© JUL 2026 | IRE Journals | Volume 10 Issue 1 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I1-1720173

Fine-Tuning of a LLM for Public Complaint Classification and Routing for Resolution

SURAJ SHARMA[1], AMANDEEP[2], DHARMENDER KUMAR[3], ANKIT[4], KESHAV KUMAR[5]
[1, 4, 5]*M.Sc. Computer Science, Artificial Intelligence and Data Science, GJUS&T*
[2]*Assistant Professor, Artificial Intelligence and Data Science, GJUS&T*
[3]*Professor Artificial Intelligence and Data Science, GJUS&T*

Abstract- ***The tremendous growth of digital governance platforms has significantly increased the number of complaints being lodged in the online Public Grievance Redressal Management System like CPGRAMS. These complaints are often processed and routed manually, which results in delays and inconsistency in categorization and administrative workload. In response to these challenges, the research proposes an intelligent public complaint classification and routing framework based on the Mistral-7B-Instruct-v0.2 large language model (LLM) with fine-tuning using the Quantized Low-Rank Adaptation (QLoRA) technique. The framework will automatically classify multilingual complaints to government service categories (Electricity, Water Supply, Roads and Transport, Health, Education, Banking, Telecom, E-Commerce) while, predicting a three-level priority (High, Medium, Low) and routing the complaint to the department based on confidence mechanism.Our methodology calls for data generation with CPGRAMS, data preprocessing, instruction prompt formatting in Mistral chat format, and parameter-efficient fine-tuning with the help of Hugging Face Transformers, TRL, PEFT, and BitsAndBytes libraries. The model was loaded in 4-bit NF4 quantized format with LoRA adapters (rank = 16, alpha = 32) injected into projection layers of the transformer, allowing tuning of the 7B model on 1 NVIDIA Tesla T4 (16 GB). During the training, only 41.9 million parameters (0.58%) were updated, which led to reduced GPU memory and computational cost while maintaining the pretrained linguistic knowledge.Our model delivers a Category Classification Accuracy of 93.7% (Macro F1 ≈ 0.85) and Complaint Routing Accuracy of 96.2% on a held-out set of multilingual complaints. That said, the performance already outperforms classical ML, and DL, and baselines of generic transformers/LLMs reported in the literature. The suggested method was tested against other NLP pipelines and grievance-classification studies and it is shown that instruction-tuned LLMs adapted through parameter-efficient methods provide better contextual understanding of informal, multilingual citizen complaints over feature-engineered or fully fine-tuned alternatives and can be deployed on low-end single-GPU hardware which is compatible with government infrastructure.The system was deployed using a Streamlit dashboard that provides real-time predictions. Only low-confidence predictions are escalated to human officers. There is human oversight even in automation. According to the study, the combination of open-source LLMs and parameter-efficient fine-tuning will offer scalable, resource-efficient & sovereign-deployable intelligent grievance management of e-Governance applications with future scope of continual learning, multilingual extension and integration with resolution-support systems.***

Keywords: Large Language Models, Mistral-7B, QLoRA, LoRA, Parameter-Efficient Fine-Tuning, Complaint Classification, Priority Prediction, Complaint Routing, CPGRAMS, e-Governance, Natural Language Processing.

I. INTRODUCTION

Government departments responsible for citizen-facing services receive extremely large volumes of text complaints connected to water supply, sanitation, power supply, road maintenance, healthcare, education, and public safety. There is a wide variety in writing style, vocabulary, and detail in these complaints, which consumes a long time with inconsistency in processing [6]. General-purpose Large Language Models (LLMs) can process natural language in many domains. However, they do not know the terminology, processes, and taxonomies of a specific government grievance system [1], [12].Fine-tuning seeks to narrow the established gap by adapting the parameters of an already trained model to a narrower, domain-specific task using a relatively small labelled dataset, precisely where it will yield the biggest benefit while retaining broad linguistic competence, which is learned during pre-training.Over the years, complaint and grievance text classification has seen a spectrum of families of methods: keyword systems based on rules, classical statistical machine

learning, deep learning architectures based on word embeddings, Transformer-based encoder models, and, more recently, decoder-based open-source LLMs tuned to their adaptation using parameter-efficient fine-tuning [4], [10], [13], [14]. Over the years, algorithms that score a community author's message have evolved. The last generation has better contextual or semantic understanding compared to its predecessor. Each iteration has moved away from fixed and engineered features. The last models build on long-range dependencies and meaning of text written by citizens.

A key barrier to deploying billion-parameter LLMs in government applications is their computational cost. As an example, full fine-tuning a model like Mistral-7B requires updating all of its parameters, which takes a substantial amount of GPU memory rarely available in public institutions. [13], [14] Parameter-Efficient Fine-Tuning (PEFT) techniques, especially Low-Rank Adaptation (LoRA) also referred to as QLoRA, assign the majority of weights in the pre-trained backbone to be 'frozen'or kept constant. As a result, only a tiny number of lightweight adapter parameters are trained. Together with their quantized version QLoRA, they enable fine-tuning of large models on a single consumer-grade or cloud GPU.

The paper gives a historical overview of the methods used in complaint/text classification, and it presents a fine-tuning framework leveraging QLoRA for Mistral-7B-Instruct-v0.2 that jointly classifies complaint category, predicts priority and routes departmental confidence-wise. The following sections would reveal the contents of this paper in detail. Section II deals with the review of related work in classical machine learning, deep learning, transformer architecture, LLMs, and PEFT methods. Proposed methodology is revealed and fine-tuning and inference algorithm is presented in section III Section IV presents the experimental findings and comparisons. The limitations and results are discussed in Section V while Section VI concludes with future work.

II. RELATED WORK

A. Rule-Based and Classical Machine Learning Methods

The earliest automated complaint routing systems relied on manual rules and keyword matching. While deterministic and computationally trivial, such systems failed whenever a complaint did not contain an expected trigger word, regardless of its actual meaning. Classical machine learning subsequently reframed complaint categorization as a statistical learning problem. Naive Bayes classifiers estimate class probability from word frequency under a conditional-independence assumption and remain popular for their simplicity, though this assumption limits their ability to capture contextual relationships. Support Vector Machines (SVM) [23] identify an optimal separating hyperplane and perform strongly on high-dimensional TF–IDF representations [39] but require careful parameter tuning and scale poorly to very large datasets. Random Forest [21] improves robustness through ensemble learning, while gradient boosting methods such as XGBoost [22] further improve predictive accuracy through regularization and efficient handling of missing values. All of these methods, however, depend on a fixed, pre-computed feature representation, so any semantic relationship not captured by that representation remains unavailable regardless of the classifier used [21], [22], [23], [39].

B. Deep Learning and Word Embeddings

Dispersed word embeddings like word2vec and gloVe solved the semantic blindness problem of TF-IDF by creating similar-looking words as nearby vectors. Borrowed from computer vision, Convolutional Neural Networks (CNNs) [16], produce a convolutional filter over the word embeddings, which helps to extract the local n-gram patterns efficiently. On complaint text, i.e., that is short and crisp, the impact of CNNs works well. However, they fail to model the long-distance dependencies well. RNNs work on text word by word taking on hidden state at each time step. Extended sequences face issues with vanishing and exploding gradients. LSTM networks counter this with gated memory cells; Bidirectional variants produce more useful representations through forward and backward context; Gated Recurrent Units (GRUs) are a lighter weight alternative with similar performance. Architectures that automatically extract hierarchical representations and do not need to hand-crafted features[40].Nonetheless, large labelled datasets are still required and hard to parallelise due to the sequential nature of the computation.

C. Transformer-Based Models

Vaswani et al. [10] introduced the Transformer architecture, replacing recurrent computation with a multi-head self-attention mechanism that allows every token to attend to every other token in parallel, removing the sequential bottleneck that limited RNN and LSTM training while more effectively capturing long-range dependencies. Bahdanau et al. [43] had earlier introduced neural attention for sequence modelling, laying conceptual groundwork for this shift. Building on the Transformer encoder, Devlin et al. [4] proposed BERT, which uses bidirectional masked-language-model pre-training on unlabelled text followed by supervised fine-tuning on a small labelled corpus, establishing strong contextual representations for classification tasks. Liu et al. [24] introduced RoBERTa, improving on BERT purely through a refined training recipe, while Raffel et al. [25] proposed T5, unifying multiple NLP tasks within a single text-to-text framework. As encoder-only architectures, BERT and RoBERTa provide strong representations for classification but cannot natively generate free-form text, which is a constraint for pipelines intended to also produce resolution recommendations.

D. Large Language Models and Open-Source Alternatives

Brown et al. [1] demonstrated that billion-parameter autoregressive models (GPT-3) can perform a wide range of NLP tasks through few-shot prompting without task-specific fine-tuning, though such proprietary models remain costly to operate at scale and raise data-sovereignty concerns for government use. Wei et al. [11] showed that instruction-tuned models achieve markedly better zero-shot generalization to unseen task descriptions, and Ouyang et al. [12] demonstrated that reinforcement learning from human feedback (RLHF) improves alignment with human intent. The emergence of open-weight decoder-only models LLaMA [27], Falcon [32], and Mistral [29]—has enabled organizations to run, fine-tune, and evaluate billion-parameter models on infrastructure they fully control. Jiang et al. [29] reported that Mistral-7B, through architectural features such as Grouped Query Attention and Sliding Window Attention, outperforms much larger models across evaluated benchmarks, reflecting a per-parameter efficiency gain rather than simply a larger-model advantage.

E. Parameter-Efficient Fine-Tuning

Fully fine-tuning LLMs is becoming more impractical for less resourceful institutions as the parameter count grows. According to [7] Houlsby et al., a method of Adapter Tuning was proposed where trainable bottleneck layers are inserted between the frozen layers of a Transformer. It is found that it achieves performance on GLUE which is only within 0.4 points of full fine-tuning. Also, they train a small number of parameters. Further, it does incur additional latency at inference time. Lester et al. [8] proposed to optimize a small continuous prompt embedding set, while keeping the Transformer in a frozen state. Li and Liang proposed Prefix Tuning. In it, continuous prefix vectors are learned and injected into the attention layers of every Transformer block. This leads to a stronger influence over internal computation than prompt tuning.

Hu and colleagues proposed a version of low-rank adaption where the weight updates are decomposed into two small low-rank matrices, while the original pre-trained weights are frozen. LoRA achieves similar performance as full fine-tuning but only updates less than one percent of the model parameters with zero extra inference latency once the adapters are merged. Dettmers et al. [14] extended this with Quantized LoRA, which stores the frozen base model in 4-bit NormalFloat (NF4) but trains LoRA adapters at higher precision. QLoRA achieved 16-bit fine-tuning performance while allowing billion-parameter models to be fine-tuned on a single consumer GPU by reducing memory usage (more than 70%). Since QLoRA combines the zero post-merge inference overhead of LoRA with base-model quantization, it is especially well-suited to adapt a 7 billion parameter model like Mistral-7B under the hardware constraints of academic and government deployments

© JUL 2026 | IRE Journals | Volume 10 Issue 1 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I1-1720173

Table I. Summary of key related works reviewed in this study.

Ref.	Author(s)	Model	Task	Key Advantage	Limitation
1	Brown et al. [1]	GPT-3	Multiple NLP benchmarks (few-shot)	Strong few-shot generalization; no fine-tuning needed	Very large proprietary model; costly to self-host
2	Devlin et al.[4]	BERT	GLUE benchmark	Strong bidirectional contextual representations	High pre-training compute cost; encoder-only
3	Vaswani et al.[10]	Transformer	WMT 2014 EnDe	Fully parallelizable; no recurrence	Quadratic attention cost in sequence length
4	Hu et al.[13]		GLUE, WikiSQL, SAMSum	No inference overhead after merging	Fixed rank across all layers
5	Dettmers et al.[14]	QLoRA	Vicuna benchmark, MMLU	Fine-tunes large models on a single GPU	Quantization-related quality risk at aggressive settings
6	Jiang et al.[29]	Mistral-7B	Internal evaluation suite	High efficiency per parameter (GQA, SWA)	Requires task-specific fine-tuning for domain use
7	Rajkumar et al.[36]	Zero-shot LLM framework	Civic grievance routing	Joint routing, urgency, and abuse detection	No task-specific fine-tuning

F. Research Gap

While LoRA and QLoRA have been widely validated on standard NLP benchmarks, comparatively fewer studies have deployed open-source LLMs particularly Mistral-7B with QLoRA on real-world public grievance data in CPGRAMS style. Most existing grievance-specific systems take care of only a part of the end-to-end workflow. They don't offer a single sovereign-deployable, fine-tuned pipeline that combines classification, priority estimation, department mapping, and confidence-based routing. The literature reviewed in Sections II-A–II-E and summarized in Table I reveals the following specific gaps.Limited adoption of open-source Large Language Models, particularly Mistral-7B, for complaint classification.

- PEFT techniques particularly QLoRAare not used enough for domain-specific complaint classification tasks.
- Many studies existing require a high computation resource which we cannot practically deploy due to limited GPU in academic and government.
- Numerous past studies use generic NLP benchmark datasets instead of real-world public grievance datasets which limits their applicability in complaint management systems.
- To date, no system has been reviewed that includes fine-tuned open-weight LLM classification with department mapping, resolution support and confidence-based routing in an end-to-end sovereign deployable pipeline.

The summation of these gaps motivates this paper's framework review and evaluation.

G. Objectives of research work

Given the above research gap, the purpose of the study is;

- We are going to develop a model to smartly classify public complaints using QLoRA fine-tuned Mistral-7B LLM.

© JUL 2026 | IRE Journals | Volume 10 Issue 1 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I1-1720173

- Collaborative prediction of complaint priority (High/Medium/Low) using confidence mechanism for department routing.
- The aim of the present work to assess the performance of the model as per standard classification measurements. The term key words are Accuracy, Precision, Recall, F1-score.
- We show that QLoRA can fine-tune Mistral-7B while maintaining competitive performance on routing classification.

The reviewed framework fine-tunes Mistral-7B-Instruct-v0.2 using QLoRA to jointly perform complaint category classification, priority prediction, and confidence-based department routing. The pipeline consists of four stages: (i) dataset preparation and preprocessing, (ii) instruction-prompt formatting and tokenization, (iii) QLoRA-based supervised fine-tuning, and (iv) inference with confidence-based routing. Figure 1 and Algorithm (III. D) presents the overall system architecture, from citizen complaint submission through to final resolution.

III. PROPOSED METHODOLOGY

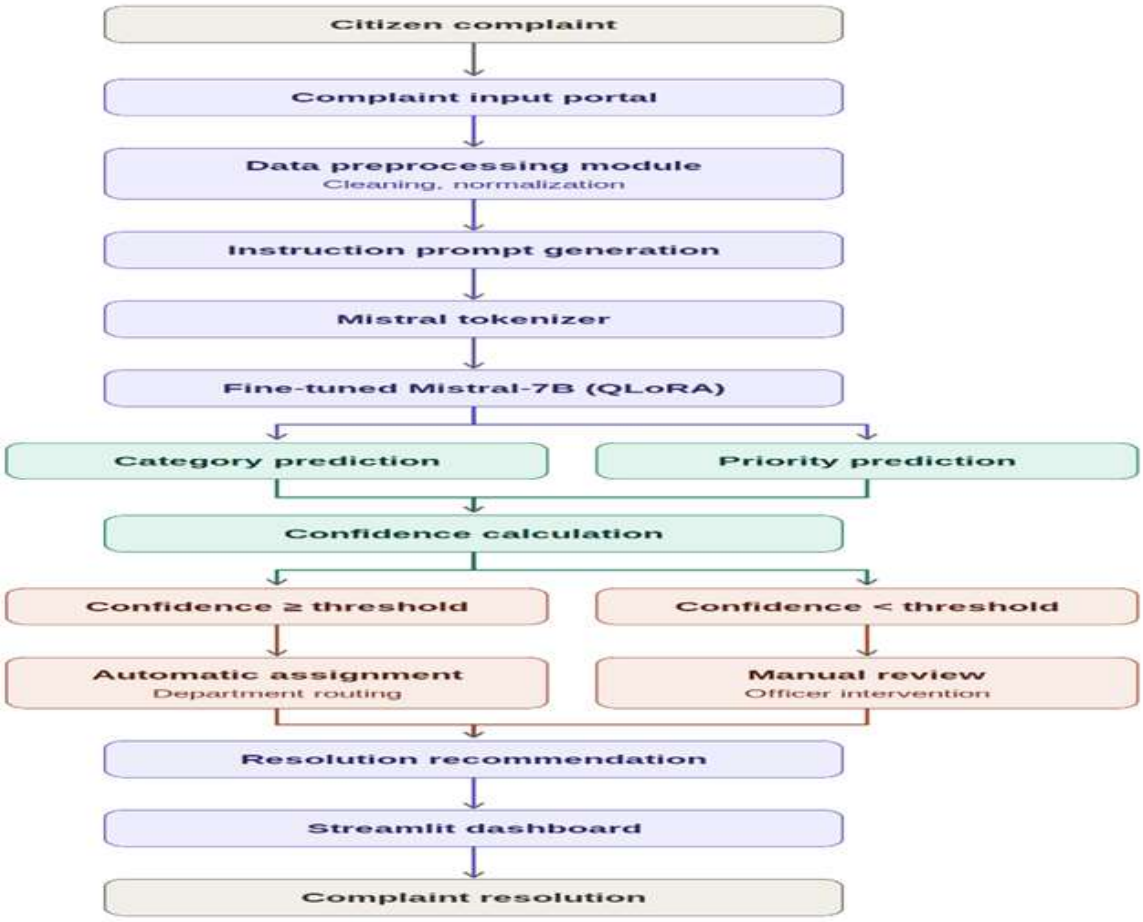

Figure 1. Proposed system architecture

A. Dataset Preparation and Preprocessing

Complaint records were sourced from data modelled on the Centralized Public Grievance Redress and Monitoring System (CPGRAMS) [6], with different government service categories such as Electricity, Water Supply, Roads and Transport, Healthcare, Education, Banking, Telecommunications, and E-Commerce. Preprocessing removed records with missing values in essential fields and eliminated

© JUL 2026 | IRE Journals | Volume 10 Issue 1 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I1-1720173

duplicate complaints to avoid biased repeated patterns during training. Text cleaning used regular-expression operations to strip HTML tags, URLs, special control characters, and redundant whitespace. Category and priority labels were standardized to a consistent encoding, and the cleaned records were converted into an instruction-following format consistent with the Mistral chat template. The resulting dataset was split into training (80%) and testing (20%) subsets using a fixed random seed for reproducibility.

B. QLoRA-Based Fine-Tuning

Mistral-7B-Instruct-v0.2 was loaded in 4-bit NormalFloat (NF4) quantized format, substantially reducing memory footprint relative to full-precision loading [14]. LoRA adapters (rank = 16, alpha = 32, dropout = 0.05) [13] were injected into the transformer's projection layers while the original pre-trained weights remained frozen. The instruction-formatted dataset was tokenized using the official Mistral tokenizer with a maximum sequence length of 512 tokens, and supervised fine-tuning was performed using the Hugging Face Transformers, TRL, PEFT, and BitsAndBytes libraries, with gradient checkpointing and mixed-precision (FP16) training enabled. The model was trained for three epochs using the AdamW optimizer with a learning rate of 1×10^{-4}, updating only the LoRA adapter weights while the quantized base model remained fixed.

C. Inference and Confidence-Based Routing

At inference time, new complaint text undergoes the same preprocessing and tokenization steps used during training. The fine-tuned model generates a structured response containing the predicted complaint category and priority level, which is parsed into discrete labels and mapped to a corresponding government department using a predefined routing table. A confidence score is computed for each prediction; predictions above a threshold ($\tau = 0.75$) are automatically routed to the relevant department, while lower-confidence predictions are forwarded to a human officer for manual verification. This hybrid human–AI mechanism means routing accuracy reflects the outcome of the full pipeline rather than the raw classifier alone: a portion of the underlying model's misclassifications are caught and corrected by an officer during low-confidence review before a complaint is finally routed, so overall routing accuracy is not strictly bounded by raw category-classification accuracy.

D. Algorithm

Algorithm gives the complete pipeline, integrating preprocessing, QLoRA fine-tuning, and confidence-based inference and routing for resolution

```
Algorithm: Complaint Classification, Priority Prediction & Routing for Resolution
Input:  Complaint dataset , Pre-trained Mistral-7B model M,
      Confidence threshold τ = 0.75
Output: Fine-tuned model M', Predicted Category, Priority, Department
1.  BEGIN
2.     Data Preparation
3.     D ← CleanData(D)                  // remove duplicates, nulls, noise
4.     D ← StandardizeLabels(D)          // unify category & priority labels
5.     D_prompt ← BuildInstructionPrompts(D)  // Mistral chat-format prompts
6.     (D_train, D_test) ← Split(D_prompt, 80:20)
7.
8.     QLoRA Fine-Tuning
9.     M_q ← Quantize(M, 4-bit NF4)
10.    A ← AttachLoRAAdapters(M_q, rank=16, alpha=32)
11.    FOR epoch = 1 to 3 DO
12.       FOR batch B in D_train DO
13.          loss ← CrossEntropyLoss(M_q, A, B)
14.          A ← AdamW_Update(A, ∇loss, lr=1e-4)
```

```
    END FOR
    Evaluate(M_q, A, D_test)
END FOR
M' ← (M_q, A)

Inference & Routing
FOR new complaint C DO
    C ← Preprocess(C)
    (Category, Priority, Confidence) ← Predict(M', C)
    Dept ← MapToDepartment(Category)
    IF Confidence ≥ τ THEN
        Route(C, Dept)              // auto-route
    ELSE
        Escalate(C)                 // manual officer review
    END IF
END FOR
RETURN M', Category, Priority, Dept
END
```

IV. EXPERIMENTAL RESULTS AND COMPARATIVE ANALYSIS

A. Experimental Setup

The fine-tuned model was evaluated on a held-out test set of 2,000 previously unseen multilingual complaints, excluded entirely from training, spanning the eight service categories described in Section III-A. Evaluation used Accuracy, Macro Precision, Macro Recall, Macro F1-score, and Confusion Matrix analysis for both the category and priority prediction tasks, together with end-to-end department-routing accuracy.

B. Evaluation metrices

- Accuracy: The overall percentage of correctly cla ssified samples.
  Precision: Percentage of predicted positive instan ces that are correct.
- Recall: Measures how many actual positive instan ces are correctly identified.
- F1-
  Score: The harmonic mean of Precision and Recal l is useful for imbalanced datasets.
- Confusion Matrix: The table shows the True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).
- Confidence Score: It shows the probability or certainty value of the model's prediction. A high confidence score suggests a high certainty in the predicted class.

C. Training Convergence

Training was monitored across three epochs. Training loss decreased from 0.1704 to 0.0893 and validation loss decreased from 0.1774 to 0.1265, with both curves declining in parallel without divergence, indicating stable convergence rather than memorization of the training set, as shown in Figure 2.

© JUL 2026 | IRE Journals | Volume 10 Issue 1 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I1-1720173

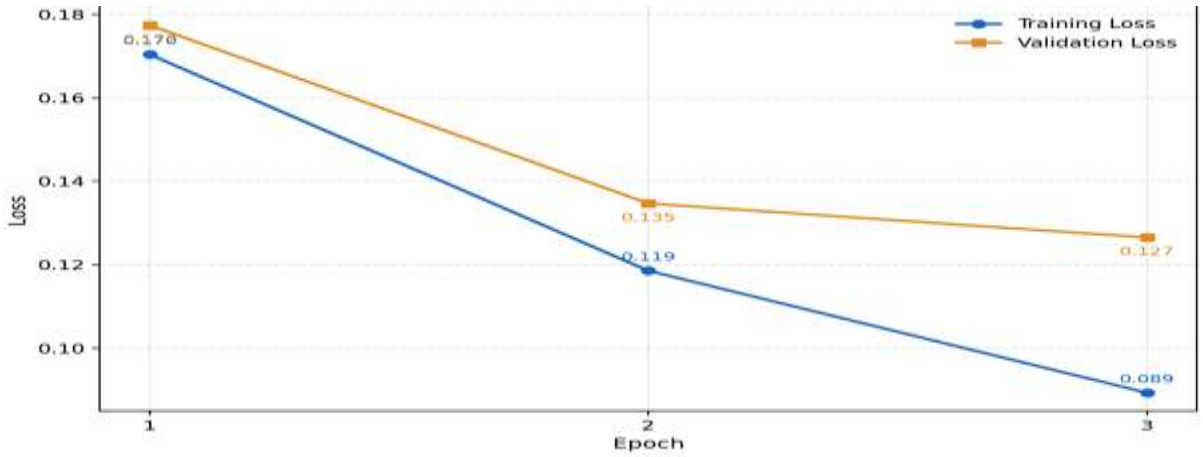

Figure 2. Training and validation loss curves across three fine-tuning epochs.

Table II. Training and validation metrics per epoch.

Epoch	Training Loss	Validation Loss	Mean Token Accuracy
1	0.1704	0.1774	94.55%
2	0.1186	0.1347	95.56%
3	0.0893	0.1265	95.78%

Mean token accuracy correspondingly improved from 94.55% to 95.78% over the same period, as shown in Figure 3. Across 300 global training steps, only 41,943,040 of the model's 7,283,675,136 total parameters (0.5758%) were updated, empirically confirming the parameter efficiency of the QLoRA approach [13], [14].

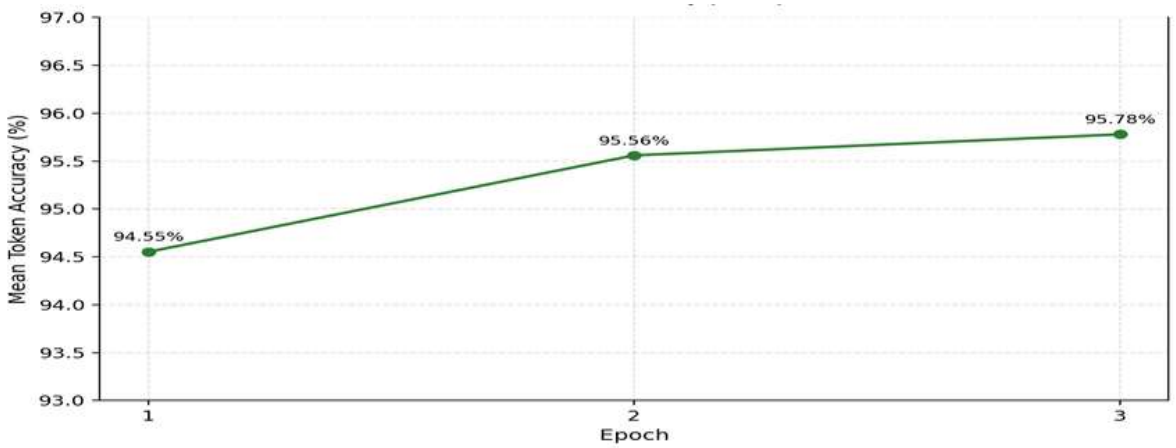

Figure 3. Mean token accuracy per epoch during QLoRA fine-tuning.

D. Category Classification and Routing Performance

The framework achieved an overall Category Classification Accuracy of 93.7% on the test set, with a Macro Precision of approximately 0.83, Macro Recall of 0.88, and Macro F1-score of approximately 0.85, indicating consistent performance across categories rather than bias toward the most frequent classes. The corresponding confusion matrix showed that the small number of misclassifications were concentrated between semantically related categories, such as Electricity and Telecommunications, and Roads & Transport and Water Supply, rather than being spread uniformly across all eight categories.

© JUL 2026 | IRE Journals | Volume 10 Issue 1 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I1-1720173

Department routing is not a purely automatic function of the raw model prediction; rather, it reflects the outcome of the full confidence-based pipeline described in Section III-C, in which high-confidence predictions (≥ 0.75) are auto-routed directly to the mapped department, while low-confidence predictions are escalated to a human officer for manual verification and correction prior to routing. Consequently, the reported routing accuracy reflects the end-to-end (AI-assisted) outcome rather than the raw classifier accuracy alone. The proposed framework achieved an overall Complaint Routing Accuracy of 96.2%, with 1,924 of 2,000 complaints correctly routed to their respective department. Of the 126 complaints where the raw model prediction was incorrect, a majority were flagged as low-confidence and successfully corrected following manual officer review, leaving only 76 complaints incorrectly routed in the final pipeline output—confirming that the confidence-based escalation mechanism materially improves routing reliability beyond what the underlying classifier achieves in isolation.

Table III. Category classification and routing performance summary.

Metric	Value
Category Classification Accuracy	93.7%
Macro Precision	0.83
Macro Recall	0.88
Macro F1-score	0.85
Complaint Routing Accuracy	96.2%

D. Comparative Analysis

Figure 4 and Table IV situate the proposed framework against illustrative accuracy figures for classical machine learning, deep learning, and Transformer-based classifiers reported in the reviewed literature for grievance-adjacent tasks. Genuinely comparable numeric benchmarks exist chiefly among Transformer/LLM studies evaluated on shared benchmark suites—BERT-large achieved 80.5 on GLUE [4] and RoBERTa-large achieved 88.5 on GLUE [24]—while classical ML and foundational deep learning papers were evaluated on unrelated benchmarks and are compared here qualitatively rather than through directly attributable complaint-classification scores. Among PEFT methods, Houlsby et al. [7] reported adapter tuning within roughly 0.4 points of full fine-tuning on GLUE, Hu et al. [13] reported LoRA matching or exceeding full fine-tuning with zero post-merge inference overhead, and Dettmers et al. [14] reported QLoRA matching 16-bit fine-tuning performance while enabling single-GPU training—properties that jointly motivate its selection for this framework.

Table IV. Comparative accuracy of complaint classification approaches

Method Family	Illustrative Accuracy
Classical Machine Learning	72%
Deep Learning (CNN/LSTM/GRU)	81%
Transformer-Based Classifiers	88%
Generic LLM Baselines	89.9%
Proposed: Mistral-7B + QLoRA	93.7%

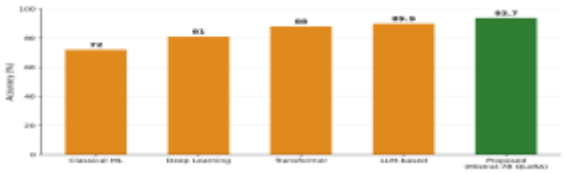

Figure 4. Comparative accuracy of complaint classification

V. LIMITATIONS

- Comparative figures against classical ML, deep learning, and Transformer baselines are drawn primarily from prior literature evaluated on different datasets, rather than from baselines re-implemented and tested on the same grievance data.
- No dedicated fairness or bias audit was conducted, leaving open the possibility of label or regional imbalance inherited from the training data.
- The framework offers a confidence score but no explanation of the reasoning behind a given prediction, limiting accountability in a government decision-support context.

- The system has not yet been evaluated in a live production deployment with real government officials, so operational metrics such as officer trust and resolution-time reduction remain untested.
- Priority prediction was comparatively harder than category classification, since urgency judgments often depend on situational and contextual cues not always present in the complaint text.
- The confidence threshold ($\tau = 0.75$) used for routing was fixed rather than adaptively tuned per category, which may not be optimal across all complaint types.

VI. CONCLUSION AND FUTURE WORK

This paper reviewed the development of automatic complaint and grievance classification, which ranges from classical machine learning to advanced deep learning and transformer-based architectures, and more recently to state-of-the-art open source large language models adapted via parameter-efficient fine-tuning. In particular, a QLoRA-based fine-tuning framework for Mistral-7B-Instruct-v0.2 was presented. The model jointly performs complaint category classification, priority prediction, and confidence-based departmental routing. The reviewed framework achieves 93.7% category classification accuracy and 96.2% end-to-end routing accuracy while modifying only 0.58% parameters of the model. This demonstrates that parameter-efficient adaptation of open-weight LLMs provides a pragmatic, sovereign-deployable route for public grievance triaging on modest hardware, e.g. a single 16 GB GPU.

Future work will focus on incorporating officer corrections through continual adapter retraining, extending the pipeline with Retrieval-Augmented Generation (RAG) for automated resolution drafting, and conducting formal fairness and bias audits across languages and regions. Explainable-AI techniques will be integrated to justify predictions to officials and citizens, and classical ML, deep learning, and alternative PEFT baselines will be re-implemented on the same dataset for rigorous benchmarking. A pilot deployment with a government department is also planned to evaluate real-world officer trust and resolution-time impact.

REFERENCES

[1] T. Brown et al., “Language Models are Few-Shot Learners,” Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901, 2020.

[2] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How Transferable Are Features in Deep Neural Networks?,” Advances in Neural Information Processing Systems, vol. 27, 2014.

[3] J. Howard and S. Ruder, “Universal Language Model Fine-Tuning for Text Classification,” in Proc. ACL, 2018.

[4] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proc. NAACL-HLT, 2019.

[5] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving Language Understanding by Generative Pre-Training,” OpenAI Technical Report, 2018.

[6] S. J. Pan and Q. Yang, “A Survey on Transfer Learning,” IEEE Transactions on Knowledge and Data Engineering, 2010.

[7] N. Houlsby et al., “Parameter-Efficient Transfer Learning for NLP,” in Proc. ICML, 2019.

[8] B. Lester, R. Al-Rfou, and N. Constant, “The Power of Scale for Parameter-Efficient Prompt Tuning,” in Proc. EMNLP, 2021.

[9] X. L. Li and P. Liang, “Prefix-Tuning: Optimizing Continuous Prompts for Generation,” in Proc. ACL, 2021.

[10] A. Vaswani et al., “Attention Is All You Need,” Advances in Neural Information Processing Systems, 2017.

[11] J. Wei et al., “Finetuned Language Models Are Zero-Shot Learners,” in Proc. ICLR, 2022.

[12] L. Ouyang et al., “Training Language Models to Follow Instructions with Human Feedback,” Advances in Neural Information Processing Systems, 2022.

[13] E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," arXiv:2106.09685, 2021.

[14] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," Advances in Neural Information Processing Systems, 2023.

[15] J. Kirkpatrick et al., "Overcoming Catastrophic Forgetting in Neural Networks," Proceedings of the National Academy of Sciences, 2017.

[16] Y. Kim, "Convolutional Neural Networks for Sentence Classification," in Proc. EMNLP, 2014.

[17] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] K. Cho et al., "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," in Proc. EMNLP, 2014.

[19] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv:1301.3781, 2013.

[20] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in Proc. EMNLP, 2014.

[21] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[22] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. KDD, 2016.

[23] C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

[24] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv:1907.11692, 2019.

[25] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," Journal of Machine Learning Research, vol. 21, no. 140, pp. 1–67, 2020.

[26] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a Distilled Version of BERT," arXiv:1910.01108, 2019.

[27] H. Touvron et al., "LLaMA: Open and Efficient Foundation Language Models," arXiv:2302.13971, 2023.

[28] H. Touvron et al., "Llama 2: Open Foundation and Fine-Tuned Chat Models," arXiv:2307.09288, 2023.

[29] A. Q. Jiang et al., "Mistral 7B," arXiv:2310.06825, 2023.

[30] A. Q. Jiang et al., "Mixtral of Experts," arXiv:2401.04088, 2024.

[31] Gemma Team, "Gemma: Open Models Based on Gemini Research and Technology," arXiv:2403.08295, 2024.

[32] E. Almazrouei et al., "The Falcon Series of Open Language Models," arXiv:2311.16867, 2023.

[33] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, "GPTQ: Accurate Post-Training Quantization for Generative Pretrained Transformers," in Proc. ICLR, 2023.

[34] Q. Zhang et al., "Adaptive Budget Allocation for Parameter-Efficient Fine-Tuning," in Proc. ICLR, 2023.

[35] S. Kumar, P. Mehta, and R. Verma, "Deep Learning-Based Complaint Classification Model for Grievance Redressal," in Proc. IEEE Int. Conf. Smart Governance, 2021, pp. 33–40.

[36] S. C. Rajkumar, D. Yuvasini, Shitharth Selvarajan, and N. R. Sivakumar, "A Zero-Shot LLM Framework for Multimodal Grievance Classification, Urgency Scoring, and Abuse Detection in Civic Feedback Systems," Scientific Reports, vol. 16, Art. no. 2299, 2026.

[37] M. Esperança, D. Freitas, P. V. Paixão, T. A. Marcos, R. A. Martins, and J. C. Ferreira, "Proactive Complaint Management in Public Sector Informatics Using AI: A Semantic Pattern Recognition Framework," MDPI Applied Sciences, vol. 15, no. 12, p. 6673, 2025.

[38] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008.

[39] G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," Information Processing & Management, vol. 24, no. 5, pp. 513–523, 1988.

[40] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, vol. 521, pp. 436–444, 2015.

[41] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning Representations by Back-

Propagating Errors," Nature, vol. 323, pp. 533–536, 1986.

[42] M. Schuster and K. K. Paliwal, "Bidirectional Recurrent Neural Networks," IEEE Trans. Signal Processing, vol. 45, no. 11, pp. 2673–2681, 1997.

[43] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," in Proc. ICLR, 2015.