

“Whose Side Are You On?”

Estimating Ideology of Political and News Content Using Large Language Models and Few-shot Demonstration Selection

Muhammad Haroon

University of California, Davis
mharoon@ucdavis.edu

Magdalena Wojcieszak

University of California, Davis
mwojcieszak@ucdavis.edu

Anshuman Chhabra

University of South Florida
anshumanc@usf.edu

Abstract

The rapid growth of social media platforms has led to concerns about radicalization, filter bubbles, and content bias. Existing approaches to classifying ideology are limited in that they require extensive human effort, the labeling of large datasets, and are not able to adapt to evolving ideological contexts. This paper explores the potential of Large Language Models (LLMs) for classifying the political ideology of online content through in-context learning (ICL). Our extensive experiments involving demonstration selection in label-balanced fashion, conducted on three datasets comprising news articles and YouTube videos, reveal that our approach outperforms zero-shot and traditional supervised methods. Additionally, we evaluate the influence of metadata (e.g., content source and descriptions) on ideological classification and discuss its implications. Finally, we show how providing the source for political and non-political content influences the LLM’s classification.

1 Introduction

Social media platforms are one of the leading sources of information for the majority of Americans and have grown in centrality over the past few years (Barthel et al., 2020). Observers worry that the content recommended and seen on social media platforms can radicalize users, cause extremism (Tufekci, 2018; Roose, 2019), and facilitate the formation of filter bubbles that restrict exposure to diverse viewpoints (Pariser, 2011).

However, empirical evidence for these worries is inconsistent (Haroon et al., 2023; HosseiniMardi et al., 2021; Chen et al., 2023; Nyhan et al., 2023) in part because scholars use different methodologies and approaches (audits versus observational data) and in part because we lack *robust* (i.e., highly accurate across domains) and *scalable* (i.e., requiring minimal human supervision) methods for estimating political ideology of the content users

consume.¹ Systematic assessments of whether (and the extent to which) social media platforms indeed expose users to ideologically biased or politically radical content are not possible without *robust* and *scalable* political ideology estimation methods.

Prior approaches to ideology estimation have utilized manual expert annotations (Otero, 2019), supervised classification models (Baly et al., 2020), and social media sharing patterns (Le et al., 2017; Barberá, 2015; Lai et al., 2024; Wojcieszak et al., 2023; Eady et al., 2020). Manual annotations are highly accurate but resource-intensive and difficult to scale, especially given the growing volume of online content. Supervised classification models, while offering some automation, face challenges such as a reliance on extensive labeled data and difficulty adapting to dynamic ideological contexts. Social media-based methods estimate ideology by analyzing content sharing patterns on platforms such as Twitter or Reddit, but these approaches are hindered by data access limitations (Murtfeldt et al., 2024), reliance on widely shared content, and being agnostic to the content itself. These limitations highlight the need for scalable, adaptable, and robust solutions to ideology estimation.

In this paper, we advance on literature showing that Large Language Models (LLMs) can be applied to ideology estimation of text in the context of the two-party US political spectrum ranging from democrat/liberal to republican/conservative, and can address the aforementioned challenges of scalability and robustness. More specifically, pre-trained LLMs can learn new classification tasks *in-context* at test-time (Hardt and Sun, 2024), which enables a few-shot learning paradigm that only requires a few annotated samples. Furthermore, LLMs can handle complex language

¹We recognize the need to also estimate political ideology of the users themselves, which has been done using variety of methods (Barberá, 2015) and which is outside the scope of this work.

subtleties necessary for accurate/robust ideology estimation. They can also adapt to new and evolving information by simple adjustments of the provided prompt. Thus, LLMs can serve as a compelling choice for robust and scalable ideology estimation across diverse text domains and platforms.

Past applications of LLMs in ideology estimation (Heseltine et al., 2024; Törnberg, 2023; Gilardi et al., 2023) have primarily focused on zero-shot approaches, utilizing pre-trained LLMs without employing more nuanced few-shot in-context learning (ICL) strategies. Our work advances these prior studies by proposing a modified version of coverage-based ICL that selects demonstrations in a class-balanced manner, substantially improving performance over simple zero-shot prompting and coverage-based ICL. Moreover, whereas existing works primarily classify political texts with explicit ideology (e.g., tweets from senators), we demonstrate the effectiveness of our approach across diverse content categories, including political news, political non-news, and non-political news content.

In sum, we make the following contributions:

- We conduct an extensive study on the capabilities of differently sized LLMs (Llama2-13B, Mistral-7B, and GPT-4o) in estimating the ideology of text. Through rigorous experimentation on three distinct datasets previously labeled by domain experts, we demonstrate these models' accuracy, adaptability, and robustness in capturing ideological cues across multiple data. Our findings position LLMs as a robust and scalable solution for effective ideology estimation.
- We propose a demonstration selection algorithm that utilizes coverage-based ICL and class-balanced selection to achieve state-of-the-art LLM performance on ideology estimation. We show that our approach improves upon zero-shot prompting used in past work (Heseltine et al., 2024; Törnberg, 2023; Gilardi et al., 2023).
- We analyze the effect of additional metadata, such as content source and description, on LLM ideology predictions, especially for neutral or non-political content. Our results show that metadata can influence model predictions, at times aligning the predicted ideology with that of the source or introducing bias when source and content ideologies diverge. This analysis provides critical insights into the limitations and potentials of LLMs for nuanced

ideological analysis.

2 Related Work

Prior approaches to ideology estimation have generally fallen under three types (1) manual expert annotations (Otero, 2019), (2) content analysis through supervised classification models (Baly et al., 2020), and (3) social media sharing approaches (Le et al., 2017; Barberá, 2015; Lai et al., 2024; Wojcieszak et al., 2023; Eady et al., 2020).

Manual annotations by experts. Several efforts have been taken by independent parties to identify the political ideology of online news articles and sources such as Ad Fontes (Otero, 2019), All-Sides², Media Bias/Fact Check³, etc through a manual annotation process involving experts. Whereas manual annotation is accurate and robust (albeit depending on the number and quality of experts employed), it is highly resource-intensive and not scalable. Expert annotation becomes especially challenging as the volume of social media content increases consistently.

Training supervised classifiers. In turn, supervised ideology classification models can provide some form of automation. The advent of deep learning techniques revolutionized this field by enabling more sophisticated models capable of capturing complex linguistic patterns. Iyyer et al. (2014) introduced deep neural networks, such as LSTMs and CNNs, which outperformed traditional models in detecting ideology from unstructured text data. In a similar vein, Baly et al. (2020) used a BERT model trained on political news articles to show improvements over existing state-of-the-art supervised learning approaches. However, they are limited by a need for large volumes of human-labeled data and have difficulty adapting to dynamic and ever evolving ideological contexts (e.g., cross-platform estimation, emerging news events, changes in the political environment).

Utilizing social media data. Social media sharing ideology estimation approaches utilize data from external platforms such as Twitter (Le et al., 2017; Barberá, 2015; Wojcieszak et al., 2023; Eady et al., 2020) or Reddit (Lai et al., 2024) to assess how a given piece of content is shared among users. By identifying left- and right-leaning elites (e.g., politicians, news media) those users follow, those approaches first estimate the ideology of the users,

²<https://allsides.com>

³<https://mediabiasfactcheck.com>

and then estimate the ideology of the shared content by aggregating the ideology of the sharing users. This approach is limited in several ways. First, as it relies on third-party platforms (such as Twitter/X), changes to APIs and data acquisition systems can render this approach untenable. For instance, since Twitter transitioned to X in 2023, it has become challenging for researchers to collect sharing data at scale without incurring significant costs (Murtfeldt et al., 2024; Poudel and Weninger, 2024). Second, to accurately estimate the ideology of content (e.g., a YouTube video), the content needs to be shared extensively, which might not happen for niche videos with small viewership. Finally, this approach is agnostic to the nature of the content itself and only looks at how it is shared by users. Owing to these limitations, these approaches may not be robust and reliable.

Using Large Language Models. LLMs have achieved state-of-the-art performance for several tasks across multiple language domains (e.g., legal transcripts (Colombo et al., 2024) and medical data (Yang et al., 2022)) and contexts (e.g., identifying misinformation (Hoes et al., 2023), hate speech (Huang et al., 2023), sentiment (Heseltine et al., 2024), or relevance, topics, and frames in tweets and news articles (Gilardi et al., 2023)).

As aforementioned, prior works have utilized LLMs to classify the ideology of text (Heseltine et al., 2024; Törnberg, 2023; Gilardi et al., 2023). As we show in Table 1, our work differs from these in several key ways. First, all these approaches are zero-shot methods that only utilize the pre-trained LLMs as is, and do not employ optimal and efficient few-shot in-context learning (ICL) approaches to improve accuracy and adapt LLMs to multiple domains. We propose a modified version of coverage-based ICL that selects demonstrations in a class-balanced manner, and improves performance over simple zero-shot prompting and coverage-based ICL. Second, these methods focus on classifying texts from sources with clearly distinct ideology (e.g. tweets from liberal or conservative politicians). We focus on content with differential levels of ideological signal to demonstrate that our LLM-based approach can estimate ideology of *political news*, *political content but not news*, and *non-political news content* as well. We also analyze the potential limitations for ideology estimation of each of these content subcategories. Finally, prior works do not undertake extensive benchmarking of performance across multiple LLMs and diverse

| Method | Zero-shot | Few-shot (ICL) | Non-political content |
|-------------------------|-----------|----------------|-----------------------|
| Gilardi et al. (2023) | ✓ | ✗ | ✗ |
| Törnberg (2023) | ✓ | ✗ | ✗ |
| Heseltine et al. (2024) | ✓ | ✗ | ✗ |
| Ours | ✓ | ✓ | ✓ |

Table 1: Comparison of our work with prior LLM-based approaches.

datasets. We run experiments on three different datasets (two of news articles and one consisting of YouTube videos) and three different LLMs (open-source and closed-source) of different sizes. Our results shed light on which LLMs are better suited for ideology classification and what data/information augments LLM performance, i.e. whether additional video descriptions or channel information in prompts increases/reduces performance.

3 Problem Formulation

Our research problem can be defined as follows: *Given content that has an underlying ideological leaning (e.g., a YouTube video, a news article), we aim to determine if the content is liberal, neutral, or conservative.* Next, we describe our proposed approach where we utilize set coverage and balanced shot selection for ideology classification.

3.1 Proposed Approach

Algorithm 1 provides our proposed in-context learning (ICL) framework. Given a LLM \mathcal{L} , a training set of content with ideology class labels (Liberal, Neutral, Conservative) \mathcal{T} , and number of demonstrations k , our task is to classify a given query x (an online post, video, or news article) as either liberal, neutral, or conservative. Our approach first builds a smaller candidate set \mathcal{C} of potential demonstrations and computes a relevance score for each demonstration in the candidate set. From the sorted candidate set, we select k demonstrations equally from across all three ideology classes, i.e., only $\frac{k}{3}$ demonstrations can be selected from each ideology class. Thus, we skip over demonstrations with higher scores if $\frac{k}{3}$ demonstrations have already been selected from the class they belong to. The instruction prompts used can be found in Appendix A.

3.2 Test Query Specifications

We specify the query x to be classified as an online post, video, or news article that has a title, source, and a description. An LLM can leverage

Algorithm 1 Estimating Ideology using Set Cover and Label-Balanced Shot Selection

```
1: Input:
   • Test Instance:  $x$ 
   • LLM:  $\mathcal{L}$ 
   • Prompt:  $\mathcal{P}$ 
   • Training Set:  $\mathcal{T}$ 
   • Number of Demonstrations:  $k$ 
2: Output: Ideology classification for  $x$ 
3: Initialize:
   •  $\mathcal{D} \leftarrow \emptyset$   $\triangleright$  Selected Demonstrations
   •  $\mathcal{C} \leftarrow \text{SET}(\mathcal{T})$   $\triangleright$  Candidate set using coverage
   •  $\text{counts}(s) \leftarrow 0 \forall s \in \text{ideologies}$   $\triangleright$  Label counts
4: Sort  $\mathcal{C}$  (descending) by SET-BSR over  $x$ 
5:  $i \leftarrow 0$ 
6: while  $\sum \text{counts} \leq n$  do
7:    $s \leftarrow \text{ideology}(\mathcal{C}_i)$ 
8:
9:   if  $\text{counts}(s) \leq \frac{n}{3}$  then  $\triangleright$  Ensure label balance
10:     $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{C}_i$   $\triangleright$  Add demonstration to  $\mathcal{D}$ 
11:     $\text{counts}(s) \leftarrow \text{counts}(s) + 1$ 
12:   end if
13:    $i \leftarrow i + 1$   $\triangleright$  Move to next element in  $\mathcal{C}$ 
14: end while
15: return  $\mathcal{L}(\mathcal{P} \cup \mathcal{D}, x)$   $\triangleright$  Classify  $x$  with prompt  $\mathcal{P}$  and
   selected demonstrations  $\mathcal{D}$ 
```

information from all three components to aid in classification, however, limitations such as context-window sizes and longer compute times for longer text make it worthwhile to explore the minimum amount of information that provides the best accuracy. As such, our experiments rely on four combinations: title; title and source; title and description; and title, source, and description.

3.3 Demonstration Selection

We select k demonstrations for in-context learning (ICL) to augment the LLM’s performance. We utilize SET-COVER (Vazirani, 2003) to maximize coverage across the candidate pool, where BertScore-Recall (BSR) is chosen as the coverage metric. This strategy is known as SET-BSR, and has been shown to be successful for ICL with multiple classes (Gupta et al., 2023). Additionally, we ensure that an equal number of demonstrations is selected from each ideology class. A comparison of random and balanced shot selection can be found in Appendix G to validate this design choice.

4 Results and Evaluation

4.1 Experimental Setup

4.1.1 Datasets

For our evaluation, we rely on three different datasets, which have been used in prior work on

Sample Text

Trump tweets COVID-19 video so outrageous it was banned from Facebook and Twitter

Selected Demonstrations

Twitter removes Trump retweet that falsely claimed COVID-19 ‘has a cure’

Reporter to sue Michigan Gov. Whitmer for allegedly not releasing COVID-19 nursing home data

AU Health masters COVID-19 therapy but loses \$20 million

Facebook to label COVID-19 vaccine posts to combat misinformation, help people find where to get the shot

Trump Gives Delusional COVID-19 Press Conference Like A Broken Man

Florida Scientist Says She Was Fired For Not Manipulating COVID-19 Data

Wisconsin pharmacist arrested for sabotaging COVID-19 vaccines

Facebook openly conspiring with FBI to punish Trump supporters for wrongthink

Table 2: Example of demonstration selection using Algorithm 1 for a given text. Colors show how demonstrations cover salient aspects of the samples.

ideology classification.⁴

YouTube Slant Dataset. We sample 645,187 videos from the YouTube dataset by Haroon et al. (2023) which estimates the ideology of videos using the Twitter-based slant estimation approach (Le et al., 2017; Barberá, 2015; Wojcieszak et al., 2023) on a scale of -1 (most liberal) to +1 (most conservative) and was collected between 2021 and 2022. We use a BERT-based political classifier from Askari et al. (2024) to further identify 49,894 political videos to create our YouTube slant dataset. Videos with slant ≤ -0.33 as were labelled as liberal, $\geq +0.33$ as conservative, and everything else as neutral. The training set used for selection consists of 48,895 videos from which 20,000 candidates were selected and ranked using the SET-BSR approach. The test set consists of 999 videos sampled as to maintain an equal balance of ideologies. Overall, we have videos from over 8,523 channels across the entire training and testing sets (See Appendix I for where the slant cut-offs lie respective to popular news channels.).

Ad Fontes News Articles. This dataset consists of news articles that have been labeled by expert human annotators on Ad Fontes Media Bias Panel

⁴While we use a conservatively large candidate set size, we show in Appendix F that even a smaller candidate set size yields similar results.

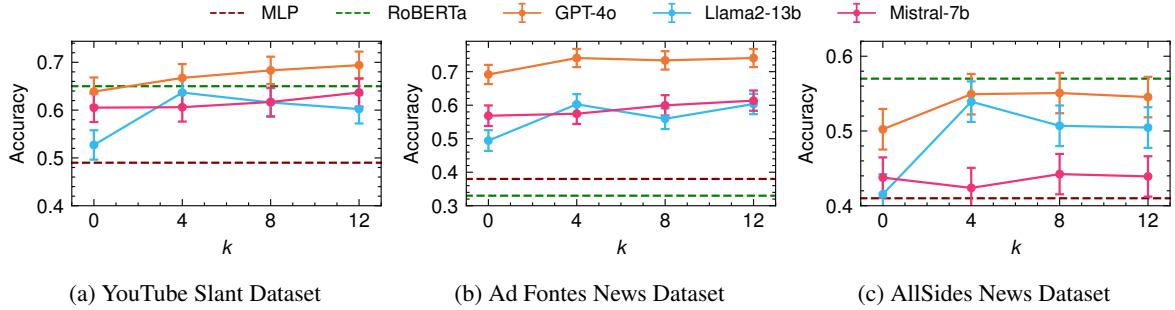


Figure 1: Improvement in accuracies of the GPT-4o, Llama2, and Mistral models across the three datasets by increasing the number of ICL demonstrations k . The baseline MLP and RoBERTa models are also shown. We see that a higher number of demonstrations leads to better predictions, and that the models generally outperform the baselines (even though the RoBERTa and MLP models have access to the full training set, and the LLMs only see k demonstrations in-context). The error bars show a 95% confidence interval.

(Otero, 2019) from “most extreme left” to “most extreme right.” Each article is rated by at least three experts from across the political spectrum on a scale of -42 to +42 corresponding to extremely left and extremely right news articles, respectively. Additionally, Ad Fontes labels consider articles with scores lower than -14.0 to skew left and articles with scores greater than +14.0 as skewing right. Adopting these cut-offs, we label articles with scores lower than -14.0 as Liberal, scores greater than +14.0 as Conservative, and everything in-between as Neutral. The training set consists of 33,396 news articles from which we build a candidate set of 8,000 using the SET-BSR approach. The test set contains 999 news articles with a balanced distribution of ideology.

AllSides News Articles. The dataset contains expert human annotations from AllSides⁵, which relies on several human annotation approaches to rank news articles such as blind surveys, expert reviews (on a scale of -6 to +6), and independent reviews. We use the training and test set of news articles used by Baly et al. (2019) from this dataset. The training set consists of 26,590 news articles, from which we build a candidate set of 8,000 using the SET-BSR approach. The test set consists of 1,300 videos.

4.1.2 LLMs

We use three different publicly available LLMs for the experiments.

OpenAI GPT-4o. We use the pre-trained GPT-4o model by OpenAI (Hurst et al., 2024) using a paid API subscription.

Meta Llama-2. We use the Hugging Face library

to run Meta’s Llama-2 model with 13B (Touvron et al., 2023) parameters on a local machine with a RTX A6000 GPU, 256GB memory, and 48 cores.

MistralAI Mistral. We use the Hugging Face library to run MistralAI’s Mistral model with 7B parameters (Jiang et al., 2023) on the same local machine with a RTX A6000 GPU, 256GB memory, and 48 cores.

4.1.3 Baselines

We use two standard classification approaches as our supervised learning baselines.

Multilayer Perceptron (MLP). This model is a fully connected neural network that takes a 384-dimension text embedding as input to an initial dense hidden layer of 512 neurons with the tanh activation function, and a three output softmax final layer (for the three ideologies).

RoBERTa. As in Baly et al. (2020), we use the base RoBERTa transformer model and fine-tune on each of the three full training datasets for 50 epochs. Unlike the LLM which only receives a small number of k demonstrations, the RoBERTa model sees the full dataset(s).

4.2 Results

Baselines. The baseline MLP model performed the worst, achieving an accuracy of 0.49, 0.38, and 0.41 across the YouTube, Ad Fontes, and AllSides datasets. The RoBERTa baseline achieved a score of 0.65 on the YouTube Dataset and 0.57 on the AllSides News dataset, marked improvements from the MLP model, but fared worse on the Ad Fontes dataset, with accuracy of 0.33. This clearly shows that simpler models are not well-suited to the task of ideological estimation, especially when compared to more advanced transformer-based architectures.

⁵<http://www.allsides.com/media-bias/media-bias-rating-methods>

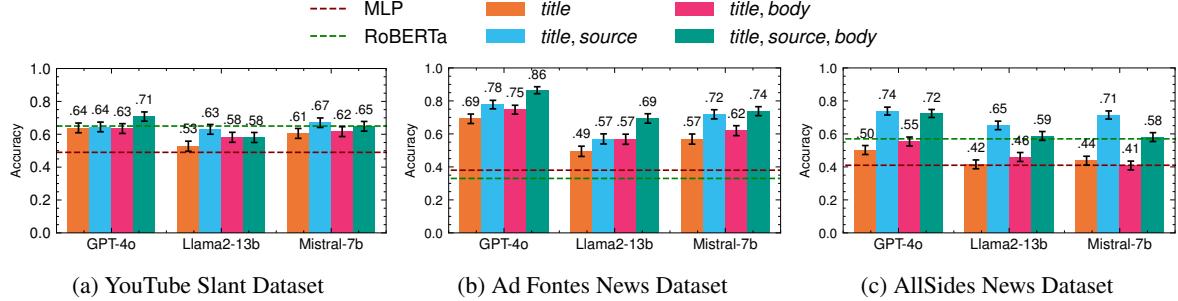


Figure 2: Improvement in the accuracies of the GPT-4o, Llama2, and Mistral models by testing different combinations of the title, source, and description from the three datasets in the zero-shot setting. We see that providing source and description immensely improves the accuracy of the models across the Ad Fontes and AllSides news datasets where the news sources are more likely to be already known by the model. The YouTube dataset, on the other hand, sees a degradation in performance albeit a slight improvement in the case of Mistral. The error bars show a 95% confidence interval.

Zero-Shot Learning. Figure 1 shows the performance of the three LLMs and baselines across the three datasets. Focusing on the zero-shot setting, we notice that GPT-4o outperformed both Mistral and Llama2, achieving an accuracy of 0.64, 0.69, and 0.50 across the YouTube, Ad Fontes, and AllSides datasets. Mistral performed better than Llama2, achieving a 0.59 accuracy on the YouTube dataset closer to GPT-4o whereas Llama2’s accuracy was very low at 0.44. On the Ad Fontes dataset, GPT-4o performed 14% better than Mistral at 0.69 accuracy compared to Mistral’s 0.55. Finally, all LLMs performed poorly on the AllSides dataset with GPT-4o performing the best at 0.50, 7% lower than the baseline RoBERTa model. Comparing with the baselines, RoBERTa performed slightly better than GPT-4o at 0.65 and 0.57 on the YouTube Slant and AllSides datasets, respectively. In contrast, the MLP performed the worst in the Ad Fontes and AllSides datasets but fared better than Llama2 in the YouTube slant dataset.

Few-Shot ICL Comparison. Providing the set of demonstrations \mathcal{D} selected using the SET-BSR approach to the prompt \mathcal{P} , Figure 1 also shows how the accuracy of the models improved by increasing the number of demonstrations k . The trend of GPT-4o outperforming Llama2 and Mistral persisted as the number of demonstrations increased. GPT-4o consistently achieved the highest scores across all datasets, suggesting that it is particularly well-suited for ideological estimation tasks.

Mistral-7B and GPT-4o showed a steady increase in performance as the number of demonstrations increased, both seeing a 5% increase in accuracy going from $k = 0 \rightarrow 12$ (0.64 to 0.69 for GPT-4o, and 0.59 to 0.64 for Mistral-7B).

Llama2-13B saw marked increases in accuracy across all datasets from $k = 0 \rightarrow 4$, but grew worse as $k = 4 \rightarrow 12$.

Providing Additional Metadata. Augmenting the test instance x to include source and description increased the performance of the models across all the datasets in the zero-shot setting. Figure 2 shows how model performance was affected by altering x to include the title, the title and source, the title and description, and the title, source, and description. For the YouTube dataset, performance for GPT-4o increased by 7% when the source and description was provided along with the title. Llama2 saw an increase of 5% when the description was provided and Mistral saw an increase of 8% when only the source was provided. When augmenting the title with the source and description, we see higher increases for Ad Fontes, with the performance of GPT-4o increasing by 17%, Llama2 by 21%, and Mistral by 19%. Similar increases are seen for AllSides, where adding just the source pushed the accuracy of the models further. In short, when provided the source information, all three LLMs were able to surpass the baseline RoBERTa model in the zero-shot setting.

Classifying non-political content. Our findings so far have focused on exclusively political content. The improvement in the accuracy of the LLMs when the source is provided suggests that the source carries substantial weight in determining the ideology of content, where the title alone was previously misclassified. This improvement is more evident in the Ad Fontes and AllSides news datasets, where the sources are well-known news organizations, and less so in the YouTube slant dataset, where the sources can be any YouTube

channel. To pinpoint the relative role of the title and source in determining the classification, we resample the entire YouTube slant dataset to procure three datasets: 1) political content from well-known news channels (political news), 2) non-political content from well-known news channels (non-political news), and 3) political news from non-news channels, such as influencers or pundits (political non-news). We use the same training set for the YouTube slant dataset and estimate ideology using GPT-4o.

| | Political news | Political non-news | Non-political news |
|--------------|----------------|--------------------|--------------------|
| Liberal | 314 | 309 | 125 |
| Neutral | 435 | 436 | 207 |
| Conservative | 251 | 255 | 76 |
| Total | 1,000 | 1,000 | 408 |

Table 3: Number of samples in the political news, political non-news, and non-political news test sets taken from the YouTube slant dataset. There are fewer examples in the non-political news datasets because that was the maximum number of samples that met this criteria. The training and candidate sets were the same as for the larger YouTube slant dataset.

Breaking this down by the class labels, Figure 3 shows the change in prediction accuracy after adding source information using a heatmap of changes across the three datasets. The heatmap shows which corresponding class was a given label misclassified as before providing the source allowing us to analyze the influence of the source when estimating ideology. Coupling this with qualitative examples from Appendix H provides insight into the influence of the source in classifying ideology.

For political news, we see an improvement in accuracy of content previously misclassified as neutral to liberal (+16%) and conservative (+7%), becoming almost 70% more accurate in both cases. When looking at the kinds of titles re-classified in Table 16 in Appendix H, we see that the titles themselves do not indicate the any ideological leaning. For example, “*Haberman says Trump has been diminished but isn’t dead politically*” was classified as neutral but when the LLM was informed it was from CNN, it was re-classified as liberal. Similarly, the video titled “*Transgender Clinical Psychologist on people de-transitioning | Dr. Erica Anderson*” was re-classified as conservative when the LLM learned it came from Newsmax, a conservative outlet, despite the title itself indicating

no such political leaning. Cases where the label changed from liberal to conservative or vice-versa showed more nuance. The video titled “*Democrats introduce bill to block Trump from holding office*” was classified as liberal but then re-classified as conservative when the source Newsmax was provided which was consistent with the content of the video. This is true for the non-political news dataset as well, where the titles do not depict any political inclination as can be seen in Table 17 in Appendix H but the model re-classified it according to the ideology of the news source. Finally, for the political non-news videos, we do not see much movement as the LLM is likely unfamiliar with the sources provided. The cases where we do see movement as seen in Table 18, it is also driven by the source of the video as well. We see some movement towards the neutral as well. For example, the videos titled “*The Trouble with the Electoral College*” and “*Is capitalism actually broken?*” are liberal in ideology but when the model learns they are from educative channels CGP Grey and TED-Ed, it re-classifies them as neutral.

4.3 Discussion

Few-shot ICL increases performance. With higher number of shots, we see that accuracy improves across the board for all three LLMs across all three datasets. Except for Llama2-13B which performs best at 4-shots (likely due to its shorter context window length), both GPT-4o and Mistral-7B see continued improvement in accuracy at an average of 5% and 3% across all three datasets.

Supplementing prompt with additional metadata improves performance. The additional metadata consist of the source and description of the content. Just adding the content source increased the accuracy of all the models, suggesting that the models are familiar with the ideological positions of the sources used and leverage that information to improve their ideology estimation. In contrast, the description actually lowers the accuracy when added, likely because it affects the context-window of the input poorly and may not provide more context than the title itself.

Leveraging few-shot ICL and additional metadata for best performance. Combining the few-shot ICL $k = 12$ and incorporating the source resulted in the highest gains in GPT-4o’s performance across all three datasets. The model achieved a 10% increase in the YouTube slant dataset ($0.64 \rightarrow 0.74$), 18% increase in the Ad

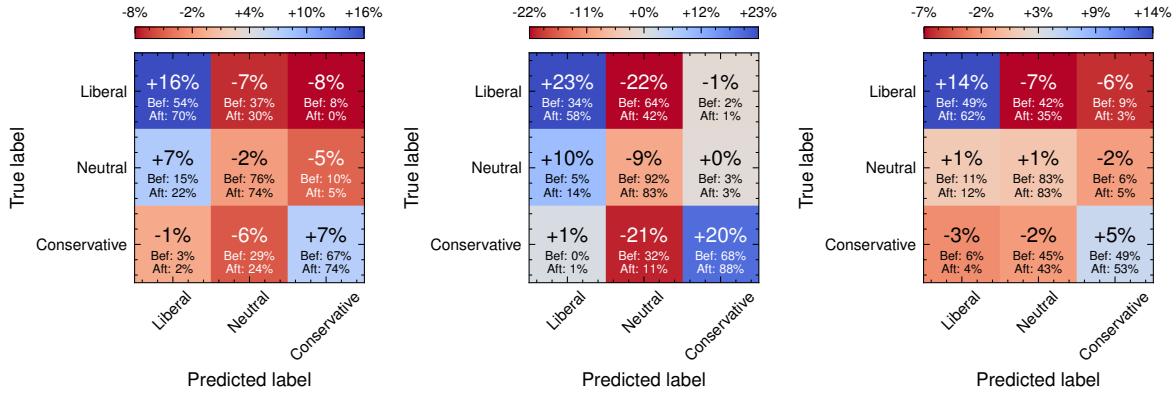


Figure 3: Heatmap showing changes in the performance of GPT-4o when the channel name is provided for a) political news, b) non-political news, and c) political non-news YouTube videos. Positive values along the main diagonal represent an increase in accuracy. For example, +16% in the top-left cell means providing the channel name increased the accuracy of classifying the Liberal class from 54% to 70%. We see more accurate prediction (increase in top-left and bottom-right) of news content (3a, 3b) when the channel name is provided showing how the LLM is already familiar with the partisan leanings of online news channels.

Fontes dataset ($0.69 \rightarrow 0.87$), and 37% increase in the AllSides dataset ($0.50 \rightarrow 0.87$). These improvements demonstrate how GPT-4o leverages details from relevant demonstrations and its existing knowledge of news sources to more accurately reason about the ideology of content.

Multi-modal classification using YouTube thumbnails does not improve GPT-4o’s performance. We experiment with adding video thumbnails as part of the prompt to GPT-4o. Despite the potential of multi-modal classification methods to enhance model performance by incorporating visual information, our experiments reveal that using YouTube thumbnails as an additional modality does not improve GPT-4o’s performance. When comparing the results of GPT-4o with and without the inclusion of YouTube thumbnails, the accuracy of the model dropped to 0.35 from 0.64 on the YouTube slant dataset. This result indicates that, for the problem of ideology estimation from YouTube content, visual information does not contribute additional meaningful signals beyond what is captured in the text and instead can be misleading for the model. However, more extensive analysis and experiments are needed for incorporating multi-modal information for ideology classification and can be undertaken in future work.

Chain-of-thought reasoning is ineffective. We apply chain-of-thought (CoT) reasoning (i.e., prompting the model to think through the task step-by-step) to GPT-4o for our task and see that it doesn’t improve accuracy despite its known benefits (Kojima et al., 2024). On the YouTube slant dataset, the

model accuracy dropped from 0.64 to 0.62 when CoT reasoning was used. One explanation for this finding is that ideology estimation is primarily reliant on identifying patterns in content that are explicit rather than requiring deep logical steps. The task might not demand the same kind of multi-step reasoning that CoT typically helps with in applications, such as arithmetic or procedural problem-solving.

Accuracy of LLMs is comparable with human annotation. The high accuracy of GPT-4o is comparable to human annotation on the AllSides and Ad Fontes News datasets, both of which are human-annotated. On these, GPT-4o achieved high accuracy, closely aligning with human judgments, exceeding 0.80 in some configurations. For additional validation, a domain expert annotated a sample of 561 videos in the YouTube dataset, finding that GPT-4o agreed with the human annotation 79% of the time. This suggests that GPT-4o is able to effectively mirror human-level ideological classification, even in more informal content. The model’s performance highlights its capacity to generalize patterns and signals used by human annotators, making it a valuable tool for automated ideological estimation.

5 Conclusion

This study explored the efficacy of few-shot learning using Large Language Models for the task of ideology estimation across diverse textual datasets, including news articles and YouTube videos. Our findings demonstrate that LLMs, particularly GPT-4o, achieve state-of-the-art performance, surpassing traditional supervised baselines. By leveraging

advanced few-shot in-context learning methods, we addressed challenges of scalability and robustness, showcasing the ability of LLMs to adapt to evolving contexts with minimal supervision.

6 Limitations

Despite the promising results and contributions outlined in this paper, several limitations must be acknowledged to identify areas for future work. The performance of our approach is heavily reliant on the capabilities of pre-trained LLMs such as GPT-4o, Llama2-13B, and Mistral-7B. Variability in model updates or differences in underlying pre-training data could affect reproducibility and consistency across different deployments. In addition, while incorporating metadata, such as source information improved performance, it also introduced the potential for biased predictions. Models may disproportionately rely on known ideological leanings of sources, potentially misclassifying nuanced or neutral content. This issue might be further exacerbated by the pre-processing guardrails that serve to filter data, but inadvertently possess their own robustness and bias concerns (Achara and Chhabra, 2025; Yang et al., 2024).

Furthermore, our experiments focus on text, and extending these methods to classify ideology in non-textual data such as video or audio remains an open challenge and an important direction for future work. Lastly, we acknowledge that the datasets used for evaluation, while diverse, may not fully capture the breadth of content and ideological nuance present in real-world social media platforms. For instance, the balance of liberal, neutral, and conservative content in the datasets may not reflect natural distributions, and also the majority of content consumed by users is generated by other users themselves, not by news media organizations or YouTube channels, suggesting the need to adapt our approach to user-generated posts and comments. Finally, while beyond the scope of our work, LLMs can themselves have several robustness issues (Askari et al., 2025; Chhabra et al., 2024) that can result in potential downstream evaluation issues.

7 Ethics Statement

Automated ideology estimation carries inherent risks, including misuse for surveillance or censorship. Furthermore, reliance on LLMs can perpetuate biases present in training data, which might

influence predictions in unintended ways. While the study evaluates multiple datasets, the generalizability of the approach across all social media platforms, with varying content types and user behaviors, is yet to be fully established.

Addressing these limitations could involve refining LLM prompting techniques, improving model transparency, exploring alternative datasets, and incorporating mechanisms for bias mitigation and real-time adaptability. Future research should also emphasize ethical frameworks for deploying such systems responsibly, ensuring that advancements in technology contribute positively to societal discourse without exacerbating existing challenges.

References

Akshit Achara and Anshuman Chhabra. 2025. Watching the AI watchdogs: A fairness and robustness analysis of AI safety moderation classifiers. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 253–264.

Hadi Askari, Anshuman Chhabra, Muhamo Chen, and Prasant Mohapatra. 2025. Assessing llms for zero-shot abstractive summarization through the lens of relevance paraphrasing. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2187–2201.

Hadi Askari, Anshuman Chhabra, Bernhard Clemm von Hohenberg, Michael Heseltine, and Magdalena Wojcieszak. 2024. Incentivizing news consumption on social media platforms using large language models and realistic bot accounts. *PNAS nexus*, 3(9):pgae368.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991, Online. Association for Computational Linguistics.

Ramy Baly, Georgi Karadzhov, Abdelrhman Saleh, James Glass, and Preslav Nakov. 2019. Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2109–2116, Minneapolis, Minnesota. Association for Computational Linguistics.

Pablo Barberá. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis*, 23(1):76–91.

Michael Barthel, Amy Mitchell, Dorene Asare-Marfo, Courtney Kennedy, and Kirsten Worden. 2020. Measuring news consumption in a digital era. <https://pewrsr.ch/2VUtTFe>.

Annie Y Chen, Brendan Nyhan, Jason Reifler, Ronald E Robertson, and Christo Wilson. 2023. Subscriptions and external links help drive resentful users to alternative and extremist youtube channels. *Science Advances*, 9(35):eadd8080.

Anshuman Chhabra, Hadi Askari, and Prasant Mohapatra. 2024. Revisiting zero-shot abstractive summarization in the era of large language models from the perspective of position bias. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 1–11.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre FT Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, et al. 2024. Saullm-7b: A pioneering large language model for law. *arXiv preprint arXiv:2403.03883*.

Gregory Eady, Richard Bonneau, Joshua A Tucker, and Jonathan Nagler. 2020. News sharing on social media: Mapping the ideology of news media, politicians, and the mass public.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Shivanshu Gupta, Matt Gardner, and Sameer Singh. 2023. Coverage-based example selection for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13924–13950, Singapore. Association for Computational Linguistics.

Moritz Hardt and Yu Sun. 2024. Test-time training on nearest neighbors for large language models. In *The Twelfth International Conference on Learning Representations*.

Muhammad Haroon, Magdalena Wojcieszak, Anshuman Chhabra, Xin Liu, Prasant Mohapatra, and Zubair Shafiq. 2023. Auditing youtube’s recommendation system for ideologically congenial, extreme, and problematic recommendations. *Proceedings of the National Academy of Sciences*, 120(50):e2213020120.

Michael Heseltine et al. 2024. Large language models as a substitute for human experts in annotating political text. *Research & Politics*, 11(1):20531680241236239.

Emma Hoes, Sacha Altay, and Juan Bermeo. 2023. Using chatgpt to fight misinformation: Chatgpt nails 72% of 12,000 verified claims. *PsyArXiv*. 3.

Homa Hosseinmardi, Amir Ghasemian, Aaron Clauset, Markus Mobius, David M Rothschild, and Duncan J Watts. 2021. Examining the consumption of radical content on YouTube. *Proceedings of the National Academy of Sciences*, 118(32).

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion proceedings of the ACM web conference 2023*, pages 294–297.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1113–1122, Baltimore, Maryland. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2024. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS ’22*, Red Hook, NY, USA. Curran Associates Inc.

Angela Lai, Megan A. Brown, James Bisbee, Joshua A. Tucker, Jonathan Nagler, and Richard Bonneau. 2024. Estimating the ideology of political youtube videos. *Political Analysis*, page 1–16.

Huyen Le, Zubair Shafiq, and Padmini Srinivasan. 2017. Scalable News Slant Measurement Using Twitter. In *AAAI Conference on Web and Social Media*, volume 11.

Ryan Murtfeldt, Naomi Alterman, Ihsan Kahveci, and Jevin D. West. 2024. Rip twitter api: A eulogy to its vast research contributions. *Preprint*, arXiv:2404.07340.

Brendan Nyhan, Jaime Settle, Emily Thorson, Magdalena Wojcieszak, Pablo Barberá, Annie Y Chen, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, et al. 2023. Like-minded sources on facebook are prevalent but not polarizing. *Nature*, 620(7972):137–144.

Vanessa Otero. 2019. Ad fontes media’s first multi-analyst content analysis ratings project.

Eli Pariser. 2011. *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Press.

Amrit Poudel and Tim Weninger. 2024. [Navigating the post-api dilemma](#). In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 2476–2484, New York, NY, USA. Association for Computing Machinery.

Kevin Roose. 2019. The Making of a YouTube Radical. The New York Times. <https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html>.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Zeynep Tufekci. 2018. YouTube, the Great Radicalizer. The New York Times. <https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html>.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *Preprint*, arXiv:2304.06588.

Vijay V. Vazirani. 2003. *Set Cover*, pages 15–26. Springer Berlin Heidelberg, Berlin, Heidelberg.

Magdalena Wojcieszak, Rong-Ching (Anna) Chang, and Ericka Menchen-Trevino. 2023. Political content and news are polarized but other content is not in youtube watch histories. *Journal of Quantitative Description: Digital Media*, 3.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. 2022. A large language model for electronic health records. *NPJ digital medicine*, 5(1):194.

Yahan Yang, Soham Dan, Dan Roth, and Insup Lee. 2024. Benchmarking llm guardrails in handling multilingual toxicity. *arXiv preprint arXiv:2410.22153*.

Appendix

A Prompts used for the LLM

Here, we present the prompts that are used for querying the models.

Prompt for querying with just the title.

Classify the following news article titles as ideologically liberal, neutral, or conservative. Titles with no ideological content are classified as neutral. Only respond with the final answer.

Prompt for querying with the title and source.

*Classify the following news article titles as ideologically liberal, neutral, or conservative. Titles with no ideological content are classified as neutral. **The news source is also specified for additional context.** Only respond with the final answer.*

Prompt for querying with the title, description, and source.

*Classify the following news article titles as ideologically liberal, neutral, or conservative. Titles with no ideological content are classified as neutral. **The news description is also specified for additional context.** Only respond with the final answer.*

B Statistical significances

We use the McNemar test to determine the statistical differences between the accuracies of two specific classifiers. Tables 4, 5, and 6 show the significance values for the models trained on the YouTube Slant, Ad Fontes news, and AllSides news datasets respectively.

C Sources can be misleading

Our prior results showed that providing the source information substantially increases the accuracy. However, not all content from a given source may share the ideology of the source itself (Haroon et al., 2023), as some liberal articles can come from conservative outlets. To highlight this particular scenario, we identify cases in our test set where the ideology of the article or video is different from the ideology of its source for all three datasets (e.g., a liberal article from Fox News or a conservative article from MSNBC). We rely on AllSides’ me-

dia bias chart⁶ for source ideology and map these sources to their corresponding news articles and YouTube videos. We then compare the accuracy of GPT-4o when 1) only title is provided, 2) only source is provided, and 3) both title and source are provided.

Table 7 lists the names of the left and right sources identified and mapped to the three datasets. In the YouTube test set, we identified 40 cases where an neutral/conservative article from a liberal source and 21 cases where a liberal/neutral article came from a conservative source. The Ad Fontes news test set had 38 cases of neutral/conservative articles from liberal sources and 11 cases of liberal/neutral articles from conservative sources. Finally, the AllSides news test set had 98 neutral/conservative articles from liberal sources and 0 liberal/neutral articles from conservative sources.

Table 8 reports the accuracy of the LLMs on these specific test cases for liberal and conservative sources respectively. We see that the LLMs suffer from incorrect classifications when only the source is considered as the ground truth contradicts the ideology of the source. Here, providing the title increases the accuracy of the model especially in the conservative sources where the accuracy on just the source is 0% for the YouTube and Ad Fontes datasets. In fact, providing the title for the YouTube test set in this case still does not improve the accuracy seeing as the model accuracy is still 0%. This suggests that the LLM overtly relies on the source in this case completely disregarding title wherein accuracy on just the title was close to 45%. Similar trends can be seen for the Ad Fontes news test set where the accuracy of source alone was 38% and title was 58%, and both was 70%. We observe that the AllSides news test set still has better performance when the source is provided at a 100%. On closer inspection, we observe that in the test set sampled, there is a high correlation between the source and article ideology which likely leads to such a high performance when source is provided.

D Few Shot Results for all Datasets

Figure 4 shows the increases in accuracy across all datasets by increasing the number of shots.

⁶<https://www.allsides.com/media-bias/media-bias-chart>

| GPT | | | | Llama | | | | Mistral | | | | |
|---------|-------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| | T | T/S | T/B | T/S/B | T | T/S | T/B | T/S/B | T | T/S | T/B | T/S/B |
| GPT | T | — | 0.13 | 0.06 | 33.06** | 34.18** | 0.21 | 11.05** | 10.75** | 4.50* | 4.21* | 2.34 |
| | T/S | 0.13 | — | 0.52 | 27.95** | 35.79** | 0.69 | 12.10** | 12.30** | 4.72* | 2.79 | 3.31 |
| | T/B | 0.06 | 0.52 | — | 42.29** | 30.30** | 0.06 | 10.25** | 10.16** | 2.96 | 5.33* | 1.70 |
| | T/S/B | 33.06** | 27.95** | 42.29** | — | 89.28** | 22.45** | 59.74** | 58.72** | 41.44** | 7.87** | 45.75** |
| Llama | T | 34.18** | 35.79** | 30.30** | 89.28** | — | 37.09** | 9.03** | 8.62** | 19.19** | 58.08** | 23.16** |
| | T/S | 0.21 | 0.69 | 0.06 | 22.45** | 37.09** | — | 7.81** | 8.57** | 1.75 | 5.46* | 0.56 |
| | T/B | 11.05** | 12.10** | 10.25** | 59.74** | 9.03** | 7.81** | — | 0.00 | 1.71 | 26.06** | 4.14* |
| | T/S/B | 10.75** | 12.30** | 10.16** | 58.72** | 8.62** | 8.57** | 0.00 | — | 1.83 | 26.38** | 4.38* |
| Mistral | T | 4.50* | 4.72* | 2.96 | 41.44** | 19.19** | 1.75 | 1.71 | 1.83 | — | 18.21** | 0.43 |
| | T/S | 4.21* | 2.79 | 5.33* | 7.87** | 58.08** | 5.46* | 26.06** | 26.38** | 18.21** | — | 13.56** |
| | T/B | 2.34 | 3.31 | 1.70 | 45.75** | 23.16** | 0.56 | 4.14* | 4.38* | 0.43 | 13.56** | — |
| | T/S/B | 0.39 | 0.04 | 0.88 | 22.90** | 43.91** | 1.20 | 16.75** | 17.32** | 7.90** | 2.79 | 9.75** |

Table 4: Statistical significance values using the McNemar test for the YouTube slant dataset. T is Title, S is Source, and B is Body of the post. ** if $p < 0.01$ and * if $p < 0.05$.

| GPT | | | | Llama | | | | Mistral | | | | |
|---------|-------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| | T | T/S | T/B | T/S/B | T | T/S | T/B | T/S/B | T | T/S | T/B | T/S/B |
| GPT | T | — | 34.40** | 20.11** | 449.51** | 96.77** | 32.68** | 38.99** | 224.78** | 51.50** | 2.29 | 18.13** |
| | T/S | 34.40** | — | 6.67** | 435.86** | 200.31** | 132.25** | 140.91** | 218.99** | 136.48** | 19.91** | 82.16** |
| | T/B | 20.11** | 6.67** | — | 440.16** | 163.22** | 81.73** | 103.88** | 233.55** | 104.43** | 3.12 | 62.26** |
| | T/S/B | 259.15** | 373.77** | 318.02** | — | 68.03** | 125.03** | 124.95** | 107.43** | 130.00** | 286.31** | 175.39** |
| Llama | T | 96.77** | 200.31** | 163.22** | 433.77** | — | 21.99** | 21.16** | 236.90** | 18.25** | 138.91** | 40.15** |
| | T/S | 32.68** | 132.25** | 81.73** | 475.86** | 21.99** | — | 0.00 | 274.00** | 0.00 | 85.23** | 5.80* |
| | T/B | 38.99** | 140.91** | 103.88** | 468.49** | 21.16** | 0.00 | — | 271.24** | 0.00 | 81.82** | 8.23** |
| | T/S/B | 252.24** | 383.38** | 327.86** | 107.43** | 79.93** | 148.70** | 144.11** | — | 136.57** | 313.88** | 184.20** |
| Mistral | T | 51.50** | 136.48** | 104.43** | 451.99** | 18.25** | 0.00 | 0.00 | 237.93** | — | 86.05** | 9.88** |
| | T/S | 2.29 | 19.91** | 3.12 | 465.55** | 138.91** | 85.23** | 81.82** | 249.35** | 86.05** | — | 33.00** |
| | T/B | 18.13** | 82.16** | 62.26** | 453.61** | 40.15** | 5.80* | 8.23** | 240.67** | 9.88** | 33.00** | 295.82** |
| | T/S/B | 240.52** | 343.16** | 295.33** | 83.12** | 56.78** | 104.94** | 107.75** | 7.11** | 113.84** | 263.56** | 161.16** |

Table 5: Statistical significance values using the McNemar test for the AdFontes news dataset. T is Title, S is Source, and B is Body of the post. ** if $p < 0.01$ and * if $p < 0.05$.

E Ablations on different parameters

In this section, we report on the various ablations of title, source, description, and number of shots for the various baseline and LLM models.

E.1 MLP Baseline

Table 9 shows the accuracy of the MLP model under different title, source, and description ablations. The multi-layer perceptron model achieved good accuracy when the source was provided. In the AllSides news dataset, wherein we observe that the source is highly correlated with article ideology, the model achieves 69% accuracy when just the source is provided. Similarly, MLP achieves 64% accuracy on the YouTube slant dataset when title, source, and article description are provided. Finally, the model achieves a high accuracy of 62% on the Ad Fontes news dataset when just the title and source are provided.

E.2 RoBERTa baseline

Table 10 shows the accuracy of the RoBERTa model under different title, source, and descrip-

tion ablations. The RoBERTa model performed well on the YouTube slant dataset achieving an accuracy of 65% when title, source, and description were all provided. The performance was worse on the news datasets, with a maximum of 48% on the Ad Fontes news dataset and 58% on the AllSides news dataset.

E.3 GPT-4o

Table 11 shows the accuracy of the GPT-4o model under different ablations of title, source, and description across the three datasets and varying number of demonstrations k considered. We see that the model performs the best when given a higher number of demonstrations for the YouTube slant and Ad Fontes news datasets. For the YouTube dataset, the model performed best when title, source, and description were all provided whereas for the Ad Fontes news dataset, only title and source were sufficient. The AllSides news dataset showed a different trend in that zero demonstrations gave the best score when only the source was provided. We observed in the test set for this dataset that source

| GPT | | | | Llama | | | | Mistral | | | | | |
|---------|-------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| | | T | T/S | T/B | T/S/B | T | T/S | T/B | T/S/B | T | T/S | T/B | T/S/B |
| GPT | T | — | 277.85** | 32.27** | 255.80** | 23.89** | 60.47** | 6.33* | 25.05** | 17.31** | 186.27** | 47.23** | 33.34** |
| | T/S | 277.85** | — | 195.62** | 3.90* | 258.18** | 26.08** | 212.25** | 81.25** | 272.09** | 4.23* | 357.08** | 139.18** |
| | T/B | 32.27** | 195.62** | — | 192.83** | 56.81** | 26.50** | 30.13** | 3.88* | 52.94** | 118.53** | 107.43** | 4.20* |
| | T/S/B | 255.80** | 3.90* | 192.83** | — | 240.60** | 17.80** | 200.62** | 70.24** | 253.02** | 0.56 | 342.79** | 124.91** |
| Llama | T | 23.89** | 258.18** | 56.81** | 240.60** | — | 247.06** | 12.31** | 132.97** | 2.17 | 260.91** | 0.13 | 80.77** |
| | T/S | 60.47** | 26.08** | 26.50** | 17.80** | 247.06** | — | 164.01** | 24.63** | 162.56** | 17.64** | 151.72** | 16.56** |
| | T/B | 6.33* | 212.25** | 30.13** | 200.62** | 12.31** | 164.01** | — | 140.34** | 2.27 | 204.32** | 10.70** | 50.39** |
| | T/S/B | 25.05** | 81.25** | 3.88* | 70.24** | 132.97** | 24.63** | 140.34** | — | 89.82** | 71.34** | 108.30** | 0.17 |
| Mistral | T | 17.31** | 272.09** | 52.94** | 253.02** | 2.17 | 162.56** | 2.27 | 89.82** | — | 292.91** | 4.31* | 87.31** |
| | T/S | 186.27** | 4.23* | 118.53** | 0.56 | 260.91** | 17.64** | 204.32** | 71.34** | 292.91** | — | 339.67** | 138.56** |
| | T/B | 47.23** | 357.08** | 107.43** | 342.79** | 0.13 | 151.72** | 10.70** | 108.30** | 4.31* | 339.67** | — | 175.10** |
| | T/S/B | 33.34** | 139.18** | 4.20* | 124.91** | 80.77** | 16.56** | 50.39** | 0.17 | 87.31** | 138.56** | 175.10** | — |

Table 6: Statistical significance values using the McNemar test for the AllSides news dataset. T is Title, S is Source, and B is Body of the post. ** if $p < 0.01$ and * if $p < 0.05$.

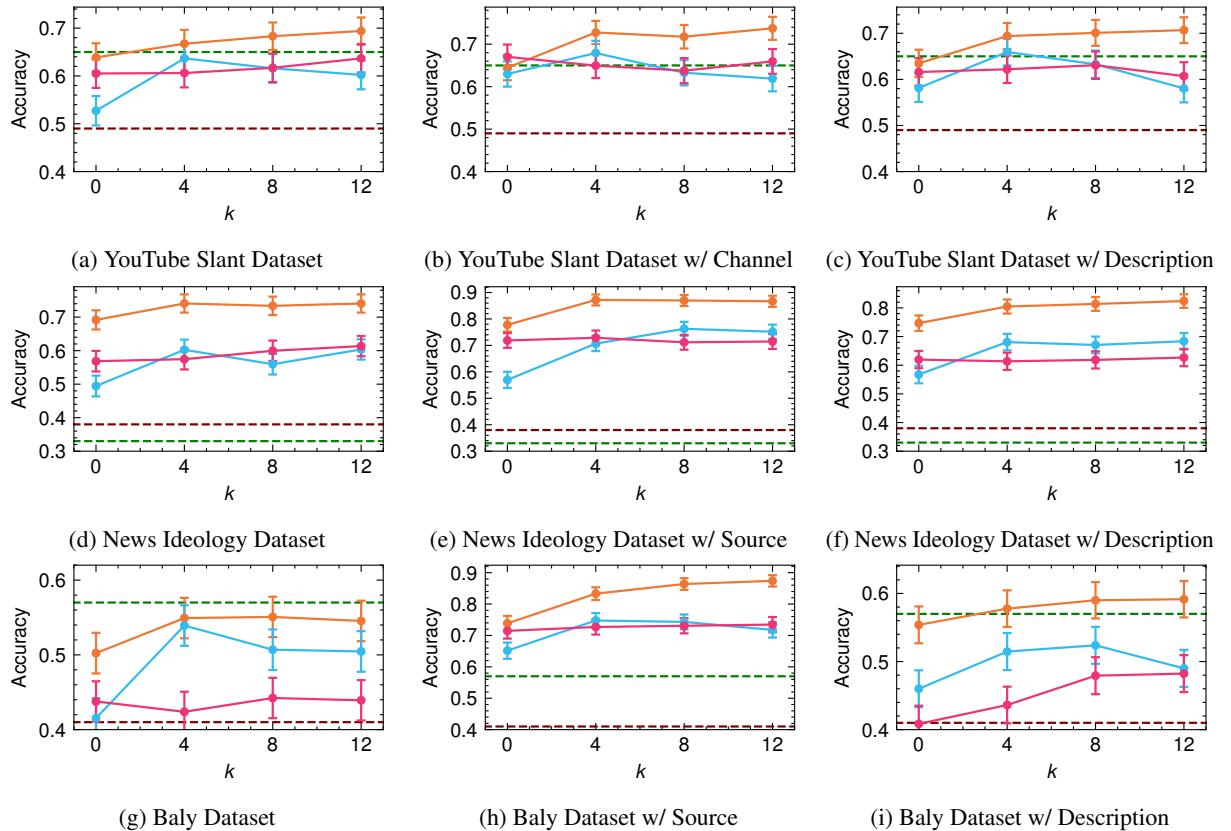


Figure 4: Increases in accuracy across all datasets by increasing the number of ICL shots.

and target label were highly correlated (i.e., a video from Fox, a well-known conservative outlet, was almost always labeled conservative) making it simpler for the model to just assign the source ideology as the target ideology. In Appendix C, we investigate this further by considering edge cases in the test set.

E.4 Llama2

Table 12 shows the accuracy of the Llama2 model under different ablations of title, source, and description across the three datasets and varying num-

ber of demonstrations k considered. We see that the model performs the best at 4 or fewer demonstrations indicating that more demonstrations leads to a drop in performance likely due to a limited context window. For the YouTube and Ad Fontes news datasets, the model performed best when title, source, and only four demonstrations were provided. The AllSides news dataset showed similar trends as GPT-4o in that the source was sufficient without any demonstrations to achieve the best accuracy.

| Liberal Sources | Conservative Sources |
|--------------------|----------------------------|
| ABC News | Breitbart |
| Associated Press | CBN |
| CBC News | Daily Caller |
| CBS News | Daily Mail |
| CNN | Daily Wire |
| Daily Beast | Fox Business |
| HuffPost | Fox News |
| Jacobin | New York Post |
| MSNBC | Newsmax |
| Mother Jones | The American Conservative |
| NPR | The American Spectator |
| Slate | The Blaze |
| The Atlantic | The Daily Caller |
| The Guardian | The Epoch Times |
| The Intercept | The Federalist |
| The Nation | The Federalist Society |
| The New York Times | The Post Millennial |
| The New Yorker | The Washington Free Beacon |
| Vox | Washington Examiner |

Table 7: Liberal and conservative sources considered for the misleading sources experiment.

| Non-liberal articles from liberal sources | | | | | |
|---|--------|---------------|----------------|---------------|--|
| Title | Source | YouTube Slant | Ad Fontes News | AllSides News | |
| ✓ | | 0.63 | 0.58 | 0.49 | |
| | ✓ | 0.47 | 0.38 | 1.00 | |
| ✓ | ✓ | 0.66 | 0.70 | 1.00 | |
| Non-conservative articles from conservative sources | | | | | |
| Title | Source | YouTube Slant | Ad Fontes News | AllSides News | |
| ✓ | | 0.45 | 0.57 | — | |
| | ✓ | 0.00 | 0.00 | — | |
| ✓ | ✓ | 0.00 | 0.52 | — | |

Table 8: Accuracy of GPT-4o on test set where only non-liberal/non-conservative articles were considered from liberal/conservative sources.

| Title | Source | Body | YouTube | Ad Fontes | AllSides |
|-------|--------|------|---------|-----------|----------|
| | | | Slant | News | News |
| ✓ | | | 0.51 | 0.38 | 0.43 |
| | ✓ | | 0.61 | 0.55 | 0.69 |
| | | ✓ | 0.55 | 0.38 | 0.47 |
| ✓ | ✓ | | 0.63 | 0.62 | 0.57 |
| ✓ | | ✓ | 0.57 | 0.43 | 0.46 |
| ✓ | ✓ | ✓ | 0.62 | 0.60 | 0.58 |
| ✓ | ✓ | ✓ | 0.64 | 0.61 | 0.52 |

Table 9: Ablations on the baseline multi-layer perceptron model on the three datasets.

E.5 Mistral

Table 13 shows the accuracy of the Mistral model under different ablations of title, source, and description across the three datasets and varying number of demonstrations k considered. We see that the

| Title | Source | Body | YouTube | Ad Fontes | AllSides |
|-------|--------|------|---------|-----------|----------|
| | | | Slant | News | News |
| ✓ | | | 0.57 | 0.35 | 0.52 |
| | ✓ | | 0.45 | 0.48 | 0.48 |
| | | ✓ | 0.62 | 0.35 | 0.50 |
| ✓ | ✓ | | 0.65 | 0.44 | 0.58 |
| ✓ | | ✓ | 0.64 | 0.38 | 0.53 |
| | ✓ | ✓ | 0.65 | 0.44 | 0.58 |
| ✓ | ✓ | ✓ | 0.65 | 0.44 | 0.55 |

Table 10: Ablations on the baseline RoBERTa model on the three datasets.

| k | Title | Source | Body | YouTube | Ad Fontes | AllSides |
|-----|-------|--------|------|---------|-----------|----------|
| | | | | Slant | News | News |
| 0 | ✓ | | | 0.64 | 0.69 | 0.50 |
| 0 | | ✓ | | 0.63 | 0.80 | 0.88 |
| 0 | | | ✓ | 0.56 | 0.64 | 0.44 |
| 0 | ✓ | ✓ | | 0.64 | 0.78 | 0.74 |
| 0 | ✓ | | ✓ | 0.63 | 0.75 | 0.55 |
| 0 | | ✓ | ✓ | 0.59 | 0.82 | 0.85 |
| 0 | ✓ | ✓ | ✓ | 0.71 | 0.86 | 0.72 |
| 4 | ✓ | | | 0.67 | 0.74 | 0.55 |
| 4 | ✓ | ✓ | | 0.73 | 0.87 | 0.83 |
| 4 | ✓ | | ✓ | 0.69 | 0.80 | 0.58 |
| 4 | ✓ | ✓ | ✓ | 0.72 | 0.86 | 0.82 |
| 8 | ✓ | | | 0.68 | 0.73 | 0.55 |
| 8 | ✓ | ✓ | | 0.72 | 0.87 | 0.86 |
| 8 | ✓ | | ✓ | 0.70 | 0.81 | 0.59 |
| 8 | ✓ | ✓ | ✓ | 0.71 | 0.86 | 0.84 |
| 12 | ✓ | | | 0.69 | 0.74 | 0.55 |
| 12 | ✓ | ✓ | | 0.74 | 0.87 | 0.87 |
| 12 | ✓ | | ✓ | 0.71 | 0.82 | 0.59 |
| 12 | ✓ | ✓ | ✓ | 0.71 | 0.87 | 0.84 |

Table 11: Ablations on the GPT-4o on the three datasets. Here, k is the number of demonstrations.

model performs the best at zero demonstrations indicating that more demonstrations leads to a drop in performance likely due to a limited context window. For the YouTube and Ad Fontes news datasets, the model performed best when title, source, and description were provided with zero demonstrations. The AllSides news dataset again showed similar trends as GPT-4o in that the source was sufficient without any demonstrations to achieve the best accuracy.

F Candidate set size can be smaller

One might argue that the candidate set \mathcal{C} from which demonstrations are selected is similar to the training set used for training supervised models and, thus, a large amount of annotated data is also required for LLMs. While the size of the candidate set was configured high across all three datasets, we further analyze whether the need for annotated

| k | Title | Source | Body | YouTube Slant | Ad Fontes News | AllSides News |
|-----|-------|--------|------|---------------|----------------|---------------|
| 0 | ✓ | | | 0.53 | 0.49 | 0.42 |
| 0 | | ✓ | | 0.61 | 0.52 | 0.82 |
| 0 | | | ✓ | 0.52 | 0.49 | 0.37 |
| 0 | ✓ | ✓ | | 0.63 | 0.57 | 0.65 |
| 0 | ✓ | | ✓ | 0.58 | 0.57 | 0.46 |
| 0 | | ✓ | ✓ | 0.48 | 0.65 | 0.71 |
| 0 | ✓ | ✓ | ✓ | 0.58 | 0.69 | 0.59 |
| 4 | ✓ | | | 0.64 | 0.60 | 0.54 |
| 4 | ✓ | ✓ | | 0.68 | 0.79 | 0.75 |
| 4 | ✓ | | ✓ | 0.66 | 0.68 | 0.51 |
| 4 | ✓ | ✓ | ✓ | 0.64 | 0.79 | 0.68 |
| 8 | ✓ | | | 0.60 | 0.56 | 0.51 |
| 8 | ✓ | ✓ | | 0.63 | 0.76 | 0.74 |
| 8 | ✓ | | ✓ | 0.63 | 0.67 | 0.52 |
| 8 | ✓ | ✓ | ✓ | 0.60 | 0.78 | 0.67 |
| 12 | ✓ | | | 0.60 | 0.60 | 0.50 |
| 12 | ✓ | ✓ | | 0.62 | 0.75 | 0.72 |
| 12 | ✓ | | ✓ | 0.58 | 0.68 | 0.49 |
| 12 | ✓ | ✓ | ✓ | 0.56 | 0.78 | 0.67 |

Table 12: Ablations on the Llama2 on the three datasets. Here, k is the number of demonstrations.

| k | Title | Source | Body | YouTube Slant | Ad Fontes News | AllSides News |
|-----|-------|--------|------|---------------|----------------|---------------|
| 0 | ✓ | | | 0.61 | 0.57 | 0.44 |
| 0 | | ✓ | | 0.61 | 0.72 | 0.83 |
| 0 | | | ✓ | 0.52 | 0.59 | 0.34 |
| 0 | ✓ | ✓ | | 0.67 | 0.72 | 0.71 |
| 0 | ✓ | | ✓ | 0.62 | 0.62 | 0.41 |
| 0 | | ✓ | ✓ | 0.57 | 0.73 | 0.66 |
| 0 | ✓ | ✓ | ✓ | 0.65 | 0.74 | 0.58 |
| 4 | ✓ | | | 0.61 | 0.57 | 0.42 |
| 4 | ✓ | ✓ | | 0.65 | 0.73 | 0.73 |
| 4 | ✓ | | ✓ | 0.62 | 0.61 | 0.44 |
| 4 | ✓ | ✓ | ✓ | 0.63 | 0.68 | 0.60 |
| 8 | ✓ | | | 0.62 | 0.59 | 0.44 |
| 8 | ✓ | ✓ | | 0.64 | 0.71 | 0.73 |
| 8 | ✓ | | ✓ | 0.63 | 0.62 | 0.48 |
| 8 | ✓ | ✓ | ✓ | 0.63 | 0.68 | 0.61 |
| 12 | ✓ | | | 0.64 | 0.61 | 0.44 |
| 12 | ✓ | ✓ | | 0.66 | 0.71 | 0.73 |
| 12 | ✓ | | ✓ | 0.61 | 0.63 | 0.48 |
| 12 | ✓ | ✓ | ✓ | 0.63 | 0.69 | 0.61 |

Table 13: Ablations on the Mistral on the three datasets. Here, k is the number of demonstrations.

| LLM | YouTube Slant | Ad Fontes News | AllSides News |
|---------|---------------|----------------|---------------|
| GPT-4o | 0.66 | 0.74 | 0.52 |
| Llama2 | 0.60 | 0.57 | 0.50 |
| Mistral | 0.63 | 0.59 | 0.42 |

Table 14: Accuracy of the three LLMs across all three datasets with candidate set size $|\mathcal{C}| = 500$ and $k = 8$ demonstrations.

data can be alleviated by considering a smaller candidate set size as well. Specifically, we set the

candidate set size to only 500 samples, a fraction of what is generally required to train supervised models. Testing this across all three LLMs and all three datasets in Table 14, we observe that the accuracy is not majorly effected by this change. This is likely because the SET-BSR approach already selects the most optimal demonstrations that maximize coverage.

G Balanced demonstration selection is better than random

We experiment with two demonstration selection techniques namely 1) **Random** where the demonstrations are randomly selected from the candidate set and 2) **Balanced BERTScore** where the BERTScore-Recall metric is used to select demonstrations in a label-balanced manner (equal liberal, neutral, and conservative demonstrations). Comparing the performance of Mistral with 8 demonstrations, we see in Table 15 that there is an average increase of 7% in accuracy across the datasets when using Balanced BERTScore selection over random.

| Dataset | Random | Balanced BERTScore |
|----------------|--------|--------------------|
| YouTube Slant | 0.52 | 0.61 |
| Ad Fontes News | 0.49 | 0.57 |
| AllSides News | 0.39 | 0.42 |

Table 15: Comparison of random demonstration selection and Balanced BERTScore demonstration selection using Mistral in a 8-shot setting. There is an average increase of 7% in accuracy from random.

H Qualitative results for the political news analysis for YouTube videos

Qualitative results for political news videos, non-political news videos, and political non-news videos that changed predictions as a result of adding information about the source are provided in Tables 16, 17, and 18 respectively.

I Placement of popular news channels based on the slant cut-offs

Figure 5 shows the placement of popular news channels based on the YouTube slant dataset. Corresponding cut-offs at -0.33 and +0.33 are also shown for the liberal and conservative respectively.

| Title | Channel |
|--|--------------------------|
| Liberal to Neutral | |
| How welcome are Russian deserters in Georgia? Focus on Europe | DW News |
| Ukrainians ditch Russian traditions and celebrate Christmas with the West | Sky News |
| Asian Americans Debate Model Minority & Asian Hate VICE Debates | VICE |
| Secret Service under scrutiny ahead January 6 committee's final report | PBS NewsHour |
| White House Correspondent Jeff Mason on his tense exchanges with Trump | Reuters |
| Liberal to Conservative | |
| Democrats introduce bill to block Trump from holding office | Newsmax |
| LIVE: Jan. 6 Committee Votes on Criminal Referrals, Final Report | NTD |
| New CRIMINAL REFERRALS Against Former President Trump Unveiled by Jan. 6th Committee | BlazeTV |
| Tonight's Tipping Points: Hate Crimes, Planned Parenthood, & Trans Military Members! | One America News Network |
| Trolling #AMFEST2022 Attendees as Left-Winger 'Garth Griffin' of Edge News | Project Veritas |
| Neutral to Liberal | |
| Haberman says Trump has been diminished but isn't dead politically | CNN |
| Kevin McCarthy's Path To Speaker Is In Peril | MSNBC |
| Anger grows in Russia over Ukraine's New Year's Day strike on Russian troops | CNN |
| Rep. Adam Kinzinger Delivers Final Speech As Congressman | MSNBC |
| 'Shocked': Russia expert explains why Putin's change of word is significant | CNN |
| Neutral to Conservative | |
| UK Weather: What's behind Britain's big freeze? UK snow chaos explained | Daily Mail |
| Transgender Clinical Psychologist on people de-transitioning Dr. Erica Anderson | Newsmax |
| Actor slams crypto as 'biggest Ponzi scheme ever' during Senate hearing | Fox News |
| Rep. Spartz explains her reasoning for voting 'present' for Speaker | Fox News |
| Internet users allege Bryan Kohberger attended Idaho victims' vigil | Fox News |
| Conservative to Liberal | |
| Trump Supporter: Too Much Diversity Will Kill America | The Young Turks |
| Russian foreign minister: Ukraine must give up occupied territories | CNN |
| These Trump Supporters Are CONVINCED "Sleepy Joe Poopy Pants Isn't The Acting President" | The Young Turks |
| Steve Bannon reacts to Trump's 'major announcement' | CNN |
| Trump adviser tells reporter 'the magic is gone' in new article about re-election bid | CNN |
| Conservative to Neutral | |
| Justin Amash APPLAUDS Revolt Against McCarthy, Says Ruling Oligarchy CRUSHES Dissent | The Hill |
| State Of Emergency Declared At Texas Border | NBC News |
| Watch Trump claim the army retook airports during the Revolutionary War in July Fourth speech | The Globe and Mail |
| Trump Calls Jan. 6 Committee A 'Kangaroo Court' After Criminal Referrals | NBC News |
| 'They shot up the WRONG house!' Sheriff Grady Judd on young kids shot at while in bed in Polk County | WFLA News Channel 8 |

Table 16: Sample of political videos from news channels on YouTube whose predicted ideology changed as a result of adding information about the source.

J Complete set of results for the various LLM configurations

Table 19 shows the final set of results for all configurations.

K Code and Reproducibility

Our code and data used for prompting the models and analyzing the results can be found here: <https://github.com/Haroon96/Ideology-Estimation-LLMs>

| Title | Channel |
|--|---|
| Liberal to Neutral | |
| TRIPLEDEMIC Or Free Market Failure To Blame For Medicine SHORTAGE? STRUGGLE To Find Children's Meds Watch: Jan. 6 Committee Presents Key Video Evidence Compilation | The Hill NBC News |
| Neutral to Liberal | |
| Top U.S. & World Headlines — December 27, 2022 Top U.S. & World Headlines — January 3, 2023 Raskin On Cancer Diagnosis: 'Totally My Plan To Make It Through This Thing' Watch The 11th Hour Highlights: Dec. 20 Plans For Zelenskyy D.C. Trip Came Together In A 'Matter Of Days' | Democracy Now! Democracy Now! MSNBC MSNBC MSNBC |
| Neutral to Conservative | |
| McCarthy speakership battle: What happens if he concedes? The yule log Christmas tradition "Woman" Is Named Word of the Year What is the controversy surrounding Meghan Markle and her sister? 'DOUBLE WHAMMY': Real estate expert issues pricing warning | Fox Business Fox News DailyWire+ Fox News Fox Business |
| Conservative to Liberal | |
| All Gestures, All Shock Opera: DeSantis Launches His Newest Battle Christian Influencer Makes Surprising Confession About Trans Porn The War on Christmas RAGES ON! Alex Jones Says He's Paying His Personal Trainer \$100,000 A Week 'Look at where he is today': Bash on Kinzinger's farewell speech | MSNBC The Young Turks The Young Turks The Young Turks CNN |
| Conservative to Neutral | |
| Cancelling the F-22 Fighter is Worse than You Think | Task & Purpose |

Table 17: Sample of non-political videos from news channels on YouTube whose predicted ideology changed as a result of adding information about the source.

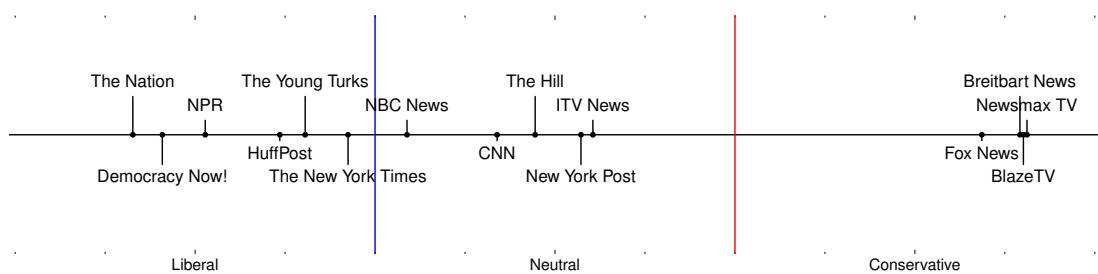


Figure 5: Placement of channels based on the slant cut-offs for the YouTube dataset.

| Title | Channel |
|--|---|
| Liberal to Neutral | |
| Is capitalism actually broken? Update from Ukraine Ukraine is getting ready for the winter Attack Ruzzia will lose it Van Lathan: A Jewish Producer Confronted Kanye at TMZ when Ye Said He Loves Hitler (Part 2) The Trouble with the Electoral College Those Times Sam Harris Made Us Slap Our Knees | TED-Ed Denys Davyдов djvlad CGP Grey Enlightainment |
| Liberal to Conservative | |
| The New Beginning of Same-Sex Marriage Meet Kamala Harris's 6-Year-Old Speechwriter Pro-Choice Bosses - A Generous Offer To Our Pregnant Employees Confronting Charlie Kirk and Tim Pool at Turning Point USA America Fest I Showed Up To An Anti-Gun Protest In Los Angeles Let's Talk | Ben Shapiro The Babylon Bee Live Action Alex Stein James Klug |
| Neutral to Liberal | |
| Meanwhile... Pantone's Color of the Year Macron Makes Condoms Free In France Grandma Arrested For Feeding The Homeless?! Danbury Court Case Clarification + GIVEAWAY! LIVE: EMERGENCY January 6 Committee hearing - Day 6 The Funniest Donald Trump Stories On The Graham Norton Show Part One Prison Officer BUSTED After Massive Nazi Gang Attack Caught | The Late Show with Stephen Colbert Long Island Audit Brian Tyler Cohen The Graham Norton Show The Damage Report |
| Neutral to Conservative | |
| Winning With The Hand You're Dealt with Rick Warren Fighting the Darkest Criminal Enterprise on the Planet STORE WARS: Target, Walmart & Home Depot go to war against Organized Retail theft Warner MAY BE SUED over Amber Heard and Aquaman 2?! NEW Hollywood Lawsuit! Historical alteration on the Ukraine Map: Russian 1st Corps withdrawing en masse from 4 locations! | Saddleback Church The 700 Club News For Reasonable People ThatUmbrellaGuy Divine Justice |
| Conservative to Liberal | |
| Elon Musk Finally CRACKS In Ultimate Outburst Over Free Speech Chris Wallace: This Would Never Have Happened Under Nancy Pelosi Mike Lindell Snaps After Hannity Admits There Was No Fraud Trump ATTACKS the FBI and DOJ and Loses his Mind in NEW POSTS BREAKING: Kari Lake slammed with brutal legal news | The Damage Report The Late Show with Stephen Colbert Tommy Campbell MeidasTouch Brian Tyler Cohen |
| Conservative to Neutral | |
| Why US Let Russia Become the Biggest Nuclear Power The CIA's Biggest Scam Exposed Cop Pulls Over "Legal Expert" And SCHOOLS HIM On The Law Why US Enemies are Scared of Ghosts (Elite Special Forces) "INSPIRATIONAL HOLIDAY VIDEO" — A Bad Lip Reading of Joe Biden | The Infographics Show KiraTV Audit the Audit The Infographics Show Bad Lip Reading |

Table 18: Sample of political videos from non-news channels on YouTube whose predicted ideology changed as a result of adding information about the source.

| Model | Shots | Metadata | | | Dataset | | | |
|------------------------------------|-------|----------|--------|------|-----------------|----------------|---------------|--|
| | | Title | Source | Body | YouTube Dataset | Ad Fontes News | AllSides News | |
| MLP | — | ✓ | | | 0.49 | 0.38 | 0.41 | |
| RoBERTa | — | ✓ | | | 0.65 | 0.33 | 0.57 | |
| Baly et al. (2020) | — | ✓ | | | — | — | 0.46 | |
| GPT-4o | 0 | ✓ | | | 0.64 | 0.69 | 0.50 | |
| Llama2-13B | 0 | ✓ | | | 0.44 | 0.46 | 0.40 | |
| Mistral-7B | 0 | ✓ | | | 0.59 | 0.55 | 0.44 | |
| GPT-4o | 0 | ✓ | ✓ | | 0.52 | 0.78 | 0.74 | |
| Llama2-13B | 0 | ✓ | ✓ | | 0.43 | 0.54 | 0.64 | |
| Mistral-7B | 0 | ✓ | ✓ | | 0.67 | 0.72 | 0.71 | |
| GPT-4o | 0 | ✓ | | ✓ | 0.57 | 0.75 | 0.55 | |
| Llama2-13B | 0 | ✓ | | ✓ | 0.49 | 0.55 | 0.46 | |
| Mistral-7B | 0 | ✓ | | ✓ | 0.59 | 0.62 | 0.41 | |
| GPT-4o | 0 | ✓ | ✓ | ✓ | 0.71 | 0.86 | 0.72 | |
| Llama2-13B | 0 | ✓ | ✓ | ✓ | 0.47 | 0.67 | 0.58 | |
| Mistral-7B | 0 | ✓ | ✓ | ✓ | 0.63 | 0.74 | 0.58 | |
| GPT-4o | 4 | ✓ | | | 0.67 | 0.74 | 0.55 | |
| Llama2-13B | 4 | ✓ | | | 0.64 | 0.60 | 0.54 | |
| Mistral-7B | 4 | ✓ | | | 0.61 | 0.57 | 0.42 | |
| GPT-4o | 4 | ✓ | ✓ | | 0.73 | 0.87 | 0.83 | |
| Llama2-13B | 4 | ✓ | ✓ | | 0.68 | 0.70 | 0.75 | |
| Mistral-7B | 4 | ✓ | ✓ | | 0.65 | 0.73 | 0.73 | |
| GPT-4o | 4 | ✓ | | ✓ | 0.69 | 0.80 | 0.58 | |
| Llama2-13B | 4 | ✓ | | ✓ | 0.63 | 0.68 | 0.51 | |
| Mistral-7B | 4 | ✓ | | ✓ | 0.60 | 0.61 | 0.43 | |
| GPT-4o | 4 | ✓ | ✓ | ✓ | 0.72 | 0.86 | 0.82 | |
| Llama2-13B | 4 | ✓ | ✓ | ✓ | 0.58 | 0.77 | 0.68 | |
| Mistral-7B | 4 | ✓ | ✓ | ✓ | 0.60 | 0.68 | 0.59 | |
| GPT-4o | 8 | ✓ | | | 0.68 | 0.73 | 0.55 | |
| Llama2-13B | 8 | ✓ | | | 0.61 | 0.56 | 0.51 | |
| Mistral-7B | 8 | ✓ | | | 0.62 | 0.60 | 0.44 | |
| GPT-4o | 8 | ✓ | ✓ | | 0.72 | 0.87 | 0.86 | |
| Llama2-13B | 8 | ✓ | ✓ | | 0.62 | 0.76 | 0.74 | |
| Mistral-7B | 8 | ✓ | ✓ | | 0.64 | 0.71 | 0.73 | |
| GPT-4o | 8 | ✓ | | ✓ | 0.70 | 0.81 | 0.59 | |
| Llama2-13B | 8 | ✓ | | ✓ | 0.57 | 0.67 | 0.52 | |
| Mistral-7B | 8 | ✓ | | ✓ | 0.57 | 0.61 | 0.47 | |
| GPT-4o | 8 | ✓ | ✓ | ✓ | 0.71 | 0.86 | 0.84 | |
| Llama2-13B | 8 | ✓ | ✓ | ✓ | 0.47 | 0.78 | 0.67 | |
| Mistral-7B | 8 | ✓ | ✓ | ✓ | 0.54 | 0.67 | 0.59 | |
| GPT-4o | 12 | ✓ | | | 0.69 | 0.74 | 0.55 | |
| Llama2-13B | 12 | ✓ | | | 0.60 | 0.57 | 0.50 | |
| Mistral-7B | 12 | ✓ | | | 0.64 | 0.61 | 0.44 | |
| GPT-4o | 12 | ✓ | ✓ | | 0.74 | 0.87 | 0.87 | |
| Llama2-13B | 12 | ✓ | ✓ | | 0.54 | 0.72 | 0.72 | |
| Mistral-7B | 12 | ✓ | ✓ | | 0.66 | 0.71 | 0.73 | |
| GPT-4o | 12 | ✓ | | ✓ | 0.71 | 0.82 | 0.59 | |
| Llama2-13B | 12 | ✓ | | ✓ | 0.46 | 0.62 | 0.49 | |
| Mistral-7B | 12 | ✓ | | ✓ | 0.53 | 0.61 | 0.46 | |
| GPT-4o | 12 | ✓ | ✓ | ✓ | 0.71 | 0.87 | 0.84 | |
| Llama2-13B | 12 | ✓ | ✓ | ✓ | 0.40 | 0.77 | 0.67 | |
| Mistral-7B | 12 | ✓ | ✓ | ✓ | 0.51 | 0.67 | 0.59 | |

Table 19: Results showing the accuracy of all experiments with GPT-4o, Llama2, Mistral across all three datasets and number of shots. Shown at the top are baseline results including the results from [Baly et al. \(2020\)](#).