Final Project Outline: Predicting Potato Late Blight in Huancavelica, Peru Using LLaMA 2-7B

---

1. Project Title

Development of a Predictive Model for Potato Late Blight Outbreaks in Huancavelica, Peru Using LLaMA 2-7B

---

2. Background and Motivation

- Potato late blight (*Phytophthora infestans*) is one of the most devastating plant diseases, causing significant economic losses and threatening food security.
- Huancavelica's high-altitude climate creates ideal conditions for blight outbreaks due to:
    - Variable weather patterns
    - Limited agricultural infrastructure

Goal: Predict outbreaks in advance to enable more effective disease management, reduce crop losses, and enhance food security.

Inspiration from LLM Research: Large language models (LLMs) can outperform human experts in predicting complex outcomes by synthesizing patterns across diverse data sources.

Objective: Adapt that methodology to agriculture for predicting potato blight outbreaks.

Outcome: Create an accurate, resource-efficient prediction model that can operate in low-resource settings. Explore the use of synthetic data to enhance predictive capabilities for specific scenarios.

---

3. Objectives

- Develop a machine learning model using LLaMA 2-7B to predict potato blight outbreaks based on weather, crop, and disease data.

- Optimize the model for deployment on a Beelink Mini S12 Pro.
- Fine-tune the model using LoRA to enhance predictive performance on agricultural data.
- Implement active inheritance to strategically select synthetic data for improved model performance.
- Establish a benchmark for potato blight prediction.
- Evaluate model performance in terms of:
    - Predictive accuracy
    - Confidence calibration
    - Generalization
- Compare performance with human agricultural experts.
- Analyze model integration of information from different data sources.
- Explore the potential of augmenting real-world data with targeted synthetic data to improve prediction accuracy.

---

4. Methodology

4.1. Model Selection

Use LLaMA 2-7B due to:

- High efficiency on small-scale hardware
- Open-source availability (customization and transparency)
- Proven effectiveness in knowledge-intensive domains

Fine-tune using LoRA to improve performance while minimizing computational load.

4.2. Data Collection and Preprocessing

Data sources:

- Historical weather data (temperature, rainfall, humidity)
- Crop data (planting dates, variety)
- Disease occurrence data for Huancavelica

Preprocessing steps:

- Clean data
- Handle missing values
- Standardize formats
- Split data into training, validation, and test sets

Data-to-Text Transformation:

- Develop a pipeline to transform numeric time-series weather data into textual or token-based inputs suitable for LLaMA 2-7B.
- Specify how often the model sees new data (e.g., weekly batch updates vs. continuous streaming) and how overlapping intervals in textual inputs are managed.
- Examples:
    - *"Week 1: Temperature averaged 18°C, rainfall was 5mm, and humidity was 80%."*
    - Use a structured format that LLaMA can interpret (e.g., table-like text representation).

4.3. Synthetic Data Generation and Active Inheritance

- Generate synthetic data using a larger pre-trained LLM (e.g., LLaMA 2 or Mixtral-8x7B).

Targeted prompting:

- Coverage of specific weather patterns conducive to blight
    - *Example:* "Generate data for a scenario with high humidity, low sunlight, and temperatures between 15-20°C."
- Generate multiple samples per prompt (e.g., 10 samples)

Selection of the most relevant sample based on:

- Keywords
- Patterns
- Data alignment with real outbreak conditions

Synthetic Data Validation:

- Involve local agronomists and extension officers to review and remove unrealistic scenarios.
- Domain-informed filtering to mitigate spurious correlations
- Expert Review Protocols: Use a short standard questionnaire vs. open-ended comments to identify unrealistic data.

4.4. Integration of Real-World and Synthetic Data

Mixing strategy:

- Fixed ratio (e.g., 70% real, 30% synthetic)
- Adapt ratio based on validation performance

Adjustments:

- Weight real-world data

- Filter synthetic data based on alignment with real data

4.5. Model Training and Evaluation

Training:

- Fine-tune using LoRA on mixed dataset

Configuration:

- Epochs: 10-20 (based on validation performance)
- Learning rate: Cosine annealing (initial learning rate of 1e-4)
- Batch size: 64 (adjustable based on hardware limits)

Benchmark Creation:

- Generate synthetic test cases
- Modify historical data to create counterfactual scenarios:
    - *Example:* "Increase rainfall by 20% in August, decrease temperature by 5°C in July"

Evaluation Metrics:

- Prediction Accuracy: ≥ 80%
- Precision: ≥ 0.75
- Recall: ≥ 0.75
- F1-Score: ≥ 0.75
- AUC-ROC: ≥ 0.85
- Inference Speed: ≤ 1 second (on Beelink Mini S12 Pro)

Calibration Metrics:

- Expected Calibration Error (ECE) and reliability diagrams
- Document how reliability diagrams will be generated (e.g., using a validation set)

Integration Analysis:

- Remove data types to test:
    - Weather-only cases
    - Crop-only cases
    - Disease-only cases
- Measure impact on accuracy and confidence

Model Profiling:

- Test for regional bias
- Test sensitivity to local agricultural practices

Human Benchmarking:

- Compare model predictions with expert opinions
- Identify areas of agreement and disagreement
- Formalize the "expert vs. model" comparison with prospective evaluations
- Run periodic forecasting "contests" with local extension teams to monitor drift

Feasibility Experiment:

- Compare LLaMA-based approach to a standard time-series model (e.g., XGBoost or LSTM)

4.6. Model Deployment

Deploy on Beelink Mini S12 Pro

Optimizations:

- Model quantization
- Use optimized inference libraries (e.g., ONNX Runtime)

Deployment Trial:

- Small pilot test with a typical week's data to validate inference speed meets ≤1s target

Automation:

- Set up automated pipelines for real-time data updates

Cloud Backup:

- Periodic re-training on Microsoft Azure

---

5. Performance Metrics

- Prediction Accuracy: ≥ 80%
- Precision: ≥ 0.75
- Recall: ≥ 0.75
- F1-Score: ≥ 0.75
- AUC-ROC: ≥ 0.85
- Inference Speed: ≤ 1 second
- Improved performance on synthetic benchmark

- Integration of diverse data sources
- Alignment with expert predictions
- Reduced potential biases

---

6. Challenges and Risks

Data Quality:

- Inconsistent weather and crop reporting
- *Mitigation:* Cross-validation, automated outlier detection

Hardware Limitations:

- Limited VRAM on Beelink Mini S12 Pro
- *Mitigation:* Reduce LoRA rank, adjust batch size, quantize model

Overfitting and Generalization:

- Risk of overfitting to specific regions
- *Mitigation:* Data augmentation, regularization, diverse test cases

Environmental Variability:

- Climate changes affecting patterns
- *Mitigation:* Periodic fine-tuning

Effectiveness of Synthetic Data:

- Synthetic data may not fully capture real-world complexities
- *Mitigation:* Careful prompt engineering, active inheritance

---

7. Timeline

Phase	Duration	Tasks
Data Collection and Cleaning	1 month	Data acquisition, preprocessing
Model Training and Fine-Tuning	2 months	Initial training, LoRA fine-tuning, active inheritance implementation
Benchmark Creation and Evaluation	1 month	Synthetic test cases, performance evaluation, model profiling
Model Deployment	1 month	Deployment, automatic updates, optimization
Performance Monitoring and Adjustment	Ongoing	Continuous refinement based on real-world feedback

---

8. Budget

Item	Cost (USD)	Notes
Beelink Mini S12 Pro	$300	Hardware
Cloud Services (Azure)	$100/month	Fine-tuning, periodic retraining
Labor	$5,000	Data scientist, ML engineer
Total	~$5,500	Initial investment

---

9. Expected Outcomes

- 80%+ prediction accuracy
- Deployment on low-cost hardware
- Scalable framework for other crops and regions
- Active inheritance improves prediction
- Benchmark for future models

- Insights into model's data integration capacity
- Identification of model biases
- Human expert comparison

This version incorporates all the feedback and provides a well-defined plan!