# Detecting Foundational Narratives in Parliament Speeches

**Matti Wiegmann**
University of Kassel
matti.wiegmann@uni-kassel.de

**Jürgen Neyer**
European University Viadrina
neyer@europa-uni.de

**Benno Stein**
Bauhaus-Universität Weimar
benno.stein@uni-weimar.de

## Abstract

Narratives are a common tool for expressing and consolidating the identity and beliefs of groups of people. In politics, foundational narratives describe deeply held beliefs about power and collective identity which political actors often use tacitly to convince supporters and consolidate power. Although these narratives are highly relevant to understand political discourse, their algorithmic detection and quantification is a new venture in NLP research, as it requires knowledge of, and competent reasoning about, political discourse. In this paper, we formulate the task of foundational narrative detection as a one-class classification problem. We present a new dataset comprising English speeches from the European Parliament that have been annotated by expert political scientists. We also propose an effective classification approach, evaluating the impact of model selection, task formulation and explicit knowledge on classification accuracy. However, due to the complexity of the task and the influence of biases and subjectivity, we do not consider direct classification, whether by human or LLM, to be the ultimate solution for labelling political texts according to narratives. As an alternative, we present two prompt strategies with comparable performance that classify political texts *indirectly* by reasoning over generated pro and con arguments.

Our data is published at https://doi.org/10.5281/zenodo.19821274

## 1 Introduction

Political discourse is shaped by narratives. These narratives emerge as stories in which agents act towards their goals in a series of events, serving as a sense-making tool and rhetorical device to frame political ideas. Narratives in political discourse are collections of persistent beliefs and interpretations of past events that shape the identity of their supporters.[1] Groth (2019) offers the example of "Pizzagate" (Tuters et al., 2018), an online conspiracy story which sees Democratic politicians performing occult rituals in the (non-existent) basement of a pizza restaurant. The persistent narrative behind Pizzagate is that of the "deep state", "overly big government", and the "corrupt insiders". Foundational narratives are the most far-reaching narratives, those of nations. They establish a common identity and idea of sovereignty (e.g. European unity) by framing the history and values of people and so foster belief in the distribution of power.

Narratives are used in political discourse to inscribe political interests into the narrative as a persuasive argument (Groth, 2019).

As narratives are well known and widely believed within their respective groups, they become powerful argumentative tools. Framing a political idea as part of the narrative or alluding to it convinces its supporters. For example, the Trump administration garnered support for the removal of civil service protections under "Schedule F" by framing policy-influencing civil servants to be part of the "deep state", using the so-called "drain the swamp" narrative. According to this narrative, "unaccountable, policy-determining federal employees [...] put their own interests ahead of the American people's" and allow "corruption to fester in agencies".[2]

Political actors often unconsciously rely on foundational narratives. As these narratives shape the identity and preferences of political actors at both the individual and collective levels, they also provide frameworks for navigating the complexities of EU politics. Consequently, they become deeply embedded in political identities and ways of reasoning. Consequently, such narratives are used in political argumentation without drawing on stereotypical narrative elements.

[1] See Section 2 for an overview of terms and definitions.

[2] Notably, the "Section F" fact sheet (The White House, 2025) explicitly lists "Drain the Swamp" as a section title.

**Definition (minimized):** The supranational narrative refers to a form of political integration in which the European Commission, the European Parliament, and the European Court of Justice are key political actors that shape the policy.

**Positive instance:** The Commission did not hesitate to act in its role of guardian of the treaties against a national legislation that restricted the right to freedom of association of civil society organisations by limiting their access to funding, in breach of EU law and of the Charter, as confirmed in 2020 by the Court of Justice.

**Reasoning:** The paragraph clearly belongs to the supranational narrative. The speaker refers to two of the three main actors in the supranational narrative: the EU Commission and the European Court of Justice. The EU Commission acts in its role as guardian of the treaties and monitors compliance with them, supported by the European Court of Justice, which has issued a judgment in this legal dispute.

**Negative instance:** We had a rule of law that had evolved since the days of Alfred The Great, but English law is now superseded by EU law. Impartial English courts are now overridden by the political courts of the European Court of Justice and the European Court of Human Rights. Our ancient protections of individual liberties such as Habeas corpus have been swept away by such measures as the European Arrest Warrant.

**Reasoning:** The paragraph does not correspond to the supranational narrative. The speaker refers to one of the three main actors in the supranational narrative: the European Court of Justice. The speaker questions the primacy of EU law over national, British law and sees the European Court of Justice as a threat to the rule of law and individual freedoms.

Figure 1: Positive and negative example passages annotated for the supranational narrative including a manually formulated reasoning for use in the prompts.

Tracking the use of foundational narratives is of great interest to political scientists as it enables them to gather empirical evidence for diachronic analyses, for example, of narrative alignment to policy areas or stages of the policy process, an increase in narrative-driven debate over factual debate and rising Euroscepticism.

However, detecting foundational narratives in a body of text is challenging. Unlike in storytelling, where characters and events are explicitly described and can be extracted using established techniques, foundational narratives are often implicit and only tacitly in a speakers mind. Furthermore, subtle political discourse rarely states or references narratives explicitly as in the "drain the swamp" example above. An NLP system designed to detect foundational narratives requires knowledge of the real world, the ability to interpret and contextualize ill-defined narratives, and the ability to deal with subjective, subtle political discourse full of rhetoric. We see a lack of research on how NLP systems handle such political texts.

In this paper, we quantify the problem of identifying foundational narratives in a corpus of political texts. We propose to formulate this task as a one-class classification problem: *"Does a given text use a given, defined narrative or not?"* We then use this formulation to analyze the ability of LLMs to reason about political questions.

Our contributions are as follows. (1) A dataset containing 800 speech passages from the European Parliament (EP), annotated by expert political scientists for the strategic use of two narratives on European sovereignty: the dominant supranational narrative *"EU institutions should have power"*, and the national counter-narrative *"nation states should have power"* (Section 3). (2) A classification study to quantify the efficacy of external knowledge about the narratives, to quantify the consistency of model decisions, and to create effective and useful prompts. We find that, with our most effective prompt, models can solve this task well ($0.8 - 0.9$ balanced accuracy) but show inconsistent or poor behavior when they are given a lot of external knowledge (Section 4). (3) An evaluation of two indirect, argument-based prompting strategies using `Persuasiveness` and `Argument Quality`. We show that models can have a preference or bias towards one narrative or another and propose to use indirect prompting as alternative. We show that these strategies can reduce such biases with comparable performance (Section 5).

## 2 Related Work

Narratives are a common and ambiguous concept, and what constitutes a "narrative" in research depends on the field and object of study. The most simple definition of a narrative is "the representation of an event or a series of events" (Abbott, 2008), an assessment shared by Shenhav (2015), who notes that, although many more narrative elements can be identified (e.g. agents, temporality, setting, and perspective as categorized by Piper et al. (2021)), these are often optional. In the linguistic theory of Genette (1980), these structural elements are part of the *Story*, which is explicitly written and thus parsable. Genette (1980) adds two more aspects of meaning to narratives, *Discourse* and *Narration*, and Shenhav (2015) adds a fourth, *Multiplicity*, for narratives in politics.

The distinction of narratives between the ex-

plicit sense of a "story" and the tacit sense of "foundational narrative" is a source of confusion, since the term *"narrative"* is often used for either. From a literary science perspective, Abbott (2008) distinguishes these two conceptions of narratives as "compact and definable" versus "loose and generally recognizable", offering the term "micro-narrative" for the former. The second conception of narratives can be defined as the stories of nations, religions, or cultures, "embraced by a group that also tells, in one way or another, something about that group." (Shenhav, 2015). Various names have been proposed for these tacit narratives, such as "narrative beliefs" (Piper et al., 2021), "deep stories" (Hochschild, 2018), "social narratives" (Shenhav, 2015), or "metanarrative" (Lyotard, 1994; Kaplan et al., 2022). In this work, we refer to explicit micro-narratives as *"stories"* and to the tacit conception as *"narrative"*. Our work concerns foundational narratives of the European Union (EU).

Most works in computer science are interested in either the generation and evaluation of literary narratives ("storytelling") (Xie et al., 2023), which is unrelated to our work, or in narrative understanding (Piper et al., 2021), the modelling and extraction of (primarily) structural elements of a story in a structured form. Examples of the latter, outside of the literary domain, include Graaf et al. (2016)'s analysis of health-related stories in advertisements, Falk and Lapesa (2023)'s collection of "argumentative narratives", stories used as a stylistic device for their persuasiveness, and (Korenčić et al., 2024) dataset of conspiracy theories annotated with structural elements (actors, objectives, victims, etc.).

Our work on narrative detection differs in that the narrative and its elements are known a priori to the speaker and the question is whether the text uses or relies on a narrative, especially when stereotypical actors or events are not explicitly mentioned. Consequentially, we use a simplified, binary annotation scheme instead of the span-based annotations more often seen in the related work.

Finally, Otmakhova and Frermann (2025) relate narratives and media framing by describing the relation between the use of narrative elements and stories in a political text and that it can be used for framing as defined by Card et al. (2015). Our work on tacit narratives also relates to framing: In both cases, the speaker is aware of the concept and uses it to tacitly shape the understanding of the text. A detector has to overcome this disconnect between speaker intent and textual expression. The difference is that the narrative is defined over elements (actors, events, ...) known to the reader, which is not necessarily the case for frames.

| | Query | | Random | |
|---|---|---|---|---|
| | Supranat. | National | Supranat. | National |
| Annotated | 200 | 200 | 200 | 200 |
| Positive | 152 | 55 | 132 | 12 |
| Negative | 30 | 130 | 34 | 167 |
| Disagreement | 18 | 15 | 34 | 21 |
| Krippendorff's $\alpha$ | 0.67 | 0.79 | 0.55 | 0.48 |

Table 1: Our dataset contains 800 annotated passages of EP speeches across two narratives. The passages were sampled using two strategies, Query and Random. Examples with disagreement were later discarded.

## 3 Data

We analyzed 800 political texts to detect the strategic use of `National` and `Supranational` narratives. The data consists of two manually annotated test datasets, sampled from the source data in different ways. Table 1 shows the data and annotation statistics and Figure 1 shows positive and negative example passages.

### 3.1 Narratives

We focus on foundational narratives about sovereignty in the EU (De Vries, 2023), as these structure the political discourse and influence many key positions. Their frequent and implicit use makes them ideal political narratives to study. The two selected narratives, `National` and `Supranational`, are oppositional in terms of allocation of power and should cause the least ambiguity in annotation.

The political scientist collaborating on the project derived definitions of the narratives based on the relevant related work (see Sweet and Sandholtz (1997); Rittberger (2003); Mancini (1998) for supranational and Moravcsik (2013); Tallberg (2008); Pollack (2003) for national). For each definition, a condensed version was created that summarizes the idea of the narrative in one sentence. Figure 10 (Appendix B) lists all definitions.

### 3.2 Dataset Construction

All example texts are selected from the $137,844$ speeches given in the EP by its members, covering five election periods from 2006 to 2022 (Husmann and Wiegmann, 2025). The speeches are divided into $474,545$ passages with an average of $70.94$ words ($\sigma = 40.03$), in order to homogenize the unit length and to isolate individual political expressions for analysis. We use the pre-existing passage separation provided by the EP.

**Dataset Sampling** We sample two test sets of passages for annotation from the sources, where each set is only annotated for the narrative it is sampled for. The passages in the first set are selected via keyword queries to maximize the probability of sampling true positives. The passages in the second set are randomly drawn from the source data to avoid any potential bias from the query terms.

The query for each narrative is formulated in conjunctive normal form, containing a clause for each of the three common narrative elements (actors, events, purposes) and with three terms per clause. The terms are selected based on the narrative definitions (see Table 2, Appendix A).

**Annotation** Both test sets are annotated by two trained graduate students of political science who have been employed and paid for the duration of the campaign. During the training phase, the annotators and the expert political scientist collaborated to design the annotation guidelines using held-out examples. During the annotation phase, each annotator labeled whether or not the designated narrative is used. Note that each example was only annotated for the narrative it was sampled for and there are no overlapping examples. While there are some cases where a speech expresses both narratives, we found this to be so rare it could be ignored during annotation.

### 3.3 Results

As shown in Table 1, the test sets contain 367 examples of type (`Query`) and 345 examples of type (`Random`). Examples with disagreement are excluded from the test data.

As expected, the data is unbalanced for both narratives, with more positive examples for the mainstream supranational narrative and more negative examples for the counter-narrative. Notably, the ratio of positive examples is higher for examples in the `Query` test set, but not enough to balance the classes. Thus, query-based sampling is an improvement, but more effective pre-sampling methods could be applied. The observation supports the thesis that the majority of narrative usage is implicit and that lexical methods are unsuitable for recognizing narratives.

The table also shows the annotator agreement, which is high (Krippendorff's $\alpha : 0.67 - 0.79$) for the `Query` test set and acceptable ($\alpha : 0.48 - 0.55$) for the `Random` test set.

The causes of disagreement is likely the expected subjectivity in interpreting political text and the descriptive nature of the annotation guidelines, which leaves decisions about edge-cases to the discretion of the expert annotators. The annotators especially reported difficulties annotating examples where an argument for a positive annotation could be made but which are not clearly covered by the narrative definitions.

The differences in agreement between the data sets may be caused by (dis)-similarity to the definitions and differences in example length. Query sampling retrieves examples that are more similar to the narrative definitions in topic and terminology, as the keywords are derived from the definitions, so it may be more clear when an example is positive or negative. Random sampling selects more short examples (`Random` token length: $\mu = 82, \sigma = 42$; `Query`: $\mu = 125, \sigma = 80$), which may be more difficult to interpret.

## 4 Classification Study

We present two experiments to evaluate the competence of reasoning language models in detecting narrative in political texts. The first experiment assesses different design parameters and task consistency. The second experiment assesses differences between models and narratives in both realistic and optimized scenarios.

### 4.1 Detecting Narrative Use

Detecting narrative use is a knowledge-dependent, one-class reasoning task, as the positive class definition must be known, the negative class is ill-defined, and the classification decision is subjective and often requires an explanation. Five design questions follow from these challenges: (1) Definition: What knowledge about the narratives must be provided in the prompts? (2) ICL: What knowledge can (only) be learned in-context? (3) Classes: Can an opt-out decision for uncertain cases increase model effectiveness for the remaining classes? (4) Narratives: Are there differences between narratives? (5) Consistency: As the task is subjective, are the models consistent, or does the sampled generation of the reasoning affect the outcome? In addition, the typical question in LLM classification tasks are also relevant: which model is most effective and are there scaling effects?

### 4.2 Experimental Design

This classification study can be implemented via a configuration system, in which the various parts

of a prompt are included according to parameter options. Figure 9 (Appendix B) shows the prompt templates for all parameters. A parameter can then be considered influential if it differs significantly across the tested configurations.

The definition is tested with two parameters: as `Full`, the long definition provided by the political scientists, or as `Condensed`, the one-sentence summary. The ICL effectiveness is tested by (not) adding example cases with manually created reasoning (see Figure 1). The classes are tested with two parameters: as `Binary`, which asks a 'yes or no' question, or as `Opt-out`, which allows a third 'unsure' option to simulate the one-class paradigm. 'Unsure' examples are removed from the evaluation. It should be noted that 'if unsure, answer no' would represent this paradigm better but the tested models performed poorly on this on preliminary tests. Narratives are tested independently for `National` and `Supranational`. Consistency is tested by repeatedly testing every configuration five times to estimate the variance. Model differences are tested using five models: Llama-8B as a non-reasoning model, Phi-4-B as a small model, Qwen3 in 8B and 32B, and GPT-5.1-mini.

Testing the full search space of 400 model runs would be too expensive. However, by splitting the study into two smaller experiments, it is possible to assess the differences between the parameters at a much lower cost. Experiment 1 tests the definition, ICL, classes, narratives and consistency using the cheaper Llama and Qwen 8B models only. Experiment 2 tests all five models for both narratives on two configurations selected based on Experiment 1's results: The `Realistic` configuration {`Binary`, `Full`, `ICL`} and the `Optimal` configuration {`Opt-out`, `Condensed`, `ICL`}.

Note that we only analyze the `Query` test dataset in depth. Figure 5 shows only small differences between the dataset for classification, so we assume that our findings generalize to the `Random` dataset.

**Evaluation Metric** The effectiveness is evaluated via balanced accuracy (BA), defined as the average recall between the positive and negative classes:

$$\text{BA} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP}+\text{FP}} + \frac{\text{TN}}{\text{TN}+\text{FN}}\right)$$

The BA is preferable to Accuracy because the dataset is imbalanced. Although the $F_1$-score would be generally adequate in our case, however, the harmonic mean increasingly punishes small precision or recall values. Since some classes in the dataset (i.e. positive national examples in the `Random` sample) have very few examples, a single deviating example can load to large score difference. As LLMs with CoT-prompting or reasoning have an element of randomness, such deviations are to be expected by chance. The BA score was chosen as it is more robust in such cases.

### 4.3 Results and Discussion

The results show that the models are effective at narrative detection (BA: $0.8$ – $0.9$) using our prompt configurations, given that the problem is subjective even for experts ($\alpha$ : $0.67$ – $0.79$). While it was expected that larger models would be more effective, it is surprising that adding more knowledge to the prompt can sometimes decrease performance and that narratives differ in their susceptibility to model and prompt choice. The results of the parameter comparison are shown in detail in Figure 2 and aggregated across configuration parameters in Figure 3. The results of the model comparison are shown in Figure 4.

**Explicit Knowledge: Definitions and ICL** The hypothesis that adding more external knowledge via the prompt increases model effectiveness must be partially rejected. Generally, using the `Condensed` definition is significantly better (Welch's $t$: $p = 0.027$) than using the full definitions, although the effect size is small. This is surprising because the working hypothesis is that adding more external knowledge should be more effective. Using ICL is significantly more effective ($t$: $p < 0.001$) than not using it, which is often the case with LLMs in general. The effect of explicit knowledge varies depending on the narrative. For the supranational narrative, using ICL and `Full` definitions is significantly more effective. However, combining both has no additional effect. For the national narrative, there is no significant difference between ICL usage. Additionally, the `Condensed` definition is significantly more effective with a large effect (ca. 0.1 BA). One possible explanation is that the models have a preconception of the national narrative from the pre-training data and adding a conflicting (scientific) definition confuses the model. It is unclear whether this effect is specific to the national narrative or whether it would similarly occur for other narratives.

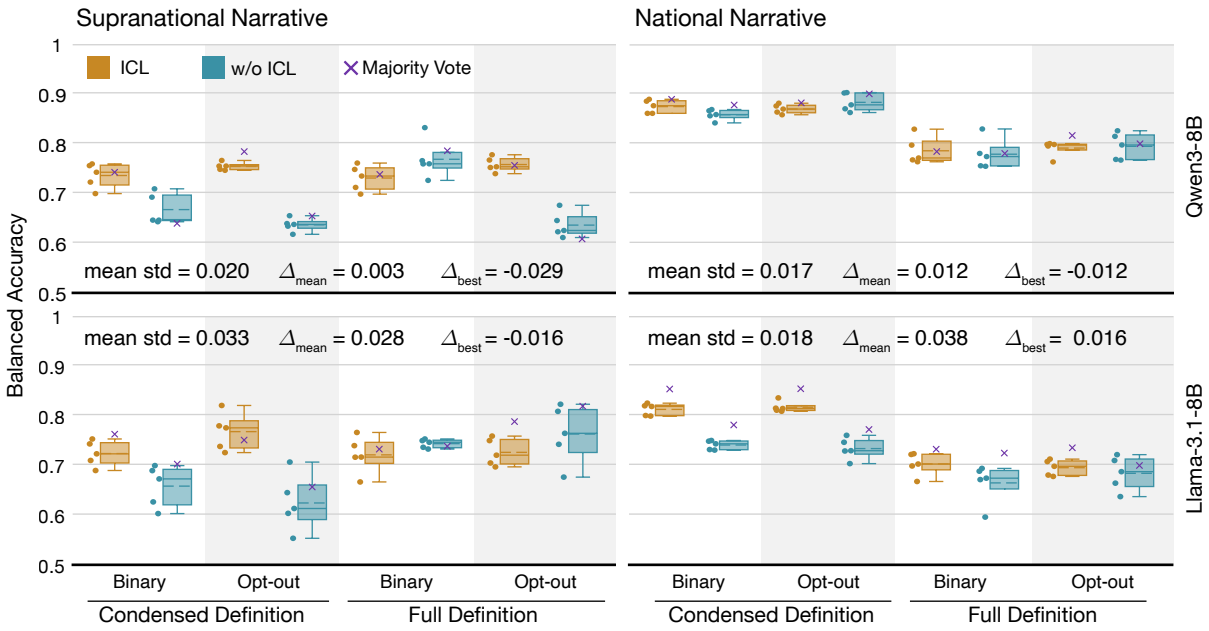


Figure 2: Variance of the BA scores across five repetitions for all configurations tested in Experiment 1. Also listed are the mean standard deviation of the configurations, the mean increase of a majority voting ensemble of the repetitions to the mean score $\Delta_{\text{mean}}$ and to the best individual score $\Delta_{\text{best}}$.

The limited utility of explicit knowledge is not easily explained by model bias (Appendix A.2) as measured via positive rate, i.e. a configurations' prior preference for accepting either narrative. There is no significant difference in positive rates between different definitions or ICL use, as opposed to the inherent bias of models and narratives.

For future experiments, the conclusion is to always add ICL, use the `Condensed` definitions for the `Optimized` configuration and the `Full` definition for the `Realistic` configuration.

**Classes** The hypothesis that the opt-out mechanism encodes the one-class principle and thus increases the scores must be rejected. There is no significant difference in BA when using the `Binary` or `Opt-out` settings, neither in general nor for the narratives individually. However, many of the top five most effective configurations (see Table 4 (Appendix A)) use the `Opt-out`, both individually (70%) and as a majority voting ensemble over the consistency repetitions (60%). Thus, we include the opt-out in the `Optimal` configuration, since the resulting combination is among the best, independent of the model or narrative, and because we cannot disregard interactions between multiple parameters. It should be noted that the more effective model, Qwen, uses the opt-out class rarely (2% of cases), as opposed to Llama (ca. 20%).

**Narratives** The the models are significantly more effective (0.7 BA) for the national narrative, although there is not conclusive explanation in our experiments. It should be noted that the positive rate of the national class is generally higher than in the ground truth and vice versa for the supranational class, resulting in a large and significant difference in positive rates relative to the ground truth of 0.36 (Figure 7 (Appendix A)). This bias is inverse to the class balance, so random misclassification is a likely explanation. For any analytical purpose, this means that the models will underestimate the mainstream narrative and overestimate the counter-narrative. This does not affect the BA measure and so does not explain the differences in model effectiveness.

**Consistency** The hypothesis that a self-consistency mechanism via a majority voting ensemble is significantly better must be rejected. Different repetitions of the same configuration have a high variance (STD: $0.008 - 0.035$ BA) and the difference between a good and a bad run can be as much as 10 percentage points. On average, the balanced accuracy of a majority voting ensemble is $0.003 - 0.038$ BA better than the configuration mean, but usually worse than the best run. However, neither the difference to the mean nor maximum are significant ($t$: $p > 0.05$), likely because of the small sample size. Nonetheless, self-consistency may be a means to avoid bad predictions at the price of runtime cost.

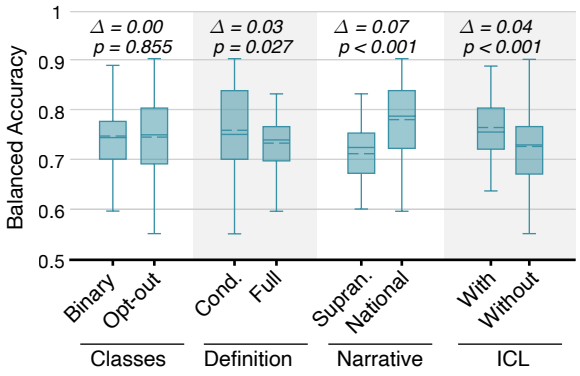


Figure 3: Variance of four configuration parameters across all configuration tested in experiment 1. Shown are the Welch's-$t$ statistics for unequal variances.

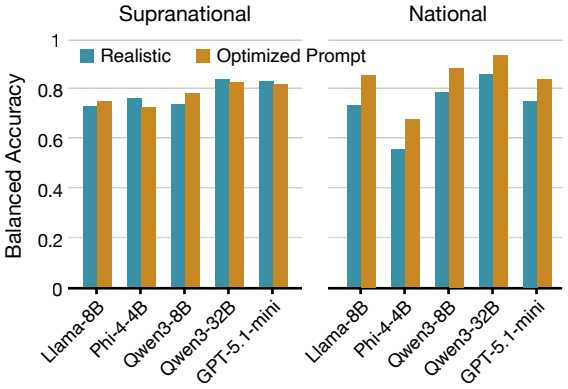


Figure 4: Effectiveness of five models for two prompt configurations and both narratives on the `Query` dataset.

**Models** Figure 4 shows that larger models (Qwen-32B and GPT-5.1-mini) are generally more effective than the smaller models, which is not surprising. The most notable difference is between narratives. There is a large variance in the national narrative between models (0.05 – 0.3 BA) and between configurations (about 0.1 BA), in contrast to the small differences in the supranational narrative. It is unclear if this variance is caused by model bias, so the model relies more on its "knowledge" than the instructions, or by characteristics of the narrative, such as clarity of the definition or how the narrative is expressed in the text. Consequentially, the `Optimized` configuration scores highly (0.93 BA) on the national test dataset, while scoring only good (0.84 BA) on the supranational data.

Models can have a bias towards over- or underestimating narrative support in general. Figure 7 (Appendix A) shows that there is a large and significant difference between relative positive rates between the models. As opposed to the narratives, this bias follows the class distribution and can not be a side effect of misclassification.

## 5 Argument-based Prompt Strategies

The large variance across repetitions and hyperparameters for the national narrative can be a result of model bias, resulting in a systematic deviation from the annotators. We propose two prompt strategies to avoid direct classification and instead classify via proxy measures: the persuasiveness and quality of pro and con arguments. Our focus is on evaluating classification effectiveness when using generated arguments.

### 5.1 Strategies

We propose two different strategies, `Persuasiveness` and `Quality`. Both strategies avoid direct questions (i.e. "Is the text using the narrative?") and instead analyze generated pro and con arguments "The text does (not) use the narrative, because [...]".

In the `Persuasiveness` strategy (see Figure 12, Appendix B), a model is given the narrative definition, passage text, and both arguments. It is then asked to decide which argument is more persuasive. The hypothesis is that a model produces weak arguments when the stance is not grounded in the text. For example, weak arguments tend to be shorter and lack references to the text or supporting statements. A second model should then be able to decide which argument is stronger.

In the `Quality` strategy (see Figure 13, Appendix B), a model is given the narrative definition, passage text, and one argument. It is then asked to rate the argument quality as *{low, medium, high}*, for each of Wachsmuth et al. (2017)'s nine argument quality criteria, as quantified by (Mirzakhmedova et al., 2024). The predicted scores are mapped to corresponding numerical values and summed. The hypothesis is that, while the individual scores may not be highly reliable, better arguments receive higher scores more often, which becomes visible in the sum.

The baseline strategy is `Classification` with the `Realistic` configuration as presented in Section 4.

**Argument Generation** All arguments are generated a priori, so they remain constant throughout the experiments. As shown in Figure 11 (see Appendix B), the generator model is given the narrative definition, the passage text, and the manually created ICL examples (both text and reason-

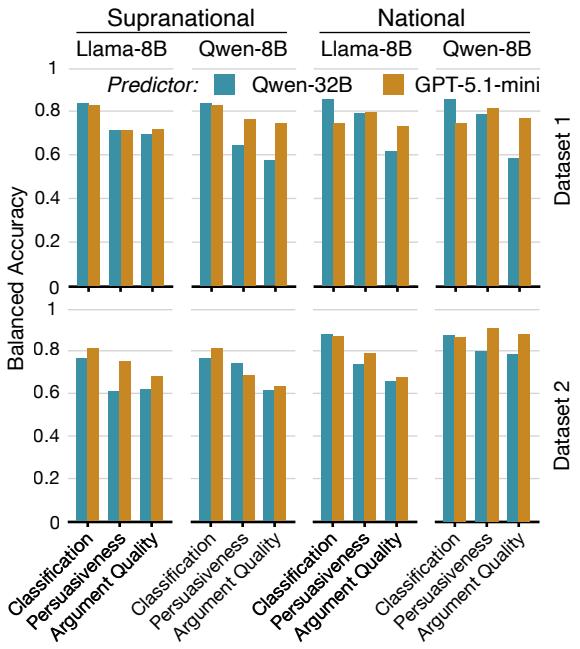


Figure 5: Effectiveness across three prompting strategies: Classification (Section 4), deciding based on the persuasiveness of generated pro and con arguments, and based on argument quality measures. Shown are two predictor models (Qwen, GPT) and two generator models (Llama, Qwen).

ing). The generator models used are Llama-3.1-8B, Qwen3-8B, and GPT-5.1-mini. All generated texts are manually cleaned to avoid model self-preferences and other spurious effects: (1) normalize quote and dash characters, (2) remove markup characters, such as * and #, (3) remove headlines and transform list items to inline lists, and (4) remove repetition of the input passage and reasoning-alike paragraphs that are not part of the argument.

### 5.2 Experimental Design

We designed two experiments to evaluate the effectiveness of the strategies. The first experiment compares the effectiveness of the `Persuasiveness` strategy using four predictor models (GPT-5.1-mini, Llama-3.1-8B, Qwen3-8B and 32B) across each of the three sets of generated arguments. This selection covers differences in model size and family. The second experiment assesses the effectiveness of all three strategies across two predictor models (Qwen3-32B and GPT-5.1-mini) and two generator models (Llama-8B and Qwen-8B), in order to cover both in-family and cross-family combinations. The results are evaluated as before.

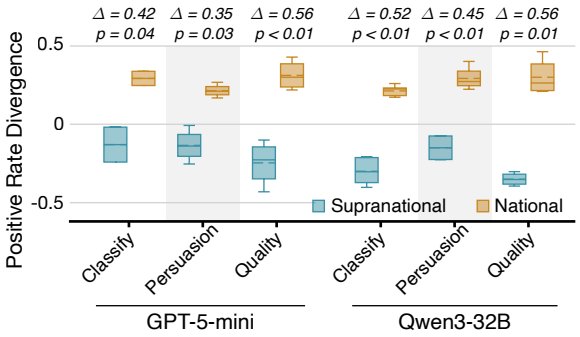


Figure 6: Variance of the positive rate divergence between the three strategies tested in experiment 2 across both datasets and the Kruskal-Wallis $H$ test scores.

### 5.3 Results and Discussion

The results show that the indirect strategies are competitive when the strongest predictor models are used, but generally lose performance. Notably, the predictor model is the only parameter with a significant impact.

**Model Selection** Figure 8 (Appendix''A) shows that the most significant difference is between the predictor models (Kruskal-Wallis $p < 0.001$), where larger models are generally more effective, with GPT being the most effective. The general recommendation for this strategy is thus to use the best possible model as predictor. There are no significant differences between the generator models ($p = 0.17$), narratives ($p = 0.11$), datasets ($p = 0.75$), or between using the same or a different model family for generation and prediction ($p = 0.34$). In practice, it seems conceivable to use a small, efficient model for argument generation.

**Strategy Comparison** Figure 5 shows that there are significant differences between strategies (Kruskal-Wallis $p < 0.001$) and narratives (Welch t $p = 0.006$), but not between datasets ($p = 0.64$), predictor ($p = 0.09$) and generator models ($p = 0.48$). Generally, the most effective strategy is `Classification`, except for the national narrative with GPT predictor, where `Persuasiveness` is slightly more effective.

**Bias Reduction** Figure 6 shows that the `Persuasiveness` strategy reduces the positive rate divergence (as a measure of bias, see A.2) compared to the baseline `Classification` by 0.07 due to a significant reduction in the bias of the national narrative. However, there is still a significant difference between the two narratives. The `Quality` strategy increases the class bias instead.

## 6 Conclusion

This paper addresses the problem of recognizing the strategic use of a predefined narrative in a political text. We formulate this problem as a one-class reasoning task and create a dataset of 800 examples annotated by experienced political scientists with good agreement ($\alpha : 0.5 - 0.8$).

Our first research question concerns the configuration space of narrative detection, evaluating model effectiveness and consistency, and the necessity of adding external knowledge. A notable finding is that prior knowledge (bias) of the LLMs is likely affecting the classification performance for some narratives. The effects can be observed, firstly, when adding knowledge through expert definitions or examples, as this can reduce the effectiveness of the model, and secondly because there is a large score difference between models and configurations. However, we only observe these effects for the national narrative, which we argue is more widely discussed outside of expert circles. Consequently, the models may have preconceptions that conflict with the given knowledge. It is also possible that the national narrative, as it is a counter-narrative, is therefore discussed more controversially in the training data.

Our second research question concerns the effectiveness of indirect prompting strategies that do directly classify political text as a means to avoid class biases. We propose using generated pro and con arguments as a proxy for the classification decision, either by establishing which argument is more convincing or by measuring the quality of the argumentation. We find that the indirect strategies are comparable in performance when using the best models, but never better, and that the `Persuasiveness` strategy reduces class bias compared to the baseline, but the `Quality` strategy does not. Notably, the predictor model has the greatest impact, while there are no significant differences between the models used to generate arguments, narratives or datasets.

## Limitations

The selection of the two narratives of sovereignty is the main limitation of this work. This means that the results may not be applicable to all narratives of sovereignty of the EU, it's nation states, narratives of non-EU nations, or political narratives in general. For example, we excluded "Intergovernmentalism" (nation state collaborate directly without EU institutions) as a (theoretical) third EU narrative of sovereignty since our annotators often found it exceedingly difficult to distinguish it from national or supranational narratives.

Additionally, the methodology is limited in that in one situation (the positive class of the national narrative with random sampling), the dataset contains too few examples ($< 30$) to be statistically stable. However, since all our results also replicate on the other dataset, this limitation does not diminish their validity.

Finally, the generated arguments in Section 5 were manually post-processed to eliminate side-effects from model idiosyncrasies. This limits the scalability of the proposed method and the effectiveness of automated post-processing is not clear.

## References


H. Porter Abbott. 2008. *The Cambridge Introduction to Narrative*. Cambridge Introductions to Literature. Cambridge University Press.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *ACL (2)*, pages 438–444. The Association for Computer Linguistics.

Catherine E. De Vries. 2023. How foundational narratives shape European Union politics. *JCMS: Journal of Common Market Studies*, 61(4):867–881.

Neele Falk and Gabriella Lapesa. 2023. StoryARG: a corpus of narratives and personal experiences in argumentative texts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2350–2372.

Gérard Genette. 1980. *Narrative Discourse: An Essay in Method*, volume 3. Cornell University Press.

Anneke de Graaf, José Sanders, and Hans Hoeken. 2016. Characteristics of narrative interventions and health effects: A review of the content, form, and context of narratives in health-related narrative persuasion research. *Review of communication research*, 4:88–131.

Stefan Groth. 2019. Political narratives / narrations of the political: An introduction. *Narrative Culture*, 6(1):1–18.

Arlie Russell Hochschild. 2018. *Strangers in their own land: Anger and mourning on the American right*. The New Press.

Sven Husmann and Matti Wiegmann. 2025. European parliament speeches 2002-2022.

Yael R Kaplan, Tamir Sheafer, and Shaul R Shenhav. 2022. Do we have something in common? understanding national identities through a metanarrative analysis. *Nations and Nationalism*, 28(4):1152–1172.

Damir Korenčić, Berta Chulvi, Xavier Bonet Casals, Alejandro Toselli, Mariona Taulé, and Paolo Rosso. 2024. What distinguishes conspiracy from critical narratives? a computational analysis of oppositional discourse. *Expert Systems*, 41(11):e13671.

Jean-François Lyotard. 1994. The postmodern condition. *The postmodern turn: new perspectives on modern theory*, pages 27–38.

G Federico Mancini. 1998. Europe: The case for statehood. *Eur. LJ*, 4:29.

Nailia Mirzakhmedova, Marcel Gohsen, Chia Hao Chang, and Benno Stein. 2024. Are Large Language Models Reliable Argument Quality Annotators? In *1st International Conference on Recent Advances in Robust Argumentation Machines (RATIO 2024)*, volume 14638, pages 129–146. Springer.

Andrew Moravcsik. 2013. *The choice for Europe: Social purpose and state power from Messina to Maastricht*. Routledge.

Yulia Otmakhova and Lea Frermann. 2025. Narrative media framing in political discourse. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9167–9196, Vienna, Austria. Association for Computational Linguistics.

Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative theory for computational narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 298–311. Association for Computational Linguistics.

Mark A. Pollack. 2003. *The engines of European integration: Delegation, agency, and agenda setting in the EU*. OUP Oxford.

Berthold Rittberger. 2003. The creation and empowerment of the european parliament. *JCMS: Journal of Common Market Studies*, 41(2):203–225.

Shaul Shenhav. 2015. *Analyzing Social Narratives*. Routledge.

Alec Stone Sweet and Wayne Sandholtz. 1997. European integration and supranational governance. *Journal of European public policy*, 4(3):297–317.

Jonas Tallberg. 2008. Bargaining power in the european council. *JCMS: journal of common market studies*, 46(3):685–708.

The White House. 2025. Fact sheet: President Donald J. Trump creates new federal employee category to enhance accountability. https://www.whitehouse.gov/fact-sheets/2025/04/fact-sheet-president-donald-j-trump-creates-new-federal-employee-category-to-enhance-accountability/. Accessed: 12.12.2025.

Marc Tuters, Emilija Jokubauskait?, and Daniel Bach. 2018. Post-truth protest: How 4chan cooked up the pizzagate bullshit. *M/C Journal*, 21(3).

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational Argumentation Quality Assessment in Natural Language. In *15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, pages 176–187.

Zhuohan Xie, Trevor Cohn, and Jey Han Lau. 2023. The next chapter: A study of large language models in storytelling. In *Proceedings of the 16th International Natural Language Generation Conference, INLG 2023, Prague, Czechia, September 11 - 15, 2023*, pages 323–351. Association for Computational Linguistics.

# A Additional Results

## A.1 Confounding Variables

We investigate the impact of confounding variables on the classification experiments from Section 4 and Section 5, as classification effectiveness in similar tasks (e.g. in argumentation and persuasiveness analysis) is often influenced by them.

**Method** As potential confounders, we investigate the length, type-token ratio (ttr), stop word ratio (swr), and fraction of named entity tokens, in the passage text and in the arguments. We then measure the difference in these values, firstly, between passage texts that are positive or negative for the respective narrative, and secondly, between arguments that align with the passage annotation (e.g. a pro argument when the passage is positive) and arguments that do not align (e.g. a pro argument when the passage is negative). The analysis is split by narrative (as the confounders may depend on it), and by generator model for the arguments. We assume a confounder when there is significant difference between the groups and we assume a harder significance level of $0.025$ as this is multi-testing.

**Argument Confounders** The results show that there are no significant differences in features for the arguments. This is expected for for ttr and swr, as machine generated text will is homogeneous. However, this also means that strong and weak arguments are not trivially separable by length or the reuse of entities found in the passage text.

**Passage Text Confounders** The results show that some feature measures are significantly different between the passage texts of the supranational texts: the ttr ($p = 0.003$) and the entity frequency ($p = 0.008$) are higher for positive passages and the swr ($p = 0.011$) is higher for negative ones. However, the differences in these values are small and an SVM trained on these features scores random performance, so we assume these confounders have no notable impact on classification models.

## A.2 Class Biases

Certain models or prompt configurations may show a preference towards accepting certain narratives and thus systematically classify a narrative more often as positive. We test for these biases via the *positive rate divergence*, which is the measured positive rate discounted by the true positive rate of the test dataset.

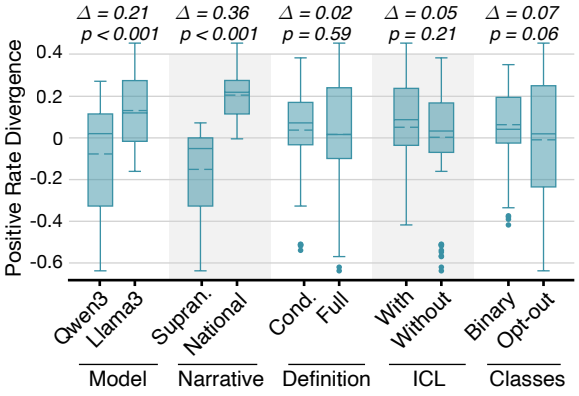


Figure 7: Variance of the positive rate divergence between models, narratives, definitions, ICL usage, and class options all configuration tested in experiment 1. Shown are the Welch's-$t$ statistics for unequal variances and the mean difference $\Delta$.

Figure 7 shows that there are significant biases between the tested models, narratives, and class options, although the the mean positive rate notably diverges from zero only for models and narratives. We can assume that Llama3 is biased towards accepting that a speech expresses a narrative and Qwen3 towards rejecting it. Further, we can assume that the supranational narrative is much more often accepted, and the national narrative slightly more often rejected.

These biases are not universal and a good configuration shows very little divergence. Figure 6 shows that a well-performing configuration with a strong model (Qwen-3-32B and GPT-5-mini) shows very little bias.

Furthermore, there is no significant difference in positive rate divergence between the three strategies discussed in Section 5 and, thus, we find no evidence that these indirect prompting strategies reduce existing biases. However, the tested configuration shows little bias to begin with, so the test might be inappropriate.

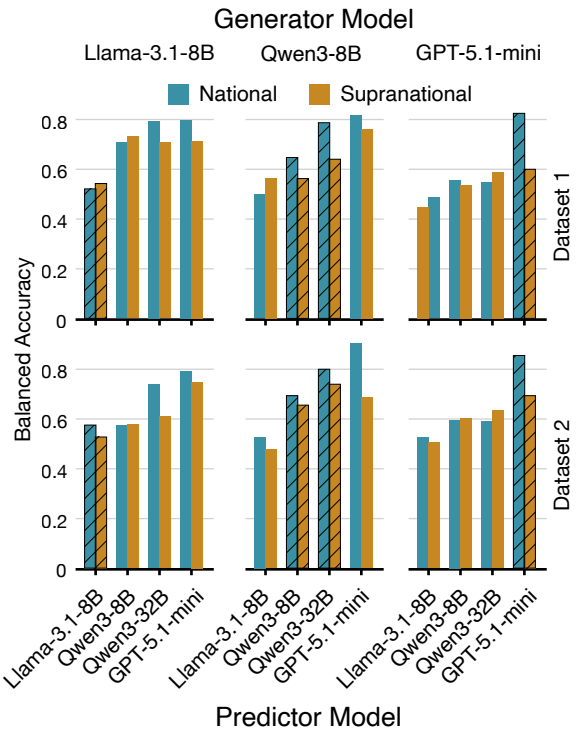


Figure 8: Effectiveness of the `Persuasiveness` strategy across four predictor and three generator models. Highlights indicate that predictor and generator are from the same model family and may contain self-recognition biases.

### A.3 Additional Tables

| | **Supranational** | **National** |
|---|---|---|
| Actor | European Commission | Member State |
| | European Court of Justice | Nation |
| | European Parliament | Government |
| Events | Law | Control |
| | Market | Protect |
| | Liberalization | Limit |
| Purpose | Integration | Sovereignty |
| | Human Rights | Identity |
| | Freedom | Culture |
| Hits | 1,778 | 677 |

Table 2: Terms used for the `Query` sampling strategy described in Section 3. Passages containing the term *"international"* are excluded so that the passages more often refer to EU domestic affairs.

| | **National** | | **Supranational** | |
|---|---|---|---|---|
| | $\rho$ | $p$ | $\rho$ | $p$ |
| Balanced Accuracy | -0.762 | <0.001 | -0.001 | 0.99 |
| False Negative Rate | -0.749 | <0.001 | -0.729 | <0.001 |
| Positive Rate | 0.739 | <0.001 | 0.674 | <0.001 |
| Precision | -0.633 | <0.001 | -0.648 | <0.001 |
| Recall | 0.0715 | 0.66 | 0.649 | <0.001 |

Table 3: Correlation between the frequency of the opt-out ('unsure' class) and various effectiveness metrics across all configurations of the first classification experiment.

## B Prompts

You classify political narratives in speeches. For a given passage from a political speech, decide if the speech supports the **[narrative]** narrative.

**Narrative Definition:**
**[definition]**

**Classification Task:**
Steps for analyzing and classifying speech passages:
1. Decide if the passage is unsubstantial, such as greetings, interludes, or unclear statements. In this case, answer with 'no'.
2. Decide if the passage has nothing to do with the narrative or it's actors. In this case, answer with 'no'.
3. Draft an explanation for if the passage supports the narrative or not, citing the passage content where appropriate.
4. Give your answer as **[classes]**.

Consider the following examples:

**Passage:**
**[ICL text]**
**Explanation:**
**[ICL reasoning]**
The answer is **[ICL truth]**

Now analyze and classify this speech. Remember to finish your answer by stating the class as **[classes]**.

**Passage:**
**[passage text]**

Figure 9: Prompt template for the `Classification` strategy used in Section 4 and Section 5. Bold text is variable. In-context (ICL) reasoning is manually created. Definitions and classes are shown in Figure 10.

| Supranational | | | | | National | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Score | Model | Definition | ICL | Classes | Score | Model | Definition | ICL | Classes |
| *Best Run* | | | | | | | | | |
| 0.77 | Qwen | Full | No | Binary | 0.88 | Qwen | Minimal | No | Opt-out |
| 0.77 | Llama | Minimal | Yes | Opt-out | 0.87 | Qwen | Minimal | Yes | Binary |
| 0.76 | Llama | Full | No | Opt-out | 0.87 | Qwen | Minimal | Yes | Opt-out |
| 0.76 | Qwen | Full | Yes | Opt-out | 0.86 | Qwen | Minimal | No | Binary |
| 0.75 | Qwen | Minimal | Yes | Opt-out | 0.82 | Llama | Minimal | Yes | Opt-out |
| *Best Majority Vote* | | | | | | | | | |
| 0.82 | Llama | Full | No | Opt-out | 0.9 | Qwen | Minimal | No | Opt-out |
| 0.79 | Llama | Full | Yes | Opt-out | 0.89 | Qwen | Minimal | Yes | Binary |
| 0.78 | Qwen | Full | No | Binary | 0.88 | Qwen | Minimal | Yes | Opt-out |
| 0.78 | Qwen | Minimal | Yes | Opt-out | 0.88 | Qwen | Minimal | No | Binary |
| 0.76 | Llama | Minimal | Yes | Binary | 0.85 | Llama | Minimal | Yes | Opt-out |

Table 4: Top 5 configurations by balanced accuracy all configurations of the first classification experiment. Also shown are top configurations of a 5-run majority voting ensemble.

**Definitions**

**National:** The national narrative champions the nation-states and the European Council as the leading European actors. Member states are portrayed as the leading political institutions and the drivers of European politics. The national narrative favors cooperation and negotiation among the member states and these member states control the supranational institutions. Whilst cooperation and negotiation among the governments is deemed crucial to the legitimate functioning of the EU, the control of the supranational institutions is an essential purpose. This narrative is underpinned by the assumption that preserving national sovereignty and borders is critical to maintaining cultural diversity and national identity.

**National Minimized:** The national narrative refers to a form of (non-)integration in which the member states shape policy sovereignly and independently of the EU institutions.

**Supranational:** The supranational narrative employs the supranational institutions as the main actors: the European Commission, the European Parliament, and the European Court of Justice. The European Parliament is the strongest protagonist of supranational democracy that aims for ever closer integration of states in the EU. Since it is directly elected, it guarantees that EU policies are democratically legitimated. The Commission acts as the 'Guardian of the Treaties', making fair rules for all and identifying violations. The European Court of Justice settles legal disputes and imposes fines in case of ongoing violations of the legal order of the EU. In the supranational narrative, these institutions collaborate to deepen market integration and strengthen the rule of law, democracy, and human rights.

**Supranational Minimized:** The supranational narrative refers to a form of political integration in which the European Commission, the European Parliament, and the European Court of Justice are key political actors that shape the policy.

**Classes**

**Binary:** 'yes' or 'no'.
**Opt-out:** 'yes', 'no', or 'unsure'. If you are not entirely certain, then answer with 'unsure'.

Figure 10: Narrative and class definitions used in the prompt templates.

You are given a passage from a speech in the EU parliament and the **[narrative]** political narrative. You argue that passage supports the narrative (PRO).

**Narrative Definition:**
**[definition]**

**Passage:**
**[passage text]**

Examples for CON argumentation:
**Passage:** **[ICL text]**

**Argumentation:**
This passage supports the **[narrative]** narrative. **[ICL reasoning]**

Write a correct and persuasive argumentation why this passage supports the narrative (PRO).

Figure 11: Prompt templated used to generate pro (and equivalently con) arguments used in the prompting strategies presented in Section 5.

Below is a passage from a speech in the EU parliament and two arguments. The PRO argument reasons if and why the passage supports the **[narrative]** narrative. The CON argument reasons that the passage does not support the **[narrative]** narrative. You decide which argument is more correct and persuasive.

**Narrative Definition:**
**[definition]**

**Passage:**
**[passage text]**

***PRO argument:***
**[pro argument]**

***CON argument:***
**[con argument]**

If the PRO argument is correct and more persuasive, then answer with PRO. Otherwise, answer with CON.

Figure 12: Prompt template for the `Persuasiveness` strategy presented in Section 5.

Below is a passage from a speech in the EU parliament and an argumentation for if the passage supports the **[narrative]** narrative or not (PRO or CON). You rate the quality of the argumentation according to the provided quality measure.

**Narrative Definition:**
**[definition]**

**Passage:**
**[passage text]**

**Argumentation:**
**[argument]**

**Quality Measure:**
**[measure]**: **[measure definition]**

Rate the **[measure]** of the author's argumentation. Choose one of the options below and explicitly provide the rating number:
3 - High
2 - Medium
1 - Low

Figure 13: Prompt template for the `Quality` strategy presented in Section 5. The prompt was used for each of the nine measures individually.