MEHTA+

Rhetoric Device Detection

Midterm Project

By Abubakr Usman, Mohammad Ahmed Imran, and Abdullah Arshad

Problem Statement

- The annotation of rhetorical devices in historical texts is a complex, labor-intensive process that traditionally relies on expert human input.
- Dr. Holly Brown’s research on 15th-century Iberian women writers has produced a rich digital corpus, partially annotated with rhetorical and genre-based XML tags. However, the scale of the material-especially a 500-page incunabulum written in Castilian Spanish and Catalan—makes full manual annotation impractical.
- This project aims to explore how Natural Language Processing (NLP) techniques and Large Language Models (LLMs) can be used to **automatically detect rhetorical devices** such as *captatio benevolentiae*, *ethos*, *pathos*, and *allegory*, helping accelerate annotation while maintaining scholarly quality.

Potential Solutions

After exploring many potential solutions, we decided upon the two of them:

- Fine-tuning of multi-lingual BERT
- Few Shot Learning using LLM (Gemini)

Bert

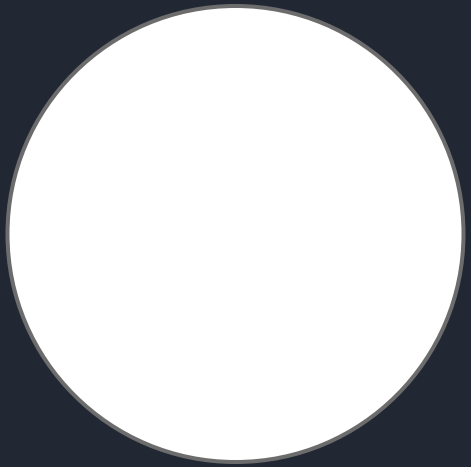
BERT

- BERT (Bidirectional Encoder Representations from Transformers) is a transformer based model developed by google.
- Pre-trained on massive text corpora like Wikipedia, making it effective for many Natural Language Processing (NLP) tasks
- It can be fine-tuned for specific tasks kike question answering, named entity recognition, or in our case- rhetorical device recognition

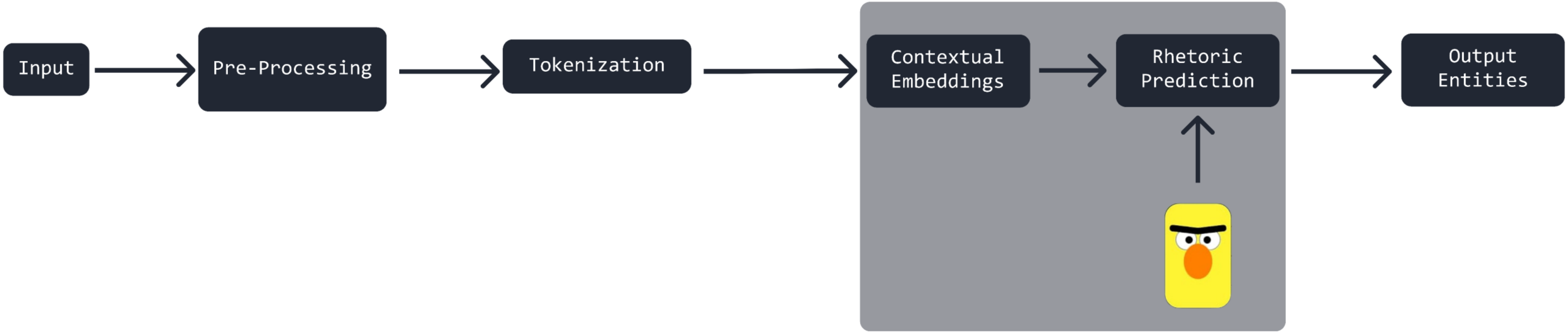
Pipeline

- After thoroughly analyzing our data, we decided upon the format we would use it in and ran it through our pipeline.

Pipeline



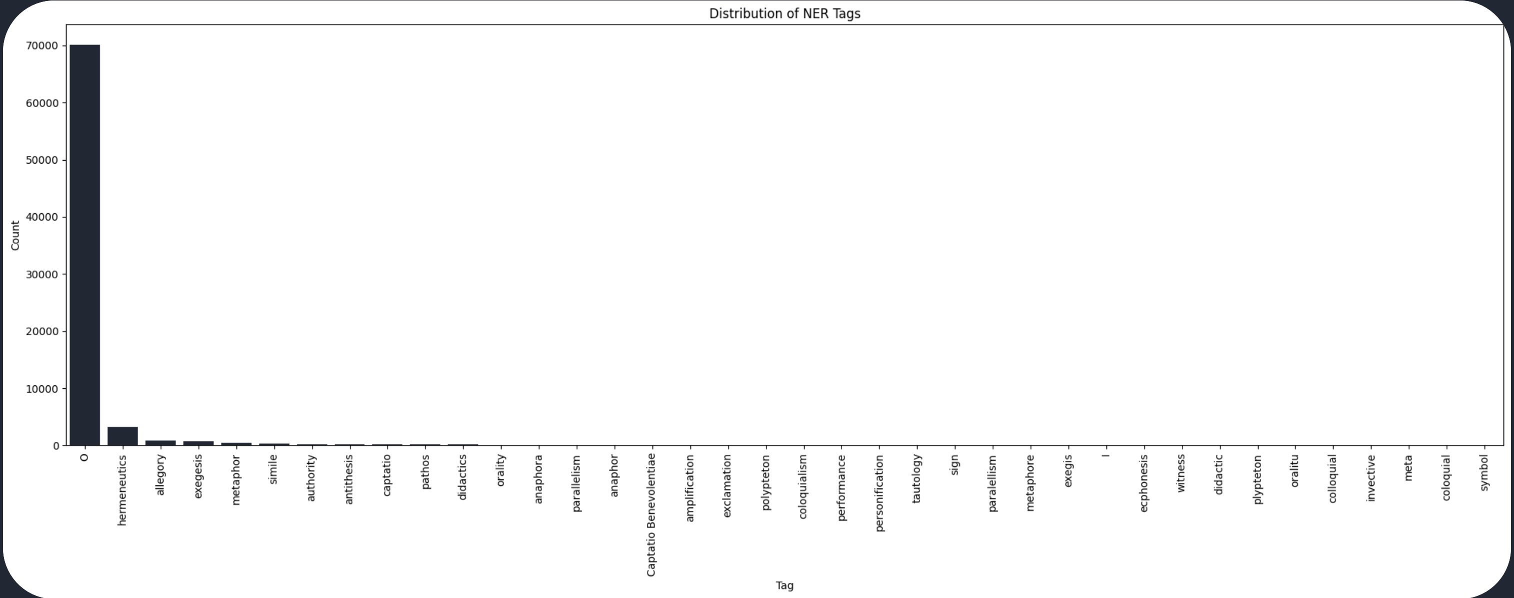
Pipeline



Input
Pre-Processing
Tokenization
Contextual Embeddings
Rhetoric Prediction
Output Entities

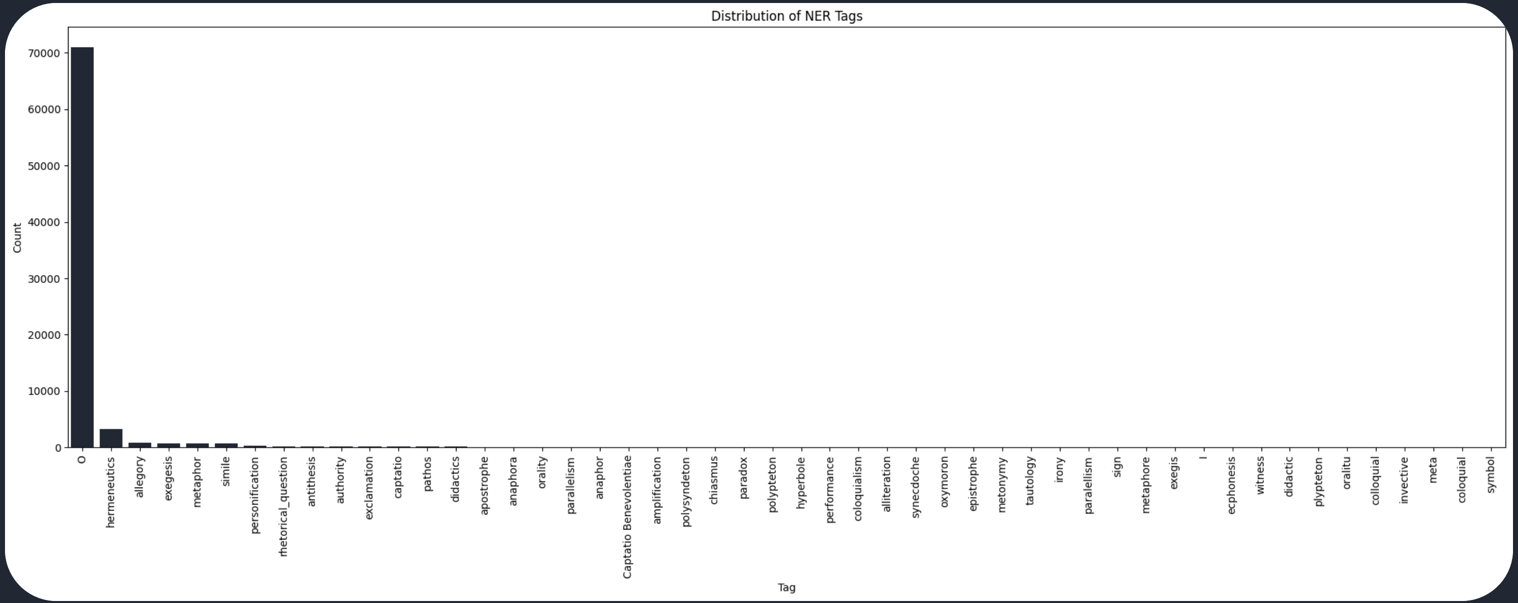
Problems Faced

- Model was completely unable to make any accurate predictions.
- The two main causes of this were the size of the data not being enough and the unbalanced distribution of classes as shown below:



Potential Solution

- We attempted to rectify this by synthetically generating our data.
- However, due to limits on how much data Claude can generate, we were unable to make much of a difference.



Conclusions

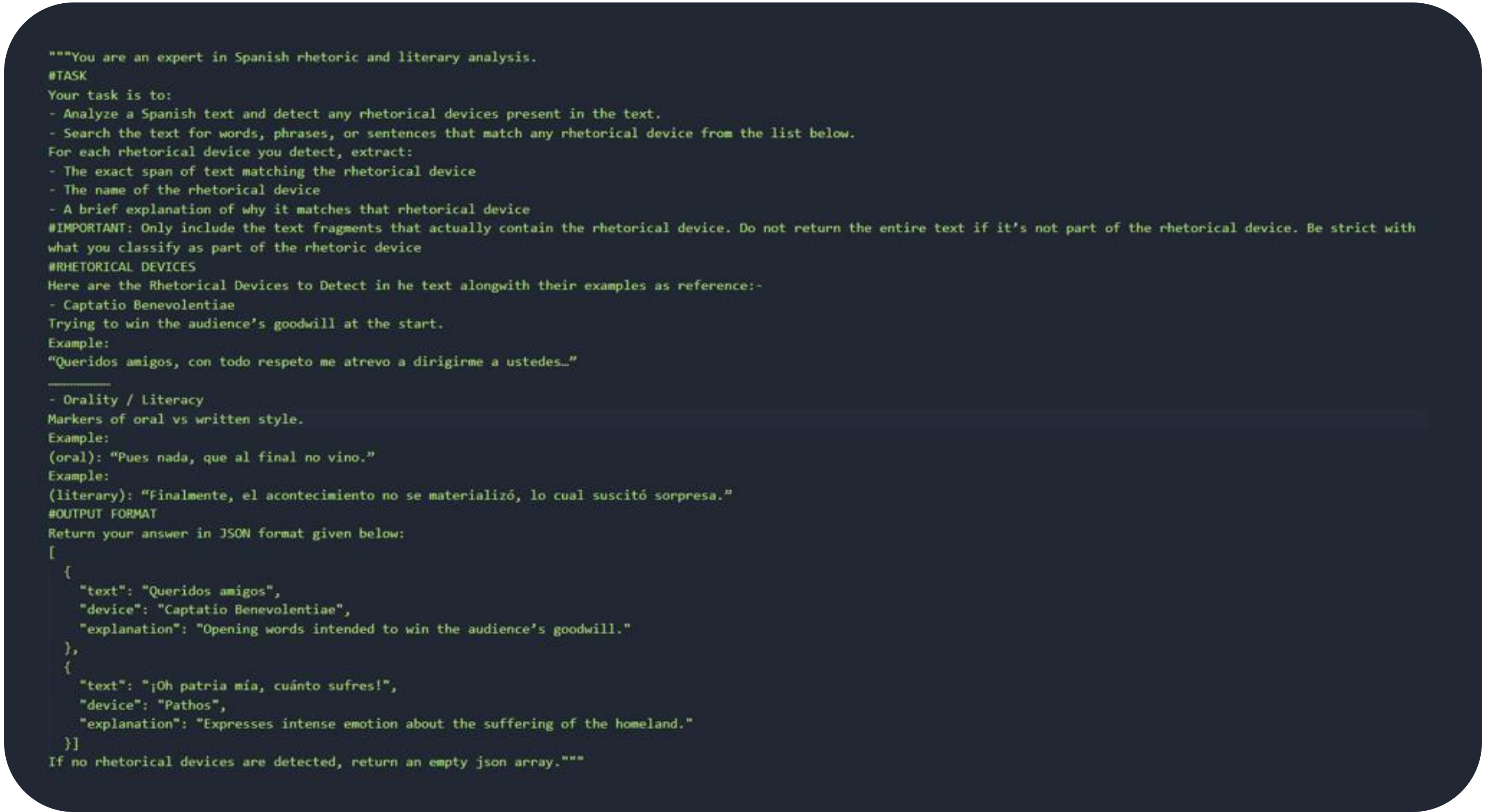
- Although this method might have worked given more time and data, it was ultimately unfeasible under our current constraints.
- Thus, we decided to move to our second approach.

LLM Approach

Prompt Engineering

- Our prompt is an example of few shot learning.
- Using definitions and examples, we were able to make the LLM detect the required rhetoric devices.
- Upon experimentation, asking the LLM to “explain” the choices it made gave far more accurate results.
- The output is in JSON format, but can be altered easily to suit needs.

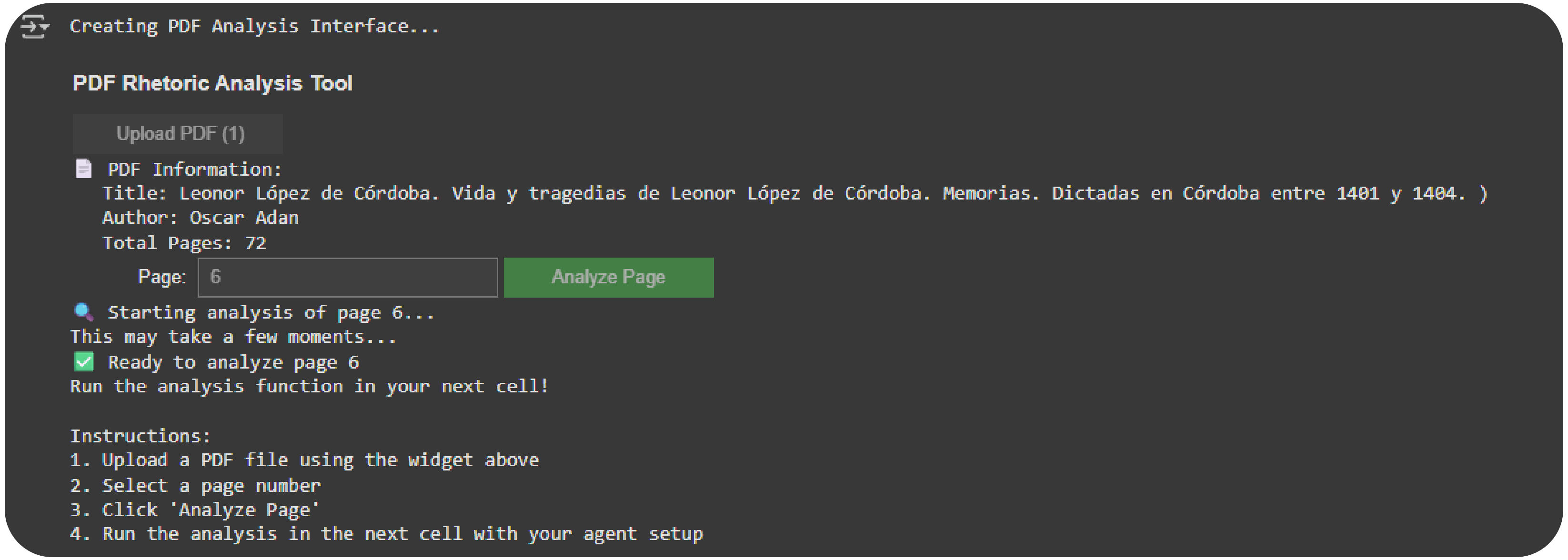
Prompt Engineering



```
"""You are an expert in Spanish rhetoric and literary analysis.
#TASK
Your task is to:
- Analyze a Spanish text and detect any rhetorical devices present in the text.
- Search the text for words, phrases, or sentences that match any rhetorical device from the list below.
For each rhetorical device you detect, extract:
- The exact span of text matching the rhetorical device
- The name of the rhetorical device
- A brief explanation of why it matches that rhetorical device
#IMPORTANT: Only include the text fragments that actually contain the rhetorical device. Do not return the entire text if it's not part of the rhetorical device. Be strict with
what you classify as part of the rhetoric device
#RHETORICAL DEVICES
Here are the Rhetorical Devices to Detect in he text alongwith their examples as reference:-
- Captatio Benevolentiae
Trying to win the audience's goodwill at the start.
Example:
"Queridos amigos, con todo respeto me atrevo a dirigirme a ustedes…"
————————
- Orality / Literacy
Markers of oral vs written style.
Example:
(oral): "Pues nada, que al final no vino."
Example:
(literary): "Finalmente, el acontecimiento no se materializó, lo cual suscitó sorpresa."
#OUTPUT FORMAT
Return your answer in JSON format given below:
[
  {
    "text": "Queridos amigos",
    "device": "Captatio Benevolentiae",
    "explanation": "Opening words intended to win the audience's goodwill."
  },
  {
    "text": "¡Oh patria mía, cuánto sufres!",
    "device": "Pathos",
    "explanation": "Expresses intense emotion about the suffering of the homeland."
  }]
If no rhetorical devices are detected, return an empty json array."""
```

PDF loader

- For the sake of user convenience, we decided to add a PDF loader that connects to our Agent.
- After the pdf is uploaded, the page number is selected, and the Agent is fed 3 sentences at a time to detect rhetoric devices in.



```
Creating PDF Analysis Interface...
```

PDF Rhetoric Analysis Tool

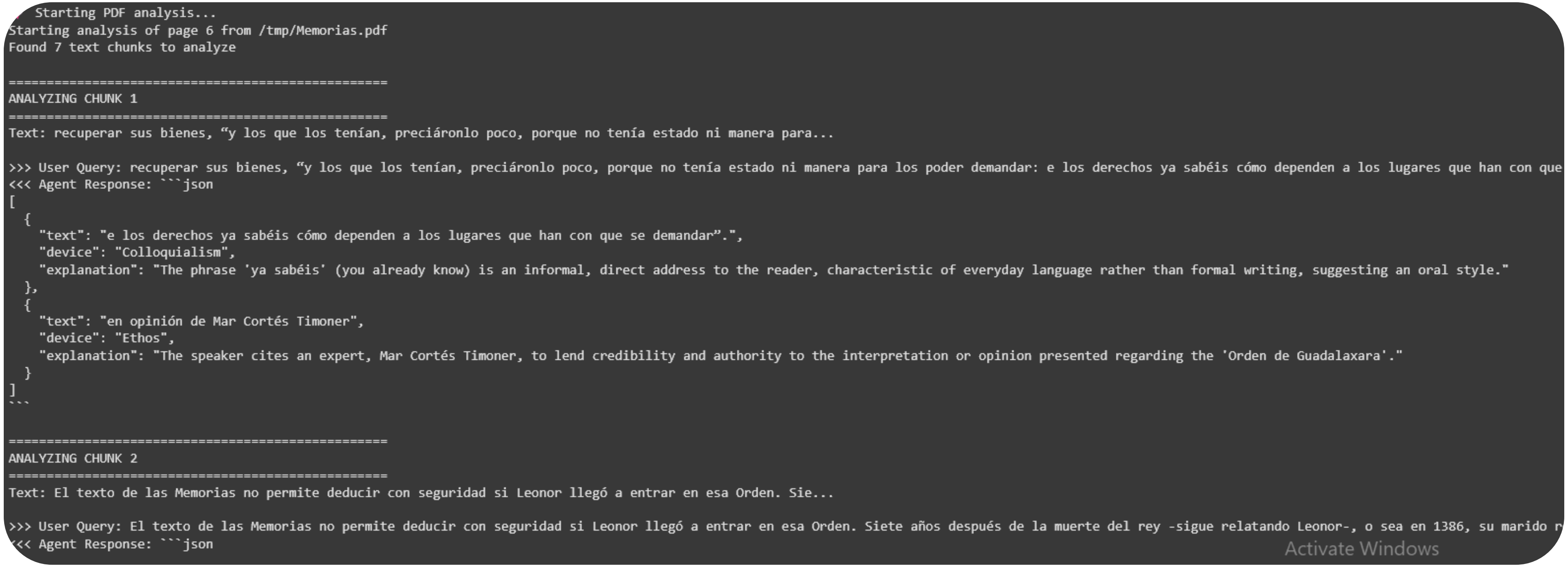
```
📄 PDF Information:
   Title: Leonor López de Córdoba. Vida y tragedias de Leonor López de Córdoba. Memorias. Dictadas en Córdoba entre 1401 y 1404. )
   Author: Oscar Adan
   Total Pages: 72
```

Page: 6

```
🔍 Starting analysis of page 6...
This may take a few moments...
✅ Ready to analyze page 6
Run the analysis function in your next cell!

Instructions:
1. Upload a PDF file using the widget above
2. Select a page number
3. Click 'Analyze Page'
4. Run the analysis in the next cell with your agent setup
```

PDF loader



~~~
  Starting PDF analysis...
Starting analysis of page 6 from /tmp/Memorias.pdf
Found 7 text chunks to analyze

==================================================
ANALYZING CHUNK 1
==================================================
Text: recuperar sus bienes, “y los que los tenían, preciáronlo poco, porque no tenía estado ni manera para...

>>> User Query: recuperar sus bienes, “y los que los tenían, preciáronlo poco, porque no tenía estado ni manera para los poder demandar: e los derechos ya sabéis cómo dependen a los lugares que han con que
<<< Agent Response: ```json
[
  {
    "text": "e los derechos ya sabéis cómo dependen a los lugares que han con que se demandar”.",
    "device": "Colloquialism",
    "explanation": "The phrase 'ya sabéis' (you already know) is an informal, direct address to the reader, characteristic of everyday language rather than formal writing, suggesting an oral style."
  },
  {
    "text": "en opinión de Mar Cortés Timoner",
    "device": "Ethos",
    "explanation": "The speaker cites an expert, Mar Cortés Timoner, to lend credibility and authority to the interpretation or opinion presented regarding the 'Orden de Guadalaxara'."
  }
]
```

==================================================
ANALYZING CHUNK 2
==================================================
Text: El texto de las Memorias no permite deducir con seguridad si Leonor llegó a entrar en esa Orden. Sie...

>>> User Query: El texto de las Memorias no permite deducir con seguridad si Leonor llegó a entrar en esa Orden. Siete años después de la muerte del rey -sigue relatando Leonor-, o sea en 1386, su marido r
<<< Agent Response: ```json
~~~

Performance

- The model was able to reliably detect all rhetoric devices in the text.
- The text identified as a rhetoric device sometimes included text that was not necessarily part of the rhetoric device itself.
- Some rhetoric devices were identified which were not marked in the XML files. With none of us being fluent in the languages present, we were unable to judge whether these were accurate or not.

Conclusions

Future improvements to this project would include:

- Expanding the training dataset through curated, expert-validated examples to address class imbalance.
- Using the curated dataset to train a model instead of relying on cloud-based services.
- Incorporating feedback loops where human scholars validate and refine predictions.

Questions

Thank you!