NextGen Scientific Publication
American International Journal of Computer Science and Technology

Volume 7 Issue 6, Pg. No. 169- 184, AIJCST-V7I6P117, 2025
https://doi.org/10.63282/3117-5481/AIJCST-V7I6P117

Original Article

Designing a Multi-Criteria Decision Framework for Enterprise Language Model Optimization

***Karan Nisar**
Independent Researcher, USA.

Abstract:

Enterprises do not choose among prompt engineering, retrieval-augmented generation (RAG), fine-tuning, and hybrids in the abstract. They choose deployable combinations of model, context, behaviour control, orchestration, and serving under task-specific constraints. This paper presents E-LADF 2.0, a design-science decision artefact for configuration-level feasibility, value functions, and robustness. It begins with a registered acceptance test, diagnoses the locked baseline, freezes a causally plausible configuration shortlist, applies noncompensatory feasibility gates, and then combines anchored 0-to-100 value functions with analytic hierarchy process weights. A bounded uncertainty analysis perturbs criterion weights and performance values. A configuration is recommended only when its worst-case advantage exceeds a predeclared practical difference; otherwise the framework returns 'not separated' and prescribes a targeted pilot. Five constructed scenarios directly supply every partial value, criterion weight, and uncertainty input used in the arithmetic and, by design, exercise prompt, RAG, PEFT, hybrid, and abstention outcomes. They demonstrate execution, traceability, and edge-case handling, not empirical effectiveness. The contribution is a configuration-level procedure that separates constraints from preferences, replaces universal archetype scores with case-specific partial values, and treats uncertainty as a decision outcome.

Keywords:

Multi-Criteria Decision Analysis, Enterprise AI, Retrieval-Augmented Generation, Parameter-Efficient Fine-Tuning, Analytic Hierarchy Process, Robustness, AI Governance.

Article History:

Received: 05.11.2025

Revised: 28.11.2025

Accepted: 09.12.2025

Published: 20.12.2025

1. Introduction

Enterprise language-model adaptation is a systems decision. Retrieval can externalise changing knowledge and provide addressable evidence [1], while fine-tuning can alter stable task behaviour. Comparative studies show that the relative advantage depends on the type of knowledge, the task, the data, and the evaluation protocol rather than on a universal hierarchy [2], [3]. Retrieval-aware tuning further shows that context and weight adaptation can be combined instead of treated as mutually exclusive [4].

The practical choice has also expanded beyond the four labels used in early enterprise guidance. Context may be supplied as bounded long context, standard retrieval, or adaptive and self-reflective retrieval [5]-[9]. Behaviour can be shaped by prompt optimisation, constrained decoding, or parameter-efficient adaptation [10]-[19]. Routing, caching, compression, batching, and adapter multiplexing can change the cost and latency of the same apparent model choice [20]-[25]. A decision framed only as 'prompt versus RAG versus fine-tuning' therefore hides the configuration that must actually pass a security review, a load test, and a release gate.

Existing decision aids commonly mix three different questions: whether an option is allowed, whether it is likely to address the diagnosed failure, and whether stakeholders prefer its trade-offs. When a legal prohibition or hard latency ceiling is placed in the same weighted sum as quality or cost, a strong score elsewhere can compensate for a condition that is not tradeable. Fixed archetype scores also create false precision because the score for 'RAG' changes with the retriever, corpus, entitlement filter, reranker, model, serving region, and refresh process.

Copyright by the Author(s). This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International (CC BY-NC-ND 4.0) License (https://creativecommons.org/licenses/by-sa/4.0/)

This paper asks: RQ1, how can enterprises compare concrete language-model adaptation configurations in a repeatable and auditable manner? RQ2, how should hard constraints, stakeholder preferences, measurement uncertainty, and abstention be combined without double counting? RQ3, what information must be retained so that a recommendation can be reproduced and revisited?

The revised artefact makes four contributions:

- A configuration grammar that separates model, context, behaviour, orchestration, and serving levers.
- Configuration-specific pass, fail, or unresolved gates applied before any preference score.
- Case-specific 0-to-100 value functions and AHP weights with complete consistency and sensitivity reporting.
- A robustness and abstention rule that replaces a scale-dependent percentage margin with an absolute, predeclared practical difference.

2. Related Work and Decision Scope

2.1. Adaptation Mechanisms have become Composable

RAG combines parametric generation with external memory [1]. Later work demonstrates that retrieval can be conditional, iterative, or self-critical [5], [6], and that long context and retrieval are complementary design options whose performance varies by corpus and query [7]-[9]. Evidence comparing knowledge injection by retrieval and tuning cautions against treating weight modification as a simple substitute for current or obscure factual content [2], [3]. RAFT illustrates a hybrid recipe in which tuning teaches a model to use retrieved evidence and ignore distractors [4]. An enterprise-oriented synthesis further distinguishes prompting, retrieval, PEFT, and hybrid architectures by where task knowledge resides and which component is mutable at run time [26]. E-LADF extends that descriptive taxonomy to complete, versioned configurations.

The behaviour lever is equally varied. Automatic prompt optimisation and compiled prompting pipelines can improve an inference-only baseline [10]-[12]. Grammar-constrained generation can guarantee syntactic membership or enforce structured output without changing model weights [13]-[15]. LoRA and QLoRA reduce the resources needed to train task adapters [16], [17], but fine-tuning may still change previously learned behaviour and create a versioned model asset that must be tested, served, and rolled back [18], [19]. The framework therefore requires constrained generation to be tried before interpreting schema errors as evidence for tuning.

Cost and latency are configuration properties. Frugal routing, preference-trained routers, prompt-state reuse, prompt compression, efficient structured execution, and multi-adapter serving can materially change the operating frontier [20]-[25]. An analytical cost-of-ownership model likewise shows that fine-tuning break-even depends on request volume, input-token price, prompt-length reduction, and serving commitments, while retrieval is ordinarily justified by freshness, provenance, or accuracy rather than intrinsic savings [27]. E-LADF represents these choices on separate levers so that a serving optimisation is not mistaken for a new adaptation archetype.

2.2. Evaluation and Security Alter the Feasible Set

RAG evaluation is modular: retrieval, faithfulness, evidence coverage, citation entailment, and answer utility can fail independently [28]-[31]. Model-based judges can be useful but require bias-aware validation and should not silently replace the task owner's acceptance test [32], [33]. A held-out set, temporal separation, and a failure taxonomy are therefore part of the decision record rather than a post-selection evaluation plan.

External context creates an attack surface as well as an update mechanism. Knowledge-base poisoning and blocker documents can steer or jam retrieval, while indirect prompt injection can cross the boundary between retrieved content and model instructions [34]-[36]. The security gate consequently evaluates the exact ingestion, retrieval, tool, and privilege path. 'Uses RAG' is neither a security pass nor a security failure.

2.3. Governance and Operations

The NIST AI Risk Management Framework is voluntary and lifecycle-oriented, and its generative-AI profile and secure-development profile provide risk and control prompts rather than a universal architecture mandate [37]-[39]. The EU AI Act assigns obligations by role and system or model category [40]. The July 2025 general-purpose AI guidance and Code of Practice further clarify

provider-facing expectations [41], [42]. E-LADF does not infer a blanket citation obligation from these instruments. Evidence requirements are activated by the applicable role, use, law, professional rule, contract, or registered acceptance test.

Operational literature consistently treats production ML as a sociotechnical system with data, deployment, monitoring, lineage, ownership, change-management, and workflow-standardisation dependencies [43]-[48]. E-LADF converts those general lessons into configuration evidence: a data-flow record, asset inventory, load test, threat model, monitoring owner, rollback procedure, and re-entry triggers.

2.4. Scope

The unit of analysis is a versioned configuration for one task, deployment boundary, workload, and decision horizon. The framework does not rank foundation models in general, interpret legal obligations for the reader, or estimate causal deployment outcomes. It structures a decision after the organisation has identified the applicable requirements and produced evidence for the candidate configurations.

3. Research Method and Design Objectives

This study follows the design-science research tradition [49], [50]. Its contribution is an artefact, not an empirically validated prediction model. The process comprises problem identification, definition of design objectives, construction of E-LADF 2.0, demonstration through composite cases, and analytical evaluation of internal consistency, traceability, and boundary-condition handling. MCDA supplies the distinction between alternatives, criteria, value functions, and stakeholder weights [51], [52], while sensitivity analysis addresses the dependence of a rank on uncertain judgments [53].

Six design objectives govern the artefact:

- Begin with a measurable acceptance test rather than a preferred technology.
- Represent alternatives as implementable configurations, not broad category labels.
- Apply legal, contractual, safety, and operational requirements as noncompensatory feasibility constraints.
- Separate stakeholder preferences from measured or explicitly assumed configuration performance.
- Propagate uncertainty and decline to select when evidence does not distinguish alternatives.
- Make the decision revisable when models, prices, workload, sources, requirements, or organisational capabilities change.

4. The E-LADF 2.0 Procedure

E-LADF 2.0 contains five layers and a lifecycle loop (Figure 1). Layer 0 fixes the decision record and acceptance test. Layer 1 diagnoses the locked baseline and attributes its failures. Layer 2 generates and freezes a causally plausible configuration shortlist. Layer 3 applies configuration-specific feasibility gates. Layer 4 applies the anchored value model, AHP weights, uncertainty analysis, and stopping rule. This order prevents preferences from influencing candidate generation and prevents a value score from restoring a failed or unresolved configuration.

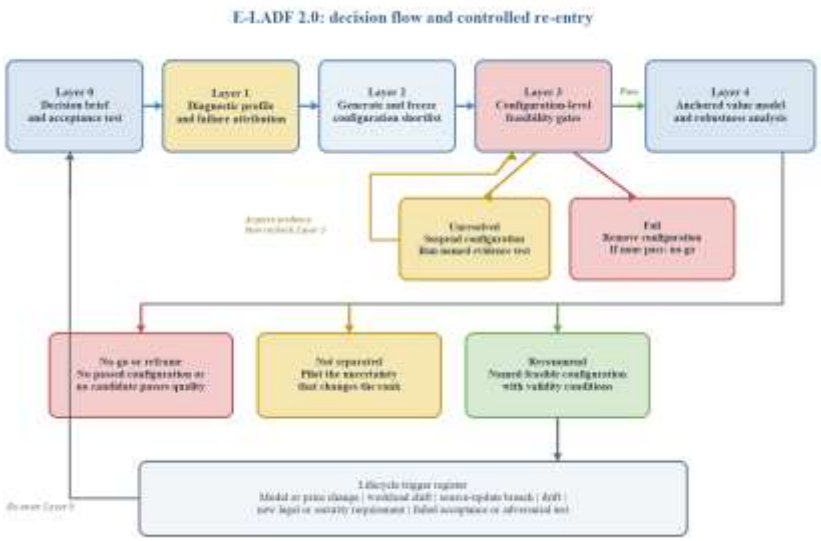


Figure 1. E-LADF 2.0 Decision Flow. Layer 3 Removes Failed Configurations and Suspends Unresolved Configurations Pending a Named Evidence-Acquisition Test. Only Configurations that Pass Every Activated Gate Enter Layer 4. Lifecycle Triggers Return the Decision to Layer 0

4.1. Configuration-Level Alternatives

A candidate configuration is represented as follows:

$$x = (M, C, B, O, S)$$

Here M is the base model and immutable version, C is the context mechanism, B is the behaviour-control mechanism, O is the orchestration mechanism, and S is the serving arrangement. Data sources, training sets, indexes, tools, deployment locations, retention settings, and update processes are attached to the record. A hybrid is any configuration that activates more than one adaptation lever, not a fifth indivisible archetype.

Table 1. Configuration Morphology and Minimum Versioned Record

Lever	Permitted values	Required record
Model (M)	Named model and immutable version	Provider, licence, context limit, release date, safety controls, evaluation date
Context (C)	C_0 no external context; C_1 bounded or long context; C_2 standard retrieval; C_3 adaptive or iterative retrieval	Corpus, chunking, index, entitlement, refresh, reranking, evidence linkage
Behaviour (B)	B_0 base; B_1 prompt or few-shot; B_2 constrained generation; B_3 PEFT; B_4 full tuning	Prompt, schema or grammar, training set, objective, adapter or checkpoint
Orchestration (O)	O_0 single call; O_1 deterministic workflow or tools; O_2 router; O_3 bounded agent loop	Tools, permissions, stopping rules, retry policy, trace
Serving (S)	Managed shared or private; dedicated tenancy; self-hosted; hybrid; declared optimisations	Region, hardware, retention, capacity, caching, batching, compression, adapter multiplexing, rollback



Figure 2. Four Orthogonal Configuration Levers. The Model Version is Fixed Separately, And Only Causally Plausible Combinations Enter the Shortlist

The full Cartesian product is not evaluated. The prompt-and-inference baseline remains in the set, after which one lever is added at a time when it has a plausible causal relationship to a registered failure. Missing or changing knowledge introduces a context-bearing candidate. Residual stable behavioural failure introduces PEFT if learning assets are adequate. Deterministic tools precede an agent loop when the required action can be specified. Serving optimisations enter when measured latency, throughput, or cost supplies the reason.

4.2. Layer 0: Decision Brief and Acceptance Test

Layer 0 creates the record before any candidate is scored. A candidate that misses a hard threshold is not deployable even if it later has the highest aggregate value. Evaluation data used to configure a prompt, retriever, or adapter cannot also serve as the held-out test set.

- Business objective, accountable decision owner, user population, operating context, and prohibited uses.
- Held-out evaluation set with inclusion, exclusion, temporal-split, adjudication, and leakage rules.
- Primary task metric plus hard quality, behavioural, safety, evidence, latency, capacity, and availability thresholds.
- Forecast workload, concurrency, input and output length, traffic shape, and decision horizon.
- Data and source inventory with sensitivity, ownership, permitted purposes, retention, residency, deletion, and update cadence.
- Locked inference-only baseline including permitted few-shot examples and constrained generation.
- Failure taxonomy separating missing knowledge, stale knowledge, retrieval, behaviour, schema, safety, orchestration, and serving failures.
- Predeclared value anchors, stakeholder roles, weighting hierarchy, practical-difference threshold, and uncertainty set.

4.3. Layers 1 and 2: Diagnostic Profile, Candidate Generation, and Shortlist Freeze

Diagnostics generate configurations but do not determine the winner and are not entered again as preference criteria. For each authoritative source s, the knowledge profile reports three anchored components: material-change frequency U_s, required propagation time L_s, and the consequence C_s of acting on superseded content. If a scalar is useful for triage, it may be recorded per source as $KVI_s = (U_s + L_s + C_s)/3$. The component vector remains primary, and unrelated sources are not averaged because a benign source could conceal a safety-critical one.

Behavioural specificity is measured only after the locked inference-only baseline receives all facts needed for the task and uses permitted prompt optimisation and constrained decoding. Blinded reviewers classify residual schema, style or register, and procedural failures. The reported behavioural-failure proportion includes sample size, confidence interval, reviewer agreement, and adjudication rule. A schema failure corrected by grammar-constrained generation is not evidence for tuning.

Table 2. Diagnostic Anchors Used to Generate, Not Rank, Candidate Configurations

Score	Material-change interval U	Required propagation time L	Staleness consequence C	Residual behavioural-failure rate
0	>365 days	>90 days	No material decision effect	≤2%
25	91-365 days	31-90 days	Minor and readily reversible	>2% to 5%
50	31-90 days	8-30 days	Material but reversible	>5% to 10%
75	2-30 days	1-7 days	Major financial, service, or professional consequence	>10% to 20%
100	≤1 day or continuous	<24 hours	Safety, legal, or difficult-to-reverse consequence	>20%

For demonstration, define $K_s = 1$ when $\max(U_s, L_s, C_s) \geq 50$. Define $B_s = 1$ when the behavioural-failure score is at least 50 or its 95% confidence interval crosses the 5% boundary. P is always generated, R is added when $K_s = 1$, F when $B_s = 1$, and H when both triggers are active. Feasibility gates are applied afterward. These midpoints are transparent design inputs, not empirically validated universal cutoffs.

Layer 2 converts each generated code into a complete versioned record x = (M, C, B, O, S), attaches the relevant data, index, adapter, region, retention, telemetry, and rollback records, and freezes the shortlist before weights or values are revealed. In the constructed demonstration, P, R, F, and H remain templates and all gate results are explicitly labelled assumptions. A field decision cannot gate an incomplete configuration record.

4.4. Layer 3: Configuration-Specific Feasibility Gates

Each gate returns pass, fail, or unresolved. Any fail excludes the configuration before weighting. An unresolved gate suspends it pending a named evidence-acquisition test. Unknown is never treated as pass. The gates are noncompensatory relative to preferences but need not be evaluated in a fixed order.

Table 3. Configuration-Level Noncompensatory Feasibility Gates

Gate	Noncompensable question	Minimum evidence
G1 Data purpose and boundary	Is every inference, embedding, logging, caching, evaluation, and parameter-update flow permitted and inside the required residency, retention, and deletion boundary?	Dataset-by-purpose rights matrix plus an end-to-end data-flow record
G2 Evidence and provenance	Can the exact configuration provide and retain the source, version, entitlement, and claim linkage required by the decision record?	Citation-coverage and citation-entailment test, if activated
G3 Update and erasure service level	Can a correction, withdrawal, deletion, or policy change stop affecting outputs within the registered interval?	Measured refresh, purge, rollback, or retraining procedure
G4 Assets and operational controls	Do the lawful corpus, training and test sets, versioning, monitoring, rollback, incident response, and accountable owners exist before release?	Dataset cards, leakage checks, control checklist, tested rollback, named owners
G5 Prototype hard thresholds	Does the complete deployed path meet task, safety, security, p95 latency, peak-throughput, and availability thresholds?	Configuration-specific evaluation, threat model, adversarial tests, and load test

G2 is activated only when the registered task has a source-evidence obligation. Plain prompting can pass if the request supplies an authoritative record and the system links claims to it. Fine-tuning can pass when tuning uses a separate lawful behavioural dataset and current knowledge remains addressable at runtime. Conversely, a nominally self-hosted RAG system can fail G1 if its embedding, telemetry, backup, or logging path crosses a prohibited boundary.

4.5. Layer 4a: Anchored Value Functions

Only configurations that pass all gates and hard acceptance thresholds enter the preference model. The fixed 1-to-5 archetype suitability table is replaced by measured or explicitly assumed case-specific performance transformed to a cardinal 0-to-100 partial value.

$$\text{Benefit: } v_i(z) = 0 \text{ if } z \le b_i;\ 100(z - b_i)/(a_i - b_i) \text{ if } b_i < z < a_i;\ 100 \text{ if } z \ge a_i$$
$$\text{Cost: } v_i(z) = 100 \text{ if } z \le a_i;\ 100(b_i - z)/(b_i - a_i) \text{ if } a_i < z < b_i;\ 0 \text{ if } z \ge b_i$$

For a benefit criterion, b_i is the worst feasible reference and a_i the aspiration. For a cost criterion, a_i is the aspiration and b_i the worst feasible reference. Nonlinear or categorical functions may be used when approved before results are revealed. Highly correlated criteria are combined or one is removed, and the additive model is used only after stakeholders accept preferential independence.

In a field application, each partial value must be derived from a raw measurement, declared anchors, and a preapproved transformation or complete categorical rubric. The constructed demonstration begins at the partial-value stage instead. Its 0-to-100 values are disclosed synthetic inputs used to demonstrate aggregation, sensitivity, and abstention; they are not transformations of observed organisational measurements and do not validate the measurement-to-value step.

Table 4. Reference Criterion Register for Field Applications. The Constructed Demonstration Supplies Partial Values Directly

ID	Criterion	0 Anchor	100 Anchor
C1	Task quality	Minimum deployable result	Predeclared stretch target
C2	Behavioural conformance	Minimum acceptable schema, style, taxonomy, or procedure compliance	Stretch conformance rate
C3	Change agility	Maximum permitted change-to-deployment interval	Preferred interval
C4	Traceability and failure diagnosability	Minimum evidence; failures cannot be localized	Material outputs are traceable and failures can be localized
C5	Governance and security controllability	Only mandatory gate controls are demonstrated	Full weighted control set is independently verifiable
C6	Latency headroom	p95 equals the hard ceiling	p95 meets or improves on the preferred target
C7	Annual operating cost	Approved maximum annual cost	Preferred cost target
C8	Time to production	Latest acceptable date	Preferred production date
C9	Ongoing operating burden	Maximum available annual operating capacity	Preferred operating-effort target

4.6. Layer 4b: AHP Weighting and Aggregate Value

AHP elicits criterion weights, not alternative performance. Stakeholders compare criteria within a predeclared hierarchy. For a positive reciprocal matrix A, the priority vector p is the normalised principal right eigenvector [51]:

$$Ap = \lambda_{max} p, \quad \sum_j p_j = 1$$
$$CI = (\lambda_{max} - n)/(n - 1), \quad CR = CI/RI_n$$

A field matrix with CR > 0.10 returns to stakeholders for review. When multiple stakeholders participate, their matrices are aggregated entrywise by weighted geometric mean before priority extraction, using role weights fixed before elicitation:

$$a^{G}_{ij} = \prod_h (a^{(h)}_{ij})^{(q_h)}, \quad \sum_h q_h = 1$$

The constructed demonstration does not simulate elicitation. It begins with declared synthetic criterion weights and defines unrounded reciprocal matrices by $a_{ij} = w_i/w_j$. These matrices are exactly consistent by construction, so $Aw = 9w$, $\lambda_{max} = 9$, and CR = 0. This identity demonstrates reproducibility only; it is not evidence of stakeholder agreement or judgment quality.

$$V(x) = \sum_i w_i v_i(z_{ix}), \qquad \sum_i w_i = 1$$

Table 5. Constructed Scenario Weights. The Corresponding Unrounded Direct-Ratio Matrices are Exactly Consistent

Scenario	C1	C2	C3	C4	C5	C6	C7	C8	C9	CR
Manufacturing	.20	.10	.05	.05	.10	.13	.13	.13	.11	.000
Banking	.21	.07	.17	.19	.15	.08	.05	.04	.04	.000
Healthcare	.25	.23	.04	.04	.10	.14	.13	.03	.04	.000
Legal	.25	.19	.13	.16	.10	.05	.03	.03	.06	.000
Retail	.24	.15	.15	.07	.08	.08	.08	.07	.08	.000

C1-C9 are defined in Table 4. Each row sums to 1.00. A rounded display of the Manufacturing reciprocal matrix appears in Appendix A; all matrices use unrounded $a_{ij} = w_i/w_j$ ratios.

4.7. Layer 4c: Uncertainty, Robustness, and Abstention

The smallest practically meaningful difference ε is declared in absolute points on the anchored value scale before results are inspected. The demonstration uses ε = 3.0. Each weight varies by a multiplier m_i in [0.8, 1.2] and is renormalised:

$$w_i(m) = w_i^0 m_i / \sum_j w_j^0 m_j, \qquad m_i \in [0.8, 1.2]$$

Each criterion value has a symmetric scenario-specific uncertainty half-width u_s and remains within [0, 100]. For nominal leader t and challenger b, the conservative advantage is:

$$D^-_{t,\beta} = \min_m \sum_i w_i(m)[\max(0, v_{ti} - u_s) - \min(100, v_{\beta i} + u_s)]$$

The clipping operators preserve the declared value range. Because the clipped pairwise criterion differences are fixed for a given scenario, the objective remains linear-fractional in the weight multipliers. Evaluation of all $2^9 = 512$ corners therefore bounds the continuous multiplier box.

The leader is recommended only if $D^-_{t,\beta} > \varepsilon$ for every feasible challenger. Otherwise the output is 'not separated,' and the pilot varies the attribute responsible for the rank reversal. A percentage of the leading score is not used because it depends on the arbitrary scale origin.

Table 6. Defined Decision States and Required Outputs

Condition	Required output
Baseline passes and no alternative is robustly better by more than ε	Retain baseline under the Progressive Adaptation Principle
One configuration clears every gate and robustly dominates	Recommend the named configuration with validity conditions
Two or more are not robustly separated	Pilot the configurations that differ on the rank-changing uncertainty
Only one is feasible	Report 'sole demonstrated feasible configuration'; deploy only if it passes acceptance
A gate is unresolved	Acquire the specified evidence before scoring
No candidate is feasible or feasible candidates fail acceptance	No-go or change the task, requirements, deployment boundary, data, or candidate set
All candidates are practically equivalent	Report an equivalence set; any lower-commitment tie-break is a policy choice, not an MCDA result

4.8. Lifecycle Loop and Progressive Adaptation Principle

Every recommendation records an expiry condition and trigger register. Re-entry occurs when volume crosses a switch point, model or token price changes, a base model changes, an update or erasure service level is missed, the task distribution shifts, a new legal or security requirement applies, an adversarial test exposes a failure, or a diagnostic component crosses a registered threshold.

> **Progressive Adaptation Principle.** Begin with the lowest-commitment configuration that could plausibly satisfy the acceptance test. Escalate only after a measured, reproducible failure has been assigned to a specific lever. Re-evaluate the complete configuration after each escalation because context, behaviour, orchestration, and serving choices interact in operation.

5. Analytical Demonstration

The framework is demonstrated on five constructed enterprise scenarios. These synthetic records are designed to expose the inputs, calculations, gate behaviour, and stopping rule. They are not observations of deployed systems, do not estimate industry prevalence, and do not establish comparative effectiveness or empirical validity. All assumed techniques were available by the 31 October 2025 cutoff.

5.1. Common Protocol and Candidate Configurations

Each scenario fixes the acceptance test and workload, locks and diagnoses the baseline, generates and freezes candidates, applies G1-G5, assigns partial values only to configurations that pass every activated gate, and then executes the value and robustness calculations. The four classes in Table 7 are configuration templates rather than products. For each synthetic scenario, the model and serving identifiers are held fixed, and Table 10 reports constructed gate assumptions. A field application would replace every code with a complete vendor, model version, region, data-flow, corpus, index, adapter, and serving record.

Table 7. Configuration Templates Used in the Constructed Demonstration

Code	Constructed configuration template
P	Versioned prompt, few-shot examples, schema or grammar-constrained generation; task records may be supplied directly; no managed retrieval or parameter update
R	P plus access-controlled lexical and vector retrieval, reranking, source identifiers, claim-level citations where required, and incremental index refresh
F	P plus a separately versioned PEFT adapter trained on curated task examples; no runtime retrieval
H	R plus a PEFT adapter trained with retrieved contexts, including negative or distractor contexts; index and adapter remain independently replaceable

Table 8. Constructed Scenario Conditions and Hard Acceptance Thresholds

Scenario	Constructed conditions
Manufacturing shift handover	25,000 requests/month; p95 ≤2.0 s; 99.5% schema validity; no external-source citation obligation. PEFT G4 threshold: 2,000 training and 500 temporally separated tuning-validation examples; 150 and 30 available. u_s = 2.0.
Banking policy QA	300,000 requests/month; p95 ≤3.5 s; policy propagation ≤4 h; source and citation entailment ≥99% for every material answer. PEFT G4 threshold: 5,000 training and 1,000 tuning-validation examples; 8,000 and 1,200 available. u_s = 1.0.
Healthcare registry abstraction	1.2 million abstractions/year; private-tenancy PHI processing; p95 ≤1.0 s; macro-F1 ≥96%; schema validity ≥99.8%. PEFT G4 threshold: 20,000 training and 5,000 tuning-validation examples; 100,000 and 20,000 available. u_s = 1.5.
Legal contract review	60,000 contracts/year; p95 ≤8 s; issue recall ≥93%; section-level evidence; playbook changes ≤24 h. PEFT G4 threshold: 10,000 training and 2,000 tuning-validation examples; 30,000 and 5,000

	available. u_s = 1.0.
Retail product support	2 million interactions/month; p95 ≤2.5 s; task success ≥91%; policy and tone conformance ≥98%; daily catalogue updates. PEFT G4 threshold: 10,000 training and 2,000 tuning-validation examples; 24,000 and 6,000 available. u_s = 1.0.

PEFT counts are predeclared constructed inputs for lawful, representative, versioned, and temporally separated training and tuning-validation sets. They exclude the held-out acceptance-test set and are not universal sample-size minima. u_s is the partial-value uncertainty half-width.

Table 9. Constructed Diagnostic Profiles and Generated Candidate Sets

Scenario	U	L	C	KVI	Residual Behavioural Failure after Locked P Baseline	B Score	Generated Set
Manufacturing	25 180 d	50 14 d	50 material, reversible	41.7	29/600 = 4.8% 95% CI 3.4-6.9%	25*	P, R, F, H
Banking	75 14 d	100 4 h	75 major service	83.3	80/1,000 = 8.0% 95% CI 6.5-9.8%	50	P, R, F, H
Healthcare	25 180 d	25 60 d	25 minor, reversible	25.0	140/1,000 = 14.0% 95% CI 12.0-16.3%	75	P, F
Legal	75 14 d	100 12 h	75 major professional	83.3	136/800 = 17.0% 95% CI 14.6-19.8%	75	P, R, F, H
Retail	100 1 d	100 12 h	50 material, reversible	83.3	156/1,200 = 13.0% 95% CI 11.2-15.0%	75	P, R, F, H

*KVI = (U + L + C)/3. P is always generated; R follows a knowledge trigger, F a behaviour trigger, and H both. *Manufacturing scores 25 by its point estimate, but the 95% interval crosses the 5% boundary, so PEFT-bearing candidates are retained. Gates are applied only after generation.*

Table 10. Constructed Feasibility Assumptions. Failed or Non-Generated Configurations Receive No Numerical Score

Scenario	P	R	F	H	Reason for exclusion
Manufacturing	Pass	Pass	Fail G4	Fail G4	150 training and 30 tuning-validation examples are below the predeclared 2,000 and 500 PEFT thresholds
Banking	Fail G2/G3	Pass	Fail G2/G3	Pass	P and F cannot meet clause-level evidence or four-hour propagation
Healthcare	Pass	Not generated	Pass	Not generated	The diagnostic profile activates behaviour adaptation but not external context
Legal	Fail G2	Pass	Fail G2/G3	Pass	P and F cannot meet section evidence; F also cannot update within 24 hours
Retail	Pass	Pass	Pass	Pass	All configurations satisfy the constructed gates

5.2. Complete Criterion Values and Results

Table 11 reports every synthetic partial value used in the demonstration. These values are direct assumptions on the anchored 0-to-100 scale, not measurements transformed from operating data. Consequently, the nominal totals and robustness calculations are

reproducible, but the demonstration does not evaluate or validate the transformation from raw evidence to partial value. Table 12 reports the nominal and worst-case pairwise gaps.

Table 11. Complete Partial-Value Matrix and Reproducible Weighted Totals for Feasible Configurations

Scenario	Cfg	C1	C2	C3	C4	C5	C6	C7	C8	C9	V
Manufacturing	P	86	92	90	70	88	92	95	95	94	90.20
Manufacturing	R	84	92	88	85	78	70	72	68	65	76.90
Banking	R	91	85	96	96	92	72	72	72	70	88.46
Banking	H	94	95	91	94	78	58	48	38	40	81.58
Healthcare	P	73	76	92	70	88	88	80	92	90	80.09
Healthcare	F	95	98	68	60	80	94	96	60	74	89.81
Legal	R	80	65	95	96	90	66	62	60	65	79.92
Legal	H	99	99	93	97	85	58	52	42	48	88.27
Retail	P	65	70	60	50	84	90	88	92	90	73.20
Retail	R	89	84	96	90	88	68	70	72	68	83.22
Retail	F	78	92	50	40	74	88	86	58	65	71.92
Retail	H	97	98	94	92	80	66	66	55	60	84.13

V is the weighted total using Table 5. Totals are rounded to two decimals. No value is shown for a configuration that failed or was not generated in Table 10.

Table 12. Outcome of the Bounded Robustness Analysis

Scenario	Feasible set	Leader	Runner-up	Gap	Robust gap	Decision
Manufacturing	P, R	P	R	13.30	6.94	Prompt + constraints
Banking	R, H	R	H	6.88	3.17	RAG
Healthcare	P, F	F	P	9.72	3.91	PEFT
Legal	R, H	H	R	8.35	3.41	Hybrid
Retail	P, R, F, H	H	R	0.91	-2.56	Not separated: pilot R/H

ε = 3.0. Score uncertainty half-width u_s is 2.0 for Manufacturing, 1.0 for Banking, 1.5 for Healthcare, 1.0 for Legal, and 1.0 for Retail.

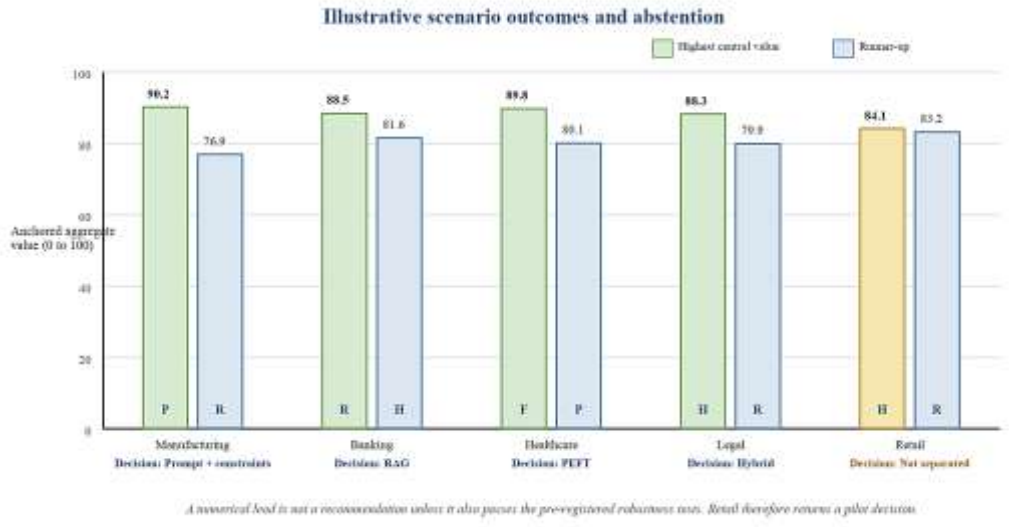


Figure 3. Nominal Leader and Runner-Up Values. Retail's Numerical Leader is not recommended because its Robust Gap is below E

5.3. Scenario Interpretations

- Manufacturing: prompt and constrained generation: The task is stable, structured, and supplied with the shift record in the request. The diagnostic uncertainty generates all four templates. F and H then fail G4 because 150 training and 30 tuning-validation examples miss the predeclared 2,000 and 500 PEFT thresholds. P and R pass. P scores 90.20 against R at 76.90. With u_s = 2.0, its minimum combined-uncertainty advantage is 6.94, so P clears ε. The validity condition is stable task behaviour; a sustained terminology or schema failure reopens B3.
- Banking: entitlement-aware RAG: P and F fail G2 and G3 because they cannot demonstrate complete clause-level evidence or guaranteed four-hour policy propagation. R and H remain feasible. R leads 88.46 to 81.58, and its worst-case advantage is 3.17 with u_s = 1.0. The recommendation is R, narrowly above ε. A larger uncertainty range or a material increase in residual behavioural error would change the output to a pilot rather than silently preserve the rank.
- Healthcare: PEFT for stable registry abstraction: This is structured abstraction from the note supplied in the request, not guideline question answering. Current external knowledge and claim citations are not part of the acceptance test, while latency, macro-F1, schema validity, and annual volume are decisive. The diagnostic profile therefore generates P and F; both pass. F scores 89.81 against P at 80.09. With u_s = 1.5, its robust advantage is 3.91. F is recommended with adapter lineage, memorisation tests, general-capability regression, and rollback retained under G4 and G5.
- Legal: retrieval plus PEFT: Contract review combines a changing playbook with stable firm-specific issue taxonomy and fallback behaviour. P fails the evidence gate; F fails both the evidence and update gates. R and H pass because knowledge remains in an addressable playbook, and H tunes only lawful behavioural examples. H scores 88.27 against R at 79.92. Its robust advantage is 3.41 with u_s = 1.0, so the hybrid is recommended. The index and adapter remain independently replaceable, and playbook content is excluded from the adapter unless separately authorised.
- Retail: not separated: All four configurations pass. H leads R by only 0.91 nominal points. Under the [0.8, 1.2] weight box, the H-minus-R difference changes sign; after u_s = 1.0, its conservative advantage is -2.56. Both remain ahead of P and F, but neither clears ε. The output is 'not separated: pilot R against H.' The pilot holds the base model, retrieval stack, traffic, and evaluation set constant while varying only the adapter. If practical equivalence remains, choosing R is an explicit lower-commitment policy tie-break rather than an MCDA finding.

The demonstration exercises prompt plus constrained generation, RAG, PEFT, hybrid, and abstention outcomes by construction. This is a coverage test of artefact behaviour. It does not show that the framework predicts deployment success or that these configuration classes occur with equal frequency.

6. Discussion

6.1. The Decision Object Changes the Conclusion

The central correction is to evaluate concrete configurations. 'Fine-tuning is prohibited' may be false when the restricted corpus is only retrieved and the adapter learns from a separate lawful behavioural set. 'Self-hosted RAG is allowed' may also be false when embedding telemetry or backups cross a boundary. The configuration record exposes the relevant data use, location, control, and update path.

Separating levers also improves diagnosis. Constrained decoding can solve syntax without changing weights [13]-[15]. Adaptive retrieval can avoid indiscriminate retrieval for simple queries [5], [6]. Prompt caching, compression, and structured runtimes can move a candidate inside a latency or cost boundary [22], [24], [25]. These are not after-the-fact implementation details because they can change feasibility and value.

6.2. Constraints, Diagnostics, and Preferences Must not be Counted Twice

A hard legal or safety requirement appears in a gate. Its residual quality within the feasible range may appear in a value function, but the missing mandatory control is never assigned a low compensable score. Knowledge and behavioural diagnostics generate candidates, then leave the preference model. Generic organisational maturity is not both a gate and a criterion; G4 asks whether the actual controls required by the configuration exist, and C9 measures residual operating burden after they do.

6.3. Abstention is a Substantive Result

A point-score model always produces a first place, even when the difference is smaller than measurement and preference uncertainty. E-LADF's absolute ε and bounded sensitivity test converts that uncertainty into an action. The Retail pilot is specific

because the configurations differ only in the adapter. A pilot that changes the model, retriever, traffic, and prompt simultaneously would not resolve the rank-changing uncertainty.

6.4. Governance Implications as of October 2025

The governance gate is deliberately role- and use-specific. NIST guidance helps structure risk management but is voluntary [37]-[39]. EU obligations depend on whether the organisation is a provider, deployer, importer, distributor, or another actor and on the relevant system or model category [40]-[42]. The framework therefore requires the organisation to register the applicable obligation and the evidence that tests it. It does not convert broad regulatory concern into an automatic RAG preference.

Security evidence must also follow the architecture. A retrievable corpus supports correction and provenance but introduces poisoning, jamming, and instruction-channel risks [34]-[36]. A tuning pipeline avoids runtime retrieval but introduces training-data rights, leakage, memorisation, versioning, and rollback concerns. A hybrid inherits both control surfaces and cannot be scored as an average of its components.

7. Limitations and Research Agenda

E-LADF 2.0 remains a design-science artefact demonstrated analytically. The five cases are intentionally contrasting synthetic records. They cannot estimate how often a configuration would be selected, whether the recommendations outperform expert judgement, or whether the procedure improves realised quality, cost, latency, incident, or governance outcomes.

The diagnostic anchors and candidate-generation thresholds are operational definitions, not validated measurement instruments. Inter-rater reliability, convergent validity, sensitivity to task sampling, and predictive association with successful configurations remain open. The value model depends on criterion choice, anchors, stakeholder judgments, and preferential independence. AHP consistency detects some contradictions but does not establish that a judgment is correct.

The uncertainty model is transparent but stylised. A 20% independent multiplier box and symmetric value half-widths are suitable for demonstrating mechanics, not for every decision. Field work should use bootstrap intervals for held-out metrics, repeated load-test distributions, workload and price scenarios, stakeholder-specific weight intervals, and documented dependence among criteria. Structural uncertainty from an incomplete candidate set remains outside any numeric interval.

The morphology is broader than a four-archetype taxonomy but remains bounded. New retrieval, test-time computation, routing, tool-use, and serving mechanisms may require new lever values. Every lifecycle re-entry therefore includes a candidate-set completeness review rather than forcing a new method into an old category.

A suitable empirical programme would apply the artefact retrospectively to documented decisions, compare it with unaided expert recommendations, use independent raters for gates and diagnostics, preregister prospective recommendations, and track realised outcomes over time. Until then, E-LADF should be presented as a transparent decision procedure and research proposition, not a validated universal selection rule.

8. Conclusion

Enterprise language-model adaptation is not a contest among four labels. It is a choice among versioned configurations that place knowledge, behaviour, control, and operational commitment in different parts of a system. E-LADF 2.0 makes that choice inspectable: define acceptance first, diagnose the locked baseline, freeze a causally plausible shortlist, eliminate failed or unresolved configurations, transform case evidence through anchored values, elicit consistent stakeholder weights, and recommend only when the result survives declared uncertainty.

The constructed demonstration shows the procedure returning prompt, RAG, PEFT, hybrid, and an abstention without scoring excluded alternatives or claiming empirical superiority. Its governing principle is progressive adaptation: make the smallest commitment that can satisfy the acceptance test, and escalate only on measured failure attributed to a specific lever.

References

[1] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Advances in Neural Information Processing Systems 33, 2020, pp. 9459-9474. Proceedings
[2] O. Ovadia, M. Brief, M. Mishaeli, and O. Elisha, "Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs," in Proc. 2024 Conf. Empirical Methods in Natural Language Processing, 2024, pp. 237-250. doi: 10.18653/v1/2024.emnlp-main.15
[3] H. Soudani, E. Kanoulas, and F. Hasibi, "Fine Tuning vs. Retrieval Augmented Generation for Less Popular Knowledge," in Proc. 2024 ACM SIGIR Conf. Research and Development in Information Retrieval in the Asia Pacific Region, 2024, pp. 12-22. doi: 10.1145/3673791.3698415
[4] T. Zhang, S. G. Patil, N. Jain, S. Shen, M. Zaharia, I. Stoica, and J. E. Gonzalez, "RAFT: Adapting Language Model to Domain Specific RAG," in Proc. 1st Conf. Language Modeling, 2024. OpenReview
[5] S. Jeong, J. Baek, S. Cho, S. J. Hwang, and J. Park, "Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity," in Proc. 2024 NAACL-HLT, vol. 1, 2024, pp. 7036-7050. doi: 10.18653/v1/2024.naacl-long.389
[6] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection," in Proc. 12th Int. Conf. Learning Representations, 2024. OpenReview
[7] Z. Li, C. Li, M. Zhang, Q. Mei, and M. Bendersky, "Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach," in Proc. 2024 EMNLP Industry Track, 2024, pp. 881-893. doi: 10.18653/v1/2024.emnlp-industry.66
[8] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the Middle: How Language Models Use Long Contexts," Transactions of the Association for Computational Linguistics, vol. 12, pp. 157-173, 2024. doi: 10.1162/tacl_a_00638
[9] Y. Bai et al., "LongBench v2: Towards Deeper Understanding and Reasoning on Realistic Long-Context Multitasks," in Proc. 63rd Annual Meeting of the Association for Computational Linguistics, vol. 1, 2025, pp. 3639-3664. doi: 10.18653/v1/2025.acl-long.183
[10] R. Pryzant, D. Iter, J. Li, Y. T. Lee, C. Zhu, and M. Zeng, "Automatic Prompt Optimization with 'Gradient Descent' and Beam Search," in Proc. 2023 Conf. Empirical Methods in Natural Language Processing, 2023, pp. 7957-7968. doi: 10.18653/v1/2023.emnlp-main.494
[11] C. Yang, X. Wang, Y. Lu, H. Liu, Q. V. Le, D. Zhou, and X. Chen, "Large Language Models as Optimizers," in Proc. 12th Int. Conf. Learning Representations, 2024. OpenReview
[12] O. Khattab et al., "DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines," in Proc. 12th Int. Conf. Learning Representations, 2024. OpenReview
[13] S. Geng, M. Josifoski, M. Peyrard, and R. West, "Grammar-Constrained Decoding for Structured NLP Tasks without Finetuning," in Proc. 2023 Conf. Empirical Methods in Natural Language Processing, 2023, pp. 10932-10952. doi: 10.18653/v1/2023.emnlp-main.674
[14] L. Beurer-Kellner, M. Fischer, and M. Vechev, "Guiding LLMs the Right Way: Fast, Non-Invasive Constrained Generation," in Proc. 41st Int. Conf. Machine Learning, PMLR, vol. 235, 2024, pp. 3658-3673. PMLR
[15] D. Banerjee, T. Suresh, S. Ugare, S. Misailovic, and G. Singh, "CRANE: Reasoning with Constrained LLM Generation," in Proc. 42nd Int. Conf. Machine Learning, PMLR, vol. 267, 2025, pp. 2836-2857. PMLR
[16] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in Proc. 10th Int. Conf. Learning Representations, 2022. OpenReview
[17] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," in Advances in Neural Information Processing Systems 36, 2023. doi: 10.52202/075280-0441
[18] X. Zhang and J. Wu, "Dissecting Learning and Forgetting in Language Model Finetuning," in Proc. 12th Int. Conf. Learning Representations, 2024. OpenReview
[19] D. Biderman et al., "LoRA Learns Less and Forgets Less," Transactions on Machine Learning Research, 2024. OpenReview
[20] L. Chen, M. Zaharia, and J. Zou, "FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance," Transactions on Machine Learning Research, 2024. OpenReview
[21] I. Ong, A. Almahairi, V. Wu, W.-L. Chiang, T. Wu, J. E. Gonzalez, M. W. Kadous, and I. Stoica, "RouteLLM: Learning to Route LLMs with Preference Data," in Proc. 13th Int. Conf. Learning Representations, 2025. OpenReview
[22] I. Gim, G. Chen, S.-S. Lee, N. Sarda, A. Khandelwal, and L. Zhong, "Prompt Cache: Modular Attention Reuse for Low-Latency Inference," Proceedings of Machine Learning and Systems, vol. 6, 2024. MLSys
[23] Y. Sheng et al., "S-LoRA: Serving Thousands of Concurrent LoRA Adapters," Proceedings of Machine Learning and Systems, vol. 6, 2024. MLSys
[24] Z. Pan et al., "LLMLingua-2: Data Distillation for Efficient and Faithful Task-Agnostic Prompt Compression," in Findings of the Association for Computational Linguistics: ACL 2024, 2024, pp. 963-981. doi: 10.18653/v1/2024.findings-acl.57
[25] L. Zheng et al., "SGLang: Efficient Execution of Structured Language Model Programs," in Advances in Neural Information Processing Systems 37, 2024. doi: 10.52202/079017-2000
[26] K. Nisar, "Prompting, Retrieval, and Fine-Tuning: Foundations of Enterprise Language Model Adaptation," International Journal of Engineering & Extended Technologies Research, vol. 6, no. 6, pp. 9310-9319, 2024. doi: 10.15662/IJEETR.2024.0606030
[27] K. Nisar, "Quantifying the Cost and Risk of Enterprise LLM Adaptation Strategies," International Journal of Research Publications in Engineering, Technology and Management, vol. 8, no. 3, pp. 12162-12169, 2025. Journal article
[28] S. Es, J. James, L. Espinosa Anke, and S. Schockaert, "RAGAs: Automated Evaluation of Retrieval Augmented Generation," in Proc. 18th Conf. European Chapter of the Association for Computational Linguistics: System Demonstrations, 2024, pp. 150-158. doi: 10.18653/v1/2024.eacl-demo.16

[29] J. Saad-Falcon, O. Khattab, C. Potts, and M. Zaharia, "ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems," in Proc. 2024 NAACL-HLT, vol. 1, 2024, pp. 338-354. doi: 10.18653/v1/2024.naacl-long.20
[30] J. Chen, H. Lin, X. Han, and L. Sun, "Benchmarking Large Language Models in Retrieval-Augmented Generation," Proc. AAAI Conf. Artificial Intelligence, vol. 38, no. 16, pp. 17754-17762, 2024. doi: 10.1609/aaai.v38i16.29728
[31] C. Niu, Y. Wu, J. Zhu, S. Xu, K. Shum, R. Zhong, J. Song, and T. Zhang, "RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models," in Proc. 62nd Annual Meeting of the Association for Computational Linguistics, vol. 1, 2024, pp. 10862-10878. doi: 10.18653/v1/2024.acl-long.585
[32] L. Zheng et al., "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," in Advances in Neural Information Processing Systems 36, 2023, pp. 46595-46623. Proceedings
[33] P. Liang et al., "Holistic Evaluation of Language Models," Transactions on Machine Learning Research, 2023. OpenReview
[34] W. Zou, R. Geng, B. Wang, and J. Jia, "PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models," in Proc. 34th USENIX Security Symp., 2025, pp. 3827-3844. USENIX
[35] A. Shafran, R. Schuster, and V. Shmatikov, "Machine Against the RAG: Jamming Retrieval-Augmented Generation with Blocker Documents," in Proc. 34th USENIX Security Symp., 2025, pp. 3787-3806. USENIX
[36] S. Abdelnabi, K. Greshake, S. Mishra, C. Endres, T. Holz, and M. Fritz, "Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection," in Proc. 16th ACM Workshop on Artificial Intelligence and Security, 2023, pp. 79-90. doi: 10.1145/3605764.3623985
[37] E. Tabassi, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1. Gaithersburg, MD, USA: National Institute of Standards and Technology, 2023. doi: 10.6028/NIST.AI.100-1
[38] C. Autio et al., Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1. Gaithersburg, MD, USA: National Institute of Standards and Technology, 2024. doi: 10.6028/NIST.AI.600-1
[39] H. Booth, M. Souppaya, A. Vassilev, M. Ogata, M. Stanley, and K. Scarfone, Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: An SSDF Community Profile, NIST SP 800-218A. Gaithersburg, MD, USA: National Institute of Standards and Technology, 2024. doi: 10.6028/NIST.SP.800-218A
[40] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence," Official Journal of the European Union, L 2024/1689, 12 Jul. 2024. Official text
[41] European Commission, "Guidelines on the Scope of the Obligations for General-Purpose AI Models Established by Regulation (EU) 2024/1689 (Artificial Intelligence Act)," draft approved by the Commission, Brussels, Belgium, 18 Jul. 2025. Official July 2025 document
[42] European Commission, "General-Purpose AI Code of Practice," 10 Jul. 2025. Official code
[43] S. Amershi et al., "Software Engineering for Machine Learning: A Case Study," in Proc. 41st IEEE/ACM Int. Conf. Software Engineering: Software Engineering in Practice, 2019, pp. 291-300. doi: 10.1109/ICSE-SEIP.2019.00042
[44] D. Kreuzberger, N. Kühl, and S. Hirschl, "Machine Learning Operations (MLOps): Overview, Definition, and Architecture," IEEE Access, vol. 11, pp. 31866-31879, 2023. doi: 10.1109/ACCESS.2023.3262138
[45] M. M. John, H. H. Olsson, and J. Bosch, "An Empirical Guide to MLOps Adoption: Framework, Maturity Model and Taxonomy," Information and Software Technology, vol. 183, art. 107725, 2025. doi: 10.1016/j.infsof.2025.107725
[46] M. Zarour, H. Alzabut, and K. T. Al-Sarayreh, "MLOps Best Practices, Challenges and Maturity Models: A Systematic Literature Review," Information and Software Technology, vol. 183, art. 107733, 2025. doi: 10.1016/j.infsof.2025.107733
[47] A. Paleyes, R.-G. Urma, and N. D. Lawrence, "Challenges in Deploying Machine Learning: A Survey of Case Studies," ACM Computing Surveys, vol. 55, no. 6, art. 114, pp. 1-29, 2023. doi: 10.1145/3533378
[48] K. Nisar, "Bridging Prototyping and Production: A Systems-Level Study of ML Workflow Standardization in Enterprise AI Adoption," International Journal of Computational and Experimental Science and Engineering, vol. 11, no. 4, pp. 4741-4757, 2025. doi: 10.22399/ijcesen.5364
[49] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design Science in Information Systems Research," MIS Quarterly, vol. 28, no. 1, pp. 75-105, 2004. doi: 10.2307/25148625
[50] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A Design Science Research Methodology for Information Systems Research," Journal of Management Information Systems, vol. 24, no. 3, pp. 45-77, 2007. doi: 10.2753/MIS0742-1222240302
[51] T. L. Saaty, "How to Make a Decision: The Analytic Hierarchy Process," European Journal of Operational Research, vol. 48, no. 1, pp. 9-26, 1990. doi: 10.1016/0377-2217(90)90057-I
[52] V. Belton and T. J. Stewart, Multiple Criteria Decision Analysis: An Integrated Approach. Boston, MA, USA: Kluwer Academic Publishers, 2002. doi: 10.1007/978-1-4615-1495-4
[53] E. Triantaphyllou and A. Sánchez, "A Sensitivity Analysis Approach for Some Deterministic Multi-Criteria Decision-Making Methods," Decision Sciences, vol. 28, no. 1, pp. 151-194, 1997. doi: 10.1111/j.1540-5915.1997.tb01306.x

APPENDIX A. REPRODUCIBILITY RECORD

The score and robustness calculations, beginning at the partial-value stage, can be reconstructed from Tables 5, 10, and 11. For each scenario, the unrounded relation $a_{ij} = w_i/w_j$ generates an exactly reciprocal matrix satisfying $Aw = 9w$, $\lambda_{max} = 9$, and CR = 0. Table A1

displays entries to three decimals, so its printed values do not preserve exact reciprocity. All calculations use the unrounded ratios, and CR is reported as 0.000 to three decimal places.

TABLE A1. ROUNDED DISPLAY OF THE DIRECT-RATIO AHP MATRIX FOR MANUFACTURING.

	C1	C2	C3	C4	C5	C6	C7	C8	C9
C1	1.000	2.000	4.000	4.000	2.000	1.538	1.538	1.538	1.818
C2	.500	1.000	2.000	2.000	1.000	.769	.769	.769	.909
C3	.250	.500	1.000	1.000	.500	.385	.385	.385	.455
C4	.250	.500	1.000	1.000	.500	.385	.385	.385	.455
C5	.500	1.000	2.000	2.000	1.000	.769	.769	.769	.909
C6	.650	1.300	2.600	2.600	1.300	1.000	1.000	1.000	1.182
C7	.650	1.300	2.600	2.600	1.300	1.000	1.000	1.000	1.182
C8	.650	1.300	2.600	2.600	1.300	1.000	1.000	1.000	1.182
C9	.550	1.100	2.200	2.200	1.100	.846	.846	.846	1.000

The nominal Manufacturing prompt score is:
V(P) = .20(86) + .10(92) + .05(90) + .05(70) + .10(88) + .13(92) + .13(95) + .13(95) + .11(94) = 90.20

For each leader-challenger pair, the audit multiplies every one of the 512 vectors $m \in \{0.8, 1.2\}^9$ by the base weights, renormalises, and evaluates the clipped lower-minus-upper criterion values. None of the reported pairwise values reaches a 0 or 100 clipping boundary, so this is numerically equivalent here to subtracting $2u_s$ from each weighted difference. The smallest result is compared with $\varepsilon = 3.0$. The robust gaps are 6.9400, 3.1742, 3.9121, 3.4063, and -2.5630 for Manufacturing, Banking, Healthcare, Legal, and Retail respectively.

APPENDIX B. MINIMUM FIELD DECISION RECORD

A field application should retain the following artefacts as a single versioned decision package:
Decision owner, task boundary, prohibited uses, acceptance test, held-out set, and temporal cutoff.
Configuration identifiers for model, prompt, context, behaviour, orchestration, serving, data, tools, index, and adapter.
Gate status with evidence, owner, test date, and a named evidence-acquisition action for every unresolved gate.
Raw measurements, uncertainty, value-function anchors, transformed values, and acceptance-threshold results.
Individual and group AHP matrices, role weights, consistency ratios, final priorities, and preferential-independence review.
Nominal values, full sensitivity set, pairwise robust gaps, practical-difference threshold, and abstention rationale.
Recommendation or no-go state, pilot design, validity conditions, rollback plan, expiry date, and lifecycle triggers.