The Future of Evidence Synthesis and Generation in HEOR: A Secure Framework for Domain Specific Adaptation and Fine-tuning Of Large Language Models

MSR37

Rajdeep Kaur, Mrinal Mayank, Ruhi, Shubhram Pandey, Gagandeep Kaur, Barinder Singh
Pharmacoevidence , Mohali, India

Pharmaco Evidence
Science-led HEOR & Tech-driven AI

INTRODUCTION

- Large language models (LLM) can support writing-intensive healthcare and HEOR tasks, but general-purpose models often lack the specific tone, structure, terminology, and quality required for evidence-generation documents
- This proof-of-concept study aimed to validate a secure, cloud-based framework for fine-tuning LLM to support HEOR workflows including the development of dossiers, reports, protocols, and health technology assessment (HTA) related content

METHODS

- A modular proof-of-concept framework was developed to support domain adaptation and fine-tuning of LLM for HEOR evidence-generation workflows within a secure, cloud-based AWS SageMaker AI environment. The framework comprised three phases:
- Phase 1 Data preparation: Domain-specific HEOR source documents, including published literature, HTA guidelines, dossiers, study reports, and protocols, were collected, cleaned, normalized, chunked, tokenized, and securely stored. The curated dataset was split into 80% training data and 20% testing data, as shown in **Figure 1**
- Phase 2 Model fine-tuning: An open-source LLM, LLaMA-2-7B, was fine-tuned using multiple adaptation strategies, including QLoRA, LoRA, prefix tuning, and adapter tuning. Based on performance and feasibility, selected strategies were further applied to Gemma 4 E4B for HEOR writing tasks
- Phase 3 Output evaluation and validation: Outputs generated from the testing dataset were assessed for writing quality, tone, scientific structure, style adaptation, and HEOR relevance. Subject matter experts (SMEs) reviewed the outputs to confirm alignment with HTA and regulatory-style evidence communication. Key value outcomes are summarized in **Figure 2**

RESULTS

- Fine-tuning open-source LLMs with domain-specific HEOR content demonstrated feasibility for generating submission-quality draft outputs across dossier, report, and protocol writing tasks
- Model performance improved after fine-tuning, with reduced training and validation loss and stable convergence, supporting successful domain adaptation
- LoRA produced the strongest output quality, with improved tonal consistency, clearer scientific structure, and better HTA alignment
- QLoRA supported rapid prototyping and efficient model adaptation, making it useful for early-stage experimentation
- Fine-tuned Gemma 4 E4B demonstrated stronger HEOR writing performance than LLaMA-2-7B based on SME evaluation across dossier, report, and protocol prompts

Figure 1. Workflow for fine-tuning a domain-adapted AI model for HEOR evidence-generation tasks.

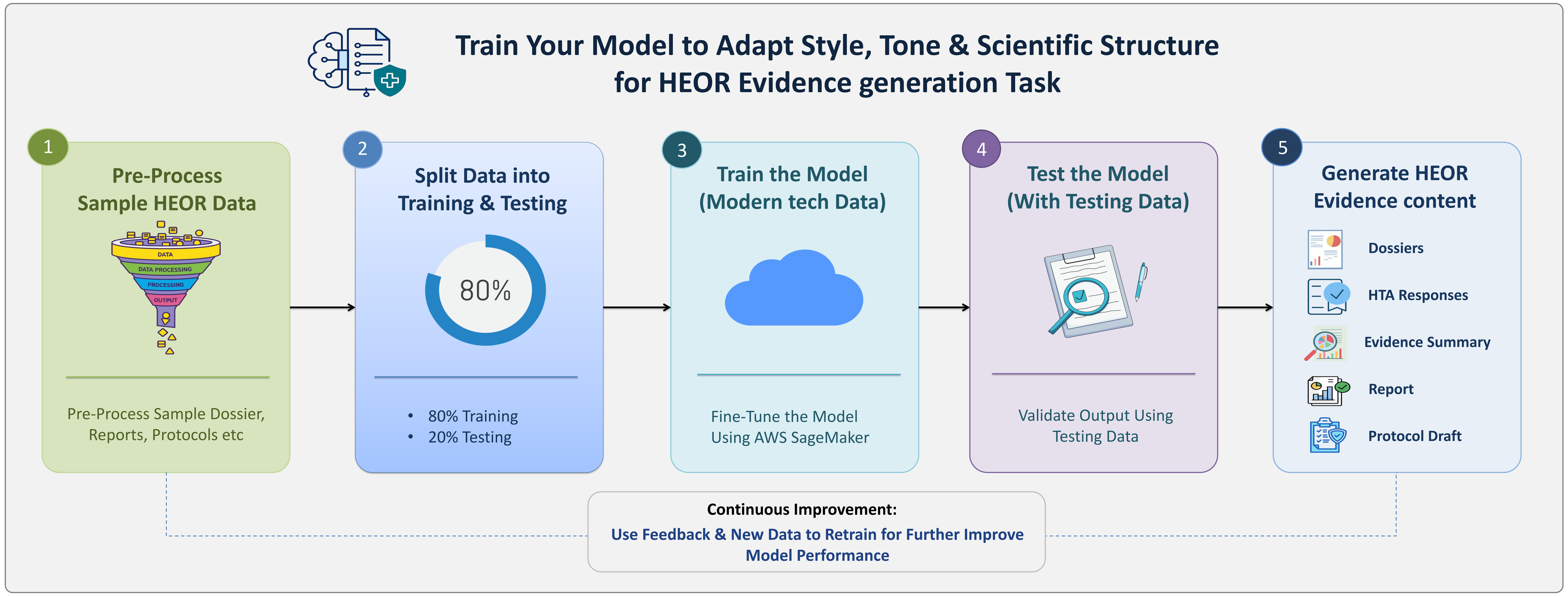
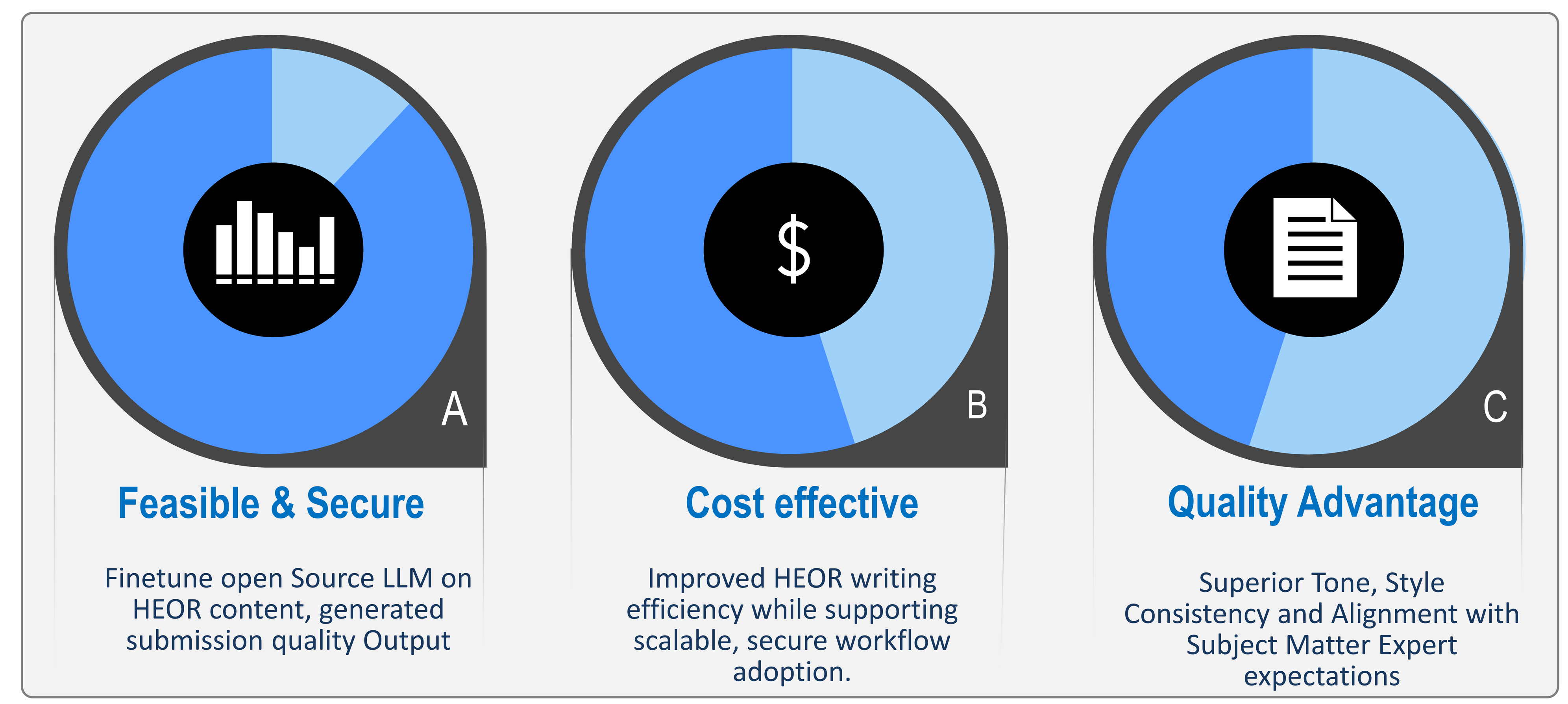


Figure 2. From Generic LLM Output to Fine-tuned HEOR specific Model Evidence Generation

KEY MODEL ATTRIBUTES	GENERIC LLM	FINE-TUNED HEOR LLM
LANGUAGE & TONE	General purpose wording	HEOR-specific terminology and scientific terms
OUTPUT STRUCTURE	Inconsistent Format	Dossier, report & protocol ready structure
CONTEXT FIT	Limited heor context alignment	Better relevance to HEOR specific task
QUALITY READINESS	Needs more manual editing	More consistent, reliable & review ready
WORKFLOW VALUE	Lower efficieny for speacilized task	Faster scalable evidence generation

Figure 3. Impact of secure LLM fine-tuning on HEOR evidence-generation workflows



CONCLUSIONS

- This study demonstrates the feasibility of a secure, cloud-based framework for fine-tuning large language models, presenting a novel application within HEOR evidence-generation workflows
- Fine-tuned open-source LLMs adapted style, tone, and scientific structure for HEOR evidence generation tasks across dossiers, reports, and protocols
- LoRA-adapted Gemma 4 E4B showed stronger tone consistency and HTA alignment than LLaMA-2-7B, supporting secure and scalable HEOR AI adoption
- Future work will focus on enterprise-scale deployment and integration of domain-adapted models into multi-agent HEOR workflows

ISPOR, May 17-20, 2026, Philadelphia, PA, United States