POLI 176 Final Project Memo

Energy Independence and the Green Transition in EU Political Discourse: A Mixed-Methods Text-as-Data Study Across Major Energy Shocks

by

Micah Roye

Professor Roberts

University of California San Diego

December 2025

Research Question & Significance

This project analyzes how the European Council, the European Union's (EU) highest political body responsible for settling strategic priorities, frames energy independence and fossil fuel needs in relation to renewable energy during a period of major geopolitical and climate-related shocks. The corpus consists of energy-related statements drawn from EU Council deliberations between 2010 and 2016. These texts represent the collective positions of Member States at the agenda-setting level; therefore, they are a critical lens for understanding the way the EU interprets, prioritizes, and narrates the green energy transition. Council conclusions often blend heavily technical, political, and economic reasoning thus providing opportunity to measure shifts in rhetorical emphasis across changing political environments. The original EU Council Deliberations corpus was filtered using a keyword-based trimming procedure creating a more focused corpus of energy-relevant sentences, which were later split into fixed-length (200 word) units for supervised and Large Language Model (LLM) classification.

The central research question guiding the study is:

How do major geopolitical and climate shocks between 2010 and 2016 influence the European Council's rhetorical alignment between energy independence and renewable energy?

This question is particularly important because in the face of climate change and environmental degradation, the EU's long-term climate ambitions rely heavily on sustained political commitment to decarbonization through the use of renewable and green energy even during crises. If renewable energy is framed consistently as a core solution to energy security, this suggests that it is embedded deep in political integration of climate goals. If, instead, pro-green framing declines during periods of geopolitical instability, economic pressure, or dramatic change, this may reveal stark vulnerabilities in the EU's climate governance architecture while also shedding light on the conditions under which Member States choose to deprioritize decarbonization. Related scholarship examines the structural tension between EU climate and energy policy.

Dupont and Oberthur (2012) argue that the EU's institutional architecture has historically produced policy competition rather than corporations making it so that climate and energy objectives often advance together but are not fully integrated. Major international agreements are shown to create top-down pressure for integration leading to calls for increased climate conscious energy strategies but a lack of explicit deployment. Further examination of EU energy discourse comes from Herranz-Surralles (2016), who documents major discursive shifts in how the EU talks about energy security and strategic partnership. Geopolitical disruptions are shown to trigger rhetorical recalibrations across EU institutions pushing policy makers to reassess energy strategies. From here, Szulecki (2019) further strengthens the significance of studying rhetorical dynamics by arguing that the EU's discourse is not merely symbolic but critical as a foundational component of EU energy governance. This article highlights persistent tension between security-focused narratives and sustainability focused narratives.

Hand Coding

To answer the research question, this project uses both supervised learning and LLM based on a hand-coded categorization scheme to capture meaningful variation in political support for renewables. To guide consistent and replicable coding, this detailed codebook defines both categories, lists positive and negative indicators, and clarifies how coders should treat ambiguous or mixed-content sentences. This codebook distinguishes between Pro-Renewables / Pro-Green framing (Code 1) and Non-Renewables / Not Pro-Green framing (Code 0). Special guidance was provided for specific borderline cases involving terms such as "efficiency", "innovation", and "sustainability", which are buzzwords that can often appear in both green and non-green contexts. Coders were generally instructed to treat these terms as non-green unless they were explicitly tied to renewable or decarbonization goals. With this structure, reliability and clarity are improved to minimize subjective interpretation.

For hand coding, the unit of analysis was a 200 word chunk. From the trimmed energy corpus of approximately 6598 paragraphs, from the original 12937 paragraphs, the corpus was left with 2640 200 word chunks. A sample of 200 chunks was then constructed using stratified sampling for manual hand coding. Chunks were tagged as “likely positive” if they contained explicit renewable or decarbonization terminology. This produced 266 likely-positive chunks and 6,332 other energy chunks. This was done to avoid an overwhelmingly negative training set and to ensure that coders saw a mix of green and non-green frames when applying the codebook. Then, 100 chunks were taken from the likely positive stratum and 100 chunks were taken from the remaining energy chunks. This design balances substantive relevance and class coverage (ensuring many plausible positives so that the classifier would have enough to train off of).

The sample size is large enough to train an interpretable classifier, yet manageable for independent coders to manually hand code within the project’s time constraints. An improved study would include more independent coders and a larger sample size. Two coders then independently coded all 200 sampled chunks according to the codebook and without any outside help, including being unable to consult each other. Following coding, the two sets of labels were compared and disagreements were reviewed strictly using the codebook definitions. The raw percentage of agreement was 90% across the 200 chunks. The set was then restricted to the 180 chunks that were agreed on to create a “gold label”.

Intercoder reliability was assessed using Cohen’s Kappa, a standard metric for measuring agreement beyond chance with two coders. The two coders achieved $K = 0.712$ ($p < 0.001$), indicating that there was substantial agreement. This level of reliability is strong given that the meaning of statements in the corpus is often context-dependent and sentences may contain implicit or mixed signals. The higher Kappa scores reflects clarity of the categories and effectiveness of the codebook in guiding coder judgments. Substantial agreement demonstrates that the concept of pro-renewables framing is reliability interpretable by human coders and that the resulting gold-standard labels are sufficient for training a supervised model. The coders are able to consistently identify and distinguish between genuinely pro-green rhetoric from neutral, procedural, unrelated, or fossil-oriented content.

Most disagreements arose around ambiguous or technical sentences with buzzwords. Especially those referring to "sustainability", "efficiency", or "innovation" without specifically linking to decarbonization were often confusing. Coders also differed on mixed-motivation statements, for example: text emphasizing economic competitiveness while briefly mentioning environmental goals. Finally, weak or implicit pro-renewable language that was not explicitly articulated could have been taken out of its initial context creating some conflicts. These disagreements highlight complex yet natural interpretation boundaries present in political elite communication where climate and energy language is often steeped in subtlety in broader policy objectives.

Understanding these disagreements provides insights into how the codebook could be further improved. Additional examples that explicitly illustrate mixed-content cases and tightening rules for symbolic mentions could be beneficial. Clarifying that mere references to climate agreements or goals is not sufficient for a 1 unless the chunk presents renewables as a central solution should be necessary. Additionally, a third "ambiguous" category could be included in future work to avoid forcing coders into sharp 0/1 choice with relatively ambiguous and complex text.

Table 1: Codebook

Codebook: Classifying EU Council Energy Independence Discourse in 200-Word Chunk Unit of Analysis: ● A fixed-length 200-word chunk, following Grimmer, Roberts & Stewart (2022), who note that long political texts may be divided into "chunks of a few hundred words" to obtain consistent units and reliable model fits.		
Code	Label	Definition

1	Pro-Renewables / Pro-Green	A 200-word chunk is coded as 1 when its dominant thrust clearly and positively supports decarbonization by: • Advocating renewable energy expansion (solar, wind, geothermal, hydro, bioenergy). • Supporting electrification, “green transition,” or low-carbon technologies (hydrogen, heat pumps, grids). • Positively endorsing emissions-reduction targets or climate goals in a way linked to energy policy. • Explicitly framing renewables or green technologies as solutions, priorities, or central to energy security/independence. The entire chunk does not need to be exclusively pro-green, it just needs a meaningful part that advances a pro-renewable argument as a central component, not a passing comment.
0	Non-Renewables / Fossil Fuels	A 200-word chunk is coded as 0 when it does not clearly advocate a renewable-oriented transition. This includes: • Focus on natural gas,

		LNG, coal, oil, nuclear (unless clearly framed as a green/low-carbon solution). ● Discussions of security, diversification, affordability, competitiveness, etc., without linking them to renewables. ● Symbolic name-drops of “Paris Agreement,” “energy transition,” or “climate goals” without substantive endorsement of green actions. ● Purely procedural, institutional, legal, or economic content with no meaningful pro-green argument. ● Mixed content where renewables are mentioned but not advocated, or subordinated to fossil-centered strategies.

Measurement - Naive Bayes Supervised Classifier

Following the creation of the hand coded set, to extend the hand-coded stratified sample to the full corpus of 6,598 energy chunks, a supervised classification model was trained using the 180 gold-labeled chunks. A multinomial Naive Bayes (NB) model was implemented for three reasons: small labeled sample set, alignment with course material, and relative simplicity. First, in regards to the small labeled sample set, only 180 labeled chunks but thousands of potential word features from the document-feature matrix made the NB seem well-suited because of the strong independence assumptions; therefore, variance would be kept low in high-dimensional settings. Second, the NB model is covered in the Text-as-Data (Grimmer, Roberts, and Stewart, 2022) and seemed like a transparent baseline for the text classification. Finally, the model allowed for a relatively limited amount of code to create a model that would integrate well. Each labeled chunk was represented using a bag-of-words dfm. The NB model is fitted on training folds and then applied to held-out chunks from the larger energy chunks corpus to evaluate performance.

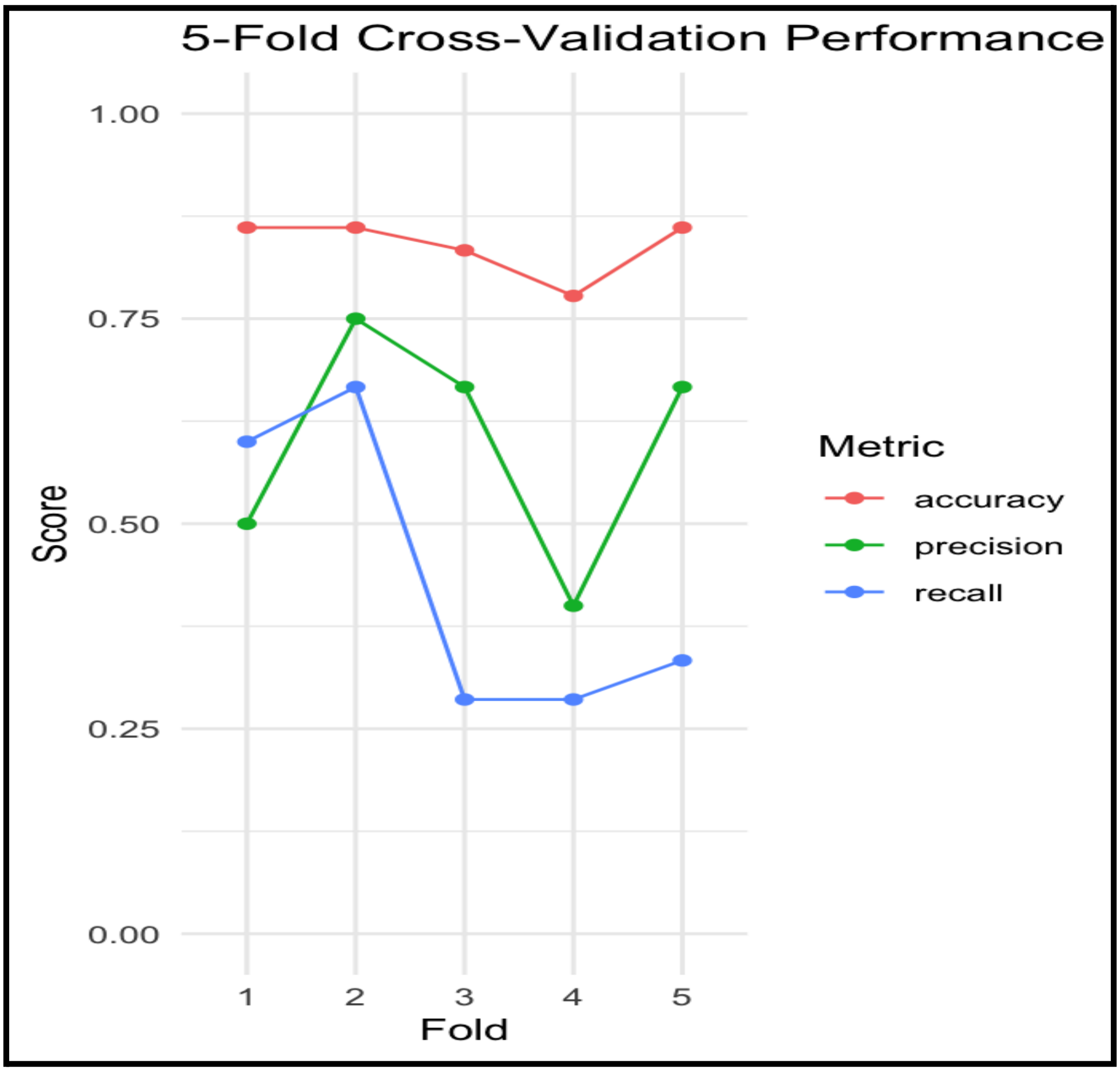
First, a simple train-test split using 80% of labeled chunks for training and 20% of testing was applied. The NB model achieved test accuracy of 0.75 but predicted no positives in the test set. Due to the rarity of the positive class (label 1), the classifier seemed to default to obtaining high accuracy by always predicting the majority class; thus, the model initially seemed useless for actually detecting pro-renewable framing. To obtain a more stable and accurate evaluation of the model, a 5-fold cross-validation test was applied on all 180 labeled chunks from the “gold label” set. In each fold, 80% of the data was used to train NB and the held out 20% was used to evaluate. For each fold, accuracy, precision, and recall were calculated for label 1. Averaging across folds the mean cross validation accuracy was around 0.84, the mean precision (label 1) was around 0.60, and then mean recall (label 1) was 0.43. Based on these figures, the NB model is relatively good at not hallucinating pro-green chunks, based on its decent precision; however, it misses a fair portion of true positives (with recall below 0.5). In other words, among chunks the model flags as pro-renewables, a majority are truly pro-renewable, but it is somewhat conservative in its guesses and fails to catch the more borderline pro-renewable cases.

Table 2: Accuracy, Precision, & Recall of NB Across 5-Fold Validation

Fold #	Accuracy	Precision (Class 1)	Recall (Class 1)
1	0.8611	0.50	0.60
2	0.8611	0.75	0.67

3	0.8333	0.67	0.29
4	0.7778	0.40	0.29
5	0.8611	0.67	0.33

Figure 1: Accuracy, Precision, & Recall of NB Across 5-Fold Validation



Next, the classifier was applied to all of the energy related chunks. A document feature matrix (dfm) was constructed for all these chunks, matching its features to the training vocabulary and predicting a label 0 or 1 and the probability of any given chunk to be in class 1. In the entirety of the corpus, 181 chunks were predicted to be pro-renewables; meanwhile, 6,417 chunks were predicted to be not pro-green / renewable. This implies that only 2.7% of energy-related chunks are classified as clearly pro-renewable. Furthermore, this minor percentage reinforces that strong green framing is rare even in energy discussions.

Measurement - Large Language Model (LLM) Classifier

In addition to the NB classifier, a LLM (Google's Gemini 2.5 Flash Lite) was used to label the energy-related text hunks as either pro-green (1) or not clearly pro-green (0). The LLM operates directly on the same 200-word chunk unit of analysis used in the hand-coding and Naive Bayes stages. A strict, format oriented prompt was used so that Gemini acts as a coder without being a summarizer. The prompt asked Gemini to classify each chunk based on a definition very similar to the codebook definition. By matching the conceptual distinction, found in the codebook, of pro-renewable vs non pro-renewable, the model was able to clearly identify. Additionally, the prompt included decision rules, to focus on whether the spark speaks clearly and directly in support of renewable energy or if only in abstract.

To evaluate the LLM as a measurement strategy the prompt was also applied to the 180 "gold-labeled" chunks that resulted from the initial hand coding. Using these "gold labels" as ground truth, Gemini's binary predictions could be compared to the human consensus labels.

Figure 2: LLM Performance on "Gold-Label" Energy Chunks

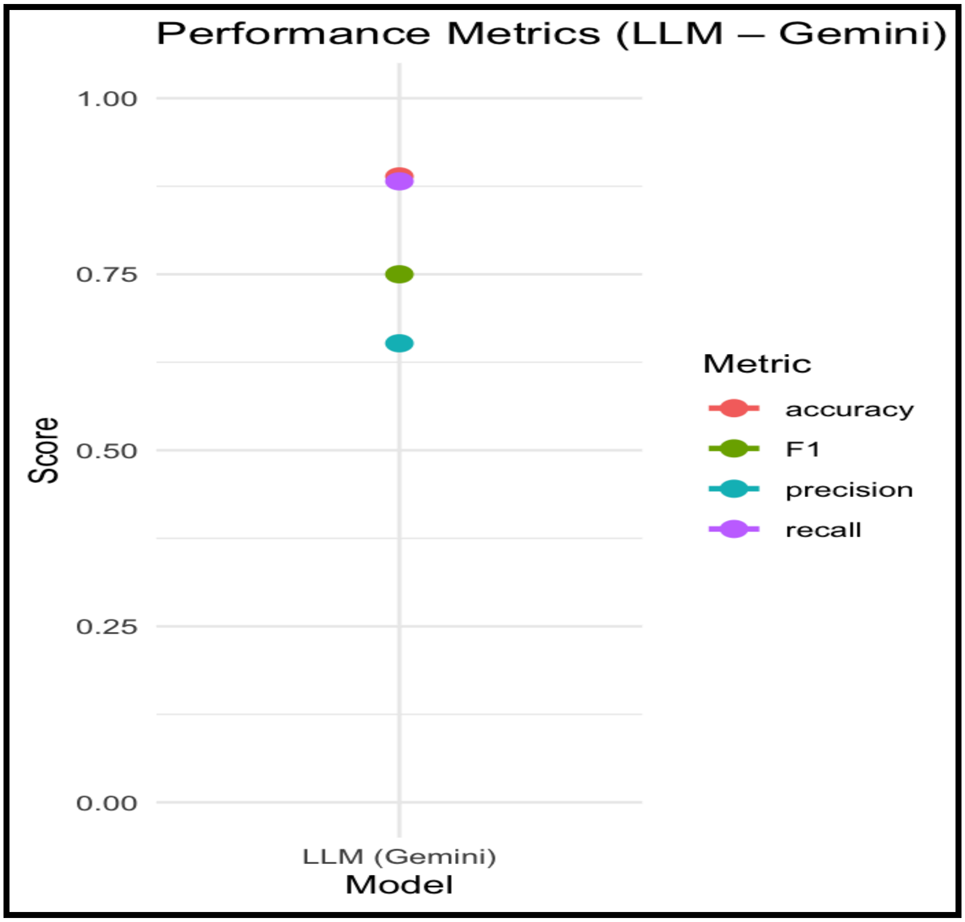
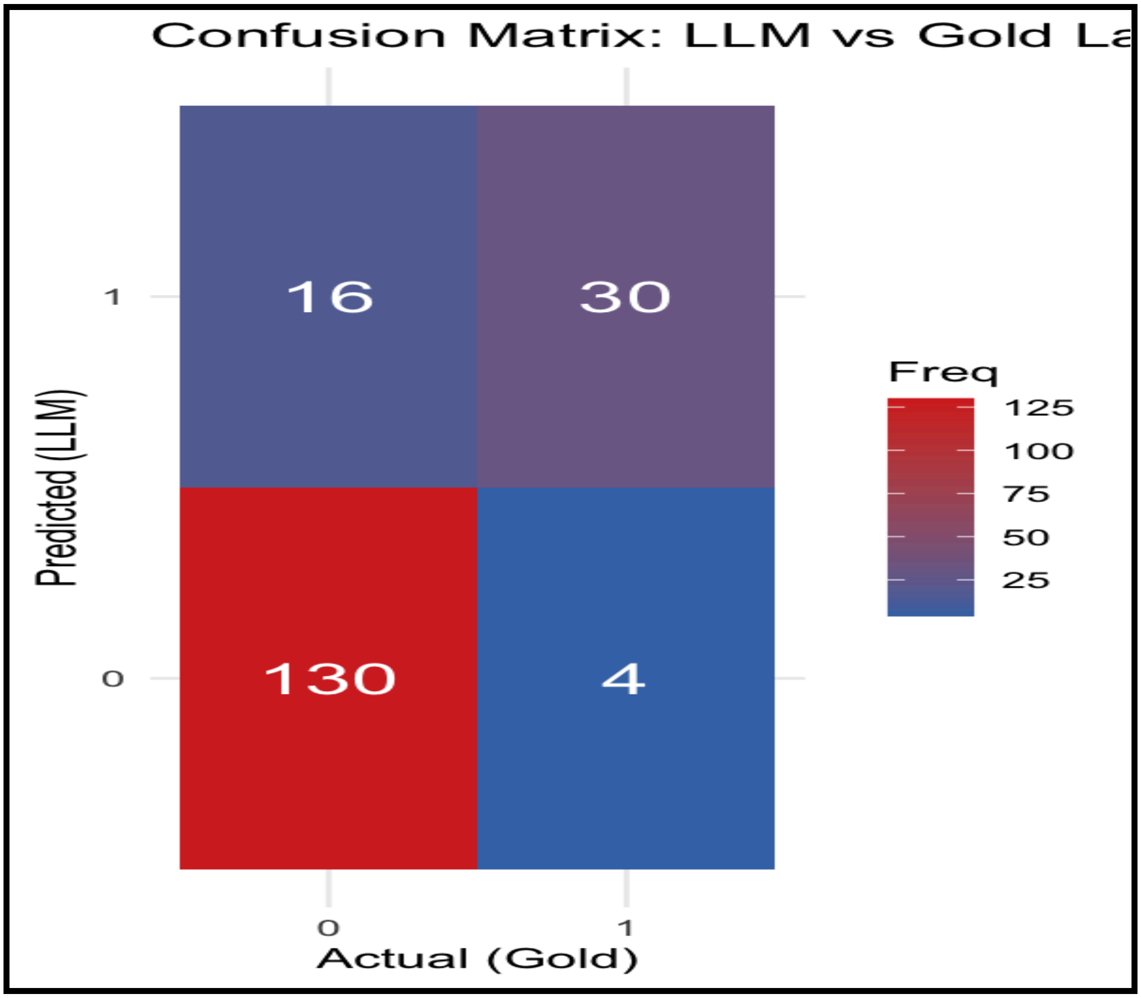


Figure 3: LLM vs Gold Label Confusion Matrix

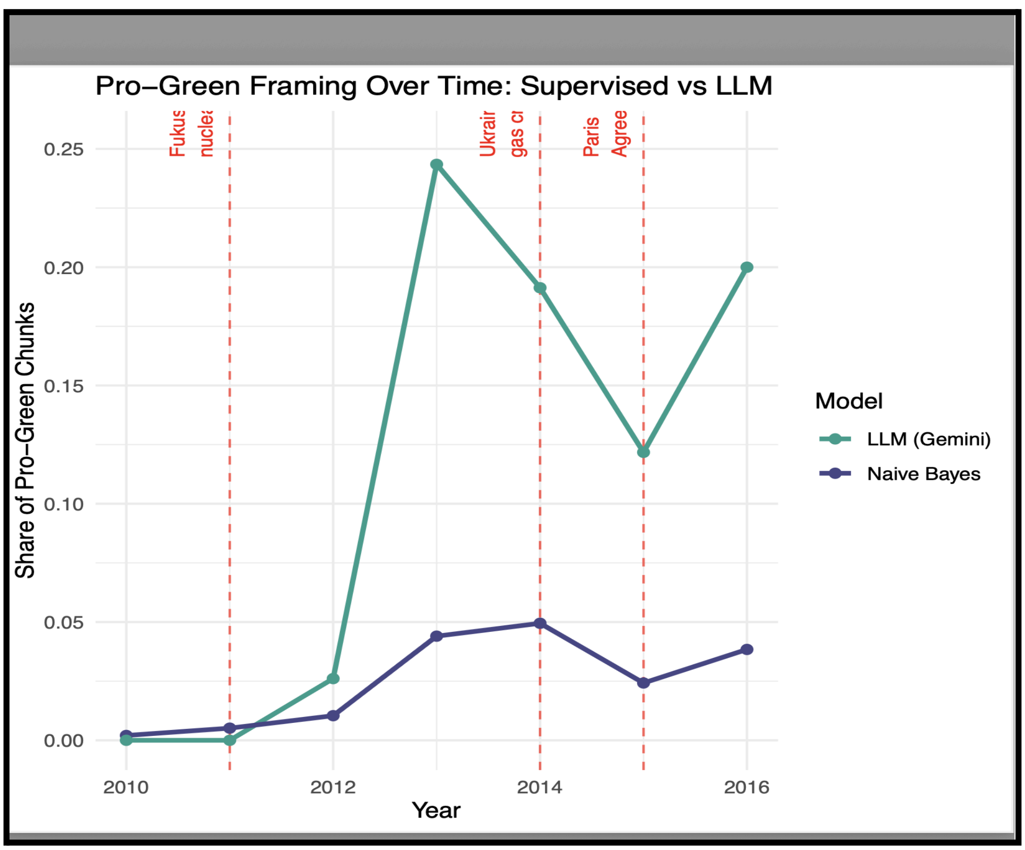


Based on the model's High recall (0.88) the LLM captured the vast majority of hand-coded pro-green chunks: only 4/34 positives were missed. The precision (0.65) is markedly lower, likely because of the presence of ambiguous and buzzword heavy energy chunks that even human coders found challenging (chunks with references to sustainability but no direct link to decarbonization). The LLM's recall (0.882) shows the presence of genuinely pro-renewable framing (according to the prompted definition) meaning that true positives were not systematically underestimated; therefore, green rhetoric in the "gold-label" corpus was accurately located. The F1 score (0.75) indicates that the LLM maintains a strong balance between not missing true pro-green chunks and not over-labeling ambiguous content as pro-green.Overall, the LLM seems to behave like a slightly cautious human coder.

The LLM model was then used to focus on the research question. To control costs, based on API rates, a stratified sample of 763 chunks from the 6,598-chunk energy corpus was taken, drawing roughly 115 chunks per year from 2010-2016. These chunks were sent to Gemini in batches of 40 chunks per API call. The LLM-coded labels in the 763-chunk sample were categorized by year from 2010-2016 creating strong indication that major geopolitical and climate shocks are associated with stronger rhetorical alignment between energy independence and the green transition in Council energy debates. As seen in Figure 3, following the Ukraine Gas Cut Off Crisis, renewable energy language decreased, likely due to a need for more presently secure forms of energy. Then, following the Paris Agreement, language around renewable energy again began to increase.

Figure 4: LLM Run 1

(The Shocks represented are 1. 2011 Fukushima Nuclear Disaster, 2. 2014 Ukraine Gas Cut Off Crisis, 3. 2015 Paris Climate Agreement)

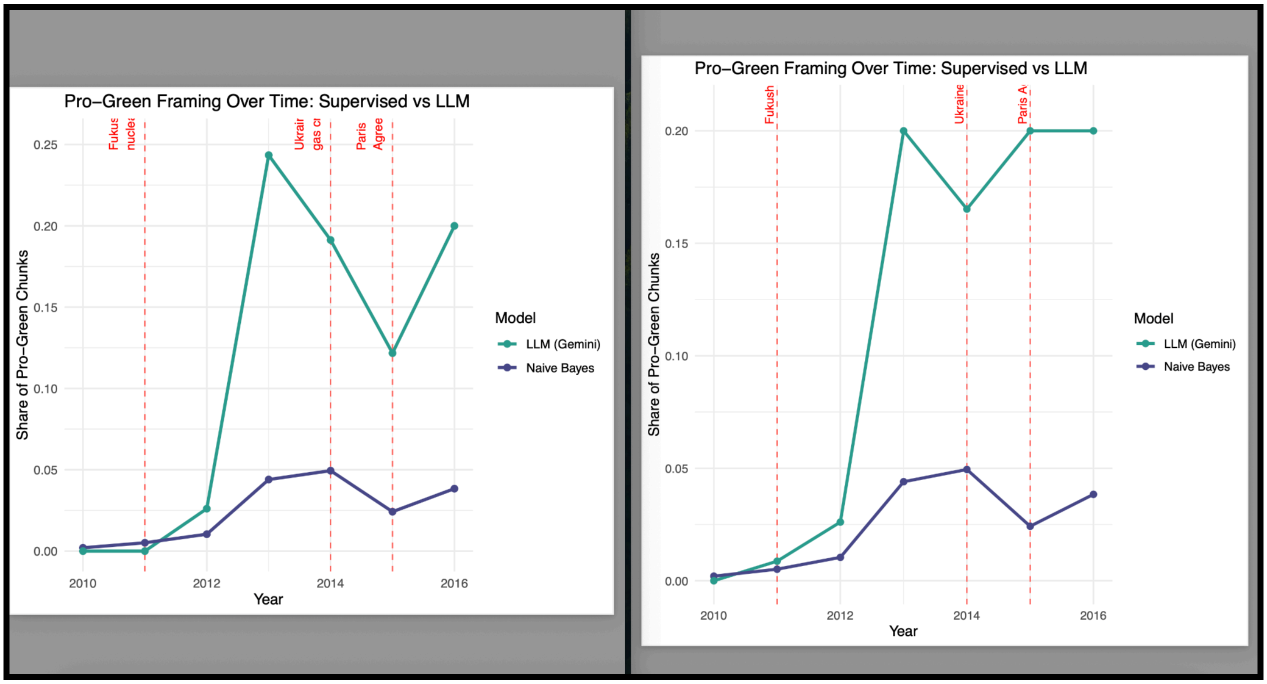


The LLM-based measure has several important limitations that must be clarified. First, Gemini is not a perfect surrogate for a human coder and may struggle with more nuanced or borderline cases. Additionally, while the LLM performs well according to F1, a small number of misclassifications could shift estimates of trend magnitudes due to the limited evaluation of 180 gold-labeled chunks and the positive class being relatively rare. If errors are systematic, for example late period speeches have more climate buzzwords, then labeling may be inaccurate. Third, the LLM was applied only to a stratified sample of 763 chunks, not the entire corpus due to API cost and rate limits. A fully LLM-coded corpus might reveal more granular differences across Council configurations or specific debates. As shown in Figure 2., when the LLM test was run again the results were similar but showing slightly less prevalent changes around events. Finally, as with the NB model, the LLM operates on fixed 200-word chunks. While this choice was implemented in the project to necessarily break up speeches, it may have mixed multiple policy themes in a single chunk. If pro-renewable language is heavily dependent on surrounding contexts, the models may have misclassified the content.

Figure 5: Comparison of Separate LLM Runs

(Left: Run 1, Right: Run 2)

(The Shocks represented are 1. 2011 Fukushima Nuclear Disaster, 2. 2014 Ukraine Gas Cut Off Crisis, 3. 2015 Paris Climate Agreement)



Measurement - Model Comparison

To assess the measurement strategy the NB supervised classifier and LLM classifier must be compared. Both models aim to code a binary measure of pro-green framing to the EU Council energy discourse. However, they differ substantially in the effectiveness. The NB model has high accuracy, moderate precision, but low recall; thus, it has a conservative tendency to predict the majority class (0: non-renewable energy). In contrast, the LLM classifier achieves high accuracy, precision, and recall indicating that the LLM captures nearly all gold-standard pro-green chunks and is overall a much more consistent classifier. The NB model provides a stricter lower-bound estimate of clear renewable energy advocacy; meanwhile, the LLM provides a broader higher-recall measure that better captures both explicit and moderately implicit pro-renewable language Therefore, the two models provide a complementary structure by strengthening the hypothesis.

Bibliography

Dupont, Claire, and Sebastian Oberthür. “Insufficient Climate Policy Integration in EU Energy Policy: A Case of Policy Competition?” Journal of European Integration, vol. 34, no. 1, 2012, pp. 1–18.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. Text as Data: A New Framework for Machine Learning and the Social Sciences. Princeton University Press, 2022.

Herranz-Surrallés, Anna. “An Emerging EU Energy Diplomacy? Discursive Shifts, Enduring Practices.” Journal of European Public Policy, vol. 23, no. 9, 2016, pp. 1386–1405.

Szulecki, Kacper. “The EU’s Energy Union: A Policy in Search of a Polity?” Energy Research & Social Science, vol. 47, 2019, pp. 1–7.