Finding the Plea: Evaluating the Ability of LLMs to Identify Rhetorical Structure in Swedish and English Historical Petitions

Ellinor Lindqvist **Eva Pettersson** **Joakim Nivre**
Uppsala University
Dept. of Linguistics and Philology
firstname.lastname@lingfil.uu.se

Abstract

Large language models (LLMs) have shown impressive capabilities across many NLP tasks, but their effectiveness on fine-grained content annotation, especially for historical texts, remains underexplored. This study investigates how well GPT-4, Gemini, Mixtral, Mistral, and LLaMA can identify rhetorical sections (*Salutatio*, *Petitio*, and *Conclusio*) in 100 English and 100 Swedish petitions using few-shot prompting with varying levels of detail. Most models perform very well, achieving F1 scores in the high 90s for *Salutatio*, though *Petitio* and *Conclusio* prove more challenging, particularly for smaller models and Swedish data. Cross-lingual prompting yields mixed results, and models generally underestimate document difficulty. These findings demonstrate the strong potential of LLMs for assisting with nuanced historical annotation while highlighting areas for further investigation.

1 Introduction

In recent years, large language models (LLMs) have demonstrated remarkable capabilities across many NLP tasks, including translation, summarisation, and question answering. However, their performance on fine-grained, content-aware text annotation tasks, particularly those involving historical texts and moderately resourced languages, remains a relatively unexamined area.

In this paper, we investigate to what extent LLMs can be used to analyse and annotate a specific type of historical document, the petition. In pre-modern and pre-democratic societies, petitions allowed ordinary people to seek redress or support from those in positions of authority — such as courts, parliaments, landlords, or monarchs (Houston, 2014). Despite their potential to shed light on the everyday lives, concerns, and ways of navigating authority in the past, petitions have been relatively neglected in both historical and computational research.

The interdisciplinary project *Speaking to One's Superiors: Petitions as Cultural Heritage and Sources of Knowledge*, led by Uppsala University's *Gender and Work* (GaW) research project and funded by the Swedish Research Council, investigates 18th-century Swedish petitions.[1] Thousands of documents have been digitised, annotated, and made publicly accessible to shed light on how women and men in Early Modern Sweden made a living and asserted their rights.

Petitions in early modern Europe often followed a classical rhetorical structure, typically divided into five sections: *Salutatio*, *Exordium*, *Narratio* (including *Argumentatio*), *Petitio*, and *Conclusio* (Dodd, 2011; Israelsson, 2016). This study explores the use of LLMs to automatically identify such sections, with the aim of supporting information extraction for historians and other scholars working with petitions. We focus in particular on three key components: the greeting (*Salutatio*), the request (*Petitio*), and the ending (*Conclusio*).

To support this work, we have created a dataset of 100 historical petitions in Swedish and 100 in English, each annotated to mark the locations of the targeted rhetorical sections. Each document is also assigned a difficulty score, reflecting the level of annotation difficulty, allowing us to evaluate the relationship between model confidence and human-perceived difficulty, for a more nuanced assessment of LLM performance on complex historical texts.

With our experiments, we evaluate how effectively LLMs can find and annotate rhetorical components in historical petitions using few-shot prompting. We have four research goals:

1. Comparing model performance across a range of commercial and open-source LLMs.
2. Studying how prompt complexity (less vs. more detailed instructions) and one vs three

[1] https://gaw.hist.uu.se/petitions/

Proceedings of the First Workshop NLP and Language Models for Digital Humanities associated with RANLP 2025, pages 86–101, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-106-6-008

output examples affect annotation accuracy.

3. Assessing cross-linguistic generalisation by testing both English and Swedish, and varying the prompt language used on Swedish data.
4. Comparing human and model uncertainty by analysing how well the performance of the models and their self-assigned difficulty scores align with human difficulty ratings.

The models under investigation include a diverse mix of architectures and scales: GPT-4 (OpenAI), Gemini 1.5 Pro (Google DeepMind), Mistral 7B and Mixtral 8x22B (Mistral AI), and LLaMA 3 (Meta AI) in both 8B and 70B configurations. We evaluate the model outputs and also analyse how closely model confidence aligns with human difficulty judgments. By combining comparative evaluation, prompt design variation, and confidence modeling, this work aims to illuminate both the capabilities and the limitations of LLMs in performing nuanced annotation tasks on historical texts across diverse settings.

2 Related Work

Applying LLMs to structured tasks like rhetorical analysis depends critically on methods used to guide the model's output, a field broadly known as prompt engineering. Sahoo et al. (2024) emphasise that obtaining accurate and structured information from LLMs is a non-trivial challenge that requires carefully designed interaction strategies. This can be argued to be particularly true for annotation and extraction tasks, where the desired output is not free-form text but a structured representation. Cheng et al. (2024) address this for Named Entity Recognition (NER) by proposing a standardised prompting method. They demonstrate that a combination of a clear task definition, illustrative few-shot examples, and a strict output format specification is essential for improving the reliability of structured data extraction in a few-shot context. In the context of rhetorical analysis, Maekawa et al. (2024) tackle discourse parsing by translating traditional parsing strategies into effective prompts for a decoder-only LLM. By combining this with parameter-efficient fine-tuning (QLoRA), they achieve state-of-the-art results with strong generalisation, demonstrating that LLMs can model complex rhetorical hierarchies.

The potential of using LLMs to process historical data is gaining attention, offering informative new tools for fields from behavioral science to the digital humanities (Varnum et al., 2024). This development has led to the application of LLMs across all phases of historical research. At the most foundational level, researchers are using LLMs to overcome long-standing barriers, such as transcribing handwritten historical documents to unlock previously inaccessible archives (Humphries et al., 2025). Moving beyond data preparation to analysis, Cohen et al. (2025) investigate the potential of BERT and GPT-4o models to detect irony in 19th-century Latin American newspaper texts, demonstrating how LLMs can be used in context-dependent tasks given historical linguistic changes. Overall, this body of research indicates that LLMs could be highly effective for automated rhetorical annotation of historical texts, a task that to the best of our knowledge has not been explored previously.

Test Set	Period	# Docs	# Toks	Avg Toks/Doc
Swedish	1709–1800	100	24,904	249 ± 116
English	1692–1799	100	28,831	288 ± 172

Table 1: Overview of the Swedish and English test sets: time period, document count, total token count, average and standard deviation of tokens per document.

3 The Petition Data Sets

An overview of the test set statistics is presented in Table 1. Below, we describe the composition and annotation process for each dataset in more detail.

3.1 The Swedish Data Set

The Swedish petition data set consists of 100 petitions from the 18th century, transcribed by a historian. These petitions were originally written between 1719 and 1800 and submitted to the regional administration in Örebro, Sweden. We also make use of an additional 10 petitions as a development set, used when developing the code and prompts to our experiments.

3.2 The English Data Set

The English dataset is drawn from the *London Lives 1690–1800* archive[2] (Hitchcock et al., 2012), a large digital collection of legal and social records focusing on everyday Londoners. We use a digitised subset of these materials curated for the London Lives Petitions Project (Howard, 2016).[3] The

[2] https://www.londonlives.org/

[3] https://github.com/sharonhoward/llpp?tab=readme-ov-file

whole digitised collection includes around 10,000 petitioning documents submitted to magistrates and courts, from which we select petitions addressed to the courts of the Old Bailey and Middlesex Sessions and City of London. The petitions were originally transcribed using a double rekeying process, where two (non-academic) typists transcribe text, the two versions are compared and only discrepancies are manually checked. We randomly select 100 petitions from this collection in a stratified manner based on court for our English test set and 10 petitions for a development set.

3.3 Rhetorical Structure of Petitions

In many parts of premodern Europe, the structure of petitions followed a classical rhetorical framework, typically comprising five or six sections (Hansson, 1988; Sokoll, 2006; Israelsson, 2016):

1. *Salutatio*: Formal salutation to the addressee.
2. *Exordium*: Brief opening phrase appealing to the recipient's greatness or capacity to help.
3. *Narratio*: Narration of the circumstances leading to the petition, often mixed with arguments (4. *Argumentatio*).
5. *Petitio*: Specific request or plea being made.
6. *Conclusio*: Final phrase(s) of courtesy and/or inferiority, often including a signature.

In this study, we focus both manual annotation and model evaluation on the sections *Salutatio*, *Petitio*, and *Conclusio*. The *Exordium* is excluded, as it is typically a brief phrase, may be absent from some texts, and its identification is often more subjective. The sections *Narratio* and *Argumentatio* are likewise omitted, as they are frequently intertwined and difficult to distinguish reliably. As a result, these parts remain unannotated, and the majority of unmarked content in the corpus should correspond to one or both of these rhetorical functions. Our experiments thus primarily test the ability of language models to identify the three selected sections.

3.4 Annotated Gold Data Sets

To identify the targeted rhetorical sections in both the Swedish and English datasets, we manually inserted start and end tags for each section. Three annotators carried out the work, with each petition annotated by two of them. The data was divided into batches, and after each round, the specific disagreements were resolved and general principles agreed upon to support consistency in later batches.

Swedish Dataset				
Diff	Section	%Exact	κ	TokDist
1.60	Overall	48.0	0.82	5.76
	Salutatio	100.0	1.00	0.00
	Petitio	48.0	0.48	16.70
	Conclusio	98.0	0.92	0.60

English Dataset				
Diff	Section	%Exact	κ	TokDist
1.49	Overall	68.0	0.88	2.15
	Salutatio	99.0	0.99	0.02
	Petitio	69.0	0.68	5.84
	Conclusio	95.0	0.95	0.59

Table 2: Inter-annotator agreement for Swedish and English datasets. Diff = average difficulty score, %Exact = percent exact matches, κ = Cohen's kappa, TokDist = mean token distance.

During the annotation process, each document was also assigned a difficulty score ranging from 0 to 2. Scoring was based on annotator agreement to reflect the level of annotation difficulty. Documents that received a score of 1 often exhibit mild ambiguities, such as blended rhetorical sections or unusual phrasing, leading to minor disagreements, which were typically resolved quickly. A score of 2 was assigned to cases that required extended discussion to resolve disagreement. These documents often present interpretive challenges due to older/non-standard orthography, incomplete phrases, or heavy use of abbreviations, which complicates clear identification of rhetorical boundaries. In particular, separating *Petitio* from *Narratio/Argumentatio* was frequently experienced as challenging in these cases. Examples of annotation agreements and the provided difficulty scores can be found in the Appendix. By including difficulty annotations, we can assess if and how model confidence aligns with human-perceived complexity thereby enriching the evaluation of LLMs on historically and linguistically complex texts.

Table 2 presents the average difficulty scores and inter-annotator agreement scores for each dataset. The exact match score (%Exact) measures the percentage of petitions or sections where the two annotations are token identical, while the κ score is Cohen's kappa. The token distance measure (TokDist) is the average number of tokens that differ between the two annotations. It is worth noting that *Petitio* emerges as the most challenging petition segment to annotate, as indicated by its lowest percentage of identical tags and Cohen's kappa scores, as well as the highest mean token distance scores. This is particularly evident in the Swedish dataset.

4 Method

With our experiments, we aim to evaluate how well LLMs can annotate rhetorical sections of historical petitions using few-shot prompting. The key components of our method are presented below.

Assess model performance on the annotation task across several leading LLMs, including both commercial and open-source systems.

Investigate the role of prompt design by comparing less vs. more detailed instructions and by providing either one or three output examples to understand how prompt complexity influences tagging accuracy.

Evaluate cross-linguistic generalisation by comparing results on English (a high-resource language) and Swedish (a moderately resourced language in the LLM training ecosystem). To further explore cross-lingual effects, we prompt the Swedish data set using both Swedish and English instructions in the prompts (apart from the given output examples), assessing how the prompt language influences model tagging performance.

Compare human and model uncertainty by assessing how well models can self-estimate uncertainty in comparison to human judgments of difficulty. Each text in the dataset has not only been annotated for rhetorical sections but has also been assigned a difficulty score by the human annotators, on a scale from 0 (easy) to 2 (difficult) (see more details in Section 3.4). To compare with human difficulty judgments, we instruct the language models to return a difficulty score alongside each predicted annotation.

4.1 Prompting Settings and Variations

To test whether and how model performance is affected by prompt design, we developed three prompt variations:

Prompt 1: short 1-shot This prompt includes a less detailed description of the task, a list of the tags to be used, the required output format, and one example output showing an annotated petition.

Prompt 2: long 1-shot Similar to Prompt 1 but with a more detailed and dataset-specific description, providing additional context and clarification about the task.

Prompt 3: long 3-shot Extends Prompt 2 by including three example outputs of annotated petitions, giving the model more extensive demonstrations of the expected tagging and formatting.

These variations were designed to evaluate how the level of detail and the number of examples influence the ability of the models to perform the tagging task accurately. For the Swedish dataset, we also examine whether the language of the prompt influences model performance by testing each prompt in both Swedish and English. Examples of prompts for both datasets can be found in the Appendix.

4.2 Models

We evaluate six contemporary LLMs with varying architectures and sizes: GPT-4 (Achiam et al., 2023) from OpenAI, Gemini Pro 1.5 (Team et al., 2024) from Google DeepMind, two LLaMA open-weight transformers from Meta AI in both 8B and 70B configurations (Touvron et al., 2023) and the open-source models Mixtral 8x22B (Jiang et al., 2024) and Mistral 7B (Jiang et al., 2023) from Mistral AI. All models are accessed via APIs (e.g., OpenAI, Google, Mistral), and we ensure consistent prompt formatting and settings across runs for comparability. To promote deterministic generation and reproducibility, all prompts are run with a temperature setting of 0.

4.3 Evaluation Procedure

To evaluate the performance of the LLMs on the rhetorical annotation task, we make use of a unified evaluation framework applicable across all models and languages. Evaluation is conducted separately on English and Swedish data sets to compare model performance on a high-resource versus a moderately resourced language. For the Swedish data set, we evaluate both the use of Swedish instructions and English instructions. We perform an evaluation per prompt type, to analyse the impact of prompt complexity by comparing shorter versus more detailed instructions.

Annotation The models are instructed to annotate texts using predefined rhetorical tags, as described in Section 3.4. We evaluate model predic-

Results for the English Dataset

Prompt	Model	salutatio				petitio				conclusio			
		P	R	F1	TD	P	R	F1	TD	P	R	F1	TD
short 1-shot prompt	GPT-4	68.4	**100.0**	81.2	0.46	94.9	97.3	96.1	0.11	97.2	97.6	**97.4**	0.03
	Gemini	95.5	**100.0**	97.7	0.06	94.7	97.2	96.0	0.12	98.2	91.0	94.5	0.06
	Mixtral	98.5	**100.0**	99.2	0.03	91.2	96.7	93.9	0.17	89.3	91.6	90.5	0.05
	Mistral	98.5	99.9	99.2	0.03	74.3	86.0	79.8	0.40	87.7	72.6	79.4	0.23
	LLaMA 70B	99.7	**100.0**	99.8	0.02	95.0	96.7	95.8	0.13	89.3	91.9	90.6	0.09
	LLaMA 8B	98.8	99.9	99.4	0.02	90.6	90.5	90.5	0.18	64.4	21.2	31.8	0.68
long 1-shot prompt	GPT-4	77.9	97.7	86.7	0.36	96.1	97.1	96.6	0.09	91.6	**98.7**	95.0	0.03
	Gemini	98.6	**100.0**	99.3	0.02	96.3	96.1	96.2	0.15	98.1	89.4	93.6	0.04
	Mixtral	98.8	**100.0**	99.4	0.03	89.7	96.5	93.0	0.25	82.2	91.9	86.8	0.07
	Mistral	98.5	99.9	99.2	0.03	68.3	92.3	78.5	0.43	89.4	84.8	87.1	0.11
	LLaMA 70B	99.4	**100.0**	99.7	0.03	96.4	97.1	96.7	0.18	97.1	94.3	95.7	0.06
	LLaMA 8B	98.9	**100.0**	99.5	0.03	88.7	93.1	90.9	0.19	74.3	49.7	59.6	0.48
long 3-shot prompt	GPT-4	89.8	**100.0**	94.6	0.15	95.6	**97.8**	96.7	0.08	89.9	97.5	93.5	0.04
	Gemini	99.6	**100.0**	99.8	0.01	97.2	94.4	95.8	0.16	**100.0**	83.6	91.0	0.06
	Mixtral	98.1	**100.0**	99.0	0.03	90.6	96.2	93.3	0.29	87.0	84.8	85.9	0.06
	Mistral	98.5	99.9	99.2	0.03	59.4	91.1	71.9	0.53	83.9	63.7	72.4	0.21
	LLaMA 70B	**99.8**	**100.0**	**99.9**	0.02	**97.3**	97.5	**97.4**	0.15	81.7	95.1	87.9	0.09
	LLaMA 8B	98.3	99.9	99.1	0.02	86.4	94.5	90.3	0.25	88.2	67.6	76.6	0.25

Table 3: Results for English data across three prompt types and six models. Scores includes token-level precision (P), recall (R), and F1 in percentage, and mean token-level edit distance for each predicted span.

tions against gold annotations using two metrics. First, we compute token-level precision, recall, and F1-score by collecting all tokens that occur inside predicted spans and comparing them to all tokens inside the corresponding gold spans for each rhetorical tag. Second, we calculate the mean token-level edit distance: for each predicted span, we compute the minimum normalised edit distance (Levenshtein distance over whitespace-tokenised words) to any gold span of the same tag. Distances are averaged across all predicted spans, including perfect matches (where distance = 0). A tag that is missing in both the model prediction and in the gold annotation is scored as a perfect match.

Difficulty Estimation To evaluate the alignment between model-assigned and human-assigned difficulty ratings, we calculate the mean error (ME) as the average difference between model-predicted and human-assigned difficulty scores:

$$\text{Error} = \text{Model} - \text{Human} \qquad (1)$$

A positive mean error indicates that the model systematically rates documents as more difficult than human annotators, whereas a negative mean error indicates that the model rates documents as easier than humans do. We also use Spearman's rank correlation coefficient (ρ) (Spearman, 1904) to measure how well the order of difficulties assigned by the model agrees with the order assigned by humans.

5 Results and Discussion

The results for the English dataset are presented in Table 3, those for the Swedish dataset with Swedish prompts in Table 4, and for the Swedish dataset with English prompts in Table 5.

5.1 Results on the Annotation Task

Although results vary between specific models, prompts, and datasets, an overall view suggests that the models generally perform the task of annotating the petitions very well, with many F1 scores reaching into the high 90s. In particular, *Salutatio* shows very strong results, aligning well with the high inter-annotator agreement among human annotators.

Results for the English data For English petitions, *Salutatio* results were consistently strong. Surprisingly, although *Petitio* posed the greatest challenge for human annotators, models more often had difficulty with *Conclusio*, particularly smaller models such as LLaMA 8B and Mistral 7B, and to some extent Mixtral.

LLaMA 70B and Gemini performed consistently well across all parts. Interestingly, GPT-4, while very strong on *Petitio* and *Conclusio*, showed weaker performance on *Salutatio* than other models, often including paragraphs presenting the petitioner (which usually followed the *Salutatio* in the English dataset). By contrast, the smaller models,

Results for the Swedish Dataset - using Swedish Instructions

Prompt	Model	salutatio				petitio				conclusio			
		P	R	F1	TD	P	R	F1	TD	P	R	F1	TD
short 1-shot prompt	GPT-4	98.7	99.9	99.3	0.01	82.5	79.2	80.8	0.29	92.3	92.5	**92.4**	0.10
	Gemini	94.0	**100.0**	96.9	0.02	80.8	80.7	80.7	0.35	**100.0**	62.1	76.6	0.34
	Mixtral	93.8	99.6	96.6	0.02	64.5	66.1	65.3	0.55	79.8	64.0	71.0	0.34
	Mistral	85.3	99.8	92.0	0.04	46.4	83.9	59.7	0.67	60.5	62.1	61.3	0.39
	LLaMA 70B	98.7	99.9	99.3	0.01	82.1	83.6	82.8	0.38	83.0	88.2	85.5	0.19
	LLaMA 8B	98.8	96.1	97.5	0.14	73.8	44.9	55.8	0.58	82.6	53.1	64.6	0.49
long 1-shot prompt	GPT-4	98.7	**100.0**	99.3	0.01	85.3	80.8	83.0	0.27	85.6	93.4	89.3	0.10
	Gemini	94.0	**100.0**	96.9	0.02	**86.5**	80.1	83.2	0.31	**100.0**	68.5	81.3	0.32
	Mixtral	95.7	99.7	97.7	0.02	81.6	62.1	70.5	0.47	91.0	58.2	71.0	0.35
	Mistral	80.1	99.4	88.7	0.05	53.2	**89.3**	66.7	0.55	95.3	59.4	73.2	0.31
	LLaMA 70B	98.7	99.9	99.3	0.01	84.5	83.0	83.7	0.36	81.2	82.9	82.1	0.21
	LLaMA 8B	98.8	96.7	97.7	0.12	75.3	42.1	54.0	0.56	87.2	54.6	67.1	0.47
long 3-shot prompt	GPT-4	98.7	**100.0**	99.3	0.01	85.7	82.8	**84.2**	0.22	86.2	**95.0**	90.4	0.08
	Gemini	**100.0**	**100.0**	**100.0**	0.00	**86.5**	81.2	83.8	0.30	**100.0**	71.6	83.4	0.28
	Mixtral	94.1	99.8	96.9	0.02	79.6	57.9	67.0	0.53	86.7	51.6	64.7	0.35
	Mistral	**100.0**	99.1	99.5	0.04	47.7	82.6	60.5	0.62	88.7	38.4	53.6	0.36
	LLaMA 70B	97.6	99.8	98.7	0.01	86.2	77.8	81.8	0.36	92.0	75.7	83.0	0.25
	LLaMA 8B	98.1	97.7	97.9	0.10	61.3	52.5	56.6	0.57	72.3	71.7	72.0	0.35

Table 4: Results for Swedish data across three Swedish prompt types and six models. Scores includes token-level precision (P), recall (R), and F1 in percentage, and mean token-level edit distance for each predicted span.

though competitive on *Salutatio*, mostly underperformed on *Petitio* and *Conclusio*.

Manual inspection, focusing on *Conclusio* errors from Mixtral, LLaMA 70B, and LLaMA 8B, highlighted different sources of difficulty. Beyond minor punctuation mismatches, some models omitted tags entirely or hallucinated content, such as fabricating a full *Conclusio* where none existed in the gold annotation, or adding phrases not present in the text. LLaMA 8B's particularly low scores for *Conclusio* were further explained by malformed outputs, where tags were placed after the relevant span instead of correctly wrapping it as specified in the prompt.

Results for the Swedish data For the Swedish petitions, as with the English data, strong results were observed for *Salutatio*. Unlike in the English dataset, GPT-4 did not struggle with annotating *Salutatio* in Swedish, achieving one of the highest F1 scores among the models. However, compared to the English data, the models generally found both *Petitio* and *Conclusio* more challenging to annotate in the Swedish Petitions, though performance varied across models.

When comparing models, larger models generally outperformed smaller ones on the Swedish petitions, with GPT-4 being the top-performing model for most petition parts and prompt types, followed by Llama 70B. Gemini and Mixtral also produced several high results, whereas the smaller models, Llama 8B and Mistral, received lower scores, especially for *Petitio* and *Conclusio*.

5.2 The Effect of Prompt Complexity

Across all models and datasets, there is no consistent pattern indicating that prompt length or the number of examples (short 1-shot vs. long 1-shot vs. long 3shot) systematically influences performance. While minor differences appear for specific models or rhetorical sections, these variations do not suggest a general advantage of more detailed or example-rich prompts for this annotation task.

5.3 Cross-Lingual Prompting

Comparing the results for Swedish petitions using Swedish versus English prompts reveals barely any consistent patterns. LLaMA 70B generally performed better with Swedish prompts, suggesting some advantage for this model, while other models showed similar results regardless of prompt language. There was a slight advantage for English prompts on *Salutatio*, whereas *Conclusio* saw marginally better performance with Swedish prompts. However, these differences were small, and the lack of a clear preference overall suggests that prompt language had little systematic effect.

5.4 Difficulty Ratings

Results for model performance and difficulty ratings in comparison to human ratings are presented in Table 6. Looking at the mean error (ME) scores,

Results for the Swedish Dataset - using English Instructions

Prompt	Model	`salutatio`				`petitio`				`conclusio`			
		P	R	F1	TD	P	R	F1	TD	P	R	F1	TD
short 1-shot prompt	GPT-4	98.7	**100.0**	99.3	0.01	84.1	78.9	81.4	0.29	87.3	**92.3**	89.7	0.11
	Gemini	98.7	**100.0**	99.3	0.01	80.1	82.2	81.1	0.34	**100.0**	59.0	74.2	0.36
	Mixtral	96.0	99.9	97.9	0.02	74.3	58.4	65.4	0.55	89.4	58.7	70.9	0.37
	Mistral	95.9	98.4	97.1	0.04	51.6	77.2	61.8	0.65	85.1	54.0	66.1	0.40
	LLaMA 70B	98.7	99.9	99.3	0.01	82.3	74.5	78.2	0.40	87.5	77.9	82.4	0.24
	LLaMA 8B	99.2	97.2	98.2	0.11	72.9	32.4	44.8	0.53	85.6	56.7	68.2	0.46
long 1-shot prompt	GPT-4	98.7	**100.0**	99.3	0.01	**87.1**	**84.3**	**85.6**	0.24	96.7	91.1	**93.9**	0.08
	Gemini	**100.0**	**100.0**	**100.0**	0.00	86.4	82.5	84.4	0.28	**100.0**	63.7	77.9	0.34
	Mixtral	95.8	99.9	97.8	0.01	84.8	68.1	75.5	0.40	86.4	55.5	67.6	0.34
	Mistral	95.7	98.4	97.1	0.04	55.5	72.5	62.8	0.64	85.8	48.0	61.6	0.41
	LLaMA 70B	82.9	60.1	69.7	0.55	78.1	38.0	51.2	0.77	73.7	37.8	50.0	0.74
	LLaMA 8B	99.2	97.4	98.3	0.10	77.1	38.8	51.6	0.48	84.4	53.2	65.3	0.44
long 3-shot prompt	GPT-4	97.7	96.4	97.1	0.06	84.8	80.0	82.3	0.28	85.8	88.5	87.1	0.16
	Gemini	**100.0**	**100.0**	**100.0**	0.00	86.4	81.9	84.1	0.30	99.9	64.3	78.2	0.33
	Mixtral	94.1	99.8	96.9	0.02	80.4	57.5	67.0	0.49	85.1	51.6	64.3	0.31
	Mistral	93.9	98.6	96.2	0.05	55.3	53.9	54.6	0.70	84.3	43.3	57.2	0.39
	LLaMA 70B	83.1	58.6	68.7	0.57	77.4	30.4	43.6	0.81	77.7	37.2	50.3	0.77
	LLaMA 8B	98.8	98.3	98.5	0.09	75.8	50.5	60.6	0.51	83.8	56.4	67.4	0.40

Table 5: Results for Swedish data across three English prompt types and six models. Scores includes token-level precision (P), recall (R), and F1 in percentage, and mean token-level edit distance for each predicted span.

the overwhelming majority of negative values indicates that models, with few exceptions, rate documents as less difficult than humans do. An interesting observation is that model difficulty ratings align most closely with human ratings when models are given detailed prompts with several examples (*long 3-shot*), as reflected by generally lower ME scores in this condition.

The Spearman's correlation coefficients further illustrate the relationship between model and human difficulty assessments. Across prompts and models, correlations ranged from -0.45 to +0.27, with most values being negative. This suggests that passages rated as more difficult by humans tended to be rated as easier by the models. Even the few positive correlations were weak, indicating minimal alignment in the ranking of document difficulty between models and humans.

When comparing model performance to human difficulty ratings, clearer trends are harder to discern. For both languages, model performance tends to be lowest on documents that humans rated as most difficult to annotate (Difficulty 2), but there is considerable variation as indicated by high standard deviations, and there is no clear difference between levels 0 and 1.

6 Conclusion and Future Work

This study has demonstrated that LLMs can perform remarkably well in annotating rhetorical sections within historical petitions, with many models achieving high F1 scores, particularly for *Salutatio*. The results highlight both the capabilities and limitations of current LLMs: while models generally perform strongly across datasets and prompt types, performance varies by section, with *Petitio* and *Conclusio* proving more challenging — particularly for the Swedish data and generally for smaller models. Additionally, although model performance and difficulty ratings correlate to some extent with human ratings, models tend to underestimate document difficulty, suggesting that while they can produce relative difficulty assessments, their ratings may not fully align with human judgments of annotation complexity.

Looking ahead, several avenues for future research emerge from these findings. Firstly, although few-shot prompting yields strong results, training or fine-tuning models specifically on rhetorical annotation tasks may further enhance performance, particularly for more challenging sections such as *Petitio* and *Conclusio*. Fine-tuning on domain-specific data could also improve model calibration, reducing the gap between model and human difficulty ratings. Secondly, future work should explore how segmentation and rhetorical annotation affect downstream tasks such as information retrieval, entity extraction, or social network reconstruction from historical petitions. Given that petitions often embed requests, narrations describ-

Model	Language	Prompt	Mean Err	Spearman ρ	p-value (ρ)	Difficulty 0	Difficulty 1	Difficulty 2
GPT-4	English	short 1-shot	-0.26	-0.05	0.64	91.6 ± 7.1	92.2 ± 6.4	89.4 ± 9.0
		long 1-shot	-0.27	-0.15	0.14	94.5 ± 5.1	94.7 ± 3.8	89.5 ± 8.8
		long 3-shot	**-0.22**	-0.12	0.25	97.2 ± 4.6	98.1 ± 2.9	90.9 ± 12.1
	Swedish	sv short 1-shot	-0.55	-0.40	< 0.01	92.5 ± 12.9	90.4 ± 16.2	81.0 ± 16.6
		sv long 1-shot	-0.37	-0.44	< 0.01	94.1 ± 11.1	92.1 ± 12.3	81.0 ± 16.5
		sv long 3-shot	-**0.03**	-0.36	< 0.01	94.0 ± 11.2	92.0 ± 15.8	86.1 ± 13.9
		eng short 1-shot	-0.56	-0.43	< 0.01	93.4 ± 12.2	90.8 ± 13.9	79.9 ± 16.9
		eng long 1-shot	-0.50	-0.43	< 0.01	94.2 ± 11.8	94.2 ± 10.4	82.6 ± 16.7
		eng long 3-shot	**-0.18**	-0.32	< 0.01	90.6 ± 16.9	86.9 ± 17.6	83.1 ± 19.8
Gemini 1.5 Pro	English	short 1-shot	**0.03**	-0.25	0.01	98.1 ± 3.1	98.2 ± 1.9	92.0 ± 11.2
		long 1-shot	0.08	-0.45	< 0.01	98.3 ± 3.8	97.4 ± 3.3	90.8 ± 11.0
		long 3-shot	0.26	-0.42	< 0.01	98.6 ± 2.6	96.9 ± 6.2	89.7 ± 12.0
	Swedish	sv short 1-shot	**-0.09**	-0.40	< 0.01	91.0 ± 10.5	87.5 ± 12.5	79.1 ± 14.8
		sv long 1-shot	-0.37	-0.20	0.04	89.8 ± 8.9	89.4 ± 10.4	82.7 ± 13.8
		sv long 3-shot	0.14	-0.07	0.49	87.3 ± 14.4	90.9 ± 8.4	86.6 ± 12.2
		eng short 1-shot	**-0.06**	-0.27	0.01	88.3 ± 14.4	88.0 ± 14.0	82.5 ± 12.0
		eng long 1-shot	-0.19	-0.30	< 0.01	90.6 ± 9.1	89.1 ± 10.5	82.4 ± 13.1
		eng long 3-shot	0.15	-0.12	0.22	88.9 ± 10.6	88.4 ± 10.5	86.0 ± 11.6
Mixtral 8x22B	English	short 1-shot	-0.36	-0.38	< 0.01	97.6 ± 4.5	96.2 ± 3.7	89.4 ± 11.6
		long 1-shot	-0.32	-0.39	< 0.01	97.0 ± 6.0	94.9 ± 4.5	92.9 ± 6.6
		long 3-shot	**-0.18**	-0.38	< 0.01	97.7 ± 3.8	95.1 ± 9.0	88.4 ± 12.4
	Swedish	sv short 1-shot	-0.33	-0.21	0.04	80.1 ± 20.7	75.1 ± 18.0	74.6 ± 14.6
		sv long 1-shot	-0.56	-0.33	< 0.01	83.9 ± 16.9	82.8 ± 16.1	71.2 ± 17.1
		sv long 3-shot	**-0.07**	-0.22	0.03	79.5 ± 18.1	72.3 ± 18.5	70.3 ± 21.1
		eng short 1-shot	-0.67	-0.27	0.01	80.9 ± 17.4	75.0 ± 21.9	69.0 ± 19.4
		eng long 1-shot	-0.66	-0.27	0.01	83.6 ± 16.7	87.9 ± 9.5	71.6 ± 19.7
		eng long 3-shot	**-0.13**	-0.23	0.02	79.8 ± 18.0	79.5 ± 18.0	68.4 ± 20.8
Mistral 7B	English	short 1-shot	-0.35	-0.13	0.21	89.7 ± 13.2	93.7 ± 9.0	79.4 ± 18.4
		long 1-shot	-0.31	-0.22	0.03	88.3 ± 13.2	90.3 ± 9.4	79.5 ± 13.5
		long 3-shot	**0.25**	0.04	0.71	82.7 ± 13.8	88.8 ± 8.0	79.6 ± 16.6
	Swedish	sv short 1-shot	-0.25	0.27	0.17	67.5 ± 17.6	78.1 ± 16.9	77.2 ± 15.1
		sv long 1-shot	-0.39	-0.21	0.28	81.8 ± 13.9	79.0 ± 15.9	75.8 ± 13.6
		sv long 3-shot	**-0.05**	-0.08	0.43	70.7 ± 24.5	64.9 ± 17.5	68.8 ± 20.1
		eng short 1-shot	-0.79	-0.03	0.74	69.8 ± 20.9	73.5 ± 18.0	66.9 ± 21.4
		eng long 1-shot	-0.78	-0.07	0.48	71.0 ± 20.0	77.9 ± 15.2	64.8 ± 23.4
		eng long 3-shot	**-0.10**	-0.37	< 0.01	72.8 ± 20.3	69.9 ± 19.9	50.5 ± 21.6
LLaMA-3 70B	English	short 1-shot	-0.34	-0.44	< 0.01	97.6 ± 5.8	98.4 ± 1.8	88.5 ± 13.5
		long 1-shot	-0.34	-0.45	< 0.01	98.7 ± 3.8	98.4 ± 2.1	92.5 ± 9.4
		long 3-shot	**-0.33**	-0.41	< 0.01	98.6 ± 4.4	98.9 ± 1.5	95.8 ± 4.3
	Swedish	sv short 1-shot	-0.55	-0.26	0.01	91.1 ± 11.6	87.5 ± 15.3	86.7 ± 11.6
		sv long 1-shot	-0.41	-0.22	0.03	89.8 ± 13.7	91.6 ± 7.5	87.1 ± 11.1
		sv long 3-shot	**0.01**	-0.23	0.02	89.8 ± 11.7	87.6 ± 11.9	84.8 ± 12.9
		eng short 1-shot	-0.67	-0.19	0.06	88.7 ± 14.0	82.3 ± 19.4	84.5 ± 17.7
		eng long 1-shot	-0.59	-0.02	0.86	54.0 ± 28.2	47.5 ± 23.4	54.7 ± 28.7
		eng long 3-shot	**-0.33**	-0.01	0.96	49.5 ± 25.6	46.4 ± 23.1	50.9 ± 27.5
LLaMA-3 8B	English	short 1-shot	**-0.06**	-0.18	0.07	93.2 ± 8.0	94.8 ± 5.2	83.9 ± 14.1
		long 1-shot	-0.20	-0.25	0.01	94.1 ± 7.7	95.9 ± 3.1	83.7 ± 14.4
		long 3-shot	0.37	-0.37	< 0.01	94.9 ± 7.7	94.7 ± 5.1	84.4 ± 11.5
	Swedish	sv short 1-shot	**0.02**	-0.19	0.06	72.3 ± 20.4	70.7 ± 15.7	62.3 ± 22.2
		sv long 1-shot	0.07	-0.11	0.27	68.6 ± 20.3	71.9 ± 22.0	61.6 ± 22.8
		sv long 3-shot	0.40	-0.19	0.05	76.1 ± 23.3	72.3 ± 24.2	67.6 ± 18.1
		eng short 1-shot	-0.30	-0.18	0.07	67.2 ± 20.7	69.4 ± 21.3	56.7 ± 23.6
		eng long 1-shot	-0.37	-0.24	0.01	71.6 ± 20.0	73.9 ± 20.8	57.4 ± 22.1
		eng long 3-shot	**-0.03**	-0.19	0.06	75.9 ± 22.0	71.0 ± 22.6	67.5 ± 21.9

Table 6: Comparison of model difficulty rating vs. human ratings expressed in Mean Error (Mean Err) and Spearman's correlation coefficients, including p-values, for each prompt type, together with performance of models on different human difficulty ratings.

ing the everyday lives of people, and expressions of social positioning within specific rhetorical sections, improved segmentation may directly enhance the accuracy and interpretability of subsequent analyses.

Finally, while this study has tested different prompt designs varying in length and number of examples, these variations do not yield systematic differences in model performance. It is possible that the different prompt types we employed do not differ substantially enough to impact results, suggesting that LLM outputs for this rhetorical annotation task may be relatively robust to prompt complexity. Moreover, although this work focuses on English and Swedish petitions, expanding to additional languages, including those with even fewer NLP resources, could further illuminate the limitations of current models and the potential for cross-lingual transfer. Together, such investigations would support the development of robust computational workflows for historical document analysis, enabling fine-grained, content-aware annotation at scale to advance humanities research.

Acknowledgments

We warmly thank our colleagues in the Petitions and GaW projects for the inspiring and engaging discussions we shared: Jezzica Israelsson, Örjan

Kardell, Jonas Lindström, Sofia Ling, Maria Ågren, Linda Oja and Fredrik Wahlberg. We are also deeply grateful to Sharon Howard for providing a plain text version of the London Lives dataset and for generously assisting us with our questions. Finally, we thank the anonymous reviewers for their constructive and valuable feedback. The research reported in this paper was supported by a grant from the Swedish Research Council (grant number 2018-06159).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Qi Cheng, Liqiong Chen, Zhixing Hu, Juan Tang, Qiang Xu, and Binbin Ning. 2024. A novel prompting method for few-shot NER via LLMs. *Natural Language Processing Journal*, 8:100099.

Kevin Cohen, Laura Manrique-Gómez, and Rubén Manrique. 2025. Historical ink: Exploring large language models for irony detection in 19th-century Spanish. *arXiv preprint arXiv:2503.22585*.

Gwilym Dodd. 2011. Writing wrongs: the drafting of supplications to the crown in later fourteenth-century England. *Medium Aevum*, 80(2):217–246.

Stina Hansson. 1988. *Svensk brevskrivning: teori och tillämpning*, volume 18. Göteborgs universitet.

Tim Hitchcock, Robert Shoemaker, Sharon Howard, Jamie McLaughlin, et al. 2012. London Lives, 1690–1800. https://www.londonlives.org. Version 1.1, 24 April 2012.

Rab Houston. 2014. *Peasant petitions: social relations and economic life on landed estates, 1600-1850*. Springer.

Sharon Howard. 2016. The London Lives Petitions Project. https://www.londonlives.org. Version 2.0, 2016, based on data from London Lives.

Mark Humphries, Lianne C Leddy, Quinn Downton, Meredith Legace, John McConnell, Isabella Murray, and Elizabeth Spence. 2025. Unlocking the archives: Using large language models to transcribe handwritten historical documents. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, pages 1–19.

Jezzica Israelsson. 2016. In consideration of my meagre circumstances: The language of poverty as a tool for ordinary people in early modern Sweden. Master's thesis, Uppsala Universitet.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Aru Maekawa, Tsutomu Hirao, Hidetaka Kamigaito, and Manabu Okumura. 2024. Can we obtain significant success in RST discourse parsing by using large language models? *arXiv preprint arXiv:2403.05065*.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Thomas Sokoll. 2006. Writing for relief: Rhetoric in English pauper letters, 1800–1834. In Andreas Gestrich, Steven King, and Lutz Raphael, editors, *Being Poor in Modern Europe: Historical Perspectives 1800–1940*, pages 91–112. Peter Lang, Bern.

Charles Spearman. 1904. The proof and measurement of association between two things. *American Journal of Psychology*, 15:72–101.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Michael EW Varnum, Nicolas Baumard, Mohammad Atari, and Kurt Gray. 2024. Large language models based on historical text could offer informative tools for behavioral science. *Proceedings of the National Academy of Sciences*, 121(42):e2407639121.

A Annotation Examples

We include two annotation examples to illustrate how difficulty scores were assigned during annotation. Text included only by Annotator 1 is shown in *purple italics*, while text where both annotators agreed is shown in **green bold**.

Example 1: LMSMPS505520040_1765, disagreement in Petitio, difficulty score 1

<salutatio>To the Worshipfull his Majesty's Justices of the Peace for the County of Middlesex in their General Sessions of the Peace Assembled</salutatio>

The Humble Petition and Appeal of the Churchwardens and Overseers of the Poor of the Parish of Saint Mary le bone in the said County of Middlesex

Sheweth That by Virtue of a Pass Warrant or Order under the Hands & Seals of George Wrighte and Thomas Edwards Esquires two of his Majesty's Justices of the Peace for the City and Liberty of Westminster [...] (one whereof being of the Quorum) bearing Date the 12th. Day of August 1765 Elizabeth Gibson Wife of Bignall Gibson gone [...] James their Child were removd from the Parish of Saint James within the Liberty of Westminster in the said County to the Parish of Saint Mary Le Bone in the said County as the Place of the Last Legal Settlement of the said Bignall Gibson Wife and Child *<petitio>Whereby Your Petitioners Think themselves aggrieved and Appeal to this Court against the same*

<petitio>And Therefore humbly pray this Court will Please to Appoint a Time in this Sessions for hearing and determining the said Appeal And that all Persons removed may then attend.</petitio>

<conclusio>And your Petitioners shall ever Pray Etc</conclusio>

Example 2: LMSMPS502350016_1726, disagreement in Conclusio, difficulty score 2

<salutatio>To the Honble Bench of Justices Novemr. att Hickes Hall</salutatio>

The Humble Petition of the

Churchwarden and Overseer of the poor and Other Anchant Inhabitants of the Hamblett of Mile and New Term in the parish of Stepney on the County of Middxss:

<petitio>Humbly Sheweth that your petitioners begs the favour of this Honble: Bench that they would not Discharge John Bloom now in Custody at the Keeper of Bridwell</petitio> for that he being a Loose Idle Disorderly person and Absenting himself from his familey whereby the Said Hamblet has bin att great Expence and Charge for the two [...] last past for the Supps of the child

<conclusio>[...] Duty on [...]

[...] April [...] } Overseer of the

<conclusio>Joe Mills John Turner [...] } [...] Stable [...]</conclusio>

B Prompt Examples for the English Dataset

```
You are an expert on analysing historical texts. Your task is to identify and label rhetorical sections in
     petitions from the 18th century using three specific tags.

### Tags to apply:
1. <salutatio>...</salutatio> - opening formal greeting to the recipient(s) of the petition
2. <petitio>...</petitio> - main request(s) being made
3. <conclusio>...</conclusio> - final phrase(s) of courtesy and/or inferiority, often including a signature

At the end, provide an overall score (0-2) for how difficult the text was to tag.
   * **0 (Easy to tag):** All sections are clear and easily identifiable.
   * **1 (Somewhat difficult):** Some sections may be a bit blended or phrasing may be unusual, requiring
        careful judgment.
   * **2 (Very difficult):** The text is irregular or difficult to interpret, making identification more
        speculative. The distinction between narrative and request (petitio) can be ambiguous.

### Output Format:
Return only the following two sections. Do not add any explanations, comments, or other text before, between
     , or after the sections. Use the exact following headings and formatting:

### TAGGED TEXT:
<salutatio>...</salutatio> [any untagged text goes here] <petitio>...</petitio> [any untagged text goes here
     ] <conclusio>...</conclusio>

### DIFFICULTY SCORE:
X

### Example Output:
### TAGGED TEXT:
<salutatio>To the Worshipfull his Majestys Justices of the Peace for the County of Middlesex in their
     General [---] Sessions of the Peace Assembled</salutatio>

The Humble Petition and Appeal of the Churchwardens and Overseers of the Poor of the Parish of Enfield in
     the County of Middlesex

Sheweth That by Virtue of a Pass Warrant or order of Removal under the Hands and Seals of John of Hesse and
     Saunders Welch Esquires two of his Majestys Justices of the Peace for the County of Middlesex (One
     whereof being of the Quorum) bearing Date the Twenty Sixth Day of October 1774 Robert Pearpoint and
     Elizabeth his Wife were removed and Conveyed from and out of the Parish of Paddington in the said
     County to the said Parish of Enfield in the said County of Middlesex as the Place of their last Legal
     Settlement Whereby your Petitioners conceive themselves to be agrieved

<petitio>Therefore humbly pray your Worships to appoint a Short Day in this present Session to hear and
     determine their said Appeal</petitio> <conclusio>And your Petitioners shall ever pray Etc

I Smart and Son Attorneys for Appellrs.</conclusio>

### DIFFICULTY SCORE:
0

### Now tag the following petition:
```

Figure 1: Prompt 1 for the English dataset, with less detailed instructions and one given example output.

```
You are an expert on analysing historical texts. Your task is to identify and label rhetorical sections in
     petitions from the 18th century using three specific tags.

### Tags to apply:
1. <salutatio>...</salutatio> - opening formal greeting to the recipient(s) of the petition
2. <petitio>...</petitio> - main request(s) being made
3. <conclusio>...</conclusio> - final phrase(s) of courtesy and/or inferiority, often including a signature

### Core Instructions
1. **Preserve Original Text:** Do NOT add, remove, or change any words, spelling, or punctuation in the
     original text.
2. **Tag Application:** Only apply tags where the content matches one of the three categories in the schema.
3. **Handle Missing Sections:** Sometimes a tag may be missing, though this should be rare.
4. **Handle Multiple Sections:** Tags may appear more than once, especially <petitio>...</petitio>, though
     this should be rare.
5. **Identify Functional Boundary:** When tagging the text segments, the functional boundaries should be
     prioritised over grammatical and/or syntactical boundaries if in conflict. Especially for petitio, this
      means separating circumstances or arguments from the request itself, e.g. "That your Petitioner
     conceives himself to be aggrievd by the said Conviction and humbly <petitio>appeals against the same</
     petitio>".
6. **Difficulty Score:** At the end, provide an overall score (0-2) for how difficult the text was to tag.
   * **0 (Easy to tag):** All sections are clear and easily identifiable.
   * **1 (Somewhat difficult):** Some sections may be a bit blended or phrasing may be unusual, requiring
        careful judgment.
   * **2 (Very difficult):** The text is irregular or difficult to interpret, making identification more
        speculative. The distinction between narrative and request (petitio) can be ambiguous.

### Output Format:
Return only the following two sections. Do not add any explanations, comments, or other text before, between
     , or after the sections. Use the exact following headings and formatting:

### TAGGED TEXT:
<salutatio>...</salutatio> [any untagged text goes here] <petitio>...</petitio> [any untagged text goes here
     ] <conclusio>...</conclusio>

### DIFFICULTY SCORE:
X

### Example Output:
/.../

### Now tag the following petition:
```

Figure 2: Prompt 2 for the English dataset, with more detailed instructions and one given example output (though the example text is left out in this figure).

C Swedish Prompt Examples for the Swedish Dataset

```
Du är expert på att analysera historiska texter. Din uppgift är att identifiera och märka upp retoriska
     segment i svenska suppliker från 1700-talet med hjälp av tre specifika taggar.

### Taggar att använda:
1. <salutatio>...</salutatio> - inledande formell hälsning till mottagaren av suppliken
2. <petitio>...</petitio> - framställning av den huvudsakliga begäran
3. <conclusio>...</conclusio> - avslutande artighets- och/eller underdånighetsfras, ofta inkluderande en
     signatur

Avslutningsvis, ange en övergripande svårighetsgrad (0-2) för hur svår texten var att tagga.
  * **0 - Lätt att taggga**: Alla sektioner är tydliga och lätta att identifiera.
  * **1 - Något svår**: Vissa sektioner kan flyta ihop något eller vissa formuleringar kan vara ovanliga,
       vilket kräver noggrant omdöme.
  * **2 - Mycket svår**: Texten är oregelbunden eller bitvis svårtolkad, vilket gör identifieringen mer
       spekulativ. Distinktionen mellan berättelse och begäran (petitio) kan vara tvetydig.

### Outputformat:
Returnera enbart följande två sektioner. Lägg inte till några förklaringar, kommentarer eller annan text fö
     re, mellan eller efter sektionerna. Använd exakt följande rubriker och formatering:

### TAGGAD TEXT:
<salutatio>...</salutatio> [eventuell otaggad text här] <petitio>...</petitio> [eventuell otaggad text här]
     <conclusio>...</conclusio>

### SVÅRIGHETSGRAD:
X

### Exempel på output:
### TAGGAD TEXT:
<salutatio>Högwälborne H Baron och Landshöfdinge
Nådige Herre</salutatio>


Inför Eders Nåde ähr iag fattige änkia högst Nödsakat mig att beswära, och ödmiukeligast tillkiänna gifwa
     huru som iag långt för detta dehlat om besittningen af 1/8 dehl uti helgiärds hemmanet bregården i
     Carlskouga sochn, med min swåger Oluf Larsson därstädes hwilken mig der ifrån trängt, oacktat hwad rätt
      iag der till äger och hoos gående högl kongl Bergs Collegii bref af d 8 Julij A 1711, samt det höga
     Landshöfdinge Embetets Resolutioner af d 1 och 10 Julij A 1717, mig nåd rättwiseligen tillägga uppå hög
      bem Kongl Collegii bref och dhe i mine Suppliqwer anförde skiähl, sedan hafwer och denna saak wähl
     warit före uti den wähl låfl lagmans tings rätten d 22 Aprill nästl, Men efter den war Incamminerat så
     i högl Kongl Bergs Collegium som och wyd detta Canceliet, Ty ähr den ej till afgiörande företagen
     worden wydare än hoosgående Resolution förmår och utwysar. Wetandes iag ej hwad för Resolution Oluf
     Larsson kunnat sig utwärka i Augusti månad A 1717. Ty så wyda han hållit sig intill Sanningen med sine
     berättelser som Eders Nåde täcktes skåda af min hoosfougade Documenter äro grundade På, så har han
     sannerligen intet Någon annan lydande Resolution kunnat utfå än iag; <petitio>Bönfaller för denskull
     till Eders höga Nåde iag alldra ödmiukast, at, I anseende till min anförde rättmätiga skiähl till be
     min hemmans dehl blifwa restituerat,</petitio> <conclusio>hwar öfwer, en nådig resolution afwacktandes
     deremot iag förblifwer.
Eders Nåds
Alldra ödmiukaste
tienarinna
Margreta Andersdotter
i österwyk.</conclusio>

### SVÅRIGHETSGRAD:
0

### Tagga nu följande supplik:
```

Figure 3: Swedish Prompt 1 for the Swedish dataset, with less detailed instructions and one given example output.

```
Du är expert på att analysera historiska texter. Din uppgift är att identifiera och märka upp retoriska
     segment i svenska suppliker från 1700-talet med hjälp av tre specifika taggar.

### Taggar att använda:
1. <salutatio>...</salutatio> - inledande formell hälsning till mottagaren av suppliken
2. <petitio>...</petitio> - framställning av den huvudsakliga begäran
3. <conclusio>...</conclusio> - avslutande artighets- och/eller underdånighetsfras, ofta inkluderande en
     signatur

### Huvudinstruktioner
1. **Bevara originaltexten:** Lägg INTE till, ta bort eller ändra några ord, stavningar eller skiljetecken i
      originaltexten.
2. **Taggtillämpning:** Använd taggar endast där innehållet matchar en av de tre kategorierna.
3. **Hantera saknade segment:** Det kan förekomma texter där någon eller några av segmenten saknas. Detta bö
     r dock vara sällsynt.
4. **Hantera flera segment:** Taggar kan förekomma mer än en gång, särskilt <petitio>. Även detta bör vara s
     ällsynt.
5. **Semantik vs syntax** Din taggning ska ta stor hänsyn till semantik, inte enbart till grammatik. När du
     taggar <petitio>, inkludera inte omgivande satser eller fraser som endast utgör argument eller
     bakgrundsinformation. Undantag: Du ska inkludera korta bindeord eller fraser i form av kausala markörer,
      som "därför" och "av detta skäl", om de direkt inleder eller avslutar själva begäran.
6. **Ange svårighetsgrad:** Avslutningsvis, ange en övergripande svårighetsgrad (0-2) för hur svår texten
     var att tagga.
   * **0 - Lätt att taggga**: Alla sektioner är tydliga och lätta att identifiera.
   * **1 - Något svår**: Vissa sektioner kan flyta ihop något eller vissa formuleringar kan vara ovanliga,
       vilket kräver noggrant omdöme.
   * **2 - Mycket svår**: Texten är oregelbunden eller bitvis svårtolkad, vilket gör identifieringen mer
       spekulativ. Distinktionen mellan berättelse och begäran (petitio) kan vara tvetydig.

### Outputformat:
Returnera enbart följande två sektioner. Lägg inte till några förklaringar, kommentarer eller annan text fö
     re, mellan eller efter sektionerna. Använd exakt följande rubriker och formatering:

### TAGGAD TEXT:
<salutatio>...</salutatio> [eventuell otaggad text här] <petitio>...</petitio> [eventuell otaggad text här]
     <conclusio>...</conclusio>

### SVÅRIGHETSGRAD:
X

### Exempel på output:
/.../

### Tagga nu följande supplik:
```

Figure 4: Swedish Prompt 2 for the Swedish dataset, with more detailed instructions and one given example output (though the example text is left out in this figure).

D English Prompt Examples for the Swedish Dataset

```
You are an expert on analysing historical texts. Your task is to identify and label rhetorical sections in
     Swedish petitions from the 18th century using three specific tags.

### Tags to apply:
1. <salutatio>...</salutatio> - opening formal greeting to the recipient(s) of the petition
2. <petitio>...</petitio> - main request(s) being made
3. <conclusio>...</conclusio> - final phrase(s) of courtesy and/or inferiority, often including a signature

At the end, provide an overall score (0-2) for how difficult the text was to tag.
   * **0 (Easy to tag):** All sections are clear and easily identifiable.
   * **1 (Somewhat difficult):** Some sections may be a bit blended or phrasing may be unusual, requiring
        careful judgment.
   * **2 (Very difficult):** The text is irregular or difficult to interpret, making identification more
        speculative. The distinction between narrative and request (petitio) can be ambiguous.

### Output Format:
Return only the following two sections. Do not add any explanations, comments, or other text before, between
     , or after the sections. Use the exact following headings and formatting:

### TAGGED TEXT:
<salutatio>...</salutatio> [any untagged text goes here] <petitio>...</petitio> [any untagged text goes here
     ] <conclusio>...</conclusio>

### DIFFICULTY SCORE:
X

### Example Output:
### TAGGED TEXT:
<salutatio>Högwälborne H Baron och Landshöfdinge
Nådige Herre</salutatio>


Inför Eders Nåde ähr iag fattige änkia högst Nödsakat mig att beswära, och ödmiukeligast tillkiänna gifwa
     huru som iag långt för detta dehlat om besittningen af 1/8 dehl uti helgiärds hemmanet bregården i
     Carlskouga sochn, med min swåger Oluf Larsson därstädes hwilken mig der ifrån trängt, oacktat hwad rätt
      iag der till äger och hoos gående högl kongl Bergs Collegii bref af d 8 Julij A 1711, samt det höga
     Landshöfdinge Embetets Resolutioner af d 1 och 10 Julij A 1717, mig nåd rättwiseligen tillägga uppå hög
      bem Kongl Collegii bref och dhe i mine Suppliqwer anförde skiähl, sedan hafwer och denna saak wähl
     warit före uti den wähl låfl lagmans tings rätten d 22 Aprill nästl, Men efter den war Incamminerat så
     i högl Kongl Bergs Collegium som och wyd detta Canceliet, Ty ähr den ej till afgiörande företagen
     worden wydare än hoosgående Resolution förmår och utwysar. Wetandes iag ej hwad för Resolution Oluf
     Larsson kunnat sig utwärka i Augusti månad A 1717. Ty så wyda han hållit sig intill Sanningen med sine
     berättelser som Eders Nåde täcktes skåda af min hoosfougade Documenter äro grundade På, så har han
     sannerligen intet Någon annan lydande Resolution kunnat utfå än iag; <petitio>Bönfaller för denskull
     till Eders höga Nåde iag alldra ödmiukast, at, I anseende till min anförde rättmätiga skiähl till be
     min hemmans dehl blifwa restituerat,</petitio> <conclusio>hwar öfwer, en nådig resolution afwacktandes
     deremot iag förblifwer.
Eders Nåds
Alldra ödmiukaste
tienarinna
Margreta Andersdotter
i österwyk.</conclusio>

### DIFFICULTY SCORE:
0

### Now tag the following petition:
```

Figure 5: English Prompt 1 for the Swedish dataset, with less detailed instructions and one given example output.

```
You are an expert on analysing historical texts. Your task is to identify and label rhetorical sections in
     Swedish petitions from the 18th century using three specific tags.

### Tags to apply:
1. <salutatio>...</salutatio> - opening formal greeting to the recipient(s) of the petition
2. <petitio>...</petitio> - main request(s) being made
3. <conclusio>...</conclusio> - final phrase(s) of courtesy and/or inferiority, often including a signature

### Core Instructions
1. **Preserve Original Text:** Do NOT add, remove, or change any words, spelling, or punctuation in the
     original text.
2. **Tag Application:** Only apply tags where the content matches one of the three categories in the schema.
3. **Handle Missing Sections:** Sometimes a tag may be missing, though this should be rare.
4. **Handle Multiple Sections:** Tags may appear more than once, <petitio>...</petitio>, though this should
     be rare.especially <petitio>, though this should be rare.
5. **Semantics over Syntax:** Your tagging should be guided primarily by semantics, not just grammar. When
     tagging <petitio>, do not include surrounding clauses or phrases that only provide arguments or
     background information. Exception: You should include short linking words or phrases that act as
     anaphoric causal markers, like "therefore" ("därför") and "for this reason" ("av detta skäl") if they
     directly introduce or conclude the actual request.
6. **Difficulty Score:** At the end, provide an overall score (0-2) for how difficult the text was to tag.
   * **0 (Easy to tag):** All sections are clear and easily identifiable.
   * **1 (Somewhat difficult):** Some sections may be a bit blended or phrasing may be unusual, requiring
        careful judgment.
   * **2 (Very difficult):** The text is irregular or difficult to interpret, making identification more
        speculative. The distinction between narrative and request (petitio) can be ambiguous.

### Output Format:
Return only the following two sections. Do not add any explanations, comments, or other text before, between
     , or after the sections. Use the exact following headings and formatting:

### TAGGED TEXT:
<salutatio>...</salutatio> [any untagged text goes here] <petitio>...</petitio> [any untagged text goes here
     ] <conclusio>...</conclusio>

### DIFFICULTY SCORE:
X

### Example Output:
/.../

### Now tag the following petition:
```

Figure 6: English Prompt 2 for the Swedish dataset, with more detailed instructions and one given example output (though the example text is left out in this figure).