International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 3, May-June 2025||

DOI:10.15662/IJRPETM.2025.0803012

Quantifying the Cost and Risk of Enterprise LLM Adaptation Strategies

Karan Nisar

Independent Researcher, USA

ABSTRACT: Enterprises choosing between prompt engineering, retrieval-augmented generation (RAG) and parameter-efficient fine-tuning are well served by the accuracy literature and poorly served by everything else. Published comparisons establish which technique injects knowledge more effectively in a given domain, but none reports what the alternatives cost to build and run, how they differ in latency, what infrastructure each obliges an organisation to own, or how their risk exposures differ in kind rather than degree. This paper supplies that missing quantitative and operational account. We first consolidate the comparative accuracy evidence and state plainly what it does and does not support. We then develop an adaptation cost of ownership model and derive a closed-form break-even condition for fine-tuning relative to prompting. The condition has three properties that are counter-intuitive and consequential: output-token cost cancels entirely, so the economic case rests only on input tokens removed per request; the break-even volume is inversely proportional to input token price, so the same task amortises very differently on frontier and small models; and it is inversely proportional to the achieved prompt-length reduction, which is a design variable rather than a constant. Under illustrative parameters the break-even sits near 0.64 million requests per month for a 4,000-token prompt, well above the volume of most internal enterprise applications. The practical conclusion is that fine-tuning is a behavioural instrument rather than a cost-reduction strategy, and that RAG lowers cost at no volume relative to the prompt it augments, so its business case must rest on freshness, provenance and accuracy. We close with a latency decomposition showing that decode time dominates, and a risk analysis showing that the archetypes do not sit on a single risk gradient: fine-tuning concentrates data risk, retrieval concentrates system risk, and prompting concentrates assurance risk.

KEYWORDS: Large language models; total cost of ownership; break-even analysis; retrieval-augmented generation; fine-tuning; inference latency; risk assessment; enterprise AI economics.

I. INTRODUCTION

The mechanisms by which an enterprise adapts a general-purpose language model to a specific task are, by now, well described. Prompting conditions a frozen model through its input, retrieval attaches an external non-parametric memory, and parameter-efficient fine-tuning modifies a small set of weights on curated data; these differ structurally in where task knowledge is stored and which component is mutable at run time [1]. What remains poorly established is what any of them costs.

This gap is not for want of comparison. Several studies place the techniques side by side, but they compare accuracy, on academic or narrowly domain-bound datasets, using different base models and metrics. Ovadia et al. found supervised fine-tuning a weaker mechanism for injecting new facts than retrieval, with unsupervised continued pre-training weaker still [2]. Balaguer et al., in agricultural question answering, found the two contributed along partly independent dimensions and combined well [3]. Soudani et al. found the ranking conditional on entity popularity, with retrieval dominating for less popular knowledge [4]. None of these reports a dollar figure, a latency budget, or a maintenance burden.

Enterprise decisions, however, are not made on accuracy alone. They are made under simultaneous constraints on build budget, per-request cost at forecast volume, latency ceilings, data sensitivity, auditability and the availability of teams who can operate the result. Work on standardising the path from prototype to production makes the same point from the workflow side: the difficulty in enterprise machine learning concentrates outside the modelling step [5], and the same is true of adaptation economics.

This paper addresses three questions. **RQ1**: what does the published comparative evidence actually support, and with what caveats? **RQ2**: under what volume, price and prompt-length conditions is each archetype the lower-cost option,

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 3, May-June 2025||

DOI:10.15662/IJRPETM.2025.0803012

and can that condition be stated in closed form? **RQ3**: how do the archetypes differ in operational burden and risk exposure, and are those differences of degree or of kind?

Section II consolidates the accuracy evidence. Section III develops the cost model and break-even condition. Section IV covers latency and infrastructure. Section V analyses risk. Sections VI to VIII discuss, bound and conclude.

II. WHAT THE COMPARATIVE EVIDENCE SUPPORTS

Table 1 summarises the direct comparisons available. Four claims are reasonably well supported.
First, for injecting knowledge the base model lacks, retrieval outperforms supervised fine-tuning, with continued pre-training on raw corpus text weaker still [2]. Second, that advantage is conditional on how well represented the target entities already are, so retrieval's margin is largest exactly where enterprise value usually sits, in long-tail proprietary content [4]. Third, fine-tuning contributes along a dimension partly orthogonal to retrieval, so combinations can exceed either alone [3]. Fourth, training a generator on retrieval-augmented inputs that include distractors improves its use of evidence at inference time [6]. One widely held claim is *not* supported: that fine-tuning on domain documents installs domain facts. Where gains appeared, they required fact-based restructuring of the training data rather than exposure to raw text [7]. A related finding tempers expectations in the other direction: low-rank adaptation learns less than full fine-tuning on the target task but forgets less of the base model's general capability [8]. A caveat applies to all of it. Each study covers one or two domains, with different base models, retrievers and metrics, and none reports cost, latency or maintenance outcomes. This literature should be used to fix the *direction* of effects. Magnitudes must be established locally, against a written acceptance test, using instruments that separate retrieval quality from generation faithfulness [9] and probe robustness under retrieval noise [10].

TABLE 1. Published comparative findings on knowledge injection and behavioural adaptation. All entries are accuracy-oriented; none of the cited studies reports cost, latency or maintenance outcomes.

Study	Setting	Comparison	Principal finding	Limitation for enterprise transfer
Ovadia et al. [2]	Multiple-choice knowledge tasks, current events	Base vs. unsupervised fine-tuning vs. RAG	RAG superior for knowledge injection; fine-tuning weaker	Multiple-choice format; no cost or latency measurement
Balaguer et al. [3]	Agricultural question answering	Prompting vs. RAG vs. fine-tuning vs. combined	Contributions partly independent; combination strongest	Single domain; small evaluation set
Soudani et al. [4]	QA stratified by entity popularity	RAG vs. fine-tuning across model sizes	Retrieval dominates for less popular knowledge	Popularity proxy imperfect; open-domain corpora
Mecklenburg et al. [7]	New-knowledge injection by supervised fine-tuning	Naive vs. fact-based augmentation	Gains require fact-based data construction	Narrow task family
Zhang et al. [6]	Domain-specific RAG	Standard vs. retrieval-aware fine-tuning	Training with retrieved context and distractors improves evidence use	Requires both a corpus and labelled data
Biderman et al. [8]	Continued training, code and mathematics	LoRA vs. full fine-tuning	LoRA learns less but forgets less; acts as a regulariser	General-domain rather than enterprise evaluation
Chen et al. [10]	RAG benchmark under retrieval noise	Multiple base models	Noise robustness and negative rejection both imperfect	Benchmark rather than deployment evidence
Barnett et al. [11]	Field study of RAG deployments	Observational	Seven recurring failure points, mostly at pipeline seams	No quantitative outcome measures

International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 3, May-June 2025||

DOI:10.15662/IJRPETM.2025.0803012

III. AN ADAPTATION COST OF OWNERSHIP MODEL

A. The Model

Cost comparisons in the practitioner literature are usually anecdotal. We therefore state a model explicitly. Let V be monthly request volume; p_i and p_o the input and output token prices; n_p and n_f the input tokens required per request under prompting and after fine-tuning respectively; n_s a short system instruction; $n_c = k \cdot l$ the retrieved context tokens for k chunks of l tokens; n_o the output tokens; c_r the per-request retrieval compute cost; F_R the monthly fixed cost of index and ingestion pipeline; C_B the one-off build cost of fine-tuning, dominated in practice by data curation and annotation rather than GPU time; T the amortisation horizon in months; and H the monthly floor for dedicated serving. Then:

$$M_{\text{prompt}}(V) = V(n_p\, p_i + n_o\, p_o)$$
$$M_{\text{RAG}}(V) = V((n_s + n_c)\, p_i + n_o\, p_o + c_r) + F_R$$
$$M_{\text{FT}}(V) = V(n_f\, p_i + n_o\, p_o) + C_B/T + H$$
$$M_{\text{hybrid}}(V) = V((n_f + n_c)\, p_i + n_o\, p_o + c_r) + F_R + C_B/T + H$$

B. The Break-Even Condition

Setting $M_{\text{FT}} = M_{\text{prompt}}$ and solving gives the break-even volume

$$\boldsymbol{V^* = (C_B/T + H) \,/\, ((n_p - n_f)\, p_i)}$$

Three properties of this expression deserve emphasis because each contradicts a common assumption.

Output cost cancels. The economic case for fine-tuning depends entirely on how many *input* tokens it removes from every request. Business cases built on total token spend, rather than on the input-side delta, are measuring the wrong quantity.

***V** is inversely proportional to input token price.** The same task can be economically fine-tuned an order of magnitude earlier on an expensive frontier model than on a cheap small one. Archetype choice is therefore not separable from model-tier choice.

***V** is inversely proportional to the prompt-length reduction**, which is a design variable rather than a constant. A task whose prompt-only baseline requires a long rubric and many exemplars amortises fine-tuning far sooner than one needing a two-line instruction.

C. Illustrative Magnitudes

With the parameters in Table 2, V^* is approximately 1.49 million requests per month at a 2,000-token prompt, 0.64 million at 4,000 tokens, and 0.30 million at 8,000 tokens (Fig. 1(a)). At a frontier-tier input price five times higher, the 4,000-token case falls to roughly 128,000.

These are order-of-magnitude figures, but the conclusion is robust across wide variation in the inputs: **at the volumes typical of internal enterprise applications, fine-tuning is not a cost-reduction strategy.** It becomes one only in high-volume, customer-facing, long-prompt settings. Elsewhere it must be justified behaviourally.

Fig. 1(b) makes a second point that is routinely missed in business cases: relative to the short instruction it augments, RAG reduces cost at *no* volume, because retrieved evidence inflates the request rather than shortening it. Every request carries the retrieved context, the retrieval compute and a share of the fixed pipeline cost. The RAG curve in Fig. 1(b) does fall below the prompting baseline above $V = F_R / ((n_p - n_s - n_c)\, p_i - c_r) \approx 0.64$ million requests per month, but that crossing is a substitution effect rather than a retrieval saving: it appears only because the retrieval-augmented request ($n_s + n_c = 3{,}100$ tokens) happens to be shorter than the 4,000-token engineered prompt whose rubric and exemplars the evidence displaces, and it vanishes whenever $n_s + n_c \geq n_p$. Retrieval's justification is freshness, provenance and accuracy. Saying so explicitly prevents a common category error, in which a project promising cost savings is later judged against a criterion it was never capable of meeting.

TABLE 2. Illustrative parameters for the cost model. These are modelled planning values chosen to be broadly representative of hosted and self-hosted price points at the time of writing. They are not measurements; a reader applying the model should substitute current organisation-specific figures.

Parameter	Symbol	Illustrative value	Notes
Input token price	p_i	US$1.00 per million	Mid-tier hosted or self-hosted equivalent
Output token price	p_o	US$3.00 per million	Cancels in the break-even condition
Prompt-only input tokens	n_p	4,000	System prompt, rubric and six

International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 3, May-June 2025||

DOI:10.15662/IJRPETM.2025.0803012

			exemplars
Post-fine-tuning input tokens	n_f	500	Short instruction only
System instruction (RAG)	n_s	600	
Retrieved context tokens	n_c	2,500	k = 5 chunks of l = 500 tokens
Output tokens	n_o	300	
Retrieval compute per request	c_r	US$0.0002	Embedding, ANN search and re-rank
Index and pipeline fixed cost	F_R	US$450 per month	
Fine-tuning build cost	C_B	US$28,000	Dominated by curation and annotation
Amortisation horizon	T	36 months	
Dedicated serving floor	H	US$1,460 per month	
Derived: break-even volume	V^*	**≈ 639,000 requests / month**	At n_p = 4,000
Derived: break-even at n_p = 8,000	V^*	≈ 300,000 requests / month	
Derived: payback at V = 3 million / month		≈ 3.1 months	Net monthly saving ≈ US$9,040

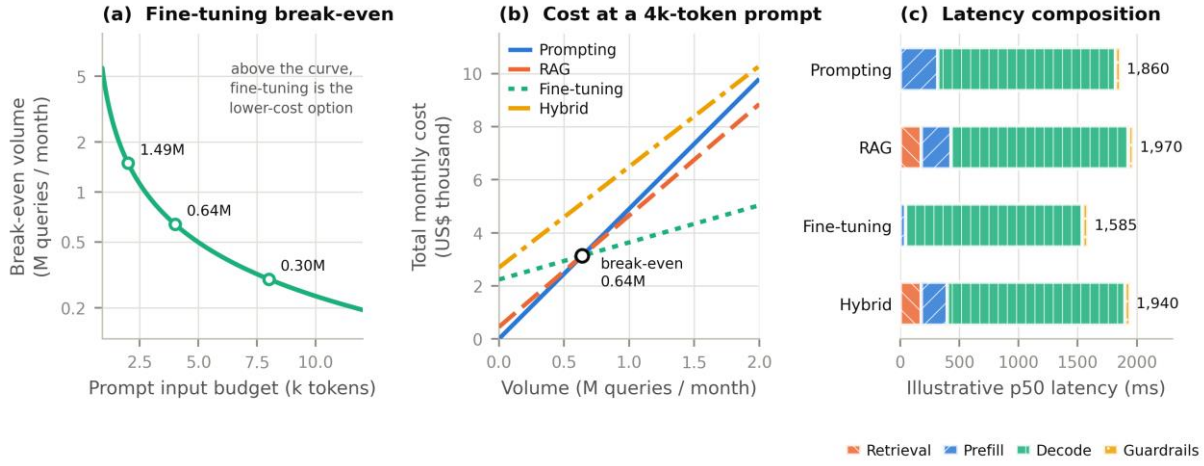

Fig. 1. Cost and performance analysis. (a) Break-even monthly volume above which fine-tuning is the lower-cost archetype, as a function of the prompt-only input budget. (b) Total monthly cost against volume for the four archetypes at a 4,000-token prompt. (c) Illustrative decomposition of p50 response latency, showing that decode time dominates and that the fine-tuning advantage is a prefill effect. All values follow from the model and parameters of Section III and are illustrative rather than measured.

IV. LATENCY, THROUGHPUT AND OPERATIONAL BURDEN

A. Latency

Latency decomposes into retrieval, prefill, decode and guardrail time (Fig. 1(c)). Two conclusions follow.

Decode dominates, so inter-archetype differences are proportionally smaller than practitioners assume. Adding 180 ms of retrieval to a request whose generation takes 1.5 s changes total latency by roughly 10 per cent. Latency objections to retrieval should therefore be measured rather than assumed.

The latency advantage of fine-tuning is a *prefill* advantage obtained by removing instruction tokens, so it is largest exactly where the prompt-only baseline is longest, which is the same condition under which the cost case is strongest. Serving-side techniques act on the dominant term rather than on archetype choice: paged key-value caches with prefix sharing [12] and speculative decoding [13] improve all four archetypes and should be assumed present before archetype choice is used to buy latency.

Scalability differs in kind rather than degree. Prompting and RAG scale horizontally, with an index that scales sub-linearly in corpus size under graph-based approximate nearest-neighbour search [14]. Fine-tuning scales across a *portfolio*, because small adapters allow one base deployment to serve many tasks.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 3, May-June 2025||

DOI:10.15662/IJRPETM.2025.0803012

B. Infrastructure and Ownership

Table 3 records what each archetype obliges an enterprise to own and operate. The salient point is not the volume of the burden but its distribution: RAG is predominantly a data-engineering and information-retrieval obligation, while fine-tuning is predominantly a machine-learning-platform obligation. Organisations frequently possess one capability and not the other. This is an under-recognised determinant of the correct choice, and it compounds the more general finding that machine-learning systems accrue maintenance debt disproportionate to their code volume [15] and that deployment difficulty concentrates outside modelling [16], [5].

TABLE 3. Infrastructure and operational requirements by archetype.

Requirement	Prompting	RAG	Fine-tuning	Hybrid
Accelerators for training	None	None	Yes (single node feasible with QLoRA [17])	Yes
Dedicated serving capacity	Optional	Optional	Usually required	Usually required
Embedding model and pipeline	No	Yes	No	Yes
Vector and lexical index	No	Yes (HNSW [14] or IVF-PQ [18])	No	Yes
Re-ranking component	No	Recommended [19]	No	Recommended
Document ingestion and refresh jobs	No	Yes	No	Yes
Access-control propagation to retrieval	No	Yes	Not applicable	Yes
Labelling and annotation capability	No	No	Yes	Yes
Model registry and adapter versioning	Prompt registry only	Prompt registry only	Yes	Yes
Evaluation harness	Task rubric	Rubric plus retrieval metrics [9]	Rubric plus regression suite	All of the above
Primary owning function	Application engineering	Data engineering and IR	ML platform	Both

V. RISK EXPOSURE

Fig. 2 assesses risk by archetype. The structural finding is that the archetypes do not sit on a single risk gradient, so "which is riskier" is not a well-formed question.

Fine-tuning concentrates data risk. Models memorise and can be induced to emit training data [20], and memorisation scales with capacity, duplication and context length [21]. Because the exposure is embedded in weights, it is not revocable: an erasure obligation cannot be met by deleting a source record, and the artefact carries a documentation burden for training-data provenance [22], [23].

Retrieval concentrates system risk. It relocates rather than removes exposure. The index becomes an attack surface and retrieved content an injection vector, so an integrated application can be compromised by instructions planted in documents it retrieves [24], [25]. Recurring engineering failures cluster at pipeline seams [11].

Prompting concentrates assurance risk. The least is written down, and behaviour can drift silently under a prompt edit or a base-model version change, because output quality is sensitive to surface form [26].

The practical implication is that the archetypes require different *controls*, not different amounts of the same control. A control catalogue written for one archetype will systematically under-protect another. Regulatory frameworks reinforce this: an obligation to exhibit a human-inspectable source for an assertion is discharged far more cheaply by retrieval, which produces citable evidence as a by-product, than by fine-tuning, whose knowledge is diffused across weights [22], [27]. Fine-tuning creates a regulated *artefact* whose training data must be documented; retrieval creates a regulated *process*.

International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 3, May-June 2025||

DOI:10.15662/IJRPETM.2025.0803012

Relative risk exposure by adaptation archetype

L = low, M = medium, H = high exposure relative to the other archetypes; n/a = mechanism not present

	Prompting	RAG	Fine-tuning	Hybrid
Unsupported assertion	**H**	M	M	M
Stale knowledge	**H**	L	**H**	L
Training-data leakage	n/a	n/a	**H**	**H**
Prompt injection	L	**H**	n/a	**H**
Entitlement bypass	n/a	M	**H**	M
Quality regression	**H**	M	M	M
Pipeline failure	L	**H**	L	**H**
Erasure infeasible	n/a	L	**H**	M
Documentation burden	L	M	**H**	**H**

Fig. 2. Relative risk exposure by adaptation archetype. Ratings are relative to the other archetypes in the same row rather than absolute, and "n/a" marks a mechanism that is not present in that archetype.

VI. DISCUSSION

Fine-tuning is a behavioural instrument, not a cost instrument. This is the most immediately actionable result here, because it contradicts a widespread business-case pattern. The break-even condition depends only on the fixed cost and the input tokens removed per request, and at internal-application volumes it is rarely met. Enterprises that fine-tune to reduce cost are usually mis-specifying a behavioural requirement as an economic one, and the resulting projects are then judged against the wrong success criterion.

Retrieval's business case is freshness and provenance, never savings. Relative to the prompt it augments, retrieval increases per-request cost monotonically. Where a RAG deployment does undercut a prompt-only baseline, as in Fig. 1(b) at high volume, the saving comes from retiring a long rubric and exemplar block, not from retrieval itself, and it disappears for any baseline shorter than the retrieved context. That is not an argument against retrieval; it is an argument for describing its benefits accurately.

Cost and risk are not separable from organisational capability. Table 3 shows the operational burden falling on different functions, and firm-level evidence indicates that value from AI is mediated by organisational capability rather than determined by technology choice [28], [29]. An archetype an organisation cannot staff is not cheaper for being theoretically optimal.

VII. LIMITATIONS AND FUTURE WORK

The cost model is deliberately simple and its parameters are illustrative. It omits engineering labour beyond C_B, caching effects, batching efficiency, reserved-capacity discounts, and the portfolio effect by which one base deployment amortises across many adapters. All of these move V^*, some substantially. Caching deserves particular note: the derivation charges the full prompt cost on every request, so any mechanism serving a repeated prefix at reduced cost lowers the prompting baseline and therefore *raises* V^*, strengthening rather than weakening the paper's conclusion. The model should be read as establishing the structure of the trade-off, particularly the cancellation of output cost and the inverse dependence on prompt-length reduction, rather than as a calculator.

The risk ratings are analytic, not measured. They are ordinal judgements derived from the cited literature and are intended for comparison across a row, not as absolute exposure levels.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 3, May-June 2025||

DOI:10.15662/IJRPETM.2025.0803012

Parameters will move, in identifiable directions. Per-token inference prices have tended to fall, which raises V^*. Prefix caching has the same directional effect. Context windows continue to lengthen, which shifts constants without disturbing the structure of the trade-off. Readers should recompute rather than quote.

Two developments fall outside this analysis rather than merely recalibrating it. Iterative or agent-directed retrieval is not represented by the single pre-generation retrieval step assumed here. And systems that expend substantial computation at inference time to reason before answering break the assumption that cost and latency track prompt length and output length, introducing a term that scales with task difficulty. Extending the cost model to price inference-time reasoning explicitly is the most valuable next step.

VIII. CONCLUSION

The literature comparing enterprise adaptation strategies measures accuracy and stops. This paper has supplied the economic and operational account that decisions actually require. A closed-form break-even condition shows that the case for fine-tuning rests entirely on input tokens removed per request, is inversely proportional to both token price and prompt-length reduction, and is rarely satisfied at internal enterprise volumes; fine-tuning should therefore be selected and defended as a behavioural instrument. Retrieval never reduces cost relative to the prompt it augments, and any apparent saving against a long engineered baseline is evidence displacing exemplars rather than retrieval paying for itself; its business case must be made on freshness, provenance and accuracy instead. Latency differences between archetypes are smaller than commonly assumed because decode time dominates. And risk does not increase monotonically across the archetypes: fine-tuning concentrates irreversible data exposure, retrieval concentrates system and injection exposure, and prompting concentrates assurance exposure, so each requires a distinct control set. Together these establish the quantitative basis on which a defensible selection procedure can be built.

REFERENCES

[1] K. Nisar, "Prompting, Retrieval, and Fine-Tuning: Foundations of Enterprise Language Model Adaptation," *International Journal of Engineering & Extended Technologies Research (IJEETR)*, vol. 6, no. 6, pp. 9310–9319, Nov. 2024, doi: 10.15662/IJEETR.2024.0606030.

[2] O. Ovadia, M. Brief, M. Mishaeli, and O. Elisha, "Fine-tuning or retrieval? Comparing knowledge injection in LLMs," arXiv preprint arXiv:2312.05934, Dec. 2023.

[3] A. Balaguer, V. Benara, R. L. de Freitas Cunha, R. de M. Estevão Filho, T. Hendry, D. Holstein, J. Marsman, N. Mecklenburg, S. Malvar, L. O. Nunes, R. Padilha, M. Sharp, B. Silva, S. Sharma, V. Aski, and R. Chandra, "RAG vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture," arXiv preprint arXiv:2401.08406, Jan. 2024.

[4] H. Soudani, E. Kanoulas, and F. Hasibi, "Fine tuning vs. retrieval augmented generation for less popular knowledge," arXiv preprint arXiv:2403.01432, Mar. 2024.

[5] 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, 2025, doi: 10.22399/ijcesen.5364.

[6] 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," arXiv preprint arXiv:2403.10131, Mar. 2024.

[7] N. Mecklenburg, Y. Lin, X. Li, D. Holstein, L. Nunes, S. Malvar, B. Silva, R. Chandra, V. Aski, P. K. R. Yannam, T. Aktas, and T. Hendry, "Injecting new knowledge into large language models via supervised fine-tuning," arXiv preprint arXiv:2404.00213, Mar. 2024.

[8] D. Biderman, J. Portes, J. J. Gonzalez Ortiz, M. Paul, P. Greengard, C. Jennings, D. King, S. Havens, V. Chiley, J. Frankle, C. Blakeney, and J. P. Cunningham, "LoRA learns less and forgets less," arXiv preprint arXiv:2405.09673, May 2024.

[9] 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*, Mar. 2024, pp. 150–158, doi: 10.18653/v1/2024.eacl-demo.16.

[10] J. Chen, H. Lin, X. Han, and L. Sun, "Benchmarking large language models in retrieval-augmented generation," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 38, no. 16, 2024, pp. 17754–17762, doi: 10.1609/aaai.v38i16.29728.

[11] S. Barnett, S. Kurniawan, S. Thudumu, Z. Brannelly, and M. Abdelrazek, "Seven failure points when engineering a retrieval augmented generation system," in *Proc. IEEE/ACM 3rd Int. Conf. AI Engineering: Software Engineering for AI (CAIN)*, 2024, pp. 194–199, doi: 10.1145/3644815.3644945.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 3, May-June 2025||

DOI:10.15662/IJRPETM.2025.0803012

[12] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica, "Efficient memory management for large language model serving with PagedAttention," in *Proc. 29th ACM Symp. Operating Systems Principles (SOSP)*, 2023, pp. 611–626, doi: 10.1145/3600006.3613165.
[13] Y. Leviathan, M. Kalman, and Y. Matias, "Fast inference from transformers via speculative decoding," in *Proc. 40th Int. Conf. Machine Learning (ICML)*, PMLR vol. 202, 2023, pp. 19274–19286.
[14] Y. A. Malkov and D. A. Yashunin, "Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 4, pp. 824–836, Apr. 2020, doi: 10.1109/TPAMI.2018.2889473.
[15] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison, "Hidden technical debt in machine learning systems," in *Advances in Neural Information Processing Systems 28*, 2015, pp. 2503–2511.
[16] 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.
[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, pp. 10088–10115.
[18] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, Jul. 2021, doi: 10.1109/TBDATA.2019.2921572.
[19] R. Nogueira and K. Cho, "Passage re-ranking with BERT," arXiv preprint arXiv:1901.04085, Jan. 2019.
[20] N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, Ú. Erlingsson, A. Oprea, and C. Raffel, "Extracting training data from large language models," in *Proc. 30th USENIX Security Symposium*, 2021, pp. 2633–2650.
[21] N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramèr, and C. Zhang, "Quantifying memorization across neural language models," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2023.
[22] National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, Gaithersburg, MD, USA, Jan. 2023, doi: 10.6028/NIST.AI.100-1.
[23] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, Dec. 2021, doi: 10.1145/3458723.
[24] K. Greshake, S. Abdelnabi, 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 (AISec)*, 2023, pp. 79–90, doi: 10.1145/3605764.3623985.
[25] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly," *High-Confidence Computing*, vol. 4, no. 2, art. 100211, Jun. 2024, doi: 10.1016/j.hcc.2024.100211.
[26] M. Sclar, Y. Choi, Y. Tsvetkov, and A. Suhr, "Quantifying language models' sensitivity to spurious features in prompt design, or: How I learned to start worrying about prompt formatting," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2024.
[27] European Parliament and Council of the European Union, *Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*, OJ L, 2024/1689, 12 Jul. 2024.
[28] I. M. Enholm, E. Papagiannidis, P. Mikalef, and J. Krogstie, "Artificial intelligence and business value: A literature review," *Information Systems Frontiers*, vol. 24, no. 5, pp. 1709–1734, Oct. 2022, doi: 10.1007/s10796-021-10186-w.
[29] P. Mikalef and M. Gupta, "Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance," *Information & Management*, vol. 58, no. 3, art. 103434, Apr. 2021, doi: 10.1016/j.im.2021.103434.