluation. Sizing it as one unit forces a false choice between a brittle number and an inflated range. The framework instead decomposes such components to leaf level, classifying each leaf on its own terms before recombining the estimates. This granularity adds modest effort during proposal authorship. It repays that cost by producing commitments the delivery team recognizes as honest, which practitioner experience links to smoother handoff and fewer disputed change requests. Table I summarizes the classification criteria and examples.

TABLE I.  
DETERMINISTIC VS. PROBABILISTIC COMPONENT CLASSIFICATION  
CRITERIA AND EXAMPLES

| Attribute              | Deterministic Components                                                                                                       | Probabilistic Components                                                                                 |
|------------------------|--------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------|
| Requirement stability  | Fully specifiable during pre-sales                                                                                             | Unresolved until implementation supplies evidence                                                        |
| Effort driver          | Documented precedent and similar prior work                                                                                    | Data quality, model behavior, evaluation thresholds                                                      |
| Typical examples       | Platform integration, identity and entitlement setup, data pipeline construction, infrastructure provisioning, security review | Model selection, prompt engineering iteration, retrieval pipeline tuning, evaluation harness development |
| Typical effort share   | Roughly 40 to 60 percent of engagement                                                                                         | Roughly 40 to 60 percent of engagement                                                                   |
| Sizing method          | Single-point deterministic estimate                                                                                            | Three-point range with Monte Carlo aggregation                                                           |
| Misclassification risk | If treated as probabilistic: inflated, uncompetitive bid                                                                       | If treated as deterministic: brittle, unmeetable commitment                                              |

### C. Three-Point Estimation

Once components are classified, the probabilistic set is sized with ranges rather than points. Each component receives an optimistic, expected, and pessimistic estimate, capturing the spread that a single figure would conceal. Across enterprise AI work these ranges commonly span a factor of two to five between the optimistic and pessimistic ends, reflecting how much data and model behavior can vary. The three-point form has a long lineage in project scheduling and cost analysis [4]. Its value here is honesty about what is not yet known. A range states the uncertainty explicitly instead of burying it inside a confident number the architect cannot defend.

Aggregating these ranges demands more than adding their expected values. Summing expected values understates total uncertainty and ignores how component risks combine. Monte Carlo simulation samples each component from a triangular or beta distribution fitted to its three points, then repeats across many iterations to build a distribution of total effort [5]. The result is a range with associated likelihoods rather than a lone figure, which supports pricing decisions grounded in confidence levels. Practitioner experience indicates that stakeholders engage more readily with a defensible eightieth-percentile figure than with a bare mean. The simulation also exposes which components drive most of the total variance.

### D. Assumption Documentation

Every probabilistic estimate rests on assumptions, and the framework requires stating them explicitly. Each assumption is written down, classified as factual or conjectural, and linked to the components it governs. A factual assumption records a condition believed true and verifiable, such as an available document corpus. A conjectural assumption records a working guess, such as expected retrieval quality against that corpus. This documentation is not administrative overhead. It becomes the evidentiary backbone of the estimate, showing reviewers and clients exactly what the numbers depend upon and where the exposure sits. Practitioner experience indicates that this written basis, more than any single number, is what lets a later conversation about scope stay factual rather than adversarial.

Documented assumptions carry directly into commercial protection. The assumption set enters the proposal scope language and the contract change management protocol, so that a later shift is handled as an anticipated event rather than a dispute. Practitioner experience across multi-tier procurement chains suggests a sharp asymmetry in how assumption changes are received. A change traced to a documented assumption reads as expected risk materializing, and renegotiation proceeds calmly. A change with no documented basis reads as estimation incompetence, straining the relationship. The discipline costs little during proposal authorship. It preserves both margin and trust when reality diverges from the plan, as it regularly does.

### E. Multi-Phase Engagement Structure

Pricing follows the same logic that shaped sizing, refusing to force uncertain and certain work into one commercial container. The framework structures the engagement as sequential phases. A discovery phase establishes scope precision by resolving the largest unknowns. A pilot phase validates capability against real data and stakeholder judgment. A scale phase delivers the production system. Each phase is independently valuable and carries its own scope, price, and gating criteria, so a client can evaluate progress and decide before further commitment. This structure mirrors staged contracting practice and aligns payment with demonstrated value rather than with promises made before evidence exists [6], [22].

Phasing changes the character of the commitment the architect must make at bid time. Rather than pricing the entire uncertain engagement at once, the provider prices the discovery phase firmly and the later phases with progressively better information. Gating criteria give both parties a defined decision point, which reduces the exposure that a single large fixed price would impose. Practitioner experience generating qualified pipeline through phased proposals suggests that clients often prefer this structure once its risk logic is explained. The early phase is affordable and concrete. The later phases are sized against evidence the pilot produced rather than against pre-sales conjecture. Table II sets out the phase structure.

TABLE II.  
MULTI-PHASE PRICING STRUCTURE

| Phase     | Purpose                                                    | Pricing approach                                          | Gating criteria                                   |
|-----------|------------------------------------------------------------|-----------------------------------------------------------|---------------------------------------------------|
| Discovery | Establish scope precision and resolve the largest unknowns | Firm fixed price on bounded scope                         | Documented scope and assumptions before pilot     |
| Pilot     | Validate capability against real data and stakeholders     | Firm fixed price with documented expected outcomes        | Demonstrated capability meeting agreed thresholds |
| Scale     | Deliver the production system                              | Ranged pricing, or risk-loaded fixed with credit protocol | Pilot outcomes confirm scale assumptions          |

### F. Pricing the Probabilistic Components

Within this structure, the probabilistic components still need a pricing mechanism, and each candidate distributes risk differently. Time-and-materials transfers risk entirely to the client, which is honest but often uncompetitive and unsettling for procurement. Fixed-price transfers risk entirely to the provider, which looks attractive to buyers yet turns brittle when effort lands near the pessimistic end. The framework's primary recommendation is ranged pricing with disclosed assumptions, which splits risk between the parties in proportion to who can influence each driver. The range communicates genuine uncertainty. The disclosed assumptions make the range auditable rather than arbitrary. Practitioner experience across enterprise procurement

suggests that buyers accept a range far more readily when its underlying assumptions are visible and specific.

Some clients cannot accommodate ranges because their procurement rules demand a single figure. For these cases the framework offers risk-loaded fixed pricing as a deliberate fallback. The component is priced at or near its pessimistic estimate, and a refund or credit protocol returns value for effort not consumed. The critical condition is that this protocol is specified upfront in the contract, not negotiated after delivery when incentives have diverged. Practitioner experience closing enterprise engagements suggests that a pre-agreed credit mechanism preserves trust even under conservative pricing. The client accepts a higher ceiling knowing the mechanism will return unused margin transparently, which keeps a conservative bid from reading as opportunism.

### G. Peer Review

The framework treats peer review as a required control rather than an optional courtesy. A solo architect working under bid pressure tends toward optimism, because an aggressive estimate wins internal approval and improves the apparent competitiveness of the bid. This bias is systematic, not a personal failing, and self-review rarely catches it. A senior architect reviewing the estimate examines the assumption set, the three-point ranges, and the dependency structure that links components. The reviewer asks where optimism has crept in and which assumptions lack support. Practitioner experience suggests that this second perspective reliably surfaces exposure the original author could not see. The economics of the review favor it strongly. A structured review consumes roughly two to four hours of a senior architect's time per engagement. The benefit is a measurable reduction in the estimation errors that cause margin loss and delivery strain, an improvement practitioner experience associates with better delivery economics within the first year of adoption. The review also transfers judgment from experienced architects to less experienced ones, functioning as training alongside quality control. Framing it as mandatory removes the pressure to skip it when a deadline looms. The discipline is cheap. Its absence is expensive in ways that surface only after commitments are signed.

### H. Tracking and Learning

Individual estimates improve only when the organization learns from their outcomes, which requires instrumented tracking. The framework records, for each engagement, the original estimates, the documented assumptions, and the realized effort by component. This record turns estimation from a repeated one-off act into a dataset. Without it, the same errors recur because no one can see them accumulate. Practitioner experience across solutioning engagements indicates that memory alone does not correct estimation bias. Structured records make patterns visible that intuition consistently misses, especially patterns that span many engagements rather than appearing within any single one. The

record therefore becomes an organizational asset that outlasts any individual architect and the memory of any single bid. Periodic analysis of that record identifies systematic error rather than random noise. It may reveal that a particular component type is routinely underestimated, or that certain client conditions, such as poor data governance, act as reliable effort multipliers. These findings feed back into future three-point ranges and classification decisions, tightening the framework over time. Sustaining the discipline requires executive sponsorship and a defined owner, because a practice belonging to no one lapses under delivery pressure. This mirrors the co-creation view of consulting value, where measurement and shared accountability underpin durable improvement [20]. The payoff compounds. Each engagement makes the next estimate better grounded.

### III. RESULT AND DISCUSSION

#### A. Worked Example

A worked illustration shows the framework applied to a mid-market enterprise building a generative AI customer service assistant. The figures are the author's own pedagogical values, not measured results, and are labeled accordingly. Scope decomposes into deterministic and probabilistic components. On the deterministic side sit CRM integration at two hundred hours, identity and entitlement setup at eighty hours, a data pipeline that is mostly deterministic at three hundred twenty hours, retrieval infrastructure at one hundred sixty hours, final integration testing at two hundred hours, and documentation and handover at one hundred twenty hours. These sum to roughly one thousand eighty hours of work the delivery team can commit to with confidence.

The data pipeline carries one probabilistic pocket worth isolating. Data quality discovery, the effort to profile and remediate real enterprise records, spans forty to one hundred twenty hours depending on what the source systems actually contain. Treating this as a fixed line would repeat the deterministic error the framework warns against. Separating it keeps the confident portion confident while flagging the genuinely uncertain remainder. Practitioner experience suggests that data quality is among the most common drivers of AI delivery overrun, which justifies the explicit range. The classification is granular by design, so the estimate reflects where knowledge actually stops. That honesty about the boundary of knowledge is exactly what a deterministic single figure would have concealed.

The purely probabilistic components receive three-point ranges aggregated by simulation. Retrieval tuning spans one hundred twenty to four hundred hours, model selection and prompt engineering span two hundred to six hundred hours, and evaluation harness development spans one hundred to three hundred hours. Aggregated through Monte Carlo rather than summed at extremes, these yield a probabilistic total in the region of five hundred to fifteen hundred hours. Combined with the deterministic base, the full engagement falls between roughly one thousand five hundred eighty and two thousand

five hundred eighty hours, with an expected value near two thousand eighty hours. Table III records the full breakdown.

TABLE III.  
WORKED EXAMPLE: ILLUSTRATIVE ENGAGEMENT SIZING AND PRICING  
BREAKDOWN

| Component                              | Classification | Hour range                       | Pricing approach                   |
|----------------------------------------|----------------|----------------------------------|------------------------------------|
| CRM integration                        | Deterministic  | 200                              | Fixed price, standard rate         |
| Identity and entitlement setup         | Deterministic  | 80                               | Fixed price, standard rate         |
| Data pipeline (stable portion)         | Deterministic  | 320                              | Fixed price, standard rate         |
| Data quality discovery                 | Probabilistic  | 40 to 120                        | Discovery phase, firm              |
| Retrieval infrastructure               | Deterministic  | 160                              | Fixed price, standard rate         |
| Final integration testing              | Deterministic  | 200                              | Fixed price, standard rate         |
| Documentation and handover             | Deterministic  | 120                              | Fixed price, standard rate         |
| Retrieval tuning                       | Probabilistic  | 120 to 400                       | Pilot and scale, ranged            |
| Model selection and prompt engineering | Probabilistic  | 200 to 600                       | Pilot and scale, ranged            |
| Evaluation harness development         | Probabilistic  | 100 to 300                       | Scale, ranged or risk-loaded fixed |
| Deterministic subtotal                 | Deterministic  | ~1,080                           | Fixed price                        |
| Probabilistic subtotal                 | Probabilistic  | 500 to 1,500                     | Phased and ranged                  |
| Total engagement                       | Mixed          | 1,580 to 2,580 (expected ~2,080) | Phased structure                   |

Pricing maps onto the phased structure. The deterministic components are quoted as fixed price at the standard rate, since their scope is settled. The probabilistic work is staged. A discovery phase of roughly two hundred hours, priced firmly, resolves the largest data and retrieval unknowns. A pilot phase of roughly four hundred hours, also priced firmly with documented expected outcomes, validates capability. A scale phase then carries ranged pricing, or risk-loaded fixed pricing where procurement requires a single figure. Each phase gates the next. The client commits incrementally against evidence, and the provider avoids pricing the whole uncertain span before the pilot has reduced it.

#### B. Sector-Specific Applications

The framework calibrates rather than changes across sectors, adjusting range width, documentation depth, and pricing form to local constraints. Regulated financial services

and healthcare impose compliance and audit obligations on probabilistic components, since model behavior must be explainable and defensible to supervisors. Effort ranges widen because validation and documentation absorb time that unregulated work does not. Assumption documentation becomes stronger still, as regulators and auditors may later inspect the basis for design choices. Practitioner experience with enterprise procurement suggests that under-documenting assumptions in these sectors converts ordinary estimation variance into compliance exposure. The framework's insistence on written, classified assumptions aligns naturally with these demands.

Public sector engagements strain the pricing side more than the sizing side. Procurement frameworks frequently cannot accommodate ranged pricing, requiring a single figure fixed in advance. Risk-loaded fixed pricing with a pre-agreed credit protocol becomes the practical instrument, letting the provider price conservatively while returning unused value transparently. Pre-sales cycles run longer, and discovery phases tend to be more elaborate because requirements pass through layered approval. Practitioner experience indicates that investing in a thorough discovery phase pays off disproportionately here, since the cost of a mispriced fixed commitment is harder to renegotiate once public procurement rules have closed around it.

Manufacturing and industrial settings introduce uncertainty from a different source. Integration with operational technology systems, which govern physical processes, carries risk that pure information systems do not, because behavior under real plant conditions is hard to predict from specifications alone. The framework responds by making the pilot phase explicit and load-bearing, using it to test operational technology integration before any scale-phase price is set. Practitioner experience suggests that integration surprises in operational technology are best absorbed in a bounded pilot rather than discovered during production rollout. The gating criteria between pilot and scale carry particular weight, since they protect the provider from committing to a scale price on unvalidated assumptions.

Across sectors the constant is the diagnostic that precedes calibration. The architect asks which components are genuinely uncertain, what constrains how that uncertainty may be priced, and where documentation must be strongest. The answers differ, but the questions do not. This portability is deliberate, since a framework that required reinvention for each sector would not survive contact with a working pre-sales team. Table IV summarizes the calibrations. Practitioner experience across enterprise solutioning suggests that the framework's value lies less in any single rule than in the consistent discipline of asking these questions before a number reaches a proposal. Calibration follows the diagnosis, never the other way around, which keeps the method coherent as contexts change.

TABLE IV.  
SECTOR-SPECIFIC FRAMEWORK CALIBRATION

| Sector                            | Key constraint                                                  | Recommended adaptation                                                           |
|-----------------------------------|-----------------------------------------------------------------|----------------------------------------------------------------------------------|
| Financial services and healthcare | Compliance and audit obligations on model behavior              | Widen probabilistic ranges; strengthen assumption documentation for auditability |
| Public sector                     | Procurement cannot accommodate ranged pricing; layered approval | Risk-loaded fixed pricing with pre-agreed credit; elaborate discovery phase      |
| Manufacturing and industrial      | Operational technology integration uncertainty                  | Explicit load-bearing pilot to validate integration before scale pricing         |

### C. Limitations

The framework presumes architects experienced enough to classify components and construct defensible ranges. Less experienced architects may misjudge which work is genuinely uncertain, producing either brittle commitments or inflated bids from the same tool. The recommended remedy is pairing them with reviewers, which the peer review discipline already supplies. That pairing doubles as a training mechanism, transferring judgment through practice rather than instruction alone. Practitioner experience suggests that estimation skill grows fastest when a novice's ranges are challenged by an experienced reviewer on real engagements. The framework does not remove the need for judgment. It structures where judgment is applied and how it is checked.

The approach also presumes clients willing to engage with ranged or multi-phase pricing. Some procurement functions cannot process a range or a phased commitment, and for them the risk-loaded fixed fallback is the intended accommodation rather than an afterthought. That fallback carries a cost, since conservative pricing can raise the headline figure and weaken a bid against a competitor willing to gamble on a low fixed number. The framework accepts this tradeoff openly. It holds that a defensible higher price with a credit protocol serves both parties better than an indefensible lower one that fails in delivery. Practitioner experience closing enterprise bids suggests that clients remember a delivery that held far longer than they remember a headline price.

Guidance on specific probability distributions remains limited. The framework recommends triangular or beta distributions as reasonable defaults for three-point aggregation, and these perform adequately in practice [5]. It does not specify which distribution best fits which component type, because the empirical basis for such a recommendation does not yet exist in the pre-sales AI setting. The tracking discipline described earlier is the intended path toward that evidence. As organizations accumulate records of estimated against realized effort, distribution choices can become empirical rather than assumed. Until then, the defaults are offered as defensible starting points, not as validated optima. Practitioner experience suggests that even approximate distribution choices outperform the implicit point assumptions they replace.

The framework itself has not been validated through controlled study. Its basis is accumulated practitioner experience across enterprise AI solutioning engagements, which grounds its claims but does not substitute for systematic evidence. The numeric figures in the worked example are illustrative rather than measured, chosen to demonstrate method rather than to report findings. A controlled evaluation comparing framework-sized engagements against conventionally sized ones on delivery accuracy and margin outcomes would test the claims made here. That study is identified as future work, and its absence is stated plainly so that the framework is read as disciplined guidance rather than established fact.

## V. CONCLUSION

Sizing AI-driven enterprise engagements under uncertainty is a learnable discipline rather than an intractable problem. The framework offered here gives pre-sales architects a structured path from raw scope to a defensible commitment: decompose components by nature, size the uncertain ones with ranges aggregated through simulation, price through phases that allocate risk deliberately, and hand delivery a clean and honest set of expectations. Each step addresses a specific way that deterministic estimation fails when applied to nondeterministic work. The method is more demanding than attaching single numbers to everything. The alternative, as practitioner experience repeatedly shows, produces worse outcomes for provider and client alike.

The operational disciplines distinguish a durable capability from a one-time technique. Peer review corrects the optimism that bid pressure induces, and instrumented tracking converts individual engagements into an improving dataset. Together they let estimation accuracy compound across engagements rather than resetting with each proposal. Practitioner experience across enterprise solutioning suggests that organizations adopting these disciplines gradually earn a reputation for commitments they actually meet. That reputation feeds client renewal and referral, which are the commercial returns the framework ultimately targets. The disciplines cost real effort. They repay it in the currency pre-sales organizations most need, which is trust that survives delivery.

The framework's portability across sectors and its explicit fallbacks are meant to keep it usable rather than theoretical. Regulated, public, and industrial settings each bend the approach without breaking it, and clients unable to accept ranges are served by risk-loaded fixed pricing with transparent credit protocols. This adaptability reflects its origin in working pre-sales practice rather than in abstraction. Practitioner experience closing multi-tier enterprise engagements informs every calibration offered. The framework does not claim to eliminate uncertainty, which no method can. It claims to size and price that uncertainty honestly, which is what the deterministic tradition failed to do. That honest framing, rather than any promise of certainty, is what earns the trust the framework is built to protect.

What the framework needs next is evidence beyond accumulated experience. A controlled study tracking framework-sized engagements against conventionally sized ones would convert practitioner conviction into validated finding, and the tracking discipline described here provides the instrumentation such a study would require. The distribution recommendations left open could likewise be grounded empirically as records accumulate. Offered as practitioner guidance rather than settled science, the framework aims to make a difficult judgment more disciplined and more teachable. Its measure of success is straightforward: engagements that deliver on the commitments made when they were sized, engagement after engagement. Practitioner conviction points the way, but disciplined measurement is what will ultimately confirm or correct it.

## REFERENCES

- [1] S. McConnell, *Software Estimation: Demystifying the Black Art*. Redmond, WA: Microsoft Press, 2006.
- [2] D. D. Galorath and M. W. Evans, *Software Sizing, Estimation, and Risk Management: When Performance Is Measured Performance Improves*. Boca Raton, FL: Auerbach Publications, 2006.
- [3] J. Ryseff and A. Narayanan, "Why AI Projects Fail," RAND Corporation, Rep. PT-A2680-1, 2025. doi: 10.7249/PTA2680-1.
- [4] Project Management Institute, *A Guide to the Project Management Body of Knowledge (PMBOK Guide)*, 7th ed. Newtown Square, PA: PMI, 2021.
- [5] R. Canesi, L. Gabrielli, and G. Marella, "Probabilistic risk assessment framework for cost overruns predictions in infrastructure projects using randomized simulations," *Computer-Aided Civil and Infrastructure Engineering*, vol. 40, no. 27, pp. 4774-4796, 2025. doi: 10.1111/mice.70100.
- [6] O. Roll and J. L. Geerties, "Unlocking value-based pricing: the moderating roles of pricing capabilities and contingency factors in B2B markets," *Journal of Revenue and Pricing Management*, vol. 24, pp. 232-253, 2025. doi: 10.1057/s41272-025-00530-z.
- [7] C. H. Rashid, I. Shafi, J. Ahmad, et al., "Software Cost and Effort Estimation: Current Approaches and Future Trends," *IEEE Access*, vol. 11, pp. 99268-99288, 2023. doi: 10.1109/ACCESS.2023.3312716.
- [8] R. R. Putri, D. Siahaan, and C. Fatichah, "A Comparative Study on COCOMO II Model for Cost Estimation," in *Proc. 13th IEEE Int. Conf. ICCSCE*, Penang, 2023, pp. 226-231. doi: 10.1109/ICCSCE58721.2023.10237162.
- [9] M. Fernandez-Diego, E. R. Mendez, et al., "An Update on Effort Estimation in Agile Software Development: A Systematic Literature Review," *IEEE Access*, vol. 8, pp. 166768-166800, 2020. doi: 10.1109/ACCESS.2020.3021664.
- [10] M. Fu and C. Tantithamthavorn, "GPT2SP: A Transformer-Based Agile Story Point Estimation Approach," *IEEE Transactions on Software Engineering*, vol. 49, no. 2, pp. 611-625, 2023. doi: 10.1109/TSE.2022.3158252.
- [11] A. S. Cheemaa, M. Azhar, F. Arif, et al., "EGPT-SPE: story point effort estimation using improved GPT-2 by removing inefficient attention heads," *Applied Intelligence*, vol. 55, art. 994, 2025. doi: 10.1007/s10489-025-06824-4.
- [12] Mizanur Rahman et al., "Review and Empirical Analysis of Machine Learning-Based Software Effort Estimation," *IEEE Access*, 2024. doi: 10.1109/ACCESS.2024.3404879.
- [13] M. Awais et al., "A Comparative Analysis of Popular Software Effort Estimation Techniques," in *Proc. IEEE Conf.*, 2023.
- [14] M. Abadeer and M. Sabetzadeh, "Machine Learning-based Estimation of Story Points in Agile Development: Industrial Experience and Lessons Learned," in *Proc. 2021 IEEE 29th Int. Requirements*

- Engineering Conf. Workshops (REW), 2021, pp. 106-115. doi: 10.1109/REW53955.2021.00022.
- [15] K. J. W. Arachchi and C. R. J. Amalraj, "An Agile Project Management Supporting Approach for Estimating Story Points in User Stories," in *Proc. 2023 Int. Conf. ICITR*, 2023. doi: 10.1109/ICITR61062.2023.10382930.
- [16] M. M. I. Shamim, A. B. B. A. Hamid, T. E. Nyamasvisva, and N. S. B. Rafi, "Advancement of Artificial Intelligence in Cost Estimation for Project Management Success," *Modelling*, vol. 6, no. 2, art. 35, 2025. doi: 10.3390/modelling6020035.
- [17] J. M. Lekraik and O. S. Ojekemi, "Advancing Digital Project Management Through AI: An Interpretable POA-LightGBM Framework for Cost Overrun Prediction," *Systems*, vol. 13, no. 12, art. 1047, 2025. doi: 10.3390/systems13121047.
- [18] S. Salimimoghadam et al., "The Rise of Artificial Intelligence in Project Management: A Systematic Literature Review," *Buildings*, vol. 15, no. 7, art. 1130, 2025. doi: 10.3390/buildings15071130.
- [19] E. Pereira, A. W. Graylin, and E. Brynjolfsson, "The Enterprise AI Playbook: Lessons from 51 Successful Deployments," Stanford Digital Economy Lab, Stanford Univ., Apr. 2026.
- [20] S. Oesterle et al., "Investigating the Co-Creation of IT Consulting Service Value," *Electronic Markets*, vol. 32, no. 2, pp. 571-597, 2022. doi: 10.1007/s12525-020-00426-3.
- [21] Minhui Hu, "Fixed-price or performance-based? Contract design with renegotiation for IT outsourcing," *Journal of Management Science and Engineering*, vol. 11, no. 1, 2026.
- [22] J. Keranen, M. Kienzler, A. Salonen, H. Terho, and D. Totzek, "Gain-sharing in performance-based contracting," *Industrial Marketing Management*, vol. 115, pp. 1-13, 2023.