OpenHirsch: Detecting Reactionary and Progressive Narratives with Large Language Models

Yann Billeter
Manthan Gadhia

January 29, 2026

1 Introduction

The fabric of political discourse is intricately woven with narratives that shape public opinion and policymaking. Understanding the underlying narratives within political texts is crucial for grasping how ideologies influence governance and societal dynamics. Albert O. Hirschman's "The Rhetoric of Reaction"[7] provides a profound analytical framework for identifying reactionary and progressive narratives, traditionally applied in narrow, qualitative studies. However, the ever-growing volume of political content calls for a scalable solution.

With advancements in artificial intelligence, large language models (LLMs) have demonstrated capabilities nearing human reasoning. This development prompts us to explore whether LLMs can effectively automate the application of Hirschman's framework. In this study, we leverage these models to systematically detect political narratives at scale by engineering prompts that guide the models to discern the nuanced rhetoric in political texts. We base our prompt engineering efforts on the concept of argumentation schemes, which are analytical templates for study of arguments. Argumentation schemes have a long tradition in the study of rhetoric and are also frequently applied in the computational studies using natural language processing techniques. This approach, combining traditional analytical techniques with modern computational tools, facilitates large-scale processing of political content and employs a "big data" strategy in narrative analysis that may uncover previously hidden patterns and provide new tools for researchers and policymakers.

To demonstrate the utility of our framework, we apply it to the Manifesto Corpus [8] – a comprehensive collection of electoral programmes from over 60 countries in almost

40 languages. This application showcases our method's potential to transcend linguistic and cultural boundaries, offering a robust tool for global political analysis.

We find that the LLM-based multi-class classifier is quite well-suited to the task of detecting Hirschmanian narrative rhetorics, showing a strong overall performance with an accuracy of 0.94. The model is also particularly good at predicting the "none" class – where none of the theses are present – with this class having the highest F1-Score at 0.97 as well as the highest precision and recall scores, also showing a disproportionally larger number of instances compared to the other three classes. This is followed by a rather robust performance for detecting "futility" and "perversity" theses, but less so for "jeopardy" which showed a notably low recall value. This relatively poor performance may be attributable in part to the small number of instances corresponding to each thesis in the training, and ultimately the test set.

Applying the classifier to a different dataset with more explicit examples of each thesis may yield better performance, perhaps data of electoral campaign or parliamentary speeches. Additionally, there remains room for improving the prompt engineering as the LLM appeared to have a narrow understanding of the word "cost", limited to monetary and not social, and also struggled to distinguish between criticisms of past actions, often overlooking them in favour of the corrective future actions proposed by the party in their programme.

2 Literature Review

Albert O. Hirschman's "The Rhetoric of Reaction: Perversity, Futility, Jeopardy", published in 1991 [7], offers a critical examination of the conservative reactions to progressive reforms throughout history. Hirschman identifies three principal rhetorical arguments used by conservatives: perversity, futility, and jeopardy. The perversity thesis suggests that any action to improve society will instead exacerbate the conditions it intends to remedy. The futility thesis argues that attempts at social change will be useless because they will produce no significant effects. Lastly, the jeopardy thesis warns that proposed changes will endanger previous hard-won achievements. Hirschman analyses these arguments not only to contest progressive policies, but also as a rhetorical framework that can be observed across different times and political contexts. This framework provides a lens for understanding how conservative rhetoric resists change, shaping both public discourse and policy outcomes.

The three reactionary rhetorical arguments are complemented by three progressive counter-arguments: the synergy thesis, the imminence thesis, and the opportunity thesis. The synergy thesis contends that social reforms address the specific issues they target and generate a network of beneficial effects that reinforce and amplify the initial improvements, highlighting the potential for reforms to create a virtuous cycle of progress. The imminence thesis, on the other hand, emphasises the urgent need for reform by arguing that failure to act promptly will lead to immediate and severe (negative) consequences. This reasoning makes the case for nothing short of swift, decisive actions to address the problems at hand. Finally, the opportunity thesis suggests that embracing reforms would open new avenues for growth and development, positioning society ideally to seize emerging opportunities and adapt to future challenges. Hirschman does not, however, herald these theses as actually resulting in meaningful social progress, noting that they can be overly optimistic or simplistic, failing to account for the complexities and unintended consequences that can accompany social change.

While Hirschman's theses have had, and keep on having, a large impact on theoretical studies of reactionary politics and populism, their practical application so far has been limited to small, narrowly scoped studies.

Hirschman's framework has, *inter alia*, been employed to understand the opposition of peers to an elected House of Lords in the United Kingdom [12] and municipal reform debates in New South Wales, Australia [5]. In [12], the author interviewed a representative sample of peers and subsequently analysed their responses using Hirschman's framework. Similar analyses have been carried out in [14, 1]. It was also found that Hirschman's triad was missing an important fourth temporal element. In [5], the authors apply Hirschman's typology to municipal amalgamation debates in New South Wales. Similar to [12], they find that the typology lacks an element, in this case one to "accommodate debate involving disputed empirical evidence". This critique can be traced back to earlier applications of Hirschman's typology to Australian politics [2, 4].

While narrative analysis using LLMs has not been performed up to this point, similar studies have been performed to detect argumentative fallacies: For instance, [13] treat the problem of detecting fallacious argumentative schemes in real-world conditions using various LLMs. Similarly, [9] fine-tune a GPT-4 model to "understand", identify, and reproduce examples of logical fallacies. So far, the most significant diverging factor between this study and previous ones along similar lines is the use of larger supervised training examples, and the possible challenges that may arise from a given LLM's limited prior access to political texts when being trained.

3 Data

This study utilizes the Manifesto corpus version *2024-1* [8], which contains approximately 3000 machine-readable electoral programmes from more than 60 different countries in almost 40 languages. Almost 2000 of these have been unitized into quasi-sentences and coded according to the Manifesto Coding Scheme. This scheme encodes a party's position on a range of issues.

Of the 3000 available machine-readable electoral programmes, 1626 coded electoral programmes from between 1946 and 2022 have been translated into English using [3]. While it would, in principle, have been possible to use the original versions due to modern LLM's multilingual capabilities, using English as the common language facilitates standardised prompt engineering and validation.

3.1 Training Set

While Manifesto coding provides useful contextual information, the electoral programmes are not labelled according to their use of Hirschman's rhetorical arguments. To optimize our prompts on a "training set" that is as close as possible to the "test set" (i.e. the Manifesto corpus), we select and manually label some programmes from the corpus. For each programme, we label every section (as identified by the presence of a header or title) with one of Hirschman's theses or "none" if no thesis applies.

To further extend our training set and reduce labeller bias, we collect labelled examples of Hirschman's rhetorical arguments from literature [14, 4, 1, 5, 12, 6]. Here Table 1 provides a detailed overview of our training set.

	Total	From Literature	From Manifesto
narrative			
futility	11	9	2
jeopardy	13	12	1
none	22	1	21
perversity	26	21	5

Table 1: Distribution of theses in the training set. Literature examples were labelled by the authors of the respective publications, while samples from the Manifesto corpus were labelled by us.

Beyond the statistics provided in table 1, it is worth highlighting that the text lengths vary significantly, since text length is an indicator of the amount of context available to the language model. Figure 1 shows the text length distribution of the training set.

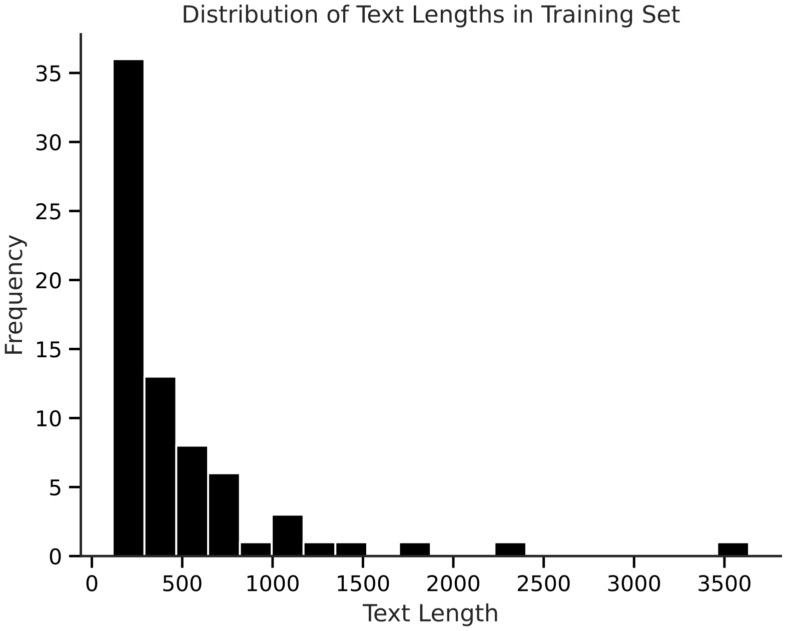

Figure 1: Distribution of text lengths (number of characters) in the training set.

3.2 Test Set

The Manifesto corpus, excluding the programmes used in the training set, acts as our test set. As it is infeasible to manually label all remaining programmes, we select a geographically and ideologically diverse subset. Manifestos were split by sections, as identified by heading annotations in the Manifesto corpus. Sections were manually annotated according to their use of arguments, according to Hirschman's typology. In case multiple theses are applicable, all applicable theses were recorded as labels. The parties are listed in table 2 and the distribution of samples and theses per party are shown in fig. 2.

Party	Country	Classification
Golden Dawn	Greece	Nationalist
Flemish Interest	Belgium	Nationalist
The New Right	Denmark	Nationalist
True Finns	Finland	Nationalist
Sweden Democrats	Sweden	Nationalist
Party of Freedom	Netherlands	Nationalist
Right Answer 21	Netherlands	Nationalist
Alternative for Germany	Germany	Nationalist
Soldiers of Destiny	Ireland	Conservative
Homeland Union	Lithuania	Conservative
Chile we can do more	Chile	Conservative
For the Good of Brazil	Brazil	Nationalist

Table 2: Parties included in the test set along with their classification according to the Manifesto Project and the year of publication of the program analysed.

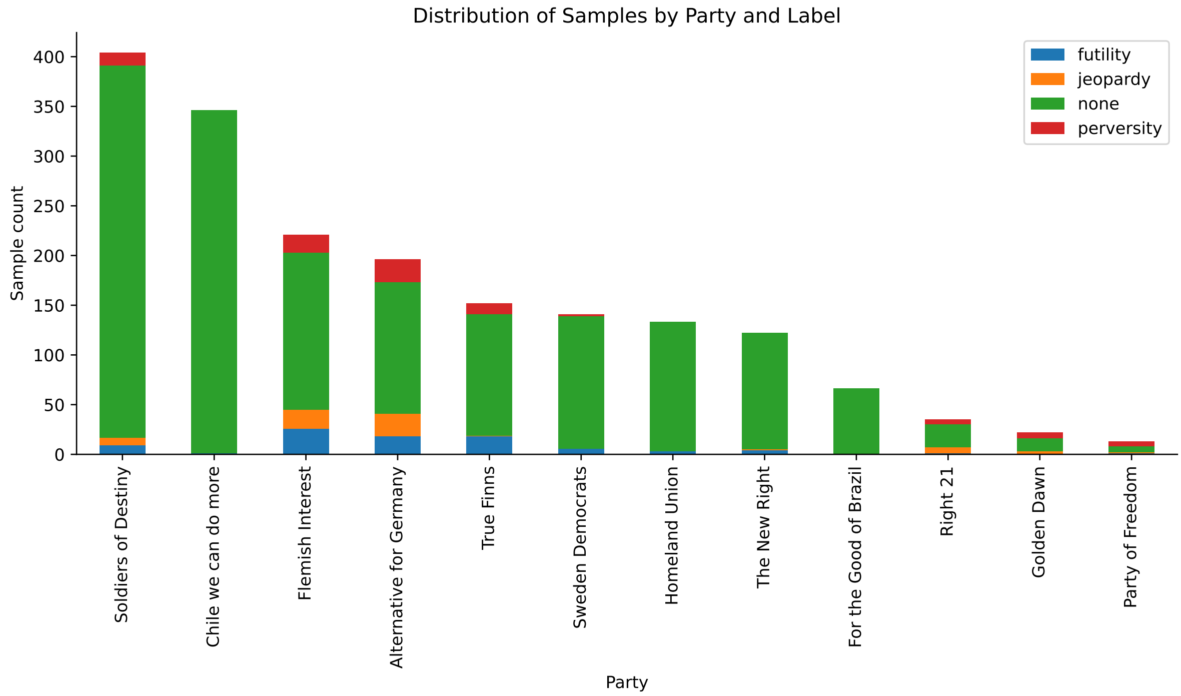

Figure 2: Distribution of samples by party and thesis.

4 Methods

Our research utilizes prompt engineering as the central methodological approach to detect reactionary rhetoric within political texts. By crafting targeted prompts, we guide an LLM to identify argumentation schemes that correspond to Hirschman's theses on reactionary rhetoric. Initially, we will provide a comprehensive overview of Hirschman's theoretical framework to establish a foundation for our methodological choices. Subsequently, we provide details on the design of our prompts and describe the specific LLM used in this study. Finally, we outline our evaluation criteria, explaining how we assess the effectiveness of our prompts in accurately detecting reactionary rhetoric.

4.1 Theoretical framework

As described previously, Hirschman identifies three rhetorical arguments used by conservatives: perversity, futility, and jeopardy. In the following, we provide a detailed description of the three theses, highlighting their analytical features, characteristics, and structure.

Perversity Hirschman summarizes the perversity argument as follows: *the attempt to push society in a certain direction will result in its moving all right but in the opposite direction* [7]. Central to the perversity argument is the assertion that a policy will not only not achieve its intended goal, but exacerbate the very condition it attempts to resolve. Hirschman further notes that such perverse effects are extreme cases of unintended side effects, of which there are both positive and negative examples, attributable to a failure of foresight of human actors. In so far, the perversity thesis requires the acknowledgement of action's effect, and may even be accompanied by mentions of "good intentions" and the use of terms such as "well-meaning" and "well-intentioned".

Futility The futility argument holds that *any alleged change is, was, or will be largely surface, faced, cosmetic, hence illusory, as the "deep" structures of society remain wholly untouched* [7]. At the heart of this thesis is the idea of a rigid, deeper structure that underpins the current state of affairs. The subtlety of the futility thesis lies in its suggestion that proponents of change are naïve for not recognizing these deep-seated structures. While a policy might not achieve its intended goals, it often still benefits certain groups, which leads Hirschman to contend that the futility thesis also alleges that initiatives may be self-serving. This notion is frequently encapsulated in metaphors like "mask", "veil", or "disguise".

Jeopardy The jeopardy thesis asserts that *the proposed change, though perhaps desirable in itself, involves unacceptable costs or consequences of one sort or another*[7]. Central to the jeopardy thesis is its grounding in historical context, requiring an awareness of specific historical settings and events. It is predicated on the notion that the past accomplishments and the proposed action are fundamentally incompatible. Crucially, this does not suggest a zero-sum game; rather, the jeopardy thesis insinuates that the loss of previous accomplishments outweighs the gains from the new action or policy.

4.2 Prompt Engineering

As a basis for our prompts, we create argumentation schemes for Hirschman's theses. We then combine these prompts with chain-of-thought prompting to construct two types of classifiers for Hirschman's theses. In the following, we first revisit the theses and highlight their analytical features, before providing matching argumentation schemes, and ultimately describing the prompts used to classify texts according to their use of reactionary rhetoric.

4.2.1 Argumentation Schemes

Broadly speaking, an argument consists of a series of premises leading to a single conclusion. The premises are intended to provide the supporting evidence and rationale upon which the conclusion is justified. Consequently, the conclusion represents the proposition that the argument aims to either substantiate or refute. An argumentation scheme represents a template for an argument.

For each of Hirschman's theses, we formulate an argumentation scheme that encapsulates the thesis within a two-premise argumentative structure. The conclusion of this structure is represented by the thesis itself. Common to all these schemes is the initial premise: a past, present or planned initiative or policy designed to alter a pre-existing condition.

The second premise varies depending on the thesis in question. For the perversity thesis, it involves the assertion that the initiative will actually worsen the condition it intends to address. In the case of the futility thesis, it concerns the claim that the initiative will be ineffective. Finally, for the jeopardy thesis, it revolves around the presence of a prior achievement and its incompatibility with the initiative."

Section 4.2.1 provides an overview of the argumentation schemes. The subsequent sections will detail the prompts engineered based on these schemes.

Premise 1: A past, present or planned initiative or policy designed to alter a pre-existing condition		
Perversity Thesis	**Futility Thesis**	**Jeopardy Thesis**
The initiative will exacerbate the targeted condition.	The initiative will be ineffective.	There exists a prior accomplishment that is incompatible with the initiative.
Conclusion: Corresponds to the respective thesis itself		

Table 3: Argumentation Schemes for Hirschman's Theses

Prompt: [Task Explanation] **Instructions:**

1. **Identify the Elements of the following argumentation scheme:**
    - **Premise 1** [Premise 1 instructions] Let's think step by step.
    - **Premise 2** [Premise 2 instructions] Let's think step by step.
    - **Conclusion:** [Conclusion Instructions] Let's think step by step.
2. **Conclude whether the argumentation scheme applies or not**
    - **Present:** The text matches both premises and the conclusion.
    - **Absent:** The text does not match one of the premises or the conclusion.

Output: Produce JSON output containing the classification and a detailed breakdown of the analysis:

- **"classification"**: A string value, either "Present" or "Absent."
- **"details"**:
    - **"premises"**:
        - **"premise_1"**: { "justification": ... }
        - **"premise_2"**: { "justification": ... }
        - **"conclusion"**: { "justification": ... }

Figure 3: Prompt template for the 1-vs-rest classifier. Text in brackets is substituted with the respective descriptions and, if applicable, additional information.

4.2.2 One-vs-Rest Classifier

In our initial approach, we engineered a distinct prompt for each of Hirschman's theses. This strategy effectively transforms the language model (LLM) into a one-vs-rest classifier. Specifically, the model is tasked with determining the presence or absence of a particular thesis within a given argument. Each prompt is designed to enable the model to differentiate whether the specific thesis –perversity, futility, or jeopardy – is applicable.

Figure 3 presents the prompt template used. The template starts with a general explanation of the task, followed by instructions for the model to execute two subtasks: first, the identification of the elements of the argumentation scheme, and second, the classification of whether the thesis is present or absent. As a final step, we instruct the model to generate JSON output that summarizes and explains its analysis in a standardized format for later evaluation.

Crucially, we avoid using terms directly associated with Hirschman's theses in the prompts; the theses are not named explicitly, and terms like 'perversity,' 'futility,' and 'jeopardy' are substituted with alternative expressions not directly linked to Hirschman's work. This precaution is taken to prevent confounding effects from the LLM's potential prior exposure to Hirschman's *Rhetoric of Reaction*.

Additionally, we enhance the descriptions of the second premises with features detailed in section 4.1. For instance, the perversity thesis prompt highlights the concept of unintended side effects and well-intentioned actions. We also include negative instructions to clarify scenarios that do not meet the thesis criteria, such as mentioning a policy's ineffectiveness, which contradicts the notion of a contrary effect characteristic of the perversity thesis. Prompts for all theses are provided in appendix A.

4.2.3 Multiclass Classifier

As an alternative to the one-vs-rest classification method previously described, we developed a prompt that transforms the task into a single multiclass classification problem. Initially, the LLM is instructed to verify the presence of the first premise in the text. Subsequently, the LLM must assess which, if any, of the three secondary premises corresponds to the given text. As in the previously described approach, the LLM then has to produce JSON output that summarizes and explains its analysis.

We again add features described in section 4.1 to the prompt. Additionally, we integrate negative examples and provide explicit instructions for the LLM to evaluate the other premises if such negative examples are identified.

Figure 4 presents the prompt template, the prompts are provided in appendix A.

4.3 Model

We use the `GPT-4o` model [10] across all our experiments through the OpenAI API[1]. To ensure reproducibility, we set the text generation temperature to 0, and use the model checkpoint on May 13, 2024, i.e. *gpt-4o-2024-05-13*. `GPT-4o` was chosen due to its comparable intelligence to and efficiency gains over `GPT-4 Turbo` [11].

4.4 Evaluation and Validation

In this study, to evaluate classification performance, four primary metrics are used: precision, recall, F1-score, and support. *Precision* measures the accuracy of positive predictions by calculating the ratio of true positive observations to the total predicted positives, with a high resulting value indicating few false positive classifications. *Recall* (otherwise known as sensitivity) measures the classifier's ability to identify all relevant instances by calculating the ratio of true positive predictions to all actual positive-labelled instances in the dataset. A high recall implies good model performance when it comes to the relevant cases, with few false negative classifications. The *F1-score* provides a balance between precision and recall (in the form of a harmonic mean), and is particularly useful when facing positive-negative class imbalances. Finally, *support* refers to the number of actual occurrences of each class in the dataset, importantly providing context for interpreting the other computed metrics and shedding light on the reliability of the performance evaluation for each category.

---

[1] `https://platform.openai.com/docs/api-reference`

[Task Explanation]

Premise 1: Initiative

[Permise 1 description]

Premise 2

Which of the following premises applies to the text? Let's think step by step. Remember that if there is both a new and an old action described in the text, we are interested in whether one of the following premises fits the description of the **old** one. Think step by step.

1) Unacceptable Cost or Endangerment

[Premise 2 for Jeopardy Thesis]

2) Ineffectiveness

[Premise 2 for Futility Thesis]

3) Contrary Effect

[Premise 2 for Perversity Thesis]

4) None of the above

Select this premise of no other premise applies.

Output

Produce JSON output containing the classification and a detailed breakdown of the analysis:

- "reasoning": "Briefly explain your reasoning. Think step by step. Be precise. Don't write a long text."
- "premise_1": "true if applies, else false"
- "premise_2": "1, 2, 3, or 4, depending on your answer for premise 2"

Figure 4: Prompt template for the multiclass classifier. Text in brackets is substituted with the respective descriptions and, if applicable, additional information.

Following this, to summarise overall performance across multiple classes in both classification setups, three aggregate metrics are used: micro-, macro-, and weighted-average. Micro-average aggregates the contributions of all classes by considering the total number of true positives, false positives, and false negatives, computing the average performance metric across all instances, giving equal weight to each instance (particularly useful when there are large differences in class distribution). Macro-average computes the performance metric independently for each class and then averages these metrics, treating all classes equally regardless of their size, and provides insight into the classifier's capacity to perform across all possible classes in a uniform manner. Finally, the weighted-average computes the metric for each class, weighting the average values by the proportion of available instances per class across the entire dataset, thereby balancing the influence of classes relative to how much they are represented in the data (also useful for cases with imbalanced class sizes, providing insight into the classifier's effectiveness in proportion to class-size).

In addition to evaluation metrics, the prompts also contain instructions to produce explanations for their classifications, which are evaluated to qualitatively gauge the classifiers' performance.

5 Results

5.1 Performance on Training Set

We adapted the prompting schemes presented in section 4.2 until they achieved satisfactory performance. Overall, the results indicate that both 1-vs-rest and multiclass classifiers perform well, however the *F1-scores* indicate that the multiclass classifier displays a more consistent performance over all the categories.

The 1-vs-rest classifier achieved precision scores of $1.00, 1.00$, and 0.71, indicating high accuracy in positive class predictions. However, the recall scores are comparatively low, which implies poor performance in identifying relevant instances. In comparison, the multiclass classifier achieves well-balanced scores and notably higher recall.

This insight is further supported by the confusion matrices displayed in section 5.1, which strongly suggest that the 1-vs-rest classifier primarily makes false negative errors. While this property is desirable in some scenarios, here we wish to maximize the discovery of reactionary rhetoric, as their manual detection is more consuming than verifying the classifiers explanations. In such a scenario, the multiclass classifier appears more suitable, as it makes comparatively few errors in detecting the presence of reactionary rhetoric. Rather, the multiclass classifier confuses the individual theses.

Category	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Support
Futility	1.00	0.18	0.31	0.64	0.64	0.64	11
Jeopardy	0.67	0.31	0.42	0.50	0.77	0.61	13
None	N/A	N/A	N/A	0.84	1.00	0.91	21
Perversity	0.71	0.46	0.56	0.94	0.58	0.71	26
Overall Scores				**Overall Scores**			
Micro Avg	0.85	0.56	0.67	0.74	0.74	0.74	72
Macro Avg	0.87	0.59	0.66	0.58	0.60	0.57	72
Weighted Avg	0.83	0.56	0.63	0.77	0.74	0.73	72
Samples Avg	0.53	0.56	0.54	0.74	0.74	0.74	72

Table 4: Left: Classification scores for the 1-vs-rest classifier. Right: Classification scores for the multiclass classifier. Micro and Macro Avg. represent performance results over all classes.

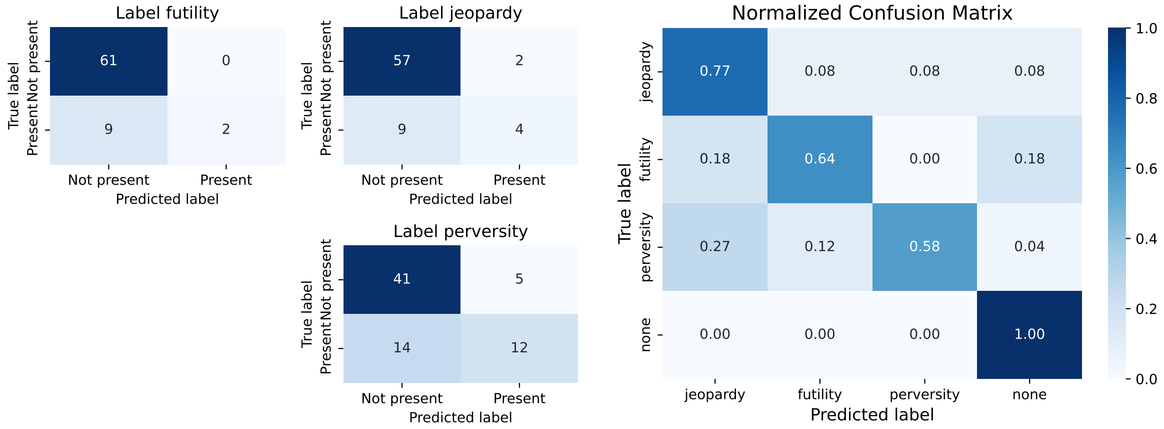

Figure 5: Left: Confusion matrices for the individual 1-vs-rest classifiers. From top to bottom, from left to right: True negatives, false negatives, false positives, and true positives. Right: Confusion matrix for the multiclass classifier.

5.2 Performance on Test Set

In the preceding section, we assessed the performance of our multi-class classifier using the training dataset. We now extend this evaluation to the test set, with a detailed performance overview presented in Table 5. Clearly, this approach performs very well in predicting the 'none' class, which also has a significantly higher number of instances (support) compared to other classes. We also observe good performance for futility and perversity, but less so for jeopardy. This discrepancy is largely driven by the significantly lower recall for the jeopardy theses. Upon closer investigation, it was found that the LLM frequently mistakes the “cost”-aspect of the jeopardy thesis for a monetary cost.

Another, more general pattern of error was found to be that the LLM often disregarded critiques of conditions resulting from past actions, and instead focused on the recommended actions to be taken in the future. In these cases it often correctly judged that there was no thesis applicable to the recommendation even though there was one applicable to the past action being rebutted. Consider the following example of the Belgian *Vlaams Belang* party manifesto calling for ends to various past policies concerning matters of anti-discrimination and the dissolution of Unia (the Belgian interferederal centre of equal opportunity) because of their large social and monetary costs, which the LLM disregards.

> **Manifesto text:** There must be a halt to the encroachment on civil liberties under anti-discrimination policies: no field tests or mystery calls. With regard to hiring and rental of housing, citizens must once again be given the opportunity to have their choices guided by their own views rather than having to conform to morals imposed by a government. Unia should be dissolved and in the meantime, its annual million-dollar subsidy should be discontinued. Free speech should be restricted only when it calls for the use of force. Subsidization of worship services is discontinued: citizens should not be required to pay taxes for the religious beliefs of others.
> **Model prediction:** None.
> **Model reasoning:** The text describes several intended actions aimed at transforming or significantly improving conditions related to civil liberties, free speech, and financial policies. These actions include halting certain anti-discrimination policies, dissolving Unia, discontinuing subsidies for Unia and worship services, and restricting free speech only when it calls for the use of force. The text does not argue that these actions come with unacceptable costs, are ineffective, or have contrary effects. Therefore, none of the specific premises 1, 2, or 3 apply.

Despite this issue, the model accuracy is high, though improvements could be made, in particular by enhancing the recall for classes with fewer instances or more balanced class distributions. This may be achieved, in part, by applying this classifier to a different category of text such as electoral campaign speeches or parliamentary speeches where we would expect more arguments *against* proposed policies rather than party manifestos' focus on what the party in question *would do*.

Inspection of the confusion matrix in fig. 6 provides us with further insight into the errors made by the classifier. Clearly, this approach's main weakness is insensitivity: The majority of errors made are predictions of the class none, when a thesis would have been applicable.

Class	Precision	Recall	F1-Score	Support
Jeopardy	0.68	0.55	0.61	58
Futility	0.92	0.78	0.84	86
Perversity	0.84	0.66	0.74	86
None	0.96	0.98	0.97	1619
Accuracy		0.94		
Macro Avg	0.85	0.74	0.79	1849
Weighted Avg	0.94	0.94	0.94	1849

Table 5: Evaluation of the multi-class classifier on the test.

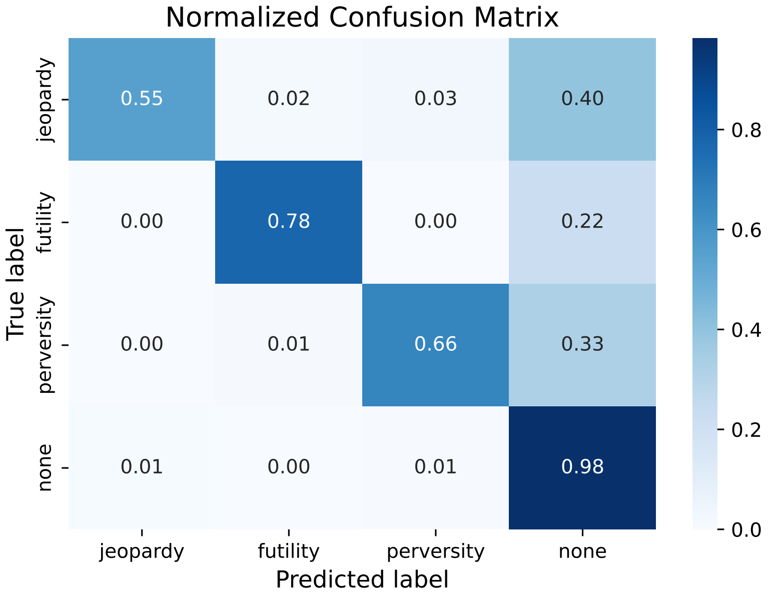

Figure 6: Confusion matrix for the multiclass classifier on the test set.

6 Conclusion

In this study, an innovative approach is developed and employed to leverage frontier LLMs to detect political rhetoric in text data at scale, based on Albert O. Hirschman's "Rhetoric of reaction" theses. For this particular study, we use OpenAI's GPT-4o model.

The test results indicate good model performance with an accuracy of 0.94, as well as high macro-average and weighted-average aggregate scores across classes. A notable limitation was the imbalance of class instances across the Manifesto Corpus dataset with a vast majority of samples not pertaining to any of Hirschman's theses. Applying this classifier to a different dataset may help alleviate this particular problem. Furthermore, the model appeared to struggle with following instructions where the language of the prompt contained some semantic ambiguity, e.g. the model would often consider "cost" to be only monetary and overlook other interpretations, and secondly it often also appeared to struggle with considering both past and future actions as distinct when discussed in the same text. Finally, it is noteworthy that not only are the training and test sets small, but that the manual labelling process in each case is not itself without bias, resulting at least in part from the complexity of identifying non-trivial rhetorical patterns.

The scope for future work lies in exploring the possibility for refining the prompt-engineering to enhance the model's robustness and further guide its understanding of the task at hand, as well as testing this novel approach with other (non-Hirschmanian) rhetorical frameworks to gain insight into the viability of this approach. In a best-case scenario, this technology can be envisioned to provide real-time analysis of political discourse across nations, allowing for better understandings of political narratives as they affect policies and overall governance over time.

References

[1] Alexander Brink. "Hirschman's Rhetoric of Reaction: U.S. and German Insights in Business Ethics". In: *Journal of Business Ethics* 89.1 (Dec. 2008), pp. 109–122. ISSN: 1573-0697. DOI: `10.1007/s10551-008-9988-5`. URL: `http://dx.doi.org/10.1007/s10551-008-9988-5`.

[2] Therese Burton, Brian Dollery, and Joe Wallis. "A Note on the Debate over 'Economic Rationalism' in Australia: An Application of Albert Hirschman's ¡i¿Rhetoric of Reaction¡/i¿". In: *History of Economics Review* 36 (1 Jan. 2002), pp. 1–9. ISSN: 1037-0196. DOI: `10.1080/10370196.2002.11733375`.

[3] DeepL GmbH. *DeepL Translator*. 2024. URL: `https://www.deepl.com/translator` (visited on 06/17/2024).

[4] Brian Dollery and Lin Crase. "Rhetorical patterns in the Australian debate over war with Iraq". In: *Prometheus* 21 (3 Sept. 2003). ISSN: 0810-9028. DOI: `10.1080/0810902032000113497`.

[5] Joseph Drew, Bligh Grant, and Nicole Campbell. "Progressive and reactionary rhetoric in the municipal reform debate in New South Wales, Australia". In: *Australian Journal of Political Science* 51 (2 Apr. 2016), pp. 323–337. ISSN: 1363030X. DOI: `10.1080/10361146.2016.1154926`. URL: `https://www.tandfonline.com/doi/abs/10.1080/10361146.2016.1154926`.

[6] Joseph Drew, Eran Razin, and Rhys Andrews. "Rhetoric in municipal amalgamations: a comparative analysis". In: *Local Government Studies* 45.5 (Oct. 2018), pp. 748–767. ISSN: 1743-9388. DOI: `10.1080/03003930.2018.1530657`. URL: `http://dx.doi.org/10.1080/03003930.2018.1530657`.

[7] ALBERT O. HIRSCHMAN. *The Rhetoric of Reaction: Perversity, Futility, Jeopardy.* Harvard University Press, Mar. 1991. ISBN: 9780674768680. DOI: `10.2307/j.ctvjnrs9q`. URL: `http://dx.doi.org/10.2307/j.ctvjnrs9q`.

[8] Pola Lehmann et al. *Manifesto Corpus.* Version 2024-1. Berlin: WZB Berlin Social Science Center and Göttingen: Institute for Democracy Research (IfDem), 2024. DOI: `Insert DOI if available`. URL: `https://manifesto-project.wzb.eu/`.

[9] Yanda Li et al. *Reason from Fallacy: Enhancing Large Language Models' Logical Reasoning through Logical Fallacy Understanding.* 2024. arXiv: `2404.04293 [cs.CL]`.

[10] OpenAI. *Hello GPT-4o.* 2024. URL: `https://openai.com/index/hello-gpt-4o/`.

[11] OpenAI et al. *GPT-4 Technical Report.* 2024. arXiv: `2303.08774`.

[12] Richard Reid. "Understanding the opposition of peers to an elected House of Lords through Hirschman's Rhetoric of Reaction". In: *British Politics* 13 (2 June 2018), pp. 234–247. ISSN: 17469198. DOI: `10.1057/S41293-017-0050-Z/METRICS`. URL: `https://link.springer.com/article/10.1057/s41293-017-0050-z`.

[13] Ramon Ruiz-Dolz and John Lawrence. "Detecting Argumentative Fallacies in the Wild: Problems and Limitations of Large Language Models". In: ed. by Milad Alshomary et al. Association for Computational Linguistics, Dec. 2023, pp. 1–10. DOI: `10.18653/v1/2023.argmining-1.1`. URL: `https://aclanthology.org/2023.argmining-1.1`.

[14] Joe L. Wallis and Brian E. Dollery. "Understanding the Rhetorical Patterns that Emerge During a Process of Paradigmatic Policy Change: The Case of New Public Management". In: *International Review of Public Administration* 4.2 (Dec. 1999), pp. 57–66. ISSN: 2331-7795. DOI: `10.1080/12294659.1999.10804933`. URL: `http://dx.doi.org/10.1080/12294659.1999.10804933`.

A Full Prompts

A.1 1-vs-rest: Perversity

```
**Prompt:**

Given a text discussing a political, social, or economic initiative,
classify the text according to the provided argumentation scheme.

**Instructions:**

1. **Identify the Elements of the following argumentation scheme:**
   - **Premise 1 (Initiative):** A past, present, or intended action (Action A) that
   aims to transform or significantly improve a condition (Condition C) in the
   political, social, or economic order. The condition may be implicit.
   - **Premise 2 (Contrary Effect):** Implementing Action A not only fails to
   improve Condition C but explicitly worsens it, as stated directly in the text.
   - **Conclusion:** Action A should not be undertaken, stopped, or
   significantly altered because it worsens Condition C.

2. **Classification:**
   - **Present:** If the text aligns with the premises and conclusion
   of the scheme above.
   - **Absent:** If the text does not align with the premises or conclusion
   of the argumentation scheme.

**Output:**
Produce JSON output containing the classification and a detailed breakdown
of the analysis:

- **"classification"**: A string value, either "Present" or "Absent."
- **"details"**:
  - **"premises"**:
    - **"initiative"**: Assessment of whether the text matches Premise 1.
    - **"endangerment"**: Assessment of whether the text matches Premise 2.
    - **"conclusion"**: Assessment of whether the text matches the conclusion..
```

A.2 1-vs-rest: Futility

```
**Prompt:**
```

```
Given a text discussing a political, social, or economic initiative,
classify the text according to the provided argumentation scheme.

**Instructions:**

1. **Identify the Elements of the following argumentation scheme:**
   - **Premise 1 (Initiative):** A past, present, or intended action (Action A) that
   aims to transform or significantly improve a condition (Condition C) in the
   political, social, or economic order. The condition may be implicit.
   - **Premise 2 (Ineffectiveness):** Action A, despite its intentions, will fail
   to produce any significant change or improvement in Condition C. The action may be
   ineffective due to various reasons, but the text should clearly suggest its
   ineffectiveness. Negative consequences or worsening of the condition are not
   sufficient to match this premise.
   - **Conclusion:** Determine if the text concludes or implies that Action A should
   not be undertaken, stoppped, or significantly altered because
   of its ineffectiveness.

2. **Classification:**
   - **Present:** If the text aligns with the premises and conclusion
   of the scheme above.
   - **Absent:** If the text does not align with the premises or conclusion
   of the argumentation scheme.

**Output:**
Produce JSON output containing the classification and a detailed breakdown
of the analysis:

- **"classification"**: A string value, either "Present" or "Absent."
- **"details"**:
  - **"premises"**:
    - **"initiative"**: Assessment of whether the text matches Premise 1.
    - **"endangerment"**: Assessment of whether the text matches Premise 2.
    - **"conclusion"**: Assessment of whether the text matches the conclusion..
```

A.3 1-vs-rest: Jeopardy

```
**Prompt:**
```

```
Given a text discussing a political, social, or economic initiative,
classify the text according to the provided argumentation scheme.

**Instructions:**

1. **Identify the Elements of the following argumentation scheme:**
   - **Premise 1 (Initiative):** A past, present, or intended action (Action A) that
   aims to transform or significantly improve a condition (Condition C) in the
   political, social, or economic order. The condition may be implicit.
   - **Premise 2 (Endangerment):** Implementing Action A will endanger or compromise
   an established and valuable Accomplishment B. Accomplishment B may be implicit, bu
   it should be clear what is at risk. The accomplishment must be distinct from
   condition C. Merely mentioning no or an insufficient improvement in Condition C
   is not enough to match this premise.
   - **Conclusion:** Action A should not be undertaken, stopped, or significantly
   altered because it will be endanger a valuable Accomplishment B.

2. **Classification:**
   - **Present:** If the text aligns with the premises and conclusion
   of the scheme above.
   - **Absent:** If the text does not align with the premises or conclusion
   of the argumentation scheme.

**Output:**
Produce JSON output containing the classification and a detailed breakdown
of the analysis:

- **"classification"**: A string value, either "Present" or "Absent."
- **"details"**:
  - **"premises"**:
    - **"initiative"**: Assessment of whether the text matches Premise 1.
    - **"endangerment"**: Assessment of whether the text matches Premise 2.
    - **"conclusion"**: Assessment of whether the text matches the conclusion..
```

A.4 Multiclass

```
# Instructions
```

```
Given a text discussing a political, social, or economic initiative,
analyze the text according to the provided argumentation scheme. Think step
by step.

## Premise 1: Initiative
Does the text describes a past, present, or intended action (Action A) that aims to transform
or significantly improve a condition (Condition C) in the political, social, or economic order?
Condition C may be broad (e.g. the economic situation or social injustice).
If you can identify the category of the action (e.g. economic or environmental policy), this is sufficient for Condition C.
The text might describe both an existing action that was implemented by an incumbent party and a new one proposed by
the authors of the text. In this case, focus on the **old** one.

## Premise 2
Which of the following premises applies to the text? Let's think step by step.
Remember that if there is both a new and an old action described in the text,
we are interested in whether one of the following premises fits the description of
the **old** one. Think step by step.

### 1) Unacceptable Cost or Endangerment
Action A comes with an unacceptable high cost or unacceptable consequences.
This is usually the case when Action A endangers or threatens a previous accomplishment.
The unacceptable consequences have to be distinct from Condition C.
Merely mentioning insufficient improvement in Condition C is not enough to match this premise.
Worsening Condition C does not match this premise, evalute premise 3) for this.

### 2) Ineffectiveness
Action A, despite its intentions, will fail to produce any significant change or improvement in Condition C.
The action may be ineffective due to various reasons, but the text should clearly suggest its ineffectiveness.
Ineffectiveness is not the same as negative consequences or worsening the Condition C, for which you should
evaluate premise 3).
```

```
### 3) Contrary Effect
The text argues or implies that implementing Action A explicitly worsens Condition C.
Instead of improving Condition C, it achieves the opposite and worsens it.
Merely describing broad negative consequences of Action A is not enough to match this premise,
evaluate premise 2) again.
If Action A has a negative consequences that are not directly related to Condition C,
evaluate premise 1) again.


### 4) None of the above
Select this premise of no other premise applies.

## Output
Produce JSON output containing the classification and a detailed breakdown of the analysis:

- "reasoning": "Briefly explain your reasoning. Think step by step. Be precise. Don’t write a long text."
- "premise_1": "true if applies, else false"
- "premise_2": "1, 2, 3, or 4, depending on your answer for premise 2"
```