Fine-tuning Small Language Models for Automated Healthcare Prior Authorization:

A Proof-of-Concept Study in Saudi Arabian Insurance Compliance

Dr. Tariq Alturkestani

Insurance Solutions
Tariq@insurance-solutions.co
25/1/2026

Abstract

Prior authorization (PA) in healthcare insurance is a labor-intensive process requiring clinical expertise and regulatory knowledge. This paper presents a proof-of-concept study demonstrating the feasibility of fine-tuning small language models (1.5B-7B parameters) for automated PA decisions under Saudi Arabia's Council of Cooperative Health Insurance (CCHI) regulations. Using synthetic training data (10,000 examples) and Low-Rank Adaptation (LoRA) fine-tuning on consumer Apple Silicon hardware, we achieved 86.7% accuracy through an ensemble approach combining a base model's strength in identifying coverage exclusions (90% DENY accuracy) with a fine-tuned model's ability to recognize approval criteria (90% APPROVE accuracy) and pending documentation requirements (80% PENDING accuracy). While our proof-of-concept used small models with synthetic data, we expect substantially better performance with larger parameter models (70B+) trained on real-world PA decisions. Our results suggest that meaningful automation of prior authorization is achievable, with significant potential for reducing administrative burden while maintaining regulatory compliance.

Keywords: prior authorization, healthcare AI, CCHI compliance, LoRA fine-tuning, small language models, ensemble methods, Saudi Arabia, MLX

1. Introduction

Prior authorization (PA) is a utilization management process requiring healthcare providers to obtain approval from insurance payers before delivering specific medical services. While intended to control costs and ensure appropriate care, PA creates significant administrative burden. The American Medical Association reports that physicians complete an average of 45 prior authorizations per week, with 35% of practices employing staff dedicated solely to PA tasks (AMA, 2023). This administrative overhead contributes to physician burnout, delayed patient care, and increased healthcare costs.

In Saudi Arabia, the Council of Cooperative Health Insurance (CCHI) establishes mandatory coverage requirements for all cooperative health insurance policies. The CCHI Unified Policy specifies complex regulatory rules including: emergency service coverage requirements (the '60-minute rule'), annual benefit limits (SAR 1,000,000), service-specific sub-limits (e.g., SAR 15,000 for maternity, SAR 50,000 for psychiatric services, SAR 15,000 for bariatric surgery), explicit

coverage exclusions (cosmetic procedures, fertility treatments, pre-existing conditions in initial waiting periods), and medical necessity criteria for elective procedures.

Recent advances in large language models (LLMs) have demonstrated remarkable capabilities in healthcare applications, including clinical decision support, medical documentation, and regulatory compliance tasks. However, deploying state-of-the-art models with hundreds of billions of parameters presents significant challenges for healthcare organizations: computational infrastructure costs, data privacy requirements mandating on-premise deployment, and regulatory concerns about cloud-based AI systems processing protected health information.

This paper investigates the feasibility of fine-tuning small language models (SLMs) with 1.5-7 billion parameters for automated prior authorization decision-making. We demonstrate that through careful training data design, parameter-efficient fine-tuning using LoRA, and ensemble methods combining multiple models, meaningful automation can be achieved even with significant resource constraints. While our proof-of-concept uses synthetic data and small models, we discuss expected performance improvements with larger models and real-world training data.

2. Background and Related Work

2.1 CCHI Regulatory Framework

The CCHI, established in 1999, regulates cooperative health insurance in Saudi Arabia under the supervision of the Saudi Arabian Monetary Authority (SAMA). The CCHI Unified Health Insurance Policy mandates minimum coverage standards including: comprehensive coverage up to SAR 1,000,000 annually, emergency services without prior authorization within 60 minutes of presentation, specific benefit sub-limits for categories including maternity, psychiatric care, and surgical procedures, and explicit exclusions for cosmetic procedures, fertility treatments, and experimental therapies.

2.2 Parameter-Efficient Fine-Tuning

Low-Rank Adaptation (LoRA), introduced by Hu et al. (2021), enables efficient fine-tuning by training low-rank decomposition matrices rather than full model weights. This approach reduces trainable parameters by 10,000x while maintaining competitive performance on downstream tasks. QLoRA (Dettmers et al., 2023) extends this to quantized models, enabling fine-tuning of larger models on consumer hardware. Our work uses LoRA with 4-bit quantized models via Apple's MLX framework.

2.3 Small Language Models

Recent research demonstrates that carefully trained small models can approach larger model performance on specific tasks. The Qwen2.5 family (Alibaba, 2024) provides instruction-tuned models from 0.5B to 72B parameters with strong multilingual capabilities including Arabic. MLX (Apple, 2024) enables efficient inference and training on Apple Silicon, making SLM deployment accessible on consumer hardware.

3. Methodology

3.1 Decision Framework

We defined three decision classes for prior authorization requests: APPROVE (service is covered, medically necessary, and within benefit limits), DENY (service is explicitly excluded, not medically necessary, or exceeds benefit limits), and PENDING (request requires additional documentation or clinical information before determination). This three-class framework reflects real-world PA workflows where many requests require follow-up rather than immediate approval or denial.

3.2 Synthetic Training Data Generation

We developed a programmatic data generator producing 10,000 prior authorization scenarios (9,500 training, 500 validation) with controlled class distribution: APPROVE (45%), DENY (30%), and PENDING (25%). Each example includes structured patient demographics, procedure and diagnosis codes (CPT/ICD-10), clinical context (symptoms, duration, severity, prior treatments), and coverage status (annual limits, specific benefit utilization).

Scenario generators were designed to capture specific regulatory patterns: emergency presentations (unconditional approval), imaging requests with/without conservative treatment documentation, bariatric surgery with BMI thresholds and program completion requirements, cosmetic procedure exclusions, fertility treatment exclusions, benefit limit exhaustion, and seven distinct PENDING scenarios including missing weight management documentation, absent physical therapy notes, required specialist evaluations, incomplete clinical information, and missing psychological clearances.

3.3 Model Selection and Training

We selected Qwen2.5-Instruct models in two sizes: 1.5B and 7B parameters, using 4-bit quantized versions (mlx-community variants) for memory efficiency. Training was performed using MLX-LM's LoRA implementation on an Apple MacBook Pro with M3 Pro chip (18GB unified memory).

Table 1: Training Configuration

Parameter	1.5B Model	7B Model
Base Model	Qwen2.5-1.5B-Instruct-4bit	Qwen2.5-7B-Instruct-4bit
LoRA Layers	16	16
Learning Rate	5e-5	2e-5
Training Iterations	2,000	2,000
Batch Size	1	1
Training Examples	9,500	9,500
Training Time	~100 minutes	~195 minutes
Peak Memory	2.6 GB	7.3 GB
Final Validation Loss	0.085	0.086
Trainable Parameters	5.3M (0.34%)	11.5M (0.15%)

3.4 Ensemble Architecture

Initial evaluation revealed complementary model strengths. The base 7B model (without fine-tuning) excelled at identifying coverage exclusions and denials (90% DENY accuracy) due to its pre-training knowledge of medical terminology and insurance concepts, but could not recognize PENDING cases (0% accuracy) as this requires understanding our specific documentation requirements. The fine-tuned 7B model learned to detect PENDING cases (80% accuracy) and approval criteria (90% APPROVE accuracy) but became overly permissive, approving cases that should be denied (only 30% DENY accuracy).

We developed an ensemble combining both models with the following decision logic: (1) If the fine-tuned model predicts PENDING, use PENDING (the fine-tuned model is the only component capable of detecting documentation deficiencies); (2) If the base model predicts DENY, trust the denial regardless of the fine-tuned model's prediction (leveraging its 90% precision on exclusions); (3) Otherwise, use the fine-tuned model's prediction for approvals. This simple rule-based ensemble captures the best performance of each component.

4. Results

4.1 Evaluation Dataset

We constructed a balanced evaluation set of 30 test cases (10 APPROVE, 10 DENY, 10 PENDING) designed to test specific regulatory scenarios. APPROVE cases included emergency presentations (MI, stroke), imaging after documented conservative treatment, bariatric surgery with complete documentation, routine maternity care, and psychiatric evaluations. DENY cases included cosmetic procedures (rhinoplasty, liposuction, breast augmentation), fertility treatments (IVF), excluded conditions (acne, hair loss), and benefit limit exhaustion. PENDING cases included missing weight management documentation, absent physical therapy notes, incomplete specialist evaluations, and missing psychological clearances.

4.2 Model Performance

Table 2: Model Performance Comparison (n=30)

Model	Overall	APPROVE	DENY	PENDING
1.5B Base	0.0%	0.0%	0.0%	0.0%
1.5B + Fine-tuned	33.3%	60.0%	20.0%	20.0%
7B Base	53.3%	70.0%	90.0%	0.0%
7B + Fine-tuned	66.7%	90.0%	30.0%	80.0%
Ensemble	86.7%	90.0%	90.0%	80.0%

The ensemble achieved 86.7% overall accuracy, representing a +33.3 percentage point improvement over the base model alone and +20.0 points over the fine-tuned model alone. Critically, the ensemble captured the best performance of each component across all three decision classes: 90% APPROVE accuracy (matching fine-tuned), 90% DENY accuracy (matching base), and 80% PENDING accuracy (from fine-tuned).

4.3 Error Analysis

Four errors occurred in the 30-case evaluation. One APPROVE case (bariatric surgery with complete documentation) was incorrectly denied due to the base model over-triggering on bariatric requests. One DENY case (psychiatric services with exhausted benefit limit) was incorrectly approved because neither model learned the benefit-limit-exhaustion pattern well from synthetic data. Two PENDING cases were misclassified when the fine-tuned model failed to produce a parseable decision, causing fallback to base model predictions.

5. Discussion

5.1 Limitations and Expected Improvements

This proof-of-concept study has significant limitations that affect generalizability. **We expect substantially better performance in production deployments for the following reasons:**

- **Model Size:** Our 7B parameter models are small compared to state-of-the-art systems. Models with 70B+ parameters would better capture regulatory nuances, reduce parsing failures, and provide more consistent outputs. Inference costs for larger models are increasingly manageable with quantization and efficient serving frameworks.
- **Training Data:** Our synthetic training data, while carefully designed, cannot capture the full variability of real-world prior authorization requests. Training on actual PA decisions (with appropriate de-identification) would expose the model to genuine edge cases, regional variations, and complex multi-condition scenarios.
- **Evaluation Scale:** Our 30-case test set, while carefully balanced, cannot represent the true distribution of PA requests. Production evaluation would require thousands of cases across diverse specialties and patient populations.
- **RAG Integration:** Retrieval-augmented generation incorporating policy documents, formularies, and clinical guidelines at inference time would improve regulatory accuracy without requiring all knowledge to be captured in model weights.

5.2 Deployment Considerations

For production deployment, we recommend: confidence thresholds routing uncertain cases to human review (the ensemble naturally provides confidence signals when component models disagree); comprehensive audit logging for regulatory compliance and continuous improvement; regular model retraining as regulations evolve; and API integration with existing claims management systems. The ensemble architecture provides interpretable decision paths useful for regulatory review.

5.3 Broader Implications

This work demonstrates that healthcare administrative AI is accessible to organizations without massive computational resources. The techniques demonstrated—synthetic data generation, LoRA fine-tuning, and ensemble methods—provide a template for similar applications in other regulatory domains. As model efficiency continues to improve, the gap between proof-of-concept and production-ready systems will narrow.

6. Conclusion

This proof-of-concept study demonstrates that small language models can be fine-tuned for healthcare prior authorization decisions with meaningful accuracy. Our ensemble approach achieved 86.7% accuracy on a challenging test set requiring differentiation between approvals, denials, and pending documentation requests under Saudi Arabia's CCHI regulations.

While production deployment would require larger models and real-world training data—where we expect substantially improved performance—our results establish the feasibility of automated prior authorization. The demonstrated techniques provide a foundation for healthcare organizations seeking to reduce administrative burden while maintaining regulatory compliance.

Future work should focus on validation with real prior authorization data, integration with clinical decision support systems, extension to other regulatory frameworks, and longitudinal studies of human-AI collaboration in PA workflows.

References

American Medical Association. (2023). 2023 AMA prior authorization physician survey. Retrieved from https://www.ama-assn.org

Apple Inc. (2024). MLX: An array framework for Apple silicon. https://github.com/ml-explore/mlx

Council of Cooperative Health Insurance. (2023). Unified Health Insurance Policy. Saudi Arabia.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. arXiv preprint arXiv:2305.14314.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

Qwen Team. (2024). Qwen2.5: A party of foundation models. Alibaba Group. https://qwenlm.github.io