LLM Experimentation Report

Step 3: Model Explainability using SHAP & LIME

In this step, we used SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations)
to interpret our trained model's predictions.

1. SHAP Explanation

- SHAP provides insights into feature importance by computing the contribution of each feature in the model's decision-making.
- We generated SHAP summary plots and bar charts to visualize which features had the most impact.
- This helps us understand how different symptoms influence mental health condition predictions.

2. LIME Explanation

- LIME helps interpret individual predictions by approximating the model with an interpretable one.
- We selected a random test sample and used LIME to see which features contributed most to that specific prediction.
- This approach is useful for debugging and understanding local model behavior.

Key Findings:

- Features such as "work interference," "family history," and "benefits" had a significant impact on predictions.
- SHAP visualizations confirmed the importance of workplace-related factors in mental health

conditions.

- LIME provided insight into individual cases, explaining why a particular prediction was made.

Conclusion:

Using SHAP and LIME, we gained deeper interpretability of our model. These techniques help ensure that our AI-based mental health model is transparent and trustworthy.