BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

Efficient Model Compression: A Teacher – Student Distillation Framework for Cost-Efficient Customer Support Automation

Piyush berwal and Prof.Swarupa kamble

Department of Artificial Intelligence and Data Science
Ajeenkya DY Patil School of Engineering, Pune, Maharashtra, India
Savitribai Phule Pune University | Technical Seminar (ELC342CSD), A.Y. 2026-27

Abstract

Customer-support automation needs intent classifiers that approach the accuracy of frontier large language models (LLMs) without paying their per-ticket cost and latency. This paper is a literature-based analytical study of three compression techniques-knowledge distillation, parameter-efficient fine-tuning (LoRA/QLoRA) and post-training quantization (GPTQ/AWQ)-together with intent-classification benchmarks and cost-aware LLM cascading, centred on ten primary studies published between 2015 and 2025. A criterion-driven analysis reveals that all three techniques are mainly evaluated individually on general benchmark tests, and no one of the studied papers evaluates the joint effect of the techniques on support-ticket classification. To address the detected deficiency, we suggest the concept of a teacher-student-quantization framework with a confidence escalation strategy and describe the evaluation methodology for it. Our contribution here is conceptual: no dataset has been created, no model has been trained, and no performance metrics have been evaluated yet.

Keywords : Customer-support automation, intent classification, knowledge distillation, LoRA, model compression, post-training quantization, QLoRA, teacher-student learning.

I. INTRODUCTION

Customer support sits at the heart of operations for platform-driven companies. These teams deal with an endless stream of repetitive tickets—late deliveries, missing items, refund requests, and account access headaches. Intent classification steps in as the first line of defense, sorting each ticket into predefined buckets so support teams can tackle issues more efficiently. Large language models (LLMs) handle this task easily, adapting to all sorts of requests. But there's a catch: every time you query a cutting-edge LLM, you pay a fee and wait for a response. That adds up fast—both in dollars and time.

So, companies are hunting for ways to squeeze LLM performance into smaller, cheaper packages. Three core approaches jump out in recent research. First, knowledge distillation lets you train a smaller "student" model by mimicking a larger "teacher." Next up, parameter-efficient fine-tuning techniques like Low-Rank Adaptation (LoRA) and its quantized cousin QLoRA, make it easier and cheaper to specialize these student models. Finally, post-training quantization reduces the numerical precision of model weights, cutting down on memory use and speeding up inference. These methods have strong

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

backing individually, but researchers haven't really explored how they work together in customer support settings.

Companies need a lean, cost-effective classifier that delivers top-notch accuracy on high-volume, repetitive tickets. It has to avoid the high costs and slowdowns of querying big frontier models for every ticket, and steer clear of inflexible rule-based systems.

This paper approaches the topic as an analytical review grounded in the literature. We haven't trained or measured a model ourselves; instead, we synthesize existing studies and draft a conceptual framework, with each part rooted in prior research. The full impact of combining these pieces remains unanswered. Here's what we bring:

A structured review of ten studies covering distillation, LoRA/QLoRA, quantization, intent classification, and LLM cascading. A comparison based on clear criteria to pinpoint gaps in current research. A conceptual teacher–student–quantization framework, plus a confidence-based escalation mechanism and suggested evaluation protocol for future work.

Section II describes the methodology, Section III gives background, Section IV reviews the literature, Section V analyses the research gap, Section VI presents the framework, and Sections VII–IX cover feasibility, limitations and conclusions.

II. RESEARCH METHODOLOGY

This study combines a literature review with conceptual framework development. No dataset was collected, no model was trained, and no accuracy, latency or cost figure was measured for this paper. There are 10 main studies examined on distillation [1]–[3], [11], distillation parameter-efficient fine-tuning [4], [5], quantization [6], [7], intent classification [8], [9] and cost-aware cascading study [10] according to a coordinated set of parameters: accuracy in reference to an uncompressed baseline; latency or throughput of the model; training and inference expenditure; model size and memory requirements; usage of a particular method of compression, and information on whether the deployment is feasible. The background literature which includes calibration study [12], [13], sequence level distillation [14], dataset shift study [15], and integer quantization [16] is used to aid in discussing the framework.

Noting the differences in the tasks and benchmarks used among the studies, the results will be cited as reported by the authors and will not be aggregated into a single figure. It is essential to differentiate (a) established concepts, (b) results reported by the quoted source, (c) the framework of the current research, and (d) the future empirical study.

III. THEORETICAL BACKGROUND

A. Customer-Support Intent Classification

Intent classification maps a ticket's free text to a category describing what the customer needs, such as a refund request or an account-access issue [8]. In production, the classifier has to decide on the mode of routing, whether an auto-response is to be used, and whether human involvement is needed.

Two factors are critical for pooling. First, we operate on the assumption that a couple of intents are responsible for the majority of tickets, which indicates that a simple classifier can serve the purpose. Second, it is as important to identify an out-of-scope query (i.e. a query that relates to no intent) as being

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

able to classify in-scope queries since wrongly classifying a new query can lead to an inappropriate automated reaction.

B. Knowledge Distillation and Teacher–Student Architecture

Distillation trains a small student to reproduce the output distribution of a larger, already-trained teacher instead of only the hard labels [1]. The teacher's temperature-softened probabilities reveal which incorrect classes it considers plausible, so a ticket labelled "refund request" with some probability on "wrong item" conveys genuine ambiguity that a hard label hides. With temperature T, softmax σ and mixing weight λ, the student's loss is a weighted combination [1]:

$$L = (1 - \lambda)\cdot \mathrm{CE}(y, \sigma(z_s)) + \lambda \cdot T^2 \cdot \mathrm{KL}(\sigma(z_t/T) \,\|\, \sigma(z_s/T)) \quad (1)$$

where z_t and z_s Teacher and student logits are involved in the process of distillation.

Distillation can either be offline, with the static teacher labeling the data once, or it can also be done online, where both the teacher and the student are trained. Offline approach of distillation particularly helps when the teacher is an advanced LLM and can only be accessed via an API, as it separates the costly labeling from the less expensive student training process; the student is therefore trained using the outputs given by the teacher [14]. Matching the layers as in TinyBERT [3] necessitates the access to the black-box teacher provided by the commercial LLMs. Moreover, while distillation transfers the outputs from the teacher, it does not transfer the thought processes and reasoning behind the process [1].

C. LoRA and QLoRA

The full fine-tuning method modifies every parameter in a way that demands high memory and computational capabilities. The method called LoRA fixes the weight matrix W and adds a low-rank modification calculated using the formula [4]

$$W' = W + (\alpha/r)\cdot BA, \quad r \ll d \quad (2)$$

where A and B are low-rank learning matrices with a rank r. The LoRA adds the modification to W allowing only a low percentage of parameters to be retrained while using the same technique as full fine-tuning.

The matrix B is initialized in a zero state, whereas BA is joined to W before the model goes into the inference phase.

The rank, scaling factor and adapted matrices must be tuned per task [4]. QLoRA quantizes the frozen base model to 4-bit NormalFloat (NF4), keeps the adapters in higher precision, and adds double quantization and paged optimizers; the authors report fine-tuning a 65-billion-parameter model on a single 48 GB GPU while preserving 16-bit fine-tuning performance [5]. Whether this carries over unchanged to the compact students considered here is not established by [5].

D. Post-Training Quantization: GPTQ and AWQ

Quantization represents weights, and optionally activations, in 8-bit or 4-bit integers instead of 16-bit floats. It can be applied after training or simulated during training [16]; post-training quantization needs no retraining and only a small calibration set, which makes it suitable for a student that is already fine-tuned [6], [7]. GPTQ quantizes layer by layer using approximate second-order information and is

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

reported to reach 3–4 bits with a minimal accuracy drop, but was validated mainly on generative tasks [6]. AWQ protects a small fraction of weight channels identified as salient from activation statistics while quantizing the rest more aggressively, and is reported to be faster and more accurate than earlier 4-bit methods [7]. Both rely on a calibration sample; the accuracy loss of any method grows with how aggressively precision is reduced [6], [7].

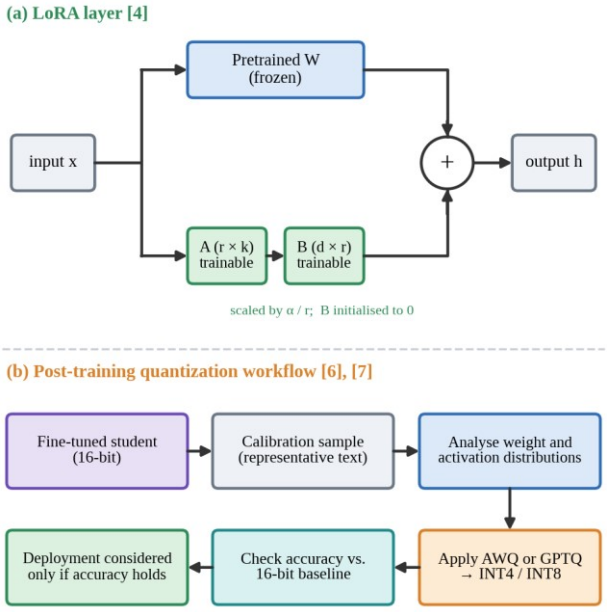


Fig. 1. Established techniques redrawn conceptually: (a) structure of a LoRA-adapted layer [4]; (b) post-training quantization workflow for a fine-tuned student [6], [7].

IV. LITERATURE REVIEW

Ten studies most relevant to the framework were examined for method, main finding as reported by the original authors, and reported limitations. Table I condenses this information; all figures are as stated in the cited sources and none was produced for this paper.

TABLE I

SUMMARY OF THE REVIEWED STUDIES

Ref., Year	Approach and main reported finding	Reported limitation
[1], 2015	Student trained on temperature-softened teacher outputs; retains most teacher accuracy at much smaller size.	Validated on small-scale vision and speech tasks, not LLMs.
[2], 2019	DistilBERT: 6-layer student from a 12-layer BERT via triple loss; 40% smaller, 60% faster, ~97% of GLUE performance.	Some accuracy loss on more complex language tasks.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

Ref., Year	Approach and main reported finding	Reported limitation
[3], 2019	TinyBERT: two-stage, layer-wise distillation; outperforms other compressed BERT variants of similar size.	Requires access to the teacher's intermediate layers.
[4], 2021	LoRA: frozen weights plus a low-rank update; <1% trainable parameters, matches full fine-tuning, no added latency once merged.	Rank and adapted matrices need per-task tuning.
[5], 2023	QLoRA: 4-bit NF4 base model plus LoRA adapters; 65B model fine-tuned on one 48 GB GPU at 16-bit-level performance.	Evidence is at large scale; compact students not tested.
[6], 2022	GPTQ: layer-wise post-training quantization using second-order information; accurate 3–4-bit weights with minimal accuracy drop.	Validated mainly on generative tasks.
[7], 2023	AWQ: protects activation-salient weight channels; faster and more accurate than earlier 4-bit methods.	Depends on a calibration set.
[8], [9], 2020, 2019	Lightweight dual encoders are competitive on intent detection [8]; CLINC150 adds out-of-scope queries [9].	General-purpose, not support-specific datasets.
[10], 2023	FrugalGPT: a cascade routes queries across LLMs using a scorer; matches the best single LLM with up to 98% cost reduction.	Savings come from routing, not from compressing one model.
[11], 2025	DeepSeek-R1: reasoning data distilled into dense 1.5B–70B students; the 32B student beats a comparable closed model.	Reasoning-focused; cost not directly quantified.

Read across the rows, distillation [1]–[3], [11] mainly reduces model size while retaining most teacher accuracy; LoRA and QLoRA [4], [5] mainly reduce the memory and hardware cost of adaptation rather than inference cost; GPTQ and AWQ [6], [7] mainly reduce inference memory and, typically, latency; and FrugalGPT [10] reduces aggregate cost through adaptive routing.

The intent-classification studies [8], [9] suggest the task suits lightweight models and stress out-of-scope handling. No single technique addresses all comparison criteria, which motivates combining them.

V. COMPARATIVE ANALYSIS AND RESEARCH GAP

Each study was rated against six criteria derived from Section II. The ratings in Fig. 2 are our qualitative reading of the studies as summarised in Table I, not a quantitative benchmark.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

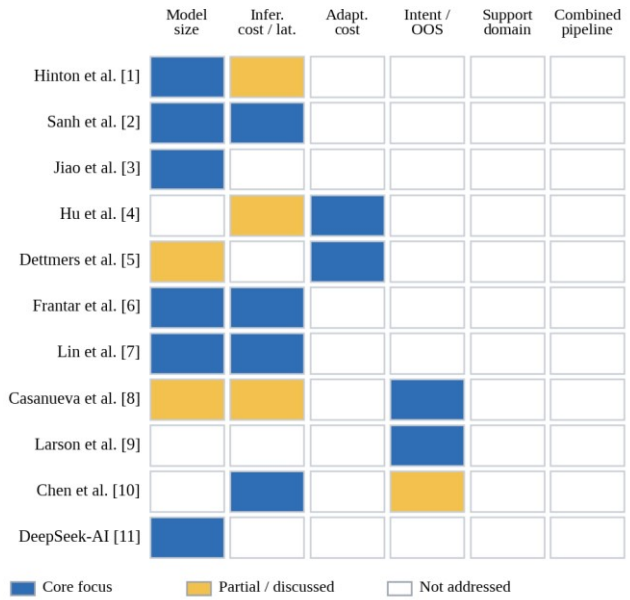


Fig. 2. Coverage of six design criteria by the reviewed studies (qualitative rating; OOS = out-of-scope). Reference [11] denotes DeepSeek-AI.

The matrix supports the following observations.

- **G1 – Isolated validation.** Each technique is evaluated on its own benchmarks; no study measures the compounded effect of distillation, LoRA/QLoRA and quantization on one model [1]–[7].
- **G2 – Domain mismatch.** Benchmarks are general language-understanding, generative or general-purpose intent datasets, not production support tickets [8], [9].
- **G3 – Unknown additivity.** It is not established whether the accuracy losses from distillation and quantization compound linearly, sub-linearly or interact, nor whether the benefits are fully compatible.
- **G4 – Escalation with compressed scorers.** The cascade in [10] used a single lightweight scorer (DistilBERT); its behaviour when driven by a distilled, LoRA-tuned and quantized student is unstudied, and confidence estimates may be poorly calibrated [12], [13].

Two cautions apply. The studies used different tasks, datasets and model families, so Table I is source-specific and no "best" technique can be derived from it. Most importantly, no reported result describes the combined, end-to-end pipeline for customer-support intent classification; this is the central gap addressed next.

VI. PROPOSED CONCEPTUAL FRAMEWORK

To address G1–G4 together, we outline the teacher–student–quantization (TSQ) framework in Fig. 3. It is a synthesis of mechanisms from the literature; it has not been implemented or evaluated, and no experimental results are claimed.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

A. Offline Teacher–Student–Quantization Pipeline

Raw, unlabelled support tickets are passed once to a frontier teacher LLM, which returns an intent label and, where the API exposes them, soft scores or a short reasoning trace, in the spirit of the data used for distillation in [11]. Following offline distillation [1], [14], these outputs form a labelled dataset for fine-tuning a compact student (low single-digit billions of parameters) with LoRA or QLoRA instead of full fine-tuning [4], [5]. The fine-tuned student is then quantized to INT4 or INT8 with GPTQ or AWQ [6], [7] and deployed. Because a commercial teacher is typically API-only, the proposal follows the output-only route of DistilBERT [2] rather than the layer-wise route of TinyBERT [3].

B. Confidence-Based Inference and Escalation

Modelled on FrugalGPT's cascading strategy [10], a new ticket is pre-processed and classified by the quantized student (Fig. 4). If the student's confidence p reaches a threshold τ, the ticket is auto-routed or auto-answered; otherwise it is escalated to the teacher or a human agent. Low confidence also serves as the signal for out-of-scope queries [9], with maximum softmax probability as the simplest baseline [13]. Escalated and low-confidence cases are logged as new labelled examples and periodically re-labelled by the teacher, closing the feedback loop of Fig. 3. Whether this loop works as intended, and where τ should be set, are empirical questions left open.

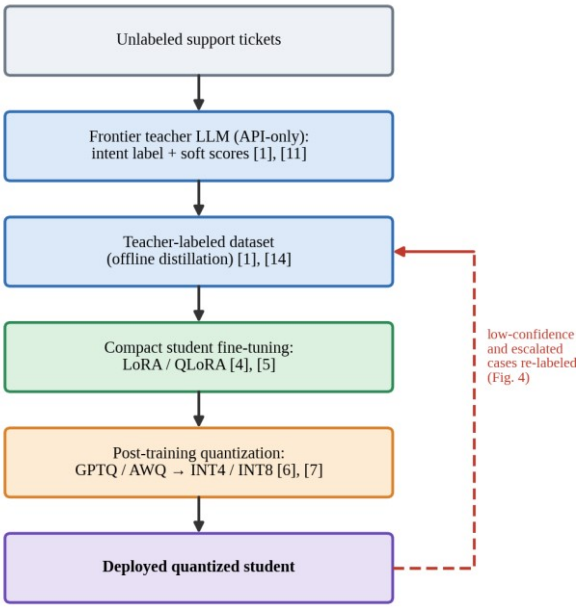


Fig. 3. Proposed offline teacher–student–quantization (TSQ) pipeline (conceptual framework; not implemented or evaluated).

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

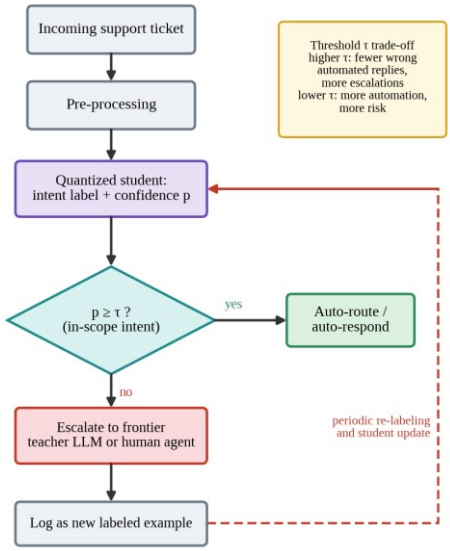


Fig. 4. Proposed inference-time decision flow of the lightweight student classifier (conceptual framework).

C. Proposed Evaluation Protocol

The framework should be compared with the frontier teacher, a distilled but unquantized student, and a DistilBERT-style student [2], using the measures in Table II. Ablations that add one stage at a time (distillation, then LoRA/QLoRA, then quantization) would show directly whether losses compound (G3).

TABLE II

PROPOSED EVALUATION METRICS

Component	Proposed metrics (source)
Classification	Accuracy, precision, recall and F1 against the teacher on held-out tickets
Out-of-scope	In-scope accuracy and out-of-scope recall on CLINC150-style data [9]
Efficiency	Latency, training and inference cost, memory footprint, model size [4]–[7]
Confidence	Calibration error [12]; max-softmax baseline [13]
Escalation	Escalation volume and overall system accuracy as functions of τ [10]
Robustness	Accuracy as the ticket mix changes over time [15]

VII. FEASIBILITY AND DEPLOYMENT ANALYSIS

The analysis relies on the findings from the reviewed literature and does not include in the results any data from any practical implementation.

Hardware, cost, and latency. QLoRA fine-tuned a 65 billion-parameter model on a single 48 GB GPU, which means that fine-tuning a much smaller student on one single workstation's GPU is feasible. Based on the reports in the literature, it is feasible to conduct inference with a relatively modest hardware

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

utilization. Although teacher labelling has a little sporadic cost per learning example, in the case of cascading teacher latencies with tickets being escalated, there is still no available information on how much it all costs.

Calibration and thresholds. The efficacy of the escalation gate is dependent on the student's level of confidence. Modern neural networks suffer from issues regarding calibration [12], and maximum softmax value should be treated as an indicator rather than an assured criterion that can be effective for flagging errors or out-of-distribution inputs [13]. Setting threshold τ represents a tradeoff between extent of automation and the probability of wrong automated answer conducted; the work does not tackle the question of threshold determination.

Calibration for quantum states. The GPTQ and AWQ algorithms illustrate the importance of making use of a small calibration dataset [6], [7]; however, calibration should be carried out with real instead of generic calibration data.

Transition of distribution. The teacher's labels show the mix of tickets at the time of labelling. Change in the mix like a seasonal increase in delivery complaints may lead to drop in the accuracy of students if no closed feedback loop is in action; however, there are a lot of ways to detect the transition in the datasets [15] that may trigger re-labelling.

Humans involved in work. The papers reviewed do not provide us with much useful guidance on how to assign the responsibility of automated decision making versus human intervention; the diagram shown in the Figure 4 does not give us the right answer. In practical terms it is not enough just to make use of monitoring, versioning, and budgeting for teacher re-labelling.

VIII. LIMITATIONS AND FUTURE SCOPE

The review covers ten primary studies, mostly outside the support domain, so conclusions about the literature as a whole should be treated cautiously. The framework is unimplemented, and the figures in Table I are not comparable across studies. Future empirical work should: (1) implement the labelling, fine-tuning and quantization stages as a working system; (2) collect a representative support-ticket dataset with out-of-scope examples in the spirit of CLINC150 [9]; (3) fine-tune the student with LoRA or QLoRA on teacher labels [4], [5]; (4) apply INT4 or INT8 quantization with GPTQ or AWQ [6], [7]; (5) compare the quantized student with the frontier teacher on held-out tickets using Table II; (6) evaluate how escalation volume and system accuracy vary with τ; and (7) test behaviour under changing ticket mixes [15]. Only after this work would it be appropriate to report accuracy, latency or cost figures; this paper deliberately stops short of that point.

IX. CONCLUSION

This paper reviewed knowledge distillation, teacher–student architectures, LoRA/QLoRA and post-training quantization, together with intent-classification benchmarks and LLM cascading. Across ten studies, each technique applied individually substantially reduces model size, training cost or inference cost while retaining most of an uncompressed model's accuracy on the tasks studied [1]–[7], [10], [11]. We proposed a conceptual teacher–student–quantization pipeline with confidence-based escalation, presented explicitly as a proposal and not as an evaluated system. The central gap is that these techniques are evaluated separately rather than as one pipeline for production-representative support classification; establishing the combined accuracy, latency, cost and failure modes is the next stage of this research.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

REFERENCES

[1]G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv:1503.02531, 2015.

[2]Y. Kim and A. M. Rush, "Sequence-level knowledge distillation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Austin, TX, USA, 2016, pp. 1317–1327.

[3]C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 1321–1330.

[4]D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, arXiv:1610.02136.

[5]B. Jacob et al., "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 2704–2713.

[6]V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," arXiv:1910.01108, 2019.

[7]X. Jiao et al., "TinyBERT: Distilling BERT for natural language understanding," arXiv:1909.10351, 2019.

[8]S. Larson et al., "An evaluation dataset for intent classification and out-of-scope prediction," in *Proc. EMNLP-IJCNLP*, Hong Kong, China, 2019, pp. 1311–1316.

[9]S. Rabanser, S. Günnemann, and Z. C. Lipton, "Failing loudly: An empirical study of methods for detecting dataset shift," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019.

[10]I. Casanueva, T. Temčinas, D. Gerz, M. Henderson, and I. Vulić, "Efficient intent detection with dual sentence encoders," arXiv:2003.04807, 2020.

[11]E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," arXiv:2106.09685, 2021.

[12]E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, "GPTQ: Accurate post-training quantization for generative pre-trained transformers," arXiv:2210.17323, 2022.

[13]T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," arXiv:2305.14314, 2023.

[14]J. Lin et al., "AWQ: Activation-aware weight quantization for LLM compression and acceleration," arXiv:2306.00978, 2023.

[15]L. Chen, M. Zaharia, and J. Zou, "FrugalGPT: How to use large language models while reducing cost and improving performance," arXiv:2305.05176, 2023.

[16]DeepSeek-AI, D. Guo et al., "DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning," arXiv:2501.12948, 2025.