Drafted Use Cases for Fine-Tuning Large Language Models in Enterprise Healthcare Applications:

As the adoption of Large Language Models (LLMs) in healthcare continues to expand, enterprises require specialized fine-tuning approaches to enhance accuracy, reliability, and trustworthiness in medical AI applications. This document outlines key use cases where Supervised Fine-Tuning (SFT) and Low-Rank Adaptation (LoRA) can be applied to optimize LLMs for enterprise healthcare tasks. Each use case highlights the problem statement, datasets, training strategy, and evaluation metrics, ensuring that LLMs can effectively reduce hallucinations, detect misinformation, analyze sentiment, and power interactive AI healthcare assistants.

The following sections detail four core use cases demonstrating how LLMs can be fine-tuned and adapted for real-world medical applications.

---

1. Hallucination Detection in Biomedical QA Systems

Objective:
Fine-tune LLMs to **reduce and detect hallucinations** in medical QA.

Datasets:

- **PubMedQA** – Biomedical QA from PubMed abstracts.
- **MedQA (USMLE)** – Medical board exam Q&A.
- **MedicineQuAD** – Medication-focused QA dataset.

Task Breakdown:

1. **Train models** on **gold-standard medical QA datasets**.
2. **Fine-tune using SFT and LoRA** to enhance factual consistency.
3. **Develop hallucination detection mechanisms**, including:
    - **Fact-checking models** (e.g., retrieval-augmented generation).
    - **Confidence scoring system** to assign uncertainty ratings.
    - **External knowledge retrieval** (PubMed, clinical guidelines).
4. **Evaluate hallucination rates** using human experts and automated fact-checking.

Models to Use:

- **Llama-2 7B, Llama-2 13B, Falcon 7B, Mistral 7B, Mistral 8×7B, GPT-3.5-Turbo, Gemini Pro 1.0, Gemini 1.5 Pro, Gemini 1.5 Flash, Gemini 1.5 Flash 8B, Gemini 2.0 Flash Experimental, Flan-T5 Large, Flan-T5 XL**

Evaluation Metrics:

- **Faithfulness Score** (Factual correctness % based on retrieval).
- **ROUGE/BERTScore** (Semantic similarity to benchmark answers).
- **Hallucination Rate** (% of unsupported/generated content).

2. Medical Misinformation Detection in LLM Responses

Objective:
Fine-tune LLMs to **detect misleading or incorrect medical advice**.

Datasets:

- **HealthFact** – Medical misinformation dataset.
- **SciFact** – Scientific fact-checking dataset.
- **COVID-19 Fake News Dataset** – Misinformation and fact-checked claims.

Task Breakdown:

1. **Train models** on **fact-checked medical Q&A datasets**.
2. **Fine-tune for misinformation classification** (true, false, misleading).
3. **Develop a confidence scoring model** for **reliability measurement**.
4. **Evaluate models' ability to detect misleading medical claims**.

Models to Use:

- **Llama-2 7B, Llama-2 13B, Falcon 7B, Mistral 7B, Mistral 8×7B, GPT-3.5-Turbo, Gemini Pro 1.0, Gemini 1.5 Pro, Gemini 1.5 Flash, Gemini 1.5 Flash 8B, Gemini 2.0 Flash Experimental, Flan-T5 Large, Flan-T5 XL**

Evaluation Metrics:

- **Accuracy** in **misinformation classification**.
- **False Positive Rate** (how often true statements are wrongly flagged).
- **Trustworthiness Score** (alignment with verified medical sources).

---

3. Personalized Healthcare QA System (Chatbot)

Objective:
Develop a **healthcare chatbot** that **answers patient inquiries on medications, symptoms, and treatments** while **reducing hallucinations**.

Datasets:

- **MedicineQuAD** – Medication-focused QA dataset.
- **MedQA (USMLE)** – Medical board exam Q&A.
- **PubMedQA** – Biomedical QA dataset.

Task Breakdown:

1. **Train models** on **medical QA datasets**.
2. **Fine-tune for patient-friendly responses** (simplified, clear language).

3. **Ensure context-aware, regulatory-compliant answers**:
    - Implement **FDA/TGA guideline alignment**.
    - Develop **a retrieval-based validation** for generated answers.
4. **Deploy as a chatbot interface** (React-based UI + API integration).
5. **Implement real-time fact-checking**:
    - **Confidence score visualization** (green = high confidence, yellow = medium, red = low confidence).
    - Integration with **PubMed and trusted medical sources**.

Models to Use:

- **Llama-2 7B, Llama-2 13B, Falcon 7B, Mistral 7B, Mistral 8×7B, GPT-3.5-Turbo, Gemini Pro 1.0, Gemini 1.5 Pro, Gemini 1.5 Flash, Gemini 1.5 Flash 8B, Gemini 2.0 Flash Experimental, Flan-T5 Large, Flan-T5 XL**

Evaluation Metrics:

- **Medical factual accuracy** (expert validation).
- **Answer clarity and patient-friendliness** (user study assessment).
- **Trustworthiness Score** (alignment with FDA/TGA guidelines).

---

4 - Sentiment Analysis in Healthcare

Objective:

Fine-tune LLMs to analyze **patient and doctor sentiment** from medical records, reviews, and patient-doctor conversations.

Datasets:

- **Patient Experience Dataset (NHS, MIMIC-III, or MIMIC-IV clinical notes)**.
- **i2b2 Sentiment Analysis Dataset** (annotated electronic health records).
- **HealthReview (Medical product & service reviews)**.
- **Twitter Health Sentiment Dataset** (health-related tweets with sentiment labels).

Task Breakdown:

1. **Fine-tune models** for **medical sentiment analysis**:
    - **Detect emotions** (positive, neutral, negative).
    - **Classify sentiment intensity** (e.g., very negative to very positive).
2. **Develop a sentiment classifier** that:
    - Identifies **patient dissatisfaction in healthcare settings**.

    - Detects **negative sentiment towards medications/treatments**.
    - Analyzes **doctor-patient communication** for tone assessment.
3. **Evaluate accuracy** across different healthcare contexts.

Models to Use:

- **All specified models**.

Evaluation Metrics:

- **F1 Score** (classification performance).
- **Precision-Recall** (handling class imbalance).
- **False Positive/Negative Rate** (error analysis).
- **Sentiment Consistency** (repeatability across different models).

Implementation Strategy

1. Fine-Tuning Approach

- Use **Supervised Fine-Tuning (SFT)** and **LoRA** for efficient training.
- Train each model separately, evaluating before and after fine-tuning.

2. Dataset Preprocessing

- **Tokenization & data augmentation** for question-answer pairs.
- **Alignment filtering** to remove conflicting data.

3. Model Training Pipeline

- **Step 1**: Train models with **pre-existing medical QA datasets**.
- **Step 2**: Implement **LoRA-based fine-tuning** to inject domain-specific expertise.
- **Step 3**: Evaluate **hallucination/misinformation rates** pre- and post-fine-tuning.
- **Step 4**: Deploy **chatbot interface** for real-time healthcare QA.

4. Chatbot Deployment for Personalized Healthcare QA

- **Frontend**: React-based UI.
- **Backend**: Flask/FastAPI + model hosting on cloud GPUs.
- **Integration**: Fact-checking module with **real-time verification**.
- **Confidence Display**: Answers are color-coded based on accuracy scores.

Expected Outcomes

1. **Hallucination Reduction**:
    - Models will **generate factually accurate answers** and **self-identify hallucinations**.
    - Real-time **confidence scores** will warn users about potentially misleading responses.
2. **Misinformation Detection**:
    - LLMs will classify medical claims as **true, false, or misleading**.
    - **Trustworthiness scores** will indicate reliability.
3. **AI-Powered Healthcare Chatbot**:
    - **Patient-friendly responses** with **regulatory compliance**.
    - **Real-time fact-checking** for **accurate health guidance**.

Data Selection and Training Strategy for each Task

1. Hallucination Detection in Biomedical QA Systems

Datasets Suggested:

- **PubMedQA** – Biomedical QA dataset from PubMed abstracts.
- **MedQA (USMLE)** – Medical board exam Q&A dataset.
- **MedicineQuAD** – Medication-focused QA dataset (from TGA data).

Training Strategy:

Use all three datasets, but with different roles:

- **Fine-tuning:** MedicineQuAD + MedQA (to improve factual accuracy).
- **Evaluation:** PubMedQA (to test hallucination rates).

Why?

- MedicineQuAD ensures **medication-focused accuracy**.
- MedQA helps **general medical QA accuracy**.
- PubMedQA provides **external validation** for hallucination detection.

---

2. Medical Misinformation Detection in LLM Responses

Datasets Suggested & Their Role in Training

Dataset	Purpose	Training or Evaluation?
HealthFact	General medical misinformation detection	Training + Evaluation
SciFact	Fact-checking scientific claims	Training + Evaluation
COVID-19 Fake News Dataset	Identifying pandemic-related misinformation	Training + Evaluation

Training Strategy

Use all three datasets, but **train models in stages**:

1. **Pre-train on HealthFact + SciFact** → General fact-checking ability.
2. **Fine-tune on COVID-19 Fake News Dataset** → Improve misinformation classification.

Model Training Pipeline

Step 1: Data Preprocessing

- Convert datasets into **true/false/misleading** classification format.
- **Tokenize** using **Hugging Face's tokenizer** for each model.

- **Balance dataset** to prevent overfitting on "misleading" labels.

Step 2: Fine-Tuning Process

- **Supervised Fine-Tuning (SFT)**: Train models on labeled misinformation.
- **LoRA Adaptation**: Efficient fine-tuning to avoid memory issues.

Confidence Scoring System

- Implement a **"Trust Score"** that **assigns a confidence level** to each response.
- **Use Retrieval-Augmented Generation (RAG)**:
    - Retrieve **scientific references** from trusted medical sources.
    - Compare LLM-generated answers to retrieved information.
    - Flag **low-confidence responses** for review.

How it Works

- If an LLM says:
  "Vitamin C prevents COVID-19."
- The system **retrieves** PubMed articles.
- If **conflicting evidence** is found, **response confidence drops**.

---

3. Personalized Healthcare Chatbot

Datasets Suggested:

- **MedicineQuAD** – Medication-focused QA dataset.
- **MedQA (USMLE)** – Medical board exam Q&A dataset.
- **PubMedQA** – Biomedical QA dataset.
- **MedDialog** – Doctor-patient conversation dataset.

Training Strategy:

Use different datasets for different chatbot functions:

- **For general medical QA** → Use **MedQA + PubMedQA**.
- **For medication-specific queries** → Use **MedicineQuAD**.
- **For conversational style training** → Use **MedDialog**.

Why?

- **MedDialog** fine-tunes the chatbot's conversational **style**.
- **MedQA and PubMedQA** ensure **medical factual accuracy**.
- **MedicineQuAD** specializes in **pharmaceutical knowledge**.

➞ Recommendation:
Use all datasets, but weight them **differently**:

- **Core Training**: MedQA + MedicineQuAD.
- **Evaluation**: PubMedQA + MedDialog.

---

4. Sentiment Analysis in Healthcare

Datasets Suggested:

- **Patient Experience Dataset (MIMIC-III, NHS surveys)** – Real patient reviews.
- **i2b2 Sentiment Analysis Dataset** – Annotated sentiment labels from EHRs.
- **HealthReview (Medical product & service reviews)** – Healthcare product sentiment.
- **Twitter Health Sentiment Dataset** – Sentiment from health-related tweets.

Training Strategy:

Pick datasets based on the use case:

- **For patient feedback analysis** → Use **MIMIC-III + NHS surveys**.
- **For doctor-patient interactions** → Use **i2b2 dataset**.
- **For medical product reviews** → Use **HealthReview dataset**.
- **For social media health trends** → Use **Twitter Health Sentiment Dataset**.

Why?

- Training on **all datasets** might reduce specialization.
- Using a **combination** of domain-specific datasets is best.

➞ Recommendation:
Choose 2-3 datasets depending on **whether you are analyzing clinical sentiment or patient experience**.

Examples for Each Task in Fine-Tuning Large Language Models for Enterprise Healthcare Applications

Below are **real-world examples** of how each use case could be implemented in a healthcare setting using fine-tuned Large Language Models (LLMs).

---

1. Hallucination Detection in Biomedical QA Systems

Example Scenario:
A **hospital's AI-powered knowledge assistant** is trained to provide clinicians with evidence-based responses to medical inquiries. However, the model **hallucinates** information, leading to incorrect advice.

Use Case:
A doctor asks the AI:
❌ **Incorrect (Hallucinated) Response:**
"Aspirin is recommended for pregnant women to prevent high blood pressure."

✅ **Correct (Fine-Tuned) Response:**
"Aspirin use during pregnancy is sometimes recommended at low doses to prevent preeclampsia, but only under a doctor's supervision. According to the American College of Obstetricians and Gynecologists (ACOG), low-dose aspirin is advised for high-risk patients."

How It Works:

- The fine-tuned model **retrieves clinical guidelines from sources like PubMed or the FDA**.
- A **confidence score is assigned**, flagging high-risk hallucinations.
- The system **cross-references trusted sources** before responding.

---

2. Medical Misinformation Detection in LLM Responses

Example Scenario:
A healthcare chatbot used by **patients and medical students** provides answers about **disease prevention and treatment**. Some responses include **misleading or false claims**.

Use Case:
A patient asks:
"Can garlic cure high blood pressure?"

❌ **Incorrect (Misinformation from Baseline LLM):**
"Yes, garlic can completely cure high blood pressure, so you don't need medications."

✅ **Correct (Fine-Tuned Model Response with Misinformation Detection):**
*"Garlic may help lower blood pressure slightly due to its allicin content, but it is **not a replacement for prescribed medication**. Clinical studies suggest that while it can support heart health, patients with hypertension should follow doctor-recommended treatments."*

How It Works:

- The fine-tuned LLM is trained on **HealthFact and SciFact datasets** to **classify responses as True, False, or Misleading**.

- A **Trust Score is generated**, flagging potential misinformation.
- Retrieval-Augmented Generation (**RAG**) checks the model's claims against **verified sources like WHO or PubMed**.

---

3. Personalized Healthcare QA System (Chatbot)

Example Scenario:
A **pharmacy chain launches an AI chatbot** that provides **medication guidance** to customers. Patients ask about potential side effects, interactions, and usage instructions.

Use Case:
A customer asks:
"Can I take ibuprofen if I'm already on aspirin for heart disease?"

❌ **Incorrect (Generic Chatbot Response):**
"Yes, you can take both without any concerns."

✅ **Correct (Fine-Tuned Chatbot Response):**
*"Caution is advised. **Ibuprofen may interfere with aspirin's blood-thinning effects**, reducing its cardiovascular benefits. The FDA recommends consulting a healthcare provider before combining these medications."*

How It Works:

- The chatbot is fine-tuned on **MedicineQuAD and MedDialog datasets** for patient-friendly medical responses.
- **Context-aware retrieval** ensures answers align with **TGA, FDA, or WHO guidelines**.
- **Confidence levels** are displayed in color codes (**Green = High Confidence, Yellow = Moderate, Red = Low**).

---

4. Sentiment Analysis in Healthcare

Example Scenario:
A hospital analyzes **patient reviews** from **Google, NHS surveys, and social media** to identify areas for improvement in **patient care**.

Use Case:
A patient review states:
"The nurses were great, but I had to wait over six hours in the ER before seeing a doctor. It was a frustrating experience."

❌ **Incorrect Sentiment Analysis from Untrained Model:**

- **Overall Sentiment:** Positive 😀

✅ **Correct (Fine-Tuned Sentiment Model):**

- **Overall Sentiment:** Negative 😡
- **Sentiment Breakdown:**

- **Staff Friendliness:** Positive ✅
- **Wait Time Experience:** Negative ❌
- **Overall Experience:** Negative ❌

How It Works:

- The **i2b2 dataset** helps classify **doctor-patient interactions**.
- The **MIMIC-III dataset** enables **analysis of hospital reviews**.
- **Aspect-based sentiment classification** breaks down **specific concerns (wait time, staff friendliness, etc.)**.