Technical Test: Senior Generative AI Applied Scientist

High level guidance:

This test is reflective of the type of projects you will work on at Sirona, as well as our expectations of what you should be able to accomplish within a given timeframe. You are free to use whatever packages or LLM coding assistants you think would be helpful. Evaluation will be focused on technical competence and creativity, as well as rationale for decisions (e.g. choices around evaluation methodologies that would actually be relevant for the end-user of model outputs; in this case a practicing radiologist). Plan to dedicate a **maximum** 5-7 hours on the test. An OpenAI API key will be provided. Keep API spend at ≤$100 (this budget will likely affect your modeling choices).

Objective:
Demonstrate your practical skill in segmenting data, fine-tuning a generative AI model, and evaluating its effectiveness in automatically generating medical impressions based on radiology findings ("autoimpressions"). The Impression is the radiologist's concise synthesis of the report's Findings—the part clinicians read first. It prioritizes diagnoses and key takeaways, conveys certainty/uncertainty, and recommends actions (when appropriate). Unlike the Findings (objective observations), the Impression is interpretive: it integrates relevant history and priors to answer "What matters, how sure are we, and what should happen next?" Radiologists tend to prefer model outputs that mirror their language and style, including which findings they would normally include or exclude in their summaries. For this reason, they tend to prefer fine-tuned models to general ones.

Data Provided:

- De-identified radiology reports dataset (~30K reports total).
    - Reports include various clinics and modalities (CT, MRI, X-ray, etc.), and anatomical regions (within the reports themselves).
- OpenAI platform API key

Tasks (Suggested time: 5–7 hours):

1. Exploratory Data Analysis (1–2 hours):

- Briefly analyze and visualize key differences in reports across dimensions such as:
    - Clinic
    - Modality
    - Anatomy

- Identify clearly and briefly document key insights guiding your segmentation strategy.

2. Data Segmentation and Modeling Strategy (1 hour):

- Based on your EDA insights, choose and clearly justify the optimal approach to segment the dataset for fine-tuning. Ideally, we could fine-tune models that could handle every scenario, but that is impractical for this exercise. Your sampling and modeling strategy should aim to balance quality with scalability, though there will likely be tradeoffs. Many strategies will likely work, but please make sure to justify your choice with pros and cons.

3. Fine-tuning and Evaluation (3–4 hours):

- Fine-tune a generative AI model to automatically generate impressions from findings in radiology reports for at least one selected segment (~1-5K reports).

- Use practical and relevant evaluation metrics clearly aligned with the clinical use-case. The metrics should align to what radiologists, as consumers of model outputs, may care about. Generated impressions will be consumed by radiologists in the course of their reporting workflow, so things like latency matter.

- Evaluate your fine-tuned model's performance clearly against a baseline foundational generative model.

- Assume a modest fine-tuning budget (≤$100 API spend). You are not expected to exhaustively hyperparameter-tune, but should demonstrate a working pipeline.

4. Scaling Recommendations (brief, ~30 min.):

- Provide concise recommendations on scaling your chosen segmentation and fine-tuning strategy across many radiologists or clinics.

Deliverable:
A concise (4–6 pages or equivalent) notebook with clearly annotated code and a short written summary (or slides). Please make sure to include:

- Key EDA findings
- Code
- Segmentation strategy and justification
- Fine-tuning approach and results

- Evaluation methodology, metrics, and comparison to baseline
- A link to the final fine-tuned model(s)
- Brief strategic scaling recommendations