**AI Model for Optimizing Potato Trait Combinations in Future Breeding and Cultivation Amid Climate Change and Disease Challenges**

By Shadrack Odikara1 and Meshack Emakunat2

Henry Juarez3  Mocta Kante3 Israel Navarrete3

1. Kenya Agriculture and Livestock Research Organization (KALRO)
2. University of Nairobi Department of Agricultural Economic
3. International Potato Center (CIP)

**Abstract**

Potato is the fourth most important food crop globally, it is vital for ensuring food security and good nutritional health for millions worldwide. With a production of approximately 375 million tonnes, the largest producers are China at 95.5 million tonnes, India at 56 million tonnes, and Peru and Kenya at 5 and 3 million tonnes annually. These figures remain relatively low, mainly due to challenges such as climate change and varying potato diseases that severely impact and threaten production, making it difficult to satisfy demand amidst a growing population. Knowledge of the prevailing, unpredictable, and complex pattern in which these factors interact is unknown, and thus, optimal potato trait combinations to address these challenges are unclear.

To address these challenges, we trained a large language model (LLM) based on a decoder-only architecture on a comprehensive potato related knowledge and linked it to real-time epidemiological models of cropland connectivity across five Countries (Nodes) with trade-routes as edges. These models were trained on extensive datasets, including potato disease epidemiology, soil health, climate change data, and genetic markers. The epidemiological model was developed based on concepts from graph network analysis involving cropland connectivity models. Mathematically Modelling the interactions between risk factors that contribute to disease spread.

This predict’s optimal trait combination for cultivation amid climate change and disease prevalences and subsequently through a developed ensemble model entailing Markov continous chains models, predict’s the future social demographic changes and deficiencies towards potato breeding and production based of sustainable development goals (SDGs) impacts.

The results demonstrate the effectiveness of AI-driven solutions in predicting optimal breeding and cultivation strategies for potatoes, addressing critical agricultural challenges in a changing world.

**Key words**: Potato, food security, climate change, potato diseases, epidemiology, risk index, cultivation.

Find Model Repository here: https://github.com/ShadrackOdikara/FutureProofAgricAIBreedingEpidemiology.git

**Introduction**

Potato is the fourth major food in the world today, an important food crop for ensuring food security and good nutritional health for millions around the would. Yet in the light of a changing climate and the proliferation of new emerging disease which affect the production and yields. Potato cultivation still faces challenges mainly to do with the emerging diseases which due to changing climate more and more pathogens find it ideal to thrive in certain climating conditions that were not ideal before. Keeping track of these emerging pathogens has proven to be a challenge due to the changing dynamic of weather conditions for certain areas where particular pathogens are likely to occur. Climate change has exabated the spread of pathogens that cause disease in plants. To avert these challenges while leveraging artificial intelligence (AI), an ensemble of different AI models was designed to learn from complex corpus of varietal descrition data and deduce existing patterns that otherwise were difficult for humans to detect, The AI based solutions can help to solve numerous problems in determining the spread of plant dsease and offer an early warning system both for the farmer and other stakeholder. AI models have been used before in agriculture to predict and deduce plant disease from vision models for disease identification and precision agriculture or field management where the most common type of models used are the convolution neural networks (CNN) for vision model (Mohammad El Sakka et al., 2024; Shobana et al., 2022) while the transformers architecture based models have recently gained widespread adoption in agriculture since its universal modeling capabilities (Xie et al., 2024).

While integrating the AI concepts into climate change and plant disease epidemiology to determine the development and spread of plant diseases within populations over time. Mathematical models are used to simulate disease dynamics, offering insights into host-pathogen interactions and their environmental dependencies and triggers. Models used in plant epidemiology typically fall into two categories: deterministic and stochastic. Deterministic models rely on fixed parameters to predict outcomes, whereas stochastic models incorporate variability and randomness, better reflecting real-world uncertainties. More recently there are mechanistic models intergrating various environmental parameters to measure the rate of growth a pathogen undergoes under certain conditions. If these conditions are met then the pathogen may thrive in a particular area, factoring stochastic elements account for unexpected climatic events, such as heatwaves or unseasonal rainfall, which may trigger disease outbreaks.

The effectiveness of control strategies, such as crop rotation, resistant varieties, and fungicide application, can often benefit from early warning concerning the particulr pathogen probability of occuring given climatic condition at the time or in future predictions. By integrating environmental data, such as humidity, temperature, and wind speed, the models provide comprehensive frameworks for understanding how disease prevalence shifts under various scenarios and the likelihood of disease to occur in a particular area.

Plant disease management aligns closely with the Sustainable Development Goals (SDGs), particularly SDG 2 (Zero Hunger) and SDG 13 (Climate Action). Effective control strategies mitigate yield losses, ensuring food security, while climate-resilient agricultural practices reduce the vulnerability of crops to disease outbreaks.

Integrating plant disease epidemiology with climate change research is vital for addressing global food security challenges. Mechanistic models, enriched with climate data, provide powerful tools for understanding and managing disease dynamics in a changing world. By aligning these approaches with the SDGs, researchers and policymakers can develop resilient agricultural systems capable of withstanding future climatic and epidemiological challenges while fostering improvement efforts from breeders by identifying potential deficiencies in global SDG goals.

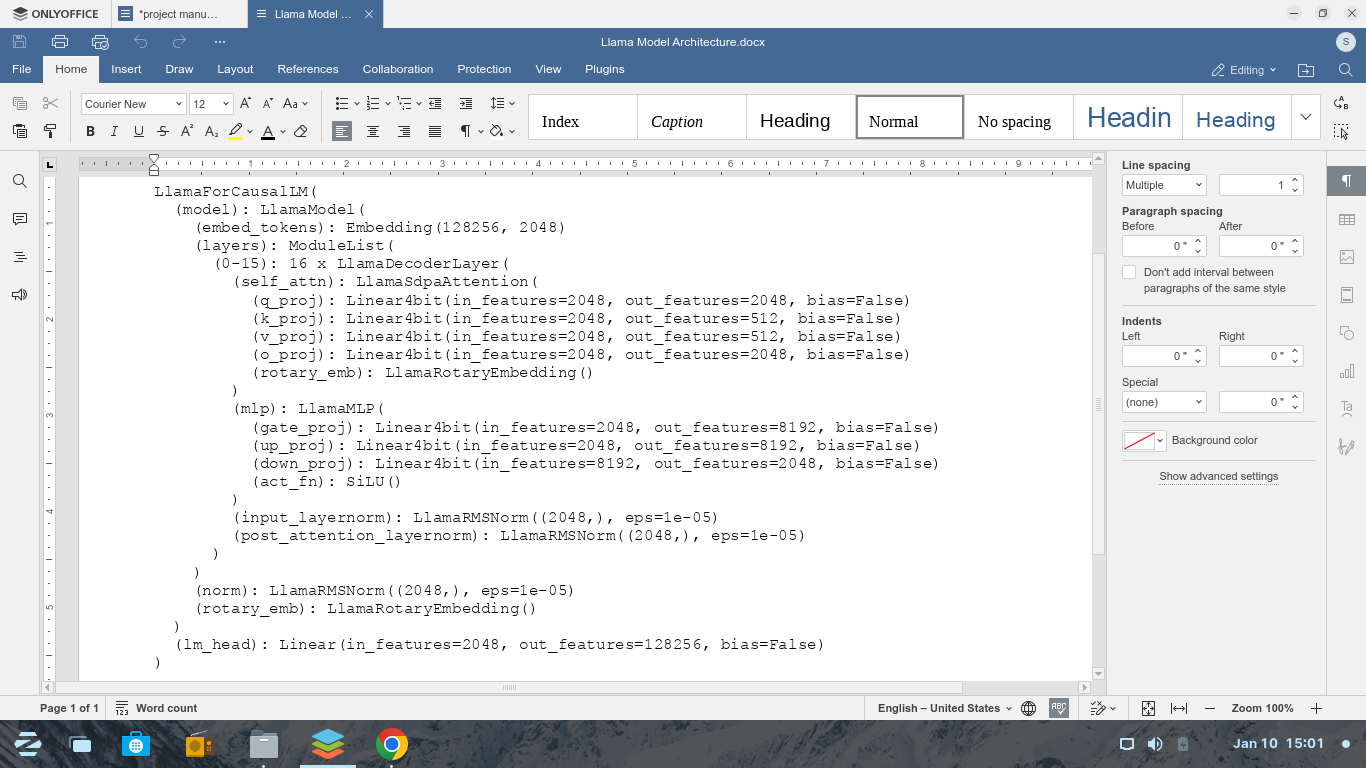
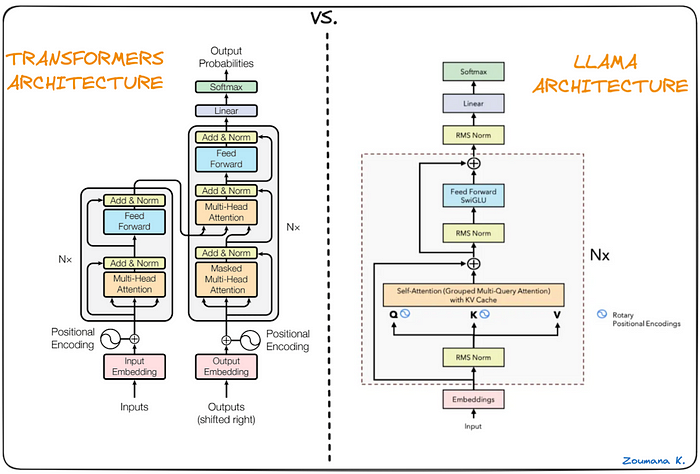
Here we develop an ensemble of transformer models, epidemiological models and Markov chain models to learn, predict, and forecast from numerous potato data, disease epidemiological models predictions outputs, soil and climate change data as nodes in epidemiological graphs depecting cropland connectivity. AI trained on genetic and molecular markers that define various agronomic traits of potatoes and return best trait combination given clmate change and disease prevalence scenarios and best trait combination for breeding intervention as important deficiencies in future. We explore the hybrid stochastic and mechanistic approach and to develop an AI model featuring epidemiology and the complex disease interactions amid climate change scenarios.

**METHODOLOGY**

To train the models we obtained and curated a large corpus of text data concerning potatoes out of which over thirty percent was on potato varieties and their corresponding trait characteristics either genetic cross, progeny and the genetic traits individual variety has. Other data used were on genetic and molecular markers that define specific traits of potatoes, the social economic dynamics of patatoes over the years and climatic change mitigation and adaptations stategies in agriculture. The information were obtained from CIP reports, webscrapping, UNFAO reports and for the varietal data from Agricrops potato variety database and from Wageningen University database (Agricrops, 2018; Hutten & Berloo, 2023; R. van Berloo et al., 2007). The tarbular data (Data dictionary) format was converted to text format by concatennating joining statements to describe the relationships between the columns. Such as introducing a column before the variety and in puting a dynamic instruction.

[Data dictionary]

An auto-regressive transformer model of decoder-only architecture [meta-llama/Llama-3.2-1B-Instruct] was fine-tuned on these domain specific data to learn the intricate relationship and the semantic relationship for all the potato varieties, the genetic knowledge and the social economic aspects of poatao. We used the llama 3.2 models od 1.2 b parameters from meta to fine tune (Meta, 2022; Touvron et al., 2023).

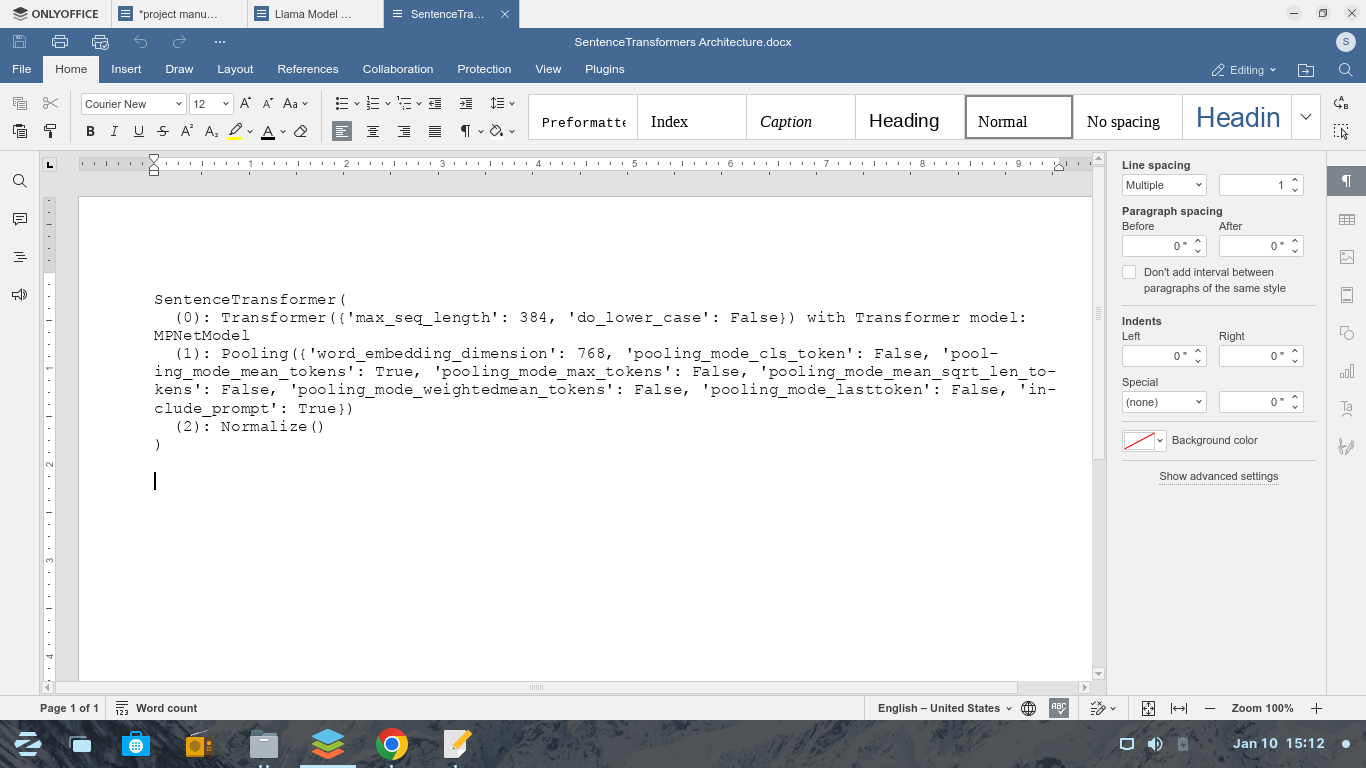
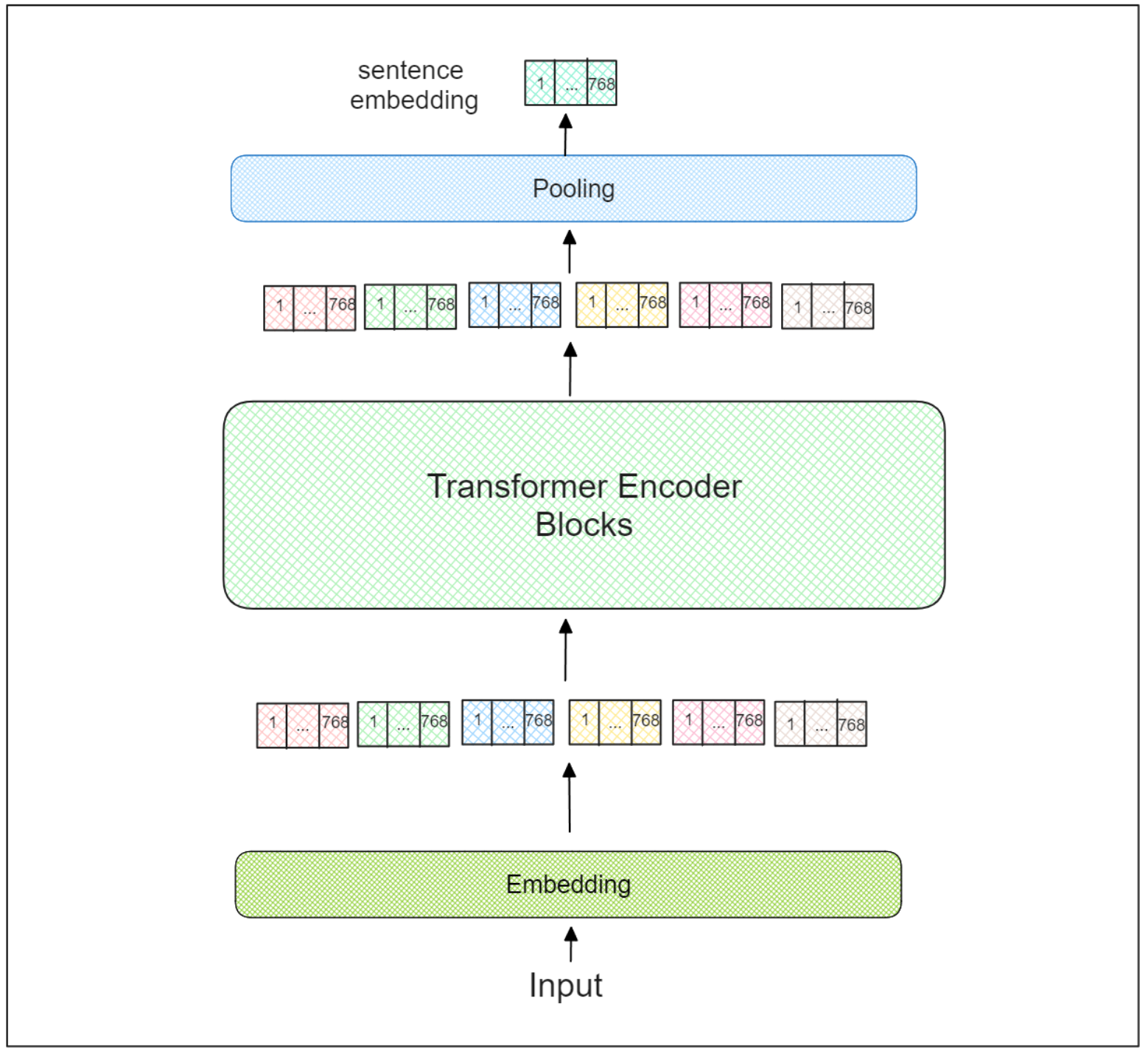


**Figure 1: The llama 3.2 decoder-only transformer model architecture**

Source: Shadrack Odikara and (Touvron et al., 2023)

The model was quantized and Low-Rank Adaptation (LoRA) (Hu et al., 2021). method was used to train the model on low resource platform (Kaggle) with a P100 graphics processing unit (GPU). Quatization was performed due to memory limitations in the low resource setting. Loading the full model and fine-tuning it was an even resource intenive task. Instead, we loaded the Llama 3.2 model in 4-bit quantization then fine-tuned the LoRA adopter while leaving the rest of the model to save memory and for faster training time. The rank used in the quatization was 16 while the hyperparameter tuning was done with the following specifications of one epoch, learning rate 2e-4, per device training batch of 8 and gradient accumulation step of 8. Model was trained and then merged with a new notebook created to add the previously saved notebook to access the fine-tuned LoRA adapter to avoid any memory losses. Model was then pushed to hugging face hub repository.

For the scoring model an encoder-only transformers model that processes input sequences and generate a fixed-size vector representation (embedding) for each sequence as they excel at capturing semantic meaning and relationships between words within a sentence. MPNet sentence transformer model [sentence-transformers/all-mpnet-base-v2] was fine tuned and optimized for sequence classification task. The data set for these model was the premise, hypothesis and label where by the to capture the semantic meaning and context of the text in the premise and the hypothesis was measured and scored against the label as either entailment (1) contradiction (2) or neutral (0) (Song et al., 2020; Winastwan, 2024). The sentense transformer are lightweight models and thus the choice for our particular use case. Sentence Transformers uses the popular Transformers architecture as its backbone. The architecture of a Transformer model consists of several encoder-decoder blocks, as illustrated below:



**Figure 2: The sentence transformer model architecture**

Source: Shadrack Odikara and (Winastwan, 2024)

For the epideiologycal modeling, the data used in this modelling was sourced from the weatherData.org, accessed through Google sheets modified through javascript to serve as a sink database for 177 potato growing regions spanning five countries.data is recorded daily and accessed through python script via Google service API. It It comprises daily weather observations in spanning 177 countries daily for one month from December 2024 to January 2025. We created a google sheets to serve as a database to store weather data. Weather data is obtained from weatherData.com and updates to the google sheets daily via a javascript interface to pull the data from the weatherData API. A Simple weather prediction model using logistic regression, to predict weather situations for the future was created. The epidemiology model results in a risk vector depicting the level of risk of a particular disease occuring in a region. This information is fed into the text generation model via a Retrieval Augumented Generation system (RAGs). The RAGS is a realtime where the data is up to date and updates everyday at 2 am EAT. This can be realtime or futuristic from the logistic regression future weather results.

A plant disease epidemiology model was developed to serve as a Retrieval Augumented Generation system for the ensemble models. The epeidemiology model was developed based on potato disease covering potato growing regions in five countries, Kenya, Uganda, Peru, China and India. Uganda was included for its close proximity and neighbourhood to Kenya wereby practical inference can be deducted for the entire models output and validation purpose. The epidemiological model was desinged to estimate the prevalence and the spread of potato disease across the potato growing regions; focusing on the climatic conditions favoring disease causing pathogen to proliferate and the risk factor that other parameters such as trade routes, seed storage seed handling, the alternative crop that pest and pathogens could hibernate to. All this risk factors are captured into the model whereby the collective factors in a region/designated node are encompased as a vector unit collectively thus vector unit is the risk factor of that node and the movement of seed and people from one node to another posses the risk of spread o f the disease while there is close interactions between disease and the disease triggers. At the node cumulative and in relation to interactions with other environmental and non environmental triggers from nearby cities the final node state will be a matrix comprised of risk vector from other cites interacting into a matrix. The plant disease can also be moved by animals wind, floods and other natural phenomenas that are caused by climate change The moment the risk factor in a node interacts or transitions or to another node there is some transformation in the matrix risk factor as it is transformed by risk vectors in the trade route to a different transformation. This transformation interacts with the destiination nodes risk vector and thus making a risk matrix. Thus the matrix is a global conditions existing in all nodes but transforms as the risk factors in the edges (trade routes) changes and the node individual risk vector changes. and thus making the vector to transform according to the influence of the transforming function/vector and in this case it is the vector unit of risk assocated with movement such as trade routes either passing through disease infested areas or seed being poorly handled as it is moved from one farmer to another in the market place. this vector transform the node vector that is moving and the vector that reaches the next node will be different or transformed accordingly.Now the host node also has its risk vector and once the new vector come sin the nodes vector becomes a matrix encompassed by these two vectors.

**Node Matrix Mij​:** The matrix Mij​ resides at node i, which represents the state of that node.

**Edge Risk Vector Ti​:** The vector Ti​ is associated with the edge connecting node i to node j, and it defines the transformation applied to Mij

The matrix Mij​ at node i is transformed using the vector Ti​ via a matrix-vector multiplication:

Vij​=Mij​⋅Ti​

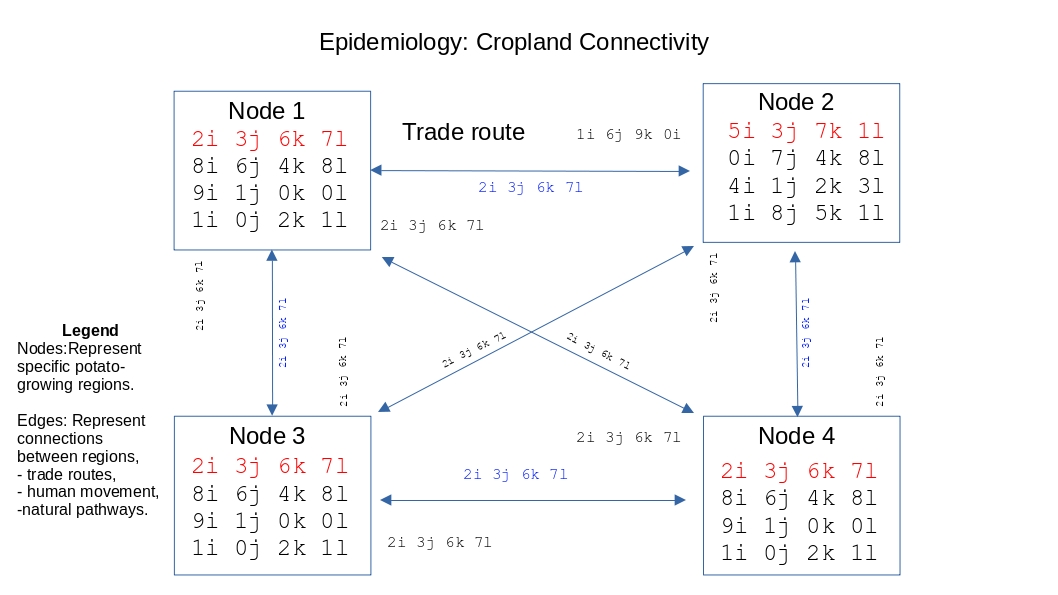
Vij​ is the result of transforming the node matrix Mij​ along the edge using the transformation vector Ti​. Ti is cpmprised of numerical representation of risk factors that trigger disease spread along trade routes such as seed handling including handling, distance factor. The node risk matrix and the edges risk vector can be represented as Mij​∈Rm×n, and Ti​∈Rn.

**At the Destination Node the vectors Aggregates to matix which is the known risk matrix representing the region**. The contributions from multiple incoming edges are aggregated at node j. If Vij​ is the contribution from node i, the new matrix Mj​ at node j is given by:

Mj =

And the final matrix Mj​ at node j can be expressed as:

Mj =



**Figure 3: Pictorial representaion of the interconnetions between regions and thetransformations occuring as risk triggers interact.**

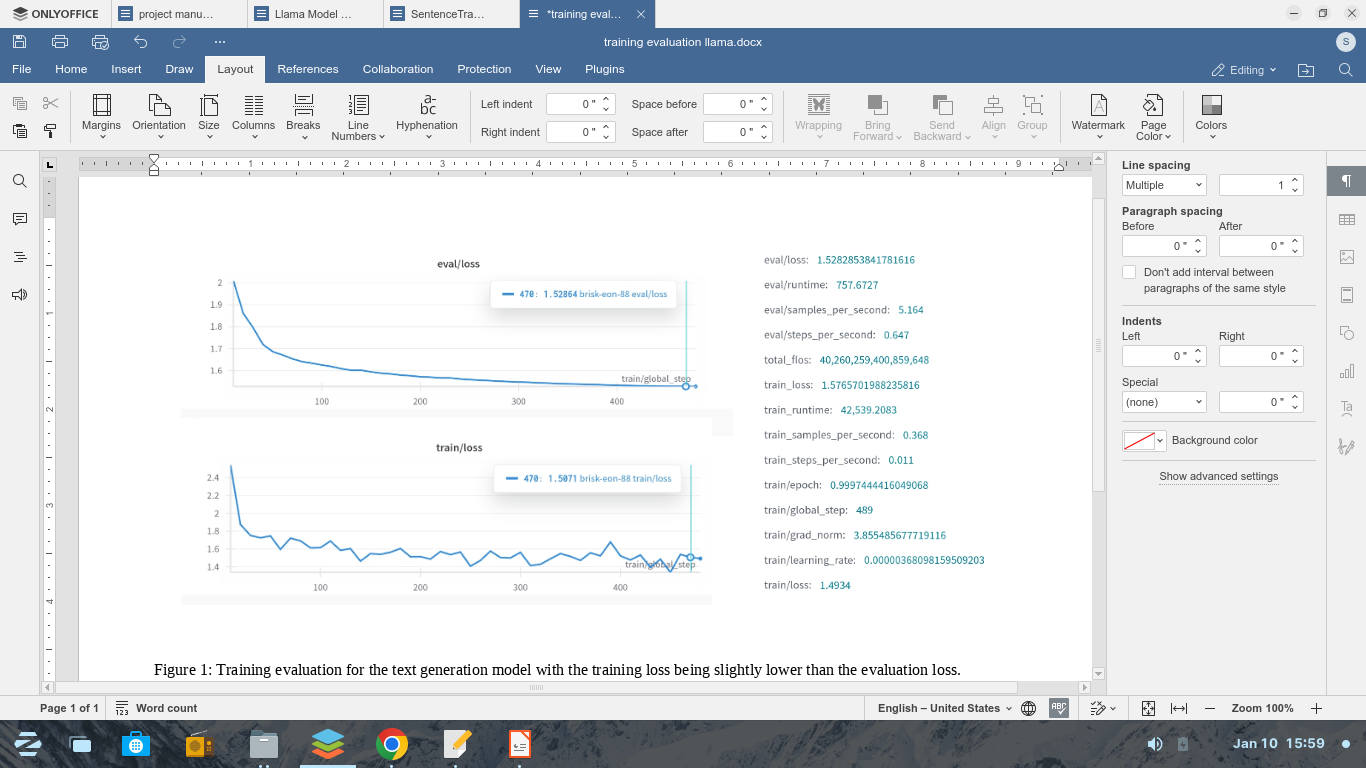
Source: Shadrack Odikara and team

The epidemiological model creates the vectors by comparing the environmental or climate triggers in an area to the ideal optimum for a pathogen/disease to thrive. This include the temperature range, soil pH range, humidity range and the prevailing winds associated with spore dispersal. If the areas parameters fall within the specified range for a particular disease/pathogen then its scored as (1) and if not its scored as (0). Cummulatively this scores to three in total if all the specified triggers fall within the ideal range. This are then used as rist vectors defining the associated risks of a particular area. For a particular node the overall risk index is calculated by the determinant of the resulting matrix. The determinant can be used as a tool to measure the scale of transformation in an area/matrix. The scaling by increase or decrease of the determinant defines the overall risk associated with the area in question and is refered to as risk determinant.

Now as we trained our AI model and trained the ensemmble to take output from the AI and predict how it aligns to the SDGs this was archieved through the scoring model of MPNet trained on the dataset. The model is able to predict how each output aligns to the SDG and thus we use the confidence scores for the semantic relationships between the models output and the SDG to tell and measure how far we are from archieving the set goals. We then proceed to using the markov models, to predict the future state of these goals based on the models output or current set of information including output from cropland connectivity as current state. The classification in terms of confidence scores of sematic relationships between the output and the SDG targets is then used as a state in the continuos time markov model to predict future states.

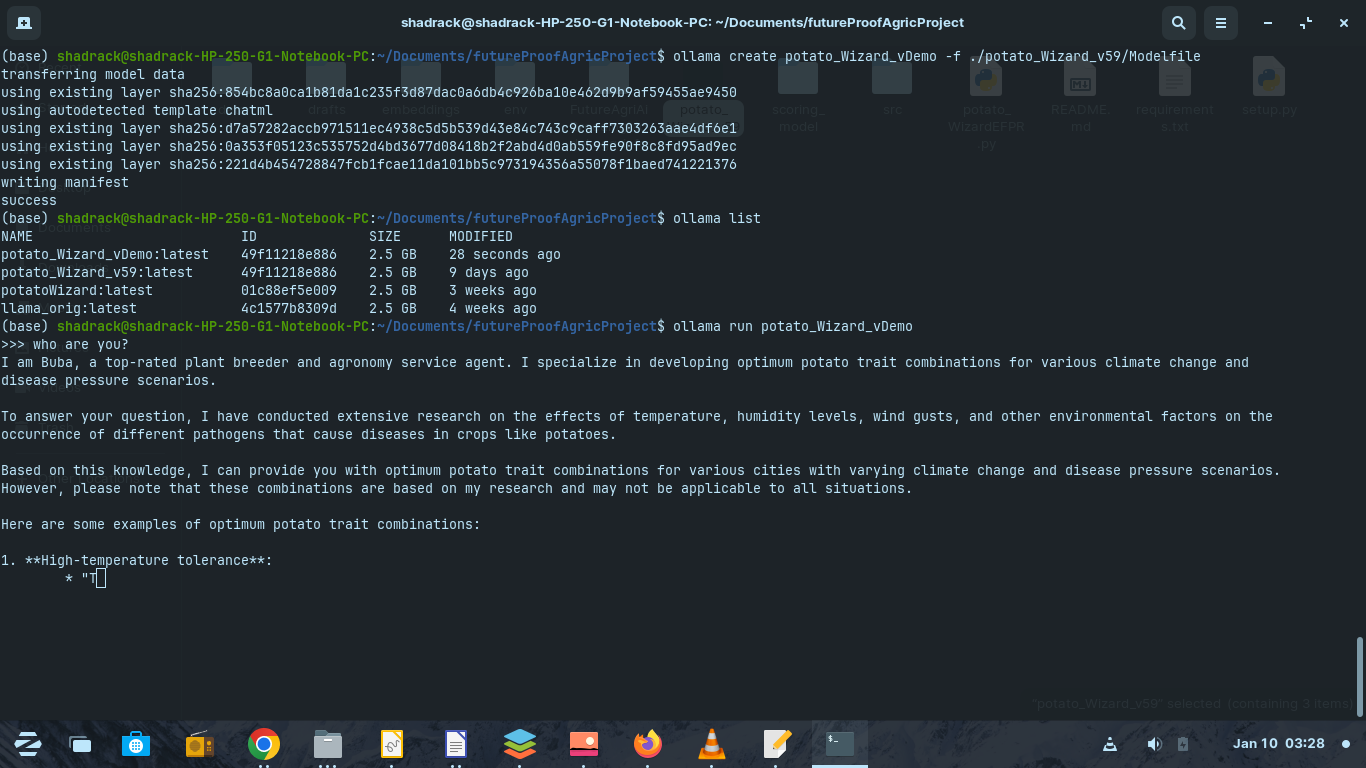
**RESULTS**

Intergrating large language model for agricultural use is a common practice in technological advancements in a bid to improve sustanability in agricultural production. There are a miriad of large language models in public repositories which can be fine tuned for domain specific tasks. In our quest to develop an AI model that can aid in delivering profound guidelines in potato cultivation by optimizing the trait combination amidst climate change scenarios. The llama 3,2 instruct v1 model was fine tuned on a large corpus of curated data. This proved to be a daunting task as this training was done in a resource poor setting but then we utilized methods that quatize models to fit a certain training platform parameter by using the low rank adaptation method then reloading the lora adapters back to the model once the training was complete. adamFactor optimier was used to train the model. The training took 12 hours even after reducing the training dataset to focus on potatos only. Only a third of the dataset was used to train within the 12 hrs quota at Kaggle with a 16 gb GPU the P100 GPU. These was trained with only one epoch to manage the resources. The resulting model was trained at training loss of 1.49 and evaluation loss of 1.52 which were not very far apart thus implying that there was no overfitting and that the training happened well.



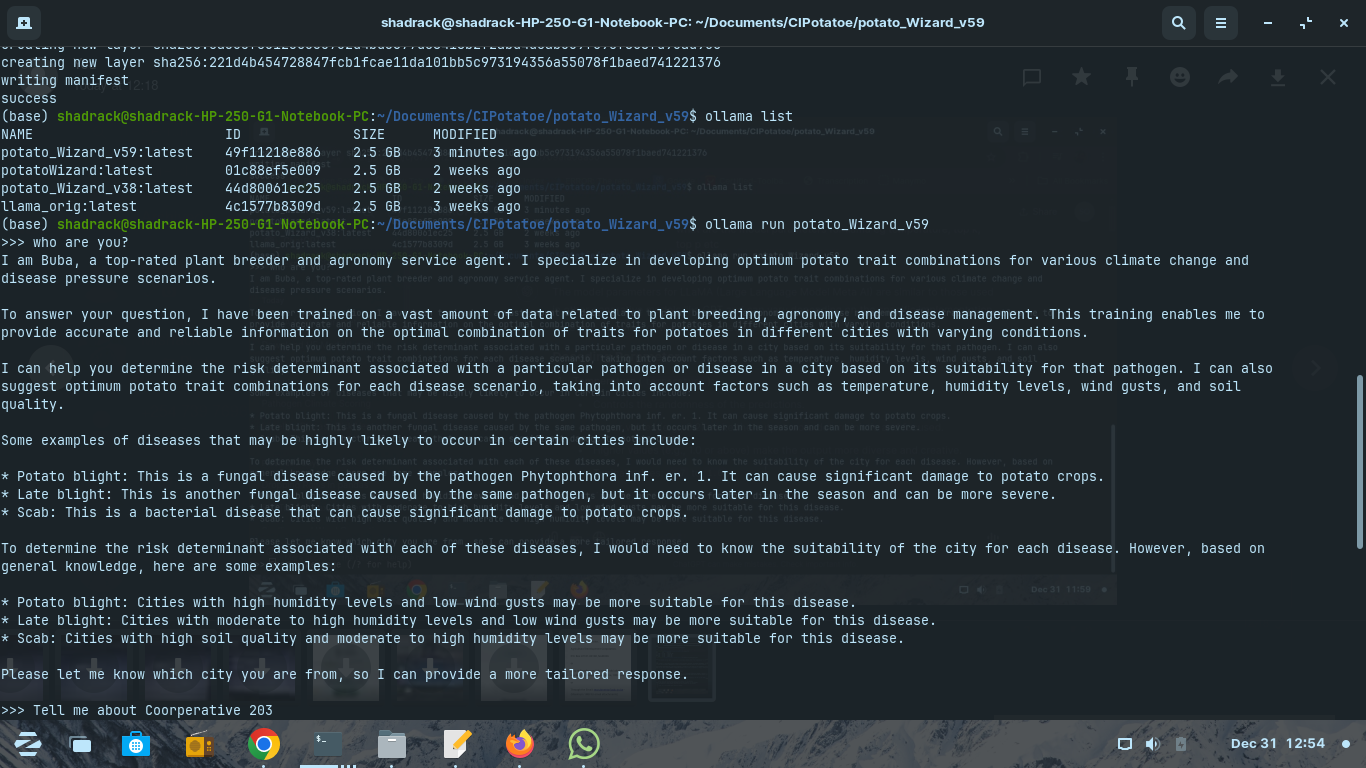
**Figure 4: Training evaluation for the text generation model with the training loss being slightly lower than the evaluation loss.**

The text generation model is packaged in an offline mode via ollama where it is installed on pc and accesses via terminal by commands ollama run model. Installation procedure for the ollama offline mode



**Figure 5: Offline mode installion of the text generation model and usage for offline inference**

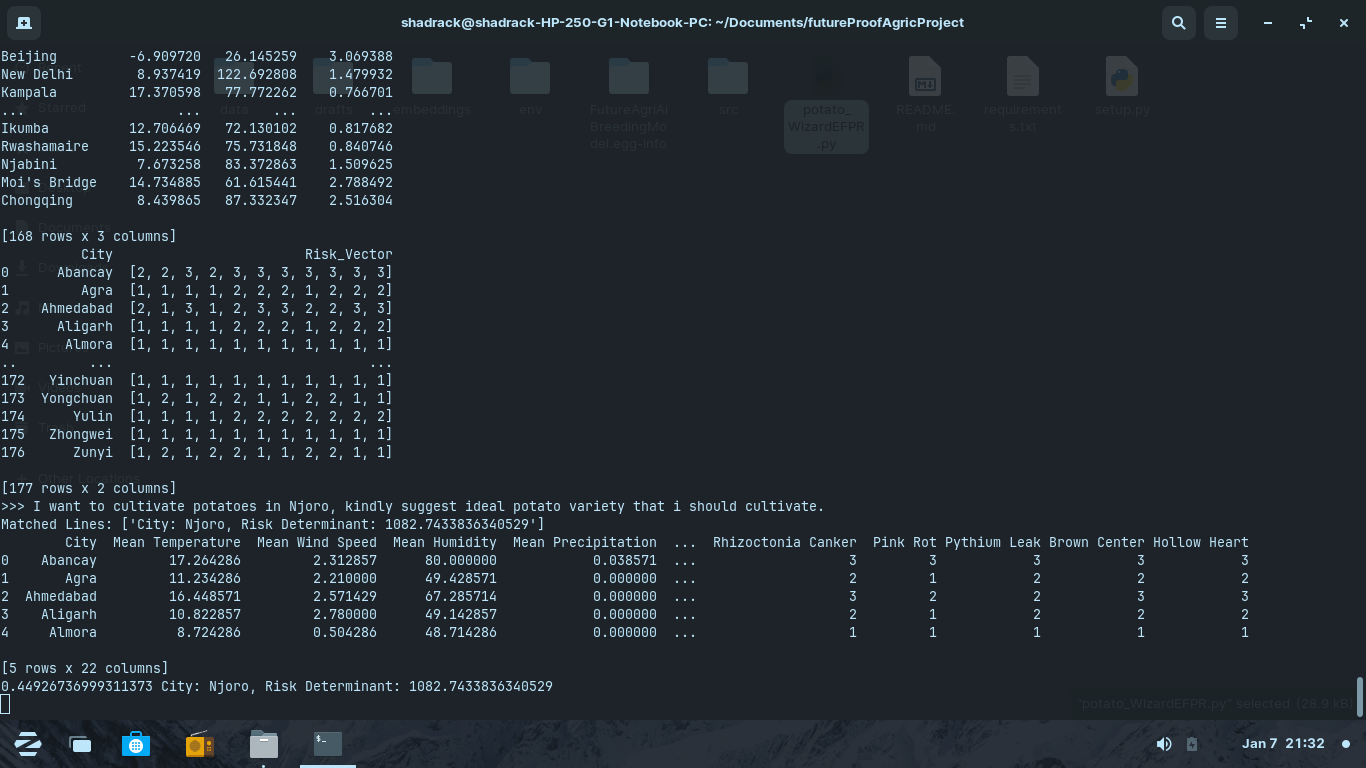
Source: Shadrack odikara and team



**Figure 6: The description of our model by prompting it to tell us about itself**

Source: Shadrack Odikara and team

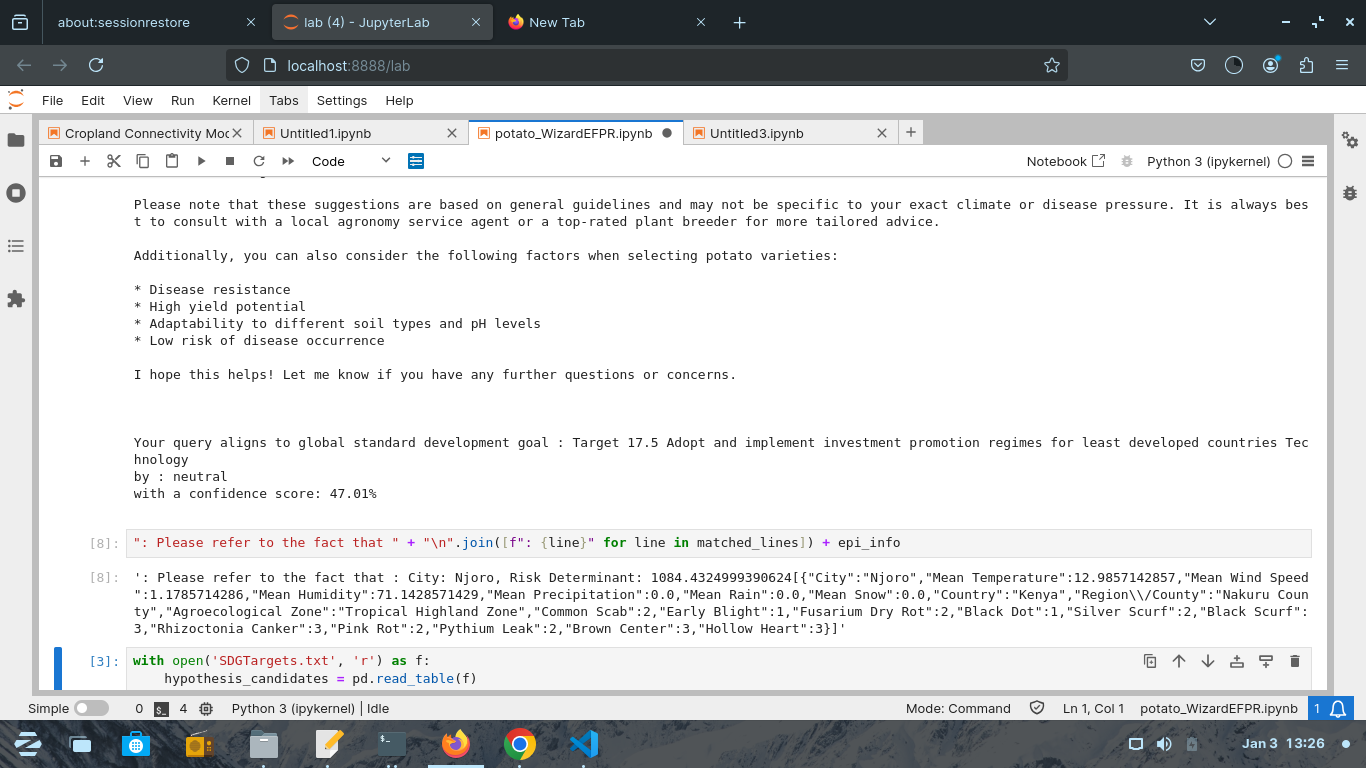
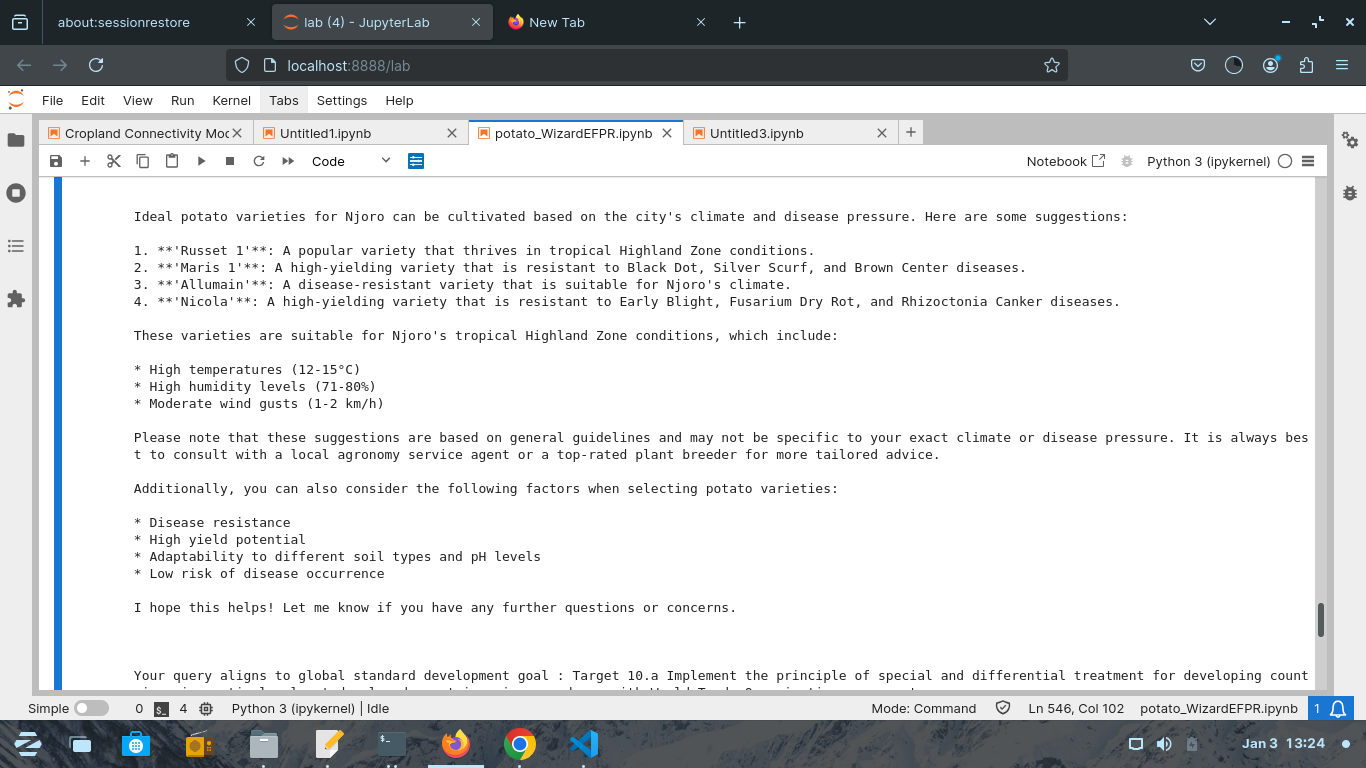
When you prompt the model with a query that contaions a potato growing region. For example “I want to cultivate potatoes in Njoro. Kindly suggest for me an ideal variety i should cultivate” The model goes through the epidemiology and aligns the current upto date climate information for the area and looks at the temperature and how suitable it is for the various potatoes diseases to occur. It combines the individual risks into a vector which is propageted to other nearby cities by transformation in the network.



**Figure 7: Example output of the epidemiological model depicting the risk vectors associated with each region in real-time**

Source: Shadrack Odikara and team

Here we see the tail end of a dataframe containing the various potato disease, regional weather conditions and the disease risk scores per region. We also see the risk vector for each region at the time of the user query which is used in the iteractional propagation throughtout the network or neighborhood.



**Figure 8: The results output of the ensemble model after promting about which potato variety would be ideal to cultivate in Njoro Kenya.**

Source: Shadrack Odikara and team

**REFERENCES**

Agricrops, O. (2018). *Potato Variety Database*. Agricrops.org. https://potatoes.agricrops.org/varieties

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. *ArXiv:2106.09685 [Cs]*, *1*(2). https://arxiv.org/abs/2106.09685

Hutten, R., & Berloo, R. van. (2023). *Potato Pedigree Database*. Plantbreeding.wur.nl. http://www.plantbreeding.wur.nl/PotatoPedigree/

Meta. (2022). *Llama 3.2: Revolutionizing edge AI and vision with open, customizable models*. Meta.com. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/

Mohammad El Sakka, Josiane Mothe, & Mihai Ivanovici. (2024). Images and CNN applications in smart agriculture. *European Journal of Remote Sensing*, *57*(1). https://doi.org/10.1080/22797254.2024.2352386

R. van Berloo, Hutten, B., Eck, van, & Visser, F. (2007). An Online Potato Pedigree Database Resource. *Potato Research*, *50*(1), 45–57. https://doi.org/10.1007/s11540-007-9028-3

Shobana, M., S, V., C, G. P., SP, P. K., P, M. K., C, N., & N, K. (2022). Plant Disease Detection Using Convolution Neural Network. *2022 International Conference on Computer Communication and Informatics (ICCCI)*, *1*(2). https://doi.org/10.1109/iccci54379.2022.9740975

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020). *MPNet: Masked and Permuted Pre-training for Language Understanding*. ArXiv.org. https://arxiv.org/abs/2004.09297v2

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models*. https://arxiv.org/pdf/2302.13971

Winastwan, R. (2024). *All-Mpnet-Base-V2: Enhancing Sentence Embedding with AI - Zilliz blog*. Zilliz.com. https://zilliz.com/learn/all-mpnet-base-v2-enhancing-sentence-embedding-with-ai

Xie, W., Zhao, M., Liu, Y., Yang, D., Huang, K., Fan, C., & Wang, Z. (2024). Recent advances in Transformer technology for agriculture: A comprehensive survey. *Engineering Applications of Artificial Intelligence*, *138*(1), 109412. https://doi.org/10.1016/j.engappai.2024.109412