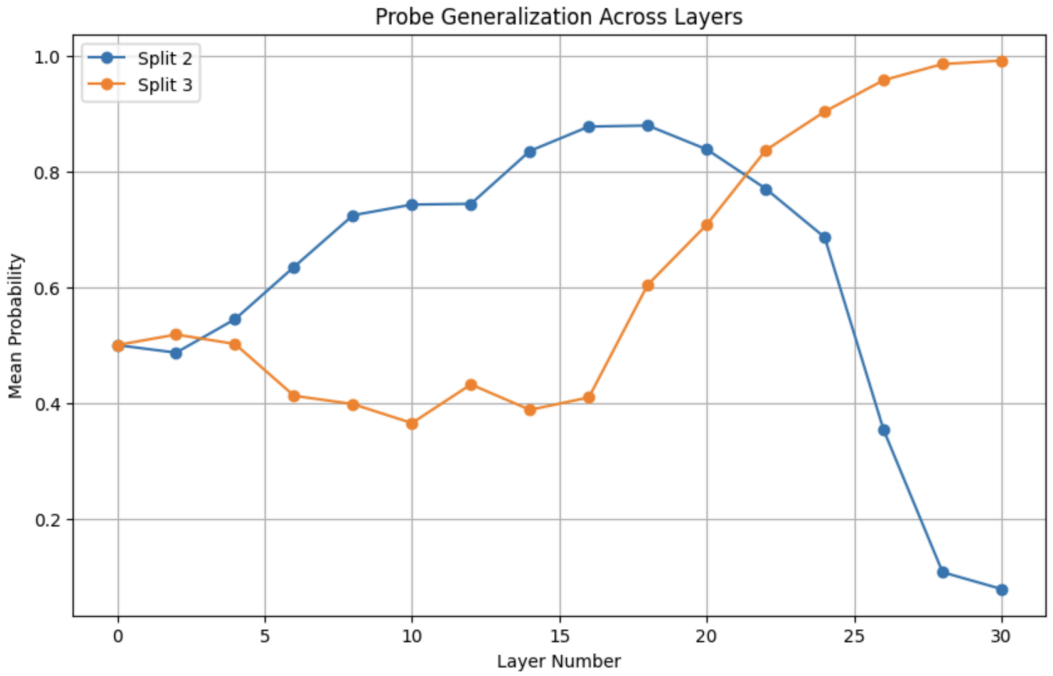
Report on LLM Probe Generalization

Experiment Overview: The objective of this experiment was to train a probe that could generalize based on either the input sentence's subject or its language. I varied the layer at which the probe is inserted into the language model to observe changes in generalization performance.

Access Notebook [here]

Experiment Results:



Layer-wise Analysis:

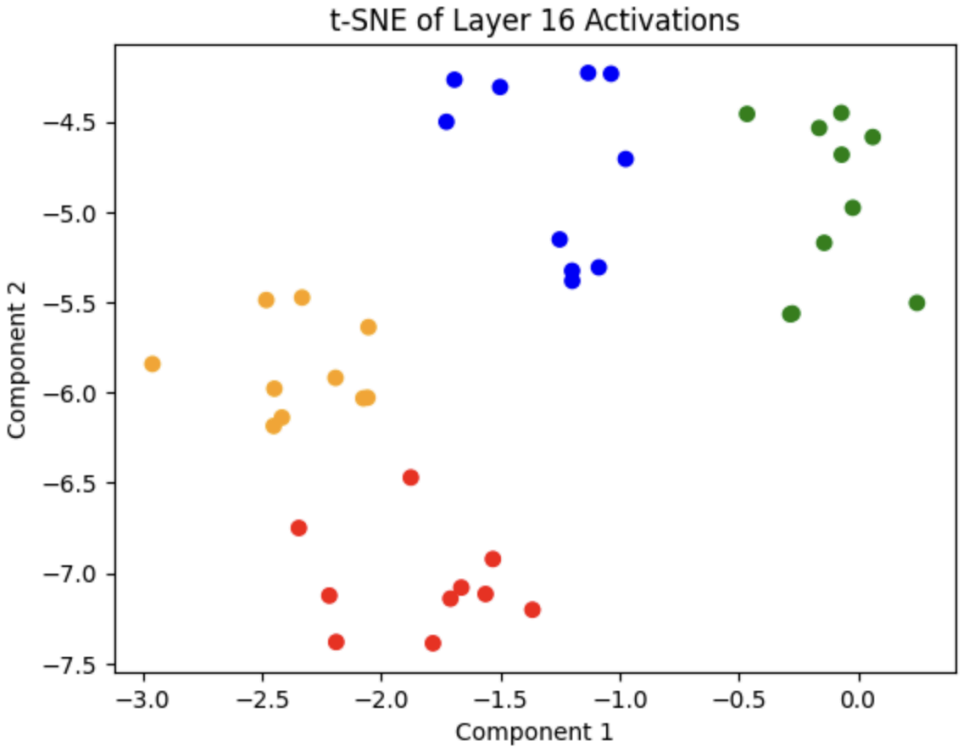
- **Rising Trend:** Both test splits show an increase in mean probability through the middle layers. Split 2 (Weather in English) peaks around layer 18, while Split 3 (Economics in Spanish) continues to rise.
- **Split Divergence:**
    - **Split 2 (Weather in English):** Shows a decline after layer 18, suggesting a loss in generalization for weather content in a different language.
    - **Split 3 (Economics in Spanish):** Increases, indicating a focus on linguistic features specific to Spanish in deeper layers.

Conclusion:

- **Optimal Layer for Generalization:** Around layer 22, where the model balances subject matter and language features, making it optimal for cross-lingual tasks.
- **Deep Layer Specialization:** Deeper layers specialize in linguistic features over subject matter, beneficial for language-specific tasks but less so for subject matter generalization.

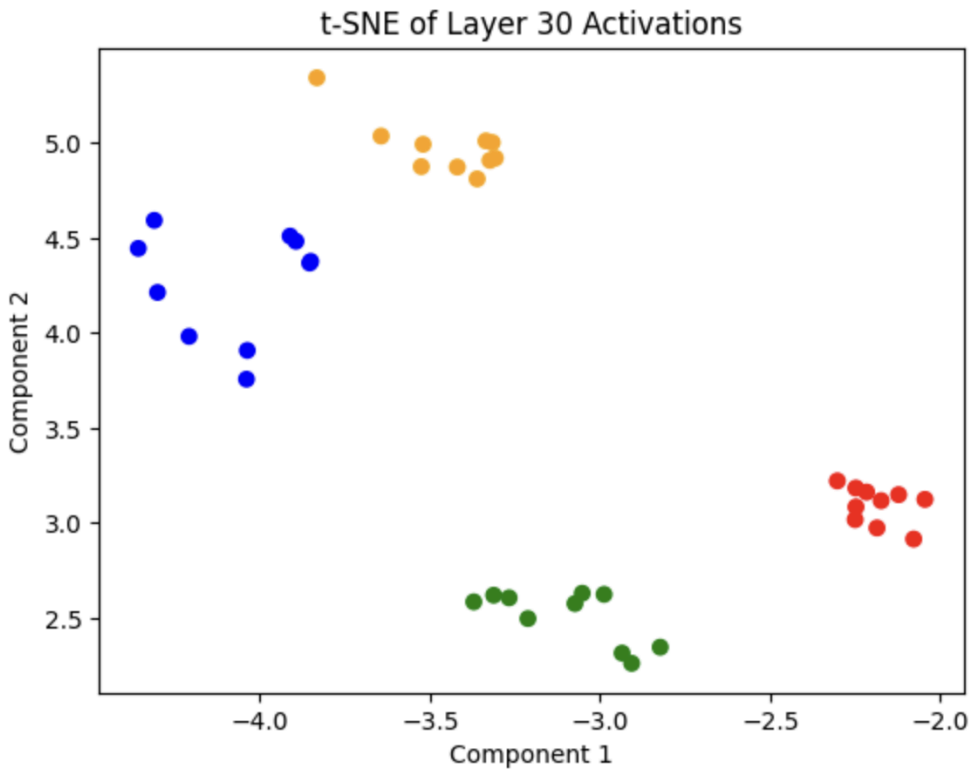
t-SNE Plots Analysis:

1. **Layer 16 Activations:**



   - **Observation:** Clusters appear more intermixed but with a discernible pattern where blue (weather in Spanish) and green (weather in English) are closer together. This suggests a commonality based on subject matter (weather).

2. **Layer 30 Activations:**



   - **Observation:** The separation is clearer, with green (weather in English) and red (economics in English) forming one group, distinct from orange (weather in Spanish) and blue (economics in Spanish). This shows a grouping based on language.

Justification for choice of layers for t-SNE plots:

- **Layer 16:** Chosen to visualize mid-level representations, where the model begins to differentiate features based on the subject matter.
- **Layer 30:** Selected to illustrate deeper layer specialization, showing clear distinctions based on language.

Conclusion:

- **Divergence of Splits Later On:** It was noted that as we probe deeper layers, the model's ability to generalize based on language becomes more pronounced. This suggests that deeper layers capture more linguistic features.
- **t-SNE Plots Reinforce Quantitative Findings:** The visual separation in t-SNE plots aligns with the quantitative results from logistic regression, providing a clear visual confirmation of how the model distinguishes between different languages and subjects at various layers.

Time Taken: Around 1.5 hours

Difficulties Encountered:

- **Access to the Model:** Initial access to the model took a few minutes.
- **RAM Issues:** Encountered RAM limitations, needed an upgrade to Colab Pro.

Improvements Made:

1. **Visualization for Logistic Regression Probabilities:** Added detailed visualizations to better illustrate the probabilities predicted by the logistic regression model.
2. **Enhanced Explanation and Analysis:** Provided more in-depth explanations and analysis of the results.
3. **Cluster Plots for Visualization:** Added t-SNE cluster plots to visually confirm the findings from the logistic regression analysis.