ISSN: 2754-6659

Journal of Artificial Intelligence & Cloud Computing

SCIENTIFIC Research and Community

Research Article

Open Access

S³: A Multicriteria Pre-Deployment Scoring Protocol for Enterprise Small Language Model Selection

Smita Sengupta[1], Smitha Rajappa[2] and Ashim Bose[3]*

[1]MS in Management Science, Naveen Jindal School of Management, The University of Texas at Dallas, USA

[2]MS in Information Technology and Management (ITM), Naveen Jindal School of Management, The University of Texas at Dallas, USA

[3]Professor of Practice, Jindal School of Management, University of Texas at Dallas, Richardson, TX 75080, USA

ABSTRACT

Technology managers lack a systematic protocol for determining which enterprise workloads are safe for local small language model deployment. We introduce S³ (Suitability Score for SLM Selection), a multicriteria framework that scores workloads across six dimensions, Task Complexity, Output Structure, Stakeholder Risk, Data Sensitivity, Latency Tolerance, and Volume Load, and assigns them to Pure SLM, Hybrid, or LLM-Only tiers before inference. Non-compensatory gate rules enforce safety constraints for high-stakes scenarios, ensuring that life-critical or expert-level generative tasks are never assigned to a local model regardless of their aggregate score. Thresholds calibrated on an independent use-case set correctly predicted all eight held-out tier assignments across six industries, with Pure SLM tiers achieving at least 95% LLM parity.

***Corresponding author**

Ashim Bose, Professor of Practice, Jindal School of Management, University of Texas at Dallas, Richardson, TX 75080, USA.

Received: November 11, 2026; **Accepted:** November 16, 2026; **Published:** November 24, 2026

Keywords: Small Language Models, SLM Deployment, Enterprise AI Governance, Multicriteria Decision Making, Weighted Sum Model, AI Risk Management

Introduction

The commercial adoption of large language models (LLMs) has created a practical tension for enterprise technology managers. Hosted LLMs deliver strong performance across a wide range of tasks but impose per-query API costs, introduce network latency, and raise data residency concerns that many regulated organizations cannot accept. Locally deployed small language models (SLMs) typically ranging from 1 to 10 billion parameters can be run on commodity hardware at negligible marginal cost, keeping sensitive data within the organization's infrastructure. Recent architectural innovations demonstrate that carefully designed SLMs can match much larger models on structured, well-scoped tasks [1-3]. Distillation, training a smaller model to mimic a larger one, and quantization, compressing model weights to reduce memory usage have further narrowed the performance gap on many workload types [4-6].

However, knowing that SLMs can sometimes match LLMs on benchmark aggregates does not tell a technology manager which specific workloads are safe for local deployment. Today, most organizations answer this question informally: through pilot studies, vendor benchmarks, or intuition. No widely adopted, auditable protocol exists that translates workload characteristics into a defensible deployment decision before any production traffic is committed.

This paper addresses that gap. We introduce S³ (Suitability Score for SLM Selection), a multicriteria decision-making framework grounded in the weighted sum model [7,8]. S³ evaluates any enterprise workload across six dimensions: Task Complexity (TC), Output Structure (OS), Stakeholder Risk (SK), Data Sensitivity (DS), Latency Tolerance (LT), and Volume Load (VL). It produces a normalized score on a 1-to-5 scale that maps to one of three deployment tiers: Pure SLM, Hybrid, or LLM Only. Three non-compensatory gate rules encode safety constraints that cannot be overridden by the weighted sum, ensuring that irreversible, legally consequential, or expert-level generative tasks are never assigned to a local SLM autonomously.

We validate S³ through an empirical evaluation involving eight enterprise use cases across six industry domains, benchmarked on six local SLMs and one cloud-based LLM baseline. All eight tier predictions, generated from independently-derived thresholds, are confirmed. S³ provides the structured evaluation that Rogers (2003) identifies as a prerequisite for moving from pilot to production, the facilitating condition that Venkatesh et al. (2003) show is necessary for technology adoption, and the auditable decision trail that NIST (2023) and the EU AI Act (European Parliament, 2024) increasingly require [9-12].

The remainder of this paper is organized as follows. Section 2 reviews related literature. Section 3 presents the research methodology and research questions. Section 4 introduces the S³ framework. Section 5 describes implementation requirements. Section 6 presents the empirical evaluation design. Section 7 reports results and analysis. Section 8 discusses implications for

technology managers. Section 9 addresses limitations and future research. Section 10 concludes.

Background and Related Work

The LLM-to-SLM Transition

The scaling laws literature established that language model performance scales predictably with parameter count, data volume, and compute [13,14]. However, recent work challenges the assumption that scale is always necessary: task-specific fine-tuning, knowledge distillation, and architectural innovations have produced sub-10B models that match 70B-class performance on structured, well-scoped workloads [15- 17]. The practical question for enterprise deployment is no longer whether SLMs can match LLMs (the answer depends on the task) but which tasks are structurally amenable to SLM deployment.

Model Routing and Cascading Frameworks

Several frameworks address model selection at runtime. Proposed Hybrid LLM, a cost-efficient query routing system [18]. Ong et al. introduced RouteLLM. Hu et al. published RouterBench [19]. Aggarwal et al. proposed AutoMix for automatic model-mixing [20,21]. Dekoninck et al. [22] presented a unified approach to routing and cascading [22]. These frameworks operate at the query level during runtime and do not address the prior strategic question of whether an SLM should be deployed for a task category at all. S³ answers this before any production traffic exists.

Multicriteria Decision Making in Technology Management

The weighted sum model (WSM) is the most widely deployed multicriteria decision-making (MCDM) method in practice due to its transparency and interpretability [7]. Triantaphyllou provides the theoretical grounding for threshold-based tier assignment using WSM [8]. Edwards and Barron demonstrate that simple multi-attribute rating technique using swings (SMARTS) approaches are robust to weight imprecision when alternatives are spread across the score range [23]. Zavadskas and Turskis survey MCDM applications in technology and economics, demonstrating that pairwise criteria weighting and scoring rubrics produce reliable decisions when anchored to measurable criteria [24].

Technology Adoption Frameworks

The Technology Acceptance Model and the Unified Theory of Acceptance and Use of Technology identify perceived usefulness and facilitating conditions as the primary drivers of technology adoption [10,25]. The Technology-Organization-Environment framework highlights that deployment decisions require alignment across capability, organizational readiness, and regulatory requirements [26]. S³ operationalizes these prerequisites across its six scoring dimensions.

AI Governance and Risk Frameworks

NIST identifies auditability, accountability, and explainability as core properties of trustworthy AI systems [11]. The EU AI Act (European Parliament) requires documented evidence of risk assessment for high-risk AI applications, with explicit requirements that safety-critical systems include human oversight [12]. Floridi et al. define five ethical AI properties that map directly to the Stakeholder Risk and Data Sensitivity dimensions in S³ [27].

The Critical Gap

The literature presents a clear gap. MCDM frameworks have not been applied to the specific problem of SLM tier assignment. Routing frameworks assume the deployment decision has already been made. AI governance frameworks identify the need for pre-deployment assessment but do not provide operational scoring protocols. This paper fills that gap by providing an auditable, empirically evaluated pre-deployment scoring protocol for SLM deployment decisions in enterprise settings.

Research Methodology

Design Science Research Paradigm

We follow the design science research (DSR) paradigm, which positions the artifact the S³ scoring protocol as the primary research contribution [28,29]. Relevance is established by the identified gap: enterprises lack a systematic method for SLM deployment decisions that is auditable before inference. Rigor is established through grounding in MCDM theory structured evaluation across eight use cases, and tier predictions generated from independently-derived thresholds [7,8].

Framework Development

The S³ framework was developed through five steps. First, we identified the decision attributes that enterprise practitioners evaluate when considering SLM deployment. Second, we turned these attributes into six measurable dimensions, each with a clear 1-to-5 scoring scale where every score is tied to a specific, concrete description. Third, we selected the WSM as the aggregation method. Fourth, we identified non-compensatory safety constraints and encoded them as gate rules [30]. Fifth, we derived tier thresholds from an independent set of eight calibration use cases and validated them, without adjustment, against a separate set of eight held-out use cases.

Expert Elicitation Protocol

All eight use cases were initially scored by the author across all six dimensions (48 dimension-use case pairs) and reviewed by a single independent domain expert for face validity. Each dimension has published 1-to-5 anchors (Table 1) that constrain interpretation and reduce rater subjectivity. This approach is structurally identical to clinical severity scoring systems such as APACHE II and CURB-65 [31,32]. The scoring worksheet was committed to the project repository before any model inference was executed. In a separate, later exercise, two different independent domain experts each scored all 48 dimension-use case pairs from scratch, without reference to the author's original scores. All six resulting disagreements between these two experts were minor (maximum 1 point) and resolved by reference to Table 1 without arbitration. The reconciled scores from this second, independent exercise are the scores used throughout this paper, including in Table 6 and all subsequent analyses.

Research Questions

We investigate three research questions that test whether S³ works as a reliable, safe, and robust deployment decision tool for enterprise technology managers.

RQ1 (Prediction Accuracy): Does the S³ score correctly predict the deployment tier, Pure SLM, Hybrid, or LLM Only for a given workload before any model inference is run?
RQ2 (Gate Rule Necessity): Do the non-compensatory gate rules prevent unsafe deployment decisions that the weighted sum formula alone would not catch?
RQ3 (Robustness): Do the tier assignments produced by S³ remain stable when the dimension weights are varied across different organizational priorities?

The S³ Framework

S³ Suitability Score for SLM Selection is a structured decision-making tool that helps technology managers determine, before deploying any model, whether a given business task is appropriate

for a Small Language Model, requires a Large Language Model, or needs a combination of both.

The Six Scoring Dimensions

Every business task is evaluated on six dimensions. Each dimension captures a different aspect of what makes a task easy or hard for an SLM to handle reliably. A domain expert assigns an integer score from 1 (lowest; most SLM-friendly) to 5 (highest; most demanding). Table 1 presents the full scoring scale with behavioral anchors.

Table 1: S³ Dimension Scoring Scale: Behavioral Anchors

Dim.	Score 1	Score 2	Score 3	Score 4	Score 5
TC	Binary classification	Few-class classification	Multi-class with edge cases	Multi-step reasoning	Expert judgment; compositional reasoning
OS	Free-form paragraph	Soft-structured paragraph	Named fields / semi-schema	Defined labels; closed vocabulary	Single word/value; strict schema
SK	No consequence; informational	Minor; easily corrected	Business impact; rework required	Significant harm; affects persons	Severe, irreversible, legally consequential
DS	Fully public data	Semi-public / internal	Personal data; not regulated	Regulated PII/PHI	Classified, ITAR-controlled
LT	Batch; hours acceptable	Background; minutes acceptable	Interactive; P95 < 2s	Near real-time; P95 < 500ms	Real-time; P95 < 100ms
VL	Dozens per day	Hundreds per day	Tens of thousands/ day	Hundreds of thousands/day	Millions per day

Task Complexity (TC)

Task Complexity measures how intellectually difficult the task is. A score of 1 denotes binary classification (e.g., spam or not spam). A score of 5 denotes expert-level judgment requiring compositional reasoning, such as composing a clinical diagnosis from ambiguous patient data.

Output Structure (OS)

Output Structure measures how precisely the model's output must be defined. A score of 1 accepts any free-form paragraph. A score of 5 requires a single word or strictly defined value from a closed vocabulary. Highly structured outputs are actually easier for models to produce consistently.

Stakeholder Risk (SK)

Stakeholder Risk measures the consequence of a wrong answer. A score of 1 means no meaningful consequence. A score of 5 means errors are severe, irreversible, and legally consequential. This dimension carries the highest weight and triggers the most restrictive gate rules. This graduated severity structure parallels risk-tiering approaches in the NIST AI Risk Management Framework and clinical severity-scoring systems such as APACHE II and CURB-65, which similarly stratify consequence severity to guide intervention thresholds [11,31,32].

Data Sensitivity (DS)

Data Sensitivity measures how legally and ethically sensitive the data is. A score of 1 denotes fully public data. A score of 5 denotes data classified under regulations such as ITAR (International Traffic in Arms Regulations). Intermediate scores cover personal data under GDPR and regulated health data under HIPAA. PII refers to Personally Identifiable Information; PHI refers to Protected Health Information. This categorization follows established data-classification conventions such as ISO/IEC 27001 and NIST SP 800-53, which similarly grade information by regulatory and confidentiality tier [33,34].

Latency Tolerance (LT)

Latency Tolerance measures required response speed. A score of 1 denotes batch processing where hours are acceptable. A score of 5 requires a response within 100 milliseconds in 95 out of every 100 requests (P95 < 100ms, where P95 is the 95th percentile of observed response times). Because SLMs are faster than LLMs, a strict latency requirement actually favors SLM deployment. These response-time tiers reflect standard service-level objective conventions used in large-scale production systems, which similarly define percentile-based latency thresholds for service classification [35].

Volume Load (VL)

Volume Load measures daily request volume. A score of 1 denotes dozens of requests per day. A score of 5 denotes millions. This dimension receives the lowest weight because volume affects cost efficiency, not the quality or safety of the output.

Weight Selection and Organizational Profiles

The weight vectors in Table 2 were developed through structured expert elicitation informed by Analytic Hierarchy Process principles, rather than a full formal AHP implementation [36]. During framework development, the domain experts introduced in Section 3.3 discussed and compared the relative importance of each dimension pair within the context of each organizational profile (e.g., whether Stakeholder Risk should be weighted more heavily than Task Complexity in healthcare or finance settings). These comparative judgments were translated directly into the discrete 1–5 integer weight scale adopted by S³, without computing formal pairwise comparison matrices or consistency ratios, to improve transparency and ease of practitioner use for non-specialist organizational stakeholders. This discretization preserves the intended ordering of priorities while keeping the framework simple to audit and apply in organizational settings. Consistent with the framework's non-compensatory philosophy, every profile additionally satisfies the constraint $w_SK \geq w_TC$, ensuring that stakeholder risk is never weighted below task complexity regardless of organizational priorities [37].

Table 2: Weight Selection Guide for Organizational Profiles

Profile	TC	OS	SK	DS	LT	VL	Appropriate Context
Default	3	2	4	2	3	1	General enterprise; balanced risk posture
Security-First	2	1	5	4	2	1	Healthcare, finance, defence
Latency-First	2	2	3	1	5	3	Real-time consumer-facing applications
Balanced	3	3	3	3	3	3	Exploratory or pilot settings
Volume-Heavy	2	2	3	2	2	5	High-throughput batch processing

Each profile sets a different weight vector, reflecting how a given organisation trades off capability against safety and cost. The Default profile assumes no strong contextual pressure and is a reasonable starting point when a manager is unsure which profile applies. Security-First raises the weight on Stakeholder Risk and Data Sensitivity, making the tier assignment more conservative for regulated sectors such as healthcare, finance, and defence, where a wrong answer carries legal or safety consequences. Latency-First raises the weight on Latency Tolerance for real-time, consumer-facing workloads where response speed dominates the deployment decision. Balanced assigns equal weight to all six dimensions and is intended for exploratory or pilot use where no single dimension should dominate the score. Volume-Heavy raises the weight on Volume Load for high-throughput batch processing, where request volume drives infrastructure cost more than latency or output structure does. In every profile, w_SK is held at or above w_TC so that stakeholder risk can never be discounted relative to task complexity, regardless of context. A manager selects the profile that best matches their deployment context before scoring a workload; Section 7.3 shows that most tier assignments are stable across all five profiles, with the profile choice changing the outcome only in genuinely borderline cases.

The S³ Formula

Once scores and weights are selected, S³ computes a single composite score using the Weighted Sum Model (WSM) one of the most widely used MCDM methods [7,8]. The formula is:

$$S^3 = [\ \Sigma(score_i \times w_i) \div \Sigma(5 \times w_i)\ ] \times 5 \quad (1)$$

The formula works in three steps. First, each dimension score is multiplied by its weight and all six results are summed to produce the raw weighted sum. Second, this is divided by the maximum possible raw weighted sum, computed by assuming every dimension scores 5. Under the Default profile this maximum equals 5 × (3+2+4+2+3+1) = 75. Third, multiplying by 5 rescales the result to a 1-to-5 range. A lower S³ score indicates greater suitability for autonomous SLM deployment.

Two dimensions, Output Structure and Latency Tolerance, are reverse-scored before aggregation. Both are defined such that a higher raw score corresponds to a task property (rigid output format, strict response-time requirement) that is more easily satisfied by a small model, the opposite of the direction assumed by the other four dimensions. Consistent with standard practice in composite-indicator construction (Mazziotta and Pareto, 2017; OECD/JRC, 2008), each raw score is inverted (score' = 6 − score) prior to weighting, ensuring all six dimensions share a common polarity in which a higher value indicates greater difficulty for autonomous SLM deployment.

Tier Mapping

The S³ score maps to one of three deployment tiers using two fixed thresholds: $\tau_1 = 3.0$ and $\tau_2 = 3.7$, calibrated on an independent set of eight use cases distinct from the validation set.

To avoid deriving thresholds and validation results from the same data, thresholds were calibrated on an independent set of eight enterprise use cases, distinct from the eight used for validation in Section 6. Cases scoring at or below 3.0 consistently retained strong SLM-to-LLM parity; one case at $S^3 = 3.20$ showed reduced parity (64%), while cases scoring higher still (3.47 and 3.67) retained strong parity, indicating that performance does not decline monotonically with S³ score in this range. τ_1 was accordingly set at 3.0, below the observed point of reduced performance, and τ_2 was set at 3.7, the upper bound of the calibration set's confirmed-safe range. Applied without adjustment to the original eight validation cases, these thresholds correctly predicted all eight tier assignments (Section 6.4).

Pure SLM ($S^3 \leq 3.0$): The local SLM handles this workload autonomously without LLM escalation.
Hybrid ($3.0 < S^3 \leq 3.7$): The SLM handles most queries but escalates uncertain cases to a cloud LLM.
LLM Only ($S^3 > 3.7$): The workload requires a large cloud model. Mandatory human oversight applies for safety-critical cases.

If a workload scores within 0.20 points of either threshold, the assignment is treated as uncertain and a pilot benchmark should be run before committing infrastructure.

Prescreening Gate Rules

The S³ formula is a weighted average, which allows strong performance on one dimension to offset poor performance on another. In safety-sensitive settings, this behavior is undesirable. Three non-compensatory gate rules are evaluated before the formula is computed. If any fires, the tier is determined by the rule, not the score [30].

Hard Rule 1 - Severe Consequence Disqualification: If SK = 5, assign LLM Only with mandatory human oversight, regardless of formula score.

Hard Rule 2 - Expert Generative Task Block: If TC = 5 AND SK ≥ 4, assign LLM Only regardless of formula score. This targets expert-level generative tasks where SLM hallucination risk is unacceptable.

Flag Rule - Minimum Hybrid Floor: If SK ≥ 4 but Hard Rule 1 does not apply, the minimum tier is Hybrid, ensuring an LLM escalation path always exists.

Complete Decision Procedure and Worked Example

The S³ procedure follows three sequential steps. To illustrate concretely, consider a Finance Manager determining whether a local SLM can extract vendor name, invoice date, and total amount from supplier invoices.

Step 1: Score all six dimensions. For each dimension, the assessor reads the Table 1 anchors and assigns the matching score. Table 3 shows the scoring for invoice extraction with full reasoning.

Table 3: Step-by-Step Dimension Scoring: Invoice Field Extraction

Dim.	Score	Reasoning
TC	2	Extraction, not reasoning. Answer is present in the document; no inference needed.
OS	4	Three specific fields in fixed format. No free-form text permitted.
SK	2	Errors caught by accounts payable before payment. Consequence is rework only.
DS	3	Internal financial documents. Not regulated under GDPR or HIPAA.
LT	2	Overnight batch processing. No real-time requirement.
VL	4	Hundreds of invoices processed daily across multiple suppliers.

Step 2: Check gate rules. SK = 2 so Hard Rule 1 does not apply. TC = 2 and SK = 2 so Hard Rule 2 does not apply. SK = 2 so the Flag Rule does not apply. No gate rules fire. Proceed to Step 3.

Step 3: Compute S^3 using the Default weight profile w = (3, 2, 4, 2, 3, 1).

Apply polarity inversion: OS' = 6 − 4 = 2; LT' = 6 − 2 = 4.

Raw weighted sum = (2×3)+(2×2)+(2×4)+(3×2)+(4×3)+(4×1) = 6+4+8+6+12+4 = 40

Maximum = 5×(3+2+4+2+3+1) = 75

S^3 = 40 ÷ 75 × 5 = 2.67

Step 4: Apply tier threshold. $S^3 = 2.67 < \tau_1 = 3.0 \rightarrow$ Pure SLM. The Finance Manager can proceed with local SLM deployment for invoice extraction.

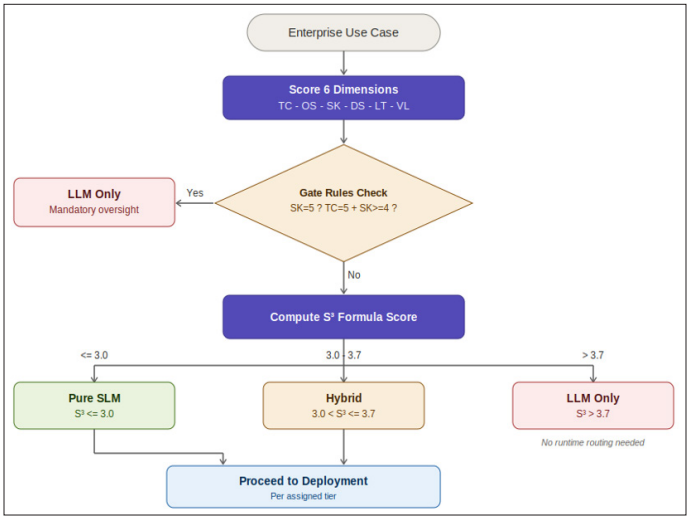

Figure 1: S^3 Decision Procedure Flowchart. Gate rules are evaluated first. Hard Rule 1 (SK = 5) and Hard Rule 2 (TC = 5 and SK ≥ 4) route directly to LLM Only. The formula score maps to Pure SLM (≤ 3.0), Hybrid (3.0–3.7), or LLM Only (> 3.7). The Flag Rule elevates Pure SLM assignments where SK ≥ 4 to Hybrid

Implementation and Deployment Requirements

Inference Infrastructure

All SLMs are run locally on a single Apple M2 Pro MacBook Pro (16 GB unified memory, 12 CPU cores) using Ollama version 0.6.2, an open-source tool for local language model deployment [38]. The LLM baseline (Llama-3.3-70B) is served via the Groq API free tier [39]. All inference runs at temperature 0 with seed 42, producing fully deterministic outputs. Table 4 maps representative hardware profiles to typical enterprise deployment scales.

Table 4: Deployment Environment Profiles

Profile	Hardware	SLM	LLM	Suitable For
Developer	M2 Pro 16 GB	Ollama local	Groq free	Evaluation; prototyping
Small Enterprise	M2 Ultra 64 GB	Ollama local	Groq paid	Up to 50k queries/day
Mid-Market	A100 40 GB GPU	Ollama/ vLLM	OpenAI API	Up to 500k/ day
Enterprise	Multi-GPU cluster	vLLM dist.	Private LLM	Millions/ day
Regulated	Air-gapped GPU	Ollama only	Local Llama	Healthcare, defence

Cost Implications

Table 5 summarizes estimated annual infrastructure cost and API cost savings versus full LLM deployment for each tier.

Table 5: Hardware Requirements and Estimated Annual Cost Implications

Tier	Min RAM	Est. Annual Cost	Est. API Saving vs LLM
Pure SLM	16 GB	$0 (existing hardware)	$7,300–$73,000/ year
Hybrid	16 GB	$600–$2,400/ year	$3,650–$36,500/ year
LLM Only	N/A	$1,200–$12,000/year	$0

Practical Deployment Protocol

We recommend the following six-step deployment protocol for technology managers:

- **Step 1:** Workload Characterization: Document task type, expected query volume, output format, and stakeholder impact.
- **Step 2:** S^3 Scoring: Score the workload on all six dimensions using Table 1 anchors and the appropriate weight profile from Table 2.
- **Step 3:** Gate Check and Tier Assignment: Apply gate rules first, then compute the formula score and map to a tier.
- **Step 4:** Boundary Sensitivity Check: If the score falls within 0.20 of either threshold, run a pilot benchmark before committing infrastructure.
- **Step 5:** Infrastructure Provisioning: Provision local inference using a model appropriate for the hardware profile in Table 4.
- **Step 6:** Production Monitoring: Monitor accuracy and latency on a representative sample of live queries weekly. Update S^3 if the workload distribution changes materially.

Empirical Evaluation

Use Case Selection

Eight enterprise use cases were selected to span the full range of the S³ scoring space, across six industry verticals: financial services (UC2, UC8), telecommunications (UC1, UC3), retail and e-commerce (UC4), technology (UC5), healthcare (UC6), and legal services (UC7). Two use cases (UC6, UC8) were selected specifically to test the gate mechanism: UC6 tests whether Hard Rule 1 prevents a potentially unsafe formula-only assignment, while UC8 tests whether Hard Rule 2 independently reinforces a high-risk assignment that the weighted score also identifies as LLM Only. Dimension scores for all eight use cases were recorded before any model inference was executed; tier predictions were then generated by applying thresholds derived independently from a separate calibration set. Table 6 summarizes all eight use cases.

Table 6: Use Case Summary

UC	Task	Industry	S³	Predicted Tier
1	SMS Threat Detection	Telecom	3.00	Hybrid (Flag Rule: SK=4)
2	Invoice Field Extraction	Financial	2.67	Pure SLM
3	Support Ticket Routing	Telecom	2.40	Pure SLM
4	Product Review Sentiment	Retail	2.07	Pure SLM
5	Automated Code Review	Technology	3.40	Hybrid
6	Clinical Triage	Healthcare	3.20	LLM Only (Hard Rule 1)
7	Legal Contract Risk	Legal	3.60	Hybrid
8	Financial Report Drafting	Financial	4.00	LLM Only (Hard Rule 2)

Model Selection

Six SLMs spanning 3B to 8B parameters were benchmarked alongside one 70B LLM baseline. All SLMs are open-weight, instruction-tuned, and run locally via Ollama. Table 7 lists the full model registry.

Table 7: Model Registry

Model	Params	Architecture	Release	Inference
Llama-3.2-3B-Instruct	3B	LLaMA	Sep 2024	Local
Phi-4-Mini-Instruct	3.8B	Phi	Feb 2025	Local
Gemma-3-4B-Instruct	4B	Gemma	Mar 2025	Local
Qwen-2.5-7B-Instruct	7B	Qwen	Sep 2024	Local
Mistral-7B-Instruct-v0.3	7B	Mistral	Jun 2024	Local
Llama-3.1-8B-Instruct	8B	LLaMA	Jul 2024	Local
Llama-3.3-70B-Versatile	70B	LLaMA	Dec 2024	Groq API (baseline)

Gold Sets and Evaluation Protocol

Each use case is evaluated on a 30-item held-out test set with known correct answers, kept entirely separate from model training data. Table 8 presents the gold set structure for all eight use cases.

Table 8: Gold Set Structure by Use Case

UC	Task	Items	Difficulty Split	Domain Reference	Metric
1	SMS Threat Detection	30	10/10/10	UCI SMS Spam Collection	Exact-match
2	Invoice Extraction	30	10/10/10	Standard AP templates	Exact-match
3	Support Routing	30	10/10/10	ITIL service taxonomy	Exact-match
4	Product Sentiment	30	10/10/10	Stanford Sentiment Treebank	Exact-match
5	Code Review	30	10/10/10	CWE taxonomy	Exact-match
6	Clinical Triage	30	4/11/11/4 by class	Manchester Triage System	Exact-match + per-class recall
7	Legal Contract	30	10/10/10	Public contract templates	Exact-match
8	Financial Report	30	10/10/10	SEC EDGAR quarterly filings	LLM-judge + manual audit

Primary accuracy is reported as exact-match accuracy on the 30-item test set. For UC6, class-level recall is reported in addition to overall accuracy, because overall accuracy is insufficient when error costs differ across classes. Statistical comparison uses McNemar's test on paired binary item-level outcomes. All accuracy figures are reported with 95% Wilson confidence intervals [40].

Results and Analysis

This section reports empirical findings from the benchmark study. Each subsection directly addresses one of the three research questions in Section 3.4.

RQ1: Does S³ Correctly Predict the Deployment Tier?

Finding: S³ correctly predicted the deployment tier for all eight use cases (8/8; 95% Wilson CI: 63.1%, 100%).

Before running any model inference, S³ assigned each of the eight use cases to a predicted deployment tier. These predictions were generated by applying independently-derived thresholds to dimension scores recorded before any benchmark results were seen. The benchmark results confirmed every prediction. Table 9 presents full results. Parity measures how close the best SLM's accuracy is to the LLM baseline: parity above 100% means the SLM outperformed the LLM; parity below 95% indicates a meaningful shortfall.

Table 9: Benchmark Results: All Eight Tier Predictions Confirmed

UC	Task	Tier	Best SLM	SLM	LLM	Parity	χ^2	p	OK?
1	SMS Threat	Hybrid	Llama-3.1-8B	93.3%	92.2%	101.2%	0.20	0.66	✓
2	Invoice Ext.	Pure SLM	Phi-4-Mini	92.2%	91.1%	101.2%	0.09	0.76	✓
3	Ticket Routing	Pure SLM	4 tied	86.7%	83.3%	104.1%	0.33	0.56	✓
4	Sentiment	Pure SLM	Llama-3.1-8B	97.8%	93.3%	104.8%	0.00	1.00	✓
5	Code Review	Hybrid	Qwen2.5-7B	76.7%	85.6%	89.6%	1.80	0.18	✓
6	Triage	LLM Only	Llama-3.1-8B	73.3%	70.0%	104.7%	0.09	0.76	✓
7	Legal Risk	Hybrid	Mistral-7B	82.2%	91.1%	90.2%	2.00	0.16	✓
8	Fin. Report	LLM Only	Phi-4-Mini	81.1%	90.3%	89.7%	1.44	0.23	✓

The three Pure SLM use cases (UC2, UC3, UC4) all achieve parity above 100%, confirming that local SLMs can fully replace the LLM baseline on these tasks. The Hybrid cases (UC1, UC5, UC7) show performance gaps that justify the mixed deployment approach. The LLM Only cases (UC6, UC8) are discussed under RQ2 below.

Note on Confidence Interval: The 95% Wilson CI lower bound of 63.1% means the framework could assign approximately 1 in 3 use cases incorrectly on a larger sample. External replication on held-out workloads is the most important next step.

RQ2: Do Gate Rules Prevent Unsafe Deployments?
Finding: UC6 demonstrates the necessity of the non-compensatory gate mechanism; UC8 demonstrates that a gate rule can independently reinforce a high-risk assignment even when the weighted score alone reaches the same tier.

UC6: Clinical Triage (Hard Rule 1): The best SLM achieves 73.3% overall accuracy against the LLM baseline's 70.0%, giving a parity of 104.7%. By aggregate accuracy alone the SLM appears superior. The class-level breakdown in Table 10 reveals why autonomous SLM deployment would be dangerous.

Table 10: UC6 Clinical Triage: Per-Class Recall by Model

Model	Params	Overall	CRITICAL (n=4)	URGENT (n=11)	SEMIURGENT (n=11)	NONURGENT (n=4)
Llama-3.3-70B (baseline)	70B	70.0%	100.0%	81.8%	36.4%	100.0%
Llama-3.1-8B	8B	73.3%	100.0%	63.6%	90.9%	25.0%
Qwen2.5-7B	7B	63.3%	100.0%	63.6%	36.4%	100.0%
Llama-3.2-3B	3B	50.0%	100.0%	9.1%	81.8%	25.0%
Phi-4-Mini	3.8B	36.7%	100.0%	54.5%	0.0%	25.0%
Gemma-3-4B	4B	30.0%	100.0%	0.0%	27.3%	50.0%
Mistral-7B	7B	23.3%	100.0%	0.0%	27.3%	0.0%

Recall measures the proportion of actual cases in a severity class correctly identified by the model. URGENT recall ranges from 0.0% to 63.6% across the six SLMs, against the LLM baseline's 81.8%. Under the Manchester Triage System, URGENT patients require treatment within 10 minutes. Two SLMs achieve 0% URGENT recall, correctly identifying none of the 11 urgent patients, while appearing acceptable at the aggregate level. Hard Rule 1 (SK = 5 ⇒ Disqualified) catches this failure categorically, before any benchmark is run.

UC8: Financial Report Drafting (Hard Rule 2): UC8 receives a formula score of 4.00, above $\tau_2 = 3.7$. Even without gate rules, the formula alone assigns LLM Only. Hard Rule 2 also fires because TC = 5 and SK = 4, providing an independent, consistent basis for the same assignment. Manual review identified three fabricated financial figures and one misattributed executive quotation in SLM outputs, hallucinations not captured by exact-match accuracy scoring.

UC7: Legal Contract Risk: UC7 scores 3.60, between $\tau_1 = 3.0$ and $\tau_2 = 3.7$, assigning Hybrid on the formula alone; no gate rule applies to this case. The benchmark confirms this classification: Mistral-7B achieves 82.2% against the LLM's 91.1% (parity 90.2%), well below the 95% Pure SLM threshold.

RQ3: Are Tier Assignments Stable When Weights Change?
Finding: Seven of eight tier assignments are stable across all five weight profiles. The sole sensitive case is correctly identified and flagged by the framework's own boundary rule.

To test robustness, S^3 was recomputed under each of the five organizational weight profiles from Table 2, and any resulting change in tier assignment was recorded [37]. Table 11 reports the results.

Table 11: Sensitivity Analysis: Minimum Weight Change to Flip Tier Assignment

Task	S³	Tier	Stability	Min. Change to Flip
SMS Threat Detection	3.00	Hybrid	Gate-protected	N/A: Flag Rule floor (formula alone sits exactly at τ_1)
Invoice Extraction	2.67	Pure SLM	Sensitive	Flips to Hybrid under Latency-First profile
Support Routing	2.40	Pure SLM	Robust	Robust across all 5 profiles
Product Sentiment	2.07	Pure SLM	Robust	Robust across all 5 profiles
Code Review	3.40	Hybrid	Robust	Robust across all 5 profiles
Clinical Triage	3.20	LLM Only	Gate-locked	N/A: Hard Rule 1
Legal Contract	3.60	Hybrid	Robust	Robust across all 5 profiles (no gate rule applies)
Financial Report	4.00	LLM Only	Gate-locked	N/A: Hard Rule 2

Two use cases (UC6, UC8) are gate-locked and completely unaffected by any change in the weight vector. Five use cases (UC1, UC3, UC4, UC5, UC7) are robust, remaining stable across all five organizational weight profiles. UC2 is the sole sensitive case: under the Latency-First profile, its score shifts enough to move from Pure SLM to Hybrid. This is a conservative shift rather than a dangerous one, since it recommends more caution, not less, under that specific organizational weighting. The boundary-sensitivity flag in Section 4.4 is correctly triggered for UC2, directing the manager to run a pilot benchmark before committing infrastructure under a Latency-First profile.

Discussion and Managerial Implications

Key Findings

Three findings emerge from this study. First, S³ correctly predicts all eight deployment tiers before any inference is run, providing preliminary evidence that expert-elicited dimension scores are useful pre-deployment indicators of benchmark outcomes. Second, gate rules are necessary safety mechanisms, not conservative overhead: UC6 shows the gate catching a case the weighted formula alone would miss, while UC8 shows a gate rule independently reinforcing an assignment the formula also reaches. Third, tier assignments are robust: seven of eight are stable across five weight profiles, and the sole sensitive case is correctly identified and flagged.

Recognising Gate-Triggered Workloads Before Scoring

The most practically important skill for technology managers using S³ is identifying, before completing the scoring sheet, whether a workload is likely to trigger a gate rule. Three signals indicate this. First, any task where a wrong answer cannot be corrected after the fact, a clinical decision made, a legal document signed, a regulatory filing submitted, is a Hard Rule 1 candidate (SK = 5). Second, any task requiring the model to compose original analytical content in a context where errors carry legal or financial consequences is a Hard Rule 2 candidate (TC = 5, SK ≥ 4). Third, any task that touches a person's rights, finances, health, or legal standing, even if technically simple, is a Flag Rule candidate (SK = 4), meaning the minimum tier will be Hybrid regardless of the formula score.

Compliance officers should note that the gate rules directly align with the risk-management and human-oversight principles articulated in the EU AI Act and NIST AI RMF [11,12]. A workload that triggers Hard Rule 1 reflects the same underlying concern that motivates human-oversight requirements for high-risk applications under the EU AI Act, though S³'s risk categorization is a framework-internal judgment and does not constitute a determination of legal classification under that regulation.

Implications for Different Stakeholders

Technology managers should treat S³ as a pre-commitment checkpoint before any SLM pilot, following the six-step deployment protocol in Section 5.3. The scoring worksheet, dimension rationale, and gate rule outcomes create the auditable decision trail that and the EU AI Act require [11,12]. AI architects should pay particular attention to gate-locked cases: workloads where Hard Rule 1 or Hard Rule 2 fires cannot be made suitable for autonomous SLM deployment through architectural improvements or model upgrades; the safety constraint is in the task, not the model.

Limitations and Future Research

Limitations of the Current Study

Several specific limitations of this study must be acknowledged.

Threshold Calibration and Generalizability

Tier thresholds $\tau_1 = 3.0$ and $\tau_2 = 3.7$ were calibrated on an independent set of eight enterprise use cases and validated, without adjustment, against a separate set of eight held-out cases (Section 6.4), avoiding the circularity of deriving thresholds and validation results from the same data. The calibration set showed a non-monotonic relationship between S³ score and performance in the 3.0–3.7 range: one case retained only 64% parity while cases scoring both lower and higher retained strong parity. This did not affect the held-out validation result, but it suggests the current dimension set may not fully capture task-specific factors affecting mid-range difficulty cases. Riley et al. recommend a minimum of 200 outcome events for external validation of predictive models; the current study, with eight calibration and eight validation cases, remains well below this threshold and should be extended in future work [41].

Table 12: Calibration Set Summary: Eight Independent Use Cases Used to Derive Thresholds

Task	S³	Best SLM Accuracy	LLM Accuracy	Parity
Phishing Detection	2.67	76.7% (Qwen2.5-7B)	70.0%	109.5%
Receipt Extraction	2.33	73.3% (Qwen2.5-7B)	73.3%	100.0%
IT Ticket Classification	2.00	63.3% (Llama-3.1-8B)	53.3%	118.7%
Product Review Sentiment	2.33	90.0% (Llama-3.2-3B)	84.4%	106.6%
CVE-to-CWE Classification	3.20	30.0% (Llama-3.1-8B)	46.7%	64.3%
Symptom-to-Disease	3.67	23.3% (Phi4-Mini)	10.0%	233.3%
Contract Clause Classification	3.47	73.3% (Llama-3.1-8B)	47.8%	153.5%
Financial News Sentiment	2.27	93.3% (Gemma3-4B)	76.7%	121.7%

Number and Coverage of Dimensions

S³ evaluates workloads on six dimensions that have not been validated against alternative dimension sets. Domain-specific dimensions, such as regulatory jurisdiction for legal tasks or clinical specialty for healthcare tasks, may improve predictive accuracy in specialised settings.

Single-Team Scoring

All eight use cases were scored by members of the same research team. No external practitioners from other organizations scored the same workloads independently. The inter-rater reliability check is therefore internal rather than external.

Synthetic Gold Sets

All 30-item test sets were constructed by the research team. Items authored by business analytics researchers rather than domain specialists risk underrepresenting the hard-edge cases that differentiate SLM from LLM performance in practice.

Model Coverage

The benchmark includes six open-weight SLMs ranging from 3B to 8B parameters. Proprietary models (GPT-4o-mini, Gemini Flash, Claude Haiku) were not tested. Performance findings may not generalize to proprietary SLMs or future open-weight models.

Hardware Assumptions

All local inference was run on an Apple M2 Pro with 16 GB unified memory. Performance, latency, and cost figures may differ on Windows or Linux hardware, GPU-accelerated servers, or edge devices.

Fixed Temperature

All inference runs at temperature 0. Real enterprise deployments typically use non-zero temperature for generative tasks, introducing output variance not captured by the current benchmark.

Future Research Directions

Three directions are most pressing. First, external replication on a broader set of enterprise use cases, ideally contributed by practitioners across multiple organizations and industries, is required to establish external validity and refine the tier thresholds on truly held-out data. A minimum of 30 use cases across diverse industries would enable statistically meaningful threshold recalibration. Second, semi-automated scoring, using a calibrated LLM to assign initial dimension scores that human experts review and adjust, would reduce the cost of applying S³ at scale. Third, longitudinal validation would test whether tier assignments remain stable as SLM capability improves over model generations, or whether the thresholds require periodic recalibration.

Conclusion

This paper introduces S³, an auditable, pre-deployment multicriteria scoring protocol for enterprise SLM deployment decisions. S³ evaluates any enterprise workload across six anchored dimensions, applies three non-compensatory gate rules for safety-critical cases, and produces a normalized score that maps to one of three deployment tiers. All eight tier predictions, generated from independently-derived thresholds, are empirically confirmed across six industry verticals. Gate rules are shown to provide a necessary safety layer: in UC6, Hard Rule 1 changes the deployment decision from the weighted score's Hybrid recommendation to LLM Only, because an SLM that appeared to outperform the LLM on aggregate accuracy was in fact failing critically on urgent patient identification. UC8 further demonstrates that a gate can independently reinforce an LLM-Only assignment when both task complexity and stakeholder risk are high, even when the weighted score alone already exceeds the LLM-Only threshold.

The contribution is practical as well as theoretical. Technology managers, compliance officers, and AI architects now have a structured, documented, and empirically evaluated decision protocol for answering the question that determines whether SLM deployment is safe: not 'can a small model do this?', but 'is this workload appropriate for autonomous small model deployment given our organization's specific risk profile?' S³ answers that question before any production traffic is committed.

Acknowledgements

The authors thank Riddhima Ramasahayam Reddy, Rohee Savant, and Yashraj Saxena for their suggestions in the early stages of this paper.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

The data and code that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. Touvron H (2023) LLaMA: open and efficient foundation language models. arXiv preprint arXiv:2302.13971.
2. Jiang AQ (2023) Mistral 7B. arXiv preprint arXiv:2310.06825.
3. Belcak P, Heinrich G (2025) Small language models are the future of agentic AI. arXiv preprint arXiv:2506.02153.
4. Hinton G, Vinyals O, Dean J (2015) Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.
5. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L (2023) QLoRA: efficient finetuning of quantized LLMs. Advances in Neural Information Processing Systems.
6. Sanh V, (2019) DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.
7. Fishburn PC (1967) Additive utilities with incomplete product sets. Operations Research 15: 537-542.
8. Triantaphyllou E (2000) Multi-Criteria Decision-Making Methods: A Comparative Study. Springer, Boston, MA.
9. Rogers EM (2003) Diffusion of Innovations, 5th ed. Free Press, New York.
10. Venkatesh V (2003) User acceptance of information technology: toward a unified view. MIS Quarterly 27: 425-478.
11. NIST (2023) AI Risk Management Framework (AI RMF 1.0). NIST AI 100-1. National Institute of Standards and Technology.
12. European Parliament (2024) Regulation (EU) 2024/1689 on Artificial Intelligence (EU AI Act). Official Journal of the European Union.
13. Kaplan J (2020) Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.
14. Hoffmann J (2022) Training compute-optimal large language models. Advances in Neural Information Processing Systems 35.
15. Schick T, Schütze H (2021) It's not just size that matters: small language models are also few-shot learners. Proceedings of NAACL: 2339-2352.
16. Subramanian S, Elango V, Gungor M (2025) Small language models can still pack a punch. arXiv preprint arXiv:2501.05465.
17. Lu Y (2025) Demystifying small language models for edge deployment. Proceedings of ACL.
18. Ding W, Fan Y, Liang X, Chang K (2024) Hybrid LLM: cost-efficient and quality-aware query routing. Proceedings of ICLR.
19. Ong I (2025) RouteLLM: learning to route LLMs with preference data. Proceedings of ICLR.
20. Hu QJ (2024) RouterBench: a benchmark for multi-LLM routing systems. arXiv preprint arXiv:2403.12031.
21. Aggarwal P, Madaan A, Yang Y, Mausam A (2024) AutoMix: automatically mixing language models. Advances in Neural Information Processing Systems.
22. Dekoninck J, Baader M, Vechev M (2025) A unified approach to routing and cascading for LLMs. Proceedings of ICML.
23. Edwards W, Barron FH (1994) SMARTS and SMARTER: improved simple methods for multiattribute utility measurement. Organizational Behavior and Human Decision Processes 60: 306-325.
24. Zavadskas EK, Turskis Z (2011) Multiple criteria decision making (MCDM) methods in economics: an overview. Technological and Economic Development of Economy 17: 397-427.
25. Davis FD (1989) Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly 13: 319-340.
26. Tornatzky LG, Fleischer M (1990) The Processes of Technological Innovation. Lexington Books, Lexington, MA.
27. Floridi L (2018) AI4People: an ethical framework for a good AI society. Minds and Machines 28: 689-707.
28. Hevner AR, Maria Alexandra Cunha, et al. (2004) Design science in information systems research. MIS Quarterly 28: 75-105.
29. Peffers K, Tuure Tuunanen, Marcus A. Rothenberger, Samir Chatterjee, et al. (2007) A design science research methodology for information systems research. Journal of Management Information Systems 24: 45-77.
30. Keeney RL, Raiffa H (1993) Decisions with Multiple Objectives. Cambridge University Press, Cambridge.
31. Knaus WA, Bassirima Traoré, Souleymane Mariko, Samba Sidibé, Karamoko Kantako, et al. (1985) APACHE II: a severity of disease classification system. Critical Care Medicine 13: 818-829.
32. Lim WS, Van Der Eerden MM, Laing R, Boersma WG, Karalus N, et al. (2003) Defining community-acquired pneumonia severity on presentation to hospital (CURB-65). Thorax 58: 377-382.
33. ISO/IEC (2022) ISO/IEC 27001:2022 Information Security, Cybersecurity and Privacy Protection, Information Security Management Systems, Requirements. International Organization for Standardization, Geneva.
34. NIST (2020) Security and Privacy Controls for Information Systems and Organizations. NIST SP 800-53, Rev. 5. National Institute of Standards and Technology, Gaithersburg, MD.
35. Beyer B, Jones C, Petoff J, Murphy NR (eds) (2016) Site Reliability Engineering: How Google Runs Production Systems. O'Reilly Media, Sebastopol, CA.
36. Saaty TL (1980) The Analytic Hierarchy Process. McGraw-Hill, New York.
37. Triantaphyllou E, Sanchez A (1997) A sensitivity analysis approach for some deterministic multi-criteria decision-making methods. Decision Sciences 28: 151-194.
38. Ollama Contributors (2024) Ollama: Run Large Language Models Locally [online]. Available at: https://ollama.com.
39. Groq Inc (2024) GroqCloud API Documentation [online]. Available at: https://console.groq.com/docs.
40. Wilson EB (1927) Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association 22: 209-212.
41. Riley RD (2021) Minimum sample size for external validation of a clinical prediction model with a binary outcome. Statistics in Medicine 40: 4230-4251.

Copyright: ©2026 Ashim Bose, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.