ML Engineer Practical Test - Questa Engine Project
Time Limit: 90 minutes

OBJECTIVE:
Develop a text anonymization system that can process very long documents, detect sensitive information using NER/PII models, replace it with placeholders, and generate statistics. Include comprehensive unit tests.

REQUIREMENTS:
1. Use transformer-based NER/PII detection models (HuggingFace)
2. Handle documents of any length (memory and token limit considerations)
3. Replace sensitive data with appropriate placeholders
4. Generate detailed statistics about anonymization
5. Write unit tests covering all functionality

DELIVERABLES:
1. Complete anonymization system implementation
2. Unit test suite with good coverage
3. Brief documentation explaining your approach

LIBRARIES YOU CAN USE:
- transformers (HuggingFace)
- torch/tensorflow
- pytest or unittest
- Any other Python standard libraries

==

TECHNICAL CHALLENGE:

You need to build a system that takes a text input (potentially very long - think 100+ pages of documents) and anonymizes it by:

1. DETECTION: Using transformer models to identify:
   - Person names (PER)
   - Organizations (ORG)
   - Locations (LOC)
   - Email addresses (EMAIL)
   - Phone numbers (PHONE)
   - Other sensitive information (MISC)

2. ANONYMIZATION: Replace detected entities with placeholders:
   - "John Smith" -> "[PER_1]"
   - "Apple Inc" -> "[ORG_1]"
   - "john@email.com" -> "[EMAIL_1]"
   - etc.

3. STATISTICS: Generate counts and details:
   - Total entities found
   - Count by category
   - Optionally: entity mapping for de-anonymization

4. HANDLE LONG DOCUMENTS: Deal with:
   - Memory constraints

   - Model token limits
   - Processing efficiency
   - Maintaining entity consistency across chunks

==

EXPECTED OUTPUT FORMAT:

Your system should return something like:

```
{
    "anonymized_text": "Dear [PER_1], Thank you for contacting [ORG_1]...",
    "statistics": {
        "total_entities": 15,
        "by_category": {
            "PER": 5,
            "ORG": 3,
            "EMAIL": 4,
            "PHONE": 2,
            "LOC": 1
        }
    },
    "entity_mapping": {
        "[PER_1]": "John Smith",
        "[ORG_1]": "Acme Corporation",
        ...
    }
}
```

==

UNIT TESTING REQUIREMENTS:

Write comprehensive tests that cover:

1. BASIC FUNCTIONALITY:
   - Simple text anonymization
   - Correct placeholder generation
   - Statistics accuracy

2. EDGE CASES:
   - Empty text
   - Text with no entities
   - Duplicate entities
   - Overlapping entities
   - Very long texts

3. MODEL BEHAVIOR:
   - Low confidence predictions
   - Different confidence thresholds
   - Model failures/errors

4. PERFORMANCE:
   - Memory usage with large texts

   - Processing time benchmarks
   - Batch processing efficiency

==

EVALUATION CRITERIA:

1. ARCHITECTURE DESIGN (25%)
   - How you handle long documents
   - Memory efficiency approach
   - Code organization and modularity
   - Scalability considerations

2. NER/PII IMPLEMENTATION (25%)
   - Correct use of transformer models
   - Entity detection accuracy approach
   - Confidence threshold handling
   - Entity consistency across chunks

3. ANONYMIZATION LOGIC (20%)
   - Placeholder generation strategy
   - Text replacement accuracy
   - Statistics generation
   - Entity mapping maintenance

4. TESTING QUALITY (20%)
   - Test coverage and completeness
   - Edge case handling
   - Mock usage for models
   - Performance testing approach

5. CODE QUALITY (10%)
   - Clean, readable code
   - Proper error handling
   - Documentation quality
   - Type hints usage

==

BONUS POINTS:

- Implement entity de-anonymization capability
- Handle different text formats (preserve formatting)
- Implement configurable entity types
- Add performance benchmarking
- Include integration tests with real models
- Implement entity relationship preservation
- Add logging and monitoring capabilities

==

SUBMISSION INSTRUCTIONS:

1. Create your implementation from scratch

2. Structure it as you would for a production system
3. Include a README explaining your approach
4. Make sure all tests pass
5. Be prepared to explain your design choices

SAMPLE TEST DOCUMENT (use this for testing):

"Dear John Smith and Mary Johnson,

Thank you for your interest in our services at Acme Corporation.
Please contact us at info@acme.com or call our office at 555-123-4567.

Our headquarters are located in New York City, with additional offices
in San Francisco and London. We serve major clients including
Microsoft, Google, and Amazon.

For technical support, reach out to support@acme.com or call
1-800-SUPPORT. You can also visit our website at www.acme.com.

Best regards,
Robert Davis
CEO, Acme Corporation
robert.davis@acme.com
Direct: 555-987-6543"

Expected entities in this sample:
- PER: John Smith, Mary Johnson, Robert Davis
- ORG: Acme Corporation, Microsoft, Google, Amazon
- LOC: New York City, San Francisco, London
- EMAIL: info@acme.com, support@acme.com, robert.davis@acme.com
- PHONE: 555-123-4567, 1-800-SUPPORT, 555-987-6543

Good luck!